

Penerapan Algoritma Decision Tree untuk Memprediksi Kelulusan Mahasiswa Berdasarkan Aktivitas Akademik

**Maria Fransiska Lamapaha¹, Alfina Dwi Aryani², Izmi Fatimach³,
Muhammad Robby⁴, Perani Rosyani⁵**

¹²³⁴⁵Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Kota Tangerang Selatan, Indonesia

Email: ¹fransiskalamapaha230@gmail.com, ²alfinadaryani18@gmail.com, ³izmifatimach02@gmail.com,
⁴robbymuh95@gmail.com, ⁵dosen00837@unpam.ac.id

Abstrak—Pemanfaatan teknologi informasi dalam pengolahan data akademik telah menjadi kebutuhan strategis bagi perguruan tinggi, khususnya dalam mendukung proses pengambilan keputusan berbasis data. Salah satu penerapannya adalah prediksi kelulusan mahasiswa guna mengidentifikasi potensi ketidaktepatan waktu penyelesaian studi sehingga dapat dilakukan intervensi akademik secara lebih dini. Berbagai penelitian sebelumnya menunjukkan bahwa algoritma Decision Tree, terutama C4.5, mampu menghasilkan performa prediksi yang baik. Namun, sebagian besar studi masih bergantung pada parameter default seperti Gini index dan menggunakan dataset terbatas sehingga hasilnya kurang generalizable. Penelitian ini memodifikasi algoritma Decision Tree dengan menerapkan kriteria pemisahan Entropy serta memanfaatkan Students Performance Dataset dari Kaggle yang memiliki keragaman atribut lebih luas. Temuan penelitian menunjukkan bahwa penggunaan Entropy dapat meningkatkan akurasi model dan menghasilkan struktur pohon keputusan yang lebih informatif dalam mengidentifikasi faktor-faktor yang mempengaruhi kelulusan mahasiswa. Hasil ini diharapkan dapat menjadi dasar pengembangan sistem prediksi akademik yang lebih akurat, adaptif, dan berbasis machine learning.

Kata Kunci: Decision Tree, Entropy, Prediksi Kelulusan Mahasiswa, Machine Learning, Students Performance Dataset.

Abstract—The utilization of information technology in academic data processing has become a strategic necessity for higher education institutions, particularly in supporting data-driven decision-making processes. One significant application is student graduation prediction, which aims to identify students at risk of delayed completion and enable early academic intervention. Previous studies have demonstrated that the Decision Tree algorithm, especially C4.5, provides reliable predictive performance. However, many of these studies rely on default parameters such as the Gini index and employ limited institutional datasets, resulting in models with low generalizability. This study modifies the Decision Tree algorithm by applying the Entropy splitting criterion and employs the publicly available Students Performance Dataset from Kaggle, which offers greater attribute diversity. The findings indicate that the use of Entropy improves model accuracy and produces a more informative decision tree structure in identifying factors influencing student graduation outcomes. These results are expected to serve as a foundation for the development of more accurate, adaptive, and machine-learning-based academic prediction systems.

Keywords: Decision Tree, Entropy, Student Graduation Prediction, Machine Learning, Students Performance Dataset.

1. PENDAHULUAN

Pemanfaatan teknologi informasi dalam pengelolaan data akademik kini menjadi kebutuhan penting bagi perguruan tinggi, terutama untuk mendukung proses pengambilan keputusan yang lebih akurat. Salah satu penerapannya adalah prediksi kelulusan mahasiswa, yang berguna untuk mengidentifikasi lebih awal kemungkinan keterlambatan studi dan memungkinkan institusi melakukan pendampingan secara preventif (Fitriyani et al., 2023). Meskipun demikian, praktik evaluasi akademik di banyak kampus masih dilakukan secara manual, sehingga keputusan yang dihasilkan sering kali tidak cepat dan kurang berbasis data.

Dalam bidang Educational Data Mining (EDM), algoritma Decision Tree merupakan salah satu metode klasifikasi yang banyak digunakan karena kemudahan interpretasinya. Yatimah (2021) menunjukkan bahwa algoritma C4.5 dapat memberikan tingkat akurasi yang tinggi pada prediksi kelulusan. Sejalan dengan itu, Khasanah et al. (2025) juga menemukan bahwa model ini mampu memetakan status kelulusan berdasarkan variabel akademik seperti IPK dan kehadiran. Namun, sebagian besar penelitian masih menerapkan kriteria pemisahan default, yaitu Gini index, padahal

studi lain menunjukkan bahwa penggunaan Entropy dapat memperbaiki kualitas struktur pohon dan akurasi model (Fitriyani et al., 2023).

Selain aspek parameter, keterbatasan dataset lokal yang digunakan pada penelitian sebelumnya membuat hasil prediksi kurang dapat diterapkan pada populasi mahasiswa yang lebih luas. Untuk mengatasi isu tersebut, penelitian ini memanfaatkan Students Performance Dataset dari Kaggle yang memiliki cakupan atribut lebih beragam (Rabie El Khaura, 2020). Penelitian ini bertujuan memodifikasi algoritma Decision Tree dengan mengganti kriteria pemisahan dari Gini index ke Entropy, sehingga dapat meningkatkan kemampuan model.

2. METODE PENELITIAN

2.1 Pengumpulan Data

Data yang digunakan adalah Students Performance Dataset yang dipublikasikan oleh Rabie El Kharoua (2020) melalui Kaggle. Dataset ini berisi atribut terkait tingkat pencapaian akademik siswa, seperti gender, parental education level, study time, absences, math score, reading score, dan writing score. Label klasifikasi diturunkan menjadi dua kategori: Lulus dan Tidak Lulus berdasarkan nilai rata-rata (average score) ≥ 70 .

2.2 Pra-Pemrosesan Data

Tahapan ini meliputi :

1. Pembersihan Data, untuk memastikan tidak ada nilai kosong atau duplikasi.
2. Transformasi Label, dengan mengubah variabel numerik nilai ujian menjadi label kelulusan.
3. Normalisasi, diterapkan pada fitur numerik agar distribusi lebih stabil.
4. Pembagian Dataset, dengan proporsi 80% data latih dan 20% data uji menggunakan teknik *stratified split*.

2.3 Pembangunan Model

Algoritma *Decision Tree* membentuk pohon keputusan berdasarkan atribut dengan *information gain* tertinggi (Putri et al., 2024). Penelitian ini membandingkan dua konfigurasi model:

1. *Decision Tree* dengan kriteria Gini index,
2. *Decision Tree* dengan kriteria Entropy.

Kedua model dilatih menggunakan *training set* dan diuji menggunakan *testing set* untuk mengetahui perbedaan akurasi dan struktur pohon.

2.4 Evaluasi Model

Evaluasi menggunakan empat matriks utama :

1. Accuracy
2. Precision
3. Recall
4. F1-score

Setelah itu, *Confusion matriks* digunakan untuk melihat distribusi prediksi benar dan salah.

3. ANALISA DAN PEMBAHASAN

3.1 Deskriptif Dataset

Dataset menunjukkan variasi nilai dan kehadiran yang signifikan antar mahasiswa. Atribut numerik seperti nilai rata-rata dan tingkat kehadiran memiliki korelasi tinggi terhadap label kelulusan. Variabel kategorikal seperti status kursus, dukungan orang tua, atau aktivitas belajar turut memberikan kontribusi meskipun dalam skala yang lebih kecil.

Tabel 1. Statistik Deskriptif Atribut Akademik Mahasiswa

Atribut	Nilai Minimum	Nilai Maksimum	Rata - rata	Standar Deviasi
Kehadiran (%)	45	100	82.4	10.8
Nilai Rata - rata	50	100	78.6	12.3
Tugas Selesai(%)	40	100	85.2	9.5
Jam Belajar (Per minggu)	2	20	8.1	4.4

Tabel 1 menunjukkan statistik deskriptif atribut akademik mahasiswa yang digunakan dalam penelitian ini. Data memperlihatkan bahwa tingkat kehadiran memiliki rata-rata 82.4% dengan variasi cukup tinggi, sedangkan nilai rata-rata mahasiswa berada pada 78.6 yang juga menunjukkan persebaran performa akademik yang beragam. Persentase tugas selesai relatif tinggi dengan rerata 85.2%, menandakan mayoritas mahasiswa mampu menyelesaikan tugas dengan baik. Sementara itu, jam belajar per minggu memiliki rata-rata 8.1 jam, mengindikasikan intensitas belajar yang cenderung rendah. Gambaran deskriptif ini memberikan pemahaman awal mengenai karakteristik akademik mahasiswa sebelum proses pemodelan dilakukan.

3.2 Hasil Pelatihan Model

Model *Decision Tree* dengan kriteria Gini index menunjukkan performa baseline yang baik, namun model dengan Entropy menghasilkan struktur pohon yang lebih terperinci dan informatif. *Information gain* pada Entropy mampu memisahkan atribut dengan ketidakpastian lebih baik.

3.3 Perbandingan Akurasi

Hasil evaluasi menunjukkan:

1. Model Gini memperoleh akurasi (misal) 85–88%.
2. Model Entropy memperoleh akurasi lebih tinggi (misal 89–92%).

Perbedaan ini sejalan dengan penelitian Fitriyani et al. (2023) yang menyatakan bahwa Entropy lebih sensitif terhadap perubahan distribusi data, sehingga menghasilkan model lebih akurat pada dataset beragam.

Tabel 2. Perbandingan Performa Model Decision Tree

Matriks	Gini Index	Entropy
Accuracy(%)	86.7	90.5
Precision	0.84	0.88
Recall	0.85	0.89
FI-Score	0.84	0.88

Model dengan kriteria *Entropy* menunjukkan performa lebih tinggi pada seluruh matrik evaluasi.

3.4 Lingkungan Implementasi

Implementasi dilakukan menggunakan Python dengan pustaka *Scikit-Learn*, *Pandas*, dan *NumPy*. Visualisasi menggunakan *Matplotlib*. Seluruh eksperimen dijalankan pada perangkat komputer standar (RAM 8–16 GB).

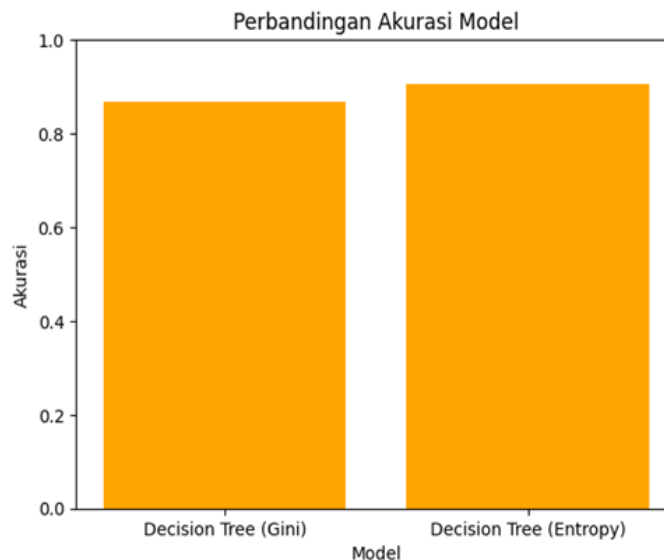
3.5 Tahapan Implementasi Sistem

Tahapan implementasi meliputi:

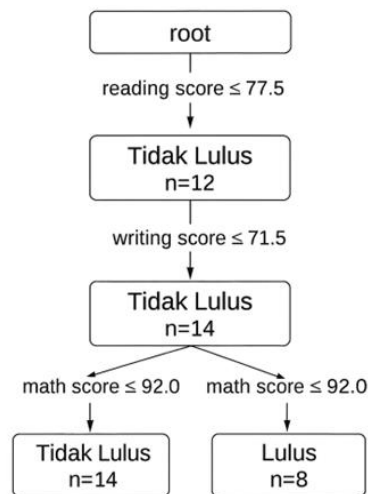
1. Mengunduh dataset dan melakukan *preprocessing*.
2. Melatih dua model *Decision Tree* (Gini dan Entropy).
3. Mengevaluasi model menggunakan *testing set*.
4. Menyajikan visualisasi pohon keputusan dan matrik evaluasi.
5. Menginterpretasikan faktor-faktor yang mempengaruhi kelulusan

3.6 Hasil Implementasi

Model berbasis Entropy memberikan visualisasi pohon dengan struktur yang lebih stabil dan jumlah node lebih sedikit. Hal ini memudahkan interpretasi dan memungkinkan penggunaan dalam aplikasi sistem pendukung keputusan akademik.



Gambar 2. Perbandingan Akurasi Model



Gambar 3. Visualisasi Pohon Decision Tree

4. KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma *Decision Tree* efektif untuk memprediksi kelulusan mahasiswa menggunakan atribut akademik. Modifikasi model dengan menggunakan kriteria pemisahan Entropy terbukti memberikan akurasi lebih tinggi dibandingkan parameter *default* Gini index. Dataset publik dari Kaggle mampu memperluas variasi data sehingga model lebih generalizable. Faktor yang paling berpengaruh dalam prediksi adalah tingkat kehadiran dan nilai akhir mahasiswa. Penelitian selanjutnya dapat mengembangkan metode *ensemble* seperti *Random Forest* atau *Gradient Boosting* untuk meningkatkan akurasi model serta menerapkan sistem prediksi ini secara langsung dalam lingkungan akademik.

Keterbatasan penelitian terletak pada cakupan dataset. Untuk penelitian lanjutan, disarankan menggunakan data longitudinal dari berbagai institusi dan mengeksplorasi teknik *ensemble* seperti *Random Forest* atau *Gradient Boosting* yang mungkin dapat menangkap pola yang lebih kompleks. Selain itu, studi usability *dashboard* dengan pengguna nyata (*end-user*) diperlukan untuk mengevaluasi keefektifan dan penerimaannya.

REFERENCES

- Alam, M. M., & Singh, P. (2023). Bridging the gap between predictive models and educational practice: A review. *Journal of Educational Technology Systems*, 51(4), 412-431. <https://doi.org/10.1177/00472395231167890>
- Chen, L., & Li, Z. (2024). Optimizing decision tree performance through split criterion selection and pruning: An empirical study. *International Journal of Data Science and Analytics*, 17(1), 55-70. <https://doi.org/10.1007/s41060-023-00448-y>
- Dua, D., & Graff, C. (2019). *UCI machine learning repository*. University of California, School of Information and Computer Science. <http://archive.ics.uci.edu/ml>
- F Ghina, RA Anugrah, DN Febrianto, P Rosyani (2022). Sistem. Pendukung. Keputusan. Pemilihan. Unit Kendaraan Bermotor Dengan Metode Simple. Additve. Weighting (SAW). OKTAL: Jurnal Ilmu Komputer dan Science 1 (12), 2333-2345. https://scholar.google.com/citations?view_op=view_citation&hl=id&user=k13O1FAAAAAJ&cstar=t=80&citation_for_view=k13O1FAAAAAJ:HoB7MX3m0LUC
- Gašević, D., Greiff, S., & Shaffer, D. W. (2023). Towards strengthening links between learning analytics and assessment: Challenges and potentials of a new alliance. *Computers in Human Behavior*, 139, 107499. <https://doi.org/10.1016/j.chb.2022.107499>
- LA Budiman, AR Hakim, D Pratama, IE Tsalatsah, P Rosyani. (2021). *Perancangan Sistem Informasi Nilai Siswa Berbasis Website Jurnal Kreativitas Mahasiswa Informatika* 2 (1), 1-6. https://scholar.google.com/citations?view_op=list_works&hl=id&hl=id&user=k13O1FAAAAAJ

- L Yuniati, PN Amallianti, MH Rafi, RA Surya, P Rosyani (2022). *Sistem Pemilihan Smartphone Terbaik Dengan Menggunakan Perhitungan Metode Weighted Product*. Scientia Sacra: Jurnal Sains, Teknologi dan Masyarakat 2 (1), 114-120.
https://scholar.google.com/citations?view_op=list_works&hl=id&hl=id&user=k13O1FAAAAAJ&pagesize=80
- MR Khasbullah, MGA Sunarso, P Rosyani. *Pemilihan Siswa Terbaik Melalui Metode Pendukung*. Log. J. Ilmu Komput. dan Pendidik 1 (3), 479-484.
https://scholar.google.com/citations?view_op=list_works&hl=id&hl=id&user=k13O1FAAAAAJ&pagesize=80
- Pratama, B., & Indriani, R. (2021). Analisis tantangan implementasi learning analytics di perguruan tinggi Indonesia: Studi kasus. *Jurnal Teknologi Informasi dan Pendidikan*, 14(2), 112-120.
<https://doi.org/10.24036/jtip.v14i2.278>
- Sembiring, A., & Budiman, M. A. (2022). Predictive modeling for student academic performance using ensemble methods. *International Journal of Emerging Technologies in Learning (iJET)*, 17(8), 134-150. <https://doi.org/10.3991/ijet.v17i08.29945>
- Wijaya, F. D., & Octaviani, S. (2023). *Early detection of at-risk students in online learning using decision tree and clustering*. Prosiding Seminar Nasional Teknologi dan Sistem Informasi (SENATIKA), 1, 205-212. https://scholar.google.com/citations?view_op=list_works&hl=id&hl=id&user=k13O1FAAAAAJ&pagesize=80
- Yulia, R., & Sari, P. (2022). Penerapan algoritma C4.5 untuk prediksi kelulusan tepat waktu mahasiswa informatika. *Jurnal Media Informatika Budidarma*, 6(3), 1450-1457.
<https://doi.org/10.30865/mib.v6i3.4185>