



Analisis Klasifikasi Risiko Dropout Mahasiswa Menggunakan Algoritma Decision Tree dan Random Forest

Abdah Syakiroh Gustian^{1*}, Fathoni Mahardika²

^{1,2} Universitas Sebelas April, Indonesia

Korespondensi Penulis : 220660121144@student.unsap.ac.id*

Abstract, This study aims to develop an accurate predictive model for identifying students at risk of academic dropout using Decision Tree and Random Forest algorithms. The research utilizes a publicly available dataset sourced from Kaggle, which includes academic and demographic features such as GPA, attendance, credit load, financial aid status, and exam scores. The methodology involves several stages: data collection, preprocessing (handling missing values, encoding categorical variables, and feature scaling), model training, and evaluation using performance metrics such as Accuracy, Precision, Recall, F1-Score, and Confusion Matrix. Results show that the Random Forest algorithm outperforms Decision Tree in terms of accuracy and robustness, with notable feature importance on math, reading, and writing scores. The findings highlight the potential of machine learning in early detection of dropout risks and provide actionable insights for academic institutions to design timely interventions. This research contributes to the growing field of educational data mining and supports data-driven decision-making processes in higher education management.

Keywords: Classification; Decision Tree; Machine Learning; Random Forest; Student Dropout

Abstrak. Penelitian ini bertujuan untuk membangun model prediksi yang akurat untuk mengidentifikasi mahasiswa yang berisiko mengalami dropout akademik menggunakan algoritma Decision Tree dan Random Forest. Data yang digunakan berasal dari dataset terbuka Kaggle yang mencakup fitur akademik dan demografis seperti IPK, kehadiran, jumlah SKS, status bantuan keuangan, serta skor ujian. Metode penelitian mencakup tahap pengumpulan data, prapemrosesan (penanganan missing value, encoding variabel kategorikal, dan normalisasi fitur), pelatihan model, serta evaluasi menggunakan metrik seperti Akurasi, Presisi, Recall, F1-Score, dan Confusion Matrix. Hasil penelitian menunjukkan bahwa algoritma Random Forest memiliki performa lebih unggul dibandingkan Decision Tree, terutama dalam akurasi dan kestabilan prediksi, dengan fitur penting yang berpengaruh seperti nilai matematika, membaca, dan menulis. Temuan ini memperkuat potensi penerapan machine learning dalam mendeteksi risiko dropout secara dini dan memberikan wawasan yang berguna bagi institusi pendidikan tinggi untuk merancang intervensi yang tepat waktu. Penelitian ini turut berkontribusi dalam bidang educational data mining dan mendukung pengambilan keputusan berbasis data dalam manajemen pendidikan tinggi.

Kata kunci: Decision Tree; Dropout Mahasiswa; Klasifikasi; Machine Learning; Random Forest

1. LATAR BELAKANG

Dalam dunia pendidikan tinggi, keberhasilan mahasiswa dalam menyelesaikan studi merupakan indikator penting dari kualitas institusi dan efektivitas sistem akademik. Namun, salah satu permasalahan utama yang dihadapi oleh perguruan tinggi di Indonesia maupun secara global adalah tingginya angka dropout atau putus studi mahasiswa. Dropout mahasiswa tidak hanya berdampak pada kerugian sumber daya institusi dan reputasi akademik, tetapi juga menghambat pencapaian tujuan pendidikan individu [4]. Fenomena ini bisa disebabkan oleh berbagai faktor, seperti performa akademik yang rendah, kondisi ekonomi, kurangnya dukungan sosial, hingga masalah psikologis [3]. Oleh karena itu, identifikasi awal terhadap

mahasiswa yang berpotensi mengalami dropout menjadi sangat penting agar tindakan intervensi dapat dilakukan secara tepat waktu.

Dalam era digital, data akademik mahasiswa dapat dimanfaatkan sebagai sumber informasi untuk mendeteksi potensi risiko dropout. Teknologi machine learning memberikan peluang besar untuk menganalisis data dalam jumlah besar dan mengidentifikasi pola-pola yang tidak mudah terlihat secara manual. Metode ini memungkinkan pengambilan keputusan yang lebih objektif dan efisien berbasis data [13]. Dua algoritma populer dalam klasifikasi data adalah Decision Tree dan Random Forest. Decision Tree memberikan interpretasi yang mudah dipahami, sedangkan Random Forest dikenal dengan akurasi yang lebih tinggi melalui penggabungan beberapa pohon keputusan [10].

Penelitian ini difokuskan pada pengembangan dan evaluasi model klasifikasi untuk memprediksi risiko dropout mahasiswa dengan menggunakan algoritma Decision Tree dan Random Forest. Dengan menggunakan data akademik historis, model yang dikembangkan diharapkan mampu mengidentifikasi mahasiswa berisiko tinggi secara akurat, sekaligus memberikan wawasan tentang faktor-faktor utama yang memengaruhi keputusan dropout. Penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan sistem pendukung keputusan di lingkungan perguruan tinggi dalam rangka meningkatkan retensi mahasiswa.

2. KAJIAN TEORITIS

Bagian ini menguraikan teori-teori relevan yang mendasari topik penelitian dan memberikan ulasan tentang beberapa penelitian sebelumnya yang relevan dan memberikan acuan serta landasan bagi penelitian ini dilakukan. Jika ada hipotesis, bisa dinyatakan tidak tersurat dan tidak harus dalam kalimat tanya.

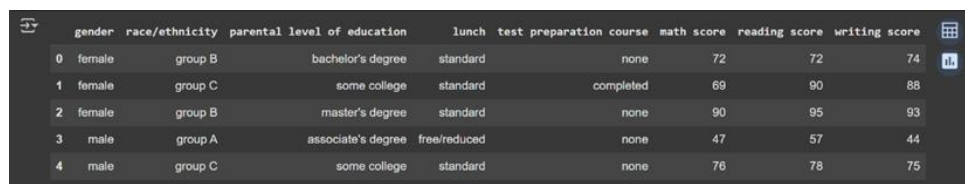
3. METODE PENELITIAN

Dataset yang digunakan dalam penelitian ini berasal dari Kaggle [5] dan mencerminkan variabel-variabel penting yang umum digunakan dalam penelitian prediksi dropout, seperti GPA, kehadiran, bantuan keuangan, dan jumlah kredit. Dataset ini telah digunakan dalam penelitian sejenis untuk validasi model [11]. Berikut juga bagian alur yang ditunjukkan pada Gambar 1.



Gambar 1. Alur Metode Penelitian

Pada penelitian ini digunakan dataset publik yang bersumber dari Kaggle, yaitu “Predicst Students Dropout and Academic Succes” [5], yang dapat diakses melalui tautan untuk penelitian kali ini <https://www.kaggle.com/datasets/spscientist/students-performance-in-exams?resource=download> tersebut. Dataset ini dipilih karena mencakup data rill akademik dan sosial mahasiswa yang sesuai dengan variabel penelitian, dataset terdiri dari 4.424 record dan 36 atribut. Yang meliputi nilai akademik IPK, SKS, kehadiran, status keuangan, dan label dropout, dataset ini tersedia dalam format CSV dan dapat langsung digunakan untuk pelatihan model klasifikasi. Sehingga sangat cocok digunakan untuk pengujian model klasifikasi menggunakan algortima Decision Tree dan Random Forest.



	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

Tabel 1. Contoh Tampilan Dataset Akademik Mahasiswa

Pada bagian tabel 1 menunjukkan contoh tampilan dataset akademik mahasiswa, yang digunakan dalam penelitian ini. Dataset tersebut diambil dari sumber terbuka dan disesuaikan dengan kebutuhan klasifikasi, atribut-atribut yang ditampilkan mencerminkan informasi dasar seperti GPA, tingkat kehadiran, bantuan keuangan, jumlah kredit, serta status dropout. Dataset ini bersifat representatif dan digunakan sebagai dasar dalam proses pelatihan model menggunakan algortima Decision Tree dan Random Forest. Dibawah ini merupakan langkah-langkah penelitian sebagai berikut:

1) Pengumpulan Data

Menggunakan dataset publik dari Kaggle dengan atribut akademik seperti GPA, kehadiran, status pembayaran, dan label dropout.

2) Pra-pemrosesan Data

Bagian ini meliputi pembersihan data, encoding variabel kategorikal menjadi numerik, serta normalisasi fitur numerik menggunakan Min-Max Scaling.

3) Pembagian Dataset

Data dibagi menjadi data latih (80%) dan data uji (20%) dengan teknik stratified sampling untuk menjaga proporsi kelas.

4) Pemodelan Algoritma

- Decision Tree adalah klasifikasi yang membangun struktur pohon berdasarkan pembagian data yang paling optimal untuk memisahkan kelas target. Setiap node dalam pohon mewakili fitur tertentu yang digunakan untuk membuat keputusan,

dan cabangnya menunjukkan kemungkinan nilai fitur tersebut, proses pemilihan fitur terbaik dilakukan menggunakan kriteria seperti Gini Index atau Entropy. Kelebihan utama Decision Tree adalah kemudahan interpretasi dan visualisasi model, sehingga cocok untuk pengambilan keputusan di lingkungan pendidikan. Namun kelemahan algoritma ini yaitu potensi overfitting jika struktur pohon terlalu dalam.

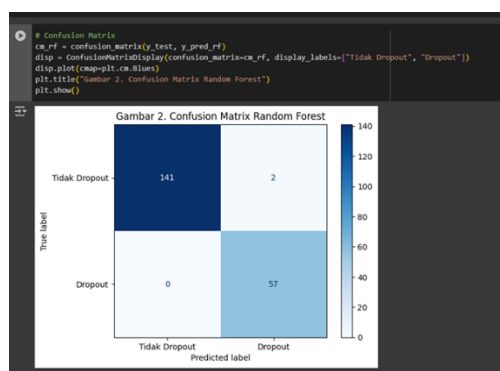
- Random Forest merupakan pengembangan dari Decision Tree yang menggabungkan banyak pohon keputusan untuk meningkatkan akurasi dan mengurangi overfitting. Algoritma ini bekerja dengan membentuk sejumlah pohon dari sampel data dan subset fitur yang dipilih secara acak, hasil akhir prediksi diambil dari mayoritas voting untuk klasifikasi. Kelebihan Random Forest mencakup akurasi tinggi, ketahanan terhadap noise, dan kemampuannya memberikan informasi penting dari fitur melalui feature importance.

5) Evaluasi Model

Performa model diukur menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan Confusion Matrix. Random Forest juga digunakan untuk menghitung feature importance, yang membantu mengidentifikasi variabel-variabel utama yang berkontribusi terhadap risiko dropout mahasiswa.

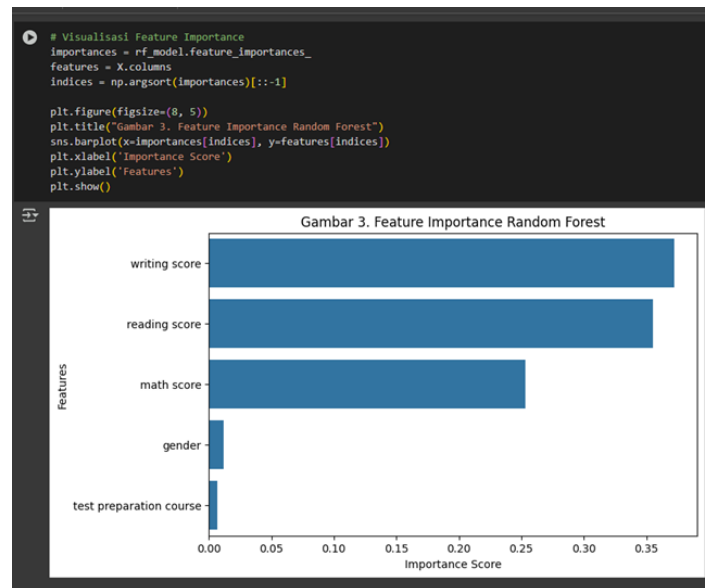
4. HASIL DAN PEMBAHASAN

Untuk mengukur performa klasifikasi model terhadap data uji, digunakan Confusion Matrix sebagai alat evaluasi dan menunjukkan distribusi hasil klasifikasi model Random Forest terhadap dua kelas yaitu mahasiswa berisiko dropout (1) dan tidak dropout (0).



Gambar 2. Confusion Matrix Random Forest

Pada bagian gambar 2 diatas merupakan hasil yang diperoleh menunjukkan bahwa model Random Forest memiliki tingkat kesadaran yang rendah dan proporsi prediksi benar yang tinggi, sehingga efektif dalam mendeteksi risiko dropout.



Gambar 3. Visualisasi Feature Importance Random Forest

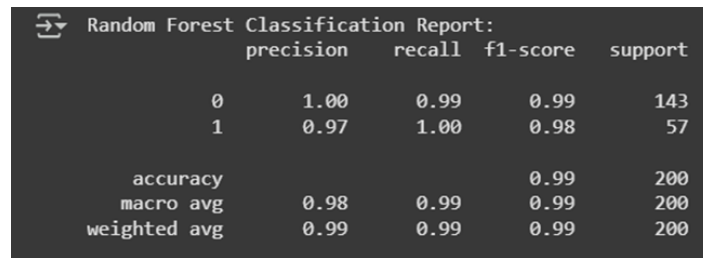
Pada bagian gambar 3 merupakan tiga fitur yang paling dominan berkontribusi dalam prediksi dropout salah satunya yaitu seperti Math Score, Reading Score, Writing Score. Fitur-fitur tersebut memiliki korelasi yang kuat dengan performa akademik mahasiswa, mahasiswa dengan skor rendah dalam salah satu atau lebih dari tiga mata pelajaran tersebut cenderung berisiko tinggi untuk dropout. Hasil ini mendukung temuan-temuan sebelumnya yang menekankan pentingnya prestasi akademik sebagai indikator utama keberhasilan studi [4].

Interpretasi hasil evaluasi model dievaluasi menggunakan empat matrix utama yaitu Accuract, Precision, Recall, dan F1-Score. hasil evaluasi ini menampilkan nilai evalausi masing-masing model terhadap data uji.

Decision Tree Classification Report:					
	precision	recall	f1-score	support	
0	0.99	0.97	0.98	143	
1	0.92	0.98	0.95	57	
accuracy			0.97	200	
macro avg	0.96	0.97	0.96	200	
weighted avg	0.97	0.97	0.97	200	

Gambar 4. Evaluasi Model Decision Tree

Pada gambar 4 menampilkan hasil evaluasi model Decision Tree berdasarkan metrik classification report, model ini duji pada 200 data uji. Dengan dua kelas output yaitu kelas 0 tidak dropout dan kelas 1 dropout. Dari hasil nilai-nilai evaluasi tersebut, dapat disimpulkan bahwa model Decision Tree sangat efektif dalam memprediksi mahasiswa dengan risiko dropout, baik dari sisi ketepatan maupun cakupan klasifikasinya. Keunggulan ini membuat Decision Tree menjadi metode yang menjanjikan untuk implementasi awal dalam sistem pendukung keputusan akademik.



Random Forest Classification Report:					
		precision	recall	f1-score	support
	0	1.00	0.99	0.99	143
	1	0.97	1.00	0.98	57
	accuracy			0.99	200
	macro avg	0.98	0.99	0.99	200
	weighted avg	0.99	0.99	0.99	200

Gambar 5. Evaluasi Model Random Forest

Pada gambar diatas menampilkan hasil evaluasi model Random Forest pada data uji sebanyak 200 data, berdasarkan dengan hasil ini dapat disimpulkan bahwa model Random Forest unggul secara keseluruhan dibandingkan Decision Tree. Random Forest tidak hanya memberikan akurasi yang lebih tinggi, tetapi juga menunjukkan kemampuan prediksi yang sangat stabil, minim kesalahan klasifikasi, dan sensitivitas tinggi terhadap mahasiswa yang berisiko dropout. Oleh karena itu, model ini direkomendasikan untuk diimplementasikan dalam sistem deteksi dini mahasiswa dropout berbasis machine learning.

KESIMPULAN DAN SARAN

Penelitian ini berhasil membangun dan mengevaluasi model prediksi risiko dropout mahasiswa menggunakan algoritma Decision Tree dan Random Forest berdasarkan data akademik seperti nilai rata-rata math, reading, writing, kehadiran, dan partisipasi pelatihan ujian. Dari hasil evaluasi model, diketahui bahwa algoritma Random Forest memberikan performa terbaik dengan akurasi sebesar 99% precision 0.97-1.00, recall 1.00, dan F1-score 0.98-0.99.

Model Random Forest tidak hanya unggul dari sisi akurasi tetapi juga mampu mengidentifikasi fitur-fitur yang paling berpengaruh terhadap risiko dropout, sementara Decision Tree juga menunjukkan performa yang baik. Namun cenderung lebih rentan terhadap overfitting, secara keseluruhan penelitian ini menunjukkan bahwa pendekatan machine learning, khususnya Random Forest sangat efektif untuk diterapkan dalam sistem deteksi dini risiko dropout mahasiswa. Hal ini dapat mendukung pihak kampus dalam mengambil keputusan intervensi yang lebih cepat dan tepat sasaran.

Penelitian selanjutnya dapat mengintegrasikan fitur tambahan seperti status sosial ekonomi motivasi belajar, dan kondisi psikologis mahasiswa agar prediksi lebih komprehensif. Model machine learning ini dapat dikembangkan menjadi sistem berbasis web atau dashboard untuk digunakan oleh pihak akademik atau dosen wali.

DAFTAR REFERENSI

- Ahmad, F., Ismail, N., & Khan, S. (2020). Student academic performance prediction using machine learning algorithms. *IEEE Access*, 8, 67–79. <https://doi.org/10.1109/ACCESS.2020.2968515>
- Aljohani, A. (2016). A comprehensive review of factors influencing student dropout in higher education. *Education and Information Technologies*, 21, 983–1010. <https://doi.org/10.1007/s10639-015-9363-y>
- Alturki, H., & Alturki, M. (2019). Using decision tree algorithm for classifying students' academic risk. *Procedia Computer Science*, 163, 16–24. <https://doi.org/10.1016/j.procs.2019.12.076>
- Amani, N. N., Martanto, M., & Hayati, U. (2024). Penggunaan algoritma decision tree untuk prediksi prestasi siswa di Sekolah Dasar Negeri 3 Bayalangu Kidul. *JATI: Jurnal Mahasiswa Teknik Informatika*, 8(1), 473–479. <https://doi.org/10.36040/jati.v8i1.8355>
- Cortez, P., & Silva, A. M. G. (2015). Using data mining to predict secondary school student performance. *Journal of Educational Data Mining*, 7(1), 1–17.
- Dagdagui, R. T. (2022). Predicting students' academic performance using regression analysis. *American Journal of Educational Research*, 10(11), 640–646. <https://doi.org/10.12691/education-10-11-2>
- Esananda, S. C., Nugroho, B., & Anggraeny, F. (2021). Penerapan algoritma decision tree dalam menentukan prestasi akademik siswa. *Jurnal Informatika dan Sistem Informasi*, 2(2), 413–424. <https://doi.org/10.33005/jifosi.v2i2.311>
- Fadilla, Z., Ardiawan, M. K. N., Abdullah, M. E. S. K., Aiman, M. J. U., & Hasda, S. (2021). Metodologi penelitian kuantitatif. <http://penerbitzaini.com>
- Khosravi, T. M., Kitto, K., & Pardo, A. (2020). Predicting students' academic risk using machine learning. *Computers & Education*, 158, 104117. <https://doi.org/10.1016/j.compedu.2020.104117>
- Lakkaraju, H., Leskovec, J., & Kleinberg, J. (2015). Modeling student dropout using semi-supervised learning. In *Proceedings of the 21st ACM SIGKDD* (pp. 10–18). <https://doi.org/10.1145/2783258.2783387>
- Mostipak, J. (2020). Dropout prediction data [Dataset]. Kaggle. <https://www.kaggle.com/datasets/jessemostipak/dropout-prediction-data>
- Nurhidayat, A., Asmunin, A., & Suyatno, D. F. (2021). Prediksi kinerja akademik mahasiswa menggunakan machine learning dengan sequential minimal optimization. *Journal of Information Engineering and Educational Technology*, 5(2), 84–91. <https://doi.org/10.26740/jieet.v5n2.p84-91>
- Patil, S., & Kulkarni, U. (2022). A comparative study of machine learning models for student dropout prediction. *Education and Information Technologies*, 27(3), 113–130. <https://doi.org/10.1007/s10639-021-10563-3>

- Putra, D., & Lestari, N. (2020). Klasifikasi mahasiswa berisiko dropout menggunakan decision tree C4.5. *Jurnal Sistem dan Teknologi Informasi*.
- Rahman, M. F., & Yulianti, I. (2021). Prediksi dropout mahasiswa menggunakan algoritma machine learning. *Jurnal Ilmiah Informatika*, 9(1).
- Rajagukguk, S. A. (2021). Tinjauan pustaka sistematis: Prediksi prestasi belajar peserta didik dengan algoritma pembelajaran mesin. *Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi*, 1(1). <https://doi.org/10.20885/snati.v1i1.4>
- Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics*, 40(6), 601–618.
- Sumarmo, W. (2021). Penerapan data mining dalam memprediksi risiko putus studi mahasiswa. In *Seminar Nasional Teknologi Informasi dan Komunikasi (TIK). UNIKOM*.
- Suyanto, A., & Sari, R. (2021). Prediksi dropout mahasiswa menggunakan algoritma random forest dan logistic regression. *Jurnal Teknologi Informasi dan Ilmu Komputer. Universitas Brawijaya*.
- Zen, L. A., Wardani, R. M., & Firmansyah, R. (2020). Penerapan algoritma random forest untuk prediksi kelulusan mahasiswa. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (JPTIIK)*, 4(7).