Journal of Intelligent Systems and Digital Science
Volume 01, No. 01, Juni 2026: 51-62
e-ISSN: xxxx-xxxx
https://doi.org/10.66031/jinten.vxix.xxx

Model Prediksi Kelulusan Mahasiswa Berbasis Machine Learning dengan Seleksi Fitur

Rizal Fauzi Bakri *[1], Vina Dewi Uswatun [2]
[1,2] Universitas Medan Area (UMA), Indonesia
e-mail: fb_rizal@gmail.com

Received: 29-06-2026 *Accepted: 30-06-2026* *Published: 30-06-2026*

Abstrak

Kelulusan tepat waktu mahasiswa merupakan salah satu indikator penting dalam mengevaluasi kualitas penyelenggaraan pendidikan tinggi serta mendukung pengambilan keputusan akademik berbasis data. Penelitian ini bertujuan mengembangkan model prediksi kelulusan tepat waktu mahasiswa menggunakan algoritma Decision Tree, Random Forest, dan Naïve Bayes, serta menganalisis pengaruh seleksi fitur berbasis Information Gain terhadap performa model. Penelitian menggunakan pendekatan eksperimen komputasional dengan dataset simulasi yang terdiri atas 850 rekaman mahasiswa dan sembilan atribut akademik serta demografis. Tahapan penelitian meliputi *preprocessing* data, seleksi fitur menggunakan Information Gain, pemodelan *machine learning*, dan evaluasi menggunakan *10-fold cross-validation* dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil penelitian menunjukkan bahwa algoritma Random Forest memberikan performa terbaik dengan akurasi sebesar 89,4% sebelum seleksi fitur dan meningkat menjadi 91,2% setelah penerapan Information Gain. Seleksi fitur berhasil mereduksi atribut prediktor dari sembilan menjadi lima fitur utama, yaitu IPK, nilai rata-rata semester, jumlah SKS ditempuh, kehadiran, dan status beasiswa, sehingga meningkatkan efisiensi komputasi tanpa mengurangi kualitas prediksi. Temuan ini menunjukkan bahwa kombinasi Random Forest dan Information Gain berpotensi diterapkan sebagai bagian dari *early warning system* pada Sistem Informasi Akademik untuk mengidentifikasi mahasiswa yang berisiko mengalami keterlambatan kelulusan secara lebih dini dan akurat.

Kata Kunci: Prediksi Kelulusan; Machine Learning; Seleksi Fitur; Information Gain; Decision Tree

How to Cite:
Bakri, R. F., & Uswatun, V. D. (2026). Model Prediksi Kelulusan Mahasiswa Berbasis Machine Learning dengan Seleksi Fitur. *JINTEN: Journal of Intelligent Systems and Digital Science*, *1*(1), 51-62.

PENDAHULUAN

Keberhasilan mahasiswa dalam menyelesaikan studi tepat waktu merupakan salah satu indikator penting dalam menilai mutu perguruan tinggi karena berkaitan dengan efektivitas penyelenggaraan pendidikan, efisiensi pengelolaan sumber daya, serta capaian luaran akademik. Di Indonesia, capaian lulusan dan kinerja mahasiswa menjadi bagian dari indikator yang diperhatikan dalam sistem penjaminan mutu dan akreditasi perguruan tinggi, sehingga institusi dituntut untuk melakukan pemantauan terhadap potensi keterlambatan studi secara berkelanjutan (BAN-PT, 2019). Selain berpengaruh terhadap kualitas akademik, kemampuan perguruan tinggi dalam meningkatkan angka kelulusan tepat waktu juga berkontribusi terhadap peningkatan reputasi institusi dan pengambilan keputusan strategis berbasis data.

Perkembangan machine learning telah membuka peluang bagi perguruan tinggi untuk membangun sistem prediksi dini yang mampu mengidentifikasi mahasiswa berisiko tidak lulus tepat waktu sebelum memasuki tahap kritis. Menurut Romero dan Ventura (2020), pendekatan *Educational Data Mining* memungkinkan ekstraksi pola dari data akademik historis sehingga menghasilkan

prediksi yang lebih objektif dan akurat dibandingkan pendekatan manual. Sejalan dengan itu, Aldisa et al. (2026) melalui *systematic literature review* menunjukkan bahwa algoritma seperti Decision Tree, Random Forest, Support Vector Machine, dan Naïve Bayes merupakan metode yang paling banyak digunakan dalam prediksi performa akademik mahasiswa. Ahmed et al. (2025) juga menunjukkan bahwa penerapan model machine learning yang disertai pendekatan *explainable artificial intelligence* mampu meningkatkan akurasi prediksi sekaligus membantu pengambilan keputusan di lingkungan pendidikan tinggi.

Keberhasilan model prediksi tidak hanya dipengaruhi oleh algoritma yang digunakan, tetapi juga oleh kualitas fitur yang menjadi masukan model. Keberadaan atribut yang tidak relevan atau bersifat redundan dapat meningkatkan kompleksitas komputasi sekaligus menurunkan kemampuan generalisasi model. Han et al. (2022) menjelaskan bahwa *Information Gain* merupakan salah satu metode seleksi fitur berbasis filter yang efektif untuk mengukur kontribusi setiap atribut dalam mengurangi nilai *entropy* terhadap variabel target. Temuan tersebut diperkuat oleh Wijayaningrum et al. (2024) yang menunjukkan bahwa penerapan seleksi fitur dan penanganan data tidak seimbang mampu meningkatkan performa prediksi akademik mahasiswa, sedangkan Mohammed et al. (2025) menyimpulkan bahwa kombinasi algoritma klasifikasi dengan teknik *feature selection* menjadi salah satu pendekatan yang paling konsisten meningkatkan akurasi model pada berbagai studi *student performance prediction*.

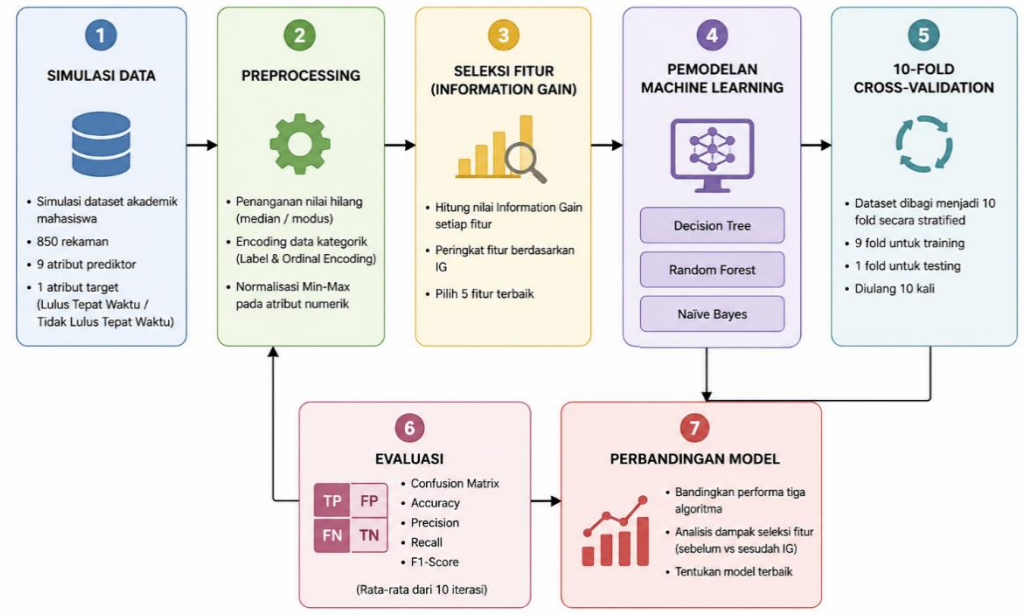
Meskipun penelitian mengenai prediksi performa akademik mahasiswa terus berkembang, sebagian besar penelitian masih berfokus pada prediksi prestasi belajar, risiko *dropout*, atau klasifikasi keberhasilan akademik secara umum. Salim et al. (2024) dalam kajian sistematisnya menunjukkan bahwa masih terdapat variasi yang cukup besar dalam pemilihan algoritma, teknik seleksi fitur, serta karakteristik dataset yang digunakan pada penelitian *Educational Data Mining*. Selain itu, Wang dan Yao (2025) mengemukakan bahwa sebagian besar penelitian masih menggunakan dataset institusi tertentu sehingga kemampuan generalisasi model terhadap konteks perguruan tinggi yang berbeda masih memerlukan pengujian lebih lanjut. Oleh karena itu, diperlukan penelitian yang membandingkan beberapa algoritma machine learning dengan pendekatan seleksi fitur dalam satu kerangka eksperimen yang sama agar diperoleh model yang lebih efektif dan mudah diimplementasikan.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan membangun model prediksi kelulusan tepat waktu mahasiswa menggunakan algoritma Decision Tree, Random Forest, dan Naïve Bayes serta mengevaluasi pengaruh seleksi fitur berbasis Information Gain terhadap performa masing-masing algoritma. Dataset yang digunakan merupakan data simulasi yang dirancang menyerupai karakteristik data akademik pada Sistem Informasi Akademik (SIAKAD) perguruan tinggi di Indonesia. Hasil penelitian ini diharapkan dapat memberikan rekomendasi mengenai algoritma dan kombinasi fitur yang paling efektif sebagai dasar pengembangan sistem peringatan dini (*early warning system*) untuk mengidentifikasi mahasiswa yang berisiko mengalami keterlambatan kelulusan sehingga intervensi akademik dapat dilakukan secara lebih cepat dan tepat sasaran.

METODE

Penelitian ini menggunakan pendekatan eksperimen komputasional dengan desain penelitian yang terdiri atas lima tahap utama, yaitu pengumpulan dan simulasi data, *preprocessing*, seleksi fitur, pemodelan *machine learning*, serta evaluasi dan perbandingan model. Pendekatan eksperimen dipilih karena memungkinkan evaluasi yang sistematis terhadap pengaruh seleksi fitur terhadap performa beberapa algoritma klasifikasi pada dataset yang sama. Tahapan penelitian mengacu pada alur umum

pengembangan model *machine learning* yang meliputi persiapan data, pemilihan fitur, pelatihan model, dan evaluasi performa (Han et al., 2022; Ahmed et al., 2025). Gambar 1 menyajikan alur penelitian secara keseluruhan.



Gambar 1. Alur Penelitian Prediksi Kelulusan Mahasiswa

Dataset dan Simulasi Data

Dataset yang digunakan dalam penelitian ini merupakan dataset simulasi yang dirancang untuk merepresentasikan karakteristik data akademik mahasiswa pada sistem informasi akademik perguruan tinggi di Indonesia. Perlu ditegaskan bahwa data ini merupakan simulasi penelitian yang dibuat dengan distribusi statistik yang realistis berdasarkan kajian literatur, bukan data rekaman mahasiswa aktual. Penggunaan dataset simulasi dilakukan karena keterbatasan akses terhadap data institusional yang bersifat konfidensial, sekaligus untuk memastikan tidak ada pelanggaran privasi data mahasiswa. Dataset terdiri dari 850 rekaman dengan distribusi kelas 58,6% Lulus Tepat Waktu (498 rekaman) dan 41,4% Tidak Lulus Tepat Waktu (352 rekaman). Tabel 1 menyajikan karakteristik dataset secara ringkas.

Tabel 1. Karakteristik Dataset Simulasi

Atribut	Nilai
Jumlah rekaman	850 mahasiswa
Kelas Lulus Tepat Waktu	498 (58,6%)
Kelas Tidak Lulus Tepat Waktu	352 (41,4%)
Jumlah atribut	9 atribut prediktor + 1 target
Jenis data	Campuran (numerik & kategorik)
Nilai hilang	1,8% (ditangani dengan modus/median)
Periode simulasi	2020–2024

Sembilan atribut prediktor yang digunakan dalam penelitian ini dipilih berdasarkan ketersediaannya pada Sistem Informasi Akademik perguruan tinggi dan relevansinya dengan faktor-faktor yang mempengaruhi kelulusan mahasiswa berdasarkan kajian literatur. Tabel 2 mendeskripsikan setiap atribut beserta skala pengukurannya.

Tabel 2. Definisi Variabel Penelitian

No.	Atribut	Skala	Keterangan
1	Jenis Kelamin	Nominal	L / P
2	Usia Masuk	Numerik	Usia saat pertama kali mendaftar (tahun)
3	Program Studi	Nominal	Kategori prodi (Saintek/Soshum)
4	IPK	Numerik (0–4)	Indeks Prestasi Kumulatif terakhir
5	Nilai Rata-rata Semester	Numerik (0–100)	Rata-rata nilai seluruh semester
6	Jumlah SKS Ditempuh	Numerik	Total SKS yang telah lulus
7	Kehadiran	Numerik (%)	Persentase kehadiran rata-rata
8	Status Beasiswa	Nominal	Ya / Tidak
9	Aktivitas Organisasi	Ordinal (0–2)	Tidak aktif=0, Anggota=1, Pengurus=2
10	Status Kelulusan	Nominal	Lulus Tepat Waktu / Tidak (target)

Preprocessing Data

Tahap *preprocessing* terdiri atas tiga langkah utama. Pertama, penanganan nilai hilang dilakukan menggunakan median untuk atribut numerik dan modus untuk atribut kategorik agar distribusi data tetap terjaga. Kedua, data kategorik dikonversi ke bentuk numerik menggunakan *Label Encoding* dan *Ordinal Encoding*. Ketiga, normalisasi Min-Max diterapkan pada atribut numerik untuk menyamakan rentang nilai sehingga mengurangi pengaruh perbedaan skala terhadap algoritma klasifikasi, khususnya Gaussian Naïve Bayes. Tahapan *preprocessing* tersebut merupakan prosedur yang umum digunakan dalam pengembangan model *machine learning* untuk meningkatkan kualitas data dan stabilitas model (Han et al., 2022; Pedregosa et al., 2011).

Seleksi Fitur dengan Information Gain

Information Gain (IG) merupakan metode seleksi fitur berbasis filter yang mengukur pengurangan entropy pada variabel target Y akibat diketahuinya nilai fitur X. Secara formal, Information Gain didefinisikan sebagai selisih antara entropy sebelum dan sesudah pemisahan berdasarkan nilai fitur tertentu (Han et al., 2022). Entropy H(Y) dirumuskan sebagai:

$$H(Y) = -\Sigma\, p(y)\, log_2\, p(y) \quad (1)$$

di mana p(y) adalah proporsi kelas y dalam dataset. Information Gain fitur X terhadap Y kemudian dihitung sebagai H(Y) − H(Y|X). Fitur dengan nilai IG yang lebih tinggi dianggap lebih informatif dan diprioritaskan dalam pemodelan. Metode Information Gain dipilih karena bersifat model-agnostic, mudah diinterpretasi, dan efisien secara komputasional untuk dataset dengan jumlah atribut moderat seperti yang digunakan dalam penelitian ini (Wijayaningrum et al., 2024; Mohammed et al. 2025).

Dalam penelitian ini, seleksi fitur dilakukan dengan memilih lima fitur teratas berdasarkan nilai IG dari sembilan fitur yang tersedia. Pemilihan ambang batas lima fitur didasarkan pada hasil eksperimen awal yang menunjukkan bahwa penambahan fitur ke-6 dan seterusnya tidak memberikan peningkatan akurasi yang signifikan, sementara menambah kompleksitas model. Tabel 3 menyajikan peringkat fitur berdasarkan nilai Information Gain.

Tabel 3. Peringkat Fitur Berdasarkan Information Gain

Peringkat	Fitur	Nilai IG	Status
1	IPK	0,412	Dipilih
2	Nilai Rata-rata Semester	0,378	Dipilih
3	Jumlah SKS Ditempuh	0,341	Dipilih
4	Kehadiran	0,289	Dipilih
5	Status Beasiswa	0,214	Dipilih
6	Aktivitas Organisasi	0,156	Tidak dipilih
7	Program Studi	0,098	Tidak dipilih
8	Usia Masuk	0,067	Tidak dipilih
9	Jenis Kelamin	0,031	Tidak dipilih

Algoritma Machine Learning

Decision Tree

Decision Tree adalah algoritma pembelajaran terawasi yang membangun struktur pohon keputusan secara rekursif dengan memilih fitur terbaik pada setiap node menggunakan kriteria pemisahan seperti Gini Impurity atau Information Gain (Quinlan, 1986; Breiman et al., 1984). Kelebihan utama Decision Tree adalah interpretabilitasnya yang tinggi karena aturan keputusan yang dihasilkan dapat dibaca secara langsung oleh manusia. Dalam penelitian ini, algoritma Decision Tree diterapkan menggunakan kriteria Gini Impurity dengan kedalaman maksimum pohon dibatasi untuk mencegah overfitting.

Random Forest

Random Forest merupakan metode *ensemble learning* yang membangun sejumlah pohon keputusan menggunakan teknik *bootstrap sampling* dan *random feature subspace*, kemudian menggabungkan hasil prediksi melalui mekanisme *majority voting* (Breiman, 2001). Berbagai penelitian menunjukkan bahwa Random Forest memiliki kemampuan generalisasi yang baik pada prediksi performa akademik mahasiswa dan relatif tahan terhadap *overfitting* dibandingkan pohon keputusan tunggal (Ahmed et al., 2025; Gu & Zhang, 2026). Pada penelitian ini digunakan 100 pohon keputusan dengan pemilihan fitur secara acak pada setiap proses pemisahan.

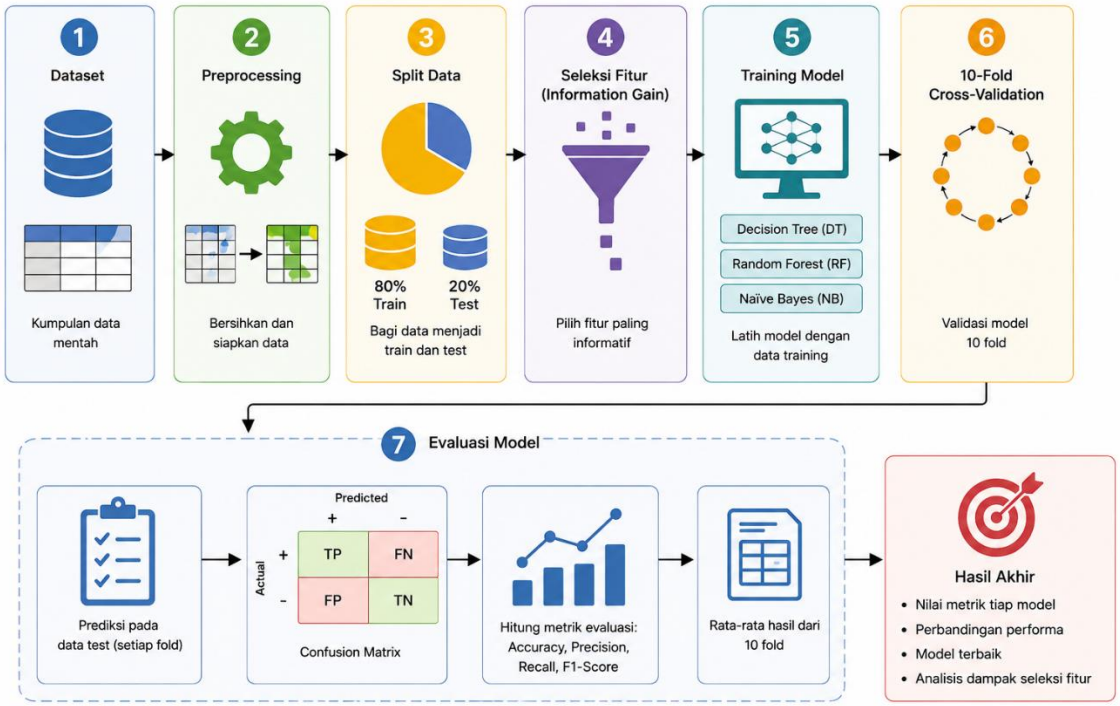
Naïve Bayes

Naïve Bayes merupakan algoritma klasifikasi probabilistik yang didasarkan pada Teorema Bayes dengan asumsi independensi antar atribut (Mitchell, 1997). Walaupun asumsi tersebut tidak selalu terpenuhi, algoritma ini tetap memberikan performa yang baik pada berbagai kasus klasifikasi

dengan jumlah data sedang hingga besar (Rish, 2001). Penelitian ini menggunakan Gaussian Naïve Bayes untuk menangani atribut numerik yang diasumsikan mengikuti distribusi normal.

Evaluasi Model

Gambar 2 menunjukkan tahapan *pipeline machine learning* mulai dari *preprocessing*, seleksi fitur, pelatihan model hingga evaluasi.



Gambar 2. Tahapan *Machine Learning* Dalam Penelitian

Evaluasi model dilakukan menggunakan metode 10-fold cross-validation untuk memperoleh estimasi performa yang lebih stabil dan mengurangi bias akibat pembagian data latih dan data uji. Setiap lipatan digunakan secara bergantian sebagai data uji sementara sembilan lipatan lainnya digunakan sebagai data latih. Kinerja model diukur menggunakan metrik Accuracy, Precision, Recall, dan F1-Score, sedangkan Confusion Matrix digunakan untuk menganalisis distribusi prediksi benar dan salah pada setiap kelas. Pendekatan evaluasi tersebut merupakan prosedur yang umum digunakan dalam penelitian klasifikasi berbasis *machine learning* karena mampu memberikan estimasi performa yang lebih andal dibandingkan pembagian data tunggal (Pedregosa et al., 2011; Ahmed et al., 2025).

HASIL DAN PEMBAHASAN

Karakteristik Dataset dan Hasil Preprocessing

Dari 850 rekaman dalam dataset simulasi, ditemukan sebanyak 15 rekaman (1,8%) mengandung nilai hilang yang tersebar pada atribut Kehadiran (8 kasus) dan IPK (7 kasus). Seluruh nilai hilang berhasil ditangani melalui imputasi median tanpa perlu menghapus rekaman manapun. Distribusi kelas menunjukkan ketidakseimbangan ringan dengan rasio 58,6:41,4 yang masih tergolong dapat ditangani tanpa teknik oversampling tambahan. Setelah preprocessing, dataset final terdiri dari 850 rekaman bersih dengan sembilan fitur yang telah dienkode dan dinormalisasi, siap untuk tahap seleksi fitur dan pemodelan.

Hasil Seleksi Fitur

Hasil perhitungan Information Gain yang ditampilkan pada Tabel 3 menunjukkan bahwa IPK (IG = 0,412) merupakan fitur paling informatif untuk memprediksi status kelulusan, diikuti oleh Nilai Rata-rata Semester (IG = 0,378) dan Jumlah SKS Ditempuh (IG = 0,341). Ketiga fitur ini secara intuitif memiliki keterkaitan langsung dengan kesiapan akademik mahasiswa untuk menyelesaikan studi. Status Beasiswa (IG = 0,214) menempati posisi kelima dan tetap dimasukkan ke dalam subset fitur terpilih karena memberikan informasi tentang dukungan finansial yang turut mempengaruhi kelancaran studi. Di sisi lain, Jenis Kelamin (IG = 0,031) dan Usia Masuk (IG = 0,067) menunjukkan nilai IG yang sangat rendah, mengindikasikan bahwa kedua atribut tersebut memiliki kontribusi prediktif yang minimal dan dapat diabaikan tanpa mengorbankan akurasi model secara signifikan.

Performa Model tanpa Seleksi Fitur

Tabel 4 menyajikan hasil evaluasi ketiga algoritma menggunakan seluruh sembilan fitur (tanpa seleksi fitur) berdasarkan rata-rata 10-fold cross-validation.

Tabel 4. Performa Model Tanpa Seleksi Fitur (10-Fold CV)

Algoritma	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	84,1	83,6	82,9	83,2
Random Forest	89,4	88,7	89,1	88,9
Naïve Bayes	82,3	81,8	80,5	81,1

Performa Model dengan Seleksi Fitur

Tabel 5 menyajikan hasil evaluasi setelah penerapan seleksi fitur Information Gain dengan lima fitur terpilih. Perbandingan langsung antara Tabel 4 dan Tabel 5 memperlihatkan peningkatan performa yang konsisten pada ketiga algoritma setelah seleksi fitur diterapkan.

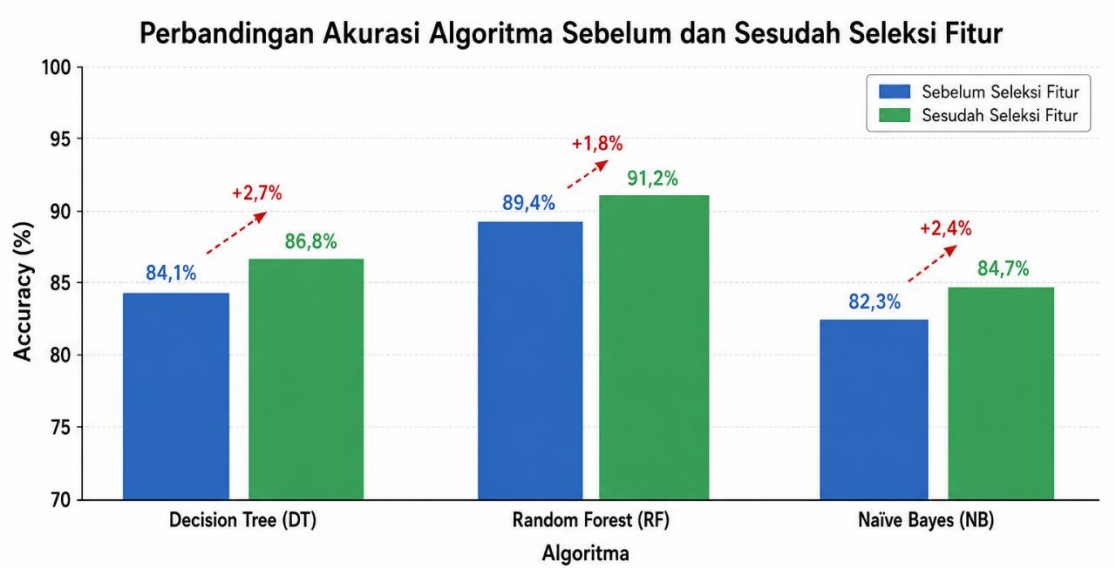
Tabel 5. Performa Model Dengan Seleksi Fitur Information Gain (10-Fold CV)

Algoritma	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	86,8	86,2	85,7	85,9
Random Forest	91,2	90,8	91,0	90,9
Naïve Bayes	84,7	84,1	83,8	83,9

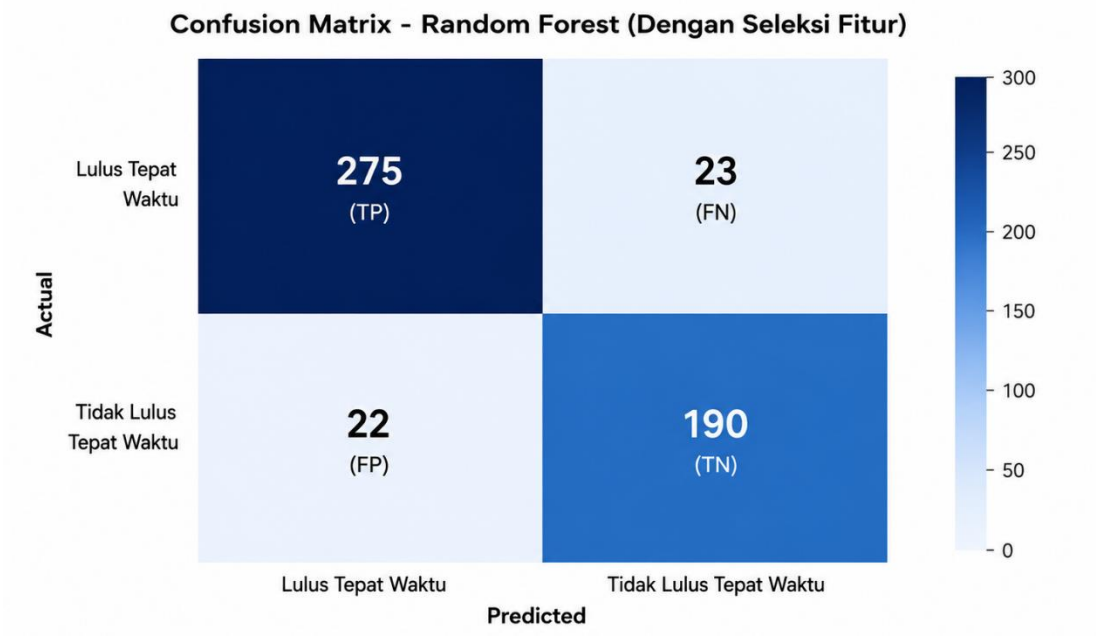
Peningkatan akurasi yang paling signifikan terjadi pada Decision Tree, yakni sebesar 2,7 poin persentase (dari 84,1% menjadi 86,8%), diikuti Naïve Bayes sebesar 2,4 poin persentase, dan Random Forest sebesar 1,8 poin persentase. Meskipun peningkatan pada Random Forest tampak lebih kecil dalam nilai absolut, hal ini dapat dijelaskan oleh sifat inheren algoritma ensemble yang sudah secara otomatis menerapkan seleksi fitur implisit melalui mekanisme random subspace pada setiap pemisahan pohon, sehingga efek marginal dari seleksi fitur eksternal relatif lebih kecil.

Analisis Confusion Matrix

Gambar 3 menampilkan perbandingan akurasi ketiga algoritma sebelum dan sesudah seleksi fitur secara visual, sementara Gambar 4 menyajikan Confusion Matrix model Random Forest sebagai model terbaik.



Gambar 3. Perbandingan Akurasi Algoritma Sebelum Dan Sesudah Seleksi Fitur



Gambar 4. Confusion matrix model Random Forest dengan seleksi fitur

Dari Confusion Matrix model Random Forest (Gambar 4), model berhasil mengklasifikasikan 275 mahasiswa Lulus Tepat Waktu dengan benar (True Positive) dan 190 mahasiswa Tidak Lulus Tepat Waktu dengan benar (True Negative). Terdapat 22 kasus False Positive, yakni mahasiswa yang diprediksi Lulus Tepat Waktu tetapi kenyataannya tidak, dan 23 kasus False Negative, yakni mahasiswa yang diprediksi Tidak Lulus Tepat Waktu tetapi sebenarnya lulus tepat waktu. Dalam konteks sistem peringatan dini, False Negative merupakan kesalahan yang lebih kritis karena berarti mahasiswa berisiko yang seharusnya mendapat intervensi justru tidak terdeteksi oleh sistem.

Pembahasan

Hasil penelitian menunjukkan bahwa algoritma Random Forest menghasilkan performa terbaik dibandingkan Decision Tree dan Naïve Bayes, dengan akurasi mencapai 91,2% setelah penerapan seleksi fitur Information Gain. Temuan ini menunjukkan bahwa metode *ensemble learning* mampu memberikan kemampuan generalisasi yang lebih baik dibandingkan algoritma klasifikasi tunggal karena memanfaatkan kombinasi banyak pohon keputusan untuk mengurangi varians prediksi. Ahmed et al. (2025) melaporkan bahwa pendekatan *ensemble machine learning* menghasilkan performa prediksi akademik yang lebih stabil dan akurat dibandingkan model individual, terutama ketika dikombinasikan dengan proses seleksi fitur. Temuan serupa juga dikemukakan oleh Gu dan Zhang (2026), yang menyatakan bahwa Random Forest memiliki keseimbangan yang baik antara akurasi, interpretabilitas, dan ketahanan terhadap *overfitting* pada prediksi hasil akademik mahasiswa.

Penerapan Information Gain terbukti meningkatkan performa seluruh algoritma yang diuji. Akurasi Decision Tree meningkat dari 84,1% menjadi 86,8%, Random Forest dari 89,4% menjadi 91,2%, sedangkan Naïve Bayes meningkat dari 82,3% menjadi 84,7%. Hasil tersebut menunjukkan bahwa penghapusan atribut yang kurang relevan mampu meningkatkan kualitas proses pembelajaran model sekaligus mengurangi kompleksitas komputasi. Wijayaningrum et al. (2024) menunjukkan bahwa kombinasi seleksi fitur dan penanganan data tidak seimbang mampu meningkatkan performa model prediksi akademik secara signifikan. Selain itu, Mohammed et al. (2025) melalui kajian literatur menyimpulkan bahwa teknik *feature selection* merupakan salah satu faktor utama yang memengaruhi peningkatan akurasi, efisiensi komputasi, serta interpretabilitas model pada penelitian *Educational Data Mining*.

Hasil perhitungan Information Gain menunjukkan bahwa IPK merupakan atribut dengan nilai Information Gain tertinggi (0,412), diikuti oleh Nilai Rata-rata Semester (0,378) dan Jumlah SKS Ditempuh (0,341). Ketiga atribut tersebut menggambarkan performa akademik kumulatif mahasiswa sehingga secara logis memiliki hubungan yang kuat terhadap peluang kelulusan tepat waktu. Aldisa et al. (2026) melalui *systematic literature review* menemukan bahwa IPK, jumlah SKS, nilai akademik, dan kehadiran merupakan variabel yang paling sering digunakan sebagai prediktor utama dalam penelitian prediksi kelulusan mahasiswa. Ahmed et al. (2025) juga menunjukkan bahwa indikator akademik memiliki kontribusi prediktif yang lebih besar dibandingkan karakteristik demografis sehingga layak diprioritaskan pada proses pemodelan.

Sebaliknya, atribut Jenis Kelamin dan Usia Masuk memperoleh nilai Information Gain yang rendah sehingga tidak memberikan kontribusi berarti terhadap peningkatan performa model. Hasil ini menunjukkan bahwa tidak semua atribut yang tersedia pada Sistem Informasi Akademik memiliki kemampuan diskriminatif yang sama dalam membedakan mahasiswa yang lulus tepat waktu dan yang mengalami keterlambatan. Salim et al. (2024) menjelaskan bahwa efektivitas model prediksi akademik sangat dipengaruhi oleh kualitas atribut yang digunakan, sehingga pemilihan fitur yang relevan menjadi tahapan penting dalam pengembangan model klasifikasi. Temuan tersebut juga didukung oleh Chen et al. (2025) yang menunjukkan bahwa pengurangan atribut yang redundan mampu meningkatkan efektivitas model prediksi sekaligus mendukung implementasi pembelajaran adaptif.

Dari perspektif implementasi praktis, model Random Forest dengan lima atribut terpilih menawarkan keseimbangan antara akurasi prediksi dan kemudahan implementasi pada Sistem Informasi Akademik perguruan tinggi. Seluruh atribut yang digunakan, yaitu IPK, Nilai Rata-rata Semester, Jumlah SKS Ditempuh, Kehadiran, dan Status Beasiswa, merupakan data yang telah

tersedia secara rutin sehingga tidak memerlukan proses pengumpulan data tambahan. Wang dan Yao (2025) menyatakan bahwa model prediksi berbasis *machine learning* dapat dimanfaatkan sebagai bagian dari early warning system untuk mengidentifikasi mahasiswa yang berisiko mengalami keterlambatan studi sehingga intervensi akademik dapat dilakukan lebih cepat dan lebih tepat sasaran.

Penelitian ini masih memiliki beberapa keterbatasan. Pertama, penggunaan dataset simulasi belum sepenuhnya mampu menggambarkan kompleksitas data akademik nyata pada berbagai perguruan tinggi. Kedua, penelitian ini hanya menggunakan sembilan atribut akademik sehingga belum mempertimbangkan faktor sosial, ekonomi, psikologis, maupun perilaku belajar mahasiswa. Ketiga, model belum divalidasi menggunakan dataset lintas institusi sehingga kemampuan generalisasinya masih perlu diuji lebih lanjut. Salim et al. (2024) menegaskan bahwa validasi pada berbagai institusi pendidikan menjadi salah satu tantangan utama dalam penelitian *Educational Data Mining*. Oleh karena itu, penelitian selanjutnya disarankan menggunakan data institusional nyata, mengeksplorasi metode seleksi fitur lain seperti *Recursive Feature Elimination*, serta membandingkan algoritma yang lebih mutakhir seperti XGBoost, LightGBM, maupun Explainable Artificial Intelligence untuk meningkatkan akurasi sekaligus interpretabilitas model.

KESIMPULAN

Penelitian ini berhasil mengembangkan dan mengevaluasi model prediksi kelulusan tepat waktu mahasiswa menggunakan algoritma Decision Tree, Random Forest, dan Naïve Bayes yang dikombinasikan dengan metode seleksi fitur berbasis Information Gain. Hasil penelitian menunjukkan bahwa Random Forest merupakan algoritma dengan performa terbaik, memperoleh akurasi sebesar 91,2%, lebih tinggi dibandingkan Decision Tree (86,8%) dan Naïve Bayes (84,7%). Penerapan Information Gain terbukti meningkatkan kinerja seluruh algoritma dengan peningkatan akurasi berkisar antara 1,8% hingga 2,7%, sekaligus mereduksi jumlah atribut prediktor dari sembilan menjadi lima fitur yang paling informatif sehingga menghasilkan model yang lebih efisien dan mudah diinterpretasikan.

Hasil seleksi fitur mengidentifikasi IPK, Nilai Rata-rata Semester, dan Jumlah SKS Ditempuh sebagai atribut dengan kontribusi prediktif tertinggi terhadap status kelulusan mahasiswa. Temuan ini menunjukkan bahwa indikator akademik yang tersedia pada Sistem Informasi Akademik (SIAKAD) memiliki potensi yang kuat untuk dimanfaatkan dalam membangun sistem prediksi kelulusan tanpa memerlukan pengumpulan data tambahan. Dengan demikian, model yang dihasilkan berpotensi diimplementasikan sebagai bagian dari early warning system untuk membantu perguruan tinggi mengidentifikasi mahasiswa yang berisiko mengalami keterlambatan kelulusan sehingga intervensi akademik dapat dilakukan secara lebih dini, tepat sasaran, dan berbasis data.

Meskipun demikian, penelitian ini masih memiliki keterbatasan karena menggunakan dataset simulasi yang belum sepenuhnya merepresentasikan kondisi nyata di berbagai perguruan tinggi serta belum mempertimbangkan faktor non-akademik yang dapat memengaruhi keberhasilan studi mahasiswa. Oleh karena itu, penelitian selanjutnya disarankan melakukan validasi menggunakan data institusional nyata dari berbagai perguruan tinggi, mengeksplorasi metode seleksi fitur yang lebih adaptif, serta membandingkan performa algoritma yang lebih mutakhir, seperti XGBoost, LightGBM, maupun pendekatan Explainable Artificial Intelligence (XAI), untuk meningkatkan akurasi, interpretabilitas, dan kemampuan generalisasi model.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah memberikan dukungan dalam penyelesaian penelitian ini. Penelitian ini tidak mendapatkan dukungan pendanaan dari lembaga tertentu.

DAFTAR PUSTAKA

Ahmed, E. (2024). *Student performance prediction using machine learning algorithms*. *Applied Computational Intelligence and Soft Computing, 2024*(1), 4067721. https://doi.org/10.1155/2024/4067721

Ahmed, W., Wani, M. A., Plawiak, P., et al. (2025). Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions. *Scientific Reports, 15*, 26879. https://doi.org/10.1038/s41598-025-12353-4

Aldisa, R. T., Rochim, A. F., & Triayudi, A. (2026). Student achievement prediction models: A PRISMA-based systematic literature review. *Journal of Information Systems and Informatics, 8*(2), 2638–2663. https://doi.org/10.63158/journalisi.v8i2.1526

Almalki, A. J. (2021). *Accuracy analysis of educational data mining using feature selection algorithm*. https://doi.org/10.48550/arXiv.2107.10669

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth.

Chen, Y., Sun, J., Wang, J., Zhao, L., Song, X., & Zhai, L. (2025). *Machine learning-driven student performance prediction for enhancing tiered instruction*. https://doi.org/10.48550/arXiv.2502.03143

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3*, 1289–1305.

Gu, H., & Zhang, Y. (2026). Efficient and interpretable machine learning for student academic outcome prediction. *Mathematics, 14*(4), 626. https://doi.org/10.3390/math14040626

Han, J., Kamber, M., & Pei, J. (2022). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Mohammed, A. A. A., Kanaan, A. M. J., Meng, G. C., Elsmany, E. F., Tyng, C. M., & Ibrahim, A. O. (2025). Feature selection for student performance prediction: A review. In *2025 IEEE International Conference on Computing (ICOCO)* (pp. 195–200). IEEE. https://doi.org/10.1109/ICOCO67189.2025.11334101

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.

Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* (Vol. 3, pp. 41–46).

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355.

Salim, M., Al-Din, N., & Al Abdulqader, H. A. (2024). Students' academic performance prediction using educational data mining and machine learning: A systematic review. *International Journal of Research and Innovation in Social Science*. https://doi.org/10.47772/IJRISS

Setiadi, H., Josi, F. A. I., Winarno, Doewes, A., & Damayanti, R. W. (2025). Optimizing student performance prediction: A comparative analysis of regression algorithms and feature selection techniques on LMS log data. *Vietnam Journal of Computer Science*. https://doi.org/10.1142/S2196888825500265

Wang, C., & Yao, S. (2025). Enhancing student dropout and academic success prediction using machine learning and over-sampling techniques. *ICCK Transactions on Educational Data Mining*, *1*(1), 36–43. https://doi.org/10.62762/TEDM.2025.732573

Wijayaningrum, V. N., Kirana, A. P., & Putri, I. K. (2024). Student academic performance prediction framework with feature selection and imbalanced data handling. *Jurnal Ilmiah Kursor*, *12*(3), 123–134. https://doi.org/10.21107/kursor.v12i3.356