

Prediksi Risiko Performa Akademik Siswa Menggunakan Random Forest, XGBoost, CatBoost, SMOTENC dan Explainable AI

Giri Reksa Guritno*, Angger Shalahuddin Al Ayyubi, Gandung Triyono

Program Studi Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta Selatan, Indonesia
Email: 2411600998@student.budiluhur.ac.id, 2411601103@student.budiluhur.ac.id, Gandung.triyono@budiluhur.ac.id
Email Penulis Korespondensi: 2411600998@student.budiluhur.ac.id
Submitted: 26/07/2026; Accepted: 20/09/2026; Published: 30/09/2026

Abstrak- Prediksi risiko performa akademik diperlukan untuk membantu institusi pendidikan mengidentifikasi siswa yang membutuhkan intervensi lebih awal. Tujuan penelitian ini adalah mengevaluasi pengaruh penanganan data tidak seimbang dan interpretabilitas model terhadap prediksi risiko performa akademik siswa. Penelitian ini menggunakan dataset sekunder Student Performance Factors berjumlah 6.377 data dengan target biner Berisiko dan Tidak Berisiko. Penelitian membandingkan Random Forest, XGBoost, dan CatBoost pada dua skenario, yaitu tanpa SMOTENC dan dengan SMOTENC, serta menambahkan explainable AI melalui feature importance, SHAP, dan analisis arah pengaruh fitur. Evaluasi dilakukan menggunakan accuracy, precision macro, recall macro, F1 macro, F1 weighted, recall per kelas, confusion matrix, stratified 5-fold cross-validation, dan uji signifikansi. Hasil pengujian menunjukkan bahwa CatBoost dengan SMOTENC memperoleh performa terbaik dengan accuracy 94,83%, precision macro 93,06%, recall macro 93,06%, dan F1 macro 93,06%. Penerapan SMOTENC meningkatkan recall kelas Tidak Berisiko pada CatBoost dari 85,76% menjadi 89,56%. Hasil interpretasi menunjukkan bahwa Attendance, Hours_Studied, Tutoring_Sessions, Access_to_Resources, dan Previous_Scores merupakan faktor dominan, dengan nilai yang lebih tinggi cenderung menurunkan prediksi Berisiko. Temuan ini menunjukkan bahwa kombinasi ensemble learning, penanganan imbalanced data, validasi statistik, dan explainable AI dapat mendukung sistem peringatan dini berbasis data.

Kata Kunci: Performa Akademik; CatBoost; Explainable AI; Imbalanced data; SMOTENC

Abstract- Academic performance risk prediction is needed to help educational institutions identify students who require early intervention. The objective of this study is to evaluate the effect of imbalanced data handling and model interpretability on academic performance risk prediction. This study uses the secondary Student Performance Factors dataset consisting of 6,377 records with a binary target of At Risk and Not At Risk. The study compares Random Forest, XGBoost, and CatBoost under two scenarios, namely without SMOTENC and with SMOTENC, and adds explainable AI through feature importance, SHAP, and feature-effect direction analysis. Evaluation was conducted using accuracy, macro precision, macro recall, macro F1, weighted F1, class-level recall, confusion matrix, stratified 5-fold cross-validation, and significance testing. The experimental results show that CatBoost with SMOTENC achieves the best performance with 94.83% accuracy, 93.06% macro precision, 93.06% macro recall, and 93.06% macro F1. SMOTENC improves the recall of the Not At Risk class in CatBoost from 85.76% to 89.56%. The interpretation results indicate that Attendance, Hours_Studied, Tutoring_Sessions, Access_to_Resources, and Previous_Scores are dominant factors, with higher values generally reducing the prediction toward the At Risk class. These findings show that ensemble learning, imbalanced data handling, statistical validation, and explainable AI can support data-driven academic early warning systems.

Keywords: Academic Performance; CatBoost; Explainable AI; Imbalanced data; SMOTENC

1. PENDAHULUAN

Pemanfaatan data pendidikan menjadi salah satu pendekatan penting untuk mendukung pengambilan keputusan akademik karena aktivitas belajar, kehadiran, dukungan keluarga, motivasi, dan capaian sebelumnya dapat direkam sebagai data yang dapat dianalisis. Kajian komprehensif terbaru menunjukkan bahwa machine learning, deep learning, dan explainable AI banyak digunakan untuk prediksi performa siswa, dengan tujuan membantu intervensi akademik berbasis data [1]. Penelitian benchmark machine learning juga memperlihatkan bahwa indikator seperti riwayat akademik, waktu belajar, kegagalan sebelumnya, dan karakteristik keluarga dapat digunakan untuk memprediksi performa akademik serta mendukung identifikasi siswa berisiko [2].

Dalam konteks sekolah dan perguruan tinggi, prediksi performa akademik tidak hanya bertujuan mengetahui nilai akhir, tetapi juga membantu mengidentifikasi siswa yang membutuhkan perhatian lebih awal. Pendekatan prediksi dini yang berpusat pada kebutuhan pengguna pendidikan dinilai penting karena hasil model harus dapat diterjemahkan menjadi keputusan yang dapat ditindaklanjuti [3]. Penelitian mengenai *at-risk student* juga menekankan bahwa deteksi dini dapat membantu institusi memberikan dukungan sebelum siswa mengalami kegagalan akademik atau *dropout* [4].

Perkembangan penelitian lima tahun terakhir menunjukkan bahwa prediksi performa akademik telah dilakukan dengan berbagai algoritma *machine learning*. Beberapa kajian mengevaluasi model klasik untuk melihat performa prediksi akademik pada data pendidikan [5]. Kajian terbaru juga membandingkan beberapa model *machine learning* untuk memperoleh model yang lebih stabil dalam memprediksi performa akademik siswa [6].

Selain model klasik, pendekatan ensemble learning mulai banyak digunakan karena mampu menangkap hubungan nonlinier pada data tabular. Penelitian terbaru yang menggabungkan feature selection, Gradient Boosting, Random Forest, dan SHAP memperlihatkan bahwa model berbasis pohon dapat memperoleh performa

kompetitif sekaligus tetap interpretabel dalam prediksi outcome akademik [7]. Penelitian lain pada education big data juga menunjukkan bahwa machine learning dapat membantu analisis performa akademik berdasarkan atribut pembelajaran dan lingkungan pendidikan [8].

Penelitian prediksi performa akademik juga berkembang pada lingkungan pembelajaran digital. Pada pembelajaran berbantuan video conference, kombinasi SMOTE, Random Forest, dan SHAP digunakan untuk memprediksi performa siswa serta menjelaskan atribut yang berhubungan dengan keberhasilan pembelajaran daring [9]. Pendekatan Hybrid-PSP juga menggunakan feature importance dan beberapa algoritma machine learning untuk memprediksi performa siswa, sehingga menegaskan pentingnya pemilihan fitur dan komparasi model dalam educational data mining [10].

Kebutuhan sistem peringatan dini semakin penting karena performa akademik berkaitan dengan risiko keterlambatan pembelajaran dan dropout. Berbagai pendekatan, termasuk knowledge distillation serta model tabular dan graph-based, telah digunakan untuk mendeteksi siswa berisiko lebih awal [11], [12]. Namun, ketidakseimbangan kelas masih menjadi masalah karena dapat membuat model lebih dominan pada kelas mayoritas. Penelitian sebelumnya menunjukkan bahwa SMOTE dan teknik generatif dapat membantu meningkatkan keseimbangan prediksi, meskipun efektivitasnya bergantung pada karakteristik data dan algoritma yang digunakan [13], [14].

Keterbatasan lain adalah interpretabilitas model. Model prediksi akademik yang digunakan untuk mendukung keputusan pendidikan sebaiknya tidak hanya menghasilkan label prediksi, tetapi juga menjelaskan faktor yang menyebabkan siswa diprediksi berisiko. Penelitian berbasis Random Forest, XGBoost, SHAP, dan LIME menunjukkan bahwa explainable AI dapat memperjelas faktor akademik yang memengaruhi prediksi risiko, baik pada interpretasi global maupun lokal [15]. Evaluasi terhadap explainers pada prediksi keberhasilan siswa juga menunjukkan bahwa interpretasi model penting agar hasil prediksi dapat dipercaya oleh pemangku kepentingan pendidikan [16].

Dalam domain pendidikan, interpretabilitas juga berkaitan dengan keadilan dan keberterimaan model. Fairness metric pernah dibahas dalam prediksi performa akademik untuk memastikan model tidak hanya akurat, tetapi juga tidak merugikan kelompok tertentu [17]. Kajian explainable AI untuk evaluasi siswa berbasis ujian menunjukkan bahwa faktor sosial, akademik, dan ekonomi dapat dianalisis secara kausal dan prediktif melalui model yang dapat dijelaskan [18].

Penelitian terdahulu juga menekankan bahwa rekayasa fitur berperan penting dalam meningkatkan performa prediksi akademik. *Feature engineering* pada data learning management system digunakan untuk mengoptimalkan model prediksi performa siswa [19]. Pada konteks lain, transfer learning digunakan untuk meningkatkan kemampuan adaptasi model prediksi performa siswa pada kondisi data yang berbeda [20].

Berdasarkan telaah tersebut, gap penelitian ini terletak pada perlunya pengujian komparatif algoritma *ensemble learning* pada data performa akademik yang memiliki karakteristik campuran numerik dan kategorikal, disertai penanganan *imbalanced data* dan interpretasi explainable AI. Random Forest dipilih sebagai baseline ensemble yang stabil, XGBoost digunakan sebagai metode gradient boosting yang kuat pada data tabular, sedangkan CatBoost dipilih karena mendukung karakteristik fitur kategorikal [21], [22]. Penelitian ini juga menggunakan SMOTENC karena data yang digunakan memuat fitur numerik dan kategorikal sehingga lebih sesuai dibandingkan *oversampling* numerik biasa [23].

Tujuan penelitian ini adalah membandingkan performa Random Forest, XGBoost, dan CatBoost dalam memprediksi risiko performa akademik siswa, mengevaluasi pengaruh SMOTENC pada data tidak seimbang, serta mengidentifikasi faktor dominan melalui feature importance dan SHAP. Kebaruan penelitian terletak pada integrasi komparasi tiga algoritma ensemble learning tersebut dengan SMOTENC, stratified k-fold cross-validation, uji signifikansi, dan explainable AI dalam satu alur eksperimen. Berbeda dari penelitian sebelumnya yang umumnya menekankan akurasi model, penelitian ini mengevaluasi performa menggunakan accuracy, macro F1, recall per kelas, kestabilan hasil antar fold, serta perbedaan performa antar model melalui uji statistik. Selain itu, feature importance dan SHAP digunakan untuk menjelaskan faktor dominan dan mendukung interpretasi hasil sebagai dasar pengembangan sistem peringatan dini akademik.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian disusun mengikuti alur kerja eksperimen machine learning mulai dari pengumpulan dataset sampai interpretasi model. Gambar 1 memperlihatkan enam tahap utama penelitian, sedangkan penjelasan rinci mengenai tujuan, proses, dan implementasi setiap tahap disajikan pada subbab berikutnya.



Gambar 1. Tahapan Penelitian

Gambar 1 menunjukkan bahwa penelitian dimulai dari pengumpulan data sekunder, kemudian dilanjutkan dengan eksplorasi data dan pembentukan target. Setelah itu dilakukan *preprocessing* dan penanganan *imbalanced data*, pemodelan, evaluasi, serta interpretasi explainable AI. Alur ini digunakan agar pembahasan pada bagian hasil tetap konsisten dengan tahapan metodologi.

2.2 Pengumpulan Data Sekunder

Tahap pertama adalah pengumpulan dataset sekunder *Student Performance Factors* dari Kaggle [24]. Penggunaan data sekunder publik memungkinkan eksperimen direplikasi oleh peneliti lain karena struktur data dan atributnya dapat diperiksa kembali. Dataset terdiri atas 6.377 data dan 20 atribut yang menggambarkan faktor belajar, faktor lingkungan, faktor sosial-ekonomi, serta skor ujian siswa. Penggunaan data pendidikan untuk prediksi performa telah banyak dilakukan dalam educational data mining karena data tersebut dapat merepresentasikan pola belajar dan risiko akademik [1].

Untuk memperjelas transparansi dataset, data diakses pada 20 Juni 2026 melalui halaman Kaggle Student Performance Factors. Dataset ini tidak memuat atribut identitas langsung siswa, seperti nama atau nomor induk, sehingga analisis dilakukan pada level observasi anonim. Target awal dataset adalah Exam_Score, sedangkan target Risk_Performance dibentuk secara operasional dalam penelitian ini. Kolom Exam_Score hanya digunakan sebagai dasar pembentukan target dan selanjutnya dikeluarkan dari fitur prediktor untuk mencegah target leakage. Setelah proses tersebut, model menggunakan 19 fitur prediktor yang terdiri atas 6 fitur numerik dan 13 fitur kategorikal, dengan jumlah missing value sebesar 0 pada seluruh kolom.

Rincian atribut yang digunakan dalam penelitian ditampilkan pada Tabel 1. Tabel ini penting karena menjelaskan variabel masukan yang dipakai model sebelum dilakukan pembentukan target risiko.

Tabel 1. Atribut Dataset *Student Performance Factors*

No	Atribut	Keterangan
1	Hours_Studied	Jumlah jam belajar siswa
2	Attendance	Persentase kehadiran siswa
3	Parental_Involvement	Tingkat keterlibatan orang tua
4	Access_to_Resources	Akses terhadap sumber daya belajar
5	Extracurricular_Activities	Keikutsertaan dalam kegiatan ekstrakurikuler
6	Sleep_Hours	Durasi tidur siswa
7	Previous_Scores	Nilai akademik sebelumnya
8	Motivation_Level	Tingkat motivasi belajar
9	Internet_Access	Ketersediaan akses internet
10	Tutoring_Sessions	Jumlah sesi bimbingan belajar
11	Family_Income	Tingkat pendapatan keluarga
12	Teacher_Quality	Kualitas guru
13	School_Type	Jenis sekolah
14	Peer_Influence	Pengaruh teman sebaya
15	Physical_Activity	Aktivitas fisik siswa
16	Learning_Disabilities	Kondisi kesulitan belajar
17	Parental_Education_Level	Tingkat pendidikan orang tua
18	Distance_from_Home	Jarak dari rumah ke sekolah
19	Gender	Jenis kelamin
20	Exam_Score	Kategori nilai ujian awal

Berdasarkan Tabel 1, dataset memiliki kombinasi atribut numerik dan kategorikal. Kondisi ini menjadi pertimbangan utama dalam pemilihan metode, terutama karena algoritma berbasis pohon relatif sesuai untuk data tabular dan SMOTENC dapat digunakan pada data campuran numerik-kategorikal [23].

2.3 Eksplorasi Data dan Pembentukan Target

Tahap kedua adalah eksplorasi data untuk melihat struktur atribut, kelengkapan nilai, dan distribusi target. Target awal dataset adalah Exam_Score dengan empat kategori, yaitu *Poor*, *Acceptable*, *Good*, dan *Excellent*. Namun, fokus penelitian ini adalah prediksi risiko performa akademik, sehingga target dibentuk ulang menjadi dua kelas,

yaitu Berisiko dan Tidak Berisiko. Pendekatan pembentukan target risiko sejalan dengan penelitian at-risk student yang menempatkan prediksi sebagai dasar intervensi akademik [25]. Konsep prediksi ini juga digunakan dalam penelitian *dropout risk* karena model perlu memberi sinyal sebelum masalah akademik menjadi lebih serius [26].

Pemetaan target dilakukan dengan mengelompokkan *Poor* dan *Acceptable* sebagai Berisiko, sedangkan *Good* dan *Excellent* sebagai Tidak Berisiko. Pengelompokan ini dipilih agar keluaran model lebih mudah diterjemahkan menjadi kebutuhan intervensi akademik. Dengan demikian, model tidak hanya memprediksi kategori nilai, tetapi memberikan informasi apakah siswa berada pada kondisi performa yang perlu diperhatikan.

Pembentukan target tersebut merupakan definisi operasional penelitian, bukan label risiko yang tersedia secara langsung pada dataset. Kelas Berisiko merepresentasikan observasi dengan kategori *Exam_Score Poor* atau *Acceptable*, sedangkan kelas Tidak Berisiko merepresentasikan *Good* atau *Excellent*. Karena batas ini ditetapkan untuk kebutuhan eksperimen *early warning*, hasil model harus dibaca sebagai prediksi risiko relatif terhadap skema pengelompokan tersebut dan bukan sebagai diagnosis akademik individual.

2.4 Preprocessing dan SMOTENC

Tahap ketiga adalah preprocessing data. Proses ini meliputi pemeriksaan missing value, pemisahan fitur dan target, identifikasi atribut numerik serta kategorikal, encoding fitur kategorikal, dan pembagian data latih serta data uji. Kerangka machine learning terbaru dalam prediksi performa akademik menekankan pentingnya preprocessing, pemilihan fitur, dan evaluasi yang transparan agar hasil model dapat direplikasi dan digunakan sebagai dasar pengambilan keputusan pendidikan [27]. Dalam penelitian ini, data dibagi menjadi data latih dan data uji dengan proporsi 80:20 menggunakan `random_state 42` agar hasil eksperimen lebih stabil.

Untuk mencegah target *leakage*, kolom *Exam_Score* yang menjadi dasar pembentukan *Risk_Performance* dikeluarkan dari himpunan fitur sebelum pelatihan model. Setelah kolom tersebut dikeluarkan, model menggunakan 19 prediktor yang terdiri atas 6 fitur numerik dan 13 fitur kategorikal. Proses encoding dipelajari hanya dari data latih: *OrdinalEncoder* di-fit pada `X_train` dan selanjutnya digunakan untuk mentransformasi `X_test`. SMOTENC diterapkan hanya pada data latih setelah pembagian *stratified 80:20*, sehingga data uji tidak mengandung sampel sintetis dan tetap merepresentasikan distribusi asli.

Karena target memiliki distribusi tidak seimbang, penelitian menerapkan dua skenario, yaitu tanpa SMOTENC dan dengan SMOTENC. SMOTE pada dasarnya menghasilkan sampel sintetis untuk kelas minoritas [28]. Namun, karena dataset ini memiliki fitur numerik dan kategorikal, penelitian menggunakan SMOTENC yang dirancang untuk data campuran nominal dan kontinu [23]. Kajian teoretis mengenai SMOTE menunjukkan bahwa distribusi sampel sintetis perlu dipahami agar efek oversampling tidak ditafsirkan secara berlebihan [29]. Selain itu, pemetaan terbaru mengenai class imbalance menegaskan bahwa pemilihan teknik balancing harus disesuaikan dengan konteks data, algoritma, dan metrik evaluasi yang digunakan [30].

Alur preprocessing CatBoost dan SMOTENC dilakukan secara berurutan untuk menjaga validitas eksperimen. Pertama, target *Risk_Performance* dibentuk dari *Exam_Score*, kemudian *Exam_Score* dihapus dari fitur prediktor. Kedua, data dibagi menggunakan *stratified train-test split 80:20*. Ketiga, fitur kategorikal diidentifikasi sebagai indeks *categorical_features* pada SMOTENC dan *cat_features* pada CatBoost. Keempat, *OrdinalEncoder* hanya di-fit pada data latih dan digunakan untuk mentransformasi data uji. Kelima, pada skenario dengan SMOTENC, oversampling diterapkan hanya pada data latih, sedangkan data uji tetap mengikuti distribusi asli. Pada proses *stratified 5-fold cross-validation*, tahapan encoding dan SMOTENC diulang di setiap fold sehingga data validasi tidak digunakan dalam pembentukan sampel sintetis.

2.5 Pemodelan Random Forest, XGBoost, dan CatBoost

Tahap keempat adalah pemodelan menggunakan tiga algoritma *ensemble learning*. Random Forest digunakan sebagai *baseline* karena membangun banyak pohon keputusan dan menggabungkan prediksi untuk meningkatkan stabilitas model [31]. Algoritma ini cocok digunakan sebagai pembandingan karena sederhana, relatif kuat terhadap variasi data, dan dapat menghasilkan informasi *feature importance*.

XGBoost digunakan sebagai model *gradient boosting* yang banyak diterapkan pada data tabular karena efisien dan memiliki kemampuan regularisasi [21]. Metode ini membangun model secara bertahap dengan memperbaiki kesalahan model sebelumnya. Dalam penelitian prediksi performa akademik, pendekatan *machine learning* berbasis komparasi model juga banyak digunakan untuk memperoleh model terbaik [6].

CatBoost digunakan karena dataset memuat banyak fitur kategorikal, seperti *Parental_Involvement*, *Access_to_Resources*, *Motivation_Level*, *Family_Income*, *Teacher_Quality*, *School_Type*, *Peer_Influence*, dan *Distance_from_Home*. CatBoost dikembangkan untuk menangani fitur kategorikal dan mengurangi bias prediksi melalui *ordered boosting* [22]. Karena karakteristik dataset campuran, CatBoost diperkirakan mampu memberikan performa yang kompetitif dibandingkan Random Forest dan XGBoost. Pengaturan hyperparameter ditampilkan pada Tabel 2 agar eksperimen dapat direplikasi.

Tabel 2. Hyperparameter dan Pengaturan Eksperimen

Model	Hyperparameter
Random Forest	n_estimators=120; max_depth=None; random_state=42; n_jobs=-1
XGBoost	n_estimators=120; learning_rate=0,08; max_depth=3; subsample=0,9; colsample_bytree=0,9; random_state=42; eval_metric=logloss; tree_method=hist; n_jobs=2
CatBoost	iterations=150; learning_rate=0,08; depth=5; random_seed=42; verbose=False; cat_features=fitur kategorikal; thread_count=2
SMOTENC	categorical_features=fitur kategorikal; random_state=42; k_neighbors=5
Stratified K-Fold	n_splits=5; shuffle=True; random_state=42

Tabel 2 menunjukkan konfigurasi yang digunakan pada seluruh eksperimen, termasuk SMOTENC dan stratified k-fold. Eksperimen dilakukan menggunakan Google Colab dengan Python 3.13.15. Library utama yang digunakan meliputi scikit-learn 1.6.1, imbalanced-learn 0.14.2, XGBoost 3.4.1, CatBoost 1.2.10, dan SHAP 0.52.0. Seluruh eksperimen menggunakan `random_state=42` untuk mendukung reproduktibilitas hasil.

2.6 Evaluasi Model

Tahap kelima adalah evaluasi model. Evaluasi tidak hanya menggunakan accuracy, tetapi juga precision macro, recall macro, F1 macro, F1 weighted, recall per kelas, dan *confusion matrix*. Pada data tidak seimbang, accuracy dapat memberi gambaran terlalu optimistis jika model dominan memprediksi kelas mayoritas. Oleh karena itu, macro F1 dan recall per kelas digunakan agar performa model pada kelas Berisiko dan Tidak Berisiko dapat dianalisis lebih seimbang. Penggunaan metrik tambahan relevan dengan kajian *imbalanced classification* yang menekankan bahwa model perlu dievaluasi dari kemampuan mengenali kelas minoritas [13].

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

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

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

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

Formula evaluasi yang digunakan ditunjukkan pada persamaan (1) sampai (4). TP, TN, FP, dan FN masing-masing merepresentasikan *true positive*, *true negative*, *false positive*, dan *false negative*. Dalam penelitian ini kelas Berisiko dikodekan sebagai kelas positif (1), sedangkan Tidak Berisiko sebagai kelas negatif (0). Dengan definisi tersebut, TP adalah siswa Berisiko yang diprediksi Berisiko; FN adalah siswa Berisiko yang diprediksi Tidak Berisiko; FP adalah siswa Tidak Berisiko yang diprediksi Berisiko; dan TN adalah siswa Tidak Berisiko yang diprediksi Tidak Berisiko.

Selain evaluasi holdout stratified 80:20, penelitian ini menambahkan stratified 5-fold cross-validation untuk menguji kestabilan performa model pada pembagian data yang berbeda. Pada skenario dengan SMOTENC, oversampling hanya diterapkan pada data latih di setiap fold sehingga data validasi tidak mengandung sampel sintetis. Perbedaan performa antar kombinasi model diuji menggunakan Friedman Test, kemudian model terbaik dibandingkan dengan pembandingnya menggunakan Wilcoxon Signed-Rank Test berdasarkan nilai F1 macro per fold.

2.7 Explainable AI

Tahap keenam adalah interpretasi model terbaik menggunakan feature importance dan SHAP. Feature importance digunakan untuk melihat kontribusi global fitur pada model CatBoost, sedangkan SHAP digunakan dalam dua bentuk, yaitu mean absolute SHAP untuk menunjukkan besarnya kontribusi fitur dan tanda nilai SHAP untuk membaca arah pengaruh fitur terhadap kelas Berisiko. Nilai SHAP positif menunjukkan bahwa suatu fitur mendorong prediksi ke arah Berisiko, sedangkan nilai SHAP negatif menunjukkan bahwa fitur tersebut menurunkan prediksi Berisiko [32]. Pendekatan ini sejalan dengan studi XAI pada prediksi performa akademik yang menekankan pentingnya transparansi, interpretasi global-lokal, serta kestabilan faktor dalam learning analytics [15], [16], [33].

3. HASIL DAN PEMBAHASAN

3.1 Karakteristik Data dan Target Risiko

Hasil eksplorasi data menunjukkan bahwa target awal *Exam_Score* memiliki empat kategori dengan distribusi yang tidak seimbang. Rincian jumlah data pada setiap kategori ditampilkan pada Tabel 3. Tabel ini digunakan untuk memperlihatkan kondisi target sebelum dilakukan penyederhanaan menjadi label risiko. Pemeriksaan kualitas data menunjukkan bahwa 6.377 observasi yang digunakan tidak memiliki missing value. Setelah *Exam_Score* dikeluarkan dari fitur prediktor, komposisi data terdiri atas 6 fitur numerik dan 13 fitur kategorikal. Karakteristik campuran ini mendukung penggunaan skema preprocessing yang membedakan perlakuan fitur numerik dan kategorikal serta menjadi dasar pemilihan SMOTENC pada skenario penyeimbangan data.

Tabel 3. Distribusi Target Awal *Exam_Score*

No	Kategori <i>Exam_Score</i>	Jumlah
1	Acceptable	4731
2	Good	1546
3	Poor	66
4	Excellent	34

Tabel 3 menunjukkan bahwa kelas Acceptable menjadi kelas mayoritas dengan 4.731 data, sedangkan kelas Excellent hanya 34 data dan Poor hanya 66 data. Ketimpangan ini menunjukkan bahwa klasifikasi empat kelas berpotensi membuat model sulit mempelajari kelas yang jumlahnya sangat sedikit. Oleh karena itu, target penelitian difokuskan menjadi dua kelas risiko agar lebih sesuai dengan tujuan *early warning system*.

Setelah dilakukan pembentukan target, *Poor* dan *Acceptable* dikelompokkan sebagai Berisiko, sedangkan *Good* dan *Excellent* dikelompokkan sebagai Tidak Berisiko. Hasil pembentukan target risiko ditampilkan pada Tabel 4.

Tabel 4. Distribusi Target *Risk_Performance*

No	Kategori Risiko	Jumlah
1	Berisiko	4797
2	Tidak Berisiko	1580

Tabel 4 menunjukkan bahwa kelas Berisiko berjumlah 4.797 data dan kelas Tidak Berisiko berjumlah 1.580 data. Meskipun target sudah disederhanakan menjadi dua kelas, distribusi data masih tidak seimbang. Kondisi tersebut memperkuat alasan penggunaan SMOTENC dan metrik evaluasi yang tidak hanya bergantung pada accuracy.

3.2 Dampak Penanganan *Imbalanced data*

Tahap preprocessing menghasilkan dua skenario pengujian, yaitu tanpa SMOTENC dan dengan SMOTENC. Skenario tanpa SMOTENC digunakan untuk melihat kemampuan model pada data asli, sedangkan skenario dengan SMOTENC digunakan untuk melihat dampak penyeimbangan data latih. Pendekatan ini mengikuti pandangan bahwa strategi balancing harus diuji secara empiris karena tidak selalu meningkatkan seluruh metrik pada semua model [13], [14], [30]. Pada penelitian ini, SMOTENC hanya diterapkan pada data latih agar data uji tetap merepresentasikan distribusi asli. Cara ini dilakukan untuk menghindari kebocoran data dan menjaga validitas evaluasi.

Pada pembagian stratified 80:20, data latih terdiri atas 5.101 observasi, yaitu 3.837 Berisiko dan 1.264 Tidak Berisiko, sedangkan data uji terdiri atas 960 Berisiko dan 316 Tidak Berisiko. Setelah SMOTENC diterapkan hanya pada data latih, kelas minoritas bertambah menjadi 3.837 sehingga data latih hasil resampling berjumlah 7.674 observasi dengan komposisi seimbang 3.837:3.837. Distribusi data uji tidak diubah. Hasil ini menunjukkan bahwa skenario SMOTENC mengubah ruang belajar model tanpa mengubah basis evaluasi.

3.3 Komparasi Model

Setelah *preprocessing*, tiga algoritma diuji pada dua skenario sehingga terdapat enam kombinasi eksperimen. Hasil pengujian ditampilkan pada Tabel 5. Tabel ini memuat *accuracy*, *precision macro*, *recall macro*, *F1 macro*, *F1 weighted*, dan waktu pelatihan untuk melihat performa model secara menyeluruh.

Tabel 5. Hasil Komparasi Model

Skenario	Model	Accuracy	Precision Macro	Recall Macro	F1 Macro	F1 Weighted	Training (s)
Dengan SMOTENC	CatBoost	94.83%	93.06%	93.06%	93.06%	94.83%	1.1523
Tanpa SMOTENC	CatBoost	94.59%	93.64%	91.63%	92.58%	94.53%	0.7615
Tanpa SMOTENC	XGBoost	94.12%	94.01%	89.94%	91.74%	93.98%	0.1686
Dengan SMOTENC	XGBoost	93.34%	90.77%	91.54%	91.15%	93.37%	0.1642
Dengan SMOTENC	Random Forest	92.01%	90.12%	88.00%	88.99%	91.90%	1.2611
Tanpa SMOTENC	Random Forest	91.69%	92.34%	84.93%	87.84%	91.30%	1.2371

Berdasarkan Tabel 5, model terbaik adalah CatBoost dengan SMOTENC yang memperoleh *accuracy* 94,83%, *precision macro* 93,06%, *recall macro* 93,06%, dan *F1 macro* 93,06%. Pada skenario tanpa SMOTENC, CatBoost memperoleh *accuracy* 94,59% dan *F1 macro* 92,58%. Kenaikan tersebut menunjukkan bahwa penerapan SMOTENC memberikan perbaikan pada CatBoost meskipun peningkatannya tidak terlalu besar pada *accuracy*.

XGBoost tanpa SMOTENC memperoleh *accuracy* 94,12% dan *F1 macro* 91,74%, sedangkan XGBoost dengan SMOTENC memperoleh *accuracy* 93,34% dan *F1 macro* 91,15%. Hasil ini menunjukkan bahwa SMOTENC tidak selalu meningkatkan seluruh metrik pada XGBoost. Random Forest menghasilkan performa paling rendah dibandingkan dua model *boosting*, yaitu *accuracy* 91,69% tanpa SMOTENC dan 92,01% dengan SMOTENC. Meskipun demikian, penerapan SMOTENC tetap meningkatkan *F1 macro* Random Forest dari 87,84% menjadi 88,99%.

Jika perubahan metrik dianalisis dalam poin persentase, CatBoost memperoleh peningkatan *accuracy* 0,24 poin dan *F1 macro* 0,48 poin setelah SMOTENC; *recall* Tidak Berisiko naik 3,80 poin, sementara *recall* Berisiko turun 0,94 poin. Pada Random Forest, *recall* Tidak Berisiko meningkat 8,54 poin dengan kenaikan *F1 macro* 1,15 poin. Sebaliknya, XGBoost mengalami penurunan *accuracy* 0,78 poin dan *F1 macro* 0,59 poin, meskipun *recall* Tidak Berisiko meningkat 6,32 poin. Pola ini menegaskan adanya trade-off antara sensitivitas terhadap kelas minoritas dan performa agregat; karena itu, efektivitas *balancing* tidak layak dinilai hanya dari *accuracy* [29], [31].

3.4 Validasi Silang dan Uji Signifikansi

Validasi stratified 5-fold cross-validation dilakukan untuk memperkuat hasil evaluasi model yang sebelumnya diperoleh melalui pembagian holdout 80:20. Validasi ini penting karena hasil satu kali pembagian data dapat dipengaruhi oleh komposisi data latih dan data uji. Melalui lima fold yang berbeda, setiap kombinasi model dan skenario diuji secara berulang sehingga kestabilan performa dapat diamati dari nilai rata-rata dan standar deviasi. Hasil rata-rata *accuracy* dan *F1 macro* pada enam kombinasi eksperimen ditampilkan pada Tabel 6.

Tabel 6. Hasil Stratified 5-Fold Cross-Validation

Skenario	Model	Accuracy Mean	Accuracy Std	F1 Macro Mean	F1 Macro Std
Dengan SMOTENC	CatBoost	94,31%	0,61%	92,47%	0,76%
Tanpa SMOTENC	CatBoost	94,10%	1,24%	91,78%	1,77%
Dengan SMOTENC	XGBoost	92,65%	0,99%	90,27%	1,28%
Tanpa SMOTENC	XGBoost	92,83%	1,04%	89,92%	1,45%
Dengan SMOTENC	Random Forest	91,23%	0,69%	88,01%	0,96%
Tanpa SMOTENC	Random Forest	90,90%	1,05%	86,90%	1,51%

Berdasarkan Tabel 6, CatBoost dengan SMOTENC memperoleh performa rata-rata tertinggi dengan *accuracy* 94,31% dan *F1 macro* 92,47%, serta standar deviasi *F1 macro* 0,76% yang menunjukkan performa relatif stabil antar fold. CatBoost tanpa SMOTENC berada di bawahnya dengan *accuracy* 94,10% dan *F1 macro* 91,78%. Hasil

ini menunjukkan bahwa SMOTENC memberikan peningkatan moderat pada CatBoost, sementara XGBoost dan Random Forest memperoleh performa lebih rendah.

Karena nilai F1 macro antar model relatif berdekatan, dilakukan uji signifikansi menggunakan F1 macro per fold. Friedman Test digunakan untuk menguji perbedaan performa secara umum, sedangkan Wilcoxon Signed-Rank Test membandingkan model terbaik dengan model pembanding. Hasil pengujian ditampilkan pada Tabel 7.

Tabel 7. Hasil Uji Signifikansi Statistik

Pengujian	Statistik	p-value	Kesimpulan
Friedman Test seluruh kombinasi	23,4000	0,000283	Signifikan
Wilcoxon model terbaik vs pembanding	0,0000-4,0000	0,0625-0,4375	Tidak signifikan

Tabel 7 menunjukkan bahwa Friedman Test signifikan ($p = 0,000283$), sehingga terdapat perbedaan performa secara umum antar kombinasi model. Namun, Wilcoxon Signed-Rank Test menghasilkan p-value 0,0625–0,4375, sehingga perbedaan CatBoost dengan SMOTENC terhadap masing-masing pembanding belum signifikan pada taraf 0,05. Dengan demikian, CatBoost dengan SMOTENC unggul secara numerik, tetapi keunggulan tersebut perlu ditafsirkan secara hati-hati.

3.5 Analisis Recall dan Confusion matrix

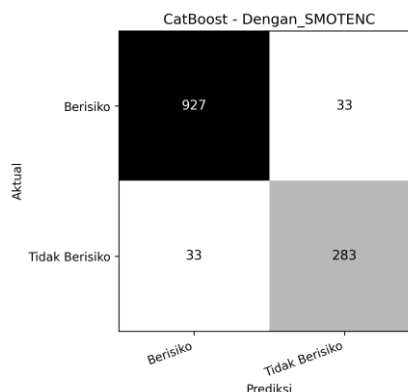
Selain melihat metrik global, penelitian ini menganalisis recall per kelas karena tujuan utama model adalah mengidentifikasi risiko performa akademik. Rincian recall kelas Berisiko dan Tidak Berisiko ditampilkan pada Tabel 8.

Tabel 8. Recall Per Kelas pada Setiap Model

Skenario	Model	Recall Berisiko	Recall Tidak Berisiko
Dengan SMOTENC	CatBoost	96.56%	89.56%
Tanpa SMOTENC	CatBoost	97.50%	85.76%
Tanpa SMOTENC	XGBoost	98.23%	81.65%
Dengan SMOTENC	XGBoost	95.10%	87.97%
Dengan SMOTENC	Random Forest	95.94%	80.06%
Tanpa SMOTENC	Random Forest	98.33%	71.52%

Tabel 8 menunjukkan bahwa CatBoost dengan SMOTENC memiliki recall Berisiko sebesar 96,56% dan recall Tidak Berisiko sebesar 89,56%. Dibandingkan CatBoost tanpa SMOTENC, recall Tidak Berisiko meningkat dari 85,76% menjadi 89,56%. Hal ini berarti SMOTENC membantu model mengenali kelas Tidak Berisiko dengan lebih baik, meskipun recall Berisiko sedikit lebih rendah dibandingkan skenario tanpa SMOTENC.

Untuk memperjelas pola kesalahan klasifikasi model terbaik, Gambar 2 menampilkan *confusion matrix* CatBoost dengan SMOTENC. Gambar ini digunakan untuk melihat jumlah prediksi benar dan salah pada kedua kelas.



Gambar 2. Confusion matrix CatBoost dengan SMOTENC

Secara numerik, *confusion matrix* model terbaik mencatat 927 *true positive*, 283 *true negative*, 33 *false positive*, dan 33 *false negative*. Dengan total 1.276 data uji, sebanyak 1.210 observasi diklasifikasikan dengan

benar dan 66 observasi mengalami salah klasifikasi. Nilai *false negative* yang berjumlah 33 penting diperhatikan karena kasus ini merepresentasikan siswa yang secara operasional termasuk Berisiko tetapi diprediksi Tidak Berisiko. Dalam konteks *early warning*, jenis kesalahan tersebut dapat menyebabkan siswa yang membutuhkan perhatian tidak terdeteksi. Sebaliknya, *false positive* juga berjumlah 33 dan dapat meningkatkan beban intervensi karena siswa Tidak Berisiko ditandai sebagai Berisiko. Keseimbangan kedua jenis kesalahan tersebut mendukung penggunaan recall per kelas bersama macro F1, bukan *accuracy* saja.

3.6 Interpretasi Faktor Dominan

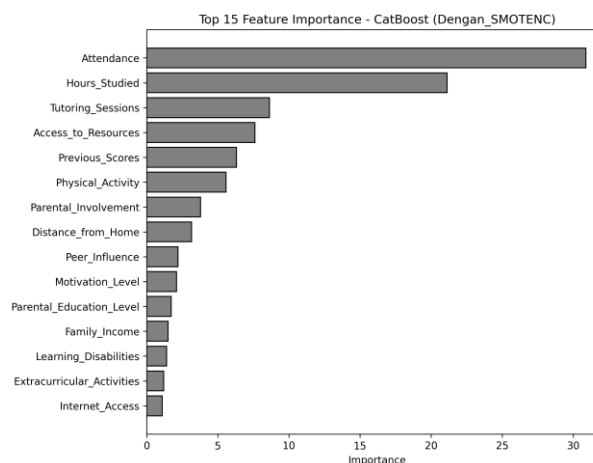
Setelah model terbaik ditentukan, penelitian melakukan interpretasi faktor dominan menggunakan *feature importance*. Hasil lima belas fitur terpenting ditampilkan pada Tabel 9. Tabel ini menjelaskan atribut yang paling banyak berkontribusi dalam pembentukan keputusan model CatBoost dengan SMOTENC.

Tabel 9. Lima Belas Fitur Terpenting pada Model Terbaik

Peringkat	Fitur	Importance
1	Attendance	30.8752
2	Hours_Studied	21.1141
3	Tutoring_Sessions	8.6218
4	Access_to_Resources	7.6119
5	Previous_Scores	6.3157
6	Physical_Activity	5.5770
7	Parental_Involvement	3.7906
8	Distance_from_Home	3.1434
9	Peer_Influence	2.1950
10	Motivation_Level	2.0922
11	Parental_Education_Level	1.7231
12	Family_Income	1.4958
13	Learning_Disabilities	1.3956
14	Extracurricular_Activities	1.1863
15	Internet_Access	1.0811

Berdasarkan Tabel 9, Attendance menjadi fitur paling berpengaruh dengan nilai importance 30,8752, diikuti Hours_Studied sebesar 21,1141, Tutoring_Sessions sebesar 8,6218, Access_to_Resources sebesar 7,6119, dan Previous_Scores sebesar 6,3157. Urutan ini menunjukkan bahwa faktor kehadiran dan intensitas belajar memiliki kontribusi paling besar dalam memprediksi risiko performa akademik siswa. Temuan ini sejalan dengan penelitian terbaru yang menempatkan riwayat akademik, waktu belajar, dan atribut pembelajaran sebagai indikator penting dalam prediksi performa siswa [2], [27].

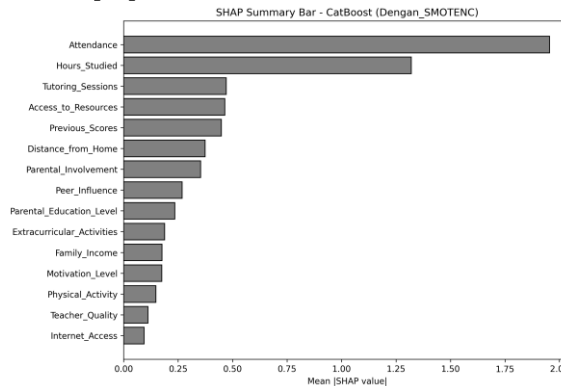
Visualisasi *feature importance* disajikan pada Gambar 3 untuk memperjelas perbandingan kontribusi antar fitur. Grafik ini memudahkan pembaca melihat perbedaan dominasi fitur secara lebih cepat dibandingkan tabel numerik.



Gambar 3. *Feature importance* Model Terbaik

Gambar 3 memperlihatkan bahwa Attendance dan Hours_Studied jauh lebih dominan dibandingkan fitur lainnya. Secara praktis, hal ini dapat diartikan bahwa sistem peringatan dini akademik sebaiknya tidak hanya melihat hasil nilai, tetapi juga memantau kehadiran, jam belajar, dan dukungan pembelajaran. Faktor tutoring dan akses sumber daya juga penting karena dapat menggambarkan ketersediaan bantuan belajar bagi siswa.

Selain *feature importance*, penelitian menggunakan SHAP untuk memperkuat interpretasi model. Visualisasi SHAP ditampilkan pada Gambar 4. SHAP digunakan karena mampu menjelaskan kontribusi fitur terhadap prediksi model secara lebih konsisten [33].



Gambar 4. SHAP Summary Bar Model Terbaik

Gambar 4 menunjukkan pola yang sejalan dengan *feature importance*, yaitu Attendance, Hours_Studied, dan Tutoring_Sessions muncul sebagai fitur dominan. Kesesuaian antara *feature importance* dan SHAP memperkuat keandalan interpretasi model. Pada konteks pendidikan, interpretasi semacam ini penting karena hasil prediksi dapat diarahkan menjadi rekomendasi tindakan, misalnya pemantauan kehadiran, dukungan belajar tambahan, dan perbaikan akses sumber daya belajar [34], [35].

Untuk menunjukkan arah pengaruh fitur, nilai SHAP dianalisis terhadap kelas Berisiko. Nilai SHAP positif menunjukkan fitur mendorong prediksi ke arah Berisiko, sedangkan nilai SHAP negatif menunjukkan fitur menurunkan prediksi Berisiko. Ringkasan arah pengaruh fitur dominan ditampilkan pada Tabel 10.

Tabel 10. Arah Pengaruh Fitur Berdasarkan SHAP

Fitur	Nilai/Kategori yang Meningkatkan Risiko	Nilai/Kategori yang Menurunkan Risiko
Attendance	Nilai rendah	Nilai tinggi
Hours_Studied	Nilai rendah	Nilai tinggi
Tutoring_Sessions	Nilai rendah	Nilai tinggi
Access_to_Resources	Low	High
Previous_Scores	Nilai rendah	Nilai tinggi
Distance_from_Home	Far	Near
Parental_Involvement	Low	High
Peer_Influence	Negative	Positive

Berdasarkan Tabel 10, nilai Attendance, Hours_Studied, Tutoring_Sessions, dan Previous_Scores yang lebih tinggi cenderung menurunkan prediksi Berisiko. Pada fitur kategorikal, Access_to_Resources rendah, Distance_from_Home jauh, Parental_Involvement rendah, dan Peer_Influence negatif cenderung meningkatkan prediksi Berisiko. Hasil ini menunjukkan bahwa SHAP tidak hanya mengurutkan fitur dominan, tetapi juga memperjelas arah kontribusi fitur terhadap keputusan model.

3.7 Pembahasan dan Perbandingan dengan Penelitian Terdahulu

Hasil penelitian memperlihatkan bahwa model boosting lebih unggul dibandingkan Random Forest pada dataset *Student Performance Factors*. CatBoost dengan SMOTENC menjadi model terbaik karena menghasilkan *accuracy*, *recall macro*, dan *F1 macro* tertinggi. Temuan ini sesuai dengan karakteristik CatBoost yang dirancang untuk menangani fitur kategorikal [22]. Pada dataset yang memiliki banyak fitur kategori, kemampuan tersebut memberi keuntungan dibandingkan model yang memerlukan transformasi kategori secara lebih umum.

Dibandingkan penelitian prediksi performa akademik yang hanya berfokus pada akurasi model, penelitian ini menambahkan analisis *imbalanced data* dan interpretasi faktor dominan. Kajian mengenai prediksi performa siswa melalui data mining menunjukkan bahwa pemilihan model perlu disesuaikan dengan karakteristik data [5]. Penelitian evaluasi model prediksi akademik juga menunjukkan pentingnya membandingkan beberapa algoritma agar model terbaik tidak dipilih secara asumsi [6].

Penerapan SMOTENC memberikan dampak paling jelas pada peningkatan recall kelas Tidak Berisiko pada CatBoost. Hal ini menunjukkan bahwa balancing dapat membantu model mengenali pola kelas yang jumlahnya lebih kecil. Namun, hasil pada XGBoost menunjukkan bahwa balancing tidak selalu menaikkan accuracy dan F1 macro. Kondisi tersebut sejalan dengan penelitian terbaru mengenai data tidak seimbang yang menyatakan bahwa efektivitas SMOTE, variasinya, dan pendekatan generatif bergantung pada karakteristik data, algoritma, dan metrik evaluasi yang digunakan [13], [14], [30].

Dari sudut pandang sistem peringatan dini, hasil penelitian ini relevan karena model tidak hanya mengklasifikasikan siswa ke dalam kelas risiko, tetapi juga menunjukkan faktor penyebabnya. Penelitian *human-centered early prediction* menekankan bahwa model prediktif seharusnya dapat digunakan dalam konteks nyata dan mendukung keputusan manusia [3]. Penelitian mengenai *AI-based identification of at-risk students* juga menunjukkan bahwa prediksi risiko dapat menjadi dasar dukungan akademik yang lebih terarah [4].

Interpretasi model menunjukkan bahwa Attendance dan Hours_Studied merupakan faktor paling dominan. Temuan ini dapat dijelaskan karena kehadiran mencerminkan keterpaparan siswa terhadap materi pembelajaran, sedangkan jam belajar menunjukkan intensitas usaha akademik. Hasil ini konsisten dengan studi terbaru yang menunjukkan bahwa faktor akademik, kebiasaan belajar, dan riwayat performa merupakan prediktor penting dalam model prediksi akademik [2], [27]. Selain itu, faktor Access_to_Resources dan Tutoring_Sessions menunjukkan bahwa dukungan pembelajaran eksternal turut memengaruhi risiko performa akademik.

Kontribusi utama penelitian ini adalah integrasi antara komparasi ensemble learning, penanganan imbalanced data, dan explainable AI pada dataset performa akademik. Penelitian sebelumnya telah membahas explainability pada prediksi performa siswa [15], integrasi XAI dalam educational data mining [16], serta stabilitas interpretasi model dalam learning analytics [33]. Namun, penelitian ini menempatkan explainable AI sebagai bagian akhir dari alur eksperimen sehingga hasil model dapat langsung ditafsirkan menjadi faktor-faktor akademik beserta arah pengaruhnya terhadap prediksi Berisiko. Perbandingan posisi penelitian terhadap studi terdahulu disajikan pada Tabel 11.

Tabel 11. Perbandingan dengan Penelitian Terdahulu

Penelitian	Dataset/Metode	Fokus Temuan	Posisi Penelitian Ini
Boujmiraz et al. [1]	Review ML, DL, dan XAI pada student performance prediction	XAI mulai penting untuk keputusan pedagogis, tetapi masih terbatas penerapannya	Mengimplementasikan XAI pada model terbaik dan menambah arah pengaruh SHAP
Al-Shanableh et al. [2]	Benchmark ML pada data pendidikan UCI	Gradient boosting unggul dan feature importance membantu membaca faktor utama	Membandingkan RF, XGBoost, CatBoost serta menilai risiko biner pada data tidak seimbang
Gu dan Zhang [7]	Feature selection, Gradient Boosting, Random Forest, dan SHAP	Model dapat efisien sekaligus interpretabel pada prediksi outcome akademik	Menambahkan SMOTENC, recall per kelas, CV, uji statistik, dan SHAP direction
Chachoui et al. [13]; Alnassar et al. [14]	SMOTE dan teknik balancing pada dataset pendidikan	Balancing membantu mengatasi bias kelas, tetapi efeknya bergantung pada data dan model	Menggunakan SMOTENC untuk fitur numerik-kategorikal dan mengevaluasi dampaknya pada tiga model
Hidayatulloh et al. [15]; Tiukhova et al. [33]	XAI dengan SHAP/LIME dan learning analytics	Interpretasi global-lokal dan stabilitas faktor penting untuk mendukung keputusan pendidikan	Menyajikan feature importance, SHAP magnitude, dan arah pengaruh fitur terhadap kelas Berisiko
Penelitian ini	RF, XGBoost, CatBoost, SMOTENC, CV, Wilcoxon, SHAP	Acc 94,83%; F1 holdout 93,06%; F1 CV 92,47%	Komparasi, balancing, validasi statistik, dan arah pengaruh fitur dalam satu alur eksperimen

Tabel 11 memperlihatkan bahwa penelitian ini memperkuat penelitian terdahulu melalui integrasi komparasi model, penanganan data tidak seimbang, validasi statistik, dan interpretasi arah pengaruh fitur berbasis SHAP. Dibandingkan studi yang hanya menekankan akurasi atau hanya menggunakan XAI sebagai interpretasi global, penelitian ini menyajikan evaluasi holdout, stratified cross-validation, uji signifikansi, recall per kelas, dan arah pengaruh fitur dalam satu alur eksperimen. Dengan demikian, kontribusi penelitian tidak hanya berada pada nilai performa model, tetapi juga pada transparansi proses eksperimen dan keterjelasan faktor yang dapat digunakan untuk mendukung sistem peringatan dini akademik.

Keterbatasan penelitian ini terletak pada penggunaan dataset sekunder publik sehingga konteks data belum tentu sepenuhnya merepresentasikan kondisi sekolah di Indonesia. Selain itu, target risiko dibentuk berdasarkan pengelompokan *Exam_Score* sehingga belum memuat validasi langsung dari kebijakan akademik tertentu. Penelitian selanjutnya dapat menggunakan data institusi lokal, menambahkan validasi ahli pendidikan, dan membandingkan metode balancing lain agar hasil model lebih kuat secara praktis dan metodologis. Validasi tambahan menunjukkan bahwa CatBoost dengan SMOTENC tetap unggul secara numerik, tetapi hasil Wilcoxon belum menunjukkan perbedaan signifikan terhadap pembanding pada taraf 0,05.

4. KESIMPULAN

Penelitian ini membandingkan Random Forest, XGBoost, dan CatBoost untuk memprediksi risiko performa akademik siswa dengan SMOTENC dan explainable AI. CatBoost dengan SMOTENC memperoleh performa terbaik secara numerik dengan accuracy 94,83% dan F1 macro 93,06% pada holdout 80:20, serta F1 macro rata-rata 92,47% pada stratified 5-fold cross-validation. Friedman Test menunjukkan perbedaan performa secara umum antar kombinasi model, sedangkan Wilcoxon belum menunjukkan perbedaan signifikan terhadap masing-masing pembanding pada taraf 0,05. Feature importance dan SHAP menunjukkan bahwa Attendance, Hours_Studied, Tutoring_Sessions, Access_to_Resources, dan Previous_Scores merupakan faktor dominan. Analisis arah SHAP menunjukkan bahwa nilai Attendance, Hours_Studied, Tutoring_Sessions, dan Previous_Scores yang lebih tinggi cenderung menurunkan prediksi Berisiko, sedangkan Access_to_Resources rendah, Distance_from_Home jauh, Parental_Involvement rendah, dan Peer_Influence negatif cenderung meningkatkan prediksi Berisiko. Dengan demikian, kombinasi ensemble learning, SMOTENC, validasi silang, uji signifikansi, dan XAI dapat mendukung sistem peringatan dini akademik yang akurat, stabil, dan dapat dijelaskan.

REFERENCES

- [1] S. Boujmiraz, H. Darhmaoui, dan A. D. el maliani, "Predicting Student Performance: A Comprehensive Review of Machine Learning, Deep Learning, and Explainable AI Approaches," *Comput. Educ. Artif. Intell.*, vol. 10, hal. 100548, 2026, doi: 10.1016/j.caeai.2026.100548.
- [2] N. Al-Shanableh *et al.*, "Forecasting Students' Academic Performance in Educational Data Using Machine Learning Techniques," *Int. J. Inf. Commun. Technol. Educ.*, vol. 22, no. 1, hal. 1–23, 2026, doi: 10.4018/IJICTE.399756.
- [3] H. Zhang, Y. Ren, P. S. Nurius, J. Mankoff, dan A. K. Dey, "Towards Human-Centered Early Prediction Models for Academic Performance in Real-World Contexts," *Proc. ACM Human-Computer Interact.*, vol. 9, no. 7, hal. 1–41, 2025.
- [4] B. B. Alkan, S. Kuzucuk, N. Alkan, dan A. Sinan, "Using Machine Learning to Predict Student Outcomes for Early Intervention and Formative Assessment," *Sci. Rep.*, vol. 15, hal. 39797, 2025, doi: 10.1038/s41598-025-23409-w.
- [5] D. Khairy, N. Alharbi, M. A. Amasha, M. F. Areed, S. Alkhalaf, dan R. A. Abougalala, "Prediction of Student Exam Performance Using Data Mining Classification Algorithms," *Educ. Inf. Technol.*, vol. 29, no. 16, hal. 21621–21645, 2024, doi: 10.1007/s10639-024-12619-w.
- [6] A. G. R. Sandeepa dan S. Mohottala, "Evaluation of Machine Learning Models in Student Academic Performance Prediction," in *2025 5th International Conference on Advanced Research in Computing (ICARC)*, 2025, hal. 1–6. doi: 10.1109/ICARC64760.2025.10963104.
- [7] H. Gu dan Y. Zhang, "Efficient and Interpretable Machine Learning for Student Academic Outcome Prediction," *Mathematics*, vol. 14, no. 4, hal. 626, 2026, doi: 10.3390/math14040626.
- [8] C. Wang, J. Chen, Z. Xie, dan J. Zou, "Research on education big data for student's academic performance analysis based on machine learning," in *Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Education Digitalization and Computer Science*, 2024, hal. 223–227.
- [9] E. Miranda, M. Aryuni, M. I. Rahmawati, S. E. Hiererra, dan A. V. D. Sano, "Machine learning's model-agnostic interpretability on the prediction of students' academic performance in video-conference-assisted online learning during the COVID-19 pandemic," *Comput. Educ. Artif. Intell.*, vol. 7, hal. 100312, 2024, doi: 10.1016/j.caeai.2024.100312.
- [10] S. John dan A. Mohamed, "Hybrid-PSP: Hybrid Machine Learning Techniques for Predicting Student Performance," *Procedia Comput. Sci.*, vol. 282, hal. 2844–2853, 2026, doi: 10.1016/j.procs.2026.06.052.
- [11] S. Leelaluk, C. Tang, V. Švábenský, dan A. Shimada, "Knowledge distillation in rnn-attention models for early prediction of student performance," in *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, 2025, hal. 64–73.
- [12] M. Nagy dan R. Molontay, "Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention," *Int. J. Artif. Intell. Educ.*, vol. 34, hal. 274–300, 2024, doi: 10.1007/s40593-023-00331-8.
- [13] Y. Chachoui, N. Azizi, R. Hotte, dan T. Bensebaa, "Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning," *Comput. Educ. Artif. Intell.*, vol. 6, hal. 100222, 2024, doi: 10.1016/j.caeai.2024.100222.
- [14] F. M. Alnassar, T. Blackwell, E. Homayounvala, dan M. Yee-king, "Addressing Class Imbalance in Predicting Student Performance Using SMOTE and GAN Techniques," *Appl. Sci.*, vol. 16, no. 7, hal. 3274, 2026, doi: 10.3390/app16073274.
- [15] W. Hidayatulloh, F. Mahardika, dan D. I. Junaedi, "Explainable artificial intelligence-based model for student

- academic performance prediction,” *J. Inf. Syst. Explor. Res.*, vol. 4, no. 1, hal. 31–40, 2026, doi: 10.52465/joiser.v4i1.624.
- [16] M. M. Islam, F. H. Sojib, M. F. H. Mihad, M. Hasan, dan M. Rahman, “The Integration of Explainable AI in Educational Data Mining for Student Academic Performance Prediction and Support System,” *Telemat. Informatics Reports*, vol. 18, hal. 100203, 2025, doi: 10.1016/j.teler.2025.100203.
- [17] M. Abdalkareem dan N. Min-Allah, “Explainable Models for Predicting Academic Pathways for High School Students in Saudi Arabia,” *IEEE Access*, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3369586.
- [18] W. Ahmed, M. A. Wani, P. Plawiak, S. Meshoul, A. Mahmoud, dan M. Hammad, “Machine Learning-Based Academic Performance Prediction with Explainability for Enhanced Decision-Making in Educational Institutions,” *Sci. Rep.*, vol. 15, hal. 26879, 2025, doi: 10.1038/s41598-025-12353-4.
- [19] K. Hubbard dan S. Amponsah, “Feature Engineering on LMS Data to Optimize Student Performance Prediction,” in *New Frontiers in Data Science*, H. Han dan J. Stamey, Ed., Cham: Springer Nature Switzerland, 2026, hal. 125–142.
- [20] Y. Zhao, H. Wang, dan P. Wang, “Improvement of Applicability in Student Performance Prediction Based on Transfer Learning,” in *2024 5th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*, 2024, hal. 308–314. doi: 10.1109/AIEA62095.2024.10692712.
- [21] T. Chen dan C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, hal. 785–794. doi: 10.1145/2939672.2939785.
- [22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, dan A. Gulin, “CatBoost: unbiased boosting with categorical features,” *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.
- [23] M. Mukherjee dan M. Khushi, “SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features,” *Appl. Syst. Innov.*, vol. 4, no. 1, hal. 18, 2021.
- [24] L. Nguyen, “Student Performance Factors,” <https://www.kaggle.com/datasets/lainguy123/student-performance-factors>. Diakses: 20 Juni 2026. [Daring]. Tersedia pada: <https://www.kaggle.com/datasets/lainguy123/student-performance-factors>
- [25] R. A. Saputri, Asrianda, dan L. Rosnita, “A Random Forest-Based Predictive Model for Student Academic Performance: A Case Study in Indonesian Public High Schools,” *J. Appl. Informatics Comput.*, vol. 9, no. 3, hal. 1042–1049, 2025.
- [26] B. Ujkani, D. Minkovska, dan N. Hinov, “Course Success Prediction and Early Identification of At-Risk Students Using Explainable Artificial Intelligence,” *Electronics*, vol. 13, no. 21, hal. 4157, 2024, doi: 10.3390/electronics13214157.
- [27] M. N. Gul, W. Abbasi, M. Z. Babar, A. Aljohani, dan M. Arif, “Data driven decisions in education using a comprehensive machine learning framework for student performance prediction,” *Discov. Comput.*, vol. 28, no. 1, hal. 153, 2025, doi: 10.1007/s10791-025-09585-3.
- [28] N. V Chawla, K. W. Bowyer, L. O. Hall, dan W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, hal. 321–357, 2002, doi: 10.1613/jair.953.
- [29] D. Elreedy, A. F. Atiya, dan F. Kamalov, “A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning,” *Mach. Learn.*, vol. 113, no. 7, hal. 4903–4923, 2024, doi: 10.1007/s10994-022-06296-4.
- [30] G. S. Hida dan A. C. A. Do Nascimento, “Overview of machine learning in class imbalance scenarios: Trends, challenges, and approaches,” *Expert Syst. Appl.*, vol. 298, hal. 129592, 2026, doi: 10.1016/j.eswa.2025.129592.
- [31] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, hal. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [32] S. M. Lundberg dan S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, 2017, hal. 4765–4774. [Daring]. Tersedia pada: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- [33] E. Tiukhova *et al.*, “Explainable Learning Analytics: Assessing the stability of student success prediction models by means of explainable AI,” *Decis. Support Syst.*, vol. 182, hal. 114229, 2024, doi: 10.1016/j.dss.2024.114229.
- [34] F. T. Johora, M. N. Hasan, A. Rajbongshi, M. Ashrafuzzaman, dan F. Akter, “An explainable AI-based approach for predicting undergraduate students academic performance,” *Array*, vol. 26, hal. 100384, 2025.
- [35] M. Delprato, “Identifying the post-pandemic determinants of low performing students in Latin America through interpretable Machine Learning methods,” *Eng. Appl. Artif. Intell.*, vol. 179, hal. 115154, 2026.