

Analisis Komparatif KNN dan SVR dengan Optimasi Parameter untuk Prediksi Produktivitas Siswa Berbasis Distraksi Digital

Dekka Andri Cahya¹, Fathoni^{*2}

¹Jurusan Sistem Informasi, Fakultas Ilmu Komputer, Universitas Sriwijaya, Indonesia

²Fakultas Ilmu Komputer, Universitas Sriwijaya, Indonesia

Email: ¹dekkaac02@gmail.com, ²fathoni@unsri.ac.id

Abstrak

Distraksi digital seperti penggunaan ponsel, media sosial, YouTube, dan gim daring terbukti berdampak negatif terhadap produktivitas siswa, namun studi komparatif berbasis machine learning yang secara spesifik membandingkan kemampuan prediksi KNN Regressor dan Support Vector Regression (SVR) menggunakan variabel distraksi digital sebagai prediktor tunggal belum pernah dilakukan, sehingga belum diketahui algoritma mana yang lebih optimal untuk kasus prediksi skor produktivitas yang bersifat kontinu. Penelitian ini bertujuan membandingkan performa KNN Regressor dan SVR dengan optimasi parameter menggunakan Grid Search Cross-validation untuk memprediksi *productivity_score* siswa berdasarkan empat variabel distraksi digital, serta mengidentifikasi algoritma yang menghasilkan prediksi paling akurat berdasarkan metrik MAE, RMSE, dan R^2 . Dataset bersumber dari Kaggle yang memuat 20.000 data siswa diproses menggunakan Python dengan library Scikit-learn, dinormalisasi melalui Min-Max Scaling, dibagi dengan proporsi 80:20, dan dioptimasi melalui 5-Fold Cross-validation. Hasil evaluasi menunjukkan SVR dengan kernel linear, $C = 1$, dan $\epsilon = 0,5$ secara konsisten mengungguli KNN Regressor ($k = 15$, metrik manhattan) pada seluruh metrik dengan MAE 12,4461, RMSE 15,1753, dan R^2 0,1060, sedangkan KNN menghasilkan MAE 12,8300, RMSE 15,6960, dan R^2 0,0436; keunggulan SVR pada R^2 mencapai 143% lebih tinggi dibanding KNN. Nilai R^2 yang rendah (0,1060) menunjukkan keterbatasan model, yang mengindikasikan bahwa variabel distraksi digital yang digunakan belum mampu menjelaskan variasi produktivitas siswa secara menyeluruh.

Kata kunci: *distraksi digital, KNN, produktivitas, SVR*

Comparative Analysis of KNN and SVR with Parameter Optimization for Student Productivity Prediction Based on Digital Distraction

Abstract

Digital distractions, including smartphone use, social media, YouTube, and online gaming, have been proven to negatively impact student productivity; however, a comparative machine learning study specifically examining the predictive capabilities of KNN Regressor and Support Vector Regression (SVR) using digital distraction variables as sole predictors has not been conducted, leaving it unclear which algorithm is more optimal for predicting continuous productivity scores. This study aims to compare the performance of KNN Regressor and SVR with parameter optimization using Grid Search Cross-validation to predict student *productivity_score* based on four digital distraction variables, while identifying the algorithm that yields the most accurate predictions based on MAE, RMSE, and R^2 metrics. A Kaggle dataset of 20,000 student records was processed using Python with the Scikit-learn library, normalized via Min-Max Scaling, split at an 80:20 ratio, and optimized through 5-Fold Cross-validation. Evaluation results show that SVR with a linear kernel, $C = 1$, and $\epsilon = 0.5$ consistently outperforms KNN Regressor ($k = 15$, Manhattan metric) across all metrics with MAE 12.4461, RMSE 15.1753, and R^2 0.1060, while KNN yields MAE 12.8300, RMSE 15.6960, and R^2 0.0436; SVR's advantage in R^2 reaches 143% higher than KNN. The low R^2 value (0.1060), however, reflects a limitation of the model, indicating that the digital distraction variables used were not sufficient to fully explain the variation in student productivity.

Keywords: *digital distraction, KNN, productivity, SVR*

1. PENDAHULUAN

Penggunaan perangkat digital seperti ponsel, media sosial, platform video, dan gim daring telah mengubah pola interaksi sosial dan kebiasaan belajar secara fundamental. Meskipun teknologi ini menawarkan berbagai

manfaat edukatif, penggunaannya yang berlebihan dan tidak terkontrol justru menimbulkan ancaman serius bagi produktivitas akademik siswa[1], [2]. Distraksi digital yang bersumber dari jejaring sosial, platform video hiburan, dan gim daring telah terbukti secara empiris mengganggu konsentrasi, mengurangi motivasi belajar intrinsik, serta menurunkan kualitas pengelolaan waktu pelajar secara keseluruhan[3], [4], [5].

Permasalahan distraksi digital semakin kompleks seiring meningkatnya durasi penggunaan berbagai platform secara bersamaan. Data menunjukkan bahwa rata-rata pelajar generasi Z menghabiskan beberapa jam per hari pada aktivitas non-akademik di perangkat digital mereka, mencakup penggunaan media sosial, menonton konten video, hingga bermain gim daring[6], [7]. Kondisi ini menimbulkan dampak yang dapat diukur terhadap skor produktivitas siswa, di mana peningkatan intensitas distraksi digital berkorelasi negatif secara signifikan dengan capaian akademik dan efektivitas belajar, bahkan tanpa memandang tahun studi mahasiswa yang bersangkutan[8], [9]. Ironisnya, meskipun fenomena ini telah luas diamati, pemahaman kuantitatif mengenai seberapa besar masing-masing variabel distraksi digital seperti durasi penggunaan ponsel, media sosial, YouTube, dan gim memengaruhi skor produktivitas secara numerik masih sangat terbatas, sehingga rekomendasi intervensi yang dihasilkan sering bersifat generik dan kurang tepat sasaran.

Sejumlah penelitian terdahulu telah mengkaji hubungan antara distraksi digital dan produktivitas dari berbagai pendekatan. Martin et al.[1] melakukan tinjauan sistematis terhadap 26 artikel dan mengidentifikasi bahwa distraksi teknologi, khususnya jejaring sosial dan kecanduan internet, menjadi penyebab utama hilangnya fokus akademik. Göz et al.[2] dalam studinya terhadap 1.532 mahasiswa menemukan korelasi negatif yang signifikan antara tingkat distraksi digital dengan kepuasan umum dan pembelajaran yang dirasakan. Di ranah machine learning, algoritma KNN dan SVR telah dibandingkan di berbagai domain seperti energi[10], pertanian[11], dan keuangan[12], [13]. SVR umumnya lebih unggul pada data dengan pola linear, dan performanya meningkat signifikan ketika parameternya dioptimasi secara sistematis[14], [15].

Meskipun demikian, terdapat celah penelitian yang signifikan dan belum terjawab. Pertama, belum ada studi yang secara spesifik membandingkan performa KNN dan SVR dalam tugas regresi untuk memprediksi skor produktivitas siswa secara numerik menggunakan variabel-variabel distraksi digital yang terukur, seperti durasi penggunaan ponsel, media sosial, platform video, dan gim daring. Sebagian besar penelitian yang ada menggunakan pendekatan survei persepsi atau korelasi bivariat, bukan pemodelan prediktif berbasis data numerik[1], [8]. Kedua, meskipun studi-studi komparatif KNN versus SVR telah dilakukan di berbagai domain seperti energi, pertanian, dan keuangan[10], [11], [16], belum ada yang mengeksplorasi perbandingan ini dalam konteks prediksi produktivitas siswa berbasis distraksi digital. Ketiga, optimasi parameter yang sistematis mencakup penentuan nilai k optimal pada KNN serta pemilihan kernel, parameter C , dan epsilon pada SVR masih jarang diimplementasikan secara eksplisit dalam penelitian prediksi produktivitas pendidikan, padahal Daulay[17] dan Su et al.[15] menunjukkan bahwa optimasi parameter dapat meningkatkan performa model secara substantif.

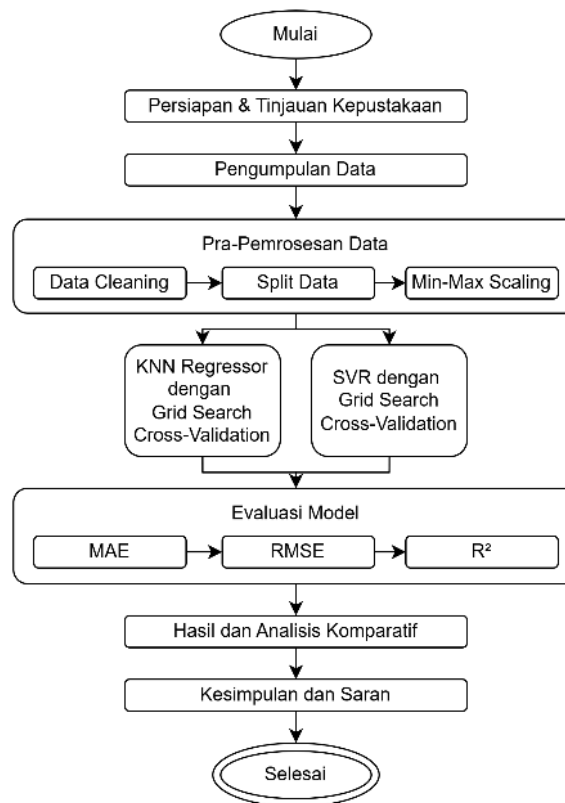
Penelitian ini diusulkan untuk menjawab celah tersebut dengan melakukan komparasi algoritma KNN Regressor dan Support Vector Regression (SVR) secara sistematis, disertai optimasi parameter pada masing-masing model, untuk memprediksi skor produktivitas siswa secara numerik berbasis variabel distraksi digital. Dataset yang digunakan bersumber dari Kaggle, yang memuat fitur-fitur terukur berupa *phone_usage_hours*, *social_media_hours*, *youtube_hours*, dan *gaming_hours* sebagai prediktor terhadap variabel target *productivity_score* yang bersifat kontinu. Pendekatan komparatif ini memungkinkan identifikasi algoritma yang paling akurat dan andal dalam memperkirakan nilai produktivitas siswa berdasarkan pola konsumsi digitalnya. Optimasi parameter dilakukan secara terstruktur mencakup pencarian nilai k optimal pada KNN serta penentuan kernel, parameter C , dan epsilon pada SVR agar hasil perbandingan mencerminkan potensi maksimal masing-masing algoritma, bukan sekadar konfigurasi default yang mungkin tidak optimal[14], [17].

Tujuan utama penelitian ini, pertama membandingkan performa algoritma KNN Regressor dan SVR setelah optimasi parameter dalam memprediksi skor produktivitas siswa berdasarkan variabel distraksi digital, menggunakan metrik evaluasi MAE, RMSE, dan R^2 . Kedua mengidentifikasi algoritma mana yang menghasilkan prediksi paling akurat dan konsisten terhadap nilai *productivity_score* pada dataset yang digunakan. Kontribusi penelitian ini secara teoretis, penelitian ini memperkaya literatur komparasi KNN dan SVR dalam konteks baru prediksi produktivitas berbasis distraksi digital dengan pendekatan regresi. Secara praktis, temuan penelitian ini dapat menjadi acuan bagi peneliti dan pengembang sistem dalam memilih algoritma yang paling sesuai apabila hendak membangun model prediksi produktivitas pelajar di era digital.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimental komparatif untuk membandingkan performa algoritma K-Nearest Neighbor (KNN) Regressor dan Support Vector Regression (SVR) dalam memprediksi skor produktivitas siswa berdasarkan variabel distraksi digital. Kedua algoritma dilatih pada data yang identik, dioptimasi menggunakan Grid Search Cross-validation, dan dievaluasi menggunakan metrik

yang sama untuk memastikan perbandingan yang adil dan objektif. Rangkaian tahapan penelitian mencakup pengumpulan dan pra-pemrosesan data, implementasi kedua algoritma beserta formulasi matematisnya, optimasi parameter, serta validasi hasil prediksi. Seluruh implementasi dilakukan menggunakan bahasa pemrograman Python dengan library Scikit-learn pada Google Colab. Keseluruhan alur penelitian digambarkan pada Gambar 1.



Gambar 1. Alur Penelitian

2.1. Dataset dan Pengumpulan Data

Dataset yang digunakan dalam penelitian ini diperoleh dari publik Kaggle dengan nama "Student Productivity & Behavior Dataset (20K)", yang terdiri dari 20.000 baris data rekam perilaku penggunaan perangkat digital siswa secara kuantitatif. Variabel prediktor yang digunakan meliputi *phone_usage_hours* (durasi penggunaan ponsel per hari dalam jam), *social_media_hours* (durasi penggunaan media sosial per hari dalam jam), *youtube_hours* (durasi menonton YouTube per hari dalam jam), dan *gaming_hours* (durasi bermain gim daring per hari dalam jam). Variabel target yang diprediksi adalah *productivity_score*, yaitu skor produktivitas siswa yang bersifat kontinu dengan rentang nilai 0 hingga 100. Meskipun dataset memuat 18 kolom secara keseluruhan, variabel kategorikal gender dan kolom-kolom di luar cakupan distraksi digital seperti *study_hours_per_day*, *sleep_hours*, *stress_level*, dan *focus_score* tidak diikutsertakan sebagai prediktor karena fokus penelitian ini adalah menguji sejauh mana durasi penggunaan platform digital sebagai bentuk distraksi yang dapat diukur langsung mampu memprediksi produktivitas siswa. Penggunaan 4 variabel prediktor distraksi digital diharapkan akan mendapatkan hasil yang lebih mudah diinterpretasikan dan langsung menjawab celah penelitian. Hal ini diperkuat dengan penelitian sebelumnya yang telah dilakukan oleh Martin et al.[1] dan Pérez-Juárez et al.[8].

2.2. Pra-Pemrosesan Data

Sebelum data dimasukkan ke dalam model, dilakukan serangkaian tahap pra-pemrosesan. Pertama, pemeriksaan missing values dan duplikasi dilakukan untuk memastikan integritas data. Kedua, dataset kemudian dibagi menggunakan teknik holdout dengan proporsi 80:20, yaitu 16.000 data (80%) sebagai data pelatihan dan 4.000 data (20%) sebagai data pengujian. Pembagian dilakukan secara acak menggunakan fungsi `train_test_split` dengan nilai `random_state` yang tetap guna memastikan replikabilitas eksperimen. Ketiga, normalisasi fitur dilakukan menggunakan metode Min-Max Scaling yang menyetarakan seluruh variabel prediktor ke dalam rentang [0, 1]. Normalisasi ini penting karena algoritma KNN sangat sensitif terhadap perbedaan skala antar fitur-fitur

berskala besar akan mendominasi perhitungan jarak jika tidak dinormalisasi[17]. Persamaan Min-Max Scaling dirumuskan sebagai berikut:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Di mana x adalah nilai asli, x' adalah nilai yang telah dinormalisasi, $x_{\min}$ adalah nilai minimum fitur, dan $x_{\max}$ adalah nilai maksimum fitur. MinMaxScaler dilatih hanya pada data pelatihan kemudian diterapkan ke data pengujian untuk menghindari kebocoran data (data leakage).

2.3. Algoritma K-Nearest Neighbor (KNN) Regressor

K-Nearest Neighbor (KNN) adalah algoritma pembelajaran berbasis instance (instance-based learning) yang membuat prediksi berdasarkan kedekatan suatu titik data baru terhadap titik-titik dalam data pelatihan. Dalam tugas regresi, KNN memprediksi nilai output dengan menghitung rata-rata nilai target dari k tetangga terdekat[10], [17]. Algoritma ini tidak membangun model eksplisit selama fase pelatihan, melainkan menyimpan seluruh data pelatihan dan melakukan komputasi saat prediksi diperlukan.

Jarak antara titik data baru x^z dan setiap titik data pelatihan dihitung menggunakan metrik jarak Euclidean:

$$d(x_i, x_j) = \sqrt{\sum_{n=1}^p (x_{in} - x_{jn})^2} \quad (2)$$

Di mana x^i dan x^j adalah dua titik data dalam ruang fitur berdimensi p . Setelah jarak dihitung terhadap seluruh data pelatihan, algoritma memilih k titik terdekat. Nilai prediksi $\hat{y}$ dihitung sebagai rata-rata nilai target dari k tetangga tersebut:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^k y_i \quad (3)$$

Di mana y^i adalah nilai target tetangga ke- i dan k adalah jumlah tetangga yang dipilih. Parameter k merupakan hiperparameter utama KNN: nilai k terlalu kecil menyebabkan overfitting, sementara nilai k terlalu besar menyebabkan underfitting. Penentuan nilai k optimal dilakukan melalui Grid Search Cross-validation.

2.4. Algoritma Support Vector Regression (SVR)

Support Vector Regression (SVR) merupakan pengembangan Support Vector Machine untuk tugas regresi. SVR berusaha menemukan fungsi $f(x)$ yang memprediksi nilai target dengan deviasi maksimum ε (epsilon) dari nilai aktual, sekaligus mempertahankan fungsi tetap sedatar mungkin[13], [14]. Fungsi regresi SVR berbentuk:

$$f(x) = \langle w, x \rangle + b \quad (4)$$

Di mana w adalah vektor bobot dan b adalah bias. SVR mencari w dan b dengan meminimalkan fungsi objektif berikut:

$$\text{Minimize: } \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (5)$$

Dengan kendala: $y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i$ dan $\langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^*$, di mana $\xi_i, \xi_i^* \geq 0$. Variabel slack ξ_i dan ξ_i^* mengukur deviasi di luar batas epsilon, C adalah parameter regularisasi yang mengontrol trade-off antara kerataan fungsi dan toleransi terhadap kesalahan, serta ε mendefinisikan lebar zona toleransi (epsilon-tube) di mana prediksi yang berada dalam zona ini tidak dihitung sebagai kesalahan.

Untuk menangani hubungan non-linier antara variabel distraksi digital dan skor produktivitas, SVR menggunakan fungsi kernel Radial Basis Function (RBF) yang memetakan data ke ruang berdimensi lebih tinggi:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (6)$$

Di mana γ (gamma) mendefinisikan jangkauan pengaruh setiap titik data pelatihan. Parameter C , ϵ , dan γ merupakan hiperparameter utama SVR yang dioptimasi melalui Grid Search Cross-validation.

2.5. Grid Search Cross-validation

Optimasi parameter merupakan tahap krusial yang menentukan kualitas akhir model prediksi. Penelitian ini menggunakan Grid Search dikombinasikan dengan K-Fold Cross-validation (Grid Search CV) untuk mencari kombinasi hiperparameter terbaik secara sistematis [14], [15]. Grid Search mendefinisikan ruang pencarian berupa kumpulan nilai kandidat untuk setiap hiperparameter, kemudian mengevaluasi seluruh kombinasi yang mungkin.

Untuk setiap kombinasi hiperparameter, performa model diestimasi menggunakan 5-Fold Cross-validation, di mana data pelatihan dibagi menjadi 5 lipatan (fold) yang sama besar. Model dilatih pada 4 lipatan dan dievaluasi pada 1 lipatan secara bergantian, sehingga setiap data pelatihan digunakan sebagai data validasi tepat satu kali. Skor validasi yang digunakan sebagai kriteria pemilihan adalah negatif Mean Squared Error (neg-MSE) kombinasi hiperparameter dengan nilai MSE terkecil dipilih sebagai konfigurasi optimal. Ruang pencarian hiperparameter untuk masing-masing algoritma disajikan pada Tabel 1.

Tabel 1. Ruang Pencarian Hiperparameter Grid Search CV

Algoritma	Hiperparameter	Nilai Kandidat
KNN Regressor	n_neighbors (k)	{3, 5, 7, 9, 11, 13, 15}
KNN Regressor	weights	{‘uniform’, ‘distance’}
KNN Regressor	metric	{‘euclidean’, ‘manhattan’}
SVR	C	{0.1, 1, 10, 100}
SVR	epsilon (ϵ)	{0.01, 0.1, 0.5, 1}
SVR	kernel	{‘rbf’, ‘linear’}
SVR	gamma (γ)	{‘scale’, ‘auto’, 0.01, 0.1}

Seluruh proses *Grid Search CV* diimplementasikan menggunakan library Scikit-learn pada Python melalui kelas GridSearchCV, yang secara otomatis mengevaluasi seluruh kombinasi parameter dan mengembalikan model dengan performa validasi terbaik.

2.6. Evaluasi dan Validasi Model

Evaluasi performa model dilakukan pada data pengujian menggunakan tiga metrik regresi yang diaplikasikan secara bersamaan untuk memberikan gambaran komprehensif tentang kualitas prediksi masing-masing algoritma. Ketiga metrik tersebut adalah Mean Absolute Error (MAE), Root Mean Square Error (RMSE), dan Koefisien Determinasi (R^2).

MAE mengukur rata-rata selisih absolut antara nilai prediksi dan nilai aktual. MAE bersifat tahan terhadap outlier dan mudah diinterpretasikan dalam satuan yang sama dengan variabel target:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (7)$$

RMSE mengukur akar kuadrat dari rata-rata kuadrat selisih prediksi dan nilai aktual. RMSE memberikan penalti lebih besar terhadap kesalahan prediksi yang besar sehingga lebih sensitif terhadap outlier dibandingkan MAE:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (8)$$

R^2 mengukur proporsi variansi variabel target yang berhasil dijelaskan oleh model, dengan rentang nilai 0 hingga 1 di mana nilai yang mendekati 1 menandakan model yang sangat baik:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (9)$$

Di mana y^l adalah nilai aktual, $\hat{y}_i$ adalah nilai prediksi, $\bar{y}$ adalah rata-rata nilai aktual, dan n adalah jumlah data pengujian. Algoritma yang lebih optimal ditentukan berdasarkan nilai MAE dan RMSE yang lebih kecil serta nilai R^2 yang lebih besar. Rangkuman metrik dan kriteria perbandingan disajikan pada Tabel 2.

Tabel 2. Metrik Evaluasi dan Kriteria Model Terbaik

Metrik	Rumus	Kriteria Model Terbaik
MAE	$MAE = \frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $	Nilai lebih kecil lebih baik
RMSE	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$	Nilai lebih kecil lebih baik
R^2	$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$	Nilai lebih besar lebih baik (maks. 1)

3. HASIL DAN PEMBAHASAN

3.1. Karakteristik Dataset dan Statistik Deskriptif

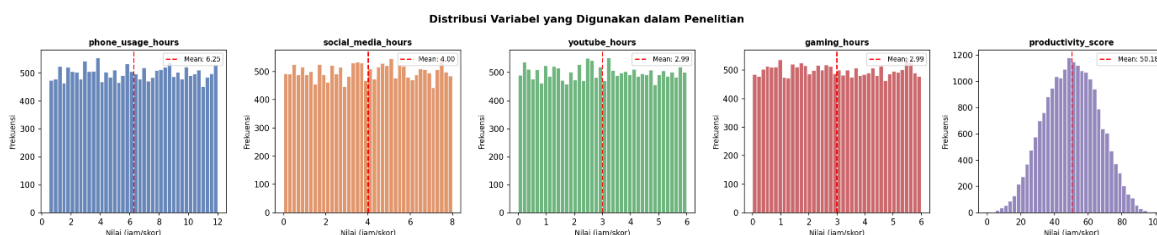
Dataset yang digunakan terdiri dari 20.000 data siswa dengan lima variabel yang relevan terhadap penelitian ini. Pemeriksaan awal menunjukkan bahwa dataset tidak memiliki nilai hilang (missing values) maupun duplikasi data, sehingga seluruh 20.000 baris dapat langsung digunakan tanpa penghapusan data. Statistik deskriptif seluruh variabel yang digunakan disajikan pada Tabel 3.

Tabel 3. Statistik Deskriptif Variabel Penelitian

Variabel	Min	Maks	Rata-rata	Std Dev	Median
<i>phone_usage_hours</i>	0,50	12,00	6,25	3,31	6,24
<i>social_media_hours</i>	0,00	8,00	4,00	2,31	4,01
<i>youtube_hours</i>	0,00	6,00	2,99	1,73	2,98
<i>gaming_hours</i>	0,00	6,00	2,99	1,73	2,97
<i>productivity_score</i>	0,00	100,00	50,18	16,09	50,24

Berdasarkan Tabel 3, variabel *phone_usage_hours* memiliki rentang nilai paling luas (0,50–12,00 jam) dengan rerata 6,25 jam per hari, menjadikannya variabel distraksi digital dengan intensitas penggunaan tertinggi di antara keempat prediktor. Variabel *social_media_hours* memiliki rata-rata 4,00 jam, sedangkan *youtube_hours* dan *gaming_hours* keduanya memiliki rerata yang hampir identik yaitu sekitar 2,99 jam per hari. Variabel target *productivity_score* terdistribusi cukup simetris dengan rerata 50,18 dan median 50,24, menunjukkan distribusi yang mendekati normal di sekitar nilai tengah rentang (0–100). Standar deviasi *productivity_score* sebesar 16,09 mengindikasikan variasi produktivitas yang cukup besar antar siswa dalam dataset ini.

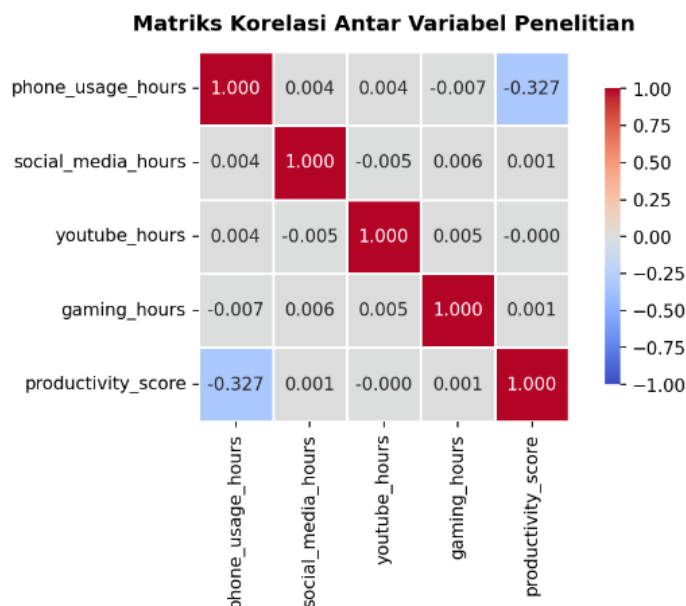
Distribusi keempat variabel prediktor divisualisasikan pada Gambar 2. Distribusi *phone_usage_hours* cenderung datar (uniform) dengan nilai yang tersebar merata dari 0 hingga 12 jam. Distribusi yang serupa juga terlihat pada *social_media_hours*, *youtube_hours*, dan *gaming_hours*, yang masing-masing tersebar merata dalam rentangnya. Pola distribusi yang seragam pada keempat variabel prediktor ini mengindikasikan bahwa data siswa dalam dataset terwakili secara proporsional di semua tingkat intensitas penggunaan digital. Sebaliknya, variabel *productivity_score* menunjukkan distribusi yang lebih mendekati normal dengan puncak di sekitar nilai 50.



Gambar 2. Distribusi Variabel yang Digunakan dalam Penelitian

3.2. Analisis Korelasi Antar Variabel

Sebelum membangun model prediksi, dilakukan analisis korelasi untuk memahami hubungan linear antara masing-masing variabel prediktor dengan variabel target. Hasil matriks korelasi disajikan pada Gambar 3.



Gambar 3. Matriks Korelasi Antar Variabel Penelitian

Gambar 3 mengungkapkan temuan yang sangat penting mengenai hubungan antar variabel dalam dataset ini. Dari keempat variabel distraksi digital, hanya *phone_usage_hours* yang memiliki korelasi bermakna terhadap *productivity_score*, yaitu sebesar -0,327. Nilai korelasi negatif ini menunjukkan bahwa semakin tinggi durasi penggunaan ponsel per hari, semakin rendah skor produktivitas siswa, sebuah temuan yang konsisten dengan penelitian terdahulu tentang dampak distraksi digital [7], [8]. Tiga variabel lainnya *social_media_hours* ($r = 0,001$), *youtube_hours* ($r = -0,000$), dan *gaming_hours* ($r = 0,001$) menunjukkan korelasi yang mendekati nol terhadap *productivity_score*, mengindikasikan hubungan linear yang hampir tidak ada antara ketiga variabel tersebut dengan produktivitas dalam dataset ini.

Di sisi lain, korelasi antar variabel prediktor sendiri juga sangat rendah (semua di bawah 0,01), yang berarti tidak terdapat masalah multikolinearitas. Temuan korelasi ini memberikan indikasi awal bahwa kemampuan prediksi model kemungkinan akan terbatas, karena sebagian besar variabel prediktor tidak memiliki hubungan linear yang kuat dengan variabel target.

3.3. Hasil Pra-Pemrosesan Data

Tahap pra-pemrosesan diawali dengan pemeriksaan missing values dan duplikasi. Hasilnya menunjukkan bahwa tidak ditemukan missing values maupun data duplikat pada kelima kolom yang digunakan, sehingga tidak ada data yang dihapus pada tahap ini. Selanjutnya, pembagian data dilakukan dengan proporsi 80:20, menghasilkan 16.000 data pelatihan dan 4.000 data pengujian. Normalisasi Min-Max Scaling kemudian diterapkan pada data pelatihan dan hasilnya digunakan untuk mentransformasi data pengujian, guna menghindari kebocoran informasi (data leakage). Setelah normalisasi, seluruh nilai fitur berada dalam rentang [0, 1].

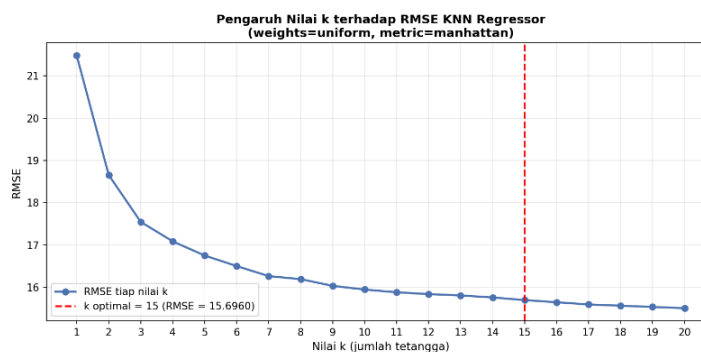
3.4. Hasil Optimasi Parameter Grid Search Cross-validation

Optimasi parameter dilakukan menggunakan Grid Search dengan 5-Fold Cross-validation pada data pelatihan. Untuk KNN Regressor, Grid Search mengevaluasi 28 kombinasi parameter (7 nilai $k \times 2$ weights $\times 2$ metric), sehingga total 140 iterasi pelatihan dilakukan. Untuk SVR, Grid Search mengevaluasi 128 kombinasi parameter (4 C $\times 4$ epsilon $\times 2$ kernel $\times 4$ gamma), sehingga total 640 iterasi pelatihan dilakukan. Hasil parameter optimal yang diperoleh disajikan pada Tabel 4.

Tabel 4. Parameter Optimal Hasil *Grid Search Cross-validation*

Algoritma	Parameter	Nilai Optimal
KNN Regressor	n_neighbors (k)	15
KNN Regressor	weights	uniform
KNN Regressor	metric	manhattan
SVR	C (regularisasi)	1
SVR	epsilon (ϵ)	0,5
SVR	kernel	linear
SVR	gamma (γ)	scale

Berdasarkan Tabel 4, KNN Regressor menghasilkan konfigurasi optimal dengan $k = 15$, pembobotan seragam (uniform), dan metrik jarak Manhattan. Nilai $k = 15$ berada di batas atas ruang pencarian yang didefinisikan ($k = 3$ hingga 15), yang sejalan dengan grafik pengaruh nilai k pada Gambar 4 yang menunjukkan tren penurunan RMSE seiring meningkatnya nilai k tanpa tanda-tanda kenaikan kembali (overfitting). Hal ini mengindikasikan bahwa dengan dataset sebesar 20.000 data, model KNN cenderung memerlukan lebih banyak tetangga untuk menghasilkan rata-rata prediksi yang lebih stabil dan kurang terpengaruh oleh noise individual.

Gambar 4. Pengaruh Nilai k terhadap RMSE KNN Regressor ($weights=uniform$, $metric=manhattan$)

Dari Gambar 4 terlihat bahwa RMSE turun drastis dari $k = 1$ ($RMSE \approx 21,4$) hingga $k = 5$ ($RMSE \approx 17,0$), kemudian penurunan melambat dan semakin landai setelah $k = 10$. Kurva yang terus menurun tanpa membentuk titik minimum yang jelas dalam rentang $k = 1$ hingga 20 ini mengisyaratkan bahwa peningkatan nilai k di atas 15 masih berpotensi menghasilkan RMSE yang lebih rendah. Kondisi ini merupakan salah satu keterbatasan dari ruang pencarian parameter yang telah ditetapkan.

Untuk SVR, konfigurasi optimal diperoleh dengan kernel linear, $C = 1$, $\epsilon = 0,5$, dan $\gamma = scale$. Pemilihan kernel linear oleh Grid Search merupakan temuan yang signifikan, karena mengkonfirmasi bahwa hubungan antara variabel distraksi digital dan skor produktivitas dalam dataset ini lebih bersifat linear daripada non-linear. Nilai $C = 1$ menunjukkan keseimbangan antara kerataan fungsi regresi dan toleransi terhadap kesalahan, sementara nilai $\epsilon = 0,5$ berarti prediksi yang berada dalam rentang $\pm 0,5$ dari nilai aktual tidak dihitung sebagai kesalahan sebuah toleransi yang relatif ketat mengingat rentang produktivitas 0–100.

3.5. Perbandingan Performa KNN Regressor dan SVR

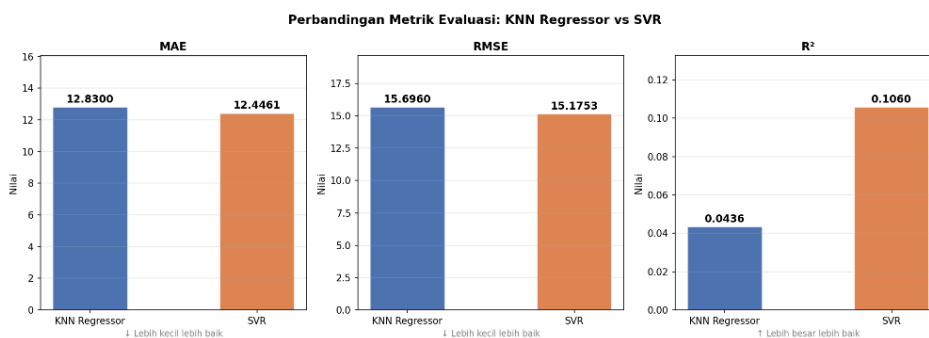
Setelah parameter optimal diperoleh, kedua model dilatih ulang menggunakan seluruh data pelatihan (16.000 data) dengan konfigurasi optimal masing-masing, kemudian dievaluasi pada data pengujian (4.000 data) yang tidak pernah dilihat selama proses pelatihan maupun validasi silang. Hasil evaluasi lengkap disajikan pada Tabel 5.

Tabel 5. Parameter Optimal Hasil *Grid Search Cross-validation*

Metrik	KNN Regressor	SVR	Model Terbaik
MAE	12,8300	12,4461	SVR ($\downarrow 3,0\%$)
RMSE	15,6960	15,1753	SVR ($\downarrow 3,3\%$)
R^2	0,0436	0,1060	SVR ($\uparrow 143,1\%$)

Berdasarkan Tabel 5, SVR mengungguli KNN Regressor pada ketiga metrik evaluasi secara konsisten. SVR menghasilkan MAE sebesar 12,4461 dibandingkan KNN sebesar 12,8300, dengan selisih 3,0%. Untuk RMSE, SVR menghasilkan nilai 15,1753 dibandingkan KNN sebesar 15,6960, dengan selisih 3,3%. Kecilnya selisih MAE

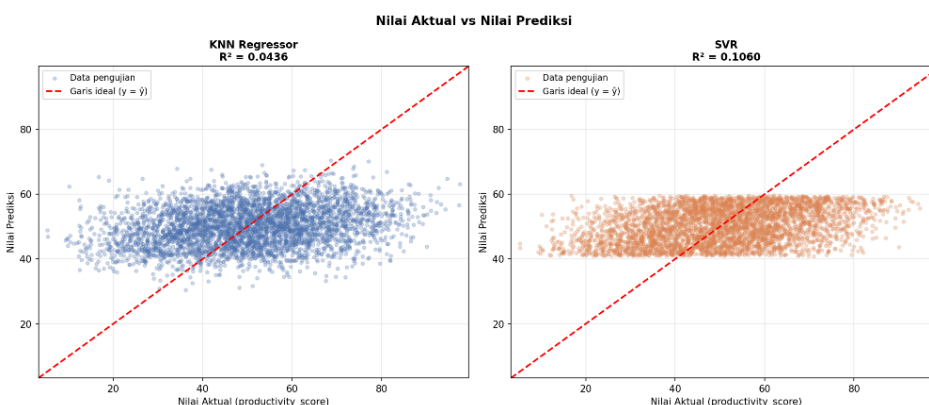
dan RMSE ini disebabkan karena nilai korelasi rendah yang dipengaruhi oleh ketidakkonsistenan dataset. Hal ini menyebabkan kedua model cenderung memprediksi nilai mendekati rata-rata populasi. Selisih yang kecil juga ditemukan pada studi komparatif KNN-SVR sebelumnya yang telah dilakukan oleh Wahyusari[10]. Perbedaan hasil yang lain terlihat pada metrik R^2 , di mana SVR menghasilkan nilai 0,1060 sementara KNN hanya 0,0436. Secara persentase SVR mampu menjelaskan proporsi variansi data yang 143,1% lebih besar dibandingkan KNN. R^2 tetap menunjukkan perbedaan yang lebih jelas karena metrik ini mengukur kemampuan model dalam mengikuti variasi data secara keseluruhan bukan sekadar seberapa besar rata-rata kesalahannya. Nilai R^2 kedua model masih tergolong rendah, yang menunjukkan bahwa kemampuan model dalam menjelaskan variasi produktivitas siswa masih terbatas dan kemungkinan terdapat faktor-faktor lain yang lebih dominan memengaruhi produktivitas siswa. Perbandingan visual ketiga metrik tersebut disajikan pada Gambar 5.



Gambar 5. Perbandingan Metrik Evaluasi KNN Regressor dan SVR (MAE, RMSE, R^2)

3.6. Analisis Nilai Aktual vs Nilai Prediksi

Analisis lebih mendalam dilakukan melalui visualisasi scatter plot nilai aktual versus nilai prediksi untuk kedua model, sebagaimana ditunjukkan pada Gambar 6.



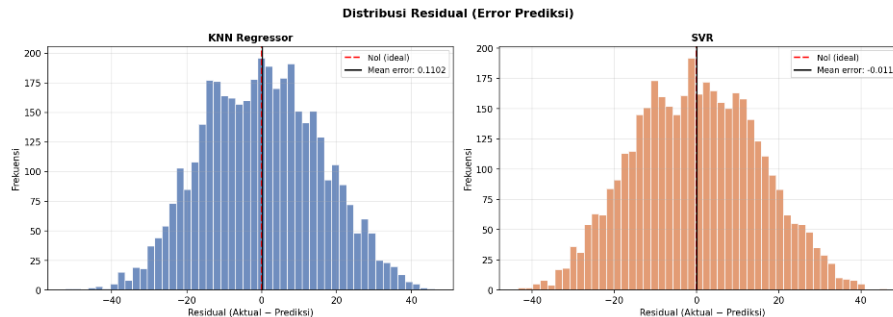
Gambar 6. Scatter plot Nilai Aktual vs Nilai Prediksi (Kiri: KNN Regressor, Kanan: SVR)

Gambar 6 mengungkapkan karakteristik prediksi kedua model secara visual. Pada KNN Regressor (panel kiri), titik-titik data membentuk pola menyebar yang cukup lebar di sekitar garis ideal ($y = \hat{y}$), namun masih menunjukkan tren naik yang terlihat model memang menangkap sebagian pola bahwa nilai aktual tinggi cenderung menghasilkan prediksi yang lebih tinggi pula. Nilai $R^2 = 0,0436$ mengkonfirmasi bahwa hanya sekitar 4,36% variansi produktivitas yang berhasil dijelaskan oleh KNN.

Pada SVR (panel kanan), pola prediksi terlihat sangat berbeda: titik-titik data membentuk pita horizontal yang relatif datar dalam rentang prediksi sekitar 40–60, hampir tidak bergerak seiring perubahan nilai aktual. Ini adalah karakteristik khas model regresi dengan R^2 yang rendah model cenderung memprediksi nilai mendekati rata-rata populasi (sekitar 50) untuk hampir semua data pengujian, tanpa mampu membedakan dengan baik antara siswa dengan produktivitas rendah (di bawah 30) dan tinggi (di atas 70). Meskipun nilai R^2 SVR (0,1060) lebih tinggi dari KNN, pola "pita datar" ini menunjukkan bahwa SVR pun memiliki keterbatasan serupa dalam menangkap variasi produktivitas yang sesungguhnya.

3.7. Analisis Distribusi Residual

Analisis residual dilakukan untuk memahami pola kesalahan prediksi masing-masing model. Distribusi residual (selisih antara nilai aktual dan nilai prediksi) kedua model ditampilkan pada Gambar 7.



Gambar 7. Scatter plot Nilai Aktual vs Nilai Prediksi (Kiri: KNN Regressor, Kanan: SVR)

Gambar 7 menunjukkan bahwa distribusi residual KNN Regressor memiliki mean error sebesar 0,1102, sangat mendekati nol, yang mengindikasikan tidak adanya bias sistematis yang signifikan dalam prediksi. Distribusi residualnya cukup simetris di sekitar nol meskipun dengan sebaran yang lebar (rentang -40 hingga +40). Pada SVR, mean error bahkan lebih kecil yaitu -0,0117, mendekati nol hampir sempurna, menandakan SVR hampir tidak memiliki bias dalam prediksi. Distribusi residual SVR juga lebih simetris dan mendekati distribusi normal dibandingkan KNN, meskipun sebarannya sama-sama lebar. Keduanya menunjukkan bahwa kesalahan prediksi bukan berasal dari bias model, melainkan dari variansi yang tinggi model tidak cukup mampu menjelaskan variasi produktivitas yang ada..

3.8. Pembahasan

Hasil penelitian ini mengungkapkan dua temuan utama yang saling berkaitan. Pertama, SVR dengan kernel linear terbukti lebih unggul dari KNN Regressor dalam memprediksi skor produktivitas siswa berdasarkan variabel distraksi digital, konsisten pada ketiga metrik evaluasi (MAE, RMSE, dan R^2). Temuan ini sejalan dengan Amellia Kharis et al.[12] dan Ghosh & Gor[13] yang menemukan SVR secara konsisten mengungguli KNN pada data yang memiliki hubungan linear, serta dengan Su et al.[15] yang menunjukkan bahwa SVR dengan optimasi parameter mampu menghasilkan prediksi yang lebih akurat dan stabil. Walau begitu, selisih MAE dan RMSE antara keduanya di bawah 4%, sehingga perbedaan yang paling bermakna justru ada pada R^2 SVR mampu menjelaskan lebih dari dua kali lipat variansi data dibanding KNN, meski keduanya masih berada di kisaran yang lemah. Keunggulan SVR dapat dijelaskan oleh pemilihan *kernel linear* yang mengkonfirmasi sifat hubungan linear dalam data, dan mekanisme *epsilon-tube* yang membuatnya lebih tahan terhadap *noise*.

Kedua, nilai R^2 yang sangat rendah pada kedua model (KNN: 0,0436; SVR: 0,1060) merupakan temuan yang lebih substansial. Nilai ini mengindikasikan bahwa hanya 4,4% (KNN) dan 10,6% (SVR) variansi *productivity_score* yang berhasil dijelaskan oleh keempat variabel distraksi digital. Hal ini dapat dipahami melalui hasil analisis korelasi pada Subbab 3.2: hanya *phone_usage_hours* yang memiliki korelasi bermakna ($r = -0,327$), sementara *social_media_hours*, *youtube_hours*, dan *gaming_hours* hampir tidak berkorelasi ($r \approx 0,001$). Raj[9] menemukan bahwa kesehatan mental, pola tidur, dan tingkat stres merupakan prediktor produktivitas yang jauh lebih kuat, sementara Martin et al.[1] menyimpulkan bahwa dampak distraksi digital dimediasi oleh berbagai faktor kontekstual yang tidak tercermin dalam durasi penggunaan semata.

Distribusi variabel prediktor yang seragam juga memberi sinyal bahwa dataset ini kemungkinan bersifat sintesis, sehingga kekuatan prediktif model yang rendah sebagian besar bersumber dari hal tersebut. Terlepas dari itu, penelitian ini memberi dua hal yang berguna. dari sisi perbandingan algoritma, SVR dengan kernel linear dan optimasi Grid Search CV lebih cocok dipilih dibanding KNN Regressor untuk kasus serupa. dan dari sisi substantif, temuan ini menunjukkan bahwa durasi penggunaan platform digital saja, kecuali penggunaan ponsel secara keseluruhan tidak cukup kuat untuk memprediksi produktivitas siswa, sehingga penelitian ke depan perlu mempertimbangkan variabel yang lebih beragam.

4. KESIMPULAN

SVR dengan konfigurasi kernel linear, $C = 1$, $\epsilon = 0,5$, dan $\gamma = \text{scale}$ terbukti lebih unggul dibandingkan KNN Regressor dengan $k = 15$, $\text{weights} = \text{uniform}$, dan $\text{metric} = \text{manhattan}$ dalam memprediksi skor produktivitas siswa berbasis distraksi digital, dengan SVR menghasilkan MAE = 12,4461, RMSE = 15,1753,

dan $R^2 = 0,1060$ dibandingkan KNN yang menghasilkan $MAE = 12,8300$, $RMSE = 15,6960$, dan $R^2 = 0,0436$. Namun demikian, nilai R^2 yang rendah pada kedua model mengungkapkan temuan yang lebih substansial: keempat variabel distraksi digital yang digunakan *phone_usage_hours*, *social media hours*, *youtube hours*, dan *gaming_hours* hanya mampu menjelaskan sebagian kecil variansi produktivitas siswa, sebagaimana dikonfirmasi oleh korelasi yang bermakna hanya pada *phone_usage_hours* ($r = -0,327$) sementara tiga variabel lainnya mendekati nol. Temuan ini membuktikan secara empiris bahwa produktivitas siswa merupakan konstruk multidimensi yang membutuhkan cakupan variabel yang lebih komprehensif mencakup aspek psikologis, kebiasaan belajar, dan kesehatan untuk dapat dimodelkan secara prediktif dengan tingkat akurasi yang memadai.

Penelitian ini memiliki keterbatasan pada dataset yang kemungkinan bersifat sintetis, sebagaimana ditunjukkan oleh distribusi variabel prediktor yang seragam dan korelasi antar variabel yang sangat rendah, serta ruang pencarian nilai k pada KNN yang dibatasi hingga 15 padahal RMSE masih terus menurun. Untuk itu, penelitian selanjutnya disarankan mengintegrasikan variabel yang lebih prediktif seperti *stress_level*, *sleep_hours*, dan *focus_score* ke dalam model, menggunakan dataset primer dari survei langsung, serta memperluas ruang pencarian parameter dan mengeksplorasi teknik feature selection guna memperoleh model prediksi produktivitas siswa yang lebih akurat dan representatif.

DAFTAR PUSTAKA

- [1] F. Martin, S. Long, K. Haywood, and K. Xie, "Digital distractions in education: a systematic review of research on causes, consequences and prevention strategies," *Educ. Technol. Res. Dev.*, vol. 73, no. 6, pp. 3423–3451, 2025, doi: 10.1007/s11423-025-10550-6.
- [2] B. Gö, U. Özbek, and M. B. Horzum, "Digital distraction levels of university students in emergency remote teaching," *Educ. Inf. Technol.*, vol. 28, no. 7, pp. 9149–9170, 2023, doi: 10.1007/s10639-022-11570-y.
- [3] F. Zike, A. Kamelia, B. M. Lovely, and S. Zulfani, "Analisis Perubahan Motivasi Belajar Siswa Di Tengah Tingginya Penggunaan Smartphone Pendahuluan," *KOLONI J. Multidisiplin Ilmu*, vol. 4, no. 4, pp. 39–50, 2025, [Online]. Available: <https://www.city.kawasaki.jp/500/page/0000174493.html>
- [4] Mawarni, D. Aidil Mahendra, N. Kamaly, W. Alqarni, and U. Nisa, "Pengaruh Media Sosial sebagai Pemicu Stres dan Distraksi Akademik : Studi Kasus pada Mahasiswa di Perguruan Tinggi Aceh," *J. Sos. Hukum, Polit. dan Hum.*, vol. 1, no. 2, pp. 15–29, 2025.
- [5] I. Julfan and Haifaturrahmah, "Dampak Penggunaan Gadget terhadap Konsentrasi Belajar Siswa The Impact of Gadget Use on Students ' Learning Concentration," *Action Res. J. Indones.*, vol. 7, no. 4, pp. 3310–3326, 2025, [Online]. Available: <https://journal.nahnuinisiatif.com/index.php/ARJI> Volume
- [6] N. G. Saragi, Firman, and Sahiruddin, "TIKTOK SEBAGAI DISTRAKSI, MEMPENGARUHI PRODUKTIVITAS DAN MANAJEMEN WAKTU PADA GENERASI Z," *J. Teknol. dan Sist. Inf. Univrab*, vol. 11, no. 1, pp. 1709–1718, 2026.
- [7] P. J. Barwick, S. Chen, C. Fu, and T. Li, "Digital Distractions with Peer Influence: The Impact of Mobile App Usage on Academic and Labor Market Outcomes," *Q. J. Econ.*, vol. 141, no. 1, pp. 1–49, 2026, doi: 10.1093/qje/qjaf048.
- [8] M. Á. Pérez-Juárez, D. González-Ortega, and J. M. Aguiar-Pérez, "Digital Distractions from the Point of View of Higher Education Students," *Sustain.*, vol. 15, no. 7, 2023, doi: 10.3390/su15076044.
- [9] R. Raj, "Mental Health and Productivity Among Students," *Int. J. Sci. Res. Eng. Manag.*, vol. 09, no. 05, pp. 1–9, 2025, doi: 10.55041/ijrsrem47745.
- [10] R. Wahyusari, "Perbandingan K-Nearest Neighbors (Knn) Dan Support Vector Regression (Svr) Untuk Prediksi Konsumsi Energi Listrik," *Pros. Snast*, no. November, pp. G44-50, 2024, doi: 10.34151/prosidingsnast.v1i1.5034.
- [11] B. Hamdikatama, K. Kusri, and A. Setyanto, "Comparison of The Performance of SVR, KNN and Decision Tree Methods in Predicting Rice Production," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 12, no. 1, pp. 90–105, 2025, doi: 10.35957/jatisi.v12i1.10133.
- [12] S. A. Amellia Kharis, A. H. Anna Zili, M. Malik, W. Nuryaningrum, and A. Putri, "Comparing machine learning models for Indonesia stock market prediction," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 38, no. 1, p. 508, 2025, doi: 10.11591/ijeecs.v38.i1.pp508-516.
- [13] M. Ghosh and R. Gor, "Stock Price Prediction Using Support Vector Regression and K-Nearest Neighbors: a Comparison," *Int. J. Eng. Sci. Technol.*, vol. 6, no. 4, pp. 1–9, 2022, doi: 10.29121/ijoeest.v6.i4.2022.354.
- [14] N. A. I. Lestari, Erwin Hermawan, Sahid Agustian Hudjimartu, Arif K Wijayanto, Ahmad Junaedi, and

-
- Miftahul B.R. Khamid, “Analisis Produktivitas Padi Menggunakan UAV Multispektral Dan Support Vector Regression Dengan Optimasi Grid Search Di Balumbang Jaya, Bogor Barat,” *J. Geogr. Edukasi dan Lingkung.*, vol. 9, no. 2, pp. 180–193, 2025, doi: 10.22236/jgel.v9i2.17257.
- [15] X. Su, X. He, G. Zhang, Y. Chen, and K. Li, “Research on SVR Water Quality Prediction Model Based on Improved Sparrow Search Algorithm,” *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/7327072.
- [16] Y. Karsavran, “Comparison of Random Forest, SVR and KNN Based Models in Sea Level Prediction for Erdemli Coast of Mersin,” *Celal Bayar Üniversitesi Fen Bilim. Derg.*, vol. 20, no. 2, pp. 14–18, 2024, doi: 10.18466/cbayarfbe.1384547.
- [17] R. S. Daulay, “Analisis Kritis dan Pengembangan Algoritma K-Nearest Neighbor (KNN): Sebuah Tinjauan Literatur,” *J. Pendidik. Sains dan Komput.*, vol. 4, no. 02, pp. 131–141, 2024, doi: 10.47709/jpsk.v4i02.5055.