Buletin Sistem Informasi dan Teknologi Islam
Vol 6, No 4 (2025), pp 205-213
ISSN: 2721-0901
DOI: https://dx.doi.org/10.33096/busiti.v6i4.3205

Penerapan Metode *Machine Learning* (*Random Forest & XGBoost*) untuk Memprediksi Prestasi Akademik Berdasarkan Faktor Internal dan Eksternal Siswa

Application of Machine Learning Methods (Random Forest & XGBoost) to Predict Academic Achievement Based on Students' Internal and External Factors

Risna Sari[a,1,*], Wisnu Kurniadi[a,2], dan Muhammad Salimy Ahsan[b,3]

[a]*Informatika, Universitas Cokroaminoto Palopo, Palopo, Indonesia*
[b]*Teknik Informatika, Universitas PGRI Madiun, Jawa Timur, Indonesia*
[1]*sarisnasari@gmail.com;* [2]*wisnukurniadi@uncp.ac.id;* [3]*salimy@unipma.ac.id;*

Informasi Artikel

Diserahkan : 21 November 2025
Diterima : 14 Desember 2025
Direvisi : 20 Desember 2025
Diterbitkan : 20 Desember 2025

Kata Kunci:
Machine learning
Random forest
XGBoost
Prestasi akademik

ABSTRAK

Penelitian ini bertujuan untuk menerapkan metode *machine learning*, yaitu *Random Forest* dan *XGBoost*, dalam memprediksi prestasi akademik siswa berdasarkan kombinasi faktor internal dan eksternal yang memengaruhi prestasi mereka. Data yang digunakan yakni *Performance Trends in Education* yang mencakup 6606 baris data dengan 20 kolom, terdiri dari fitur numerik dan kategorik. Dalam penelitian ini, model prediksi dilatih dengan data yang telah diproses melalui tahapan preprocessing dan pembagian data terdiri atas 80% untuk data training dan 20% untuk data testing. Hasil evaluasi menunjukkan bahwa XGBoost memiliki kinerja yang lebih baik dibandingkan dengan Random Forest, dengan nilai RMSE yang lebih rendah yaitu 1.909, MAPE yang lebih kecil yaitu 1.003%, dan R2 yang lebih tinggi yaitu 0.742. Hasil ini menunjukkan bahwa XGBoost mampu memberikan prediksi yang lebih akurat dan lebih baik dalam menjelaskan variabilitas data prestasi akademik siswa. Temuan ini memberikan kontribusi penting dalam pengembangan model prediksi untuk meningkatkan kebijakan pendidikan dan strategi pembelajaran yang lebih efektif. Penelitian ini juga menyarankan penelitian lebih lanjut untuk mengeksplorasi teknik ensemble lain yang dapat meningkatkan akurasi prediksi dan mempertimbangkan faktor eksternal lainnya yang dapat memengaruhi prestasi akademik siswa.

***Keywords*:**
Machine learning
Random forest
XGBoost
Academic achievement

ABSTRACT

This study aims to apply machine learning methods, namely Random Forest and XGBoost, to predict student academic performance based on a combination of internal and external factors influencing their achievement. The data used is from the "Performance Trends in Education" dataset, which includes 6606 rows of data with 20 columns, consisting of numerical and categorical features. In this study, the prediction models were trained using data that had been processed through preprocessing stages and data splitting 80% for training data and 20% for testing data. The evaluation results show that XGBoost performs better than Random Forest, with a lower RMSE value 1.909, a smaller MAPE 1.003%, and a higher R2 0.742. These results indicate that XGBoost provides more accurate predictions and is better at explaining the variability in student academic performance data. This finding contributes significantly to the development of predictive models to improve educational policies and more effective learning strategies. The study also suggests further research to explore other ensemble techniques that could enhance prediction accuracy and consider additional external factors that may influence student academic performance.

I. Pendahuluan

Prestasi akademik siswa merupakan salah satu indikator utama keberhasilan proses pendidikan karena menggambarkan tingkat penguasaan kompetensi, kesiapan melanjutkan studi, maupun daya saing di dunia kerja. Di tengah tuntutan era digital dan *society* 5.0, sekolah dan pemangku kepentingan pendidikan tidak lagi cukup hanya mengukur prestasi secara *post factum* melalui nilai akhir, tetapi juga perlu mengantisipasi risiko

rendahnya prestasi sedini mungkin agar dapat merancang intervensi yang tepat. Kondisi ini mendorong perlunya pendekatan prediktif yang mampu memanfaatkan data beragam dan kompleks untuk memetakan potensi keberhasilan atau kegagalan belajar siswa.

Prestasi akademik sendiri dipengaruhi oleh kombinasi faktor internal dan faktor eksternal siswa. Faktor internal meliputi motivasi belajar, minat, kemampuan kognitif, manajemen waktu, kesehatan fisik dan mental, serta kebiasaan belajar. Sementara faktor eksternal mencakup dukungan keluarga, status sosial ekonomi, iklim kelas, metode mengajar guru, fasilitas sekolah, dan lingkungan sosial di luar sekolah. Sejumlah penelitian mutakhir menegaskan bahwa capaian belajar yang rendah sering kali merupakan hasil interaksi keduanya: misalnya, motivasi tinggi tanpa dukungan lingkungan belajar yang kondusif tetap berpotensi menghasilkan prestasi suboptimal, demikian pula sebaliknya. Penelitian di sekolah menengah menunjukkan bahwa waktu belajar, motivasi, dan keaktifan siswa sebagai faktor internal, serta metode mengajar dan kualitas lingkungan belajar sebagai faktor eksternal, berkontribusi signifikan terhadap hasil belajar matematika dan prestasi akademik secara umum [1].

Dalam beberapa tahun terakhir, ketersediaan data pendidikan menjadi semakin melimpah melalui *Learning Management System* (LMS), sistem informasi akademik, dan berbagai aplikasi pendukung pembelajaran. Perkembangan ini mendorong lahirnya bidang *Educational Data Mining* (EDM) dan *Learning Analytics*, yang memanfaatkan teknik analisis data dan *machine learning* untuk mengekstraksi pola dari data siswa, memprediksi performa, mengidentifikasi siswa berisiko, dan merumuskan rekomendasi kebijakan pendidikan berbasis bukti. Berbagai kajian menunjukkan bahwa model *machine learning* mampu mengatasi hubungan *non-linier*, interaksi antarvariabel, serta kombinasi data demografis, perilaku, dan psikologis yang sulit diakomodasi oleh metode statistik tradisional [2].

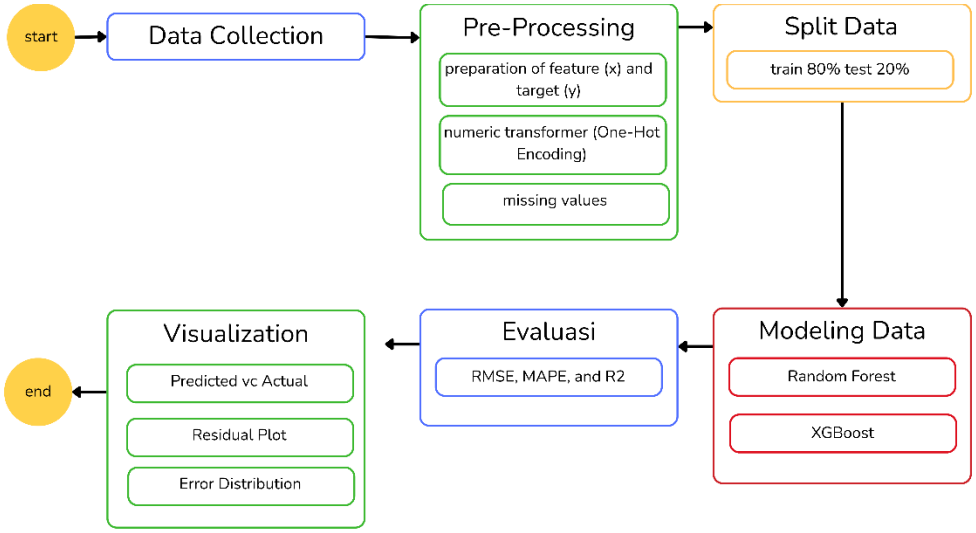
Sejumlah penelitian terdahulu telah mengaplikasikan algoritma *machine learning* seperti *Random Forest* dan *XGBoost* untuk memprediksi prestasi akademik. Studi komprehensif di tingkat pendidikan tinggi menemukan bahwa model *ensemble* (*Random Forest*, *Gradient Boosting*, *XGBoost*) umumnya memberikan performa prediksi lebih baik dibandingkan model linier klasik, dengan akurasi tinggi dalam memprediksi IPK atau kategori keberhasilan studi [3]. Penelitian lain di jenjang sekolah dasar dan menengah menunjukkan bahwa XGBoost mampu memprediksi hasil belajar sains dengan akurasi di atas 95%, sekaligus mengungkap variabel penting seperti kehadiran, nilai tugas, dan dukungan orang tua [4]. Di konteks Indonesia, beberapa studi menggunakan Random Forest untuk memprediksi kinerja akademik berdasarkan data akademik, demografis, serta sifat kepribadian, dan berhasil mengidentifikasi faktor dominan yang memengaruhi prestasi siswa [5]. Penelitian terbaru bahkan secara khusus membandingkan kinerja Random Forest dan XGBoost dalam memprediksi prestasi akademik mahasiswa, dan menemukan bahwa kedua algoritma sama-sama menjanjikan dengan variasi keunggulan pada aspek akurasi dan kemampuan generalisasi [6].

Meskipun demikian, terdapat *research gap* yang masih relevan untuk ditangani. Pertama, banyak penelitian masih berfokus pada satu kelompok faktor saja (misalnya hanya faktor akademik atau hanya faktor demografis) padahal dalam praktiknya prestasi akademik merupakan hasil sinergi faktor internal dan eksternal yang kompleks. Kedua, masih relatif sedikit studi yang secara eksplisit menggabungkan variabel psikologis (motivasi, kebiasaan belajar) dan variabel lingkungan (dukungan keluarga, fasilitas sekolah, pola interaksi dengan guru) dalam satu model prediktif yang utuh. Ketiga, studi yang secara khusus membandingkan kinerja *Random Forest* dan *XGBoost* dengan kerangka variabel internal–eksternal dalam konteks siswa di Indonesia masih terbatas, sehingga kontribusi empiris pada konteks lokal belum optimal.

Bertolak dari uraian tersebut, diperlukan suatu penelitian yang secara khusus menerapkan dan membandingkan metode *machine learning Random Forest* dan *XGBoost* untuk memprediksi prestasi akademik siswa dengan memasukkan variabel-variabel yang merepresentasikan faktor internal dan eksternal secara bersamaan. Penelitian ini diharapkan tidak hanya menghasilkan model prediksi dengan kinerja yang baik, tetapi juga memberikan gambaran faktor-faktor apa saja yang paling berpengaruh terhadap prestasi, sehingga dapat menjadi dasar penyusunan program pendampingan belajar, kebijakan sekolah, dan strategi komunikasi dengan orang tua. Dengan demikian, studi berjudul "Penerapan *Metode Machine Learning* (*Random Forest & XGBoost*) untuk Memprediksi Prestasi Akademik Berdasarkan Faktor Internal dan Eksternal Siswa" diharapkan memberikan kontribusi teoritis pada literatur EDM sekaligus kontribusi praktis bagi peningkatan mutu pendidikan di Indonesia.

II. Metode

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksplanatori, bertujuan untuk memprediksi prestasi akademik siswa berdasarkan kombinasi faktor internal dan faktor eksternal menggunakan a*lgoritma machine learning* yaitu *Random Forest* dan *XGBoost*. Proses penelitian mengikuti tahapan yang digambarkan dalam flowchart alur penelitian Gambar 1 yang mencakup beberapa langkah kritis, mulai dari pengumpulan data hingga evaluasi hasil model. Setiap tahap memiliki tujuan spesifik untuk memastikan model yang dihasilkan mampu memprediksi prestasi siswa secara akurat dan memberikan wawasan yang berguna untuk kebijakan pendidikan.



Gambar 1. Alur Penelitan

A. *Data Collection*

Tahap pertama dalam penelitian ini adalah pengumpulan data. Data yang digunakan dalam penelitian ini diperoleh dari dataset Kaggle yang berjudul Performance Trends in Education, yang dapat diakses di link https://www.kaggle.com/datasets/minahilfatima12328/performance-trends-in-education/data. Dataset ini berisi informasi terkait prestasi akademik siswa serta berbagai faktor internal dan eksternal yang memengaruhi prestasi akademik mereka.

	Hours_Studied	Attendance	Parental_Involvement	Access_to_Resources	Extracurricular_Activities	Sleep_Hours	Previous_Sco
0	23	84	Low	High	No	7	73
1	19	64	Low	Medium	No	8	59
2	24	98	Medium	Medium	Yes	7	91
3	29	89	Low	Medium	Yes	8	98
4	19	92	Medium	Medium	Yes	6	65
...	...	...	...	...	...	...	...
6602	25	69	High	Medium	No	7	76
6603	23	76	High	Medium	No	8	81
6604	20	90	Medium	Low	Yes	6	65
6605	10	86	High	High	Yes	6	91
6606	15	67	Medium	Low	Yes	9	94

Gambar 2. Sampel Dataset

Gambar 2 merupakan dataset yang digunakan terdiri dari 20 kolom dan 6606 baris data yang terdiri dari factor internal dan ekternal akademik siswa. Dalam dataset ini terdapat dua fitur yakni fitur numerik [*'Hours_Studied', 'Attendance', 'Sleep_Hours', 'Previous_Scores', 'Tutoring_Sessions', 'Physical_Activity'*] dan fitur kategorik [*'Parental_Involvement', 'Access_to_Resources', 'Extracurricular_Activities', 'Motivation_Level', 'Internet_Access', 'Family_Income', 'Teacher_Quality', 'School_Type', 'Peer_Influence', 'Learning_Disabilities', 'Parental_Education_Level', 'Distance_from_Home', 'Gender'*].

B. *Preprocessing*

Data yang terkumpul disiapkan agar siap digunakan untuk pemodelan. Pada tahap ini dilakukan pemisahan fitur X (variabel prediktor yang merepresentasikan faktor internal dan eksternal) dan y (variabel target), dengan target yang digunakan adalah Exam_Score. Selanjutnya, fitur pada X dikelompokkan menjadi dua tipe, yaitu fitur numerik (misalnya *Hours_Studied, Attendance, Sleep_Hours, Previous_Scores, Tutoring_Sessions,* dan *Physical_Activity*) dan fitur kategorik (misalnya *Parental_Involvement, Access_to_Resources, Extracurricular_Activities, Motivation_Level, Internet_Access, Family_Income, Teacher_Quality, School_Type, Peer_Influence, Learning_Disabilities, Parental_Education_Level, Distance_from_Home,* dan *Gender*).

Fitur kategorik ditangani menggunakan One-Hot Encoding, yaitu mengubah setiap kategori menjadi vektor biner (0/1). Pendekatan ini dipilih karena Random Forest dan XGBoost merupakan model berbasis pohon keputusan yang dapat bekerja efektif dengan representasi fitur biner tanpa mengasumsikan adanya hubungan ordinal antar-kategori. Selain itu, One-Hot Encoding membantu menghindari bias urutan (ordinal semu) yang

berpotensi muncul apabila menggunakan *Label Encoding* pada data kategorik nominal. Proses encoding dilakukan setelah pembagian data training–testing dalam sebuah *pipeline* untuk mencegah *data leakage*, serta menggunakan pengaturan *handle_unknown = ignore* agar kategori yang mungkin hanya muncul pada data uji tidak menyebabkan kegagalan transformasi.

Sebelum pemodelan, dilakukan pemeriksaan nilai hilang (*missing values*) pada seluruh variabel. Hasil pemeriksaan menunjukkan tidak terdapat missing values pada dataset yang digunakan, sehingga tidak diperlukan proses imputasi. Meskipun demikian, prosedur pemeriksaan nilai hilang tetap dilakukan sebagai bagian dari kontrol kualitas data dan untuk memastikan reprodusibilitas apabila dataset diperluas atau diperbarui pada penelitian selanjutnya.

C. *Split Data*

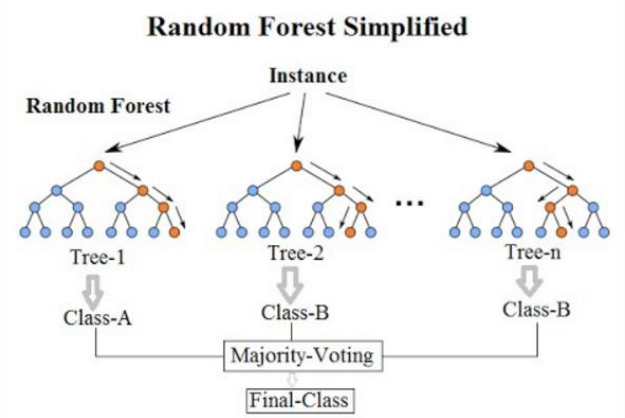
Split Data atau Pembagian Data merupakan tahapan penting dalam pemodelan. Pembagian data menjadi *data training* dan *data testing* dilakukan untuk digunakan nanti dalam membangun dan mengevaluasi pemodelan machine learning yang dibangun. Jumlah *data training* biasanya memiliki presentase yang lebih ebsar dibandingkan dengan *data testing* [7]. Untuk penelitian ini dilakukan pembagian data untuk pemodelan menggunakan algoritma Random Forest dan XGBoost yaitu pembagian 80% untuk *data training* dan 20% untuk *data testing*.

D. *Modeling Data*

Pada tahap pemodelan data, penelitian ini menggunakan dua algoritma machine learning yang populer, yaitu Random Forest dan XGBoost, untuk memprediksi prestasi akademik siswa berdasarkan faktor internal dan eksternal yang telah dipersiapkan sebelumnya.

1) *Random Forest*

Random Forest adalah meta estimator yang menyesuaikan sejumlah pengklasifikasi pohon keputusan pada berbagai sub-sampel dataset dan menggunakan rata-rata untuk meningkatkan akurasi prediktif dan mengendalikan *over-fitting*. Pohon-pohon dalam hutan seperti pada Gambar 3 menggunakan strategi pemisahan terbaik, yaitu setara dengan meneruskan *splitter="best"* ke variabel yang mendasarinya *DecisionTreeClassifier*. Ukuran sub-sampel dikontrol dengan *max_samplesparameter if bootstrap=True(default)*, jika tidak, seluruh dataset digunakan untuk membangun setiap pohon [8]. Random forest melakukan klasifikasi dengan kumpulan pohon keputusan yang hasil akhirnya diperoleh dengan melakukan vote pada *data training* yang dilatih dan melakukan testing secara acak dengan fitur yang berbeda-beda [9].



Gambar 3. Ilustrasi *Random Forest*

Secara matematis, Random Forest untuk melakukan sebuah regresi dapat dinyatan dengan bentuk persamaan sebagai berikut;

$$y = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \quad (1)$$

di mana:

- T adalah jumlah pohon keputusan,
- $f_t(x)$ adalah hasil prediksi pohon keputusan ke-t,
- y adalah hasil prediksi final [10]

2) *XGBoost*

XGBoost (*Extreme Gradient Boosting*) adalah algoritma ensemble yang menggunakan teknik boosting untuk meningkatkan akurasi model secara bertahap. Dalam boosting, pohon keputusan dibangun secara berurutan, dengan setiap pohon baru mencoba memperbaiki kesalahan yang dilakukan oleh pohon sebelumnya. *XGBoost* memiliki keunggulan dalam hal *regularization*, yang membatasi kompleksitas pohon dan membantu menghindari *overfitting* [11].

Fungsi objektif XGBoost untuk regresi dapat dituliskan sebagai berikut:

$$\mathcal{L} = \sum_{i=1}^{N} l\left(y^{(i)}, y'^{(i)}\right) + \sum_{t=1}^{T} \Omega(f_t) \quad (2)$$

di mana:

- $l\left(y^{(i)}, y'^{(i)}\right)$ adalah fungsi loss untuk prediksi $y^{(i)}$ dan $y'^{(i)}$,
- $\Omega(f_t)$ adalah komponen regularisasi pohon keputusan ke-t,
- T adalah jumlah pohon Keputusan,
- $\mathcal{L}$ adalah fungsi objektif yag akan dioptimalkan

E. Evaluasi

Untuk evaluasi model prediksi yang dibuat menggunakan MSE dan RMSE dimana kedua metode evaluasi ini merupakan metode yang mengukur error dari pada model prediksi. RMSE adalah akar kuadrat dari residu varian yang menunjukkan seberapa dekat titik data yang diamati dengan nilai yang dievaluasi yang dihasilkan oleh metode prediksi [12]. RMSE sering digunakan untuk masalah regresi. yi dalam persamaan (3) menunjukkan nilai prediksi, y menunjukkan nilai aktual, dan n adalah jumlah data.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}i - yi)^2}{n}} \quad (3)$$

Mean Absolute Percentage Eror atau biasa di singkat MAPE merupakan perhitungan nilai rata – rata selisih mutlak antara hasil prediksi dengan nilai aktual yang dinyatakan sebagai persentase dari nilai realisasi [13]. Persamaan yang digunakan untuk menghitung nilai MAPE dalam sebuah peramalan yaitu:

$$MAPE = \frac{\sum_{t=1}^{n} |(\frac{Y - Y'}{Y})| 100}{n} \quad (4)$$

Y pada persamaan (4) mewakili nilai aktual Y' mewakili nilai prediksi dan n mewakili jumlah keseluruhan data. Nilai akhir perhitungan MAPE dpaat ditentukan berdasarkan keterangan Tabel 1.

Tabel 1. Range Perhitungan MAPE

Range MAPE	Keterangan
<10%	Model Peramalan Sangat Baik
10-20%	Model Peramalan Baik
20-50%	Model Peramalan Cukup
>50%	Model Peramalan Buruk

Koefisiensi Determinasi juga sering disebut R-Square yang dilambangkan dengan R2 merupakan salah satu kriteria untuk menentukan apakah sampel yang digunakan untuk membangun fungsi regresi dugaan telah tepat memenuhi syarat. Koefisiensi Determinasi memberikan informasi terkait proporsi keragaman ataupun variasi total di sekitar nilai tengah Y yang dapat dijelaskan oleh model regresi yang digunakan. Ukuran ini sering dinyatakan dalam persentase dengan mengalikannya dengan 100. Persamaan yang dapat digunakan untuk menghitung koefisiensi determinasi ini dijelaskan pada persamaan (5)

$$R^2 = 1 - \frac{RSS}{TSS} \quad (5)$$

Keterangan untuk persamaan (5) yaitu RSS adalah jumlah kuadrat residual atau jumlah kuadrat yang dijelaskan oleh regresi. TSS adalah jumlah kuadrat total. Kisaran nilai R2 di antara 0 hingga 1 ($0 \leq r2 \leq 1$) jika dikalikan dengan 100 maka kisaran nilai R2 0% hingga 100%. Sehingga semakin besar nilai R2 maka semakin besar kemampuan model regresi yang digunakan menjelaskan keragaman data sampel [13].

F. Visualisasi

Tahap visualisasi di penelitian ini menggunakan tiga teknik utama untuk mengevaluasi hasil prediksi model. *Predicted vs Actual Plot* memperlihatkan perbandingan antara nilai prediksi dan nilai aktual untuk mengevaluasi seberapa baik model memprediksi prestasi akademik. *Residual Plot* menggambarkan perbedaan antara nilai yang diprediksi dan nilai aktual (residuals), yang digunakan untuk mengidentifikasi pola kesalahan dalam model. Terakhir, *Error Distribution* menunjukkan distribusi kesalahan di seluruh data, membantu untuk memeriksa apakah kesalahan terdistribusi secara acak atau terdapat bias dalam prediksi.

III. Hasil dan Pembahasan

Penelitian ini secara khusus menerapkan dan membandingkan metode machine learning, yaitu Random Forest dan XGBoost, untuk memprediksi prestasi akademik siswa dengan memasukkan variabel-variabel yang merepresentasikan faktor internal dan eksternal secara bersamaan. Mengacu pada alur penelitian, tahap pertama dalam penelitian ini adalah Pengumpulan Data (Data Collection), di mana data yang digunakan merupakan data numerik dari Kaggle yang berformat CSV, terdiri dari 20 kolom dan 6606 baris data yang mencakup faktor internal dan eksternal akademik siswa. Data tersebut terdiri dari dua fitur, yaitu fitur numerik dan fitur kategorik. Selanjutnya, tahap Preprocessing mempersiapkan data dengan mengekstrak fitur X dan fitur Y, dengan target yang ditetapkan adalah kolom "Exam_Score" yang akan digunakan untuk pemodelan data. Tahap berikutnya adalah pembagian data (Split Data), di mana data training mencakup 80% dari total 6606 baris data, dan sisanya, yaitu 20%, digunakan untuk data testing.

Pada tahap selanjutnya, yaitu Modeling Data, penelitian ini menggunakan dua algoritma, yakni Random Forest dan XGBoost. Kedua algoritma tersebut diimplementasikan menggunakan parameter baseline yang mengacu pada praktik umum dan rekomendasi dokumentasi algoritma untuk masalah regresi. Parameter ini dipilih untuk menjaga keseimbangan antara akurasi prediksi, risiko overfitting, dan efisiensi komputasi. Tabel 2 menunjukkan parameter yang digunakan pada kedua algoritma tersebut.

Tabel 2. Parameter Random Forest dan XGBoost

Parameter	
Random Forest	XGBoost
n_estimators = 200	n_estimators=300
random_state = 42	learning_rate=0.05
n_jobs = -1	max_depth=5
	subsample=0.8
	colsample_bytree=0.8
	random_state=42
	objective="reg:squarederror"

Parameter pada Tabel 2 digunakan sebagai konfigurasi baseline untuk membandingkan performa Random Forest dan XGBoost secara adil. Penelitian ini tidak menerapkan hyperparameter tuning sistematis (*Grid Search* atau *Random Search*) pemilihan parameter didasarkan pada praktik umum dan uji coba awal (*preliminary run*) pada data latih untuk memperoleh konfigurasi yang stabil serta efisien secara komputasi.

Bagian Random Forest, n_estimators = 200 dipilih untuk menghasilkan prediksi yang lebih stabil (variansi lebih kecil) tanpa meningkatkan waktu komputasi secara berlebihan. random_state = 42 digunakan untuk memastikan hasil eksperimen reprodusibel, sedangkan n_jobs = -1 digunakan agar proses pelatihan memanfaatkan seluruh inti prosesor sehingga lebih efisien.

Bagian XGBoost, learning_rate = 0.05 dipilih sebagai laju pembelajaran yang relatif konservatif agar proses boosting lebih bertahap sehingga membantu mengurangi risiko overfitting. max_depth = 5 digunakan untuk membatasi kompleksitas pohon sehingga model tidak terlalu mengikuti noise data. Selanjutnya, subsample = 0.8 dan colsample_bytree = 0.8 digunakan sebagai teknik regularisasi berbasis sampling untuk menurunkan korelasi antar pohon dan meningkatkan kemampuan generalisasi model. Parameter objective = "reg:squarederror" dipilih karena target yang diprediksi berupa nilai kontinu (regresi), sedangkan random_state = 42 digunakan untuk konsistensi hasil.

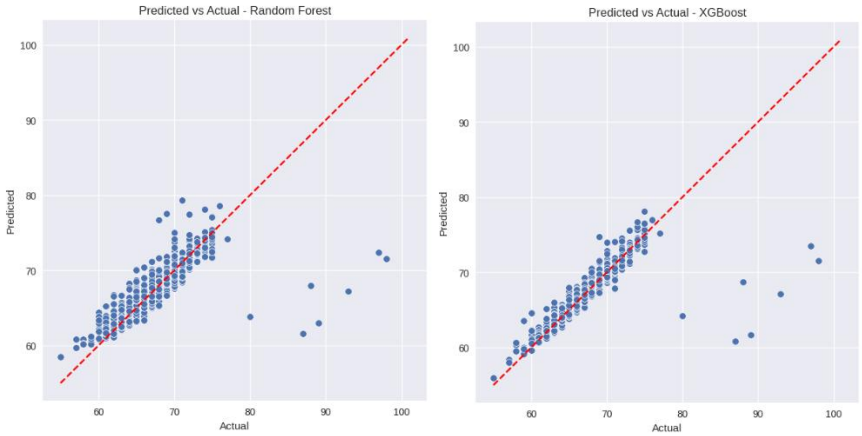
Memasuki tahap Modeling Data, di mana parameter yang telah ditentukan diterapkan, proses dimulai dengan melakukan training model menggunakan algoritma Random Forest terlebih dahulu. Setelah itu, dilakukan training model menggunakan algoritma XGBoost. Kedua model tersebut dilatih dengan menggunakan data yang telah diproses pada tahap sebelumnya, dengan tujuan untuk membandingkan kinerja keduanya dalam memprediksi prestasi akademik siswa. Proses training dilakukan dengan mengoptimalkan parameter baseline yang digunakan pada masing-masing algoritma, serta memastikan kedua model mampu

menangani data dengan baik untuk menghasilkan prediksi yang akurat. Hasil dari modeling dapat kita lihat pada Tabel 3 yang merupakan perbandingan metode evaluasi yang digunakan

Tabel 3. Hasil Evaluasi Algorima Random Forest dan XGBoost

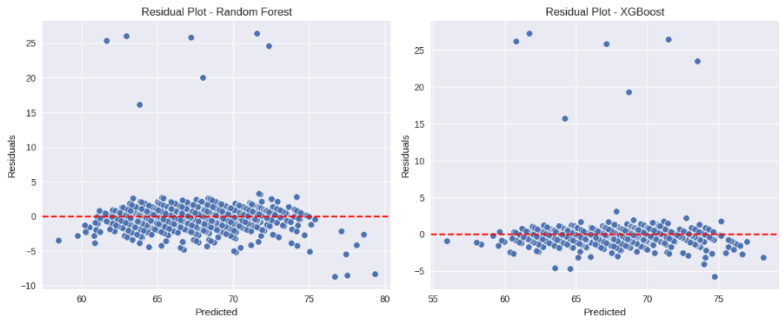
Algoritma	RMSE	MAPE	R2
Random Forest	2.164	1.576%	0.669
XGBoost	1.909	1.003%	0.742

Berdasarkan hasil yang ditampilkan pada Tabel 3, XGBoost memiliki nilai RMSE yang lebih rendah (1.909) dibandingkan dengan Random Forest (2.164), yang menunjukkan bahwa XGBoost memberikan prediksi dengan kesalahan yang lebih kecil. Selain itu, MAPE pada XGBoost juga lebih rendah (1.003%) dibandingkan dengan Random Forest (1.576%), mengindikasikan bahwa XGBoost lebih akurat dalam hal persentase kesalahan. Di sisi lain, nilai R2 XGBoost (0.742) lebih tinggi daripada Random Forest (0.669), yang menunjukkan bahwa XGBoost memiliki kemampuan yang lebih baik dalam menjelaskan variabilitas data. Secara keseluruhan, XGBoost menunjukkan performa yang lebih baik dibandingkan dengan Random Forest dalam hal prediksi prestasi akademik siswa pada dataset ini.



Gambar 4. Visualisasi data Predicted vs Actual Random Forest dan XGBoost

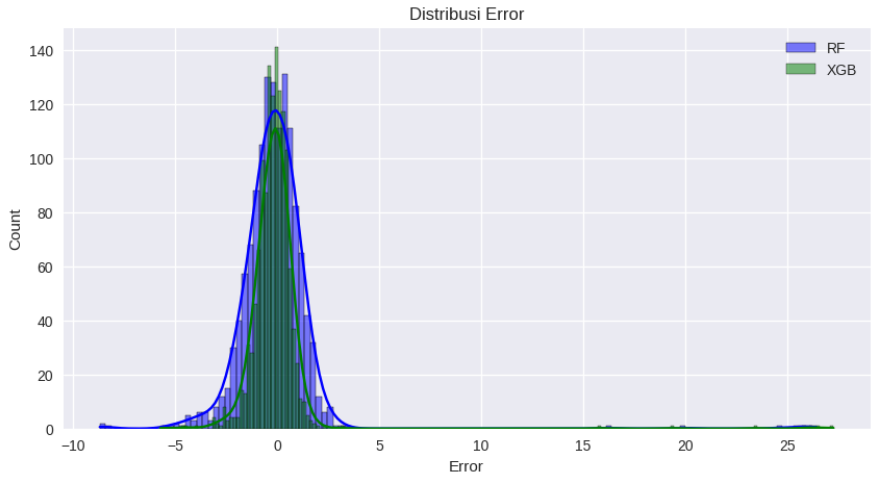
Gambar 4 di atas menunjukkan visualisasi perbandingan antara nilai prediksi dan nilai aktual untuk model Random Forest (kiri) dan XGBoost (kanan). Pada kedua grafik, garis putus-putus merah mewakili garis referensi di mana nilai prediksi dan nilai aktual seharusnya sama. Grafik untuk XGBoost menunjukkan pola yang lebih rapat di sekitar garis referensi, mengindikasikan bahwa model XGBoost memberikan prediksi yang lebih akurat dan lebih mendekati nilai aktual dibandingkan dengan Random Forest. Sebaliknya, grafik untuk Random Forest menunjukkan penyebaran yang lebih lebar, yang mengindikasikan adanya kesalahan prediksi yang lebih besar. Hal ini sesuai dengan hasil evaluasi yang menunjukkan bahwa XGBoost memiliki kinerja yang lebih baik dibandingkan dengan Random Forest dalam memprediksi prestasi akademik siswa.



Gambar 5. Visualisasi data Residual Plot Random Forest dan XGBoost

Grafik yang ditampilkan pada Gambar 5 merupakan grafik residu perbandingan antara model Random Forest (kiri) dan XGBoost (kanan) dengan residuals pada sumbu vertikal dan nilai prediksi pada sumbu horizontal. Pada grafik Random Forest, terlihat bahwa residu tersebar lebih lebar dan tidak sepenuhnya acak di sekitar garis nol, dengan beberapa titik residual yang cukup besar, mengindikasikan adanya bias atau kesalahan prediksi yang lebih signifikan pada model ini. Sebaliknya, pada grafik XGBoost, residu lebih

terpusat di sekitar garis nol, dengan penyebaran yang lebih sempit dan lebih acak, menunjukkan bahwa model XGBoost mampu menghasilkan prediksi yang lebih tepat dan dengan kesalahan yang lebih kecil dibandingkan dengan Random Forest. Jika dibandingkan dengan grafik "Predicted vs Actual" sebelumnya, XGBoost menunjukkan konsistensi yang lebih baik baik dalam hal akurasi prediksi maupun kesalahan residual, sementara Random Forest menunjukkan penyebaran yang lebih besar baik pada nilai prediksi maupun residual.



Gambar 6. Visualisasi Distribusi Error Algoritma Random Forest dan XGBoost

Grafik distribusi error yang ditunjukkan pada Gambar 6 diatas membandingkan distribusi kesalahan (error) antara model Random Forest (RF) dan XGBoost (XGB). Terlihat bahwa error pada model XGBoost (garis hijau) lebih terpusat di sekitar nol dan memiliki distribusi yang lebih sempit dibandingkan dengan model Random Forest (garis biru), yang menunjukkan bahwa kesalahan prediksi pada XGBoost lebih kecil dan lebih konsisten. Sebaliknya, error pada model Random Forest cenderung lebih tersebar dengan beberapa nilai error yang lebih besar. Hal ini mengindikasikan bahwa XGBoost menghasilkan prediksi yang lebih akurat dengan kesalahan yang lebih terkontrol, sementara Random Forest memiliki kesalahan yang lebih besar dan lebih variatif. Grafik ini menegaskan hasil sebelumnya yang menunjukkan kinerja XGBoost yang lebih baik dalam memprediksi prestasi akademik siswa.

IV. Kesimpulan dan saran

Berdasarkan hasil evaluasi dan visualisasi yang diperoleh, dapat disimpulkan bahwa model XGBoost secara keseluruhan menunjukkan kinerja yang lebih baik dibandingkan dengan Random Forest dalam memprediksi prestasi akademik siswa. Evaluasi menunjukkan bahwa XGBoost memiliki nilai RMSE yang lebih rendah (1.909), MAPE yang lebih kecil (1.003%), dan nilai R2 yang lebih tinggi (0.742) dibandingkan dengan Random Forest. Hal ini mengindikasikan bahwa XGBoost mampu menghasilkan prediksi yang lebih akurat dan lebih baik dalam menjelaskan variabilitas data. Temuan ini memiliki implikasi penting dalam pengembangan model prediksi akademik, yang dapat membantu pengambilan keputusan yang lebih baik dalam pendidikan. Untuk penelitian selanjutnya, disarankan untuk mengeksplorasi penggunaan teknik ensemble atau model lain yang dapat meningkatkan akurasi prediksi lebih lanjut, serta mempertimbangkan faktor-faktor eksternal lainnya yang mungkin berpengaruh terhadap hasil akademik siswa.

Daftar Pustaka

[1] A. W. Yanti, M. Z. Irwanto, D. A. Rofikha, A. Kautsar, S. M. H. Septiane, and S. Mas'ullah, "Analisis Faktor-Faktor yang Mempengaruhi Hasil Belajar Siswa : Studi Kasus Hasil Penilaian Tengah Semester Matematika," *J. Mercumatika J. Penelit. Mat. dan Pendidik. Mat.*, vol. 8, no. 1, Apr. 2025, doi: 10.26486/jm.v8i1.3602.

[2] T. Hongli, "Educational data mining for student performance prediction: feature selection and model evaluation," *J. Electr. Syst.*, pp. 1063–1074, 2024, doi: 10.52783/jes.3434.

[3] A. Abatal, A. Korchi, M. Mzili, T. Mzili, H. Khalouki, and M. E. K. Billah, "A Comprehensive Evaluation of Machine Learning Techniques for Forecasting Student Academic Success," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 1, pp. 1–12, Jan. 2025, doi: 10.35882/jeeemi.v7i1.489.

[4] E. Harefa, "Predicting Science Learning Outcomes Of Elementary Students In Rural Area Using Machine Learning Algorithms," *J. Soc. Sci. Res.*, vol. 4, pp. 3780–3792, 2024.

[5] B. D. Setiawan, "Predicting Student Performance Using Machine Learning for Student Management in University (Study Case Maju Jaya Sentosa University)," 2023.

[6] N. Sulistiyaningsih and R. Rismayati, "Comparison of Random Forest, Decision Tree, and XGBoost Models in Predicting Student Academic Success," *J. Artif. Intell. Softw. Eng.*, vol. 5, no. 3, pp. 920–

930, 2025, doi: 10.30811/jaise.v5i3.7138.
[7] D. Haryadi, A. R. Hakim, D. M. U. Atmaja, and S. N. Yutia, "Implementation of Support Vector Regression for Polkadot Cryptocurrency Price Prediction," *Int. J. Informatics Vis.*, vol. 6, no. 1–2, pp. 201–207, 2022, doi: 10.30630/joiv.6.1-2.945.
[8] Scikit-learn, "RandomForestClassifier — scikit-learn 1.7.2 documentation," *scikit-learn*, 2025. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed Nov. 20, 2025).
[9] R. Sari, L. Trihardianingsih, R. Firdaus Mulya, M. I. Arief, and K. Kusrini, "Analisis Index Vegetation Wilayah Terdampak Kebakaran Hutan Riau Menggunakan Citra Landsat-8 dan Sentinel-2," *CogITo Smart J.*, vol. 8, no. 2, pp. 282–294, 2022, doi: 10.31154/cogito.v8i2.439.282-294.
[10] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, "RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.
[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-August-2016, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.
[12] K. Grace and Martin, "Measures of Model Fit for Linear Regression Models," *theanalysisfactor*, 2013. https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/ (accessed Dec. 14, 2022).
[13] R. Sari, "Optimasi Metode Lstm Dan Svm Dalam Meningkatkan Performa Prediksi Harga Cryptocurrency," 2022.