

K-Means Clustering for Grouping Elementary School Facilities and Infrastructure in Indonesia

OK Muhammad Zikri Fadillah^{1*}, Abdul Halim Hasugian²

^{1,2}Ilmu Komputer, Sains Dan Teknologi, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia

Email: ^{1*}okzikri@gmail.com, ²abdulhalimhasugian@uinsu.ac.id

(*: corresponding author)

Abstract - Equitable access to quality basic education remains a national priority in Indonesia, yet the condition of elementary school facilities and infrastructure still varies considerably across regions. This research applies the K-Means Clustering algorithm to classify the condition of elementary school facilities and infrastructure at the regency/city level in Indonesia, addressing disparities in basic education facilities through a more objective, data-based mapping approach. Data were obtained from the 2025 Data Portal of the Ministry of Primary and Secondary Education, covering 514 regencies/cities and six indicators: classrooms, libraries, laboratories, school health units, boys' toilets, and girls' toilets. Each indicator was converted into a numerical score based on four condition categories and normalized using Min-Max Normalization prior to clustering. The number of clusters was determined using the Elbow Method and validated using the Davies-Bouldin Index, both supporting $K=3$. The clustering grouped the 514 regencies/cities into 117 regions with relatively low facility conditions, 302 with relatively moderate conditions, and 95 with relatively high conditions, with the low cluster concentrated in eastern Indonesia, particularly Papua and Maluku. The main contribution of this research is a data-driven cluster map that education authorities can use as an evidence-based reference for prioritizing budget allocation and infrastructure development toward regencies/cities in the low cluster, as well as a replicable scoring-normalization-clustering pipeline adaptable for future studies with additional indicators.

Keywords: Davies-Bouldin Index, Educational Infrastructure, Elementary School, Infrastructure Mapping, K-Means Clustering.

Abstrak - Pemerataan akses pendidikan dasar yang bermutu masih menjadi prioritas nasional di Indonesia, namun kondisi sarana dan prasarana sekolah dasar di berbagai wilayah masih menunjukkan perbedaan yang cukup signifikan. Penelitian ini menerapkan algoritma *K-Means Clustering* untuk mengelompokkan kondisi sarana dan prasarana sekolah dasar pada tingkat kabupaten/kota di Indonesia, guna menjawab perbedaan kondisi fasilitas pendidikan dasar antarwilayah melalui pendekatan pemetaan berbasis data yang lebih objektif. Data penelitian diperoleh dari *Data Portal* Kementerian Pendidikan Dasar dan Menengah tahun 2025, mencakup 514 kabupaten/kota dan enam indikator, yaitu ruang kelas, perpustakaan, laboratorium, usaha kesehatan sekolah, toilet peserta didik laki-laki, dan toilet peserta didik perempuan. Setiap indikator diubah menjadi skor numerik berdasarkan empat kategori kondisi, kemudian dinormalisasi menggunakan *Min-Max Normalization* sebelum diproses dengan *K-Means Clustering*. Jumlah *cluster* ditentukan menggunakan *Elbow Method* dan divalidasi menggunakan *Davies-Bouldin Index*, yang keduanya mendukung $K=3$. Hasil *clustering* membagi 514 kabupaten/kota menjadi 117 wilayah dengan kondisi relatif rendah, 302 wilayah dengan kondisi relatif sedang, dan 95 wilayah dengan kondisi relatif tinggi, dengan *cluster* rendah terkonsentrasi di kawasan timur Indonesia, khususnya Papua dan Maluku. Kontribusi utama penelitian ini adalah peta *cluster* berbasis data yang dapat digunakan oleh dinas pendidikan sebagai acuan objektif dalam memprioritaskan alokasi anggaran dan pembangunan sarana pada kabupaten/kota di *cluster* rendah, serta kerangka kerja skor-normalisasi-*clustering* yang dapat direplikasi untuk penelitian selanjutnya dengan indikator tambahan.

Kata Kunci: Indeks Davies-Bouldin, Infrastruktur Pendidikan, Klasterisasi K-Means, Sekolah Dasar, Pemetaan Infrastruktur.

1. PENDAHULUAN

Pendidikan memiliki peran penting dalam peningkatan kualitas sumber daya manusia. Dalam kerangka *Sustainable Development Goals* (SDGs), pendidikan diarahkan agar dapat berlangsung secara inklusif, merata, dan memberikan kesempatan belajar bagi seluruh masyarakat [1]. Pada tingkat sekolah dasar, pemerataan pendidikan tidak hanya ditentukan oleh kegiatan pembelajaran, tetapi juga oleh dukungan sarana dan prasarana. Ruang kelas, perpustakaan, laboratorium, usaha kesehatan sekolah, serta toilet peserta didik merupakan fasilitas yang berpengaruh terhadap kenyamanan dan kelancaran proses belajar [2]. Kondisi sarana dan prasarana sekolah dasar di Indonesia masih menunjukkan perbedaan antarwilayah. Data Dapodik Kemendikdasmen dalam publikasi Statistik Sekolah Dasar Tahun 2024/2025 memuat informasi mengenai kondisi ruang kelas dan sanitasi sekolah yang dapat digunakan untuk melihat keadaan fasilitas pendidikan dasar [3].

Perbedaan kondisi tersebut menunjukkan bahwa fasilitas sekolah dasar belum sepenuhnya merata. Wilayah perkotaan umumnya memiliki dukungan fasilitas yang lebih baik dibandingkan wilayah terpencil atau daerah dengan keterbatasan pembangunan. Ketimpangan ini dapat berdampak pada perbedaan layanan pendidikan yang diterima peserta didik [4]. Permasalahan tersebut perlu dikaji melalui pendekatan yang mampu membaca pola dari banyak indikator. Data sarana dan prasarana sekolah tidak hanya memuat satu jenis fasilitas, tetapi terdiri dari beberapa indikator dan kategori kondisi. Apabila data hanya disajikan dalam bentuk tabel deskriptif, pola

kemiripan dan perbedaan antarwilayah akan sulit terlihat. Oleh karena itu, dibutuhkan metode berbasis data yang dapat mengelompokkan wilayah berdasarkan karakteristik fasilitas yang serupa.

Salah satu metode yang dapat digunakan adalah *clustering*, yaitu teknik *data mining* untuk membentuk kelompok objek berdasarkan kemiripan karakteristik tanpa label kelas awal [5]. *K-Means Clustering* dipilih karena mampu mengelompokkan data numerik berdasarkan kedekatan objek terhadap pusat *cluster* atau *centroid*. Dalam penelitian ini, *K-Means* sesuai digunakan karena setiap kabupaten/kota direpresentasikan dalam bentuk nilai numerik dari beberapa indikator sarana dan prasarana. Selain itu, hasil *K-Means* relatif mudah dibaca melalui nilai *centroid*, sehingga karakteristik setiap kelompok wilayah dapat dijelaskan dengan lebih sederhana [6].

Beberapa penelitian sebelumnya menunjukkan bahwa metode *clustering* relevan digunakan dalam bidang pendidikan dan analisis wilayah. Irwansyah et al. menggunakan *K-Means* untuk mengklasifikasikan kualitas pendidikan sekolah dasar di Indonesia [7]. Salman et al. membandingkan *K-Means* dan *K-Medoids* dalam pengelompokan sekolah berdasarkan ketersediaan sarana dan prasarana [8]. Wardani et al. mengelompokkan provinsi berdasarkan kelayakan ruang kelas dan tenaga kependidikan sekolah dasar [2]. Emilia et al. menerapkan *K-Means* untuk mengelompokkan kabupaten/kota berdasarkan indikator pembangunan manusia [9]. Penelitian lain juga menunjukkan bahwa metode *clustering* seperti DBSCAN, *Gaussian Mixture Model*, dan *Hierarchical Clustering* dapat digunakan untuk mengelompokkan objek berdasarkan karakteristik tertentu [10]. Selain itu, Faisal et al. menerapkan *K-Means* untuk menganalisis fasilitas sekolah pada tiap provinsi di Indonesia [11], sedangkan Yanto et al. menggunakan *K-Means* untuk menentukan kualitas satuan pendidikan berdasarkan nilai internal dan eksternal [12].

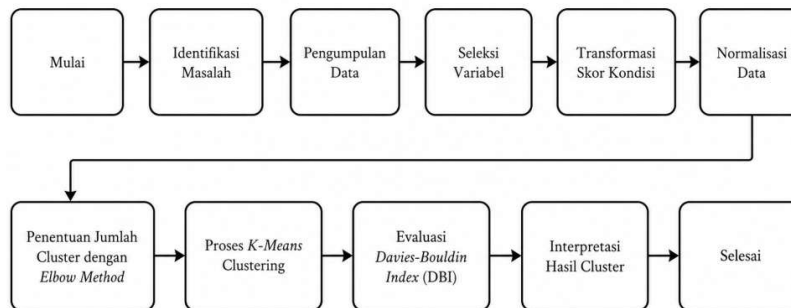
Meskipun demikian, penelitian sebelumnya masih memiliki ruang untuk dikembangkan. Sebagian penelitian dilakukan pada tingkat provinsi, sebagian lainnya hanya berfokus pada wilayah tertentu atau satuan pendidikan tertentu. Selain itu, indikator yang digunakan juga masih beragam, seperti kualitas pendidikan secara umum, kelayakan ruang kelas, tenaga kependidikan, atau indikator pembangunan manusia. Penelitian yang secara khusus mengelompokkan kondisi sarana dan prasarana sekolah dasar pada tingkat kabupaten/kota di seluruh Indonesia masih relatif terbatas. Padahal, analisis pada tingkat kabupaten/kota dapat memberikan gambaran yang lebih rinci dibandingkan tingkat provinsi karena mampu menunjukkan variasi kondisi fasilitas pendidikan antarwilayah secara lebih detail. Dibandingkan dengan penelitian terdahulu, penelitian ini memiliki perbedaan yang eksplisit.

Wardani et al. [2] dan Faisal et al. [11] melakukan pengelompokan pada tingkat provinsi, sehingga variasi kondisi antarwilayah di dalam satu provinsi belum dapat teramati, sedangkan penelitian ini menggunakan unit analisis kabupaten/kota secara nasional pada 514 wilayah sehingga menghasilkan pemetaan yang lebih rinci. Irwansyah et al. [7] mengukur kualitas pendidikan sekolah dasar secara umum, sedangkan penelitian ini berfokus khusus pada enam indikator sarana dan prasarana fisik. Salman et al. [8] membatasi cakupan penelitian pada sekolah di Provinsi Riau, sementara penelitian ini mencakup seluruh kabupaten/kota di Indonesia. Emilia et al. [9] mengelompokkan wilayah berdasarkan Indeks Pembangunan Manusia yang bersifat makro dan tidak spesifik pada fasilitas pendidikan dasar. Dengan demikian, kebaruan penelitian ini terletak pada kombinasi cakupan nasional tingkat kabupaten/kota, indikator sarana dan prasarana yang spesifik dan multidimensi, serta skema transformasi skor kondisi yang dijelaskan secara eksplisit, yang belum ditemukan secara bersamaan pada penelitian-penelitian sebelumnya.

Berdasarkan permasalahan tersebut, penelitian ini dilakukan untuk mengelompokkan sarana dan prasarana sekolah dasar kabupaten/kota di Indonesia menggunakan algoritma *K-Means Clustering*. Data yang digunakan berasal dari *Data Portal* Kementerian Pendidikan Dasar dan Menengah tahun 2025 dengan unit analisis sebanyak 514 kabupaten/kota. Indikator yang digunakan meliputi ruang kelas, perpustakaan, laboratorium, usaha kesehatan sekolah, toilet peserta didik laki-laki, dan toilet peserta didik perempuan. Hasil pengelompokan dalam penelitian ini tidak dimaksudkan sebagai penilaian mutlak terhadap baik atau buruknya suatu wilayah, melainkan sebagai pemetaan kondisi relatif berdasarkan kemiripan data. Dengan demikian, kontribusi utama penelitian ini adalah menghasilkan peta *cluster* berbasis data yang dapat digunakan oleh dinas pendidikan dalam memprioritaskan alokasi anggaran dan pembangunan sarana pendidikan dasar pada wilayah dengan kondisi relatif rendah, serta menyediakan kerangka kerja transformasi skor-normalisasi-*clustering* yang dapat direplikasi pada penelitian berikutnya dengan indikator yang lebih beragam.

2. METODE PENELITIAN

Penelitian ini dilakukan melalui beberapa tahap pengolahan data untuk menghasilkan kelompok kondisi sarana dan prasarana sekolah dasar kabupaten/kota di Indonesia. Tahapan yang dilakukan meliputi pengumpulan data, pemilihan variabel, perubahan data kondisi menjadi skor numerik, normalisasi, penentuan jumlah *cluster*, proses *K-Means Clustering*, serta evaluasi hasil. Urutan tahapan tersebut digunakan agar proses eksperimen lebih terarah dan hasil pengelompokan dapat dijelaskan secara runtut. Alur penelitian ditunjukkan pada Gambar 1.



Gambar 1. Alur Tahapan Penelitian

2.1. Pengumpulan Data

Data penelitian diperoleh dari *Data Portal* Kementerian Pendidikan Dasar dan Menengah tahun 2025 [3]. Cakupan data meliputi 514 kabupaten/kota di Indonesia, sehingga unit analisis dalam penelitian ini adalah wilayah kabupaten/kota. Data yang dipilih berhubungan dengan kondisi sarana dan prasarana sekolah dasar, meliputi ruang kelas, perpustakaan, laboratorium, usaha kesehatan sekolah, toilet peserta didik laki-laki, dan toilet peserta didik perempuan.

Setiap indikator pada data awal masih dinyatakan dalam bentuk jumlah fasilitas berdasarkan kategori kondisi. Kategori tersebut terdiri dari baik, rusak ringan, rusak sedang, dan rusak berat. Karena algoritma *K-Means* membutuhkan masukan berupa nilai numerik, data jumlah fasilitas pada setiap kategori kondisi tidak digunakan secara langsung. Data tersebut terlebih dahulu diubah menjadi skor kondisi agar setiap indikator dapat mewakili satu nilai yang dapat dihitung pada tahap berikutnya.

2.2. Transformasi Data Kondisi Menjadi Skor Numerik

Transformasi skor dilakukan untuk mengubah jumlah fasilitas pada tiap kategori kondisi menjadi satu nilai yang mewakili kondisi setiap indikator. Tahap ini diperlukan karena *K-Means* melakukan pengelompokan berdasarkan jarak antarobjek, sehingga variabel yang digunakan harus berbentuk numerik [6], [11]. Pada penelitian ini, kategori baik, rusak ringan, rusak sedang, dan rusak berat diperlakukan sebagai kategori berurutan karena menggambarkan tingkat kelayakan fasilitas dari kondisi terbaik sampai kondisi terendah [13].

Bobot diberikan berdasarkan urutan kelayakan tersebut. Kondisi baik memperoleh bobot 3 karena menunjukkan fasilitas paling layak, rusak ringan memperoleh bobot 2, rusak sedang memperoleh bobot 1, dan rusak berat memperoleh bobot 0. Dengan skema ini, jumlah fasilitas pada setiap kategori dapat digabungkan menjadi skor kondisi. Skema bobot yang digunakan ditampilkan pada Tabel 1.

Tabel 1. Bobot Kondisi Sarana dan Prasarana

Kondisi	Bobot
Baik	3
Rusak Ringan	2
Rusak Sedang	1
Rusak Berat	0

Skema bobot 3, 2, 1, dan 0 pada Tabel 1 tidak diambil dari satu standar penilaian pemerintah yang baku, mengingat Kemendikdasmen belum menerbitkan pedoman pembobotan tunggal untuk kondisi sarana dan prasarana sekolah. Oleh karena itu, bobot tersebut ditetapkan sebagai asumsi skala interval yang linier dan berjarak sama (*equal-interval*), sesuai dengan prinsip umum *ordinal encoding* yang menyatakan bahwa kategori kondisi yang berjenjang dapat dikonversi menjadi nilai numerik berurutan selama urutan tingkatannya tetap terjaga [14]. Urutan baik–rusak ringan–rusak sedang–rusak berat merepresentasikan tingkat kelayakan fasilitas secara ordinal, sehingga pemberian selisih bobot yang sama antarkategori (3, 2, 1, 0) merupakan pendekatan yang wajar dan umum digunakan ketika belum tersedia bobot empiris khusus untuk domain yang bersangkutan.

Skor kondisi dihitung dengan membandingkan nilai tertimbang dari seluruh kategori kondisi dengan nilai maksimum yang mungkin diperoleh pada indikator yang sama. Rumus skor kondisi ditunjukkan pada Persamaan (1).

$$\text{Skor kondisi} = \frac{(B \times 3) + (RR \times 2) + (RS \times 1) + (RB \times 0)}{T \times 3} \quad (1)$$

Pada Persamaan (1), B merupakan jumlah fasilitas berkondisi baik, RR rusak ringan, RS rusak sedang, RB rusak berat, dan T total fasilitas pada indikator yang sama. Nilai pembilang menunjukkan total nilai tertimbang dari seluruh kategori kondisi, sedangkan penyebut menunjukkan nilai maksimum yang mungkin diperoleh apabila seluruh fasilitas berada dalam kondisi baik. Pembagian dengan nilai maksimum dilakukan agar skor kondisi berada pada rentang 0 sampai 1.

Sebagai contoh, pada indikator ruang kelas Kabupaten Aceh Selatan terdapat 581 ruang kelas dalam kondisi baik, 658 rusak ringan, 882 rusak sedang, dan 160 rusak berat. Total ruang kelas pada wilayah tersebut adalah 2.281. Berdasarkan Persamaan (1), skor kondisi ruang kelas Kabupaten Aceh Selatan dihitung sebagai berikut.

$$\text{Skor kondisi} = \frac{(581 \times 3) + (658 \times 2) + (882 \times 1) + (160 \times 0)}{2281 \times 3} = \frac{3941}{6843} = 0,5759$$

Hasil tersebut menunjukkan bahwa skor kondisi ruang kelas Kabupaten Aceh Selatan adalah 0,5759. Proses perhitungan yang sama dilakukan pada seluruh indikator dan seluruh kabupaten/kota. Dengan demikian, data awal yang sebelumnya terdiri dari jumlah fasilitas pada kategori baik, rusak ringan, rusak sedang, dan rusak berat diubah menjadi enam variabel skor, yaitu skor ruang kelas, skor perpustakaan, skor laboratorium, skor UKS, skor toilet peserta didik laki-laki, dan skor toilet peserta didik perempuan.

2.3. Normalisasi Data

Setelah data ditransformasikan menjadi skor kondisi, tahap berikutnya adalah normalisasi data. Normalisasi dilakukan agar seluruh variabel berada pada skala yang sama sehingga tidak ada variabel yang memiliki pengaruh terlalu dominan dalam proses *clustering*. Metode normalisasi yang digunakan adalah *Min-Max Normalization*, yaitu metode yang mengubah nilai data ke dalam rentang tertentu berdasarkan nilai minimum dan maksimum pada setiap variabel [15]. Rumus *Min-Max Normalization* ditunjukkan pada Persamaan (2).

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (2)$$

Pada Persamaan (2), X' merupakan nilai hasil normalisasi, X merupakan nilai skor asli, $X_{\min}$ merupakan nilai minimum pada variabel yang sama, dan $X_{\max}$ merupakan nilai maksimum pada variabel yang sama. Hasil normalisasi berada pada rentang 0 sampai 1. Nilai hasil normalisasi inilah yang digunakan sebagai input dalam proses *K-Means Clustering*.

Tahap normalisasi dilakukan setelah seluruh indikator menghasilkan skor kondisi. Oleh karena itu, nilai minimum dan maksimum yang digunakan dalam normalisasi berasal dari hasil skor kondisi pada setiap variabel, bukan dari data jumlah fasilitas awal. Dengan demikian, hubungan antarp proses dalam penelitian ini adalah data awal pada setiap indikator diubah menjadi skor kondisi, kemudian skor kondisi tersebut dinormalisasi sebelum diproses menggunakan *K-Means*.

2.4. Penentuan Jumlah Cluster Menggunakan Elbow Method

Jumlah *cluster* ditentukan sebelum proses *K-Means* dilakukan. Penelitian ini menggunakan *Elbow Method* karena metode tersebut dapat membantu memilih jumlah *cluster* berdasarkan perubahan nilai *Within Cluster Sum of Squares* (WCSS) pada beberapa nilai K [6]. WCSS menggambarkan total jarak kuadrat antara data dan *centroid* pada *cluster* masing-masing. Nilai WCSS yang lebih kecil menunjukkan bahwa data dalam *cluster* semakin dekat dengan pusat kelompoknya.

Pengujian dilakukan pada nilai $K=2$ sampai $K=10$. Jumlah *cluster* tidak dipilih hanya berdasarkan nilai WCSS terkecil, karena semakin banyak *cluster* biasanya akan menurunkan WCSS tetapi dapat membuat hasil menjadi terlalu terpecah. Oleh karena itu, jumlah *cluster* ditentukan melalui titik siku, yaitu titik ketika penurunan WCSS mulai melambat. Nilai K pada titik tersebut digunakan sebagai dasar proses *clustering*.

2.5. Proses K-Means Clustering

Setelah jumlah *cluster* ditetapkan, data hasil normalisasi diproses menggunakan algoritma *K-Means*. Algoritma ini bekerja dengan mengukur kedekatan antara data dan *centroid*, kemudian memasukkan data ke *cluster* dengan jarak terdekat [6], [11]. Tahapan yang dilakukan meliputi penentuan jumlah *cluster*, pembentukan *centroid* awal, perhitungan jarak setiap kabupaten/kota terhadap *centroid*, pengelompokan data berdasarkan jarak terdekat, pembaruan *centroid* dari rata-rata anggota *cluster*, serta pengulangan proses sampai anggota *cluster* tidak berubah atau batas iterasi tercapai.

Pengukuran jarak dalam penelitian ini menggunakan *Euclidean Distance*. Ukuran jarak ini digunakan karena sesuai untuk menghitung kedekatan antarobjek pada data numerik multidimensi [6]. Rumus *Euclidean Distance* ditunjukkan pada Persamaan (3).

$$d(x, c) = \sqrt{\sum_{i=1}^n (x_i - c_i)^2} \quad (3)$$

Pada Persamaan (3), $d(x, c)$ merupakan jarak antara data x dan *centroid* c , x_i merupakan nilai data pada variabel ke- i , dan c_i merupakan nilai *centroid* pada variabel ke- i . Data akan dimasukkan ke *cluster* dengan jarak paling dekat terhadap *centroid*. Proses ini berulang hingga tidak terjadi perubahan anggota *cluster* atau hingga mencapai batas iterasi yang ditentukan.

2.6. Evaluasi Cluster Menggunakan Davies-Bouldin Index

Evaluasi hasil *clustering* dilakukan dengan *Davies-Bouldin Index* (DBI). DBI digunakan untuk menilai kualitas *cluster* dengan memperhatikan kedekatan data dalam *cluster* dan jarak antar*cluster* [15]. Semakin kecil nilai DBI, semakin baik kualitas pengelompokan karena anggota dalam *cluster* lebih rapat dan antar*cluster* lebih terpisah.

Pada penelitian ini, DBI digunakan sebagai evaluasi pendukung setelah jumlah *cluster* diperoleh melalui *Elbow Method*. Pemilihan jumlah *cluster* tidak hanya mempertimbangkan nilai evaluasi terkecil, tetapi juga kemudahan interpretasi hasil. Oleh karena itu, apabila nilai DBI pada dua jumlah *cluster* tidak berbeda jauh, maka pemilihan *cluster* tetap mempertimbangkan tujuan penelitian, yaitu menghasilkan kelompok kondisi relatif rendah, sedang, dan tinggi.

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil pengolahan data sarana dan prasarana sekolah dasar pada tingkat kabupaten/kota di Indonesia. Pembahasan disusun mulai dari deskripsi data, transformasi skor, normalisasi, penentuan jumlah *cluster*, evaluasi *cluster*, sampai interpretasi hasil akhir. Penyusunan tersebut dilakukan agar hubungan antara data awal, proses pengolahan, dan hasil *clustering* dapat dipahami dengan jelas.

3.1. Deskripsi Data Penelitian

Data yang dianalisis pada penelitian ini mencakup kondisi sarana dan prasarana sekolah dasar dari 514 kabupaten/kota di Indonesia. Enam indikator digunakan sebagai dasar pengelompokan, yaitu ruang kelas, perpustakaan, laboratorium, usaha kesehatan sekolah, toilet peserta didik laki-laki, dan toilet peserta didik perempuan. Pada setiap indikator, kondisi fasilitas dibedakan menjadi empat kategori, yaitu baik, rusak ringan, rusak sedang, dan rusak berat.

Data awal pada setiap indikator belum berbentuk skor tunggal, tetapi masih berupa jumlah fasilitas dalam masing-masing kategori kondisi. Oleh sebab itu, data perlu dikonversi terlebih dahulu menjadi skor kondisi agar dapat digunakan dalam proses pengolahan. Tabel 2 digunakan untuk menunjukkan contoh bentuk data awal pada indikator ruang kelas, sedangkan keseluruhan data 514 kabupaten/kota tetap digunakan dalam proses analisis.

Tabel 2. Contoh Data Awal Kondisi Ruang Kelas Sekolah Dasar

No	Kota/Kab	Baik	Rusak Ringan	Rusak Sedang	Rusak Berat
1	Kab. Aceh Selatan	581	658	882	160
2	Kab. Aceh Tenggara	458	664	476	142
3	Kab. Aceh Timur	1505	774	806	451
...	...	...	...	...	...
514	Kab. Sorong	492	240	152	84

3.2. Hasil Transformasi Skor Kondisi

Hasil transformasi skor menunjukkan perubahan data dari bentuk jumlah fasilitas menjadi nilai kondisi pada rentang 0 sampai 1. Nilai yang mendekati 1 menunjukkan bahwa proporsi fasilitas berkondisi baik lebih besar, sedangkan nilai yang mendekati 0 menunjukkan bahwa kondisi fasilitas relatif lebih rendah. Skor ini diperlukan agar setiap indikator dapat digunakan dalam perhitungan jarak pada *K-Means* [6], [11].

Pada indikator ruang kelas Kabupaten Aceh Selatan, jumlah fasilitas berkondisi baik adalah 581, rusak ringan 658, rusak sedang 882, dan rusak berat 160. Berdasarkan Persamaan (1), nilai skor kondisi yang diperoleh adalah 0,5759. Nilai tersebut merupakan hasil penggabungan kategori kondisi melalui pembobotan, bukan nilai yang langsung berasal dari data mentah. Dengan nilai tersebut, kondisi ruang kelas Kabupaten Aceh Selatan berada di antara nilai minimum dan maksimum, sehingga dapat dikategorikan berada pada tingkat menengah.

Proses yang sama diterapkan pada seluruh indikator dan seluruh kabupaten/kota. Setelah transformasi selesai, setiap wilayah memiliki enam skor kondisi, yaitu skor ruang kelas, perpustakaan, laboratorium, UKS,

toilet peserta didik laki-laki, dan toilet peserta didik perempuan. Skor inilah yang kemudian digunakan sebagai dasar pada tahap normalisasi.

3.3. Hasil Normalisasi Data

Normalisasi dilakukan agar seluruh variabel memiliki skala yang sebanding sebelum diproses menggunakan K-Means. Tahap ini diperlukan karena perhitungan jarak dapat dipengaruhi oleh perbedaan rentang nilai antarvariabel. Variabel dengan rentang nilai lebih besar dapat memberi pengaruh lebih dominan terhadap hasil *clustering* apabila tidak dinormalisasi [15]. Sebelum normalisasi dilakukan, nilai minimum dan maksimum pada setiap variabel skor ditentukan dari seluruh data kabupaten/kota. Hasil nilai minimum dan maksimum tersebut ditampilkan pada Tabel 3.

Tabel 3. Nilai Minimum dan Maksimum Setiap Variabel

Variabel	X Min	Kota/Kab (X Min)	X Max	Kota/Kab (X Max)
Ruang Kelas	0.346043	Kab. Lanny Jaya	0.946729	Kota Adm. Jakarta Utara
Toilet Siswa	0.285714	Kab. Puncak	0.951720	Kota Adm. Jakarta Utara
Toilet Siswi	0.309979	Kab. Simeulue	0.962376	Kota Adm. Jakarta Utara
Uks	0.000000	Kab. Intan Jaya	1.000000	Kab. Jayawijaya
Laboratorium	0.333333	Kab. Deiyai	0.974359	Kota Sabang
Perpustakaan	0.295238	Kab. Lanny Jaya	0.935780	Kota Adm. Jakarta Utara

Berdasarkan Tabel 3, setiap variabel memiliki nilai minimum dan maksimum yang berbeda. Perbedaan tersebut menunjukkan bahwa kondisi sarana dan prasarana antarwilayah tidak sama pada setiap indikator. Sebagai contoh, nilai skor kondisi ruang kelas Kabupaten Aceh Selatan adalah 0,5759. Berdasarkan Tabel 3, nilai minimum pada variabel ruang kelas adalah 0,346043 dan nilai maksimum adalah 0,946729. Dengan menggunakan Persamaan (2), nilai normalisasi ruang kelas Kabupaten Aceh Selatan dihitung sebagai berikut:

$$X^I = \frac{0,5759 - 0,346043}{0,946729 - 0,346043} = \frac{0,229857}{0,600686} = 0,382685$$

Hasil tersebut mendekati nilai normalisasi ruang kelas Kabupaten Aceh Selatan pada Tabel 4, yaitu 0,382685. Proses perhitungan yang sama dilakukan pada seluruh variabel dan seluruh kabupaten/kota. Setelah nilai minimum dan maksimum diperoleh, seluruh skor kondisi dinormalisasi ke dalam rentang 0 sampai 1. Data hasil normalisasi inilah yang digunakan sebagai input dalam proses *K-Means Clustering*. Contoh hasil normalisasi data ditunjukkan pada Tabel 4.

Tabel 4. Contoh Hasil Normalisasi Data

No	Kab/Kota	Ruang Kelas	Toilet Siswa	Toilet Siswi	UKS	Lab	Perpus
1	Kab. Aceh Selatan	0.382685	0.351369	0.348312	0.624650	0.454177	0.423036
2	Kab. Aceh Tenggara	0.437448	0.340998	0.308137	0.751244	0.356296	0.395293
3	Kab. Aceh Timur	0.501905	0.438084	0.427995	0.762411	0.373176	0.551221
.....							
514	Kota Sorong	0.632364	0.376490	0.546734	0.750000	0.780000	0.452051

3.4. Penentuan Jumlah Cluster dengan Elbow Method

Penentuan jumlah *cluster* dilakukan dengan menjalankan *K-Means* pada beberapa nilai K, yaitu K=2 sampai K=10. Setiap percobaan menghasilkan nilai WCSS yang menunjukkan akumulasi jarak kuadrat data terhadap *centroid* pada *cluster* masing-masing. Nilai tersebut dihitung menggunakan Persamaan (4), sehingga WCSS pada Tabel 5 merupakan hasil perhitungan dari seluruh data kabupaten/kota untuk setiap jumlah *cluster* yang diuji.

$$WCSS = \sum_{i=1}^k \sum_{x \in C_i} ||x - C_i|| \quad (4)$$

Pada Persamaan (4), k menunjukkan jumlah *cluster*, C_i menunjukkan *cluster* ke-i, x merupakan data anggota *cluster*, dan c_i merupakan *centroid* pada *cluster* tersebut. Pada K=3, nilai WCSS yang diperoleh adalah 25,5790. Nilai ini mengalami penurunan dari K=2 yang memiliki WCSS 34,6870. Perhitungannya adalah sebagai berikut:

$$\text{Penurunan WCSS} = 34,6870 - 25,5790 = 9,1080$$

Penurunan berikutnya dari K=3 ke K=4 lebih kecil, yaitu dari 25,5790 menjadi 21,1412. Perhitungannya adalah sebagai berikut:

$$\text{Penurunan WCSS} = 25,5790 - 21,1412 = 4,4378$$

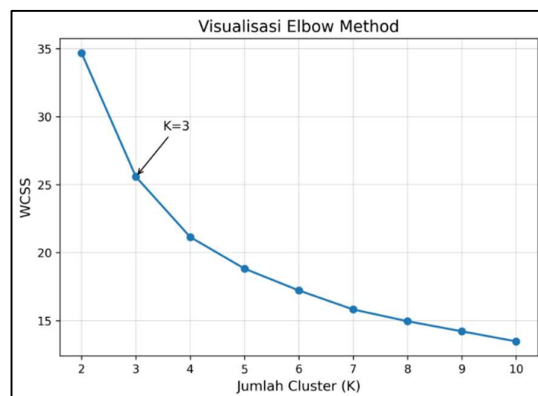
Perbandingan tersebut menunjukkan bahwa penurunan WCSS mulai melambat setelah K=3. Oleh karena itu, K=3 dipilih sebagai jumlah *cluster* karena berada pada titik siku *Elbow Method*. Hasil perhitungan WCSS ditampilkan pada Tabel 5.

Tabel 5. Nilai WCSS pada K=2 sampai K=10

Jumlah Cluster (K)	Nilai WCSS	Penurunan dari K Sebelumnya
2	34.6870	-
3	25.5790	9.1080← (titik siku)
4	21.1412	4.4378
5	18.8146	2.3266
6	17.2097	1.6048
7	15.8110	1.3987
8	14.9458	0.8652
9	14.2033	0.7425
10	13.4616	0.7417

Berdasarkan Tabel 5, nilai WCSS terus menurun seiring bertambahnya jumlah *cluster*. Penurunan terbesar terjadi dari K=2 ke K=3, yaitu sebesar 9,1080. Setelah K=3, penurunan WCSS mulai mengecil, misalnya dari K=3 ke K=4 hanya menurun sebesar 4,4378 dan terus menurun lebih kecil pada K berikutnya. Pola tersebut menunjukkan bahwa titik siku berada pada K=3.

Visualisasi *Elbow Method* digunakan untuk memperjelas perubahan nilai WCSS pada setiap jumlah *cluster*. Pada grafik tersebut, titik siku terlihat pada K=3 karena setelah nilai tersebut penurunan WCSS mulai melambat. Visualisasi hasil *Elbow Method* ditunjukkan pada Gambar 2.



Gambar 2. Visualisasi *Elbow Method*

3.5. Evaluasi Cluster Menggunakan Davies-Bouldin Index

Setelah jumlah *cluster* diperoleh melalui *Elbow Method*, hasil pengelompokan dievaluasi kembali menggunakan *Davies-Bouldin Index* (DBI). Evaluasi ini digunakan untuk melihat kualitas *cluster* berdasarkan kepadatan anggota di dalam *cluster* dan jarak antar*cluster*. Nilai DBI yang lebih rendah menunjukkan bahwa data dalam *cluster* lebih rapat dan pemisahan antar*cluster* lebih baik [16].

Rumus *Davies-Bouldin Index* ditunjukkan pada Persamaan (5).

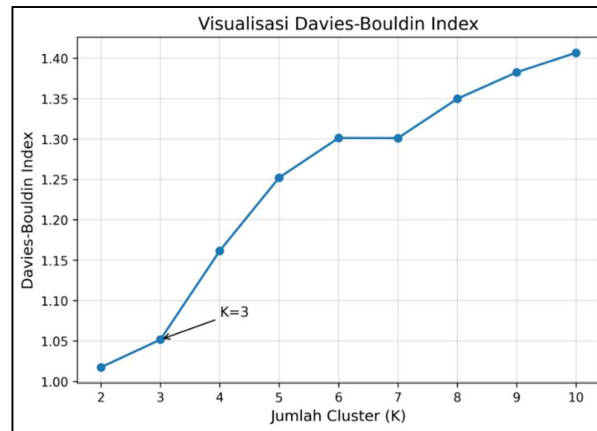
$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (5)$$

Pada Persamaan (5), k merupakan jumlah *cluster*, S_i merupakan rata-rata jarak data dalam *cluster* ke-i terhadap *centroid cluster* tersebut, S_j merupakan rata-rata jarak data dalam *cluster* ke-j terhadap *centroid cluster* tersebut, dan M_{ij} merupakan jarak antara *centroid cluster* ke-i dan *cluster* ke-j. Nilai DBI diperoleh dengan membandingkan kepadatan dalam *cluster* dan jarak antar*cluster*. Semakin kecil nilai DBI, semakin baik kualitas *cluster* yang terbentuk. Pada penelitian ini, nilai DBI untuk K=2 adalah 1,0175, sedangkan K=3 menghasilkan nilai 1,0517. Selisih kedua nilai tersebut dihitung sebagai berikut:

$$\text{Selisih DBI} = 1,0517 - 1,0175 = 0,0342$$

Selisih sebesar 0,0342 menunjukkan bahwa kualitas *cluster* pada $K=3$ tidak berbeda jauh dari $K=2$. Oleh karena itu, $K=3$ tetap digunakan karena didukung oleh titik siku pada *Elbow Method* dan lebih mudah ditafsirkan menjadi tiga kelompok kondisi relatif, yaitu rendah, sedang, dan tinggi.

Perubahan nilai DBI pada setiap jumlah *cluster* divisualisasikan untuk memperjelas perbandingan kualitas hasil pengelompokan. Visualisasi tersebut ditampilkan pada Gambar 3.



Gambar 3. Visualisasi Davies-Bouldin Index

Berdasarkan Gambar 3, nilai DBI paling rendah berada pada $K=2$, sedangkan $K=3$ memiliki nilai yang sedikit lebih tinggi. Meskipun demikian, $K=3$ tetap digunakan karena selisihnya kecil dan hasil *Elbow Method* menunjukkan titik siku pada $K=3$. Selain itu, penggunaan tiga *cluster* lebih sesuai dengan kebutuhan interpretasi, yaitu membedakan wilayah ke dalam kondisi relatif rendah, sedang, dan tinggi.

3.6. Hasil Clustering K-Means

Setelah jumlah *cluster* ditetapkan sebesar $K=3$, data hasil normalisasi dikelompokkan menggunakan K-Means. Proses ini membagi kabupaten/kota ke dalam tiga kelompok berdasarkan kedekatan pola nilai pada enam indikator sarana dan prasarana. Karakteristik masing-masing kelompok dilihat dari *centroid* akhir, yaitu nilai pusat *cluster* yang menunjukkan kecenderungan rata-rata setiap variabel [6]. Rumus *centroid* ditunjukkan pada Persamaan (6).

$$C_i = \frac{1}{n_i} \sum_{x \in C_i} x \quad (6)$$

Pada Persamaan (6), c_i menunjukkan *centroid* pada *cluster* ke- i , n_i menunjukkan jumlah data dalam *cluster* tersebut, dan x menunjukkan data anggota *cluster*. Dengan demikian, nilai pada Tabel 6 menggambarkan rata-rata setiap variabel dari wilayah yang menjadi anggota *cluster* yang sama.

Tabel 6. Centroid Akhir K-Means

Cluster	Ruang Kelas	Toilet Siswa	Toilet Siswi	UKS	Laboratorium	Perpustakaan	Rata-rata
Cluster 1 (Rendah)	0,3324	0,3157	0,3110	0,6343	0,3521	0,3611	0,4011
Cluster 2 (Sedang)	0,4840	0,4942	0,4957	0,7094	0,5162	0,5096	0,5349
Cluster 3 (Tinggi)	0,6951	0,6890	0,6920	0,7829	0,6858	0,6864	0,7052

Rata-rata pada Tabel 6 digunakan untuk membantu membaca kecenderungan umum setiap *cluster*. Sebagai contoh, *Cluster 2* memiliki rata-rata *centroid* sebesar 0,5349 yang diperoleh dari enam variabel, yaitu ruang kelas, toilet siswa, toilet siswi, UKS, laboratorium, dan perpustakaan. Perhitungannya adalah sebagai berikut:

$$\text{Rata-rata Cluster 2} = \frac{0,4840 + 0,4942 + 0,4957 + 0,7094 + 0,5162 + 0,5096}{6} = 0,5349$$

Nilai tersebut berada di antara rata-rata *Cluster 1* dan *Cluster 3*, sehingga *Cluster 2* menggambarkan kelompok dengan kondisi relatif sedang. Berdasarkan Tabel 6, *Cluster 1* memiliki rata-rata *centroid* sebesar 0,4011 sehingga dikategorikan sebagai kelompok dengan kondisi relatif rendah. *Cluster 2* memiliki rata-rata

centroid sebesar 0,5349 sehingga mewakili kelompok dengan kondisi relatif sedang. Sementara itu, *Cluster 3* memiliki rata-rata *centroid* sebesar 0,7052 dan menjadi kelompok dengan kondisi relatif tinggi.

Label rendah, sedang, dan tinggi digunakan untuk membaca posisi relatif antar*cluster*. Label tersebut bukan penilaian mutlak terhadap kualitas suatu wilayah, tetapi menunjukkan perbandingan kondisi berdasarkan data yang dianalisis. Dengan demikian, wilayah pada *cluster* tinggi hanya berarti memiliki nilai sarana dan prasarana yang relatif lebih baik dibandingkan dua *cluster* lainnya. Selain nilai *centroid*, hasil *clustering* juga dilihat melalui jumlah anggota pada setiap *cluster*. Distribusi ini digunakan untuk mengetahui sebaran kabupaten/kota pada masing-masing kelompok kondisi. Hasil distribusi kabupaten/kota per *cluster* ditunjukkan pada Tabel 7.

Tabel 7. Distribusi Kabupaten/Kota per *Cluster*

<i>Cluster</i>	Label Kondisi	Jumlah	Persentase
<i>Cluster 1</i>	Kondisi Relatif Rendah	117 kab/kota	22,8%
<i>Cluster 2</i>	Kondisi Relatif Sedang	302 kab/kota	58,8%
<i>Cluster 3</i>	Kondisi Relatif Tinggi	95 kab/kota	18,5%
Total		514 kab/kota	100%

Nilai persentase pada Tabel 7 diperoleh dari perbandingan jumlah anggota pada setiap *cluster* terhadap total data, yaitu 514 kabupaten/kota. Perhitungan persentase tersebut ditunjukkan pada Persamaan (7).

$$\text{Persentase Cluster} = \frac{\text{Jumlah anggota cluster}}{\text{Total data}} \times 100\% \quad (7)$$

Contoh perhitungan persentase pada *Cluster 1* disajikan sebagai berikut:

$$\text{Persentase Cluster 1} = \frac{117}{514} \times 100\% = 22,8\%$$

Hasil pada Tabel 7 menunjukkan bahwa *Cluster 2* memiliki jumlah anggota paling banyak, yaitu 302 kabupaten/kota atau 58,8% dari total data. *Cluster 1* terdiri dari 117 kabupaten/kota atau 22,8%, sedangkan *Cluster 3* berisi 95 kabupaten/kota atau 18,5%. Komposisi tersebut menunjukkan bahwa sebagian besar wilayah berada pada kelompok kondisi relatif sedang. Dengan demikian, peningkatan fasilitas sekolah dasar tetap perlu diperhatikan, terutama pada wilayah yang termasuk dalam *cluster* rendah, serta tetap mempertimbangkan wilayah pada *cluster* sedang.

Selain dilihat dari nilai *centroid* dan jumlah anggota, hasil *clustering* juga ditelaah berdasarkan sebaran wilayahnya untuk melihat apakah kondisi sarana dan prasarana yang relatif rendah cenderung terkonsentrasi pada kawasan tertentu. Penelusuran dilakukan dengan mengelompokkan provinsi asal setiap kabupaten/kota anggota *cluster* ke dalam kawasan Indonesia Barat, Tengah, dan Timur, dengan kawasan Indonesia Timur mencakup Papua (seluruh provinsi pemekaran), Maluku, Maluku Utara, dan Nusa Tenggara. Secara keseluruhan, wilayah Indonesia Timur hanya menyumbang 95 dari 514 kabupaten/kota (18,5%) pada data penelitian ini, sehingga proporsi tersebut dapat dijadikan acuan pembandingan pada tiap *cluster*.

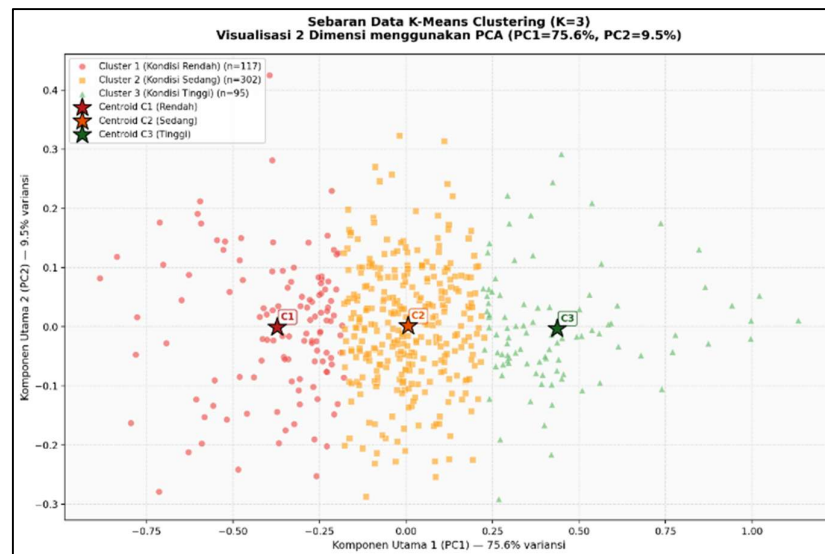
Tabel 8. Proporsi Anggota *Cluster* yang Berasal dari Kawasan Indonesia Timur

<i>Cluster</i>	Label Kondisi	Wilayah Indonesia Timur	Proporsi Total Anggota <i>Cluster</i>
1	Rendah	37 dari 117 kab/kota	31,6%
2	Sedang	47 dari 302 kab/kota	15,6%
3	Tinggi	11 dari 95 kab/kota	11,6%

Berdasarkan Tabel 8, wilayah Indonesia Timur menyumbang 31,6% anggota pada *Cluster 1* (Rendah), jauh di atas proporsi rata-ratanya secara nasional sebesar 18,5%, sementara proporsinya menurun menjadi 15,6% pada *Cluster 2* (Sedang) dan hanya 11,6% pada *Cluster 3* (Tinggi). Pola ini menunjukkan bahwa kabupaten/kota di kawasan Indonesia Timur cenderung lebih banyak berada pada kelompok kondisi sarana dan prasarana yang relatif rendah dibandingkan proporsinya secara nasional, meskipun tidak seluruh anggota *cluster* rendah berasal dari kawasan tersebut karena sebagian lainnya tersebar di Sumatera dan Sulawesi. Temuan ini konsisten dengan disparitas pembangunan pendidikan antara kawasan barat dan timur Indonesia yang telah dilaporkan pada tingkat provinsi [17], sehingga hasil pengelompokan pada tingkat kabupaten/kota dalam penelitian ini memperkuat

gambaran ketimpangan tersebut dengan cakupan wilayah yang lebih rinci. Pendekatan klasterisasi K-Means yang dipadukan dengan visualisasi seperti pada penelitian ini juga diterapkan pada domain lain, seperti pengelompokan data musik digital [18], dan analisis tren lowongan pekerjaan menggunakan *clustering* dan *social network analysis* [19], yang menegaskan fleksibilitas metode ini pada berbagai jenis data.

Visualisasi pada Gambar 4 digunakan untuk memperlihatkan pola pemisahan hasil *clustering* dalam ruang dua dimensi. Setiap titik merepresentasikan kabupaten/kota, sedangkan warna atau simbol menunjukkan kelompok *cluster* yang terbentuk. Berdasarkan Gambar 4, data membentuk tiga kelompok utama sesuai hasil *K-Means*. *Cluster 1* berada pada kelompok dengan nilai kondisi lebih rendah, *Cluster 2* berada pada kelompok menengah, dan *Cluster 3* berada pada kelompok yang relatif lebih tinggi. Beberapa titik tampak berdekatan karena terdapat wilayah dengan karakteristik fasilitas yang mirip. Namun, secara umum visualisasi tersebut menunjukkan bahwa proses *clustering* mampu membentuk pemisahan kelompok berdasarkan kemiripan data.



Gambar 4. Sebaran Data K-Means Clustering

4. KESIMPULAN

Penelitian ini menghasilkan pengelompokan kondisi sarana dan prasarana sekolah dasar pada 514 kabupaten/kota di Indonesia menggunakan algoritma *K-Means Clustering*. Data awal yang terdiri dari kategori baik, rusak ringan, rusak sedang, dan rusak berat terlebih dahulu diubah menjadi skor numerik, kemudian dinormalisasi menggunakan *Min-Max Normalization*. Jumlah *cluster* ditentukan menggunakan *Elbow Method*, yang menunjukkan titik siku pada $K=3$. Evaluasi menggunakan *Davies-Bouldin Index* menunjukkan bahwa nilai $K=3$ tidak berbeda jauh dari $K=2$, sehingga $K=3$ tetap digunakan karena lebih sesuai untuk interpretasi kondisi relatif rendah, sedang, dan tinggi.

Hasil akhir menunjukkan bahwa 117 kabupaten/kota berada pada *cluster* kondisi relatif rendah, 302 kabupaten/kota berada pada *cluster* kondisi relatif sedang, dan 95 kabupaten/kota berada pada *cluster* kondisi relatif tinggi. Sebagian besar wilayah berada pada kelompok sedang, sehingga peningkatan sarana dan prasarana sekolah dasar masih diperlukan, terutama pada wilayah dalam *cluster* rendah. Hasil pengelompokan ini tidak dimaksudkan sebagai penilaian mutlak terhadap kualitas wilayah, melainkan sebagai pemetaan berbasis kemiripan data. Penelitian berikutnya dapat menambahkan indikator lain, seperti jumlah peserta didik, jumlah guru, akreditasi sekolah, akses internet, atau kondisi geografis, serta membandingkan hasil dengan metode *clustering* lain. Secara keseluruhan, kontribusi utama penelitian ini adalah menghasilkan peta *cluster* kondisi sarana dan prasarana sekolah dasar berbasis data yang dapat digunakan sebagai acuan objektif oleh dinas pendidikan dalam memprioritaskan alokasi anggaran dan pembangunan sarana pada kabupaten/kota di *cluster* rendah, serta menyediakan kerangka kerja transformasi skor-normalisasi-*clustering* yang dapat direplikasi pada penelitian berikutnya.

DAFTAR PUSTAKA

- [1] S. A. Nurfatimah, S. Hasna, and D. Rostika, "Membangun Kualitas Pendidikan di Indonesia dalam Mewujudkan Program Sustainable Development Goals (SDGs)," *J. Basicedu*, vol. 6, no. 4, pp. 6145–6154, 2022, doi:

- <https://doi.org/10.31004/basicedu.v6i>.
- [2] M. T. Wardani, V. O. Ramadhani, N. Anggreani, Sumanto, and A. D. Kuswanto, "Pengelompokan Provinsi Berdasarkan Kelayakan Ruang Kelas dan Tenaga Kependidikan Sekolah Dasar Menggunakan Algoritma K-Means: Analisis Data Periode 2023-2024," *Riau J. Tek. Inform.*, vol. 4, no. 2, pp. 163–172, 2025, doi: 10.30606/rjti.v4i2.3389.
 - [3] "Statistik Sekolah Dasar SD Tahun 2024/2025," Kementerian Pendidikan Dasar dan Menengah. Accessed: Jun. 25, 2026. [Online]. Available: <https://www.kemendikdasmen.go.id/en>
 - [4] A. Edo and M. Yasin, "Dampak Kesenjangan Akses Pendidikan dan Faktor Ekonomi Keluarga terhadap Mobilitas Sosial," *J. Ilmu Pendidik. Sos.*, vol. 2, no. 3, pp. 317–326, 2024, doi: 10.71382/sinova.v2i3.175.
 - [5] N. Wijaya, "Performa Algoritma K-Means dan Fuzzy C-Means Dalam Analisis Kluster Pendidikan di Tingkat Kecamatan Jakarta," *Jurnal Komputer dan Informatika*. Accessed: Jun. 25, 2026. [Online]. Available: <https://journal.untar.ac.id/index.php/JKI/article/view/34485>
 - [6] S. Maryam, R. Astuti, and F. M. Basysyar, "Optimalisasi Jumlah Cluster Data Sekolah Dasar (Sd) Menggunakan Algoritma K-Means Clustering," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 7, no. 6, pp. 3640–3646, 2024, doi: 10.36040/jati.v7i6.8246.
 - [7] M. A. Irwansyah, Y. N. Alinda, R. N. Aini, and I. A. Alfitriah, "Analisis K-Means Dengan Rapidminer Untuk Klasifikasi Kualitas Pendidikan Sekolah Dasar Di Indonesia," *Zo. J. Sist. Inf.*, vol. 7, no. 1, pp. 1008–1020, 2025, doi: 10.31849/a0hhsc57.
 - [8] M. D. Salman *et al.*, "Perbandingan Kinerja Algoritma Clustering K-Means dan K-Medoids dalam Pengelompokan Sekolah di Provinsi Riau Berdasarkan Ketersediaan Sarana dan Prasarana," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. 3, pp. 797–806, 2025, doi: 10.57152/malcom.v5i3.1950.
 - [9] K. Emilia, A. S. Rahayu, D. P. Yuliani, D. A. Prasetya, and P. A. Riyantoko, "Application of K-Means Clustering for Regency/City Clustering in East Java Based on 2024 Human Development Index Indicators," *J. Apl. Sains Data*, vol. 1, no. 2, pp. 85–95, 2025, doi: 10.33005/jasid.v1i2.21.
 - [10] K. E. Setiawan and A. Kurniawan, "Pengelompokan Rumah Sakit di Jakarta Menggunakan Model DBSCAN, Gaussian Mixture, dan Hierarchical Clustering," *J. Inform. Terpadu*, vol. 9, no. 2, pp. 149–156, 2023, doi: <https://doi.org/10.54914/jit.v9i2.995>.
 - [11] M. Faisal, J. Ilham Saputro, and W. Sri Utami, "Analysis Cluster K-Means Terhadap Fasilitas Sekolah Pada Tiap Provinsi Republik Indonesia," *Bull. Inf. Technol.*, vol. 6, no. 4, pp. 408–416, 2025, doi: 10.47065/bit.v5i2.2340.
 - [12] D. Yanto, H. Susanto, K. Zulkifli, and F. R. Gupron, "Penerapan Algoritma K-Means Dalam Menentukan Kualitas Satuan Pendidikan Berdasarkan Nilai Internal Dan Eksternal," *JEECOM J. Electr. Eng. Comput.*, vol. 5, no. 2, pp. 319–328, 2023, doi: 10.33650/jeecon.v5i2.6946.
 - [13] F. Bolikulov, R. Nasimov, A. Rashidov, F. Akhmedov, and Y. I. Cho, "Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms," *Mathematics*, vol. 12, no. 16, pp. 1–21, 2024, doi: 10.3390/math12162553.
 - [14] P. Koukaras and C. Tjortjis, "Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices," *AI*, vol. 6, no. 10, pp. 1–40, 2025, doi: 10.3390/ai6100257.
 - [15] R. R. Laska and A. M. Yolanda, "A Comparative Study of Z-Score and Min-Max Normalization for Rainfall Classification in Pekanbaru," *J. Data Sci.*, vol. 2024, no. 4, pp. 1–8, 2024, doi: 10.61453/jods.v2024no04.
 - [16] Y. Hasan, "Pengukuran Silhouette Score dan Davies-Bouldin Index pada Hasil Cluster K-Means dan Dbscan," KAKIFIKOM (Kumpulan Artikel Karya Ilmiah Fakultas Ilmu Komputer). Accessed: Jun. 25, 2026. [Online]. Available: <https://ejournal.ust.ac.id/index.php/KAKIFIKOM/article/view/3938>
 - [17] R. S. T. Sihombing, "Pemerataan Pendidikan: Studi Kasus 34 Provinsi di Indonesia," *Parahyangan Econ. Dev. Rev.*, vol. 1, no. 2, pp. 143–151, 2023, doi: 10.26593/pedr.v1i2.6670.
 - [18] B. Hakim, F. J. Kaunang, C. Susanto, J. Salim, and R. Indradjaja, "Implementasi Machine Learning Dalam Pengelompokan Musik Menggunakan Algoritma K-Means Clustering," *IDEALIS Indones. J. Inf. Syst.*, vol. 8, no. 1, pp. 74–83, 2025, doi: 10.36080/idealis.v8i1.3357.
 - [19] T. A. J. Harahap and C. B. Santoso, "Analisis Tren Lowongan Pekerjaan Software Engineering Di Indonesia Dengan Clustering Dan Social Network Analysis," *IDEALIS Indones. J. Inf. Syst.*, vol. 8, no. 2, pp. 160–171, 2025, doi: 10.36080/idealis.v8i2.3518.