Journal of Information System Research (JOSH)
Volume 7, No. 2, January 2026, pp 298−305
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9099

Optimalisasi Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree CART

Afiani Agus Abdillah*, Yono Cahyono, Teti Desyani, Perani Rosyani

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Jl. Raya Puspitek No.46, Serpong, Kota Tangerang Selatan, Banten 15316, Indonesia
Email: [1,*]dosen01364@unpam.ac.id, [2]dosen00843@unpam.ac.id, [3]dosen00839@unpam.ac.id, [4]dosen00837@unpam.ac.id
Email Penulis Korespondensi: dosen01364@unpam.ac.id
Submitted: **04/01/2026**; Accepted: **12/01/2026**; Published: **14/01/2026**

Abstrak−Kelulusan mahasiswa tepat waktu merupakan salah satu indikator penting dalam menilai mutu dan efektivitas penyelenggaraan pendidikan tinggi. Penelitian ini bertujuan untuk mengoptimalkan prediksi kelulusan mahasiswa menggunakan algoritma Decision Tree berbasis Classification and Regression Tree (CART) dengan memanfaatkan variabel akademik dan non-akademik. Data yang digunakan berasal dari dataset terbuka *Student Graduation Dataset* yang terdiri dari 379 data mahasiswa dengan status kelulusan sebagai variabel target. Tahapan penelitian meliputi pra-pemrosesan data melalui penanganan *missing value* menggunakan *mean imputation*, transformasi variabel kategorikal, pembagian data latih dan data uji dengan rasio 80:20, serta pembangunan dan optimalisasi model CART melalui pengaturan *hyperparameter* sebagai bentuk *post-pruning*. Evaluasi kinerja model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*. Hasil penelitian menunjukkan bahwa model CART yang dioptimalkan mampu mencapai tingkat akurasi sebesar 92,1%, dengan nilai *F1-score* di atas 90% pada kedua kelas kelulusan, serta keseimbangan yang baik antara *precision* dan *recall*. Struktur pohon keputusan yang dihasilkan juga bersifat sederhana dan mudah diinterpretasikan. Dengan demikian, algoritma CART yang dioptimalkan dinilai efektif dan relevan untuk diterapkan sebagai *early warning system* dalam mendukung pengambilan keputusan akademik di perguruan tinggi.

Kata Kunci: Kelulusan Mahasiswa; Decision Tree; CART; Klasifikasi; Machine Learning

Abstract−Timely student graduation is a key indicator of higher education quality and institutional effectiveness. This study aims to optimize student graduation prediction using a Decision Tree algorithm based on Classification and Regression Tree (CART) by integrating academic and non-academic variables. The dataset used in this study is the open-source *Student Graduation Dataset* obtained from Kaggle, consisting of 379 student records with graduation status as the target variable. The research stages include data preprocessing through *mean imputation* for missing values, categorical variable transformation, data splitting with an 80:20 ratio, and model optimization using CART *hyperparameter tuning* as a form of post-pruning. Model performance was evaluated using accuracy, precision, recall, F1-score, and a confusion matrix. The experimental results show that the optimized CART model achieved an accuracy of 92.1%, with F1-scores above 0.90 for both graduation classes and a balanced trade-off between precision and recall. Furthermore, the resulting decision tree structure is relatively simple and highly interpretable. These findings indicate that the optimized CART algorithm is effective and suitable for implementation as an early warning system to support academic decision-making in higher education institutions.

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

1. PENDAHULUAN

Kelulusan mahasiswa tepat waktu merupakan salah satu indikator kinerja utama perguruan tinggi yang berpengaruh langsung terhadap mutu akademik, efisiensi penyelenggaraan pendidikan, serta penilaian akreditasi institusi [1]. Rendahnya persentase kelulusan tepat waktu tidak hanya berdampak pada meningkatnya beban biaya pendidikan mahasiswa, tetapi juga menjadi sinyal adanya permasalahan pada proses pembelajaran, manajemen akademik, maupun faktor pendukung lainnya [2]. Dalam praktiknya, perguruan tinggi sering kali baru menyadari risiko keterlambatan kelulusan ketika mahasiswa telah memasuki akhir masa studi, sehingga peluang intervensi menjadi terbatas dan kurang efektif [3], [4].

Permasalahan kelulusan mahasiswa bersifat kompleks dan dipengaruhi oleh berbagai faktor akademik maupun non-akademik [5]. Faktor akademik meliputi indeks prestasi semester, indeks prestasi kumulatif, serta konsistensi capaian belajar pada semester menengah hingga akhir. Sementara itu, faktor non-akademik seperti status pekerjaan, usia, dan kondisi sosial mahasiswa juga berpotensi memengaruhi kemampuan mahasiswa dalam menyelesaikan studi sesuai waktu ideal. Pendekatan konvensional yang mengandalkan evaluasi manual cenderung bersifat subjektif, tidak terstruktur, dan kurang mampu menangkap pola tersembunyi dalam data akademik berskala besar [6].

Sebagai solusi atas permasalahan tersebut, pendekatan *data mining* dan *machine learning* menawarkan mekanisme analisis berbasis data historis yang lebih sistematis dan objektif. Salah satu metode yang banyak digunakan dalam prediksi kelulusan mahasiswa adalah algoritma *Decision Tree[7]*, karena memiliki kemampuan klasifikasi yang baik serta keunggulan dalam hal interpretabilitas aturan keputusan. Algoritma *Classification and Regression Tree* (CART) [8], secara khusus menggunakan pemisahan biner berbasis *Gini Index* yang efektif dalam menangani data numerik maupun kategorikal, serta menghasilkan struktur pohon keputusan yang mudah dipahami oleh pengambil kebijakan akademik [9].

Sejumlah penelitian terdahulu telah membahas prediksi kelulusan mahasiswa menggunakan berbagai pendekatan *data mining* dan *machine learning*. Hendra *et al.* menerapkan algoritma *Decision Tree* yang

Journal of Information System Research (JOSH)
Volume 7, No. 2, January 2026, pp 298−305
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9099

dikombinasikan dengan *Particle Swarm Optimization* (PSO) untuk meningkatkan akurasi prediksi kelulusan mahasiswa dan melaporkan adanya peningkatan akurasi dibandingkan penggunaan *Decision Tree* tanpa optimasi [10]. Meskipun demikian, pendekatan tersebut menghasilkan model yang lebih kompleks dan kurang menekankan aspek interpretabilitas aturan keputusan, yang justru penting dalam konteks pengambilan kebijakan akademik.

Penelitian lain oleh Nurhasanah *et al.* [1], menggunakan algoritma *Decision Tree* C4.5 berbasis data akademik dengan validasi *10-fold cross validation* dan menunjukkan performa klasifikasi yang cukup tinggi dengan akurasi di atas 88%. Studi tersebut menegaskan bahwa variabel akademik seperti IPK dan nilai ujian memiliki kontribusi dominan terhadap kelulusan mahasiswa. Namun, fokus penelitian lebih diarahkan pada pengujian performa algoritma C4.5, tanpa mengkaji secara mendalam keseimbangan antara *precision* dan *recall* pada masing-masing kelas kelulusan, khususnya untuk mahasiswa berisiko terlambat lulus.

Pendekatan yang lebih kompleks juga ditunjukkan dalam penelitian oleh Rahman *et al.*[3], yang menggunakan metode cart klasifikasi waktu kelulusan mahasiswa dan memperoleh peningkatan akurasi yang signifikan. Meskipun hasil yang diperoleh sangat baik dari sisi akurasi, penggunaan model *ensemble* cenderung menghasilkan sistem yang bersifat *black-box*, sehingga sulit diinterpretasikan oleh pengelola akademik dalam memahami faktor-faktor utama yang memengaruhi kelulusan mahasiswa.

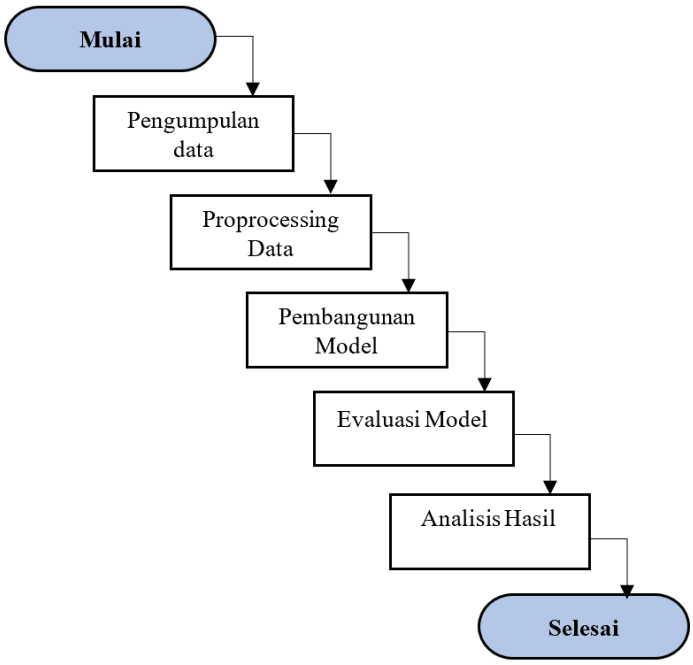
Oleh karena itu, penelitian ini mengusulkan penggunaan algoritma Decision Tree berbasis Classification and Regression Tree (CART) dengan fokus pada optimalisasi performa klasifikasi melalui pengendalian kompleksitas pohon keputusan (post-pruning). Optimalisasi dilakukan dengan pengaturan hyperparameter utama CART, meliputi pembatasan kedalaman maksimum pohon (maximum tree depth), jumlah minimum data pada node pemisah (minimum samples per split), dan jumlah minimum data pada node daun (minimum samples per leaf). Pendekatan ini bertujuan menghasilkan model yang tidak hanya akurat dan stabil, tetapi juga tetap mudah diinterpretasikan sebagai sistem peringatan dini dalam pengelolaan akademik perguruan tinggi..

2. METODOLOGI PENELITIAN

Metode penelitian disusun untuk menggambarkan alur sistematis dalam menyelesaikan permasalahan prediksi kelulusan mahasiswa berbasis data akademik dan non-akademik. Tahapan penelitian dirancang mulai dari pengumpulan data, praproses, pembangunan model, hingga evaluasi kinerja model klasifikasi. Pendekatan ini bertujuan memastikan bahwa metode yang diterapkan mampu menghasilkan model prediksi yang akurat, stabil, serta mudah diinterpretasikan sebagai dasar pengambilan keputusan akademik.

2.1 Tahapan Penelitian

Tahapan penelitian pada studi ini disusun secara berurutan untuk memastikan proses penerapan metode *machine learning* berjalan sistematis dan dapat direplikasi. Alur penelitian dimulai dari identifikasi permasalahan hingga evaluasi hasil model prediksi. Secara umum, tahapan penelitian yang digunakan ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian Prediksi Kelulusan Mahasiswa

Berdasarkan Gambar 1, tahapan penelitian dijelaskan sebagai berikut.

a. Pengumpulan Data

Journal of Information System Research (JOSH)
Volume 7, No. 2, January 2026, pp 298−305
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9099

Data yang digunakan dalam penelitian ini merupakan data historis mahasiswa yang diperoleh dari dataset terbuka "Student Graduation Dataset" yang tersedia pada platform Kaggle. Dataset ini berisi 379 data mahasiswa dengan 15 atribut, yang mencakup variabel akademik dan non-akademik, seperti jenis kelamin, usia, status pernikahan, status mahasiswa, indeks prestasi semester (IPS 1–IPS 8), indeks prestasi kumulatif (IPK), serta status kelulusan sebagai variabel target. Status kelulusan diklasifikasikan ke dalam dua kelas, yaitu lulus tepat waktu dan lulus terlambat. Dataset ini merepresentasikan data mahasiswa jenjang pendidikan tinggi dari institusi pendidikan di luar Indonesia, sehingga digunakan sebagai data simulasi untuk menguji performa dan karakteristik algoritma CART dalam konteks prediksi kelulusan mahasiswa.

b. Pra-pemrosesan Data
Tahap pra-pemrosesan bertujuan meningkatkan kualitas data sebelum digunakan dalam pemodelan. Proses ini mencakup pembersihan data, penanganan *missing value[11]*, penghapusan atribut yang tidak relevan, serta transformasi variabel kategorikal ke dalam bentuk numerik. Setelah seluruh tahapan pra-pemrosesan selesai, dataset dibagi menjadi data latih dan data uji dengan rasio 80:20, di mana 80% data digunakan untuk proses pelatihan model dan 20% data digunakan untuk pengujian. Pembagian data ini dilakukan untuk memastikan evaluasi kinerja model dilakukan secara objektif terhadap data yang tidak dilibatkan dalam proses pelatihan.[12]

c. Pembangunan Model Klasifikasi
Pada tahap ini, model klasifikasi dibangun menggunakan algoritma Decision Tree dengan pendekatan Classification and Regression Tree (CART). Proses pelatihan dilakukan dengan memanfaatkan **data latih** untuk mempelajari pola hubungan antara variabel prediktor dan status kelulusan mahasiswa..

d. Evaluasi dan Pengujian Model
Model yang telah dibangun dievaluasi menggunakan metrik kinerja klasifikasi, seperti *accuracy*, *precision*, *recall*, dan *F1-score*.[13] Selain itu, *confusion matrix* digunakan untuk menganalisis kesalahan klasifikasi yang terjadi. Evaluasi ini bertujuan menilai kemampuan model dalam melakukan generalisasi terhadap data baru.[14]

e. Analisis Hasil
Tahap akhir penelitian berfokus pada analisis hasil prediksi dan interpretasi model. Analisis *feature importance* digunakan untuk mengidentifikasi variabel yang paling berpengaruh terhadap kelulusan mahasiswa. Hasil analisis ini diharapkan dapat memberikan wawasan yang relevan bagi pengelola akademik sebagai dasar pengambilan keputusan dan perancangan intervensi dini.

2.2 Algoritma Decision Tree CART

Algoritma *Decision Tree* merupakan salah satu metode klasifikasi dalam *machine learning* yang membangun model prediksi berbentuk struktur pohon keputusan [15]. Setiap node merepresentasikan atribut, setiap cabang menunjukkan hasil pengujian atribut, dan setiap *leaf node* menyatakan kelas hasil prediksi. Keunggulan utama algoritma ini terletak pada kemampuannya menghasilkan aturan keputusan yang mudah dipahami dan diinterpretasikan.

Pada penelitian ini digunakan algoritma Classification and Regression Tree (CART), yang melakukan proses pemisahan data secara biner pada setiap node dengan menggunakan Gini Index sebagai kriteria pemilihan atribut. Nilai Gini Index yang lebih kecil menunjukkan tingkat ketidakmurnian data yang lebih rendah [16], sehingga atribut dengan nilai *Gini Index* minimum dipilih sebagai pemisah pada node tertentu [17].

Proses pembentukan pohon CART diawali dengan penentuan *root node* yang mencakup seluruh data latih. Selanjutnya, data dipisahkan berdasarkan atribut yang menghasilkan penurunan Gini Index terbesar. Proses ini dilakukan secara rekursif hingga memenuhi kondisi berhenti tertentu, seperti batas kedalaman pohon atau jumlah minimum data pada node daun..

Dalam penelitian ini, optimalisasi performa klasifikasi dilakukan melalui pengendalian kompleksitas pohon keputusan (*post-pruning*) berbasis pengaturan hyperparameter CART. Parameter yang dikendalikan meliputi kedalaman maksimum pohon (*maximum tree depth*), jumlah minimum data pada node pemisah (*minimum samples per split*), dan jumlah minimum data pada node daun (*minimum samples per leaf*). Berdasarkan hasil pengujian awal, konfigurasi dengan kedalaman pohon maksimum sekitar 4–5 tingkat dipilih karena mampu menghasilkan model yang stabil, tidak terlalu kompleks, dan memiliki kemampuan generalisasi yang baik terhadap data uji, sekaligus tetap mempertahankan interpretabilitas model [18].

3. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil penerapan metode *Decision Tree* berbasis CART[19] dalam memprediksi kelulusan mahasiswa, serta pembahasan mendalam terhadap kinerja model yang dihasilkan. Penyajian hasil dilakukan secara sistematis mengikuti tahapan metodologi penelitian yang telah dijelaskan pada bagian sebelumnya, mulai dari hasil pra-pemrosesan data, pembentukan model, hingga evaluasi performa klasifikasi. Pembahasan difokuskan pada interpretasi hasil eksperimen, analisis pengaruh variabel prediktor, serta perbandingan temuan penelitian dengan studi sebelumnya yang relevan.

Journal of Information System Research (JOSH)
Volume 7, No. 2, January 2026, pp 298−305
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9099

3.1 Hasil Penerapan Metodologi Penelitian

Penerapan metodologi penelitian menghasilkan serangkaian keluaran yang merepresentasikan kinerja model prediksi kelulusan mahasiswa. Proses ini dimulai dari data yang telah melalui tahap pra-pemrosesan hingga menghasilkan model klasifikasi yang siap dievaluasi.

a. Hasil Pra-pemrosesan Data

Sebelum melakukan pra-pemrosesan data maka ada pengumpulan data dengan melakukan pemilihan dataset. Dataset yang digunakan dalam penelitian ini berasal dari platform kaggle yang berisi tentang prediksi kelulusan mahasiswa yang didalamnya berisi 379 baris data dengan 15 kolom fitur meliputi Nama, Jenis Kelamin, Status Mahasiswa, Umur, Status Nikah, IPS 1, IPS 2, IPS 3, IPS 4, IPS 5, IPS 6, IPS 7, IPS 8, IPK dan Status Kelulusan sebagai label atau target prediksi yang menunjukkan apakah mahasiswa dinyatakan "Terlambat" atau "Tepat" waktu dalam hal kelulusan. Dalam pra-pemrosesan terdapat cleaning data dengan menguji adanya missing values[20] dari kolom-kolom data yang merupakan kategorikal dan numerik[21].setelah dilakukan pengecekan maka diketahui terdapat missing value pada kolom IPS 8 dan IPK sehingga total baris sebanyak 379 namun pada IPS 8 hanya terbaca sebanyak 372 nilai sehingga ada data kosong sebanyak 7 data. Sedangkan pada kolom IPK terbaca sebanyak 376 data sehingga ada data kosong sebanyak 3 data. Dapat dilihat dari gambar 2 hasil missing values dari data yang ada.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 379 entries, 0 to 378
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   JENIS KELAMIN     379 non-null    object
 1   STATUS MAHASISWA  379 non-null    object
 2   UMUR              379 non-null    int64
 3   STATUS NIKAH      379 non-null    object
 4   IPS 1             379 non-null    float64
 5   IPS 2             379 non-null    float64
 6   IPS 3             379 non-null    float64
 7   IPS 4             379 non-null    float64
 8   IPS 5             379 non-null    float64
 9   IPS 6             379 non-null    float64
 10  IPS 7             379 non-null    float64
 11  IPS 8             372 non-null    float64
 12  IPK               376 non-null    float64
 13  STATUS KELULUSAN  379 non-null    object
dtypes: float64(9), int64(1), object(4)
memory usage: 41.6+ KB
```

Gambar 2. Missing Values

Untuk data yang kosong ini maka dilakukan pendekatan mean imputation[22] yang telah di buktikan beberapa penelitian, mean mempertahankan akurasi klasifikasi lebih baik dibanding pengisian nol atau penghapusan data. Sehingga hasil dapat kita lihat pada Gambar 3.

```
JENIS KELAMIN       0
STATUS MAHASISWA    0
UMUR                0
STATUS NIKAH        0
IPS 1               0
IPS 2               0
IPS 3               0
IPS 4               0
IPS 5               0
IPS 6               0
IPS 7               0
IPS 8               0
IPK                 0
STATUS KELULUSAN    0
dtype: int64
```

Gambar 3. Hasil penerapan inputan nilai mean

Setelah melakukan missing values, selanjutnya adalah melakukan label encoding , karena variabel kategorikal perlu dilakukan encode ke nilai numerik sebelum pelatihan model. Langkah yang dilakukan berikutnya adalah

Journal of Information System Research (JOSH)
Volume 7, No. 2, January 2026, pp 298−305
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9099

pemisahan fitur (X) sebagai input dan target (Y) sebagai 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 dengan melabelkan status kelulusan.

b. Pembagian Data Latih dan Data Uji
Dataset dibagi menjadi dua bagian, yaitu data latih dan data uji dengan rasio 80:20. Pembagian ini bertujuan untuk mengevaluasi kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Data latih digunakan untuk membangun struktur pohon keputusan, sedangkan data uji digunakan sebagai dasar evaluasi performa model.

c. Pembentukan Model Decision Tree CART
Pada tahap pembangunan model menggunakan algoritma CART (Classification and Regression Tree)[23] dimana algoritma ini digunakan sebagai model klasifikasi karena struktur pohonnya mudah dipahami dan diterapkan. Algoritma ini bekerja dengan membentuk pohon keputusan dari himpunan data, dimana setiap cabang merepresentasikan aturan-aturan keputusan. Pohon dibangun secara rekursif, dan setiap node berisi pertanyaan atau kondisi berdasarkan atribut, hingga mencapai kedalaman tertentu atau tidak ada pemisahan yang signifikan dengan Leaf node sebagai prediksi kelas akhirnya. Model *Decision Tree* dibangun menggunakan pendekatan CART dengan kriteria pemisahan berbasis *Gini Index*. Proses pelatihan menghasilkan struktur pohon keputusan yang mampu mengelompokkan mahasiswa ke dalam dua kelas utama, yaitu lulus tepat waktu dan terlambat lulus. Struktur pohon yang terbentuk menunjukkan bahwa tidak seluruh variabel memiliki kontribusi yang sama dalam menentukan hasil prediksi.

3.2 Implementasi dan pengujian model

Tahap implementasi dan pengujian difokuskan pada evaluasi kinerja model *Decision Tree* CART dalam memprediksi kelulusan mahasiswa. Evaluasi dilakukan menggunakan beberapa metrik klasifikasi untuk memperoleh gambaran performa model secara menyeluruh.

3.2.1 Pembentukan Model

Berdasarkan hasil evaluasi menggunakan confusion matrix dan metrik klasifikasi, diperoleh performa model CART sebagaimana ditunjukkan pada Tabel 1

Tabel 1. Hasil Akurasi pemodelan

Kelas	Precision	Recall	F1-Score	Support
0 (Tepat)	0.89	0.98	0.93	43
1 (Terlambat)	0.97	0.85	0.90	33
Accuracy			0.92	76
Macro Avg	0.93	0.91	0.92	76
Weighted Avg	0.93	0.92	0.92	76

Berdasarkan Tabel 1, berikut adalah penjelasan detail mengenai performa model:

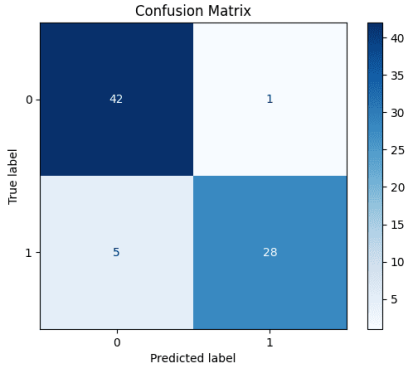
a. Prediksi Kelulusan "TEPAT" (Kelas 0):
Pada kelas kelulusan tepat waktu (Kelas 0), model menghasilkan nilai precision sebesar 89% dan recall sebesar 98%. Hal ini menunjukkan bahwa sebagian besar mahasiswa yang diprediksi lulus tepat waktu sesuai dengan kondisi aktual, serta hampir seluruh mahasiswa yang benar-benar lulus tepat waktu berhasil teridentifikasi oleh model. Nilai F1-score sebesar 93% mengindikasikan performa klasifikasi yang stabil pada kelas ini.

b. Prediksi Kelulusan "TERLAMBAT" (Kelas 1):
Pada kelas kelulusan terlambat (Kelas 1), model memperoleh nilai precision sebesar 97% dan recall sebesar 85%. Hasil ini menunjukkan bahwa prediksi mahasiswa berisiko terlambat lulus memiliki tingkat ketepatan yang sangat tinggi, meskipun masih terdapat sebagian kecil mahasiswa terlambat yang belum terdeteksi. Nilai F1-score sebesar 90% mencerminkan keseimbangan yang baik antara ketepatan dan sensitivitas model.

3.2.2 Confusion Matrix

Confusion matrix dapat digunakan untuk mengevaluasi keakuratan dari hasil klasifikasi. Matriks ini memungkinkan hasil analisis seberapa tepat classifier mengidentifikasi masing- masing kelas. Evaluasi kinerja model Decision Tree berbasis CART dilakukan menggunakan *confusion matrix* dan metrik klasifikasi yang meliputi *accuracy*, *precision*, *recall*, dan *F1-score*. Berdasarkan hasil evaluasi, model mencapai tingkat akurasi sebesar 92%, yang menunjukkan kemampuan prediksi yang baik secara keseluruhan. Pada kelas kelulusan terlambat sebagai kelas positif, model menghasilkan nilai *precision* sebesar 96,5%, *recall* sebesar 84,8%, dan *F1-score* sebesar 90,2%, yang mengindikasikan bahwa model memiliki tingkat ketepatan prediksi yang tinggi serta kemampuan yang cukup baik dalam mengidentifikasi mahasiswa berisiko mengalami keterlambatan kelulusan. Kombinasi nilai metrik tersebut menunjukkan bahwa model CART yang dioptimalkan melalui pengaturan hyperparameter mampu menghasilkan performa klasifikasi yang seimbang dan stabil dalam mendeteksi kedua kelas kelulusan.

Journal of Information System Research (JOSH)
Volume 7, No. 2, January 2026, pp 298−305
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9099



Gambar 3. Confusion Matrix

3.3 Pembahasan

Hasil penelitian menunjukkan bahwa algoritma Decision Tree berbasis Classification and Regression Tree (CART) mampu memberikan performa klasifikasi yang baik dalam memprediksi kelulusan mahasiswa. Pencapaian performa ini diperoleh melalui optimalisasi pengaturan hyperparameter CART sebagai bentuk post-pruning, yang berperan dalam mengendalikan kompleksitas model dan meningkatkan kemampuan generalisasi. Pada model akhir, parameter yang digunakan meliputi kedalaman maksimum pohon keputusan (*maximum tree depth*) sekitar 4–5 tingkat, serta pengaturan jumlah minimum data pada node pemisah dan node daun, sehingga dihasilkan model yang tidak terlalu kompleks namun tetap stabil. Model menghasilkan akurasi sebesar 92%, dengan nilai *precision*, *recall*, dan *F1-score* yang seimbang pada kedua kelas kelulusan.

Pada kelas kelulusan tepat waktu (Kelas 0), model menghasilkan nilai *precision* sebesar 89% dan *recall* sebesar 98%, yang menunjukkan bahwa sebagian besar mahasiswa yang diprediksi lulus tepat waktu sesuai dengan kondisi aktual serta hampir seluruh mahasiswa yang benar-benar lulus tepat waktu berhasil teridentifikasi. Nilai *F1-score* sebesar 93% mengindikasikan performa klasifikasi yang stabil pada kelas ini.

Sementara itu, pada kelas kelulusan terlambat**,** model menghasilkan nilai *precision* sebesar 97%, yang menunjukkan tingkat keyakinan prediksi yang sangat tinggi ketika model mengklasifikasikan mahasiswa sebagai terlambat lulus. Nilai *recall* sebesar 85% mengindikasikan bahwa sebagian besar mahasiswa yang benar-benar mengalami keterlambatan kelulusan berhasil teridentifikasi oleh model, meskipun masih terdapat sejumlah kecil data yang tidak terdeteksi. Nilai *F1-score* sebesar 90% menunjukkan bahwa model memiliki keseimbangan yang baik antara ketepatan prediksi dan sensitivitas terhadap kelas berisiko, sehingga dapat diandalkan dalam konteks identifikasi mahasiswa yang membutuhkan perhatian akademik lebih lanjut.

Analisis *confusion matrix* memperkuat temuan tersebut, di mana jumlah prediksi benar (*true positive* dan *true negative*) jauh lebih dominan dibandingkan kesalahan prediksi (*false positive* dan *false negative*). Jumlah kesalahan klasifikasi yang relatif kecil menunjukkan bahwa model memiliki kemampuan generalisasi yang baik terhadap data uji. Kesalahan prediksi yang masih muncul dapat dipahami sebagai konsekuensi dari kompleksitas faktor yang memengaruhi kelulusan mahasiswa, di mana tidak seluruh aspek akademik maupun non-akademik dapat direpresentasikan secara lengkap dalam dataset yang digunakan.

Jika ditinjau dari sisi metodologi, penerapan tahapan pra-pemrosesan data, khususnya penanganan *missing value* menggunakan pendekatan *mean imputation* dan transformasi variabel kategorikal ke dalam bentuk numerik, tidak menurunkan performa model secara signifikan. Dikombinasikan dengan optimalisasi parameter CART melalui post-pruning, model yang dihasilkan memiliki struktur pohon keputusan yang relatif sederhana dan mudah diinterpretasikan, sehingga memberikan nilai tambah dibandingkan model klasifikasi yang bersifat *black-box*.

Secara keseluruhan, pembahasan ini menegaskan bahwa algoritma Decision Tree CART yang dioptimalkan melalui pengaturan hyperparameter tidak hanya mampu menghasilkan tingkat akurasi yang tinggi, tetapi juga memberikan interpretasi yang jelas terhadap pola kelulusan mahasiswa. Dengan karakteristik tersebut, model yang dihasilkan berpotensi diterapkan sebagai alat bantu pengambilan keputusan akademik, khususnya dalam upaya meningkatkan persentase kelulusan mahasiswa tepat waktu melalui intervensi yang lebih terarah dan berbasis data.

Dengan demikian, optimalisasi model CART melalui pengendalian parameter pohon keputusan tidak hanya meningkatkan performa klasifikasi, tetapi juga menghasilkan struktur model yang lebih sederhana dan mudah diinterpretasikan. Hal ini memperkuat potensi penerapan model sebagai sistem peringatan dini (*early warning system*) dalam mengidentifikasi mahasiswa berisiko mengalami keterlambatan kelulusan.

4. KESIMPULAN

Penelitian ini bertujuan untuk mengoptimalkan prediksi kelulusan mahasiswa dengan memanfaatkan algoritma Decision Tree berbasis Classification and Regression Tree (CART) melalui integrasi data akademik dan non-akademik. Berdasarkan hasil eksperimen dan evaluasi kinerja model, dapat disimpulkan bahwa algoritma CART

Journal of Information System Research (JOSH)
Volume 7, No. 2, January 2026, pp 298−305
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9099

yang dioptimalkan mampu menghasilkan performa klasifikasi yang sangat baik dengan tingkat akurasi sebesar 92,1%**.** Selain itu, model menunjukkan keseimbangan performa antar kelas, dengan nilai F1-score di atas 90% pada kelas kelulusan tepat waktu maupun terlambat, yang menandakan kemampuan prediksi yang stabil dan andal. Optimalisasi prediksi dalam penelitian ini dicapai melalui beberapa tahapan utama, yaitu penerapan pra-pemrosesan data yang tepat melalui penanganan *missing value* menggunakan mean imputation, transformasi variabel kategorikal ke dalam bentuk numerik, serta pengendalian kompleksitas model CART melalui pengaturan hyperparameter sebagai bentuk post-pruning, khususnya pembatasan kedalaman pohon keputusan dan jumlah minimum data pada node. Pendekatan ini terbukti meningkatkan kemampuan generalisasi model tanpa mengorbankan interpretabilitas struktur pohon keputusan. Dengan karakteristik tersebut, model CART yang dihasilkan tidak hanya unggul dari sisi performa kuantitatif, tetapi juga memiliki nilai praktis yang tinggi sebagai sistem peringatan dini (early warning system) untuk mendukung pengambilan keputusan akademik, khususnya dalam mengidentifikasi mahasiswa yang berisiko mengalami keterlambatan kelulusan. Meskipun demikian, penelitian ini masih memiliki keterbatasan, terutama pada penggunaan dataset dari satu sumber terbuka dengan jumlah variabel non-akademik yang relatif terbatas, serta belum dilakukannya perbandingan dengan algoritma klasifikasi lain. Oleh karena itu, penelitian selanjutnya disarankan untuk memperluas sumber data, menambahkan variabel yang lebih beragam, serta melakukan evaluasi komparatif dengan metode machine learning lainnya guna memperoleh model prediksi kelulusan mahasiswa yang lebih komprehensif dan adaptif.

REFERENCES

[1] N. * Risky, D. Setiyawan, D. Hermawan, and O. Herdiyanto, “Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree C4.5 Berbasis Data Akademik dengan Validasi 10-Fold,” *TIN: Terapan Informatika Nusantara*, vol. 6, no. 6, pp. 670–678, Nov. 2025, doi: 10.47065/tin.v6i6.8662.

[2] J. Wang, Y. He, L. Yan, S. Chen, and K. Zhang, “Predicting Osteoporosis and Osteopenia by Fusing Deep Transfer Learning Features and Classical Radiomics Features Based on Single-Source Dual-energy CT Imaging,” *Acad Radiol*, vol. 31, no. 10, pp. 4159–4170, Oct. 2024, doi: 10.1016/j.acra.2024.04.022.

[3] G. A. Rahman, K. A. Notodiputro, B. Sartono, and L. Surimi, “CART and Random Forest Analysis on Graduation Status of Halu Oleo University Students,” *Inferensi*, vol. 8, no. 3, p. 271, Nov. 2025, doi: 10.12962/j27213862.v8i3.23336.

[4] F. Ariska, V. Sihombing, and I. Irmayani, “Student Graduation Predictions Using Comparison of C5.0 Algorithm With Linear Regression,” *SinkrOn*, vol. 7, no. 1, pp. 256–266, Feb. 2022, doi: 10.33395/sinkron.v7i1.11261.

[5] T. H. Hasibuan and D. Mahdiana, “Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma C4.5 Pada Uin Syarif Hidayatullah Jakarta,” *SKANIKA: Sistem Komputer dan Teknik Informatika*, vol. 6, pp. 61–74, 2023, Accessed: Jan. 13, 2026. [Online]. Available: https://jom.fti.budiluhur.ac.id/SKANIKA/article/view/2976

[6] D. Kurniawan, A. Anggrawan, and H. Hairani, “Graduation Prediction System On Students Using C4.5 Algorithm,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 19, no. 2, pp. 358–365, May 2020, doi: 10.30812/matrik.v19i2.685.

[7] A. S. R. Siregar, Y. S. Siregar, and M. Khairani, “Implementation Of The Data Mining Cart Algorithm In The Characteristic Pattern Of New Student Admissions,” *Journal of Computer Networks, Architecture and High Performance Computing*, vol. 5, no. 1, pp. 263–275, Feb. 2023, doi: 10.47709/cnahpc.v5i1.1975.

[8] S. Sarbaini and F. Ulfa, “STUDENT GRADUATION PREDICTION USING DECISION TREE METHOD WITH C4.5 ALGORITHM,” *Jurnal Diferensial*, vol. 6, no. 1, pp. 9–15, Jan. 2024, doi: 10.35508/jd.v6i1.12287.

[9] D. A. Wagner, T. Nair, A. Thapa, and A. Kumar, “The Gini Learning Index: A new framework to measure learning inequality across contexts,” *Int J Educ Dev*, vol. 119, pp. 1–12, Nov. 2025, doi: 10.1016/j.ijedudev.2025.103433.

[10] M. Abdul Azis and P. Studi Ilmu Komputer STMIK Nusa Mandiri, “Analisis Prediksi Kelulusan Mahasiswa Menggunakan Decission Tree Berbasis Particle Swarm Optimization,” *Sisfokom : Sistem Informasi dan Komputer*, vol. 09, pp. 102–107, doi: 10.32736/sisfokom.v9.i1.

[11] A. Tashk, K. M. Sørensen, S. B. Engelsen, K. S. Pedersen, and C. E. Eskildsen, “MIPLS2: Exploiting PLS2 to impute missing values in a two-block system with multiple response variables,” *Anal Chim Acta*, vol. 1364, pp. 1–10, Aug. 2025, doi: 10.1016/j.aca.2025.344134.

[12] B. Baydil, V. H. de la Peña, H. Zou, and H. Yao, “Unbiased estimation of the Gini coefficient,” *Stat Probab Lett*, vol. 222, pp. 1–9, Jul. 2025, doi: 10.1016/j.spl.2025.110376.

[13] I. Irumas and J. N. Utamajaya, “Penerapan Metode EUCS Untuk Evaluasi Tingkat Kepuasan Pengguna Aplikasi PNM Digi Karyawan,” *Journal of Computer System and Informatics (JoSYC)*, vol. 4, no. 1, pp. 101–108, Dec. 2022, doi: 10.47065/josyc.v4i1.2492.

[14] S. Kim, T. H. Lee, and J. Lee, “TMF-GNN: Temporal matrix factorization-based graph neural network for multivariate time series forecasting with missing values,” *Expert Syst Appl*, vol. 275, pp. 1–11, May 2025, doi: 10.1016/j.eswa.2025.127001.

[15] Y. B. L. Kintomonho, M. N. Atchadé, and D. Daddah, “Decision tree-based statistical learning and quantile regression adjustment: Insights from pregnant women in Benin,” *Sci Afr*, vol. 29, pp. 1–10, Sep. 2025, doi: 10.1016/j.sciaf.2025.e02832.

[16] A. J. F. Martin and T. M. Conway, “Using the Gini Index to quantify urban green inequality: A systematic review and recommended reporting standards,” Feb. 01, 2025, *Elsevier B.V.* doi: 10.1016/j.landurbplan.2024.105231.

[17] J. M. Gavilan-Ruiz, Á. Ruiz-Gándara, F. J. Ortega-Irizo, and L. Gonzalez-Abril, “Some Notes on the Gini Index and New Inequality Measures: The nth Gini Index,” *Stats (Basel)*, vol. 7, no. 4, pp. 1354–1365, Dec. 2024, doi: 10.3390/stats7040078.

Journal of Information System Research (JOSH)
Volume 7, No. 2, January 2026, pp 298−305
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.9099

[18] M. Mesran and D. P. Indini, "Analisis Dalam Pendukung Keputusan Seleksi Content Creator Mahasiswa Terbaik Menerapkan Metode EDAS dan ROC," *Journal of Computer System and Informatics (JoSYC)*, vol. 4, no. 4, pp. 912–921, 2023, doi: 10.47065/josyc.v4i4.4093.

[19] T. H. Wu, P. Y. Chen, C. C. Chen, M. J. Chung, Z. K. Ye, and M. H. Li, "Classification and Regression Tree (CART)-based estimation of soil water content based on meteorological inputs and explorations of hydrodynamics behind," *Agric Water Manag*, vol. 299, pp. 1–17, Jun. 2024, doi: 10.1016/j.agwat.2024.108869.

[20] N. J. Downing, "Missing value imputation in environmental, social, and governance data: an impact on emissions scores," *Financ Res Lett*, vol. 85, pp. 1–10, Nov. 2025, doi: 10.1016/j.frl.2025.107818.

[21] W. Li, G. Subašić, I. Korolija, R. Guida, and S. M. Hong, "Data imputation methods for missing U-values of building envelopes in building performance database," *Journal of Building Engineering*, vol. 118, pp. 1–6, Jan. 2026, doi: 10.1016/j.jobe.2025.115046.

[22] A. Kumar, S. Bhushan, R. Pokhrel, A. I. Al-Omari, A. R. A. Alanzi, and S. S. Alshqaq, "Imputation of missing data for domain mean estimation using simple random sampling," *Kuwait Journal of Science*, vol. 52, no. 4, pp. 1–10, Oct. 2025, doi: 10.1016/j.kjs.2025.100461.

[23] H. Tian *et al.*, "Classification and regression tree (CART) for predicting cadmium (Cd) uptake by rice (Oryza sativa L.) and its application to derive soil Cd threshold based on field data," *Ecotoxicol Environ Saf*, vol. 285, pp. 1–8, Oct. 2024, doi: 10.1016/j.ecoenv.2024.117125.