DATAMATIKA

Tersedia secara daring di:
https://conference.ipb.ac.id/datamatika

IMPROVING XGBOOST PERFORMANCE FOR CLASSIFYING IMPORTANT VARIABLES OF SNBP STUDENT GPA AT IPB USING EASYENSEMBLE

PENINGKATAN KINERJA XGBOOST DALAM KLASIFIKASI PEUBAH PENTING INDEKS PRESTASI MAHASISWA SNBP IPB UNIVERSITY BERBASIS *EASYENSEMBLe*

Salsabila Fayiza‡, Utami Dyah Syafitri, Aulia Rizki Firdawanti

[1]Program Studi Statistika dan Sains Data, SSMI, IPB University, Indonesia
‡corresponding author: fyzsalsabila@apps.ipb.ac.id

Abstract

The implementation of the Merdeka Curriculum emphasizes flexible and competency based learning, resulting in students with increasingly diverse academic backgrounds. This condition is closely related to the SNBP admission scheme, which selects students based on academic performance without a written examination. This study aims to develop an integrated approach by combining XGBoost, EasyEnsemble, and 3^{4-1} Fractional Factorial Design (FFD) based hyperparameter tuning to improve the classification of important variables affecting the first semester GPA of SNBP students at IPB University. GPA data are categorized into three class and four class schemes and evaluated using K-Fold Cross Validation. Model performance is measured using accuracy, precision, recall, F1-score, and G-Mean, while feature importance based on XGBoost gain values is used to identify influential variables. The results show that the three class XGBoost model with EasyEnsemble achieves the best performance, with an accuracy of 90.55% and G-Means of 58.18%, and effectively handles class imbalance, particularly for minority classes. The most influential variables include overall average grades, average required grades, study program, school province, and “PPKU” cluster. Overall, the proposed approach demonstrates strong potential for handling imbalanced classification problems and supporting data driven decision making.

Keywords: Class Imbalance, Classification, EasyEnsemble, Feature Importance, XGBoost.

Abstrak

Implementasi Kurikulum Merdeka menekankan pembelajaran yang fleksibel dan berbasis kompetensi, sehingga menghasilkan mahasiswa dengan latar belakang akademik yang semakin beragam. Kondisi ini berkaitan erat dengan skema penerimaan SNBP yang menyeleksi mahasiswa berdasarkan prestasi akademik tanpa melalui ujian tertulis. Penelitian ini bertujuan untuk mengembangkan pendekatan terintegrasi dengan mengombinasikan metode XGBoost, EasyEnsemble, serta tuning hyperparameter berbasis Fractional Factorial Design 3^{4-1} untuk meningkatkan kinerja klasifikasi peubah penting yang memengaruhi IP semester 1 mahasiswa SNBP di IPB University. Data IP dikategorikan ke dalam skema tiga kelas dan empat kelas, serta dievaluasi menggunakan K-Fold Cross Validation. Kinerja model diukur menggunakan metrik accuracy, precision, recall, F1-score, dan G-Mean, sementara feature importance berdasarkan nilai gain XGBoost digunakan untuk mengidentifikasi peubah yang berpengaruh. Hasil penelitian menunjukkan bahwa model XGBoost dengan EasyEnsemble pada skema tiga kelas memberikan kinerja terbaik, dengan nilai accuracy sebesar 90,55% dan G-Mean sebesar 58,18%, serta mampu menangani ketidakseimbangan kelas dengan baik, khususnya pada kelas minoritas. Peubah yang paling berpengaruh meliputi nilai rata-rata keseluruhan, rata-rata nilai mata pelajaran wajib, program studi, provinsi sekolah, dan klaster PPKU. Secara keseluruhan, pendekatan yang diusulkan menunjukkan potensi yang kuat dalam menangani permasalahan klasifikasi tidak seimbang serta mendukung pengambilan keputusan berbasis data.

Kata Kunci: EasyEnsemble, Ketidakseimbangan Kelas, Klasifikasi, Peubah Penting, XGBoost.

1. Pendahuluan

Perubahan kebijakan pendidikan dari Kurikulum 2013 ke Kurikulum Merdeka merupakan upaya pemerintah untuk menyesuaikan pembelajaran dengan kebutuhan peserta didik yang semakin beragam. Kurikulum Merdeka memberikan fleksibilitas pembelajaran, penguatan kompetensi esensial, serta pengembangan karakter, sehingga menghasilkan lulusan dengan latar belakang akademik yang lebih beragam. Sejak tahun ajaran 2022/2023, Kurikulum Merdeka mulai diterapkan secara luas di satuan pendidikan di Indonesia (Addini *et al.,* 2024). Perubahan kurikulum ini berkaitan dengan sistem Seleksi Nasional Berdasarkan Prestasi (SNBP) yang menilai capaian akademik siswa berdasarkan nilai rapor selama pendidikan menengah. Mahasiswa SNBP angkatan 2025 IPB University dipilih sebagai objek penelitian karena mayoritas berasal dari sekolah yang telah menerapkan Kurikulum Merdeka. Berdasarkan laporan Kemendikbudristek (2024), lebih dari 90% satuan pendidikan formal di

Indonesia telah mengimplementasikan Kurikulum Merdeka pada tahun ajaran 2024/2025. Keberagaman latar belakang akademik lulusan ini berpotensi memengaruhi capaian akademik mahasiswa di perguruan tinggi, yang dalam penelitian ini diukur melalui Indeks Prestasi (IP). Data IP mahasiswa memiliki karakteristik yang kompleks dan berpotensi tidak seimbang antar kelas, sehingga diperlukan pendekatan *machine learning* yang mampu menangani kondisi tersebut secara efektif (Abdullah dan Prasetyo 2020).

XGBoost merupakan algoritma *ensemble learning* yang banyak digunakan dalam klasifikasi karena efisiensi komputasi yang baik serta kemampuannya menangani data berdimensi tinggi dan mengidentifikasi kontribusi setiap peubah terhadap hasil prediksi, sehingga sangat berguna dalam analisis peubah penting yang memengaruhi IP mahasiswa (Mahesh *et al.,* 2022). Namun, XGBoost memiliki keterbatasan dalam menangani data yang tidak seimbang karena model cenderung lebih berpihak pada kelas mayoritas, sehingga menurunkan kinerja klasifikasi pada kelas minoritas (Wiens *et al.,* 2025). Salah satu pendekatan untuk mengatasi ketidakseimbangan kelas adalah *ensemble learning* (Zhang *et al.,* 2022). Teknik yang umum digunakan untuk menangani ketidakseimbangan kelas adalah SMOTE, namun metode ini berpotensi menghasilkan data tambahan yang kurang mencerminkan kondisi sebenarnya dan meningkatkan *overfitting* (Zhang *et al.,* 2022). Sebagai alternatif, *EasyEnsemble* merupakan metode *ensemble* berbasis *undersampling* pada kelas mayoritas tanpa menambah data baru, sehingga lebih stabil dalam meningkatkan sensitivitas model terhadap kelas minoritas (Abdullah dan Prasetyo 2020). Penelitian sebelumnya oleh (Syifa *et al.,* 2025), menggunakan data mahasiswa jalur SNBP dengan penanganan ketidakseimbangan kelas dilakukan menggunakan SMOTE, namun kinerja XGBoost yang dihasilkan belum optimal, khususnya dalam mengklasifikasikan kelas minoritas. Sejalan dengan permasalahan tersebut, penelitian (Rais *et al.,* 2024), membandingkan metode SMOTE dan *EasyEnsemble* dalam menangani data tidak seimbang dan menunjukkan bahwa *EasyEnsemble* menghasilkan kinerja model yang lebih baik. Temuan ini menunjukkan bahwa *EasyEnsemble* berpontensi menjadi metode yang lebih efektif untuk meningkatkan kemampuan model dalam mengklasifikasikan kelas minoritas. Oleh karena itu, *EasyEnsemble* dipilih untuk mendukung peningkatan kinerja XGBoost dalam klasifikasi peubah penting yang memengaruhi IP mahasiswa. Penelitian ini bertujuan untuk menerapkan metode *EasyEnsemble* untuk meningkatkan kinerja model XGBoost pada data tidak seimbang dan mengidentifikasi peubah-peubah yang penting terhadap klasifikasi Indeks Prestasi Semester 1 mahasiswa. Hasil penelitian ini diharapkan dapat mendukung pengembangan sistem seleksi mahasiswa baru IPB University.

2. Metodologi

2.1 Data

Data yang digunakan merupakan data primer IP mahasiswa program sarjana angkatan 62 jalur SNBP pada tahun 2025 yang diperoleh dari Direktorat Administrasi Pendidikan dan Penerimaan Mahasiswa Baru (DAPPMB) IPB University. Penjelasan untuk masing-masing peubah yang digunakan pada penelitian ini disajikan pada Tabel 1.

Tabel 1: Peubah yang digunakan pada penelitian

Notasi	Nama Peubah	Jenis	Keterangan Peubah
Y	IP mahasiswa IPB 62	Numerik	IP semester 1 mahasiswa, dengan rentang $0 \leq Y \leq 4$
X_1	Rataan rapor total	Numerik	Nilai rata-rata seluruh mata pelajaran rapor SMA/MA, dengan rentang $0 \leq X_1 \leq 100$
X_2	Rataan rapor syarat	Numerik	Nilai rata-rata mata Pelajaran yang menjadi syarat program studi yang dipilih, dengan rentang $0 \leq X_2 \leq 100$
X_3	Jenis sekolah	Kategorik	SMA, MA
X_4	Jurusan sekolah	Kategorik	IPA, IPS, Umum
X_5	Kurikulum sekolah	Kategorik	K13, Merdeka
X_6	Asal provinsi sekolah	Kategorik	Jawa Barat, DKI Jakarta, Banten, Jawa Selainnya, Sumatera Barat, Sumatera Utara, lainnya
X_7	Klaster sekolah	Kategorik	1, 2, 3, Tidak Terdaftar (Anggraini *et al.,* 2024)
X_8	Jenis kelamin	Kategorik	Laki-laki, Perempuan
X_9	Program studi diterima	Kategorik	A01, A11, A21, A31, A41, B01, B11, C11, C21, C31, C41, C51, D11, D21, D31, E11, E21, E31, E41, F11, F21, F31, F41, G21, G31, G41, G71, G81, H11, H21, H31, H41, H51, I11, I21, I31, K11, L11, M01, M02, M03, M04
X_{10}	Klaster PPKU	Kategorik	ST, SS, SE

2.2 *Extreme Gradient Boosting (*XGBoost*) Classifier*

XGBoost merupakan algoritma *ensemble learning* berbasis *gradient boosting* yang menggunakan pohon keputusan. XGBoost membangun model secara bertahap dan menggabungkan sejumlah pohon untuk memperbaiki kesalahan prediksi sebelumnya, sehingga mampu meningkatkan prediksi dan efisiensi komputasi. Proses pembelajaran pada

XGBoost dilakukan melalui optimalisasi fungsi objektif yang terdiri atas *loss function* dan komponen regularisasi untuk mengendalikan kompleksitas model dan mengurangi risiko *overfitting*, sebagaimana disajikan pada persamaan (1) (Chen dan Guestrin 2016):

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_{i=1}^{t} \Omega(f_t) \quad (1)$$

dimana $l\ (y_i, \hat{y}_i^{(t-1)})$ merupakan *loss function* yang mengukur selisih antara nilai aktual (y_i) dan nilai prediksi $(\hat{y}_i)$. Nilai $f_t(x_i)$ merupakan fungsi pohon regresi pada iterasi ke-*t*. Bentuk aditif ini menunjukkan setiap pohon baru berperan memperbaiki kesalahan prediksi sebelumnya. Untuk menjaga agar model tidak terlalu kompleks, XGBoost menambahkan fungsi regularisasi yang ditunjukkan pada persamaan (2).

$$\Omega(f_t) = \gamma T_t + \frac{1}{2}\lambda \parallel w_t \parallel^2 \quad (2)$$

dengan γ berfungsi mengontrol jumlah daun T_t pada pohon regresi, sedangkan λ merupakan parameter regularisasi L2 terhadap bobot daun w_t. Regularisasi ini membantu menghasilkan model yang lebih stabil dan memiliki kemampuan generalisasi yang lebih baik.

2.3 *EasyEnsemble*

EasyEnsemble merupakan metode penanganan ketidakseimbangan data berbasis *ensemble undersampling* yang diperkenalkan oleh Liu *et al.,* (2009). Metode ini dirancang untuk meningkatkan kinerja klasifikasi pada data tidak seimbang dengan mengurangi kesalahan model terhadap kelas mayoritas serta meningkatkan kemampuan model dalam mengenali kelas minoritas (Siregar dan Arifin, 2024).

EasyEnsemble bekerja dengan membentuk beberapa subset data seimbang melalui proses *undersampling* pada kelas mayoritas, dimana setiap subset dikombinasikan dengan seluruh data kelas minoritas. Selanjutnya, model dilatih pada masing-masing subset secara terpisah, kemudian hasil prediksi dari seluruh model digabungkan untuk memperoleh keputusan akhir. Pendekatan ini memungkinkan pemanfaatan data secara lebih optimal tanpa menghilangkan informasi penting dari kelas mayoritas.

2.4 Metode Penelitian

Penelitian dilakukan menggunakan *software Rstudio* dengan tahapan analisis sebagai berikut.

1. Praproses data
   a. Pemeriksaan data hilang dan ketidaksesuaian data.
   b. Pengelompokan IP
      Data IP dikelompokkan ke dalam menjadi 2 skema berikut :
      1) Skema pertama melakukan pengelompokan IP dalam tiga kelas

sebagaimana yang tertera pada Tabel 2.

Tabel 2: Kategorisasi IP skema pertama (3 kelas)

Kelas	Rentang Indeks Prestasi (IP)
0	$2.10 \leq IP\ Semester\ 1 < 2.75$
1	$2.75 \leq IP\ Semester\ 1 < 3.50$
2	$IP\ Semester\ 1 \geq 3.50$

2) Skema kedua melakukan pengelompokan IP dalam empat kelas sebagaimana yang tertera pada Tabel 3.

Tabel 3: Kategorisasi IP skema kedua (4 kelas)

Kelas	Rentang Indeks Prestasi (IP)
0	$2.10 \leq IP\ Semester\ 1 < 2.75$
1	$2.75 \leq IP\ Semester\ 1 < 3.00$
2	$3.00 \leq IP\ Semester\ 1 < 3.50$
3	$IP\ Semester\ 1 \geq 3.50$

2. Eksplorasi data
   Eksplorasi dilakukan secara deskriptif dan visual untuk memahami karakteristik data, sebaran IP mahasiswa, serta proporsi kelas pada setiap skema pengkategorian IP untuk mengidentifikasi adanya ketidakseimbangan kelas.
3. *Hyperparameter Tuning* dengan 3^{4-1} *Fractional Factorial Design*
   Fractional Factorial Design (FFD) merupakan metode dalam *design of experiments* yang digunakan untuk mengurangi jumlah kombinasi percobaan dari *full factorial design* tanpa menghilangkan informasi utama. Dalam penelitian ini, FFD digunakan untuk menentukan *hyperparameter* yang berpengaruh signifikan terhadap kinerja model klasifikasi XGBoost, dengan memperlakukan setiap *hyperparameter* sebagai faktor dengan beberapa level dan membentuk kombinasi *hyperparameter* berdasarkan rancangan FFD sehingga proses *tuning* dapat dilakukan secara lebih efisien (Montgomery 2013). Rancangan yang digunakan adalah fraksi $\frac{1}{3}$ dari FFD 3^4, sehingga menghasilkan 27 kombinasi *hyperparameter* yang lebih efisien secara komputasi. Empat *hyperparameter* utama, yaitu *nrounds, max_depth, eta, dan gamma*, masing-masing dengan tiga level, sementara tiga *hyperparameter* turunan ditetapkan pada satu nilai untuk menjaga efisiensi dan fokus analisis. Setiap kombinasi dijalankan sebanyak tiga kali pengulangan untuk meningkatkan kestabilan estimasi galat. Nilai *hyperparameter* yang digunakan sebagai berikut:
   - *Nrounds* dengan nilai 100, 300, dan 500.
   - Max_depth dengan nilai 3, 6, dan 9.
   - Eta dengan nilai 0,01; 0,1; dan 0,3.
   - Gamma dengan nilai 0, 1, dan 5.

Respon yang dianalisis adalah akurasi model XGBoost dan hasil FFD dievaluasi menggunakan *Analysis of Variance* (ANOVA) untuk menentukan *hyperparameter* yang berpengaruh signifikan serta memperoleh kombinasi *hyperparameter* terbaik.

4. Pembagian data
   Pembagian data dilakukan menggunakan *K-Fold Cross-Validation* menjadi 5 bagian (*k*=5). Pada setiap iterasi, 4 *fold* sebagai data latih dan 1 *fold* sisanya sebagai data uji.
5. Penanganan ketidakseimbangan data dengan *EasyEnsemble*
   Metode ini bekerja dengan cara sebagai berikut:
   a. Mengambil sebagian data dari kelas mayoritas secara acak untuk membentuk beberapa subset data latih.
   b. Menggabungkan setiap subset data mayoritas dengan seluruh data dari kelas minoritas sehingga terbentuk data latih yang seimbang.
   c. Membangun model klasifikasi dasar pada setiap subset data seimbang.
   d. Menggabungkan hasil prediksi dari seluruh model melalui mekanisme *ensemble* untuk memperoleh prediksi akhir.
6. Perbandingan pemodelan Klasifikasi XGBoost
   Pemodelan klasifikasi dilakukan menggunakan metode dengan *hyperparameter* terbaik hasil FFD untuk membentuk empat model, yaitu:
   a. Model 1 : skema 1 (3 kelas) tanpa penanganan ketidakseimbangan data.
   b. Model 2 : skema 1 (3 kelas) dengan *EasyEnsemble*.
   c. Model 3 : skema 2 (4 kelas) tanpa penanganan ketidakseimbangan data.
   d. Model 4 : skema 2 (4 kelas) dengan *EasyEnsemble*.
7. Evaluasi model
   Tingkat kebaikan model dilihat berdasarkan matriks konfusi pada data uji. Kinerja diukur berdasarkan nilai akurasi, presisi, sensitivitas, skor F1, serta *G-Mean* yang digunakan untuk menilai keseimbangan model.
8. Identifikasi *Feature Importance*
   Feature importance merupakan ukuran untuk menilai kontribusi relatif setiap peubah terhadap hasil prediksi model klasifikasi (Tian *et al.,* 2022). Pengukuran ini didasarkan pada nilai *gain*, yaitu besarnya penurunan *loss function* ketika suatu peubah digunakan sebagai pemisah. Semakin besar nilai *gain*, semakin besar penting peran peubah tersebut dalam meningkatkan kinerja model. Identifikasi *Feature Importance* dilakukan menggunakan model XGBoost terbaik, dengan peubah dengan nilai *gain* tertinggi sebagai yang paling penting dalam klasifikasi IP mahasiswa.
9. Penarikan Kesimpulan

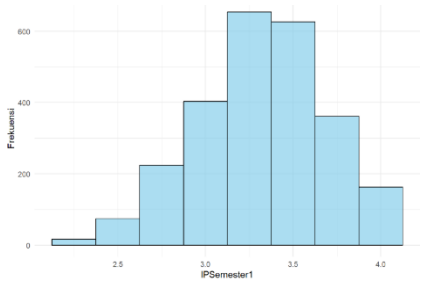
3. Hasil dan Pembahasan

3.1 Praproses Data

Praproses data dalam penelitian ini diawali dengan pemeriksaan kelengkapan dan kesesuaian data. Hasil pengecekan menunjukkan bahwa tidak terdapat data hilang maupun ketidaksesuaian data. Selain itu, berdasarkan peubah jurusan, dibentuk peubah kurikulum sekolah, dengan mayoritas mahasiswa berasal dari Kurikulum Merdeka sebesar 64.1%. Selanjutnya, dilakukan pengelompokan peubah IP ke dalam dua skema, yaitu skema tiga kelas dan empat kelas sesuai dengan rentang nilai yang telah ditentukan.

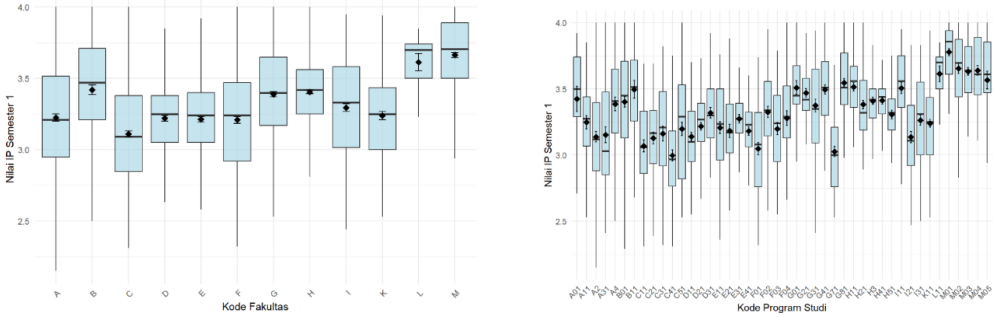
3.2 Eksplorasi Data

Sebaran IP semester 1 mahasiswa angkatan 2025 program S1 jalur masuk SNBP ditampilkan pada Gambar 1.



Gambar 1. Histogram sebaran IP semester 1 mahasiswa S1 SNBP angkatan 2025

Terlihat dari histogram tersebut bahwa mahasiswa S1 SNBP angkatan 2024 dominan memiliki IP semester 1 di antara 3.00 sampai 3.50. Selain itu, didapatkan pula bahwa rata-rata dari IP semester 1 untuk masing-masing Fakultas dan Kode Program Studi Diterima dapat dilihat pada Gambar 2.



(a) (b)

Gambar 2. (a) Sebaran IP semester 1 berdasarkan fakultas dan (b) program studi diterima

Berdasarkan boxplot pada Gambar 2 (a), dapat dilihat bahwa kode fakultas M (Sekolah Sains Data, Matematika, dan Informatika) merupakan fakultas dengan mahasiswa yang cenderung memiliki capaian akademik

awal paling tinggi. Hal tersebut dapat dilihat dari median dan rata-rata IP fakultas M yang berada di sekitar angka 3.8. Sementara itu, fakultas C merupakan fakultas yang memiliki capaian akademik awal terendah dengan median dan rata-rata berada di sekitar angka 3.1. Selain itu, berdasarkan Kode Program Studi terlihat bahwa program studi terlihat bahwa program studi M01 (Statistika dan Sains Data) merupakan program studi dengan mahasiswa yang memiliki IP semester 1 tertinggi. Disamping itu, keempat program studi di lingkup fakultas M juga sangat baik dalam nilai IP semester 1. Hal ini mendukung eksplorasi sebaran IP semester 1 berdasarkan kode fakultas sebelumnya. Di sisi lain, kode program studi C41 (Teknologi dan Manajemen Perikanan Tangkap), G71 (Fisika), dan F01 (Teknik Pertanian dan Biosistem) merupakan tiga program studi dengan rata-rata IP semester 1 terendah.

Eksplorasi juga dilakukan berdasarkan peubah numerik yang digunakan, yaitu nilai rataan total dan rataan syarat terhadap IP semester 1. Kedua peubah nilai rataan tersebut memiliki hubungan linear positif terhadap IP semester 1. Korelasi positif ini menunjukkan bahwa semakin tinggi rataan siswa, maka cenderung diikuti oleh peningkatan IP semester 1. Namun kekuatan hubungan keduanya masih tergolong lemah dengan nilai korelasi 0.195 dan 0.248 untuk rataan syarat.

3.3 Pemodelan Klasifikasi dengan XGBoost

3.3.1 Pengkategorian Peubah Target dan Pembagian Data dengan *Cross Validation*

Peubah target yaitu IP semester 1 dikategorikan menjadi 2 skema seperti yang tertulis pada bagian prosedur analisis. Selanjutnya, pembagian data dilakukan dengan metode 5-*fold Cross-Validation*. Data dibagi menjadi 5 *fold* di mana dalam setiap iterasi, 4 *fold* digunakan sebagai data latih, sementara 1 *fold* lainnya digunakan sebagai data uji.

3.3.2 *Hyperparameter Tuning* dengan 3^{4-1} *Fractional Factorial Design*

Hasil signifikansi *hyperparameter tuning* berdasarkan ANOVA dapat dilihat pada Tabel 4.

Tabel 4: Hasil signifikansi *hyperparameter*

		Hyperparameter			
		nrounds	*max_depth*	*eta*	*gamma*
P-Value	Skema 1	0.2933	0.9710	0.1503	0.0075**
	Skema 2	0.4732	0.3430	0.5978	0.0162*

Berdasarkan hasil pada Tabel 4, sebagian *hyperparameter* yang diuji memiliki nilai *p-value* lebih besar dari 0.05, sehingga tidak memberikan pengaruh yang signifikan terhadap performa model. Pada skema 1, hanya gamma yang menunjukkan pengaruh signifikan dengan *p-value* sebesar 0.0062, sedangkan *hyperparameter* lainnya tidak signifikan. Sementara itu,

pada skema 2 seluruh *hyperparameter* yang diuji tidak berpengaruh signifikan terhadap performa model. Hal tersebut menunjukkan bahwa model XGBoost telah mampu membentuk pola klasifikasi yang relatif stabil pada berbagai kombinasi *hyperparameter* yang digunakan. Selanjutnya dilakukan pencarian kombinasi *best hyperparameter* untuk kedua skema yang ditampilkan pada Tabel 5.

Tabel 5: Hasil signifikansi *hyperparameter*

		Hyperparameter			
		nrounds	*max_depth*	*eta*	*gamma*
Nilai Terbaik	Skema 1	100	9	0.3	1
	Skema 2	200	6	0.1	1

3.3.3 Penanganan Ketidakseimbangan Data dengan *EasyEnsemble*

Data IP semester 1 yang digunakan merupakan data yang tidak seimbang, di mana kelas mayoritas terletak pada kelas 1 skema 1 dan terletak pada kelas 2 pada skema 2. Kondisi tersebut berpotensi menyebabkan model lebih dominan mempelajari pola pada kelas mayoritas. Oleh karena itu, dilakukan penanganan ketidakseimbangan data menggunakan metode *EasyEnsemble* melalui proses *undersampling* pada kelas mayoritas. Tabel 6 menunjukkan komposisi jumlah dan persentase data latih sebelum dan sesudah penerapan metode *EasyEnsemble*.

Tabel 6: Komposisi data latih sebelum dan sesudah *EasyEnsemble*

Kelas IP Semester 1	Skema 1		Kelas IP Semester 1	Skema 2	
	Sebelum	Sesudah		Sebelum	Sesudah
0	143 (7.09%)	143 (33.33%)	0	143 (7.09%)	143 (25%)
1	1194 (59.23%)	143 (33.33%)	1	243(12.04%)	143 (25%)
2	679 (33.68%)	143 (33.33%)	2	952(47.18%)	143 (25%)
			3	680(33.70%)	143 (25%)

Berdasarkan Tabel 6, sebelum dilakukan *EasyEnsemble*, distribusi data latih pada kedua skema masih didominasi oleh kelas mayoritas. Setelah dilakukan *EasyEnsemble*, jumlah data pada setiap kelas menjadi seimbang karena proses *undersampling* mengikuti jumlah data pada kelas minoritas. Pada skema 1, setiap kelas memiliki proporsi sebesar 33,33%, sedangkan pada skema 2 masing-masing kelas memiliki proporsi sebesar 25%. Hal tersebut menunjukkan bahwa metode *EasyEnsemble* mampu menghasilkan distribusi data latih yang lebih seimbang pada setiap kelas.

3.3.4 Pelatihan Model XGBoost dan Evaluasi

Pelatihan model dilakukan pada keempat model yang ditetapkan sebelumnya. XGBoost bekerja dengan prinsip *boosting*, yaitu membangun model secara bertahap dengan memperbaiki kesalahan dari model sebelumnya, sehingga mampu meningkatkan akurasi prediksi. Proses klasifikasian dilakukan menggunakan nilai *hyperparameter* terbaik yang telah diperoleh untuk masing-masing skema. Selanjutnya, kinerja model dievaluasi menggunakan metrik yang telah disebutkan sebelumnya untuk menilai kemampuan prediksi terhadap kelas IP semester 1. Evaluasi model dilakukan dengan mempertimbangkan beberapa metrik performa, yang dihitung berdasarkan rataan seluruh fold. Hasil evaluasi setiap model disajikan pada Tabel 7.

Tabel 7: Metrik evaluasi

Model	Akurasi	Presisi	Sensitivitas	Skor F1	*G-Mean*
1	84.96%	47.83%	43.37%	47.38%	44.12%
2	90.55%	42.79%	47.44%	39.98%	58.18%
3	70.75%	39.80%	33.79%	37.57%	39.47%
4	77.54%	34.21%	37.18%	33.22%	53.34%

Berdasarkan Tabel 7, model 2 yang merupakan skema 1 (3 kelas) dengan penerapan *EasyEnsemble* menunjukkan performa terbaik dibandingkan model lainnya. Hal ini terlihat dari nilai akurasi tertinggi sebesar 90.55%, yang menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar. Selain itu, model 2 juga memiliki nilai *G-Mean* tertinggi sebesar 58.18%, yang menunjukkan bahwa model mampu menjaga keseimbangan performa klasifikasi antar kelas, termasuk pada kelas minoritas. Nilai sensitivitas sebesar 47.44% yang juga menjadi tertinggi menunjukkan bahwa model lebih baik dalam mengenali data pada kelas yang sebenarnya. Meskipun skor F1 model 2 sebesar 39.98% bukan yang tertinggi, nilai tersebut tetap menunjukkan keseimbangan yang cukup baik antara presisi dan sensitivitas.

Di sisi lain, model 1 memiliki nilai skor F1 tertinggi yaitu 47.38% serta presisi tertinggi sebesar 47.83%. Hal ini menunjukkan bahwa model 1 lebih baik dalam hal ketepatan prediksi. Namun, nilai *G-Mean* yang lebih rendah sebesar 44.12% menunjukkan bahwa performanya kurang seimbang antar kelas, terutama dalam menangani kelas minoritas. Sementara itu, model 3 memiliki nilai yang lebih rendah pada hampir semua metrik dibandingkan model lainnya, sehingga dapat dikatakan bahwa performanya masih kurang optimal. Model 4 juga menggunakan *EasyEnsemble* pada skema 4 kelas menunjukkan adanya peningkatan dibandingkan model 3, terutama pada *G-Mean* yang mencapai 53.34%. Hal ini menunjukkan bahwa model mulai mampu manangani ketidakseimbangan data dengan lebih baik. Namun, nilai presisinya yang paling rendah sebesar 34.21%, menunjukkan bahwa ketepatan prediksi model masih perlu ditingkatkan.

Tingkat kebaikan model juga diukur dari *confusion matrix* yang terbentuk pada data uji. Model yang dipilih untuk identifikasi peubah

penting di tahap selanjutnya adalah model 2. Adapun *confusion matrix* dari model 2 dapat dilihat pada Tabel 8 berikut.

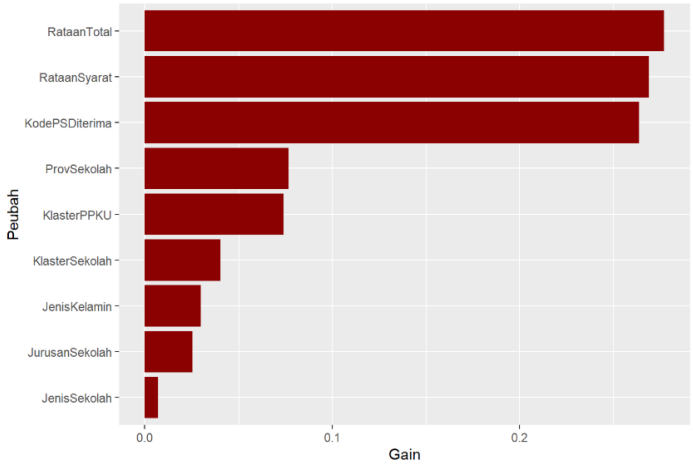
Tabel 8: *Confusion matrix* hasil prediksi dari model 2.

Kelas Aktual	Kelas Prediksi		
	0	1	2
0	16	10	10
1	87	123	88
2	21	41	108

Berdasarkan *confusion matrix* pada Tabel 8, model 2 mampu mengklasifikasikan data dengan cukup baik, terutama pada kelas 1 dan kelas 2 yang memiliki jumlah prediksi benar tertinggi, masing-masing sebanyak 123 dan 108 data. Namun, pada kelas 0 performa model masih lebih rendah dengan hanya 16 prediksi benar, serta masih terdapat cukup banyak kesalahan klasifikasi antar kelas, khususnya pada kelas 1 yang sering tertukar dengan kelas lain. Secara keseluruhan, model sudah cukup baik dalam mengenali pola pada kelas mayoritas, tetapi masih perlu peningkatan dalam membedakan kelas minoritas agar hasil klasifikasi lebih seimbang.

3.3.5 Tingkat Kepentingan Peubah

Model yang dipilih untuk mengidentifikasi peubah penting dalam klasifikasi IP Semester 1 mahasiswa S1 SNBP 2025 adalah model 2. Berikut merupakan peubah penting yang diperoleh dari model 2.



Gambar 4. *Feature importance* dalam klasifikasi IP Semester 1

Berdasarkan grafik *feature importance* pada Gambar 4, terlihat bahwa peubah rataan total, rataan syarat, dan program studi diterima merupakan peubah yang paling penting dalam proses pengklasifikasian IP Semester 1 pada model XGBoost. Kode Program Studi Diterima memiliki nilai gain tertinggi, yang menunjukkan bahwa karakteristik program studi memberikan kontribusi terbesar dalam pembentukan keputusan model. Sementara itu,

tingginya nilai gain pada rataan syarat dan rataan total mengindikasikan bahwa kemampuan akademik sebelum memasuki perguruan tinggi masih menjadi faktor utama yang memengaruhi capaian IP Semester 1 mahasiswa. Di sisi lain, peubah seperti provinsi sekolah, klaster PPKU, dan klaster sekolah memiliki pengaruh yang relatif lebih kecil, sedangkan jenis kelamin, jurusan sekolah, dan jenis sekolah menunjukkan kontribusi yang rendah terhadap performa model. Secara keseluruhan, hasil *feature importance* menunjukkan bahwa faktor akademik lebih dominan dibandingkan faktor demografis dalam menentukan klasifikasi IP Semester 1 mahasiswa SNBP IPB University.

4. Simpulan dan Saran

Model yang dipilih untuk mengidentifikasi peubah penting dalam klasifikasi IP semester 1 mahasiswa S1 SNBP IPB angkatan 2025 adalah model klasifikasi XGBoost 3 kelas dengan penerapan *EasyEnsemble.* Model tersebut memberikan performa terbaik dibandingkan model lainnya dengan akurasi sebesar 90.55% dan nilai *G-Mean* tertinggi sebesar 58.18%. Hasil ini menunjukkan bahwa penerapan *EasyEnsemble* mampu membantu model dalam menangani ketidakseimbangan data serta meningkatkan kemampuan klasifikasi pada setiap kelas, termasuk kelas minoritas. Selain itu, tiga peubah yang paling penting dalam klasifikasi IP semester 1 mahasiswa adalah rataan total, rataan syarat, dan program studi diterima. Hasil penelitian ini menunjukkan bahwa kombinasi metode *EasyEnsemble* dan XGBoost dapat digunakan sebagai pendekatan yang baik untuk menangani permasalahan klasifikasi pada data yang tidak seimbang. Penelitian selanjutnya disarankan untuk mengeksplorasi *hyperparameter* yang lebih beragam agar performa model dapat dioptimalkan lebih lanjut. Selain itu, metode penanganan ketidakseimbangan data lainnya dapat digunakan sebagai pembanding terhadap metode *EasyEnsemble* untuk memperoleh model klasifikasi yang lebih baik.

Daftar Pustaka

Berikut adalah daftar pustaka tersebut yang telah diubah dari format Harvard ke format **APA (American Psychological Association) 7th Edition**:

Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia. (2024). *Implementasi Kurikulum Merdeka dan Peningkatan Kemampuan Literasi dan Numerasi Peserta Didik*. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi.

Abdullah, S., & Prasetyo, G. (2020). Easy Ensemble with Random Forest to handle imbalanced data in classification. *Jurnal Fundamental and Applied Mathematics*, 3(1), 39–46. https://doi.org/10.14710/jfma.v3i1.7415

Addini, P. F., Dakhi, K. R. S., & Tarigan, P. C. (2024). Identifikasi faktor-faktor yang mempengaruhi prestasi mahasiswa teknik informatika dan bisnis digital menggunakan Regresi Logistik Binary. *Jurnal Sains dan Teknologi*, 5(3), 922–925. https://doi.org/10.55338/saintek.v5i1.2792

Anggraini, E. D., Masjkur, M., & Syafitri, U. D. (2024). *Klasterisasi sekolah pada penerimaan mahasiswa baru IPB University jalur rapor menggunakan algoritma K-Prototypes* [Skripsi sarjana, IPB University]. Repository IPB.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Dewantari, N. K. M., Syafitri, U. D., & Alamudi, A. (2021). *Penggerombolan sekolah pada penerimaan mahasiswa baru jalur SNMPTN di IPB menggunakan metode Two-Step Cluster* [Skripsi sarjana, IPB University]. Repository IPB.

Khairunnisa, S., Syafitri, U. D., & Mualifah, L. N. A. (2025). *Identifikasi peubah penting dalam klasifikasi indeks prestasi mahasiswa SNBP IPB University angkatan 2024 menggunakan XGBoost* [Skripsi sarjana, IPB University]. Repository IPB.

Liu, X.-Y., Wu, J., & Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2), 539–550. https://doi.org/10.1109/TSMCB.2008.2007853

Mahesh, T. R., Vinoth Kumar, V., Muthukumaran, V., Shashikala, H. K., Swapna, B., & Guluwadi, S. (2022). Performance analysis of XGBoost ensemble methods for survivability with the classification of breast cancer. *Journal of Sensors*, 2022, Article 4649510. https://doi.org/10.1155/2022/4649510

Montgomery, D. C. (2013). *Design and Analysis of Experiments* (8th ed.). John Wiley & Sons, Inc.

Rais, Soleh, A. M., & Susetyo, B. (2024). *Perbandingan kinerja rotation double random forest dengan algoritma ensemble tree lain pada klasifikasi biner data tidak seimbang* [Tesis magister / Skripsi, IPB University]. Repository IPB.

Siregar, A. Y., & Arifin, A. S. (2024). Enhancing XGBoost classification with SVM-SMOTE & EasyEnsemble for imbalanced telemedicine sentiment data. *Journal of Computer Science and Informatics Engineering (JERICO)*, 5(10), 3893–3902.

Tian, J., Jiang, Y., Zhang, J., Wang, Z., Rodríguez-Andina, J. J., & Luo, H. (2022). High-performance fault classification based on feature importance ranking-XGBoost approach with feature selection of redundant sensor data. *Current Chinese Science*, 2(5), 450–461.

Wiens, M., Verone-Boyle, A., Henscheid, N., Podichetty, J. T., & Burton, J. (2025). A tutorial and use case example of the eXtreme Gradient Boosting (XGBoost) artificial intelligence algorithm for drug development applications. *Clinical and Translational Science*, 18(3), e70172. https://doi.org/10.1111/cts.70172

Zhang, R., Song, H., Chen, Q., Wang, Y., Wang, S., & Li, Y. (2022). Comparison of ARIMA and LSTM for prediction of hemorrhagic fever at different time scales in China. *PLoS ONE*, 17(1), e0262009. https://doi.org/10.1371/journal.pone.0262009