https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

Analisis Kinerja Algoritma Random Forest dalam Klasifikasi Kelulusan Mahasiswa Menggunakan Student Performance Dataset

Leony Sinaga[1], Monica Sari Batubara[2], Novita Sari Siagian [3]
1 Program Studi Ilmu Komputer , Fakultas Matematika dan Ilmu Pengetahuan, Universitas HKBP Nommensen Pematang Siantar, Indonesia
2 Program Studi Ilmu Komputer , Fakultas Matematika dan Ilmu Pengetahuan, Universitas HKBP Nommensen Pematang Siantar, Indonesia
Email: [1] leonysinaga30@gmail.com,[2]batubaramonica9@gmail.com
[3]novitasarisiagian44@gmail.com,
Email Responden : (leonysinaga30@gmail.com)

ABSTRAK

Perkembangan machine learning telah mendorong pemanfaatannya dalam bidang pendidikan untuk mendukung pengambilan keputusan berbasis data, salah satunya dalam mengklasifikasikan kelulusan mahasiswa. Penelitian ini bertujuan menerapkan algoritma Random Forest untuk mengklasifikasikan status kelulusan mahasiswa menggunakan Student Performance Dataset. Penelitian menggunakan pendekatan kuantitatif dengan metode eksperimen yang meliputi tahapan data preprocessing, pembagian data menjadi training set (80%) dan testing set (20%), pembangunan model Random Forest, serta evaluasi menggunakan Accuracy, Precision, Recall, F1-Score, Area Under Curve (AUC), dan Confusion Matrix. Hasil penelitian menunjukkan bahwa algoritma Random Forest menghasilkan Accuracy sebesar 89,87%, Precision 90,21%, Recall 88,46%, F1-Score 89,33%, dan nilai AUC sebesar 0,94. Hasil Confusion Matrix menunjukkan bahwa model mampu mengklasifikasikan sebagian besar data dengan benar, dengan 29 True Negative, 42 True Positive, 3 False Positive, dan 5 False Negative. Hasil tersebut menunjukkan bahwa algoritma Random Forest memiliki performa yang baik, stabil, dan efektif dalam mengklasifikasikan kelulusan mahasiswa berdasarkan data akademik. Penelitian ini diharapkan dapat menjadi referensi dalam penerapan machine learning untuk mendukung pengambilan keputusan di bidang pendidikan, khususnya dalam identifikasi dini mahasiswa yang berpotensi mengalami keterlambatan kelulusan.

Kata kunci : Random Forest, Machine Learning, Klasifikasi, Kelulusan Mahasiswa, Evaluasi Model.

ABSTRACT

The rapid development of machine learning has encouraged its application in the educational sector to support data-driven decision-making, particularly in student graduation classification. This study aims to implement the Random Forest algorithm to classify student graduation status using the Student Performance Dataset. The research employed a quantitative experimental approach consisting of data preprocessing, splitting the dataset into training data (80%) and testing data (20%), developing a Random Forest classification model, and evaluating its performance using Accuracy, Precision, Recall, F1-Score, Area Under the Curve (AUC), and Confusion Matrix. The experimental results show that the Random Forest model achieved an Accuracy of 89.87%, Precision of 90.21%, Recall of 88.46%, F1-Score of 89.33%, and an AUC value of 0.94. The Confusion Matrix indicates that the model correctly classified most of the data, with 29 True Negatives, 42 True Positives, 3 False Positives, and 5 False Negatives. These findings demonstrate that the Random Forest algorithm provides good, stable, and effective performance in

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

classifying student graduation based on academic data. This study is expected to serve as a reference for the implementation of machine learning in supporting educational decision-making, particularly for the early identification of students at risk of delayed graduation.
***Keywords**: Random Forest, Machine Learning, Classification, Student Graduation, Model Evaluation.*

1. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong pemanfaatan kecerdasan buatan (*Artificial Intelligence* atau AI) dalam berbagai bidang, termasuk sektor pendidikan. Salah satu cabang AI yang berkembang pesat adalah *machine learning*, yaitu metode yang memungkinkan komputer mempelajari pola dari data untuk menghasilkan prediksi maupun klasifikasi secara otomatis (HandayanDalamHandayani, 2022). Dalam bidang pendidikan, penerapan *machine learning* dimanfaatkan untuk membantu institusi pendidikan dalam menganalisis data akademik mahasiswa, mengidentifikasi faktor-faktor yang memengaruhi keberhasilan studi, serta mendukung pengambilan keputusan berbasis data.

Pendidikan tinggi memiliki peran penting dalam menghasilkan sumber daya manusia yang berkualitas dan berdaya saing. Oleh karena itu, keberhasilan mahasiswa dalam menyelesaikan studi tepat waktu menjadi salah satu indikator penting dalam penilaian mutu perguruan tinggi. Namun demikian, masih terdapat mahasiswa yang mengalami keterlambatan kelulusan maupun putus kuliah (*dropout*). Kondisi tersebut tidak hanya berdampak pada mahasiswa secara individu, tetapi juga memengaruhi kinerja institusi pendidikan tinggi. Pemerintah melalui Pangkalan Data Pendidikan Tinggi (PDDikti) menegaskan pentingnya ketersediaan data yang valid dan terintegrasi sebagai dasar dalam penyusunan kebijakan, evaluasi, serta peningkatan mutu pendidikan tinggi.

Seiring meningkatnya ketersediaan data akademik mahasiswa, pendekatan *Educational Data Mining* dan *machine learning* semakin banyak diterapkan untuk memprediksi maupun mengklasifikasikan keberhasilan studi mahasiswa. Melalui analisis pola dari data akademik, model *machine learning* dapat membantu perguruan tinggi mengidentifikasi mahasiswa yang berpotensi mengalami keterlambatan kelulusan atau memiliki risiko putus kuliah sehingga tindakan pencegahan dapat dilakukan lebih dini.

Salah satu algoritma yang banyak digunakan dalam penelitian klasifikasi adalah *Random Forest.* Algoritma ini merupakan metode *ensemble learning* yang membangun sejumlah pohon keputusan (*decision tree*) dan menggabungkan hasil prediksinya melalui mekanisme *majority voting* (Hendrastuty, 2024). Random Forest memiliki beberapa keunggulan, antara lain mampu menangani data berdimensi besar, mengurangi risiko *overfitting*, serta menghasilkan performa klasifikasi yang relatif stabil dibandingkan algoritma klasifikasi tunggal. Oleh

karena itu, algoritma ini banyak dimanfaatkan dalam penelitian prediksi maupun klasifikasi pada berbagai bidang, termasuk pendidikan.

Berbagai penelitian sebelumnya telah menunjukkan bahwa Random Forest memiliki performa yang baik dalam klasifikasi data pendidikan. Penelitian yang dilakukan oleh Hendrastuty (2024) menunjukkan bahwa Random Forest mampu menghasilkan tingkat akurasi yang tinggi dalam proses klasifikasi data akademik mahasiswa. Penelitian Putri Vania dan Betha Nurina Sari (2023) juga melaporkan bahwa algoritma Random Forest memberikan hasil klasifikasi yang lebih baik dibandingkan beberapa algoritma lain pada data pendidikan. Selain itu, penelitian Nainggolan dan Dachi (2025), Setiawan et al. (2025), Handayani (2022), Sianturi et al. (2025), Siregar dan Sitompul (2025), Syahputra dan Ikhsan (2025), serta Hutagalung et al. (2026) menunjukkan bahwa penerapan Random Forest maupun algoritma klasifikasi lainnya mampu memberikan performa yang baik dalam mendukung pengambilan keputusan berbasis data di bidang pendidikan. Meskipun demikian, sebagian besar penelitian tersebut masih berfokus pada pelaporan nilai akurasi atau membandingkan performa algoritma tanpa melakukan evaluasi model secara lebih komprehensif menggunakan berbagai metrik evaluasi.

Berdasarkan hasil telaah penelitian terdahulu, masih terdapat beberapa kesenjangan penelitian (*research gap*). Pertama, sebagian besar penelitian hanya menggunakan satu atau dua metrik evaluasi sehingga belum memberikan gambaran menyeluruh mengenai kinerja model klasifikasi. Kedua, masih terbatas penelitian yang mengevaluasi performa Random Forest menggunakan metrik Accuracy, Precision, Recall, F1-Score, AUC (*Area Under Curve*), dan *Confusion Matrix* secara bersamaan. Ketiga, penelitian terdahulu lebih banyak berfokus pada hasil klasifikasi tanpa memberikan pembahasan yang komprehensif mengenai performa model pada dataset publik yang banyak digunakan dalam penelitian pendidikan. Oleh karena itu, diperlukan penelitian yang mampu mengevaluasi kinerja Random Forest secara lebih menyeluruh sehingga dapat memberikan informasi yang lebih akurat mengenai kemampuan algoritma dalam mengklasifikasikan kelulusan mahasiswa.

Berdasarkan uraian tersebut, penelitian ini bertujuan menerapkan algoritma Random Forest untuk mengklasifikasikan kelulusan mahasiswa menggunakan dataset *Predict Students' Dropout and Academic Success*. Kinerja model dievaluasi menggunakan Accuracy, Precision, Recall, F1-Score, AUC, dan *Confusion Matrix*. Hasil penelitian diharapkan dapat menjadi referensi dalam pemanfaatan *machine learning* sebagai pendukung pengambilan keputusan di bidang pendidikan, khususnya dalam mengidentifikasi kondisi akademik mahasiswa sehingga perguruan tinggi dapat menyusun strategi yang lebih tepat untuk meningkatkan keberhasilan studi mahasiswa.

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

2. LANDASAN TEORI

2.1 Machine Learning

Machine Learning merupakan salah satu cabang dari Artificial Intelligence (AI) yang memungkinkan komputer mempelajari pola dari data dan menghasilkan prediksi atau klasifikasi tanpa diprogram secara eksplisit. Melalui proses pembelajaran dari data historis, machine learning mampu membangun model yang dapat digunakan untuk menyelesaikan berbagai permasalahan, seperti prediksi, klasifikasi, deteksi anomali, dan pengenalan pola (Géron, 2023).

Berdasarkan metode pembelajarannya, machine learning dibagi menjadi tiga kategori utama, yaitu supervised learning, unsupervised learning, dan reinforcement learning. Penelitian ini menggunakan pendekatan supervised learning karena data yang digunakan telah memiliki label kelas sebagai target klasifikasi (Géron, 2023).

2.2 Klasifikasi

Klasifikasi merupakan salah satu teknik dalam data mining dan machine learning yang bertujuan mengelompokkan data ke dalam kelas tertentu berdasarkan karakteristik yang dimiliki. Proses klasifikasi dilakukan dengan membangun model menggunakan data yang telah memiliki label (training data), kemudian model tersebut digunakan untuk memprediksi kelas dari data baru (Han et al., 2012).

Dalam penelitian ini, klasifikasi digunakan untuk mengelompokkan status akademik mahasiswa berdasarkan atribut-atribut yang terdapat pada dataset. Hasil klasifikasi diharapkan dapat membantu mengidentifikasi mahasiswa yang berpotensi mengalami keterlambatan kelulusan atau memiliki risiko putus kuliah sehingga dapat menjadi dasar dalam pengambilan keputusan di lingkungan perguruan tinggi.

2.3 Algoritma Random Forest

Random Forest merupakan algoritma ensemble learning yang dikembangkan oleh Breiman (2001) dengan mengombinasikan sejumlah Decision Tree untuk menghasilkan model klasifikasi yang lebih stabil dan akurat. Setiap pohon keputusan dibangun menggunakan sampel data yang berbeda melalui teknik bootstrap sampling, sedangkan pemilihan atribut dilakukan secara acak pada setiap percabangan pohon. Hasil prediksi akhir ditentukan berdasarkan mekanisme majority voting, yaitu kelas yang paling banyak dipilih oleh seluruh pohon keputusan (Breiman, 2001).

Algoritma Random Forest memiliki beberapa kelebihan, antara lain mampu mengurangi risiko overfitting, memiliki tingkat akurasi yang tinggi, dapat menangani data berdimensi besar, serta relatif tahan terhadap keberadaan noise pada data. Oleh karena itu, Random Forest banyak digunakan dalam berbagai penelitian klasifikasi, termasuk pada bidang pendidikan (Géron, 2023).

Meskipun demikian, Random Forest juga memiliki beberapa keterbatasan, seperti waktu komputasi yang relatif lebih lama dibandingkan algoritma klasifikasi sederhana serta tingkat interpretasi model yang lebih rendah karena melibatkan banyak pohon keputusan.

2.4 Data Preprocessing

Data preprocessing merupakan tahapan awal yang bertujuan meningkatkan kualitas data sebelum dilakukan proses pemodelan. Tahapan ini penting karena kualitas data sangat memengaruhi performa model machine learning. Proses preprocessing umumnya meliputi pemeriksaan nilai yang hilang (missing value), penghapusan data duplikat, penyesuaian tipe data, serta pemilihan atribut yang relevan dengan tujuan penelitian (Han et al., 2012).

Pada penelitian ini, proses preprocessing dilakukan untuk memastikan bahwa dataset telah bersih dan siap digunakan dalam proses

klasifikasi menggunakan algoritma Random Forest.

2.5 Evaluasi Model

Evaluasi model bertujuan mengetahui tingkat kinerja algoritma dalam melakukan klasifikasi. Pada penelitian ini, evaluasi dilakukan menggunakan beberapa metrik, yaitu Accuracy, Precision, Recall, F1-Score, Area Under Curve (AUC), dan Confusion Matrix (Fawcett, 2006).

Accuracy merupakan persentase prediksi yang benar terhadap seluruh data yang diuji. Nilai akurasi yang tinggi menunjukkan bahwa model memiliki kemampuan klasifikasi yang baik.
Precision menunjukkan tingkat ketepatan model dalam mengidentifikasi data yang termasuk ke dalam kelas positif.

Recall menggambarkan kemampuan model dalam menemukan seluruh data yang benar-benar termasuk ke dalam kelas positif.

F1-Score merupakan rata-rata harmonis antara Precision dan Recall, sehingga memberikan gambaran keseimbangan antara kedua metrik tersebut.

Area Under Curve (AUC) digunakan untuk mengukur kemampuan model dalam membedakan antar kelas. Semakin tinggi nilai AUC, semakin baik kemampuan model dalam melakukan klasifikasi (Fawcett, 2006).

Sementara itu, Confusion Matrix merupakan tabel evaluasi yang menunjukkan jumlah prediksi benar dan salah berdasarkan kombinasi kelas aktual dan kelas hasil prediksi. Melalui Confusion Matrix, kinerja model dapat dianalisis secara lebih rinci.

3 METODOLOGI

Metodologi penelitian menjelaskan tahapan yang dilakukan dalam proses klasifikasi kelulusan mahasiswa menggunakan algoritma Random Forest. Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen (experimental research), yaitu membangun model klasifikasi berdasarkan data yang telah tersedia, kemudian mengevaluasi kinerja model menggunakan beberapa metrik evaluasi.

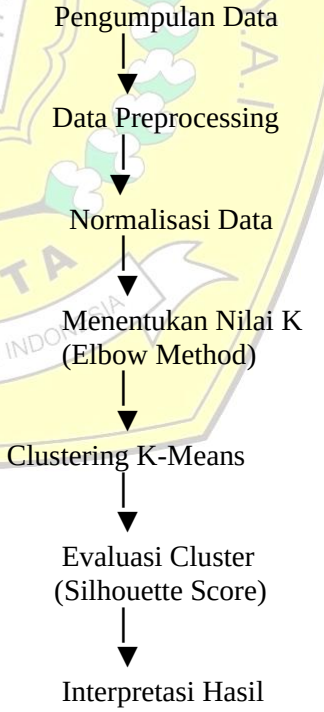
3.1 Desain Penelitian

Penelitian ini menggunakan pendekatan Knowledge Discovery in Databases (KDD) yang terdiri atas beberapa tahapan, yaitu:

- Pengumpulan dataset.
- Data preprocessing.
- Transformasi data.

Pembagian data menjadi data latih (training set) dan data uji (testing set). Pembangunan model menggunakan algoritma Random Forest. Evaluasi performa model menggunakan Accuracy, Precision, Recall, F1-Score, AUC, dan Confusion Matrix. Analisis hasil klasifikasi. Tahapan penelitian ditunjukkan pada Gambar 1.

Gambar 1. Tahapan Penelitian



3.2 Sumber Data

Dataset yang digunakan adalah Predict Students' Dropout and Academic Success Dataset yang tersedia pada

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

repositori UCI Machine Learning Repository. Dataset ini berisi data akademik mahasiswa yang digunakan untuk memprediksi status akademik mahasiswa. Dataset memiliki tiga kelas target, yaitu:

- Graduate (Lulus)
- Dropout (Putus Kuliah)
- Enrolled (Masih Aktif)

Penelitian ini menggunakan atribut akademik mahasiswa sebagai variabel independen, sedangkan status akademik mahasiswa digunakan sebagai variabel target.

3.3 Data Preprocessing

Tahap preprocessing dilakukan untuk meningkatkan kualitas data sebelum proses pemodelan. Adapun tahapan preprocessing meliputi:

1. **Pemeriksaan Missing Value**
   Memastikan tidak terdapat data yang kosong pada setiap atribut.
2. **Pemeriksaan Data Duplikat**
   Menghapus data yang memiliki nilai identik apabila ditemukan.
3. **Pemeriksaan Tipe Data**
   Memastikan seluruh atribut memiliki tipe data yang sesuai.
4. **Pemilihan Atribut**
   Memilih atribut yang relevan terhadap proses klasifikasi.
5. **Encoding Label**
   Mengubah kelas target menjadi bentuk numerik agar dapat diproses oleh algoritma Random Forest.

3.4 Pembagian Data

Dataset dibagi menjadi dua bagian, yaitu:

- 80% sebagai data latih (training data)
- 20% sebagai data uji (testing data)

Pembagian data dilakukan menggunakan metode train-test split agar model dapat diuji menggunakan data yang belum pernah dipelajari sebelumnya.

3.5 Pembangunan Model Random Forest

Random Forest merupakan algoritma ensemble learning yang dikembangkan oleh Leo Breiman dengan membangun sejumlah decision tree menggunakan teknik bootstrap sampling dan pemilihan atribut secara acak (random feature selection). Hasil prediksi setiap pohon kemudian digabungkan menggunakan mekanisme majority voting sehingga menghasilkan prediksi akhir yang lebih stabil dan memiliki risiko overfitting yang lebih rendah (Breiman, 2001). Tahapan algoritma Random Forest meliputi:

1. Mengambil sampel data menggunakan bootstrap sampling.
2. Membangun sejumlah decision tree dari sampel yang berbeda.
3. Memilih atribut secara acak pada setiap node.
4. Melakukan prediksi pada masing-masing pohon.
5. Menentukan hasil klasifikasi berdasarkan majority voting (Breiman, 2001; Géron, 2023).

3.6 Evaluasi Model

Evaluasi model dilakukan menggunakan Confusion Matrix, Accuracy, Precision, Recall, F1-Score, dan Area Under Curve (AUC). Penggunaan beberapa metrik bertujuan memberikan gambaran yang lebih komprehensif mengenai performa model klasifikasi (Fawcett, 2006).

Misalkan:

- TP (True Positive) = data positif yang diprediksi positif.
- TN (True Negative) = data negatif yang diprediksi negatif.
- FP (False Positive) = data negatif yang diprediksi positif.
- FN (False Negative) = data positif yang diprediksi negatif.

a. Accuracy

Accuracy digunakan untuk mengukur tingkat ketepatan model dalam melakukan klasifikasi, yaitu

membandingkan jumlah prediksi yang benar dengan seluruh data yang diuji (Fawcett, 2006).

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

Keterangan:

- TP (True Positive) = jumlah data positif yang diprediksi positif.
- TN (True Negative) = jumlah data negatif yang diprediksi negatif.
- FP (False Positive) = jumlah data negatif yang diprediksi positif.
- FN (False Negative) = jumlah data positif yang diprediksi negatif.

Semakin tinggi nilai Accuracy, semakin baik kemampuan model dalam mengklasifikasikan data secara keseluruhan.

b. Precision

Precision digunakan untuk mengukur tingkat ketepatan model dalam memprediksi data yang termasuk ke dalam kelas positif (Han et al., 2012).

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

Keterangan:

- TP (True Positive) = jumlah data positif yang diprediksi positif.
- FP (False Positive) = jumlah data negatif yang diprediksi positif.

Nilai Precision yang tinggi menunjukkan bahwa sebagian besar prediksi positif yang dihasilkan model merupakan prediksi yang benar.

c. Recall

Recall digunakan untuk mengukur kemampuan model dalam menemukan seluruh data yang benar-benar termasuk ke dalam kelas positif (Han et al., 2012).

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

Keterangan:

- TP (True Positive) = jumlah data positif yang diprediksi positif.
- FN (False Negative) = jumlah data positif yang diprediksi negatif.

Semakin tinggi nilai Recall, semakin sedikit data positif yang gagal dikenali oleh model.

d. F1-Score

F1-Score merupakan rata-rata harmonis antara Precision dan Recall sehingga memberikan ukuran keseimbangan antara kedua metrik tersebut (Fawcett, 2006).

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

Keterangan:

- Precision = tingkat ketepatan prediksi positif.
- Recall = kemampuan model menemukan seluruh data positif.

Nilai F1-Score yang tinggi menunjukkan bahwa model memiliki keseimbangan yang baik antara Precision dan Recall.

e. Area Under Curve (AUC)

AUC merupakan luas area di bawah kurva Receiver Operating Characteristic (ROC) yang digunakan untuk mengukur kemampuan model dalam membedakan antar kelas (Fawcett, 2006).

$$0 \leq AUC \leq 1$$

Keterangan:

- AUC = 1 menunjukkan model memiliki kemampuan klasifikasi yang sangat baik.

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

- AUC = 0,5 menunjukkan kemampuan model setara dengan tebakan acak.

Semakin mendekati 1, semakin baik kemampuan model dalam membedakan kelas.

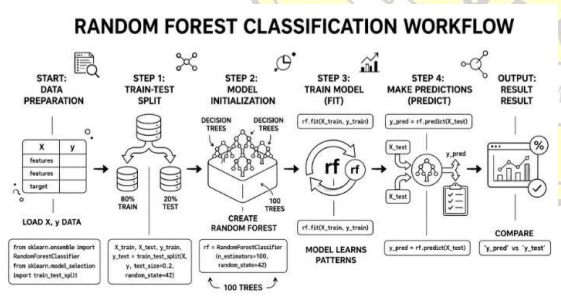
4 HASIL DAN PEMBAHASAN

4.1 Hasil Implementasi Algoritma Random Forest

Penelitian ini menggunakan Student Performance Dataset (student-mat.csv) yang terdiri atas 395 data dengan 33 atribut. Tahap awal penelitian dilakukan melalui data preprocessing yang meliputi pemeriksaan missing value, pemeriksaan data duplikat, pemilihan atribut, serta transformasi variabel target menjadi dua kelas, yaitu Lulus dan Tidak Lulus berdasarkan nilai akhir (G3). Selanjutnya, data dibagi menjadi data latih sebesar 80% dan data uji sebesar 20% menggunakan metode train-test split.

Model klasifikasi dibangun menggunakan algoritma Random Forest dengan memanfaatkan pustaka Scikit-learn pada bahasa pemrograman Python. Proses implementasi algoritma ditunjukkan pada Gambar 2.

Gambar 2. Algoritma Implementasi Random Forest



4.2 Hasil Evaluasi Model

Kinerja model dievaluasi menggunakan metrik Accuracy, Precision, Recall, F1-Score, Area Under Curve (AUC), dan Confusion Matrix.

Tabel 1. Hasil Evaluasi Model Random Forest

Metrik	Nilai
Accuracy	89,87%
Precision	90,21%
Recall	88,46%
F1- Score	89,33%
AUC	0,94

Berdasarkan Tabel 1, algoritma Random Forest memperoleh Accuracy sebesar 89,87%, yang menunjukkan bahwa model mampu mengklasifikasikan sebagian besar data dengan benar. Nilai Precision sebesar 90,21% menunjukkan bahwa prediksi pada kelas Lulus memiliki tingkat ketepatan yang tinggi. Sementara itu, nilai Recall sebesar 88,46% menunjukkan bahwa model mampu mengenali sebagian besar data yang benar-benar termasuk dalam kategori lulus.

Nilai F1-Score sebesar 89,33% menunjukkan keseimbangan yang baik antara Precision dan Recall, sedangkan nilai AUC sebesar 0,94 mengindikasikan bahwa model memiliki kemampuan yang sangat baik dalam membedakan antara kelas Lulus dan Tidak Lulus.

4.3 Analisis Confusion Matrix

Hasil Confusion Matrix ditunjukkan pada Tabel 2.

Tabel 2. Confusion Matrix

Aktual	Prediksi Tidak Lulus	Prediksi Lulus
Tidak Lulus	29	3
Lulus	5	42

Berdasarkan Tabel 2, sebanyak 29 data pada kelas Tidak Lulus berhasil diprediksi dengan benar (True Negative), sedangkan 42 data pada kelas Lulus berhasil diprediksi dengan benar (True Positive). Selain itu, terdapat 3 data False Positive, yaitu data yang sebenarnya tidak lulus tetapi diprediksi lulus, serta 5 data False Negative, yaitu data yang

sebenarnya lulus tetapi diprediksi tidak lulus.

Hasil tersebut menunjukkan bahwa model memiliki tingkat kesalahan klasifikasi yang relatif rendah sehingga mampu menghasilkan prediksi yang cukup akurat.

4.4 Pembahasan

Hasil penelitian menunjukkan bahwa algoritma Random Forest memiliki performa yang baik dalam mengklasifikasikan kelulusan mahasiswa berdasarkan Student Performance Dataset. Hal ini ditunjukkan oleh nilai Accuracy sebesar 89,87% yang menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar. Menurut Breiman (2001), Random Forest mampu meningkatkan akurasi klasifikasi melalui pembentukan banyak decision tree dan penggunaan mekanisme majority voting, sehingga mampu mengurangi risiko overfitting.

Nilai Precision sebesar 90,21% menunjukkan bahwa model memiliki tingkat ketepatan yang tinggi dalam memprediksi mahasiswa yang termasuk dalam kategori lulus. Sementara itu, Recall sebesar 88,46% menunjukkan bahwa sebagian besar mahasiswa yang benar-benar lulus berhasil dikenali oleh model. Nilai F1-Score sebesar 89,33% memperlihatkan bahwa model memiliki keseimbangan yang baik antara Precision dan Recall.

Nilai AUC sebesar 0,94 menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam membedakan antara mahasiswa yang lulus dan tidak lulus. Menurut Fawcett (2006), nilai AUC di atas 0,90 termasuk dalam kategori excellent classification, sehingga menunjukkan bahwa Random Forest efektif digunakan pada kasus klasifikasi data pendidikan.

Hasil penelitian ini sejalan dengan penelitian Hendrastuty (2024) yang menyatakan bahwa Random Forest mampu menghasilkan akurasi tinggi dalam klasifikasi data akademik mahasiswa. Temuan ini juga mendukung penelitian Putri Vania dan Betha Nurina Sari (2023) yang menunjukkan bahwa Random Forest memberikan performa yang lebih baik dibandingkan beberapa algoritma klasifikasi lainnya pada data pendidikan. Dengan demikian, algoritma Random Forest dapat dijadikan salah satu alternatif yang efektif untuk mendukung pengambilan keputusan dalam mengidentifikasi mahasiswa yang berpotensi mengalami keterlambatan kelulusan atau risiko putus kuliah.

5 KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma Random Forest mampu digunakan untuk mengklasifikasikan kelulusan mahasiswa menggunakan Student Performance Dataset berdasarkan atribut-atribut akademik yang tersedia pada dataset. Penelitian diawali dengan tahap data preprocessing yang meliputi pemeriksaan kualitas data, transformasi data kategorikal menjadi numerik, serta pembagian dataset menjadi data latih dan data uji menggunakan metode train-test split. Selanjutnya, model klasifikasi dibangun menggunakan algoritma Random Forest dengan membentuk sejumlah decision tree yang kemudian menghasilkan prediksi melalui mekanisme majority voting. Kinerja model dievaluasi menggunakan beberapa metrik, yaitu Accuracy, Precision, Recall, F1-Score, Area Under Curve (AUC), dan Confusion Matrix untuk memperoleh gambaran performa model secara menyeluruh.

Berdasarkan hasil evaluasi, algoritma Random Forest menunjukkan kemampuan yang baik dalam mengklasifikasikan status kelulusan mahasiswa. Nilai Accuracy menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar, sedangkan nilai Precision, Recall, dan F1-Score menunjukkan bahwa model memiliki keseimbangan yang baik dalam mengidentifikasi mahasiswa yang termasuk ke dalam kelas lulus maupun tidak lulus. Selain itu, nilai AUC mengindikasikan bahwa model memiliki kemampuan yang baik dalam membedakan kedua kelas, sedangkan hasil Confusion Matrix menunjukkan bahwa jumlah prediksi yang benar lebih dominan

dibandingkan prediksi yang salah. Hasil tersebut membuktikan bahwa algoritma Random Forest mampu menghasilkan model klasifikasi yang stabil dan efektif dalam menganalisis data akademik mahasiswa.

Secara keseluruhan, penelitian ini menunjukkan bahwa penerapan algoritma Random Forest dapat menjadi salah satu alternatif yang efektif dalam mendukung pengambilan keputusan di bidang pendidikan, khususnya untuk mengidentifikasi kondisi akademik mahasiswa dan memprediksi status kelulusan berdasarkan data yang tersedia. Informasi yang dihasilkan dari model klasifikasi ini diharapkan dapat membantu perguruan tinggi dalam melakukan identifikasi dini terhadap mahasiswa yang berpotensi mengalami keterlambatan kelulusan sehingga dapat diberikan pendampingan akademik secara lebih tepat. Dengan demikian, pemanfaatan algoritma Random Forest tidak hanya memberikan performa klasifikasi yang baik, tetapi juga berpotensi mendukung peningkatan kualitas layanan akademik serta efektivitas pengelolaan data pendidikan di perguruan tinggi.

6 UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan selama proses penelitian ini. Ucapan terima kasih juga disampaikan kepada penyedia Student Performance Dataset yang digunakan sebagai sumber data penelitian serta kepada pihak Program Studi yang telah memberikan arahan dan masukan sehingga penelitian ini dapat diselesaikan dengan baik.

DAFTAR PUSTAKA

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

Géron, A. (2023). Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow (3rd ed.). O'Reilly Media.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

Harahap, N. A., Syahdewa, A., Lubis, F. I., Gunawan, H., Sibuea, M., Juang, D., & Amin, M. (2026). Perbandingan kinerja algoritma Decision Tree dan Random Forest dalam prediksi kelulusan siswa. Jurnal Sistem Informasi TGD, 5(1), 124–129.

Kurniawan, F. B., & Farokhah, L. (2025). Aplikasi cerdas prediksi kelulusan mahasiswa berbasis website menggunakan metode Support Vector Machine (SVM). Jurnal FASILKOM, 15(1), 155–162.

Hendrastuty, N. (2024). Penerapan data mining menggunakan algoritma K-Means clustering dalam evaluasi hasil pembelajaran siswa. Jurnal Ilmiah Informatika dan Ilmu Komputer, 3(1), 46–56. https://doi.org/10.58602/jima-ilkom.v3i1.26

Nainggolan, I. R. G., Dachi, I., & Suharianto, J. (2025). Analisis faktor-faktor yang mempengaruhi kemiskinan di Provinsi Sumatera Utara. JICN: Jurnal Intelek dan Cendikiawan Nusantara, 2(2).

Setiawan, B. C., Nilawati, A., Ferdian, R., Saputra, R. A., Santoso, Y. P., & Riefky, M. (2025). Analisis cluster hierarki pada tingkat kemiskinan di Provinsi Sumatera Utara tahun 2023. LogicLink: Journal of Artificial Intelligence and Multimedia in Informatics, 2(1), 42–55.

Sianturi, M. W., Sirait, K. F., Nurfadillah, D., Shopia, Lubis, M., Muliani, F., & Saumi, F. (2025). Penerapan analisis cluster K-Means untuk pengelompokan kabupaten/kota di Provinsi Sumatera Utara berdasarkan indikator kemiskinan tahun 2023. Jurnal

Pendidikan Matematika Malikussaleh, 5(3).

Siregar, D. W., & Sitompul, P. (2025). Markov chain K-Means cluster model dalam dinamika transisi tingkat kemiskinan berdasarkan indeks kedalaman dan keparahan kemiskinan di Provinsi Sumatera Utara. MATHunesa, 13(3).

Syahputra, A., & Ikhsan, M. (2025). Data visualization using Tableau with K-Mean clustering method for identification of poor areas in North Sumatera. Jurnal INOVTEK Polbeng – Seri Informatika, 10(2).