**Building of Informatics, Technology and Science (BITS)**
Volume 7, No 1, June 2025 Page: 584−594
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7387

# Prediksi Kinerja Akademik Siswa Bimbingan Belajar Menggunakan Algoritma Extreme Gradient Boosting (XGBoost)

**Muhammad Bayu Ardi Alfarizi*, Wina Witanti, Agus Komarudin**

Teknik Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani, Cimahi, Indonesia
Email:[1,*]bayuardialparizi17@gmail.com, [2]witanti@gmail.com, [3]agus.komarudin@lecture.unjani.ac.id
Email Penulis Korespondensi: bayuardialparizi17@email.com
Submitted: **19/05/2025**; Accepted: **22/06/2025**; Published**: 25/06/2025**

**Abstrak**−Peningkatan mutu pendidikan menjadi fokus utama dalam menghadapi tantangan dunia pendidikan yang semakin kompleks. Salah satu pendekatan yang menjanjikan dalam mendukung pengambilan keputusan berbasis data adalah prediksi kinerja akademik siswa menggunakan algoritma machine learning. Penelitian ini bertujuan untuk mengembangkan model klasifikasi kinerja akademik siswa dengan memanfaatkan algoritma Extreme Gradient Boosting (XGBoost). Data yang digunakan berasal dari SMPN 1 Gunung Halu dan mencakup atribut akademik serta non-akademik siswa. Lima fitur utama yang digunakan adalah nilai awal, nilai tengah, nilai akhir, kelakuan siswa, dan absensi. Pra-pemrosesan data dilakukan melalui seleksi fitur, penghapusan data kosong, transformasi data kategorikal menggunakan label encoding, serta penyeimbangan kelas dengan metode SMOTE. Model XGBoost kemudian dilatih dengan pembagian data 80:20 dan dilakukan tuning hyperparameter menggunakan Grid Search. Hasil evaluasi menunjukkan bahwa model mencapai akurasi sebesar 84% dengan nilai F1-score yang merata di seluruh kelas. Performa model ini lebih unggul dibandingkan algoritma lain seperti Bagging dan Random Forest. Dengan akurasi dan stabilitas yang baik, model XGBoost berpotensi menjadi alat bantu dalam mengidentifikasi siswa yang membutuhkan intervensi akademik lebih lanjut. Penelitian ini memberikan kontribusi penting terhadap pengembangan sistem pendidikan berbasis kecerdasan buatan, dan dapat menjadi fondasi bagi penerapan machine learning dalam peningkatan kualitas pembelajaran di tingkat menengah.

**Kata Kunci:** Prediksi Kinerja Siswa; XGBoost; Hyperparameter; SMOTE; Data Mining; Machine Learning.

**Abstract**−Improving the quality of education has become a primary focus in addressing the increasingly complex challenges of the educational landscape. One promising approach to support data-driven decision-making is the prediction of students' academic performance using machine learning algorithms. This study aims to develop a classification model for predicting students' academic performance by leveraging the Extreme Gradient Boosting (XGBoost) algorithm. The dataset used was obtained from SMPN 1 Gunung Halu and includes both academic and non-academic attributes of students. Five key features were selected: initial grades, midterm grades, final grades, student behavior, and attendance. Data preprocessing involved feature selection, handling missing values, transforming categorical variables using label encoding, and balancing the classes using the SMOTE method. The XGBoost model was then trained using an 80:20 data split and hyperparameter tuning was performed using Grid Search. Evaluation results showed that the model achieved an accuracy of 84% with balanced F1-scores across all classes. The model outperformed other algorithms such as Bagging and Random Forest. With its strong accuracy and stability, the XGBoost model has the potential to serve as a tool for identifying students who require academic intervention. This study makes a significant contribution to the development of AI-based education systems and provides a foundation for the application of machine learning in improving the quality of secondary-level learning.

**Keywords**: Academic Performance Prediction; Hyperparameter; SMOTE; Data Mining; Machine Learning.

## 1. PENDAHULUAN

Pendidikan merupakan pilar utama dalam pembangunan suatu bangsa, karena berperan penting dalam mencetak sumber daya manusia (SDM) yang berkualitas, produktif, serta adaptif terhadap perkembangan zaman. Dalam era globalisasi dan kemajuan teknologi yang pesat, tantangan yang dihadapi dunia pendidikan semakin kompleks. Pendidikan tidak lagi hanya berfokus pada penguasaan materi akademik semata, tetapi juga pada kemampuan untuk bertahan dan berinovasi di tengah perubahan sosial dan ekonomi. Oleh karena itu, peningkatan mutu pendidikan menjadi prioritas utama bagi berbagai pihak, baik pemerintah, sekolah, hingga masyarakat.

Salah satu permasalahan yang masih kerap dijumpai dalam sistem pendidikan di Indonesia adalah tingginya angka putus sekolah dan rendahnya retensi siswa. Hal ini berdampak langsung pada kualitas lulusan serta ketimpangan capaian pendidikan antarwilayah. Dalam konteks ini, lembaga bimbingan belajar memiliki peran strategis sebagai pelengkap pendidikan formal. Lembaga ini hadir untuk membantu siswa dalam memahami materi pelajaran, meningkatkan motivasi belajar, serta memberikan dukungan tambahan bagi mereka yang mengalami kesulitan belajar [1]. Dengan pendekatan belajar yang lebih fleksibel dan personal, bimbingan belajar dapat menjadi solusi efektif dalam membantu siswa mencapai prestasi akademik yang lebih optimal.

Seiring dengan perkembangan teknologi informasi, penerapan teknik analisis data dalam bidang pendidikan menjadi semakin umum. Salah satu pendekatan yang menjanjikan adalah penggunaan data mining, yaitu proses eksplorasi data besar untuk menemukan pola atau informasi tersembunyi yang berguna dalam pengambilan keputusan. Dalam dunia pendidikan, data mining dapat dimanfaatkan untuk memprediksi kinerja akademik siswa, mengidentifikasi siswa berisiko rendah, serta membantu merancang strategi pembelajaran yang lebih tepat sasaran [2]. Proses prediksi kinerja siswa ini sangat penting, terutama untuk lembaga pendidikan yang ingin meningkatkan kualitas peserta didik secara menyeluruh.

**Building of Informatics, Technology and Science (BITS)**
Volume 7, No 1, June 2025 Page: 584−594
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7387

Salah satu algoritma data mining yang telah terbukti efektif dalam melakukan prediksi adalah Extreme Gradient Boosting (XGBoost). XGBoost merupakan pengembangan dari metode Gradient Boosting yang dirancang untuk meningkatkan kecepatan pemrosesan data, mengurangi risiko overfitting, serta mengatasi permasalahan pada dataset yang besar, kompleks, dan tidak seimbang [3]. Keunggulan lainnya adalah kemampuan XGBoost untuk melakukan parallel processing, penanganan data hilang secara otomatis, serta fitur regularisasi untuk meningkatkan generalisasi model.

Berbagai penelitian terdahulu telah menunjukkan keunggulan XGBoost dalam melakukan klasifikasi dan prediksi pada berbagai bidang. Dalam studi prediksi adopsi hewan peliharaan, algoritma ini mampu mencapai akurasi hingga 95% [4]. Sementara itu, pada penelitian yang memprediksi kualitas udara, XGBoost berhasil memperoleh akurasi sebesar 98,14%, jauh lebih tinggi dibandingkan metode Naïve Bayes yang hanya mencapai 92% [5][6]. Dalam konteks pendidikan, algoritma ini telah digunakan untuk memprediksi keberhasilan studi mahasiswa dengan hasil akurasi sebesar 76,8%, yang tetap konsisten meskipun diuji dengan beberapa strategi validasi [7].

Selain itu, penelitian yang membandingkan kinerja beberapa algoritma klasifikasi seperti SVM, Random Forest, dan XGBoost, menunjukkan bahwa XGBoost mampu memberikan nilai akurasi tertinggi sebesar 82%, dengan nilai recall mencapai 70%, dan precision sebesar 92% [8]. Penelitian lain juga mengonfirmasi bahwa XGBoost lebih unggul dibandingkan AdaBoost, terutama dalam menghadapi dataset yang kompleks. XGBoost mampu mencapai akurasi sebesar 92%, sedangkan AdaBoost hanya memperoleh 40% dengan precision sebesar 39%, karena keterbatasan pada struktur pohon keputusannya [9].

Berdasarkan temuan-temuan tersebut, dapat disimpulkan bahwa XGBoost merupakan salah satu algoritma yang sangat potensial untuk diterapkan dalam sistem prediksi kinerja akademik siswa. Kemampuan algoritma ini dalam mengolah data besar dan kompleks menjadikannya relevan untuk digunakan dalam lingkungan pendidikan yang dinamis. Dengan mengadopsi pendekatan ini, sekolah maupun lembaga pendidikan lainnya dapat memperoleh informasi yang lebih akurat mengenai kondisi siswa dan memberikan intervensi yang sesuai secara tepat waktu.

Penelitian ini bertujuan untuk mengimplementasikan algoritma XGBoost dalam memprediksi dan menganalisis data siswa dari SMPN 1 Gunung Halu. Tujuan utama dari sistem yang dikembangkan adalah untuk membantu pihak sekolah dalam mengidentifikasi siswa yang membutuhkan perhatian lebih atau dukungan khusus berdasarkan data historis akademik mereka. Dengan sistem ini, diharapkan proses pengambilan keputusan di tingkat satuan pendidikan dapat dilakukan secara lebih cepat, tepat, dan berbasis data. Lebih jauh, sistem yang dibangun juga diharapkan dapat menjadi acuan awal dalam pengembangan sistem berbasis kecerdasan buatan untuk sektor pendidikan di tingkat menengah. Penerapan model seperti XGBoost tidak hanya mendukung perbaikan proses pembelajaran secara teknis, tetapi juga memperkuat budaya pengambilan keputusan berbasis bukti (evidence-based education) yang sedang didorong oleh banyak institusi pendidikan di Indonesia. Hasil akhir dari penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pemetaan masalah akademik secara preventif serta sebagai referensi untuk penelitian lanjutan dalam pengembangan sistem prediktif pendidikan. Lebih dari sekadar alat bantu prediksi, penerapan algoritma XGBoost dalam dunia pendidikan juga mencerminkan kemajuan pemanfaatan teknologi kecerdasan buatan dalam meningkatkan kualitas pembelajaran. Di era transformasi digital saat ini, integrasi machine learning dalam proses akademik bukan lagi sekadar tren, melainkan kebutuhan. Dengan algoritma yang canggih dan kemampuan analitis yang kuat, XGBoost dapat membantu lembaga pendidikan mengatasi tantangan akademik secara proaktif dan efisien. Dengan demikian, penelitian ini diharapkan tidak hanya memberikan solusi praktis bagi sekolah, tetapi juga menjadi landasan bagi pengembangan sistem pembelajaran adaptif di masa depan yang lebih responsif terhadap kebutuhan individu siswa.

# 2. METODOLOGI PENELITIAN

## 2.1 Prediksi Kinerja Siswa

Prediksi kinerja akademik siswa merupakan aspek penting dalam peningkatan mutu pendidikan, karena memungkinkan pendidik untuk mengidentifikasi siswa yang memerlukan bantuan sejak dini. Dengan memahami kemampuan siswa dalam menyerap materi, intervensi yang lebih tepat dapat diberikan. Selain berdampak pada proses belajar, pencapaian akademik juga berkontribusi terhadap peluang karier siswa di masa depan, menjadikannya faktor penting tidak hanya dalam pendidikan, tetapi juga dalam kesiapan menghadapi dunia kerja [1].

## 2.2 Data Mining

Data mining merupakan proses inti dalam *Knowledge Discovery in Databases (KDD)* yang bertujuan mengekstraksi pola tersembunyi dari kumpulan data besar dan kompleks [10]. Melalui penerapan algoritma tertentu, data mining memungkinkan terbentuknya pengetahuan baru dari data yang telah ada, yang kemudian dapat dimanfaatkan untuk mendukung pengambilan keputusan. Teknik yang digunakan mencakup klasifikasi, klasterisasi, regresi, dan asosiasi, yang secara otomatis membantu mengidentifikasi pola-pola penting dalam basis data [11] [12].

## 2.3 Machine Learning

Machine learning adalah cabang dari kecerdasan buatan yang memungkinkan sistem belajar dari data dan secara otomatis meningkatkan kinerjanya. Dalam konteks pendidikan, teknologi ini digunakan untuk mempersonalisasi

**Building of Informatics, Technology and Science (BITS)**
Volume 7, No 1, June 2025 Page: 584−594
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7387

pembelajaran, memantau kemajuan siswa, serta mendukung guru dalam mengevaluasi performa akademik melalui sistem pemantauan berbasis data [13].

### 2.4 XGBoost

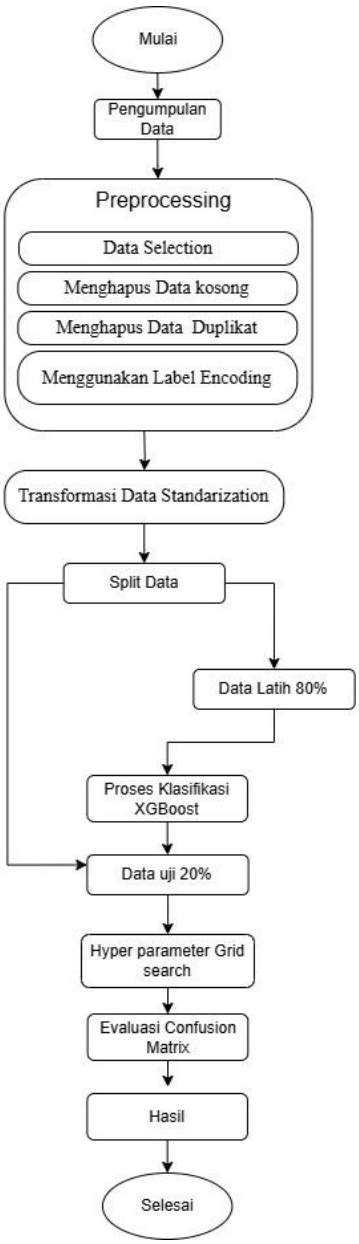
XGBoost merupakan algoritma machine learning berbasis boosting yang membangun pohon keputusan secara berurutan untuk memperbaiki kesalahan prediksi sebelumnya [14]. Algoritma ini unggul dalam efisiensi komputasi, mencegah overfitting melalui regularisasi, dan efektif untuk tugas klasifikasi, regresi, maupun ranking [15]. Prediksinya dihitung berdasarkan gabungan skor dari K pohon dengan fungsi objektif yang mengintegrasikan loss function dan regularisasi, sehingga menghasilkan model yang akurat dan tidak terlalu kompleks [16].

### 2.5 Hyperparameter

Hyperparameter adalah parameter yang ditetapkan sebelum pelatihan model dan digunakan untuk mengontrol proses pembelajaran dalam algoritma machine learning. Pada XGBoost, pengaturan hyperparameter sangat memengaruhi performa model, termasuk kecepatan pelatihan, tingkat akurasi, dan kemampuan generalisasi terhadap data baru [17].

### 2.6 Tahapan Penelitian

Penelitian ini bertujuan untuk membangun model klasifikasi guna memprediksi kinerja akademik siswa menggunakan algoritma XGBoost. Tahapan penelitian dirancang untuk memastikan proses pelatihan dan evaluasi model dilakukan secara sistematis dan akurat. Adapun tahapan penelitian ini digambarkan pada Gambar 1.




**Gambar 1.** Alur Metode Penelitian

**Building of Informatics, Technology and Science (BITS)**
Volume 7, No 1, June 2025 Page: 584−594
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7387

### 2.6.1 Pengumpulan Data

Data diperoleh dari SMPN 1 Gunung Halu sebanyak 1000 entri siswa yang terdiri atas atribut akademik (nilai awal, tengah, akhir), perilaku, absensi, dan partisipasi ekstrakurikuler. Namun, hanya lima fitur utama yang digunakan untuk pelatihan model: nilai awal, nilai tengah, nilai akhir, kelakuan siswa, dan absensi.

**Tabel 1.** Daftar Variabel Dataset

| No | Variabel | Deskripsi |
|---|---|---|
| 1 | NISN | Nomor Induk Siswa Nasional, digunakan sebagai identitas unik setiap siswa. |
| 2 | Nama Siswa | Nama lengkap siswa yang terdaftar dalam sistem. |
| 3 | Jenis Kelamin | Jenis kelamin siswa, dapat berupa "Laki-laki" atau "Perempuan". |
| 4 | Nilai Awal | Nilai akademik siswa di awal periode pembelajaran. |
| 5 | Nilai Tengah | Nilai akademik siswa di pertengahan periode pembelajaran. |
| 6 | Nilai Akhir | Nilai akhir siswa setelah seluruh proses pembelajaran selesai. |
| 7 | Kelakuan Siswa | Penilaian terhadap perilaku siswa selama proses belajar berlangsung. |
| 8 | Absensi | Jumlah kehadiran siswa selama periode pembelajaran. |
| 9 | PMR | Partisipasi siswa dalam kegiatan Palang Merah Remaja (PMR). |
| 10 | Kepramukaan | Partisipasi siswa dalam kegiatan Pramuka. |
| 11 | Seni Tari | Partisipasi siswa dalam kegiatan Seni Tari. |
| 12 | Volly Ball | Partisipasi siswa dalam kegiatan Bola Voli. |
| 13 | Futsal | Partisipasi siswa dalam kegiatan Futsal. |
| 14 | PASKIBRA | Partisipasi siswa dalam kegiatan Pasukan Pengibar Bendera (PASKIBRA). |
| 15 | Sepak Bola | Partisipasi siswa dalam kegiatan Sepak Bola. |

### 2.6.2 Pra-pemrosesan Data (Pre-processing)

Tahap pra-pemrosesan data merupakan bagian penting dalam analisis data mining, yang bertujuan untuk mempersiapkan data agar siap digunakan dalam proses analisis. Proses ini mencakup beberapa langkah utama seperti seleksi data untuk memilih atribut yang relevan, penanganan nilai kosong yang umum ditemukan dalam dataset empiris, serta transformasi data kategorikal menjadi numerik melalui encoding. Selain itu, outlier yang berpotensi mengganggu pelatihan model juga diidentifikasi dan diatasi. Dengan tahapan ini, kualitas data dapat ditingkatkan agar model dapat dilatih secara lebih akurat dan efisien[18] [19] [20].

### 2.6.3 Transformasi Data

Setelah data dibersihkan, dilakukan standardisasi agar seluruh fitur numerik memiliki skala yang seragam. Proses ini bertujuan untuk mencegah bias dalam pelatihan model akibat perbedaan rentang nilai antar fitur. Distribusi kelas target dalam data awal menunjukkan ketidakseimbangan. Untuk mengatasi hal ini, digunakan metode Synthetic Minority Oversampling Technique (SMOTE) guna menambah data sintetis pada kelas minoritas, sehingga distribusi antar kelas menjadi seimbang [21].

### 2.6.4 Pembagian Data (Split Dataset)

Dataset yang telah seimbang dibagi menjadi dua bagian, yaitu 80% sebagai data latih dan 20% sebagai data uji. Teknik stratified sampling diterapkan agar proporsi setiap kelas tetap seimbang pada kedua subset, memastikan hasil evaluasi model lebih objektif [22],

### 2.6.5 Evaluasi Model

Evaluasi model dalam penelitian ini dilakukan dengan menggunakan teknik Confusion Matrix, yang digunakan untuk menilai sejauh mana ketepatan klasifikasi model dalam konteks data mining. Teknik ini membantu mengidentifikasi keakuratan prediksi model terhadap data uji dengan cara membandingkan hasil klasifikasi yang dihasilkan model dengan label sebenarnya. Dalam proses evaluasi ini, terdapat beberapa indikator penting yang digunakan, yaitu True Positive (TP) yang menunjukkan jumlah data positif yang diklasifikasikan dengan benar, True Negative (TN) yang mencerminkan jumlah data negatif yang juga diklasifikasikan secara tepat, False Positive (FP) yang menggambarkan data negatif yang salah diklasifikasikan sebagai positif, dan False Negative (FN) yang merupakan data positif yang justru salah diklasifikasikan sebagai negatif. Keempat komponen ini menjadi dasar dalam menghitung berbagai metrik evaluasi, seperti akurasi, presisi, recall, dan F1-score, untuk menilai performa model secara menyeluruh [23].

Akurasi mengukur proporsi prediksi yang benar dari seluruh prediksi yang dilakukan model yang dapat dilihat pada Persamaan (1). Akurasi cocok digunakan saat distribusi kelas seimbang.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Presisi menunjukkan seberapa banyak dari prediksi positif yang benar-benar positif yang dapat dilihat pada Persamaan (2). Presisi penting saat biaya kesalahan prediksi positif cukup tinggi, misalnya dalam diagnosis penyakit atau deteksi penipuan.

**Building of Informatics, Technology and Science (BITS)**
Volume 7, No 1, June 2025 Page: 584−594
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7387

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall mengukur seberapa banyak dari total kasus positif yang berhasil dikenali oleh model yang dapat dilihat pada Persamaan (3). Recall penting ketika kesalahan dalam mengabaikan kasus positif (FN) perlu diminimalkan, seperti pada deteksi kanker atau sistem alarm.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1-score adalah rata-rata harmonik antara presisi dan recall. Metode ini memberikan keseimbangan ketika penting untuk mempertimbangkan keduanya yang dapat dilihat pada Persamaan (4). F1-score berguna pada kasus kelas tidak seimbang, karena memberikan gambaran lebih adil dibanding hanya menggunakan akurasi.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

# 3. HASIL DAN PEMBAHASAN

## 3.1 Pengumpulan Dataset

Data yang digunakan dalam penelitian ini diperoleh dari SMPN 1 Gunung Halu dengan pendekatan dokumentasi, yaitu melalui pengumpulan data yang telah tersimpan dalam sistem sekolah. Data terdiri atas informasi akademik dan non-akademik siswa, yang mencakup atribut seperti NISN/NIS, nama siswa, jenis kelamin, nilai awal, nilai tengah, nilai akhir, kelakuan siswa, jumlah absensi, serta partisipasi dalam berbagai kegiatan ekstrakurikuler seperti PMR, kepramukaan, seni tari, voli, futsal, paskibra, dan sepak bola.

Gambaran awal data dapat dilihat pada Gambar 2. Data disajikan dalam format tabel dengan setiap baris mewakili satu entri siswa. Atribut-atribut yang digunakan tidak seluruhnya memiliki nilai lengkap karena adanya variasi partisipasi siswa terhadap kegiatan ekstrakurikuler. Oleh karena itu, dilakukan tahapan pembersihan data dan seleksi fitur sebelum dimasukkan ke dalam proses modeling machine learning.

| | URUT | NISN/NIS | NAMA SISWA | L/P | NILAI AWAL | NILAI TENGAH | NILAI AKHIR | KELAKUAN SISWA | ABSENSI | PMR | KEPRAMUKAAN | SENI TARI | VOLLY BALL | FUTSAL | PASKIBRA | SEP/ BO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 8789734382/33405058 | AGUS RAMDANI | L | 46 | 78 | 80 | Cukup | 7 | B | SB | NaN | B | B | C | N |
| 1 | 2 | 4695035453/26676690 | Ahmad Akmal Rifa'i | L | 57 | 77 | 78 | Cukup | 14 | B | C | B | C | B | NaN | |
| 2 | 3 | 1088762719/47390904 | Ahmad Febrian Maulana | L | 42 | 62 | 70 | Kurang | 1 | SB | SB | SB | B | NaN | NaN | |
| 3 | 4 | 1333940684/17072129 | Alika Lutfi Lutfiah | P | 45 | 65 | 70 | Cukup | 1 | B | SB | NaN | SB | NaN | C | |
| 4 | 5 | 4919154512/18836764 | BILQIS AULIA DESTIYANI | P | 78 | 80 | 81 | Cukup | 1 | NaN | SB | NaN | C | NaN | B | |
| 5 | 6 | 6583708059/15440400 | DELIAN NUR ANNISA | P | 74 | 80 | 81 | Cukup | 4 | C | C | SB | C | NaN | NaN | |
| 6 | 7 | 9468221396/78604956 | DEVIA SRI ASTUTI | P | 81 | 84 | 85 | Kurang | 18 | SB | B | B | NaN | NaN | NaN | |
| 7 | 8 | 5683957904/56869686 | Diran Bagas Permana | L | 62 | 82 | 85 | Cukup | 6 | B | NaN | B | C | C | B | |
| 8 | 9 | 9511293670/69073591 | M. Ihsan Husaen | L | 45 | 65 | 78 | Cukup | 3 | SB | C | SB | SB | SB | SB | N |

**Gambar 2.** Dataset

Gambar 2 menunjukkan sebagian sampel data siswa dengan informasi lengkap yang mencerminkan kombinasi atribut numerik dan kategorikal, yang kemudian diolah untuk keperluan prediksi kinerja akademik siswa. Dataset yang digunakan dalam penelitian ini diperoleh dari SMPN 1 Gunung Halu, yang terdiri atas data akademik dan non-akademik siswa. Terdapat 15 variabel awal yang mencakup informasi seperti NISN, nama siswa, jenis kelamin, nilai awal, nilai tengah, nilai akhir, kelakuan siswa, absensi, serta partisipasi dalam kegiatan ekstrakurikuler seperti PMR, pramuka, seni tari, bola voli, futsal, paskibra, dan sepak bola. Dari keseluruhan fitur tersebut, dilakukan seleksi fitur untuk memilih variabel yang paling relevan dengan kinerja akademik siswa. Lima variabel yang dipilih untuk dijadikan input dalam model klasifikasi adalah nilai awal, nilai tengah, nilai akhir, kelakuan siswa, dan absensi. Pemilihan fitur ini didasarkan pada pertimbangan kontribusi langsung terhadap pencapaian akademik siswa serta ketersediaan data yang lengkap. Proses seleksi fitur ini ditampilkan pada Gambar 3, yang memperlihatkan subset data yang akan digunakan dalam proses pemodelan.

**Building of Informatics, Technology and Science (BITS)**
Volume 7, No 1, June 2025 Page: 584−594
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7387

| | NILAI AWAL | NILAI TENGAH | NILAI AKHIR | KELAKUAN SISWA | ABSENSI | KINERJA SISWA |
|---|---|---|---|---|---|---|
| 0 | 46 | 78 | 80 | Cukup | 7 | Cukup |
| 1 | 57 | 77 | 78 | Cukup | 14 | Cukup |
| 2 | 42 | 62 | 70 | Kurang | 1 | Cukup |
| 3 | 45 | 65 | 70 | Cukup | 1 | Cukup |
| 4 | 78 | 80 | 81 | Cukup | 1 | Baik |
| 5 | 74 | 80 | 81 | Cukup | 4 | Baik |
| 6 | 81 | 84 | 85 | Kurang | 18 | Cukup |
| 7 | 62 | 82 | 85 | Cukup | 6 | Cukup |
| 8 | 45 | 65 | 78 | Cukup | 3 | Cukup |
| 9 | 86 | 86 | 88 | Sangat Baik | 0 | Sangat Baik |

**Gambar 3.** Lima Variabel yang Digunakan

### 3.2 Pra-pemrosesan Data (Pre-processing)

Setelah dilakukan seleksi fitur, diperoleh lima atribut yang paling relevan terhadap label target, yaitu nilai awal, nilai tengah, nilai akhir, kelakuan siswa, dan jumlah absensi. Seperti terlihat pada Gambar 3, data ditampilkan dalam format tabular dengan nama-nama kolom yang digunakan dalam model prediksi.

Langkah pertama dalam tahap pra-pemrosesan adalah memastikan tidak ada data kosong (missing value). Pemeriksaan dilakukan dengan metode isna().sum() pada setiap kolom, seperti ditunjukkan pada Gambar 4. Hasilnya menunjukkan bahwa seluruh fitur yang digunakan tidak memiliki nilai kosong, sehingga proses imputasi tidak diperlukan. Ini menunjukkan bahwa kualitas data pada tahap awal sudah baik dan siap untuk tahap pemrosesan selanjutnya.

Cek Apakah Ada Data Kosong Atau Tidak

```
print("\nCek Missing Value:")
print(df_model.isna().sum())
```

```
Cek Missing Value:
NILAI AWAL        0
NILAI TENGAH      0
NILAI AKHIR       0
KELAKUAN SISWA    0
ABSENSI           0
KINERJA SISWA     0
dtype: int64
```

**Gambar 4.** Pra-pemrosesan (Missing Value)

### 3.3 Transformasi Data

Langkah berikutnya adalah melakukan transformasi data. Pada tahap ini, dilakukan identifikasi terhadap fitur-fitur yang bersifat kategorikal. Berdasarkan Gambar 5, diketahui bahwa fitur "KELAKUAN SISWA" dan "KINERJA SISWA" masih berupa tipe data objek.

| | NILAI AWAL | NILAI TENGAH | NILAI AKHIR | KELAKUAN SISWA | ABSENSI | KINERJA SISWA |
|---|---|---|---|---|---|---|
| 0 | 46 | 78 | 80 | Cukup | 7 | Cukup |
| 1 | 57 | 77 | 78 | Cukup | 14 | Cukup |
| 2 | 42 | 62 | 70 | Kurang | 1 | Cukup |
| 3 | 45 | 65 | 70 | Cukup | 1 | Cukup |
| 4 | 78 | 80 | 81 | Cukup | 1 | Baik |
| 5 | 74 | 80 | 81 | Cukup | 4 | Baik |
| 6 | 81 | 84 | 85 | Kurang | 18 | Cukup |
| 7 | 62 | 82 | 85 | Cukup | 6 | Cukup |
| 8 | 45 | 65 | 78 | Cukup | 3 | Cukup |
| 9 | 86 | 86 | 88 | Sangat Baik | 0 | Sangat Baik |

**Gambar 5.** Fitur-Fitur yang Bersifat Kategorikal

Oleh karena itu, dilakukan encoding terhadap kedua fitur ini menggunakan metode Label Encoding sebagaimana ditunjukkan pada Gambar 6. Setelah proses encoding, seluruh fitur telah berada dalam format numerik dan siap untuk digunakan dalam pelatihan model. Label numerik dari "KINERJA SISWA" terdiri dari: Kelas 0: siswa

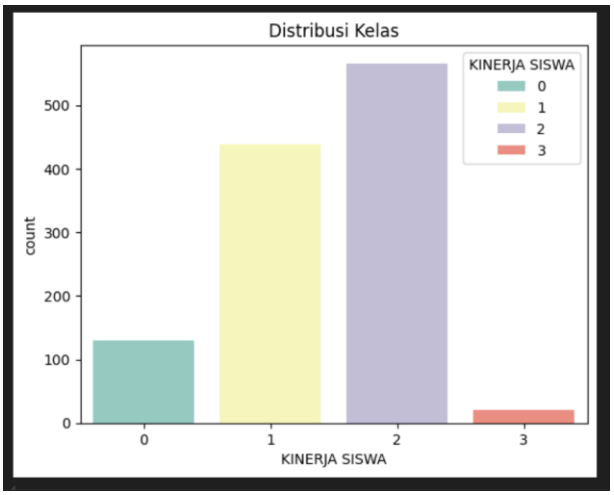
**Building of Informatics, Technology and Science (BITS)**
Volume 7, No 1, June 2025 Page: 584−594
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7387

dengan kinerja "Baik", Kelas 1: siswa dengan kinerja "Cukup", Kelas 2: siswa dengan kinerja "Kurang", Kelas 3: siswa dengan kinerja "Sangat Baik".

| | NILAI AWAL | NILAI TENGAH | NILAI AKHIR | KELAKUAN SISWA | ABSENSI | KINERJA SISWA |
|---|---|---|---|---|---|---|
| 0 | 46 | 78 | 80 | 0 | 7 | 1 |
| 1 | 57 | 77 | 78 | 0 | 14 | 1 |
| 2 | 42 | 62 | 70 | 1 | 1 | 1 |
| 3 | 45 | 65 | 70 | 0 | 1 | 1 |
| 4 | 78 | 80 | 81 | 0 | 1 | 0 |
| 5 | 74 | 80 | 81 | 0 | 4 | 0 |
| 6 | 81 | 84 | 85 | 1 | 18 | 1 |
| 7 | 62 | 82 | 85 | 0 | 6 | 1 |
| 8 | 45 | 65 | 78 | 0 | 3 | 1 |
| 9 | 86 | 86 | 88 | 2 | 0 | 3 |

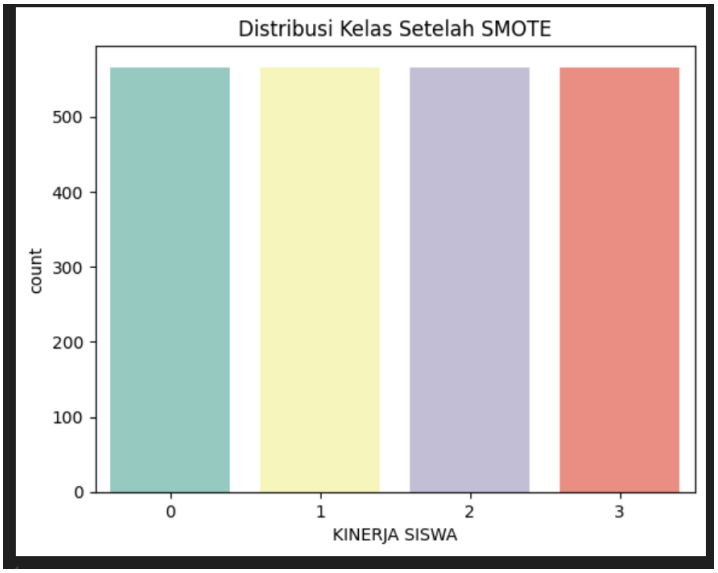
**Gambar 6.** Setelah Proses Encoding (Numerik)

Selanjutnya, dilakukan analisis distribusi kelas target yaitu "KINERJA SISWA" yang divisualisasikan pada Gambar 7.




**Gambar 7.** Distribusi Kelas Tidak Seimbang

Grafik tersebut menunjukkan adanya ketidakseimbangan jumlah data antar kelas, di mana kelas mayoritas adalah kelas 1 dan 2, sedangkan kelas 0 dan 3 berjumlah jauh lebih sedikit. Untuk mengatasi ketidakseimbangan ini, digunakan metode Synthetic Minority Oversampling Technique (SMOTE) guna menyeimbangkan distribusi kelas. Hasil distribusi setelah SMOTE dapat dilihat pada Gambar 8, di mana jumlah sampel pada keempat kelas menjadi seimbang.




**Gambar 8.** Distribusi Kelas Setelah Melakukan SMOTE

**Building of Informatics, Technology and Science (BITS)**
Volume 7, No 1, June 2025 Page: 584−594
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7387

### 3.4 Split Data

Setelah proses penyeimbangan kelas dilakukan menggunakan teknik SMOTE, data yang semula tidak seimbang antara kelas mayoritas dan minoritas telah menjadi seimbang, sebagaimana ditunjukkan pada Gambar 8. Hal ini memastikan bahwa model tidak bias terhadap kelas tertentu dan mampu belajar secara proporsional dari semua kelas target. Langkah selanjutnya adalah memisahkan dataset ke dalam dua bagian, yaitu data latih dan data uji. Proses ini dilakukan untuk menguji performa model secara objektif terhadap data yang belum pernah dilihat sebelumnya. Pemisahan dilakukan dengan rasio 80:20, di mana 80% dari total data digunakan sebagai data latih (training data) dan sisanya 20% digunakan sebagai data uji (testing data).

### 3.5 Klasifikasi XGBoost

Proses klasifikasi dalam penelitian ini dilakukan menggunakan algoritma Extreme Gradient Boosting (XGBoost). XGBoost merupakan salah satu metode ensemble learning yang berbasis pada teknik boosting, yaitu proses membangun model secara berurutan untuk memperbaiki kesalahan dari model sebelumnya. Setiap pohon keputusan (decision tree) yang dibentuk akan fokus pada observasi yang sebelumnya diklasifikasikan secara salah, sehingga akurasi model meningkat secara bertahap.

Algoritma XGBoost dikenal karena kemampuannya dalam menangani dataset berskala besar, mendukung fitur kategorikal maupun numerik, serta mampu mencegah overfitting melalui teknik regularisasi. XGBoost juga memiliki keunggulan dalam efisiensi komputasi dan fleksibilitas dalam pengaturan hyperparameter.

Dalam implementasinya, digunakan library XGBClassifier dari pustaka xgboost untuk membangun model klasifikasi. Model ini dilatih menggunakan data hasil pra-pemrosesan dan penyeimbangan sebelumnya, yaitu X_train dan y_train. Selain itu, untuk memperoleh hasil yang optimal, dilakukan juga tuning hyperparameter menggunakan metode Grid Search seperti dijelaskan pada subbab berikutnya.

Proses pelatihan model XGBoost dilakukan setelah pembagian data latih dan uji. Model yang telah dibentuk kemudian digunakan untuk melakukan prediksi terhadap data uji (X_test) dan hasil prediksi dibandingkan dengan label aktual (y_test) untuk menghitung performa model.

### 3.6 Hyperparamater

Pada tahap ini, dilakukan penyesuaian nilai hyperparameter untuk memperoleh kinerja terbaik dari model klasifikasi XGBoost. Hyperparameter merupakan parameter eksternal yang tidak dipelajari secara langsung dari data, melainkan harus ditentukan sebelum proses pelatihan model dimulai. Pemilihan nilai hyperparameter yang tepat sangat berpengaruh terhadap akurasi, efisiensi pelatihan, serta kemampuan generalisasi model. Dalam penelitian ini, proses tuning hyperparameter dilakukan dengan metode Grid Search, yaitu teknik pencarian sistematis pada ruang kombinasi parameter untuk menemukan konfigurasi terbaik berdasarkan metrik performa tertentu, dalam hal ini akurasi. Beberapa hyperparameter utama yang disesuaikan antara lain adalah n_estimators yang menentukan jumlah pohon yang dibentuk, learning_rate sebagai laju pembelajaran, max_depth yang mengatur kedalaman maksimum setiap pohon keputusan, subsample untuk menentukan proporsi data yang digunakan pada tiap iterasi, serta colsample_bytree yang mengatur proporsi fitur yang digunakan dalam pembuatan setiap pohon. Proses tuning dilakukan menggunakan validasi silang sebanyak tiga lipatan (3-fold cross-validation) guna menghindari overfitting dan memastikan model teruji secara menyeluruh. Setelah melalui proses eksplorasi terhadap berbagai kombinasi parameter, diperoleh konfigurasi terbaik yaitu n_estimators = 200, learning_rate = 0.1, max_depth = 5, subsample = 0.8, dan colsample_bytree = 1.0. Konfigurasi ini digunakan dalam pelatihan akhir model karena menghasilkan akurasi tertinggi pada data latih dan performa yang stabil pada data uji.

### 3.7 Evaluasi Model

Evaluasi model dilakukan untuk mengukur sejauh mana kinerja model XGBoost dalam mengklasifikasikan kinerja akademik siswa ke dalam empat kategori. Proses evaluasi dilakukan dengan memisahkan data uji sebesar 20% dari total dataset setelah dilakukan proses SMOTE. Model yang telah dilatih menggunakan kombinasi hyperparameter terbaik dari Grid Search kemudian diuji terhadap data uji untuk mendapatkan metrik performa. Metrik evaluasi yang digunakan meliputi akurasi, precision, recall, dan F1-score yang dapat dilihat pada Gambar 9.

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.88 | 0.82 | 0.85 | 113 |
| 1 | 0.81 | 0.76 | 0.79 | 113 |
| 2 | 0.81 | 0.89 | 0.85 | 114 |
| 3 | 0.88 | 0.90 | 0.89 | 113 |
| accuracy | | | 0.84 | 453 |
| macro avg | 0.84 | 0.84 | 0.84 | 453 |
| weighted avg | 0.84 | 0.84 | 0.84 | 453 |

**Gambar 9.** Evaluasi Model

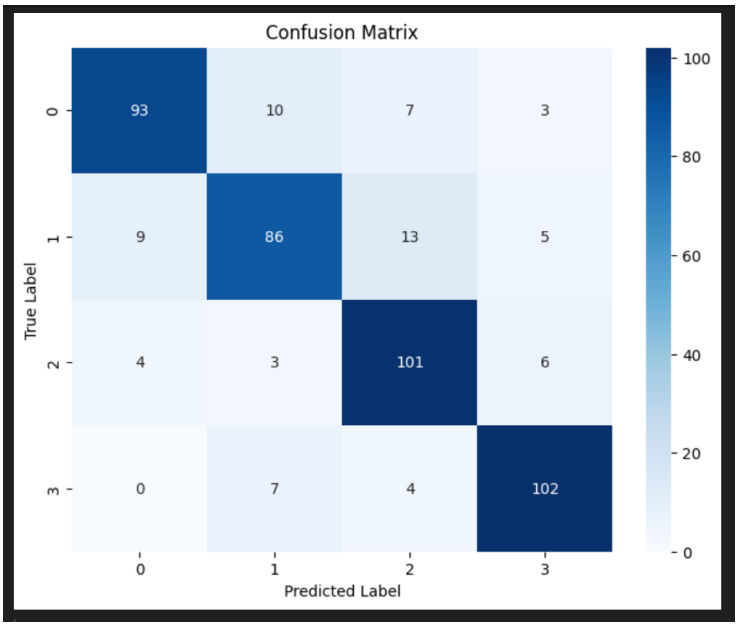
**Building of Informatics, Technology and Science (BITS)**
Volume 7, No 1, June 2025 Page: 584−594
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7387

Akurasi menunjukkan proporsi keseluruhan prediksi yang benar terhadap seluruh data uji, sedangkan precision, recall, dan F1-score digunakan untuk mengevaluasi performa pada masing-masing kelas. Berdasarkan hasil yang diperoleh, model mencapai nilai akurasi sebesar 84%. Nilai macro average untuk precision, recall, dan F1-score masing-masing sebesar 0.84, yang mengindikasikan bahwa model memiliki performa yang seimbang di semua kelas target.

Lebih lanjut, hasil classification report menunjukkan bahwa kelas 0 (Baik) memiliki precision sebesar 0.88 dan recall 0.82, sedangkan kelas 1 (Cukup) memiliki precision 0.81 dan recall 0.76. Untuk kelas 2 (Kurang), precision mencapai 0.81 dengan recall 0.89, dan kelas 3 (Sangat Baik) menunjukkan hasil terbaik dengan precision dan recall masing-masing 0.88 dan 0.90. F1-score tertinggi terdapat pada kelas 3 sebesar 0.89, sedangkan F1-score terendah terdapat pada kelas 1 sebesar 0.79. Hal ini menunjukkan bahwa meskipun terjadi perbedaan performa antar kelas, model tetap menunjukkan akurasi dan stabilitas yang cukup baik secara keseluruhan. Hasil ini mengindikasikan bahwa algoritma XGBoost dengan konfigurasi optimal mampu mengenali pola yang relevan dalam data akademik dan non-akademik siswa, serta memberikan prediksi yang cukup akurat terhadap kinerja siswa dalam kategori yang seimbang.

### 3.8 Confusion Matriks

Untuk mendapatkan gambaran yang lebih rinci mengenai performa model XGBoost dalam mengklasifikasikan data, digunakan confusion matrix sebagai alat visualisasi evaluasi. Confusion matrix memungkinkan untuk mengetahui jumlah prediksi benar dan salah pada masing-masing kelas, serta menunjukkan kesalahan klasifikasi antar kategori yang dapat dilihat pada Gambar 10.




**Gambar 10.** Confusion Matrix

Pada penelitian ini, model diminta untuk memetakan siswa ke dalam empat kelas kinerja akademik, yaitu kelas 0 (Kurang), kelas 1 (Cukup), kelas 2 (Baik), dan kelas 3 (Sangat Baik). Hasil confusion matrix menunjukkan bahwa sebagian besar prediksi berada di sepanjang diagonal utama, yang mengindikasikan bahwa model mampu memprediksi label dengan benar. Misalnya, dari 113 data siswa yang sebenarnya termasuk kelas 0, sebanyak 93 siswa berhasil diklasifikasikan dengan benar, sementara sisanya keliru diklasifikasikan ke kelas lain.

Begitu pula pada kelas 3 (Sangat Baik), sebanyak 102 dari 113 data diklasifikasikan dengan benar, dan hanya sebagian kecil yang salah diprediksi ke kelas 1 dan 2. Pola yang serupa terlihat juga pada kelas 2 (Baik) dan kelas 1 (Cukup), meskipun terlihat ada kecenderungan model untuk sedikit tertukar dalam membedakan antara dua kelas tersebut. Kesalahan ini wajar mengingat nilai-nilai fitur antar kelas 1 dan 2 bisa jadi cukup berdekatan. Meskipun terdapat kesalahan klasifikasi, performa keseluruhan tetap baik dan distribusi kesalahan tergolong kecil serta tidak menunjukkan bias terhadap salah satu kelas tertentu. Hal ini menjadi indikator penting bahwa model yang dikembangkan tidak hanya memiliki akurasi tinggi, tetapi juga proporsional dalam melakukan klasifikasi terhadap seluruh kategori kinerja siswa.

### 3.9 Hasil

Hasil yang diperoleh menunjukkan bahwa model XGBoost sangat efektif dalam mengklasifikasikan kinerja akademik siswa. Kombinasi antara seleksi fitur, transformasi data, penyeimbangan kelas, dan tuning hyperparameter memberikan kontribusi besar terhadap akurasi dan stabilitas model. Dalam konteks lembaga pendidikan, model ini dapat digunakan untuk mengidentifikasi siswa yang perlu mendapat perhatian lebih dini, sehingga intervensi pembelajaran dapat dilakukan secara tepat sasaran.

**Building of Informatics, Technology and Science (BITS)**
Volume 7, No 1, June 2025 Page: 584−594
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7387

Tabel 2. menyajikan perbandingan hasil model XGBoost dalam penelitian ini dengan algoritma lain yang digunakan dalam penelitian sebelumnya.

**Tabel 2.** Penelitian Sebelumnya

| Model | Recall | Akurasi | F1-Score |
|---|---|---|---|
| XGBoost (Penelitian Ini) | 0.84 | 0.84 | 0.84 |
| Bagging [24] | 0.812 | 0.8125 | 0.809 |
| Random Forest [25] | 0.746 | 0.746 | 0.721 |

Dari tabel di atas, terlihat bahwa algoritma XGBoost menghasilkan akurasi yang lebih tinggi dibandingkan metode sebelumnya. Hal ini menunjukkan keunggulan XGBoost dalam menangani data klasifikasi kinerja siswa, terutama ketika dikombinasikan dengan teknik balancing (SMOTE) dan hyperparameter tuning.

Keunggulan utama XGBoost terletak pada kemampuannya menggabungkan regularisasi L1 dan L2, menangani missing value, serta fleksibilitas dalam mengolah data kategorikal dan numerik. Namun demikian, terdapat beberapa keterbatasan yang perlu dicatat. Pertama, jumlah fitur yang digunakan relatif sedikit dan belum mencakup aspek non-akademik yang lebih luas seperti latar belakang keluarga atau kondisi psikologis siswa. Kedua, model hanya dilatih pada satu sumber data dari satu sekolah, sehingga generalisasi model terhadap populasi yang lebih luas masih perlu diuji lebih lanjut. Penelitian lanjutan disarankan untuk memperluas cakupan data dan mencoba pendekatan ensemble learning yang lebih kompleks atau mengombinasikan XGBoost dengan teknik deep learning untuk meningkatkan akurasi dan kemampuan generalisasi.

Secara keseluruhan, penelitian ini membuktikan bahwa algoritma XGBoost dengan pendekatan data mining mampu memberikan solusi efektif untuk prediksi kinerja akademik siswa. Dengan implementasi yang tepat, pendekatan ini dapat menjadi salah satu strategi penting dalam upaya peningkatan mutu pendidikan berbasis data.

Penelitian ini dilakukan dengan tujuan untuk membangun model prediksi kinerja akademik siswa dengan memanfaatkan algoritma Extreme Gradient Boosting (XGBoost). Permasalahan utama yang diangkat dalam penelitian ini adalah bagaimana mengklasifikasikan siswa ke dalam kategori kinerja akademik yang berbeda, agar lembaga pendidikan dapat memberikan intervensi pembelajaran yang tepat. Berdasarkan hasil evaluasi, model XGBoost yang telah dioptimasi melalui proses Grid Search dan penyeimbangan kelas dengan teknik SMOTE menunjukkan performa yang memuaskan, dengan akurasi mencapai 84% serta nilai F1-score yang merata di keempat kelas.

Hasil ini menunjukkan bahwa XGBoost tidak hanya mampu melakukan klasifikasi dengan akurasi tinggi, tetapi juga mempertahankan keseimbangan dalam mengenali berbagai kategori kinerja siswa, termasuk pada kelas minoritas yang sebelumnya kurang terwakili. Temuan ini memberikan kontribusi signifikan terhadap proses pengambilan keputusan berbasis data dalam ranah pendidikan, terutama pada bimbingan belajar yang membutuhkan strategi intervensi personal.

Namun demikian, penelitian ini memiliki beberapa keterbatasan. Pertama, jumlah fitur yang digunakan masih terbatas dan belum mencakup aspek psikososial atau lingkungan belajar siswa. Kedua, data yang digunakan hanya berasal dari satu institusi pendidikan, sehingga generalisasi hasil ke sekolah lain perlu dikaji lebih lanjut. Oleh karena itu, penelitian selanjutnya disarankan untuk memperluas cakupan fitur dan sumber data agar model yang dikembangkan dapat lebih akurat dan aplikatif secara luas dalam konteks pendidikan nasional.

# 4. KESIMPULAN

Penelitian ini berhasil membangun model klasifikasi untuk memprediksi kinerja akademik siswa menggunakan algoritma Extreme Gradient Boosting (XGBoost), dengan menerapkan tahapan seleksi fitur, transformasi data, penyeimbangan kelas menggunakan SMOTE, serta tuning hyperparameter melalui Grid Search, yang menghasilkan akurasi sebesar 84% dan F1-score yang merata pada seluruh kelas target. Evaluasi menggunakan confusion matrix menunjukkan bahwa model mampu melakukan klasifikasi secara proporsional dan akurat, termasuk pada kelas minoritas. Keunggulan XGBoost dalam kestabilan, efisiensi, dan kemampuannya menangani data kompleks menjadikannya lebih unggul dibanding algoritma lain seperti Bagging dan Random Forest, serta potensial sebagai sistem pendukung keputusan di bidang pendidikan. Namun demikian, keterbatasan seperti jumlah fitur yang terbatas dan cakupan data dari satu institusi menjadi catatan penting, sehingga penelitian lanjutan disarankan untuk memperluas sumber data dan mempertimbangkan fitur tambahan yang mencakup aspek psikososial maupun lingkungan belajar siswa. Secara keseluruhan, pendekatan ini memberikan kontribusi terhadap strategi preventif dalam peningkatan mutu pendidikan berbasis data.

# REFERENCES

[1] T. Gori, A. Sunyoto, and H. Al Fatta, "Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.

[2] O. Ojajuni *et al.*, "*Predicting Student Academic Performance Using Machine Learning*," in *Lecture Notes in Computer Science*, Springer International Publishing, 2021, pp. 481–491. doi: https://doi.org/10.1007/978-3-030-87013-3_36.

[3] P. Septiana Rizky, R. Haiban Hirzi, and U. Hidayaturrohman, "Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 15, no. 2, pp. 228–236, 2022, doi:

**Building of Informatics, Technology and Science (BITS)**
Volume 7, No 1, June 2025 Page: 584−594
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7387

10.36456/jstat.vol15.no2.a5548.
[4] G. Dwilestari, "Prediksi Adopsi Hewan Peliharaan Menggunakan Metode Xgboost," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 4, pp. 7470–7477, 2024, doi: 10.36040/jati.v8i4.10337.
[5] A. A. Nababan, M. Jannah, M. Aulina, and D. Andrian, "Prediksi Kualitas Udara Menggunakan Xgboost Dengan Synthetic Minority Oversampling Technique (Smote) Berdasarkan Indeks Standar Pencemaran Udara (Ispu)," *JTIK (Jurnal Tek. Inform. Kaputama)*, vol. 7, no. 1, pp. 214–219, 2023, doi: 10.59697/jtik.v7i1.66.
[6] B. P. Salsabila, P. Belva, C. Trana, N. Ramadhani, A. P. Sari, and N. Bayes, "Penerapan Algoritma Naive Bayes Terhadap Kualitas Udara Di Jakarta dan Rekomendasi Aktivitas Masyarakat," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 6, pp. 11732–11738, 2024, doi: 10.36040/jati.v8i6.11592.
[7] E. J. Sudarman and S. Budi, "Pengembangan Model Kecerdasan Mesin Extreme Gradient Boosting untuk Prediksi Keberhasilan Studi Mahasiswa," *J. Strateg.*, vol. 5, no. 2, pp. 297–314, 2023.
[8] M. R. Givari, M. R. Sulaeman, and Y. Umaidah, "Perbandingan Algoritma SVM, Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit," *J. Nuansa Inform.*, vol. 16, no. 1, pp. 141–149, 2022, doi: 10.25134/nuansa.v16i1.5406.
[9] C. E. Sukmawati, A. Fitri, N. Masruriyah, and A. R. Juwita, "Efektivitas Algoritma AdaBoost dan XGBoost pada Dataset Obesitas Populasi Dewasa," *Jambura J. Informatics*, vol. 6, no. 2, pp. 101–111, 2024, doi: 10.37905/jji.
[10] M. T. I. Rahmayani, "Analisis Clustering Tingkat Keparahan Penyakit Pasien Menggunakan Algoritma K-Means," *J. Inov. Tek. Inform.*, vol. 1, no. 2, pp. 40–44, 2018.
[11] E. Muningsih and S. Kiswati, "Penerapan Metode K-Means Untuk Clustering Produk Online Shop Dalam Penentuan Stok Barang," *J. Bianglala Inform.*, vol. 1, no. 3, pp. 10–17, 2015.
[12] C. Romero and S. Ventura, "*Educational Data Mining and Learning Analytics: An Updated Survey*," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov. Min. Knowl. Discov.*, vol. 10, no. 3, pp. 1–21, 2020, doi: 10.1002/widm.1355.
[13] G. N. Sihombing, "Optimasi Sistem Pemantauan Akademik Siswa dengan Collaborative Filtering dan Decision Tree," *J. Sains Student Res.*, vol. 2, no. 6, pp. 369–375, 2024, doi: https://doi.org/10.61722/jssr.v2i6.2982.
[14] K. Aditya, A. Wisnu, and A. M. A. Rahim, "Analisis Perbandingan Algoritma XGBoost Dan Algoritma Random Forest Untuk Klasifikasi Data Kesehatan Mental," *Log. J. Ilmu Komput. dan Pendidik.*, vol. 2, no. 5, pp. 808–818, 2024.
[15] Y. N. Sukmaningtyas, R. M. Akbar, and G. Rohma, "Penerapan Predictive *Analytics* untuk Analisis Faktor-faktor yang Mempengaruhi Performa Akademik Siswa," *Arcitech J. Comput. Sci. Artif. Intell.*, vol. 4, no. 2, pp. 127–145, 2024, doi: http://dx.doi.org/10.29240/arcitech.v4i2.12048.
[16] K. D. K. Wardhani and M. Akbar, "*Diabetes Risk Prediction Using Extreme Gradient Boosting (XGBoost)*," *J. Online Inform.*, vol. 7, no. 2, pp. 244–250, 2022, doi: 10.15575/join.v7i2.970.
[17] D. A. Anggoro and S. S. Mukti, "*Performance Comparison of Grid Search and Random Search Methods for Hyperparameter Tuning in Extreme Gradient Boosting Algorithm to Predict Chronic Kidney Failure*," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 6, pp. 198–207, 2021, doi: 10.22266/ijies2021.1231.19.
[18] A. Charu, *Data Mining The Textbook*. Springer, 2016. doi: 10.1007/978-3-319-14142-8 ISBN.
[19] N. Syahfitri, E. Budianita, A. Nazir, and I. Afrianty, "Pengelompokan Produk Berdasarkan Data Persediaan Barang Menggunakan Metode Elbow dan K-Medoid," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 3, pp. 1668–1675, 2023, doi: 10.30865/klik.v4i3.1525.
[20] K. Boros and Z. Kmetty, *"Identifying Missing Data Handling Methods withText Mining*," *Int. J. Data Sci. Anal.*, 2024, doi: 10.1007/s41060-024-00582-1.
[21] M. Buda, A. Maki, and M. A. Mazurowski, "*A systematic Study of the Class Imbalance Problem in Convolutional Neural Networks*," *Neural Networks*, vol. 106, pp. 249–259, 2018, doi: 10.1016/j.neunet.2018.07.011.
[22] A. I. Pradana *et al.*, "Perbandingan Data Untuk Memprediksi Ketepatan Studi Berdasarkan Atribut Keluarga Menggunakan Machine Learning," *JIKA (Jurnal Informatics) Univ. Muhammadiyah Tangerang*, vol. 8, no. 2, pp. 221–228, 2024, doi: 10.31000/jika.v8i2.10752.
[23] M. R. Santoso and P. Musa, "Rekomendasi Kesehatan Janin Dengan Penerapan Algoritma C5.0 Menggunakan Classifying Cardiotocography Dataset," *J. Simantec*, vol. 9, no. 2, pp. 65–76, 2021, doi: 10.21107/simantec.v9i2.10730.
[24] M. Ardianti, O. D. Nurhayati, and B. Warsito, "Model Prediksi Kinerja Siswa Berdasarkan Data Log LMS Menggunakan Ensemble Machine Learning," *J. Sains dan Teknol.*, vol. 12, no. 3, pp. 562–571, 2024, doi: 10.23887/jstundiksha.v12i3.59816.
[25] M. Yağcı, "*Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms*," *Yağcı Smart Learn. Environ.*, vol. 9, no. 1, pp. 1–19, 2022, doi: 10.1186/s40561-022-00192-z.