

PREDIKSI KELULUSAN TEPAT WAKTU MAHASISWA MENGUNAKAN RANDOM FOREST PADA DATA TIDAK SEIMBANG

Niko Aprilyanto¹, Tri Listyorini², Endang Supriyati³

^{1, 2, 3} Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muria Kudus, Kudus,
Indonesia

e-mail: ¹nikoapil1010@gmail.com, ²trilistyorini@umk.ac.id,
³endang.supriyati@umk.ac.id

ABSTRAK

Tingkat ketepatan waktu kelulusan mahasiswa menjadi salah satu indikator dalam mengukur kualitas pendidikan tinggi. Penelitian ini menggunakan algoritma Random Forest untuk melakukan prediksi terhadap status kelulusan mahasiswa tepat waktu berdasarkan enam atribut akademik semester 1-4 (IPS Semester 1, IPS Semester 2, IPS Semester 3, IPS Semester 4, IPK Semester 4, dan SKS tempuh), dengan *Naïve Bayes* sebagai pembandingan. Data bersumber dari 1.020 mahasiswa Program Studi Teknik Informatika Angkatan 2018-2021. Setelah seleksi data, diperoleh 418 mahasiswa berstatus lulus dengan proporsi 321 mahasiswa (76,8%) tepat waktu dan 97 mahasiswa (23,2%) tidak tepat waktu. Karena data bersifat tidak seimbang, diterapkan tiga skema penanganan yaitu default, *class weighting*, dan SMOTE (Synthetic Minority Oversampling Technique). Evaluasi dilakukan melalui single *train-test split* dan 5-fold stratified *cross-validation* menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*. Hasil pengujian menunjukkan *Random Forest* dengan SMOTE mencapai *ROC-AUC* 0,743, *F1-score* kelas tepat waktu rata-rata 0,834, dan *recall* kelas tidak tepat waktu $50,6\% \pm 0,045$, lebih stabil dibandingkan *Naïve Bayes* (*F1-score* 0,859, *recall* $44,5\% \pm 0,112$). Atribut IPS Semester 3 teridentifikasi sebagai prediktor paling berpengaruh (23,3%). Hasil penelitian diimplementasikan dalam purwarupa aplikasi berbasis web menggunakan Streamlit sebagai alat bantu deteksi dini.

Kata kunci: Random Forest; *Naïve Bayes*; Kelulusan Tepat Waktu; Class Imbalance; SMOTE

ABSTRACT (11pt Bold Italic)

On-time graduation is an important indicator of higher education quality. This study applies the Random Forest algorithm to predict students' graduation status based on six academic attributes from semester 1-4 (GPA semester 1, GPA semester 2, GPA semester 3, GPA semester 4, cumulative GPA, and completed credits), with Naïve Bayes as a comparison. The data were drawn from 1,020 students of the Informatics Engineering Study Program (2018-2021 cohorts). After data selection, 418 graduates were obtained, comprising 321 students (76.8%) who graduated on time and 97 students (23.2%) who did not. Since the data is imbalanced, three handling schemes were applied: default, class weighting, and SMOTE (Synthetic Minority Oversampling Technique). Evaluation was conducted using a single train-test split and 5-fold stratified cross-validation with accuracy, precision, recall, F1-score, and ROC-AUC. Results show that Random Forest with SMOTE achieved a ROC-AUC of 0.743, an average F1-score of 0.834 for the on-time class, and $50.6\% \pm 0.045$ recall for the not on-time class, more stable than Naïve Bayes (F1-score 0.859, recall $44.5\% \pm 0.112$). The Semester 3 GPA attribute was identified

as the most influential predictor (23.3%). The results were implemented as a web-based application prototype using Streamlit as an early-detection.

Keywords: Random Forest; Naïve Bayes; On-Time Graduation; Class Imbalance; SMOTE

1. PENDAHULUAN

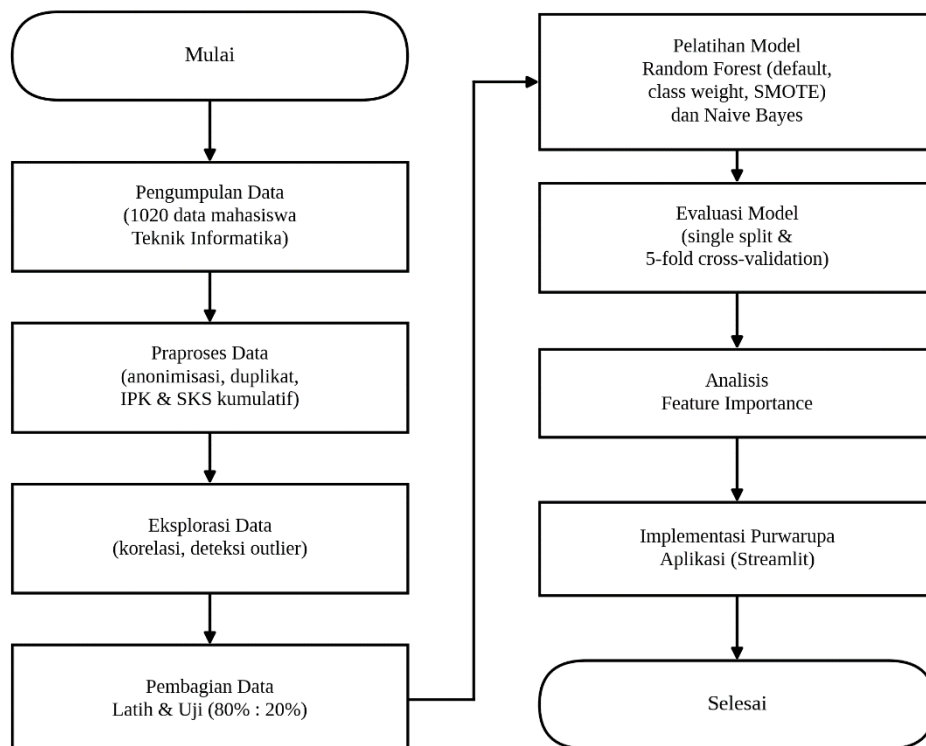
Tingkat kelulusan tepat waktu merupakan salah satu indikator penting kualitas penyelenggaraan pendidikan tinggi, yang berkaitan efisiensi biaya operasional, alokasi sumber daya akademik, dan kesiapan mahasiswa memasuki dunia kerja. Berdasarkan data akademik alumni Program Studi Teknik Informatika Angkatan 2018-2021, dari 418 mahasiswa yang telah menyelesaikan studi, sebanyak 97 mahasiswa (23,2%) tercatat lulus melewati masa tempuh kurikulum normal (8 semester)[1]. Kondisi ini menunjukkan perlunya pendekatan yang mampu mengidentifikasi mahasiswa berisiko terlambat lulus secara dini agar intervensi akademik dapat dilakukan tepat sasaran.

Perkembangan *data mining* membuka peluang pemanfaatan data akademik untuk membangun model prediktif *Random Forest*. *Random Forest* merupakan algoritma ensemble yang menggunakan banyak pohon keputusan dan banyak diterapkan untuk memprediksi kelulusan mahasiswa dengan performa akurasi, presisi, dan *recall* yang baik pada berbagai konteks institusi[2], [3], [4], [5]. Namun, sebagian besar penelitian tersebut belum secara eksplisit menangani kondisi *class imbalance* yang lazim ditemukan pada data kelulusan mahasiswa, yaitu jumlah mahasiswa yang lulus tepat waktu jauh lebih banyak dibandingkan yang terlambat sehingga akurasi keseluruhan berpotensi menyesatkan karena model dapat memperoleh nilai akurasi tinggi hanya dengan cenderung memprediksi seluruh mahasiswa sebagai lulus tepat waktu. Penelitian yang secara khusus menangani ketidakseimbangan kelas pada permasalahan serupa menunjukkan bahwa penerapan teknik penyeimbangan seperti SMOTE mampu meningkatkan performa model secara signifikan, termasuk pada domain klasifikasi medis dengan karakteristik data yang juga timpang [6], [7]. Kombinasi SMOTE dengan algoritma boosting misalnya, terbukti mampu meningkatkan *recall* kelas minoritas dari 50,0% menjadi 95,0% pada klasifikasi penyakit menular berbasis gejala klinis[8] . Tidak hanya itu, sebagian besar penelitian terdahulu masih berfokus pada evaluasi performa model tunggal tanpa perbandingan sistematis dengan algoritma lain seperti *Naïve Bayes* yang telah banyak digunakan sebagai *baseline* pada berbagai permasalahan klasifikasi karena kesederhanaan dan performanya yang kompetitif [9], [10], [11].

Merujuk pada pemaparan sebelumnya, penelitian ini bertujuan menerapkan algoritma *Random Forest* untuk memprediksi status kelulusan mahasiswa menggunakan data akademik semester 1-4, dengan *Naïve Bayes* sebagai algoritma pembanding. Penelitian ini secara eksplisit membahas penanganan *class imbalance* melalui skema *class weighting* dan SMOTE, dilengkapi analisis *feature importance* untuk mengidentifikasi atribut akademik yang paling berpengaruh terhadap ketepatan waktu kelulusan mahasiswa. Sebagai kontribusi praktis, hasil penelitian diwujudkan dalam purwarupa aplikasi berbasis web menggunakan Streamlit sebagai alat bantu deteksi dini.

2. METODE PENELITIAN

Penelitian ini menerapkan pendekatan kuantitatif dengan menggunakan metode klasifikasi berbasis *supervised learning*. Tahapan penelitian meliputi pengumpulan data, praproses, eksplorasi data, pembagian data latih-uji, pelatihan model dengan beberapa skema penanganan *class imbalance*, evaluasi model (single split dan 5-fold cross validation), analisis *feature importance*, dan implementasi purwarupa sebagaimana disajikan pada Gambar 1.



Gambar 1. Alur Penelitian

Data Penelitian

Data penelitian bersumber dari sistem Informasi akademik Program Studi Teknik Informatika, mencakup populasi awal 1.020 mahasiswa Angkatan 2018-2021 dengan berbagai status (aktif, cuti, keluar, dan lulus). Penelitian ini secara khusus menggunakan subset 418 mahasiswa yang telah lulus, dengan rincian 321 (76,8%) berstatus tepat waktu dan 97 (23,2%) tidak tepat waktu. Atribut yang digunakan disajikan pada Tabel 1, dibatasi hanya pada data akademik hingga semester 4 agar model yang dihasilkan relevan digunakan sebagai sistem deteksi dini, yaitu saat mahasiswa masih memiliki cukup waktu menerima intervensi akademik. Periode data yang tidak seragam antarangkatan turut memengaruhi proporsi ini dan dibahas pada bagian keterbatasan penelitian.

Tabel 1. Deskripsi Atribut Penelitian

Atribut	Tipe Data	Keterangan
IPS Semester 1	Numerik	Indeks Prestasi semester 1, skala 0.00-4.00
IPS Semester 2	Numerik	Indeks Prestasi Semester 2, skala 0.00-4.00
IPS Semester 3	Numerik	Indeks Prestasi Semester 3, skala 0.00-4.00
IPS Semester 4	Numerik	Indeks Prestasi Semester 4, skala 0.00-4.00
IPK Semester 4	Numerik	IPK Kumulatif semester 1-4
SKS Tempuh Semester 4	Numerik	Total SKS kumulatif semester 1-4
Status Kelulusan	Kategorikal	Target “Ya” (tepat waktu) dan “tidak” (tidak tepat waktu)

Praproses Data

Praproses meliputi anonimisasi data (NIM menggantikan nama), penanganan 910 baris data duplikat (kombinasi NIM dan semester sama, nilai IPS berbeda) melalui perataan, serta perhitungan ulang IPK dan SKS kumulatif secara manual ($IPK = \sum(IPS \times SKS) / \sum SKS$) untuk menghindari kebocoran data, karena kolom IPK/SKS pada data mentah bersifat statis (nilai akhir kelulusan). Duplikasi ini tidak tersebar acak, seluruh 528 baris duplikat pada Angkatan 2019 terjadi pada semester 3, dan seluruh 382 baris duplikat pada Angkatan 2020 terjadi pada semester 1, mengindikasikan kemungkinan dua periode pencatatan berbeda pada sistem sumber, bukan sekedar kesalahan input acak. Perataan tetap digunakan sebagai pendekatan pragmatis, dan hal ini dicatat sebagai keterbatasan data. Perataan tersebut juga menghasilkan nilai SKS tempuh non-integer pada 50 dari 418 mahasiswa (mis. 82,5 : 86,5), yang secara konseptual tidak merepresentasikan satuan SKS asli. Nilai ini dipertahankan apa adanya pada model utama untuk menjaga konsistensi dengan skema Perataan yang sama. Pemeriksaan data lebih lanjut juga menemukan satu baris data (Angkatan 2021) dengan total SKS tempuh semester 1-4 sebesar 41 SKS namun berstatus tepat waktu, kombinasi yang secara akademik tidak konsisten dan diduga merupakan kesalahan pencatatan pada sistem sumber. Baris tersebut tetap disertakan pada model utama namun dicatat sebagai keterbatasan kualitas data. Data dibagi menjadi 80% data latih (334 data) dan 20% data uji (84 data) secara stratified, menjaga proporsi kelas pada kedua subset.

Algoritma yang Digunakan

Random Forest adalah algoritma *ensemble learning* yang membangun sejumlah pohon keputusan secara acak melalui *bootstrap sampling* dan seleksi fitur acak pada tiap split, dengan hasil klasifikasi ditentukan melalui *majority voting*[12]. Sebagai pembanding, digunakan Gaussian Naïve Bayes, algoritma probabilistik berbasis Teorema Bayes dengan asumsi independensi antar fitur, yang mengasumsikan distribusi normal pada tiap atribut numerik[13].

Penangan Class Imbalance

Tiga skema diuji pada *Random Forest* ($n_estimators=100$):

1. **Default**.
2. **class_weight='balanced'**, yang memberikan bobot lebih besar pada kelas minoritas berdasarkan frekuensinya.
3. **SMOTE**, yang menghasilkan data sintesis kelas minoritas melalui interpolasi terhadap *k-nearest neighbors* dalam ruang fitur[6] .

SMOTE diterapkan hanya pada data latih untuk mencegah kebocoran data. *Naïve bayes* turut dilatih menggunakan skema SMOTE yang untuk memastikan perbandingan yang adil.

Evaluasi Model

Evaluasi dilakukan melalui single *train-test split* dan 5-fold stratified *cross-validation*, dengan SMOTE diterapkan per-fold melalui *pipeline* untuk mencegah kebocoran data. Metrik yang digunakan meliputi *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC* (Area Under the Receiver Operating Characteristic Curve), dengan TP, TN, FP, FN berturut-turut menyatakan *true positive*, *true negative*, *false positive*, dan *false negative*:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$Precision = TP / (TP + FP) \quad (2)$$

$$Recall = TP / (TP + FN) \quad (3)$$

$$F1-score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (4)$$

Precision, *recall*, dan *F1-score* dihitung menggunakan fungsi bawaan scikit-learn dengan kelas positif mengikuti pengkodean label pada data (1 = tepat waktu), sehingga *F1-score* yang dilaporkan pada Tabel 3 dan Tabel 4 merupakan *F1-score* kelas tepat waktu (kelas mayoritas). Mengingat kondisi *class imbalance*, *recall* dan presisi kelas minoritas (tidak tepat waktu) dihitung dan dilaporkan secara terpisah, karena metrik agregat pada kelas mayoritas berpotensi menyesatkan gambaran kemampuan model mendeteksi mahasiswa berisiko[14]. *ROC-AUC* turut dilaporkan sebagai metrik pelengkap yang tidak bergantung pada ambang klasifikasi maupun pilihan kelas positif. Model dengan skema terbaik selanjutnya dilatih ulang pada seluruh data dan diimplementasikan ke purwarupa aplikasi berbasis Streamlit.

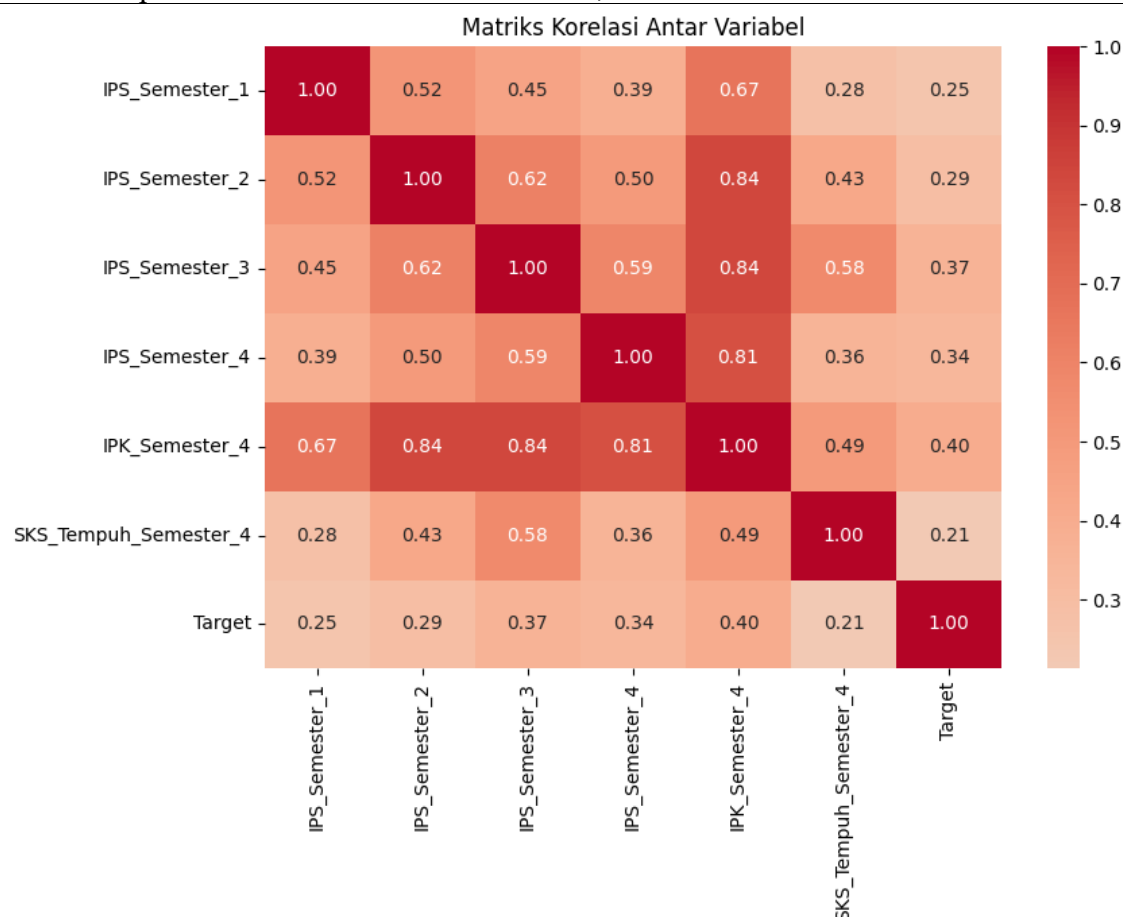
3. HASIL DAN PEMBAHASAN

Eksplorasi Data

Analisis korelasi (Tabel 2) menunjukkan seluruh atribut berkorelasi positif dengan target, tertinggi pada IPK Semester 4 (0,40) dan terendah pada SKS Tempuh (0,21) yang berarti hubungan moderat yang realistis untuk data akademik dan tidak ada korelasi yang mencapai kuat ($>0,5$). Hal ini mengindikasikan kelulusan tepat waktu dipengaruhi interaksi banyak faktor yang tidak sepenuhnya tertangkap melalui hubungan linier sederhana.

Tabel 2. Koefisien Korelasi Atribut terhadap Target

Atribut	Koefisien Korelasi
IPK Semester 4	0,40
IPS Semester 3	0,37
IPS Semester 4	0,34
IPS Semester 2	0,29
IPS Semester 1	0,25
SKS Tempuh Semester 4	0,21



Gambar 2. Matriks Korelasi

Ditemukan pula korelasi yang bervariasi antara IPK Semester 4 dan IPS per semester (0,67-0,84), dengan korelasi terendah IPS Semester 1 (0,67) dan tertinggi pada IPS Semester 3 (0,84). Korelasi yang cukup tinggi pada semester-semester akhir secara matematis wajar karena IPK merupakan turunan langsung dari rata-rata IPS. Kondisi ini relatif tidak memengaruhi *Random Forest*, namun berpotensi memengaruhi asumsi independensi atribut pada Naïve Bayes. Pada tahap pemeriksaan data, ditemukan tiga mahasiswa dengan IPS pada salah satu semester sangat rendah (0,00-0,56). SKS tempuh ketiganya bervariasi (69, 79, dan 82 SKS, dibandingkan rata-rata keseluruhan $85,9 \pm 3,9$ SKS), salah satunya (69 SKS) berada signifikan di bawah rata-rata dan mengindikasikan kemungkinan kegagalan akademik pada sejumlah matakuliah semester tersebut. Sementara dua lainnya relatif mendekati rata-rata sehingga penyebab nilai rendahnya tidak dapat dipastikan hanya dari data kuantitatif. Ketiganya tetap berlabel “tidak tepat

waktu”, konsisten dengan pola umum penelitian sehingga data tetap dipertahankan pada model utama.

Perbandingan Skema Penanganan Class Imbalance

Random Forest default menghasilkan akurasi 79,8% namun *recall* kelas “tidak tepat waktu” hanya 32% (6/19), menunjukkan bias terhadap kelas mayoritas meskipun *F1-score* kelas tepat waktu tampak tinggi (0,878). *Class weighting* tidak memberi perbaikan berarti pada *recall* minoritas (tetap 32%). SMOTE meningkatkan *recall* kelas minoritas menjadi 47%, meski akurasi turun menjadi 70,2% dan *F1-score* kelas tepat waktu ikut menurun menjadi 0,800. Ringkasan disajikan pada Tabel 3.

Tabel 3. Hasil Evaluasi (Data Uji)

Skema	Akurasi	F1-score (Tepat Waktu)	Recall Minoritas	Presisi Minoritas
Default	79,8%	0,878	32% (6/19)	60%
Class Weight Balanced	80,9%	0,886	32% (6/19)	67%
SMOTE	70,2%	0,800	47% (9/19)	38%

Presisi kelas minoritas pada skema SMOTE (38%) lebih rendah dibandingkan skema default (60%) dan balance (67%), menegaskan *trade-off* antara *recall* dan presisi. Semakin agresif model mendeteksi mahasiswa berisiko, semakin tinggi pula proporsi “alarm palsu” yang dihasilkan. *Naïve Bayes* dengan SMOTE pada skema single split menghasilkan presisi kelas minoritas sebesar 44% berada di antara default dan SMOTE pada *Random Forest*.

Validasi Cross-Validation dan Perbandingan dengan Naïve Bayes

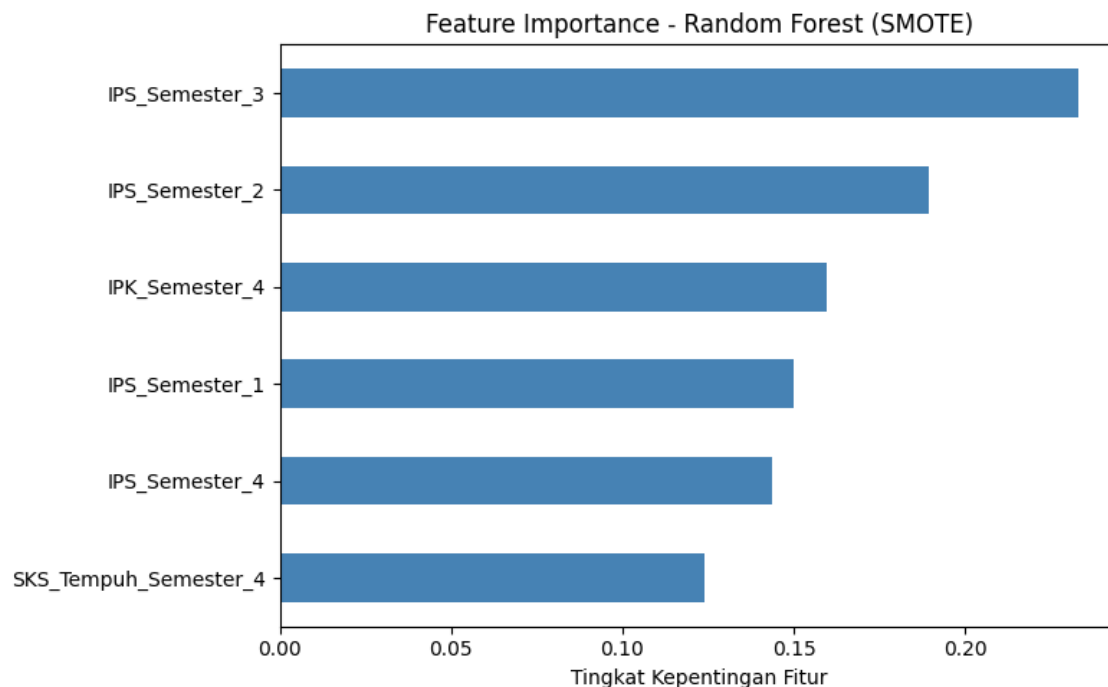
Validasi 5-fold *cross-validation* (Tabel 4) menunjukkan *Random Forest*+SMOTE menghasilkan *recall* kelas minoritas yang lebih tinggi dan lebih stabil ($50,6\% \pm 0,045$) dibandingkan *Naïve Bayes*+SMOTE ($44,5\% \pm 0,112$), meski *F1-score* kelas tepat waktu rata-rata *Naïve Bayes* sedikit lebih tinggi (0,859 berbanding 0,834). Kedua model juga menghasilkan *ROC-AUC* rata-rata yang hampir identik (*Random Forest* $0,743 \pm 0,051$, sedangkan *Naïve Bayes* $0,743 \pm 0,038$), menunjukkan kapasitas diskriminatif keseluruhan yang setara meskipun berbeda dalam *trade-off recall-presisi* pada ambang default. Standar deviasi *recall Naïve Bayes* yang jauh lebih besar mengindikasikan performa yang kurang dapat diandalkan antar-fold pada data berskala kecil, sehingga *Random Forest*+SMOTE dipilih sebagai model utama berdasarkan kestabilannya dalam mendeteksi kelas berisiko.

Tabel 4. Hasil 5-Fold *Cross-Validation*

Model	F1-score Rata-rata(Tepat Waktu)	Recall Kelas Minoritas	ROC-AUC
Random Forest+SMOTE	$0,834 \pm 0,013$	$0,506 \pm 0,045$	$0,743 \pm 0,051$
Naïve Bayes+SMOTE	$0,859 \pm 0,028$	$0,445 \pm 0,112$	$0,743 \pm 0,038$

Feature Importance

Analisis *feature importance* (Gambar 3) menunjukkan IPS Semester 3 sebagai prediktor paling berpengaruh (23,3%), diikuti IPS Semester 2 (19,0%) dan IPS Semester 4 (16,0%). Sedangkan, IPS Semester 3 bukan atribut dengan korelasi tertinggi, menunjukkan kemampuan *Random Forest* menangkap pola non-linier yang tidak teridentifikasi melalui korelasi sederhana. Interpretasi atribut yang berkontribusi terhadap keputusan model semacam ini sejalan dengan penelitian klasifikasi lain yang secara eksplisit mengevaluasi seleksi fitur, seperti *Recursive Feature Elimination*, untuk memperoleh model yang lebih ringkas dan mudah diinterpretasikan tanpa kehilangan performa secara signifikan [15].



Gambar 3. Feature Importance Model Random Forest

Implementasi Aplikasi

Model *Random Forest* dengan SMOTE dilatih ulang menggunakan keseluruhan 418 data agar memanfaatkan seluruh Informasi yang tersedia, kemudian diimplementasikan ke purwarupa aplikasi web menggunakan *framework* Streamlit (Gambar 4). Aplikasi menyediakan formulir input untuk keenam atribut (IPS Semester 1, IPS Semester 2, IPS Semester 3, IPS Semester 4, IPK Semester 4, dan SKS Tempuh Semester 4), dan menampilkan hasil prediksi status kelulusan beserta persentase tingkat keyakinan model untuk masing-masing kelas (Gambar 5). Aplikasi juga dilengkapi panel interaktif yang menampilkan tingkat kepentingan tiap atribut, agar pengguna dapat memahami dasar pertimbangan model. Sebagai purwarupa, aplikasi ini belum terintegrasi dengan sistem Informasi Akademik institusi dan dilengkapi pernyataan penyangkalan (*disclaimer*) bahwa hasil prediksi bersifat estimasi statistik yang tidak menggantikan penilaian akademik resmi.

Masukkan Data Akademik Mahasiswa

IPS Semester 1: 3,20

IPS Semester 2: 3,20

IPS Semester 3: 3,20

IPS Semester 4: 3,20

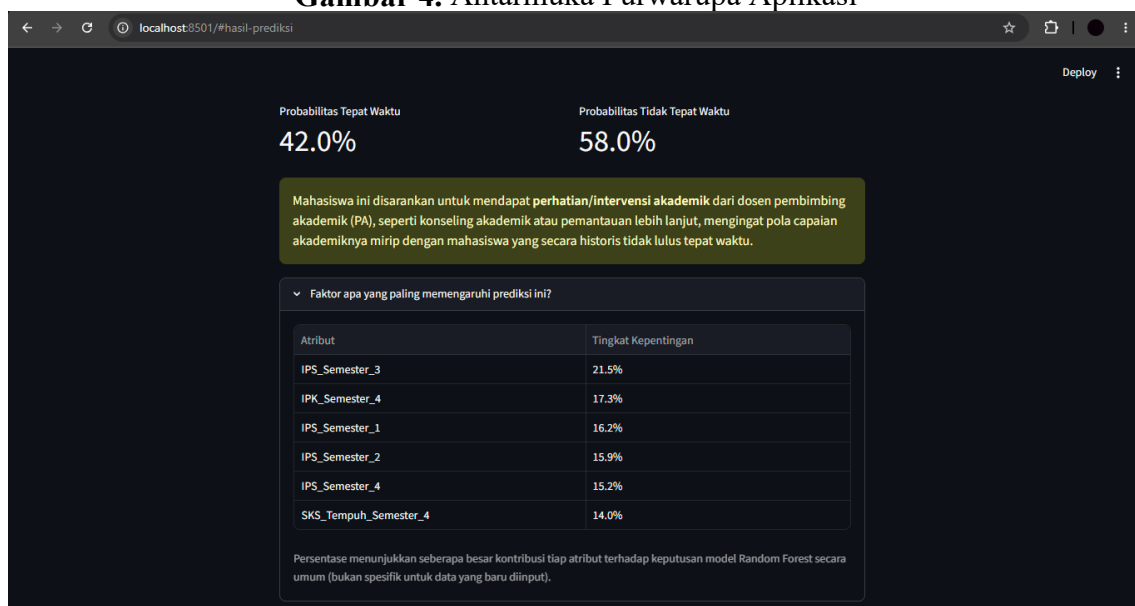
IPK Kumulatif (s.d. Semester 4): 3,20

Total SKS Tempuh (s.d. Semester 4): 85

Prediksi Status Kelulusan

Disclaimer: Prediksi ini bersifat estimasi statistik berdasarkan model machine learning yang dilatih dari data alumni Program Studi Teknik Informatika, dan tidak menggantikan penilaian atau keputusan akademik resmi dari dosen pembimbing akademik maupun pihak program studi.

Gambar 4. Antarmuka Purwarupa Aplikasi



Gambar 5. Tampilan Hasil Prediksi

Keterbatasan Penelitian

Proporsi kelas pada data penelitian (76,8% tepat waktu dan 23,2% tidak tepat waktu) perlu dibaca dengan mempertimbangkan periode pencatatan data. Data kelulusan yang digunakan tercatat hingga semester genap 2024/2025, sehingga keempat angkatan memiliki masa observasi yang berbeda. Seluruh 82 mahasiswa angkatan 2021 tercatat lulus tepat waktu karena mahasiswa dari angkatan tersebut yang berpotensi lulus terlambat belum tercatat lulus pada periode pengambilan data, sementara angkatan 2018 relatif teramat penuh dengan proporsi tidak tepat waktu mencapai 40,3%. Untuk menguji sejauh mana kondisi ini memengaruhi hasil, model *Random Forest*+SMOTE dilatih ulang menggunakan subset data tanpa angkatan 2021 (n=336, proporsi tidak tepat waktu 28,9%)

dan subset angkatan 2018-2019 saja (n=276, proporsi tidak tepat waktu 31,5%). Hasilnya disajikan pada Tabel 5

Tabel 5. Hasil *Random Forest*+SMOTE pada subset data berbeda (5-fold CV)

Subset	n	% Tidak Tepat Waktu	Recall Minoritas	ROC-AUC
Seluruh data (2018-2021)	418	23,2%	0,506 ± 0,045	0,743 ± 0,051
Tanpa Angkatan 2021	336	28,9%	0,556 ± 0,053	0,774 ± 0,047
Angkatan 2018-2019	276	31,5%	0,461 ± 0,116	0,689 ± 0,071

Recall kelas minoritas justru meningkat ketika Angkatan 2021 dikeluarkan (55,6% berbanding 50,6%), menunjukkan bahwa hasil utama pada seluruh 418 data bersifat konservatif dan bukan disebabkan oleh optimisme akibat penyertaan data yang belum teramati penuh. Meski demikian, seluruh 418 data tetap digunakan sebagai analisis utama pada penelitian ini untuk memaksimal ukuran sampel latih, dengan mempertimbangkan bahwa proporsi kelas sebenarnya pada populasi kemungkinan lebih tinggi dari 23,2% yang dilaporkan. Keterbatasan lain mencakup pola duplikasi data yang sistematis pada Angkatan 2019 dan 2020 serta satu baris data dengan kombinasi SKS dan status kelulusan yang tidak konsisten (angkatan 2021).

4. KESIMPULAN

Random Forest dengan penanganan SMOTE terbukti mampu meningkatkan kemampuan deteksi mahasiswa berisiko tidak lulus tepat waktu, dengan *recall* kelas minoritas yang lebih tinggi dan lebih stabil (50,6% ± 0,045) dibandingkan dengan *Naïve bayes* (44,5% ± 0,112) pada validasi 5-fold *cross-validation*, meskipun *F1-score* kelas tepat waktu rata-rata *Naïve Bayes* sedikit lebih tinggi (0,859 berbanding 0,834). Kedua model menghasilkan *ROC-AUC* yang setara (±0,743). IPS Semester 3 teridentifikasi sebagai prediktor paling berpengaruh (23,3%), diikuti IPS Semester 2 dan IPK Semester 4. Penanganan *class imbalance* terbukti krusial dalam penelitian ini, model tanpa penanganan khusus hanya mendeteksi 32% mahasiswa berisiko meski akurasi keseluruhan tinggi (79,8%), sementara *class weighting* tidak memberikan perbaikan berarti. Analisis sensitivitas menunjukkan bahwa proporsi kelas tidak seimbang pada data penelitian ini bersifat konservatif akibat perbedaan periode observasi antarangkatan, sehingga hasil yang dilaporkan cenderung merupakan estimasi batas bawah kemampuan model.

Penelitian selanjutnya disarankan untuk mempertimbangkan atribut non-akademik seperti ekonomi dan keaktifan organisasi, menguji algoritma ensemble lain seperti *Gradient Boosting* atau XGBoost sebagai pembanding tambahan, dan menggunakan data dengan periode antarangkatan yang lebih seragam.

DAFTAR PUSTAKA

- [1] “BERITA NEGARA REPUBLIK INDONESIA.” [Online]. Available: www.peraturan.go.id
- [2] M. Sulehu, W. Wisda, F. Wanita, and M. Markani, “Optimasi Prediksi Kelulusan Mahasiswa Menggunakan Random Forest untuk Meningkatkan Tingkat Retensi,” *Jurnal Minfo Polgan*, vol. 13, no. 2, pp. 2364–2374, Jan. 2025, doi: 10.33395/jmp.v13i2.14472.
- [3] H. Andrianof, A. P. Gusman, and O. A. Putra, “Implementasi Algoritma Random Forest untuk Prediksi Kelulusan Mahasiswa Berdasarkan Data Akademik: Studi Kasus di Perguruan Tinggi Indonesia,” *Jurnal Sains Informatika Terapan (JSIT) E-ISSN*, no. 1, p. 2025, 2025.
- [4] M. Hafizh Sytar and Ermanita, “PREDIKSI KELULUSAN MAHASISWA TEPAT WAKTU DENGAN METODE RANDOM FOREST BERDASARKAN KLASIFIKASI ALGORITMA K-MEANS,” *Jurnal Pendidikan Matematika : Judika Education*, vol. 8, no. 3, p. 2025, 2025, doi: 10.31539/judika.v8i3.0xez5165.
- [5] G. Puspa, M. F. Rachmadhan Amri, and M. P. Nugraha, “Pemanfaatan Random Forest untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa Studi Kasus: Institut Desain dan Bisnis Bali,” *JIKO (Jurnal Informatika dan Komputer)*, vol. 9, no. 2, p. 413, Jun. 2025, doi: 10.26798/jiko.v9i2.1886.
- [6] R. Bakri, N. P. Astuti, and A. S. Ahmar, “Evaluating Random Forest Algorithm in Educational Data Mining: Optimizing Graduation on-time prediction using Imbalance Methods,” *ARRUS Journal of Social Sciences and Humanities*, vol. 4, no. 1, pp. 108–116, Feb. 2024, doi: 10.35877/soshum2449.
- [7] S. Diantika, H. Nalatissifa, N. Maulidah, R. Supriyadi, and A. Fauzi, “Penerapan Teknik Random Oversampling Untuk Memprediksi Ketepatan Waktu Lulus Menggunakan Algoritma Random Forest,” 2024. [Online]. Available: <https://www.kaggle.com/>
- [8] N. Ranti Muntiari, K. Hudaiby Hanif, P. Studi Promosi Kesehatan, P. Kaltara, and K. Utara, “PENANGANAN KETIDAKSEIMBANGAN DATA PADA KLASIFIKASI PENYAKIT CAMPAK MENGGUNAKAN KOMBINASI SMOTE DAN XGBOOST,” 2026.
- [9] A. Armansyah and R. K. Ramli, “Model Prediksi Kelulusan Mahasiswa Tepat Waktu dengan Metode Naïve Bayes,” *Edumatic: Jurnal Pendidikan Informatika*, vol. 6, no. 1, pp. 1–10, Jun. 2022, doi: 10.29408/edumatic.v6i1.4789.
- [10] P. Nurawati, E. Supriyati, and T. Listyorini, “ANALISIS SENTIMEN TERHADAP PENGGEAR K-POP DI MEDIA SOSIAL TWITTER MENGGUNAKAN NAIVE BAYES (STUDI KASUS PENGGEAR GRUP BTS).”
- [11] D. Rizki Aditya, E. Supriyati, and T. Listyorini, “ANALISIS SENTIMEN PENGGUNA TWITTER TERHADAP ROKOK ELEKTRIK (VAPE) DI INDONESIA MENGGUNAKAN METODE NAÏVE BAYES.
- [12] D. Triyana, M. Muharrom, A. Haromainy, and H. Maulana, “IMPLEMENTASI METODE ENSEMBLE MAJORITY VOTE PADA ALGORITMA NAIVE BAYES DAN RANDOM FOREST UNTUK

- ANALISIS SENTIMEN TWITTER HARGA TIKET PESAWAT DOMESTIK,” 2024.
- [13] Y. Dani and M. A. Ginting, “Comparison of Iris dataset classification with Gaussian naïve Bayes and decision tree algorithms,” *International Journal of Electrical and Computer Engineering*, vol. 14, no. 2, pp. 1959–1968, 2024, doi: 10.11591/ijece.v14i2.pp1959-1968.
- [14] N. L. Fitriyani *et al.*, “A Novel Approach Utilizing Bagging, Histogram Gradient Boosting, and Advanced Feature Selection for Predicting the Onset of Cardiovascular Diseases,” *Mathematics*, vol. 13, no. 13, Jul. 2025, doi: 10.3390/math13132194.
- [15] S. Yusnaini Putri, S. rahman, N. Ramadani, N. Aprianti Ginting, L. Syalsyadilla, and D. Agustriaman Zebua, “EVALUASI RECURSIVE FEATURE ELIMINATION UNTUK KLASIFIKASI KANKER PAYUDARA MENGGUNAKAN BERBAGAI ALGORITMA MACHINE LEARNING,” 2026.