



Prediction of Final School Grades Using the Random Forest Algorithm

Prediksi Nilai Akhir Sekolah Menggunakan Algoritma Random Forest

Najwa Hanindya Putri^{1*}, Fajar Nugraha², Yudie Irawan³

^{1,2,3}Program Studi Sistem Informasi, Fakultas Teknik, Universitas Muria Kudus, Indonesia

E-Mail: ¹202253137@std.umk.ac.id, ²fajar.nugraha@umk.ac.id, ³yudie.irawan@umk.ac.id

Received Jun 05th 2026; Revised Jul 21th 2026; Accepted Aug 13th 2026; Available Online Aug 15th 2026

Corresponding Author: Najwa Hanindya Putri

Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Final school grades serve as a crucial indicator of students' academic achievement and support data-driven academic decision-making. This study employs Random Forest Regression to predict students' final academic grades at SMKN 1 Kudus using historical academic data. The research workflow follows the Cross-Industry Standard Process for Data Mining (CRISP-DM), a framework that guides the study systematically through stages, including understanding school requirements, analyzing data characteristics, data preparation and cleaning, model construction, performance evaluation, and applying the results for early academic monitoring. The dataset comprises 1,530 student records from six specialization programs, split into 80% training data and 20% testing data. The model was constructed using parameters `n_estimators=100`, `max_depth=10`, and `random_state=42`, with the average report card grade serving as the proxy for the final academic grade. Evaluation results based on the 306 test samples yielded a Mean Absolute Error (MAE) of 0.4556 and a Root Mean Square Error (RMSE) of 0.5801. Visualization of actual versus predicted values shows a distribution of points close to the diagonal line, while residual errors are centered around zero. Feature importance analysis identifies Guidance and Counseling (BK), Civics (PPKn), Mathematics, Indonesian Language, and Entrepreneurship (PKK) as the features contributing most significantly to the model. These results demonstrate that Random Forest Regression can generate stable predictions and support early academic monitoring.

Keywords: CRISP-DM, Academic Data, Feature Importance, Grade Prediction, Random Forest Regression

Abstrak

Nilai akhir sekolah merupakan indikator penting untuk menilai capaian akademik siswa dan mendukung pengambilan keputusan berbasis data. Penelitian ini menerapkan Random Forest Regression untuk mengestimasi nilai akhir akademik siswa menggunakan data akademik historis SMKN 1 Kudus. Alur penelitian mengikuti Cross-Industry Standard Process for Data Mining (CRISP-DM), yang meliputi pemahaman kebutuhan, pemahaman data, persiapan data, pemodelan, evaluasi, dan penggunaan konseptual hasil untuk pemantauan akademik dini. Dataset berisi 1.530 data siswa dari enam program keahlian dan dibagi menjadi 80% data latih serta 20% data uji. Model menggunakan `n_estimators=100`, `max_depth=10`, dan `random_state=42`, dengan Rata-Rata_Rapor sebagai proksi capaian akhir akademik. Evaluasi pada 306 data uji menghasilkan Mean Absolute Error (MAE) sebesar 0,4556 dan Root Mean Squared Error (RMSE) sebesar 0,5801. Nilai RMSE yang hanya sedikit lebih tinggi daripada MAE menunjukkan bahwa galat yang relatif besar terbatas; titik aktual-prediksi mendekati garis diagonal dan residual berpusat di sekitar nol. BK, PPKn, Matematika, B. Indonesia, dan PKK memiliki skor kepentingan fitur relatif tertinggi. Kebaruan penelitian terletak pada prediksi skor akademik kontinu, bukan hanya kelas lulus atau tidak lulus, serta penggabungan evaluasi berbasis galat dengan analisis kepentingan fitur untuk mendukung pemantauan dini yang dapat diinterpretasikan. Model diposisikan sebagai komponen penyaringan pada sistem peringatan dini akademik yang tetap memerlukan penelaahan guru, bukan sebagai dasar otomatis untuk keputusan akademik berisiko tinggi.

Kata Kunci: CRISP-DM, Data Akademik, Kepentingan Fitur, Prediksi Nilai, Random Forest Regression

1. PENDAHULUAN

Evaluasi hasil belajar merupakan bagian penting dalam sistem pendidikan karena digunakan untuk menilai ketercapaian kompetensi siswa dan menjadi dasar pengambilan keputusan akademik. Data yang tersedia dalam penelitian ini mencakup 1.530 catatan siswa SMKN 1 Kudus dari enam program keahlian dan 36 fitur masukan, dengan nilai Rata_Rata_Rapor berada pada rentang 76,50-94,00. Besarnya volume data, perbedaan kelengkapan antarfitur, serta kombinasi nilai mata pelajaran umum, mata pelajaran kejuruan, jurusan, dan ketidakhadiran menyebabkan penelaahan pola capaian secara manual menjadi kurang efisien untuk penyaringan awal. Oleh karena itu, model prediktif diperlukan untuk membantu sekolah membaca pola capaian secara lebih cepat dan konsisten, dengan tetap menempatkan guru sebagai pihak yang memverifikasi hasil dan menentukan bentuk pendampingan. Penelitian di Indonesia juga menunjukkan bahwa data nilai, kehadiran, dan riwayat akademik dapat diolah untuk mendeteksi pola performa serta mendukung keputusan akademik berbasis data [1].

Pemanfaatan pembelajaran mesin dalam bidang pendidikan semakin relevan karena data akademik yang tersimpan di sekolah dapat diolah menjadi informasi prediktif. Dalam ranah penambangan data pendidikan, model prediksi telah digunakan untuk mendukung pengambilan keputusan berbasis data, misalnya pada prediksi kelulusan mahasiswa berbasis data kunjungan dan peminjaman buku [2], prediksi kelulusan siswa Madrasah Aliyah dengan Support Vector Machine dan Random Forest [3], prediksi kelulusan siswa SMP menggunakan Decision Tree, Random Forest, dan XGBoost [4], serta prediksi ketepatan waktu kelulusan mahasiswa menggunakan Random Forest [5]. Sumber lain juga menunjukkan bahwa pemodelan prediktif dapat diterapkan pada beragam konteks, seperti prediksi konsumsi harian program makan bergizi berbasis kepuasan siswa, validasi algoritma prediksi tinggi badan berbasis sensor ultrasonik, dan sistem prediksi harga pangan berbasis web menggunakan SARIMA [6]. Perbandingan metode pada studi nasional menunjukkan bahwa Random Forest, Support Vector Machine, Naïve Bayes, Decision Tree, dan model regresi dapat menghasilkan performa berbeda sesuai distribusi data, pemilihan atribut, dan tujuan prediksi [7], [1], [8].

Pada konteks regresi, Kusworo et al. memodelkan nilai akhir semester siswa menggunakan Random Forest [6], sedangkan Sari et al. membandingkan Random Forest dan XGBoost pada data prestasi akademik yang memuat faktor internal dan eksternal [9]. Kajian yang lebih luas menunjukkan bahwa kinerja model dipengaruhi oleh karakteristik dataset, pemilihan fitur, target, dan prosedur evaluasi; tidak terdapat satu algoritma yang selalu unggul pada seluruh konteks pendidikan [10], [11]. Selain itu, penelitian terkini menekankan bahwa nilai prediksi perlu disertai penjelasan faktor yang digunakan model agar hasilnya dapat ditindaklanjuti secara bertanggung jawab dalam sistem pendukung keputusan [11]. Dengan demikian, kajian terdahulu memperlihatkan pergeseran dari sekadar klasifikasi menuju prediksi numerik dan interpretabilitas, tetapi masih terdapat kebutuhan untuk menguji pendekatan tersebut pada data sekolah kejuruan yang menggabungkan mata pelajaran umum dan kejuruan. Kajian komparatif dan literature review Indonesia menegaskan bahwa tidak ada model yang unggul secara universal; evaluasi perlu mempertimbangkan target, prapemrosesan, pembagian data, serta kebutuhan interpretasi [12], [13], [14], [15], [8].

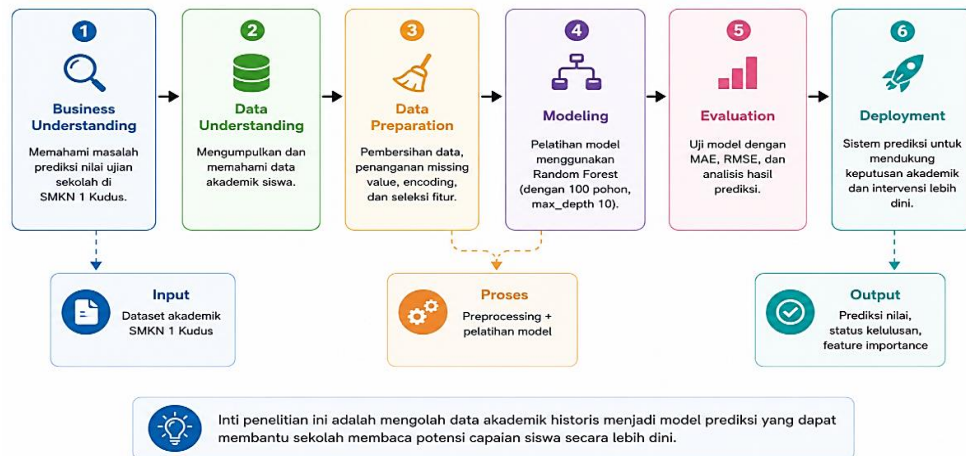
Random Forest Regression dipilih karena membangun banyak pohon pada sampel bootstrap dan subset fitur acak, kemudian merata-ratakan prediksi untuk menurunkan varians model [16]. XGBoost dan Gradient Boosting dapat memberikan kinerja lebih tinggi pada dataset tertentu, tetapi pembentukan pohonnya berlangsung secara berurutan dan umumnya memerlukan penalaan hiperparameter yang lebih intensif; hasil komparatif mutakhir juga menunjukkan bahwa keunggulan model bergantung pada struktur data [9], [11]. Dalam penelitian ini, Random Forest tidak diposisikan sebagai algoritma yang selalu terbaik, melainkan sebagai pilihan yang sesuai dengan ukuran dataset yang moderat, jumlah fitur yang relatif banyak, hubungan antarfitur yang berpotensi nonlinier, kebutuhan stabilitas, dan tujuan menyediakan kepentingan fitur secara langsung. Pemilihan tersebut merupakan pertimbangan kesesuaian metodologis, bukan klaim superioritas tanpa perbandingan eksperimen. Penelitian pada data akademik menunjukkan bahwa Random Forest relevan untuk data berdimensi banyak dan hubungan nonlinier, sedangkan seleksi fitur serta penalaan hiperparameter dapat meningkatkan kinerja model [12], [17], [18].

Berdasarkan kesenjangan tersebut, penelitian ini memprediksi nilai akhir akademik dalam bentuk skor kontinu menggunakan 36 fitur yang mencakup jurusan, ketidakhadiran, mata pelajaran umum, dan mata pelajaran kejuruan. Kebaruan penelitian terletak pada penerapan regresi pada data sekolah kejuruan dan penggabungan MAE, RMSE, visualisasi residual, serta kepentingan fitur untuk menghasilkan prediksi yang tidak hanya akurat, tetapi juga dapat diinterpretasikan sebagai bahan pemantauan awal. Tujuan penelitian adalah membangun model Random Forest Regression, mengevaluasi galat prediksi pada data uji, mengidentifikasi fitur dengan kontribusi relatif terbesar, dan merumuskan implikasi penggunaannya sebagai komponen sistem peringatan dini akademik. Perkembangan penelitian terbaru juga mengarah pada model regresi yang dapat diinterpretasikan dan pendekatan Explainable Artificial Intelligence untuk menghubungkan hasil prediksi dengan intervensi akademik [5], [19].

2. BAHAN DAN METODE

2.1. Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif berbasis penambangan data dengan tahapan *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Tahapan tersebut meliputi business understanding untuk merumuskan kebutuhan pemantauan akademik, data understanding untuk membaca struktur dan kualitas data, data preparation untuk pembersihan serta transformasi, modeling untuk membangun Random Forest Regression, evaluation untuk menilai galat pada data uji, dan deployment yang dalam penelitian ini dibatasi pada rancangan pemanfaatan hasil sebagai bahan pemantauan dini. Dengan batasan tersebut, penelitian tidak mengklaim telah menerapkan sistem produksi secara langsung di sekolah. Penggunaan CRISP-DM pada penelitian prediksi kelulusan dan prestasi akademik membantu menjaga keterlacakan proses dari pemahaman masalah hingga evaluasi model [20], [8], [21]. Alur penelitian dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian Berbasis CRISP-DM

2.2. Dataset dan Persiapan Data

Data yang digunakan dalam penelitian ini merupakan dataset akademik siswa SMKN 1 Kudus yang telah melalui proses pembersihan dan seleksi data. Dataset utama terdiri atas 1.530 baris data siswa yang berasal dari enam program keahlian. Variabel masukan yang digunakan dalam proses pemodelan mencakup informasi jurusan siswa, tingkat ketidakhadiran, nilai mata pelajaran umum, serta nilai mata pelajaran kejuruan. Variabel-variabel tersebut dipilih karena dinilai dapat merepresentasikan kondisi akademik dan keterlibatan siswa selama proses pembelajaran.

Target dalam penelitian ini adalah variabel *Rata_Rata_Rapor*, yang digunakan sebagai proksi untuk menggambarkan capaian akhir akademik siswa. Sebelum dilakukan proses pemodelan, dataset terlebih dahulu melalui tahap prapemrosesan, seperti pemeriksaan data kosong, penyesuaian format data, pengkodean variabel kategorikal, serta normalisasi atau standarisasi apabila diperlukan. Selanjutnya, data dibagi menjadi data latih dan data uji menggunakan rasio 80:20. Dari total 1.530 data bersih, sebanyak 1.224 data digunakan sebagai data latih untuk membangun model, sedangkan 306 data digunakan sebagai data uji untuk mengevaluasi kemampuan model dalam memprediksi rata-rata nilai rapor siswa pada data yang belum pernah digunakan sebelumnya. Tahapan seleksi, pembersihan, transformasi, dan pembagian data perlu didokumentasikan karena perubahan prapemrosesan dapat memengaruhi hasil prediksi [22],[23].

Tabel 1. Ringkasan Dataset Penelitian

Komponen	Nilai	Keterangan
Data bersih	1.53	Catatan akademik setelah pembersihan dan penyusunan satu baris untuk setiap siswa
Data latih	1.224	80% dari dataset setelah prapemrosesan
Data uji	306	20% dari dataset yang digunakan untuk evaluasi model
Fitur masukan	36	Jurusan, ketidakhadiran, mata pelajaran umum, dan mata pelajaran kejuruan
Target	Rata - Rata Rapor	Proksi nilai akhir akademik siswa
Algoritma	Random Forest Regression	Model ansambel berbasis banyak pohon keputusan

Tahapan prapemrosesan dilakukan dengan menghapus baris kosong dan data duplikasi. Nilai yang hilang pada fitur numerik ditangani melalui imputasi, sedangkan variabel nominal seperti Jurusan dikonversi menjadi numerik menggunakan Label Encoding. Proses ini diperlukan agar seluruh fitur dapat diproses oleh

algoritma Random Forest Regression. Pemisahan data dilakukan menggunakan `random_state=42` agar eksperimen dapat direplikasi. Gori et al. menunjukkan bahwa prapemrosesan dan pemilihan fitur dapat meningkatkan kinerja klasifikasi, sedangkan Hidayah et al. memperlihatkan pentingnya penalaan hiperparameter pada Random Forest [23], [18].

2.3. Pemodelan Regresi Random Forest

Regresi Random Forest membangun sejumlah pohon keputusan pada subset data dan subset fitur yang berbeda. Prediksi akhir diperoleh dari rata-rata prediksi seluruh pohon sehingga model lebih stabil dan dapat mengurangi risiko overfitting dibandingkan pohon keputusan tunggal. Pendekatan ansambel berbasis Random Forest banyak digunakan dalam studi prediksi akademik karena mampu menangani fitur numerik dan kategorik serta menyediakan informasi kepentingan fitur untuk membantu interpretasi hasil [9], [24], [16]. Parameter model yang digunakan dalam penelitian ini ditunjukkan pada Tabel 2. Studi pada data akademik di Indonesia menunjukkan bahwa model ansambel dapat menangkap relasi yang kompleks, tetapi tetap perlu dibandingkan dengan model lain dan diuji pada data yang belum pernah dilihat [12], [8], [25]. Secara matematis, prediksi Random Forest Regression untuk data siswa x dirumuskan pada Persamaan (1).

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(x) \quad (1)$$

Pada Persamaan (1), $\hat{y}$ menyatakan nilai akhir yang diprediksi, B menyatakan jumlah pohon keputusan, $T_b(x)$ menyatakan prediksi pohon ke- b untuk data siswa x , dan x menyatakan vektor fitur siswa. Penelitian ini menggunakan $B=100$ dengan parameter `n_estimators=100`. Pada setiap pohon, pemisahan simpul dipilih dengan memaksimalkan penurunan galat kuadrat rata-rata yang dirumuskan pada Persamaan (2).

$$MSE(t) = \frac{1}{N_t} \sum_{i \in t} (y_i - \bar{y}_t)^2 \quad (2)$$

Pada Persamaan (2), $MSE(t)$ adalah galat kuadrat rata-rata pada simpul t , N_t adalah jumlah data dalam simpul, y_i adalah nilai aktual siswa ke- i , dan $\bar{y}_t$ adalah rata-rata target pada simpul tersebut. Pemisahan dengan penurunan MSE terbesar dipilih sebagai pemisahan terbaik.

Tabel 2. Konfigurasi Regresi Random Forest

Parameter	Nilai	Fungsi
<code>n_estimators</code>	100	Jumlah pohon keputusan yang digunakan model
<code>max_depth</code>	10	Membatasi kedalaman pohon untuk mengurangi overfitting
<code>random_state</code>	42	Menjaga reproduibilitas eksperimen
<code>test_ratio</code>	0,2	Proporsi data uji

2.4. Evaluasi Model

Kinerja model dievaluasi menggunakan *Mean Absolute Error (MAE)* dan *Root Mean Square Error (RMSE)*. MAE menunjukkan rata-rata selisih absolut antara nilai aktual dan prediksi, sedangkan RMSE memberi penalti lebih besar terhadap kesalahan prediksi yang besar. Penggunaan metrik galat seperti MAE dan RMSE sesuai untuk model regresi karena target penelitian berupa nilai numerik, sebagaimana digunakan pula pada penelitian prediksi prestasi atau nilai akademik berbasis pembelajaran mesin [10], [16]. Selain itu, penelitian ini menggunakan visualisasi nilai aktual terhadap prediksi, distribusi galat residual, peta panas korelasi, dan kepentingan fitur untuk memperjelas interpretasi hasil model. MAE dan RMSE telah digunakan dalam penelitian regresi akademik untuk mengukur besarnya galat pada skala target; RMSE lebih sensitif terhadap galat besar sehingga kedua metrik sebaiknya dibaca secara bersama [14].

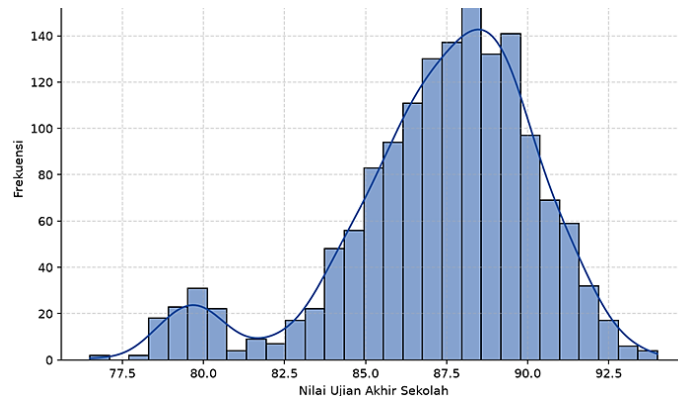
3. HASIL DAN PEMBAHASAN

3.1. Deskripsi Dataset

Nilai Rata_Rata_Rapor berada pada rentang 76,50-94,00 dengan rata-rata 87,2781 dan simpangan baku 3,0116. Rata-rata PPKn sebesar 88,9527, B. Indonesia 87,0791, Matematika 84,9190, B. Inggris 84,5556, dan PKK 87,8845. Tabel 3 menampilkan statistik sebelum imputasi; karena itu, jumlah observasi PPKn dan PKK lebih kecil daripada 1.530. Konsentrasi target pada rentang nilai menengah-tinggi menunjukkan variasi yang cukup untuk regresi, tetapi distribusi yang tidak merata perlu diperhatikan ketika menafsirkan galat keseluruhan.

Tabel 3. Statistik Deskriptif Fitur Utama

Fitur	Jumlah	Rata-rata	Simpangan Baku	Minimum	Median	Maksimum
PPKn	1.289	889,527	39,204	76,00	89,00	99,00
B. Indo	1.53	870,791	47,371	63,00	87,00	98,00
Matematika	1.53	849,190	43,549	40,00	85,00	98,00
B. Inggris	1.53	845,556	38,703	60,00	85,00	96,00
PKK	1.428	878,845	38,105	60,00	88,00	98,00
Rata Rata Rapor	1.53	872,781	30,116	76,50	87,75	94,00
Sakit	1.082	0,8845	15,992	-1,00	0,00	18,00
Izin	1.33	28,667	138,215	0,00	0,00	87,27
Alpa	1.228	26,434	143,257	0,00	0,00	87,09

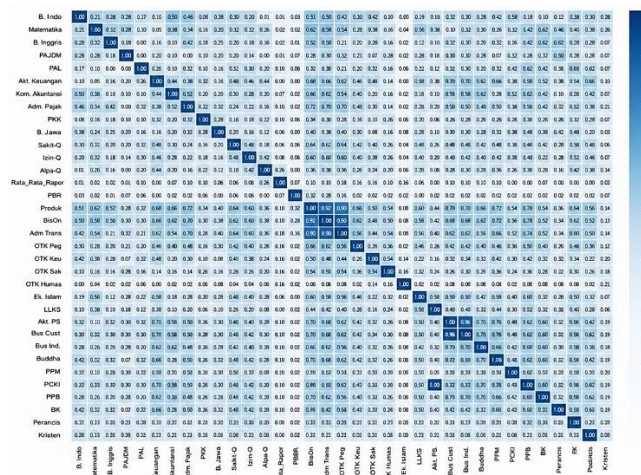
**Gambar 2.** Distribusi Nilai Akhir Akademik Siswa

Gambar 2 menunjukkan bahwa distribusi nilai akhir akademik siswa cenderung terkonsentrasi pada rentang nilai 85 hingga 90. Kondisi tersebut mengindikasikan bahwa sebagian besar siswa dalam dataset memiliki capaian akademik yang tergolong baik hingga sangat baik. Konsentrasi nilai pada rentang tersebut juga memperlihatkan adanya kecenderungan performa akademik yang relatif homogen pada mayoritas siswa, khususnya pada kelompok dengan hasil belajar tinggi.

Meskipun demikian, masih terdapat sejumlah siswa yang memiliki nilai pada rentang yang lebih rendah. Variasi nilai tersebut menunjukkan bahwa tingkat pencapaian akademik siswa tidak sepenuhnya seragam. Perbedaan ini dapat dipengaruhi oleh berbagai faktor, seperti jurusan, tingkat ketidakhadiran, penguasaan mata pelajaran umum, maupun capaian pada mata pelajaran kejuruan. Oleh karena itu, model prediksi tetap perlu mempelajari hubungan antarvariabel secara menyeluruh agar mampu menghasilkan prediksi nilai *Rata_Rata_Rapor* yang akurat, tidak hanya untuk siswa dengan nilai tinggi, tetapi juga untuk siswa yang berada pada kelompok nilai menengah dan rendah.

3.2. Korelasi Antarfitur Akademik

Peta panas korelasi ditunjukkan pada Gambar 3.

**Gambar 3.** Peta Panas Korelasi Antarfitur Akademik

Beberapa mata pelajaran menunjukkan hubungan positif dengan rata-rata rapor, sedangkan fitur ketidakhadiran cenderung memiliki kontribusi lebih kecil dibandingkan nilai akademik. Hasil korelasi ini tidak digunakan sebagai satu-satunya dasar keputusan, tetapi membantu memahami struktur data sebelum pemodelan. Karena Random Forest mampu menangkap pola nonlinier dan interaksi antarfitur, model tetap dapat memanfaatkan informasi yang tidak sepenuhnya terlihat melalui korelasi linear sederhana.

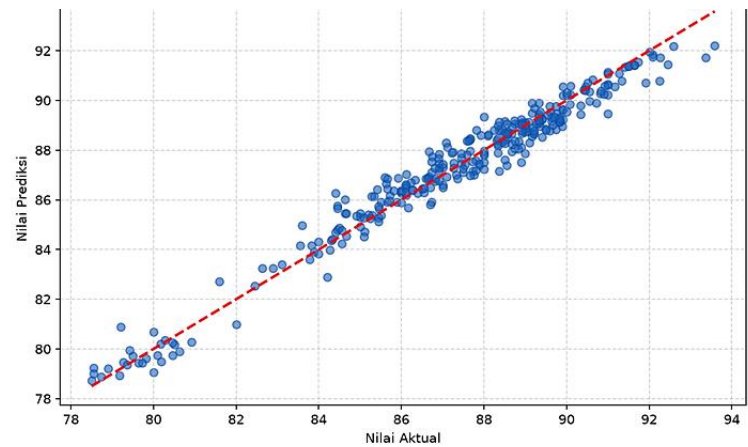
3.3. Hasil Evaluasi Model

Model *Random Forest Regression* yang digunakan dalam penelitian ini dibangun menggunakan konfigurasi sebanyak 100 pohon keputusan ($n_estimators = 100$) dengan kedalaman maksimum pohon sebesar 10 ($max_depth = 10$). Konfigurasi tersebut dipilih untuk menghasilkan model yang mampu mempelajari pola hubungan antara variabel masukan, seperti jurusan, tingkat ketidakhadiran, nilai mata pelajaran umum, dan nilai mata pelajaran kejuruan, terhadap variabel target berupa *Rata_Rata_Rapor*. Setelah proses pelatihan dilakukan menggunakan 1.224 data latih, kinerja model kemudian diuji menggunakan 306 data uji yang sebelumnya tidak digunakan dalam proses pembentukan model.

Hasil evaluasi menunjukkan bahwa model memperoleh nilai *Mean Absolute Error* (MAE) sebesar 0,4556 dan *Root Mean Squared Error* (RMSE) sebesar 0,5801. Nilai MAE tersebut menunjukkan bahwa rata-rata selisih absolut antara nilai prediksi model dan nilai aktual siswa adalah sekitar 0,46 poin. Sementara itu, nilai RMSE sebesar 0,58 mengindikasikan bahwa kesalahan prediksi yang lebih besar masih berada pada tingkat yang relatif rendah. Kedua nilai evaluasi tersebut berada di bawah satu poin pada skala nilai akademik, sehingga dapat disimpulkan bahwa model memiliki tingkat kesalahan prediksi yang rendah. Dengan demikian, model *Random Forest Regression* dinilai mampu memberikan prediksi nilai akhir akademik siswa secara cukup akurat dan dapat digunakan sebagai pendekatan pendukung dalam melakukan estimasi capaian akademik siswa.

Tabel 4. Hasil Evaluasi Model

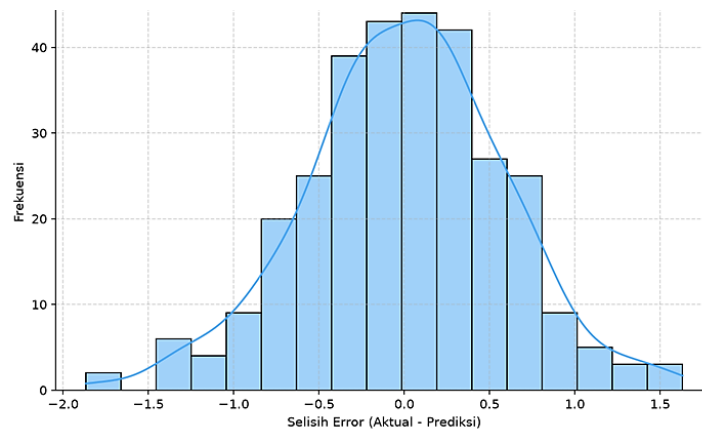
Metrik	Nilai	Interpretasi
MAE	0,4556	Rata-rata galat absolut prediksi sekitar 0,46 poin
RMSE	0,5801	Galat prediksi besar relatif jarang karena nilai RMSE rendah



Gambar 4. Perbandingan Nilai Aktual dan Nilai Prediksi

Gambar 4 menunjukkan bahwa titik-titik hasil prediksi tersebar mendekati garis diagonal. Kondisi ini mengindikasikan bahwa nilai prediksi yang dihasilkan oleh model memiliki tingkat kedekatan yang baik dengan nilai aktual. Semakin dekat posisi titik terhadap garis diagonal, maka semakin kecil selisih atau kesalahan antara nilai prediksi dan nilai sebenarnya. Pola sebaran titik yang relatif rapat di sekitar garis diagonal juga menunjukkan bahwa model mampu melakukan prediksi secara cukup akurat. Hal ini sejalan dengan hasil evaluasi menggunakan MAE dan RMSE yang rendah, sehingga dapat disimpulkan bahwa model memiliki performa prediksi yang baik.

Galat residual dihitung dari selisih antara nilai aktual dan nilai prediksi. Gambar 5 memperlihatkan bahwa residual cenderung terpusat di sekitar nol, yang berarti kesalahan prediksi model relatif seimbang dan tidak menunjukkan bias berlebih ke arah prediksi terlalu tinggi atau terlalu rendah. Distribusi residual seperti ini mendukung interpretasi bahwa model memiliki generalisasi yang baik pada data uji.



Gambar 5. Distribusi Galat Residual Prediksi

3.4. Analisis Kepentingan Fitur (Feature Importance)

Kepentingan fitur (feature importance) digunakan untuk mengidentifikasi variabel masukan yang memiliki kontribusi relatif terbesar terhadap pembentukan prediksi model. Pada Random Forest Regression, skor kepentingan diperoleh dari akumulasi penurunan ketidakmurnian atau Mean Squared Error (MSE) ketika sebuah fitur dipilih sebagai pemisah pada seluruh pohon keputusan, kemudian dinormalisasi sehingga total skor seluruh fitur bernilai 1. Oleh sebab itu, semakin besar skor suatu fitur, semakin besar perannya dalam membantu model menurunkan galat prediksi. Penggunaan kepentingan fitur untuk menginterpretasikan model Random Forest dalam prediksi hasil belajar juga digunakan pada penelitian prediksi nilai ujian dan nilai akhir semester siswa [11], [9]. Dalam penelitian Explainable Artificial Intelligence, feature importance dan SHAP digunakan untuk menerangkan kontribusi fitur agar hasil prediksi lebih transparan bagi pengambil keputusan [5].

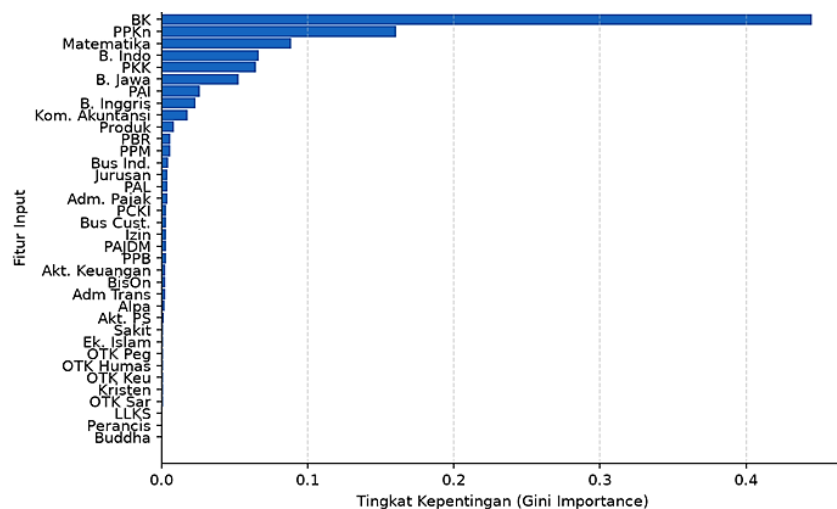
Skor kepentingan bersifat relatif di dalam model serta dipengaruhi oleh pola data, konfigurasi model, dan keterkaitan antarfitur. Dengan demikian, skor BK sebesar 0,4447 dapat dibaca sebagai sekitar 44,47% dari total kepentingan relatif yang dihitung model pada dataset ini. Skor tersebut tidak menunjukkan bahwa BK secara langsung menyebabkan 44,47% perubahan nilai akhir. Karena target Rata_Rata_Rapor merupakan agregasi nilai mata pelajaran, tingginya kepentingan fitur akademik perlu dipahami sebagai kedekatan statistik antara variabel masukan dan target, bukan sebagai bukti hubungan kausal tunggal [16], [24]. Puspangara serta Christian et al. menekankan bahwa kontribusi fitur merupakan penjelasan terhadap perilaku model, bukan bukti hubungan kausal [5].

Secara praktis, fitur dengan skor tinggi dapat digunakan sebagai indikator awal untuk menelaah pola capaian siswa dan mendukung perencanaan pendampingan yang lebih terarah. Namun, keputusan akademik tidak sebaiknya hanya bertumpu pada satu skor kepentingan fitur. Interpretasi hasil model perlu dikonfirmasi melalui evaluasi guru, konteks kurikulum, dan pengujian pada data periode lain agar rekomendasi yang dihasilkan tetap relevan bagi sekolah [24], [11]. Karena itu, hasil interpretasi sebaiknya digunakan sebagai dasar penelaahan lanjutan, bukan sebagai keputusan akademik otomatis [1], [5], [21].

Tabel 5. Sepuluh Fitur Teratas Berdasarkan Kepentingan Fitur

Peringkat	Fitur Masukan	Skor Kepentingan
1	BK	0,4447
2	PPKn	0,1599
3	Matematika	0,0885
4	B. Indo	0,0661
5	PKK	0,0639
6	B. Jawa	0,0521
7	PAI	0,0256
8	B. Inggris	0,0225
9	Kom. Akuntansi	0,0175
10	Produk	0,0081

Tabel 5 dan Gambar 6 menunjukkan bahwa BK memiliki skor kepentingan tertinggi (0,4447), diikuti PPKn (0,1599), Matematika (0,0885), B. Indonesia (0,0661), dan PKK (0,0639). Peringkat tersebut menunjukkan urutan kontribusi relatif fitur dalam model, bukan peringkat mutu mata pelajaran atau ukuran pengaruh kausal. Secara praktis, hasil ini dapat digunakan sekolah sebagai dasar pemantauan awal dan bahan diskusi bersama guru untuk meninjau pola capaian pada fitur yang paling sering digunakan model dalam membentuk prediksi [16], [9].



Gambar 6. Skor Kepentingan Fitur Random Forest

3.5. Pembahasan

Galat yang rendah pada penelitian ini konsisten dengan temuan Kusworo et al. bahwa Random Forest dapat memodelkan nilai akhir semester [6] dan dengan penelitian Sari et al. yang menunjukkan kemampuan model pohon ansambel menangkap hubungan antara faktor akademik dan prestasi [9]. Kesamaan tersebut dapat dijelaskan oleh kemampuan Random Forest menangani hubungan nonlinier, interaksi antarfitur, dan korelasi di antara nilai mata pelajaran. Namun, hasil numerik tidak dapat dibandingkan secara langsung karena penelitian terdahulu menggunakan dataset, target, skala nilai, fitur, dan prosedur evaluasi yang berbeda. Studi multi-model juga menunjukkan bahwa model linear, boosting, atau ensemble lain dapat mengungguli Random Forest pada kondisi tertentu [10], [16]. Temuan ini sejalan dengan studi perbandingan nasional yang menunjukkan bahwa Random Forest dapat unggul pada sebagian dataset, sedangkan SVM, Decision Tree, regresi, atau metode lain dapat lebih sesuai pada karakteristik data yang berbeda [20], [26], [13], [14], [8].

Implikasi praktis penelitian adalah penggunaan model sebagai komponen prediktif dalam sistem peringatan dini. Secara konseptual, nilai dan ketidakhadiran yang tersedia pada suatu periode diproses untuk menghasilkan estimasi capaian; hasil kemudian dibandingkan dengan ambang risiko yang ditetapkan sekolah; siswa yang terindikasi memerlukan perhatian ditelaah kembali oleh wali kelas, guru mata pelajaran, dan guru BK; selanjutnya sekolah menentukan pendampingan yang proporsional. Alur tersebut menempatkan prediksi sebagai alat penyaringan dan prioritasasi, bukan keputusan otomatis. Analisis kepentingan fitur dapat membantu menjelaskan alasan umum model, sedangkan penjelasan individual dan keputusan intervensi tetap memerlukan data kontekstual serta pertimbangan profesional [12]. Studi pada siswa SMK dan mahasiswa menempatkan prediksi sebagai alat identifikasi awal agar pendampingan dapat diberikan kepada kelompok yang berisiko [1], [5], [21].

Penelitian ini memiliki beberapa batasan. Pertama, Rata_Rata_Rapor digunakan sebagai proksi nilai akhir akademik dan dibentuk dari nilai mata pelajaran yang sebagian menjadi fitur, sehingga terdapat risiko information leakage apabila model ditafsirkan sebagai ramalan nilai masa depan. Kedua, data berasal dari satu sekolah dan satu pembagian holdout, sehingga generalisasi lintas periode dan institusi belum diuji. Ketiga, penelitian belum membandingkan Random Forest dengan baseline linear, Gradient Boosting, XGBoost, atau model regresi lain dalam kerangka validasi yang sama. Keempat, impurity-based feature importance memiliki keterbatasan interpretasi pada fitur yang berkorelasi. Penelitian berikutnya perlu menggunakan nilai ujian aktual atau target pada periode berikutnya, menerapkan validasi silang dan pemisahan temporal, membandingkan beberapa algoritma, serta menggunakan permutation importance atau SHAP untuk memperkuat transparansi [10]-[12]. Selain itu, impurity-based feature importance perlu dilengkapi dengan permutation importance atau SHAP untuk mengurangi risiko interpretasi yang terlalu sederhana pada fitur-fitur yang saling berkorelasi [5].

4. KESIMPULAN

Penelitian ini menerapkan Random Forest Regression untuk mengestimasi nilai akhir akademik siswa berdasarkan 1.530 catatan historis dari enam program keahlian. Model menggunakan 1.224 data latih dan 306 data uji dengan parameter $n_estimators=100$, $max_depth=10$, dan $random_state=42$. Hasil evaluasi menghasilkan MAE 0,4556 dan RMSE 0,5801. Nilai RMSE yang sekitar 1,27 kali MAE, sebaran aktual-prediksi yang dekat dengan garis diagonal, dan residual yang berpusat di sekitar nol menunjukkan kinerja yang baik pada data uji, meskipun kestabilan lintas periode dan sekolah masih memerlukan validasi. Kontribusi praktis penelitian adalah menyediakan komponen prediktif untuk sistem peringatan dini akademik: model

dapat membantu menyaring siswa yang memerlukan penelaahan lebih lanjut agar pendampingan dapat dilakukan lebih awal.

BK, PPKn, Matematika, B. Indonesia, dan PKK memiliki kontribusi relatif tertinggi dalam model. Hasil tersebut harus dipahami sebagai pola statistik, bukan pengaruh kausal, dan tidak boleh digunakan sebagai dasar tunggal keputusan akademik. Pengembangan berikutnya perlu menggunakan target pada periode mendatang atau nilai ujian aktual, melakukan validasi silang dan temporal, membandingkan beberapa algoritma, serta menerapkan teknik interpretasi seperti permutation importance atau SHAP. Dalam penerapan sekolah, setiap keluaran model harus dikonfirmasi oleh guru dan konselor agar sistem peringatan dini tetap transparan, proporsional, dan berorientasi pada dukungan siswa.

REFERENSI

- [1] A. Fatunnisa and H. Marcos, "Prediksi Kelulusan Tepat Waktu Siswa SMK Teknik Komputer Menggunakan Algoritma Random Forest Prediction of On-Time Graduation for Computer Engineering Vocational School Students Using the Random Forest Algorithm," vol. 14, no. April, pp. 101–111, 2024.
- [2] I. Jurnal *et al.*, "Prediksi Kelulusan Mahasiswa Berdasarkan Data Kunjung dan Peminjaman Buku menggunakan Rapid Miner dengan Metode C . 45 dan Random Forest," vol. 7, no. 2, pp. 14–20, 2023.
- [3] J. M. Polgan *et al.*, "Pola Prediksi Kelulusan Siswa Madrasah Aliyah Swasta dengan Support Vector Machine dan Random Forest," vol. 12, pp. 387–400, 2023.
- [4] A. Frananda, A. Naibaho, and A. Zahra, "Prediksi Kelulusan Siswa Sekolah Menengah Pertama Menggunakan Machine Learning," vol. 11, no. 3, 2023.
- [5] G. Puspa, M. Febrian, and R. Amri, "Pemanfaatan Random Forest untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa Prediction of The Student Graduation Time Using Random Forest Case Study : Bali Institute Design and Business," no. 2, pp. 413–426, 2025, doi: 10.26798/jiko.v9i2.1886.
- [6] A. Fikri, N. Hidayasari, and Z. Efendi, vol. 11, no. 1, pp. 1489–1502, 2026.
- [7] F. M. Almufqi, A. Voutama, S. Informasi, I. Komputer, U. S. Karawang, and T. Timur, "Perbandingan Metode Data Mining Untuk Memprediksi Prestasi," vol. 15, no. 1, pp. 61–66, 2023, doi: 10.30736/jt.v15i1.929.
- [8] U. Indahyanti *et al.*, "Perbandingan Algoritma Machine Learning dalam," vol. 11, no. 2, pp. 96–105, 2025.
- [9] S. Kusworo, N. A. Santoso, and R. D. Kurniawan, "Prediksi Nilai Akhir Semester Siswa Menggunakan Algoritma Random Forest," no. 2, 2024.
- [10] R. Sari and W. Kurniadi, "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," vol. 6, no. 4, pp. 205–213, 2025.
- [11] R. Amin, A. Siwi, and F. Utami, "Prediksi Nilai Ujian Berdasarkan Kebiasaan Siswa Menggunakan Algoritma Random Forest Regressor," vol. 10, no. 2, pp. 149–158, 2025.
- [12] N. Maftucha, S. Salma, N. Rahmayuna, and N. Wakhidah, "Perbandingan Algoritma Machine Learning Dalam Memprediksi Kelulusan Siswa," vol. 19, no. 2, pp. 116–128.
- [13] L. Javier, P. Charle-cu, A. A. Gado, N. Ousmane, C. Luc, and L. Noem, "Using Machine Learning to Fight Child Acute Malnutrition and Predict Weight Gain During Outpatient Treatment with a Simplified Combined Protocol," pp. 1–13, 2024.
- [14] R. Rismaya, D. Yuniarto, and D. Setiadi, "Penerapan Algoritma Machine Learning dalam Prediksi Prestasi Akademik Mahasiswa," 2025.
- [15] A. Wijaya, D. Permana, and N. Rahmawati, "Pengembangan Sistem Informasi Absensi Pegawai Berbasis Web Menggunakan Laravel Framework," *J. Rekayasa Teknol. Inf.*, vol. 9, no. 1, pp. 32–39, 2023.
- [16] M. S. Amin, E. Sekar, and R. Forest, "Analisis Prediktif Terhadap Kinerja Siswa dalam Ujian Menggunakan Algoritma Random Forest dan K-Nearest," no. 4, pp. 1–13, 2025.
- [17] R. Rasendriya, A. O. Marundrury, Y. L. Jumadi, and A. D. Kuswanto, "MEMPREDIKSI KELULUSAN SISWA MENGGUNAKAN ORANGE," pp. 60–71, 2025.
- [18] R. B. Hidayah, K. Latifah, and B. A. Herlambang, "Prediksi gaya belajar menggunakan algoritma random forest dengan optimasi hyperparameter tuning berbasis web 1)," vol. 10, no. 2, pp. 864–873, 2025.
- [19] N. Fitriyani, "Predicting Students ' Mathematics Scores from Reading Scores Using Supervised Learning," vol. 6, no. April, pp. 494–503, 2026.
- [20] S. Junaidi, R. V. Anggela, and D. Kariman, "Journal Of Applied Computer Science And Technology (JACOST) Klasifikasi Metode Data Mining untuk Prediksi Kelulusan Tepat Waktu Mahasiswa dengan Algoritma Naïve Bayes , Random Forest , Support Vector Machine (SVM) dan Artificial Neural Network (ANN)," vol. 5, no. 1, pp. 109–119, 2024.

- [21] M. A. Rachman *et al.*, “Aplikasi Prediksi Tingkat Kelulusan Mahasiswa Berdasarkan Data Akademik dan Demografi Menggunakan Algoritma Klasifikasi Random Jurnal Media Informatika [JUMIN],” vol. 7, no. 1, pp. 247–255, 2026.
- [22] M. R. Forest, “Klasifikasi prestasi akademik mahasiswa menggunakan metode”.
- [23] T. Gori *et al.*, “Preprocessing Data Dan Klasifikasi Untuk Prediksi Kinerja Data Preprocessing And Classification For Predicting Student,” vol. 11, no. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.
- [24] H. Andrianof, A. P. Gusman, O. A. Putra, and A. Info, “Jurnal Sains Informatika Terapan (JSIT),” pp. 24–28, 2025.
- [25] M. Anggito, H. Hadisuwarno, and K. T. Martono, “Komparasi performa model machine learning algoritma XGBoost dan Random Forest pada studi kasus mendeteksi stunting,” vol. 22, no. 2, pp. 266–278, 2025.
- [26] N. R. Muntiari, K. H. Hanif, and L. Herawati, “Deteksi Stanting Pada Balita Dengan Menggunakan Perbandingan Diterima : Diterbitkan :,” vol. 13, no. 1, pp. 17–23, 2025.