

Klasifikasi Prestasi Akademik Mahasiswa Menggunakan Metode Random Forest

Sofiyatus Zawiyah¹, *Lailatul Qodriyah², Moh. Badri Tamam³
^{1,2,3}) Teknik Informatika, Fakultas Teknik, Universitas Islam Madura
JL. Pondok Peantren Miftahul Ulum Bettet, Pamekasan, Jawa Timur
Email: ¹sfyatszwyh@gmail.com, ²l.qodriyah1227@gmail.com, ³badri.uimadura@gmail.com

ABSTRACT

Student academic achievement is the main indicator in evaluating the quality of higher education. This study aims to classify student academic achievement using the Random Forest algorithm by utilizing academic data such as mid-term exam scores, final exam scores, assignments, attendance, notes, and final grades. The methods used include data collection from Kaggle, data preprocessing, data sharing, model training using Random Forest, and model performance evaluation. The results showed that the Random Forest model was able to classify student academic achievement with a high accuracy of 98.2%. Further analysis revealed that mid-term exam scores and final exam scores were the dominant factors in influencing the prediction of students' final grades. The conclusion of this study shows that the Random Forest algorithm is an effective tool in analyzing student academic data because it is able to provide accurate classification results and provide important insights for educational institutions to improve the quality of learning.

Keywords : Random Forest; Academic Performance; Algorithm; Higher Education; Classification

ABSTRAK

Prestasi akademik mahasiswa merupakan indikator utama dalam mengevaluasi kualitas pendidikan tinggi. Penelitian ini bertujuan untuk mengklasifikasikan prestasi akademik mahasiswa menggunakan algoritma Random Forest dengan memanfaatkan data akademik seperti nilai UTS, UAS, tugas, kehadiran, catatan, dan nilai akhir. Metode yang digunakan meliputi pengumpulan data dari Kaggle, preprocessing data, pembagian data, pelatihan model menggunakan Random Forest, dan evaluasi kinerja model. Hasil penelitian menunjukkan bahwa model Random Forest mampu mengklasifikasikan prestasi akademik mahasiswa dengan akurasi yang tinggi sebesar 98.2%. Analisis lebih lanjut mengungkapkan bahwa nilai UTS dan UAS merupakan faktor dominan dalam memengaruhi prediksi nilai akhir mahasiswa. Kesimpulan dari penelitian ini menunjukkan bahwa algoritma Random Forest merupakan alat yang efektif dalam analisis data akademik mahasiswa karena mampu memberikan hasil klasifikasi yang akurat serta memberikan wawasan penting bagi institusi pendidikan untuk meningkatkan kualitas pembelajaran.

Kata kunci : Random Forest; Prestasi Akademik; Algoritma; Pendidikan Tinggi; Klasifikasi

1. PENDAHULUAN

Prestasi akademik mahasiswa menjadi salah satu indikator utama untuk mengevaluasi kualitas sistem pendidikan tinggi [1]. Menurut Datu dan Buenconsejo (2021), keterlibatan akademik dan pencapaian berperan penting dalam memprediksi adaptabilitas karier mahasiswa. Di era persaingan pendidikan yang semakin ketat, institusi pendidikan memerlukan pemahaman yang mendalam tentang faktor-faktor yang memengaruhi prestasi akademik untuk merancang intervensi yang efektif [2].

Seiring kemajuan teknologi informasi, data akademik mahasiswa kini lebih mudah diakses, mencakup variabel seperti nilai ujian tengah semester (UTS), ujian akhir semester (UAS), tugas, catatan, dan Nilai akhir [3]. Namun, pemanfaatan data ini masih menghadapi tantangan dalam proses analisis dan ekstraksi informasi yang berguna. Untuk mengatasi tantangan ini, metode berbasis pembelajaran mesin (*machine learning*) telah banyak diterapkan dalam bidang pendidikan, salah satunya adalah *algoritma Random Forest* [4].

Beberapa penelitian sebelumnya telah membuktikan efektivitas algoritma *Random Forest* dalam mengolah data pendidikan. Yuda et al. (2022) menunjukkan bahwa penerapan *Random Forest* untuk mengklasifikasikan kelulusan mahasiswa berdasarkan Indeks Prestasi Kumulatif (IPK) dan jumlah SKS yang ditempuh memberikan tingkat akurasi sebesar 98% [5]. Selain itu, Jawad et al. (2022) memanfaatkan *Random Forest* dalam memprediksi kinerja akademik mahasiswa di lingkungan belajar *virtual (Virtual Learning Environment/VLE)* menggunakan dataset *Open University Learning Analytics (OULAD)*. Studi ini menyoroti kemampuan *Random Forest* dalam memprediksi kinerja akademik berdasarkan skor penilaian dan keterlibatan siswa [6].

Meskipun demikian, masih terdapat celah penelitian dalam eksplorasi lebih lanjut mengenai bagaimana algoritma ini dapat dioptimalkan untuk mengolah variabel-variabel spesifik dalam pendidikan tinggi, seperti keterkaitan antara tugas, catatan, dan nilai akhir [7]. Selain itu, pengintegrasian *algoritma Random Forest* dengan data akademik yang beragam belum banyak dieksplorasi

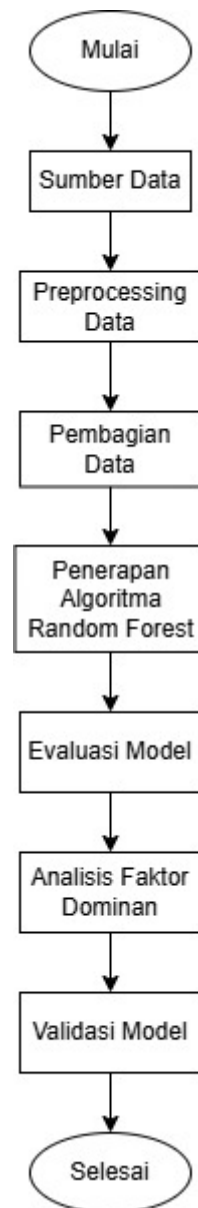
dalam konteks pendidikan di Indonesia, yang memiliki karakteristik data yang berbeda dibandingkan dengan studi internasional sebelumnya [8].

Penelitian ini bertujuan untuk mengklasifikasikan prestasi akademik mahasiswa menggunakan metode *Random Forest* dengan mengolah data akademik seperti nilai ujian tengah semester (UTS), ujian akhir semester (UAS), tugas, kehadiran, catatan, dan Nilai akhir. Pendekatan ini diharapkan dapat menghasilkan analisis yang lebih mendalam mengenai faktor-faktor dominan yang memengaruhi prestasi mahasiswa serta memberikan kontribusi terhadap pengembangan metode berbasis data untuk meningkatkan kualitas pendidikan di tingkat institusi pendidikan tinggi.

2. METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan *algoritma Random Forest* untuk mengklasifikasikan prestasi akademik mahasiswa berdasarkan dataset yang diunduh dari *platform Kaggle*. Proses prediksi dan implementasi algoritma dilakukan menggunakan *Orange*, sebuah *software* untuk analisis data dan pembelajaran mesin berbasis visual [9].

Berikut adalah alur tahapan penelitian yang dilakukan:



Gambar 1. Alur Penelitian

Berdasarkan *flowchart* yang terdapat pada Gambar 1, berikut adalah tahapan proses pemodelan yang dilakukan:

2.1. Sumber Data

Data yang digunakan dalam penelitian ini diambil dari dataset publik yang tersedia di Kaggle dengan judul “Nilai Mahasiswa Semester 2023”. Dataset ini mencakup 134 sampel data mahasiswa dengan variabel nilai ujian tengah semester (UTS), ujian akhir semester (UAS), tugas, kehadiran, catatan, dan Nilai akhir.

2.2. Preprocessing Data

Proses *preprocessing* dilakukan dalam *Orange* untuk menyiapkan dataset yang siap digunakan dalam model pembelajaran mesin, meliputi: *Handling Missing Values*: Menggunakan *widget Data Table* untuk mengidentifikasi dan mengimputasi nilai yang hilang. *Normalisasi Data*: Menggunakan *widget Normalize* untuk menormalisasi data numerik ke rentang [0,1].

2.3. Pembagian Data

Dataset dibagi menggunakan *widget Data Sampler* di *Orange* menjadi dua subset:

Data Latih

Data Uji

Proses pembagian ini dilakukan secara acak dengan menggunakan

teknik *stratified sampling* untuk memastikan distribusi yang seimbang dari label target. Dataset dibagi dengan rasio 80:20, di mana 80% data digunakan untuk pelatihan (*training*) dan 20% sisanya untuk pengujian (*testing*).

2.4. Penerapan Algoritma Random Forest

Model *Random Forest* diterapkan menggunakan *widget Random Forest* dalam *Orange*.

Parameter utama yang digunakan dalam proses pelatihan meliputi:

Jumlah pohon keputusan (number of trees): 100.

Kedalaman maksimum pohon (max depth): 10.

Kriteria pemisahan (criterion): Entropy.

Model dilatih menggunakan data latih dan dievaluasi dengan data uji untuk mengukur akurasi model.

2.5. Evaluasi Model

Untuk evaluasi model, metrik yang digunakan meliputi:

Akurasi: Diukur melalui *widget Test & Score* di *Orange*.

Precision, *Recall*, dan *F1-Score*:

Digunakan untuk menilai kinerja model dalam klasifikasi multikategori.

Confusion Matrix: Diperoleh menggunakan *widget Confusion Matrix* di *Orange* untuk menganalisis kesalahan klasifikasi pada masing-masing kelas.

2.6. Analisis Faktor Dominan

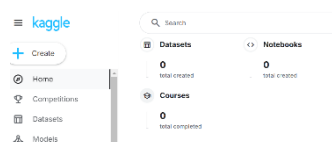
Analisis dilakukan untuk mengidentifikasi variabel-variabel yang paling berpengaruh terhadap prediksi prestasi akademik mahasiswa. Fitur penting dihitung menggunakan *widget Feature Statistics* yang tersedia di *Orange*.

2.7. Validasi Model

Untuk memastikan keandalan hasil, dilakukan validasi silang (*cross-validation*) dengan menggunakan *widget Cross Validation* dalam *Orange*, dengan $k = 5$. Teknik ini digunakan untuk menghindari *overfitting* dan memastikan kestabilan model.

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Dataset



Gambar 2. Sumber Dataset

Penelitian ini menggunakan dataset yang diunduh dari *Kaggle*. Dataset tersebut mencakup 334

sampel data mahasiswa yang melibatkan beberapa variabel penting dalam penilaian akademik, yaitu UTS, UAS, Tugas, Kehadiran, Catatan, dan Nilai Akhir. Variabel-variabel tersebut digunakan untuk memprediksi prestasi akademik mahasiswa berdasarkan klasifikasi nilai akhir [10].

3.1.1 Proses Preprocessing

Sebelum digunakan dalam pemodelan, dataset tersebut melalui beberapa tahapan *preprocessing* menggunakan *Orange*, untuk memastikan data siap digunakan dalam model pembelajaran mesin. Tahapan Preprocessing yang dilakukan meliputi Missing Value dan Normalisasi Data:

1. Penanganan Nilai yang Hilang (*Missing Values*): Dataset diperiksa untuk mengetahui apakah terdapat nilai yang hilang.

Gambar 3. Pemeriksaan Nilai Hilang

Hasilnya menunjukkan bahwa tidak ada nilai yang hilang dalam

variabel (UTS, UAS, Tugas, Kehadiran, Nilai Akhir, dan Catatan).

2. Normalisasi Data: Variabel numerik dinormalisasi ke dalam rentang [0,1] menggunakan teknik *min-max normalization* dengan widget "Normalize" di Orange.

	NILAI AKHIR	Nama	No	UTS	UAS	TUGAS
2	Baik	ABYAN KHARIRI	0.007519	0.940065	0.873027	0.802434
3	Baik	ALDEVA	0.015038	0.940065	0.873022	0.802434
4	Cukup	ALLYA AZZAHRI	0.022556	0.709172	0.67476	0.703809
5	Baik	ANNISA NUR	0.030075	0.742157	0.714412	0.785996
6	Baik	CARLICO	0.037294	0.827609	0.833369	0.977496
7	Baik	DIBLY AYLA	0.040113	0.742157	0.853796	0.884621
8	Baik	FACHRI	0.052632	0.709172	0.773891	0.720246

Gambar 4. Normalisasi Data

Teknik ini penting untuk menghindari perbedaan skala antarvariabel yang dapat memengaruhi hasil pemodelan.

3.2 Hasil Evaluasi Model

Model *Random Forest* diterapkan untuk mengklasifikasikan prestasi akademik mahasiswa menggunakan data latih dan data uji. Hasil evaluasi model menunjukkan performa yang baik berdasarkan metrik berikut: *akurasi (CA) sebesar 98.2%, precision sebesar 98.2%, recall sebesar 98.2%, dan F1-score sebesar 98.2%*. Selain itu, nilai *AUC* sebesar 0.999 menunjukkan kemampuan diskriminasi model yang sangat baik.

Model	AUC	CA	F1	Prec	Recall
Random Forest	0.999	0.982	0.982	0.982	0.982

Gambar 5. Hasil Evaluasi Model

3.2.1 Akurasi Model

Hasil evaluasi menunjukkan bahwa model *Random Forest* menghasilkan akurasi sebesar 98.2%, yang menunjukkan bahwa model mampu dengan tepat mengklasifikasikan prestasi akademik mahasiswa dalam hampir seluruh kasus. Ini menunjukkan bahwa model berhasil mengklasifikasikan prestasi akademik mahasiswa dengan tingkat kesalahan yang sangat rendah, mengingat performa ini didukung oleh metrik lain seperti *precision*, *recall*, dan *F1-score* yang juga mencapai nilai 98.2%, menunjukkan konsistensi yang sangat baik dalam kinerja model.

Jika dibandingkan dengan penelitian terdahulu, Yuda et al. (2022) juga menggunakan model

Random Forest untuk mengklasifikasikan kelulusan mahasiswa berdasarkan Indeks Prestasi Kumulatif (IPK) dan jumlah SKS yang ditempuh, dengan tingkat akurasi sebesar 98%. Hasil ini menunjukkan bahwa model Random Forest memiliki kinerja yang sangat baik dalam mengklasifikasikan data akademik mahasiswa, baik pada penelitian ini maupun pada penelitian sebelumnya.

3.2.2 Precision, Recall, dan F1-Score

Untuk mengevaluasi kinerja model lebih lanjut, metrik lain seperti *precision*, *recall*, dan *F1-score* dihitung sebagai berikut:

Precision: 98.2%

Recall: 98.2%

F1-score: 98.2%

Nilai-nilai ini menunjukkan bahwa model *Random Forest* memiliki keseimbangan yang cukup baik antara ketepatan (*precision*) dan kemampuan untuk menangkap semua kategori yang relevan (*recall*), dengan *F1-score* yang mendekati nilai *precision* dan *recall*, mencerminkan stabilitas performa model.

3.2.3 Confusion Matrix

Berikut adalah *confusion matrix* yang menggambarkan hasil prediksi model *Random Forest* terhadap kelas yang telah ditentukan:

		Predicted			
		Baik	Cukup	Kurang	Σ
Actual	Baik	136	2	0	138
	Cukup	2	152	2	156
	Kurang	0	2	38	40
Σ		138	156	40	334

Gambar 6. Confusion Matrix

Interpretasi:

Kategori "Baik":

Dari data aktual dalam kategori "Baik," sebanyak 136 data berhasil diklasifikasikan dengan benar.

Sebanyak 2 data salah diklasifikasikan sebagai "Cukup."

Tidak ada data "Baik" yang salah diklasifikasikan sebagai "Kurang."

Kesimpulan: Model menunjukkan tingkat akurasi yang sangat tinggi untuk kategori ini, dengan kesalahan minimal.

Kategori "Cukup":

Sebanyak 152 data dari kategori aktual "Cukup" diklasifikasikan dengan benar.

2 data salah diklasifikasikan sebagai "Baik," dan 2 data lainnya salah diklasifikasikan sebagai "Kurang."

Kesimpulan: Kategori ini memiliki performa klasifikasi yang sangat baik,

meskipun ada sedikit kesalahan klasifikasi ke kategori lain.

Kategori "Kurang":

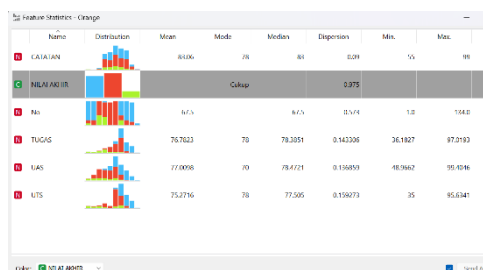
Dari data aktual dalam kategori "Kurang," 38 data berhasil diklasifikasikan dengan benar.

Sebanyak 2 data salah diklasifikasikan sebagai "Cukup," dan tidak ada data yang salah diklasifikasikan sebagai "Baik."

Kesimpulan: Model memiliki performa yang baik dalam mengenali data dari kategori ini, dengan tingkat kesalahan yang rendah.

3.3 Analisis Faktor Dominan

Analisis dilakukan untuk mengidentifikasi variabel-variabel yang paling memengaruhi prediksi Nilai Akhir mahasiswa.



Gambar 7. Analisis Faktor Dominan

Berdasarkan hasil perhitungan dari *widget Feature Statistics* pada *Orange*, variabel-variabel yang memiliki pengaruh utama adalah sebagai berikut: UTS dengan rata-rata nilai 75,27 dan median 77,08.

UAS dengan rata-rata nilai 77,55 dan median 77,95.

TUGAS dengan rata-rata nilai 77,31 dan median 78,03.

CATATAN dengan rata-rata nilai 82,74 dan median 83.

Dari hasil ini, terlihat bahwa semua fitur, terutama UTS, UAS, dan TUGAS, menunjukkan variasi yang signifikan dan berperan penting dalam menentukan prediksi Nilai Akhir mahasiswa. Variabel CATATAN meskipun memiliki nilai rata-rata lebih tinggi, berkontribusi dengan penyebaran data yang lebih kecil.

3.4 Pembahasan

Hasil evaluasi menunjukkan bahwa model Random Forest memiliki performa yang sangat baik dalam mengklasifikasikan Nilai Akhir mahasiswa dengan akurasi sebesar 98.2%. Tingginya nilai akurasi ini mencerminkan kemampuan algoritma Random Forest dalam menangani data akademik yang melibatkan variabel numerik, seperti UTS, UAS, Tugas, dan Catatan, serta variabel kategorikal seperti kategori Nilai Akhir.

Namun, meskipun metrik akurasi, precision, recall, dan F1-score menunjukkan hasil yang konsisten, beberapa kesalahan klasifikasi tetap

ditemukan, terutama pada kategori Cukup dan Kurang. Kesalahan ini diduga disebabkan oleh distribusi data yang tidak seimbang, di mana sebagian besar data lebih terkonsentrasi pada kategori Baik. Distribusi yang tidak merata dapat mengurangi sensitivitas model terhadap kategori dengan jumlah sampel yang lebih kecil. Untuk mengatasi masalah ini, pendekatan seperti class balancing atau oversampling pada data kategori minoritas dapat diterapkan. Alternatif lainnya adalah dengan menggunakan teknik seperti SMOTE (Synthetic Minority Oversampling Technique) untuk meningkatkan representasi data pada kategori yang kurang terwakili.

Hasil analisis faktor dominan menunjukkan bahwa variabel UTS dan UAS memiliki pengaruh paling besar terhadap prediksi Nilai Akhir. Hal ini wajar mengingat nilai ujian sering kali menjadi indikator utama dalam menilai prestasi akademik mahasiswa. Selain itu, variabel Tugas dan Catatan juga memberikan kontribusi signifikan, meskipun tidak sebesar pengaruh nilai ujian. Variabel Catatan, misalnya, memiliki nilai rata-rata yang tinggi, namun dengan penyebaran data yang lebih kecil, sehingga kontribusinya

terhadap variasi model menjadi lebih stabil.

Secara keseluruhan, penelitian ini menunjukkan bahwa algoritma Random Forest tidak hanya mampu memberikan hasil klasifikasi yang akurat, tetapi juga menyediakan wawasan penting mengenai faktor-faktor utama yang memengaruhi prestasi akademik mahasiswa. Model ini tidak hanya mampu memberikan hasil yang akurat, tetapi juga dapat memberikan wawasan yang berguna bagi institusi pendidikan. Informasi ini dapat digunakan untuk merancang strategi pembelajaran yang lebih efektif, seperti memberikan perhatian lebih pada peningkatan nilai ujian melalui metode pengajaran yang lebih inovatif dan efisien.

4. KESIMPULAN

Setelah seluruh tahapan penelitian ini selesai dilaksanakan, dapat disimpulkan bahwa penelitian ini berhasil mengklasifikasikan prestasi akademik mahasiswa menggunakan *algoritma Random Forest* dengan akurasi yang tinggi, yaitu sebesar 98.2%. Model ini menunjukkan kemampuan yang baik dalam menangani data akademik mahasiswa yang melibatkan variabel numerik seperti UTS, UAS,

Tugas, dan Catatan, serta variabel kategorikal seperti kategori Nilai Akhir. Variabel UTS dan UAS ditemukan sebagai faktor dominan yang paling memengaruhi hasil prediksi Nilai Akhir mahasiswa, menunjukkan pentingnya nilai ujian dalam menentukan prestasi akademik.

5. DAFTAR PUSTAKA

- [1] Dima, A., Kleden, M. A., & Atti, A. "Analisis Faktor-Faktor yang Mempengaruhi Prestasi Siswa Menggunakan Metode Structural Equation Modeling (SEM)." *Statistika*, vol. 23, no. 2, 2023, pp. 132–146. <https://doi.org/10.29313/statistika.v23i2.2642>.
- [2] Datu, J. A. D., and Buenconsejo, J. U. "Academic Engagement and Achievement Predict Career Adaptability." *The Career Development Quarterly*, vol. 69, 2021, pp. 34–48. <https://doi.org/10.1002/cdq.12247>.
- [3] Elsa, D. A., and Hayati, I. "Pengaruh Penggunaan Pembayaran Digital terhadap Minat Mahasiswa Menggunakan Quick Response Code Indonesian Standard (QRIS) pada Program Studi Manajemen Bisnis Syariah UMSU." *Journal of Economic, Business and Accounting (COSTING)*, vol. 7, no. 3, 2024, pp. 5031–5037. <https://doi.org/10.31539/costing.v7i3.9274>.
- [4] Rozi, K., Muhsi, M., Anwari, A., and Tamam, M. B. "Klasifikasi Penyakit Tanaman Padi Menggunakan Convolutional Neural Network (CNN) dan Random Forest." *Jurnal Mnemonic*, vol. 7, no. 2, 2024, pp. 241–248. <https://doi.org/10.36040/mnemonic.v7i2.10187>.
- [5] Oon Wira Yuda, Darmawan Tuti, Lim Sheih Yee, and Susanti. "Penerapan Data Mining untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Random Forest." *SATIN - Sains dan Teknologi Informasi*, vol. 8, no. 2, 2022, pp. 122–131. <https://doi.org/10.33372/stn.v8i2.885>.
- [6] Jawad, K., Shah, M. A., and Tahir, M. "Students' Academic Performance and Engagement Prediction in a Virtual Learning Environment Using Random Forest with Data Balancing." *Sustainability*, vol. 14, no. 22, 2022, article 14795. <https://doi.org/10.3390/su142214795>.
- [7] Sulistiyawati, A., and Supriyanto, E. "Implementasi Algoritma K-Means Clustering dalam Penentuan Siswa Kelas Unggulan." *Jurnal Tekno Kompak*, vol. 15, no. 2, 2021, pp. 25–36. <https://doi.org/10.33365/jtk.v15i2.1162>.
- [8] Sheykhmousa, Mohammadreza, et al. "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020): 6308–6325.

DOI:10.1109/JSTARS.2020.302
6724

- [9] Orange Data Mining. *Orange Documentation: Data Mining and Machine Learning Toolkit*. 2023. Orange Biolab, <https://orange.biolab.si>.
- [10] <https://www.kaggle.com/>