Journal of Digital Community Services
Volume 3, Nomor 2, Tahun 2026
Halaman: 375 - 383
First published: 14 – 11 - 2026

Clustering Pola Tidur dan Kebiasaan Belajar Siswa untuk Segmentasi Strategi Belajar dan Prediksi Performa Ujian

Dedy Wahyudi[1], Moh. Zainuddin[2], Sunu Jatmika[3]
[1,2]Teknik Informatika, Fakultas Teknologi dan Desain, Institut Teknologi dan Bisnis Asia Malang, Indonesia
[3]Sistem Komputer, Fakultas Teknologi dan Desain, Institut Teknologi dan Bisnis Asia Malang, Indonesia
*Email Author: dedykresnawahyudi@gmail.com[1]**

Kata Kunci:
Data Mining Pendidikan; K-Means; Prediksi Performa; Pola Tidur; Kebiasaan Belajar

Abstrak

Penelitian ini bertujuan mengidentifikasi segmentasi pola tidur dan kebiasaan belajar serta menguji kontribusinya terhadap prediksi performa ujian. Dataset terdiri atas 1.000 observasi dengan 10 atribut, tanpa nilai hilang dan duplikasi. Clustering K-Means diterapkan pada tujuh fitur perilaku dengan k=2–6 yang dievaluasi menggunakan silhouette, Calinski–Harabasz, dan Davies–Bouldin. Prediksi nilai dibandingkan antara mean baseline, regresi linear, Ridge, random forest, dan gradient boosting dengan validasi silang 10-fold. Solusi k=5 memperoleh silhouette tertinggi sebesar 0,125 dan Davies–Bouldin 1,869, menghasilkan lima profil: belajar rendah–tidur pendek; belajar intensif–stres dan media tinggi; tidur cukup–stres tinggi; tidur panjang–media sosial tinggi; dan belajar intensif–gaya hidup terkendali. Nilai antarkelompok berbeda signifikan (ANOVA F=66,459; p<0,001; η^2=0,211), tetapi silhouette rendah menunjukkan tumpang tindih yang kuat. Ridge memberikan prediksi terbaik (MAE=3,845; RMSE=4,442; R^2=0,567), sedangkan penambahan cluster secara fold-local tidak meningkatkan prediksi (RMSE=4,452; R^2=0,565). Dengan demikian, clustering relevan untuk menyusun strategi belajar tersegmentasi, tetapi label cluster tidak menambah informasi prediktif di luar fitur asli. Hasil ditafsirkan secara eksploratif karena provenance dataset tidak tersedia.

***Key Word**:*
Educational Data Mining; K-Means; Performance Prediction; Sleep Patterns; Study Habits

Abstract

This study aims to identify segments of sleep patterns and study habits and to examine their contribution to exam performance prediction. The dataset consists of 1,000 observations with 10 attributes, with no missing values or duplicates. K-Means clustering was applied to seven behavioral features with k=2–6, evaluated using the silhouette, Calinski–Harabasz, and Davies–Bouldin indices. Score prediction was compared across a mean baseline, linear regression, Ridge, random forest, and gradient boosting using 10-fold cross-validation. The k=5 solution obtained the highest silhouette (0.125) and a Davies–Bouldin index of 1.869, yielding five profiles: low study–short sleep; intensive study–high stress and media use; adequate sleep–high stress; long sleep–high social media use; and intensive study–controlled lifestyle. Scores differed significantly among clusters (ANOVA F=66.459; p<0.001; η^2=0.211), although the low silhouette indicates strong overlap. Ridge gave the best prediction (MAE=3.845; RMSE=4.442; R^2=0.567), while adding fold-local cluster labels did not improve prediction (RMSE=4.452; R^2=0.565). Clustering is therefore useful for designing segmented study strategies, but cluster labels add no predictive information beyond the original features. Results should be interpreted exploratorily because the provenance of the dataset is unavailable.

Copyright © 2026 Dedy Wahyudi
*This work is licensed under a **Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)***

PENDAHULUAN

Performa ujian tidak hanya dipengaruhi oleh kemampuan kognitif, tetapi juga oleh pola perilaku yang membentuk kesiapan belajar. Durasi belajar, keteraturan revisi, waktu belajar yang disukai, penggunaan pembelajaran daring, paparan media sosial, tingkat stres, dan tidur merupakan faktor yang saling berhubungan. Dalam praktik pendidikan, dua siswa dengan jumlah jam belajar yang sama dapat memperoleh hasil berbeda apabila salah satunya tidur cukup dan melakukan revisi secara teratur, sedangkan yang lain belajar pada kondisi stres tinggi dan mengorbankan waktu tidur.

Tidur berperan dalam perhatian, konsolidasi memori, regulasi emosi, dan pemulihan fisiologis. Kajian empiris menunjukkan adanya hubungan antara durasi atau kualitas tidur dan performa akademik, meskipun kekuatan dan arahnya dapat berbeda menurut desain penelitian, populasi, dan cara pengukuran (Musshafen et al., 2021; Zeek et al., 2015). Perbedaan tersebut menegaskan bahwa tidur tidak seharusnya ditafsirkan sebagai determinan tunggal. Tidur perlu dianalisis bersama kebiasaan belajar dan faktor gaya hidup agar profil kesiapan ujian lebih kontekstual.

Kebiasaan belajar juga bersifat multidimensional. Jam belajar yang panjang belum tentu efektif bila tidak disertai manajemen usaha, revisi terstruktur, pemilihan sumber belajar, dan pengendalian

**Corresponding author's e-mail: dedykresnawahyudi@gmail.com*

distraksi. Bickerdike et al. (2016) menemukan bahwa manajemen usaha serta pola belajar terorganisasi berkaitan dengan capaian akademik, sementara aktivitas jejaring sosial dapat berinteraksi dengan strategi belajar. Strategi belajar yang terorganisasi, pengelolaan usaha, dan revisi terjadwal merupakan indikator yang lebih informatif daripada total jam belajar saja (Bickerdike et al., 2016; Walck-Shannon et al., 2021). Media sosial dapat menyediakan sumber belajar dan interaksi, namun durasi yang berlebihan berpotensi menggantikan waktu tidur atau belajar. Hal ini mengarahkan penelitian dari sekadar mencari korelasi tunggal menuju identifikasi kombinasi perilaku.

Zeek et al. (2015) melaporkan keterkaitan pola tidur dengan performa pada mahasiswa, sedangkan tinjauan Musshafen et al. (2021) menunjukkan bahwa kualitas tidur dapat lebih konsisten berasosiasi dengan performa dibanding durasi tidur semata. Studi populasi besar terbaru juga menekankan bahwa hubungan tidur dan capaian dapat berbeda antarmata pelajaran (Wang et al., 2025). Literatur ini mendukung penggunaan durasi tidur sebagai variabel penting, tetapi sekaligus menunjukkan keterbatasan dataset penelitian ini karena tidak memuat kualitas tidur, konsistensi waktu tidur, kronotipe, atau pengukuran objektif.

Educational data mining menyediakan dua sudut pandang yang saling melengkapi. Pertama, pembelajaran tanpa supervisi dapat menemukan kelompok siswa yang memiliki pola perilaku serupa tanpa menggunakan nilai ujian sebagai penentu cluster. Kedua, pembelajaran terawasi dapat mengukur seberapa jauh pola tersebut memprediksi nilai. K-Means membagi observasi ke dalam k kelompok dengan meminimalkan jumlah kuadrat jarak dalam cluster (MacQueen, 1967), dan kualitas hasilnya dapat diukur menggunakan koefisien silhouette (Rousseeuw, 1987) serta indeks Calinski–Harabasz dan Davies–Bouldin. K-Means banyak digunakan karena sederhana, transparan, dan mudah diterjemahkan menjadi profil intervensi; dalam konteks pendidikan, clustering digunakan untuk mengenali pola keterlibatan atau performa serta mendukung personalisasi (Alawi et al., 2023; Mohd Talib et al., 2023). Di sisi prediksi, tinjauan sistematis menunjukkan penggunaan regresi, decision tree, random forest, support vector machine, dan jaringan saraf untuk identifikasi siswa berisiko (Albreiki et al., 2021; Zhang et al., 2021). Namun, akurasi tinggi tidak otomatis menjamin manfaat pendidikan; interpretabilitas, validasi eksternal, dan pencegahan bias tetap diperlukan.

Masalah yang sering muncul dalam penelitian sejenis ialah cluster langsung diberi label “baik” atau “buruk” tanpa menilai validitas internal, sedangkan model prediksi dilaporkan hanya dari satu pembagian data. Praktik tersebut berisiko menghasilkan segmentasi yang rapuh dan estimasi performa yang terlalu optimistis. Penelitian ini menggunakan evaluasi beberapa nilai k, standardisasi, one-hot encoding, pengujian stabilitas inisialisasi, perbandingan beberapa baseline prediksi, dan validasi silang 10-fold. Nilai ujian dikeluarkan dari proses clustering agar profil terbentuk dari perilaku, bukan dari hasil yang hendak dijelaskan.

Kontribusi penelitian ini meliputi enam hal. Pertama, mengintegrasikan segmentasi tanpa supervisi dan prediksi terawasi dalam satu kerangka analisis perilaku tidur–belajar. Kedua, membentuk cluster tanpa memasukkan nilai ujian sehingga interpretasi profil tidak mengalami kebocoran target. Ketiga, menguji k=2 sampai k=6 dengan tiga indeks internal dan menilai stabilitas label terhadap perubahan seed. Keempat, membandingkan mean baseline, regresi linear, Ridge, random forest, dan gradient boosting menggunakan validasi silang 10-fold. Kelima, menguji secara khusus apakah label cluster memberi peningkatan prediksi di luar fitur asli melalui pembentukan cluster pada setiap fold. Keenam, menerjemahkan profil cluster menjadi strategi belajar yang bersifat pendampingan, bukan pelabelan deterministik.

Pertanyaan penelitian adalah: “Bagaimana penerapan metode clustering berdasarkan pola tidur dan kebiasaan belajar siswa dapat menghasilkan segmentasi strategi belajar yang relevan serta mendukung prediksi performa ujian siswa?” Pertanyaan tersebut dioperasionalkan menjadi empat sasaran: menentukan jumlah cluster yang paling defensibel; mendeskripsikan karakter tiap cluster; membangun model prediksi nilai; dan menilai kontribusi tambahan cluster terhadap prediksi.

METODE

Data Penelitian

Dataset aktual yang dianalisis terdiri atas 1.000 observasi dan 10 atribut: usia, gender, jam belajar, frekuensi revisi, waktu belajar yang disukai, penggunaan pembelajaran daring, jam media sosial, jam tidur, stres ujian, dan nilai ujian terakhir. Peran setiap variabel dalam analisis dirangkum pada Tabel 1. Rentang usia 17–24 tahun menunjukkan populasi pelajar akhir sekolah hingga pendidikan tinggi.

e-ISSN: 3046-7527

Dengan keterbatasan metadata tersebut, kata "siswa" digunakan dalam arti generik dan hasil tidak diklaim mewakili seluruh siswa SMA.

Tabel 1. Variabel Dataset dan Perannya dalam Analisis

Variabel	Jenis	Rentang/kategori	Peran
age	Numerik	17–24	Prediktor kontrol
gender	Kategorikal	3 kategori	Prediktor kontrol
study_hours_per_day	Numerik	1–8	Cluster + prediksi
revision_frequency	Kategorikal	4 kategori	Cluster + prediksi
preferred_study_time	Kategorikal	4 kategori	Cluster + prediksi
uses_online_learning	Biner	Yes/No	Cluster + prediksi
social_media_hours_per_day	Numerik	0–6	Cluster + prediksi
sleep_hours_per_day	Numerik	4–9	Cluster + prediksi
exam_stress_level	Ordinal	1–5	Cluster + prediksi
last_exam_score_percent	Numerik	55,4–92,2	Target/evaluasi eksternal

Sumber: Dataset penelitian

Definisi Operasional

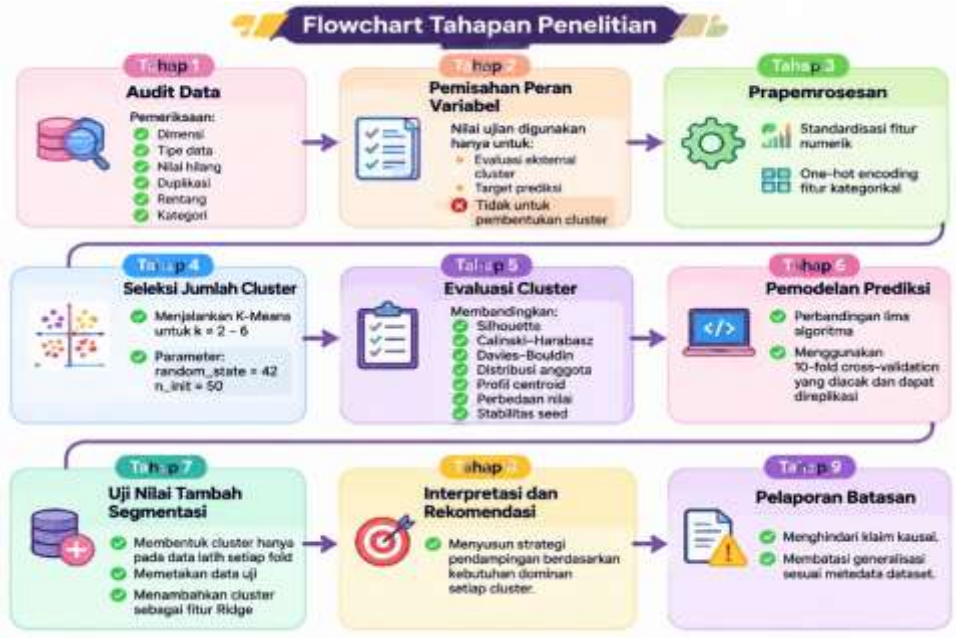
Tabel 2 merangkum definisi operasional setiap elemen penelitian, mulai dari pola tidur, kebiasaan belajar, hingga relevansi hasil segmentasi.

Tabel 2. Definisi Operasional Elemen Penelitian

Elemen	Definisi operasional
Pola tidur	Durasi tidur per hari (4–9 jam).
Kebiasaan belajar	Jam belajar, frekuensi revisi, waktu belajar pilihan, dan pembelajaran daring.
Konteks perilaku	Jam media sosial dan stres ujian.
Segmentasi	Kelompok K-Means berdasarkan fitur perilaku yang telah diskalakan/dikodekan.
Performa ujian	Persentase nilai ujian terakhir sebagai target kontinu.
Relevansi	Cluster dapat diinterpretasikan, berbeda secara deskriptif, dan menghasilkan rekomendasi yang masuk akal tanpa klaim kausal.

Sumber: Penulis

Tahapan Penelitian



Gambar 1. Flowchart Tahapan Penelitian

e-ISSN: 3046-7527

Landasan Metode

Untuk observasi x_i dan pusat cluster μ_j, K-Means meminimalkan fungsi objektif pada Persamaan (1) (MacQueen, 1967).

$$J = \Sigma_i \min_j \|x_i - \mu_j\|^2 \quad (1)$$

Dalam Persamaan (1), J adalah fungsi objektif yang diminimalkan, x_i adalah observasi ke-i, dan μ_j adalah pusat cluster ke-j. Karena jarak Euclidean sensitif terhadap skala, fitur numerik distandardisasi menurut Persamaan (2).

$$z = (x - \mu) / \sigma \quad (2)$$

Pada Persamaan (2), z adalah nilai terstandardisasi, x adalah nilai asli fitur, μ adalah rata-rata fitur, dan σ adalah simpangan baku fitur. Variabel kategorikal diubah menjadi indikator biner. Kombinasi ini memungkinkan implementasi yang praktis, walaupun jarak Euclidean atas one-hot encoding tidak selalu ideal bagi data campuran; keterbatasan ini menjadi alasan pengembangan berikutnya menggunakan k-prototypes atau Gower distance.

Silhouette untuk observasi i dirumuskan pada Persamaan (3).

$$s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\} \quad (3)$$

Pada Persamaan (3), $a(i)$ adalah jarak rata-rata dalam cluster dan $b(i)$ adalah jarak minimum ke cluster lain. Nilai mendekati 1 menunjukkan pemisahan baik, nilai sekitar 0 menunjukkan tumpang tindih, dan nilai negatif mengindikasikan kemungkinan salah penempatan (Rousseeuw, 1987). Calinski–Harabasz yang lebih tinggi dan Davies–Bouldin yang lebih rendah umumnya lebih baik.

Prediksi dievaluasi menggunakan MAE, RMSE, dan R^2. MAE mudah dibaca sebagai galat absolut rata-rata dalam poin nilai; RMSE memberi penalti lebih besar pada galat ekstrem; R^2 menyatakan proporsi variasi target yang dijelaskan relatif terhadap prediksi rata-rata. Ridge meminimalkan jumlah kuadrat galat ditambah penalti L2, sehingga koefisien lebih stabil ketika prediktor saling berkorelasi (Hoerl & Kennard, 1970).

Implementasi dan Perangkat Pengujian

Analisis dilakukan dengan Python dan pustaka ilmiah sumber terbuka, termasuk scikit-learn (Pedregosa et al., 2011). Seluruh proses menggunakan seed yang dinyatakan, sehingga dapat direplikasi pada file yang sama. Perangkat keras tidak menjadi faktor kritis karena skala data kecil; pengujian dapat dijalankan pada komputer standar dengan memori 8 GB. Model yang dibandingkan meliputi mean baseline, regresi linear, Ridge, random forest (Breiman, 2001), dan gradient boosting. Konfigurasi lengkap setiap tahap disajikan pada Tabel 3.

Tabel 3. Konfigurasi Implementasi

Tahap	Konfigurasi
Lingkungan	Python 3.12; pandas, NumPy, SciPy, scikit-learn, Matplotlib, Seaborn.
Fitur cluster	Jam belajar, jam media sosial, jam tidur, stres, frekuensi revisi, waktu belajar pilihan, pembelajaran daring.
Fitur prediksi	Seluruh fitur non-target: usia, gender, dan tujuh fitur perilaku.
K-Means	k=2–6; n_init=50; random_state=42.
Validasi prediksi	KFold 10 bagian; shuffle=True; random_state=42.
Model	Mean baseline, linear, Ridge α=1, random forest 500 pohon, gradient boosting 200 estimator.
Metrik	Silhouette, CH, DB; ARI; ANOVA/Kruskal–Wallis; MAE, RMSE, R^2.

Sumber: Penulis

Skema fold-local pada uji nilai tambah cluster mencegah data uji ikut menentukan pusat cluster. Preprocessor dan K-Means hanya dipasang pada bagian latih; data uji ditransformasikan dengan parameter latih dan dipetakan ke centroid terdekat. Dengan demikian, perbandingan Ridge tanpa dan dengan cluster lebih mendekati penggunaan nyata.

e-ISSN: 3046-7527

HASIL DAN PEMBAHASAN

Karakteristik Data

Berkas aktual berisi 1.000 observasi dan 10 variabel. Audit menemukan 0 nilai hilang dan 0 baris duplikat. Rata-rata usia 20,429 tahun; jam belajar 4,649 jam/hari; jam tidur 6,556 jam/hari; penggunaan media sosial 2,946 jam/hari; stres ujian 2,990 pada skala 1–5; dan nilai ujian 74,204±6,755 dengan rentang 55,4–92,2.

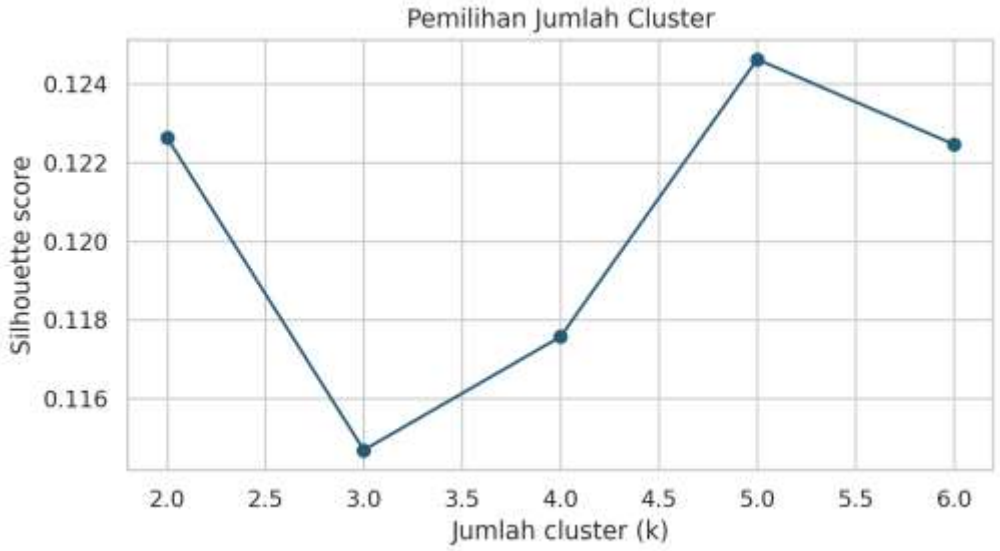
Hasil Clustering

Hasil evaluasi jumlah cluster untuk k=2–6 ditunjukkan oleh Tabel 4, sedangkan perbandingan nilai silhouette antar-k divisualisasikan pada Gambar 2.

Tabel 4. Evaluasi Jumlah Cluster Berdasarkan Indeks Validitas Internal

k	Silhouette	Calinski–Harabasz	Davies–Bouldin
2	0,123	148,27	2,566
3	0,115	134,35	2,222
4	0,118	126,96	2,075
5	0,125	122,05	1,869
6	0,122	114,35	1,889

Sumber: Hasil olah data penulis



Gambar 2. Perbandingan silhouette untuk k=2–6

Nilai silhouette tertinggi diperoleh pada k=5 (0,125), sedikit di atas k=2 (0,123) dan k=6 (0,122). Davies–Bouldin terbaik juga terdapat pada k=5 (1,869), sedangkan Calinski–Harabasz tertinggi pada k=2. Karena perbedaan antarsolusi kecil dan seluruh silhouette rendah, k=5 dipilih untuk granularitas interpretasi, tetapi tidak diklaim sebagai struktur cluster yang kuat. Stabilitas terhadap 20 seed menghasilkan adjusted Rand index rata-rata 0,827, minimum 0,563, dan maksimum 0,983; artinya hasil cukup stabil secara rata-rata, tetapi sebagian inisialisasi masih mengubah keanggotaan secara substansial.

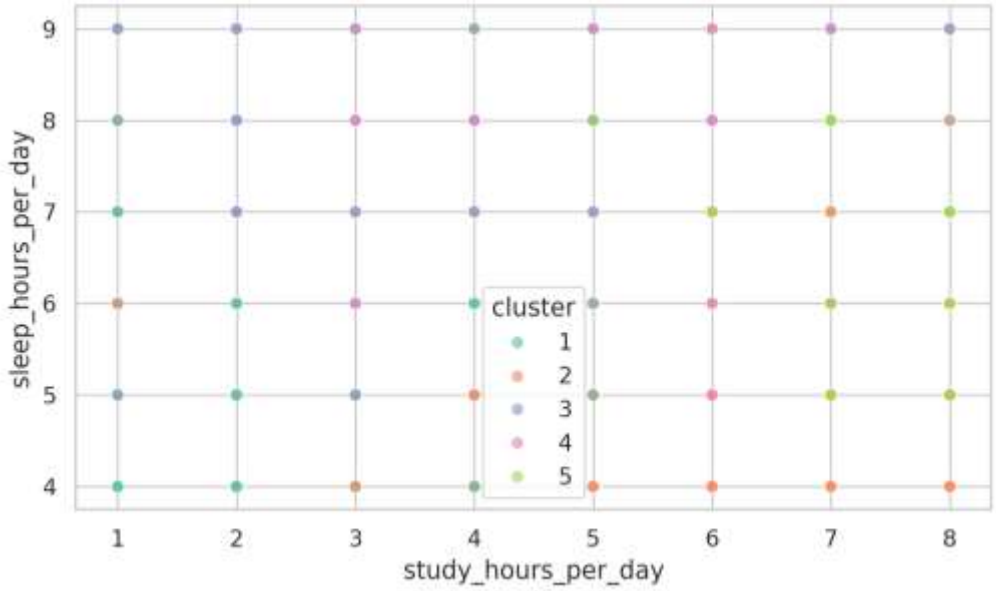
Karakteristik kelima cluster pada solusi k=5 dirangkum pada Tabel 5, dan sebaran cluster pada dimensi jam belajar dan jam tidur ditampilkan pada Gambar 3.

e-ISSN: 3046-7527

Tabel 5. Profil Cluster pada Solusi k=5 (Rata-rata per Cluster)

Cluster	n	Profil interpretatif	Belajar	Tidur	Media	Stres	Nilai
C1	190	Belajar rendah–tidur pendek	2,17	5,45	2,60	1,94	69,92
C2	213	Belajar intensif–stres dan media tinggi	5,94	5,18	4,18	4,00	72,91
C3	218	Tidur cukup–stres tinggi	3,69	7,36	1,58	4,31	73,48
C4	192	Tidur panjang–media sosial tinggi	4,81	7,95	5,06	2,38	75,50
C5	187	Belajar intensif–gaya hidup terkendali	6,65	6,87	1,33	2,00	79,55

Sumber: Hasil olah data penulis



Gambar 3. Proyeksi cluster pada jam belajar dan jam tidur

Cluster 1 memiliki nilai rata-rata terendah, jam belajar rendah, dan tidur relatif pendek. Strategi yang disarankan adalah membangun rutinitas belajar bertahap, target harian kecil, serta menambah durasi tidur. Cluster 2 belajar intensif tetapi mengalami stres tinggi, tidur pendek, dan penggunaan media sosial tinggi; intervensi lebih tepat berupa manajemen stres, pembatasan distraksi, dan perlindungan waktu tidur daripada menambah jam belajar.

Cluster 3 memiliki tidur relatif cukup dan media sosial rendah, tetapi stres tertinggi serta pola revisi dominan menjelang ujian. Strategi utamanya adalah mengubah revisi mendadak menjadi spaced practice dan menyediakan dukungan pengelolaan kecemasan. Cluster 4 tidur panjang dan stres lebih rendah, namun media sosial sangat tinggi serta revisi dominan jarang; strategi diarahkan pada peningkatan kualitas jam belajar dan penggunaan media sosial secara terjadwal. Cluster 5 menunjukkan kombinasi jam belajar tinggi, tidur mendekati cukup, media sosial rendah, dan stres rendah; profil ini memiliki nilai rata-rata tertinggi dan dapat diarahkan pada pemeliharaan kebiasaan, latihan pengambilan kembali, dan pencegahan kelelahan.

Perbedaan nilai antarkelompok signifikan secara statistik: ANOVA $F=66{,}459$, $p<0{,}001$, $\eta^2=0{,}211$; Kruskal–Wallis $H=209{,}485$, $p<0{,}001$. Eta squared menunjukkan sekitar 21,1% variasi nilai berkaitan dengan keanggotaan cluster secara deskriptif. Hasil ini bukan bukti kausal karena nilai tidak diacak dan variabel perancu tidak lengkap.

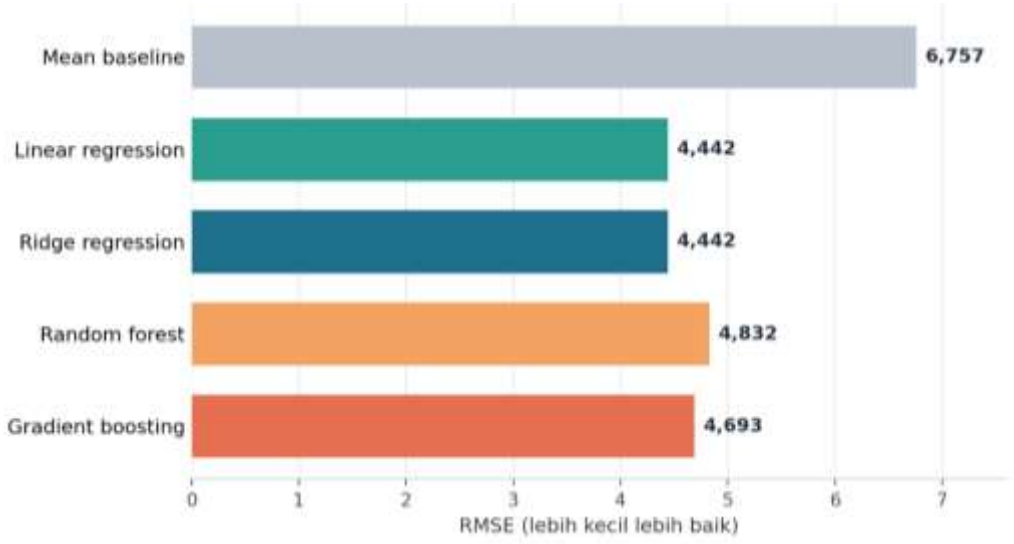
Hasil Prediksi

Kinerja kelima model prediksi pada validasi silang 10-fold dibandingkan pada Tabel 6, dan nilai RMSE setiap model divisualisasikan pada Gambar 4.

Tabel 6. Kinerja Model Prediksi pada Validasi Silang 10-Fold

Model	MAE	RMSE	R^2
Mean baseline	5,502	6,757	−0,002
Linear regression	3,845	4,442	0,567
Ridge regression	3,845	4,442	0,567
Random forest	4,112	4,832	0,488
Gradient boosting	4,034	4,693	0,517

Sumber: Hasil olah data penulis



Gambar 4. RMSE model pada validasi silang 10-fold

Ridge memberikan hasil terbaik dengan MAE 3,845, RMSE 4,442, dan R^2 0,567. Regresi linear hampir identik, menunjukkan hubungan dalam dataset terutama bersifat aditif dan mendekati linear. Random forest dan gradient boosting tidak mengungguli Ridge. Dibanding mean baseline, Ridge menurunkan RMSE sekitar 34,3%.

Uji fold-local menunjukkan Ridge tanpa cluster menghasilkan RMSE 4,442 dan R^2 0,567; setelah label cluster ditambahkan, RMSE menjadi 4,452 dan R^2 0,565. Dengan demikian, label cluster tidak meningkatkan akurasi dan justru menurunkan R^2 sekitar 0,002. Cluster tetap relevan untuk komunikasi segmentasi, tetapi informasi prediktifnya telah tercakup oleh fitur asli.

Permutation importance pada model Ridge menempatkan jam belajar sebagai fitur terpenting, diikuti jam tidur, frekuensi revisi, jam media sosial, penggunaan pembelajaran daring, stres ujian, dan waktu belajar pilihan. Usia serta gender memberi kontribusi sangat kecil. Korelasi bivariat dengan nilai adalah 0,531 untuk jam belajar, 0,329 untuk jam tidur, −0,194 untuk media sosial, dan −0,181 untuk stres. Korelasi tidak boleh ditafsirkan sebagai efek sebab-akibat.

Analisis Baseline

Mean baseline merepresentasikan sistem yang selalu memprediksi nilai rata-rata. RMSE 6,757 dan R^2 negatif menunjukkan bahwa baseline tidak dapat menjelaskan variasi individual. Regresi linear dan Ridge meningkatkan performa secara substansial, sedangkan ensemble nonlinier tidak lebih baik. Pola ini dapat terjadi ketika target pada dataset terbentuk terutama dari kombinasi linear beberapa fitur atau ketika ukuran data belum cukup untuk memanfaatkan fleksibilitas ensemble.

Baseline clustering perlu dibaca lebih hati-hati. Solusi k=5 unggul tipis pada silhouette dan Davies–Bouldin, tetapi silhouette 0,125 merupakan indikasi tumpang tindih tinggi. Karena data perilaku pendidikan lazim membentuk spektrum kontinu, tidak selalu terdapat batas alami yang tajam. Maka, cluster sebaiknya digunakan sebagai ringkasan navigasi untuk konselor atau guru, bukan sebagai dasar tunggal untuk keputusan berisiko tinggi.

Perbandingan dengan Penelitian Terdahulu

State of the art dalam educational data mining bergerak dari prediksi satu tahap menuju analitik yang dapat ditindaklanjuti, personalisasi, pembentukan fitur berbasis representasi, dan evaluasi yang

mencegah kebocoran data (Albreiki et al., 2021; Mohd Talib et al., 2023; Wu, 2026; Zhang et al., 2021). Penelitian ini berada pada jalur tersebut melalui integrasi cluster dan prediksi, namun menghasilkan temuan penting: segmentasi yang mudah dijelaskan tidak otomatis meningkatkan akurasi prediksi. Temuan negatif ini memperjelas perbedaan tujuan antara deskripsi kelompok dan optimasi galat.

Dibanding penelitian yang mengelompokkan siswa berdasarkan nilai, rancangan ini lebih kuat secara interpretatif karena nilai dikeluarkan dari pembentukan cluster. Profil kemudian divalidasi terhadap nilai secara eksternal. Dibanding penelitian prediksi yang hanya melaporkan satu model, lima baseline dan validasi silang memberi estimasi lebih dapat dipercaya. Uji cluster fold-local juga lebih ketat daripada menambahkan label cluster yang dibentuk dari seluruh data.

Namun, penelitian belum dapat menyamai pendekatan mutakhir yang menggunakan data longitudinal, log aktivitas, wearable sleep tracking, model temporal, explainable AI tingkat individu, serta validasi lintas institusi. Dataset hanya menangkap durasi harian dalam satu baris per siswa dan tidak mengandung waktu pencatatan relatif terhadap ujian. Oleh karena itu, kata "menjelang ujian" tidak dapat diverifikasi dari file aktual dan sebaiknya digunakan hanya jika dokumentasi sumber mengonfirmasi periode tersebut.

Implikasi Praktis

Implikasi praktis dari kelima profil, berupa risiko utama dan strategi belajar yang disarankan, dirangkum pada Tabel 7.

Tabel 7. Risiko Utama dan Strategi Belajar yang Disarankan per Profil

Profil	Risiko utama	Strategi belajar yang disarankan
C1	Belajar dan tidur rendah	Jadwal mikro 25–30 menit, target mingguan, sleep hygiene, pemantauan progres.
C2	Stres, media sosial, dan kurang tidur	Time blocking, batas gawai malam, latihan relaksasi, tanpa menambah jam belajar.
C3	Revisi mendadak dan stres tinggi	Spaced practice, low-stakes quiz, konseling kecemasan ujian.
C4	Media sosial tinggi dan revisi jarang	Pomodoro, aplikasi pemblokir distraksi, active recall, audit waktu layar.
C5	Perilaku relatif adaptif	Pemeliharaan kebiasaan, latihan soal bertingkat, pencegahan burnout.

Sumber: Hasil olah data penulis

Rekomendasi harus disampaikan sebagai pilihan dukungan dan dievaluasi ulang, bukan sebagai hukuman atau pengelompokan permanen. Data gender tidak digunakan untuk memberi strategi berbeda karena kontribusinya sangat kecil dan berpotensi menimbulkan stereotip. Siswa perlu diberi kesempatan mengoreksi data serta memahami tujuan analitik.

Keterbatasan dan Ancaman Validitas

Penelitian ini memiliki sejumlah keterbatasan. Provenance dataset, lokasi, teknik sampling, dan status data aktual atau sintetis tidak tersedia dalam file CSV. Rentang usia 17–24 tidak identik dengan populasi siswa SMA sehingga validitas eksternal terhadap SMA Indonesia belum terbukti. Data bersifat potong lintang dan self-report, sehingga hubungan yang ditemukan bukan kausal. Durasi tidur tidak mewakili kualitas, keteraturan, waktu tidur, atau kronotipe, dan tidak ada variabel kemampuan awal, kondisi sosial ekonomi, kesehatan, mata pelajaran, dan kualitas pengajaran. K-Means dengan one-hot encoding merupakan pendekatan praktis, tetapi bukan metrik ideal untuk data campuran. Silhouette rendah menunjukkan batas cluster lemah, dan nama profil merupakan interpretasi peneliti. Terakhir, nilai ujian terakhir mungkin tidak sebanding antarresponden bila berasal dari ujian yang berbeda.

KESIMPULAN

Penerapan K-Means pada pola tidur dan kebiasaan belajar menghasilkan lima profil: belajar rendah–tidur pendek; belajar intensif–stres dan media tinggi; tidur cukup–stres tinggi; tidur panjang–media sosial tinggi; serta belajar intensif–gaya hidup terkendali. Kelima cluster berbeda signifikan pada nilai ujian dan memiliki implikasi strategi belajar yang berbeda. Namun, silhouette 0,125 menandakan pemisahan lemah, sehingga cluster lebih tepat dipakai sebagai alat komunikasi dan eksplorasi daripada klasifikasi siswa yang kaku. Untuk prediksi, Ridge merupakan model terbaik dengan MAE 3,845, RMSE 4,442, dan R^2 0,567 pada validasi silang 10-fold. Jam belajar dan jam tidur menjadi dua fitur terpenting. Penambahan cluster tidak memperbaiki prediksi; RMSE meningkat dari 4,442 menjadi 4,452

dan R^2 turun dari 0,567 menjadi 0,565. Jawaban atas pertanyaan penelitian adalah bahwa clustering mampu menghasilkan segmentasi strategi belajar yang dapat ditafsirkan, tetapi dukungannya terhadap prediksi bersifat tidak langsung: fitur pembentuk cluster berguna untuk prediksi, sedangkan label cluster tidak menambah akurasi di luar fitur asli. Saran yang diajukan meliputi validasi kembali instrumen dan provenance data sebelum menggeneralisasi hasil; penggunaan cluster sebagai pemicu asesmen lanjutan oleh guru atau konselor, bukan keputusan otomatis; pengumpulan data siswa SMA Indonesia secara longitudinal pada beberapa minggu sebelum ujian; penambahan kualitas tidur, konsistensi jadwal, kronotipe, efikasi diri, kemampuan awal, dan konteks sekolah; serta pelaporan ketidakpastian prediksi dan audit fairness antarkelompok sebelum implementasi. Penelitian berikutnya dapat membandingkan k-prototypes, hierarchical clustering berbasis Gower distance, Gaussian mixture, HDBSCAN, dan latent profile analysis. Data longitudinal memungkinkan pemodelan trajectory atau sequence model untuk melihat perubahan tidur dan belajar mendekati ujian. Wearable atau sleep diary dapat mengurangi bias self-report. Pada sisi prediksi, nested cross-validation, kalibrasi interval prediksi, SHAP, serta external validation lintas sekolah diperlukan. Eksperimen terkontrol juga dapat menguji apakah strategi yang dipersonalisasi berdasarkan profil benar-benar meningkatkan performa dibanding rekomendasi umum.

DAFTAR REFERENSI

Alawi, S. J. S., Shaharanee, I. N. M., & Jamil, J. M. (2023). Clustering student performance data using K-means algorithms. *Journal of Computational Innovation and Analytics, 2*(1), 41–55. https://doi.org/10.32890/jcia2023.2.1.3

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. *Education Sciences, 11*(9), 552. https://doi.org/10.3390/educsci11090552

Bickerdike, A., O'Deasmhunaigh, C., O'Flynn, S., & O'Tuathaigh, C. (2016). Learning strategies, study habits and social networking activity of undergraduate medical students. *Advances in Health Sciences Education, 21*, 1021–1031. https://doi.org/10.1007/s10459-016-9689-0

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium* (pp. 281–297).

Mohd Talib, N. I., et al. (2023). Identification of student behavioral patterns in higher education using clustering and classification. *Applied Sciences, 13*(5), 3267. https://doi.org/10.3390/app13053267

Musshafen, L. A., et al. (2021). Associations between sleep and academic performance in US adolescents: A systematic review and meta-analysis. *Sleep Medicine, 83*, 71–82. https://doi.org/10.1016/j.sleep.2021.04.015

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Walck-Shannon, E. M., Rowell, S. F., & Frey, R. F. (2021). To what extent do study habits relate to performance? *CBE—Life Sciences Education, 20*(1). https://doi.org/10.1187/cbe.20-05-0091

Wang, M., et al. (2025). Sleep duration and subject-specific academic performance among adolescents. *npj Science of Learning*. https://www.nature.com/articles/s41539-025-00361-y

Wu, M. (2026). K-Means clustering-based feature generation for student performance prediction. *Transactions on Educational Data Mining*. https://www.icck.org/article/abs/tedm.2026.716076

Zeek, M. L., et al. (2015). Sleep duration and academic performance among student pharmacists. *American Journal of Pharmaceutical Education, 79*(5), 63. https://doi.org/10.5688/ajpe79563

Zhang, Y., et al. (2021). Educational data mining techniques for student performance prediction. *Frontiers/PMC*. https://pmc.ncbi.nlm.nih.gov/articles/PMC8688359/