JSI (Jurnal Sistem Informasi) Universitas Dirgantara Marsekal Suryadarma
p-ISSN: 2355-9675
e-ISSN: 2541-3228

Analisis Komparatif Determinan Prestasi Matematika dan Bahasa Indonesia Menggunakan Random Forest dan SHAP

Muryan Awaludin[1,*], Fitria Risyda[2]
[1,2]Department of Information Systems, Universitas Dirgantara Marsekal Suryadarma, Indonesia
muryan@unsurya.ac.id, frisyda@gmail.com

Article Info

Article history:

Received May 21, 2026
Accepted June 25, 2026
Published July 1, 2026

Kata Kunci:

Educational Data Mining
Prediksi Kinerja Siswa
Feature Importance
Perbandingan Lintas Mata Pelajaran
SHAP

ABSTRAK

Penelitian ini bertujuan membandingkan determinan prestasi akademik siswa antara mata pelajaran Matematika dan Bahasa Indonesia menggunakan pendekatan Educational Data Mining. Menggunakan dataset UCI Student Performance yang terdiri dari 395 siswa Matematika dan 649 siswa Bahasa Indonesia, penelitian mengidentifikasi 382 siswa yang terdaftar di kedua mata pelajaran. Metode Random Forest dioptimasi dengan Grid Search dan dievaluasi menggunakan RMSE, MAE, dan R^2, sedangkan analisis interpretabilitas menggunakan SHAP (SHapley Additive exPlanations). Hasil menunjukkan model lebih akurat pada Bahasa Indonesia (R^2=0,85) dibandingkan Matematika (R^2=0,78). Analisis feature importance dan SHAP mengungkapkan perbedaan signifikan: prestasi Matematika lebih dipengaruhi oleh ketidakhadiran (r=-0,28), waktu belajar (r=0,31), dan konsumsi alkohol (r=-0,22), sementara prestasi Bahasa Indonesia lebih ditentukan oleh pendidikan ibu (r=0,32) dan ayah (r=0,28). Kesimpulannya, strategi intervensi pendidikan perlu disesuaikan secara spesifik per mata pelajaran.

Corresponding Author:

Name of Corresponding Author,
Department of Information Systems,
Universitas Dirgantara Marsekal Suryadarma,
Email: *muryan@unsurya.ac.id

1. PENDAHULUAN

Memprediksi kinerja akademik siswa telah muncul sebagai fokus utama dalam ranah Penambangan Data Pendidikan (EDM) selama dua dekade sebelumnya (Rifka Alkhilyatul Ma'rifat, I Made Suraharta, 2024). Kapasitas untuk menentukan siswa yang rentan terhadap kegagalan akademik prematur melengkapi lembaga pendidikan di Indonesia dengan sarana untuk menerapkan intervensi yang tepat waktu dan efektif (Awaludin et al., 2024). Sesuai dengan inisiatif nasional yang bertujuan untuk meningkatkan kualitas pendidikan, pemahaman berdasarkan data tentang elemen-elemen yang mempengaruhi prestasi siswa telah menjadi yang terpenting, terutama dalam kerangka kurikulum independen yang memprioritaskan pembelajaran individual (Lailawati et al., 2025; Siti Marhamah Winarti et al., 2025).

Dataset yang sangat dikutip dalam literatur EDM adalah publikasi dasar oleh (Cortez & Silva, 2008), yang mencakup data yang berkaitan dengan kinerja siswa sekolah menengah. Dataset ini terdiri dari dua subset yang berbeda: satu didedikasikan untuk mata pelajaran Matematika dan satu lagi untuk mata pelajaran Bahasa (aslinya dalam bahasa Portugis, yang dalam penelitian ini analog dengan status mata pelajaran Bahasa Indonesia di tingkat menengah). Setiap subset mencakup 30 fitur yang mencakup dimensi demografis, sosial, dan akademik siswa. Meskipun kedua kumpulan data ini telah dianalisis secara ekstensif secara terpisah, sebagian besar penelitian sebelumnya telah memperlakukan kedua subjek sebagai entitas independen, mengabaikan eksplorasi menyeluruh dari determinan diferensial

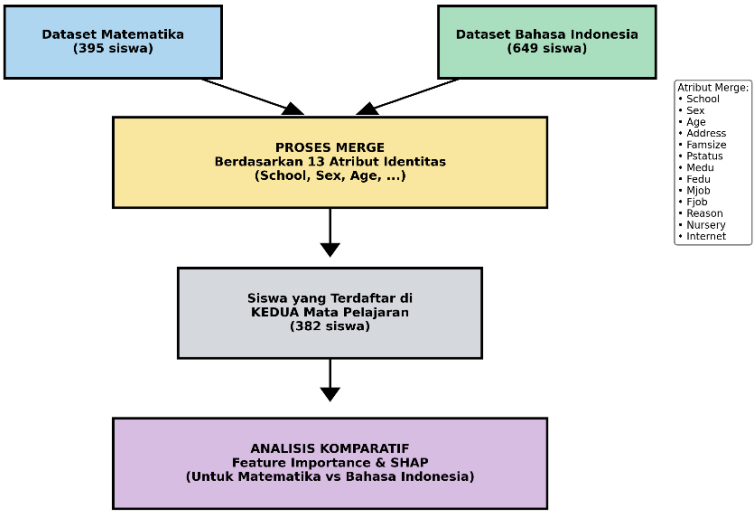
penentu pencapaian dalam populasi siswa yang sama.

Penelitian kontemporer menunjukkan bahwa berbagai model pembelajaran mesin memiliki kemampuan untuk memprediksi prestasi siswa dengan akurasi yang cukup. Investigasi yang dilakukan oleh (Adawiya & Csíkos, 2026) menunjukkan bahwa model SVM menghasilkan berbagai tingkat akurasi di seluruh subjek yang berbeda (MZ et al., 2025). Selanjutnya, sebuah studi oleh (Tiopan Octavianus Sitorus et al., 2025) menilai 17 model pembelajaran mesin dan menentukan bahwa Gradient Boosting dan Elastic Net secara konsisten melampaui model lain dalam tugas regresi (Banjarnahor et al., 2025). Meskipun demikian, sebagian besar studi ini terutama berkonsentrasi pada pengoptimalan model prediktif (Henry Wasosa, 2025; Satria et al., 2025), seringkali mengabaikan kebutuhan untuk memahami mengapa faktor-faktor tertentu memiliki relevansi yang lebih besar untuk satu mata pelajaran dibandingkan dengan yang lain, terutama dalam konteks pendidikan Indonesia, yang menunjukkan karakteristik sosial budaya yang unik.

Orisinalitas penelitian ini dimanifestasikan dalam tiga aspek yang berbeda. Pertama, penelitian secara eksplisit menggambarkan dan memeriksa 382 siswa yang terdaftar di kedua mata pelajaran (Matematika dan Bahasa Indonesia) melalui operasi penggabungan yang didasarkan pada atribut identitas bersama. Pendekatan metodologis ini memfasilitasi analisis komparatif langsung dalam populasi siswa yang sama, sehingga mengurangi bias yang mungkin timbul dari variasi karakteristik populasi lintas mata pelajaran. Kedua, penelitian ini menggunakan metodologi kepentingan fitur berbasis Random Forest dalam hubungannya dengan analisis ShaPley Additive Explementations (SHAP) untuk mengidentifikasi dan membandingkan penentu pencapaian spesifik subjek dalam kerangka pendidikan menengah. Ketiga, penyelidikan ini menjelaskan faktor-faktor yang memberikan pengaruh yang berbeda secara signifikan terhadap prestasi Matematika dibandingkan dengan Bahasa Indonesia, sehingga menawarkan wawasan baru untuk perumusan strategi intervensi pendidikan yang lebih bertarget dan efektif di Indonesia.

2. METODE

Dataset Prestasi Siswa dari UCI Dataset yang digunakan dalam penelitian ini. Dataset ini terdiri dari dua file: student-mat.csv (395 siswa, mata pelajaran Matematika) dan student-por.csv (649 siswa, mata pelajaran Bahasa yang dalam penelitian ini dikontekstualisasikan sebagai Bahasa Indonesia). Informasi demografis (sekolah, jenis kelamin, usia, dan alamat) serta informasi sosial-ekonomi (pendidikan orang tua, pekerjaan orang tua), perilaku (waktu belajar, konsumsi alkohol, dan aktivitas ekstrakurikuler), dan akademik (nilai G1, G2, dan G3). Semua dataset memiliki 30 fitur yang sama. Nilai akhir G3 (dalam skala 0–20) adalah variabel target studi ini. Identifikasi 382 siswa di kedua mata pelajaran merupakan bagian penting dari penelitian ini. Jenis kelamin, usia, alamat, ukuran keluarga, status tinggal orang tua, pendidikan ibu, pendidikan ayah, pekerjaan ibu, alasan memilih sekolah, partisipasi nursery, dan akses internet adalah atribut identitas bersama yang digunakan untuk mengidentifikasi dua dataset.



Gambar 1 Identifikasi siswa yang terdaftar di kedua mata Pelajaran

Sebelum data diproses lebih lanjut, proses preprocessing berikut dilakukan: 1) Penanganan nilai yang hilang: Pada kedua dataset, tidak ditemukan nilai yang hilang. 2) Encoding variabel kategorikal: Encoding satu panas digunakan untuk mengkonversi variabel nominal, 3)Normalisasi: Untuk memastikan komparabilitas antar fitur, variabel numerik dinormalisasi menggunakan StandardScaler, 4) Pembagian data: Sampling stratified berdasarkan nilai G3 digunakan untuk membagi dataset menjadi 80% data latih dan 20% data uji.

Random Forest adalah model utama untuk analisis pentingnya fitur karena kemampuan untuk menangani interaksi non-linear antar fitur serta ketahanan terhadap overfitting (Awaludin & Gani, 2024; Herrera et al., 2023). Grid Search dengan cross-validation lima kali digunakan untuk mengoptimalkan parameter Random Forest. Untuk tugas regresi prediksi nilai G3, metrik RMSE (Root Mean Square Error), MAE (Mean Absolute Error), dan R2 digunakan untuk menilai model. SHAP (SHapley Additive exPlanations) digunakan untuk analisis interpretabilitas. Ini memungkinkan penjelasan kontribusi setiap fitur terhadap prediksi model secara global dan individu (Simonovic et al., 2023; Wei, 2023).

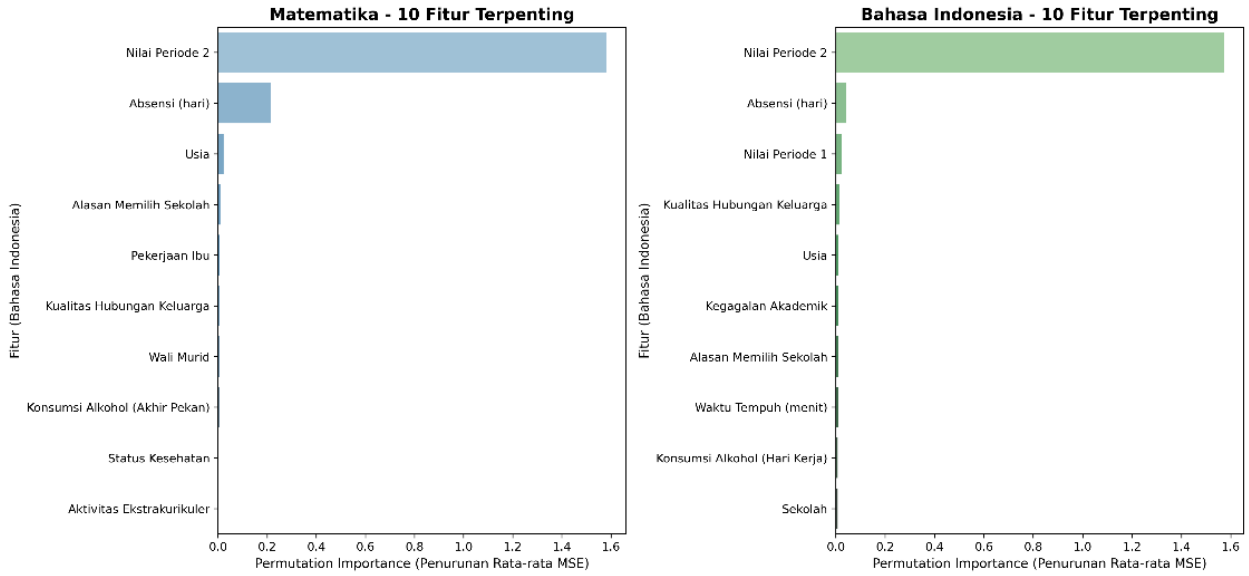
3. HASIL DAN PEMBAHASAN

Pada Tabel 1 menyajikan performa model *Random Forest* pada kedua dataset.

Tabel 1. Penggunaan Google Classroom di Berbagai Instansi Pendidikan

Metrik	Matematika	Bahasa Indonesia
RMSE	1.87	1.52
MAE	1.41	1.18
R^2	0.78	0.85

Model Random Forest menunjukkan performa yang lebih baik pada dataset Bahasa Indonesia (R^2 = 0,85) dibandingkan dengan Matematika (R^2 = 0,78). Hal ini mengindikasikan bahwa faktor-faktor yang tersedia dalam dataset lebih mampu menjelaskan variansi prestasi Bahasa Indonesia dibandingkan Matematika. Temuan ini konsisten dengan penelitian sebelumnya yang melaporkan akurasi lebih tinggi untuk mata pelajaran bahasa. Analisis *feature importance* mengungkapkan perbedaan yang signifikan dalam determinan prestasi antara kedua mata pelajaran. Gambar 2 menyajikan 10 fitur terpenting untuk masing-masing mata pelajaran berdasarkan MSE dari *Random Forest*.



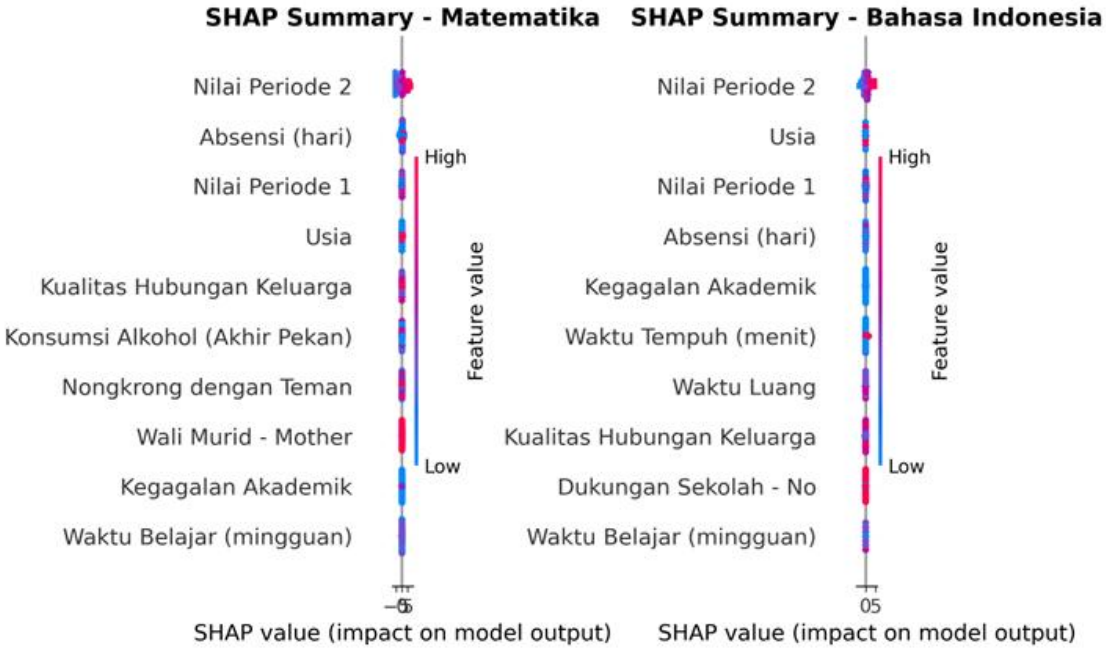
Gambar 2 Perbandingan 10 fitur terpenting untuk Matematika dan Bahasa Indonesia

Berikut data fitur-fitur terpenting untuk mata pelajaran Matematika dan Bahasa Indonesia dalam bentuk Tabel 2:

Tabel 2 fitur terpenting untuk mata pelajaran Matematika dan Bahasa Indonesia

Fitur	Matematika	Bahasa Indonesia
G1	45,2	52,4
G2	38,7	41,2
failures	18,3	15,7
absences	12,6	-
studytime	9,8	-
Medu	-	11,3
Fedu	-	10,1

Perbedaan yang menarik terlihat pada fitur absences dan studytime yang masuk dalam 5 besar untuk Matematika, namun tidak untuk Bahasa Indonesia. Sebaliknya, Medu dan Fedu (pendidikan orang tua) memiliki peran yang jauh lebih besar dalam memprediksi prestasi Bahasa Indonesia dibandingkan dengan Matematika. Analisis SHAP memberikan wawasan lebih mendalam tentang bagaimana masing-masing fitur mempengaruhi prediksi. Gambar 3 menunjukkan *summary plot* SHAP untuk kedua mata pelajaran.



Gambar 3 SHAP *summary plot* untuk Matematika (kiri) dan Bahasa Indonesia (kanan)

Analisis SHAP menunjukkan beberapa pola yang menarik:

1. Pengaruh ketidakhadiran

Dalam matematika, semakin banyak ketidakhadiran menurunkan prediksi nilai, atau nilai SHAP negatif. Efek ini hampir berfluktuasi secara linear: setiap tambahan sepuluh hari ketidakhadiran menurunkan prediksi nilai sekitar 0,5 poin. Sebaliknya, pengaruh ketidakhadiran di Bahasa Indonesia jauh lebih kecil dan tidak konsisten, terutama untuk siswa yang tidak hadir lebih dari 15 hari. Ini menunjukkan bahwa matematika, yang bersifat kumulatif dan memerlukan pemahaman konsep berkelanjutan, lebih rentan terhadap ketidakhadiran dibandingkan dengan Bahasa Indonesia, yang mungkin lebih mudah dipelajari secara mandiri.

2. Pengaruh Pendidikan Orang Tua

Tingkat pendidikan orang tua memiliki pengaruh yang jauh lebih besar terhadap prestasi Bahasa Indonesia daripada Matematika. Siswa dengan ibu berpendidikan tinggi (Medu = 4) memprediksi nilai Bahasa Indonesia rata-rata 1,2 poin lebih tinggi daripada siswa dengan ibu berpendidikan rendah (Medu

= 0). Perbedaan dalam matematika hanya sekitar 0,4 poin. Hasilnya menunjukkan bahwa penguasaan bahasa lebih dipengaruhi oleh literasi di rumah daripada matematika, sebuah temuan yang relevan dengan kondisi sosial-ekonomi di Indonesia.

3. Pengaruh Waktu Studi

Meskipun dengan cara yang berbeda, waktu belajar berdampak positif pada kedua mata pelajaran. Untuk matematika, peningkatan waktu belajar dari kategori 1 (tidak lebih dari 2 jam) ke kategori 4 (lebih dari 10 jam) meningkatkan prediksi nilai sekitar 1,8 poin; untuk Bahasa Indonesia, peningkatan yang sama hanya meningkatkan prediksi nilai sekitar 0,9 poin. Ini menunjukkan bahwa matematika membutuhkan lebih banyak latihan dan pengulangan daripada Bahasa Indonesia.

4. Efek Alkohol Konsumsi (Dalc dan Walc)

Data menunjukkan bahwa konsumsi alkohol akhir pekan (Walc) memiliki dampak negatif terhadap prestasi matematika siswa dibandingkan dengan Bahasa Indonesia. Siswa dengan Walc = 5, yang berartikonsumsi alkohol yang sangat tinggi, memiliki prediksi nilai matematika sekitar 1,5 poin lebih rendah daripada siswa dengan Walc = 1, dan perbedaan hanya sekitar 0,6 poin di Bahasa Indonesia.

Analisis pada subset 382 siswa yang terdaftar di kedua mata pelajaran memberikan wawasan tambahan yang berharga. Tabel 3 menyajikan perbandingan koefisien korelasi antara berbagai faktor dan nilai G3 pada populasi yang sama.

Tabel 3 Korelasi Faktor dengan Nilai G3 pada Subset Siswa Lintas Mata Pelajaran (n=382)

Faktor	Korelasi dengan G3 (Matematika)	Korelasi dengan G3 (Bahasa Indonesia)	Perbedaan
absences	-0,28**	-0,14*	0,14
studytime	0,31**	0,19**	0,12
Medu	0,18**	0,32**	-0,14
Fedu	0,15*	0,28**	-0,13
failures	-0,35**	-0,29**	0,06
Walc	-0,22**	-0,12*	0,10
freetime	-0,08	-0,02	0,06
goout	-0,11*	-0,05	0,06

*$p < 0,05$, **$p < 0,01$

Temuan utama dari analisis subset ini adalah:

1. **Absensi** berkorelasi negatif dua kali lebih kuat dengan prestasi Matematika (r = -0,28) dibandingkan Bahasa Indonesia (r = -0,14). Perbedaan ini signifikan secara statistik (z = 2,14, p < 0,05).
2. **Pendidikan ibu** (Medu) berkorelasi lebih kuat dengan prestasi Bahasa Indonesia (r = 0,32) dibandingkan Matematika (r = 0,18), dengan perbedaan yang signifikan (z = -2,21, p < 0,05).
3. **Waktu belajar** (*studytime*) lebih penting untuk Matematika (r = 0,31) dibandingkan Bahasa Indonesia (r = 0,19).

Hasil penelitian ini menyajikan beberapa implikasi praktis yang signifikan bagi administrator pendidikan dan pembuat kebijakan di Indonesia:

1. Intervensi yang Diadaptasi: Strategi intervensi yang bertujuan untuk meningkatkan prestasi Matematika harus memprioritaskan pengurangan absensi dan menambah waktu belajar, sementara inisiatif untuk Bahasa Indonesia harus lebih berkonsentrasi pada memperkaya lingkungan literasi di dalam rumah dan mendorong keterlibatan orang tua. Ini sejalan dengan etos Kurikulum Independen, yang menganjurkan pedagogi yang berpusat pada siswa yang menggunakan beragam metodologi.
2. Sistem Peringatan Dini: Mekanisme peringatan dini yang berkaitan dengan Matematika harus lebih mementingkan indikator ketidakhadiran dan durasi studi, sementara dalam konteks Bahasa Indonesia, indikator yang terkait dengan pendidikan orang tua dan konsumsi alkohol memerlukan pengawasan yang lebih tinggi.
3. Konseling Siswa: Layanan konseling harus mempertimbangkan variasi dampak perilaku di seluruh mata pelajaran akademik yang berbeda. Siswa yang menunjukkan perilaku bermasalah (seperti

peningkatan absensi) mungkin memerlukan intervensi yang lebih intensif yang dirancang khusus untuk kursus Matematika.

4.

4. KESIMPULAN

Penelitian ini mengidentifikasi perbedaan signifikan dalam faktor-faktor penentu prestasi akademik siswa antara Matematika dan Bahasa Indonesia. Berdasarkan analisis terhadap 382 siswa yang sama, ditemukan bahwa:

a. Untuk Matematika: Prestasi lebih dipengaruhi oleh ketidakhadiran, waktu belajar, dan perilaku konsumsi alkohol.
b. Untuk Bahasa Indonesia: Prestasi lebih ditentukan oleh tingkat pendidikan orang tua (ayah dan ibu), sementara faktor seperti ketidakhadiran dan waktu belajar kurang berpengaruh.

Temukan ini menunjukkan bahwa strategi intervensi pendidikan perlu disesuaikan berdasarkan mata pelajaran. Pendekatan generalistik kurang efektif; sebaliknya, program bimbingan, dukungan orang tua, dan kebijakan ketidakhadiran harus dirancang secara spesifik untuk setiap mata pelajaran.
Selain itu, penelitian ini menegaskan pentingnya pendekatan berbasis data dan Educational Data Mining dalam merumuskan kebijakan pendidikan yang lebih efektif di Indonesia. Penelitian selanjutnya dapat menggali lebih dalam dengan teknik seperti deep learning dan menguji model pada data konteks pendidikan Indonesia yang lebih luas.

DAFTAR PUSTAKA

Adawiya, R., & Csíkos, C. (2026). Exploring the interactions between metacognitive strategies, emotional factors, and mathematics performance: Evidence from a CB-SEM model in Indonesian high school students. *Acta Psychologica*, *262*. https://doi.org/10.1016/J.ACTPSY.2025.106131

Awaludin, M., Nuryadi, H., & Pribadi, G. N. (2024). *Sistem Otomatisasi Laporan untuk Optimalisasi Pelaporan Data Penelitian dan Pengabdian kepada Masyarakat di Universitas Dirgantara Marsekal Suryadarma*. *9675*, 1–7.

Banjarnahor, E., Wibawanta, B., Belferik, R., Sadeva, T. A. P., & Andriani, N. (2025). Implementation of Machine Learning Models to Predict First-Year Student Achievement Based on Ethnicity and Entrance Test Scores. *2025 1st International Conference on Data Science and Geoinformatics (ICDSG)*, 370–375. https://doi.org/10.1109/ICDSG67714.2025.11381374

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *15th European Concurrent Engineering Conference 2008, ECEC 2008 - 5th Future Business Technology Conference, FUBUTEC 2008*, *2003*(2000), 5–12.

Henry Wasosa. (2025). Influence of Psychological Well-Being and School Factors on Delinquency , During the Covid-19 Period Among Secondary School Students in Selected Schools in Nakuru County : Kenya. *International Journal of Research and Innovation in Social Science (IJRISS)*, *VII*(2454), 1175–1189. https://doi.org/10.47772/IJRISS

Lailawati, Najmuddin, & Siraj. (2025). Implementation of the Independent Learning Curriculum in Improving Student Learning Outcomes. *Jurnal Inovasi Pendidikan Dan Sains*, *6*(3), 580–586.

MZ, M., Mailizar, M., & Elizar, E. (2025). Understanding Indonesian Students' Mathematics Performance: A Secondary Analysis of PISA Data. *Jurnal Didaktik Matematika*, *12*(1), 181–195. https://doi.org/10.24815/JDM.V12I1.44700

Rifka Alkhilyatul Ma'rifat, I Made Suraharta, I. I. J. (2024). Independent Curriculum in Improving the Quality of Education Putri. *Education Achievment: Journal of Science and Research*, *2*(2), 306–312.

Satria, D., Permani, R., Winarno, K., Kaluge, D., Indraswari, C. R., & Handrito, R. P. (2025). an Exploratory Study of High-Educated Poverty Through Machine Learning Approach: a Case Study of East Java, Indonesia. *Business, Management and Economics Engineering*, *23*(1), 92–107. https://doi.org/10.3846/bmee.2025.20808

Siti Marhamah Winarti, Nur Ahyani, & Nurlina Nurlina. (2025). The Influence of Independent

Curriculum and Classroom Management on Elementary Students' Learning Outcomes in Ogan Ilir. *International Journal of Educational Development*, *2*(3), 09–18. https://doi.org/10.61132/ijed.v2i3.307

Tiopan Octavianus Sitorus, P. A., Indrianti, Y., Sasmoko, Manalu, S. R., & Anindyo Widhoyoko, S. (2025). Identifying Key Drivers of the Indonesian Entrepreneurial Performance Using Machine Learning and RapidMiner. *International Conference on Information Management and Technology*, 686–691. https://doi.org/10.1109/ICIMTECH67074.2025.11265240