Prediksi Performa Akademik Siswa Berdasarkan Aktivitas Harian Menggunakan Machine Learning

Adam Akbar Alamsyah   
*Sistem Informasi*  
*Universitas Komputer Indonesia*Bandung, Indonesia  
adam.10522006@mahasiswa.unikom.ac.id

Akhfa Shidqie Muttaqien  
*Sistem Informasi*  
*Universitas Komputer Indonesia*Bandung, Indonesia  
akhfa.10522028@mahasiswa.unikom.ac.id

Sandy Permana  
*Sistem Informasi*  
*Universitas Komputer Indonesia*Bandung, Indonesia  
sandy.10522013@mahasiswa.unikom.ac.id

Arya Nugraha  
*Sistem Informasi*  
*Universitas Komputer Indonesia*Bandung, Indonesia  
arya.10522016@mahasiswa.unikom.ac.id

*Abstract*— Performa akademik siswa dipengaruhi oleh banyak faktor, salah satunya adalah kebiasaan dan aktivitas harian yang dilakukan secara konsisten. Studi ini bertujuan untuk menganalisis keterkaitan antara aktivitas harian siswa dengan performa akademiknya menggunakan pendekatan machine learning. Dataset yang digunakan mencakup data aktivitas siswa seperti durasi belajar, waktu tidur, penggunaan media sosial, dan kebiasaan olahraga, serta nilai akademik sebagai variabel target. Penelitian ini membandingkan performa model klasifikasi berbasis *Decision Tree*, *Random Forest*, *Linear Regression*, *Neural Network* dan *K-Nearest Neighbors* (*kNN*) untuk mengidentifikasi pola-pola signifikan yang mempengaruhi pencapaian akademik siswa. Evaluasi dilakukan dengan matrik regresi, *Mean Absolute Error* (*MAE*), *Mean Squared Error* (*MSE*), *Root Mean Squared* *Error* (*RMS*E), *Mean Absolute Percentage Error* (*MAPE*), R² Score (*Coefficient of Determination*). Hasil menunjukkan bahwa model Linear Regression memberikan performa terbaik dengan nilai *error* terkecil di semua metrik, mengindikasikan bahwa aktivitas harian memiliki korelasi kuat terhadap hasil akademik. Temuan ini diharapkan dapat menjadi referensi bagi institusi pendidikan dalam merancang intervensi berbasis data guna meningkatkan prestasi belajar siswa.

*Keywords— Performa akademik, aktivitas harian, klasifikasi, machine learning, Linear Regression.*

# Introduction

Setiap siswa ingin sukses, begitu juga dengan pemerintah yang berusaha meningkatkan keberhasilan siswa untuk menjamin masa depan yang cerah bagi negara mereka. Jadi, peningkatan prestasi akademik siswa sangat penting dengan membawa sistem pendidikan ke tingkat yang lebih baik. Keberhasilan siswa berubah sesuai dengan beberapa kondisi.[1] Keberhasilan akademik siswa sendiri dapat berubah-ubah tergantung pada berbagai kondisi yang mempengaruhinya. Faktor internal dan eksternal merupakan aspek yang memiliki dampak signifikan terhadap kinerja akademik siswa. Faktor internal mencakup aspek seperti kesehatan, kemampuan intelektual, bakat dan minat, motivasi, serta kesiapan siswa dalam belajar.[2] Sementara itu, faktor eksternal mencakup lingkungan keluarga, kondisi sosial-ekonomi, dukungan dari sekolah, serta kebiasaan dan aktivitas harian siswa di luar kegiatan akademik formal.

Faktor internal dan eksternal yang ada pada setiap orang dapat memengaruhi prestasi akademik siswa.[3] Faktor-faktor yang memengaruhi prestasi akademik tidak hanya bersifat internal, seperti kemampuan intelektual, minat, dan motivasi belajar, tetapi juga bersifat eksternal, seperti lingkungan belajar, dukungan orang tua, serta pola penggunaan waktu luang.[4]

Seiring berkembangnya teknologi informasi dan ketersediaan data yang semakin besar di sektor pendidikan, pendekatan berbasis *machine learning* mulai banyak diterapkan untuk mendukung proses pengambilan keputusan yang lebih akurat dan berbasis data. *Machine learning* mampu mengidentifikasi pola-pola tersembunyi dari data historis dan memprediksi performa siswa berdasarkan atribut-atribut perilaku yang tersedia. Penelitian ini bertujuan untuk menganalisis hubungan antara aktivitas harian siswa dengan performa akademik menggunakan pendekatan prediksi regresi berbasis *machine learning*. Beberapa algoritma yang digunakan dalam penelitian ini meliputi *Decision Tree*, *Random Forest*, *Linear Regression*, *Neural Network* dan *kNN* dengan data yang dikumpulkan dari berbagai atribut kebiasaan siswa serta hasil akademiknya. Evaluasi dilakukan menggunakan metrik seperti regresi, untuk menilai efektivitas masing-masing model dalam melakukan prediksi. Melalui pendekatan ini, penelitian diharapkan dapat memberikan kontribusi nyata terhadap pengembangan sistem prediksi akademik yang dapat membantu institusi pendidikan, guru, maupun orang tua dalam membentuk kebiasaan belajar yang lebih baik bagi siswa.

Penelitian nasional oleh Simbolon et al.[5] yang berjudul “*Prediksi Performa Akademik Siswa Berdasarkan Kehadiran dan Aktivitas E-Learning Menggunakan Algoritma Linear Regression”*. Studi ini berfokus pada pengembangan model prediktif yang mampu memperkirakan nilai performa

akademik siswa secara kuantitatif. Evaluasi model dilakukan menggunakan *metrik regresi* seperti *Mean Absolute Error* (*MAE*), *Root Mean Squared Error* (*RMSE*), dan *R-squared* (*R²*) untuk mengukur tingkat akurasi dan kemampuan model dalam menjelaskan variansi data. Hasil yang diperoleh menunjukkan bahwa model memiliki performa yang cukup baik. Penelitian ini menekankan pentingnya pemanfaatan data log dan interaksi digital dalam sistem manajemen pembelajaran (*LMS*) sebagai sumber informasi yang berharga untuk mendukung pengambilan keputusan di bidang pendidikan. Pendekatan regresi seperti ini masih belum banyak diterapkan secara luas di lingkungan akademik digital di Indonesia, sehingga memberikan kontribusi terhadap pengembangan sistem analitik pendidikan yang lebih cerdas dan adaptif.

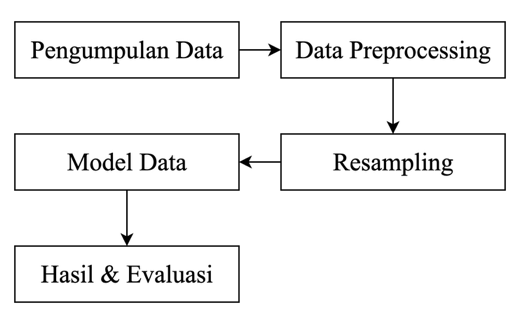
Penelitian lain oleh Sukmaningtyas et al [6] yang berjudul *“Penerapan Predictive Analytics untuk Analisis Faktor-faktor yang Mempengaruhi Performa Akademik Siswa”* memiliki tujuan yang lebih terfokus dalam mengevaluasi pengaruh fitur-fitur perilaku siswa seperti *study hours, sleep hours, previous scores,* dan *partisipasi ekstrakurikuler* terhadap capaian akademik dengan menerapkan pendekatan *machine learning* berbasis *XGBoost, Random Forest, SVR,* dan *KNN*. Hasilnya menunjukkan bahwa nilai sebelumnya memiliki korelasi sangat tinggi dengan performa akademik (r = 0,92), dan algoritma *XGBoost* menjadi model terbaik. Kontribusi penting dari penelitian ini terletak pada pemanfaatan kombinasi faktor akademik dan non-akademik dalam satu model prediksi terpadu, yang memperluas perspektif dari penelitian-penelitian sebelumnya yang cenderung hanya fokus pada aspek kognitif atau data nilai.

Adapun penelitian oleh Rusdiana Ekawati et al [7] yang berjudul *“Prediksi Ketuntasan Belajar Siswa Menggunakan Naive Bayes dengan Integrasi Data Akademik, Absensi, dan Partisipasi Ekstrakurikuler”* secara spesifik dirancang untuk mengidentifikasi sejauh mana kombinasi nilai ujian, absensi, dan partisipasi siswa dalam kegiatan ekstrakurikuler dapat digunakan untuk memprediksi ketuntasan belajar dengan algoritma *Naive Bayes*. Penelitian ini menunjukkan bahwa model tersebut mampu mencapai akurasi 85,2% dengan *F1-score* sebesar 87%, di mana fitur nilai ujian memiliki pengaruh dominan terhadap prediksi. Kontribusi yang ditawarkan studi ini adalah integrasi faktor afektif dan kehadiran sebagai fitur penting dalam model prediktif berbasis statistik probabilistik, yang membuka ruang eksplorasi terhadap integrasi data akademik formal dan aktivitas siswa yang lebih luas dalam sistem evaluasi pendidikan.

Berbeda dengan penelitian sebelumnya yang umumnya menitikberatkan pada data akademik konvensional seperti kehadiran, nilai ujian, atau interaksi dalam sistem *e-learning*, penelitian ini menawarkan pendekatan yang lebih komprehensif dengan mengintegrasikan variabel-variabel aktivitas harian siswa di luar konteks akademik formal, seperti durasi belajar, waktu tidur, penggunaan media sosial, dan kebiasaan berolahraga. Pendekatan ini diharapkan dapat memberikan kontribusi baru dalam pengembangan model prediksi performa akademik yang lebih kontekstual, personal, dan relevan terhadap dinamika kehidupan siswa masa kini.

# Metode Penelitian

Metodologi penelitian ini dilaksanakan melalui tahapan sistematis yang terdiri dari proses pengumpulan data, pembersihan dan persiapan data (*data preprocessing*), seleksi fitur yang relevan, pembagian data (*resampling*), pembangunan model (*modeling*), serta evaluasi performa model regresi. Diagram alur penelitian ditunjukkan pada Gambar 1.



Gambar . Alur proses penelitian

## Pengumpulan Data

Tahap ini melibatkan pengumpulan data yang relevan dari sumber yang sudah ditentukan[8]. Peneliti menggunakan sumber data terbuka, atau *open source*, yang dapat diakses melalui situs web Kaggle.com. Kaggle adalah *platform* terkemuka yang memungkinkan akses ke berbagai jenis dataset untuk penelitian, pembelajaran mesin, dan proyek data science lainnya [9]. Dataset yang digunakan dalam penelitian ini berjudul “*Student Habits vs Academic Performance*” yang dikembangkan oleh Jayaantanaath. Dataset ini berisi 1.000 data sintetik siswa dengan berbagai atribut kebiasaan harian seperti durasi belajar per hari, waktu tidur, durasi penggunaan media sosial dan Netflix, kualitas diet, kondisi kesehatan mental, frekuensi olahraga, partisipasi ekstrakurikuler, status pekerjaan paruh waktu, serta nilai ujian akhir sebagai variabel target.

Pemilihan dataset ini didasarkan pada keberagaman atribut yang mencakup aspek akademik maupun non-akademik, yang dinilai relevan untuk mengkaji keterkaitan antara pola kebiasaan siswa dengan pencapaian akademiknya. Dataset tersedia dalam format *CSV,* sehingga memudahkan proses analisis data dan pembangunan model prediktif secara langsung.

## Data Preprocessing

Data *Preprocessing* bertujuan untuk mengubah data mentah menjadi data yang berkualitas sehingga data layak untuk diolah pada tahapan selanjutnya [10]. Proses *preprocessing* meliputi:

## Pembersihan Data (*Data Cleaning*)

## Data *cleaning* atau pembersihan data merupakan proses yang digunakan untuk mendeteksi, memperbaiki ataupun menghapus dataset, tabel, dan database yang korup atau tidak akurat [11]. Nilai yang hilang diimputasi menggunakan rata-rata (*mean*) atau median sesuai karakteristik variabel.

## Normalisasi Data

## Standarisasi data menggunakan metode *Z-Score Normalization,* di mana setiap nilai dikurangi dengan rata-rata variabelnya, lalu dibagi dengan standar deviasi variabel tersebut. Pendekatan ini digunakan agar setiap variabel memiliki skala yang setara (rata-rata 0 dan standar deviasi 1), sehingga algoritma *K-Means* tidak bias terhadap variabel yang memiliki skala lebih besar. Hal ini penting untuk menghasilkan klasterisasi yang lebih akurat dan adil antar variabel [12]. Zscore normalization dapat dihitung menggunakan rumus berikut [13] :

x: nilai asli

x̅: rata-rata sampel/populasi

σ: deviasi standar sampel/populasi

## Encoding Variable Kategorikal

Langkah encoding variabel kategorikal dianggap salah satu langkah paling penting dalam *preprocessing* data, karena dalam proses ini data yang tidak memiliki nilai numerik akan diubah menjadi nilai yang spesifik. Hasilnya, seluruh data kategorikal akan memiliki nilai baru yang direpresentasikan dengan nomor.[14]

## Seleksi Fitur

Seleksi data adalah suatu proses di mana suatu subset atau bagian dari data yang ada dipilih untuk dianalisis, sementara yang lain diabaikan. Tujuan dari seleksi data adalah untuk memfokuskan perhatian pada informasi yang paling relevan atau signifikan, mengurangi kompleksitas, meningkatkan efisiensi analisis, dan mengurangi beban komputasional. [15]

## Resampling

## Teknik resampling merupakan cara yang paling populer untuk mengatasi masalah ketidakseimbangan kelas pada dataset. Terdapat tiga pendekatan utama untuk mengatasi *underclassification*, yaitu dengan menambah jumlah sampel pada kelas minoritas (*oversampling*), mengurangi jumlah sampel pada kelas mayoritas (*undersampling*), atau menggunakan metode *hybrid* yang merupakan kombinasi keduanya [16]. Karena adanya ketidakseimbangan yang signifikan dalam dataset, metode undersampling dan oversampling diterapkan pada beberapa model *machine learning* yang digunakan dalam sistem ini. Undersampling dilakukan dengan mengurangi jumlah sampel pada kelas mayoritas, sedangkan oversampling dilakukan dengan menambah jumlah sampel pada kelas minoritas dengan menggandakan beberapa data yang ada [17].

## Model Data

Dalam penelitian ini, digunakan sejumlah algoritma machine learning sebagai model untuk membandingkan kinerja prediksi berdasarkan data aktivitas harian siswa. Pemilihan algoritma dilakukan dengan mempertimbangkan kemampuannya dalam mengolah data bertipe numerik maupun kategorikal, serta efektivitasnya dalam menyelesaikan permasalahan regresi dan klasifikasi. Berikut adalah penjelasan mengenai algoritma-algoritma yang digunakan dalam studi ini:

* *Linear Regression*

Regresi linear adalah alat statistik yang digunakan untuk menentukan pengaruh satu atau beberapa variabel terhadap suatu variabel lainnya. Variabel yang mempengaruhi sering disebut sebagai variabel independen atau variabel penjelas, sedangkan variabel yang dipengaruhi sering disebut sebagai variabel dependen atau variabel terikat. [18]

* *Decision Tree*

*Decision Tree* adalah algoritma populer dan sangat efektif dengan melakukan pengklasifikasian dan prediksi. *Algoritma Decision Tree* dapat merepresentasikan ketentuan dari banyaknya fakta ke dalam bentuk pohon keputusan. Pohon keputusan adalah struktur yang membagi sejumlah besar data menjadi sejumlah kecil data. Atribut kelas berfungsi sebagai representasi untuk simpul daun pohon keputusan [19].

* *Random Forest*

*Random Forest* adalah algoritma *machine learning* yang menggabungkan keluaran dari beberapa pohon keputusan (*decision trees*) untuk menghasilkan satu hasil prediksi. [20]

* *Neural Network*

*Neural Network* merupakan model yang terdiri dari beberapa layer yang terdiri dari input layer, hidden layer, dan output layer. Model ini dapat memproses data tidak terstruktur dengan dimensi tinggi seperti gambar, audio, dll. [21]

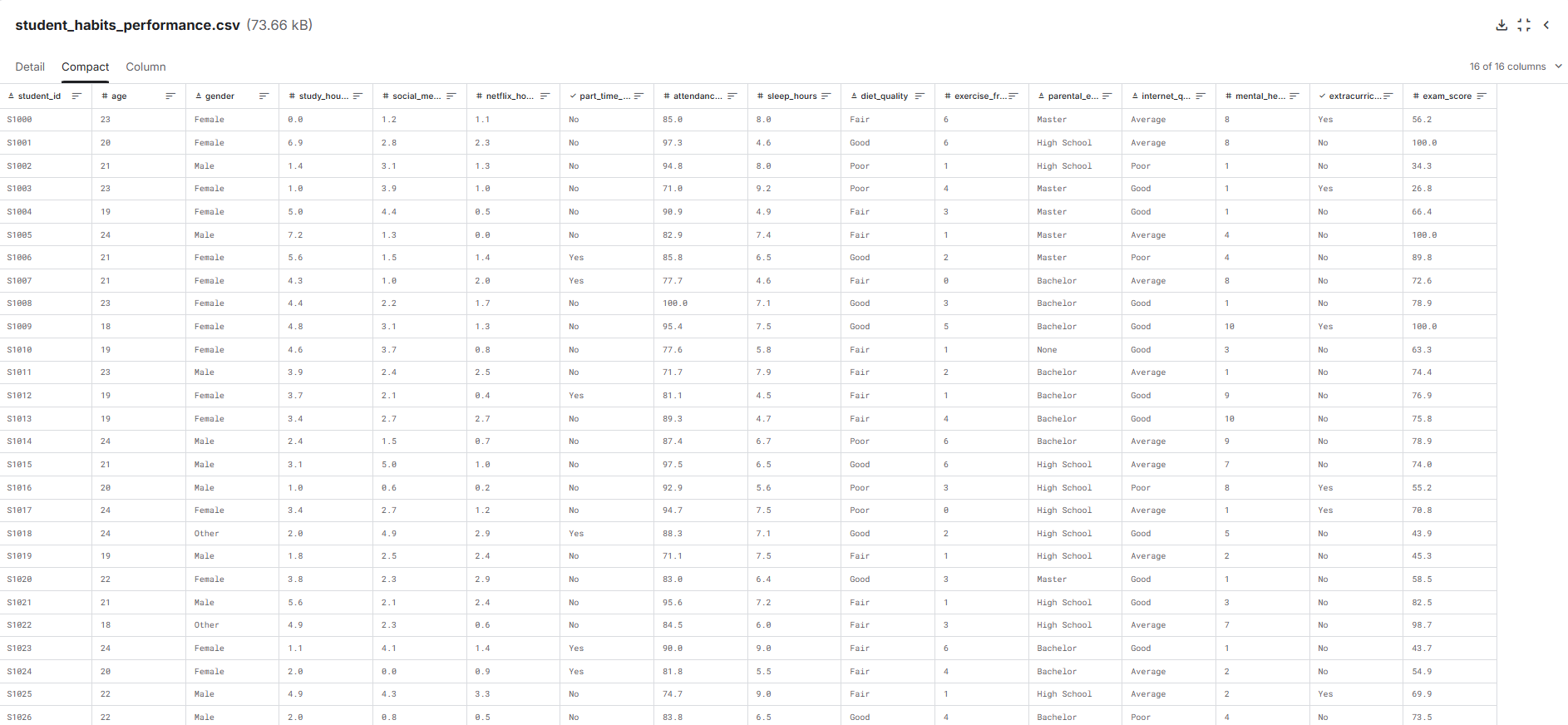
* *k-Nearest Neighbors* (*kNN*)

*Algoritma k-nearest neighbor* (*kNN*) merupakan *algoritma supervised* *machine learning* yang digunakan untuk klasifikasi. *kNN* memprediksi klasifikasi dari data tanpa label dengan mempertimbangkan fitur dan label dari training data.[22]

# Hasil dan Pembahasan

A. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini diambil dari platform Kaggle, dengan judul *Student Habits vs Academic Performance*, yang tersedia pada tautan berikut: <https://www.kaggle.com/datasets/jayaantanaath/student-habits-vs-academic-performance>. Dataset ini berasal dari salah satu sekolah menengah atas (SMA) dan terdiri dari 1000 baris data (*instances*) serta 16 kolom (*fitur*). Seluruh data telah tersaji dengan baik tanpa nilai yang hilang (n*o missing values*) dan mencakup kombinasi tipe data numerik dan kategori (*categorical*). Dalam penelitian ini, digunakan satu fitur sebagai *variabel* target, yaitu *exam\_score*, yang merepresentasikan nilai ujian akhir siswa. Karena target berupa nilai kontinu, maka pendekatan yang digunakan adalah *supervised learning* dengan model regresi. Dataset ini juga memuat satu atribut meta yang tidak dilibatkan dalam proses pelatihan model. Struktur data yang rapi dan informatif menjadikan dataset ini sangat mendukung untuk membangun model prediksi performa akademik berbasis analisis data.

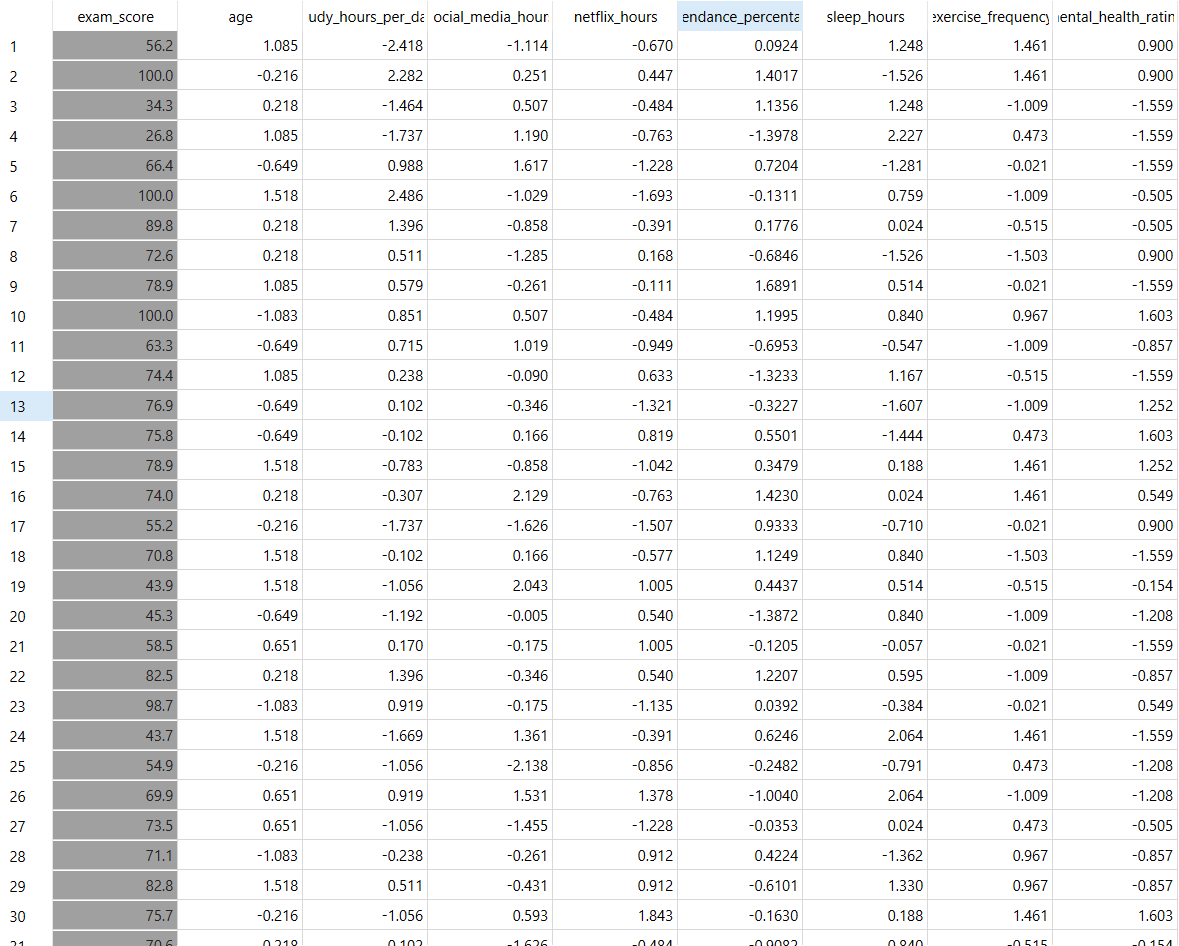


Gambar . Dataset kaggle

B. Data *Preprocessing*

*Preprocessing* merupakan tahap awal yang sangat krusial dalam proses pembangunan model prediksi regresi. Tahap ini bertujuan untuk mempersiapkan data agar dapat diproses secara optimal oleh algoritma pembelajaran mesin. Pada penelitian ini, proses preprocessing dilakukan menggunakan perangkat lunak *Orange*, yang meliputi dua langkah utama, yaitu pemilihan kolom (*select columns*) dan normalisasi data (*normalize features*). Langkah pertama yang dilakukan adalah seleksi kolom, di mana fitur-fitur yang dianggap relevan untuk proses prediksi dipilih dan dikelompokkan sebagai variabel independen (*features*). Sementara itu, satu variabel numerik yang menjadi sasaran prediksi ditetapkan sebagai variabel dependen (*target*). Atribut-atribut lain yang bersifat non-numerik atau dianggap tidak berkontribusi signifikan dalam proses prediksi dikelompokkan sebagai ignored. Pemilihan fitur ini bertujuan untuk meminimalkan kompleksitas model serta meningkatkan performa akurasi prediksi. Langkah selanjutnya adalah normalisasi data. Proses ini dilakukan untuk menyamakan skala nilai antar fitur agar model tidak berat sebelah terhadap fitur yang memiliki nilai absolut besar. Normalisasi yang digunakan dalam penelitian ini adalah metode standardization, yaitu mengubah distribusi setiap fitur agar memiliki nilai rata-rata (μ) sebesar 0 dan standar deviasi (σ) sebesar 1. Pendekatan ini dinilai efektif terutama untuk algoritma regresi, yang sensitif terhadap perbedaan skala data.

Dengan tahapan *preprocessing* ini, data telah disiapkan dalam bentuk yang optimal untuk dilanjutkan ke tahap pemodelan regresi guna memperoleh hasil prediksi yang akurat dan dapat diandalkan.



Gambar . Hasil *preprocessing*

C. *Resampling*

Setelah data melalui *preprocessing*, tahap selanjutnya adalah sampling, yang dilakukan untuk membagi data menjadi dua bagian utama: data pelatihan dan data pengujian. Mengingat dataset yang digunakan terdiri dari 1.000 data, maka teknik yang diterapkan adalah train-test split dengan proporsi 80:20. Artinya, sebanyak 80% data (800 data) digunakan untuk melatih model *machine learning*, sementara 20% sisanya (200 data) digunakan untuk mengukur performa model melalui proses pengujian.

Proses sampling ini dilakukan menggunakan fitur Data Sampler pada perangkat lunak *Orange*, dengan konfigurasi *replicable* (*deterministic*) sampling untuk memastikan konsistensi hasil yang dapat direproduksi. Pendekatan ini memungkinkan evaluasi model secara lebih objektif dan mencegah terjadinya overfitting akibat pelatihan pada keseluruhan data.

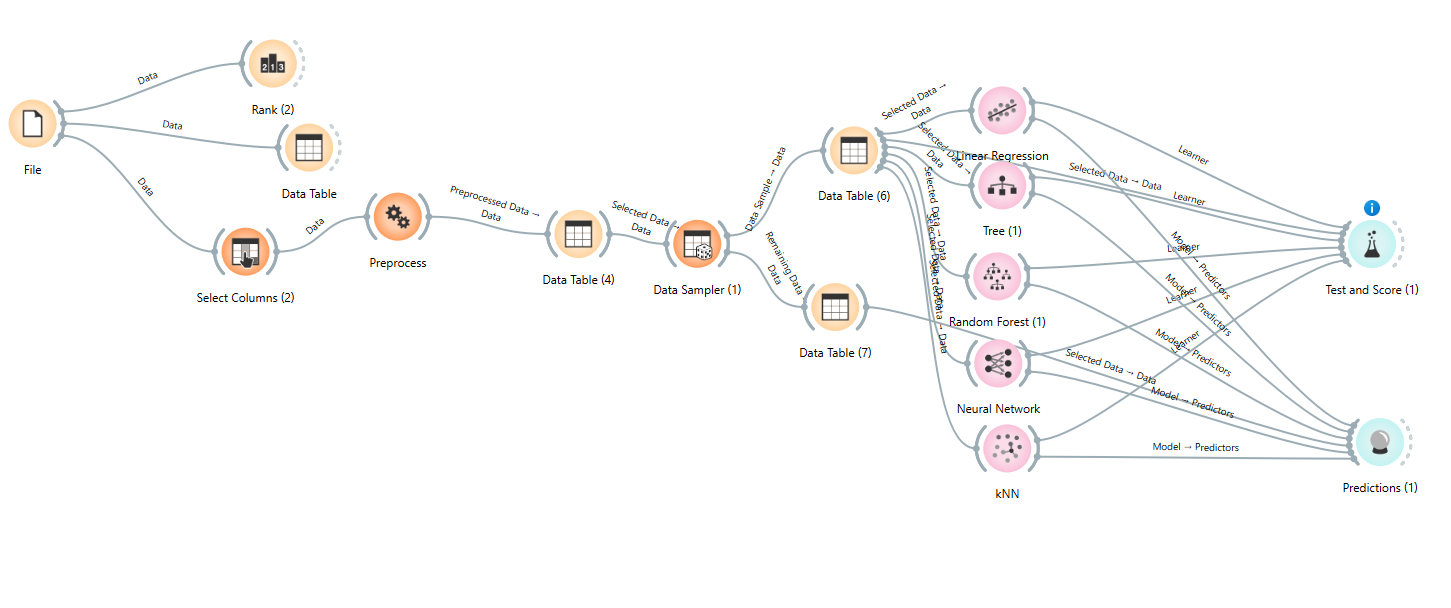
D. Model Data

Setelah proses pembagian data selesai melalui teknik sampling, langkah selanjutnya adalah pemodelan menggunakan berbagai algoritma *machine learning*. Tujuannya adalah untuk membandingkan performa masing-masing model dalam memprediksi nilai target, yaitu *exam\_score.* Dalam eksperimen ini, digunakan lima algoritma pembelajaran mesin yang berbeda, yaitu:

* *Linear Regression*
* *Decision Tree*
* *Random Forest*
* *Neural Network*
* *k-Nearest Neighbors (kNN)*

Kelima model tersebut dilatih menggunakan data hasil pelatihan (80%) dan kemudian diuji performanya menggunakan data pengujian (20%) yang telah dipisahkan sebelumnya. Proses evaluasi dilakukan dengan menggunakan *widget Test and Score*, yang menghasilkan metrik evaluasi seperti *Mean Squared Error* (*MSE*) yang berarti rata-rata dari kuadrat selisih antara nilai prediksi dan nilai aktual, *Root* *Mean Square Error* (*RMSE*) yang berarti akar kuadrat dari *MSE* untuk memberikan gambaran yang lebih realistis karena kembali ke satuan asli dari target, *Mean Absolute Error* (*MAE*) yang berarti rata-rata nilai absolut dari selisih antara prediksi dan aktual, *Mean Absolute Percentage Error* (*MAPE*) yang berarti rata-rata persentase kesalahan *absolut*, dan *R-squared* (*R²*) yang berarti seberapa baik model menjelaskan variasi dalam data dimana semakin mendekati 1 maka semakin baik.

Hasil dari setiap prediksi juga ditampilkan melalui *widget Predictions*, yang memungkinkan visualisasi nilai prediksi dibandingkan dengan nilai aktual. Proses ini penting untuk mengidentifikasi model mana yang paling akurat dan sesuai dengan karakteristik data yang dimiliki.



Gambar . *Workflow Orange*

E. Hasil & Evaluasi Model *Machine Learning*

Evaluasi kinerja model *machine learning* dilakukan terhadap lima algoritma yang berbeda, yaitu *Linear Regression*, *Decision Tree*, *Random Forest*, *Neural Network*, dan *k-Nearest Neighbors* (*kNN*). Masing-masing model dievaluasi menggunakan lima metrik utama, yaitu *Mean Squared Error* (*MSE*), *Root Mean Squared Error* (*RMSE*), *Mean Absolute Error* (*MAE*), *Mean Absolute Percentage Error* (*MAPE*), dan *Coefficient of Determination* (*R²*).

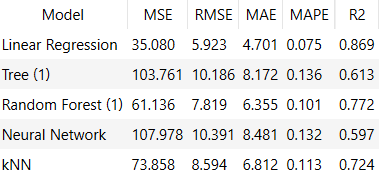
Hasil pengujian menunjukkan bahwa *Linear Regression* memberikan performa terbaik di antara kelima model yang diuji. Model ini menghasilkan nilai *MSE* sebesar 35.080, *RMSE* sebesar 5.923, *MAE* sebesar 4.701, dan *MAPE* sebesar 0.075, serta nilai *R²* tertinggi sebesar 0.869. Hal ini menunjukkan bahwa model *Linear Regression* mampu memprediksi dengan akurasi tinggi dan kesalahan yang relatif rendah dibandingkan model lainnya.

Model *Random Forest* menunjukkan kinerja yang cukup baik, dengan nilai *MSE* sebesar 61.136, *RMSE* sebesar 7.819, *MAE* sebesar 6.355, dan *MAPE* sebesar 0.101, serta nilai *R²* sebesar 0.772. Model ini mampu menjelaskan sekitar 77,2% variasi dalam data target, meskipun masih memiliki tingkat kesalahan prediksi yang lebih tinggi dibandingkan *Linear Regression.*

Sementara itu, model *k-Nearest* *Neighbors* (*kNN*) mencatat nilai *MSE* sebesar 73.858, *RMSE* sebesar 8.594, *MAE* sebesar 6.812, *MAPE* sebesar 0.113, dan *R²* sebesar 0.724. Meskipun hasilnya tidak sebaik *Random Forest*, model *kNN* masih dapat dipertimbangkan karena memberikan akurasi yang cukup baik.

Sebaliknya, model *Decision Tree* dan *Neural Network* menunjukkan performa yang kurang optimal. Model *Decision Tree* menghasilkan nilai *MSE* sebesar 103.761, *MAE* sebesar 8.172, dan R² sebesar 0.613, yang menunjukkan bahwa model ini tidak mampu menangkap pola kompleks dengan baik. Sementara itu, *Neural Network* mencatat kinerja terburuk dengan nilai *MSE* sebesar 107.978, *RMSE* sebesar 10.391, *MAE* sebesar 8.481, *MAPE* sebesar 0.132, dan nilai *R²* terendah sebesar 0.597.

Berdasarkan evaluasi tersebut, dapat disimpulkan bahwa model *Linear Regression* merupakan model paling optimal untuk digunakan dalam studi ini, baik dari sisi kesalahan prediksi yang rendah maupun dari kemampuan menjelaskan variabilitas data target. *Model Random Forest* dan *kNN* juga dapat dipertimbangkan sebagai alternatif dengan melakukan tuning parameter lebih lanjut untuk meningkatkan performa. Sebaliknya, model *Decision Tree* dan *Neural Network* memerlukan perbaikan atau penyesuaian sebelum dapat diterapkan secara efektif.



Gambar . Perbandingan performa model *machine learning*

# Kesimpulan

Berdasarkan hasil analisis untuk membangun model prediksi performa akademik siswa berdasarkan kebiasaan belajar menggunakan pendekatan supervised learning dengan algoritma regresi. Dataset yang digunakan diambil dari platform Kaggle dengan judul *Student Habits vs Academic Performance*, yang berisi data terstruktur sebanyak 1000 baris dan 16 kolom fitur, tanpa adanya nilai yang hilang. Variabel target yang diprediksi adalah *exam\_score*, yang merupakan nilai ujian akhir siswa.

Tahapan preprocessing dilakukan dengan pemilihan kolom yang relevan serta normalisasi data menggunakan metode standardization untuk memastikan kesetaraan skala antar fitur. Selanjutnya, dilakukan pembagian data menggunakan teknik *train-test* *split* dengan proporsi 80:20

untuk memastikan evaluasi model yang objektif dan menghindari *overfitting.*

Lima algoritma *machine learning* diterapkan dan dievaluasi, yaitu *Linear Regression*, *Decision Tree*, *Random Forest*, *Neural Network*, dan *k-Nearest Neighbors* (*kNN*). Evaluasi performa dilakukan dengan menggunakan metrik *MSE*, *RMSE*, *MAE*, *MAPE*, dan *R²*.

Hasil evaluasi menunjukkan bahwa *Linear Regression* merupakan model paling unggul dengan nilai error paling rendah dan nilai *R²* tertinggi (0.869), yang mengindikasikan kemampuan model dalam menjelaskan variasi data target secara signifikan. *Random Forest* dan *kNN* juga menunjukkan performa yang baik dan dapat dijadikan alternatif dengan penyempurnaan lebih lanjut melalui tuning parameter. Sementara itu, model *Decision Tree* dan *Neural Network* menunjukkan hasil yang kurang optimal dalam konteks dataset ini dan memerlukan pengembangan atau penyesuaian lebih lanjut.

Dengan demikian, model *Linear Regression* direkomendasikan sebagai pendekatan paling efektif dalam prediksi performa akademik siswa berbasis data kebiasaan belajar yang tersedia, khususnya dalam konteks dataset ini.

# Reference

[1] A. Rahman, “Klasifikasi Performa Akademik Siswa Menggunakan Metode Decision Tree dan Naive Bayes,” *Jurnal SAINTEKOM*, vol. 13, no. 1, 2023, doi: 10.33020/saintekom.v13i1.349.

[2] T. Gori, A. Sunyoto, and H. Al Fatta, “Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 1, 2024, doi: 10.25126/jtiik.20241118074.

[3] A. Jalil, A. Homaidi, and Z. Fatah, “Implementasi Algoritma Support Vector Machine Untuk Klasifikasi Status Stunting Pada Balita,” *G-Tech: Jurnal Teknologi Terapan*, vol. 8, no. 3, pp. 2070–2079, Jul. 2024, doi: 10.33379/gtech.v8i3.4811.

[4] J. M. C. Kaimarehe and Marsofiyati, “PENGARUH FAKTOR MOTIVASI BELAJAR TERHADAP PRESTASI AKADEMIK SISWA,” *Cendikia: Jurnal Pendidikan dan Pengajaran*, vol. 2, no. 6, pp. 221–227, Jun. 2024.

[5] Nanda Mulyani Pratiwi, A Julizar, Lilis Gusliyah, Miftahir Rizqa, and Mhd Khairi Azhar, “Pengaruh Penggunaan Media Sosial terhadap Hasil Belajar Siswa Kelas IX SMP Negeri 2 Batang Peranap,” *Jurnal Nakula : Pusat Ilmu Pendidikan, Bahasa dan Ilmu Sosial*, vol. 3, no. 5, pp. 290–303, Jun. 2025, doi: 10.61132/nakula.v3i5.2220.

[6] I. Simbolon, P. Aditya, and E. B. Purba, “Prediksi Performa Akademik Siswa Berdasarkan Kehadiran dan Aktivitas E-Learning Menggunakan Algoritma Decision Tree,” *Journal of Artificial Intelligence and Digital Business (RIGGS)*, vol. 4, no. 2, pp. 4899–4910, 2025, doi: 10.31004/riggs.v4i2.1352.

[7] Yanuarini Nur Sukmaningtyas, R. Makhfuddin Akbar, and G. Rohma Utami Asyafiiyah, “Penerapan Predictive Analytics untuk Analisis Faktor-faktor yang Mempengaruhi Performa Akademik Siswa,” *Arcitech: Journal of Computer Science and Artificial Intelligence*, vol. 4, no. 2, pp. 127–145, Dec. 2024, doi: 10.29240/arcitech.v4i2.12048.

[8] R. Rusdiana Ekawati, V. Atina, and J. Maulidar, “Prediksi Ketuntasan Belajar Siswa Menggunakan Naive Bayes dengan Integrasi Data Akademik, Absensi, dan Partisipasi Ekstrakurikuler,” *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 5, no. 4, pp. 1161–1173, Apr. 2025, doi: 10.52436/1.jpti.752.

[9] J. Kuswanto, H. Lukmanul, A. Info, and K. Kunci, “Penerapan Algoritma Random Forest untuk memprediksi Performa Akademik Mahasiswa,” *DECODE: Jurnal Pendidikan Teknologi Informasi*, vol. 5, no. 1, pp. 262–270, 2025, doi: 10.51454/decode.v5i1.1103.

[10] F. Alghifari and D. Juardi, “PENERAPAN DATA MINING PADA PENJUALAN MAKANAN DAN MINUMAN MENGGUNAKAN METODE ALGORITMA NAÏVE BAYES,” *JURNAL ILMIAH INFORMATIKA*, vol. 9, no. 02, 2021, doi: 10.33884/jif.v9i02.3755.

[11] N. P. A. Widiari, I. M. A. D. Suarjaya, and D. P. Githa, “Teknik Data Cleaning Menggunakan Snowflake untuk Studi Kasus Objek Pariwisata di Bali,” *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, 2020, doi: 10.24843/jim.2020.v08.i02.p07.

[12] M. D. Sianturi, M. C. Lubis, S. D. Payana, A. N. Putri, and H. R. A. Panjaitan, “Analisis Pengelompokan Jenis Kejahatan di Sumatera Utara Berdasarkan Pola Kejadian Tahunan Menggunakan Algoritma K-Means Clustering,” *AKSIOMA: Jurnal Sains Ekonomi dan Edukasi*, vol. 2, no. 6, pp. 1308–1324, 2025, [Online]. Available: https://doi.org/10.62335

[13] R. G. Whendasmoro and J. Joseph, “Analisis Penerapan Normalisasi Data Dengan Menggunakan Z-Score Pada Kinerja Algoritma K-NN,” *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 4, 2022, doi: 10.30865/jurikom.v9i4.4526.

[14] F. Bolikulov, R. Nasimov, A. Rashidov, F. Akhmedov, and Y. I. Cho, “Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms,” *Mathematics*, vol. 12, no. 16, Aug. 2024, doi: 10.3390/math12162553.

[15] R. Hadi *et al.*, “Implementasi Metode Normalisasi dan Seleksi Fitur dalam Optimasi Algoritma K-Nearest Neighbor (KNN) untuk Klasifikasi Data Bank,” *Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)*, vol. 7, no. 5, 2024.

[16] Sindrawati, D. Syaripudin, and A. Walad, “TEKNIK RESAMPLING UNTUK MENINGKATKAN NILAI AKURASI ALGORITMA RANDOM FOREST PADA DATA PREDIKSI KECACATAN PERANGKAT LUNAK,” *JICN: Jurnal Intelek dan Cendikiawan Nusantara*, vol. 1, no. 3, 2024, [Online]. Available: https://jicnusantara.com/index.php/jicn

[17] C. L. Udeze, I. E. Eteng, and A. E. Ibor, “Application of Machine Learning and Resampling Techniques to Credit Card Fraud Detection,” *Journal of the Nigerian Society of Physical Sciences*, vol. 4, no. 3, 2022, doi: 10.46481/jnsps.2022.769.

[18] H. Atthifa Ryantika, M. Parida, H. Afandi, S. Hanika Lubis, A. Afandi, and D. Febry Nugroho, “LINEAR REGRESSION METHOD APPLICATION TO PREDICT CIMORY MILK SALES,” *International Journal of Information System and Computer Science) IJISCS*, vol. 7, no. 1, 2022, [Online]. Available: https://ojs.stmikpringsewu.ac.id/index.php/ijiscs

[19] D. Septhya, K. Rahayu, S. Rabbani, V. Fitria, Y. Irawan, and R. Hayami, “Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, pp. 15–19, Apr. 2023.

[20] S. Wijaya and F. Fauziah, “Analysis of the Comparison Between Linear Regression, Random Forest, and Logistic Regression Methods in Predicting Crude Palm Oil (CPO) Price,” *Brilliance: Research of Artificial Intelligence*, vol. 3, no. 2, 2023, doi: 10.47709/brilliance.v3i2.3334.

[21] J. H. Yousif and M. J. Yousif, “Critical Review of Neural Network Generations and Models Design,” Nov. 09, 2023. doi: 10.20944/preprints202309.1149.v2.

[22] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, “Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction,” *Sci Rep*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-10358-x.