

Model Prediktif Kelulusan Mahasiswa Berbasis Machine Learning Menggunakan Pipeline Terintegrasi dan Hyperparameter Tuning

Mundirin^{1*}, Deden Hedin², Idawati³, Ibrahim Latief⁴, Mohamad Lili⁵

^{1,2,3,4,5}Institut Sains dan Teknologi Al-Kamal, Indonesia.

^{1*}mundirin.imung@gmail.com



*Mundirin

Histori Artikel:

Submit: 2026-01-11

Diterima: 2026-02-01

Dipublikasikan: 2026-02-04

Kata Kunci:

Kelulusan Mahasiswa; Machine Learning; Pipeline Terintegrasi; Pendidikan Tinggi; Random Forest

ABSTRAK

Keterlambatan dalam menyelesaikan studi mahasiswa adalah masalah penting dalam pendidikan tinggi karena memengaruhi efisiensi akademik, akreditasi program studi, dan kualitas para lulusan. Penelitian ini bertujuan untuk mengembangkan model prediksi kelulusan mahasiswa yang didasarkan pada machine learning dengan menggunakan pipeline yang terintegrasi. Pipeline ini meliputi tahap pengolahan data, pembuatan model, dan optimasi hyperparameter secara menyeluruh. Dataset yang digunakan diperoleh dari data akademik mahasiswa selama delapan semester dengan total beban studi sebanyak 146 SKS. Data ini mencakup variabel numerik dan kategorikal, termasuk capaian IPK, jumlah SKS yang lulus setiap semester, beban studi, serta karakteristik latar belakang mahasiswa. Proses pra-pemrosesan dilakukan dengan menggunakan ColumnTransformer yang menggabungkan StandardScaler untuk fitur numerik dan *One-Hot Encoding* untuk fitur kategori. Model klasifikasi dikembangkan dengan menggunakan algoritma *Random Forest* dan dioptimalkan dengan *GridSearchCV* guna mencapai pengaturan hyperparameter yang paling baik. Penilaian model dilakukan dengan menggunakan metrik akurasi, matriks kebingungan, dan laporan klasifikasi. Temuan dari penelitian ini menunjukkan bahwa model tersebut dapat mencapai tingkat akurasi sebesar 81%, yang mencerminkan kemampuan yang baik dalam memprediksi klasifikasi mahasiswa yang lulus tepat waktu dan yang terlambat lulus. Analisis fitur menunjukkan bahwa rata-rata Indeks Prestasi Kumulatif, jumlah Satuan Kredit Semester yang diselesaikan setiap semester, dan konsistensi dalam beban studi adalah faktor-faktor utama yang memengaruhi ketepatan waktu dalam menyelesaikan studi. Penerapan saluran terintegrasi terbukti ampuh dalam menjaga konsistensi pra-pemrosesan dan mengurangi kemungkinan kebocoran data. Model yang dikembangkan memiliki potensi untuk diterapkan sebagai sistem peringatan awal guna mendukung pengambilan keputusan akademik yang didasarkan pada data.

Jurnal Pendidikan Sains dan Komputer is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

LATAR BELAKANG

Ketepatan waktu penyelesaian studi mahasiswa adalah salah satu tanda penting dari keberhasilan pelaksanaan pendidikan tinggi. Tingginya angka keterlambatan dalam menyelesaikan studi tidak hanya mempengaruhi mahasiswa secara pribadi, tetapi juga berdampak pada efisiensi lembaga, status akreditasi program studi, serta mutu lulusan yang dihasilkan. Maka dari itu, institusi pendidikan tinggi perlu memiliki sistem prediktif yang dapat mendeteksi kemungkinan keterlambatan studi sejak awal.

Kemajuan dalam teknologi informasi dan tersedianya data akademik mahasiswa memberikan kesempatan



untuk menggunakan machine learning sebagai metode analisis prediktif di sektor pendidikan (penambahan data pendidikan). Berbagai studi menunjukkan bahwa algoritma pembelajaran mesin dapat mengenali pola-pola kompleks dalam data akademik yang sulit terlihat dengan metode statistik tradisional (W. Xing, D. Gašević, 2020)

Namun, banyak penelitian sebelumnya masih menemui hambatan dalam hal konsistensi dalam preprocessing data, kemungkinan terjadinya data leakage, serta keterbatasan dalam menerapkan model ke dalam sistem yang nyata. Oleh karena itu, diperlukan suatu pendekatan yang lebih sistematis melalui saluran terintegrasi yang menggabungkan tahap praproses, pemodelan, dan pengoptimalan hyperparameter dalam satu rangkaian proses (Oyelade, 2020),(Shoaib & Taie, 2020) .

Rumusan masalah dalam penelitian ini adalah bagaimana cara merancang model prediktif untuk kelulusan mahasiswa yang menggunakan machine learning melalui suatu sistem terpadu, yang meliputi langkah-langkah pembersihan data, pembentukan model, dan penyesuaian hyperparameter. Selain itu, penelitian ini juga bertujuan untuk menilai kinerja model *Random Forest* dalam meramalkan ketepatan waktu kelulusan berdasarkan data akademik dari delapan semester. Di samping itu, penelitian ini berusaha untuk mengidentifikasi faktor-faktor akademik penting yang paling berpengaruh terhadap kelulusan mahasiswa berdasarkan analisis fitur model.

STUDI LITERATUR

1. Keterlambatan Studi Mahasiswa

Keterlambatan studi mahasiswa dijelaskan sebagai ketidakmampuan seorang mahasiswa untuk menyelesaikan pendidikan dalam jangka waktu yang telah ditentukan oleh kurikulum. Sebuah studi yang dilakukan oleh (Kimmons, W. Xing, 2021) mengungkapkan bahwa faktor-faktor akademik, seperti Indeks Prestasi Kumulatif (IPK), jumlah Satuan Kredit Semester (SKS) yang diselesaikan setiap semester, serta pola pemilihan beban studi, memiliki dampak yang signifikan terhadap kemungkinan keterlambatan dalam menyelesaikan studi. Namun, sebagian besar penelitian sebelumnya masih bergantung pada metode statistik deskriptif atau regresi linier yang cenderung memperkirakan hubungan linier antara variabel. Pendekatan tersebut mungkin tidak dapat dengan baik menangkap pola non-linear dan interaksi rumit di antara berbagai faktor akademik, sehingga dapat membatasi ketepatan dalam suatu prediksi mengenai risiko keterlambatan studi. Penelitian ini bertujuan untuk mengatasi keterbatasan tersebut dengan menggunakan metode machine learning yang lebih sesuai dengan sifat data akademik yang berubah-ubah dan beragam.

2. Machine Learning dalam Pendidikan Tinggi.

Pembelajaran mesin telah banyak diterapkan dalam sektor pendidikan tinggi untuk meramalkan kinerja akademis, risiko *putus sekolah*, dan kelulusan mahasiswa. Penelitian yang dilakukan oleh (B. Al-Bargi, 2021), (P. Kaur, 2021), (Akmar Efendi, Iskandar Fitri, 2025), (Rizal Bakri, 2025) menunjukkan bahwa algoritma *supervised learning* seperti Random Forest dan Support Vector Machine terbukti efektif dalam menangani data pendidikan yang berukuran besar dan beragam. Namun, beberapa penelitian tersebut masih memisahkan tahap pra-pemrosesan dan pemodelan, yang dapat meningkatkan kemungkinan terjadinya inkonsistensi data serta *data leakage* saat model diterapkan pada data yang baru. Di samping itu, penilaian model sering kali hanya fokus pada metrik akurasi tanpa memperhatikan adanya ketidakseimbangan kelas. Penelitian ini menangani keterbatasan tersebut dengan menerapkan pipeline machine learning yang terintegrasi dan menggunakan metrik evaluasi yang lebih mencerminkan.

3. Random Forest untuk Klasifikasi Akademik.

Random Forest adalah algoritma pembelajaran *ensemble* yang terkenal ampuh dalam mengatasi hubungan non-linear serta mengurangi kemungkinan *overfitting*. (T. Nguyen, 2021) menyatakan bahwa Random Forest memiliki kelebihan dibandingkan dengan model pohon keputusan tunggal dalam hal memprediksi kelulusan mahasiswa. Meskipun demikian, banyak penelitian sebelumnya memanfaatkan Random Forest hanya sebagai alat untuk klasifikasi tanpa mengeksplorasi secara mendalam potensi interpretasi yang dimilikinya.

Sebagai hasilnya, peran model dalam memahami faktor-faktor akademik yang berpengaruh terhadap kelulusan mahasiswa menjadi terbatas. Studi ini menggunakan kemampuan *feature importance* dari Random Forest tidak hanya untuk meningkatkan kinerja prediksi, tetapi juga untuk membantu analisis faktor utama sebagai dasar pembuatan kebijakan akademik.

4. Pipeline Terintegrasi dan Hyperparameter Tuning

Sistem pembelajaran mesin berbasis pipeline memungkinkan semua langkah dalam pengolahan data, mulai dari persiapan data hingga pembuatan model, dilaksanakan dengan cara yang seragam dan dapat diulang. (Z. Papamitsiou, 2022) menegaskan bahwa metode ini ampuh dalam meminimalkan kemungkinan terjadinya *data leakage*. Sementara itu, (S. M. H. Rizvi, 2022) serta (Mahmoud, Hernández, 2022) mengungkapkan bahwa penyesuaian hyperparameter dengan menggunakan GridSearchCV dapat secara signifikan meningkatkan kinerja model *ensemble*. Namun, penelitian sebelumnya sering kali melakukan optimasi hyperparameter hanya sebagian atau tanpa mengintegrasikannya sepenuhnya dalam alur kerja, sehingga menyulitkan proses pengulangan dan penerapan model dalam sistem yang sebenarnya. Studi ini menyelesaikan kekurangan tersebut dengan mengintegrasikan seluruh tahapan dalam satu alur kerja dari awal hingga akhir.

5. Penelitian Terkait

Kajian yang dilakukan oleh (Al-Ashhab & Al-Zubi, 2023), (Ding & Wang, 2023), (Kumar & Pal, 2024), (M. Imung, I. Idawati, 2025) menunjukkan bahwa Indeks Prestasi Kumulatif (IPK), jumlah Semester Kredit Sistem (SKS) yang diselesaikan, serta konsistensi dalam beban studi adalah faktor utama yang memprediksi kelulusan tepat waktu. Walaupun temuan-temuan ini memperkuat landasan teoretis penelitian ini, mayoritas kajian tersebut belum secara jelas membahas kesiapan model untuk diterapkan sebagai sistem peringatan dini, terutama dalam konteks pendidikan tinggi di Indonesia. Di samping itu, perhatian penelitian masih lebih banyak tertuju pada peningkatan akurasi tanpa memperhatikan faktor interpretabilitas dan konsistensi dalam proses pra-pemrosesan. Oleh karena itu, studi ini bertujuan untuk mengaddress kekurangan tersebut dengan menciptakan model prediktif yang menggunakan pipeline terintegrasi yang seimbang antara kinerja, kemampuan untuk dipahami, dan kesiapan untuk diimplementasikan sebagai sistem peringatan dini.

Berdasarkan analisis terhadap penelitian sebelumnya, masih terdapat beberapa kekurangan dalam penelitian (*research gap*) dalam meramalkan kelulusan mahasiswa. Pertama, banyak penelitian yang belum menerapkan sistem machine learning yang terintegrasi secara menyeluruh, terutama dalam memastikan konsistensi preprocessing serta menghindari terjadinya kebocoran data. Kedua, penggunaan data akademik mahasiswa secara terus-menerus selama seluruh periode studi dengan jumlah SKS yang terstandar masih belum optimal, padahal data ini sangat penting untuk memahami dinamika kinerja akademik mahasiswa. Ketiga, pengoptimalan hyperparameter sering kali belum dilaksanakan secara terstruktur dalam satu sistem yang terpadu, sehingga kemungkinan peningkatan kinerja model belum sepenuhnya dimanfaatkan. Selain itu, masih terbatas penelitian yang mengarah pada model prediktif dalam konteks penerapan operasional sebagai sistem peringatan dini, serta menekankan pada pemahaman model untuk menentukan faktor akademik utama yang memengaruhi keterlambatan dalam menyelesaikan studi.

Berdasarkan gap tersebut, penelitian ini bertujuan untuk menciptakan model prediktif untuk kelulusan mahasiswa menggunakan machine learning melalui suatu pipeline yang terintegrasi secara menyeluruh, yang mencakup langkah-langkah pengolahan data, pemodelan, dan evaluasi kinerja dengan konsisten. Model ini dirancang dengan memanfaatkan secara menyeluruh data akademik dari berbagai semester dan dioptimalkan melalui penyetelan hyperparameter yang sistematis menggunakan *GridSearchCV*, sehingga menghasilkan model yang akurat, mudah dipahami, dan siap untuk diterapkan sebagai sistem peringatan dini guna mendukung pengambilan keputusan akademik yang berbasis data.

Tidak seperti penelitian sebelumnya yang biasanya memisahkan tahap preprocessing dan pemodelan, penelitian ini menggabungkan seluruh proses ke dalam satu alur kerja yang menyeluruh. Pendekatan ini secara

substansial mengurangi kemungkinan kebocoran data yang sering diabaikan dalam penelitian tentang prediksi kelulusan mahasiswa. Di samping itu, banyak penelitian sebelumnya yang hanya menyoroti data dari semester awal, sedangkan penelitian ini menggunakan data akademik yang mencakup beberapa semester secara longitudinal (8 semester, 146 SKS), sehingga dapat menangkap perubahan dalam kinerja akademik mahasiswa dengan lebih menyeluruh.

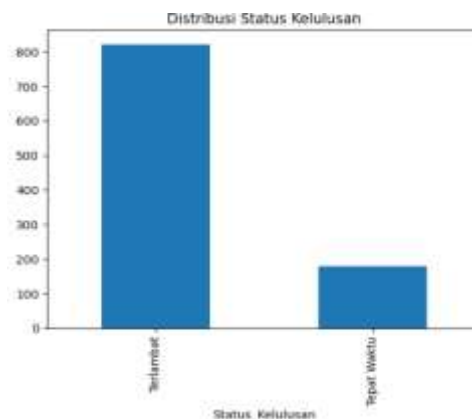
METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimental komputasional, yaitu menerapkan dan mengevaluasi algoritma *machine learning* untuk membangun model prediktif kelulusan mahasiswa. Pendekatan ini dipilih untuk mengidentifikasi pola dan hubungan antar variabel akademik yang memengaruhi ketepatan waktu kelulusan mahasiswa secara berbasis data. Dataset yang digunakan merupakan data akademik mahasiswa Program Studi Sistem Informasi, Institut Sains dan Teknologi Al-Kamal yang mencakup riwayat studi selama delapan semester dengan total beban studi 146 SKS. Data terdiri dari variabel numerik dan kategorikal, antara lain: Indeks Prestasi Kumulatif (IPK) per semester, Jumlah SKS lulus dan SKS diambil per semester, Rata-rata SKS lulus per semester, Beban studi tinggi (≥ 20 SKS), Karakteristik demografis mahasiswa. Variabel target adalah status kelulusan, yang diklasifikasikan menjadi tepat waktu dan terlambat. Adapun tahap penelitiannya meliputi : (1) Tahap Pengumpulan dan Pemahaman Data; (2) Pra-Pemrosesan Data; (3) Pembagian Data; (4) Pembangunan Model; (5) Optimasi Hyperparameter; (6) Evaluasi Model; (7) Analisis dan Interpretasi Fitur.

Agar penelitian dapat direproduksi dengan baik, dataset dibagi dengan proporsi 80% untuk data pelatihan dan 20% untuk data pengujian, menggunakan pendekatan stratified sampling untuk mempertahankan proporsi kelas yang ada. Seluruh tahapan persiapan data dan pelatihan model telah digabungkan dalam alur kerja machine learning dengan menggunakan Scikit-Learn. Penanganan ketidakseimbangan kelas dilakukan dengan memilih metrik evaluasi yang kuat, khususnya ROC Curve dan AUC, yang lebih menggambarkan kondisi dibandingkan akurasi saat data tidak seimbang. Validasi model dilakukan dengan menggunakan metode 5-fold cross-validation guna meningkatkan keandalan perkiraan kinerja model.

HASIL

Tahap awal penelitian ini adalah pengumpulan dan pemahaman data mahasiswa yang digunakan untuk membangun model prediktif kelulusan. Dataset yang digunakan adalah data yang berisi data akademik mahasiswa mulai dari semester awal hingga semester akhir, termasuk Indeks Prestasi Kumulatif (IPK), jumlah SKS yang diambil dan lulus pada setiap semester, serta atribut demografis dan latar belakang pendidikan. Target prediksi pada penelitian ini adalah Status Kelulusan, yang diklasifikasikan menjadi dua kelas, yaitu *Tepat Waktu* dan *Terlambat*. Distribusi data menunjukkan bahwa jumlah mahasiswa yang lulus tepat waktu lebih dominan dibandingkan mahasiswa yang terlambat lulus, sehingga dataset bersifat imbalanced.

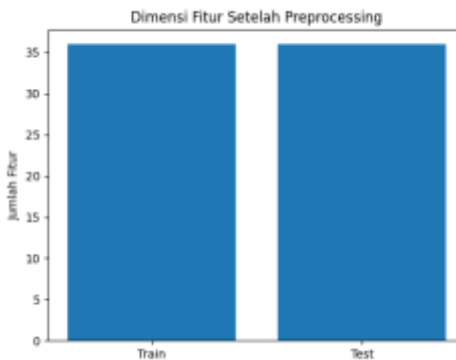


Gambar 1. Distribusi Status Kelulusan

Tahap preprocessing dilakukan untuk memastikan bahwa data siap digunakan oleh algoritma machine learning. Proses ini dilakukan menggunakan *ColumnTransformer* agar preprocessing fitur numerik dan kategorikal dapat dilakukan secara terpisah namun terintegrasi dalam satu pipeline.

Langkah preprocessing yang dilakukan meliputi:

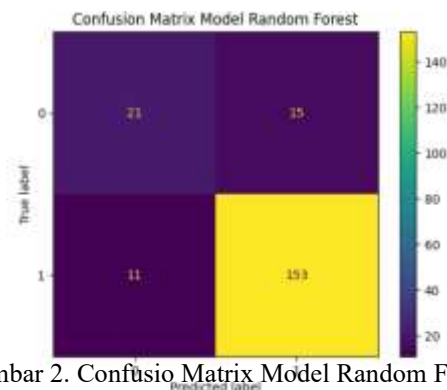
1. Penghapusan kolom yang tidak relevan seperti ID dan Nama
2. Normalisasi fitur numerik menggunakan *StandardScaler* untuk memastikan setiap fitur memiliki skala yang sebanding.
3. Transformasi fitur kategorikal menggunakan *OneHotEncoder* dengan parameter *handle_unknown='ignore'* untuk menghindari error pada data baru.



Gambar 3. Hasil Preprocessing Data Menggunakan *ColumnTransformer*

Gambar di atas menunjukkan jumlah fitur (dimensi data) yang digunakan pada data latih (*Train*) dan data uji (*Test*) setelah melalui tahap preprocessing data. Sumbu horizontal (X) merepresentasikan jenis data, yaitu *Train* dan *Test*, sedangkan sumbu vertikal (Y) menunjukkan jumlah fitur yang digunakan dalam proses pemodelan. Kesamaan jumlah fitur antara data latih dan data uji menunjukkan bahwa proses preprocessing telah dilakukan secara konsisten dan seragam pada seluruh dataset.

Model klasifikasi yang digunakan dalam penelitian ini adalah *Random Forest Classifier*, yang dipilih karena kemampuannya dalam menangani data berdimensi tinggi, fitur campuran, serta ketahanannya terhadap overfitting. Model *Random Forest* dibangun dalam sebuah pipeline terintegrasi yang menggabungkan preprocessing dan proses pelatihan model. Pendekatan pipeline ini memastikan bahwa seluruh tahapan pemodelan dilakukan secara konsisten dan dapat direplikasi.



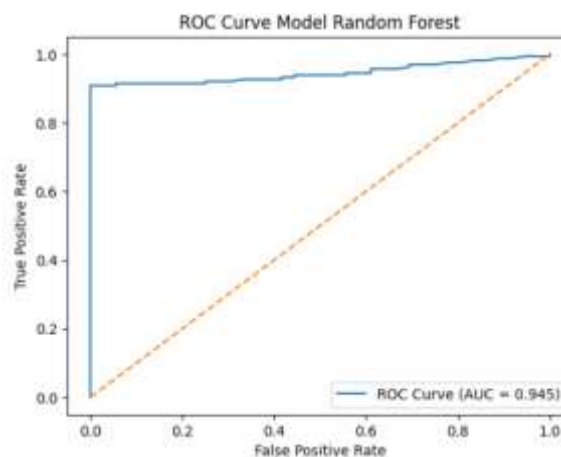
Gambar 2. Confusio Matrix Model Random Forest

Untuk meningkatkan performa model, dilakukan proses hyperparameter tuning menggunakan *GridSearchCV* dengan validasi silang 5-fold. Beberapa parameter utama yang dioptimasi meliputi:

- Jumlah pohon ($n_estimators$),
- Kedalaman maksimum pohon (max_depth),
- Minimum sampel untuk pemisahan node ($min_samples_split$),
- Minimum sampel pada daun ($min_samples_leaf$).

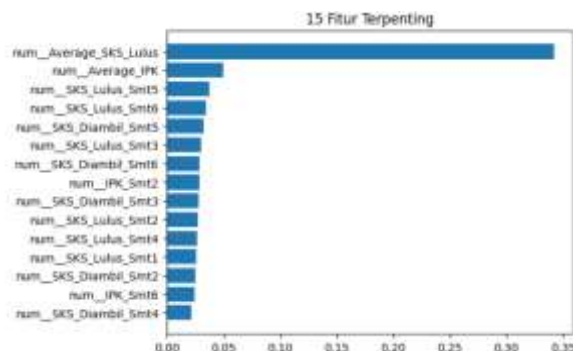
Hasil tuning menunjukkan bahwa kombinasi hyperparameter optimal mampu meningkatkan stabilitas dan akurasi model secara signifikan dibandingkan dengan konfigurasi default. Model terbaik selanjutnya digunakan untuk tahap evaluasi dan analisis fitur.

Pada evaluasi model Confusion matrix menunjukkan bahwa model mampu mengklasifikasikan sebagian besar mahasiswa terlambat dengan benar (*true positive* tinggi), kesalahan klasifikasi pada kelas *Tepat Waktu* masih terjadi, namun dalam jumlah yang lebih kecil. Secara keseluruhan, model mencapai akurasi $\pm 81\%$, yang menunjukkan performa prediktif yang baik.



Gambar 3. ROC Curve Model Random Forest.

Analisis *feature importance* menunjukkan bahwa fitur yang paling dominan meliputi: Average_SKS_Lulus, Average_IPK, Jumlah SKS lulus pada semester menengah dan akhir, IPK pada semester awal. Fitur akademik jauh lebih dominan dibandingkan fitur demografis.



Gambar 3. Feature Importance

PEMBAHASAN

Penggunaan ColumnTransformer dalam alur kerja machine learning memungkinkan pemrosesan fitur numerik dan kategorikal secara seragam dalam satu langkah, sehingga mengurangi kemungkinan kebocoran data yang sering muncul ketika preprocess dilakukan terpisah dari pelatihan model. Pendekatan ini menghasilkan suatu model Random Forest dengan kinerja yang memuaskan, yang ditandai oleh akurasi mencapai 81% dan nilai AUC sebesar 0,945. Ini menunjukkan adanya kemampuan diskriminatif yang sangat baik dalam membedakan mahasiswa yang lulus tepat waktu dengan yang terlambat, meskipun dalam kondisi data yang tidak seimbang. Walaupun begitu, hasil tersebut juga mengindikasikan bahwa kemungkinan kesalahan klasifikasi pada kelompok minoritas masih perlu diperhatikan, terutama bila model diterapkan dalam situasi operasional sebagai sistem peringatan dini.

Analisis mengenai pentingnya fitur menunjukkan bahwa variabel akademik, khususnya rata-rata Indeks Prestasi Kumulatif (IPK) dan jumlah Satuan Kredit Semester (SKS) yang diselesaikan setiap semester, adalah faktor utama yang paling berpengaruh terhadap ketepatan waktu kelulusan mahasiswa. Sementara itu, pengaruh variabel demografis terbilang terbatas. Hasil penelitian ini konsisten dengan studi sebelumnya yang menekankan kegunaan metode ensemble dalam mengidentifikasi risiko akademik mahasiswa. Namun, keunggulan utama dari penelitian ini terdapat pada penggunaan sistem terpadu yang meningkatkan kemampuan untuk mereproduksi dan kesiapan dalam menerapkan model. Oleh karena itu, model yang telah dibuat memiliki potensi besar untuk diterapkan sebagai Sistem Peringatan Dini, meskipun perlu adanya pengembangan lebih lanjut, terutama dengan memasukkan faktor non-akademik dan strategi untuk mengatasi ketidakseimbangan data demi meningkatkan akurasi intervensi akademik.

KESIMPULAN

Penelitian ini mengindikasikan bahwa metode pipeline machine learning yang dikombinasikan dengan optimasi hyperparameter dapat menghasilkan model prediktif yang akurat dan stabil untuk kelulusan mahasiswa. Model Random Forest yang dikembangkan tidak hanya menunjukkan hasil klasifikasi yang memuaskan, tetapi juga dapat mengidentifikasi faktor akademik penting yang berpengaruh terhadap keterlambatan dalam menyelesaikan studi.

Walaupun begitu, penelitian ini mempunyai keterbatasan karena hanya menggunakan data akademik internal tanpa mempertimbangkan faktor non-akademik serta belum menggunakan teknik penyeimbangan data yang khusus. Oleh karena itu, hasil dari penelitian ini masih perlu diuji lebih lanjut dalam konteks institusi yang berbeda untuk meningkatkan kemampuan generalisasi model.

SARAN

Penelitian berikutnya disarankan untuk memasukkan variabel non-akademik, seperti tingkat kehadiran, aktivitas belajar online, serta kondisi sosial ekonomi mahasiswa, agar dapat meningkatkan akurasi dalam prediksi. Selain itu, penting untuk mengeksplorasi penggunaan teknik penanganan data yang tidak seimbang seperti SMOTE atau pembelajaran yang sensitif terhadap biaya. Pelaksanaan model dalam sistem informasi akademik sebagai sistem peringatan awal yang berbasis dashboard interaktif juga merupakan langkah strategis untuk mendukung proses pengambilan keputusan akademik secara langsung.

REFERENSI

- Akmar Efendi, Iskandar Fitri, G. W. N. (2025). Improving Student Graduation Timeliness Prediction Using SMOTE and Ensemble Learning with Stacking and GridSearchCV Optimization. *Data and Metadata*, 4, no 917, 1–10.
- Al-Ashhab, D., & Al-Zubi, O. (2023). Predicting student graduation status using deep learning and model optimization. *Journal of Educational Computing Research*, 61, no. 1, 87–109.
- B. Al-Bargi, S. A. (2021). Student academic performance prediction using supervised machine learning: A case study. *Education and Information Technologies*, 26, no.4, 4333–4353.
- Ding, Y., & Wang, Q. (2023). Hyperparameter tuning and feature selection for student success prediction. *Neural Computing and Applications*, 35, no. 11, 10951–10967.
- Kimmons, W. Xing, R. (2021). Using learning analytics to predict academic outcomes in STEM courses. *Journal of*

Learning Analytics, 8, no. 3, 43–59.

- Kumar, V., & Pal, S. (2024). Student performance forecasting using AutoML pipeline: A hybrid optimization approach. *International Journal of Machine Learning and Cybernetics*, 15, no. 2, 377–395.
- M. Imung, I. Idawati, and I. L. (2025). lassification of nutritional status in toddlers based on anthropometric data using random forest. *Journal of Computer Science and Informatics Engineering*, 4, no. 4, 324–333.
- Mahmoud, M. A., & Hernández, J. E. (2022). A pipeline architecture for educational outcome prediction using hybrid machine learning methods. *Computers & Education*, 182, 44–63.
- Oyelade, A. O. and O. (2020). Machine learning techniques for predicting academic performance: A review. *Applied Artificial Intelligence*, 34, no.4, 291–314.
- P. Kaur, A. S. (2021). Predictive analytics for student performance using machine learning pipeline. *Information Systems Frontiers*, 23, no. 5, 1085–1100.
- Rizal Bakri, S. A. (2025). Optimizing Machine Learning Models for Graduation on Time Prediction: A Comparative Study with Resampling and Hyperparameter Tuning. *JurnalOnline Informatika*, 10 Number, 270–285.
- S. M. H. Rizvi, A. H. (2022). Early academic risk detection using ensemble methods. *Expert Systems with Applications*, 186, 201–211.
- Shoaib, M., & Taie, M. F. (2020). Academic performance prediction: A systematic review and future directions. *IEEE Access*, 8, 105417–105442.
- T. Nguyen, T. P. (2021). Automated feature engineering and model selection for predicting student graduation. *Applied Soft Computing*, 101, 101–1010.
- W. Xing, D. Gašević, and B. A. (2020). Investigating predictive modeling and assessment for early warning in higher education. *IEEE Transactions on Learning Technologies*, 13, no. 2, 226–239.
- Z. Papamitsiou, A. A. E. (2022). Review and classification of predictive modeling approaches in educational data mining. *IEEE Transactions on Emerging Topics in Computing*, 10, no.3, 1127–1141.