Jurnal Riset Aplikasi Ilmu Data dan Sistem 1 (1), 117–131 DOI: 0000
Juli 2026

Komparasi Metode *Random Forest*, *XGBoost*, dan *CatBoost* dalam Klasifikasi Rumpun Program Studi Berdasarkan Minat dan Nilai Akademik

Siti Aida Hanun[1], Atik Wintarti[1], Riskyana Dewi Intan Puspitasari[2,*]

[1]*Program Studi Sarjana Sains Data, Universitas Negeri Surabaya, Indonesia*
[2]*Program Studi Sarjana Kecerdasan Artifisial, Universitas Negeri Surabaya, Indonesia*

Abstrak

Pemilihan program studi yang tidak sesuai dengan minat dan kemampuan akademik dapat meningkatkan risiko salah jurusan pada mahasiswa. Penelitian ini bertujuan membandingkan performa algoritma *Random Forest*, *XGBoost*, dan *CatBoost* dalam klasifikasi subrumpun program studi berdasarkan minat dan nilai akademik siswa. Data yang digunakan merupakan data penerimaan mahasiswa baru jalur Seleksi Nasional Berdasarkan Prestasi (SNBP) Universitas Negeri Surabaya tahun 2024 yang meliputi nilai rapor dan preferensi program studi peserta. Tahapan penelitian meliputi preprocessing, *Feature Engineering*, encoding, dan penanganan ketidakseimbangan kelas menggunakan *SMOTE*. Evaluasi model dilakukan menggunakan *accuracy*, *macro precision*, *macro recall*, dan *Macro F1-score*. Hasil penelitian menunjukkan bahwa *XGBoost* memiliki performa terbaik dibandingkan *Random Forest* dan *CatBoost*. Pada kelompok IPA, *XGBoost* memperoleh *accuracy* sebesar 78,64% dan *Macro F1-score* sebesar 76,81%. Sementara itu, pada kelompok IPS, performa terbaik diperoleh menggunakan 25% fitur terpenting hasil *feature selection* dengan *accuracy* sebesar 76,35% dan *Macro F1-score* sebesar 69,67%. Hasil analisis *feature importance* menunjukkan bahwa fitur minat siswa menjadi faktor paling dominan dalam proses klasifikasi. Penelitian ini menunjukkan bahwa metode *gradient boosting* mampu memberikan performa yang baik pada klasifikasi data multikelas dan tidak seimbang berbasis data akademik siswa.

Kata Kunci *Klasifikasi Subrumpun Program Studi; Random Forest; XGBoost; CatBoost; SMOTE*

Abstract

Choosing a major that does not align with a student's interests and academic abilities can increase the risk of students ending up in the wrong major. This study aims to compare the performance of the *Random Forest*, *XGBoost*, and *CatBoost* algorithms in classifying sub-groups of majors based on students' interests and academic grades. The data used were obtained from participants of the 2024 National Selection Based on Achievement (SNBP) admission process at Surabaya State University (Universitas Negeri Surabaya), including report card grades and study program preferences. The research stages included preprocessing, *Feature Engineering*, encoding, and handling class imbalance using *SMOTE*. Model evaluation was conducted using accuracy, macro precision, macro recall, and *Macro F1-score*. The results showed that *XGBoost* performed best compared to *Random Forest* and *CatBoost*. In the Science (IPA) group, *XGBoost*

*Corresponding author. Email: atikwintarti@unesa.ac.id or riskyanapuspitasari@unesa.ac.id.

Received 3 Juli 2026; Revised 13 Juli 2026; Accepted 13 Juli 2026

achieved an accuracy of 78.64% and a *Macro F1-score* of 76.81%, while in the Social Sciences (IPS) group, the best performance was obtained using the top 25% most important features from the feature selection process, achieving an accuracy of 76.35% and a *Macro F1-score* of 69,67%. The feature importance analysis revealed that student interest was the most dominant factor in the classification process. This study demonstrates that gradient boosting methods are capable of delivering good performance in the classification of multi-class and imbalanced data based on student academic records.

Keywords *Program cluster classification; Program cluster classification; Random Forest; SMOTE; XGBoost; CatBoost*

1 Pendahuluan

Pemilihan program studi merupakan keputusan krusial yang memengaruhi masa depan akademik dan karier siswa (Gillis and Ryberg, 2021). Fenomena *mismatch major* atau salah jurusan masih sering terjadi, di mana survei Indonesia *Career Center Network* (ICCN) 2020 menunjukkan sekitar 87% mahasiswa di Indonesia merasa salah jurusan (Felldy Aslya Utama, 2020). Hal ini berdampak negatif pada penurunan motivasi, kesulitan akademik, hingga keterlambatan kelulusan (Milsom and Coughlin, 2015). Permasalahan ini kian relevan dalam Seleksi Nasional Berdasarkan Prestasi (SNBP) yang berbasis rekam akademik, mengingat pendaftar SNBP 2025 meningkat 10,6% mencapai 776.515 siswa, sehingga strategi pemilihan prodi yang tepat sangat diperlukan untuk meningkatkan peluang kelulusan (SNPMB, 2025).

Perkembangan *Artificial Intelligence* (AI) dan *Educational Data Mining* (EDM) membuka peluang analisis data akademik secara objektif (Romero and Ventura, 2020). Melalui pendekatan *machine learning*, pola nilai akademik dan preferensi siswa dapat dianalisis untuk mengidentifikasi kecenderungan bidang studi secara lebih akurat (Guevara-Reyes et al., 2025). Salah satu metode klasifikasi yang unggul adalah *ensemble learning*, yang mengombinasikan beberapa model prediktif melalui pendekatan *bagging* atau *boosting* (Zhang et al., 2021). *Random Forest* (*bagging*) dikenal stabil dan mampu mengurangi *overfitting* (Breiman, 2001). Sementara itu, *XGBoost* menyediakan efisiensi komputasi dan regularisasi yang efektif (Chen and Guestrin, 2016), dan *CatBoost* dirancang optimal untuk menangani fitur kategorikal melalui *ordered boosting* (Prokhorenkova et al., 2019).

Beberapa penelitian terdahulu telah menerapkan *data mining* dalam ranah ini. Mystere et al. (2023) menguji model tunggal dengan akurasi tertinggi pada *SVM* (70%), namun belum mengeksplorasi *ensemble learning* dan fitur minat. Prayoga and Ermatita (2024) menerapkan *Random Forest* dengan akurasi 75% untuk jurusan SMK, tetapi belum membandingkan dengan algoritma *boosting*. Di sisi lain, Alsayed et al. (2021) memanfaatkan *Random Forest*, *XGBoost*, dan *CatBoost* untuk memprediksi keberhasilan akademik, namun tidak berfokus pada klasifikasi subrumpun prodi berdasarkan integrasi minat dan nilai akademik. Oleh karena itu, studi komparatif yang spesifik membandingkan ketiga algoritma *ensemble* tersebut dalam klasifikasi subrumpun prodi masih sangat terbatas.

Selain pemilihan algoritma klasifikasi, tantangan lain yang umum ditemukan pada data pendidikan adalah ketidakseimbangan distribusi kelas *(class imbalance)*. Pada penelitian ini, program studi dikelompokkan ke dalam 31 subrumpun keilmuan dengan jumlah sampel yang tidak merata pada setiap kelas. Kondisi tersebut berpotensi menyebabkan model lebih cenderung mempelajari pola dari kelas mayoritas sehingga performa klasifikasi pada kelas minoritas menjadi kurang optimal. Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan *Synthetic Minor-*

ity Over-sampling Technique (SMOTE), yaitu metode *oversampling* yang menghasilkan sampel sintetis pada kelas minoritas berdasarkan karakteristik data yang ada (Chawla et al., 2002). Penggunaan SMOTE telah banyak digunakan untuk meningkatkan representasi kelas minoritas dan membantu model klasifikasi menghasilkan performa yang lebih seimbang pada data multikelas yang tidak seimbang (Wang et al., 2022). Oleh karena itu, penelitian ini tidak hanya membandingkan performa *Random Forest, XGBoost*, dan *CatBoost*, tetapi juga mengevaluasi pengaruh penerapan SMOTE dalam meningkatkan kemampuan model mengklasifikasikan subrumpun program studi secara lebih akurat.

Penelitian ini bertujuan membandingkan performa *Random Forest*, *XGBoost*, dan *CatBoost* dalam mengklasifikasikan siswa ke dalam subrumpun program studi berdasarkan minat dan nilai akademik. Evaluasi model dilakukan menggunakan metrik *accuracy, macro precision, macro recall, Macro F1-score*, dan *confusion matrix*. Selain itu, penelitian ini juga menganalisis kontribusi masing-masing fitur untuk mengidentifikasi faktor penentu yang paling berpengaruh dalam proses pengelompokan siswa.

2 Metode

2.1 Alur Penelitian

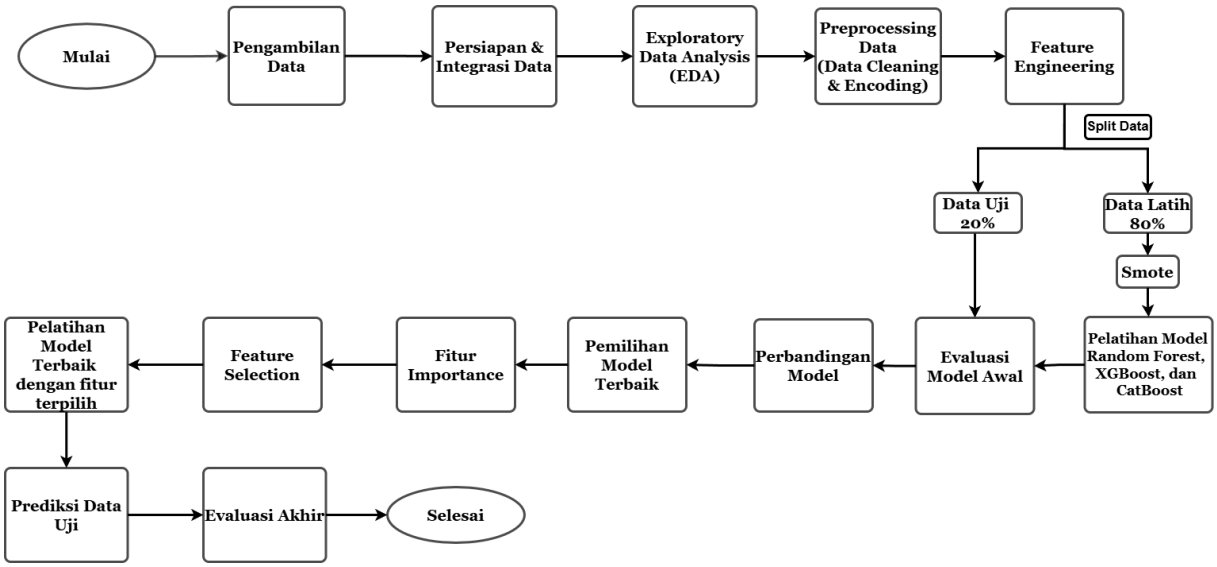

Figure 1: Alur penelitian

Fase awal meliputi pengumpulan, persiapan, integrasi, dan *Exploratory Data Analysis* (EDA). Selanjutnya, dilakukan preprocessing data dan *Feature Engineering*, yang diikuti dengan pembagian data latih dan data uji *(train-test split)*. Untuk mengatasi ketidakseimbangan kelas, teknik *SMOTE* diterapkan khusus pada data latih sebelum tahap pemodelan menggunakan algoritma *Random Forest*, *XGBoost*, dan *CatBoost*. Setelah evaluasi, model dengan performa paling optimal dipilih sebagai model terbaik. Pada tahap akhir, model terbaik tersebut dianalisis menggunakan *Feature Importance* dan *Feature Selection* untuk mengidentifikasi fitur yang paling berpengaruh terhadap hasil klasifikasi.

2.2 Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder peserta jalur SNMPTN/SNBP di Universitas Negeri Surabaya tahun 2024 yang bersumber dari basis data internal kampus. Kumpulan data awal mencakup data nilai akademik sebanyak 1.848.103 baris dari 30.579 peserta unik, data kelulusan (6.262 peserta yang diterima), data peminatan (pilihan program studi 1 dan 2), serta data referensi program studi dan mata pelajaran. Atribut pada data nilai akademik meliputi nomor pendaftaran, semester (1–5), tingkat kelas, kode mata pelajaran, dan nilai skala 100. Sementara itu, data preferensi mencakup kode program studi pilihan beserta nomor urut pilihannya.

2.3 Persiapan dan Integrasi Data

Integrasi dilakukan dengan menggabungkan data nilai pendaftar dan data kelulusan menggunakan nomor pendaftaran sebagai primary key. Kode program studi kelulusan dan kode mata pelajaran kemudian dipetakan ke nama aslinya menggunakan data referensi terkait. Setelah proses penggabungan, setiap peserta direpresentasikan ke dalam satu baris data (tabular format) yang mengintegrasikan nilai akademik semester 1–5, minat prodi, dan status kelulusan.

Untuk menjaga keseragaman fitur akademik, dilakukan penyaringan sampel dengan mengeksklusi siswa yang berasal dari SMK atau MA yang memiliki mata pelajaran tambahan di luar kurikulum reguler. Melalui proses seleksi ini, diperoleh dataset akhir sebanyak 4.797 calon mahasiswa, yang terdiri dari 2.541 siswa jurusan IPA dan 2.256 siswa jurusan IPS dengan struktur 13 mata pelajaran wajib. Sebagai variabel target (label) klasifikasi, penelitian ini mengelompokkan berbagai program studi ke dalam 31 subrumpun keilmuan yang lebih umum berdasarkan acuan Kemendikbudristek dan LLDIKTI. Pengelompokan ini bertujuan untuk mereduksi kompleksitas kelas pada tingkat program studi yang terlalu bervariasi.Rincian beberapa sampel pemetaan program studi ke dalam subrumpun keilmuan disajikan pada Tabel 1:

Table 1: Sampel Pemetaan Program Studi ke Dalam Subrumpun Keilmuan

No	**Subrumpun Keilmuan**	**Program Studi**
1	SUBRUMPUN ILMU IPA	S1 Fisika, S1 Kimia, S1 Biologi, S1 Biosains Hewan
2	SUBRUMPUN MATEMATIKA	S1 Matematika, S1 Sains Data, S1 Sains Aktuaria
3	SUBRUMPUN ILMU KOMPUTER	S1 Teknik Informatika, S1 Sistem Informasi, S1 Teknologi Informasi, S1 Kecerdasan Artifisial, D4 Manajemen Informatika
...	...	...
31	...	... *[dst. hingga 31 Subrumpun]*

2.4 Pra-pemrosesan Data dan Rekayasa Fitur

Tahapan ini bertujuan untuk meningkatkan kualitas, kebersihan, dan konsistensi data agar siap digunakan dalam pemodelan. Seluruh langkah dilaksanakan sebelum pemisahan data (*train-test split*) untuk menjamin keseragaman fitur, kecuali teknik *SMOTE* yang diterapkan khusus pada data latih guna mencegah kebocoran data (*data leakage*).

2.4.1 Penanganan Data Hilang

Sampel peserta yang memiliki *missing values* melebihi ambang batas 50% pada atribut akademik dieliminasi dari penelitian karena dianggap tidak representatif. Sementara itu, untuk fitur numerik yang memiliki data kosong di bawah 50%, diimputasi menggunakan nilai rata-rata (*mean imputation*) dari atribut terkait untuk menjaga distribusi data kontinu dalam pembelajaran mesin Zhang et al. (2021).

2.4.2 Pengodean Fitur Kategorikal *(Feature Encoding)*

Fitur kategorikal pada pilihan 1 dan pilihan 2 program studi digabungkan berdasarkan nomor pendaftaran. Nilai kosong pada pilihan *p*ilihan 1 maupun 2 diisi secara eksplisit dengan label "TIDAK_MEMILIH". Selanjutnya, fitur tersebut ditransformasikan ke bentuk numerik menggunakan metode *Label Encoding* agar menghasilkan representasi angka yang konsisten pada data latih maupun data uji.

2.4.3 Rekayasa Fitur (*Feature Engineering*)

Feature Engineering dilakukan untuk mengekstrak kemampuan akademik siswa melalui pembentukan fitur turunan berbasis perhitungan rata-rata nilai semester 1–5 Zhang et al. (2021). Fitur tambahan yang dibentuk meliputi:

- Rata-rata wajib (Matematika, Bahasa Indonesia, Bahasa Inggris).
- Rata-rata umum (Sejarah, Seni Budaya, Informatika, PPKN, PJOK, Agama, Prakarya).
- Rata-rata jurusan (IPA: Biologi, Fisika, Kimia; IPS: Ekonomi, Geografi, Sosiologi).

Selain itu, dibuat fitur selisih antar-kategori nilai (seperti selisih nilai jurusan dan nilai wajib) untuk menangkap pola kekuatan akademik siswa secara relatif (Zhang et al., 2021). Proses ini aman dilakukan sebelum pembagian data tanpa risiko *data leakage* karena hanya mengandalkan turunan aritmetika sederhana dari data asli tanpa melibatkan transformasi distribusi seperti *scaling* atau *Principal Component Analysis* (PCA).

2.4.4 *SMOTE* (*Synthetic Minority Over-sampling Technique*)

Synthetic Minority Over-sampling Technique (*SMOTE*) digunakan untuk mengatasi ketidakseimbangan kelas (*imbalanced class*) pada distribusi 31 subrumpun program studi. Teknik ini membuat sampel sintetis baru di antara sampel kelas minoritas, bukan melakukan duplikasi acak (Chawla et al., 2002). Proses pembuatan data baru dilakukan dengan mengidentifikasi tetangga terdekat (*k-Nearest Neighbors*) dari sampel kelas minoritas x_i, lalu memilih salah satu tetangga terdekat tersebut (x_{nn}) secara acak melalui persamaan berikut (Chawla et al., 2002):

$$x_{new} = x_i + \lambda(x_{nn} - x_i) \tag{1}$$

di mana x_{new} adalah data sintetis baru, x_i adalah sampel kelas minoritas, x_{nn} merupakan tetangga terdekat dari x_i, dan λ menyatakan bilangan acak pada rentang $[0, 1]$. Untuk menghindari risiko kebocoran data (*data leakage*), *SMOTE* hanya diterapkan pada data latih (*training data*) setelah proses pemisahan data selesai dilakukan (Wang et al., 2022).

2.5 Penerapan Model Klasifikasi

2.5.1 Ensemble Learning

Ensemble learning mengombinasikan beberapa *base learners* untuk meningkatkan generalisasi dan mereduksi *overfitting* (Dietterich, 2000). Pendekatan ini memanfaatkan stabilitas penggabungan prediksi melalui mekanise rata-rata atau *majority voting* (Hastie et al., 2009) Strategi utamanya meliputi *bagging*, *boosting*, dan *stacking* (lavender lili, 2025). Representasi matematis untuk strategi *bagging* dan *boosting* dijabarkan sebagai berikut (Gultom et al., 2026):

1. Bagging (Majority Voting): Melatih beberapa model secara independen dan paralel, di mana kelas prediksi akhir ($\hat{y}$) ditentukan berdasarkan akumulasi suara terbanyak dari B buah model (Gultom et al., 2026):

$$\hat{y} = \arg\max_{c \in C} \sum_{b=1}^{B} I(h_b(x) = c) \tag{2}$$

di mana arg max memilih kelas c dalam himpunan C dengan frekuensi tertinggi, I adalah fungsi indikator, dan $h_b(x)$ merupakan prediksi model ke-b.

2. Boosting: Bekerja secara sekuensial dengan melatih *weak learner* ($h_m(x)$) untuk meminimalkan residu kesalahan model tahap sebelumnya ($F_{m-1}(x)$) melalui skema *stage-wise additive model* (Gultom et al., 2026):

$$F_m(x) = F_{m-1}(x) + \gamma h_m(x) \tag{3}$$

di mana $F_m(x)$ adalah fungsi keputusan kumulatif pada iterasi ke-m, dan γ menyatakan *learning rate* penentu bobot kontribusi model baru.

2.5.2 *Random Forest*

Prediksi akhir pada algoritma *Random Forest* untuk kasus klasifikasi ditentukan melalui mekanisme *majority voting* (pemungutan suara terbanyak) dari seluruh pohon keputusan yang telah dibangun. Berdasarkan standar referensi publikasi ilmiah, penentuan label kelas prediksi akhir ini dapat diperoleh melalui fungsi agregasi indikator sesuai dengan Persamaan 4 (Brabec and Machlica, 2018):

$$y^* = \arg\max_{y \in \mathcal{Y}} \sum_{t=1}^{T} I\big(\mathcal{T}_t(x) = y\big) \tag{4}$$

di mana y^* merupakan label kelas yang diprediksi oleh *forest*, $\mathcal{Y}$ adalah ruang label *output* (himpunan kelas subrumpun keilmuan), T menyatakan jumlah seluruh pohon (*trees*) di dalam *forest*, $\mathcal{T}_t(x)$ adalah fungsi prediksi kelas dari pohon keputusan tunggal ke-t untuk sampel uji x, dan $I(\cdot)$ merupakan fungsi indikator yang bernilai 1 jika kondisi di dalam argumen terpenuhi (benar) dan bernilai 0 jika sebaliknya.

2.5.3 *XGBoost*

XGBoost (*Extreme Gradient Boosting*) merupakan algoritma *boosting* yang membangun pohon keputusan secara bertahap untuk memperbaiki kesalahan model sebelumnya melalui op-

timasi berbasis gradien dan regularisasi (Chen and Guestrin, 2016):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{5}$$

dengan $\hat{y}_i$ merupakan hasil prediksi, K adalah jumlah pohon keputusan, dan $f_k(x_i)$ menyatakan kontribusi pohon ke-k terhadap prediksi data ke-i.

2.5.4 *CatBoost*

CatBoost merupakan algoritma *gradient boosting* yang menggunakan mekanisme *ordered boosting* untuk menangani fitur kategorikal secara efektif serta mengurangi bias prediksi dan risiko *data leakage* (Prokhorenkova et al., 2019):

$$r_i = y_i - M_{\sigma(i)-1}(x_i) \tag{6}$$

dengan r_i merupakan residual pada data ke-i, y_i adalah nilai target aktual, $M_{\sigma(i)-1}$ merupakan model yang dilatih menggunakan data sebelum indeks ke-i dalam urutan permutasi σ, dan x_i adalah fitur data ke-i.

2.6 Evaluasi Model

Evaluasi dilakukan untuk mengetahui sejauh mana model mampu memprediksi kelas multikelas secara tepat pada data pendidikan (Romero and Ventura, 2020). Mengingat dataset memiliki distribusi tidak seimbang pada 31 subrumpun program studi, penggunaan akurasi saja tidak cukup objektif sehingga didukung oleh *Confusion Matrix* dan metrik berbasis *Macro-averaging* guna memberikan bobot penilaian yang setara pada seluruh kelas tanpa terpengaruh dominasi kelas mayoritas (Nurhalizah et al., 2024). Penjelasan metrik evaluasi yang digunakan dijabarkan sebagai berikut:

1. Confusion Matrix & Accuracy: *Confusion Matrix* membandingkan hasil aktual dan prediksi dalam matriks berukuran $n \times n$ untuk menganalisis pola kesalahan model melalui komponen *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) (Zhang et al., 2021). Informasi ini menjadi dasar perhitungan *Accuracy* untuk mengukur proporsi prediksi benar dari keseluruhan sampel (Fathurohman, 2021):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

2. Precision & Recall: *Precision* mengukur ketepatan model dalam mengklasifikasikan data positif guna meminimalkan kesalahan prediksi positif (FP). Sementara itu, *Recall* merepresentasikan kemampuan model mengidentifikasi seluruh sampel positif aktual untuk menekan risiko data yang tidak terdeteksi (FN) (Zhang et al., 2021). Implementasi metrik dasar beserta transformasi *Macro*-nya untuk penilaian global multikelas dirumuskan secara berdampingan sebagai berikut(Zhang et al., 2021):

$$Precision = \frac{TP}{TP + FP} \qquad Macro\ Precision = \frac{1}{K} \sum_{i=1}^{K} Precision_i \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \qquad Macro\ Recall = \frac{1}{K} \sum_{i=1}^{K} Recall_i \tag{9}$$

3. F1-Score & *Macro F1-score*: *F1-score* merupakan rata-rata harmonik yang mengukur keseimbangan antara *Precision* dan *Recall* (Zhang et al., 2021). Skor makro akhirnya dirumuskan sebagai berikut:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad Macro\ F1\text{-}Score = \frac{1}{K}\sum_{i=1}^{K} F1_i \tag{10}$$

di mana K menyatakan total jumlah kelas (31 subrumpun program studi). Perhitungan *Macro Evaluation* ini merata-ratakan skor akhir tiap subrumpun secara langsung sehingga sangat objektif untuk mengevaluasi data yang timpang (Sujon et al., 2025).

2.7 Skenario Pengujian

Eksperimen model klasifikasi dirancang secara terstruktur untuk mengevaluasi pengaruh pra-pemrosesan, rekayasa fitur, dan reduksi dimensi terhadap performa model. Metrik utama yang digunakan sebagai acuan evaluasi pada seluruh tahapan eksperimen adalah *Macro F1-score*. Metrik ini dipilih karena mampu menghitung performa setiap kelas secara terpisah kemudian merata-ratakannya, sehingga kemampuan model pada kelas minoritas tetap diperhatikan tanpa bias dari dominasi kelas mayoritas (Suhaimi et al., 2023). Alur pengujian dibagi menjadi empat tahapan eksperimen utama yang diterapkan secara terpisah pada kelompok data IPA dan IPS:

2.7.1 Eksperimen *Feature Engineering*

Tahap ini bertujuan untuk menguji dampak penambahan fitur-fitur akademik turunan—seperti rata-rata nilai wajib, umum, jurusan, dan selisih (*gap*) nilai—terhadap kemampuan prediktif algoritma. Eksperimen dilakukan dengan membandingkan nilai *Macro F1-score* antara model yang dilatih menggunakan dataset asli dengan model yang menggunakan dataset hasil *Feature Engineering*. Tahapan ini penting untuk menganalisis apakah dimensi data yang bertambah memberikan informasi baru yang relevan atau justru memicu redundansi (*noise*) yang dapat menurunkan kemampuan generalisasi model (Afshar and Usefi, 2022).

2.7.2 Eksperimen *SMOTE*

Eksperimen ini dirancang untuk mengatasi masalah ketidakseimbangan kelas (*multiclass imbalance*) melalui penerapan metode *SMOTE* pada data latih. Pengujian dilakukan dengan menerapkan variasi target *oversampling* guna menemukan proporsi distribusi kelas yang paling optimal dalam membantu algoritma mempelajari pola data. Untuk menjaga objektivitas evaluasi dan merepresentasikan kondisi data riil, proses *SMOTE* dikonfigurasikan secara eksklusif hanya pada data latih (*training data*), sementara data uji (*testing data*) tetap dipertahankan dalam kondisi asli tanpa *oversampling*.

2.7.3 Analisis Feature Importance

Setelah kombinasi model terbaik pada eksperimen sebelumnya ditemukan, dilakukan analisis *Feature Importance* untuk menginterpretasikan keputusan internal model secara ilmiah. Ekstraksi skor kepentingan fitur dilakukan menggunakan algoritma terbaik yang terpilih untuk mengidentifikasi variabel mana saja yang memberikan kontribusi keuntungan (*gain*) terbesar dalam memisahkan batas keputusan antar-subrumpun program studi, baik dari faktor nilai akademik spesifik maupun faktor preferensi minat siswa (Alsahaf et al., 2022).

2.7.4 Eksperimen Feature Selection

Sebagai tahap optimasi terakhir, dilakukan eksperimen *Feature Selection* untuk menguji pengaruh reduksi dimensi terhadap efisiensi dan performa model (Girish Chandrashekar and Ferat Sahin, 2014). Fitur-fitur yang telah diurutkan berdasarkan skor *Feature Importance* tertinggi disaring ke dalam empat skenario subset dimensi, yaitu sebesar 25%, 50%, 75%, dan 100% (seluruh fitur). Model kemudian dilatih ulang menggunakan masing-masing subset tersebut untuk membuktikan secara empiris apakah pemangkasan fitur yang kurang informatif dapat meningkatkan performa atau justru menghilangkan informasi penting yang dibutuhkan model.

3 Hasil dan Pembahasan

3.1 Persiapan Data dan Pra-pemrosesan

Dataset sekunder jalur SNBP Universitas Negeri Surabaya disaring ketat dengan memfokuskan objek pada siswa SMA demi menjaga homogenitas struktur nilai rapor. Dari 6.262 pendaftar awal, proses ini menghasilkan 4.797 data valid yang terbagi menjadi kelompok IPA (2.541 siswa) dan IPS (2.256 siswa) untuk mengklasifikasikan rekomendasi ke dalam 31 subrumpun program studi. Melalui *Exploratory Data Analysis* (EDA), ditemukan masalah *multiclass imbalance* ekstrem antar-subrumpun serta angka *missing value* pada beberapa atribut mata pelajaran yang melebihi 70%. Evaluasi distribusi nilai rapor juga mendeteksi adanya *overlap* spasial yang tinggi pada rentang median 85–92 antar-kelas, yang menegaskan kompleksitas penentuan batas keputusan klasifikasi jika hanya mengandalkan nilai absolut mentah. Langkah pra-pemrosesan data kemudian diimplementasikan secara berurutan sebagai berikut:

1. **Penanganan Missing Value:** Siswa dengan akumulasi data kosong $>$ 50% dieliminasi karena dianggap tidak representatif. Tahap ini mereduksi 27 data kelompok IPA dan 120 data kelompok IPS, menyisakan sampel bersih sebanyak 2.514 siswa IPA dan 2.136 siswa IPS. Atribut kosong di bawah ambang batas 50% diimputasi menggunakan *mean imputation* per mata pelajaran untuk mempertahankan varians asli numerik tanpa memicu bias sektoral. Meskipun terdapat kekosongan massal pada beberapa fitur secara agregat, secara individu tidak ada siswa hasil pemangkasan yang memiliki akumulasi data hilang melebihi *threshold*. Bukti empiris pemotongan distribusi ini ditunjukkan melalui kurva *kernel density estimation* pada Gambar 2, di mana sebaran persentase data hilang per siswa terhenti bersih maksimal pada titik 0,5 (50%).

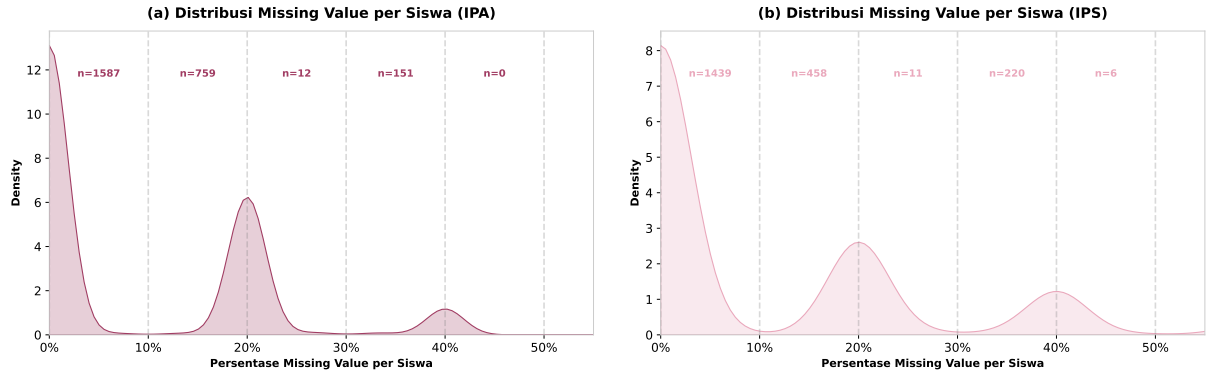

Figure 2: Distribusi Kepadatan Persentase Missing Value Per Siswa Kelompok IPA dan IPS

2. **Encoding Fitur Kategorikal:** Atribut preferensi pilihan minat pertama dan kedua siswa digabungkan menggunakan nomor pendaftaran sebagai kunci utama. Data kosong pada pilihan minat (379 pada minat 1 dan 19.010 pada minat 2) diisi dengan label *"TIDAK_MEMILIH"*. Transformasi ke bentuk numerik diselesaikan menggunakan *label encoding* yang diaplikasikan langsung pada tingkat program studi asli (menghasilkan 108 kategori unik) tanpa pemetaan awal ke subrumpun target guna mencegah risiko kebocoran data (*data leakage*).

3.2 *Feature Engineering*

Feature Engineering (FE) diterapkan untuk mengekstrak informasi performa akademis relatif siswa melalui pembentukan fitur-fitur agregat dan fitur perbandingan selisih (*gap*) antarkelompok mata pelajaran. Struktur hasil rekayasa fitur dirangkum pada Tabel 2.

Table 2: Struktur Komponen Rekayasa Fitur (*Feature Engineering*)

No	Nama Fitur	Deskripsi Atribut	Komponen Nilai Rapor (Semester 1–5)
1	*mean_wajib*	Rata-rata kelompok mata pelajaran wajib	Matematika, Bahasa Indonesia, Bahasa Inggris
2	*mean_umum*	Rata-rata kelompok mata pelajaran umum	Sejarah, Seni Budaya, Informatika, PPKN, PJOK, dll.
3	*mean_jurusan*	Rata-rata mata pelajaran rumpun spesifik	Biologi, Fisika, Kimia (IPA) / Ekonomi, Geografi, Sosiologi (IPS)
4	*diff_jur_waj*	Selisih kemampuan jurusan vs wajib	*mean_jurusan* − *mean_wajib*
5	*diff_waj_umu*	Selisih kemampuan wajib vs umum	*mean_wajib* − *mean_umum*
6–22	*mean_[mapel]*	Rata-rata per mata pelajaran tunggal	Perhitungan rata-rata kumulatif tiap-tiap dari 17 mapel utama

Dataset akhir yang telah diperkaya fitur kemudian dibagi menggunakan metode *stratified split* dengan proporsi 80% sebagai data latih dan 20% sebagai data uji untuk menjamin representasi kelas target tetap seimbang di kedua subset. Untuk mengatasi masalah *multiclass imbalance*, teknik *Synthetic Minority Over-sampling Technique* (*SMOTE*) diaplikasikan secara eksklusif hanya pada data latih, sedangkan data uji dibiarkan murni (kondisi riil) untuk menjamin

validitas pengujian bebas bias.

3.3 Hasil Eksperimen Klasifikasi

Evaluasi performa model menggunakan acuan utama metrik *Macro F1-score* demi memberikan bobot evaluasi yang setara pada kelas minoritas (Suhaimi et al., 2023). Alur pengujian mengevaluasi tiga algoritma *ensemble* utama (*Random Forest*, *CatBoost*, dan *XGBoost*) yang diintegrasikan dengan skenario penambahan fitur, penyeimbangan data (*SMOTE*), serta seleksi dimensi fitur (*Feature Selection*).

3.3.1 Dampak *Feature Engineering* dan *SMOTE*

Table 3: Perbandingan Hasil Pengujian Seluruh Model pada Rumpun IPA dan IPS

No	Skenario	Model	Rumpun IPA		Rumpun IPS	
			Akurasi	Macro F1	Akurasi	Macro F1
1	ORI	XGBoost	78,64%	0,7362	74,00%	0,5638
2	FE	XGBoost	78,84%	0,7501	73,54%	0,5542
3	FE + SMOTE	XGBoost	78,64%	**0,7681**	75,18%	0,6234
5	ORI	CatBoost	63,07%	0,4697	70,49%	0,4435
6	FE	CatBoost	62,08%	0,4344	69,56%	0,4173
7	FE + SMOTE	CatBoost	62,28%	0,4705	67,21%	0,4369
8	ORI	Random Forest	36,32%	0,1819	39,58%	0,1704
9	FE	Random Forest	33,53%	0,1505	33,26%	0,1264
10	FE + SMOTE	Random Forest	35,93%	0,2440	37,24%	0,1705

Hasil pengujian awal menunjukkan bahwa penerapan *Feature Engineering* (FE) memberikan dampak yang berbeda pada kedua rumpun data. Pada kelompok IPA, FE berhasil meningkatkan *Macro F1-score XGBoost* dari 0,7362 menjadi 0,7501. Sebaliknya, pada kelompok IPS, penambahan fitur agregasi justru memicu penurunan performa pada seluruh algoritma. Karakteristik nilai sosial-humaniora yang relatif homogen menyebabkan fitur turunan bersifat *redundant* dan menambah dimensi data tanpa memberikan informasi yang cukup signifikan, sehingga berpotensi bertindak sebagai *noise* (Afshar and Usefi, 2022). Pola serupa juga terlihat pada *CatBoost* dan *Random Forest*, di mana nilai *Macro F1-score* mengalami penurunan setelah penerapan FE. Pada kelompok IPA, *CatBoost* turun dari 0,4697 menjadi 0,4344 dan *Random Forest* turun dari 0,1819 menjadi 0,1505. Sementara itu pada kelompok IPS, *CatBoost* menurun dari 0,4435 menjadi 0,4173 dan *Random Forest* dari 0,1704 menjadi 0,1264.

Namun, ketika konfigurasi FE dikombinasikan dengan penerapan *SMOTE* pada data latih, terjadi peningkatan performa pada seluruh algoritma. *XGBoost* mencapai performa terbaik pada kelompok IPA dengan *Macro F1-score* sebesar 0,7681, sedangkan pada kelompok IPS memperoleh nilai 0,6234. Peningkatan juga terlihat pada *CatBoost*, dengan *Macro F1-score* meningkat menjadi 0,4705 pada kelompok IPA dan 0,4369 pada kelompok IPS. Sementara itu, *Random Forest* menunjukkan perbaikan menjadi 0,2440 pada kelompok IPA dan 0,1705 pada kelompok IPS. Hasil ini menunjukkan bahwa penyeimbangan distribusi kelas melalui *SMOTE* membantu seluruh model dalam mengenali pola kelas minoritas dengan lebih baik. Meskipun demikian, peningkatan performa terbesar tetap diperoleh oleh *XGBoost*, yang menunjukkan kemampuan lebih baik dalam memanfaatkan informasi tambahan hasil *Feature Engineering* dan data sintetis yang dihasilkan oleh *SMOTE*.

3.3.2 Analisis Feature Importance dan Optimasi Feature Selection

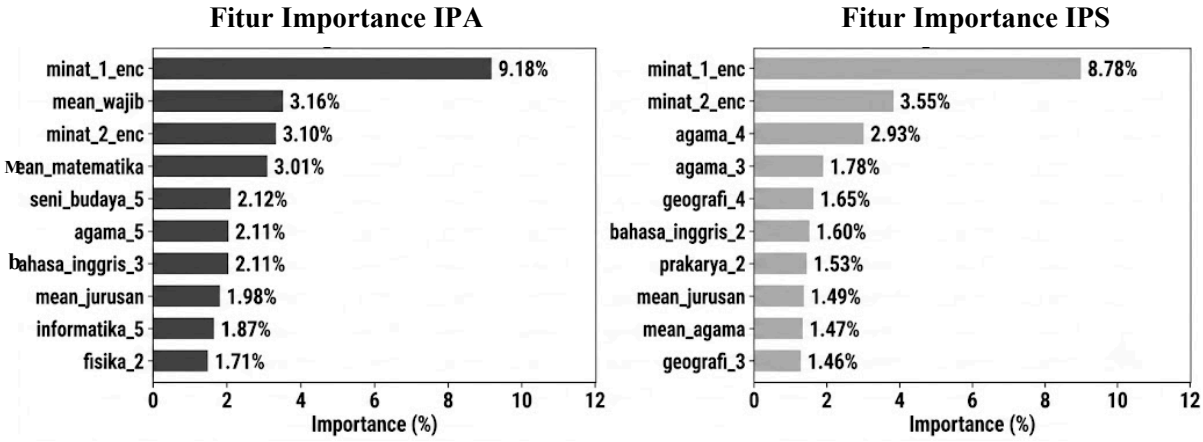

Figure 3: Fitur Importance pada Kelas Ipa dan IPS

Skor kepentingan fitur diekstrak dari model terbaik (*XGBoost*) untuk menganalisis kontribusi variabel dalam meminimalkan nilai error (Breiman, 2001). Pada kelompok IPA, variabel preferensi pilihan minat pertama siswa menjadi prediktor paling dominan dengan bobot kontribusi 9,18%, diikuti oleh fitur FE *mean_ wajib* (3,16%), pilihan minat kedua (3,10%), rata-rata Matematika (3,01%), dan Seni Budaya semester 5 (2,12%). Pada kelompok IPS, pola kontribusi juga dipimpin oleh pilihan minat pertama (8,78%) dan pilihan minat kedua (3,55%), dengan sensitivitas variabel akademik diwakili oleh nilai Agama semester 4 (2,93%), nilai Agama semester 3 (1,78%), dan Geografi semester 4 (1,65%). Temuan ini mengonfirmasi secara empiris bahwa klasifikasi rekomendasi studi yang akurat membutuhkan integrasi dinamis antara faktor preferensi subjektif dan kemampuan akademis objektif siswa (Zhang et al., 2021).

Berdasarkan urutan tingkat kepentingan tersebut, dilakukan eksperimen *Feature Selection* dengan menguji variasi ambang batas subset fitur sebesar 25%, 50%, 75%, dan 100% untuk menganalisis efisiensi model (Girish Chandrashekar and Ferat Sahin, 2014). Hasil pengujian reduksi fitur dirangkum secara komparatif pada Tabel 4.

Table 4: Hasil Eksperimen Feature Selection pada Model Terbaik (*XGBoost*)

Kelompok Data	Skenario % Fitur	Jumlah Dimensi Fitur	Akurasi (%)	*Macro F1-score*
Rumpun IPA	**100% Fitur (Optimal)**	**85**	**78,64%**	**0,768116**
	75% Fitur	63	78,24%	0,746577
	50% Fitur	42	78,04%	0,746776
	25% Fitur	21	78,44%	0,743634
Rumpun IPS	**25% Fitur (Optimal)**	**21**	**76,35%**	**0,696721**
	50% Fitur	42	75,64%	0,658849
	75% Fitur	63	75,88%	0,623500
	100% Fitur	85	75,64%	0,623427

Data Tabel 4 membuktikan adanya perbedaan karakteristik dimensi yang kontras antara kedua rumpun:

- Kelompok IPA membutuhkan representasi fitur utuh (100% fitur / 85 dimensi) untuk mempertahankan performa tertinggi. Pemangkasan dimensi justru melenyapkan informasi penting pendorong batas keputusan numerik eksakta.
- Kelompok IPS memperoleh lonjakan performa puncak secara signifikan dari *Macro F1* 62% melonjak ke 69% saat dimensi direduksi secara agresif hingga menyisakan 25% fitur terpenting (21 fitur). Langkah ini berhasil membuang tumpukan *noise* dan informasi redundant yang melekat pada karakteristik data IPS.

3.4 Analisis Komparatif Performa Model Klasifikasi

Secara keseluruhan, implementasi algoritma *XGBoost* menorehkan keunggulan performa mutlak dibandingkan *Random Forest* dan *CatBoost* di semua skenario pengujian. Model dengan terbaik disajikan pada Tabel 5.

Table 5: Performa Final Model XGBoost pada Setiap Rumpun Program Studi

Rumpun	Model	Subset Fitur	Akurasi	Macro Precision	Macro Recall	Macro F1
IPA	**XGBoost**	**100% (85 Fitur)**	**78,64%**	**82,77%**	**74,39%**	**76,81%**
IPS	**XGBoost**	**25% (21 Fitur)**	**76,35%**	**69,64%**	**66,21%**	**69,67%**

Performa XGBoost yang secara konsisten menghasilkan nilai akurasi dan Macro F1 tertinggi pada seluruh skenario pengujian di Tabel 3 mengonfirmasi landasan teoretis yang dikemukakan oleh Chen 2016 (Chen and Guestrin, 2016). Arsitektur *regularized gradient boosting* yang digunakan *XGBoost* mampu meminimalkan fungsi objektif secara sekuensial dengan memperbaiki kesalahan prediksi pada iterasi sebelumnya, sehingga menghasilkan kemampuan klasifikasi yang lebih baik dibandingkan pendekatan *bagging* yang diterapkan pada *Random Forest.* Perbedaan performa tersebut tercermin dari nilai *Macro F1 XGBoost* yang secara konsisten melampaui *CatBoost* dan *Random Forest* pada rumpun IPA maupun IPS. Temuan ini sejalan dengan penelitian ayodele 2023 yang menyatakan bahwa algoritma berbasis *boosting* memiliki kemampuan generalisasi yang lebih baik dalam merekonstruksi batas keputusan kelas minoritas dibandingkan metode berbasis *bagging* (Adefemi Ayodele, 2023). Meskipun *CatBoost* juga mengadopsi pendekatan *boosting*, hasil pengujian menunjukkan bahwa performanya masih berada di bawah *XGBoost* pada seluruh skenario. Hal ini mengindikasikan bahwa karakteristik data yang menggabungkan informasi minat dan nilai akademik lebih sesuai dimodelkan menggunakan mekanisme optimasi yang dimiliki *XGBoost.* Oleh karena itu, *XGBoost* dipilih sebagai model terbaik pada penelitian ini.

4 Kesimpulan

Penelitian ini berhasil membangun model klasifikasi rekomendasi 31 subrumpun program studi melalui *machine learning pipeline* terstruktur menggunakan teknik *mean imputation*, *label encoding*, *Feature Engineering* (85 fitur), dan penanganan *multiclass imbalance* dengan *SMOTE.* Hasil eksperimen membuktikan bahwa algoritma *XGBoost* menghasilkan performa paling superior dan adaptif dibandingkan *Random Forest* dan *CatBoost* pada kedua rumpun data.

Pada kelompok IPA, skenario terbaik dicapai dengan mempertahankan 100% fitur (85 dimensi) yang menghasilkan Akurasi 78,64% dan *Macro F1-score* 76,81%. Pada model ini, preferensi pilihan minat pertama menjadi prediktor paling dominan dengan bobot kontribusi 9,18%, diikuti minat kedua (3,10%), rata-rata Matematika (3,09%), dan Seni Budaya semester 5 (2,10%).

Sebaliknya, kelompok IPS membutuhkan *Feature Selection* agresif dengan hanya menggunakan 25% fitur terpenting (21 fitur) untuk mengeliminasi *noise* data akibat karakteristik nilai yang homogen. Strategi ini terbukti mendongkrak performa *Macro F1-score* dari 61,30% menjadi 67,21% (Akurasi 76,34%). Kontribusi fitur didominasi oleh pilihan minat pertama sebesar 7,66%, diikuti oleh nilai Agama semester 4 (4,45%), minat kedua (3,11%), Prakarya semester 4 (2,61%), dan Geografi semester 3 (2,12%). Secara keseluruhan, integrasi faktor minat subjektif dan rekam akademik terbukti saling melengkapi dalam meningkatkan akurasi generalisasi model.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Program Studi S1 Sains Data Universitas Negeri Surabaya serta seluruh pihak yang telah memberikan dukungan dan masukan selama pelaksanaan penelitian ini.

References

Adefemi Ayodele (2023). A comparative study of ensemble learning techniques for imbalanced classification problems. *World Journal of Advanced Research and Reviews*, 19(2):1633–1643.

Afshar, M. and Usefi, H. (2022). Optimizing feature selection methods by removing irrelevant features using sparse least squares. *Expert Systems with Applications*, 200:116928.

Alsahaf, A., Petkov, N., Shenoy, V., and Azzopardi, G. (2022). A framework for feature selection through boosting. *Expert Systems with Applications*, 187:115895.

Alsayed, A. O., Rahim, M. S. M., AlBidewi, I., Hussain, M., Jabeen, S. H., Alromema, N., Hussain, S., and Jibril, M. L. (2021). Selection of the Right Undergraduate Major by Students Using Supervised Learning Techniques. *Applied Sciences*, 11(22):10639.

Brabec, J. and Machlica, L. (2018). Decision-Forest Voting Scheme for Classification of Rare Classes in Network Intrusion Detection. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3325–3330, Miyazaki, Japan. IEEE.

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

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. arXiv:1603.02754 [cs].

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg. Springer.

Fathurohman, A. (2021). MACHINE LEARNING UNTUK PENDIDIKAN: MENGAPA DAN BAGAIMANA. *Jurnal Informatika Dan Tekonologi Komputer (JITEK)*, 1(3):57–62.

Felldy Aslya Utama (2020). 87 Persen Mahasiswa Indonesia Akui Salah Jurusan.

Gillis, A. and Ryberg, R. (2021). Is Choosing a Major Choosing a Career or Interesting Courses? An Investigation into College Students' Orientations for College Majors and Their Stability. *Journal of Postsecondary Student Success*, 1(2):46–71.

Girish Chandrashekar and Ferat Sahin (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28.

Guevara-Reyes, R., Ortiz-Garcés, I., Andrade, R., Cox-Riquetti, F., and Villegas-Ch, W. (2025).

Machine learning models for academic performance prediction: interpretability and application in educational decision-making. *Frontiers in Education*, 10.

Gultom, E., Darmanto, T., and Tjen, J. (2026). Sistem Pakar Pendeteksi Penyakit Pernapasan Menggunakan Gradient Boosting dan Metode CNN. *Jutisi : Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*, 15(2):782–793.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). Boosting and Additive Trees. In Hastie, T., Tibshirani, R., and Friedman, J., editors, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, pages 337–387. Springer, New York, NY.

lavender lili (2025). Metode dan Algoritma Ensemble Learning | PDF.

Milsom, A. and Coughlin, J. (2015). Satisfaction With College Major: A Grounded Theory Study. *NACADA Journal*, 35(2):5–14.

Mystere, K., Mpia, H., and Mutegheki Baraka, V. (2023). Prédiction de l'orientation des étudiants dans des filières d'études appropriées en utilisant les techniques de Data Mining. *International Journal of Innovation and Applied Studies*, 39:193–208.

Nurhalizah, R. S., Ardianto, R., and Purwono, P. (2024). Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review. *Jurnal Ilmu Komputer dan Informatika*, 4(1):61–72.

Prayoga, M. H. B. and Ermatita, E. (2024). Analisis Pemilihan Jurusan pada Calon Siswa SMK Negeri 4 Palembang Pada Faktor Penentu Pemilihan Jurusan Menggunakan Association Rule dan Random Forest. *Jurnal Pendidikan dan Teknologi Indonesia*, 4(12):537–547.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2019). CatBoost: unbiased boosting with categorical features. arXiv:1706.09516 [cs].

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

SNPMB (2025). Siaran Pers Peluncuran Seleksi Nasional Penerimaan Mahasiswa Baru (SNPMB) Tahun 2025.

Suhaimi, N. S., Othman, Z., and Yaakub, M. R. (2023). Comparative Analysis Between Macro and Micro-Accuracy in Imbalance Dataset for Movie Review Classification. In Yang, X.-S., Sherratt, S., Dey, N., and Joshi, A., editors, *Proceedings of Seventh International Congress on Information and Communication Technology*, pages 83–93, Singapore. Springer Nature.

Sujon, K. M., Hassan, R., Choi, K., and Samad, M. A. (2025). Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. *Journal of Big Data*, 12(1):268.

Wang, C.-C., Kuo, P.-H., and Chen, G.-Y. (2022). Machine Learning Prediction of Turning Precision Using Optimized XGBoost Model. *Applied Sciences*, 12(15):7739.

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., and Shang, X. (2021). Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. *Frontiers in Psychology*, 12.