JEDUMATIC

e-ISSN 3031-8408

Journal of Education and Informatics Research /Vol.7, No.2 2026

Klasterisasi Wilayah Rawan Mengulang dan Putus Sekolah pada Jenjang SD hingga SMA/SMK di Indonesia Tahun 2023-2025 Menggunakan K-Means

Diva Keishya Kirana[1*]

[1] *Program Studi Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia*
email: [1]21083010069@student.upnjatim.ac.id

ABSTRAK

Ketimpangan pendidikan masih menjadi tantangan sistemik di Indonesia meskipun anggaran pendidikan telah dialokasikan sebesar 20% dari APBN, tercermin dari tingginya variasi rasio mengulang kelas dan putus sekolah antarwilayah. Penelitian ini memetakan zona kerawanan pendidikan pada tingkat kabupaten/kota menggunakan K-Means *Clustering* terhadap delapan indikator rasio mengulang dan putus sekolah pada empat jenjang pendidikan (SD, SMP, SMA, SMK), bersumber dari basis data Dapodik Kementerian Pendidikan periode 2023–2025. Kasus ekstrem dipisahkan terlebih dahulu menggunakan ambang *interquartile range* sebelum proses *clustering*, sehingga tipologi yang dihasilkan mencerminkan pola pada mayoritas kabupaten/kota. Tiga klaster (K=3) dibentuk dan divalidasi menggunakan *Silhouette Score*, uji *Kruskal-Wallis*, skema validasi *held-out feature*, serta perbandingan dengan *Hierarchical Clustering* (*Ward's Linkage*) melalui *Adjusted Rand Index*. Hasil *clustering* menghasilkan empat kategori praktis: Kasus Ekstrem, Rawan Tinggi, Rawan Sedang, dan Rawan Rendah, dengan konsentrasi kabupaten/kota rawan tertinggi berada di Papua dan Indonesia bagian timur. Temuan ini menawarkan tipologi berbasis data yang *reproducible* bagi pemangku kebijakan pendidikan untuk merancang kebijakan intervensi yang terdiferensiasi dan berbasis (*place-based policy*).

Kata Kunci: *Data Mining*, Klasterisasi K-Means, Pengulangan Kelas, Putus Sekolah.

ABSTRACT

Educational inequality remains a persistent challenge in Indonesia despite the allocation of 20% of the national budget to education, particularly reflected in grade-repetition and drop-out rates that vary widely across regions. This study maps educational vulnerability zones at the district/city (kabupaten/kota) level using K-Means Clustering applied to eight ratio-based indicators of repetition and drop-out across four education levels (SD, SMP, SMA, SMK), sourced from the Ministry of Education's Dapodik database for 2023–2025. Extreme cases were separated using an interquartile-range threshold prior to clustering so that the resulting typology reflects the majority of districts. Three clusters (K=3) were formed and evaluated using the Silhouette Score, Kruskal-Wallis tests, a held-out feature validation scheme, and a comparison with Ward's Hierarchical Clustering through the Adjusted Rand Index. The clustering produced four practical categories: Extreme Cases, High Risk, Moderate Risk, and Low Risk, with the highest concentrations of vulnerable districts found in Papua and eastern Indonesia. These findings offer a reproducible, data-driven typology that education authorities can use to design differentiated, place-based intervention policies.

Keywords: *Data Mining, Grade Repetition, K-Means Clustering, Student Dropout.*

PENDAHULUAN

Pemerataan kualitas pendidikan dasar hingga menengah merupakan salah satu pilar utama pembangunan sumber daya manusia Indonesia, sejalan dengan target *Sustainable Development Goals* (SDGs) ke-4 dan Visi Indonesia Emas 2045. Pemerintah telah mengalokasikan anggaran pendidikan sebesar 20% dari Anggaran Pendapatan dan Belanja Negara (APBN), namun ketimpangan capaian pendidikan antarwilayah masih menjadi tantangan sistemik. Dua indikator yang secara langsung mencerminkan kerawanan pendidikan di tingkat satuan wilayah adalah rasio peserta didik yang mengulang kelas (*retention rate*) dan rasio peserta didik yang putus sekolah (*drop-out rate*), karena keduanya berkorelasi dengan kerugian ekonomi keluarga dan ketimpangan kualitas sumber daya manusia jangka panjang antarwilayah.

Sejumlah penelitian terdahulu telah menerapkan K-Means *Clustering* untuk memetakan kondisi pendidikan antarwilayah di Indonesia, misalnya pengelompokan kabupaten/kota berdasarkan indikator fasilitas pendidikan (Handitia & Sofro, 2024) dan indikator pendidikan tingkat provinsi (Hanifah & Primandari, 2023). Di sisi lain, penelitian mengenai putus sekolah dan mengulang kelas sebagian besar berfokus pada prediksi tingkat individu peserta didik dengan model *supervised learning* (Kabathova & Drlik, 2021; Christou et al., 2023; Krüger et al., 2023; Fitriana et al., 2024), sehingga belum banyak yang mengombinasikan pendekatan *unsupervised clustering* pada level agregat kabupaten/kota dengan skema validasi statistik yang ketat, termasuk penanganan eksplisit terhadap kasus ekstrem dan pengujian independensi struktur klaster melalui skema *held-out feature*.

Penelitian ini bertujuan memetakan zona kerawanan pendidikan pada 512 kabupaten/kota di Indonesia menggunakan K-Means *Clustering* terhadap delapan indikator rasio mengulang dan putus sekolah pada empat jenjang pendidikan (SD, SMP, SMA, SMK), bersumber dari basis data Data Pokok Pendidikan (Dapodik) Kementerian Pendidikan periode 2023–2025. Kontribusi utama penelitian ini adalah: (1) menghasilkan tipologi zona kerawanan pendidikan berbasis data yang dapat direproduksi untuk mendukung kebijakan intervensi yang terdiferensiasi; dan (2) mendemonstrasikan penerapan *pipeline data science end-to-end*, mencakup *feature engineering*, penanganan *outlier*, *unsupervised learning*, dan validasi statistik *non-sirkular*, sebagai bahan kajian *computational thinking* dan *educational data mining* yang relevan dengan bidang pendidikan informatika.

METODE

Penelitian ini menggunakan desain kuantitatif non-eksperimental dengan pendekatan *data mining* pada data sekunder agregat. Data bersumber dari 20 berkas mentah Data Pokok Pendidikan (Dapodik), mencakup tiga indikator (jumlah siswa total, siswa mengulang, siswa putus sekolah) pada empat jenjang pendidikan (SD/MI, SMP/MTs, SMA/MA, SMK/MAK), ditambah berkas rincian jenis kelamin untuk keperluan audit silang struktur data.

Tiga kriteria seleksi data diterapkan secara konsisten pada seluruh berkas. Pertama, periode 2023–2025 dipertahankan sebagai populasi analisis; periode 2021–2022 sengaja dikeluarkan karena berbagai faktor penyesuaian di masa pandemi menyebabkan indikator mengulang kelas tidak terdistribusi secara normal (bernilai nol secara sistemik). Kedua, baris dengan wilayah Sekolah Indonesia Luar Negeri (SILN) dieksklusi agar analisis berfokus pada kondisi domestik. Ketiga, hanya baris dengan status sekolah "Negeri dan Swasta" yang dipertahankan agar rasio merepresentasikan kondisi wilayah secara utuh. Setelah proses irisan (*inner join*) keempat jenjang, diperoleh 512 kabupaten/kota sebagai unit observasi akhir; penyusutan dari populasi awal murni disebabkan oleh dua kabupaten pemekaran baru yang belum memiliki satuan pendidikan SMK, bukan oleh data yang hilang.

Delapan fitur rasio dihitung dengan menjumlahkan nilai *numerator* dan *denominator* periode 2023–2025 per kabupaten/kota terlebih dahulu, baru kemudian dibagi, sehingga bobot setiap tahun proporsional terhadap jumlah siswa riil. Sebelum *clustering*, kabupaten/kota dengan nilai salah satu dari delapan fitur melampaui ambang Q3+3×IQR ditandai sebagai Kasus Ekstrem (75 kabupaten/kota atau 14,6%) dan dikeluarkan dari proses pembentukan klaster utama agar segmentasi mencerminkan pola pada populasi yang lebih homogen, namun tetap dianalisis dan dilaporkan terpisah sebagai daftar prioritas darurat. Standardisasi *Z-score* (*StandardScaler*) di-fit hanya pada *subset non-ekstrem* (437 kabupaten/kota), kemudian ditransformasikan ke seluruh data untuk keperluan skor risiko individu.

K-Means *Clustering* dievaluasi pada rentang K=2 hingga K=8 menggunakan Metode *Elbow* (WCSS) dan *Silhouette Score*; K=3 dipilih karena memberikan keseimbangan antara kesederhanaan interpretasi tiga tingkat kerawanan dan performa *Silhouette* yang relatif stabil pada rentang tersebut. Validitas klaster diuji melalui tiga skema: (1) uji *Kruskal-Wallis* deskriptif pada kedelapan fitur pembentuk klaster; (2) skema *held-out feature*, yaitu membentuk klaster hanya dari empat fitur rasio mengulang, kemudian menguji independensi terhadap empat fitur rasio putus sekolah yang sama sekali tidak dipakai membentuk klaster; dan (3) perbandingan dengan *Hierarchical Agglomerative Clustering* (*Ward's Linkage*) pada *subset* dan K yang sama, diukur dengan *Adjusted Rand Index* (ARI). Skor risiko individu dihitung sebagai rata-rata *Z-score* kedelapan fitur, dipakai konsisten untuk perankingan klaster maupun pengurutan daftar prioritas intervensi. Seluruh *pipeline* diimplementasikan menggunakan Python dengan pustaka *pandas*, *scikit-learn*, dan *scipy*.

HASIL DAN PEMBAHASAN

Matriks data akhir terdiri atas 512 kabupaten/kota dengan delapan indikator rasio. Statistik deskriptif (Tabel 1) menunjukkan variasi yang sangat tinggi antarwilayah, dengan koefisien variasi (CV) berkisar 124,5%–221,0%; rasio mengulang SMA memiliki CV tertinggi (221,0%), mengindikasikan sebaran yang sangat timpang antarwilayah pada jenjang tersebut.

Tabel 1. Statistik deskriptif delapan indikator rasio (N=512)

Indikator Rasio	Rata-rata (%)	Std. Dev (%)	CV (%)
Mengulang SD	0,620	0,934	150,6
Putus Sekolah SD	0,236	0,294	124,5
Mengulang SMP	0,318	0,531	166,8
Putus Sekolah SMP	0,196	0,297	151,4
Mengulang SMA	0,319	0,705	221,0
Putus Sekolah SMA	0,185	0,231	125,2
Mengulang SMK	0,597	1,187	198,7
Putus Sekolah SMK	0,380	0,675	177,4

Evaluasi K=2 hingga K=8 menunjukkan *Silhouette Score* tertinggi pada K=2 (0,320), namun K=3 dipilih (*Silhouette*=0,300) karena memberikan resolusi kebijakan tiga tingkat (Rawan Tinggi/Sedang/Rendah) yang lebih operasional bagi pengambil kebijakan dibanding hanya dua kelompok biner. K-Means dengan K=3 pada *subset non-ekstrem* (N=437) menghasilkan tiga klaster berukuran 231, 91, dan 115 kabupaten/kota.

Profil setiap zona (Tabel 2) menunjukkan pola yang konsisten: zona Rawan Tinggi memiliki skor risiko gabungan (rata-rata *Z-score* kedelapan fitur) tertinggi, diikuti Rawan Sedang, dan Rawan Rendah terendah. Menariknya, urutan pada satu fitur tunggal tidak selalu sejalan dengan urutan skor gabungan, misalnya rasio mengulang SD di Rawan Sedang (0,77%) justru lebih tinggi dibanding Rawan Tinggi (0,55%), karena label zona ditentukan dari kombinasi seluruh delapan fitur, bukan satu indikator saja; zona Rawan Tinggi justru dicirikan oleh rasio putus sekolah SMK yang jauh lebih tinggi (0,70% vs 0,28%).

Tabel 2. Profil dan karakteristik zona kerawanan pendidikan
(*subset non-ekstrem*, N=437, ditambah Kasus Ekstrem N=75)

Zona Kerawanan	N	Skor Risiko (rata-rata Z)	Karakteristik Utama
Rawan Tinggi	91	0,734	Rasio putus sekolah SMK tertinggi (0,70%) dan mengulang SMA relatif tinggi
Rawan Sedang	115	0,399	Rasio mengulang SD dan SMK tertinggi, namun rasio putus sekolah moderat
Rawan Rendah	231	-0,488	Seluruh delapan indikator rasio berada di bawah rata-rata nasional
Kasus Ekstrem (Prioritas Darurat)	75	> +2 SD*	Minimal 1 dari 8 indikator melampaui ambang Q3+3×IQR; ditangani kasus-per-kasus

Uji *Kruskal-Wallis* deskriptif menunjukkan kedelapan fitur pembentuk klaster berbeda signifikan antarklaster ($p<0,001$ pada seluruh fitur), namun hasil ini bersifat sirkular karena K-Means memang dioptimasi untuk memisahkan fitur-fitur tersebut. Bukti independen yang lebih kuat diperoleh dari skema *held-out feature*: klaster yang dibentuk murni dari empat fitur rasio mengulang tetap membedakan seluruh empat fitur rasio putus sekolah secara signifikan ($p<0,0001$), menunjukkan bahwa fenomena mengulang kelas dan putus sekolah berbagi struktur kewilayahan yang sama, bukan sekadar artefak konstruksi model. Perbandingan dengan *Hierarchical Clustering* (*Ward's Linkage*) pada *subset* dan K yang sama menghasilkan *Adjusted Rand Index* sebesar 0,469, mengindikasikan kesesuaian moderat-kuat antara dua metode berbeda dan memperkuat keandalan struktur klaster.

Secara spasial, agregasi tingkat provinsi menunjukkan konsentrasi kabupaten/kota berstatus Kasus Ekstrem dan Rawan Tinggi tertinggi berada di kawasan Indonesia timur: Papua Selatan, Papua Pegunungan, Papua, Papua Barat, dan Papua Tengah mencapai 100%, diikuti Maluku (90,9%), Maluku Utara (90,0%), Gorontalo (83,3%), Sulawesi Barat (66,7%), dan Nusa Tenggara Timur (63,6%). Pola ini konsisten dengan temuan penelitian klaster pendidikan berbasis wilayah sebelumnya yang juga menemukan disparitas tajam antara kawasan barat dan timur Indonesia (Handitia & Sofro, 2024; Hanifah & Primandari, 2023).

Temuan ini memiliki implikasi kebijakan dua jalur: kabupaten/kota berstatus Kasus Ekstrem, seperti Kabupaten Mappi dan Yalimo yang skor risikonya jauh melampaui populasi umum, memerlukan investigasi dan intervensi kasus-per-kasus, sedangkan kabupaten/kota pada klaster Rawan Tinggi/Sedang lebih tepat didekati dengan kebijakan berbasis tipologi seperti paket Bantuan Operasional Sekolah (BOS) afirmasi atau Program Indonesia Pintar (PIP) yang diseragamkan dalam satu zona. *Silhouette Score* akhir yang tergolong lemah-sedang (≈0,30) perlu dibaca sebagai batas zona yang bersifat gradual, bukan pemisahan diskrit yang tegas antarkelompok kabupaten/kota. Keterbatasan penelitian mencakup sifat data yang *cross-sectional* teragregasi, rasio yang dihitung dari status Negeri dan Swasta tergabung, ambang deteksi kasus ekstrem yang bersifat konvensional, serta representasi spasial pada tingkat provinsi akibat keterbatasan berkas geospasial resmi tingkat kabupaten/kota.

KESIMPULAN

Penelitian ini menghasilkan tipologi zona kerawanan pendidikan pada 512 kabupaten/kota di Indonesia melalui K-Means *Clustering* terhadap delapan indikator rasio mengulang dan putus sekolah periode 2023–2025, menghasilkan empat kategori praktis yaitu Kasus Ekstrem, Rawan Tinggi, Rawan Sedang, dan Rawan Rendah, dengan konsentrasi wilayah rawan tertinggi berada di Papua dan kawasan Indonesia timur lainnya. Meskipun *Silhouette Score* akhir tergolong lemah-sedang, validitas struktur klaster diperkuat oleh bukti independen dari skema *held-out feature* dan kesesuaian moderat-kuat dengan *Hierarchical Clustering* melalui *Adjusted Rand Index*, sehingga tipologi ini layak dijadikan basis kebijakan intervensi pendidikan yang terdiferensiasi dan berbasis kewilayahan bagi Kementerian

Pendidikan dan pemerintah daerah. Penelitian lanjutan disarankan mengeksplorasi ambang deteksi kasus ekstrem alternatif, memasukkan data geospasial resmi tingkat kabupaten/kota untuk pemetaan tematik yang lebih presisi, serta memperbarui model ini secara berkala setiap periode Dapodik baru dirilis agar perpindahan status wilayah antarzona dapat dipantau sebagai indikator keberhasilan program.

.

REFERENSI

Christou, V., Tsoulos, I., Loupas, V., Tzallas, A. T., Gogos, C., Karvelis, P. S., Antoniadis, N., Glavas, E., & Giannakeas, N. (2023). Performance and early drop prediction for higher education students using machine learning. Expert Systems with Applications, 225, 120079. https://doi.org/10.1016/j.eswa.2023.120079

Feng, G., & Fan, M. (2024). Research on learning behavior patterns from the perspective of educational data mining: Evaluation, prediction and visualization. Expert Systems with Applications, 237, 121555. https://doi.org/10.1016/j.eswa.2023.121555

Fitriana, S., Rinianty, & Pratama, S. A. (2024). Prediksi siswa putus sekolah dan keberhasilan akademik menggunakan machine learning. The Indonesian Journal of Computer Science, 13(6).

Handitia, N. R., & Sofro, A. (2024). Analisis klaster kecamatan di Kota Surabaya berdasarkan data pendidikan tahun 2022–2023. Jurnal Gaussian, 13(2), 351–362. https://doi.org/10.14710/j.gauss.13.2.351-362

Hanifah, S., & Primandari, A. H. (2023). Implementasi metode K-Means clustering dalam pengelompokan kabupaten/kota di Provinsi NTB berdasarkan indikator pendidikan. Emerging Statistics and Data Science Journal, 1(3), 1–10.

Haser, Ç., Doğan, O., & Kurt Erhan, G. (2022). Tracing students' mathematics learning loss during school closures in teachers' self-reported practices. International Journal of Educational Development, 88, 102536. https://doi.org/10.1016/j.ijedudev.2021.102536

Herbaut, E. (2021). Overcoming failure in higher education: Social inequalities and compensatory advantage in dropout patterns. Acta Sociologica, 64(4), 383–402. https://doi.org/10.1177/0001699320920916

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2022). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences, 622, 111–139.

Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. Applied Sciences, 11(7), 3130. https://doi.org/10.3390/app11073130

Krüger, J., Britto, A., & Barddal, J. (2023). An explainable machine learning approach for student dropout prediction. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2023.120933

Li, Y., & Zhang, H. (2024). Big data technology for teaching quality monitoring and improvement in higher education joint K-means clustering algorithm and Apriori algorithm. Systems and Soft Computing, 6, 200125.

Nur, A., Majid, A., & Sahran, S. (2023). Identification of student behavioral patterns in higher education using K-Means clustering and support vector machine. Applied Sciences, 13(5), 3267.

Patrinos, H., & Donnelly, R. (2021). Learning loss during COVID-19: An early systematic review. Covid Economics Vetted and Real-Time Papers, 1–11.

Poellhuber, L.-V., Poellhuber, B., Desmarais, M., Léger, C., Roy, N., & Vu, M. M.-C. (2023). Cluster-based performance of student dropout prediction as a solution for large scale models in a Moodle LMS. In LAK23: 13th International Learning Analytics and Knowledge Conference (pp. 592–598). ACM. https://doi.org/10.1145/3576050.3576146

Riza, L. S., Rosdiyana, R. A., Wahyudin, A., & Pérez, A. R. (2021). The K-Means algorithm for generating sets of items in educational assessment. Indonesian Journal of Science and Technology, 6(1), 93–100.

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.

Suyal, M., & Sharma, S. (2024). A review on analysis of K-Means clustering machine learning algorithm based on unsupervised learning. Journal of Artificial Intelligence and Systems, 6(1), 85–95.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236–244.

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9(1). https://doi.org/10.1186/s40561-022-00192-z

Zhong, Z., Guo, H., & Qian, K. (2024). Deciphering the impact of machine learning on education: Insights from a bibliometric analysis using bibliometrix R-package. Education and Information Technologies.