

## KLASIFIKASI PRESTASI AKADEMIK MAHASISWA MENGGUNAKAN METODE *RANDOM FOREST*

Serly Sustiana Saputri<sup>1)</sup>, Dwi Remawati<sup>2)</sup>, Teguh Susyanto<sup>3)</sup>, Wawan Laksito Yuly Saptomo<sup>4)</sup>

<sup>1,3</sup>Prodi Sistem Informasi, Fakultas Teknik, Universitas Tiga Serangkai

<sup>2</sup>Prodi Teknologi Informasi, Sekolah Vokasi, Universitas Tiga Serangkai

<sup>4</sup>Prodi Informatika, Fakultas Teknik, Universitas Tiga Serangkai

Correspondence author: SS Saputri, 123400005.serly@tsu.ac.id, Surakarta, Indonesia

### Abstract

*This study aims to classify student academic achievement using the Random forest algorithm, utilizing the Student Performance dataset from Kaggle. The main attributes used as predictors include attendance rate, weekly study duration, and class engagement. The research methodology included data preprocessing with label encoding, an 80:20 split between training and test sets, and standardized model evaluation using accuracy, precision, recall, and F1-score. The results showed an accuracy of 31.03%. This low accuracy is due to the complexity of multi-class classification and imbalanced data distribution. This research contributes to mapping student learning behavior patterns and serves as a reference for developing more optimal academic prediction models.*

**Keywords** : *classification algorithm, multi-class, academic achievement, students, random forest.*

### Abstrak

Penelitian ini bertujuan untuk mengklasifikasikan prestasi akademik mahasiswa menggunakan algoritma *Random forest* dengan memanfaatkan dataset *Students Performance* dari Kaggle. Atribut utama yang digunakan sebagai prediktor meliputi tingkat kehadiran, durasi belajar mingguan, dan keterlibatan di kelas. Metodologi penelitian mencakup pra-pemrosesan data dengan label *encoding*, pembagian data latih dan uji (80:20), serta evaluasi model yang diseragamkan menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score*. Hasil penelitian menunjukkan nilai akurasi sebesar 31,03%. Rendahnya akurasi tersebut merupakan dampak dari kompleksitas klasifikasi multi-kelas dan distribusi data yang tidak seimbang (*imbalanced data*). Penelitian ini memberikan kontribusi dalam memetakan pola perilaku belajar mahasiswa serta menjadi referensi bagi pengembangan model prediksi akademik yang lebih optimal.

**Kata Kunci** : *algoritma klasifikasi, multi-kelas, prestasi akademik, mahasiswa, random forest*

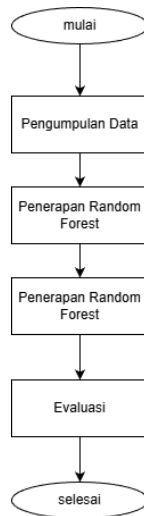
## A. PENDAHULUAN

Prestasi akademik mahasiswa merupakan salah satu indikator utama dalam menilai keberhasilan proses pembelajaran di perguruan tinggi. Tingkat prestasi yang dicapai mahasiswa mencerminkan pemahaman terhadap materi, kedisiplinan, serta keterlibatan aktif dalam kegiatan akademik (Fatunnisa & Marcos, 2024). Prestasi akademik tidak hanya dipengaruhi oleh kemampuan kognitif, tetapi juga oleh berbagai faktor pendukung lainnya seperti kebiasaan belajar, tingkat kehadiran, aktivitas fisik, serta partisipasi mahasiswa dalam kegiatan pembelajaran di kelas. Seiring dengan perkembangan teknologi informasi, institusi pendidikan menghasilkan data akademik mahasiswa dalam jumlah besar yang mencakup berbagai aspek perilaku dan aktivitas belajar. Data tersebut memiliki potensi besar untuk dianalisis guna menemukan pola-pola yang berkaitan dengan prestasi akademik mahasiswa. Pemanfaatan teknik data mining dan *machine learning* menjadi salah satu solusi yang efektif untuk mengolah data tersebut secara sistematis dan objektif. Objek penelitian dalam studi ini adalah dataset Students Performance yang memuat data mahasiswa dengan berbagai atribut, antara lain aktivitas olahraga (*sports activity*), moda transportasi, durasi belajar mingguan, tingkat kehadiran, kebiasaan membaca dan mencatat, keterlibatan dalam mendengarkan materi di kelas, partisipasi dalam kerja proyek, serta nilai akhir (*Grade*) sebagai indikator prestasi akademik. Dataset ini dianalisis untuk mengklasifikasikan tingkat prestasi akademik mahasiswa berdasarkan faktor-faktor tersebut (Almufqi & Voutama, 2023). Berbagai penelitian terdahulu telah menerapkan pendekatan *machine learning* dalam menganalisis dan memprediksi prestasi akademik mahasiswa. Algoritma yang sering digunakan antara lain *K-Nearest Neighbor* (KNN), *Naive Bayes*, *Decision Tree*, *Support Vector Machine*

(SVM), serta *Random forest* (Anofrizen et al., 2026). Penelitian oleh (Al-Shanableh et al., 2026) menunjukkan bahwa teknik data mining mampu memprediksi performa akademik siswa secara efektif. Breiman memperkenalkan *Random forest* sebagai metode *ensemble* yang mampu meningkatkan akurasi dan mengurangi risiko *overfitting* (Firdaus et al., 2026). Selanjutnya, (Kuswanto & Hakim, 2025) serta beberapa penelitian lain membuktikan bahwa *Random forest* memiliki kinerja yang baik dalam menangani data pendidikan yang kompleks dan berdimensi tinggi. Meskipun demikian, beberapa penelitian sebelumnya masih memiliki keterbatasan, seperti kurangnya pemanfaatan atribut non akademik secara menyeluruh, distribusi data yang tidak seimbang, serta minimnya analisis evaluasi model pada kasus klasifikasi multi kelas. Berdasarkan permasalahan tersebut, penelitian ini menggunakan algoritma *Random forest* untuk mengklasifikasikan prestasi akademik mahasiswa dengan memanfaatkan atribut akademik dan non akademik secara bersamaan. Tujuan dari penelitian ini adalah untuk mengklasifikasikan prestasi akademik mahasiswa menggunakan metode *Random forest* serta mengevaluasi kinerja model berdasarkan hasil akurasi dan metrik evaluasi lainnya. Hasil penelitian diharapkan dapat memberikan gambaran mengenai faktor-faktor yang memengaruhi prestasi akademik mahasiswa serta menjadi referensi dalam pengembangan sistem pendukung keputusan di bidang pendidikan berbasis data.

## B. METODE PENELITIAN

Tahapan Penelitian meliputi berbagai proses, berikut adalah alur tahapan penelitian yang dilakukan



**Gambar 1.** Alur Penelitian

### Sumber Data

Data yang digunakan dalam penelitian ini berasal dari platform Kaggle (Kharoua, 2024), khususnya dataset *Students Performance* dalam bentuk file CSV. Dataset ini bersifat terbuka dan sering dimanfaatkan dalam studi di bidang *educational data mining* serta *machine learning*. Data tersebut menggambarkan ciri-ciri dan pola belajar mahasiswa yang terkait dengan hasil akademik mereka. *Dataset Students Performance* mencakup data dari sejumlah mahasiswa dengan sembilan atribut utama, yakni *Transportation*, *Attendance*, *Sports\_activity*, *Weekly\_study\_hours*, *Reading*, *Notes*, *Listening\_in\_class*, *Project\_work*, dan *Grade*. Atribut *Grade* berfungsi sebagai variabel target (label kelas) yang menunjukkan tingkat pencapaian akademik mahasiswa, sementara atribut lainnya bertindak sebagai variabel independen. Pengumpulan data dilakukan dengan mengunduh dataset dalam format CSV, lalu memuatnya ke dalam lingkungan pemrograman Python melalui Google Colab. Selanjutnya, dataset diperiksa untuk memverifikasi struktur data, jenis atribut, dan kelengkapan informasi sebelum diterapkan pada tahap pemodelan.

**Tabel 1.** Sumber Data

| Atribut            | STU<br>DEN<br>T1 | STU<br>DEN<br>T2 | STU<br>DEN<br>T3 | STU<br>DEN<br>T4 | STU<br>DEN<br>T5 |
|--------------------|------------------|------------------|------------------|------------------|------------------|
| Student_ID         | STU<br>DEN<br>T1 | STU<br>DEN<br>T2 | STU<br>DEN<br>T3 | STU<br>DEN<br>T4 | STU<br>DEN<br>T5 |
| Student_Age        | 19-<br>22        | 19-<br>22        | 19-<br>22        | 18               | 19-<br>22        |
| Sex                | Male             | Male             | Male             | Female           | Male             |
| High_School_Type   | Other            | Other            | State            | Private          | Private          |
| Scholarship        | 50%              | 50%              | 50%              | 50%              | 50%              |
| Additional_Work    | Yes              | Yes              | No               | Yes              | No               |
| Sports_activity    | No               | No               | No               | No               | No               |
| Transportation     | Private          | Private          | Private          | Bus              | Bus              |
| Weekly_Study_Hours | 0                | 0                | 2                | 2                | 12               |
| Attendance         | Always           | Always           | Never            | Always           | Always           |
| Reading            | Yes              | Yes              | No               | No               | Yes              |
| Notes              | Yes              | No               | No               | Yes              | No               |
| Listening_in_Class | No               | Yes              | No               | No               | Yes              |
| Project_work       | No               | Yes              | Yes              | No               | Yes              |
| Grade              | AA               | AA               | AA               | AA               | AA               |

### Pembagian Data (*Data Splitting*)

Dataset yang sudah melewati proses pra-pemrosesan selanjutnya dibagi menjadi data pelatihan (*training data*) dan data pengujian (*testing data*). Pembagian ini dilakukan dengan rasio 80:20, di mana 80% dari data dijadikan data pelatihan dan 20% sisanya sebagai data pengujian (Azis, 2025; Muningsih, 2022). Data pelatihan dimanfaatkan untuk mengembangkan model klasifikasi, sementara data pengujian digunakan untuk mengukur performa model yang telah dibuat.

### Pemodelan dengan *Random forest*

Pada tahap pemodelan, algoritma *Random forest* digunakan sebagai metode klasifikasi. Model *Random forest* dibangun

menggunakan *library Scikit-Learn* dengan parameter default. Atribut input (X) terdiri dari seluruh variabel independen, sedangkan atribut target (y) adalah variabel *Grade*.

$$Gini(D) = 1 - \sum_{i=1}^C p_i^2 \quad (1)$$

Keterangan:

D: Dataset atau node saat ini.

C: Jumlah kelas (*Grade*).

$p_i$ : Probabilitas pemilihan item dengan label kelas  $i$ .

Selain itu, karena *Random forest* adalah metode *ensemble*, ditambahkan rumus *Majority Voting* untuk menentukan prediksi akhir:

$$\hat{Y} = \text{mode} \{T_1(x), T_2(x), \dots, T_n(x)\} \quad (2)$$

Keterangan:

$\hat{Y}$ : Hasil prediksi akhir.

$T_n(x)$ : Hasil prediksi dari pohon keputusan ke- $n$ .

## Evaluasi Model

### 1. Akurasi

Akurasi digunakan untuk mengukur tingkat ketepatan model dalam mengklasifikasikan data, yaitu perbandingan antara jumlah prediksi yang benar dengan total data uji. Semakin tinggi nilai akurasi, maka semakin baik performa model klasifikasi. Rumus akurasi ditunjukkan pada Persamaan (3)

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Keterangan :

TP (*True Positive*) adalah jumlah data yang diklasifikasikan dengan benar sebagai kelas tertentu.

TN (*True Negative*) adalah jumlah data yang diklasifikasikan dengan benar sebagai bukan kelas tersebut,

FP (*False Positive*) adalah jumlah data yang salah diklasifikasikan sebagai kelas tertentu,

FN (*False Negative*) adalah jumlah data yang salah diklasifikasikan sebagai bukan kelas tersebut.

### 2. Presisi

Presisi digunakan untuk mengukur tingkat ketepatan prediksi model terhadap suatu kelas tertentu. Nilai presisi menunjukkan seberapa besar proporsi data yang diprediksi sebagai suatu kelas benar-benar termasuk ke dalam kelas tersebut.

Rumus presisi ditunjukkan pada Persamaan (4):

$$\text{Presisi} = \frac{TP}{TP + FP} \quad (4)$$

### 3. Recall

Recall digunakan untuk mengukur kemampuan model dalam mengenali seluruh data yang termasuk dalam suatu kelas tertentu. Nilai recall menunjukkan seberapa besar proporsi data aktual suatu kelas yang berhasil diprediksi dengan benar oleh model. Rumus recall ditunjukkan pada Persamaan (5):

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

### 4. Confusion Matrix

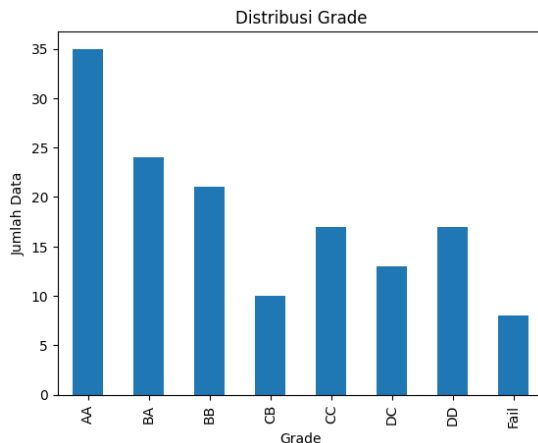
*Confusion matrix* merupakan metode evaluasi yang menyajikan perbandingan antara kelas aktual dan kelas hasil prediksi model dalam bentuk tabel. *Confusion matrix* digunakan untuk mengetahui jumlah prediksi yang benar dan salah pada setiap kelas, serta menjadi dasar perhitungan nilai akurasi, presisi, dan recall dalam model klasifikasi multi-kelas (Afandi et al., 2025).

## C. HASIL DAN PEMBAHASAN

Pembahasan meliputi eksplorasi data, proses *preprocessing*, hasil pemodelan, dan evaluasi kinerja model.

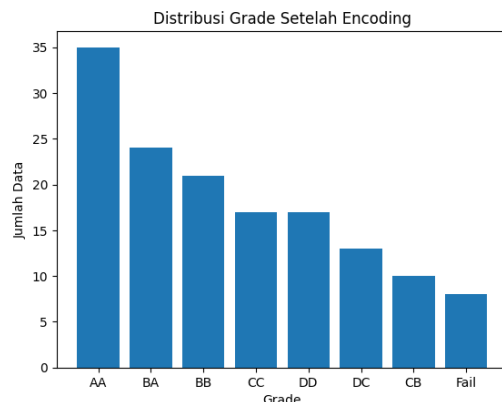
### Eksplorasi Data

Tahap eksplorasi data bertujuan untuk memahami karakteristik dataset sebelum dilakukan pemodelan. Dataset *Students Performance* terdiri dari beberapa atribut akademik dan non-akademik yang digunakan untuk memprediksi mahasiswa.



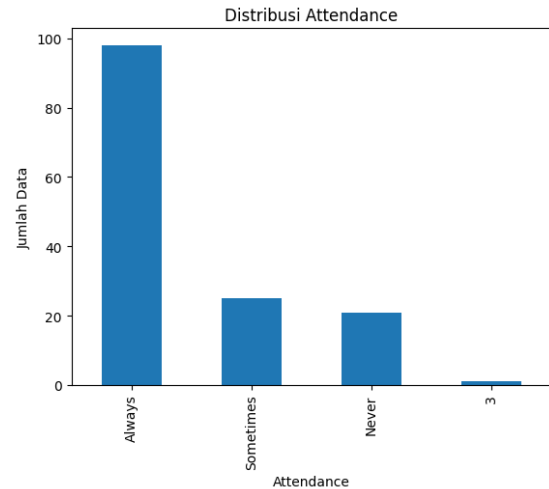
**Gambar 1.** Distribusi *Grade*

Gambar 2 menunjukkan distribusi nilai *Grade* pada *dataset Students Performance*. Terlihat bahwa jumlah data pada setiap kelas *Grade* tidak tersebar secara merata. Beberapa kelas memiliki jumlah data yang jauh lebih banyak dibandingkan kelas lainnya. Ketidakseimbangan distribusi kelas (*class imbalance*) ini dapat memengaruhi performa model klasifikasi, karena model cenderung lebih optimal dalam mempelajari pola dari kelas dengan jumlah data yang dominan.



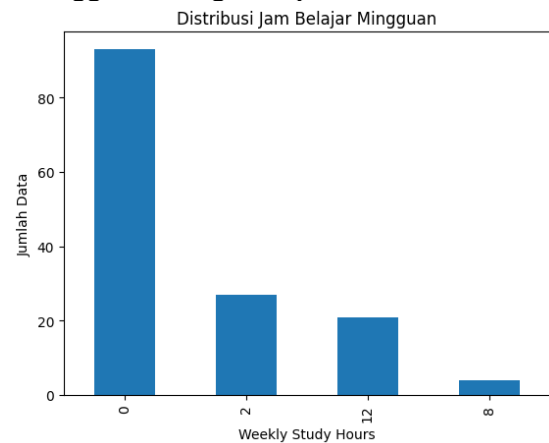
**Gambar 2.** Distribusi *Grade* setelah Encoding

Gambar 3 menunjukkan distribusi kategori *Grade* setelah dilakukan proses *encoding* (Dachi & Sitompul, 2023). Visualisasi ini digunakan untuk memastikan bahwa distribusi kelas tetap tidak seimbang, yang berpotensi memengaruhi kinerja model klasifikasi.



**Gambar 3.** Distribusi *Attendance*

Gambar 4 memperlihatkan distribusi tingkat kehadiran (*Attendance*) mahasiswa. Grafik menunjukkan bahwa sebagian besar mahasiswa memiliki tingkat kehadiran yang cukup baik. Kehadiran mahasiswa berperan penting dalam proses pembelajaran karena berkaitan langsung dengan keterlibatan mahasiswa perkuliahan, dalam menerima sehingga memengaruhi prestasi akademik.



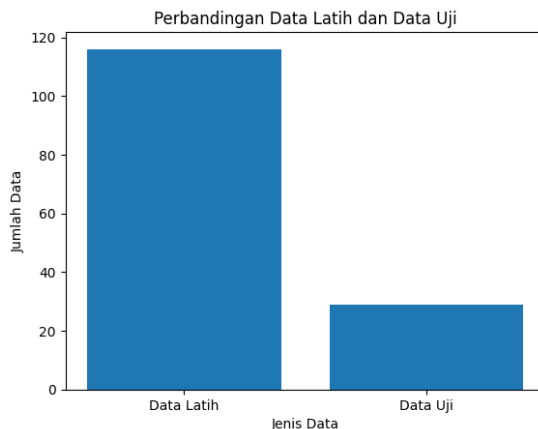
**Gambar 4.** Distribusi *Weekly Study Hours*

Gambar 5 menunjukkan distribusi jam belajar mingguan mahasiswa. Terlihat

adanya variasi durasi belajar antar mahasiswa, yang mencerminkan perbedaan kebiasaan belajar (Aprilia et al., 2021). Variasi ini dapat memengaruhi perbedaan capaian prestasi akademik, sehingga atribut *Weekly study hours* digunakan sebagai salah satu variabel input dalam proses klasifikasi.

### Splitting Dataset

*Splitting Data* memperlihatkan pembagian dataset menjadi data latih dan data uji dengan rasio 80% dan 20% yang digunakan dalam proses pelatihan dan pengujian model *Random forest* (Muningsih, 2022).



**Gambar 5.** Perbandingan Jumlah Data Latih

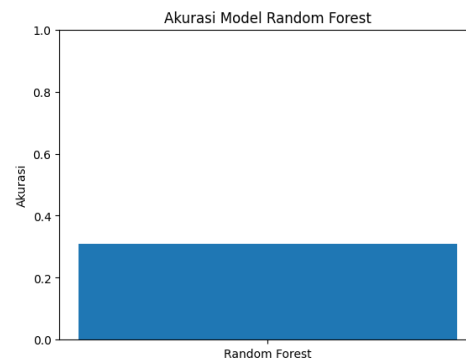
### Penerapan Algoritma *Random forest*

Algoritma *Random forest* diterapkan sebagai metode klasifikasi dalam penelitian ini. Seluruh atribut independen digunakan sebagai variabel input, sedangkan atribut *Grade* digunakan sebagai variabel target. Proses pelatihan model dilakukan menggunakan data latih yang telah dipersiapkan pada tahap sebelumnya.

Akurasi Model: 0.3103448275862069

**Gambar 6.** Hasil Akurasi Model

Setelah itu hasil akurasinya *random forest* di visualkan



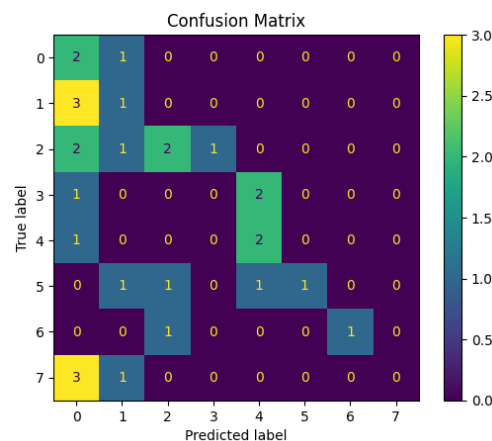
**Gambar 7.** Visual Hasil Akurasi *Random forest*

### Evaluasi Model

Evaluasi model dilakukan untuk mengukur kinerja algoritma *Random forest* dalam mengklasifikasikan prestasi akademik mahasiswa. Metode evaluasi yang digunakan dalam penelitian ini adalah *confusion matrix* dan akurasi (Jin et al., 2020).

| Classification Report: |           |        |          |         |
|------------------------|-----------|--------|----------|---------|
|                        | precision | recall | f1-score | support |
| 0                      | 0.17      | 0.67   | 0.27     | 3       |
| 1                      | 0.20      | 0.25   | 0.22     | 4       |
| 2                      | 0.50      | 0.33   | 0.40     | 6       |
| 3                      | 0.00      | 0.00   | 0.00     | 3       |
| 4                      | 0.40      | 0.67   | 0.50     | 3       |
| 5                      | 1.00      | 0.25   | 0.40     | 4       |
| 6                      | 1.00      | 0.50   | 0.67     | 2       |
| 7                      | 0.00      | 0.00   | 0.00     | 4       |
| accuracy               |           |        | 0.31     | 29      |
| macro avg              | 0.41      | 0.33   | 0.31     | 29      |
| weighted avg           | 0.40      | 0.31   | 0.29     | 29      |

**Gambar 8.** Hasil *Precision, Recal, F1-Score*



**Gambar 9.** Visual *Confussion Matrix*

#### D. PENUTUP

Berdasarkan hasil penelitian yang telah dilakukan menggunakan *dataset Students Performance* dengan pendekatan *machine learning* dan algoritma *Random forest*,

Algoritma *Random forest* dapat diterapkan untuk melakukan klasifikasi prestasi akademik mahasiswa berdasarkan atribut akademik dan non-akademik yang terdapat pada dataset.

Tahap eksplorasi data menunjukkan bahwa distribusi kelas *Grade* tidak merata, sehingga memengaruhi kemampuan model dalam melakukan klasifikasi secara optimal.

Hasil evaluasi model menggunakan *confusion matrix* menunjukkan bahwa masih terdapat kesalahan klasifikasi pada beberapa kelas *Grade*, khususnya pada kelas dengan jumlah data yang relatif sedikit.

Model *Random forest* menghasilkan nilai akurasi sebesar 31,03%, yang mengindikasikan bahwa performa model masih tergolong rendah dan memerlukan peningkatan lebih lanjut.

Analisis *feature importance* menunjukkan bahwa atribut *Attendance*, *Weekly\_study\_hours*, dan *Listening\_in\_class* merupakan faktor dominan yang memengaruhi prestasi akademik mahasiswa

Algoritma *Random forest* mampu mengklasifikasikan prestasi akademik dengan hasil akurasi 31,03%. Temuan ini menjawab permasalahan mitra (Universitas Tiga Serangkai) dalam memetakan faktor kunci keberhasilan mahasiswa, yaitu tingkat kehadiran dan durasi belajar.

#### E. DAFTAR PUSTAKA

Afandi, M. C., Indahyanti, U., Setiawan, H., & Kautsar, I. A. (2025). Perbandingan Algoritma Machine Learning dalam Memprediksi Kelulusan Mahasiswa. *SemanTIK : Teknik Informasi*, 11(2), 96–105. <https://doi.org/10.55679/semantik.v11i2>

.154

Al-Shanableh, N., Alzyoud, M. S., Khalil, A., Benlamine, M., Kraidia, I., Damrah, S., & Tabakhi, A. M. (2026). Forecasting Students' Academic Performance in Educational Data Using Machine Learning Techniques. *International Journal of Information and Communication Technology Education*, 22(1). <https://doi.org/10.4018/IJICTE.399756>

Almufqi, F. M., & Voutama, A. (2023). Perbandingan Metode Data Mining Untuk Memprediksi Prestasi Akademik Siswa. *Jurnal Teknik*, 15(1), 61–66. <https://doi.org/10.30736/jt.v15i1.929>

Anofrizen, Syaifullah, Rahmah, N., Ramadhani, S. A., Sandero, A. A., & Lubis, F. S. (2026). Literature Review: Penerapan Algoritma Machine Learning pada Sistem Informasi Akademik untuk Prediksi Kelulusan Mahasiswa Tepat Waktu. *JOECY: Journal of Innovative and Creativity*, 6(2), 29083–29097. <https://doi.org/10.31004/joecy.v6i2.11149>

Apriliyah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest. *Sistemasi: Jurnal Sistem Informasi*, 10(1), 163–171. <https://doi.org/10.32520/stmsi.v10i1.1129>

Azis, A. R. (2025). Analisis Komparasi Algoritma Machine Learning dalam Prediksi Performa Akademik Mahasiswa: Literature Review. *Jurnal Ilmu Komputer Dan Informatika*, 4(2), 143–148. <https://doi.org/10.54082/jiki.212>

Dachi, J. M. A. S., & Sitompul, P. (2023). Comparison Analysis of the XGBoost Algorithm and Random Forest Ensemble Learning Algorithm in Credit

- Decision Classification. *Jurnal Riset Rumpun Matematika Dan Ilmu Pengetahuan Alam*, 2(2), 87–103. <https://doi.org/10.55606/jurrimipa.v2i2.1470>
- Fatunnisa, A., & Marcos, H. (2024). Prediksi Kelulusan Tepat Waktu Siswa SMK Teknik Komputer Menggunakan Algoritma Random Forest. *Jurnal Manajemen Informatika (JAMIKA)*, 14(1), 101–111. <https://doi.org/10.34010/jamika.v14i1.12114>
- Firdaus, U., Alfiah, A., & Mohdo, L. (2026). Analisis Kinerja Decision Tree dan Random Forest Menggunakan Dataset Breast Cancer. *Jurnal Pendidikan Sains Dan Komputer*, 6(1), 37–42. <https://doi.org/10.47709/jpsk.v6i01.7892>
- Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., & Qiang, B. (2020). RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12343 LNCS, 503–515. [https://doi.org/10.1007/978-3-030-62008-0\\_35](https://doi.org/10.1007/978-3-030-62008-0_35)
- Kharoua, R. El. (2024). *Students Performance Dataset*. <https://www.kaggle.com/datasets/rabiee/kharoua/students-performance-dataset>
- Kuswanto, J., & Hakim, L. (2025). Penerapan Algoritma Random Forest untuk memprediksi Performa Akademik Mahasiswa. *Decode: Jurnal Pendidikan Teknologi Informasi*, 5(1), 262–270. <https://doi.org/10.51454/decode.v5i1.1103>
- Muningsih, E. (2022). Kombinasi Metode K-Means Dan Decision Tree Dengan Perbandingan Kriteria Dan Split Data. *Jurnal Teknoinfo*, 16(1), 113. <https://doi.org/10.33365/jti.v16i1.1561>