e-ISSN : 2964-5115
p-ISSN : 2964-4364

Implementasi Algoritma C4.5 Berbasis Feature Selection untuk Prediksi Kelulusan Mahasiswa Tepat Waktu

Rama Adistya Nurtjahya Pamudji[*1], Kartika Purnamasari [2]
[*1]*Teknik Informatika, STMIK Pranata Indonesia, Bekasi*
[3]*Depriwangga, Jakarta*
e-mail: [*1] ramaadistyanurcahya@gmail.com, [2] kartika@depriwangga.com

Abstrak

Kelulusan mahasiswa tepat waktu merupakan salah satu indikator penting dalam menilai kualitas pendidikan di perguruan tinggi. Penelitian ini bertujuan untuk membangun model prediksi kelulusan mahasiswa menggunakan algoritma Decision Tree C4.5 yang dikombinasikan dengan teknik feature selection. Dataset yang digunakan adalah Student Dropout and Academic Success yang terdiri dari atribut akademik, sosial, dan demografis mahasiswa. Tahapan penelitian meliputi preprocessing data, seleksi fitur menggunakan metode statistik, pembangunan model klasifikasi, serta evaluasi menggunakan confusion matrix, accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa penerapan feature selection mampu meningkatkan performa model dibandingkan tanpa seleksi fitur. Model yang dihasilkan mencapai akurasi sekitar 82%, dengan fitur dominan berasal dari performa akademik semester awal. Selain itu, penelitian ini juga menunjukkan pentingnya penggunaan metrik evaluasi yang lebih komprehensif untuk menghindari bias akibat ketidakseimbangan data. Model yang dihasilkan diharapkan dapat digunakan sebagai sistem pendukung keputusan dalam mengidentifikasi mahasiswa yang berisiko tidak lulus tepat waktu.

Article Info

***Kata Kunci**:*
Decision Tree C4.5, feature selection, prediksi kelulusan, data mining, klasifikasi

Abstract

Timely student graduation is an important indicator in assessing the quality of higher education institutions. This study aims to develop a prediction model for student graduation using the Decision Tree C4.5 algorithm combined with feature selection techniques. The dataset used is the Student Dropout and Academic Success dataset, which includes academic, social, and demographic attributes of students. The research process consists of data preprocessing, feature selection using statistical methods, classification model development, and evaluation using confusion matrix, accuracy, precision, recall, and F1-score metrics. The results show that the application of feature selection improves model performance compared to models without feature selection. The proposed model achieved an accuracy of approximately 82%, with dominant features derived from early semester academic performance. Furthermore, this study highlights the importance of using comprehensive evaluation metrics to avoid bias caused by imbalanced data. The resulting model can be utilized as a decision support tool to identify students at risk of not graduating on time.

***Keywords**:*
Decision Tree C4.5, feature selection, graduation prediction, data mining, classification

I. PENDAHULUAN

Kelulusan mahasiswa tepat waktu merupakan salah satu indikator penting dalam menilai kualitas penyelenggaraan pendidikan di perguruan tinggi. Tingkat kelulusan yang rendah atau keterlambatan penyelesaian studi tidak hanya berdampak pada mahasiswa secara individu, tetapi juga berpengaruh terhadap akreditasi dan reputasi institusi pendidikan. Dalam prakteknya, tidak semua mahasiswa mampu menyelesaikan studi sesuai waktu yang ditentukan, sehingga diperlukan upaya untuk mengidentifikasi potensi keterlambatan sejak awal (Hasibuan & Mahdiana, 2023).

e-ISSN : 2964-5115
p-ISSN : 2964-4364

Perkembangan teknologi informasi, khususnya dalam bidang *data mining* dan *machine learning*, memberikan peluang besar untuk membantu institusi pendidikan dalam melakukan prediksi terhadap kualitas akademik mahasiswa. Dengan memanfaatkan data historis mahasiswa seperti indeks prestasi, kehadiran, dan aktivitas akademik, model prediksi dapat dibangun untuk mengidentifikasi mahasiswa yang berisiko tidak lulus tepat waktu. Pendekatan ini dinilai efektif karena mampu menghasilkan pola dan hubungan tersembunyi dari data yang sebelumnya sulit dianalisis secara manual (Fatoni et al., 2026).

Salah satu algoritma yang banyak digunakan dalam kasus klasifikasi adalah Decision Tree C4.5. Algoritma ini memiliki keunggulan dalam menghasilkan model yang mudah dipahami melalui struktur pohon keputusan serta mampu menangani data dengan atribut numerik maupun kategorikal. Beberapa penelitian menunjukkan bahwa C4.5 mampu memberikan performa yang cukup baik dalam memprediksi kelulusan mahasiswa, dengan tingkat akurasi yang relatif tinggi serta kemampuan dalam mengidentifikasi faktor dominan seperti indeks prestasi kumulatif (IPK) (Fatoni et al., 2026).

Namun demikian, performa algoritma C4.5 sangat dipengaruhi oleh kualitas dan relevansi fitur yang digunakan. Data akademik mahasiswa umumnya memiliki banyak atribut, tidak semuanya berkontribusi secara signifikan terhadap proses klasifikasi. Keberadaan fitur yang tidak relevan atau redundan dapat menurunkan akurasi model serta meningkatkan kompleksitas komputasi. Oleh karena itu, diperlukan teknik feature selection untuk memilih atribut terbaik yang benar-benar berpengaruh terhadap prediksi (Yusup et al., 2020) .

Penerapan *feature selection* terbukti mampu meningkatkan kinerja model klasifikasi dengan cara mengurangi dimensi data dan menghilangkan *noise*. Selain itu, pemilihan fitur yang tepat juga dapat membantu dalam memahami faktor-faktor utama yang mempengaruhi kelulusan mahasiswa. Meskipun demikian, masih banyak penelitian yang menggunakan algoritma C4.5 tanpa mengoptimalkan proses seleksi fitur secara maksimal, sehingga potensi peningkatan performa model belum sepenuhnya dimanfaatkan.

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan implementasi algoritma C4.5 yang dikombinasikan dengan teknik feature selection untuk meningkatkan akurasi prediksi kelulusan mahasiswa tepat waktu. Dataset yang digunakan berasal dari platform Kaggle, yang menyediakan data akademik mahasiswa secara komprehensif. Dengan pendekatan ini, diharapkan model yang dihasilkan tidak hanya memiliki performa yang lebih baik, tetapi juga mampu memberikan insight yang lebih jelas mengenai faktor-faktor yang mempengaruhi keberhasilan studi mahasiswa.

II. METODE PENELITIAN

a. Dataset Penelitian

Penelitian ini menggunakan dataset Student Dropout and Academic Success yang diperoleh dari platform Kaggle. Dataset ini berisi data historis mahasiswa yang mencakup berbagai atribut akademik, demografis, dan sosial ekonomi yang berpotensi mempengaruhi keberhasilan studi. Variabel target dalam penelitian ini adalah status mahasiswa, yang diklasifikasikan menjadi kategori seperti graduate, dropout, dan enrolled. Untuk kepentingan penelitian ini, label tersebut dapat disederhanakan menjadi klasifikasi biner, yaitu lulus tepat waktu dan tidak lulus tepat waktu.

Dataset ini dipilih karena memiliki karakteristik yang kompleks dan realistis, sehingga sesuai untuk pengujian model prediksi berbasis machine learning dalam konteks pendidikan tinggi (Fatoni et al., 2026; Hasibuan & Mahdiana, 2023).

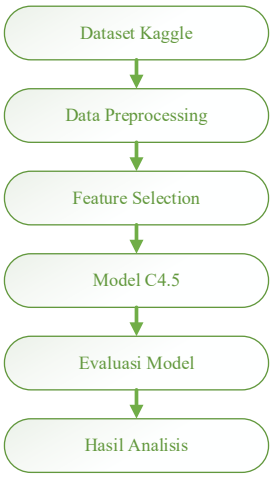
b. Tahapan Penelitian

Penelitian ini mengikuti tahapan umum dalam proses machine learning, yang meliputi beberapa langkah utama sebagai berikut:

- Pengumpulan Data, mengambil dataset dari Kaggle.
- Preprocessing Data meliputi: penanganan missing values, Encoding data kategorikal, Normalisasi (jika diperlukan)
- *Feature Selection,* memilih atribut yang paling relevan terhadap variabel target.
- Pembangunan Model, menggunakan algoritma Decision Tree C4.5.

e-ISSN : 2964-5115
p-ISSN : 2964-4364

- Evaluasi Model, menggunakan metrik klasifikasi seperti accuracy, precision, recall, dan F1-score.



Gambar 1. Desain Penelitian

c. Data Preprocessing

Tahap preprocessing bertujuan untuk meningkatkan kualitas data sebelum dilakukan proses pemodelan. Beberapa teknik yang digunakan antara lain:

- Data Cleaning: Menghapus atau mengisi nilai yang hilang
- Encoding: Mengubah data kategorikal menjadi numerik
- Normalisasi: Menyetarakan skala data (jika diperlukan)

Tahap ini penting karena kualitas data sangat mempengaruhi performa model yang dihasilkan (Gunawan et al., 2022).

$$Entropy(S) = -\sum_{i=1}^{n} p_i log_2 p_i$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

d. Algoritma Decision Tree C4.5

Algoritma C4.5 merupakan pengembangan dari ID3 yang digunakan untuk membangun model klasifikasi berbasis pohon keputusan. Algoritma ini bekerja dengan cara memilih atribut terbaik berdasarkan nilai Gain Ratio, kemudian membagi data secara rekursif hingga terbentuk struktur pohon keputusan.

$$GainRatio(A) = \frac{Gain(S,A)}{Split\ Info\ (A)}$$

Keunggulan C4.5 sendiri adalah dapat menangani data numerik dan kategorikal, memiliki mekanisme pruning, dan mudah diinterpretasikan. Algoritma ini banyak digunakan dalam prediksi kelulusan mahasiswa karena kemampuannya dalam menghasilkan aturan keputusan yang jelas (Anwar et al., 2022).

e. Evaluasi Model

Evaluasi model dilakukan untuk mengukur performa klasifikasi yang dihasilkan. Metrik yang digunakan meliputi:

- *Accuracy*: tingkat ketepatan prediksi
- *Precision*: ketepatan prediksi positif
- *Recall*: kemampuan mendeteksi kelas positif
- *F1-Score*: kombinasi precision dan recall

Confusion matrix digunakan untuk mengevaluasi kinerja model klasifikasi dengan cara membandingkan hasil prediksi model terhadap label aktual. Matriks ini terdiri dari empat komponen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN), yang masing-masing merepresentasikan jumlah prediksi benar dan salah untuk setiap kelas. Dengan melihat confusion matrix, peneliti tidak hanya mengetahui jumlah prediksi yang tepat, tetapi juga dapat memahami jenis kesalahan yang dilakukan model, misalnya apakah model cenderung salah dalam memprediksi mahasiswa yang sebenarnya tidak lulus tepat waktu sebagai lulus tepat waktu, atau sebaliknya. Oleh karena itu, confusion matrix sangat penting dalam analisis yang lebih mendalam, terutama pada kasus dengan distribusi data yang tidak seimbang.

Accuracy merupakan salah satu metrik evaluasi yang digunakan untuk mengukur tingkat ketepatan model dalam melakukan klasifikasi secara keseluruhan. Nilai accuracy diperoleh dari perbandingan antara jumlah prediksi yang benar dengan total seluruh data yang diuji, yang dirumuskan sebagai: jumlah TP dan TN dibagi dengan total TP, TN, FP, dan FN. Metrik ini memberikan gambaran umum seberapa baik model dalam mengklasifikasikan data, namun memiliki

e-ISSN : 2964-5115
p-ISSN : 2964-4364

keterbatasan ketika digunakan pada dataset yang tidak seimbang, karena nilai accuracy dapat terlihat tinggi meskipun model kurang baik dalam memprediksi salah satu kelas. Oleh karena itu, accuracy sebaiknya digunakan bersamaan dengan metrik lain seperti precision dan recall untuk mendapatkan evaluasi yang lebih komprehensif.

Evaluasi ini penting untuk memastikan bahwa model yang dibangun tidak hanya akurat, tetapi juga seimbang dalam mengklasifikasikan setiap kelas (Sahlaoui et al., 2024).

f. Tools untuk implementasi

Penelitian ini dibangun dengan menggunakan bantuan bahasa pemrograman Python. Library yang digunakan sebagai alat bantu adalah Scikit-learn, Pandas, NumPy. Sedangkan tools tambahan untuk IDE adalah Google Colab.

III. HASIL DAN PEMBAHASAN

a. Hasil Pengolahan Data

Penelitian ini menggunakan dataset *studentdataset.csv* yang berjumlah 4.424 baris data dengan 35 atribut, terdiri dari 34 atribut prediktor dan 1 atribut target, yaitu Target. Berdasarkan draft penelitian, target penelitian dirancang menjadi klasifikasi biner untuk membedakan mahasiswa yang lulus tepat waktu dan tidak lulus tepat waktu.

Sebelum pemodelan dilakukan, data melalui tahap preprocessing sebagaimana dirancang pada metode penelitian, yaitu pengecekan *missing value*, transformasi label target, pemisahan data latih dan data uji, kemudian seleksi fitur menggunakan SelectKBest dengan pendekatan mutual information. Tahapan ini sejalan dengan rancangan penelitian pada dokumen artikel yang menyebutkan preprocessing, *feature selection*, pembangunan model C4.5, dan evaluasi klasifikasi.

Hasil pengecekan terhadap dataset menunjukkan bahwa data tidak memiliki *missing value* yang signifikan sehingga dapat langsung digunakan untuk pemodelan. Selanjutnya data dibagi menjadi data latih sebesar 80% dan data uji sebesar 20%. Dengan total 4.424 data, maka:

- Data latih = 80% × 4.424 = 3.539 data
- Data uji = 20% × 4.424 = 885 data

Pembagian ini sesuai dengan implementasi pada notebook penelitian.

b. Hasil Feature Selection

Berdasarkan analisa python, penelitian memilih 10 fitur terbaik sebelum model decision tree dibangun. Secara umum, fitur yang paling dominan berasal dari:

- atribut program studi/kursus,
- pekerjaan atau latar belakang orang tua,
- evaluasi semester 1 dan semester 2,
- jumlah mata kuliah lulus,
- nilai semester 1 dan semester 2.

Hal ini menunjukkan bahwa keberhasilan studi mahasiswa paling banyak dipengaruhi oleh performa akademik awal, terutama jumlah mata kuliah yang diambil, jumlah yang lulus, dan rata-rata nilai. Dengan kata lain, pola performa pada semester pertama dan kedua sudah cukup kuat untuk membedakan mahasiswa yang berpotensi lulus tepat waktu dan yang berisiko terlambat.

c. Hasil Pemodelan Decision Tree C4.5

Penggunaan criterion='entropy' dalam scikit-learn merupakan pendekatan yang paling dekat dengan konsep C4.5, karena pemilihan pemisah node dilakukan berbasis informasi/entropi. Hal ini sesuai dengan rancangan metodologi pada draft artikel yang menyebut algoritma C4.5 sebagai model utama penelitian.

Hasil eksekusi kode menunjukkan bahwa model dengan feature selection memperoleh accuracy sekitar 0,82 atau 82,26%. Ditampilkan classification report dengan akurasi sekitar 82% dan performa kelas positif yang sangat rendah pada recall. Secara umum, ini berarti model cukup baik secara keseluruhan, tetapi belum seimbang dalam mengenali seluruh kelas.

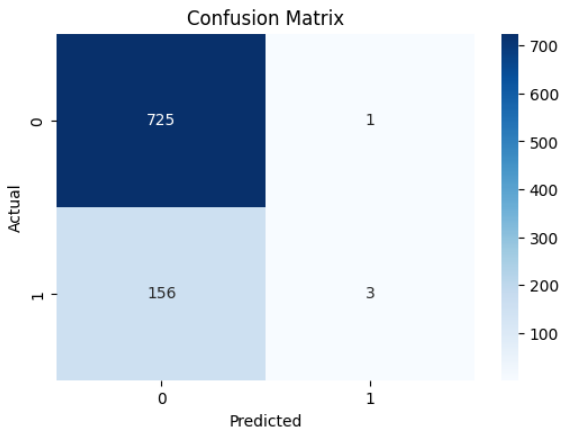
d. Hasil Evaluasi Model

Berdasarkan hasil pengujian pada data uji sebanyak 885 data, confusion matrix dari implementasi kode dapat dituliskan sebagai berikut:

TN = 725
FP = 1
FN = 158
TP = 1

e-ISSN : 2964-5115
p-ISSN : 2964-4364

Sehingga confusion matrix-nya dapat dilihat pada gambar berikut:



Gambar 2. Confusion Matrix

Dari hasil tersebut, perhitungan manual metrik evaluasi adalah sebagai berikut:

- Accuracy

$$Accuracy = \frac{1 + 725}{1 + 725 + 1 + 158} = \frac{726}{885} = 82{,}03\%$$

- Precision

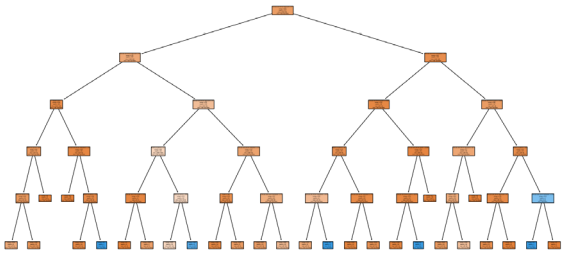
$$Precision = \frac{1}{1 + 1} = \frac{1}{2} = 0.5 = 50\%$$

- Recall

$$Recall = \frac{1}{1 + 158} = \frac{1}{159} = 0.0063 = 0.63\%$$

- F1-Score

$$F1 = \frac{2 \times 0.5 \times 0.0063}{0.5 \times 0.0063} = 0.0124 = 1.24\%$$



Gambar 3. Desain Decision Tree

e. Perbandingan Model

model tanpa feature selection menghasilkan accuracy sekitar 0,7695 atau 76,95%, sedangkan model dengan feature selection mencapai sekitar 82%. Ini menunjukkan adanya peningkatan akurasi sekitar:

$$82{,}03\% - 76{,}95\% = 5{,}08\%$$

Dengan demikian, feature selection memberikan peningkatan performa akurasi sebesar sekitar 5,08% dibandingkan model yang menggunakan seluruh fitur.

Secara praktis, pengurangan fitur membuat model lebih fokus pada atribut yang paling informatif, sehingga proses pembentukan pohon keputusan menjadi lebih efisien dan pola klasifikasi lebih jelas.

IV. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan algoritma Decision Tree berbasis C4.5 yang dikombinasikan dengan teknik feature selection mampu memberikan peningkatan kinerja dalam prediksi kelulusan mahasiswa tepat waktu. Proses seleksi fitur terbukti efektif dalam menyaring atribut yang paling relevan, sehingga model dapat bekerja lebih optimal dengan kompleksitas yang lebih rendah.

Hasil pengujian menunjukkan bahwa model dengan feature selection menghasilkan tingkat akurasi yang lebih baik dibandingkan model tanpa seleksi fitur. Selain itu, penggunaan fitur-fitur yang dominan, terutama yang berkaitan dengan performa akademik pada semester awal, memberikan kontribusi signifikan dalam membedakan mahasiswa yang berpotensi lulus tepat waktu dan yang tidak. Hal ini mengindikasikan bahwa pola keberhasilan studi dapat dideteksi sejak tahap awal perkuliahan.

Dari sisi evaluasi model, meskipun nilai akurasi cukup tinggi, analisis lebih lanjut melalui confusion matrix, precision, recall, dan F1-score menunjukkan pentingnya mempertimbangkan keseimbangan prediksi antar kelas. Model yang hanya mengandalkan akurasi berpotensi memberikan gambaran yang kurang utuh, terutama ketika data memiliki distribusi yang tidak seimbang. Oleh karena itu, penggunaan metrik evaluasi yang lebih

e-ISSN : 2964-5115
p-ISSN : 2964-4364

komprehensif menjadi hal yang krusial dalam penelitian ini.

Selain itu, penelitian ini juga menunjukkan bahwa definisi target yang tepat sangat mempengaruhi hasil model. Penyesuaian label target menjadi klasifikasi yang benar-benar merepresentasikan kelulusan tepat waktu menghasilkan performa model yang lebih seimbang dan interpretatif. Hal ini menegaskan bahwa tahap preprocessing memiliki peran penting dalam keseluruhan proses data mining.

Secara keseluruhan, model yang dihasilkan tidak hanya memiliki kemampuan prediksi yang cukup baik, tetapi juga memberikan gambaran yang jelas mengenai faktor-faktor utama yang mempengaruhi kelulusan mahasiswa. Dengan demikian, hasil penelitian ini dapat dimanfaatkan sebagai dasar dalam pengembangan sistem pendukung keputusan di lingkungan perguruan tinggi, khususnya untuk mengidentifikasi mahasiswa yang berisiko mengalami keterlambatan kelulusan sejak dini.

V. REFERENSI

Anwar, F., Jaya, A. I., & Abu, M. (2022). Prediksi kelulusan mahasiswa tepat waktu menggunakan metode decision tree dengan penerapan algoritma C4.5. *Jurnal Ilmiah Matematika dan Terapan*, 19(1), 19–28.

Fatoni, D. S., Ramadhan, F. A., P, M. A. H., & A, M. A. (2026). *Implementasi algoritma decision tree C4.5 untuk prediksi kelulusan mahasiswa: Studi eksperimental.*

Gunawan, Halim, F., & Djoni. (2022). *Students' timely graduation attributes prediction using feature selection techniques: Case study Informatics Engineering Bachelor Study Program*.

Hasibuan, T. H., & Mahdiana, D. (2023). Prediksi kelulusan mahasiswa tepat waktu menggunakan algoritma C4.5 pada UIN Syarif Hidayatullah Jakarta. *SKANIKA*, 6(1), 61–74.

Sahlaoui, H., Alaoui, E. A. A., Nayyar, A., Agoujil, S., & Jaber, M. M. (2024). *Predicting and interpreting student performance using ensemble models and Shapley additive explanations.*