**Journal of Informatics Management and Information Technology**
ISSN 2774-4744 (Media Online)
Vol 6, No 2, April 2026 | Hal 131-142
DOI: 10.47065/jimat.v6i2.950
https://hostjournals.com/jimat

# Optimasi Kinerja Algoritma Random Forest dengan SMOTE untuk Prediksi Kinerja Akademik Siswa

**Muhammad Rizky Ramadhan[1,*], Solikhun[2]**

[1] Program Studi Sistem Informasi, Sekolah Tinggi Ilmu Komputer Tunas Bangsa, Pematangsiantar, Indonesia
[2] Program Studi Teknik Informatika, Sekolah Tinggi Ilmu Komputer Tunas Bangsa, Pematangsiantar, Indonesia
Email: [1,*]mhdrizky837@gmail.com, [2]solikhun@amiktunasbangsa.ac.id
Email Penulis Korespondensi: mhdrizky837@gmail.com

**Abstrak**−Prediksi kinerja akademik siswa merupakan komponen penting dalam Educational Data Mining (EDM) untuk identifikasi dini siswa berisiko. Pendekatan ini bertujuan untuk meningkatkan akurasi prediksi serta mendukung pengambilan keputusan dalam identifikasi dini siswa berisiko. Penelitian ini mengusulkan pipeline yang dioptimalkan dengan mengintegrasikan feature engineering, Pearson Correlation-based Feature Filtering (PCFF), Synthetic Minority Over-sampling Technique (SMOTE), dan Random Forest (RF) untuk memprediksi kinerja akademik siswa. Dataset yang digunakan adalah Portuguese Student Performance Dataset (1.043 entri bersih; Pass=814, Fail=230). Empat fitur rekayasa dikonstruksi sehingga mereduksi ruang fitur menjadi 15 melalui PCFF (ambang batas |r| >= 0,1). SMOTE diterapkan secara eksklusif di dalam setiap fold pelatihan untuk mencegah data leakage. Dua model utama dievaluasi: baseline Naïve Bayes (Accuracy=90,43%) dan RF Default + SMOTE sebagai model yang diusulkan (Accuracy=93,78%, Recall=95,09%, F1=95,98%). Ten-fold stratified cross-validation menghasilkan akurasi 90,89% ± 2,08%. Fitur rekayasa G_avg mencapai skor feature importance tertinggi (0,285), melampaui fitur nilai mentah. Hasil penelitian menunjukkan bahwa integrasi SMOTE dan feature engineering secara signifikan meningkatkan deteksi kelas minoritas, dengan False Negative berkurang dari 15 (baseline) menjadi 8 (RF + SMOTE), merepresentasikan peningkatan 46,7% dalam mengidentifikasi siswa berisiko.

**Kata Kunci:** Random Forest; SMOTE; Rekayasa Fitur; Kinerja Akademik Siswa; Penambangan Data Pendidikan

**Abstract**−Student academic performance prediction is an essential component of Educational Data Mining (EDM) for the early identification of at-risk students. This approach aims to improve prediction accuracy and support decision-making for the early identification of at-risk students. This study proposes an optimized prediction pipeline that integrates feature engineering, Pearson Correlation-based Feature Filtering (PCFF), the Synthetic Minority Over-sampling Technique (SMOTE), and Random Forest (RF) to predict student academic performance. The Portuguese Student Performance Dataset (1,043 clean records; Pass = 814, Fail = 230) was used for evaluation. Four engineered features were constructed, reducing the feature space to 15 features through PCFF (threshold |r| ≥ 0.1). SMOTE was applied exclusively within each training fold to prevent data leakage. Two primary models were evaluated: a baseline Naïve Bayes model (Accuracy = 90.43%) and the proposed RF Default + SMOTE model (Accuracy = 93.78%, Recall = 95.09%, F1-score = 95.98%). Ten-fold stratified cross-validation achieved an accuracy of 90.89% ± 2.08%. The engineered feature G_avg obtained the highest feature importance score (0.285), outperforming the original grade features. The results demonstrate that integrating SMOTE and feature engineering significantly improves minority class detection, reducing False Negatives from 15 (baseline) to 8 (RF + SMOTE), representing a 46.7% improvement in identifying at-risk students.

**Keywords:** Random Forest; SMOTE; Feature Engineering; Student Performance; Educational Data Mining

## 1. PENDAHULUAN

Pendidikan merupakan fondasi utama dalam pembangunan sumber daya manusia yang berkualitas dan berdaya saing tinggi. Kinerja akademik siswa menjadi indikator penting keberhasilan proses pendidikan karena mencerminkan kemampuan siswa dalam mencapai target pembelajaran[1], [2], [3]. Rendahnya capaian akademik dan tingginya risiko putus sekolah masih menjadi tantangan serius di berbagai institusi pendidikan. Siswa dengan performa rendah cenderung memiliki risiko lebih besar mengalami kegagalan akademik dan menghentikan pendidikan sebelum menyelesaikan studinya. Oleh karena itu, diperlukan sistem prediksi yang mampu mendeteksi potensi kegagalan sejak dini agar institusi dapat melakukan tindakan preventif secara tepat dan efisien[4], [5], [6], [7].

Perkembangan teknologi informasi mendorong munculnya Educational Data Mining (EDM), yaitu bidang yang berfokus pada eksplorasi data pendidikan untuk menghasilkan informasi pendukung pengambilan keputusan akademik. EDM telah banyak dimanfaatkan untuk memprediksi performa akademik, menganalisis pola belajar, dan meningkatkan efektivitas pembelajaran[8], [9]. Dalam implementasinya, machine learning menjadi pendekatan yang dominan karena kemampuannya menangani data berdimensi tinggi dan menemukan pola hubungan yang kompleks. Berbagai algoritma seperti Naïve Bayes, Decision Tree, SVM, dan Random Forest telah diterapkan pada penelitian prediksi kinerja akademik, dengan metode ensemble learning secara konsisten menunjukkan performa superior dibandingkan algoritma tunggal[10], [11], [12], [13].

Random Forest (RF) merupakan algoritma ensemble berbasis bagging yang membangun sejumlah decision tree secara paralel dengan pemilihan fitur acak. RF memiliki kemampuan generalisasi yang baik, robust terhadap overfitting, dan mampu mengukur tingkat kepentingan fitur sehingga sering digunakan dalam penelitian klasifikasi pendidikan[14], [15], [16]. Meskipun demikian, performa RF sangat dipengaruhi oleh kualitas data pelatihan dan penanganan ketidakseimbangan kelas, sehingga diperlukan strategi yang tepat untuk memaksimalkan kemampuan prediktifnya[17], [18].

Penelitian prediksi kinerja akademik juga menghadapi permasalahan ketidakseimbangan kelas (class imbalance). Dataset yang digunakan dalam penelitian ini memiliki distribusi 814 kelas Pass (78,0%) dan 230 kelas Fail (22,0%), sehingga model berpotensi lebih cenderung memprediksi kelas mayoritas dan mengabaikan kelas minoritas. Salah satu

**Journal of Informatics Management and Information Technology**
ISSN 2774-4744 (Media Online)
Vol 6, No 2, April 2026 | Hal 131-142
DOI: 10.47065/jimat.v6i2.950
https://hostjournals.com/jimat

solusi efektif adalah Synthetic Minority Over-sampling Technique (SMOTE), yang menghasilkan sampel sintetis pada kelas minoritas secara metodologis valid tanpa risiko data leakage jika diterapkan secara terisolasi pada data latih[19], [20], [21], [22].

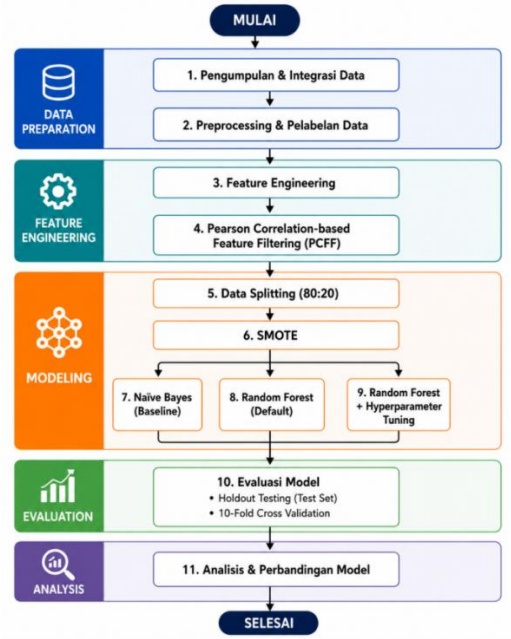
Pertama, penerapan SMOTE pada Portuguese Student Performance Dataset masih terbatas dan belum dikombinasikan dengan feature engineering berbasis domain yang terstruktur secara komprehensif. Kedua, sebagian besar penelitian belum menerapkan prosedur SMOTE yang metodologis benar, yaitu diterapkan per fold pada cross-validation untuk menghindari data leakage. Ketiga, penelitian terdahulu menggunakan Naïve Bayes dengan CFS hanya memperoleh akurasi 91,22%, sehingga masih terdapat peluang peningkatan performa yang signifikan melalui kombinasi RF dan SMOTE[23], [24], [25].

Penelitian ini mengusulkan optimasi RF menggunakan SMOTE dan feature engineering berbasis domain untuk prediksi kinerja akademik siswa. Kontribusi utama penelitian ini adalah integrasi feature engineering, PCFF, dan SMOTE yang diterapkan per fold untuk menghindari data leakage dalam satu pipeline yang komprehensif dan terdokumentasi. Berbeda dengan penelitian sebelumnya yang umumnya menerapkan SMOTE secara global sebelum pembagian data, penelitian ini memastikan SMOTE hanya diterapkan pada data latih di setiap fold secara independen, sehingga estimasi performa model tidak mengalami optimistic bias akibat kontaminasi data uji. Pendekatan ini merepresentasikan praktik metodologis yang lebih ketat dan dapat dijadikan rujukan bagi penelitian Educational Data Mining selanjutnya. Selain itu, konstruksi fitur rekayasa berbasis domain knowledge pendidikan yang dikombinasikan dengan seleksi fitur berbasis korelasi Pearson memberikan dimensi analisis yang lebih kaya dibandingkan penggunaan atribut mentah secara langsung. Tujuan penelitian: (1) membangun pipeline preprocessing yang komprehensif dengan feature engineering berbasis domain; (2) mengoptimalkan performa RF melalui penanganan ketidakseimbangan kelas menggunakan SMOTE; dan (3) mengevaluasi performa model yang diusulkan terhadap model baseline penelitian sebelumnya.

# 2. METODOLOGI PENELITIAN

## 2.1 Gambaran Umum dan Alur Penelitian

Penelitian ini menerapkan pendekatan supervised machine learning untuk mengklasifikasikan kinerja akademik siswa ke dalam dua kelas: Pass dan Fail. Alur penelitian terdiri dari sebelas tahap yang dikelompokkan dalam lima fase, sebagaimana ditunjukkan pada Gambar 1. Setiap fase dirancang secara sistematis untuk memastikan validitas metodologis, mulai dari persiapan data hingga analisis performa model secara komprehensif.




**Gambar 1.** Alur Penelitian

Fase pertama (Data Preparation) meliputi pengumpulan dan integrasi dataset serta preprocessing yang mencakup transformasi atribut kategorikal, eliminasi target leakage, dan pembersihan data duplikat. Fase kedua (Feature Engineering) mencakup konstruksi empat fitur turunan berbasis domain knowledge pendidikan, dilanjutkan seleksi fitur menggunakan Pearson Correlation-based Feature Filtering (PCFF) dengan ambang batas $|r| \geq 0{,}1$. Fase ketiga (Modeling) meliputi pembagian data dengan rasio 80:20, penanganan ketidakseimbangan kelas menggunakan SMOTE yang diterapkan secara terisolasi pada data latih, serta pembangunan dua model secara paralel: Naïve Bayes sebagai baseline dan RF + SMOTE sebagai model yang diusulkan. Fase keempat (Evaluation) menggunakan dua skema pengujian, yaitu holdout testing dan 10-Fold Stratified Cross-Validation. Fase kelima (Analysis) melakukan perbandingan performa

**Journal of Informatics Management and Information Technology**
ISSN 2774-4744 (Media Online)
Vol 6, No 2, April 2026 | Hal 131-142
DOI: 10.47065/jimat.v6i2.950
https://hostjournals.com/jimat

seluruh model berdasarkan metrik akurasi, F1-score, dan AUC-ROC. Seluruh eksperimen diimplementasikan menggunakan Python 3.10 dengan pustaka scikit-learn versi 1.3 dan imbalanced-learn versi 0.11.

### 2.2 Dataset

Dataset yang digunakan adalah Portuguese Student Performance Dataset dari UCI Machine Learning Repository, dikumpulkan dari dua sekolah menengah di Portugal. Dataset mencakup dua berkas: dataset Matematika (MAT, 395 entri) dan dataset Bahasa Portugis (POR, 649 entri), masing-masing memiliki 33 atribut demografis, sosial, dan akademik. Kedua dataset diintegrasikan menggunakan operasi concatenation. Variabel target performance dikonstruksi dari nilai akhir G3: G3 >= 10 dilabeli pass (1) dan G3 < 10 dilabeli fail (0). Statistik dataset disajikan pada Tabel 1.

**Tabel 1.** Dataset Statistik

| Atribut | Nilai | Keterangan |
|---|---|---|
| Dataset sumber | UCI ML Repository | Cortez & Silva (2008) |
| Jumlah data (raw) | 1.044 entri | Gabungan MAT + POR |
| Jumlah data (bersih) | 1.043 entri | Setelah penghapusan duplikat |
| Jumlah fitur awal | 33 fitur | Per dataset (MAT/POR) |
| Jumlah fitur setelah FE | 36 fitur | Termasuk 4 fitur rekayasa |
| Kelas Pass (1) | 814 entri | 78,0 % |
| Kelas Fail (0) | 230 entri | 22,0 % |
| Rasio class imbalance | 3,54 : 1 | Pass vs. Fail |

### 2.3 Preprocessing Data

Tahap preprocessing mencakup tiga proses. Pertama, integrasi data: kedua dataset digabungkan menghasilkan 1.044 entri. Kedua, transformasi: atribut kategorik (sex, address, Mjob, higher, dll.) dikodekan numerik menggunakan Label Encoding. Ketiga, pembersihan: atribut G3 dihapus untuk menghindari target leakage, dan satu entri duplikat dieliminasi menggunakan drop_duplicates(), menghasilkan 1.043 entri bersih. Tidak ditemukan missing values sehingga imputasi tidak diperlukan. Catatan penting: atribut school dan age tidak dihapus pada tahap ini, melainkan dipertahankan sebagai kandidat fitur yang relevansinya ditentukan oleh tahap PCFF.

### 2.4 Feature Engineering

Empat fitur turunan baru dikonstruksi berdasarkan domain knowledge pendidikan untuk menangkap pola interaksi yang tidak tersedia pada atribut mentah. Definisi matematis, justifikasi domain, dan nilai korelasi masing-masing fitur terhadap variabel target disajikan pada Tabel 2.

**Tabel 2.** Fitur Rekayasa dan Korelasi Terhadap Target

| Fitur Baru | Formula | Interpretasi | r terhadap Target |
|---|---|---|---|
| G_avg | (G1 + G2) / 2 | Rata-rata nilai dua semester | +0.665 |
| G_trend | G2 - G1 | Tren perkembangan nilai | +0.212 |
| study_fail | studytime / (failures + 1) | Efisiensi belajar relatif | +0.240 |
| goout_abs | goout x absences | Risiko sosial-akademik | -0.137 |

Fitur G_avg merepresentasikan performa kumulatif yang lebih stabil dibandingkan nilai tunggal; G_trend menangkap arah perkembangan nilai antar semester sebagai sinyal peringatan dini; study_fail mengkuantifikasi efisiensi belajar relatif terhadap riwayat kegagalan dengan pembagi *(failures+1)* untuk mencegah pembagian oleh nol; dan goout_abs menangkap efek sinergistik antara aktivitas sosial dan ketidakhadiran. Keempat fitur memperluas ruang representasi dari 33 menjadi 36 fitur. Penambahan fitur rekayasa ini bertujuan untuk memperkaya informasi yang tersedia sehingga hubungan antarvariabel dapat direpresentasikan secara lebih komprehensif dibandingkan hanya menggunakan atribut asli dataset. Dengan representasi yang lebih informatif, model diharapkan mampu mempelajari pola kinerja akademik siswa secara lebih efektif dan menghasilkan proses klasifikasi yang lebih optimal.

### 2.5 Pearson Correlation-based Feature Filtering (PCFF)

Seleksi fitur dilakukan menggunakan PCFF, yaitu metode filter yang memilih fitur berdasarkan nilai absolut koefisien korelasi Pearson terhadap variabel target. Koefisien korelasi Pearson dihitung menggunakan Persamaan (1):

$$r(X,Y) = \frac{\Sigma[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sqrt{[\Sigma(X_i - \bar{X})^2 \cdot \Sigma(Y_i - \bar{Y})^2]}} \quad (1)$$

Ambang batas ditetapkan sebesar |r| >= 0,1. Meskipun tergolong rendah, penetapan ini didasarkan pada tiga pertimbangan: (1) data pendidikan bersifat multidimensional dengan variabel-variabel yang berpengaruh kecil secara individual namun signifikan secara kolektif; (2) RF sebagai ensemble learner mampu mengeksploitasi feature interaction yang tidak terlihat pada analisis univariat; (3) keragaman fitur mendukung diversitas pohon dalam mekanisme bagging. Dari 36 fitur, diperoleh 15 fitur terpilih yang disajikan pada Tabel 3.

**Journal of Informatics Management and Information Technology**
ISSN 2774-4744 (Media Online)
Vol 6, No 2, April 2026 | Hal 131-142
DOI: 10.47065/jimat.v6i2.950
https://hostjournals.com/jimat

**Tabel 3.** Hasil Pearson Correlation-based Feature Filtering (PCFF)

| No | Fitur | Nilai Korelasi \|r\| | Deskripsi |
|---|---|---|---|
| 1 | G2 | +0.666 | Nilai semester kedua |
| 2 | G_avg | +0.665 | Rata-rata nilai G1 dan G2 (fitur rekayasa) |
| 3 | G1 | +0.614 | Nilai semester pertama |
| 4 | failures | -0.367 | Jumlah kegagalan kelas sebelumnya |
| 5 | study_fail | +0.240 | Rasio waktu belajar dan kegagalan (fitur rekayasa) |
| 6 | higher | +0.218 | Keinginan melanjutkan pendidikan tinggi |
| 7 | G_trend | +0.212 | Tren perkembangan nilai (fitur rekayasa) |
| 8 | goout_abs | -0.137 | Interaksi keluar dan absensi (fitur rekayasa) |
| 9 | age | -0.134 | Usia siswa |
| 10 | school | -0.132 | Identitas sekolah (GP atau MS) |
| 11 | absences | -0.119 | Jumlah ketidakhadiran di sekolah |
| 12 | studytime | +0.109 | Durasi belajar per minggu |
| 13 | goout | -0.107 | Intensitas keluar bersama teman |
| 14 | Medu | +0.106 | Tingkat pendidikan ibu |
| 15 | Fedu | +0.105 | Tingkat pendidikan ayah |

Fitur G2 memiliki korelasi positif tertinggi (r = +0,666) sebagai prediktor terkuat, sedangkan failures memiliki korelasi negatif terbesar (r = −0,367), mengkonfirmasi riwayat kegagalan sebagai faktor risiko dominan. Fitur school (r = −0,132) dan age (r = −0,134) dipertahankan karena keduanya memenuhi ambang batas korelasi dan berpotensi memberikan kontribusi prediktif melalui mekanisme feature interaction pada model ensemble.

### 2.6 Pembagian Data dan Penyeimbangan Kelas

Dataset yang telah melalui proses seleksi fitur dibagi menggunakan metode stratified train-test split dengan rasio 80:20 untuk mempertahankan proporsi kelas pada data pelatihan dan data pengujian. Dari 1.043 entri diperoleh 834 data latih (Fail=184, Pass=650) dan 209 data uji (Fail=46, Pass=163). SMOTE diterapkan secara eksklusif hanya pada data latih untuk menghindari data leakage. Prosedur ini menghasilkan distribusi seimbang: 650 Fail dan 650 Pass (1.300 sampel latih). Dalam konteks 10-Fold Cross-Validation, SMOTE diterapkan di dalam setiap fold pelatihan secara independen untuk mencegah optimistic bias. Secara matematis, SMOTE menghasilkan sampel sintetis menggunakan Persamaan (2):

$$x_{synthetic} = x_i + \delta \cdot (x_{neighbor} - x_i), \ \delta \sim Uniform(0,1) \quad (2)$$

di mana x_synthetic adalah sampel sintetis yang dihasilkan, $x_i$ merupakan sampel kelas minoritas yang dipilih, x_neighbor adalah salah satu tetangga terdekat dari $x_i$ dalam ruang fitur, dan δ adalah bobot acak yang diambil dari distribusi seragam pada interval [0,1] yang menentukan posisi sampel sintetis di antara kedua titik tersebut.

### 2.7 Model Random Forest

Random Forest (RF) dipilih sebagai algoritma utama karena merupakan metode ensemble learning berbasis bagging yang mampu meningkatkan kemampuan generalisasi model serta mengurangi risiko overfitting[26]. Pada proses klasifikasi, prediksi akhir ditentukan menggunakan mekanisme majority voting, yaitu memilih kelas yang memperoleh jumlah suara terbanyak dari seluruh decision tree sebagaimana ditunjukkan pada Persamaan (3).

$$y_{pred} = argmax_c \Sigma_{t=1}^{T} I(h_{t(x)} = c) \quad (3)$$

Dengan T menyatakan jumlah decision tree, $h_t(x)$ merupakan prediksi dari decision tree ke-t terhadap sampel x, sedangkan I(·) adalah fungsi indikator yang bernilai 1 jika prediksi sama dengan kelas c dan 0 jika tidak. Model RF dibangun menggunakan 15 fitur hasil Pearson Correlation-based Feature Filtering (PCFF) pada data latih yang telah diseimbangkan menggunakan Synthetic Minority Over-sampling Technique (SMOTE). Konfigurasi *default* dari scikit-learn digunakan sebagai model utama karena telah dioptimalkan secara empiris dan mampu memberikan performa yang kompetitif pada dataset berukuran menengah tanpa memerlukan penyesuaian parameter tambahan. Sebagai analisis komparatif, dilakukan pula *hyperparameter tuning* menggunakan RandomizedSearchCV dengan n_iter = 50 dan scoring = "f1_weighted" untuk mengevaluasi pengaruh optimasi parameter terhadap kinerja model. Berdasarkan hasil pencarian, konfigurasi terbaik diperoleh dengan n_estimators = 200, max_depth = 15, min_samples_split = 5, min_samples_leaf = 1, max_features = sqrt, bootstrap = False, dan class_weight = None. Selain itu, model Gaussian Naïve Bayes dibangun sebagai *baseline* menggunakan 15 fitur hasil PCFF tanpa penerapan SMOTE guna menjaga komparabilitas dengan penelitian rujukan.

### 2.8 Evaluasi Model

Performa model dievaluasi menggunakan lima metrik utama: akurasi, precision, recall, F1-Score, dan AUC-ROC, yang dihitung berdasarkan empat kuantitas pada confusion matrix[27]. Struktur confusion matrix untuk masalah klasifikasi biner disajikan pada Tabel 4.

**Journal of Informatics Management and Information Technology**
ISSN 2774-4744 (Media Online)
Vol 6, No 2, April 2026 | Hal 131-142
DOI: 10.47065/jimat.v6i2.950
https://hostjournals.com/jimat

**Tabel 4.** Struktur Confusion Matrix Klasifikasi Biner

| | | Aktual Positif (Pass) | Aktual Negatif (Fail) |
|---|---|---|---|
| Prediksi | Positif (Pass) | TP (True Positive) | FP (False Positive) |
| | Negatif (Fail) | FN (False Negative) | TN (True Negative) |

Empat kuantitas dasar pada confusion matrix adalah: True Positive (TP: prediksi pass, aktual pass); True Negative (TN: prediksi fail, aktual fail); False Positive (FP: prediksi pass, aktual fail); dan False Negative (FN: prediksi fail, aktual pass). Berdasarkan kuantitas tersebut, metrik evaluasi didefinisikan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (7)$$

Metrik AUC-ROC (Area Under the Receiver Operating Characteristic Curve) mengukur kemampuan diskriminasi model terhadap kedua kelas secara keseluruhan, dengan nilai 1,0 menunjukkan diskriminasi sempurna dan nilai 0,5 setara dengan klasifikasi acak. Selain evaluasi pada set pengujian, validasi model dilakukan menggunakan 10-Fold Stratified Cross-Validation dengan SMOTE diterapkan di dalam setiap fold pelatihan secara terpisah untuk menghindari data leakage, sehingga estimasi generalisasi model lebih andal dan tidak bias. Metrik recall diprioritaskan sebagai indikator performa kelas fail mengingat kepentingan praktisnya dalam mendeteksi siswa berisiko untuk keperluan intervensi pedagogis.

# 3. HASIL DAN PEMBAHASAN

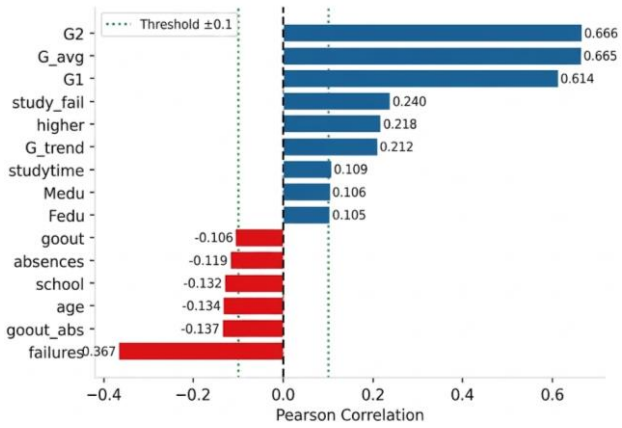
Bagian ini menyajikan hasil eksperimen penerapan Pearson Correlation-based Feature Filtering (PCFF), penanganan ketidakseimbangan kelas menggunakan SMOTE, serta evaluasi performa model Naïve Bayes dan Random Forest pada dataset performa akademik siswa, dilanjutkan dengan analisis dan pembahasan terhadap temuan-temuan tersebut beserta perbandingannya dengan penelitian terdahulu.

## 3.1 Hasil Eksperimen Model

### 3.1.1 Hasil Pearson Correlation-based Feature Filtering (PCFF)

Tahap seleksi fitur menggunakan Pearson Correlation-based Feature Filtering (PCFF) dengan ambang batas |r| >= 0,1 berhasil mereduksi dimensi data dari 36 fitur menjadi 15 fitur yang memiliki korelasi statistis bermakna terhadap variabel target performance. Reduksi ini merepresentasikan penurunan dimensi sebesar 58,3% yang secara teoritis mengurangi kompleksitas komputasi dan meminimalkan risiko curse of dimensionality.

Gambar 2 menunjukkan distribusi korelasi seluruh fitur terpilih. Fitur akademik G2, G_avg, G1—mendominasi tiga peringkat teratas dengan |r| masing-masing 0,666, 0,665, dan 0,614. Seluruh empat fitur rekayasa berhasil melewati ambang batas seleksi, mengkonfirmasi validitas pendekatan feature engineering berbasis domain yang diterapkan. Atribut failures menunjukkan korelasi negatif terbesar (|r| = 0,367), mengkonfirmasi riwayat kegagalan sebagai faktor risiko dominan. Fitur-fitur dengan korelasi rendah seperti Fedu dan Medu tetap dipertahankan karena meskipun kontribusinya kecil secara individual, keduanya berpotensi memberikan informasi tambahan melalui mekanisme feature interaction pada model ensemble.

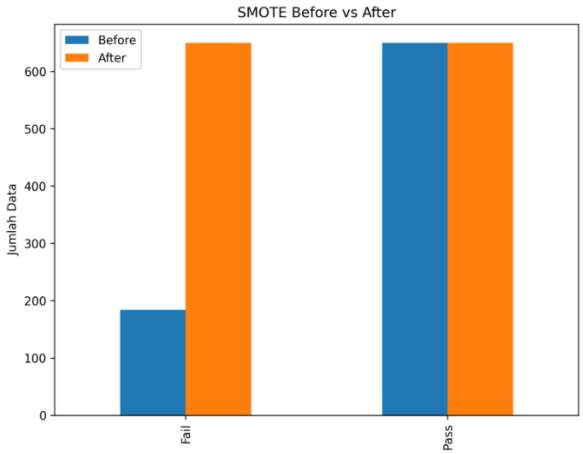



**Gambar 2.** Korelasi Fitur Terpilih

**Journal of Informatics Management and Information Technology**
ISSN 2774-4744 (Media Online)
Vol 6, No 2, April 2026 | Hal 131-142
DOI: 10.47065/jimat.v6i2.950
https://hostjournals.com/jimat

### 3.1.2 Hasil Penanganan Class Imbalance

Dataset setelah preprocessing memiliki 1.043 entri dengan distribusi 814 kelas Pass (78,0%) dan 230 kelas Fail (22,0%), membentuk rasio ketidakseimbangan 3,54:1.



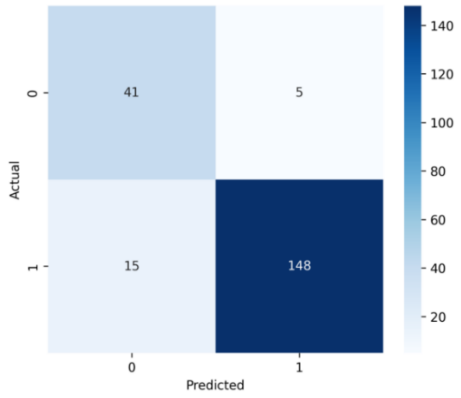

**Gambar 3.** Efek SMOTE pada Training Data

Penerapan SMOTE secara eksklusif pada data latih menghasilkan distribusi seimbang: 650 Fail dan 650 Pass (1.300 sampel latih), meningkat dari kondisi awal 184 Fail dan 650 Pass. Gambar 3 memvisualisasikan efek SMOTE pada data latih. SMOTE tidak diterapkan pada model baseline Naïve Bayes untuk menjaga komparabilitas dengan kondisi eksperimen pada penelitian rujukan.

### 3.1.3 Evaluasi Model Baseline: Naïve Bayes

Model Naïve Bayes Gaussian yang dilatih menggunakan 15 fitur terpilih dari PCFF tanpa aplikasi SMOTE menghasilkan akurasi 90,43% pada data uji (n=209). Confusion matrix model Naïve Bayes disajikan pada Gambar 4.

Dari 209 data uji, model menghasilkan 148 True Positive, 41 True Negative, 5 False Positive, dan 15 False Negative. Nilai precision sebesar 96,73% mengindikasikan selektivitas tinggi dalam prediksi pass, namun 15 False Negative menunjukkan bahwa model melewatkan 15 dari 46 siswa berisiko gagal sebuah limitasi yang signifikan dalam konteks sistem peringatan dini. Nilai AUC-ROC baseline sebesar 0,951 (95,1%) mengindikasikan kemampuan diskriminasi yang sudah baik namun masih dapat ditingkatkan.




**Gambar 4.** Confusion Matrix Naïve bayes

### 3.1.4 Evaluasi Model Random Forest

Dua konfigurasi RF dievaluasi menggunakan data latih yang diseimbangkan dengan SMOTE (1.300 sampel) dan diuji pada holdout set (n=209). Perbandingan performa seluruh model disajikan pada Tabel 5.

**Tabel 5.** Perbandingan Komprehensif Performa Model pada Data Uji (n = 209)

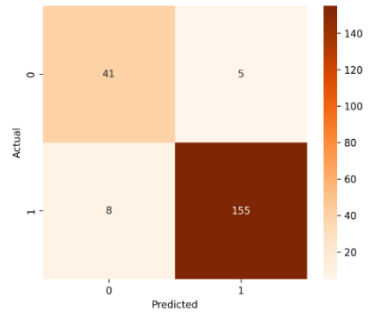
| Model | Accuracy | Recall | Precision | F1-Score |
|---|---|---|---|---|
| NB + PCFF Rujukan (Gori et al. [23], 2024) | 91.22% | 91.22% | 92.24% | 91.48% |
| NB + PCFF (this study, baseline) | 90.43% | 90.80% | 96.73% | 93.67% |
| RF Default + SMOTE (proposed) | 93.78% | 95.09% | 96.88% | 95.98% |
| RF + SMOTE + Tuning (eksplorasi) | 92.82% | 93.87% | 96.84% | 95.33% |
| 10-Fold CV RF Optimal | 90.89% ± 2.08% | — | — | — |

**Journal of Informatics Management and Information Technology**
ISSN 2774-4744 (Media Online)
Vol 6, No 2, April 2026 | Hal 131-142
DOI: 10.47065/jimat.v6i2.950
https://hostjournals.com/jimat

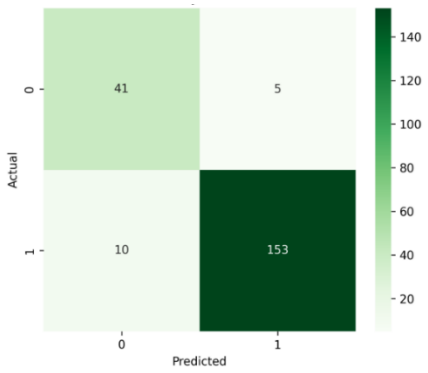
Model RF Default + SMOTE sebagai model yang diusulkan (*proposed*) menghasilkan performa terbaik dengan akurasi 93,78%, recall 95,09%, precision 96,88%, dan F1-Score 95,98%. Dibandingkan dengan model baseline Naïve Bayes pada penelitian ini, terjadi peningkatan akurasi sebesar 3,35 poin, disertai peningkatan recall sebesar 4,29 poin dan F1-Score sebesar 2,31 poin, sementara nilai precision tetap berada pada tingkat yang sangat tinggi. Hasil tersebut menunjukkan bahwa integrasi feature engineering, Pearson Correlation-based Feature Filtering (PCFF), dan SMOTE mampu menghasilkan representasi data yang lebih informatif sehingga Random Forest dapat mempelajari karakteristik kedua kelas secara lebih optimal. Konsistensi peningkatan pada seluruh metrik evaluasi mengindikasikan bahwa model tidak hanya meningkatkan kemampuan klasifikasi secara umum, tetapi juga menjaga keseimbangan antara ketepatan prediksi dan kemampuan mendeteksi siswa yang berisiko mengalami kegagalan akademik.

Gambar 5 menunjukkan confusion matrix model RF Default + SMOTE dengan 155 TP, 41 TN, 5 FP, dan 8 FN. Dibandingkan dengan baseline yang menghasilkan 15 FN, RF Default + SMOTE berhasil menekan jumlah tersebut menjadi 8, atau mengalami penurunan sebesar 46,7% dalam kasus siswa berisiko yang tidak terdeteksi. Selain meningkatkan recall, jumlah False Positive tetap berada pada nilai yang sama (5 kasus), sehingga peningkatan kemampuan mendeteksi kelas minoritas tidak diikuti oleh peningkatan kesalahan klasifikasi pada kelas mayoritas. Temuan ini menunjukkan bahwa penerapan SMOTE mampu memperbaiki distribusi data pelatihan secara efektif dan memberikan kontribusi positif terhadap kemampuan model dalam mengenali siswa berisiko tanpa mengorbankan tingkat precision yang telah tinggi.

Sebagai eksplorasi tambahan, dilakukan pengujian dengan konfigurasi hyperparameter hasil RandomizedSearchCV (RF + SMOTE + Tuning) yang menghasilkan akurasi 92,82%, recall 93,87%, precision 96,84%, dan F1-Score 95,33%. Gambar 6 menunjukkan confusion matrix-nya: 153 TP, 41 TN, 5 FP, dan 10 FN. Hasil ini menunjukkan bahwa konfigurasi tuning menghasilkan akurasi lebih rendah 0,96 poin dibandingkan konfigurasi default, dengan jumlah FN yang juga lebih tinggi (10 vs 8). Temuan ini mengkonfirmasi bahwa parameter default RF pada scikit-learn yang telah dioptimalkan secara empiris sudah cukup efektif pada dataset berukuran menengah, sehingga hyperparameter tuning tidak memberikan peningkatan performa tambahan. Nilai AUC-ROC sebesar 0,978 pada model tuning tetap mengkonfirmasi kemampuan diskriminasi model yang sangat tinggi secara global. Kurva ROC (Gambar 8) memperlihatkan kedua model RF berada konsisten lebih tinggi dari NB Baseline di seluruh rentang FPR.




**Gambar 5.** Confusion Matrix RF Default + SMOTE (Accuracy: 93.78%)




**Gambar 6.** Confusion Matrix RF Optimal (SMOTE + Tuning) (Accuracy: 92.82%)

**Tabel 6.** Laporan Klasifikasi Model RF + SMOTE + Tuning (Eksplorasi)

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Fail (0) | 0.80 | 0.89 | 0.85 | 46 |
| Pass (1) | 0.97 | 0.94 | 0.95 | 163 |
| Macro Average | 0.89 | 0.91 | 0.90 | 209 |
| Weighted Average | 0.93 | 0.93 | 0.93 | 209 |

Tabel 6 mengungkapkan perbedaan performa antar kelas. Kelas Fail memperoleh F1-Score 0,85 dengan recall 0,89 (support=46), mengindikasikan model berhasil mendeteksi 41 dari 46 siswa berisiko. Kelas Pass mencapai F1-Score 0,95

**Journal of Informatics Management and Information Technology**
ISSN 2774-4744 (Media Online)
Vol 6, No 2, April 2026 | Hal 131-142
DOI: 10.47065/jimat.v6i2.950
https://hostjournals.com/jimat

dengan recall 0,94 (support=163). Nilai weighted average F1-Score sebesar 0,93 konsisten dengan akurasi keseluruhan 92,82%.

### 3.1.5 Feature Importance Random Forest

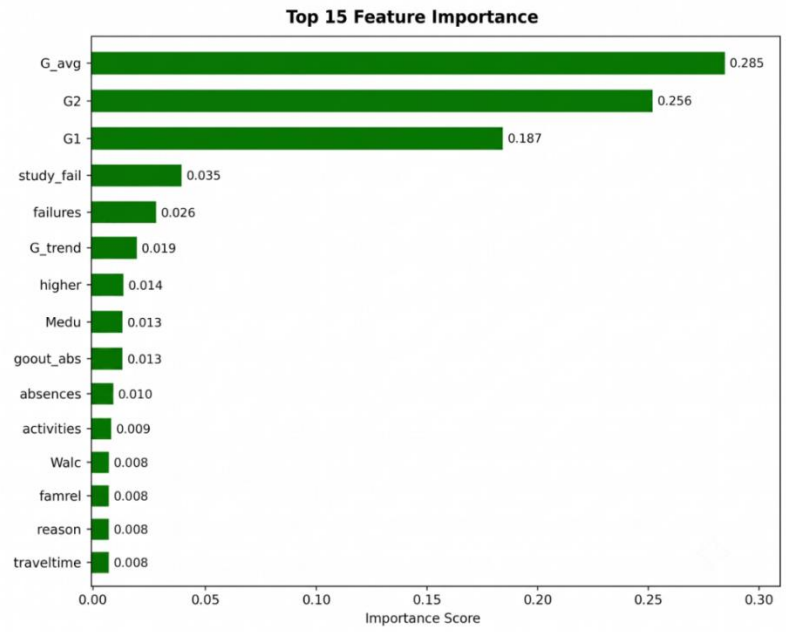
Analisis feature importance berbasis Mean Decrease Impurity (MDI) pada model RF disajikan pada Tabel 7. Gambar 7 memvisualisasikan distribusi skor kepentingan seluruh fitur.

**Tabel 7.** Fitur Importance Model Random Forest (PCFF = Top 15)

| Rank | Fitur | Importance Score | Deskripsi |
|---|---|---|---|
| 1 | G_avg | 0.285 | Rata-rata nilai G1+G2 fitur rekayasa (tertinggi) |
| 2 | G2 | 0.256 | Nilai semester kedua |
| 3 | G1 | 0.187 | Nilai semester pertama |
| 4 | study_fail | 0.035 | Efisiensi belajar relatif fitur rekayasa |
| 5 | failures | 0.026 | Riwayat kegagalan kelas sebelumnya |
| 6 | G_trend | 0.019 | Tren perkembangan nilai fitur rekayasa |
| 7 | higher | 0.014 | Keinginan melanjutkan pendidikan tinggi |
| 8 | Medu | 0.013 | Tingkat pendidikan ibu |
| 9 | goout_abs | 0.013 | Risiko sosial-akademik fitur rekayasa |
| 10 | absences | 0.010 | Jumlah ketidakhadiran di sekolah |
| 11 | activities | 0.009 | Partisipasi dalam kegiatan ekstrakurikuler |
| 12 | Walc | 0.008 | Konsumsi alkohol pada akhir pekan |
| 13 | famrel | 0.008 | Kualitas hubungan keluarga |
| 14 | reason | 0.008 | Alasan memilih sekolah |
| 15 | traveltime | 0.008 | Waktu perjalanan dari rumah ke sekolah |

Temuan paling menonjol adalah fitur rekayasa G_avg menempati posisi pertama dengan skor 0,285, melampaui fitur asli G2 (0,256) dan G1 (0,187). Tiga fitur teratas secara kolektif menyumbang 72,8% dari total feature importance, mengkonfirmasi dominasi rekam jejak nilai sebagai sinyal prediktif. Fitur rekayasa study_fail (0,035) menempati posisi keempat melampaui failures mentah (0,026) mendemonstrasikan bahwa representasi efisiensi belajar memberikan informasi lebih kaya dibandingkan atribut kegagalan tunggal. Perlu dicatat bahwa feature importance dihitung dari seluruh fitur model RF Optimal, sehingga beberapa fitur di luar 15 fitur PCFF (seperti activities, Walc, famrel, reason, traveltime) turut muncul dengan kontribusi kecil (<0,010).

Secara keseluruhan, hasil feature importance menunjukkan bahwa model Random Forest memberikan bobot terbesar pada indikator yang berkaitan langsung dengan performa akademik siswa. Dominasi fitur G_avg, G2, dan G1 menegaskan bahwa riwayat nilai sebelumnya merupakan faktor utama dalam proses prediksi, sementara fitur non-akademik berperan sebagai faktor pendukung yang membantu model meningkatkan ketepatan klasifikasi. Temuan ini sejalan dengan karakteristik dataset yang didominasi oleh variabel akademik sebagai penentu keberhasilan belajar siswa. Selain itu, hasil tersebut menunjukkan bahwa model mampu memanfaatkan kombinasi berbagai fitur untuk menghasilkan prediksi yang lebih stabil dan akurat.

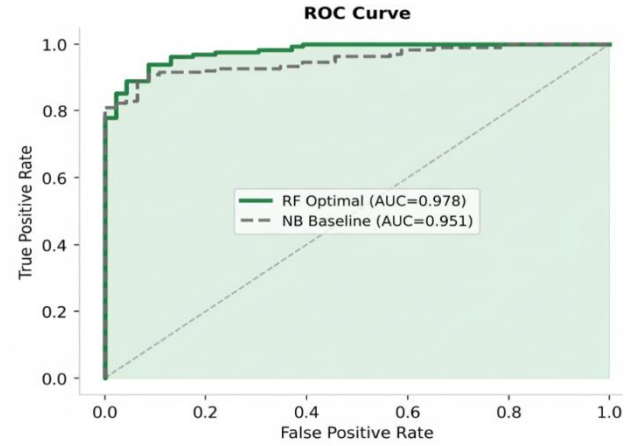



**Gambar 7.** Feature Importance RF Optimal (Top 15, MDI-based)

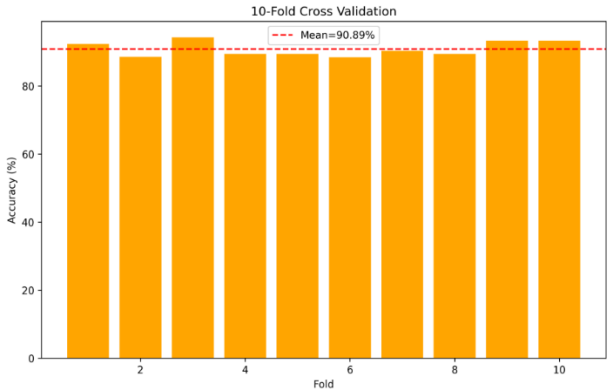
**Journal of Informatics Management and Information Technology**
ISSN 2774-4744 (Media Online)
Vol 6, No 2, April 2026 | Hal 131-142
DOI: 10.47065/jimat.v6i2.950
https://hostjournals.com/jimat

### 3.1.6 Hasil Evaluasi 10-Fold Stratified Cross-Validation

Evaluasi 10-Fold Stratified Cross-Validation dengan SMOTE diterapkan secara independen per fold menghasilkan akurasi rata-rata 90,89% dengan standar deviasi ±2,08%. Gambar 9 menunjukkan variasi akurasi lintas 10 fold, dengan beberapa fold mencapai di atas 93% dan beberapa di bawah 89%, karakteristik normal untuk dataset 1.043 entri dengan 10 partisi. Variasi tersebut menunjukkan bahwa model mampu mempertahankan performa yang relatif stabil meskipun setiap fold memiliki komposisi data pelatihan dan pengujian yang berbeda. Standar deviasi yang rendah juga mengindikasikan bahwa hasil prediksi tidak dipengaruhi secara signifikan oleh pemilihan subset data tertentu, sehingga proses pelatihan menghasilkan performa yang konsisten pada berbagai skenario pembagian data.




**Gambar 8**. ROC Curve RF Optimal (AUC=0.978) vs NB Baseline (AUC=0.951)




**Gambar 9.** 10-Fold Cross-Validation Accuracy (SMOTE applied per fold; Mean=90.89%)

Nilai rata-rata CV 90,89% berada di bawah akurasi *holdout testing* RF Default + SMOTE sebesar 93,78%, membentuk *optimism gap* sebesar 2,89 poin persentase. Selisih ini berada dalam satu standar deviasi (±2,08%), sehingga tidak mengindikasikan *overfitting* yang signifikan. Nilai AUC-ROC 0,978 dan AUC NB baseline 0,951 mengkonfirmasi superioritas kemampuan diskriminasi model RF. Secara keseluruhan, hasil 10-Fold CV mengkonfirmasi bahwa model RF Default + SMOTE memiliki kemampuan generalisasi yang andal dan konsisten, menjadikannya kandidat yang layak untuk diimplementasikan dalam sistem prediksi akademik berbasis data nyata. Selain menunjukkan performa yang tinggi pada data uji, konsistensi hasil antar fold memperlihatkan bahwa model mampu beradaptasi terhadap variasi distribusi data tanpa mengalami penurunan kinerja yang berarti. Temuan ini semakin memperkuat bahwa kombinasi feature engineering, PCFF, dan SMOTE memberikan kontribusi positif dalam membangun model prediksi yang stabil, sehingga berpotensi mendukung proses identifikasi dini siswa berisiko secara lebih akurat dan berkelanjutan.

## 3.2 Pembahasan

### 3.2.1 Analisis Performa dan Konvergensi Model

Hasil eksperimen mengkonfirmasi bahwa integrasi SMOTE dan feature engineering secara substansial meningkatkan performa prediksi dibandingkan baseline Naïve Bayes. Model RF Default + SMOTE sebagai model yang diusulkan mencapai akurasi 93,78%, melampaui baseline sebesar 3,35 poin dan melampaui target 93%. Peningkatan F1-Score dari 93,67% (baseline) menjadi 95,98% (RF Default + SMOTE) mengindikasikan perbaikan seimbang pada precision dan recall secara bersamaan pencapaian yang sulit tanpa penanganan class imbalance.

Sebagai eksplorasi tambahan, konfigurasi hyperparameter tuning menghasilkan akurasi 92,82% lebih rendah 0,96 poin dari konfigurasi default. Hal ini menunjukkan bahwa parameter default RF pada scikit-learn yang telah dioptimalkan

ISSN 2774-4744 (Media Online)
Vol 6, No 2, April 2026 | Hal 131-142
DOI: 10.47065/jimat.v6i2.950
https://hostjournals.com/jimat

secara empiris sudah cukup efektif pada dataset berukuran menengah. Perbedaan ini dapat terjadi karena proses hyperparameter tuning mengoptimalkan performa berdasarkan cross-validation, sehingga model yang dihasilkan lebih berorientasi pada kemampuan generalisasi daripada memaksimalkan akurasi pada satu pembagian data uji tertentu. Temuan ini memperkuat bahwa kontribusi utama peningkatan performa dalam penelitian ini berasal dari kombinasi SMOTE dan feature engineering, bukan dari optimasi hyperparameter.

### 3.2.2 Analisis Pengaruh SMOTE dan Feature Engineering

Efektivitas SMOTE terkuantifikasi melalui penurunan False Negative: dari 15 (baseline) menjadi 8 (RF Default + SMOTE), merepresentasikan peningkatan 46,7% dalam minority class detection. Nilai recall kelas Fail sebesar 0,89 mengindikasikan model berhasil mendeteksi 41 dari 46 siswa berisiko. Temuan ini konsisten dengan Desnelita dkk. yang mendemonstrasikan bahwa SMOTE secara signifikan meningkatkan recall kelas minoritas pada model RF tanpa mengorbankan precision secara berlebihan.

Kontribusi feature engineering terkonfirmasi secara kuantitatif: G_avg menempati posisi pertama dalam feature importance (0,285), melampaui fitur asli G2 (0,256). Ini menunjukkan rata-rata nilai dua semester menjadi representasi lebih stabil dan informatif dibandingkan nilai tunggal. Fitur study_fail menempati posisi keempat (0,035) di atas failures mentah (0,026) mendemonstrasikan bahwa rasio efisiensi belajar memberikan informasi prediktif yang tidak dapat ditangkap oleh atribut kegagalan secara terpisah.

### 3.2.3 Analisis Generalisasi Model

Estimasi generalisasi dari 10-Fold CV (90,89% ± 2,08%) lebih konservatif dari akurasi holdout (93,78% / 92,82%), membentuk optimism gap 1,93–2,89 poin dalam batas satu standar deviasi tidak mengindikasikan overfitting signifikan. AUC-ROC 0,978 merupakan metrik generalisasi yang lebih informatif pada kondisi class imbalance karena independen dari proporsi kelas. Trade-off antara akurasi point estimate dan stabilitas CV harus dipahami sebagai konsekuensi wajar dari keterbatasan ukuran dataset, bukan sebagai indikasi masalah metodologis.

### 3.2.4 Implikasi Penelitian

Dari perspektif praktis, recall kelas Fail sebesar 0,89 mengindikasikan sistem berbasis model ini mampu mengidentifikasi 41 dari 46 siswa berisiko dalam set pengujian. Delapan False Negative yang tersisa dapat ditangani dengan penyesuaian classification threshold mengeksploitasi AUC-ROC tinggi (0,978) untuk mengoptimalkan sensitivitas sesuai kebutuhan institusi. Dominasi fitur nilai (G_avg, G2, G1) dalam feature importance menegaskan pentingnya pemantauan nilai secara berkelanjutan sepanjang semester. Fitur study_fail mengindikasikan siswa dengan riwayat kegagalan tinggi relatif terhadap waktu belajarnya memerlukan perhatian prioritas. Penelitian ini juga mencontohkan prosedur SMOTE yang metodologis benar diterapkan per fold CV yang seharusnya menjadi standar dalam penelitian EDM berbasis class imbalance.

## 4. KESIMPULAN

Penelitian ini mengusulkan dan mengevaluasi pipeline yang mengintegrasikan feature engineering, Pearson Correlation-based Feature Filtering (PCFF), SMOTE, dan Random Forest untuk memprediksi kinerja akademik siswa. Berdasarkan dataset Portuguese Student Performance sebanyak 1.043 entri (Pass=814, Fail=230), model RF Default + SMOTE yang diusulkan mencapai akurasi tertinggi sebesar 93,78% dengan recall 95,09% dan F1-Score 95,98%, melampaui penelitian rujukan Naïve Bayes + CFS [23] sebesar 2,56 poin persentase. Stabilitas model dikonfirmasi melalui evaluasi 10-Fold Stratified Cross-Validation yang menghasilkan akurasi rata-rata 90,89% ± 2,08%, dengan optimism gap yang masih berada dalam batas satu standar deviasi sehingga tidak mengindikasikan overfitting yang signifikan. Penerapan SMOTE secara terisolasi pada data latih terbukti efektif menurunkan jumlah False Negative dari 15 pada model baseline menjadi 8 pada model yang diusulkan, merepresentasikan peningkatan sebesar 46,7% dalam kemampuan mendeteksi siswa berisiko gagal. Selain itu, fitur rekayasa G_avg menempati posisi pertama dalam analisis feature importance dengan skor 0,285, mengkonfirmasi bahwa konstruksi fitur berbasis domain knowledge memberikan nilai tambah signifikan dibandingkan penggunaan atribut mentah secara langsung. Sebagai analisis komparatif, hyperparameter tuning melalui RandomizedSearchCV justru menghasilkan akurasi yang sedikit lebih rendah (92,82%) dibandingkan konfigurasi default, sehingga dapat disimpulkan bahwa kontribusi utama peningkatan performa pada penelitian ini berasal dari kombinasi SMOTE dan feature engineering, bukan dari optimasi hyperparameter. Penelitian ini memiliki sejumlah keterbatasan yang perlu dicermati. Pertama, dataset yang digunakan bersumber dari dua sekolah menengah di Portugal sehingga generalisasi temuan pada konteks pendidikan dengan karakteristik demografis dan kurikulum yang berbeda, termasuk konteks pendidikan di Indonesia, masih memerlukan validasi lebih lanjut. Kedua, ukuran dataset yang tergolong menengah (1.043 entri) berpotensi memengaruhi stabilitas estimasi performa model pada skenario data yang lebih besar atau lebih kompleks. Berdasarkan keterbatasan tersebut, penelitian mendatang disarankan untuk mengeksplorasi metode oversampling lanjutan seperti ADASYN atau Borderline-SMOTE, menerapkan algoritma boosting seperti XGBoost dan LightGBM sebagai pembanding performa, serta menguji validitas pipeline yang diusulkan pada dataset pendidikan dari konteks dan negara yang berbeda. Implementasi model dalam bentuk sistem peringatan dini berbasis web juga disarankan

ISSN 2774-4744 (Media Online)
Vol 6, No 2, April 2026 | Hal 131-142
DOI: 10.47065/jimat.v6i2.950
https://hostjournals.com/jimat

agar temuan penelitian ini dapat dimanfaatkan secara praktis oleh institusi pendidikan dalam mendukung intervensi pedagogis yang lebih tepat sasaran dan tepat waktu.

# REFERENCES

[1] K. P. Pengkomputeran, I. Media, and T. Mara, "Student Dropout Prediction Using Random Forest and XGBoost Method," *INTENSIF*, vol. 9, no. 1, pp. 147–157, 2025, doi: DOI: https://doi.org/10.29407/intensif.v9i1.21191.

[2] R. A. Azizah, F. A. Bachtiar, S. Adinugroho, U. Brawijaya, and P. Korespondensi, "Klasifikasi Kinerja Akademik Siswa Menggunakan Neighbor Weighted K-Nearest Neighbor Dengan Seleksi Fitur Information Classification Of Student Academic Performance Using Neighbor Weighted K-Nearest Neighbor With Information Gain As," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 3, pp. 605–614, 2022, doi: 10.25126/jtiik.202295751.

[3] I. Riadi *et al.*, "Prediksi Kelulusan Tepat Waktu Berdasarkan Riwayat Akademik Menggunakan Metode K-Nearest Neighbor Predicted Timely Graduation Based On Academic History Before College With K-Nearest Neighbor Method," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 2, pp. 249–256, 2024, doi: 10.25126/jtiik.20241127330.

[4] E. P. Saputra *et al.*, "Komparasi Machine Learning Berbasis Pso Untuk Prediksi Tingkat Keberhasilan Belajar Berbasis E-Learning Comparation Of Pso-Based Learning Machine For E-Learning-Based," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 2, pp. 321–328, 2023, doi: 10.25126/jtiik.2023106469.

[5] U. Kristen, K. Wacana, and J. Barat, "Prediksi Mahasiswa Drop-Out Di Universitas Xyz Prediction Of Student Drop-Out At Xyz University," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 6, pp. 1345–1350, 2024, doi: 10.25126/jtiik.2024118689.

[6] C. C. F. Selection, "Comparing Correlation-Based Feature Selection and Symmetrical," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 158, pp. 542–554, 2026, doi: DOI: https://doi.org/10.29207/resti.v8i4.5911.

[7] F. A. Putra, S. Mirajdandi, B. Okmarizal, and S. Mulyanda, "Prediksi Dropout Mahasiswa : Early-Warning Berbasis Enrollment dengan Machine Learning," *J. FASILKOM*, vol. 15, no. 3, pp. 465–473, 2025, doi: DOI: https://doi.org/10.37859/jf.v15i3.10714.

[8] S. Bahri and D. M. Midyanti, "Penerapan Metode K-Medoids Untuk Pengelompokan Mahasiswa Application Of K-Medoids Method For Dropout," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 1, pp. 165–172, 2023, doi: 10.25126/jtiik.2023106643.

[9] M. H. Ariansyah, E. N. Fitri, and S. Winarno, "Improving Performance Of Students ' Grade Classification Model Uses Naïve Bayes Gaussian Tuning Model And Feature Selection," *J. Tek. Inform.*, vol. 4, no. 3, pp. 493–501, 2023, doi: DOI: https://doi.org/10.52436/1.jutif.2023.4.3.737.

[10] D. A. Rahmawati, N. A. Ibadurrahman, J. Zeniarja, N. Hendriyanto, I. Engineering, and I. Engineering, "Implementation Of The Random Forest Algorithm In Classifying The Accuracy Of Graduation Time For Computer Engineering," *J. Tek. Inform.*, vol. 4, no. 3, pp. 565–572, 2023, doi: DOI: https://doi.org/10.52436/1.jutif.2023.4.3.920.

[11] S. Gesti, A. Utami, H. Setiadi, and A. Rohmadi, "Comparative Analysis of Machine Learning Algorithms with RFE-CV for Student Dropout Prediction," *J. Tek. Inform.*, vol. 6, no. 3, pp. 1319–1338, 2025, doi: DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4695.

[12] Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang Impact of SMOTE on Random Forest Classifier Performance based on Imbalanced Data," *Matrik J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, 2022, doi: 10.30812/matrik.v21i3.1726.

[13] M. Ardianti, O. D. Nurhayati, and B. Warsito, "Model Prediksi Kinerja Siswa Berdasarkan Data Log LMS Menggunakan Ensemble Machine Learning," *J. Sains dan Teknol.*, vol. 12, no. 3, pp. 562–571, 2023, doi: https://doi.org/10.23887/jstundiksha.v12i3.59816.

[14] P. Sejati *et al.*, "Studi Komparasi Naive Bayes , K-Nearest Neighbor , Dan Random Forest Untuk Prediksi Calon Mahasiswa Yang Diterima Atau Comparative Study Of Naive Bayes , K-Nearest Neighbor , And Random Forest For The Prediction Of Prospective Students," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 7, pp. 1341–1348, 2022, doi: 10.25126/jtiik.202296737.

[15] S. Sza *et al.*, "Penerapan Decision Tree Dan Random Forest Dalam Deteksi The Application Of Decision Tree And Random Forest In Detecting Human Stress Levels Based On Sleep Conditions," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 5, pp. 1043–1050, 2024, doi: 10.25126/jtiik.2024117993.

[16] N. F. Mahing, A. L. Gunawan, A. Foresta, and A. Zen, "Klasifikasi Tingkat Stres Dari Data Berbentuk Teks Dengan Menggunakan Algoritma Support Vector Machine ( Svm ) Dan Classification Of Stress Levels From Text Data Using Support Vector Machine ( Svm ) Algorithm And Random Forest," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 5, 2024, doi: 10.25126/jtiik2024118010.

[17] A. N. Pratiwi *et al.*, "Prediksi Kinerja Akademik Matematika Siswa berdasarkan Kepribadian Big Five menggunakan Random Forest dengan Teknik Synthetic Minority Over - Sampling Personality Traits using Random Forest with Synthetic Minority Over - Sampling," *Sist. J. Sist. Inf.*, vol. 14, no. 2, pp. 985–1000, 2025, doi: DOI: https://doi.org/10.32520/stmsi.v14i2.5102.

[18] A. A. Yaqin, M. A. Barata, and N. Mahmudah, "Implementation of the Random Forest Algorithm with Optuna Optimization in Lung Cancer Classification," *Sist. J. Sist. Inf.*, vol. 14, no. 2, pp. 561–569, 2025, doi: DOI: https://doi.org/10.32520/stmsi.v14i2.4877.

[19] M. P. Pulungan, A. Purnomo, A. Kurniasih, P. Korespondensi, I. Class, and S. M. O. Technique, "Penerapan Smote Untuk Mengatasi Imbalance Class Dalam Klasifikasi Kepribadian Mbti Menggunakan Naive Bayes Application Of Smote To Overcome Class Imbalance In The Mbti Personality Classification Using The Naïve Bayes Classifier," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 5, pp. 1033–1042, 2024, doi: 10.25126/jtiik.2024117989.

[20] U. Bumigora and P. Korespondensi, "Peningkatan Kinerja Metode Svm Menggunakan Metode Knn Imputasi Dan K-Means-Smote Untuk Klasifikasi Kelulusan Improvement Performance Of Svm Method Using Knn Imputation And K-Means-Smote Method For Graduation Classification Of," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 4, pp. 713–718, 2021, doi: 10.25126/jtiik.202183428.

[21] R. Artikel, R. A. Nurdian, M. Ridwan, and A. Yusuf, "Komparasi Metode SMOTE dan ADASYN dalam Meningkatkan Performa Klasifikasi Herregistrasi Mahasiswa Baru," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. April, pp. 24–32, 2022, doi: http://dx.doi.org/10.28932/jutisi.v8i1.4004 Riwayat.

[22] N. N. Sholihah and A. Hermawan, "Implementation Of Random Forest And Smote Methods For Economic Status Classification In Cirebon City," *J. Tek. Inform.*, vol. 4, no. 6, pp. 1387–1397, 2023, doi: DOI: https://doi.org/10.52436/1.jutif.2023.4.6.1135.

ISSN 2774-4744 (Media Online)
Vol 6, No 2, April 2026 | Hal 131-142
DOI: 10.47065/jimat.v6i2.950
https://hostjournals.com/jimat

[23] T. Gori *et al.*, "Preprocessing Data Dan Klasifikasi Untuk Prediksi Kinerja Data Preprocessing And Classification For Predicting Student," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.
[24] D. S. Bhakti, A. Prasetyo, P. Arsi, C. S. Faculty, and U. A. Purwokerto, "Implementation Of Hyperparameter Tuning In Random Forest Implementasi Hyperparameter Tuning Pada Algoritma Random," *J. Tek. Inform.*, vol. 5, no. 4, pp. 63–69, 2024, doi: DOI: https://doi.org/10.52436/1.jutif.2024.5.4.2032.
[25] F. Nuraeni, D. Kurniadi, M. H. Diazki, K. Garut, P. Korespondensi, and K. Karyawan, "Algoritma K-Nearest Neighbor Pada Kasus Dataset Imbalanced K-Nearest Neighbor Algorithm And Smote Method," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 3, pp. 557–568, 2024, doi: 10.25126/jtiik.938144.
[26] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed. Sebastopol, CA: O'Reilly Media, Inc., 2022.
[27] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed. New York, NY: Springer, 2021.