p-ISSN : 2406-7733
e-ISSN : 2597-9922

PERBANDINGAN ALGORITMA *MACHINE LEARNING* UNTUK PREDIKSI NILAI UJIAN SISWA

Luthfi Luqman Fattah[1], Yuliarman Saragih[2], Budiman Faisal Tanjung[3], Naufal Akmal Nugraha[4]
Program Studi Teknik Elektro, Fakultas Teknik, Universitas Singaperbangsa Karawang
Jl. HS. Ronggo Waluyo, Telukjambe Timur, Karawang
E-mail: *2110631160048@student.unsika.ac.id[1], yuliarman@staff.unsika.ac.id[2],
2110631160005@student.unsika.ac.id[3], 2110631160055@student.unsika.ac.id[4]

Abstrak - Peningkatan kualitas pendidikan tidak hanya bergantung pada metode pembelajaran, tetapi juga pada kemampuan institusi untuk memahami dan mengantisipasi performa akademik siswa. Penelitian ini bertujuan untuk membangun dan membandingkan tiga model prediksi nilai ujian siswa berdasarkan data kebiasaan harian menggunakan algoritma Regresi Linier, Random Forest, dan XGBoost. Data diperoleh dari platform publik dengan jumlah 1000 entri siswa yang mencakup berbagai atribut seperti durasi belajar, kehadiran, aktivitas ekstrakurikuler, dan kesehatan mental. Setiap model diuji menggunakan metrik evaluasi seperti *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE), dan koefisien determinasi (R^2). Hasil penelitian menunjukkan bahwa model Regresi Linier memberikan performa terbaik dengan nilai RMSE terendah sebesar 5,139 dan R^2 tertinggi sebesar 0,897. Analisis lebih lanjut menunjukkan bahwa variabel durasi belajar memiliki kontribusi paling dominan dalam prediksi nilai ujian. Temuan ini mengindikasikan bahwa hubungan antara variabel input dan nilai ujian siswa cenderung linear dan stabil, sehingga dapat dimodelkan secara efektif menggunakan pendekatan sederhana. Penelitian ini memberikan kontribusi dalam penerapan analisis prediktif berbasis *machine learning* untuk keperluan evaluasi pendidikan dan dapat menjadi dasar bagi pengembangan sistem pendukung keputusan di lingkungan sekolah maupun perguruan tinggi.

Kata Kunci: Evaluasi Model, *Machine Learning*, Prediksi Nilai, Random Forest, XGBoost

I. PENDAHULUAN

Performa akademik siswa merupakan indikator utama dalam menilai efektivitas pembelajaran. Nilai ujian, sebagai salah satu bentuk evaluasi hasil belajar, sering digunakan sebagai tolak ukur keberhasilan pendidikan. Seiring dengan meningkatnya ketersediaan data pendidikan dan kemajuan teknologi kecerdasan buatan, penerapan *Machine Learning* (ML) dalam memprediksi capaian akademik menjadi pendekatan yang semakin relevan (Oppong, 2023). ML memungkinkan pengolahan data kompleks dan multivariabel untuk mendeteksi pola dan membuat prediksi yang akurat terhadap hasil belajar (Al-Alawi et al., 2023)

Berbagai penelitian menyebutkan bahwa performa akademik siswa dipengaruhi tidak hanya oleh faktor akademik, tetapi juga oleh kebiasaan harian seperti durasi belajar, kualitas tidur, penggunaan media sosial, serta aspek psikologis seperti kesehatan mental dan stres belajar (Prakash et al., 2024). Oleh karena itu, dibutuhkan pendekatan prediktif yang mempertimbangkan faktor-faktor tersebut secara komprehensif(Vives et al., 2024).

Model prediksi berbasis *machine learning* seperti regresi linier, Random Forest, dan XGBoost telah digunakan secara luas dalam berbagai studi. Regresi linier berganda digunakan sebagai model dasar karena sifatnya yang sederhana dan mudah diinterpretasikan, namun memiliki keterbatasan dalam menangani hubungan non-linear antar variabel. Sebaliknya, Random Forest dan XGBoost, sebagai model ensemble, memiliki kemampuan untuk menangani kompleksitas data dan seringkali menghasilkan akurasi prediksi yang lebih tinggi dalam berbagai kasus prediktif (Gurnani et al., 2021).

Penelitian ini menguji perbandingan kinerja antara regresi linier, Random Forest, dan XGBoost dalam memprediksi nilai ujian siswa berdasarkan kebiasaan harian, untuk mengetahui algoritma mana yang memberikan akurasi terbaik. Dataset yang digunakan bersumber dari platform Kaggle dan mencakup 1000 data siswa dengan atribut seperti jam belajar per hari, kehadiran, kesehatan mental, aktivitas ekstrakurikuler, dan lain-lain.

Melalui pendekatan ini, penelitian diharapkan dapat memberikan kontribusi dalam penerapan model prediksi akademik berbasis data, serta mendukung pengambilan keputusan oleh institusi pendidikan dalam hal intervensi dini dan perencanaan pembelajaran berbasis bukti..

II. TINJAUAN PUSTAKA

A. Regresi Linear

Regresi linear berganda adalah suatu metode statistik yang digunakan untuk

p-ISSN : 2406-7733
e-ISSN : 2597-9922

menganalisis hubungan antara satu variabel dependen (terikat) dengan dua atau lebih variabel independen (bebas) (Iba Z, 2024). Model ini memungkinkan peneliti untuk mengukur pengaruh relatif dari setiap variabel bebas terhadap variabel yang ingin diprediksi secara simultan. Dalam konteks penelitian ini, regresi linear berganda digunakan untuk memprediksi nilai ujian berdasarkan berbagai faktor seperti durasi belajar, kesehatan mental, partisipasi ekstrakurikuler, dan kualitas tidur. Persamaan umum regresi linear berganda adalah sebagai berikut

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \ldots(1)$$

Keterangan :

Y : Variabel dependen (nilai ujian)

$X_1, X_2, \cdots, X_n$: Variabel independent (Faktor-faktor yang mempengaruhi

β_0 : Intersep (nilai Y saat seluruh $X = 0$)

β_n : Koefisien regresi masing-masing variable independent

ε : Error atau residual

Menurut Iba dan Wardhana (2024), analisis regresi linear berganda bermanfaat untuk memahami kontribusi masing-masing faktor terhadap variabel yang diteliti dan penting untuk memperhatikan asumsi dasar regresi seperti multikolinearitas, normalitas, homoskedastisitas, dan linieritas untuk memastikan validitas model.

B. Random Forest

Random Forest merupakan algoritma ensemble yang membangun sejumlah pohon keputusan dengan menggunakan subset data dan fitur secara acak (Salman et al., 2024). Prediksi akhir diperoleh dengan menggabungkan hasil dari seluruh pohon, melalui mekanisme voting pada klasifikasi atau perataan nilai untuk regresi. Pendekatan ini membuat Random Forest lebih tangguh terhadap *overfitting* dan efektif dalam mengelola data yang kompleks dan berdimensi besar.

Keunggulan algoritma ini terletak pada fleksibilitasnya dalam menangani data numerik dan kategorikal, toleransi terhadap *missing values*, serta akurasi prediksi yang tinggi. Dalam konteks regresi, Random Forest menghitung prediksi akhir dengan merata-ratakan hasil dari semua pohon, dan menyediakan feature importance untuk menilai kontribusi masing-masing fitur.

Penggunaan parameter yang tepat seperti jumlah pohon (*ntree*) dan jumlah fitur acak yang dipilih di setiap split (*mtry*) dapat mengoptimalkan performa model (Suci Amaliah et al., 2022). Hal ini diperkuat oleh penelitian Gurnani et al. (2021) yang menunjukkan akurasi hingga 97,8% dalam kasus prediksi kebangkrutan menggunakan Random Forest.

Dalam proses pemilihan atribut terbaik di setiap node pohon, Random Forest menggunakan Gini Index untuk mengukur impuritas. Rumus Gini Index adalah:

$$Gini(S_i) = 1 - \sum_{i=1}^{c} p_i^2 \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Keterangan :

$S = subset\ data\ pada\ sebuah\ node$
$c = jumlah\ kelas$
$p_i = proporsi\ data\ dari\ kelas\ ke - i\ pada\ node$

C. XGBoost

XGBoost (*Extreme Gradient Boosting*) adalah salah satu algoritma *ensemble* yang mengembangkan pendekatan *boosting* untuk menghasilkan model prediksi yang akurat dan efisien. Algoritma ini dikembangkan dengan basis pohon keputusan (*decision tree*) yang dibentuk secara berurutan, di mana setiap pohon berikutnya bertujuan untuk memperbaiki kesalahan dari pohon sebelumnya (Chen & Guestrin, 2016).

Berbeda dengan algoritma *boosting* konvensional, XGBoost mengintegrasikan teknik regularisasi, pengendalian kompleksitas model, dan manajemen memori yang efisien sehingga menghasilkan performa yang tinggi pada data besar dan kompleks. Dalam regresi, XGBoost meminimalkan fungsi kerugian (*loss function*) melalui pendekatan berbasis turunan kedua, yang membuat konvergensi model lebih cepat dan presisi tinggi.

Dalam penelitian oleh Mubarok et al (2022), XGBoost digunakan untuk memprediksi keberlangsungan hidup pasien gagal jantung. Studi tersebut membuktikan bahwa konfigurasi *hyperparameter* yang tepat sangat menentukan kualitas model. Tiga metode penyetelan *hyperparameter* yaitu *Random Search, Tree Parzen Estimator* (TPE), dan *Grid Search* diuji, dan terbukti bahwa *tuning* dengan TPE memberikan hasil prediktif terbaik dengan nilai AUC mencapai 0,944.

p-ISSN : 2406-7733
e-ISSN : 2597-9922

Adapun parameter penting yang sering digunakan dalam konfigurasi model XGBoost antara lain:

Tabel 1. Parameter Penting dalam Konfigurasi Model XGBoost

No	Parameter	Deskripsi
1	*n_estimators*	Jumlah pohon keputusan yang akan dibangun
2	*max_depth*	Kedalaman maksimum dari setiap pohon
3	*learning_rate*	Kecepatan pembelajaran (step size shrinkage)
4	*min_child_weight*	Minimum jumlah bobot di node akhir
5	*subsample*	Rasio data yang digunakan dalam pelatihan
6	*colsample_bytree*	Rasio fitur yang digunakan di tiap pohon

D. Feature Importance

Konsep *feature importance* digunakan untuk mengevaluasi kontribusi relatif setiap variabel input dalam model prediksi. Pada algoritma Random Forest dan XGBoost, nilai *importance* dihitung berdasarkan seberapa sering dan seberapa efektif suatu fitur digunakan dalam membagi data untuk menurunkan kesalahan prediksi. Informasi ini sangat berguna untuk mengidentifikasi fitur dominan yang memengaruhi hasil prediksi dan dapat digunakan untuk *feature selection* guna meningkatkan efisiensi model.

Jia dan Jin (2023)menunjukkan bahwa dalam konteks prediksi risiko stroke, *analisis feature importance* menggunakan model regresi logistik berhasil mengidentifikasi variabel usia, hipertensi, dan kadar glukosa sebagai prediktor utama. Hal ini menunjukkan bahwa pendekatan ini mampu memberikan interpretasi yang bermakna terhadap hubungan antara fitur dan target, tidak hanya pada model berbasis pohon keputusan, tetapi juga pada model statistik klasik seperti regresi logistik.

Dengan demikian, analisis *feature importance* tidak hanya berperan dalam meningkatkan akurasi model, tetapi juga berkontribusi pada pemahaman mendalam tentang faktor-faktor kunci yang memengaruhi output, seperti performa akademik siswa dalam konteks penelitian ini.

E. Evaluasi Model

Evaluasi kinerja model dilakukan menggunakan tiga metrik utama, yaitu *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE), dan *R-squared* (R^2). Ketiga metrik ini digunakan untuk menilai sejauh mana hasil prediksi mendekati nilai actual (Ihzaniah et al., 2023).

1. Root Mean Squared Error

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

RMSE mengukur rata-rata kuadrat selisih antara nilai aktual (y_i) dan nilai prediksi ($\hat{y}_i$).

2. Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \ldots\ldots\ldots\ldots\ldots(4)$$

MAE menghitung rata-rata kesalahan absolut antara nilai aktual dan prediksi, memberikan gambaran kesalahan secara langsung dalam satuan data (Ihzaniah et al., 2023).

3. R-squared (R^2)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \hat{y})^2} \ldots\ldots\ldots\ldots\ldots\ldots(5)$$

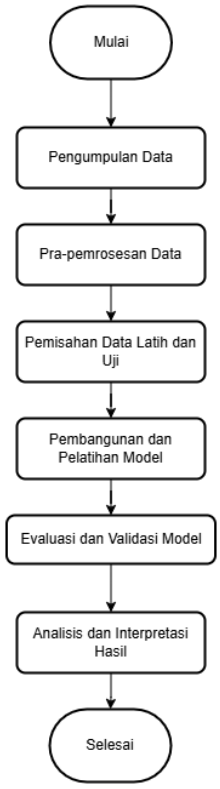
R^2 menunjukkan proporsi variasi dalam data target (y) yang dapat dijelaskan oleh model prediksi(Ihzaniah et al., 2023)

III. METODE PENELITIAN

A. Alur Penelitian

Penelitian ini dilakukan secara terstruktur mulai dari pengumpulan data, pra-pemrosesan, pemisahan data, pemodelan, hingga evaluasi. Proses ini divisualisasikan dalam Gambar 1.

p-ISSN : 2406-7733
e-ISSN : 2597-9922



Gambar 1. Diagram Alir Proses Penelitian

B. Sumber dan Karakteristik Data

Dataset yang digunakan dalam penelitian ini diperoleh dari platform Kaggle dengan judul "*Student Habits and Performance*". Dataset ini berisi 1000 entri data siswa dan terdiri atas 16 atribut yang merepresentasikan berbagai informasi mengenai siswa, mulai dari data demografis, kebiasaan belajar, aktivitas harian, kondisi kesehatan, hingga nilai ujian akhir.

Satu atribut berupa *student_id* digunakan hanya sebagai identitas siswa dan tidak digunakan dalam proses pemodelan. Atribut target yang ingin diprediksi adalah exam_score, yaitu skor ujian akhir siswa dalam rentang 0–100. Atribut lainnya berperan sebagai fitur input atau prediktor dalam model pembelajaran mesin.

Tabel 2. Deskripsi Atribut Dataset

No	Nama Atribut	Deskripsi
1	*student_id*	Identitas unik setiap siswa
2	*age*	Usia siswa dalam tahun
3	*gender*	Jenis kelamin siswa (Laki-laki atau Perempuan)
4	*study_hours_per_day*	Rata-rata jumlah jam belajar siswa per hari
5	*social_media_hours*	Rata-rata jam yang dihabiskan siswa untuk media sosial per hari
6	*Netflix_hours*	Rata-rata jam yang dihabiskan siswa untuk menonton Netflix per hari
7	*part_time_job*	Status apakah siswa memiliki pekerjaan paruh waktu (Ya atau Tidak)
8	*attendance_percentage*	Persentase kehadiran siswa di kelas
9	*sleep_hours*	Rata-rata durasi tidur siswa per hari dalam jam
10	*diet_quality*	Kualitas pola makan siswa (Buruk, Sedang, atau Baik)
11	*exercise_frequency*	Frekuensi siswa berolahraga dalam seminggu (dalam jumlah kali)
12	*parental_education_level*	Tingkat pendidikan orang tua siswa
13	*internet_quality*	Kualitas koneksi internet siswa di rumah (Buruk, Sedang, atau Baik)
14	*mental_health_rating*	Penilaian siswa terhadap kondisi kesehatan mentalnya (skala 1–10)
15	*extracurricular_participation*	Keterlibatan siswa dalam kegiatan ekstrakurikuler (Ya atau Tidak)
16	*exam_score*	Nilai ujian akhir siswa dalam skala 0–100 (target prediksi)

C. Pra-pemrosesan Data

1. Penghapusan Variabel Tidak Relevan

   Variabel identitas siswa (*student_id*) dihapus karena tidak mengandung informasi yang berguna untuk prediksi dan dapat menimbulkan bias jika disertakan

2. Penanganan Nilai Kosong (*Missing Value*)

   Terdapat satu fitur yang memiliki nilai kosong, yaitu tingkat pendidikan orang tua. Sebanyak 91 baris data (9,1% dari total) tidak memiliki isian pada kolom ini. Oleh karena itu, dilakukan imputasi dengan menggantikan nilai kosong menggunakan kategori yang paling sering muncul (modus) agar distribusi data tetap stabil.

3. Trasnformasi Fitur Kategorikal

   Terdapat tujuh fitur kategorikal dalam dataset, seperti jenis kelamin,

p-ISSN : 2406-7733
e-ISSN : 2597-9922

status pekerjaan paruh waktu, kualitas internet, dan lainnya. Fitur-fitur ini dikonversi ke bentuk numerik menggunakan metode Label Encoding karena semua fitur hanya memiliki 2–4 kategori diskrit dan tidak memerlukan teknik One-Hot Encoding yang lebih kompleks. Setelah proses ini, jumlah total fitur dalam dataset berubah dari semula 16 menjadi 15, terdiri dari 1 target dan 14 fitur input numerik.

4. Normalisasi Fitur Numerik

Untuk menyamakan skala antar fitur numerik, dilakukan penskalaan menggunakan metode StandardScaler dari pustaka *scikit-learn*. Proses ini mengubah distribusi setiap fitur numerik menjadi memiliki rata-rata 0 dan standar deviasi 1. Hal ini penting terutama untuk model yang sensitif terhadap skala seperti regresi linier dan XGBoost.

5. Pemisahan Data Latih dan Uji

Dataset kemudian dibagi menjadi dua bagian:

1) 80% data latih (800 data)
2) 20% data uji (200 data).

Pembagian ini dilakukan menggunakan fungsi *train_test_split* dengan parameter *random_state*=42 untuk memastikan bahwa hasil dapat direproduksi.

D. Pembangunan dan Pelatihan Model

Dalam penelitian ini, digunakan tiga model regresi untuk memprediksi nilai ujian siswa, yaitu Regresi Linier, Random Forest, dan XGBoost. Regresi Linier berfungsi sebagai model baseline dengan asumsi hubungan linear antar fitur dan target. Sedangkan Random Forest dan XGBoost dipilih karena kemampuannya menangani hubungan non-linear dan kompleksitas data.

Untuk Random Forest dan XGBoost, dilakukan *tuning hyperparameter* menggunakan GridSearchCV dengan validasi silang 5-fold. Parameter yang disetel mencakup jumlah estimator, kedalaman pohon, *learning rate* (khusus XGBoost), dan rasio *subsampling*. Proses *tuning* ini bertujuan mencari konfigurasi terbaik untuk meningkatkan akurasi prediksi.

Semua model dilatih menggunakan data latih yang telah diskalakan agar hasil pelatihan lebih optimal dan dapat dibandingkan secara adil. Model terbaik hasil tuning kemudian diuji pada data uji untuk evaluasi performa

E. Evaluasi dan Validasi Model

Evaluasi performa model dilakukan menggunakan metrik utama Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), dan Koefisien Determinasi (R^2). Selain pengujiam pada data uji, validasi silang 5-fold juga diterapkan untuk memastikan kestabilan dan generalisasi model.

Selain metrik numerik, dilakukan pula analisis residual pada model regresi linier untuk memeriksa pola kesalahan prediksi dan visualisai *feature importance* pada model Random Forest dan XGBoost untuk mengindentifikasi fitur-fitur paling berpengaruh dalam prediksi

IV. HASIL DAN PEMBAHASAN

A. Ringkasan Hasil Evaluasi Model

Penelitian ini menerapkan tiga algoritma regresi untuk memprediksi nilai ujian siswa berdasarkan kebiasaan belajar dan faktor pribadi, yaitu Regresi Linier, Random Forest, dan XGBoost. Evaluasi dilakukan berdasarkan metrik *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE), dan koefisien determinasi (R^2) pada data uji. Selain itu, validasi silang 5-fold digunakan untuk mengukur kestabilan performa model.

Tabel 3. Hasil Evaluasi Kinerja Model Regresi Linier, Random Forest, dan XGBoost

NO	Model	RMSE	MAE	R^2	RMSE (Cross Validation)
1.	Regresi Linier	5.139	4.169	0.897	5.380
2.	Random Forest	6.234	4.989	0.848	6.402
3.	XGBoost	5.598	4.650	0.878	5.714

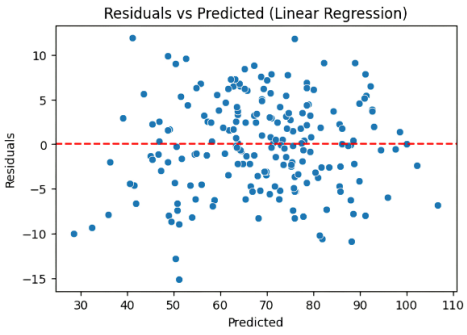
Berdasarkan hasil evaluasi pada Tabel 3, model Regresi Linier memberikan performa terbaik dengan nilai RMSE dan MAE paling rendah serta nilai R^2 tertinggi, yaitu sebesar 0,897. Ini menunjukkan bahwa model mampu menjelaskan 89,7% variasi dalam nilai ujian berdasarkan fitur-fitur yang digunakan. Model XGBoost menempati peringkat kedua, sedangkan Random Forest memiliki performa paling rendah di antara ketiganya. Seluruh model diuji pada data yang telah melalui pra-

p-ISSN : 2406-7733
e-ISSN : 2597-9922

pemrosesan dan penskalaan fitur Performa terbaik dari Regresi Linier menunjukkan bahwa hubungan antara variabel input dan output bersifat linear. Hal ini ditunjukkan oleh tingginya korelasi antara beberapa fitur utama, seperti jam belajar per hari, dengan nilai ujian. Karakteristik data yang relatif bersih, distribusi target yang normal, serta tidak adanya interaksi kompleks antar fitur membuat model regresi linier mampu bekerja lebih optimal dibandingkan model non-linear. Model XGBoost dan Random Forest, meskipun lebih kompleks secara algoritmik, tidak memberikan keunggulan signifikan dalam konteks dataset ini. Hal ini menegaskan bahwa pemilihan model harus disesuaikan dengan karakteristik data, bukan semata kompleksitas algoritma.

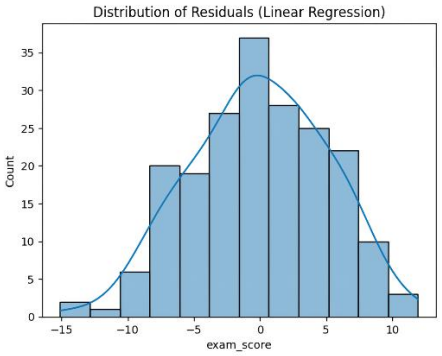
B. Visualisasi Residual dan Pemeriksaan Asusmsi

Untuk memastikan keandalan model regresi linier yang digunakan, dilakukan analisis residual guna menguji asumsi distribusi kesalahan. Visualisasi residual membantu mengidentifikasi adanya pola tertentu, *outlier*, atau pelanggaran terhadap asumsi regresi klasik, seperti homoskedastisitas dan distribusi normal error.



Gambar 2. Plot Residual terhadap Nilai Prediksi

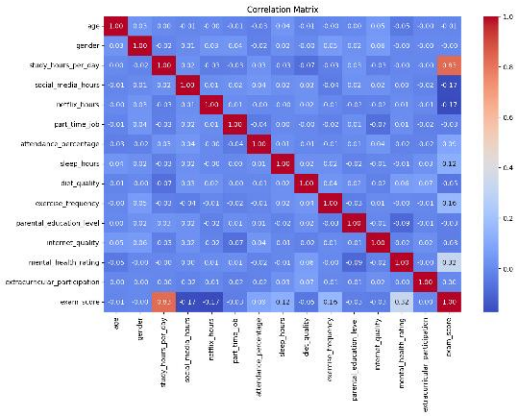
Gambar 2 menunjukkan penyebaran residual terhadap nilai prediksi. Titik-titik tersebar secara acak di sekitar garis horizontal nol tanpa pola sistematis. Pola ini mengindikasikan bahwa asumsi homoskedastisitas terpenuhi, yaitu variansi residual konstan di sepanjang rentang nilai prediksi. Tidak ditemukan pola *fan-shape* atau kurva yang menunjukkan pelanggaran asumsi linearitas atau variansi residual yang meningkat.



Gambar 3. Distribusi Residual Model Regresi Linier

Gambar 3 memperlihatkan distribusi residual model regresi linier. Bentuk distribusi residual menyerupai distribusi normal simetris, dengan puncak pada nol. Distribusi ini mendukung asumsi bahwa residual bersifat acak dan tersebar normal, yang merupakan salah satu syarat validitas model regresi linier.

C. Korelasi antar Variabel



Gambar 4. Matriks korelasi antar variabel

Analisis korelasi dilakukan untuk mengukur kekuatan dan arah hubungan linear antara setiap fitur prediktor dengan nilai ujian siswa. Hasil perhitungan korelasi Pearson yang divisualisasikan dalam bentuk heatmap menunjukkan bahwa variabel dengan korelasi tertinggi terhadap nilai ujian adalah jumlah jam belajar per hari, dengan nilai koefisien sebesar 0,83. Ini mengindikasikan adanya hubungan positif yang sangat kuat antara kebiasaan belajar dan pencapaian akademik. Selain itu, penilaian siswa terhadap kondisi kesehatan mentalnya juga menunjukkan korelasi positif sebesar 0,32, yang berarti bahwa siswa dengan kesehatan mental yang baik cenderung memiliki nilai ujian yang lebih tinggi. Variabel lain seperti frekuensi olahraga dan durasi tidur memiliki korelasi positif

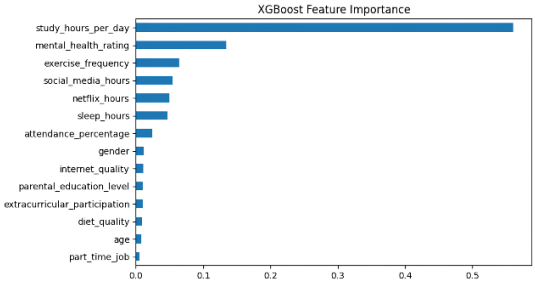
p-ISSN : 2406-7733
e-ISSN : 2597-9922

yang lebih lemah, namun tetap mendukung peningkatan hasil akademik.

Di sisi lain, ditemukan pula korelasi negatif antara penggunaan media sosial (−0,17) dan jam menonton Netflix (−0,17) dengan nilai ujian. Hal ini menunjukkan bahwa semakin banyak waktu yang dihabiskan siswa untuk aktivitas hiburan, semakin rendah kecenderungan mereka untuk memperoleh hasil akademik yang tinggi. Korelasi ini memberikan gambaran bahwa pola manajemen waktu memiliki dampak yang signifikan terhadap performa akademik siswa. Secara keseluruhan, temuan ini konsisten dengan hasil pemodelan yang menunjukkan bahwa variabel-variabel yang berkaitan dengan disiplin belajar dan kesehatan mental memiliki pengaruh paling besar terhadap nilai ujian, sementara aktivitas hiburan berlebihan justru berkorelasi negatif

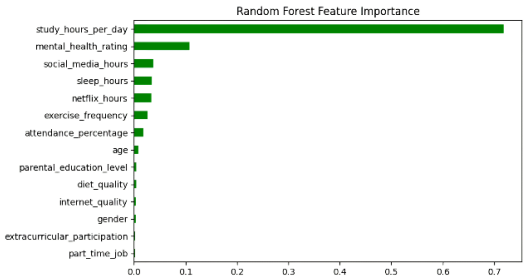
D. Kontribusi Fitur terhadap Prediksi

Analisis kontribusi fitur dilakukan untuk mengetahui seberapa besar pengaruh masing-masing variabel input dalam proses prediksi nilai ujian siswa. Dua model pembelajaran mesin yang digunakan, yaitu XGBoost dan Random Forest, memiliki mekanisme internal untuk menghitung tingkat kepentingan fitur berdasarkan frekuensi dan efektivitasnya dalam membagi data pada setiap pohon keputusan.



Gambar 5. XGBoost Feature Importamce

Hasil visualisasi pada Gambar 5 menunjukkan bahwa model XGBoost mengidentifikasi jumlah jam belajar per hari sebagai fitur dengan kontribusi paling dominan dalam prediksi. Fitur ini memberikan kontribusi lebih dari setengah dari total bobot prediksi, mengungguli fitur lainnya secara signifikan. Fitur berikutnya yang cukup penting adalah penilaian kesehatan mental, diikuti oleh frekuensi olahraga, jam penggunaan media sosial, dan jam menonton Netflix, meskipun dengan kontribusi yang jauh lebih kecil.



Gambar 6 Random Forest Feature Importance

Sementara itu, hasil pada Gambar 6 dari model Random Forest menunjukkan pola yang konsisten. Fitur jam belajar per hari kembali menjadi yang paling berpengaruh, bahkan dengan proporsi yang lebih ekstrem, mencapai lebih dari 70% dari total kontribusi. Fitur-fitur lain memiliki distribusi kontribusi yang jauh lebih rendah, termasuk kesehatan mental, durasi tidur, dan kehadiran siswa.

Konsistensi antara kedua model dalam mengidentifikasi fitur utama menegaskan bahwa kebiasaan belajar merupakan faktor yang paling berpengaruh terhadap pencapaian akademik. Selain itu, fitur-fitur pendukung seperti kesehatan mental dan olahraga tetap menunjukkan peran relevan meskipun dalam skala yang lebih kecil. Sebaliknya, variabel seperti pekerjaan paruh waktu, partisipasi ekstrakurikuler, dan kualitas internet memiliki pengaruh yang sangat rendah dalam membentuk prediksi.

Temuan ini mengindikasikan bahwa model pembelajaran mesin secara akurat mampu menangkap pola penting dalam data perilaku siswa, dan hasil kontribusi fitur ini dapat menjadi dasar dalam merancang kebijakan pendidikan berbasis data.

E. Diskusi dan Implikasi

Hasil penelitian menunjukkan bahwa model regresi linier memberikan performa terbaik dalam memprediksi nilai ujian siswa dibandingkan dengan model Random Forest dan XGBoost. Hal ini didukung oleh nilai RMSE yang paling rendah serta R^2 tertinggi, yang mengindikasikan bahwa hubungan antara variabel-variabel input dengan nilai ujian bersifat linear dan stabil. Selain itu, hasil visualisasi residual juga menunjukkan bahwa model regresi linier memenuhi asumsi dasar regresi, seperti homoskedastisitas dan distribusi normal

p-ISSN : 2406-7733
e-ISSN : 2597-9922

residual, sehingga model dapat dianggap valid secara statistik.

Salah satu temuan utama dalam penelitian ini adalah pengaruh dominan dari variabel jumlah jam belajar per hari terhadap nilai ujian. Temuan ini sejalan dengan hasil penelitian sebelumnya yang menyatakan bahwa intensitas waktu belajar merupakan faktor signifikan dalam pencapaian akademik siswa (Amir et al., 2023). Analisis korelasi dan *feature importance* secara konsisten menunjukkan bahwa variabel ini memiliki korelasi positif yang sangat kuat serta kontribusi terbesar dalam model prediksi. Kondisi ini mempertegas pentingnya disiplin belajar sebagai fondasi kesuksesan akademik.

Di samping itu, variabel-variabel non-akademik seperti penilaian kesehatan mental dan frekuensi olahraga juga menunjukkan pengaruh yang cukup signifikan, meskipun dalam skala yang lebih rendah. Hal ini mendukung pandangan bahwa keberhasilan akademik tidak hanya dipengaruhi oleh aspek kognitif semata, tetapi juga oleh faktor afektif dan fisik. Temuan ini memperkuat hasil studi seperti yang dilakukan oleh Nuranisa Ikhsani & Putranto (2024) yang menyatakan bahwa kondisi psikologis siswa memiliki hubungan signifikan dengan pencapaian akademik mereka.

Sebaliknya, variabel seperti pekerjaan paruh waktu, partisipasi ekstrakurikuler, dan kualitas internet menunjukkan kontribusi yang sangat rendah terhadap prediksi nilai ujian. Hal ini dapat disebabkan oleh variabilitas pengaruh faktor-faktor tersebut yang lebih kompleks atau bersifat tidak langsung dalam memengaruhi prestasi akademik.

Secara teoretis, hasil penelitian ini memberikan kontribusi dalam memperkuat model prediksi performa akademik berbasis data kebiasaan dan karakteristik siswa. Secara praktis, hasil ini dapat dijadikan dasar bagi institusi pendidikan untuk merancang intervensi berbasis data, seperti manajemen waktu belajar yang lebih terstruktur, dukungan kesehatan mental, dan program peningkatan kesejahteraan siswa.

V. KESIMPULAN DAN SARAN

Kesimpulan

1. Model regresi linier terbukti memberikan performa prediktif terbaik dibandingkan dengan Random Forest dan XGBoost dalam memprediksi nilai ujian siswa. Hal ini ditunjukkan oleh nilai RMSE terendah sebesar 5.139 dan R^2 tertinggi sebesar 0,897, yang mengindikasikan bahwa hubungan antara kebiasaan siswa dan performa akademik bersifat linear dan stabil.
2. Variabel jumlah jam belajar per hari menjadi fitur paling dominan dalam prediksi nilai ujian, sebagaimana dibuktikan oleh nilai korelasi sebesar 0,83 serta bobot kontribusi tertinggi dalam model XGBoost dan Random Forest.
3. Faktor non-akademik seperti kesehatan mental dan frekuensi olahraga turut memberikan kontribusi signifikan dalam model prediksi, menunjukkan bahwa aspek afektif dan fisik juga berperan dalam pencapaian akademik siswa.
4. Meskipun algoritma ensemble seperti XGBoost dan Random Forest memiliki kompleksitas dan fleksibilitas tinggi, dalam konteks dataset ini keduanya tidak mampu melampaui performa regresi linier, yang lebih cocok karena pola hubungan antar variabel yang cenderung linear.
5. Keunggulan utama penelitian ini terletak pada pendekatan evaluasi menyeluruh, yang tidak hanya mengandalkan metrik numerik (RMSE, MAE, R^2), tetapi juga menggunakan visualisasi residual, analisis korelasi, dan *feature importance* untuk memperkuat interpretasi model.
6. Keterbatasan penelitian ini terletak pada jenis data yang digunakan, yang hanya mencakup data berbasis *self-report* dari kebiasaan siswa, tanpa mempertimbangkan variabel kontekstual seperti latar belakang keluarga, lingkungan belajar, atau riwayat nilai sebelumnya.
7. Penelitian ini memiliki potensi pengembangan ke depan, seperti penambahan fitur yang lebih beragam, penggunaan teknik seleksi fitur otomatis, dan penerapan algoritma deep learning apabila tersedia dataset yang lebih besar dan kompleks.

Saran

Berdasarkan hasil dan keterbatasan penelitian ini, terdapat beberapa arah yang dapat dikembangkan untuk penelitian selanjutnya. Salah satu saran utama adalah menggabungkan data kebiasaan siswa dengan data historis nilai akademik serta latar belakang lingkungan untuk meningkatkan akurasi prediksi. Selain itu, penambahan variabel kontekstual seperti dukungan orang tua, kondisi sosial ekonomi, dan metode pengajaran yang diterima siswa juga dapat memperkaya model dan memberikan perspektif yang lebih komprehensif. Penelitian selanjutnya juga disarankan untuk menerapkan teknik *feature selection* otomatis guna menyederhanakan kompleksitas model dan meningkatkan efisiensi

p-ISSN : 2406-7733
e-ISSN : 2597-9922

komputasi. Terakhir, jika tersedia dataset yang lebih besar dan kompleks, eksplorasi terhadap model berbasis *deep learning* dapat menjadi pendekatan potensial untuk menangkap hubungan non-linear yang lebih dalam antar variabel.

.

DAFTAR PUSTAKA

Al-Alawi, L., Al Shaqsi, J., Tarhini, A., & Al-Busaidi, A. S. (2023). Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. *Education and Information Technologies*, *28*(10), 12407–12432. https://doi.org/10.1007/s10639-023-11700-0

Amir, N. S., Tikollah, M. R., & Azis, F. (2023). *P-ISSN E-ISSN Siswa pada Mata Pelajaran Ekonomi Materi Akuntansi di SMA Negeri 4 Soppeng*. *8*(1), 29–36.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Gurnani, I., Vincent, Tandian, F. S., & Anggreainy, M. S. (2021). Predicting Company Bankruptcy Using Random Forest Method. *2021 2nd International Conference on Artificial Intelligence and Data Sciences, AiDAS 2021*, *August*. https://doi.org/10.1109/AiDAS53897.2021.9574384

Iba Z, W. (2024). *Analisis Regresi dan Analisis Jalur untuk Riset Bisnis* (Vol. 19, Issue 5).

Ihzaniah, L. S., Setiawan, A., & Wijaya, R. W. N. (2023). Perbandingan Kinerja Metode Regresi K-Nearest Neighbor dan Metode Regresi Linear Berganda pada Data Boston Housing. *Jambura Journal of Probability and Statistics*, *4*(1), 17–29. https://doi.org/10.34312/jjps.v4i1.18948

Jia, G., & Jin, G. (2023). The prediction and feature importance analysis of stroke based on the machine learning algorithm. *Applied and Computational Engineering*, *18*(1), 225–229. https://doi.org/10.54254/2755-2721/18/20230994

Nuranisa Ikhsani, D., & Putranto, A. (2024). Tantangan Keluarga Broken Home (Studi Tentang Motivasi Belajar Siswa Di Smp Negeri 2 Ngantru Tulungagung). *Journal on Education*, *6*(4), 21644–21655. https://doi.org/10.31004/joe.v6i4.6313

Oppong, S. O. (2023). Predicting Students' Performance Using Machine Learning Algorithms: A Review. *Asian Journal of Research in Computer Science*, *16*(3), 128–148. https://doi.org/10.9734/ajrcos/2023/v16i3351

Prakash, P. R., Swathi, R. S. V. R., Anbalagan, N., Pattnaik, M., Varaprasad, A. M., & Yadavalli, P. K. (2024). Regression Based Modelling to Predict the Undergraduate Students Performance After Pandemic in Educational Institutions. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(7s), 57–66.

Rizky Mubarok, M., Herteno, R., Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lambung Mangkurat Jalan Ahmad Yani Km, I., & Selatan, K. (2022). Hyper-Parameter Tuning Pada Xgboost Untuk Prediksi Keberlangsungan Hidup Pasien Gagal Jantung. *Klik - Kumpulan Jurnal Ilmu Komputer*, *9*(2), 391–401. http://klik.ulm.ac.id/index.php/klik/article/view/484

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/10.58496/bjml/2024/007

Suci Amaliah, Nusrang, M., & Aswi, A. (2022). Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, *4*(3), 121–127. https://doi.org/10.35580/variansiunm31

Vives, L., Cabezas, I., Vives, J. C., Reyes, N. G., Aquino, J., Condor, J. B., & Altamirano, S. F. S. (2024). Prediction of Students' Academic Performance in the Programming Fundamentals Course Using Long Short-Term Memory Neural Networks. *IEEE Access*, *12*, 5882–5898. https://doi.org/10.1109/ACCESS.2024.3350169