Jurnal Teknologi Pendidikan Vol: 2, No 4, 2025, Page: 1-13

Analisis Prediktif Terhadap Kinerja Siswa dalam Ujian Menggunakan Algoritma *Random Forest* dan *K-Nearest Neighbors*

M. Syaiful Amin*, Deuis Nur astride, Eling Sekar

Universitas Amikom Purwokerto

Abstrak: Penelitian ini bertujuan untuk menganalisis dan memprediksi kinerja siswa dalam ujian menggunakan algoritma Random Forest dan K-Nearest Neighbors (KNN). Dataset yang digunakan berjumlah 1000 entri, mencakup nilai ujian matematika, membaca, menulis, serta variabel demografis seperti jenis kelamin, ras, tingkat pendidikan orang tua, jenis makan siang, dan partisipasi kursus persiapan ujian. Metode penelitian ini meliputi tahap preprocessing data (pembersihan, transformasi kategori menjadi numerik melalui Label Encoding, serta normalisasi dengan Min-Max Scaling), pembagian data menjadi data pelatihan (80%) dan data pengujian (20%), penerapan algoritma Random Forest dan KNN dengan optimasi parameter melalui Grid Search, serta evaluasi model menggunakan metrik mean squared error (MSE) dan R-squared. Hasil penelitian menunjukkan bahwa Random Forest memberikan kinerja prediksi yang lebih baik dibanding KNN dengan MSE sebesar 0.0025 dan R-squared 0.9149, sedangkan KNN memiliki MSE sebesar 0.0133 dan R-squared 0.5533. Analisis feature importance pada model Random Forest mengidentifikasi nilai matematika, membaca, dan tingkat pendidikan orang tua sebagai faktor utama yang mempengaruhi prediksi nilai writing score siswa. Penelitian ini memberikan kontribusi terhadap pemanfaatan machine learning dalam memprediksi kinerja akademik siswa, serta dapat dijadikan dasar untuk mengembangkan model prediktif yang lebih efektif dalam membantu guru dan pembuat kebijakan merancang intervensi pendidikan yang tepat guna meningkatkan kualitas belajar siswa.

Kata Kunci: Pendidikan, *Analisis prediktif, Random Forest, K-Nearest Neighbors (KNN), Machine learning*

DOI: https://doi.org/10.47134/jtp.v2i4.1385
*Correspondence: M. Syaiful Amin
Email: syaifulamin.amikompurwokerto.ac.id

Received: 22-04-2025
Accepted: 22-05-2025
Published: 22-06-2025

Abstract: *This study aims to analyze and predict student performance in exams using the Random Forest and K-Nearest Neighbors (KNN) algorithms. The dataset used is 1000 entries, covering math, reading, writing test scores, as well as demographic variables such as gender, race, parental education level, type of lunch, and exam preparation course participation. The research method includes data preprocessing stages (cleaning, categorical transformation to numeric through Label Encoding, and normalization with Min-Max Scaling), data division into training data (80%) and testing data (20%), application of the Random Forest and KNN algorithms with parameter optimization through Grid Search, and model evaluation using the mean squared error (MSE) and R-squared metrics. The results showed that Random Forest provided better prediction performance than KNN with an MSE of 0.0025 and an R-squared of 0.9149, while KNN had an MSE of 0.0133 and an R-squared of 0.5533. Feature importance analysis in the Random Forest model identified math, reading, and parental education levels as the main factors influencing the prediction of students' writing scores. This study contributes to the use of machine learning in predicting students' academic performance, and can be used as a basis for developing more effective predictive models to help teachers and policymakers design appropriate educational interventions to improve the quality of student learning.*

Keywords*: Education, Predictive analysis, Random Forest, K-Nearest Neighbors (KNN), Machine learning*

Pendahuluan

Pendidikan merupakan salah satu faktor utama dalam pengembangan suatu negara. Kinerja akademik siswa menjadi indikator penting dalam mengukur efektivitas sistem pendidikan, dan hasil ini tak hanya penting bagi siswa itu sendiri, tetapi juga bagi para guru, orang tua, dan pembuat kebijakan (Adnan et al, 2024). Namun, memahami dan memprediksi kinerja siswa dalam ujian adalah tantangan besar. Banyak faktor yang saling berhubungan dan mempengaruhi hasil belajar siswa, seperti latar belakang sosial-ekonomi, kebiasaan belajar, motivasi, lingkungan keluarga, serta keterlibatan dalam kegiatan akademik (Zulkifli, 2022). Kemajuan dalam bidang pendidikan tidak hanya bergantung pada kurikulum yang baik, tetapi juga pada kemampuan untuk memahami dan meningkatkan kinerja siswa. Dalam beberapa tahun terakhir, analisis data telah menjadi alat yang sangat berharga dalam berbagai bidang, termasuk pendidikan. Dengan meningkatnya ketersediaan data pendidikan, teknik analisis prediktif memberikan peluang untuk mengevaluasi dan memprediksi hasil ujian siswa dengan lebih akurat. Salah satu teknik analisis data yang sedang berkembang pesat adalah machine learning. Teknik ini dapat digunakan untuk membuat prediksi berdasarkan data historis, yang memungkinkan pengambilan keputusan yang lebih baik dan lebih tepat.

Salah satu penerapan machine learning dalam pendidikan adalah dalam analisis prediktif terhadap kinerja siswa. Prediksi kinerja siswa dapat membantu guru dan institusi pendidikan untuk mengidentifikasi siswa yang mungkin memerlukan bantuan tambahan, serta untuk merancang strategi pengajaran yang lebih efektif. Pendekatan analisis prediktif memanfaatkan berbagai algoritma pembelajaran mesin (machine learning) untuk menggali wawasan dari data besar *(big data) (*Sabrina et al, 2024). Algoritma seperti Random Forest (RF) dan K-Nearest Neighbors (KNN) menjadi pilihan yang baik karena keunggulannya dalam menangani dataset yang kompleks dan menghasilkan prediksi yang baik (Rahmadden et al, 2024). Random Forest, misalnya, mampu menangani data dengan variabel input yang besar dan mengurangi risiko overfitting dengan membangun banyak pohon keputusan *(decision trees)* secara acak (Religia et al, 2021). Di sisi lain, K-Nearest Neighbors menawarkan pendekatan berbasis kesamaan data, yang sederhana namun efektif dalam beberapa kasus aplikasi pendidikan.

Random Forest adalah algoritma berbasis pohon keputusan yang menghasilkan prediksi dengan membuat beberapa pohon keputusan dan mengkombinasikan hasilnya (Religia et al, 2021). Algoritma ini dikenal dengan kemampuannya yang baik dalam menangani data yang tidak seimbang dan dapat memberikan hasil prediksi yang akurat. Di sisi lain, K-Nearest Neighbors adalah algoritma yang memprediksi nilai berdasarkan kemiripan antara data baru dengan data yang sudah ada (Sabita & Yahfizham, 2024). Algoritma ini sederhana namun sangat efektif untuk berbagai jenis data.

Penelitian ini bertujuan untuk menganalisis kinerja siswa dalam ujian menggunakan algoritma Random Forest dan K-Nearest Neighbors. Solusi yang ditawarkan oleh penelitian ini adalah model prediktif yang dapat digunakan untuk memprediksi kinerja siswa berdasarkan data historis mereka. Penelitian ini juga membandingkan kinerja kedua

algoritma tersebut dalam konteks pendidikan, sehingga dapat memberikan wawasan yang lebih mendalam tentang keunggulan dan kelemahan masing-masing algoritma.

Selain itu, penggunaan analisis prediktif juga dapat membantu dalam pengembangan kebijakan pendidikan yang lebih efektif. Dengan memahami faktor-faktor yang mempengaruhi kinerja siswa, pembuat kebijakan dapat merancang program yang lebih tepat sasaran untuk meningkatkan kualitas pendidikan. Misalnya, data yang dianalisis melalui algoritma machine learning dapat mengidentifikasi tren atau pola yang menunjukkan siswa mana yang berisiko mengalami kesulitan akademik (Gori et al, 2024). Informasi ini dapat digunakan untuk memberikan dukungan tambahan, seperti bimbingan belajar atau program intervensi, sebelum masalah tersebut menjadi lebih serius.

Studi-studi terkait yang telah dilakukan sebelumnya menunjukkan bahwa machine learning dapat digunakan untuk memprediksi kinerja siswa dengan tingkat akurasi yang tinggi. Misalnya, penelitian oleh Li et al. menggunakan Neural Networks dan berhasil meningkatkan akurasi prediksi kinerja siswa (Cazarez & Martin, 2018). Penelitian terbaru oleh Saputra et al. dalam jurnal "Komparasi Machine Learning Berbasis PSO untuk Prediksi Tingkat Kebijakan Belajar Berbasis E-Learning" menunjukkan bahwa penggunaan algoritma machine learning berbasis Particle Swarm Optimization (PSO) dapat secara signifikan meningkatkan akurasi prediksi kinerja siswa (Saputra et al, 2023). Hasil-hasil ini menunjukkan bahwa penggunaan teknologi machine learning dapat memberikan manfaat yang signifikan dalam bidang pendidikan.

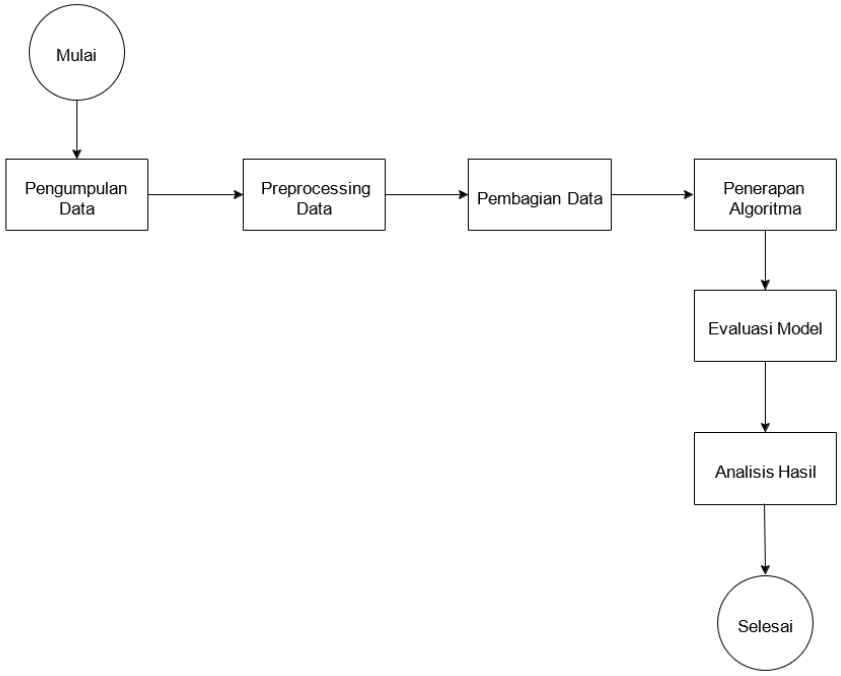
Dalam penelitian ini, data kinerja siswa akan dikumpulkan dan dianalisis menggunakan kedua algoritma tersebut. Hasil analisis akan dibandingkan untuk menentukan algoritma mana yang memberikan prediksi yang lebih akurat. Penelitian ini diharapkan dapat memberikan kontribusi yang signifikan dalam bidang pendidikan dengan menyediakan alat yang dapat digunakan oleh guru dan institusi pendidikan untuk meningkatkan kualitas pengajaran dan pembelajaran.

Dengan adanya analisis prediktif ini, diharapkan dapat membantu dalam mengidentifikasi siswa yang memerlukan bantuan tambahan lebih dini, sehingga intervensi yang tepat dapat dilakukan sebelum siswa mengalami kesulitan yang lebih serius. Selain itu, hasil dari penelitian ini juga dapat digunakan untuk mengembangkan strategi pengajaran yang lebih efektif dan efisien.

Metodologi

Tahapan Penelitian

Penelitian ini menggunakan metode kuantitatif dengan menerapkan algoritma pembelajaran mesin, yakni Random Forest dan K-Nearest Neighbors, dalam analisis prediktif terhadap kinerja siswa dalam ujian. Implementasi algoritma tersebut dilakukan melalui serangkaian uji coba dengan variasi parameter seperti jumlah pohon keputusan (n_estimators) dan kedalaman pohon (max_depth) untuk Random Forest, serta variasi nilai k dan ukuran jarak Euclidean untuk K-Nearest Neighbors. Pada tahap evaluasi, kinerja kedua algoritma dinilai menggunakan metrik mean squared error (MSE) dan R-squared. Adapun tahapan-tahapan dari penelitian ini ditunjukkan pada Gambar 1 di bawah ini:



Gambar 1. Diagram Alur Penelitian

1. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah dataset "Students Performance in Exams" yang diperoleh dari website kaggle.com dengan perintah: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams. Dataset ini mencakup berbagai informasi penting mengenai performa akademik siswa, termasuk nilai ujian dalam mata pelajaran matematika, membaca, dan menulis. Selain itu, dataset ini juga mencakup variabel demografis seperti jenis kelamin, ras, tingkat pendidikan orang tua, jenis makan siang yang diterima oleh siswa, serta partisipasi siswa dalam kursus persiapan ujian. Data ini penting untuk memahami konteks dan faktor-faktor yang mungkin mempengaruhi kinerja siswa.

2. *Preprocessing* Data

- **Pembersihan Data**

Tahap pertama dari preprocessing adalah cleaning (Efriadi et al, 2022). Data yang noise, tidak konsisten, atau tidak relevan dihilangkan melalui pembersihan data (Putri et al, 2022). Pembersihan data (cleaning) merupakan langkah awal dari proses KDD. Seluruh atribut pada dataset tersebut akan diseleksi untuk mendapatkan atribut-atribut yang berisi nilai-nilai relevan, serta memastikan tidak ada nilai yang hilang atau data yang redundant. Proses ini adalah syarat wajib dalam data mining untuk menghasilkan dataset yang siap digunakan pada tahap mining data (Setio & Prasetyaningrum, 2021).

- **Transformasi Data**

Setelah kesalahan dalam data dihilangkan, data kemudian ditransformasikan sesuai dengan tipe datanya, yang diklasifikasikan sebagai data kategorikal. Transformasi data mengacu pada serangkaian teknik atau metode untuk mengubah data dari bentuk atau distribusi awalnya menjadi bentuk atau distribusi yang sesuai dengan kebutuhan analisis

atau model. Transformasi data sering kali diperlukan untuk memenuhi asumsi statistik tertentu atau untuk meningkatkan kinerja model machine learning. Pada penelitian ini, transformasi data dilakukan dengan menggunakan teknik encoding seperti Label Encoding atau One-Hot Encoding. Transformasi ini diperlukan agar algoritma machine learning dapat memproses data dengan benar (Cumel et al, 2022).

- **Normalisasi Data**

Mengubah skala data agar berada dalam rentang yang sama, biasanya antara 0 dan 1, untuk menghindari bias dalam algoritma dan memastikan semua fitur memiliki bobot yang sama. Normalisasi dilakukan dengan menggunakan teknik seperti Min-Max Scaling atau Z-Score Standardization.

- ***Feature Selection***

Seleksi fitur adalah salah satu teknik preprocessing yang umum digunakan dalam pembelajaran mesin dan statistik untuk meningkatkan kinerja pembelajaran serta mengatasi masalah yang timbul dari data berdimensi tinggi. Proses ini melibatkan pemilihan subset fitur atau variabel dari dataset yang akan digunakan untuk membangun model atau melakukan analisis. Tujuan utamanya adalah untuk meningkatkan kinerja model dengan hanya menggunakan fitur-fitur yang paling relevan atau signifikan, sambil mengurangi kompleksitas dan beban komputasi yang terkait dengan penggunaan seluruh set fitur (Ariyoga, 2022).

3. Pembagian Data

Pembagian data adalah proses membagi dataset menjadi dua atau lebih subset yang berbeda, biasanya untuk keperluan pelatihan dan pengujian model machine learning. Tujuan utama dari pembagian data adalah untuk mengukur kinerja model pada data yang belum pernah dilihat selama proses pelatihan. Ada beberapa jenis pembagian data yang umum digunakan:

- Data Pelatihan: Data pelatihan digunakan untuk mengembangkan model dan melatih algoritma. Subset dari dataset ini berisi contoh-contoh yang diberi label atau output yang diinginkan. Model akan menggunakan data ini untuk belajar dan membuat prediksi pada data baru yang belum pernah dilihat sebelumnya.
- Data Pengujian: Ketika dihadapkan pada data baru yang tidak teridentifikasi, pengujian dilakukan untuk mengamati performa algoritma yang telah dilatih sebelumnya. Subset dari dataset ini digunakan untuk menguji kinerja model machine learning setelah model dilatih dengan data pelatihan. Data pengujian berfungsi untuk mengukur sejauh mana model mampu melakukan generalisasi pada data yang belum pernah dilihat sebelumnya, yang disebut sebagai data baru atau data yang belum terlihat (Rahmansyah et al, 2018).

4. Penerapan Algoritma

- *Random Forest*

Algoritma ini menggunakan banyak pohon keputusan yang dibuat dari subset data yang berbeda untuk meningkatkan akurasi prediksi. Parameter seperti jumlah pohon keputusan (n_estimators) dan kedalaman maksimum pohon (max_depth) dioptimalkan menggunakan teknik pencarian grid (grid search). Algoritma ini diterapkan dengan menggunakan pustaka scikit-learn pada bahasa pemrograman Python.

- *K-Nearest Neighbors (K-NN)*

Algoritma ini memprediksi kinerja siswa berdasarkan kemiripan dengan data yang ada dalam dataset. Parameter yang dioptimalkan meliputi jumlah tetangga terdekat (nilai k) dan metrik jarak (Euclidean distance) untuk mencapai kinerja terbaik. Algoritma ini juga diterapkan menggunakan pustaka scikit-learn pada bahasa pemrograman Python.

5. Evaluasi Model

Model yang telah dibangun diuji menggunakan data pengujian. Evaluasi dilakukan dengan menggunakan metrik berikut untuk menilai kinerja model:

- *Mean Squared Error (MSE):* Mengukur rata-rata kuadrat kesalahan antara nilai prediksi dan nilai aktual.
- *R-squared:* Mengukur proporsi variabilitas data yang dapat dijelaskan oleh model. Evaluasi ini dilakukan dengan menggunakan fungsi evaluasi dari pustaka scikit-learn.

6. Analisis Hasil

Hasil prediksi dari kedua algoritma dianalisis dan dibandingkan untuk menentukan algoritma yang memberikan prediksi lebih akurat. Analisis tambahan dilakukan untuk memahami faktor-faktor yang paling mempengaruhi kinerja siswa dalam ujian. Ini mencakup interpretasi feature importance dari model Random Forest dan analisis kontribusi masing-masing fitur terhadap hasil prediksi. Hasil analisis ini kemudian dibandingkan dengan hasil penelitian sebelumnya untuk memberikan konteks dan validasi tambahan.

Implementasi dan Pengujian

Implementasi algoritma Random Forest dan K-Nearest Neighbors dilakukan menggunakan pustaka scikit-learn pada bahasa pemrograman Python. Langkah-langkah implementasi adalah sebagai berikut:

1. *Random Forest:*

- Menentukan jumlah pohon keputusan yang akan digunakan dalam Random Forest berdasarkan hasil optimasi parameter. Ini dilakukan dengan menggunakan teknik pencarian grid (grid search) untuk menemukan kombinasi parameter terbaik.
- Melatih model menggunakan data pelatihan yang telah diproses. Model dilatih dengan menggunakan fungsi fit dari objek RandomForestRegressor di pustaka scikit-learn.
- Mengoptimalkan parameter seperti jumlah pohon (n_estimators) dan kedalaman maksimum pohon (max_depth) menggunakan teknik pencarian grid (grid search).

Proses ini melibatkan penggunaan fungsi GridSearchCV dari pustaka scikit-learn untuk melakukan pencarian parameter optimal secara otomatis.

2. *K-Nearest Neighbors:*

- Menentukan jumlah tetangga terdekat (nilai k) yang akan digunakan dalam algoritma K-NN berdasarkan hasil optimasi parameter. Ini juga dilakukan dengan menggunakan teknik pencarian grid (grid search) untuk menemukan nilai k yang optimal.
- Melatih model menggunakan data pelatihan yang telah diproses. Model dilatih dengan menggunakan fungsi fit dari objek KneighborsRegressor di pustaka scikit-learn.
- Mengoptimalkan parameter seperti nilai k dan metrik jarak (Euclidean) untuk mencapai kinerja terbaik. Proses optimasi ini juga dilakukan dengan menggunakan fungsi GridSearchCV.

Evaluasi Model

Evaluasi dilakukan dengan menggunakan data pengujian yang sebelumnya telah dipisahkan. Metrik yang digunakan untuk evaluasi meliputi:

- Mean Squared Error (MSE): Proporsi prediksi yang benar di antara seluruh prediksi yang dibuat.
- R-squared: Proporsi prediksi positif yang benar dibandingkan dengan total prediksi positif yang dibuat.

Evaluasi ini dilakukan dengan menggunakan fungsi evaluasi dari pustaka scikit-learn, seperti mean_squared_error dan r2_score.

Hasil dan Pembahasan

Dalam penelitian ini, dataset dibagi menjadi data pelatihan dan data pengujian dengan rasio 80:20 (data pelatihan:data pengujian). Data yang dibagi terlebih dahulu dilakukan normalisasi terhadap nilai yang hilang atau tidak wajar dengan menghapus entri yang tidak lengkap. Teknik normalisasi dilakukan dengan mengonversi semua nilai yang tidak sesuai menjadi tipe data yang benar.

Total entri data dari dataset yang digunakan adalah sebanyak 1000 data. Setiap entri data terdiri dari informasi penting terkait kinerja akademik siswa, termasuk nilai ujian dalam mata pelajaran matematika, membaca, dan menulis, serta variabel demografis seperti jenis kelamin, ras, tingkat pendidikan orang tua, jenis makan siang, dan partisipasi dalam kursus persiapan ujian. Informasi ini penting untuk memahami konteks dan faktor-faktor yang mungkin mempengaruhi kinerja siswa.

Pada tahap preprocessing, dilakukan beberapa langkah penting seperti pembersihan data, transformasi data kategori menjadi numerik menggunakan Label Encoding, serta normalisasi data numerik menggunakan teknik Min-Max Scaling. Tujuan dari preprocessing adalah untuk memastikan bahwa data yang digunakan dalam model bebas dari nilai yang hilang, konsisten, dan berada dalam rentang yang sama. Langkah pertama dalam preprocessing adalah membersihkan dataset dari nilai yang hilang atau tidak konsisten. Nilai yang hilang akan diisi menggunakan metode imputasi yang sesuai seperti

mean atau median. Data yang tidak konsisten atau outliers yang tidak sesuai dengan distribusi umum juga akan diidentifikasi dan ditangani. Langkah selanjutnya yaitu transformasi data, ata kategori seperti gender, ras, tingkat pendidikan orang tua, jenis makan siang, dan partisipasi dalam kursus persiapan ujian akan dikodekan menjadi format numerik menggunakan teknik encoding seperti Label Encoding atau One-Hot Encoding. Transformasi ini diperlukan agar algoritma machine learning dapat memproses data dengan benar. Setelah tahap transformasi data selesai selanjutnya yaitu tahap normalisasi data dengan mengubah skala data agar berada dalam rentang yang sama, biasanya antara 0 dan 1, untuk menghindari bias dalam algoritma dan memastikan semua fitur memiliki bobot yang sama. Normalisasi dilakukan dengan menggunakan teknik seperti Min-Max Scaling atau Z-Score Standardization.

```
[6]  df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   gender                       1000 non-null   object
 1   race/ethnicity               1000 non-null   object
 2   parental level of education  1000 non-null   object
 3   lunch                        1000 non-null   object
 4   test preparation course      1000 non-null   object
 5   math score                   1000 non-null   int64
 6   reading score                1000 non-null   int64
 7   writing score                1000 non-null   int64
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

Gambar 2. Fitur Yang Digunakan Dalam Dataset

dan data pengujian (20%). Pembagian ini dilakukan secara acak untuk memastikan representativitas yang baik dari setiap subset data. Data pelatihan digunakan untuk membangun model prediksi, sementara data pengujian digunakan untuk mengevaluasi kinerja model. Proses pembagian data dilakukan menggunakan fungsi train_test_split dari pustaka scikit-learn.

Dalam penelitian ini, algoritma yang digunakan adalah Random Forest dan K-Nearest Neighbors (K-NN). Algoritma Random Forest menggunakan banyak pohon keputusan yang dibuat dari subset data yang berbeda untuk meningkatkan akurasi prediksi. Parameter seperti jumlah pohon keputusan (n_estimators) dan kedalaman maksimum pohon (max_depth) dioptimalkan menggunakan teknik pencarian grid (grid search). Model dioptimalkan menggunakan parameter grid dengan n_estimators: [50, 100, 200] dan max_depth: [10, 20, None]. K-Nearest Neighbors (K-NN): Algoritma ini memprediksi kinerja siswa berdasarkan kemiripan dengan data yang ada dalam dataset. Parameter yang dioptimalkan meliputi jumlah tetangga terdekat (nilai k) dan metrik jarak (Euclidean distance). Model dioptimalkan menggunakan parameter grid dengan n_neighbors: [3, 5, 7, 9] dan metric: ['euclidean'].

Model yang telah dibangun diuji menggunakan data pengujian dan kinerja model dievaluasi menggunakan metrik mean squared error (MSE) dan R-squared. Berikut adalah hasil evaluasi dari kedua algoritma:

Tabel 1. Hasil Evaluasi Metode Random Forest

Metrik	Hasil
MSE	0.0025
R-squared	0.9149

Tabel 2. Hasil Evaluasi Metode K-Nearest Neighbors

Metrik	Hasil
MSE	0.0133
R-squared	0.5533

Secara keseluruhan, algoritma Random Forest menunjukkan kinerja yang lebih baik dibandingkan dengan K-Nearest Neighbors dalam memprediksi nilai writing score siswa. Hal ini ditunjukkan oleh nilai mean squared error (MSE) yang lebih rendah dan R-squared yang lebih tinggi pada model Random Forest.

Tabel 3. Hasil Evaluasi Metode K-Nearest Neighbors

Algoritma	MSE	R-squared
Random Forest	0.0025	0.9149
K-Nearest Neighbors	0.0133	0.5533

Berdasarkan hasil uji coba terhadap performa model dengan masing-masing algoritma, dapat disimpulkan bahwa model Random Forest memiliki kemampuan yang lebih baik dalam memprediksi kinerja akademik siswa. Keunggulan model Random Forest terletak pada kemampuannya untuk menangani data yang kompleks dan variasi yang tinggi di antara fitur-fitur yang ada.

Dari hasil prediksi yang diperoleh, analisis tambahan dilakukan untuk memahami faktor-faktor yang paling mempengaruhi kinerja siswa dalam ujian. Berdasarkan interpretasi feature importance dari model Random Forest, diketahui bahwa faktor-faktor seperti math score, reading score, dan parental level of education memiliki kontribusi yang signifikan terhadap prediksi nilai writing score. Analisis kontribusi masing-masing fitur terhadap hasil prediksi ditunjukkan dalam Tabel 4.

Tabel 4. Feature Importance Model Random Forest

Fitur	Importance Score
Reading Score	0.926167
Math Score	0.029541
Parental Level of Education	0.011973
Race/Ethnicity	0.011013
Gender	0.009651
Test Preparation Course	0.008322
Lunch	0.003333

Hasil analisis ini menunjukkan bahwa nilai matematika dan membaca, serta tingkat pendidikan orang tua adalah faktor-faktor utama yang mempengaruhi kinerja siswa dalam mata pelajaran menulis. Faktor-faktor ini memberikan wawasan berharga mengenai aspek-aspek yang perlu diperhatikan dalam upaya meningkatkan kinerja akademik siswa.

Model yang telah dibangun diuji menggunakan data pengujian yang telah dipisahkan sebelumnya. Evaluasi dilakukan dengan menggunakan metrik mean squared error (MSE) dan R-squared (R^2) untuk mengukur kinerja model dalam memprediksi nilai writing score siswa. Berdasarkan hasil pengujian, algoritma Random Forest menunjukkan performa yang lebih baik dibandingkan dengan algoritma K-Nearest Neighbors. Model Random Forest menghasilkan nilai MSE sebesar 0.0025 dan R-squared sebesar 0.9149, sedangkan K-Nearest Neighbors mencatat MSE sebesar 0.0133 dan R-squared sebesar 0.5533.

Temuan ini sejalan dengan teori yang dikemukakan oleh Yoga (Religia et al, 2021) mengenai keunggulan metode ensemble seperti Random Forest yang mampu mengurangi risiko overfitting melalui penggunaan banyak pohon keputusan (decision trees) dan agregasi hasil voting, sehingga memberikan prediksi yang lebih stabil. Selain itu, efektivitas Random Forest dalam menangani data dengan variabel input yang kompleks dan heterogen juga didukung oleh literatur machine learning yang dijelaskan oleh Hastie, Tibshirani, dan Friedman (Rahmaddeni et al, 2024), yang menyatakan bahwa metode ini sangat cocok digunakan pada data multivariat dengan interaksi yang sulit diidentifikasi secara manual.

Analisis feature importance pada model Random Forest menunjukkan bahwa variabel math score, reading score, serta tingkat pendidikan orang tua merupakan faktor yang paling signifikan dalam memengaruhi prediksi nilai writing score siswa. Hasil ini mendukung pandangan ekologi perkembangan Bronfenbrenner (Zulkifli, 2022), yang menegaskan pentingnya pengaruh lingkungan mikro, termasuk keluarga dan pendidikan orang tua, dalam membentuk perkembangan akademik anak. Hubungan erat antara nilai matematika dan membaca dengan prediksi nilai menulis juga mengonfirmasi hasil penelitian Li et al. (Cazarez & Martin, 2018) yang menyatakan bahwa terdapat transfer keterampilan lintas mata pelajaran, sehingga kemampuan belajar pada satu bidang dapat memperkuat performa akademik di bidang lain.

Dengan demikian, penelitian ini tidak hanya memperlihatkan keunggulan teknis dari algoritma Random Forest dalam memprediksi kinerja akademik siswa, tetapi juga memberikan landasan teoritis bahwa kualitas hasil belajar siswa dipengaruhi oleh kombinasi faktor internal (kemampuan dasar pada mata pelajaran terkait) maupun faktor eksternal (latar belakang pendidikan keluarga). Temuan ini dapat menjadi dasar bagi guru maupun pembuat kebijakan dalam merumuskan strategi intervensi pendidikan yang lebih tepat sasaran untuk meningkatkan kualitas proses belajar dan capaian akademik siswa secara menyeluruh.

Simpulan

Berdasarkan hasil dan pembahasan di atas, dapat disimpulkan bahwa algoritma Random Forest memiliki performa yang lebih baik dalam melakukan prediksi terhadap kinerja siswa dalam ujian dibandingkan dengan algoritma K-Nearest Neighbors. Hasil analisis feature importance menunjukkan bahwa nilai matematika dan membaca, serta tingkat pendidikan orang tua adalah faktor-faktor yang paling mempengaruhi prediksi nilai menulis siswa. Penelitian ini memberikan wawasan berharga mengenai penggunaan algoritma pembelajaran mesin untuk prediksi kinerja siswa dan dapat digunakan sebagai dasar untuk pengembangan model prediksi yang lebih baik di masa depan.

Pengujian terhadap model membuktikan bahwa prediksi terhadap nilai writing score siswa memungkinkan untuk dilakukan dengan analisis data yang ada. Hasil penelitian ini dapat dikembangkan lebih lanjut dengan mencoba algoritma lain dan optimasi parameter lebih lanjut. Hasil analisis diharapkan dapat digunakan sebagai bahan acuan dalam penelitian selanjutnya dan model dapat digunakan sebagai alat bantu dalam memprediksi kinerja akademik siswa. Uji coba terhadap model membuktikan bahwa prediksi terhadap nilai writing score siswa memungkinkan untuk dilakukan dengan analisis data yang ada. Hasil penelitian ini dapat dikembangkan lebih lanjut dengan mencoba algoritma lain dan optimasi parameter lebih lanjut. Hasil analisis diharapkan dapat digunakan sebagai bahan acuan dalam penelitian selanjutnya dan model dapat digunakan sebagai alat bantu dalam memprediksi kinerja akademik siswa.

Hasil penelitian ini memberikan dasar untuk pengembangan metode prediktif yang lebih efektif dan efisien dalam dunia pendidikan. Dengan memahami faktor-faktor utama yang mempengaruhi kinerja siswa, diharapkan dapat diambil langkah-langkah yang tepat untuk meningkatkan kualitas pendidikan dan hasil belajar siswa.

Daftar Pustaka

Adnan, A., Zohriah, A., & Mu'in, A. (2024). Evaluasi kinerja tenaga pendidik. *JIIP - Jurnal Ilmiah Ilmu Pendidikan*, *7*(2), 1463–1468. https://doi.org/10.54371/jiip.v7i2.3446

Ariyoga, D. (2022). *Perbandingan metode seleksi fitur filter, wrapper, dan embedded pada klasifikasi data nirs mangga menggunakan Random Forest dan Support Vector Machine (SVM)*.

Cazarez, R. L. U., & Martin, C. L. (2018). Neural networks for predicting student performance in online education. *IEEE Latin America Transactions*, *16*(7), 2053–2060. https://doi.org/10.1109/TLA.2018.8447376

Cumel, S., Zamri, D., & Rahmaddeni. (2022). Perbandingan metode data mining untuk prediksi banjir dengan algoritma Naïve Bayes dan KNN. *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian kepada Masyarakat*, 40–48. https://journal.irpi.or.id/index.php/sentimas/article/view/353

Efriadi, D., Rahmaddeni, R., Agustin, A., & Junadhi, J. (2022). Prediksi penambahan piutang iuran jaminan sosial ketenagakerjaan menggunakan algoritma K-Nearest Neighbor. *Edumatic: Jurnal Pendidikan Informatika*, *6*(1), 49–57. https://doi.org/10.29408/edumatic.v6i1.5255

Gori, T., Sunyoto, A., & Al Fatta, H. (2024). Preprocessing data dan klasifikasi untuk prediksi kinerja akademik siswa. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *11*(1), 215–224. https://doi.org/10.25126/jtiik.20241118074

Kumar, M. (2024). Utilizing Random Forest and XGBoost Data Mining Algorithms for Anticipating Students' Academic Performance. *International Journal of Modern Education and Computer Science*, *16*(2), 29-44, ISSN 2075-0161, https://doi.org/10.5815/ijmecs.2024.02.03

Putri, S. J., Attaqwa, Q., Pratama, A., & Rahmaddeni. (2022). Klasifikasi menentukan jadwal kerja data karyawan menggunakan algoritma C4.5 dan K-nearest Neighbor. *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian kepada Masyarakat*, 215–221. https://journal.irpi.or.id/index.php/sentimas

Rahmaddeni, S. K. M. K., Wulandari, S. K. M. K. D., Renova, M., Ramadhan, A. M. G., & Sari, R. (2024). *Machine learning*. Serasi Media Teknologi. https://books.google.co.id/books?id=owoOEQAAQBAJ

Rahmansyah, A., Dewi, O., Andini, P., Hastuti, T., Ningrum, P., & Suryana, M. E. (2018). Membandingkan pengaruh feature selection terhadap algoritma Naïve Bayes dan Support Vector Machine. *Seminar Nasional Aplikasi Teknologi Informasi*, 1907–5022.

Rajesh, P. (2021). Analysis of E-learner's Opinion Using Automated Sentiment Analysis in E-learning and Comparison with Naive Bayes Classification, Random Forest and K-Nearest Neighbour Algorithms. *Lecture Notes in Networks and Systems*, *248*, 265-277, ISSN 2367-3370, https://doi.org/10.1007/978-981-16-3153-5_30

Religia, Y., Nugroho, A., & Hadikristanto, W. (2021). Analisis perbandingan algoritma optimasi pada Random Forest untuk klasifikasi data bank marketing. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *5*(1), 187–192. https://doi.org/10.29207/resti.v5i1.2813

Sabita, S. A., & Yahfizham, Y. (2024). Penerapan algoritma klasifikasi nearest neighbor dalam mendeteksi penyakit diabetes. *Jurnal Bintang Pendidikan dan Bahasa*, *2*(1), 149–158. https://doi.org/10.59024/bhinneka.v2i1.645

Sabrina, J. A., & Mubayyinah, L. N. (2024). Optimalisasi pengambilan keputusan melalui analisis big data pada bidang kebijakan publik.

Saputra, E. P., Nurajizah, S., Maulidah, M., Hidayati, N., & Rahman, T. (2023). Komparasi machine learning berbasis PSO untuk prediksi tingkat keberhasilan belajar berbasis e-learning. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *10*(2), 321–328. https://doi.org/10.25126/jtiik.20231026469

Setio, B., & Prasetyaningrum, P. (2021). Penerapan data mining dalam mengelompokkan kunjungan wisatawan di Kota Yogyakarta menggunakan metode K-Means. *Jurnal Computer Science and Technology*, *1*(1), 27–32. https://doi.org/10.54840/jcstech.v1i1.9

Shan, K. (2025). A Study on Constraint-Related Fracture Toughness Prediction Based on Random Forest Algorithm and Data Enhancement Strategies. *Guti Lixue Xuebao Acta Mechanica Solida Sinica*, *46*(1), 105-116, ISSN 0254-7805, https://doi.org/10.19636/j.cnki.cjsm42-1250/o3.2024.044

Vural, M.S. (2025). Classification of the Heartbeats in Electrocardiograms with K-Nearest Neighbors Algorithm, Random Forests, and Support Vector Machines - A Pilot Study. *Lecture Notes in Networks and Systems*, *1202*, 177-184, ISSN 2367-3370, https://doi.org/10.1007/978-3-031-82143-1_20

Zhang, P. (2025). Predicting response to anti-VEGF therapy in neovascular age-related macular degeneration using random forest and SHAP algorithms. *Photodiagnosis and Photodynamic Therapy*, *53*, ISSN 1572-1000, https://doi.org/10.1016/j.pdpdt.2025.104635

Zulkifli, E. (2022). Pengaruh optimalisasi pembelajaran online, partisipasi mahasiswa dan gaya mengajar dosen terhadap motivasi belajar pada pembelajaran daring di tengah pandemi COVID-19 (Studi pada Kampus STIE Indonesia Jakarta).