MIND (Multimedia Artificial Intelligent Networking Database) Journal

MIND Journal | ISSN (p): 2528-0015 | ISSN (e): 2528-0902 | Vol. 11 | No. 1 | Halaman 115 - 126
DOI : https://doi.org/10.26760/mindjournal.v11i1.115-126 Juni 2026

Prediksi Retensi Mahasiswa Menggunakan Algoritma *Random Forest* dengan Optimasi Algoritma Genetika

PUDY PRIMA, AHMAD RIO ADRIANSYAH, ALFIAN NUR USYAID*

Department of Informatics, Sekolah Tinggi Teknologi Terpadu Nurul Fikri, Depok, Indonesia

Email: alfi22132ti@student.nurulfikri.ac.id

Received 20 Mei 2025 **|** *Revised* 8 Juni 2026 **|** *Accepted* 26 Juni 2026

ABSTRAK

Ketidakseimbangan kelas (imbalanced data) memicu bias mayoritas pada model konvensional dalam memprediksi retensi mahasiswa. Penelitian ini mengusulkan model peringatan dini (early warning system) dengan mengintegrasikan teknik penyeimbangan data Synthetic Minority Over-sampling Technique (SMOTE) dan pengklasifikasi Random Forest (RF). Untuk menghindari inefisiensi pencarian hyperparameter manual, Algoritma Genetika (GA) diaplikasikan guna melakukan optimasi secara global. Pengujian terhadap dataset historis mahasiswa STT Terpadu Nurul Fikri angkatan 2021 membuktikan bahwa kombinasi SMOTE dan GA-RF sangat efektif. Model hibrida ini mencapai akurasi global 99%, dengan nilai Precision 1,00 dan Recall 0,67 pada deteksi kelas minoritas (dropout). Analisis ekstraksi fitur (Feature Importance) mengungkap bahwa ketahanan studi mahasiswa didominasi oleh performa Indeks Prestasi Semester (IPS) di tahun pertama serta faktor administratif berupa jalur pendaftaran seleksi mandiri.

***Kata kunci**: prediksi Dropout, Ketidakseimbangan Data, SMOTE, Random Forest, Algoritma Genetika*

ABSTRACT

Class imbalance triggers majority bias in conventional models for predicting student retention. This study proposes an early warning model integrating the Synthetic Minority Over-sampling Technique (SMOTE) for data balancing and a Random Forest (RF) classifier. To avoid manual hyperparameter tuning inefficiencies, a Genetic Algorithm (GA) is applied for global optimization. Testing on the 2021 historical student dataset of STT Terpadu Nurul Fikri proves the effectiveness of combining SMOTE and GA-RF. The hybrid model achieved 99% global accuracy, with 1.00 Precision and 0.67 Recall in minority class (dropout) detection. Feature Importance analysis reveals that student study retention is predominantly driven by first-year Grade Point Average (GPA) performance and administrative factors, specifically the independent selection admission path.

***Keywords**: Dropout Prediction, Imbalanced Data, SMOTE, Random Forest, Genetic Algorithm*

1. PENDAHULUAN

Retensi mahasiswa merupakan salah satu fundamental dalam mengevaluasi kualitas, efektivitas, dan keberhasilan sebuah institusi pendidikan tinggi **(Moesarofah, 2021)**. Secara konsep, retensi mengacu pada kemampuan perguruan tinggi untuk mempertahankan mahasiswanya agar tetap aktif terdaftar dan berproses dari satu semester ke semester berikutnya, hingga menyelesaikan studi sesuai dengan kurikulum yang berlaku. Dalam perspektif manajemen pendidikan tinggi modern, tingkat retensi yang rendah tidak hanya mencerminkan adanya kegagalan dalam proses pembinaan akademik, tetapi juga berimplikasi langsung pada instabilitas finansial institusi akibat hilangnya potensi pendapatan biaya kuliah. Lebih jauh lagi, tingginya angka putus studi (*dropout*) dapat memicu penurunan reputasi akademik perguruan tinggi di mata publik dan lembaga akreditasi **(Campbell & Mislevy, 2009)**. Oleh karena itu, identifikasi dini terhadap mahasiswa yang berisiko tinggi mengalami kegagalan studi menjadi langkah pencegahan yang cukup penting.

Faktor-faktor yang memengaruhi retensi mahasiswa diketahui sangat kompleks dan bersifat multidimensi, mencakup faktor internal seperti motivasi dan kemampuan kognitif, serta faktor eksternal seperti dukungan finansial. Mahasiswa yang gagal melakukan integrasi akademik, yang seringkali ditandai dengan penurunan drastis pada Indeks Prestasi Semester (IPS) atau jumlah Satuan Kredit Semester (SKS) yang diselesaikan pada tahun pertama perkuliahan, memiliki probabilitas yang jauh lebih tinggi untuk masuk ke dalam kategori *dropout*. Dalam upaya mengatasi permasalahan ini secara lebih sistematis dan terukur, institusi pendidikan saat ini mulai beralih dari metode evaluasi konvensional menuju pendekatan berbasis analisis data yang dikenal sebagai *Educational Data Mining* atau EDM **(Kharis & Zili, 2022)**.

EDM merupakan disiplin ilmu interdisipliner yang memanfaatkan teknik *Data Mining* dan *Machine Learning* untuk mengeksplorasi data unik dari lingkungan pendidikan, memprediksi prestasi akademik, serta mengidentifikasi pola-pola tersembunyi yang memengaruhi perilaku pembelajar **(Indahyanti et al., 2022)**. Melalui implementasi metode *Supervised Learning*, peneliti dan institusi mampu membangun sebuah model prediktif yang dapat mengklasifikasikan status masa depan mahasiswa ke dalam kategori lulus, masih terdaftar, atau *dropout* **(Dridi, 2024)**. Model komputasi inilah yang kemudian dimanfaatkan sebagai inti dari sistem peringatan dini (*early warning system*), sebuah mekanisme yang memungkinkan pihak kampus untuk mendeteksi anomali performa akademik sejak dini dan merumuskan intervensi proaktif yang tepat sasaran sebelum mahasiswa benar-benar memutuskan untuk berhenti kuliah **(Fitriana et al., 2024)**.

Penelitian terdahulu telah banyak mengeksplorasi penggunaan berbagai algoritma *Supervised Learning* standar untuk menyelesaikan tugas klasifikasi retensi mahasiswa. Sebagai contoh, penelitian sebelumnya berhasil menerapkan algoritma *Random Forest* (RF) untuk memprediksi kelulusan dan meningkatkan retensi, namun model tersebut hanya mampu mencapai tingkat akurasi sebesar 70% dan masih terkendala oleh penanganan data yang tidak seimbang **(Sulehu et al., 2025)**. Studi lain menerapkan algoritma *Decision Tree* untuk memprediksi kelulusan dan berhasil mencapai akurasi sebesar 85%, namun model ini dinilai sangat rentan terhadap masalah *overfitting* **(Bedri et al., 2025)**. Komparasi lebih lanjut juga dilakukan dengan memanfaatkan *Support Vector Machine* (SVM) yang menghasilkan tingkat akurasi sebesar 85.06% **(M Riski Qisthiano, 2022)**.

Meskipun penelitian-penelitian dasar sebelumnya berhasil membuktikan kelayakan implementasi *Machine Learning* di ranah pendidikan, terdapat satu kesenjangan (*research gap*) yang signifikan. Kinerja algoritma *Supervised Learning* standar diketahui sangat sensitif terhadap konfigurasi *hyperparameter* bawaannya. Pencarian parameter dengan hanya

melakukan *trial-and-error* sangat tidak efisien dan seringkali membuat algoritma terjebak pada solusi yang tidak maksimal **(Vincent & Jidesh, 2023)**. Untuk mengatasi kelemahan fundamental ini, penerapan pendekatan metaheuristik, khususnya Algoritma Evolusioner, menjadi sangat relevan dan dibutuhkan **(Katoch et al., 2021)**.

Beberapa penelitian terbaru mulai mengadopsi teknik optimasi untuk meningkatkan performa model standar. Penggunaan Algoritma Genetika (GA) terbukti mampu mengatasi sensitivitas *hyperparameter* tersebut. Sebagai bukti empiris, integrasi GA pada model SVM berhasil mendongkrak tingkat akurasi secara signifikan menjadi 86.57% **(Ridwansyah et al., 2020)**. Pola keberhasilan yang sama juga ditemukan ketika GA digunakan untuk mengoptimasi model *Naïve Bayes*, di mana akurasi model meningkat dari *baseline* sebelumnya menjadi 85.43% **(Nawawi et al., 2024)**. Fakta ini mengukuhkan bahwa metode hibrida yang menggabungkan *Supervised Learning* dengan algoritma optimasi evolusioner mampu memberikan performa prediksi yang jauh lebih superior.

Meskipun optimasi menggunakan GA telah terbukti sangat efektif pada algoritma klasifikasi tunggal seperti SVM dan *Naïve Bayes*, kombinasi antara algoritma *ensemble* seperti RF dengan GA masih belum dieksplorasi secara mendalam untuk memecahkan kasus retensi mahasiswa. Selain itu, dataset akademik pada umumnya memiliki karakteristik ketidakseimbangan kelas (*imbalanced data*) yang ekstrem, yang mana representasi data mahasiswa *dropout* sangat minoritas dibandingkan dengan mahasiswa yang lulus. Apabila tidak ditangani melalui teknik pra-pemrosesan seperti *Synthetic Minority Over-sampling Technique* (SMOTE), model komputasi secanggih apa pun akan cenderung menghasilkan bias mayoritas dan gagal mengenali pola mahasiswa yang berisiko putus studi.

Mempertimbangkan celah penelitian dan urgensi penanganan data yang tidak seimbang pada penelitian sebelumnya, penelitian ini bertujuan untuk merancang, mengimplementasikan, dan mengevaluasi sebuah arsitektur model klasifikasi hibrida berbasis RF yang dioptimasi secara otomatis menggunakan GA. Penelitian ini berkontribusi dalam mengusulkan arsitektur model prediktif hibrida (GA-RF) yang mampu mendapatkan ruang pencarian *hyperparameter* secara efisien tanpa intervensi manual. Model ini mengintegrasikan SMOTE pada tahap pelatihan untuk menstabilkan sensitivitas algoritma dalam mendeteksi kelas minoritas (*dropout*). Analisis kemudian disajikan melalui interpretasi fitur (*Feature Importance*) untuk memberikan wawasan empiris kepada institusi mengenai variabel akademik tahun pertama yang paling memengaruhi probabilitas ketahanan studi mahasiswa.

2. METODOLOGI

2.1 Kerangka Kerja Penelitian (CRISP-DM)

Penelitian ini dirancang menggunakan pendekatan kuantitatif eksperimental yang terstruktur dengan mengadopsi standar metodologi *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Penggunaan CRISP-DM dipilih karena kemampuannya dalam menyediakan siklus hidup proyek *data science* yang menyeluruh, terstruktur, dan berorientasi pada pemecahan masalah dunia nyata **(Martinez-Plumed et al., 2021)**. Metodologi ini diimplementasikan melalui enam tahapan utama yang saling berkesinambungan, yaitu: (1) Pemahaman Bisnis (*Business Understanding*) untuk memformulasikan urgensi deteksi retensi mahasiswa; (2) Pemahaman Data (*Data Understanding*) melalui pengumpulan data historis akademik; (3) Persiapan Data (*Data Preparation*) yang mencakup proses *cleaning* dan penyeimbangan kelas menggunakan SMOTE; (4) Pemodelan (*Modeling*) menggunakan hibrida GA-RF; (5) Evaluasi (*Evaluation*) menggunakan metrik *Classification Report*; dan (6) Penyebaran (*Deployment*) sebagai konseptualisasi *Early Warning System* di lingkungan institusi.

2.2 Studi Pendahuluan (Preliminary Study)

Sebelum merancang dan mengimplementasikan model pada dataset historis STT Terpadu Nurul Fikri, studi pendahuluan telah dilakukan untuk menguji keandalan arsitektur hibrida GA-RF. Pengujian tahap awal ini memanfaatkan dataset publik dari repositori Kaggle (https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention), yaitu "*Predict students' dropout and academic success*" , yang terdiri dari lebih dari 4.400 *instans* data. Evaluasi performa dilakukan menggunakan 885 sampel data uji murni untuk mengukur efektivitas optimasi secara empiris.

Hasil studi pendahuluan yang disajikan pada Tabel 1 membuktikan bahwa algoritma RF standar memiliki keterbatasan, terutama dalam *F1-Score*. Setelah proses optimasi *hyperparameter* diaplikasikan menggunakan Algoritma GA, kinerja model mengalami peningkatan yang signifikan pada seluruh metrik evaluasi.

Tabel 1. Perbandingan Kinerja Model Pada Dataset Kaggle

Model	Akurasi	*F1-Score*	*Precision*	*Recall*
RF Standar	76,5%	0,74	0,75	0,73
GA-RF	82,1%	0,81	0,82	0,81

Peningkatan paling krusial dari implementasi GA-RF terlihat pada stabilitas metrik *F1-Score* yang mencapai 0,81, menandakan bahwa model jauh lebih sensitif dalam mendeteksi ancaman kelas minoritas (*dropout*). Hasil komparasi empiris pada data berskala besar ini mengonfirmasi bahwa pendekatan optimasi evolusioner menggunakan GA berhasil meningkatkan keandalan model dasar RF secara signifikan. Berdasarkan keberhasilan *proof-of-concept* tersebut, arsitektur GA-RF dinilai sangat layak dan diyakini mampu menangani kasus prediksi retensi mahasiswa pada dataset lokal STT Terpadu Nurul Fikri dengan optimal.

2.3 Pengumpulan dan Deskripsi Data

Tahap pengumpulan data (*Data Acquisition*) dilakukan dengan mengakuisisi data sekunder secara historis dari basis data akademik Sekolah Tinggi Teknologi Terpadu Nurul Fikri (STT-NF). Populasi data yang diekstraksi difokuskan pada mahasiswa angkatan 2021. Pemilihan angkatan ini didasarkan pada pertimbangan bahwa subjek penelitian telah melewati masa studi evaluasi awal yang cukup dan memiliki status akademik akhir yang definitif, mengacu pada pedoman standar dari Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi **(Pusdatin Kemendikbud, 2020)**.

Tabel 2. Deskripsi Variabel dan Atribut Dataset

Nama Atribut	Tipe Data	Keterangan
IPS Semester 1	Numerik (*Float*)	Nilai Indeks Prestasi Semester mahasiswa pada semester pertama perkuliahan
IPS Semester 2	Numerik (*Float*)	Nilai Indeks Prestasi Semester mahasiswa pada semester kedua perkuliahan
SKS Semester 1	Numerik (*Integer*)	Jumlah total Satuan Kredit Semester yang diambil pada semester pertama
SKS Semester 2	Numerik (*Integer*)	Jumlah total Satuan Kredit Semester yang diambil pada semester kedua.
Jalur Pendaftaran	Kategorikal	Jalur penerimaan mahasiswa baru saat mendaftar (SBMPTN, Seleksi mandiri)
Sistem Kuliah/Tipe Kelas	Kategorikal	Tipe kelas atau sistem perkuliahan mahasiswa (Reguler, Karyawan).

Nama Atribut	Tipe Data	Keterangan
Status Retensi sebagai Target	Numerik (Biner)	Label status akademik akhir mahasiswa (0 = *Dropout*, 1 = Aktif/Lulus)

Berdasarkan tahap seleksi fitur (*Feature Selection*) dari eksperimen awal, tidak semua atribut profil mahasiswa digunakan. Penelitian ini mereduksi dimensi data dan hanya memfokuskan parameter pada kinerja akademik awal serta administratif dasar yang secara empiris terbukti memiliki korelasi. Variabel dependen (target) dalam penelitian ini diklasifikasikan secara biner, yaitu kelas 0 untuk mahasiswa *dropout* (putus studi) dan kelas 1 untuk mahasiswa lulus/aktif.

Tabel 3. Sampel Dataset

IPS Semester 1	IPS Semester 2	SKS Semester 1	SKS Semester 2	Jalur Pendaftaran	Sistem Kuliah	Status Akhir
3.89	3.89	21	21	Seleksi Mandiri PTS	Reguler	1
3.22	2.83	21	21	Seleksi Mandiri PTS	Reguler	1
0.38	0.00	21	0	Seleksi Mandiri PTS	Reguler	0
0.16	0.00	21	0	Seleksi Mandiri PTS	Reguler	0
…	…	…	…	…	…	…

Sebagai representasi dari populasi data yang digunakan, Tabel 3 menyajikan sampel dataset operasional historis mahasiswa. Data mentah tersebut memuat nilai riil dari atribut akademik dan administratif sebelum ditransformasi menjadi format numerik (*encoding*) dan diseimbangkan menggunakan algoritma SMOTE. Cuplikan ini memperlihatkan contoh instans untuk kelas mayoritas (Status_Akhir = 1) maupun kelas minoritas (Status_Akhir = 0) yang menjadi basis pelatihan model Machine Learning dalam penelitian ini.

2.4 Pra-Pemrosesan Data dan Penanganan *Imbalance* (SMOTE)

Mengingat data operasional pendidikan seringkali bersifat *noisy* dan tidak terstruktur, pra-pemrosesan menjadi tahapan yang sangat penting. Proses ini diawali dengan pembersihan data (*data cleaning*) untuk mengeliminasi nilai yang hilang (*missing values*) dan anomali. Selanjutnya, dilakukan transformasi data numerik (*encoding*) pada variabel kategorikal seperti 'Sistem Kuliah' dan 'Jalur Pendaftaran' agar dapat diproses oleh algoritma komputasi matematis. Dataset kemudian dipartisi secara proporsional menjadi 80% data latih (*training set*) untuk proses pembelajaran algoritma, dan 20% data uji (*testing set*) sebagai data validasi (*unseen data*).

Karakteristik dataset akademik ini seringkali merepresentasikan fenomena kelas yang tidak seimbang (*imbalanced* data), jadi tantangan terbesarnya adalah klasifikasi retensi yang jumlah mahasiswa berstatus aktif/lulus mendominasi secara ekstrem. Ketimpangan ekstrem ini dapat menyebabkan algoritma klasifikasi konvensional mengalami bias, di mana model cenderung mengabaikan kelas minoritas dan hanya mengejar akurasi tinggi pada kelas mayoritas **(Fernandez et al., 2018)**. Oleh sebab itu, teknik *Synthetic Minority Over-sampling Technique* (SMOTE) diterapkan secara eksklusif pada *training set*. SMOTE bekerja dengan cara membangkitkan sampel data sintetis baru di antara sampel kelas minoritas (dropout) yang saling bertetangga menggunakan prinsip *K-Nearest Neighbors*, sehingga model tidak menjadi bias terhadap kelas mayoritas saat fase pelatihan **(Matharaarachchi et al., 2024)**.

2.5 Algoritma Klasifikasi *Random Forest*

Random Forest (RF) dipilih sebagai algoritma pengklasifikasi dasar (*baseline classifier*) karena kemampuannya yang kokoh (*robust*) dalam memodelkan hubungan non-linear yang kompleks dan ketahanannya terhadap gejala *overfitting* **(Ramdhani & Arifai, 2025)**. Sebagai metode *ensemble learning*, RF beroperasi dengan membangun sekumpulan pohon keputusan (*decision trees*) melalui mekanisme *Bootstrap Aggregating* (Bagging). Dalam proses pembagian node (percabangan kriteria), algoritma mengukur kualitas pemisahan data menggunakan metrik Gini Impurity. Formula matematis Gini Impurity dapat didefinisikan sebagai berikut **(Salman et al., 2024)**:

$$Gini(p) = 1 - \sum_{i=1}^{C} p_i^2 \quad (1)$$

C merupakan total jumlah kelas klasifikasi (dalam kasus ini $\mathrm{C} = 2$, yaitu *dropout* dan lulus), dan $\mathrm{p_i}$ merepresentasikan probabilitas atau rasio sampel yang termasuk dalam kelas i pada suatu node tertentu. Semakin rendah nilai *Gini Impurity*, semakin homogen dan baik pemisahan kelas yang dilakukan oleh pohon keputusan.

2.6 Optimasi Hyperparameter dengan Algoritma Genetika (GA)

Meskipun Arsitektur RF sangat kuat, performa optimalnya sangat dikendalikan oleh *hyperparameter* seperti jumlah pohon (*n_estimators*) dan kedalaman maksimum pohon (*max_depth*). Untuk menemukan konfigurasi terbaik tanpa metode trial-and-error yang memakan waktu komputasi besar, penelitian ini mengimplementasikan GA. GA adalah algoritma metaheuristik yang terinspirasi dari teori seleksi alam dan genetika evolusioner Darwin **(Katoch et al., 2021)**. Arsitektur optimasi GA dalam penelitian ini melalui serangkaian siklus.

Pertama, menginisialisasi populasi, di mana sistem membangkitkan populasi awal berupa kromosom (individu) acak, di mana masing-masing kromosom mempresentasikan satu paket kandidat *hyperparameter* RF.

Kedua, setiap kromosom dievaluasi kualitasnya. Kinerja model RF yang dilatih menggunakan kromosom tersebut diukur menggunakan nilai validasi silang (*cross-validation*). Fungsi kebugaran (*Fitness Function*) pada penelitian ini diformulasikan untuk memaksimalkan skor *F1-Macro* :

$$Fitness(x) = F1\text{-}Macro\left(Model_{RF(x)}\right) \quad (2)$$

Ketiga, dilakukan seleksi dengan algoritma memilih individu terbaik menggunakan metode *Tournament Selection* dengan ukuran turnamen $\mathrm{k} = 3$. Tiga individu diambil secara acak, dan individu dengan nilai fitness (skor *F1-Macro*) tertinggilah yang berhak menjadi induk (*parent*). Keempat, menggunakan teknik *Two-Point Crossover* untuk saling menukar segmen genetik antara dua kromosom induk, sehingga menghasilkan keturunan (*offspring*) dengan karakteristik campuran yang diharapkan lebih unggul.

Kelima, Bermutasi (*Mutation*), di mana mekanisme modifikasi nilai gen secara acak dengan probabilitas kecil. Operasi mutasi ini sangat krusial untuk menjaga diversitas genetik populasi dan mencegah konvergensi dini pada nilai optimal lokal (*local optima*) **(Vincent & Jidesh, 2023)**. Proses evolusi ini diulang iteratif hingga mencapai batas maksimum generasi, yang pada akhirnya akan mengekstrak satu individu terbaik sebagai set *hyperparameter* mutlak untuk membangun model final.

2.7 Metrik Evaluasi Model

Pengujian performa model final dilakukan secara objektif menggunakan 20% data uji (*testing set*) murni yang tidak pernah bersentuhan dengan proses SMOTE maupun pelatihan. Kinerja diukur menggunakan *Confusion Matrix* yang menghasilkan metrik *Precision*, *Recall*, dan *F1-Score*. Secara khusus, penelitian ini menitikberatkan evaluasi pada metrik *Recall* kelas *dropout* dan rata-rata makro (*Macro-F1 Score*) untuk memastikan bahwa model memberikan akurasi yang representatif dan tidak bias dalam mendeteksi ancaman putus studi pada dataset biner yang tidak seimbang **(Takahashi et al., 2022)**.

3. HASIL DAN PEMBAHASAN

3.1 Implementasi dan Dampak Penyeimbangan Data

Tahap eksperimen diawali dengan mengevaluasi kondisi dataset historis mahasiswa angkatan 2021. Eksplorasi awal mengonfirmasi adanya fenomena ketidakseimbangan kelas (*class imbalance*) yang sangat ekstrem. Mahasiswa berstatus lulus/aktif (kelas 1) mendominasi secara masif, sedangkan mahasiswa *dropout* (kelas 0) merepresentasikan kurang dari 5% total populasi. Jika algoritma konvensional dilatih menggunakan distribusi mentah ini, model dipastikan akan mengalami bias mayoritas ketika memprediksi seluruh mahasiswa lulus demi mengejar tingkat akurasi global yang tinggi, namun buta terhadap ancaman *dropout* yang sebenarnya. Untuk memitigasi hal ini, SMOTE diaplikasikan secara eksklusif pada 80% data latih (*training set*).

Tabel 4. Distribusi Kelas Sebelum dan Sesudah SMOTE

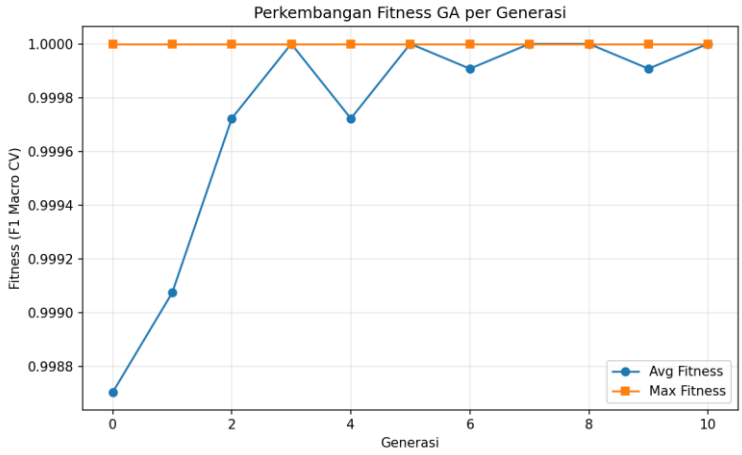
	Dropout	**Lulus/Aktif**	**Total**
Sebelum SMOTE	10	270	280
Sesudah SMOTE	270	270	540

Hasil implementasi SMOTE berhasil menyetarakan rasio kelas minoritas dan mayoritas pada ruang pelatihan menjadi proporsi 50:50 yang seimbang. Pembangkitan data sintetik ini memastikan bahwa algoritma klasifikasi dapat mengekstrak pola karakteristik mahasiswa *dropout* secara adil dan proporsional. Validitas eksperimen tetap terjaga karena 20% data uji (*testing set*) dibiarkan dalam kondisi murni tanpa intervensi sintetik.

3.2 Dinamika Evolusi Model Hibrida (GA-RF)

Pencarian *hyperparameter* optimal pada RF dieksekusi menggunakan GA guna menghindari pendekatan *trial-and-error* manual yang memakan biaya komputasi besar. Parameter GA diinisialisasi dengan ukuran populasi sebanyak 20 individu dan dievolusikan selama 10 generasi, menggunakan rata-rata *F1-Macro Score* sebagai *fitness function*.

Visualisasi evolusi pada gambar 1 merekam proses pembelajaran algoritma secara empiris. Terlihat adanya peningkatan rata-rata fitness pada generasi awal, yang kemudian diikuti oleh fluktuasi (penurunan sesaat) yang tajam pada generasi ke-4 dan ke-6. Penurunan metrik ini bukanlah kegagalan komputasi, melainkan bukti nyata beroperasinya mekanisme mutasi genetik. Algoritma secara acak mengubah struktur *hyperparameter* untuk keluar dari jebakan optimal lokal (*local optima*) dan mengeksplorasi ruang pencarian yang lebih luas. Pasca-fluktuasi, populasi kembali menguat dan mencapai titik stabil pada generasi ke-8 hingga ke-10, menghasilkan satu set konfigurasi RF yang paling tangguh (GA-RF).



Gambar 1. Perkembangan Fitness GA per Generasi

3.3 Evaluasi Kinerja Sistem Peringatan Dini

Model Hibrida optimal (GA-RF) dievaluasi menggunakan data uji murni yang berisi 71 sampel mahasiswa. Kinerja pengklasifikasi diukur menggunakan instrumen Confusion Matrix yang diringkas ke dalam *Classification Report*.

Tabel 5. Kinerja Evaluasi Model GA-RF pada Dataset Lokal (STT-NF)

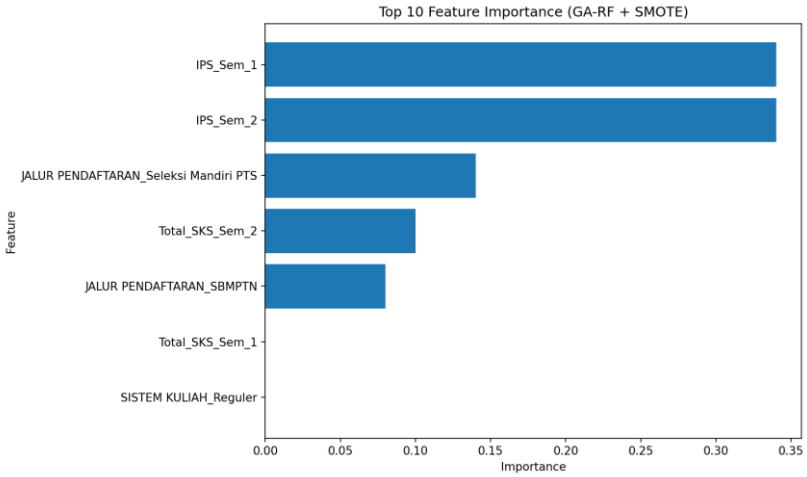
Status	*Precision*	*Recall*	*F1-Score*	*Support*
Dropout	1,00	0,67	0,80	3
Lulus/Aktif	0,99	1,00	0,99	69

Secara global, model GA-RF berhasil mencatat akurasi sebesar 99%. Namun, dalam perancangan Sistem Peringatan Dini (*Early Warning System*) untuk retensi, metrik yang paling krusial adalah kemampuan model dalam mengenali ancaman spesifik pada kelas 0 (*Dropout*). Model berhasil mencapai nilai *precision* sebesar 1,00 (100%), yang berarti model memiliki tingkat keandalan mutlak: setiap kali sistem memberi peringatan bahwa mahasiswa akan *dropout*, prediksi tersebut 100% tepat sasaran.

Di sisi lain, model mencetak skor *Recall* sebesar 0,67 (67%) untuk kelas minoritas. Dalam konteks ketersediaan data *dropout* yang sangat langka di dunia nyata, kemampuan sistem menangkap 67% dari total keseluruhan mahasiswa yang berisiko adalah pencapaian yang masif. Sinergi antara penanganan imbalanced data (SMOTE) dan optimasi parameter global (GA) terbukti berhasil melahirkan model prediktif yang akurat, dan memiliki sensitivitas operasional yang tinggi.

3.4 Ekstraksi Wawasan Penentu Retensi

Selain fungsi prediksi, arsitektur RF memiliki keunggulan analitik dalam mengekstrak bobot kepentingan (*Feature Importance*) dari setiap variabel. Ekstraksi ini memberikan wawasan empiris bagi institusi mengenai faktor apa yang paling determinan memicu putus studi. Variabel Indeks Prestasi Semester (IPS) Semester 1 dan IPS Semester 2 memimpin secara berimbang dengan kontribusi masing-masing di kisaran 34%. Temuan ini memperlihatkan fakta bahwa tahun pertama perkuliahan adalah masa kritis. Ketidakmampuan mahasiswa beradaptasi kognitif yang tercermin dari anjloknya nilai IP di semester awal adalah peringatan paling valid menuju status dropout.



Gambar 2. Grafik Bobot Kepentingan

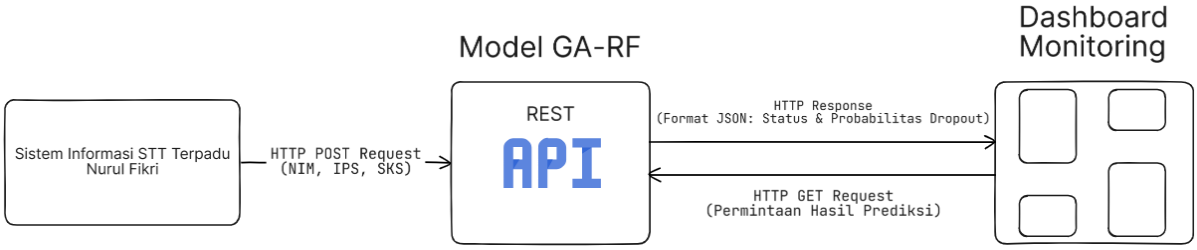
Kebaruan (novelty) dari penelitian ini terletak pada terungkapnya kontribusi administratif. Variabel "Jalur Pendaftaran_Seleksi Mandiri PTS" yang di luar dugaan menduduki peringkat ketiga dengan bobot 14%, jauh melampaui Jalur SBMPTN (8%) maupun sistem kuliah reguler (mendekati 0%). Hal ini mengindikasikan adanya korelasi implisit antara jalur masuk mandiri dengan tingkat ketahanan studi mahasiswa yang patut mendapat pengawasan khusus dari pemangku kebijakan. Variabel total SKS semester 2 memberikan pengaruh sebesar 10%. Penurunan drastis pada jumlah pengambilan beban kredit di akhir tahun pertama sering kali menjadi indikator teknis hilangnya motivasi atau kegagalan mahasiswa pada mata kuliah prasyarat.

3.5 Konseptualisasi Integrasi Deployment

Dalam batasan penelitian, tahapan ini diformulasikan sebagai rancangan konseptual untuk mengintegrasikan model ke dalam Sistem Informasi Akademik STT Terpadu Nurul Fikri di masa mendatang. Sebagai rekomendasi implementasi, model hibrida GA-RF yang telah dilatih dapat di-*deploy* sebagai layanan backend berbasis *Representational State Transfer Application Programming Interface* (REST API).

Jika institusi berencana menerapkan *Early Warning System* ini, skenario kerjanya dapat dirancang secara terotomatisasi. Sistem Informasi Akademik dapat mengirimkan data agregasi mahasiswa (seperti IPS dan SKS tahun pertama) ke layanan REST API tersebut untuk dievaluasi oleh model. Hasil probabilitas *dropout* yang dikembalikan oleh API kemudian dapat dihubungkan ke dasbor visualisasi khusus bagi Dosen Pembimbing Akademik. Rancangan sederhana ini memungkinkan pihak kampus untuk memetakan mahasiswa berisiko tinggi secara *real-time* dan melakukan intervensi preventif, tanpa harus merombak arsitektur basis data Sistem Informasi yang sudah berjalan.

Sistem Informasi STT Terpadu Nurul Fikri akan bertindak sebagai klien awal yang mengirimkan *HTTP POST Request* berisi data agregasi mahasiswa (seperti NIM, IPS, dan SKS) menuju *server* REST API Model GA-RF. Setelah model mengeksekusi prediksi, hasil probabilitas *dropout* tidak langsung dikembalikan ke sistem informasi, melainkan disediakan melalui layanan (*endpoint*) khusus. Selanjutnya, antarmuka Dashboard Monitoring yang diakses oleh Dosen Pembimbing Akademik dapat melakukan *HTTP GET Request* untuk menarik data tersebut. API kemudian memberikan respons dalam format JSON yang memuat status dan probabilitas dropout mahasiswa secara *real-time*.



Gambar 3. Arsitektur Komunikasi REST API pada Konseptualisasi Deployment

3.6 Keterbatasan Penelitian

Meskipun implementasi model hibrida GA-RF berhasil membuktikan peningkatan performa klasifikasi, penelitian ini memiliki keterbatasan yang patut dicatat secara objektif. Berdasarkan metrik evaluasi pada *testing set*, tingginya nilai *Recall* (0,67) dan *Precision* (1,00) pada deteksi kelas minoritas (*dropout*) dihasilkan dari jumlah sampel uji (*Support*) yang sangat kecil, yaitu sebanyak 3 instans. Dikarenakan secara statistik, ukuran sampel minoritas yang sangat terbatas ini membuat model rentan terhadap variansi dan dapat memengaruhi stabilitas prediksi jika dihadapkan pada distribusi data populasi berskala masif. Maka, dilakukan validasi lebih lanjut melalui studi pendahuluan pada dataset berskala besar, sehingga potensi keberhasilannya di lingkungan operasional tetap tinggi.

4. KESIMPULAN

Implementasi metode hibrida antara GA dan RF, yang disertai dengan teknik *oversampling* SMOTE, terbukti efektif dalam membangun model deteksi retensi mahasiswa pada dataset historis yang sangat tidak seimbang. Model GA-RF mencapai tingkat akurasi global sebesar 99%, dengan nilai *Precision* mutlak (1,00) dan *Recall* yang mumpuni (0,67) dalam mendeteksi kelas minoritas (*dropout*). Hal ini membuktikan bahwa sistem tidak bias terhadap kelas mayoritas dan layak dijadikan basis *Early Warning System*. Berdasarkan analisis *Feature Importance*, probabilitas mahasiswa untuk putus studi di STT Terpadu Nurul Fikri (Angkatan 2021) didominasi kuat oleh performa akademik pada tahun pertama, secara spesifik pada Indeks Prestasi Semester 1 dan 2 dengan total kontribusi kurang lebih 68%. Temuan strategis lainnya menunjukkan bahwa faktor administratif berupa Jalur Pendaftaran Seleksi Mandiri PTS memiliki pengaruh terselubung yang cukup signifikan (sekitar 14%), mematahkan asumsi bahwa ketahanan studi murni hanya bergantung pada nilai akademik.

UCAPAN TERIMA KASIH

Apresiasi dan ucapan terima kasih yang sebesar-besarnya disampaikan kepada Sekolah Tinggi Teknologi Terpadu Nurul Fikri (STT-NF) yang telah memberikan dukungan finansial untuk biaya publikasi artikel ini. Ucapan terima kasih juga ditujukan kepada pihak institusi atas dukungan fasilitas dan akses data yang memungkinkan penelitian ini terlaksana dengan baik.

DAFTAR RUJUKAN

Bedri, M. A., Putra, Y. P., & Rilvani, E. (2025). Prediksi kelulusan mahasiswa menggunakan algoritma decision tree. *Jurnal Inovasi Multidisiplin dan Teknologi Modern*. *8*(3).

Campbell, C. M., & Mislevy, J. (2009). Students' perceptions matter: Early signs of undergraduate student retention/attrition. *Harbor in the Storm: Institutional Research in the Age of Accountability*, 66–96.

Dridi, S. (2024). Supervised learning—A systematic literature review. *Open Science Framework*. https://doi.org/10.31219/osf.io/qtmcs

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192

Fitriana, S., Riniyanty, Laila, R., Pratama, S. A., & Lamasitudju, C. A. (2024). Prediksi siswa putus sekolah dan keberhasilan akademik menggunakan machine learning. *The Indonesian Journal of Computer Science*, *13*(6). https://doi.org/10.33022/ijcs.v13i6.4453

Indahyanti, U., Azizah, N. L., & Setiawan, H. (2022). Educational data mining on student academic performance prediction: A survey. *Procedia of Sciences and Humanities*.

Katoch, S., Chauhan, S. S., & Kumar, V. (2021). A review on genetic algorithm: Past, present, and future. *Multimedia Tools and Applications*, *80*(5), 8091–8126. https://doi.org/10.1007/s11042-020-10139-6

Kharis, S. A. A., & Zili, A. H. A. (2022). Learning analytics dan educational data mining pada data pendidikan. *Jurnal Riset Pembelajaran Matematika Sekolah*, *6*(1), 12–20. https://doi.org/10.21009/jrpms.061.02

Qisthiano, M. R. (2022). Klasifikasi terhadap prediksi kelulusan mahasiswa dengan menggunakan metode Support Vector Machine (SVM). *Seminar Nasional Teknologi dan Multidisiplin Ilmu (SEMNASTEKMU)*, *2*(2), 203–207. https://doi.org/10.51903/semnastekmu.v2i1.170

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Matharaarachchi, S., Domaratzki, M., & Muthukumarana, S. (2024). Enhancing SMOTE for imbalanced data with abnormal minority instances. *Machine Learning with Applications*, *18*, 100597. https://doi.org/10.1016/j.mlwa.2024.100597

Moesarofah, M. (2021). Analisis karakteristik retensi mahasiswa di perguruan tinggi. *Didaktis: Jurnal Pendidikan dan Ilmu Pengetahuan*, *21*(1). https://doi.org/10.30651/didaktis.v21i1.7005

Nawawi, I., Sugiarto, H., & Yuliandari, D. (2024). Meningkatkan akurasi prediksi kelulusan mahasiswa menggunakan metode algoritma genetika. *Jurnal Informatika*, *16*(2).

Pusdatin Kemendikbud. (2020). *Panduan penggunaan pangkalan data pendidikan tinggi (PDDikti)*. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. https://pddikti.kemdikbud.go.id

Hardiansyah, Ramdhani, I., & Mukhamad Khotib Arifai. (2025). Implementasi Algoritma Machine Learning untuk Prediksi Keberhasilan Mahasiswa di Program Studi Teknik Informatika. *Jurnal Onevision*, 1(2), 153–160. Retrieved from https://ejournal.visione.co.id/ojs/index.php/juvismi/article/view/17

Ridwansyah, R., Wijaya, G., & Purnama, J. J. (2020). Hybrid optimization method based on genetic algorithm for graduates students. *Jurnal Pilar Nusa Mandiri*, *16*(1), 53–58. https://doi.org/10.33480/pilar.v16i1.1180

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/10.58496/BJML/2024/007

Sulehu, M., Wisda, W., Wanita, F., & Markani, M. (2025). Optimasi prediksi kelulusan mahasiswa menggunakan Random Forest untuk meningkatkan tingkat retensi. *Jurnal Minfo Polgan*, *13*(2), 2364–2374. https://doi.org/10.33395/jmp.v13i2.14472

Takahashi, K., Yamamoto, K., Kuchiba, A., & Koyama, T. (2022). Confidence interval for micro-averaged F1 and macro-averaged F1 scores. *Applied Intelligence*, *52*(5), 4961–4972. https://doi.org/10.1007/s10489-021-02635-5

Vincent, A. M., & Jidesh, P. (2023). An improved hyperparameter optimization framework for AutoML systems using evolutionary algorithms. *Scientific Reports*, *13*(1), 4737. https://doi.org/10.1038/s41598-023-32027-3