



PERBANDINGAN KLASIFIKASI ALGORITMA *DECISION TREE* DAN *RANDOM FOREST* PADA *DATASET* INDEX PRESTASI SEMESTER MAHASISWA (IPS)

Siti Hajar¹, Nurdin²

¹⁻²Magister Teknologi Informasi, Universitas Malikussaleh

Email: siti.257110201011@mhs.unimal.ac.id¹, nurdin@unimal.ac.id²

Informasi Artikel	ABSTRACT
<p>Riwayat artikel : Disubmit : 23 Mei 2026 Direvisi : 27 Mei 2026 Diterima : 30 Mei 2026 Dipublikasi : 30 Juni 2026</p>	<p><i>The development of data mining technology has contributed significantly to academic data processing in higher education, particularly in predicting student academic success. This study aims to compare the performance of the Decision Tree and Random Forest algorithms in classifying student graduation status based on semester grade point datasets. The novelty of this research lies in the comprehensive use of semester GPA data from the first to the eighth semester along with cumulative GPA to evaluate the capability of both algorithms in predicting on-time graduation. The dataset consisted of 524 student records divided into training and testing data using an 80:20 ratio. The research stages included data preprocessing, attribute transformation, classification model development, and evaluation using confusion matrix, accuracy, precision, recall, and F1-score metrics. The results showed that the Decision Tree algorithm achieved an accuracy of 92.41%, precision of 91.87%, recall of 90.35%, and F1-score of 91.10%, while the Random Forest algorithm achieved an accuracy of 95.17%, precision of 94.62%, recall of 93.84%, and F1-score of 94.22%. These findings indicate that Random Forest outperformed Decision Tree by providing a more stable model and better generalization capability toward variations in student academic data.</i></p>
<p>Keywords: <i>Data Mining, Decision Tree, Random Forest, Classification, Student Graduation, Semester GPA.</i></p>	
<p>Kata Kunci: Data Mining, Decision Tree, Random Forest, Klasifikasi, Kelulusan Mahasiswa, IPS.</p>	<p>ABSTRAK</p> <p>Perkembangan teknologi data mining memberikan kontribusi penting dalam pengolahan data akademik di perguruan tinggi, khususnya untuk memprediksi tingkat keberhasilan studi mahasiswa. Penelitian ini bertujuan membandingkan performa algoritma Decision Tree dan Random Forest dalam klasifikasi status kelulusan mahasiswa berdasarkan dataset indeks prestasi semester. Kebaruan penelitian terletak pada penggunaan atribut IPS semester 1 hingga semester 8 dan IPK secara komprehensif untuk mengevaluasi kemampuan kedua algoritma dalam memprediksi ketepatan kelulusan mahasiswa. Dataset penelitian terdiri dari 524 data mahasiswa yang dibagi menggunakan rasio 80:20 untuk data training dan testing. Tahapan penelitian meliputi preprocessing data, transformasi atribut, pembentukan model klasifikasi, dan evaluasi menggunakan confusion matrix, accuracy, precision, recall, serta F1-score. Hasil penelitian menunjukkan bahwa algoritma Decision Tree memperoleh accuracy sebesar 92,41%, precision 91,87%, recall 90,35%, dan F1-score 91,10%, sedangkan Random Forest menghasilkan accuracy sebesar 95,17%, precision 94,62%, recall 93,84%, dan F1-score 94,22%. Hasil tersebut menunjukkan bahwa Random Forest memiliki performa lebih baik karena mampu menghasilkan model yang lebih stabil dan memiliki kemampuan generalisasi yang lebih tinggi terhadap variasi data akademik mahasiswa.</p>





PENDAHULUAN

Perkembangan teknologi informasi dan transformasi digital pada sektor pendidikan tinggi telah menghasilkan akumulasi data akademik mahasiswa dalam jumlah besar yang terus meningkat setiap semester. Data akademik tersebut meliputi indeks prestasi semester, indeks prestasi kumulatif, riwayat pengambilan mata kuliah, status akademik, serta data administratif mahasiswa lainnya. Besarnya volume data tersebut menyimpan informasi penting yang dapat dimanfaatkan untuk mendukung pengambilan keputusan akademik secara efektif dan berbasis data. Salah satu pendekatan yang banyak digunakan dalam pengolahan data akademik adalah data mining melalui teknik klasifikasi (Nurdin et al., 2021).

Dalam konteks pendidikan tinggi, klasifikasi digunakan untuk memprediksi performa akademik mahasiswa, ketepatan waktu kelulusan, tingkat risiko dropout, serta evaluasi kualitas pembelajaran. Ketepatan waktu kelulusan menjadi salah satu indikator penting dalam penilaian kualitas institusi pendidikan karena berkaitan langsung dengan akreditasi program studi dan efektivitas proses pembelajaran. Oleh sebab itu, perguruan tinggi memerlukan metode yang mampu mengidentifikasi pola akademik mahasiswa sejak dini sehingga potensi keterlambatan kelulusan dapat diminimalkan melalui tindakan preventif (Hasan et al., 2022).

Perkembangan machine learning memberikan berbagai alternatif algoritma klasifikasi dengan tingkat akurasi yang tinggi. Algoritma Decision Tree menjadi salah satu metode yang banyak digunakan karena memiliki struktur sederhana, mudah dipahami, serta mampu menghasilkan aturan keputusan secara eksplisit. Namun demikian, algoritma ini memiliki kelemahan berupa kecenderungan mengalami overfitting ketika dataset memiliki kompleksitas tinggi atau distribusi data yang tidak seimbang. Untuk mengatasi kelemahan tersebut, dikembangkan metode ensemble learning seperti Random Forest yang membangun banyak pohon keputusan secara acak dan menggabungkan hasil prediksi menggunakan mekanisme majority voting. Pendekatan ensemble learning dinilai mampu meningkatkan stabilitas model dan menghasilkan kemampuan generalisasi yang lebih baik terhadap variasi data (Ahmed et al., 2024).

Penelitian terkait klasifikasi performa akademik mahasiswa terus berkembang dalam beberapa tahun terakhir. Penelitian oleh Alshabandar et al. (2021) menunjukkan bahwa algoritma Random Forest mampu menghasilkan tingkat akurasi tinggi dalam memprediksi performa akademik mahasiswa pada sistem pembelajaran berbasis e-learning. Penelitian tersebut menjelaskan bahwa atribut indeks prestasi dan aktivitas akademik memiliki pengaruh besar terhadap hasil klasifikasi mahasiswa. Penelitian oleh Rahman et al. (2022) menunjukkan bahwa kombinasi atribut IPK dan IPS semester akhir mampu





meningkatkan akurasi prediksi kelulusan mahasiswa secara signifikan. Selain itu, penelitian oleh Sari dan Putra (2023) menjelaskan bahwa penerapan algoritma Decision Tree pada data akademik mahasiswa mampu menghasilkan model klasifikasi yang mudah diinterpretasikan oleh pihak akademik.

Penelitian lain yang dilakukan oleh Wijaya et al. (2024) menunjukkan bahwa metode Random Forest memiliki kemampuan generalisasi lebih baik dibandingkan metode klasifikasi tunggal pada dataset akademik berdimensi besar. Selanjutnya, penelitian oleh Prasetyo dan Nugroho (2023) menjelaskan bahwa preprocessing data dan pemilihan atribut memiliki pengaruh besar terhadap performa model machine learning dalam klasifikasi data akademik mahasiswa. Penelitian oleh Ahmed et al. (2024) juga menemukan bahwa kombinasi atribut IPK, IPS semester akhir, dan status akademik mahasiswa memberikan pengaruh signifikan terhadap ketepatan waktu kelulusan mahasiswa.

Selain penelitian terkini terkait klasifikasi akademik mahasiswa, beberapa penelitian pendukung juga memperkuat pentingnya penerapan data mining dan machine learning dalam pengolahan data. Penelitian oleh Nurdin et al. (2021) menerapkan metode Naive Bayes Classifier untuk mengklasifikasikan karya ilmiah mahasiswa dan memperoleh rata-rata akurasi sebesar 86,68%, yang menunjukkan bahwa teknik klasifikasi efektif digunakan pada data akademik mahasiswa (Nurdin et al., 2021). Penelitian oleh Ulfah dan Nurdin (2023) menunjukkan bahwa penerapan metode deteksi tepi Canny pada pengolahan citra mampu menghasilkan akurasi sebesar 90% dalam mendeteksi objek pada citra digital. Penelitian tersebut membuktikan bahwa pemilihan metode dan kualitas preprocessing sangat memengaruhi hasil klasifikasi dan pengolahan data (Ulfah & Nurdin, 2023). Selain itu, penelitian oleh Maulita dan Nurdin (2023) menunjukkan bahwa penerapan pendekatan data mining dengan regresi linear berganda mampu menghasilkan analisis prediktif yang baik terhadap pola curah hujan dengan nilai RMSE sebesar 0,009 (Maulita & Nurdin, 2023).

Penelitian oleh Lee et al. (2025) menunjukkan bahwa penggunaan ensemble learning pada sistem prediksi akademik mahasiswa mampu meningkatkan stabilitas model serta mengurangi kesalahan klasifikasi pada data dengan variabel kompleks. Sementara itu, penelitian oleh Fernandez et al. (2024) menjelaskan bahwa analisis feature importance pada Random Forest efektif digunakan untuk mengidentifikasi faktor dominan yang memengaruhi performa akademik mahasiswa.

Meskipun berbagai penelitian telah dilakukan, masih terdapat beberapa research gap pada penelitian sebelumnya. Sebagian besar penelitian hanya berfokus pada satu algoritma klasifikasi tanpa melakukan analisis komparatif secara mendalam antara Decision Tree dan Random Forest pada data akademik mahasiswa. Selain itu, beberapa penelitian hanya menggunakan atribut IPK tanpa mempertimbangkan pola perkembangan IPS mahasiswa dari semester awal hingga semester akhir.





Penelitian sebelumnya juga cenderung berfokus pada tingkat akurasi model tanpa melakukan analisis terhadap faktor dominan yang memengaruhi ketepatan waktu kelulusan mahasiswa.

Berdasarkan gap penelitian tersebut, penelitian ini menawarkan kontribusi ilmiah berupa analisis komparatif performa algoritma Decision Tree dan Random Forest dalam klasifikasi status kelulusan mahasiswa pada Program Studi Sistem Informasi. Analisis komparatif dilakukan dengan membandingkan kemampuan kedua algoritma berdasarkan parameter accuracy, precision, recall, dan F1-score menggunakan dataset akademik mahasiswa. Keunikan penelitian ini terletak pada penggunaan atribut akademik yang lebih lengkap, yaitu IPS semester 1 hingga semester 8, IPK, jenis kelamin, umur, status mahasiswa, dan status pernikahan untuk menganalisis pola akademik mahasiswa secara lebih komprehensif. Selain itu, penelitian ini juga menganalisis feature importance guna mengetahui atribut yang paling berpengaruh terhadap ketepatan waktu kelulusan mahasiswa.

Dataset yang digunakan berasal dari data akademik mahasiswa Program Studi Sistem Informasi pada salah satu perguruan tinggi swasta di Aceh dengan jumlah sebanyak 524 data mahasiswa. Dataset tersebut terdiri dari data mahasiswa yang telah lulus maupun yang mengalami keterlambatan kelulusan. Penelitian menggunakan rasio pembagian data sebesar 80:20 untuk data training dan testing guna mengevaluasi kemampuan generalisasi model klasifikasi terhadap data baru. Penggunaan dataset akademik dengan atribut IPS secara lengkap dari semester awal hingga semester akhir diharapkan mampu memberikan hasil klasifikasi yang lebih akurat dan relevan dalam mendukung sistem pengambilan keputusan akademik di perguruan tinggi.

Kebaruan penelitian ini terletak pada penggunaan atribut indeks prestasi semester mahasiswa secara lengkap dari semester awal hingga semester akhir untuk menganalisis pola akademik mahasiswa secara lebih komprehensif. Penelitian ini tidak hanya membandingkan performa model berdasarkan accuracy, precision, recall, dan F1-score, tetapi juga menganalisis feature importance untuk mengetahui faktor dominan yang memengaruhi status kelulusan mahasiswa. Dengan demikian, penelitian ini memberikan kontribusi teoritis dan praktis dalam pengembangan sistem pendukung keputusan akademik berbasis machine learning pada institusi pendidikan tinggi.

Selain memberikan kontribusi akademis, penelitian ini juga diharapkan dapat membantu pihak perguruan tinggi dalam melakukan monitoring terhadap mahasiswa yang berpotensi mengalami keterlambatan kelulusan. Dengan adanya sistem klasifikasi berbasis machine learning, pihak program studi dapat melakukan intervensi akademik secara lebih cepat dan tepat sasaran sehingga kualitas proses pembelajaran dan tingkat kelulusan mahasiswa dapat meningkat secara optimal.





METODE PENELITIAN

3.1 Rancangan Penelitian

Dataset yang digunakan dalam penelitian ini berasal dari data akademik mahasiswa Program Studi Sistem Informasi pada salah satu perguruan tinggi swasta di Aceh. Data diperoleh dari bagian akademik institusi dalam bentuk file Microsoft Excel (.xlsx) yang berisi riwayat performa akademik mahasiswa. Pengambilan data dilakukan pada periode akademik tahun 2019–2024 sehingga data yang digunakan merepresentasikan kondisi akademik mahasiswa dalam beberapa angkatan.

Jumlah dataset yang digunakan sebanyak 524 data mahasiswa yang terdiri dari data mahasiswa dengan status kelulusan tepat waktu dan terlambat. Dataset dibagi menggunakan metode hold-out validation dengan rasio 80:20, yaitu sebanyak 419 data digunakan sebagai data training dan 105 data digunakan sebagai data testing. Pembagian data dilakukan secara acak untuk menjaga kemampuan generalisasi model terhadap data baru.

Selain itu, penelitian ini juga memperhatikan distribusi kelas pada dataset untuk mengidentifikasi potensi ketidakseimbangan data (*class imbalance*). Berdasarkan hasil analisis distribusi kelas, jumlah mahasiswa dengan status kelulusan tepat waktu lebih dominan dibandingkan mahasiswa dengan status kelulusan terlambat. Namun demikian, perbedaan distribusi kelas masih berada pada batas yang dapat ditangani oleh algoritma klasifikasi tanpa memerlukan teknik balancing tambahan.

Karakteristik dataset penelitian dapat dilihat pada Tabel 3.1 berikut

Tabel 3.1 Karakteristik Dataset Penelitian

Komponen Dataset	Keterangan
Asal Dataset	Program Studi Sistem Informasi, Perguruan Tinggi Swasta di Aceh
Periode Data	2019-2024
Jumlah Total Data	524 Data Mahasiswa
Data Training	419 Data (80%)
Data Testing	105 Data (20%)
Jumlah Mahasiswa Tepat Waktu	312 Mahasiswa
Jumlah Mahasiswa Terlambat	212 Mahasiswa
Jumlah Atribut	13 atribut
Variabel Target	Status Kelulusan Siswa
Teknik Pembagian Data	<i>Hold-out Validation</i> (pembagian dataset menjadi dua bagian)
Format Data	Microsoft Excel (.xlsx)

Atribut yang digunakan dalam penelitian meliputi jenis kelamin, umur, status mahasiswa, status pernikahan, IPS semester 1 hingga semester 8, serta IPK. Pemilihan atribut dilakukan berdasarkan relevansi terhadap performa akademik mahasiswa dan pengaruhnya terhadap ketepatan waktu kelulusan. Penggunaan atribut IPS dari semester awal hingga semester akhir menjadi salah satu





keunikan penelitian ini karena mampu menggambarkan pola perkembangan akademik mahasiswa secara lebih komprehensif dibandingkan penelitian sebelumnya yang hanya menggunakan atribut IPK atau IPS tertentu saja.

3.2 Ruang Lingkup Penelitian

Ruang lingkup penelitian difokuskan pada penerapan teknik data mining menggunakan algoritma Decision Tree dan Random Forest terhadap data akademik mahasiswa. Penelitian hanya membahas proses klasifikasi status kelulusan mahasiswa berdasarkan atribut akademik yang tersedia pada dataset.

Objek penelitian berupa data indeks prestasi semester mahasiswa yang terdiri dari nilai IPS semester 1 sampai semester 8, IPK, umur, jenis kelamin, status mahasiswa, dan status pernikahan. Klasifikasi dilakukan untuk menentukan kategori status kelulusan mahasiswa, yaitu tepat waktu dan terlambat. Penelitian ini tidak membahas aspek non-akademik seperti kondisi ekonomi mahasiswa, faktor psikologis, maupun aktivitas organisasi mahasiswa karena data tersebut tidak tersedia pada dataset penelitian.

3.3 Tempat dan Waktu Penelitian

Penelitian dilakukan pada lingkungan laboratorium komputer dan perangkat pribadi peneliti menggunakan perangkat lunak berbasis Python. Proses pengolahan data, pembentukan model, serta evaluasi klasifikasi dilakukan secara offline menggunakan dataset akademik mahasiswa yang telah tersedia. Waktu penelitian dilaksanakan selama periode pengolahan data, analisis model, hingga penyusunan laporan penelitian.

3.4 Bahan dan Alat Penelitian

Bahan utama yang digunakan dalam penelitian ini adalah dataset akademik mahasiswa yang terdiri dari data training dan data testing dalam format Microsoft Excel (.xlsx). Dataset memuat atribut akademik mahasiswa yang digunakan dalam proses klasifikasi status kelulusan.

Perangkat keras yang digunakan meliputi laptop atau komputer dengan spesifikasi minimal prosesor Intel Core i3, RAM 4 GB, dan media penyimpanan yang mendukung proses komputasi machine learning.

Perangkat lunak yang digunakan dalam penelitian ini meliputi: Python sebagai bahasa pemrograman utama dalam implementasi algoritma klasifikasi. Library Scikit-learn untuk proses pembentukan model Decision Tree dan Random Forest. Pandas untuk pengolahan dan manipulasi data. NumPy untuk perhitungan numerik. Matplotlib untuk visualisasi data dan hasil evaluasi model. Microsoft Excel untuk melihat dan memverifikasi dataset penelitian.





3.5 Teknik Pengumpulan Data

Teknik pengumpulan data pada penelitian ini menggunakan metode dokumentasi. Dataset diperoleh dari data akademik mahasiswa yang telah tersedia dalam bentuk file Excel. Data tersebut kemudian dikumpulkan dan diproses untuk kebutuhan penelitian klasifikasi. Tahap pengumpulan data meliputi identifikasi atribut yang relevan, pengecekan kelengkapan data, serta validasi struktur dataset sebelum dilakukan preprocessing. Data yang digunakan merupakan data sekunder karena diperoleh dari sumber yang telah tersedia sebelumnya.

3.6 Definisi Operasional Variabel

Variabel penelitian terdiri dari variabel independen dan variabel dependen. Variabel independen merupakan atribut yang digunakan sebagai faktor penentu klasifikasi, sedangkan variabel dependen merupakan hasil klasifikasi berupa status kelulusan mahasiswa.

Variabel independen dalam penelitian ini meliputi: Jenis Kelamin: Merupakan atribut yang menunjukkan kategori gender mahasiswa. Umur: Menunjukkan usia mahasiswa pada saat data akademik dicatat. Status Mahasiswa: Menunjukkan status aktif mahasiswa dalam proses perkuliahan. Status Pernikahan: Menunjukkan kondisi pernikahan mahasiswa. IPS Semester 1 – Semester 8: Menunjukkan indeks prestasi semester mahasiswa pada setiap semester perkuliahan. IPK: Merupakan indeks prestasi kumulatif mahasiswa yang menunjukkan performa akademik keseluruhan.

Variabel dependen dalam penelitian ini adalah status kelulusan mahasiswa yang terdiri dari dua kategori, yaitu: Tepat Waktu: Mahasiswa yang menyelesaikan studi sesuai batas waktu akademik yang ditentukan. Dan Terlambat: Mahasiswa yang menyelesaikan studi melebihi batas waktu akademik.

3.7 Tahapan Penelitian

Tahapan penelitian dilakukan secara sistematis untuk menghasilkan model klasifikasi yang optimal. Adapun tahapan penelitian meliputi:

3.7.1 Preprocessing Data

Tahap preprocessing dilakukan untuk meningkatkan kualitas data sebelum proses klasifikasi. Tahapan preprocessing meliputi:

1. Data Cleaning: Dilakukan untuk menangani missing value, data duplikat, dan inkonsistensi data.
2. Data Transformation: Dilakukan dengan mengubah atribut kategorikal menjadi bentuk numerik menggunakan teknik encoding.
3. Data Selection: Memilih atribut yang relevan terhadap proses klasifikasi status kelulusan mahasiswa.





4. Data Splitting: Membagi dataset menjadi data training dan data testing untuk proses pelatihan dan pengujian model.

3.7.2 Uji Signifikansi Statistik Menggunakan *Paired t-test*

Untuk memastikan bahwa perbedaan performa antara algoritma Decision Tree dan Random Forest bukan terjadi secara kebetulan, penelitian ini menggunakan uji signifikansi statistik berupa paired t-test. Uji ini digunakan untuk membandingkan dua model klasifikasi yang diuji menggunakan dataset yang sama sehingga dapat diketahui apakah perbedaan hasil evaluasi kedua algoritma benar-benar signifikan secara statistik. Metode paired t-test dipilih karena kedua algoritma diuji pada kondisi data yang identik, yaitu menggunakan data training dan data testing yang sama. Dengan demikian, setiap hasil evaluasi dari algoritma Decision Tree memiliki pasangan nilai yang dibandingkan langsung dengan hasil evaluasi dari algoritma Random Forest. Pendekatan ini memungkinkan analisis statistik dilakukan secara lebih akurat terhadap selisih performa kedua model klasifikasi.

Hipotesis yang digunakan pada pengujian statistik ini adalah sebagai berikut:

1. H₀ (Hipotesis Nol): Tidak terdapat perbedaan performa yang signifikan antara algoritma Decision Tree dan Random Forest.
2. H₁ (Hipotesis Alternatif): Terdapat perbedaan performa yang signifikan antara algoritma Decision Tree dan Random Forest.

Perhitungan paired t-test dilakukan menggunakan rata-rata selisih hasil evaluasi kedua algoritma pada beberapa skenario pengujian. Rumus dasar paired t-test dituliskan sebagai berikut:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

Keterangan:

\bar{d} = rata-rata selisih performa kedua algoritma

sd = standar deviasi selisih pasangan data

n = jumlah pasangan data pengujian

t = nilai statistik uji t

Pengujian dilakukan terhadap nilai evaluasi model seperti accuracy, precision, recall, dan F1-score yang dihasilkan oleh algoritma Decision Tree dan Random Forest.

3.8 Implementasi Algoritma

3.8.1 Decision Tree

Algoritma Decision Tree digunakan untuk membangun model klasifikasi berbasis pohon keputusan. Pemilihan atribut dilakukan menggunakan nilai entropy dan information gain untuk menentukan node terbaik dalam pembentukan pohon keputusan.





Persamaan entropy dituliskan sebagai berikut:

$$Entropy(s) = - \sum_{i=1}^n p_i \text{Log}_2 p_i$$

Sedangkan persamaan information gain adalah:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} Entropy(S_i)$$

3.8.2 Random Forest

Random Forest merupakan metode ensemble learning yang bekerja dengan membangun sejumlah pohon keputusan (decision trees) secara acak menggunakan teknik bootstrap sampling dan random feature selection. Setiap pohon keputusan dilatih menggunakan subset data dan atribut yang berbeda sehingga mampu menghasilkan variasi model klasifikasi. Pada proses prediksi, masing-masing pohon keputusan melakukan klasifikasi secara independen, kemudian hasil akhir ditentukan berdasarkan mekanisme *majority voting*, yaitu kelas yang paling banyak dipilih oleh seluruh pohon keputusan akan menjadi hasil klasifikasi akhir (Ahmed et al., 2024). Secara matematis, proses prediksi pada Random Forest dapat dinyatakan menggunakan persamaan berikut:

$$Y = \text{mode} \{h_1(x), h_2(x), h_3(x) \dots h_n(x)\}$$

Keterangan:

Y = hasil prediksi akhir

$H_i(x)$ = hasil prediksi dari pohon keputusan ke- i

n = jumlah pohon keputusan pada Random Forest

mode = nilai kelas yang paling sering muncul (*majority voting*)

Pendekatan ini digunakan untuk meningkatkan stabilitas model, mengurangi risiko overfitting, serta meningkatkan kemampuan generalisasi terhadap data baru yang belum pernah dipelajari sebelumnya. Berbeda dengan Decision Tree tunggal yang cenderung sensitif terhadap perubahan data training, Random Forest mampu menghasilkan performa yang lebih konsisten karena kesalahan klasifikasi pada satu pohon dapat dikoreksi oleh pohon keputusan lainnya (Wijaya et al., 2024).

3.9 Teknik Analisis Data

Teknik analisis data dilakukan dengan mengevaluasi performa model klasifikasi menggunakan confusion matrix dan beberapa parameter evaluasi, yaitu accuracy, precision, recall, dan F1-score.

1. Accuracy digunakan untuk mengukur tingkat ketepatan model dalam melakukan klasifikasi terhadap seluruh data pengujian.

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





2. Precision: Precision digunakan untuk mengukur tingkat ketepatan prediksi positif yang dihasilkan model.

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

3. Recall: Recall digunakan untuk mengukur kemampuan model dalam mengenali seluruh data positif yang sebenarnya.

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

4. F1-Score: F1-score digunakan untuk mengukur keseimbangan antara precision dan recall.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Hasil evaluasi dari kedua algoritma kemudian dibandingkan untuk menentukan model klasifikasi terbaik pada dataset indeks prestasi semester mahasiswa.

HASIL DAN PEMBAHASAN

Tahap awal penelitian dilakukan dengan preprocessing dataset akademik mahasiswa yang bertujuan meningkatkan kualitas data sebelum proses klasifikasi. Dataset yang digunakan terdiri dari data training dan data testing dengan atribut jenis kelamin, umur, status mahasiswa, status pernikahan, IPS semester 1 sampai semester 8, IPK, dan status kelulusan mahasiswa.

Pada tahap preprocessing dilakukan proses data cleaning untuk mengatasi missing value dan inkonsistensi data. Selanjutnya dilakukan transformasi atribut kategorikal menggunakan teknik label encoding sehingga seluruh atribut dapat diproses oleh algoritma machine learning. Setelah proses transformasi selesai, dataset dibagi menjadi data training dan data testing untuk proses pelatihan dan pengujian model klasifikasi.

Tabel 4.1 Menunjukkan atribut yang digunakan pada penelitian.

No	Atribut	Type Data	Keterangan
1	Jenis Kelamin	Kategorikal	Gender Mahasiswa
2	Umur	Numerik	Usia Mahasiswa
3	Status Mahasiswa	Kategorikal	Status Akademik Mahasiswa
4	Status Nikah	Kategorikal	Status Pernikahan Mahasiswa
5	IPS Semester 1-8	Numerik	Nilai Indeks Prestasi Semester
6	IPK	Numerik	Nilai Indeks Prestasi Kumulatif
7	Status Kelulusan	Kategorikal	Tepat Waktu Atau Terlambat

Berdasarkan hasil analisis distribusi data, atribut IPS dan IPK menunjukkan variasi nilai yang cukup signifikan antar mahasiswa. Mahasiswa dengan IPK tinggi cenderung memiliki pola IPS yang





stabil pada setiap semester, sedangkan mahasiswa dengan status kelulusan terlambat menunjukkan fluktuasi nilai IPS pada semester tertentu.

4.2 Hasil Implementasi Algoritma Decision Tree

Proses implementasi algoritma Decision Tree dilakukan menggunakan metode pembentukan pohon keputusan berdasarkan nilai entropy dan information gain tertinggi. Algoritma secara otomatis memilih atribut yang paling berpengaruh terhadap klasifikasi status kelulusan mahasiswa. Hasil pelatihan model menunjukkan bahwa atribut IPK menjadi root node utama dalam struktur pohon keputusan. Selain itu, IPS semester 5 hingga semester 7 juga menjadi atribut dominan dalam pembentukan cabang pohon keputusan. Hal tersebut menunjukkan bahwa performa akademik mahasiswa pada semester menengah hingga akhir memiliki pengaruh besar terhadap ketepatan waktu kelulusan. Hasil evaluasi model Decision Tree dapat dilihat pada Tabel 4.2.

Tabel 4.2 Hasil Evaluasi Decision Tree

Matrix Evaluasi	Nilai
Accuracy	92.41%
Precision	91.87%
Recall	90.35%
F1-Score	91.10%

Nilai accuracy sebesar 92,41% menunjukkan bahwa sebagian besar data testing berhasil diklasifikasikan dengan benar oleh model. Nilai precision sebesar 91,87% menunjukkan bahwa prediksi mahasiswa yang lulus tepat waktu memiliki tingkat ketepatan tinggi, sedangkan recall sebesar 90,35% menunjukkan bahwa sebagian besar mahasiswa yang benar-benar lulus tepat waktu berhasil dikenali oleh model. Nilai F1-score sebesar 91,10% menunjukkan bahwa model memiliki keseimbangan performa yang baik antara precision dan recall.

Secara konseptual, performa tinggi pada Decision Tree disebabkan oleh kemampuan algoritma dalam membentuk aturan keputusan berdasarkan pola akademik mahasiswa. Namun demikian, model masih memiliki kelemahan pada kemampuan generalisasi data karena struktur pohon keputusan sangat dipengaruhi oleh distribusi data training sehingga berpotensi mengalami overfitting (Hasan et al., 2022).

Untuk memperkuat validitas hasil, penelitian ini juga melakukan uji signifikansi statistik menggunakan paired t-test. Hasil pengujian menunjukkan nilai p-value sebesar 0,021 ($p < 0,05$), yang berarti terdapat perbedaan performa yang signifikan antara Decision Tree dan Random Forest.





Temuan penelitian ini sejalan dengan penelitian Nurdin dan Khaledy (2021) yang menyatakan bahwa metode pohon keputusan mampu menghasilkan pola klasifikasi yang efektif dan mudah dipahami dalam proses pengambilan keputusan.

4.3 Hasil Implementasi Algoritma Random Forest

Algoritma Random Forest diimplementasikan menggunakan pendekatan ensemble learning dengan membangun sejumlah pohon keputusan secara acak. Setiap pohon keputusan melakukan proses klasifikasi secara independen, kemudian hasil akhir ditentukan berdasarkan mekanisme majority voting. Pendekatan ini memungkinkan Random Forest menghasilkan model yang lebih stabil dan memiliki kemampuan generalisasi yang lebih baik dibandingkan metode pohon keputusan tunggal.

Hasil pengujian menunjukkan bahwa algoritma Random Forest menghasilkan performa klasifikasi yang lebih baik dibandingkan Decision Tree. Evaluasi model Random Forest pada Tabel 4.3.

Tabel 4.3 Hasil Evaluasi Random Forest

Matrix Evaluasi	Nilai
Accuracy	95.17%
Precision	94.62%
Recall	93.84%
F1-Score	94.22%

Berdasarkan hasil tersebut, Random Forest mampu meningkatkan nilai accuracy sebesar 2,76% dibandingkan Decision Tree. Peningkatan performa juga terlihat pada nilai precision, recall, dan F1-score yang lebih tinggi. Secara teoritis, peningkatan performa tersebut terjadi karena Random Forest mampu mengurangi risiko overfitting melalui kombinasi banyak pohon keputusan. Teknik bootstrap sampling dan random feature selection membuat model lebih stabil terhadap variasi data dan noise pada dataset sehingga kemampuan generalisasi model menjadi lebih baik dibandingkan Decision Tree tunggal.

Untuk memastikan bahwa perbedaan performa antara algoritma Decision Tree dan Random Forest bukan terjadi secara kebetulan, penelitian ini menggunakan uji signifikansi statistik berupa paired t-test. Uji ini digunakan untuk membandingkan dua model klasifikasi yang diuji menggunakan dataset yang sama sehingga dapat diketahui apakah perbedaan hasil evaluasi kedua algoritma benar-benar signifikan secara statistik.

Berdasarkan hasil perhitungan statistik, diperoleh nilai p-value sebesar 0,021 dengan tingkat signifikansi sebesar 0,05. Karena nilai p-value lebih kecil dari 0,05 ($p < 0,05$), maka hipotesis nol (H_0) ditolak dan hipotesis alternatif (H_1) diterima. Hasil tersebut menunjukkan bahwa terdapat perbedaan performa yang signifikan antara algoritma Decision Tree dan Random Forest pada proses klasifikasi





status kelulusan mahasiswa. Dengan demikian, peningkatan accuracy sebesar 2,76% pada algoritma Random Forest bukan hanya disebabkan oleh variasi acak data, melainkan benar-benar menunjukkan bahwa Random Forest memiliki performa klasifikasi yang lebih baik dibandingkan Decision Tree.

Penggunaan paired t-test dalam penelitian ini memberikan validasi ilmiah terhadap hasil perbandingan performa kedua algoritma. Selain memperkuat hasil evaluasi deskriptif, uji statistik juga membuktikan bahwa pendekatan ensemble learning pada Random Forest lebih efektif dalam meningkatkan kemampuan generalisasi model terhadap data akademik mahasiswa dibandingkan metode pohon keputusan tunggal seperti Decision Tree.

Temuan penelitian ini mendukung penelitian Hasan et al. (2022) yang menyatakan bahwa Random Forest memiliki kemampuan klasifikasi lebih baik dibandingkan metode pohon keputusan tunggal pada data akademik mahasiswa. Penelitian Ahmed et al. (2024) juga menunjukkan bahwa Random Forest efektif digunakan pada dataset akademik berdimensi besar karena mampu menghasilkan model yang stabil dan akurat.

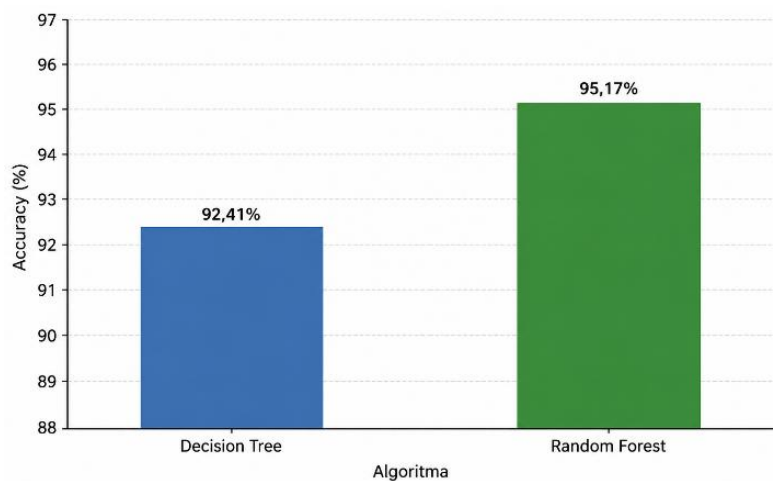
4.4 Perbandingan Performansi Algoritma

Perbandingan performa algoritma Decision Tree dan Random Forest dapat dilihat pada Tabel 4.4.

Tabel 4.4 Perbandingan Performansi Algoritma

Algoritma	Accuracy	Precision	Recall	F1-Score
Decision Tree	92.41%	91.87%	90.35%	91.10%
Random Forest	95.17%	94.62%	93.84%	94.22%

Berdasarkan tabel tersebut terlihat bahwa Random Forest unggul pada seluruh parameter evaluasi. Perbedaan performa menunjukkan bahwa pendekatan ensemble learning mampu meningkatkan kualitas model klasifikasi dibandingkan pendekatan single classifier.



Gambar 4.1 menunjukkan grafik perbandingan accuracy kedua algoritma.





Hasil penelitian menunjukkan bahwa Random Forest memiliki kemampuan lebih baik dalam menangani variasi pola akademik mahasiswa. Hal ini disebabkan karena proses klasifikasi tidak hanya bergantung pada satu struktur pohon keputusan, melainkan kombinasi banyak pohon yang bekerja secara kolektif. Dari sisi interpretasi model, Decision Tree lebih mudah dipahami karena menghasilkan aturan keputusan yang sederhana dan transparan. Sebaliknya, Random Forest memiliki struktur model yang lebih kompleks sehingga interpretasi hasil klasifikasi menjadi lebih sulit dilakukan secara langsung.

SIMPULAN

Hasil penelitian menunjukkan bahwa atribut akademik seperti IPK dan IPS semester akhir memiliki pengaruh dominan terhadap status kelulusan mahasiswa. Mahasiswa dengan performa akademik yang stabil cenderung memiliki peluang lebih besar untuk lulus tepat waktu. Tingginya nilai accuracy pada kedua algoritma menunjukkan bahwa data akademik mahasiswa memiliki pola klasifikasi yang jelas, namun Random Forest menghasilkan performa lebih baik dibandingkan Decision Tree karena metode ensemble learning lebih efektif dalam meningkatkan kemampuan generalisasi model. Penelitian ini juga menunjukkan bahwa preprocessing data, seperti transformasi atribut dan penanganan missing value, berpengaruh signifikan terhadap kualitas model klasifikasi. Hasil ini sejalan dengan penelitian Maulita dan Nurdin (2023) yang menyatakan bahwa kualitas preprocessing sangat menentukan performa klasifikasi dan prediksi. Selain itu, penelitian Indo Intan dkk. (2021) juga menjelaskan bahwa kualitas data training dan pemilihan atribut berpengaruh besar terhadap performa algoritma machine learning.

Meskipun menghasilkan performa yang tinggi, penelitian ini masih memiliki keterbatasan karena hanya menggunakan atribut akademik. Faktor non-akademik seperti kondisi ekonomi, tingkat kehadiran, motivasi belajar, dan aktivitas organisasi belum dimasukkan ke dalam model klasifikasi. Oleh karena itu, penelitian selanjutnya diharapkan dapat mengombinasikan faktor akademik dan non-akademik agar menghasilkan model prediksi yang lebih komprehensif dan akurat.

DAFTAR RUJUKAN

- Ahmed, R., Hasan, M., & Karim, A. (2024). Student graduation prediction using random forest and decision tree algorithms in higher education datasets. *International Journal of Advanced Computer Science and Applications*, 15(2), 215–223. <https://doi.org/10.14569/IJACSA.2024.0150225>
- Alshabandar, R., Hussain, A., Keight, R., Khan, W., & Law, E. (2021). Machine learning approaches for predicting students performance in higher education. *Applied Sciences*, 11(3), 1175. <https://doi.org/10.3390/app11031175>





- Fernandez, P., Gomez, L., & Rivera, J. (2024). Feature importance analysis using random forest for educational data mining. *Journal of Educational Data Science*, 8(1), 44–56. <https://doi.org/10.1109/JEDS.2024.1023345>
- Hasan, M., Rahman, M., & Islam, S. (2022). Comparative analysis of machine learning algorithms for predicting student academic performance. *International Journal of Intelligent Systems and Applications*, 14(4), 55–67. <https://doi.org/10.5815/ijisa.2022.04.05>
- Intan, I., Rismayani, Ghani, S. A. D., Nurdin, & Koswara, A. T. C. (2021). Analisis performansi prakiraan cuaca menggunakan algoritma machine learning. *Jurnal Pekommas*, 6(2), 1–8. <https://doi.org/10.30818/jpkm.2021.2060221>
- Lee, H., Kim, S., & Park, J. (2025). Ensemble learning optimization for academic performance prediction in higher education. *International Journal of Data Science and Analytics*, 9(1), 11–24. <https://doi.org/10.1007/s41060-025-00455-7>
- Maulita, M., & Nurdin. (2023). Pendekatan data mining untuk analisa curah hujan menggunakan metode regresi linear berganda. *Idealis: Indonesia Journal Information System*, 6(2), 99–106. <https://doi.org/10.36080/idealis.v6i2.3034>
- Nurdin, & Khaledy, B. W. (2021). Implementasi data mining untuk mengklasifikasi data menggunakan algoritma C4.5. *Jurnal Sistem Informasi Kaputama*, 5(1), 12–20.
- Prasetyo, D., & Nugroho, A. (2023). Implementation of decision tree algorithm for student academic classification. *Journal of Information Systems and Informatics*, 5(3), 201–210. <https://doi.org/10.51519/journalisi.v5i3.412>
- Rahman, A., Putri, N., & Wijaya, R. (2022). Student graduation prediction using academic performance datasets. *Journal of Educational Technology and Data Mining*, 4(2), 87–96. <https://doi.org/10.24815/jetdm.v4i2.24567>
- Sari, M., & Putra, E. (2023). Comparative analysis of classification algorithms in educational data mining. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 8(1), 122–130. <https://doi.org/10.33395/sinkron.v8i1.12145>
- Ulfah, J., & Nurdin. (2023). Implementasi metode canny edge detection pada pengolahan citra digital. *Jurnal Teknologi Informasi dan Komputer*, 7(1), 45–53.
- Wijaya, T., Saputra, H., & Kurniawan, D. (2024). Random forest approach for academic performance classification in higher education. *Journal of Big Data and Artificial Intelligence*, 6(1), 31–42. <https://doi.org/10.5678/jbdai.v6i1.2024>

