Penerapan Algoritma C4.5 Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa

Michelle Kinaya Sesa Rinding[1], Yanice Sampebua Teken[2]

[1,2]Universitas Papua
Jl. Gunung Salju, Amban, Manokwari, Papua Barat, 98314
e-mail: [1]202365067@student.unipa.ac.id, [2]202365005@students.unipa.ac.id

Artikel Info : Diterima : 05-05-2026 | Direvisi : 18-05-2026 | Disetujui : 01-06-2026

Abstrak - Ketepatan waktu lulus mahasiswa merupakan indikator kualitas pendidikan tinggi yang berdampak langsung pada akreditasi program studi. Penelitian ini bertujuan membangun model prediksi ketepatan waktu lulus mahasiswa Program Studi Teknik Informatika Universitas Papua memanfaatkan algoritma *Decision Tree* C4.5. Dataset terdiri dari 292 data akademik mahasiswa angkatan 2016-2021 dengan delapan fitur prediktor berupa Indeks Prestasi (IP) dan jumlah Satuan Kredit Semester (SKS) dari semester satu hingga empat. Hasil penelitian menunjukkan bahwa model C4.5 dengan teknik *RandomUnderSampler* menghasilkan performa terbaik dengan nilai *accuracy* 74,58%, *precision 77,29%, recall 74,58%, F1-Score 74,89%, dan AUC-ROC 79,11%*. Model ini diharapkan dapat menjadi instrumen pendukung keputusan bagi program studi dalam mengidentifikasi mahasiswa yang berpotensi tidak lulus tepat waktu.

Kata Kunci : *Data Mining*, Algoritma C4.5, Prediksi Kelulusan

Abstract *- Timely graduation is a key indicator of higher education quality with a direct impact on study program accreditation. This study aims to develop a timely graduation prediction model for students of the Informatics Engineering Study Program at the University of Papua using the C4.5 Decision Tree algorithm. The dataset comprises 292 academic records from students enrolled between 2016 and 2021, with eight predictor features consisting of Grade Point Average (GPA) and the number of Credit Units (CU) from the first to the fourth semester. The results show that the C4.5 model combined with the RandomUnderSampler technique achieves the best performance, with an accuracy of 74.58%, precision of 77.29%, recall of 74.58%, F1-Score of 74.89%, and AUC-ROC of 79.11%. This model is expected to serve as a decision-support tool for the study program in identifying students who are at risk of not graduating on time.*

Keywords : Data Mining, C4.5 Algorithm, Graduation Prediction

PENDAHULUAN

Salah satu tolok ukur utama dalam menilai mutu penyelenggaraan pendidikan tinggi adalah ketepatan waktu kelulusan mahasiswa. Badan Akreditasi Nasional Perguruan Tinggi menegaskan melalui matriks penilaian akreditasi program studi bahwa kemampuan mahasiswa menyelesaikan studi sesuai batas waktu yang ditetapkan mencerminkan efektivitas proses pembelajaran sekaligus memberikan dampak signifikan terhadap status akreditasi dan

citra institusi (BAN-PT, 2008). Namun demikian, pada kenyataannya masih banyak mahasiswa yang mengalami keterlambatan dalam menyelesaikan studi akibat berbagai faktor, baik akademik maupun non-akademik.

Permasalahan tersebut dapat diatasi melalui penerapan teknik prediksi kelulusan mahasiswa. Dalam hal ini, *data mining* merupakan salah satu teknik yang kerap dimanfaatkan, dengan metode klasifikasi berbasis algoritma *Decision Tree* sebagai pendekatan yang paling umum diterapkan dalam upaya prediksi kelulusan mahasiswa (Rohmawan, 2018).

Penelitian berjudul "Prediksi Kelulusan Tepat Waktu Menggunakan Pendekatan Pohon Keputusan Algoritma *Decision Tree*" memanfaatkan data dari 312 mahasiswa Program Studi Sistem Informasi pada Jurusan Teknik Informatika di Universitas Surabaya. Penelitian tersebut menghasilkan model *Decision Tree* dengan tingkat akurasi prediksi sebesar 75,95% (Moerdyanto & Nuryana, 2023).

Penelitian lain yang berjudul "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma C4.5 Pada Uin Syarif Hidayatullah Jakarta" menggunakan sebanyak 4205 data mahasiswa. Temuan dari penelitian memperlihatkan bahwa algoritma C4.5 mampu menghasilkan performa yang cukup baik dengan *accuracy* 75,52%, *precision* 75,50%, serta *recall* 75,50% (Hasibuan & Mahdiana, 2023).

Penelitian berikutnya tentang "Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Usia, Jenis Kelamin, Dan Indeks Prestasi Menggunakan Algoritma *Decision Tree*" juga menerapkan algoritma C4.5 dalam memprediksi kelulusan mahasiswa. Hasil yang diperoleh memiliki tingkat akurasi sebesar 91,51% (Romadhona et al., 2017).

Penelitian tentang "*Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques*" mengkaji perbandingan performa beberapa algoritma klasifikasi, yaitu *K-Nearest Neighbor (K-NN), Naive Bayes,* dan *Decision Tree.* Hasil dari perbandingan yang dilakukan menunjukkan bahwa algoritma *Decision Tree* cenderung memiliki tingkat akurasi yang lebih tinggi serta tingkat kesalahan yang lebih rendah dibandingkan dengan algoritma lainnya, serta lebih mudah diinterpretasikan (Jadhav & Channe, 2016).

Meskipun sejumlah penelitian tersebut telah membuktikan efektivitas algoritma *Decision Tree* C4.5 dalam prediksi kelulusan mahasiswa, sebagian besar dilakukan pada institusi dengan karakteristik data dan konteks akademik yang berbeda. Belum terdapat penelitian yang secara spesifik mengkaji prediksi kelulusan tepat waktu pada Program Studi Teknik Informatika Universitas Papua yang memiliki karakteristik dataset tersendiri dari sisi ukuran maupun distribusi kelasnya. Mempertimbangkan keunggulan yang telah dibuktikan pada penelitian-penelitian sebelumnya, algoritma *Decision Tree C4.5* dipilih sebagai metode utama dalam penelitian ini untuk membangun model prediksi kelulusan tepat waktu mahasiswa Program Studi Teknik Informatika Universitas Papua.

Tujuan penelitian ini adalah mengembangkan model prediksi kelulusan tepat waktu mahasiswa Program Studi Teknik Informatika Universitas Papua dengan memanfaatkan algoritma *Decision Tree* C4.5. Hasil yang diperoleh dari penelitian ini diharapkan dapat membantu pihak program studi dalam mengidentifikasi mahasiswa yang berpotensi mengalami keterlambatan kelulusan sehingga tindakan intervensi seperti bimbingan intensif atau konseling dapat dilakukan sedini mungkin guna menjaga kualitas dan nilai akreditasi program studi maupun universitas secara keseluruhan.

METODE PENELITIAN

Penelitian ini menerapkan pendekatan kuantitatif dengan memanfaatkan metode *data mining* untuk memprediksi kelulusan tepat waktu mahasiswa pada Program Studi Teknik Informatika Universitas Papua. Alur penelitian mencakup rancangan penelitian, pengumpulan data, pengolahan data, hingga pembangunan model klasifikasi menggunakan algoritma *Decision Tree* C4.5.

1. Mengumpulkan Data

Data dalam penelitian ini bersumber dari 292 data mahasiswa Program Studi Teknik Informatika Universitas Papua dari angkatan 2016 hingga 2021. Data yang dikumpulkan memiliki atribut berupa indeks prestasi semester 1-4, jumlah SKS semester 1-4, indeks prestasi kumulatif, dan status kelulusan yang dikategorikan sebagai tepat waktu atau tidak tepat waktu. Data tersebut dipilih karena berkaitan langsung dengan hasil belajar mahasiswa yang bisa memengaruhi kelulusan tepat waktu. Data yang digunakan merupakan data primer yang diperoleh dari sistem akademik dan telah melalui proses seleksi untuk memastikan kelengkapan data.

2. Data Mining

Data mining merupakan serangkaian proses yang bertujuan untuk menemukan pola, hubungan, dan kecenderungan yang bermakna dari sekumpulan data berukuran besar melalui penerapan metode tertentu (Ucha Putri et al., 2021). Metode penambangan data, dengan kekuatan dan otomatisitasnya, memiliki kemampuan untuk mengatasi sejumlah besar data dan mengekstraksi nilai (Kusrini et al., 2009). Adapun data yang diperoleh perlu dilakukan seleksi, pembersihan, dan normalisasi. Seleksi data dilakukan untuk memilih data yang relevan dari basis data akademik mahasiswa sesuai dengan kebutuhan penelitian.

Seleksi data dilakukan untuk memilih data yang relevan dari basis data akademik mahasiswa sesuai dengan kebutuhan penelitian. Tahap ini bertujuan untuk mereduksi kompleksitas data sekaligus meningkatkan efisiensi proses komputasi pada tahap berikutnya.

Pembersihan data adalah proses membersihkan data dari kesalahan, duplikasi, atau data yang tidak lengkap agar data menjadi rapi dan siap digunakan. Tujuannya agar hasil analisis atau sistem yang dibuat menjadi lebih akurat. Contoh kegiatan pembersihan data antara lain menghapus data ganda, memperbaiki format yang salah, mengisi nilai yang kosong, dan menyamakan format penulisan.

Normalisasi data merupakan proses transformasi nilai atribut ke dalam skala tertentu, umumnya antara 0 hingga 1. Tahap ini diperlukan agar atribut dengan rentang nilai yang besar tidak mendominasi atau mempengaruhi hasil analisis secara tidak proporsional dibandingkan atribut lainnya.

3. Penerapan Algoritma *Decision Tree* c4.5

Decision Tree merupakan salah satu metode dalam pengolahan data yang digunakan untuk membangun model prediksi berbasis struktur pohon, baik dalam bentuk klasifikasi maupun

regresi, guna menghasilkan perkiraan terhadap kondisi atau kejadian di masa mendatang (Moerdyanto & Nuryana, 2023).

Dalam penelitian ini, varian yang digunakan adalah algoritma C4.5, yakni pengembangan lanjutan dari algoritma ID3. Peningkatan yang dihadirkan oleh C4.5 mencakup penanganan *missing value*, kemampuan memproses data kontinu, serta mekanisme *pruning* untuk menyederhanakan struktur pohon. Secara umum algoritma C4.5 digunakan untuk membangun pohon keputusan adalah sebagai berikut (Kusrini et al., 2009):

a. Pemilihan atribut sebagai akar
b. Pembentukan cabang untuk setiap nilai atribut
c. Pembagian kasus ke dalam masing-masing cabang
d. Pengulangan proses pada setiap cabang hingga seluruh kasus dalam cabang memiliki kelas yang sama.

Pemilihan atribut sebagai akar didasarkan pada nilai *Information Gain* tertinggi di antara seluruh atribut yang ada. Perhitungan nilai *Gain* dilakukan menggunakan persamaan sebagai berikut:

$$Gain\ (S,A) = Entropy(S) - \sum_{i=n}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (1)$$

Keterangan:
S = himpunan kasus
A = atribut
n = jumlah partisi atribut A
|Si| = jumlah kasus pada partisi ke-i
S| = jumlah kasus dalam S

Sebelum memperoleh nilai *Gain*, terlebih dahulu diperlukan nilai *Entropy*. *Entropy* digunakan untuk mengukur tingkat ketidakpastian atau seberapa informatif suatu atribut masukan dalam menghasilkan atribut keluaran yang diharapkan. Adapun rumus dasar mencari nilai *Entropy* sebagai berikut:

$$Entropy\ (S) = \sum_{i=n}^{n} \frac{|Si|}{|S|} - pi * log2pi \quad (2)$$

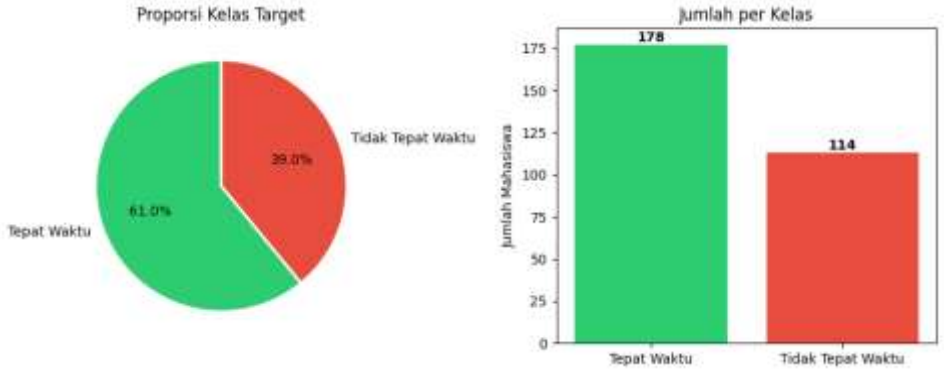
Keterangan:
S = himpunan kasus
A = Fitur
n = jumlah partisi S
pi = proporsi dari Si terhadap S

HASIL DAN PEMBAHASAN

1. Hasil

Dataset penelitian terdiri dari 292 rekaman data akademik mahasiswa program studi Teknik Informatika angkatan 2016 dengan delapan fitur prediktor, yaitu IP_SEM1 hingga IP_SEM4 dan SKS_SEM1 hingga SKS_SEM4, serta satu label target LABEL_KELULUSAN. Distribusi kelas menunjukkan 178 mahasiswa (61,0%) berlabel ‘Tepat Waktu’ dan 114

mahasiswa (39,0%) berlabel ‘Tidak Tepat Waktu’. Pemeriksaan konsistensi label terhadap jumlah semester yang ditempuh tidak menemukan data yang inkonsisten dari seluruh 292 rekaman. Pemeriksaan *outlier* menggunakan metode *Interquartile Range (IQR)* mendeteksi total 126 nilai pencilan yang tersebar di delapan fitur, dengan fitur SKS_SEM2 mencatat jumlah terbanyak yaitu 25 *outlier*. Data *outlier* ini tetap dipertahankan dalam analisis karena keterbatasan jumlah data.



Sumber: Hasil Penelitian (2026)

Gambar 1. Proporsi Kelas Target dan Jumlah per Kelas

Rata-rata IP per semester pada kelompok ‘Tepat Waktu’ lebih tinggi di setiap semester dibandingkan kelompok ‘Tidak Tepat Waktu’, dengan selisih terbesar pada IP_SEM4 (3,525 berbanding 2,915) dan terkecil pada IP_SEM1 (3,300 berbanding 2,829). Rata-rata fitur SKS antar kedua kelompok relatif tidak berbeda jauh, dengan selisih tertinggi pada SKS_SEM3 (21,253 berbanding 19,921). Korelasi fitur IP terhadap label berkisar antara -0,41 hingga -0,42, sedangkan korelasi antar sesama fitur IP tertinggi tercatat antara IP_SEM2 dan IP_SEM3 (r = 0,71) serta IP_SEM3 dan IP_SEM4 (r = 0,70). Korelasi fitur SKS terhadap label lebih rendah, berkisar antara -0,02 hingga -0,31.

Nilai *entropy* keseluruhan dataset diperoleh 0,9651. Hasil perhitungan *information gain* seluruh fitur disajikan pada Tabel 1. IP_SEM2 memperoleh *information gain* tertinggi 0,1432 dengan *threshold* 2,840 dan terpilih sebagai *root node*, diikuti IP_SEM1 (0,1429), IP_SEM4 (0,1264), dan IP_SEM3 (0,1091), sedangkan seluruh fitur SKS memperoleh nilai yang lebih rendah, yaitu di bawah 0,074.

Tabel 1. *Information Gain* dan *Threshold* Terbaik Setiap Fitur

Fitur	*Threshold*	*Information Gain*
IP_SEM2	2,840	0,1432
IP_SEM1	3,175	0,1429
IP_SEM4	3,215	0,1264
IP_SEM3	2,820	0,1091
SKS_SEM3	20,500	0,0736
SKS_SEM4	21,500	0,0670
SKS_SEM1	23,500	0,0099
SKS_SEM2	23,000	0,0099

Sumber: Hasil Penelitian (2026)

Model C4.5 dibangun dengan *criterion entropy* pada empat konfigurasi penanganan

ketidakseimbangan kelas, yaitu *Original* (tanpa *balancing*), *Undersampling (RandomUnderSampler)*, *SMOTE*, dan *Class Weight (balanced)*, masing-masing dioptimasi melalui *GridSearchCV* dengan 10-*fold Stratified Cross Validation* pada tiga skema pembagian data (70:30, 80:20, 90:10). *Split* terbaik per model dipilih berdasarkan *CV accuracy* tertinggi, yang ditampilkan pada Tabel 2.

Tabel 2. *Split* Terbaik dan Parameter Terbaik Setiap Konfigurasi Model

Model	*Balancing*	*Split* Terbaik	*CV Accuracy* (%)	*ccp_alpha*	*max_depth*	max_features	*min_samples_leaf*	*min_samples_split*
Original	Tidak Ada	70:30	79,86	0,020	4	*None*	1	2
Undersampling	*RandomUnderSampler*	80:20	78,57	0,015	6	*log2*	1	2
SMOTE	*SMOTE*	70:30	79,48	0,000	7	*sqrt*	1	20
Class Weight	*Class Weight (balanced)*	70:30	79,43	0,000	3	*log2*	1	20

Sumber: Hasil Penelitian (2026)

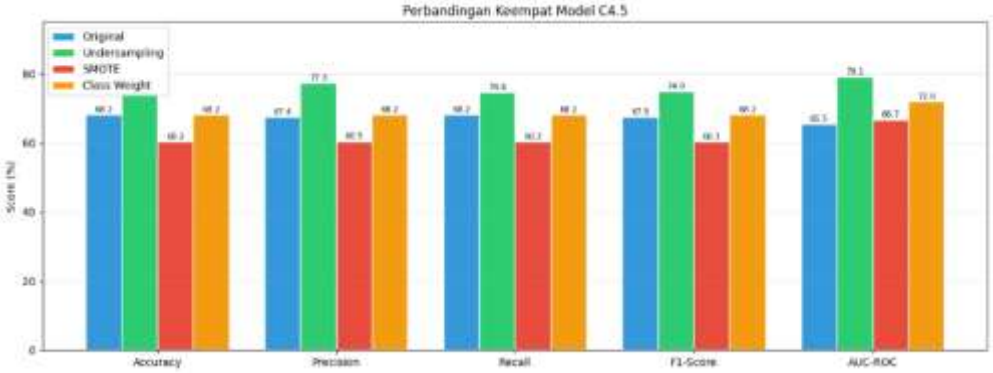
Hasil evaluasi keempat model pada data *testing* disajikan pada Tabel 3 yang memuat nilai *accuracy*, *precision*, *recall*, *F1-Score*, dan *AUC-ROC* masing-masing model beserta teknik *balancing* dan skema *split* yang digunakan.

Tabel 3. Perbandingan Hasil Evaluasi Keempat Model C4.5

Model	*Balancing*	*Split*	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*AUC-ROC (%)*
Original	Tidak Ada	70:30	68,18	67,44	68,18	67,48	65,50
Undersampling	*RandomUnderSampler*	80:20	74,58	77,29	74,58	74,89	79,11
SMOTE	*SMOTE*	70:30	60,23	60,45	60,23	60,33	66,72
Class Weight	*Class Weight (balanced)*	70:30	68,18	68,18	68,18	68,18	72,00

Sumber: Hasil Penelitian (2026)

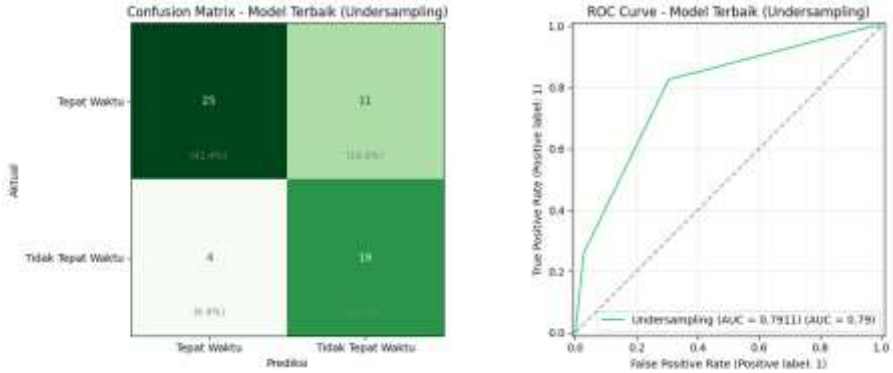
Perbandingan seluruh metrik keempat model divisualisasikan pada Gambar 2 berupa diagram batang berkelompok yang memuat nilai *accuracy*, *precision*, *recall*, *F1-Score*, dan *AUC-ROC* untuk setiap model. Berdasarkan *accuracy* tertinggi pada data *testing*, model *Undersampling* terpilih sebagai model terbaik dengan *accuracy* 74,58%, *precision* 77,29%, *recall* 74,58%, *F1-Score* 74,89%, dan *AUC-ROC* 79,11%.



Sumber: Hasil Penelitian (2026)

Gambar 2. Perbandingan Keempat Model C4.5 pada Seluruh Metrik Evaluasi

Berdasarkan *confusion matrix* pada model *Undersampling* pada Gambar 3, hasil evaluasi lanjutan pada model *Undersampling* menunjukkan bahwa dari 36 sampel aktual 'Tepat Waktu', 25 sampel diprediksi benar (42,4%) dan 11 diprediksi salah (18,6%), sedangkan dari 23 sampel aktual 'Tidak Tepat Waktu', 19 sampel diprediksi benar (32,2%) dan 4 diprediksi salah (6,8%).



Sumber: Hasil Penelitian (2026)

Gambar 3. *Confusion Matrix* dan *ROC Curve*

Learning curve model *Undersampling* mencatat *accuracy training* akhir 84,17% dan *accuracy* validasi akhir 78,57% dengan *gap* 5,60%. *Feature importance* model *Undersampling* menempatkan IP_SEM1 sebagai fitur paling berpengaruh dengan skor 0,3431, diikuti IP_SEM3 (0,2164), IP_SEM4 (0,1577), SKS_SEM4 (0,1304), SKS_SEM1 (0,0647), SKS_SEM3 (0,0597), dan SKS_SEM2 (0,0281), sedangkan hanya IP_SEM2 yang memperoleh skor 0,0000.

2. Pembahasan

Model C4.5 dengan teknik *RandomUnderSampler* menghasilkan performa terbaik di antara keempat konfigurasi yang diuji, dengan *accuracy* 74,58% dan *AUC-ROC* 79,11%. Keunggulan model *Undersampling* pada metrik *AUC-ROC* dibandingkan konfigurasi lainnya sejalan dengan penelitian yang menemukan bahwa teknik *undersampling* dapat meningkatkan kemampuan diskriminasi *classifier*, meskipun efektivitasnya bervariasi tergantung pada tingkat ketidakseimbangan data (Kim & Hwang, 2022). Peningkatan *AUC-ROC* yang diraih model *Undersampling*, dari 65,50% pada konfigurasi *Original* menjadi 79,11%,

mengindikasikan bahwa penyeimbangan kelas melalui reduksi sampel mayoritas mampu mendorong pohon keputusan untuk membentuk batas keputusan yang lebih adil terhadap kelas minoritas. Hal ini relevan dengan temuan yang menyimpulkan bahwa penanganan ketidakseimbangan kelas pada dataset pendidikan berpengaruh langsung terhadap kemampuan model untuk mengenali kelompok yang kurang terwakili, khususnya ketika rasio ketidakseimbangan berada pada kategori moderat seperti yang terjadi pada penelitian ini (61:39) (Wongvorachan et al., 2023).

Mekanisme di balik keunggulan *Undersampling* dapat dipahami melalui cara algoritma C4.5 memilih atribut menggunakan kriteria *information gain.* Ketika data tidak seimbang, kelas mayoritas mendominasi perhitungan *entropy*, sehingga pohon cenderung membentuk aturan yang bias ke arah kelas tersebut. Algoritma C4.5 yang menggunakan *entropy-based information gain* rentan terhadap bias kelas mayoritas pada dataset yang tidak seimbang karena perhitungan *entropy* pada *node* dipengaruhi oleh distribusi kelas di *data training* (Reddy & Chittineni, 2021). Dengan *RandomUnderSampler*, distribusi kelas *training* menjadi berimbang, sehingga *entropy* yang dihitung lebih merepresentasikan kedua kelas, yang pada akhirnya menghasilkan IP_SEM1 sebagai fitur paling berpengaruh dengan skor 0,3431, berbeda dari konfigurasi *Original* yang menempatkan IP_SEM2 sebagai fitur terdominant dengan skor 0,6501. Pergeseran distribusi *feature importance* antara kedua konfigurasi ini mencerminkan perbedaan mendasar dalam pola yang dipelajari oleh pohon dari data yang seimbang versus tidak seimbang.

Temuan bahwa fitur IP semester awal (IP_SEM1 dan IP_SEM3) menjadi paling berpengaruh dalam model *Undersampling* mendukung argumen bahwa prestasi akademik pada semester-semester awal merupakan prediktor kuat kelulusan tepat waktu. Sebuah penelitian menemukan bahwa fitur akademik berbasis nilai pada periode awal studi memiliki daya prediktif tertinggi dibandingkan fitur demografis dalam model prediksi performa mahasiswa, karena nilai awal mencerminkan adaptasi mahasiswa terhadap tuntutan akademik perguruan tinggi (Bilal et al., 2022). Rendahnya kontribusi fitur SKS dalam model *Undersampling* menunjukkan bahwa jumlah kredit yang ditempuh per semester tidak cukup membedakan antara mahasiswa yang lulus tepat waktu dan tidak, kemungkinan karena sebagian besar mahasiswa mengambil jumlah SKS yang relatif homogen sebagaimana tercermin dari nilai rata-rata SKS yang hampir seragam antar kelompok.

Model *SMOTE* justru menghasilkan performa terendah dengan *accuracy* 60,23%, lebih rendah dari konfigurasi *Original* maupun *Dummy Classifier*. Kondisi ini dapat dijelaskan melalui teori yang mengemukakan bahwa *SMOTE* memiliki keterbatasan pada dataset berukuran kecil karena sampel sintetis yang dihasilkan berisiko membentuk wilayah keputusan yang tumpang tindih, khususnya ketika sampel minoritas asli terlalu sedikit untuk menjadi dasar interpolasi yang representatif (Nemade et al., 2023). Dengan hanya 114 sampel kelas minoritas, interpolasi antar tetangga terdekat yang dilakukan *SMOTE* berpotensi menghasilkan sampel sintetis yang tidak mencerminkan pola data sesungguhnya, yang kemudian menyebabkan model *overfitting* pada pola sintetis tersebut dan gagal menggeneralisasi ke *data testing*. *Class Weight* juga tidak menghasilkan perbaikan berarti dibanding *Original*, yang mengindikasikan bahwa penalti berbasis bobot kelas tidak cukup mengubah batas keputusan pohon pada dataset berukuran kecil ini.

Dibandingkan dengan penelitian-penelitian terdahulu, *accuracy model* terbaik penelitian ini lebih rendah (Hasibuan & Mahdiana, 2023; Moerdyanto & Nuryana, 2023). Perbedaan ini dapat dipahami dalam kerangka yang mengemukakan bahwa performa model prediksi

kelulusan sangat dipengaruhi oleh ukuran dataset dan jumlah fitur yang tersedia, di mana dataset yang lebih besar cenderung menghasilkan model yang lebih stabil (Setiadi et al., 2024). Penelitian ini menggunakan 292 sampel dengan hanya 8 (delapan) fitur akademik, tanpa menyertakan variabel non-akademik seperti sosial-ekonomi atau keterlibatan dalam aktivitas kampus, yang kemungkinan berkontribusi pada keterbatasan performa model. Kekuatan penelitian ini terletak pada pengujian empat teknik penanganan ketidakseimbangan kelas secara sistematis dalam satu studi dan penggunaan *GridSearchCV* dengan 10-*fold Stratified Cross Validation* yang menjamin proses *tuning* yang tidak bias. Keterbatasannya adalah ukuran dataset yang kecil dan cakupan fitur yang terbatas pada data akademik kuantitatif semata, sehingga penelitian selanjutnya disarankan untuk menyertakan fitur non-akademik dan mengeksplorasi teknik *ensemble* yang lebih kuat pada dataset yang lebih besar.

KESIMPULAN

Berdasar hasil dan pembahasan yang telah dijabarkan, algoritma *Decision Tree* C4.5 dapat digunakan untuk memprediksi mahasiswa yang berpotensi mengalami keterlambatan kelulusan sehingga dapat dilakukan intervensi lebih awal oleh pihak program studi. Dari empat model yang diuji dalam penelitian ini, model C4.5 dengan teknik *RandomUnderSampler* menjadi model terbaik dengan akurasi sebesar 74,58%.

Faktor akademik yang paling berpengaruh dalam menentukan kelulusan tepat waktu mahasiswa adalah Indeks Prestasi (IP) semester awal, khususnya IP_SEM1 dan IP_SEM2, yang menunjukkan bahwa performa mahasiswa di awal perkuliahan menjadi penentu utama apakah mereka akan lulus tepat waktu atau tidak. Sebaliknya, jumlah SKS yang ditempuh per semester tidak terlalu berpengaruh karena cenderung seragam antar mahasiswa.

Meski demikian, performa model masih terbilang terbatas karena hanya menggunakan 292 data dengan 8 fitur akademik tanpa mempertimbangkan faktor non-akademik seperti kondisi sosial-ekonomi. Penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih besar serta mempertimbangkan atribut non-akademik sebagai variabel prediktor tambahan demi menghasilkan model prediksi yang lebih akurat dan representatif.

REFERENSI

BAN-PT. (2008). *Akreditasi program studi sarjana*. Badan Akreditasi Nasional Perguruan Tinggi.

Bilal, M., Omar, M., Anwar, W., Bokhari, R. H., & Choi, G. S. (2022). The role of demographic and academic features in a student performance prediction. *Scientific Reports*, *12*(1), 12508.

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

Jadhav, S. D., & Channe, H. P. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. *International Journal of Science and Research (IJSR)*, *5*(1), 1842–1845.

Kim, M., & Hwang, K. B. (2022). An empirical evaluation of sampling methods for the classification of imbalanced data. *PLOS ONE*, *17*(7), e0271260.

Kusrini, Luthfi, E. T., & Prabawati, T. A. (2009). *Algoritma data mining*. Andi.

Moerdyanto, O. P., & Nuryana, I. K. D. (2023). Prediksi kelulusan tepat waktu menggunakan pendekatan pohon keputusan algoritma decision tree. *Journal of Informatics and*

Computer Science (JINACS), *5*(01), 90–96.

Nemade, B., Bharadi, V., Alegavi, S. S., & Marakarkandy, B. (2023). A comprehensive review: SMOTE-based oversampling methods for imbalanced classification techniques, evaluation, and result comparisons. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(9s), 790–803.

Reddy, G. S., & Chittineni, S. (2021). Entropy based C4.5-SHO algorithm with information gain optimization in data mining. *PeerJ Computer Science*, *7*, e424.

Rohmawan, E. P. (2018). Prediksi kelulusan mahasiswa tepat waktu menggunakan metode decision tree dan artificial neural network. *Jurnal Ilmiah MATRIK*, *20*(1), 21–30.

Romadhona, A., Suprapedi, S., & Himawan, H. (2017). Prediksi kelulusan mahasiswa tepat waktu berdasarkan usia, jenis kelamin, dan indeks prestasi menggunakan algoritma decision tree. *Jurnal Cyberku*, *13*(1), 8.

Setiadi, H., Sanjaya, K., Wijayanto, A., Wardhani, D. W., & Cahyono, H. D. (2024). Comparative analysis of classification algorithms using feature selection techniques to predict on-time student graduation. *Ingenierie des Systemes d'Information*, *29*(4), 1365.

Putri, S. U., Irawan, E., Rizky, F., Bangsa, S. T., No, P. I. J. S. B., & Utara, S. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4. 5. *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen)*, *2*(1), 39-46.

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information*, *14*(1), 54.