JURNAL TEKNIK
Volume 24, No. 1, Juni 2026
P-ISSN : 1693-6191 E-ISSN : 2715-7660 DOI : https://doi.org/10.37031/jt.v24i1.750

Perbandingan Algoritma Supervised Learning dalam Memprediksi Jalur Seleksi Masuk Mahasiswa di Universitas Negeri Gorontalo

[1]Manda Rohandi, [2]Mukhlisulfatih Latief, [3]Mohamad Ilyas Abas
[1,2]Fakultas Teknik, Kota Gorontalo, 8114333963/Universitas Negeri Gorontalo/Indonesia
[3]Fakultas Sains dan Ilmu Komputer, Kabupaten Gorontalo, Universitas Muhammadiyah Gorontalo/Indonesia
e-mail: manda.rohandi@ung.ac.id (correspondence email)

Abstrak

Pemilihan jalur seleksi masuk (SNBT, SNBP, atau Mandiri) berkaitan erat dengan karakteristik demografis dan administratif mahasiswa, dan pemahaman terhadap pola ini bermanfaat bagi perguruan tinggi dalam merancang strategi rekrutmen yang lebih tepat sasaran. Penelitian ini bertujuan membandingkan performa empat algoritma supervised learning Decision Tree, Random Forest, Naïve Bayes, dan k-Nearest Neighbor (k-NN) dalam memprediksi jalur seleksi masuk mahasiswa Jurusan Teknik Informatika Universitas Negeri Gorontalo (UNG). Data diperoleh dari Sistem Informasi Akademik Terpadu (SIAT) UNG sebanyak 1.237 mahasiswa Program Studi Pendidikan Teknologi Informasi dan Sistem Informasi angkatan 2018–2024, yang setelah pembersihan data menjadi 1.232 sampel dengan tiga kelas jalur (SNBT = 682, SNBP = 417, Mandiri = 133). Pemilihan fitur menggunakan information gain menghasilkan lima fitur informatif: seleksi (nasional/lokal), umur, angkatan, semester awal terdaftar, dan jenis kelamin. Evaluasi dilakukan melalui tiga skenario: hold-out 80:20 dengan Synthetic Minority Oversampling Technique (SMOTE), dan 10-fold stratified cross-validation, dilanjutkan uji signifikansi Friedman. Hasil 10-fold cross-validation menunjukkan akurasi rata-rata Random Forest sebesar 64,61%, Decision Tree 64,37%, Naïve Bayes 63,47%, dan k-NN 61,03%. Uji Friedman menunjukkan tidak terdapat perbedaan performa yang signifikan antar keempat algoritma (χ^2 = 4,39; p = 0,222). Confusion matrix mengungkap bahwa seluruh algoritma memprediksi sempurna kelas Mandiri karena adanya keterkaitan definisional dengan fitur seleksi, sementara kesalahan klasifikasi terutama terjadi antara kelas SNBP dan SNBT. Penerapan SMOTE secara konsisten meningkatkan recall kelas minoritas SNBP pada keempat algoritma, namun efeknya terhadap F1-Score makro dan akurasi global bervariasi antar-algoritma dan tidak selalu positif. Penelitian ini merekomendasikan penambahan fitur perilaku akademik untuk meningkatkan kemampuan diskriminasi model antara jalur seleksi nasional.

Kata kunci: supervised learning, information gain, jalur seleksi mahasiswa, SMOTE, data mining pendidikan.

Abstract

The choice of university admission pathway (SNBT, SNBP, or independent selection/Mandiri) is closely related to students' demographic and administrative characteristics, and understanding this pattern benefits higher education institutions in designing more targeted recruitment strategies. This study aims to compare the performance of four supervised learning algorithms Decision Tree, Random Forest, Naïve Bayes, and k-Nearest Neighbor (k-NN) in predicting the admission pathway of Informatics Engineering students at Universitas Negeri Gorontalo (UNG). Data were obtained from UNG's Integrated Academic Information System (SIAT) for 1,237 students from the Information Technology Education and Information Systems study programs (2018–2024 cohorts), which after data cleaning resulted in 1,232 samples across three pathway classes (SNBT = 682, SNBP = 417, Mandiri = 133). Feature selection using information gain identified five informative features: selection type (national/local), age, cohort year, initial registration semester, and gender. Evaluation was conducted through three scenarios: 80:20 hold-out with Synthetic Minority Oversampling Technique (SMOTE), and 10-fold stratified cross-validation, followed by a Friedman significance test. The 10-fold cross-validation results show mean accuracies of 64.61% for Random Forest, 64.37% for Decision Tree, 63.47% for Naïve Bayes, and 61.03% for k-NN. The Friedman test indicated no statistically significant difference in performance among the four algorithms (χ^2 = 4.39; p = 0.222). The confusion matrix revealed that all algorithms perfectly classified the Mandiri class due to a definitional dependency with the selection-type feature, while most misclassifications occurred between the SNBP and SNBT classes. Applying SMOTE consistently improved minority-class (SNBP) recall across all four

algorithms, but its effect on macro F1-score and overall accuracy varied by algorithm and was not uniformly positive. This study recommends incorporating academic behavioral features to improve the model's discriminative ability between national selection pathways.

***Keywords**: supervised learning, information gain, student admission pathway, SMOTE, educational data mining.*

Diterima : Mei 2026
Disetujui : Mei 2026
Dipublikasi : Juni 2026

Pendahuluan

Proses penerimaan mahasiswa baru di perguruan tinggi negeri di Indonesia dilaksanakan melalui beberapa jalur seleksi, yaitu Seleksi Nasional Berdasarkan Tes (SNBT), Seleksi Nasional Berdasarkan Prestasi (SNBP), dan seleksi Mandiri yang diselenggarakan secara independen oleh masing-masing perguruan tinggi. Ketiga jalur ini memiliki karakteristik proses seleksi, kriteria penerimaan, dan profil demografis pendaftar yang berbeda-beda. Pemahaman terhadap pola dan karakteristik mahasiswa pada masing-masing jalur seleksi penting bagi perguruan tinggi dalam merancang strategi rekrutmen, alokasi kuota, dan kebijakan akademik yang lebih tepat sasaran (Matz, dkk, 2023).

Educational Data Mining (EDM) berbasis algoritma supervised learning telah banyak dimanfaatkan untuk menganalisis data historis mahasiswa guna mengidentifikasi pola-pola tersembunyi yang relevan dengan pengambilan keputusan institusional. Berbagai algoritma, seperti Decision Tree, Random Forest (Elbouknify, dkk, 2025), Naïve Bayes, K-Nearest Neighbor (Indra dan Agustinawati, 2024), dan regresi logistik, telah digunakan untuk memprediksi berbagai luaran akademik mahasiswa, mulai dari performa akademik (Yağcı, 2022) hingga risiko putus studi (Villegas-Ch, dkk, 2023). Namun, penerapan algoritma-algoritma ini untuk memprediksi jalur seleksi masuk mahasiswa berdasarkan atribut demografis dan administratif masih jarang dibahas secara eksplisit dalam literatur EDM di Indonesia.

Oleh karena itu, tujuan penelitian ini adalah membandingkan performa empat algoritma supervised learning Decision Tree, Random Forest, Naïve Bayes, dan k-NN dalam memprediksi jalur seleksi masuk mahasiswa Jurusan Teknik Informatika Universitas Negeri Gorontalo, menggunakan data operasional riil dari Sistem Informasi Akademik Terpadu (SIAT). Penelitian ini turut mengevaluasi pengaruh penanganan ketidakseimbangan kelas (class imbalance) melalui teknik Synthetic Minority Oversampling Technique (SMOTE) serta menggunakan 10-fold cross-validation dan uji

signifikansi statistik Friedman untuk memastikan validitas perbandingan performa antar algoritma.

Kebaruan (novelty) penelitian ini terletak pada tiga aspek: (1) penerapan perbandingan empat algoritma supervised learning pada konteks lokal institusi Fakultas Teknik UNG dengan target prediksi jalur seleksi masuk yang belum banyak dibahas pada studi EDM sejenis; (2) evaluasi menyeluruh melalui tiga skenario pengujian (hold-out, SMOTE, dan cross-validation) disertai uji signifikansi statistik formal; dan (3) identifikasi eksplisit terhadap dependensi definisional antar-fitur yang dapat menimbulkan bias evaluasi performa model, sebuah isu metodologis yang jarang dilaporkan secara transparan pada studi EDM di lingkup perguruan tinggi Indonesia.

Metode

Penelitian ini dilakukan dengan tujuh langkah utama, sebagai berikut:

1. Langkah Pengumpulan Data

Data mahasiswa dikumpulkan melalui ekstraksi dari aplikasi Sistem Informasi Akademik Terpadu (SIAT) Universitas Negeri Gorontalo, mencakup 1.237 mahasiswa dari dua program studi di Jurusan Teknik Informatika Pendidikan Teknologi Informasi (582 mahasiswa) dan Sistem Informasi (655 mahasiswa) angkatan 2018 sampai dengan 2024. Dataset terdiri atas 18 atribut, meliputi data demografis (jenis kelamin, agama, tanggal lahir), data akademik (angkatan, program studi, semester awal terdaftar), dan data administratif seleksi (jenis seleksi, jalur, status awal, status aktif).

2. Langkah Pembersihan Data

Pemeriksaan data tidak menemukan nilai yang hilang (missing value) maupun data duplikat berdasarkan Nomor Induk Mahasiswa. Fitur turunan UMUR dihitung dari selisih tanggal lahir terhadap tanggal referensi 31 Desember 2024. Kolom identifier yang tidak relevan untuk klasifikasi (NO, NIK, NIM, NAMA, TANGGAL LAHIR) serta kolom berkardinalitas konstan (JENJANG) dikeluarkan dari analisis. Variabel target JALUR pada data awal memiliki enam kategori (SNBT = 682, SNBP = 417, Mandiri = 133, ADik = 3, Kerjasama = 1, Jurusan = 1); tiga kategori dengan jumlah anggota di bawah 30 sampel (ADik, Kerjasama, Jurusan; total 5 sampel) dikeluarkan dari analisis karena tidak memenuhi syarat minimum untuk pengujian 10-fold stratified cross-validation maupun oversampling SMOTE, sehingga dataset final berjumlah 1.232 sampel dengan tiga kelas target: SNBT (682; 55,4%), SNBP (417; 33,8%), dan Mandiri (133; 10,8%).

3. Langkah Pemilihan Fitur Data

Fitur-fitur data diseleksi menggunakan information gain berdasarkan entropy dari masing-masing fitur terhadap variabel target. Perhitungan entropy dan information gain mengikuti persamaan (1) dan (2):

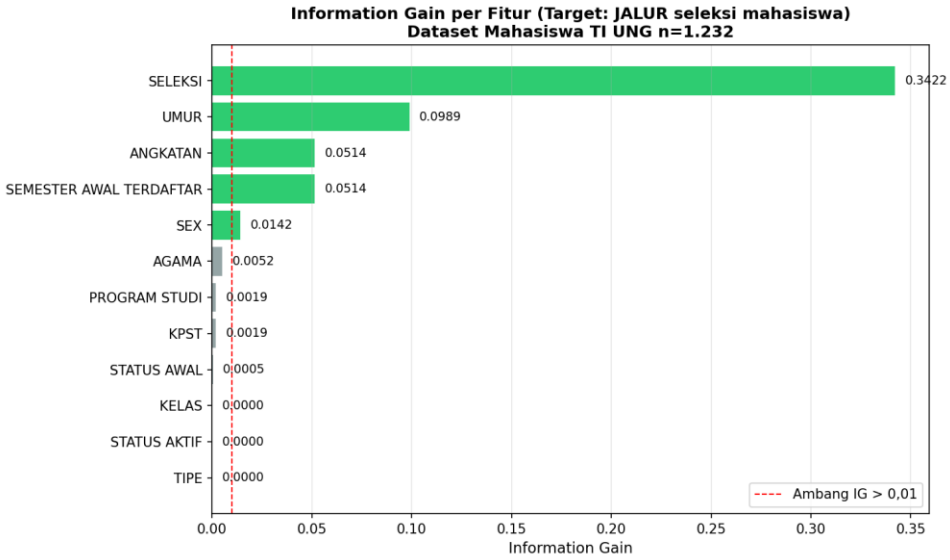
$$\text{Entropy}(S) = -\Sigma\, p(xi) \times \log 2\, p(xi) \quad ...(1)$$

Keterangan: n = jumlah nilai yang mungkin untuk kelas target; p(xi) = probabilitas kelas ke-i.

$$\text{Gain}(x,t) = \text{Entropy}(S) - \Sigma\, (|xi| / |x|) \times \text{Entropy}(Si) \quad ...(2)$$

Keterangan: x = jumlah sampel seluruh kelas; t = fitur; S = entropy seluruh fitur sebelum pemisahan; xi = jumlah sampel kelas dengan nilai = i; Si = entropy fitur untuk kelas i setelah pemisahan.

Entropy variabel target H(JALUR) = 1,348 bit. Hasil perhitungan information gain menunjukkan lima fitur dengan nilai IG > 0,01 dan dinyatakan informatif untuk proses klasifikasi: SELEKSI (0,3422), UMUR (0,0989), ANGKATAN (0,0514), SEMESTER AWAL TERDAFTAR (0,0514), dan SEX (0,0142). Fitur dengan IG mendekati nol (AGAMA, PROGRAM STUDI, KPST, STATUS AWAL, KELAS, TIPE, STATUS AKTIF) dikeluarkan dari proses klasifikasi. Rincian nilai information gain seluruh fitur disajikan pada Gambar 1.



Gambar 1. Information gain per fitur terhadap variabel target JALUR

4. Langkah Pemilahan Data Latih dan Data Uji

Data sebanyak 1.232 sampel dipisahkan menjadi data latih (985 sampel; 80%) dan data uji (247 sampel; 20%) menggunakan stratified sampling untuk menjaga proporsi masing-masing kelas target pada kedua subset. Distribusi kelas pada data latih: SNBT = 545, SNBP = 334, Mandiri = 106; pada data uji: SNBT = 137, SNBP = 83,

Mandiri = 27. Untuk mengatasi ketidakseimbangan kelas, teknik Synthetic Minority Oversampling Technique (SMOTE) diterapkan pada data latih (k=5 tetangga terdekat), menghasilkan 1.635 sampel latih dengan distribusi seimbang (545 sampel per kelas). SMOTE hanya diterapkan pada data latih; data uji tetap menggunakan distribusi asli agar evaluasi mencerminkan kondisi populasi riil.

5. Langkah Klasifikasi Data

Klasifikasi dilakukan menggunakan empat algoritma supervised learning: Decision Tree (kriteria entropy, kedalaman maksimum 10, minimal leaf size 2), Random Forest (100 trees, kriteria entropy, kedalaman maksimum 10), Naïve Bayes (Gaussian Naïve Bayes), dan k-Nearest Neighbor (k=5). Untuk algoritma Naïve Bayes dan k-NN yang sensitif terhadap skala fitur, dilakukan standardisasi (StandardScaler) sebelum pelatihan model. Evaluasi dilakukan pada tiga skenario: (a) hold-out 80:20 tanpa penyeimbangan kelas, (b) hold-out 80:20 dengan SMOTE pada data latih, dan (c) 10-fold stratified cross-validation pada keseluruhan dataset.

6. Langkah Pengujian Keakuratan

Pengujian keakuratan dilakukan menggunakan confusion matrix, dengan metrik akurasi, precision, recall, dan F1-Score (rata-rata makro) untuk mengakomodasi evaluasi yang adil pada dataset tidak seimbang.

7. Langkah Perbandingan dan Uji Signifikansi Statistik

Perbandingan performa antar algoritma pada hasil 10-fold cross-validation diuji signifikansinya menggunakan uji non-parametrik Friedman ($\alpha = 0{,}05$), yang sesuai untuk membandingkan lebih dari dua algoritma pada sampel berpasangan (paired) dari fold yang sama. Apabila uji Friedman menunjukkan perbedaan signifikan, dilanjutkan dengan uji post-hoc Wilcoxon signed-rank secara berpasangan.

Seluruh proses pengumpulan data, pembersihan, pemilihan fitur, klasifikasi, dan evaluasi dilakukan menggunakan Python 3.12 dengan pustaka scikit-learn 1.8, imbalanced-learn, dan scipy untuk uji statistik.

Hasil dan Pembahasan

Hasil

1. Hasil Klasifikasi Skenario Baseline (Hold-out 80:20 tanpa SMOTE)

Confusion matrix hasil klasifikasi keempat algoritma pada skenario baseline disajikan pada Tabel 1–4. Seluruh algoritma memprediksi dengan sempurna (27/27) kelas

Mandiri, sedangkan kesalahan klasifikasi terkonsentrasi pada pembedaan antara kelas SNBP dan SNBT.

Tabel 1. Confusion matrix klasifikasi algoritma Decision Tree (baseline)

	True Mandiri	True SNBP	True SNBT	Precision	Akurasi
Pred. Mandiri	27	0	0	100.00%	**56.68%**
Pred. SNBP	0	22	46	32.35%	
Pred. SNBT	0	61	91	59.87%	
Class Recall	100.00%	26.51%	66.42%		

Tabel 2. Confusion matrix klasifikasi algoritma Random Forest (baseline)

	True Mandiri	True SNBP	True SNBT	Precision	Akurasi
Pred. Mandiri	27	0	0	100.00%	**59.92%**
Pred. SNBP	0	24	40	37.50%	
Pred. SNBT	0	59	97	62.18%	
Class Recall	100.00%	28.92%	70.80%		

Tabel 3. Confusion matrix klasifikasi algoritma Naïve Bayes (baseline)

	True Mandiri	True SNBP	True SNBT	Precision	Akurasi
Pred. Mandiri	27	0	0	100.00%	**65.59%**
Pred. SNBP	0	29	31	48.33%	
Pred. SNBT	0	54	106	66.25%	
Class Recall	100.00%	34.94%	77.37%		

Tabel 4. Confusion matrix klasifikasi algoritma k-NN (baseline)

	True Mandiri	True SNBP	True SNBT	Precision	Akurasi
Pred. Mandiri	27	0	0	100.00%	**57.09%**
Pred. SNBP	0	20	43	31.75%	
Pred. SNBT	0	63	94	59.87%	
Class Recall	100.00%	24.10%	68.61%		

2. Hasil Klasifikasi Skenario dengan SMOTE

Setelah data latih diseimbangkan dengan SMOTE, distribusi hasil klasifikasi bergeser ke arah yang lebih merata antar kelas, sebagaimana disajikan pada Tabel 5.

Tabel 5. Confusion matrix klasifikasi keempat algoritma dengan SMOTE

Algoritma		True Mandiri	True SNBP	True SNBT	Precision	Akurasi
Decision Tree	Pred. Mandiri	27	0	0	100.00%	**60.73%**

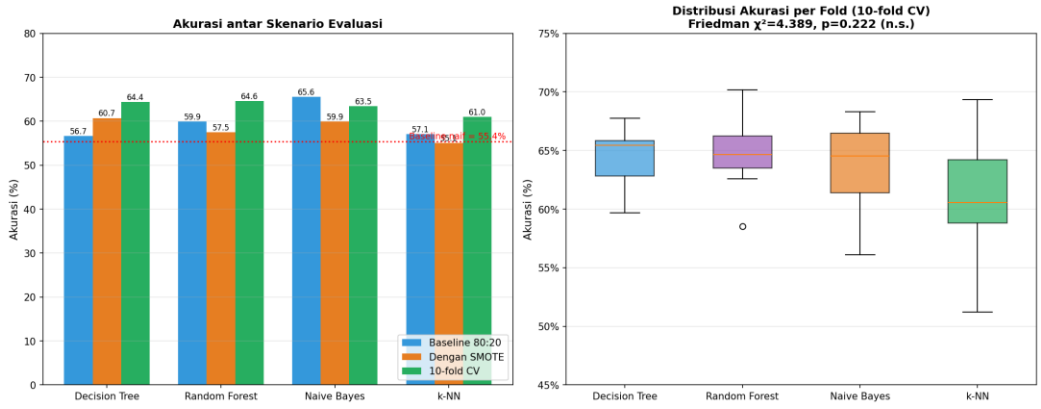
	Pred. SNBP	0	47	61	43.52%	
	Pred. SNBT	0	36	76	67.86%	
Random Forest	Pred. Mandiri	27	0	0	100.00%	**57.49%**
	Pred. SNBP	0	40	62	39.22%	
	Pred. SNBT	0	43	75	63.56%	
Naïve Bayes	Pred. Mandiri	27	0	0	100.00%	**59.92%**
	Pred. SNBP	0	52	68	43.33%	
	Pred. SNBT	0	31	69	69.00%	
k-NN	Pred. Mandiri	27	0	0	100.00%	**55.06%**
	Pred. SNBP	0	32	60	34.78%	
	Pred. SNBT	0	51	77	60.16%	

3. Hasil 10-fold Stratified Cross-Validation

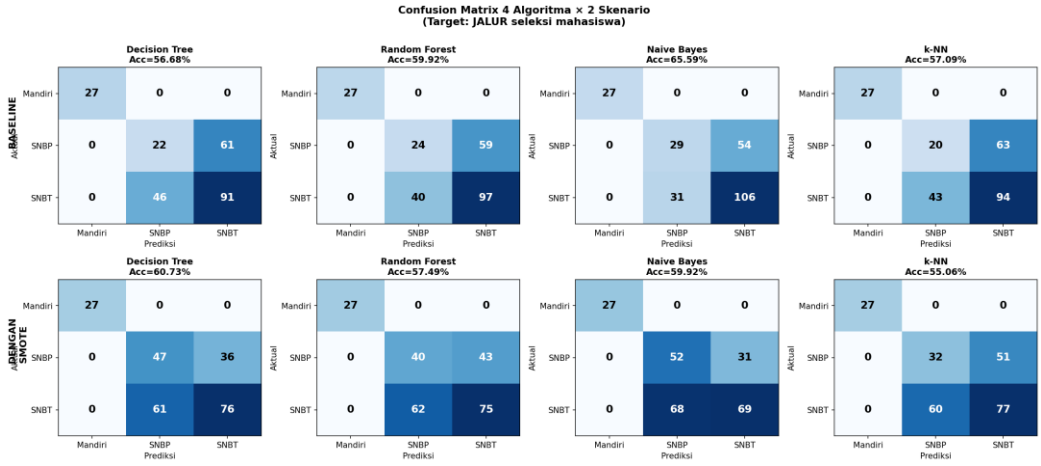
Tabel 6 menyajikan rata-rata dan simpangan baku akurasi, precision, recall, dan F1-Score (makro) hasil 10-fold cross-validation untuk keempat algoritma.

Tabel 6. Rata-rata ± simpangan baku hasil 10-fold stratified cross-validation

Algoritma	Akurasi (%)	Precision Macro (%)	Recall Macro (%)	F1 Macro (%)
Decision Tree	64.37 ± 2.48	70.84 ± 2.57	70.47 ± 2.51	70.19 ± 2.77
Random Forest	64.61 ± 2.88	70.84 ± 2.60	69.99 ± 1.90	69.70 ± 1.98
Naïve Bayes	63.47 ± 3.72	71.47 ± 2.79	71.61 ± 2.89	71.43 ± 2.88
k-NN	61.03 ± 5.26	68.30 ± 4.38	68.14 ± 4.00	68.00 ± 4.13



Gambar 2. Perbandingan akurasi antar skenario evaluasi dan distribusi akurasi per fold (10-fold cross-validation)



Gambar 3. Confusion matrix keempat algoritma pada skenario baseline dan SMOTE

4. Hasil Uji Signifikansi Statistik (Friedman)

Uji Friedman terhadap akurasi 10-fold cross-validation menghasilkan $\chi^2 = 4{,}39$ dengan p-value = 0,222 ($\alpha = 0{,}05$). Karena p-value lebih besar dari 0,05, hipotesis nol tidak dapat ditolak, yang berarti tidak terdapat bukti statistik yang cukup untuk menyatakan adanya perbedaan performa yang signifikan di antara keempat algoritma pada kasus prediksi jalur seleksi ini. Oleh karena itu, uji post-hoc Wilcoxon tidak dilanjutkan.

Pembahasan

Hasil penelitian menunjukkan bahwa keempat algoritma supervised learning menghasilkan performa yang relatif setara dalam memprediksi jalur seleksi masuk mahasiswa, dengan rentang akurasi 10-fold cross-validation antara 61,03% (k-NN) hingga 64,61% (Random Forest) selisih tertinggi-terendah hanya 3,58 poin persentase, jauh lebih kecil dibandingkan simpangan baku masing-masing algoritma antar-fold (2,48–5,26 poin). Uji Friedman mengonfirmasi bahwa selisih sekecil itu memang tidak melampaui ambang signifikansi statistik ($\chi^2 = 4{,}39$; $p = 0{,}222$), sehingga pemilihan algoritma untuk implementasi praktis pada kasus ini lebih tepat didasarkan pada pertimbangan operasional: interpretabilitas struktur pohon keputusan yang lebih mudah ditelusuri auditor akademik dibandingkan struktur ensemble Random Forest, atau beban komputasi Naïve Bayes yang jauh lebih ringan karena tidak memerlukan proses pelatihan iteratif. Pola kesetaraan performa semacam ini bukan kekhususan dataset SIAT UNG semata; Sathe dan Adamuthe (2021) mencatat gejala serupa pada dataset pendidikan berskala kecil-menengah, yang mereka kaitkan dengan terbatasnya daya pisah (separability) antar-kelas ketika jumlah fitur informatif sedikit kondisi yang persis terjadi di sini, mengingat hanya lima dari dua belas atribut awal yang lolos ambang information gain.

Temuan yang perlu digarisbawahi secara kritis adalah keberhasilan sempurna (recall dan precision 100%) seluruh algoritma dalam mengklasifikasikan kelas Mandiri. Hal ini bukan disebabkan oleh kapabilitas prediktif model yang unggul, melainkan oleh dependensi definisional antara fitur SELEKSI dan kelas Mandiri: seluruh mahasiswa jalur Mandiri di dataset ini terdaftar melalui seleksi berjenis Lokal, sementara seluruh mahasiswa jalur SNBP dan SNBT terdaftar melalui seleksi Nasional. Fitur SELEKSI dengan demikian berfungsi sebagai proxy yang hampir sempurna untuk membedakan Mandiri dari dua kelas lainnya, sebagaimana tercermin dari nilai information gain-nya yang jauh lebih tinggi (0,342) dibandingkan fitur lain. Kami mengakui bahwa temuan performa sempurna pada kelas Mandiri semestinya tidak diinterpretasikan sebagai keunggulan algoritma, melainkan sebagai karakteristik struktural dataset yang perlu ditelaah lebih lanjut sebelum model ini digunakan untuk keperluan praktis.

T antangan klasifikasi yang sesungguhnya terletak pada pembedaan antara kelas SNBP dan SNBT, yang keduanya sama-sama berjalur seleksi Nasional dan karenanya tidak dapat dibedakan oleh fitur SELEKSI. Confusion matrix pada Tabel 1–4 menunjukkan arah kesalahan yang konsisten pada keempat algoritma: proporsi mahasiswa SNBP yang justru diklasifikasikan sebagai SNBT jauh lebih besar dibandingkan arah sebaliknya pada Decision Tree misalnya, 61 dari 83 mahasiswa SNBP (73,5%) salah diklasifikasikan sebagai SNBT, sementara SNBT sendiri hanya "tersesat" ke SNBP pada 46 dari 137 kasus (33,6%). Arah bias ini bersifat sistematis, bukan acak, dan selaras dengan mekanisme yang diuraikan Oppong (2023): ketika fitur pembeda antar dua kelas tidak cukup informatif, classifier akan menyerap ambiguitas tersebut ke arah kelas yang secara proporsional lebih dominan pada data latih dalam kasus ini SNBT, yang berjumlah 545 dari 985 sampel latih (55,3%) berbanding SNBP yang hanya 334 sampel (33,9%).

Penerapan SMOTE pada data latih secara konsisten menaikkan recall kelas minoritas SNBP pada keempat algoritma dari kisaran 24,10–34,94% pada baseline menjadi 38,55–62,65% setelah SMOTE namun dengan konsekuensi penurunan recall kelas mayoritas SNBT yang besarannya bervariasi antar-algoritma; penurunan paling tajam terjadi pada Naïve Bayes (77,37% menjadi 50,36%), sedangkan pada Decision Tree penurunannya lebih moderat (66,42% menjadi 55,47%). Pola pada precision tidak seragam: pada Decision Tree dan Random Forest, precision kedua kelas sama-sama meningkat setelah SMOTE (precision SNBP: 32,35%→43,52% dan 37,50%→39,22%; precision SNBT: 59,87%→67,86% dan 62,18%→63,56%), sedangkan pada Naïve Bayes precision SNBP justru sedikit menurun (48,33% menjadi 43,33%). Karena F1-

Score makro merangkum precision dan recall kedua kelas secara seimbang, arah perubahannya juga tidak seragam: naik pada Decision Tree (64,04%→70,09%), Random Forest (66,29%→67,36%), dan k-NN (63,78%→64,89%), tetapi justru turun tipis pada Naïve Bayes (70,65%→69,82%) algoritma yang pada baseline sudah memiliki recall SNBT tertinggi (77,37%) sehingga paling banyak "kehilangan" sensitivitas pada kelas mayoritas ketika SMOTE menggeser bobot data latih ke arah kelas minoritas. Pola bercabang ini adalah konsekuensi langsung dari cara kerja SMOTE: dengan menyeimbangkan data latih menjadi 545 sampel per kelas, batas keputusan (decision boundary) tiap model bergeser untuk lebih mengakomodasi kelas minoritas SNBP, tetapi besarnya pergeseran itu dan karena itu besarnya trade-off recall SNBP versus recall SNBT bergantung pada karakteristik masing-masing algoritma dalam merespons perubahan distribusi data latih. Implikasi praktisnya sejalan dengan yang ditekankan Azizah dkk. (2024): keputusan untuk menerapkan SMOTE pada kasus serupa tidak dapat digeneralisasi sebagai "selalu menguntungkan", melainkan perlu dievaluasi per algoritma dan disesuaikan dengan metrik yang relevan terhadap tujuan penggunaan model jika tujuannya mendeteksi mahasiswa jalur SNBP untuk kepentingan evaluasi kuota, kenaikan recall SNBP yang konsisten pada keempat algoritma menjadikan SMOTE tetap relevan dipertimbangkan meskipun efeknya terhadap F1-Score makro dan akurasi global tidak seragam.

Sebagai perbandingan, model dengan strategi naif "selalu memprediksi kelas mayoritas (SNBT)" akan mencapai akurasi 55,4% (682/1.232). Akurasi rata-rata keempat algoritma pada 10-fold cross-validation (61–65%) berada 6–9 poin persentase di atas baseline naif tersebut, mengindikasikan bahwa model memang mempelajari pola yang bermakna dari fitur UMUR, ANGKATAN, SEMESTER AWAL TERDAFTAR, dan SEX meskipun pola tersebut belum cukup kuat untuk mencapai akurasi tinggi pada pembedaan SNBP–SNBT. Hal ini kemungkinan mengindikasikan bahwa atribut demografis dan administratif yang tersedia belum sepenuhnya menangkap faktor-faktor yang sesungguhnya mendorong pilihan jalur seleksi mahasiswa, seperti nilai rapor, prestasi non-akademik, atau faktor sosial-ekonomi yang menjadi dasar penilaian SNBP.

Kesimpulan

Penelitian ini membandingkan empat algoritma supervised learning Decision Tree, Random Forest, Naïve Bayes, dan k-NN dalam memprediksi jalur seleksi masuk mahasiswa Jurusan Teknik Informatika Universitas Negeri Gorontalo berdasarkan 1.232 data mahasiswa dari SIAT. Hasil 10-fold cross-validation menunjukkan Random Forest mencatat akurasi rata-rata tertinggi (64,61%), diikuti Decision Tree (64,37%), Naïve

Bayes (63,47%), dan k-NN (61,03%). Namun, uji Friedman membuktikan bahwa perbedaan performa antar keempat algoritma tersebut tidak signifikan secara statistik ($\chi^2 = 4,39$; $p = 0,222$), sehingga tidak dapat disimpulkan adanya algoritma yang secara definitif unggul untuk kasus ini.

Analisis lebih lanjut mengungkap bahwa performa sempurna pada kelas Mandiri merupakan artefak dependensi definisional dengan fitur seleksi, bukan cerminan kapabilitas prediktif model. Tantangan klasifikasi yang sesungguhnya terletak pada pembedaan kelas SNBP dan SNBT yang sama-sama berjalur seleksi Nasional. Penerapan SMOTE terbukti secara konsisten meningkatkan recall kelas minoritas SNBP pada keempat algoritma, tetapi efeknya terhadap F1-Score makro dan akurasi global tidak seragam meningkat pada tiga algoritma namun sedikit menurun pada Naïve Bayes menegaskan pentingnya evaluasi multi-metrik dan per-algoritma pada dataset tidak seimbang, alih-alih menyimpulkan SMOTE sebagai solusi tunggal yang berlaku umum.

Penelitian ini memiliki beberapa keterbatasan yang perlu diakui. Pertama, fitur yang tersedia terbatas pada atribut demografis dan administratif dari SIAT, tanpa mencakup data prestasi akademik (nilai rapor, skor ujian) yang secara substantif mendasari keputusan penerimaan SNBP dan SNBT. Kedua, tiga kelas jalur berukuran sangat kecil (ADik, Kerjasama, Jurusan; total 5 sampel) harus dikeluarkan dari analisis karena tidak memenuhi syarat minimum pengujian statistik. Penelitian selanjutnya perlu memperkaya dataset dengan fitur akademik dan perilaku (nilai rapor, skor tes, partisipasi ekstrakurikuler) untuk meningkatkan kemampuan diskriminasi model, khususnya dalam membedakan jalur seleksi Nasional (SNBP dan SNBT) yang menjadi sumber utama kesalahan klasifikasi pada penelitian ini.

Daftar Pustaka

Arévalo-Cordovilla, F. E., & Peña, M. (2024). Comparative Analysis of Machine Learning Models for Predicting Student Success in Online Programming Courses: A Study Based on LMS Data and External Factors. Mathematics, 12(20), 3272. https://doi.org/10.3390/math12203272

Azizah, Z., Ohyama, T., Zhao, X., Ohkawa, Y., & Mitsuishi, T. (2024). Predicting at-risk students in the early stage of a blended learning course via machine learning using limited data. Computers and Education: Artificial Intelligence, 7, 100261. https://doi.org/10.1016/j.caeai.2024.100261

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

Elbouknify, I., Berrada, I., Mekouar, L., Iraqi, Y., Bergou, E. H., Belhabib, H., Nail, Y., & Wardi, S. (2025). AI-based identification and support of at-risk students: A case study of the Moroccan education system [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2504.07160

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. Journal of the American Statistical Association, 32(200), 675-701. https://doi.org/10.2307/2279372

Indra, & Agustinawati, D. (2024). Perbandingan Metode Data Mining pada Prediksi Kelulusan Mahasiswa Fakultas Teknologi Informasi di Perguruan Tinggi dengan Algoritma Naive Bayes dan K-Nearest Neighbor. SIMETRIS, 15(1), 69-84.

Matz, S. C., Bukow, C. S., Peters, H., et al. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. Scientific Reports, 13, 5705. https://doi.org/10.1038/s41598-023-32484-w

Oppong, S. (2023). Predicting Students' Performance Using Machine Learning Algorithms: A Review. Asian Journal of Research in Computer Science, 16, 128-148. https://doi.org/10.9734/AJRCOS/2023/v16i3351

Sathe, M. T., & Adamuthe, A. C. (2021). Comparative Study of Supervised Algorithms for Prediction of Students' Performance. International Journal of Modern Education and Computer Science, 13(1), 1-19. https://doi.org/10.5815/ijmecs.2021.01.01

Villegas-Ch, W., Govea, J., & Revelo-Tapia, S. (2023). Improving Student Retention in Institutions of Higher Education through Machine Learning: A Sustainable Approach. Sustainability, 15, 14512. https://doi.org/10.3390/su151914512

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9, 11. https://doi.org/10.1186/s40561-022-00192-z