

Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree dengan Metode CART

Asri Sukma Pratiwi¹, Anggoro Riski², Muhamad Reza Nur Ihsan³, Perani Rosyani⁴

¹Ilmu Komputer, Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: 1psasri19@gmail.com*, 2anggororiski4@gmail.com, 3muhamadreza190505@gmail.com,
4dosen00837@unpam.ac.id
(* : coressponding author)

Abstrak—Kelulusan tepat waktu merupakan indikator penting kinerja perguruan tinggi dan keberhasilan proses akademik mahasiswa. Tingkat keterlambatan kelulusan yang masih tinggi menunjukkan perlunya pendekatan prediktif berbasis data untuk mengidentifikasi mahasiswa berisiko sejak dini. Penelitian ini bertujuan membangun model prediksi kelulusan mahasiswa menggunakan algoritma *Decision Tree* dengan pendekatan CART. Dataset yang digunakan berasal dari Kaggle dan terdiri dari 379 data mahasiswa dengan variabel akademik dan non-akademik. Tahapan penelitian meliputi preprocessing data, pemodelan, evaluasi kinerja, serta analisis *feature importance* dan reduksi fitur. Evaluasi dilakukan menggunakan *confusion matrix*, akurasi, *precision*, *recall*, *F1-score*, serta *cross-validation*. Hasil penelitian menunjukkan bahwa variabel status mahasiswa dan beberapa nilai IPS memiliki pengaruh paling signifikan terhadap prediksi kelulusan. Setelah dilakukan tuning hiperparameter dan stratified *cross-validation*, model mencapai akurasi hingga 92,1% dan menunjukkan stabilitas yang lebih baik. Model dengan fitur terpilih dinilai lebih efisien dan interpretatif sehingga berpotensi diimplementasikan sebagai sistem peringatan dini bagi institusi pendidikan.

Kata Kunci: Mesin Pembelajaran, Pohon Keputusan, Kelulusan Mahasiswa, Klasifikasi, CART

Abstract—Timely graduation is an important indicator of higher education performance and student academic success. Persistently high graduation delays indicate the need for data-driven predictive approaches to identify at-risk students early. This study aims to develop a student graduation prediction model using the Decision Tree algorithm with the CART approach. The dataset was obtained from Kaggle and consists of 379 student records including academic and non-academic variables. The research stages include data preprocessing, modeling, performance evaluation, and feature importance analysis with feature reduction. Model evaluation employed confusion matrix, accuracy, precision, recall, F1-score, and cross-validation. Results indicate that student status and several semester GPA variables are the most influential predictors. After hyperparameter tuning and stratified cross-validation, the model achieved an accuracy of up to 92.1% with improved stability. The reduced-feature model is more efficient and interpretable, making it suitable as an early warning system for higher education institutions.

Keywords: Machine Learning, Decision Tree, Student Graduation, Classification, CART

1. PENDAHULUAN

Secara umum, kelulusan merupakan pengakuan formal bahwa mahasiswa telah menyelesaikan seluruh kewajiban akademik dan administratif sesuai standar perguruan tinggi. Proses ini ditetapkan melalui verifikasi pencapaian kompetensi akhir berdasarkan Standar Kompetensi Lulusan (SKL) yang mencakup aspek pengetahuan, keterampilan, dan sikap. Dengan demikian, kelulusan tidak hanya menandai akhir masa studi, tetapi juga menjadi bentuk sertifikasi atas pencapaian standar kompetensi yang telah ditetapkan (Rahman, 2022).

Dalam konteks pendidikan tinggi, kelulusan sering dikaitkan dengan durasi penyelesaian studi atau kelulusan tepat waktu. Di Indonesia, masa studi ideal ditetapkan selama 4 tahun untuk program sarjana (S1), 3 tahun untuk diploma tiga (D3), dan 2 tahun untuk magister (S2), dengan batas maksimal penyelesaian studi S1 selama 7 tahun. Tingkat kelulusan tepat waktu memiliki pengaruh signifikan terhadap kualitas akademik dan akreditasi perguruan tinggi, di mana Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT) menetapkan standar minimal 60% lulusan per angkatan harus lulus tepat waktu. Oleh karena itu, permasalahan kelulusan tepat waktu telah berkembang menjadi isu strategis institusional yang berkaitan dengan reputasi, akuntabilitas publik, dan keberlanjutan perguruan tinggi.

Untuk mendukung pengambilan keputusan berbasis data dalam memprediksi kelulusan mahasiswa, pendekatan *machine learning* (ML) banyak digunakan. *Machine learning* merupakan cabang kecerdasan buatan yang memungkinkan sistem mempelajari pola dari data tanpa diprogram

secara eksplisit, sehingga mampu mengidentifikasi pola dan permasalahan baru secara efektif. Salah satu tugas utama dalam *machine learning* adalah klasifikasi, yaitu proses pengelompokan data ke dalam kelas tertentu berdasarkan model yang dibangun dari data latih. Model klasifikasi digunakan untuk memprediksi kelas data baru yang belum pernah diamati sebelumnya dan telah banyak diterapkan pada berbagai bidang, termasuk prediksi kelulusan mahasiswa (Fatimah & Nuryaningsih, 2018).

Salah satu algoritma klasifikasi yang banyak digunakan adalah *Decision Tree* atau pohon keputusan. Algoritma ini termasuk dalam teknik data mining yang menyajikan proses pengambilan keputusan dalam bentuk struktur pohon. Proses pembentukan pohon keputusan dimulai dari root node yang merepresentasikan seluruh data (Faisal et al., 2021). *Decision Tree* tersusun atas tiga jenis node, yaitu root node sebagai titik awal pengambilan keputusan, internal node yang berfungsi sebagai node percabangan berdasarkan kondisi tertentu, serta leaf node sebagai node terminal yang merepresentasikan hasil klasifikasi akhir (Renyut et al., 2022).

Penelitian ini menggunakan pendekatan kuantitatif dengan metode klasifikasi berbasis *machine learning* untuk memprediksi kelulusan mahasiswa. Dataset yang digunakan berisi data akademik dan non-akademik mahasiswa, meliputi umur, status mahasiswa, status pernikahan, nilai IPS per semester, dan IPK, dengan status kelulusan sebagai variabel target. Tahapan penelitian meliputi preprocessing data, penanganan missing value, transformasi data kategorikal, pembagian data latih dan data uji, serta pembangunan model menggunakan algoritma *Decision Tree* dengan pendekatan CART. Evaluasi kinerja model dilakukan menggunakan *confusion matrix* dan metrik akurasi, presisi, recall, serta F1-score, disertai analisis feature importance untuk mengidentifikasi variabel yang paling berpengaruh terhadap prediksi kelulusan.

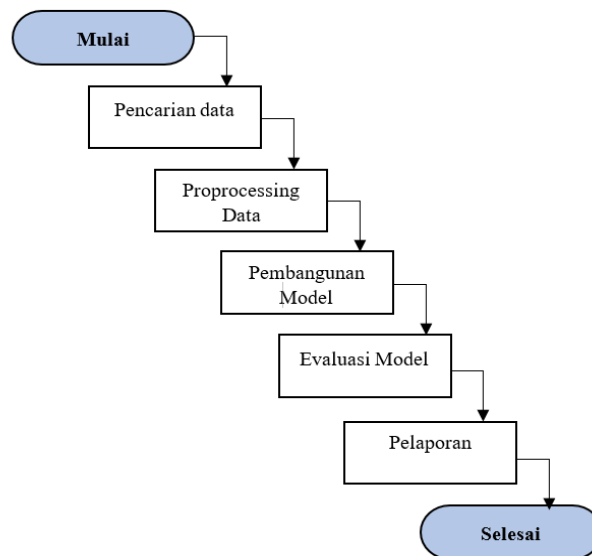
Hasil penelitian menunjukkan bahwa algoritma *Decision Tree* mampu memprediksi kelulusan mahasiswa dengan tingkat akurasi yang baik serta menghasilkan aturan keputusan yang mudah diinterpretasikan. Variabel status mahasiswa dan beberapa nilai IPS memiliki pengaruh paling signifikan terhadap ketepatan waktu kelulusan, sedangkan variabel demografis menunjukkan kontribusi yang relatif lebih kecil. Dengan demikian, model yang dihasilkan berpotensi digunakan sebagai sistem pendukung keputusan untuk membantu institusi pendidikan dalam mengidentifikasi mahasiswa yang berisiko mengalami keterlambatan kelulusan secara lebih dini.

2. METODE PENELITIAN

2.1 Metode Penelitian

Metode yang digunakan dalam penelitian ini adalah literature review. Penelitian terkait digunakan sebagai landasan dalam penyusunan penelitian ini serta untuk mengidentifikasi kesenjangan dengan penelitian sebelumnya. Melalui kajian penelitian terdahulu, dilakukan perbandingan metode, pendekatan, dan hasil penelitian, sehingga penelitian ini diharapkan memberikan kontribusi yang relevan sesuai dengan permasalahan yang diteliti.

Metode lainnya dalam penelitian ini yaitu metode waterfall yang merupakan salah satu model dalam System Development Life Cycle (SDLC) yang menerapkan pendekatan pengembangan perangkat lunak secara linear dan berurutan. Setiap tahapan pada metode waterfall harus diselesaikan terlebih dahulu sebelum melanjutkan ke tahap berikutnya, sehingga proses pengembangan menjadi terstruktur dan sistematis. Tahapan dalam metode ini meliputi pencarian data sebagai bentuk analisis kebutuhan untuk mengidentifikasi kebutuhan pengguna, tahap desain yang mencakup perancangan struktur data dan arsitektur sistem, preprocessing dan pemodelan pada tahap pengkodean untuk menerjemahkan desain ke dalam program, evaluasi model sebagai tahap pengujian untuk memastikan fungsi sistem berjalan dengan baik menggunakan fitur test dan score yang tidak hanya membagi dataset menjadi data pelatihan dan pengujian, tetapi juga memberikan berbagai metrik evaluasi untuk menilai kinerja model (Kusuma et al., 2025), serta tahap pemeliharaan untuk melakukan perbaikan dan penyesuaian sistem setelah digunakan oleh pengguna, termasuk pelaporan (Sukanto & Shalahuddin, 2014).



Gambar 1. Diagram Proses

2.2 PreProcessing

Dataset penelitian ini berasal dari platform Kaggle dan terdiri dari 379 data dengan 14 variabel setelah penghapusan atribut nama yang tidak memiliki nilai prediktif. Variabel yang digunakan mencakup data akademik dan non-akademik, yaitu umur, nilai IPS semester 1–8, IPK, jenis kelamin, status mahasiswa, dan status pernikahan, dengan status kelulusan sebagai variabel target yang diklasifikasikan menjadi tepat waktu dan terlambat. Sebelum pemodelan, dilakukan pemeriksaan struktur data, distribusi nilai, tipe data, serta penanganan missing value untuk memastikan kualitas data. Dataset ini dinilai sesuai untuk penelitian karena memiliki variabel target yang jelas, struktur data yang mendukung algoritma *Decision Tree*, serta fitur yang relevan dan mudah diinterpretasikan untuk mendukung pengambilan keputusan akademik.

Tabel 1. Struktur Dataframe

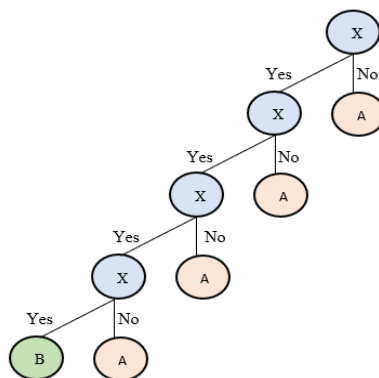
Column	Non-Null Count	Dtype	Missing Value
Jenis Kelamin	379 non-null	Object	
Status Mahasiswa	379 non-null	Object	
Umur	379 non-null	Int64	
Status Nikah	379 non-null	Object	
IPS 1	379 non-null	Float64	
IPS 2	379 non-null	Float64	
IPS 3	379 non-null	Float64	
IPS 4	379 non-null	Float64	
IPS 5	379 non-null	Float64	

Column	Non-Null Count	Dtype	Missing Value
IPS 6	379 non-null	Float64	
IPS 7	379 non-null	Float64	
IPS 8	372 non-null	Float64	7 Data
IPK	366 non-null	Float64	3 Data
Status Kelulusan	379 non-null	Object	

Tahapan preprocessing data dimulai dengan pembersihan data, yaitu menghapus variabel yang tidak relevan seperti kolom nama karena tidak memiliki nilai prediktif dan berpotensi menimbulkan noise. Selanjutnya dilakukan pemeriksaan struktur data yang menunjukkan adanya missing value pada kolom IPS 8 dan IPK, sehingga dilakukan penanganan dengan metode imputasi menggunakan nilai mean untuk menjaga stabilitas dan akurasi model. Variabel kategorikal kemudian diubah menjadi numerik melalui proses encoding agar dapat diproses oleh algoritma. Setelah itu, dilakukan pemisahan fitur dan label, di mana fitur X berperan sebagai variabel input dan label y sebagai variabel output. Dataset selanjutnya dibagi menjadi data latih dan data uji dengan rasio 80:20 untuk memastikan evaluasi kinerja model dilakukan secara objektif dan seimbang. Rasio ini dipilih karena merupakan standar umum dalam penelitian klasifikasi dan memberikan keseimbangan yang baik antara pelatihan dan evaluasi sesuai jurnal terdahulu. (Ikasari et al., 2025).

2.3 Algoritma Decision Tree

Algoritma *Decision Tree* memiliki berbagai varian, seperti ID3, C4.5, C5.0, CART, CHAID, dan Conditional Inference Trees, yang masing-masing berbeda dalam kriteria pemilihan atribut dan strategi splitting (Mienye & Jere, 2024). Pada penelitian ini pemilihan metode CART (Classification and Regression Tree) sebagai model utama karena menggunakan prinsip pemisahan biner serta Gini Index dalam menentukan split pada fitur numerik maupun kategorikal yang telah di-encode (Hakim & Rilvani, 2025). Pemilihan CART didasarkan pada kemampuannya dalam melakukan pruning untuk mencegah *overfitting*, mendukung berbagai tipe data, serta memiliki stabilitas dan akurasi yang baik pada dataset yang kompleks dengan dominasi fitur numerik.



Gambar 2. Pohon Keputusan

Visualisasi pohon keputusan yang dihasilkan menunjukkan bahwa setiap node melakukan satu pemisahan berdasarkan nilai threshold tertentu, yang membagi data ke dalam dua cabang keputusan, yaitu ya (yes) dan tidak (no), hingga mencapai node terminal. Proses pembentukan

pohon keputusan diawali dengan perhitungan nilai Entropy dan Information Gain untuk menentukan variabel yang menjadi akar pohon berdasarkan nilai Gain tertinggi, kemudian dilanjutkan dengan pemisahan biner secara berurutan pada setiap tingkat pohon (Khasanah et al., 2025).

Algoritma *Decision Tree* pada penelitian ini diimplementasikan menggunakan library Scikit-learn, yang menyediakan parameter internal untuk mengatur proses pembentukan pohon keputusan secara otomatis selama pelatihan model (Balcan & Sharma, 2024). Parameter criterion digunakan sebagai ukuran ketidakmurnian data, di mana Gini Index dipilih sebagai kriteria pemisahan atribut yang paling optimal (Rahayu et al., 2025), sedangkan parameter splitter diatur menjadi best untuk memilih pemisahan terbaik berdasarkan nilai impurity terendah. Selain parameter model, dilakukan penyetelan hiperparameter seperti max_depth, min_samples_split, min_samples_leaf, dan random_state untuk mengontrol kompleksitas model, mencegah *overfitting*, serta memastikan hasil pelatihan yang konsisten dan dapat direproduksi (Arifin & Adiyono, 2024).

Evaluasi kinerja model dilakukan menggunakan *confusion matrix*, yang merepresentasikan hasil klasifikasi dalam bentuk True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Dalam jurnal klasifikasi yang ditulis Prasetyo, Simeon Yuda dkk, dijelaskan bahwa perhitungan dalam *Confusion Matrix* memperhatikan nilai TP, TN, FP dan FN (Bangsa & Hermawan, 2021) yang menjadi dasar untuk menghitung metrik evaluasi seperti Akurasi (*Accuracy*), Presisi (*Precision*), Recall dan *F1-Score*:

a. Akurasi

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

b. *Precision*

$$Precision = \frac{(TP)}{(TP + FP)}$$

c. *Recall*

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

d. *F1-Score*

$$F1 - Score = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)}$$

3. ANALISA DAN PEMBAHASAN

3.1 Analisa Penelitian

a. Preprocessing

Preprocessing dilakukan untuk memastikan data yang digunakan bersih dan siap untuk dianalisis, yang meliputi pembersihan data, encoding kolom kategorikal hingga pembagian data (Yazid, 2024).

Pemilihan dataset diambil dari sumber data online kaggle.com bertipe *supervised learning* dimana terdapat satu atribut yang berperan sebagai kelas atau target dan sesuai untuk dimodelkan menggunakan metode klasifikasi (Darujati et al., 2012).

Import library, menggunakan Pandas menyediakan struktur data yang efisien dan fleksibel dan Numpy yang berfungsi sebagai fondasi numerik Python dengan kemampuan mengelola array multidimensi dan melakukan komputasi matematika dalam jumlah besar secara cepat dan efisien sangat sesuai digunakan dalam penelitian berbasis data mining dan *machine learning* (Candra, 2025).

Dataset yang telah diupload menampilkan dataframe dimana terdapat informasi missing values pada kolom IPS 8 dan IPK yang membutuhkan penanganan dengan cara mengisi nilai kosong dengan mean. Menurut penelitian terdahulu mean mempertahankan akurasi klasifikasi lebih baik dibanding pengisian nol atau penghapusan data (Pamuji et al., 2024). Setelah penanganan tersebut, kemudian variabel kategorikal diubah melalui encoding label menjadi numerik agar algoritma *machine learning* dapat memproses variabelnya (Zhu et al., 2024). Dan didapatkan pengkodean JENIS KELAMIN : LAKI- LAKI = 0, PEREMPUAN = 1, STATUS MAHASISWA : MAHASISWA = 0, BEKERJA = 1, STATUS NIKAH : BELUM MENIKAH = 0, MENIKAH = 1, STATUS KELULUSAN : TEPAT = 0, TERLAMBAT = 1.

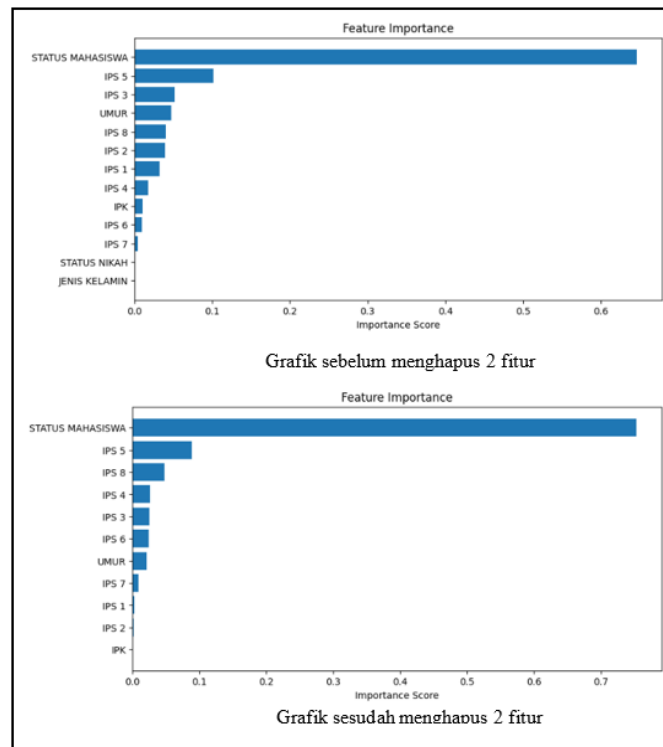
Selanjutnya dilakukan pemisahan fitur X (input) dan Y (output). Fitur X berisi variabel yang dianggap berpengaruh terhadap hasil prediksi, sementara fitur Y adalah satu kolom khusus yang memiliki label atau nilai yang ingin diprediksi.

Pada penelitian ini digunakan pembagian data secara acak dengan rasio 80% sebagai data latih dan 20% sebagai data uji, dengan pengaturan random state tertentu agar hasil pembagian bersifat konsisten. Rasio ini dipilih karena umum digunakan dan terbukti memberikan keseimbangan yang baik antara proses pelatihan dan evaluasi model dalam berbagai penelitian sebelumnya (Fahrudin et al., 2024).

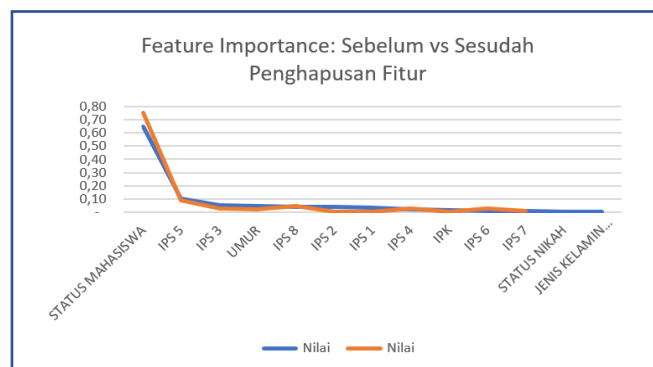
b. Membangun Model

Penelitian ini menggunakan Algoritma *Decision Tree* dengan metode CART (Classification and Regression Tree) sebagai metode klasifikasi karena struktur pohonnya mudah dipahami dan diinterpretasikan. Metode CART membangun pohon keputusan secara rekursif dengan memisahkan data berdasarkan atribut tertentu hingga mencapai node daun sebagai hasil prediksi kelas. Implementasi CART dilakukan menggunakan pustaka Scikit-learn, sebagaimana diterapkan pada penelitian terdahulu terkait klasifikasi risiko gagal studi mahasiswa (Rafika et al., 2025). Hasilnya, akurasi pada model mencapai 90,78%.

Pada bagian feature importance, diketahui bahwa terdapat 2 variabel yang kurang berpengaruh yaitu STATUS NIKAH dan JENIS KELAMIN. Feature importance menjelaskan pengurutan dilakukan berdasarkan skor terpenting (Wang et al., 2024). Mempertimbangkan variabel dalam dataset yang berhubungan dengan karakteristik pribadi seseorang dan kekhawatiran mengenai diskriminasi gender, maka peneliti memutuskan untuk menghapus variabel tersebut, dan setelah pengukuran kembali akurasi menurun sebesar 89,47%. Hal ini dapat dipahami karena saat dua fitur dibuang, importance dari fitur lain otomatis naik atau berubah. Karena total importance = 1, maka fitur yang tersisa akan dibagi ulang kontribusinya. *Decision Tree* membentuk struktur baru, setelah fitur dihapus, tree harus mencari split terbaik berdasarkan fitur yang tersisa. Nilai importance suatu fitur bisa meningkat atau menurun jika fitur lain dihapus atau dimodifikasi, karena model akan menyesuaikan pembagian data dan kontribusi setiap fitur terhadap prediksi (Rajbahadur et al., 2022).



Gambar 3. Perbandingan Featur Importance Dalam Proses Modifikasi



Gambar 4. Chart Perbandingan Featur Importance Dalam Proses Modifikasi

Tabel 2. Tabel Perbandingan Featur Importance Dalam Proses Modifikasi

Sebelum	Nilai	Sesudah Penghapusan Fitur	Nilai
Status Mahasiswa	0,65	Status Mahasiwa	0,75
IPS 5	0,10	IPS 5	0,09
IPS 3	0,05	IPS 8	0,05
Umur	0,05	IPS 4	0,03
IPS 8	0,04	IPS 3	0,03
IPS 2	0,04	IPS 6	0,02

Sebelum	Nilai	Sesudah Penghapusan Fitur	Nilai
IPS 1	0,03	Umur	0,02
IPS 4	0,02	IPS 7	0,01
IPK	0,01	IPS 1	0,00
IPS 6	0,01	IPS 2	0,00
IPS 7	0,00	IPK	0,00
Status Nikah	0,00	-	-
Jenis Kelamin	0,00	-	-

Peneliti kemudian menemukan adanya kemungkin *overfitting* pada model setelah dilakukan *cross validation* sebagai bagian dari evaluasi model (Rahayu et al., 2025), dimana angka akurasi pada test menjadi 84,9%. Sehingga peneliti membagi data menjadi beberapa fold pada *cross-validation* dapat menghasilkan distribusi target yang berbeda-beda di setiap fold menggunakan teknik stratified agar hasil stabil dan melakukan pembatasan parameter *max_depth*, *min_samples_leaf* dan *min_samples_split* dimana ketiga parameter ini adalah cara utama regularisasi *Decision Tree* agar tidak *overfit*, terutama saat dataset kecil atau fitur banyak. Dengan demikian *Cross-validation* membantu mengevaluasi performa model secara lebih robust (Qiu, 2024).

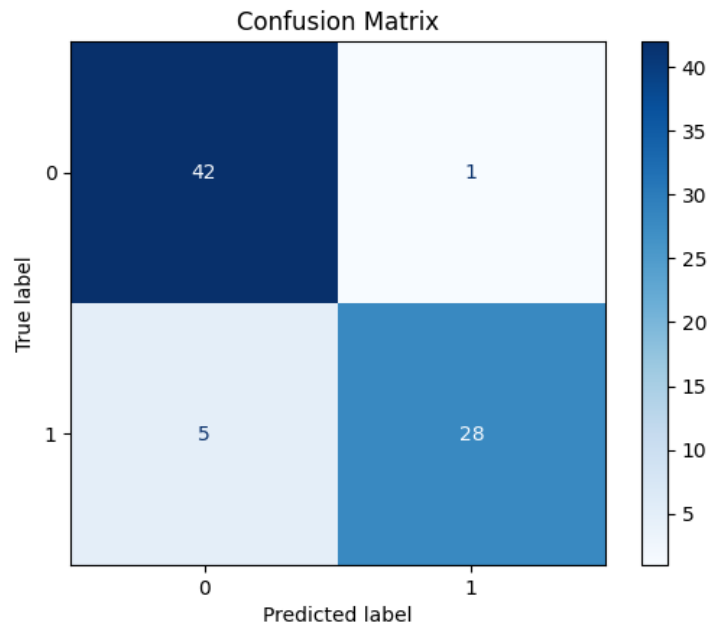
Tabel 3. Perbandingan Akurasi

	Akurasi	
Parameter	Data Train	Cross Validation
Model awal	89,4%	84,9%
Pembatasan 3 parameter	90,7%	90,2%
Stratify Sampling	92,1%	90,2%

Secara keseluruhan, kombinasi *hyperparameter tuning* dan stratify sampling berhasil menghasilkan model *Decision Tree* berbasis CART yang lebih akurat, stabil, dan memiliki kemampuan generalisasi yang lebih baik terhadap data baru.

c. Visualisasi dan Evaluasi

Melalui *confusion matrix*, peneliti mengevaluasi keakuratan dari hasil klasifikasi. Matriks ini memungkinkan hasil analisis seberapa tepat classifier mengidentifikasi masing-masing kelas (Yunita et al., 2018).



Gambar 5. Hasil confusion matrix

Dari *confusion matrix* tersebut, model menghasilkan 42 True Negative (TN), 28 True Positive (TP), 1 False Positive (FP), dan 5 False Negative (FN).

1. Akurasi

$$Akurasi = \frac{(TP + TN)}{(TP + TN + FN + FP)} = \frac{(42 + 28)}{(42 + 28 + 5 + 1)} = 92,1\%$$

Model mampu memprediksi dengan benar 92% dari seluruh data. Ini menunjukkan performa keseluruhan model yang sangat baik setelah *tuning* dan pembagian data yang optimal.

2. *Precision*

$$Precision = \frac{(TP)}{(TP + FP)} = \frac{(28)}{(28 + 1)} = 96,5\%$$

Dari semua prediksi positif (kelas 1), 96,5% adalah benar. Hal ini menunjukkan model jarang salah menandai data negatif sebagai positif.

3. *Recall*

$$Recall = \frac{(TP)}{(TP + FN)} = \frac{(28)}{(28 + 5)} = 84,8\%$$

Model berhasil menemukan 84,8% dari seluruh data positif. Ada 5 data positif yang terlewat (FN), artinya model sedikit kurang sensitif terhadap kelas positif dibanding *precision*-nya.

4. *F1-Score*

$$F1 - Score = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} = \frac{(2 \times 0,965 \times 0,848)}{(0,965 + 0,848)} = 90,2\%$$

F1-score 90,2% menunjukkan model memiliki keseimbangan yang baik antara ketepatan prediksi positif (*precision* tinggi), dan kemampuan menangkap semua data positif (*recall* cukup tinggi).

3.2 Hasil Penelitian

Evaluasi kinerja model dilakukan melalui beberapa skenario untuk menguji respons model terhadap fitur-fitur yang memiliki tingkat kepentingan tinggi. Pada skenario pertama, perubahan nilai status mahasiswa menunjukkan bahwa mahasiswa dengan status bekerja memiliki peluang lebih besar untuk diprediksi lulus terlambat dibandingkan mahasiswa penuh waktu. Skenario kedua menunjukkan bahwa umur yang lebih muda memiliki probabilitas lebih tinggi untuk lulus tepat waktu dibandingkan umur yang lebih tua. Pada skenario ketiga, peningkatan nilai IPS 5 menghasilkan peluang yang lebih besar untuk diprediksi lulus tepat waktu.

```
=====
SISTEM PREDIKSI KELULUSAN MAHASISWA
=====

Silakan masukkan data mahasiswa:
-----

Status Mahasiswa (pilih nomor):
1. BEKERJA
2. MAHASISWA

=====
DATA YANG DIINPUTKAN
=====
Status Mahasiswa : BEKERJA
Umur : 24.0
IPS 1 : 3.45
IPS 2 : 3.21
IPS 3 : 3.12
IPS 4 : 2.87
IPS 5 : 2.9
IPS 6 : 3.32
IPS 7 : 3.13
IPS 8 : 3.22
IPK : 3.15
...
Prediksi Kelulusan : TERLAMBAT
Prob Tepat Waktu : 0.1429
Prob Terlambat : 0.8571
=====
```

```
=====
SISTEM PREDIKSI KELULUSAN MAHASISWA
=====

Silakan masukkan data mahasiswa:
-----

Status Mahasiswa (pilih nomor):
1. BEKERJA
2. MAHASISWA

=====
DATA YANG DIINPUTKAN
=====
Status Mahasiswa : MAHASISWA
Umur : 24.0
IPS 1 : 3.45
IPS 2 : 3.21
IPS 3 : 3.12
IPS 4 : 2.87
IPS 5 : 2.9
IPS 6 : 3.32
IPS 7 : 3.13
IPS 8 : 3.22
IPK : 3.15
...
Prediksi Kelulusan : TEPAT WAKTU
Prob Tepat Waktu : 1.0000
Prob Terlambat : 0.0000
=====
```

Gambar 6. Perbandingan Prediksi Berdasarkan Status Mahasiswa

```
=====
SISTEM PREDIKSI KELULUSAN MAHASISWA
=====

Silakan masukkan data mahasiswa:
-----

Status Mahasiswa (pilih nomor):
1. BEKERJA
2. MAHASISWA

=====
DATA YANG DIINPUTKAN
=====
Status Mahasiswa : MAHASISWA
Umur : 18.0
IPS 1 : 3.45
IPS 2 : 3.21
IPS 3 : 3.12
IPS 4 : 2.87
IPS 5 : 2.9
IPS 6 : 3.32
IPS 7 : 3.13
IPS 8 : 3.22
IPK : 3.15
...
Prediksi Kelulusan : TEPAT WAKTU
Prob Tepat Waktu : 1.0000
Prob Terlambat : 0.0000
=====
```

```
=====
SISTEM PREDIKSI KELULUSAN MAHASISWA
=====

Silakan masukkan data mahasiswa:
-----

Status Mahasiswa (pilih nomor):
1. BEKERJA
2. MAHASISWA

=====
DATA YANG DIINPUTKAN
=====
Status Mahasiswa : MAHASISWA
Umur : 28.0
IPS 1 : 3.45
IPS 2 : 3.21
IPS 3 : 3.12
IPS 4 : 2.87
IPS 5 : 2.9
IPS 6 : 3.32
IPS 7 : 3.13
IPS 8 : 3.22
IPK : 3.15
...
Prediksi Kelulusan : TEPAT WAKTU
Prob Tepat Waktu : 0.6333
Prob Terlambat : 0.3667
=====
```

Gambar 7. Perbandingan Prediksi Berdasarkan Umur

<pre>===== SISTEM PREDIKSI KELULUSAN MAHASISWA ===== Silakan masukkan data mahasiswa: ----- Status Mahasiswa (pilih nomor): 1. BEKERJA 2. MAHASISWA ===== DATA YANG DIINPUTKAN ===== Status Mahasiswa : MAHASISWA Umur : 24.0 IPS 1 : 3.45 IPS 2 : 3.21 IPS 3 : 3.12 IPS 4 : 2.87 IPS 5 : 2.9 IPS 6 : 3.32 IPS 7 : 3.13 IPS 8 : 3.22 IPK : 3.15 ===== HASIL PREDIKSI ===== Prediksi Kelulusan : TEPAT WAKTU Prob Tepat Waktu : 1.0000 Prob Terlambat : 0.0000 =====</pre>	<pre>===== SISTEM PREDIKSI KELULUSAN MAHASISWA ===== Silakan masukkan data mahasiswa: ----- Status Mahasiswa (pilih nomor): 1. BEKERJA 2. MAHASISWA ===== DATA YANG DIINPUTKAN ===== Status Mahasiswa : MAHASISWA Umur : 24.0 IPS 1 : 3.45 IPS 2 : 3.21 IPS 3 : 2.78 IPS 4 : 2.87 IPS 5 : 0.0 IPS 6 : 3.32 IPS 7 : 3.13 IPS 8 : 3.22 IPK : 2.75 ... Prediksi Kelulusan : TEPAT WAKTU Prob Tepat Waktu : 0.8000 Prob Terlambat : 0.2000 =====</pre>
--	--

Gambar 8. Perbandingan Prediksi Berdasarkan IPS 5

Hasil ini menunjukkan bahwa model secara konsisten memanfaatkan fitur-fitur penting dalam proses pengambilan keputusan dan selaras dengan analisis *feature importance*.

4. KESIMPULAN

Penelitian ini berhasil membangun model klasifikasi kelulusan mahasiswa menggunakan algoritma Decision Tree berbasis CART dengan memanfaatkan data akademik dan non-akademik mahasiswa. Hasil pengujian menunjukkan bahwa penerapan *hyperparameter tuning* serta stratify sampling mampu meningkatkan akurasi dan stabilitas model serta mengurangi risiko *overfitting*. Analisis *feature importance* dan pengujian skenario membuktikan bahwa status mahasiswa, umur, dan nilai IPS pada semester tertentu merupakan faktor utama yang memengaruhi ketepatan waktu kelulusan. Model yang dihasilkan tidak hanya memiliki kinerja yang baik, tetapi juga mudah diinterpretasikan, sehingga berpotensi digunakan sebagai sistem pendukung keputusan bagi institusi pendidikan dalam mendeteksi mahasiswa yang berisiko mengalami keterlambatan kelulusan secara lebih dini.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan dan kontribusi dalam pelaksanaan penelitian ini, khususnya kepada dosen pembimbing atas arahan dan bimbingan yang diberikan, serta kepada pihak-pihak yang telah menyediakan dataset dan referensi yang digunakan dalam penelitian ini. Penulis juga menyampaikan apresiasi kepada institusi dan rekan-rekan yang turut membantu hingga penelitian ini dapat diselesaikan dengan baik.

REFERENCES

Arifin, M., & Adiyono, S. (2024). Hyperparameter tuning in machine learning to predict student academic achievement. *International Journal of Artificial Intelligence Research*, 8(1), 1–8.
<https://ijair.id/index.php/ijair/article/view/1214>

- Balcan, M. F., & Sharma, D. (2024). Learning accurate and interpretable decision trees. *Proceedings of Machine Learning Research*, 244, 288–307.
- Bangsa, P. D., & Hermawan, I. (2021). "Jurnal Teknologi Terpadu," *J. Teknol. Terpadu*, vol. 7, no. 1, pp. 15–22, 2021, [Online]. Available: <https://media.neliti.com/media/publications/493730-water-ph-and-turbidity-control-system-in-0a553e14.pdf>
- Candra, A. P. (2025). Analisis data menggunakan Python: Memperkenalkan Pandas dan NumPy. *Journal of Information System Education Development*, 3(1), 11–16. <https://doi.org/10.62386/jised.v3i1.118>
- Darujati, C., et al. (2012). "Pemanfaatan teknik supervised untuk klasifikasi teks bahasa Indonesia" *Text*, vol. 16, no. 1, pp. 1–8, 2012.
- Fahrudin, N. F., Putra, K. R., Umaroh, S., & Lautan, G. B. (2024). Influence of data scaling and train/test split ratios on LightGBM efficacy for obesity rate prediction. *MIND Journal*, 9(2), 220–234. <https://doi.org/10.26760/mindjournal.v9i2.220-234>
- Faisal, F., Dhika, H., & Veris, H. (2021). Penerapan algoritma decision tree dalam penjualan handphone. *JRKT (Jurnal Rekayasa Komputasi Terapan)*, 1(4). <https://doi.org/10.30998/jrkt.v1i04.6157>
- Fatimah, & Nuryaningsih. (2018). Buku ajar asuhan kebidanan pada kehamilan. Yogyakarta: KMedia.
- Hakim, A. S., & Rilvani, E. (2025). Prediksi cuaca menggunakan metode classification ID3 dan CART. *Jurnal Media Akademik*, 3(8). <https://doi.org/10.62281/v3i8.2669>
- Ikasari, I. H., Rosyani, P., & Amalia, R. (2025). Klasifikasi jenis buah menggunakan metode CNN. *RIGGS: Journal of Artificial Intelligence and Digital Business*, 4(2), 5451–5458. <https://doi.org/10.31004/riggs.v4i2.1271>
- Khasanah, N., Saputri, D. U. E., Hidayat, T., & Aziz, F. (2025). Prediksi kelulusan mahasiswa menggunakan algoritma C4.5 dengan RapidMiner: Studi kasus data akademik perguruan tinggi XYZ. *Indonesian Journal of Computer Science*, 4(2), 100–107. <https://doi.org/10.31294/ijcs.v4i2.9647>
- Kusuma, S. D. Y., Al Islami, H., & Rosyani, P. (2025). Penerapan Naive Bayes untuk klasifikasi penyakit endokrin pada pasien lansia. *KERNEL: Jurnal Riset Inovasi Bidang Informatika dan Pendidikan Informatika*, 5(2), 72–82. <https://doi.org/10.31284/j.kernel.2024.v5i2.7312>
- Mienye, I. D., & Jere, N. (2024). A survey of decision trees: Concepts, algorithms, and applications. *IEEE Access*, 12, 86716–86727. <https://doi.org/10.1109/ACCESS.2024.3416838>
- Pamuji, F. Y., Muslikh, A. R., Arief, R. M., & Muti, D. (2024). Komparasi metode mean dan KNN imputation dalam mengatasi missing value pada dataset kecil. *Jurnal Informatika Polinema*, 10(2), 257–264.
- Qiu, J. (2024). An analysis of model evaluation with cross-validation: Techniques, applications, and recent advances. *Advances in Economics, Management and Political Sciences*, 99(1), 69–72. <https://doi.org/10.54254/2754-1169/99/2024ox0213>
- Rafika, C. S., Reynanda, R. M., & Sari, A. P. (2025). Penerapan decision tree CART untuk klasifikasi risiko gagal studi mahasiswa. *Jurnal Mahasiswa Teknik Informatika*, 4(2), 37–46. <https://doi.org/10.35473/jamastika.v4i2.4145>
- Rahayu, E. S., Anugrah, O., Purnama, A., Zakaria, H., & Rosyani, P. (2025). Klasifikasi penyakit jamur pada tanaman tomat dengan algoritma SVM. *Bulletin of Computer Science Research*, 5(4), 756–762. <https://doi.org/10.47065/bulletincsr.v5i4.515>
- Rahayu, S., Rosyani, P., Saputra, R. Y., Umar, R. A., Prasdio, S., & Syach, W. A. (2025). Application of expert system in rice seedling selection based on smart data with methods: Knowledge-based system and decision tree. *International Journal of Integrated Science*, 4(1), 217–224. <https://doi.org/10.55927/ijis.v4i1.13510>
- Rahman, A. (2022). Peran prestasi non-akademik dalam membangun karakter dan keterampilan sosial siswa. *Jurnal Pendidikan dan Konseling*, 4(1980), 1349–1358.
- Rajbahadur, G. K., Wang, S., Oliva, G. A., Kamei, Y., & Hassan, A. E. (2022). The impact of feature importance methods on the interpretation of defect classifiers. *IEEE Transactions on Software Engineering*, 48(7), 2245–2261. <https://doi.org/10.1109/TSE.2021.3056941>
- Renyut, D. H., Wabula, Y., & Ferdinand, F. (2022). Prediksi kelulusan mahasiswa menggunakan algoritma C4.5 (Studi kasus: Sekolah Tinggi Ilmu Administrasi Trinitas Ambon). *Simtek: Jurnal Sistem Informasi dan Teknik Komputer*, 7(2), 80–86. <https://doi.org/10.51876/simtek.v7i2.137>
- Sukanto, & Shalahuddin. (2014). *Rekayasa perangkat lunak terstruktur dan berorientasi objek*. Bandung: Informatika.
- Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, 11(1). <https://doi.org/10.1186/s40537-024-00905-w>
- Yazid, A. S. (2024). Eksplorasi data akademik untuk memprediksi ketepatan waktu lulus mahasiswa menggunakan algoritma Naive Bayes. *Jurnal Teknik Informatika dan Sistem Informasi*, 11(4), 558–568.

- Yunita, D., Rosyani, P., & Amalia, R. (2018). Analisa prestasi siswa berdasarkan kedisiplinan, nilai hasil belajar, sosial ekonomi dan aktivitas organisasi menggunakan algoritma Naïve Bayes. Jurnal Informatika Universitas Pamulang, 3(4), 209. <https://doi.org/10.32493/informatika.v3i4.2032>
- Zhu, W., Qiu, R., & Fu, Y. (2024). Comparative study on the performance of categorical variable encoders in classification and regression tasks. arXiv. <https://arxiv.org/abs/2401.09682>