Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3622~3631
e-ISSN: 2548-6861

Analysis of Factors Affecting the Delay in Completion of Student Final Projects Using the C5.0 Decision Tree Algorithm

Chaidir Chalaf Islamy [1]*, Mochamad Choirul Anwar [2]*
* Teknik Informatika, Universitas 17 Agustus 1945 Surabaya
chaidirc@untag-sby.ac.id [1], choirulanwar1140@gmail.com [2]

Article Info

Article history:

Received 2025-07-22
Revised 2025-10-27
Accepted 2025-11-08

Keyword:

Students,
Final Project,
C5.0 Algorithm,
Decision Tree

ABSTRACT

Delays in completing final projects are a common problem faced by students and can lead to delayed graduation, increased study load, and reduced readiness to enter the workforce. This study uses a quantitative predictive approach to analyze the factors influencing delays in completing student final projects by applying the C5.0 Decision Tree classification algorithm. Data were collected through a Likert-scale questionnaire from 204 students of the Faculty of Engineering, University of 17 August 1945 Surabaya, who graduated between 2019 and 2021. The analyzed factors include time management, student motivation, campus policies, faculty support, family support, surrounding environment, and academic skills. The C5.0 algorithm was selected for its higher accuracy and efficiency compared to earlier methods such as C4.5 and CART. The results show that the Surrounding Environment factor is the most dominant, followed by Student Motivation, Time Management, and Family Support. Evaluation of the model yielded excellent classification performance, achieving an accuracy of 95.31%, precision of 96.77%, recall of 93.75%, and an F1-score of 95.24%. These results indicate that the model effectively classifies students at risk of delay with strong predictive reliability. The findings provide insights for universities to develop targeted strategies to enhance student motivation, improve time management, and create a more supportive academic environment. In conclusion, the C5.0 algorithm demonstrates a strong capability to identify dominant delay factors and supports data-driven decision-making in academic management.

I. PENDAHULUAN

Tugas akhir (TA) adalah suatu karya ilmah ataupun proyek yang disusun oleh mahasiswa untuk menuntaskan pendidikan di perguruan tinggi. Penyelesaian tugas akhir sangat penting bagi mahasiswa karena merupakan satu tahap penting dalam perjalanan akademik, yang menjadi syarat kelulusan pada program sarjana. Namun, realitas diberbagai perguruan tinggi menunjukan bahwa banyak mahasiswa yang mengalami keterlambatan dalam menyelesaikan tugas akhir mereka. Berdasarkan data dari " Kantor Fakultas Teknologi Industri Universitas Ahmad Dahlan pada tahun 2019, menunjukan data bahwa rata-rata keterlambatan studi yang dialami oleh mahasiswa Angkatan 2011, 2012, 2013, dan 2014 adalah mencapai nilai sebanyak 91,92%" [1]. Fenomena ini memiliki dampak yang serius, termasuk peningkatan biaya Pendidikan, tertundanya untuk masuk ke dunia kerja, hingga potensi menurunnya motivasi akademik mahasiswa

Keterlambatan penyelesaian TA tidak hanya dipengaruhi oleh faktor individu, tetapi juga oleh efektivitas kebijakan akademik dan dukungan dari lingkungan sekitar mahasiswa. Menurut , kurangnya fleksibilitas jadwal konsultasi dosen dan minimnya dukungan keluarga menjadi hambatan utama. Sementara itu, faktor internal seperti rendahnya motivasi dan kecemasan yang berlebihan turut memperlambat proses penyelesaian [2]. Faktor lain seperti hubungan interpersonal dengan dosen pembimbing, pengaruh teman sebaya, serta keterampilan akademik juga berperan dalam memperbesar risiko keterlambatan [3].

Pada penelitian ini, algoritma yang digunakan C5.0 untuk membentuk sebuah *Decision Tree* (DT) dengan menggunakan Bahasa pemrograman R programming. R Programming merupakan Bahasa pemrograman berorientasi objek yang terinterpretasi dan interaktif. Peneliti memilih bahasa ini karena R memiliki banyak library dan package statistik yang kuat seperti C50, rpart, dan caret, yang sangat mendukung dalam proses pengolahan dan pemodelan data. Selain itu, R juga menyediakan berbagai fitur visualisasi yang sangat berguna dalam menganalisis hasil dan mendukung interpretasi data secara menyeluruh [4].

Sedangkan *Decision Tree* (DT) adalah sebuah metode berbasiskan pohon yang digunakan untuk membuat klasifikasi berdasarkan variable-variabel independent guna untuk memprediksi variable target. Dalam data mining, metode Decision Tree (DT) menjadi salah satu metode yang sering digunakan karena kemampuannya untuk menyederhanakan pengambilan keputusan dengan menampilkan aturan *"if-else"* yang mudah untuk dipahami [5].

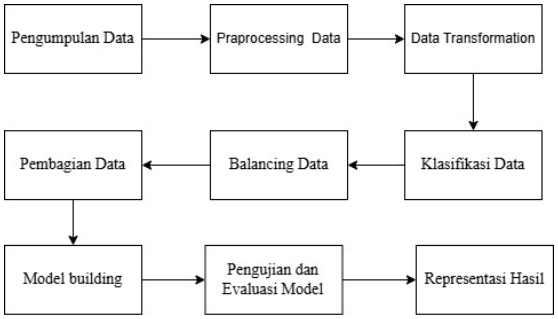
Pada penelitian sebelumnya telah menggunakan pendekatan data mining, khususnya algoritma klasifikasi, untuk menganalisis penyebab keterlambatan studi. [6] menggunakan algoritma C4.5 dan mencatat akurasi sebesar 73,48%. Sementara itu, penelitian oleh [7] mengidentifikasi bahwa lingkungan teman sebaya merupakan faktor dominan dalam keterlambatan dengan akurasi mencapai 82%.

Berdasarkan kedua penelitian sebelumnya, algoritma *decision tree* C4.5 menunjukan tingkat akurasi yang tinggi. Namun, pada algoritma C4.5 masih memiliki keterbatasan dari segi efisiensi pemrosesan, kompleksitas hasil, dan interpretabilitas model. M aka sebagai pengembangan dari penelitian sebelumnya, algoritma *decision tree* C5.0 dipilih karena dinilai lebih unggul karena menghasilkan pohon keputusan yang lebih sederhana, akurat, dan cepat. C5.0 juga memiliki kemampuan melakukan pruning otomatis untuk mencegah overfitting serta mendukung pemrosesan atribut yang hilang[8].

Fokus pada penelitian ini untuk mengidentifikasi faktor dominan penyebab keterlambatan serta memberikan rekomendasi solusi berdasarkan hasil analisis dan visualisasi model pohon keputusan. Dengan demikian, hasil penelitian ini diharapkan dapat memberikan kontribusi strategis dalam upaya peningkatan efisiensi penyelesaian tugas akhir mahasiswa dan menjadi referensi kebijakan akademik yang lebih berbasis data.

II. METODE

Penelitian yang telah dilakukan menggunakan pendekatan kuantitatif dengan metode eksploratif untuk menganalisis faktor-faktor yang memengaruhi keterlambatan penyelesaian tugas akhir mahasiswa. Algoritma C5.0 dipilih karena memiliki keunggulan dibandingkan C4.5 maupun CART, yaitu menghasilkan pohon keputusan yang lebih sederhana, proses pelatihan yang lebih cepat, serta akurasi yang lebih tinggi. Selain itu, C5.0 juga memiliki fitur *automatic pruning* untuk menghindari *overfitting*.



Gambar 1. Alur Penelitian

Penelitian dilakukan melalui beberapa tahapan sistematis, meliputi pengumpulan data melalui kuesioner skala Likert, preprocessing data, data transformation, klasifikasi data, balancing data menggunakan metode SMOTE, pembagian data dengan perbandingan 80:20, model building menggunakan algoritma Decision Tree C5.0, Pengujian dan evaluasi model, dan representasi hasil . Pengolahan data dan pembuatan model klasifikasi dilakukan dengan menggunakan bahasa pemrograman R melalui lingkungan pengembangan Kaggle Notebook. Proses ini didukung oleh berbagai pustaka (library) seperti C50, caret, dan ggplot2 untuk keperluan klasifikasi dan visualisasi data. Model dievaluasi menggunakan confusion matrix, serta metrik performa seperti akurasi, *precision, recall,* dan *F1-score*. Diagram alur penelitian dapat pada gambar 1.

A. Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan untuk menggali informasi mengenai faktor-faktor yang mempengaruhi keterlambatan penyelesaian tugas akhir mahasiswa. Teknik pengumpulan data yang digunakan adalah penyebaran kuesioner kepada mahasiswa fakultas teknik Angkatan 2019, 2020, 2021 yang berbasis skala likert, dan terdiri dari pernyataan-pernyataan terstruktur berdasarkan variabel-variabel penelitian yang diambil dari penelitian terdahulu seperti variabel manajemen waktu yang dibuat berdasarkan jurnal terdahulu penelitian yang dilakukan oleh razali [9], variabel motivasi mahasiswa yang dibuat berdasarkan penelitian yang dilakukan oleh silva[10], variabel kebijakan kampus yang dibuat berdasarkan penelitian yang dilakukan oleh berthod [11], variabel dukungan dosen yang dibuat berdasarkan penelitian yang dilakukan oleh henricson[12], variabel dukungan keluarga yang dibuat berdasarkan penelitian yang dilakukan oleh hassan [13], variabel lingkungan sekitar yang dibuat berdasarkan penelitian yang dilakukan oleh Bankole Adeyemi [14], dan variabel keterampilan akademik yang dibuat berdasarkan penelitian yang dilakukan oleh nabizadeh [15]

Instrumen penelitian yang digunakan berupa kuisioner yang terdiri dari 21 butir pernyataan yang. Pada penelitian ini

menggunakan total 7 faktor utama atau variabel bebas yang memengaruhi keterlambatan penyelesaian tugas akhir yang meliputi:

- Manajemen waktu
- Motivasi mahasiswa
- Kebijakan kampus
- Dukungan dosen
- Dukungan keluarga
- Lingkungan sekitar
- Keterampilan akademik

Jumlah populasi yang digunakan adalah sebanyak 420 mahasiswa, dan teknik pengambilan sampel yang digunakan adalah Stratified Random Sampling, yaitu metode pengambilan sampel dengan membagi populasi ke dalam strata berdasarkan tahun Angkatan. Selanjutnya, digunakan rumus Slovin untuk menentukan jumlah minimum sampel yang representatif dengan *error tolerance* 5% dengan jumlah populasi [16]. Dengan formula:

$$n = \frac{N}{1 + N \cdot e^2} \quad (1)$$

n : Jumlah sampel yang dibutuhkan
N : Jumlah populasi
e : Tingkat kepercayaan ($Error\ tolerance$)

Maka,

$$n = \frac{400}{1 + 400 \cdot (0{,}05)^2} = \frac{400}{1 + 400 \cdot 0{,}0025} = \frac{400}{2} = 200$$

B. Uji Validitas dan Reliabilitas

Uji validitas dilakukan menggunakan korelasi Pearson (item-total correlation) terhadap 22 item pernyataan. Hasil menunjukkan bahwa seluruh item memiliki nilai korelasi (r) > 0,30 dan dinyatakan valid, kecuali satu item demografis ("Tahun Angkatan") yang dikeluarkan karena bukan indikator konstruk.

Uji reliabilias menggunakan metode Cronbach's Alpha, menghasilkan nilai α = 0,925, yang berarti tingkat reliabilitas sangat tinggi. Hal ini menunjukkan bahwa instrumen kuesioner memiliki konsistensi internal yang kuat dan dapat dipercaya untuk digunakan dalam analisis faktor.

C. *Praprocessing Data*

data mentah hasil kuesioner agar dapat diolah menggunakan algoritma *decision tree* C5.0, selain pada tahap ini akan dilakukan proses pelabelan sesuai dengan tujuh variabel penelitian. Setiap variabel terdiri atas tiga indikator, dan nilai akhir masing-masing variabel diperoleh melalui perhitungan rata-rata.

D. Klasifikasi Data

Tahap klasifikasi data merupakan kelanjutan dari proses pra-pemrosesan data, di mana setiap responden akan dikelompokkan ke dalam kategori "Tepat Waktu" atau "Terlambat" dalam penyelesaian tugas akhir. Penentuan kategori ini didasarkan pada perhitungan nilai interval dari total skor rata-rata yang diperoleh dari tujuh variabel utama yang dianalisis: Manajemen waktu, motivasi mahasiswa, kebijakan kampus, dukungan dosen, dukungan keluarga, lingkungan sekitar, dan keterampilan akademik.

Proses klasifikasi ini mengadopsi pendekatan statistik deskriptif [7]. Berdasarkan perhitungan, rentang interval ditentukan sebagai berikut:

TABEL I
RENTANG INTERVAL KLASIFIKASI KATEGORI

Rentang Skala	Kategori
17.6 - 28	Tepat Waktu
7 - 17.5	Terlambat

Setiap nilai rata-rata dari masing-masing faktor atau variabel diubah menjadi tiga tingkatan kualitatif: "Rendah", "Sedang", dan "Tinggi". Kategorisasi ini dilakukan berdasarkan interval nilai rata-rata sebagai berikut:

$Rendah: mean = 1.0 \leq x < 2.0$
$Sedang: mean = 2.0 \leq x < 3.0$
$Tinggi: mean = 3.0 \leq x < 4.0$

E. *Balancing Data*

Balancing data merupakan tahapan untuk mengatasi potensi ketidakseimbangan jumlah data antar kategori, yang dapat menyebabkan bias pada model klasifikasi dan memengaruhi performa algoritma (misalnya, membuat model cenderung memprediksi kelas mayoritas dan mengabaikan kelas minoritas)[18]. Dalam penelitian ini metode *Synthetic Minority Over-sampling Technique* (SMOTE) diterapkan untuk menyeimbangkan distribusi data. Dengan formula:

$$x_{new} = x_i + \delta.(x_{zi} - x_i) \quad (2)$$

x_i : Titik data minoritas
x_{zi} : Nilai minoritas terdekat dari x_i
δ : Bilangan acak interpolasi antara dua titik [0,1]
x_{new} : Data baru yang berada di antara x_i dan x_{zi}

Dataset hasil klasifikasi, yang meliputi variabel manajemen waktu, motivasi mahasiswa, kebijakan kampus, dukungan dosen, dukungan keluarga, lingkungan sekitar, dan keterampilan akademik, akan ditingkatkan jumlah sampel pada kelas minoritas, sehingga distribusi data antar kelas dapat mencapai keseimbangan yang lebih optimal.

F. Pembagian Data

Pembagian data adalah proses membagi dataset menjadi dua segmen yaitu data pelatihan (*training data*) dan data pengujian (*testing data*). Dalam penelitian ini, pembagian

data menggunakan rasio 80:20, meliputi 80% untuk data pelatihan dan 20% untuk data pengujian. Proses pembagian data dilakukan menggunakan fungsi createDataPartition() dari pustaka *caret* pada R, sedangkan proses pelatihan dan pengujian model dilakukan dengan pustaka *C50*. Evaluasi model dilakukan melalui *confusion matrix* yang menghasilkan metrik akurasi, precision, recall, dan F1-score Proporsi ini dipilih berdasarkan penelitian sebelumnya untuk memastikan kinerja model yang optimal dan hasil evaluasi yang akurat[19].

G. *Tie Breaking*

Tie Breaking merupakan pembobotan awal sebelum data diolah. Pada penelitian ini metode *Expert Judgement* untuk menentukan pembobotan awal. Hal ini berguna untuk mencegah terjadinya nilai Gain Ratio yang sama antar faktor dan mengantisipasi kondisi di mana dua faktor memiliki tingkat informasi yang setara. Urutan prioritas faktor yang ditetapkan meliputi motivasi mahasiswa, keterampilan akademik, manajemen waktu, dukungan dosen, dukungan keluarga, lingkungan sekitar, dan kebijakan kampus.

H. *Decision Tree C5.0*

Decision Tree (DT) adalah sebuah metode yang digunakan untuk membangun model prediktif atau klasifikasi berdasarkan atribut-atribut didalam sebuah data[8]. Algoritma C5.0 merupakan metode didalam data mining yang berbasis pada teknik DT dengan algoritma klasifikasi. Algoritma C5.0 merupakan penyempurnaan dari C4.5 yang usulkan oleh Ross Quilan pada tahun 1987 yang dirancang agar lebih cepat dan lebih hemat penyimpanan[20].

Tahapan algoritma C5.0 dimulai dari semua atribut yang ada dijadikan akar dari DT. Selanjutnya dipilih atribut yang memiliki hasil nilai *Gain* tertinggi untuk dijadikan sebagai *root node*. Proses ini akan diulang hingga mendapatkan *root node* lain. Proses klasifikasi pada C5.0 bergantung pada nilai Gain dan Entrophy dalam memilih atribut terbaik untuk pembagian data pada setiap node. Dengan Formula:

$$Entrophy\ (S) = -\sum_{i=1}^{k} p_i\ .\ log_2(p_i) \quad (3)$$

S : dataset yang sedang dianalisis
k : jumlah kelas dalam dataset
p_i : Proporsi data di kelas ke $-$ i

Maka *Information Gain* dari attribute A adalah,

$$Gain\ (S, A) = Entrophy\ (S) - \sum_{i=1}^{m} \frac{|S_i|}{|S|} Entrophy(Si) \quad (4)$$

m : jumlah kategori pada atribut A
A : atribut yang digunakan untuk membagi data
S_i : Subset pembagian berdasarkan kategori atribut A

Setelah mendapatkan *Gain dan Entrophy*, selanjutnya nilai gain ratio akan dihitung yang berguna sebagai perbaikan dari Information *Gain* untuk mengatasi bias terhadap atribut dengan banyak kategori.

$$GainRatio(S, A) = \frac{Gain\ (S, A)}{\sum_{i=1}^{m} Entrophy\ (S_i)}$$

I. *Confusion Matrix*

Confusion Matrix adalah sebuah hasil evaluasi dari klasifikasi uang digambarkan dalam sebuah tabel. *confusion matrix* adalah sebuah tabel yang digunakan untuk mengevaluasi performa model klasifikasi dengan cara mencatat prediksi terhadap data aktual pada masing-masing kelas. Pada dasarnya confusion matrix dibuat untuk memberikan evaluasi yang komprehensif terhadap model, baik dalam bentuk (*True Positive* (TP)*, True Negative* (TN)*), maupun (False Positive* (FP) *, False Negative* (FN))[21]. Confusion Matrix memiliki komponen utama seperti yang digambarkan pada tabel 2.

TABEL II
CONFUSION MATRIX

		Kelas Aktual	
		True	**False**
Kelola Hasil Prediksi	***False***	FN *(False Negative)*	TN (*True Negative)*
	True	TP *(True Positive)*	FP (*False Positive)*

(TP) yaitu data aktual positif dan berhasil diprediksi positif, (TN) yaitu Data aktual negatif dan diprediksi negative, (FP) yaitu Data aktual negatif tetapi diprediksi positif (*Type I Error*). (FN) yaitu Data aktual positif tetapi diprediksi negatif (*Type II Error*). Berdasarkan komponen utama dalam *confusion matrix*, dapat dihitung sejumlah metrik evaluasi kinerja model klasifikasi sebagai berikut

1) Accuracy: didapat semua hasil dari perhitungan nilai prediksi yang benar dibagi dengan keseluruhan data.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (6)$$

2) Sensitivity (Recall): merupakan hasil dari jumlah prediksi yang benar dibagi dengan seluruh jumlah kelas yang salah

$$Sensitivity\ (Recall) = \frac{TP}{TP+FN} \quad (7)$$

3) Precision (Positive Predictive Value): merupakan hasil dari perhitungan jumlah seluruh nilai produktif positif dibagi dengan keselutuhan prediksi kelas yang benar

$$Precision = \frac{TP}{TP+FN} \quad (8)$$

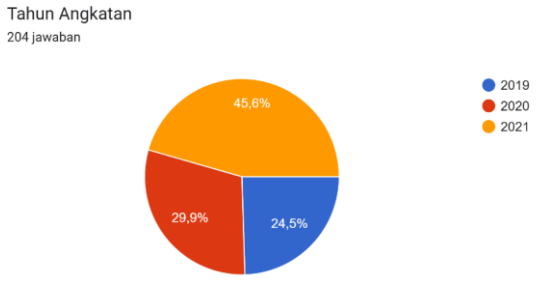
4) F-1 Score : merupakan matrik yang menutupi kekurangan pada recall dan precision didalam penilaian performa terhadap kelas positif.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

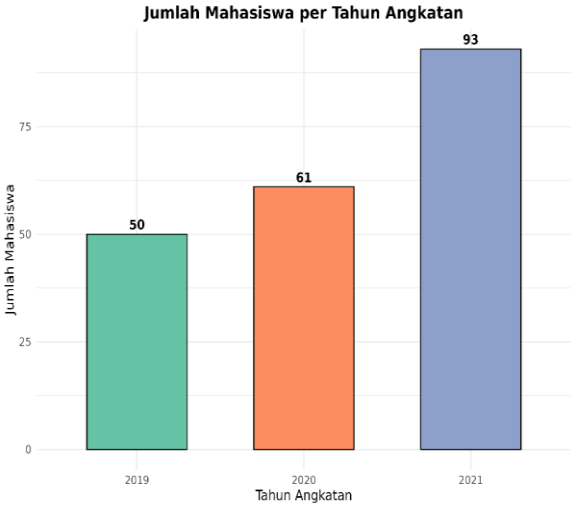
III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Tahap awal pada penelitian ini yaitu pengumpulan data. data diambil dengan penyebaran kuisioner melalui google form dengan jumlah data yang didapat sebanyak 204 responden dan distribusi data yang dapat dilihat pada gambar 2 dan 3.



Gambar 2. Piechart Distribusi Data



Gambar 3. Diagram Distribusi Data

B. Pra-Processing Data

Data awal yang diperoleh berupa hasil google form dengan format kolom yang masih menggunakan kalimat pertanyaan lengkap, sehingga perlu disederhanakan. Proses ini dimulai dengan memberikan label pendek pada setiap pertanyaan menjadi "P1", "P2",….,"Pn", seperti yang ditunjukan pada tabel 3. Untuk mengamati hasil penuh dari pergantian kolom dapat dilihat pada [22].

TABEL III
POTONGAN HASIL PERGANTIAN KOLOM

Cap waktu	Nama pengguna	Tahun Angkatan	P 1	P 2	P 3	P 4
2025/06/0	naurahzahiya	2021	4	3	4	4
2025/06/0	mohammadsa	2019	3	4	4	3
2025/06/0	arfianhilmy45	2020	2	3	3	3
2025/06/0	bagasdwiprat	2020	3	4	3	3
2025/06/0	arielmuhamm	2021	4	3	4	3
2025/06/0	maarifparlika	2021	4	3	3	2
2025/06/0	iqadzul2003@	2019	3	2	2	3

Selanjutnya, indikator yang telah dilabeli dikelompokkan Misalnya, variabel Manajemen Waktu dihitung dari rata-rata nilai P1, P2, dan P3), sehingga nilai rata-ratanya diperoleh dari jumlah ketiga nilai tersebut dibagi tiga. Seperti yang ditunjukan pada tabel 4, karena keterbatasan tempat pada table maka setiap faktor akan diringkas menjadi F1, F2 dan seterusnya. Untuk mengamati hasil penuh dari hasil rata-rata dapat dilihat pada [22].

TABEL IV
POTONGAN HASIL RATA-RATA

No. responden	Tahun Angkatan	F1	F2	F3	F4	F5	F6	F7
1	2021	3,67	3	2,33	3,33	3,33	2,33	3,33
2	2019	3,67	3,67	2,67	3,67	3,33	3,33	3
3	2020	2,67	3	2,67	3,67	2,67	3,33	2,67
4	2020	3,33	3,33	3	3,33	3,33	2,67	2,67
5	2021	3,67	3,33	3	3,33	2,67	3,67	3,67
6	2021	3,33	3	2,33	2,33	3	2,67	3
7	2019	2,33	3,33	2,67	3,33	2,33	2	2,33
8	2021	3,67	3,67	3,33	2,33	3	3,33	3
9	2021	3,33	3,33	3,33	3,33	2,33	2,67	3,33

C. Klasifikasi Data

Klasifikasi data pada penelitian ini dibagi menjadi dua kategori yaitu "Tepat Waktu" dan "Terlambat". Penentuan kategori ini dilakukan dengan menghitung nilai interval dari keseluruhan hasil perhitungan mean (rata-rata) dari ketujuh variabel yang telah dianalisis sebelumnya, yaitu Manajemen Waktu, Motivasi Mahasiswa, Kebijakan Kampus, Dukungan Dosen, Dukungan Keluarga, Lingkungan Sekitar, dan Keterampilan Akademik seperti yang ditunjukan pada tabel 5. karena keterbatasan tempat pada table maka setiap faktor akan diringkas menjadi Tepat Waktu = TW dan Terlambat = T Untuk mengamati hasil penuh dari hasil klasifikasi dapat dilihat pada [22].

TABEL V
POTONGAN HASIL KLASIFIKASI

No	F1	F2	F3	F4	F5	F6	F7	Total Score	Kategori
1	3,67	3	2,33	3,33	3,33	2,33	3,33	21,32	TW
2	3,67	3,67	2,67	3,67	3,33	3,33	3	23,34	TW
3	2,67	3	2,67	3,67	2,67	3,33	2,67	20,68	TW
4	3,33	3,33	3	3,33	3,33	2,67	2,67	21,66	TW
5	2	2,33	2	2,33	1,67	2,67	2	15	T
6	3,33	3	2,33	3,33	3	2,67	3	20,68	TW

Setelah data selesai diklasifikasikan ke dalam dua kategori utama, yaitu "Tepat Waktu" dan "Terlambat", tahap selanjutnya adalah melakukan kategorisasi nilai dari masing-masing faktor atau variabel menjadi tiga tingkatan, yaitu Rendah, Sedang, dan Tinggi. Dimana setiap nilai rata-rata faktor untuk masing-masing responden dikelompokkan berdasarkan interval seperti yang ditunjukan pada tabel 6. karena keterbatasan tempat pada table maka s akan diringkas menjadi Tinggi =T, Sedang = S, dan Rendah = R.Untuk mengamati hasil penuh dari hasil klasifikasi nilai faktor dapat dilihat pada [22].

TABEL VI
POTONGAN HASIL KLASIFIKASI NILAI FAKTOR

No	Total Score	Kategori	F1	F2	F3	F4	F5	F6	F7
1	21,32	TW	T	T	S	T	T	S	T
2	23,34	TW	T	T	S	T	T	T	T
3	20,68	TW	S	T	S	T	S	T	S
4	21,66	TW	T	T	T	T	T	S	S
5	15	T	S	S	S	R	S	S	S
6	20,68	TW	T	S	S	S	S	S	S

D. Balancing Data

Pada penelitian ini, hasil klasifikasi antara kategori " Tepat Waktu" dan "Terlambat" mengalami ketidakseimbangan sehingga terjadi data minoritas yang bisa menyebabkan overfitting pada model.

Data hasil kuesioner diperiksa untuk memastikan tidak ada nilai kosong (*missing value*). Karena seluruh variabel menggunakan skala Likert 1–5, normalisasi data tidak diperlukan. Untuk mengatasi ketidakseimbangan kelas antara mahasiswa yang tepat waktu dan terlambat.

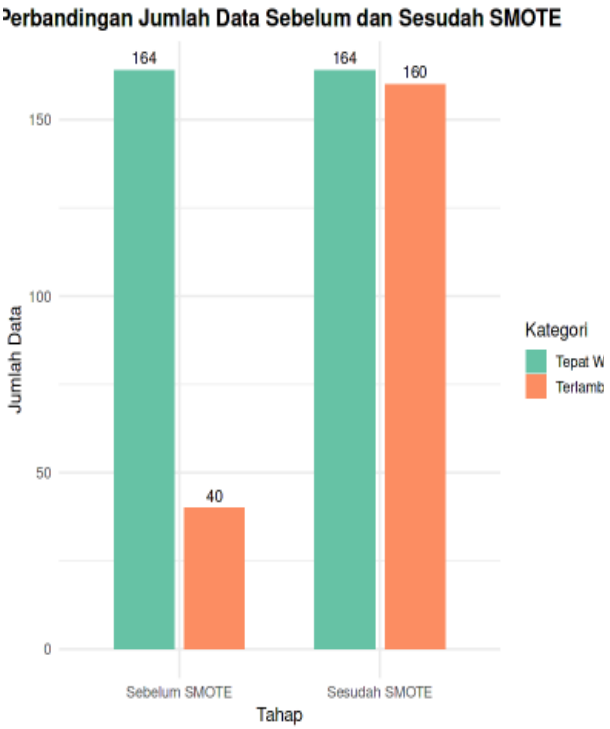
Untuk mengatasi ini data akan melalui proses SMOTE (Synthetic Minority Over-sampling Technique) untuk menyeimbangkan data dengan nilai K=20. Distribusi perbedaan sebelum dan sesudah SMOTE dapat dilihat pada gambar 4.

E. Data Training dan Testing

Setelah data memasuki proses SMOTE, didapatkan total hasil data sebanyak 324, yang akan dibagi menjadi data training dan data testing dengan proporsi 80:20, hingga didapatkan 260 data training dan 64 data testing yang akan diolah pada tahap selanjutnya.

F. Perhitungan Entrophy

Proses ini dimulai dengan menghitung entropi awal dari keseluruhan data berdasarkan jumlah mahasiswa yang "Tepat Waktu" dan "Terlambat", dan didapatkan entrophy awal dataset terhitung sebesar 0,9999 . Setelah itu, nilai entrophy dihitung untuk setiap kategori pada masing-masing fitur



Gambar 4. Diagram Perbadingan Hasil SMOTE

kategorikal yang akan dianalisis (yaitu, hasil kategorisasi dari setiap faktor), yang merefleksikan tingkat ketidakpastian atau keragaman distribusi data dalam suatu kategori. Nilai entrophy yang lebih rendah mengindikasikan homogenitas kategori terhadap salah satu kelas target, sehingga dianggap lebih informatif dalam proses klasifikasi. Sebaliknya, nilai entropi yang mendekati 1 menunjukkan distribusi data yang merata antara dua kelas, menjadikannya kurang informatif untuk pemisahan data. Hasil perhitungan entropi untuk setiap kategori faktor disajikan dalam tabel 7.

TABEL VII
HASIL PERHITUNGAN ENTROPHY

Fitur	Nilai	Jumlah_Data	Entropy
Manajemen Waktu Kategori	Tinggi	161	0,6075
Manajemen Waktu Kategori	Sedang	100	0,8415
Manajemen Waktu Kategori	Rendah	63	0
Motivasi Mahasiswa Kategori	Tinggi	165	0,5499
Motivasi Mahasiswa Kategori	Sedang	106	0,6987
Motivasi Mahasiswa Kategori	Rendah	53	0
Kebijakan Kampus Kategori	Sedang	148	0,9979
Kebijakan Kampus Kategori	Tinggi	92	0,0865
Kebijakan Kampus Kategori	Rendah	84	0,2223
Dukungan Dosen Kategori	Tinggi	124	0
Dukungan Dosen Kategori	Sedang	90	0,9641
Dukungan Dosen Kategori	Rendah	110	0,2668
Dukungan Keluarga Kategori	Tinggi	129	0,0655
Dukungan Keluarga Kategori	Sedang	108	0,8987
Dukungan Keluarga Kategori	Rendah	87	0,1579
Lingkungan Sekitar Kategori	Sedang	82	0,8722
Lingkungan Sekitar Kategori	Tinggi	142	0,2525
Lingkungan Sekitar Kategori	Rendah	100	0,2423
Keterampilan Akademik Kategori	Tinggi	97	0,199
Keterampilan Akademik Kategori	Sedang	140	0,9976
Keterampilan Akademik Kategori	Rendah	87	0,2691

G. Perhitungan Gain Ratio

Setelah nilai entropy total dari target variabel "Kategori" yang sebelumnya telah dihitung sebagai *entropy global* kemudian digunakan untuk menghitung *information gain dan juga gain ratio.* Seperti yang ditunjukan pada tabel 8.

TABEL VIII
HASIL PERHITUNGAN GAIN RATIO

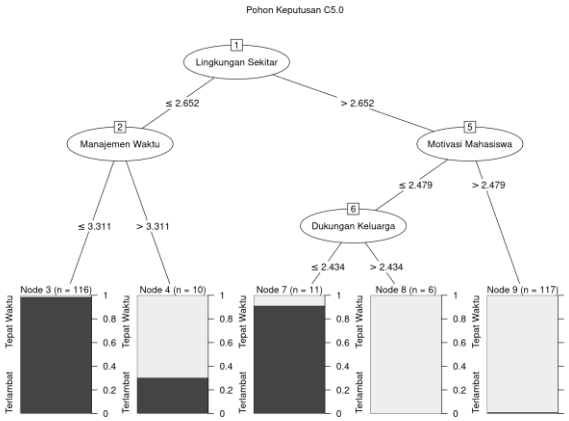
Faktor	Entropy_Total	Weighted_Entropy	Info_Gain	Split_Info	Gain_Ratio
F1	0,9999	0,3584	0,6415	1,5728	0,4079
F2	0,9999	0,368	0,6319	1,5667	0,4033
F3	0,9999	0,4062	0,5937	1,5467	0,3838
F4	0,9999	0,5086	0,4913	1,4504	0,3387
F5	0,9999	0,538	0,4619	1,537	0,3005
F6	0,9999	0,5616	0,4383	1,4842	0,2953
F7	0,9999	0,5629	0,437	1,5533	0,2813

Hasil perhitungan menunjukkan bahwa faktor Dukungan Dosen memiliki nilai tertinggi pada kedua metrik, dengan Information Gain sebesar 0,6415 dan Gain Ratio sebesar 0,4079, sehingga dipilih sebagai atribut utama (root node) dalam pembentukan pohon keputusan.

Faktor-faktor lain yang juga memiliki nilai Gain Ratio tinggi adalah Dukungan Keluarga (0,4033), Lingkungan Sekitar (0,3838), dan Motivasi Mahasiswa (0,3387), yang menunjukkan kontribusi signifikan dalam pemisahan kelas target. Sebaliknya, faktor Kebijakan Kampus (0,3005), Manajemen Waktu (0,2953), dan Keterampilan Akademik (0,2813) menunjukkan nilai Gain Ratio lebih rendah, menandakan kontribusinya lebih kecil dalam proses klasifikasi.

H. Decision Tree C5.0

Pemodelan dilakukan menggunakan bahasa R Programming dengan pustaka C50 dan caret. Dataset dibagi menjadi 80% data latih dan 20% data uji untuk menghindari



Gambar 5. Hasil Decision Tree C5.0

bias model. Pada penelitian ini visualisasi pohon keputusan yang dihasilkan dari pemodelan menggunakan algoritma C5.0. Pohon keputusan ini merepresentasikan alur logika dalam pengambilan keputusan untuk memprediksi keterlambatan penyelesaian tugas akhir mahasiswa berdasarkan beberapa atribut yang digunakan dalam penelitian. Hasil decision tree c5.0 dapat dilihat pada gambar 5.

Hasil visualisasi pohon keputusan C5.0 menunjukkan bahwa faktor Lingkungan Sekitar menjadi node akar, dengan batas keputusan ≤ 2,652. Mahasiswa dengan skor di bawah atau sama dengan batas ini cenderung diklasifikasikan sebagai Terlambat, sedangkan skor di atasnya lebih dominan Tepat Waktu. Pada cabang kiri, klasifikasi dilanjutkan dengan atribut Manajemen Waktu. Jika nilainya ≤ 3,311, mahasiswa diklasifikasikan sepenuhnya sebagai Terlambat. Sebaliknya, jika nilainya lebih tinggi, sebagian besar termasuk dalam kategori Tepat Waktu.

Pada cabang kanan, klasifikasi dilanjutkan ke atribut Motivasi Mahasiswa. Jika nilai motivasi ≤ 2,479, klasifikasi dipengaruhi oleh Dukungan Keluarga. Dukungan keluarga rendah (≤ 2,434) dominan Terlambat, sedangkan nilai lebih tinggi sepenuhnya Tepat Waktu. Untuk mahasiswa dengan motivasi tinggi (> 2,479), seluruhnya diklasifikasikan sebagai Tepat Waktu. Secara keseluruhan, urutan faktor yang paling berpengaruh terhadap keterlambatan tugas akhir adalah: Lingkungan Sekitar, Motivasi Mahasiswa, Dukungan Keluarga, dan Manajemen Waktu. Kombinasi nilai rendah pada faktor-faktor tersebut cenderung mengarah pada keterlambatan penyelesaian tugas akhir.

I. Boxplot Nilai Kategori

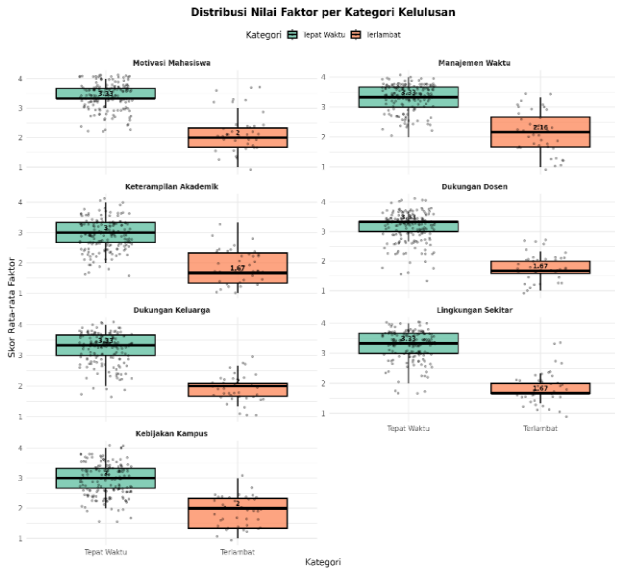
Visualiasasi boxplot pada penelitian ini berguna untuk melihat distribusi nilai rata-rata dari masing-masing faktor yang dianalisis terhadap dua kategori kelulusan mahasiswa, yaitu Tepat Waktu dan Terlambat, dengan menampilkan tujuh faktor yang dianalisis. Hasil visualisasi boxplot dapat dilihat pada gambar 6.

Visualisasi data menunjukkan bahwa mahasiswa yang lulus tepat waktu memiliki nilai median faktor-faktor penyebab keterlambatan yang lebih tinggi dibandingkan mahasiswa yang lulus terlambat. Perbedaan median paling mencolok terlihat pada Dukungan Dosen dan Lingkungan Sekitar, masing-masing dengan selisih 1,66 poin, menandakan pengaruh signifikan dari faktor eksternal terhadap ketepatan waktu kelulusan.

Faktor lain seperti Keterampilan Akademik, Dukungan Keluarga, dan Motivasi Mahasiswa juga menunjukkan selisih median yang cukup tinggi (1,33 poin), mencerminkan pentingnya dukungan internal dan eksternal mahasiswa. Manajemen Waktu memiliki selisih median 1,17 poin, tetap relevan namun tidak sebesar faktor lainnya.

Sementara itu, Kebijakan Kampus menunjukkan selisih median terendah (1,00 poin), yang mengindikasikan pengaruhnya relatif lebih kecil dalam membedakan ketepatan waktu penyelesaian tugas akhir.

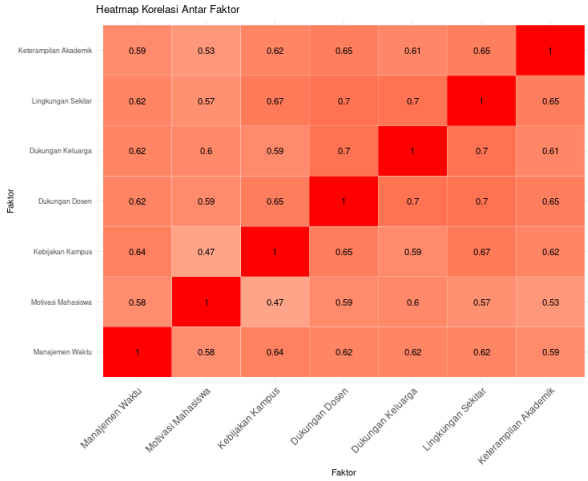
Perbedaan median ini memperkuat temuan bahwa kombinasi dukungan lingkungan eksternal dan motivasi serta keterampilan internal mahasiswa sangat berperan dalam mendukung penyelesaian tugas akhir secara tepat waktu.



Gambar 6 Hasil Boxplot Korelasi Antar Faktor

J. Heatmap Korelasi Antar faktor

Visualisasi heatmap korelasi antar faktor pada penelitian ini bertujuan untuk menunjukkan derajat hubungan linier antara masing-masing faktor yang dianalisis dalam penelitian ini. Korelasi ditampilkan dalam bentuk matriks warna, dengan nilai korelasi berkisar antara -1 hingga 1. Warna merah menunjukkan hubungan positif, sedangkan semakin mendekati biru menunjukkan hubungan negatif. Hasil visualisasi boxplot dapat dilihat pada gambar 7.

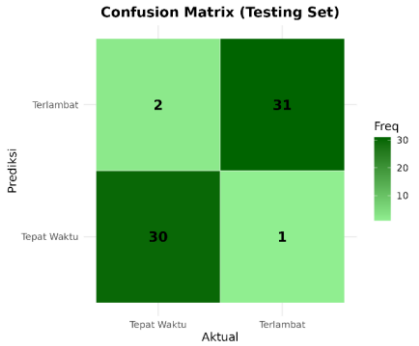


Gambar 7. Hasil Heatmap Korelasi Antar Faktor

Hasil visualisasi heatmap korelasi menunjukkan bahwa sebagian besar faktor memiliki hubungan yang cukup kuat satu sama lain. Korelasi tertinggi tercatat antara Dukungan Dosen, Dukungan Keluarga, dan Lingkungan Sekitar, dengan nilai sebesar 0,70, menandakan keterkaitan erat antar faktor eksternal. Faktor Kebijakan Kampus juga menunjukkan korelasi tinggi dengan Dukungan Dosen (0,65) dan Lingkungan Sekitar (0,67), yang mengindikasikan bahwa kebijakan kampus yang baik cenderung selaras dengan persepsi mahasiswa terhadap dukungan akademik.
Sementara itu, faktor internal seperti Motivasi Mahasiswa menunjukkan korelasi 0,58 dengan Manajemen Waktu, menandakan bahwa mahasiswa yang lebih termotivasi cenderung memiliki keterampilan manajemen waktu yang lebih baik. Beberapa korelasi lebih rendah juga tercatat, seperti antara Motivasi Mahasiswa dan Kebijakan Kampus (0,47), menunjukkan hubungan yang tidak terlalu kuat..

K. Confusion Matrix

Pada Penelitian ini confusion matrix digunakan untuk mengevaluasi performa model klasifikasi dengan membandingkan hasil prediksi terhadap nilai aktual pada data testing. Matrix ini menampilkan empat komponen utama, yaitu True Positive (TP), False Positive (FP), False Negative (FN), dan True Negative (TN) seperti yang ditunjukan gambar 8.



Gambar 8.Hasil Confusion Matrix

Dari visualisasi tersebut menunjukkan bahwa model memiliki performa yang cukup baik. Jumlah prediksi benar (TP + TN) sebanyak 61 dari total 64 data uji, yang berarti tingkat akurasi cukup tinggi. Model juga memiliki jumlah kesalahan yang rendah, yaitu hanya 1 kasus *false positive* dan 2 kasus *false negative.* Sedangkan hasil evaluasi dapat dilihat pada tabel 9.

TABEL IX
HASIL EVALUASI MODEL

Confusion Matriks	Hasil Evaluasi
Accuracy	95,31%
Precision	96,77%
Recall	93,75%
F1-Score	95,24%

Hasil evaluasi terhadap data testing pada tabel 9 menunjukkan Accuracy sebesar 95,31%, dengan jumlah prediksi benar sebanyak 61 dari total 64 data. Precision yang diperoleh sebesar 96,77% menunjukkan bahwa mayoritas prediksi "Tepat Waktu" oleh model memang akurat. Nilai Recall sebesar 93,75% menunjukkan bahwa hampir seluruh data mahasiswa yang benar-benar "Tepat Waktu" berhasil dikenali oleh model. Dengan demikian, nilai F1-score sebesar 95,24% mencerminkan bahwa model memiliki performa yang baik dalam melakukan klasifikasi terhadap data baru.

L. Hasil Analisis

Hasil analisis menggunakan algoritma Decision Tree C5.0 menunjukkan bahwa variabel "Lingkungan Sekitar" memiliki nilai Gain Ratio tertinggi dibandingkan faktor lainnya, sehingga menjadi variabel paling dominan dalam memengaruhi keterlambatan penyelesaian tugas akhir mahasiswa. Dominannya pengaruh lingkungan sekitar menunjukkan bahwa aspek eksternal non-akademik berperan besar dalam menentukan tingkat konsistensi dan fokus mahasiswa selama proses penyusunan tugas akhir.

Lingkungan sekitar mencakup berbagai elemen seperti kondisi tempat tinggal, suasana belajar, dukungan sosial, serta tingkat gangguan dari lingkungan sosial. Mahasiswa yang tinggal di lingkungan dengan tingkat kebisingan tinggi, kurangnya fasilitas belajar, atau lingkungan pergaulan yang tidak mendukung kegiatan akademik cenderung mengalami penurunan fokus dan motivasi. Kondisi tersebut dapat menyebabkan mahasiswa menunda proses penulisan atau penelitian tugas akhir sehingga memperpanjang waktu penyelesaian.

Sebaliknya, mahasiswa yang berada di lingkungan yang kondusif—seperti suasana rumah atau kos yang tenang, dukungan teman sebaya yang produktif, dan akses terhadap sumber belajar—menunjukkan kecenderungan lebih cepat dalam menyelesaikan tugas akhir. Hal ini sejalan dengan penelitian sebelumnya [14] yang menyatakan bahwa faktor eksternal seperti kondisi lingkungan belajar dan dukungan sosial memiliki kontribusi signifikan terhadap efektivitas proses akademik dan ketepatan waktu penyelesaian studi..

M. Implikasi Praktis Penelitian

Hasil penelitian ini menunjukkan bahwa faktor Lingkungan Sekitar, Motivasi Mahasiswa, dan Manajemen Waktu memiliki pengaruh dominan terhadap keterlambatan penyelesaian tugas akhir. Oleh karena itu, universitas perlu menciptakan lingkungan belajar yang kondusif, menyediakan fasilitas akademik yang memadai, serta memperkuat program pendampingan dan motivasi bagi mahasiswa tingkat akhir. serta memberikan pelatihan manajemen waktu dan dukungan psikologis juga penting dilakukan agar mahasiswa mampu mengatur jadwal kerja secara efektif dan menjaga komitmen akademik. Dengan penerapan langkah-langkah tersebut, universitas diharapkan dapat menekan tingkat keterlambatan dan meningkatkan keberhasilan penyelesaian tugas akhir mahasiswa.

Selain itu, mahasiswa juga perlu untuk membentuk komunitas belajar atau kelompok diskusi khusus mahasiswa tingkat akhir yang bersifat inklusif dan kolaboratif, serta mahasiswa juga harus lebih selektif dalam memilih circle pertemanan yang produktif, yang tujuan untuk saling mendorong dalam penyelesaian akademik.

Penelitian ini terbatas pada satu fakultas dan periode tertentu sehingga hasilnya belum dapat digeneralisasi secara luas. Namun, secara praktis, hasil ini memberikan masukan penting bagi pihak universitas untuk membangun intervensi berbasis data guna mendukung penyelesaian tugas akhir tepat waktu.

IV. KESIMPULAN

Berdasarkan hasil analisis terhadap data mahasiswa menggunakan algoritma Decision Tree C5.0, dapat disimpulkan bahwa faktor Lingkungan Sekitar merupakan variabel yang paling berpengaruh terhadap keterlambatan

penyelesaian tugas akhir. Faktor ini terbukti signifikan dalam membedakan mahasiswa yang lulus tepat waktu dan yang terlambat, baik dari struktur pohon keputusan, nilai gain ratio, maupun distribusi median. Selain itu, faktor Motivasi Mahasiswa, Manajemen Waktu, dan Dukungan Keluarga juga memiliki pengaruh kuat, sebagaimana ditunjukkan melalui keterlibatannya dalam struktur model dan nilai gain ratio yang tinggi. Meskipun Dukungan Dosen menunjukkan nilai gain ratio yang tinggi, faktor ini tidak terpilih dalam model akhir karena memiliki korelasi yang sangat tinggi dengan faktor eksternal lainnya. Adapun Keterampilan Akademik dan Kebijakan Kampus tercatat memiliki pengaruh yang lebih rendah karena nilai gain ratio-nya relatif kecil.

Model klasifikasi yang dibangun menggunakan algoritma C5.0 menunjukkan performa yang sangat baik, dengan akurasi sebesar 95,31%, precision 96,77%, recall 93,75%, dan F1-score 95,24%. Hasil evaluasi ini membuktikan bahwa algoritma C5.0 mampu mengklasifikasikan mahasiswa secara akurat dan efektif dalam mengidentifikasi faktor-faktor yang paling berpengaruh terhadap keterlambatan penyelesaian tugas akhir.

DAFTAR PUSTAKA

[1] N. D. Larasati and W. S. Jatiningrum, "Analisis Faktor pada Keterlambatan Studi Mahasiswa Teknik Industri Universitas Ahmad Dahlan," *Manaj. Pendidik.*, vol. 16, no. 2, pp. 83–96, 2021, doi: 10.23917/jmp.v16i2.12134.

[2] A. F. Ramadhan, A. Sukohar, and F. Saftarina, "Perbedaan Derajat Kecemasan Antara Mahasiswa Tahap Akademik Tingkat Awal dengan Tingkat Akhir di Fakultas Kedokteran Universitas Lampung," *Medula*, vol. 9, no. 1, pp. 78–82, 2019, [Online]. Available: https://juke.kedokteran.unila.ac.id/index.php/medula/article/view/2355/pdf

[3] R. T. Sugiharno, W. H. Ari Susanto, and F. Wospakrik, "Faktor-Faktor yang Mempengaruhi Kecemasan Mahasiswa dalam Menghadapi Tugas Akhir," *J. Keperawatan Silampari*, vol. 5, no. 2, pp. 1189–1197, 2022, doi: 10.31539/jks.v5i2.3760.

[4] P. K. Karmokar, "R Programming Unlocked : Easy Learning," no. August, 2024.

[5] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[6] A. Fatkhudin, M. Y. Febrianto, F. A. Artanto, M. W. N. Hadinata, and R. Fahlevi, "Algoritma Decision Tree C.45 Dalam Analisa Kelulusan Mahasiswa Program Studi Manajemen Informatika Umpp," *J. Ilm. Ilmu Komput.*, vol. 8, no. 2, pp. 83–86, 2022, doi: 10.35329/jiik.v8i2.240.

[7] I. Technology, "Analisis penghambat mahasiswa membuat skripsi menggunakan algoritma C4.5," *Conf. Electr. Eng. Informatics, Ind. Technol. Creat. Media*, vol. 3, pp. 838–843, 2023.

[8] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," *IEEE Access*, vol. 12, no. May, pp. 86716–86727, 2024, doi: 10.1109/ACCESS.2024.3416838.

[9] S. N. A. M. Razali, M. S. Rusiman, W. S. Gan, and N. Arbin, "The Impact of Time Management on Students' Academic Achievement," *J. Phys. Conf. Ser.*, vol. 995, no. 1, 2018, doi: 10.1088/1742-6596/995/1/012042.

[10] R. Silva, R. Rodrigues, and C. Leal, "Academic motivation scale: Development and validation for portuguese accounting and marketing undergraduate students," *Proc. Eur. Conf. Games-based Learn.*, vol. 2018-Octob, no. 11, pp. 600–607, 2018, doi: 10.5539/ijbm.v13n11p1.

[11] O. Berthod, "Global Encyclopedia of Public Administration, Public Policy, and Governance," *Glob. Encycl. Public Adm. Public Policy, Gov.*, pp. 1–5, 2020, doi: 10.1007/978-3-319-31816-5.

[12] M. Henricson, B. Fridlund, J. Mårtensson, and B. Hedberg, "The validation of the Supervision of Thesis Questionnaire (STQ)," *Nurse Educ. Today*, vol. 65, pp. 11–16, 2018, doi: 10.1016/j.nedt.2018.02.010.

[13] M. Hassan, S. Fang, A. A. Malik, T. A. Lak, and M. Rizwan, "Impact of perceived social support and psychological capital on university students' academic success: testing the role of academic adjustment as a moderator," *BMC Psychol.*, vol. 11, no. 1, pp. 1–11, 2023, doi: 10.1186/s40359-023-01385-y.

[14] F. Bankole Adeyemi, "Peer group influence on academic performance of undergraduate students in Babcock University, Ogun State," *African Educ. Res. J.*, vol. 7, no. 2, pp. 81–87, 2019, doi: 10.30918/aerj.72.19.010.

[15] S. Nabizadeh, S. Hajian, Z. Sheikhan, and F. Rafiei, "Prediction of academic achievement based on learning strategies and outcome expectations among medical students," *BMC Med. Educ.*, vol. 19, no. 1, pp. 1–11, 2019, doi: 10.1186/s12909-019-1527-9.

[16] N. I. Majdina, B. Pratikno, and A. Tripena, "Penentuan Ukuran Sampel Menggunakan Rumus Bernoulli Dan Slovin: Konsep Dan Aplikasinya," *J. Ilm. Mat. dan Pendidik. Mat.*, vol. 16, no. 1, p. 73, 2024, doi: 10.20884/1.jmp.2024.16.1.11230.

[17] P. Rahayu *et al.*, *Buku Ajar Data Mining*, vol. 1, no. January 2024. 2024.

[18] H. Guan, Y. Zhang, M. Xian, H. D. Cheng, and X. Tang, "SMOTE-WENN: Solving class imbalance and small sample problems by oversampling and distance scaling," *Appl. Intell.*, vol. 51, no. 3, pp. 1394–1409, 2021, doi: 10.1007/s10489-020-01852-8.

[19] H. Y. Taihuttu and I. S. Sitanggang, "Spatial Classification of Forest and Land Fire Risk using Decision Tree C5.0 Algorithm," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1315, no. 1, pp. 0–12, 2024, doi: 10.1088/1755-1315/1315/1/012059.

[20] N. Tanjung, D. Irmayani, and V. Sihombing, "Implementation of C5.0 Algorithm for Prediction of Student Learning Graduation in Computer System Architecture Subjects," *Sinkron*, vol. 7, no. 1, pp. 274–280, 2022, doi: 10.33395/sinkron.v7i1.11259.

[21] K. Riehl, M. Neunteufel, and M. Hemberg, "Hierarchical confusion matrix for classification performance evaluation," *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 72, no. 5, pp. 1394–1412, 2023, doi: 10.1093/jrsssc/qlad057.

[22] M. C.Anwar, "Hasil_Analisis_Decision_Tree_C5.0," kaggle. Accessed: Jul. 16, 2025. [Online]. Available: https://www.kaggle.com/datasets/cyclone4215/analisis-faktor-keterlambatan-decision-tree-c5-0