Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 3 July 2025, pp: 797-806
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Comparison of K-Means and K-Medoids Clustering Algorithm Performance in Grouping Schools in Riau Province Based on Availability of Facilities and Infrastructure

Perbandingan Kinerja Algoritma *Clustering* K-Means dan K-Medoids dalam Pengelompokan Sekolah di Provinsi Riau Berdasarkan Ketersediaan Sarana dan Prasarana

Muhammad Dzaki Salman[1*], Rahmaddeni[2], Nanda Rizki Pratama[3], M. Nakhlah Farid A[4], Ahmad Agung Setiawan[5], fenisya Zalianti[6], Isra Bil Huda[7]

[1,2,3,4,5,6,7]Program Studi Teknik Informatika, Universitas Sains dan Teknologi Indonesia, Indonesia

E-Mail: [1]muhammaddzakisalman@gmail.com, [2]Rahmaddeni@usti.ac.id, [3]rizkiperawan28@gmail.com, [4]nakhlahfarid@gmail.com, [5]bwaahmadagung@gmail.com, [6]deswita0999@gmail.com, [7]israbilhuda239@gmail.com

Received Feb 09th 2025; Revised Apr 10th 2025; Accepted May 13th 2025; Available Online Jun 19th 2025, Published Jun 22th 2025
Corresponding Author: Muhammad Dzaki Salman

Abstract

Quality education is strongly influenced by the availability of adequate facilities and infrastructure. This study aims to compare the performance of two clustering algorithms, namely K-Means and K-Medoids, in clustering 497 public schools in Riau Province consisting of elementary, junior high, high school, and vocational school levels. The data analyzed included the number of teachers, students, classrooms, laboratories, internet access, sanitation, and accreditation status. Data were obtained from the Riau Province Education Office and BPS, then analyzed through Exploratory Data Analysis (EDA), preprocessing, and dimension reduction with Principal Component Analysis (PCA). Evaluation results using Davies-Bouldin Index (DBI) with k=3 showed that K-Medoids produced more separated and better quality clusters (0.61) than K-Means (0.80). The advantage of K-Medoids lies in its resistance to outliers and uneven data distribution. The results of this study can be used as a reference in planning a more equitable and targeted education policy in Riau Province.

Keyword: Clustering, Davies-Bouldin Index, K-Means, K-Medoids, Principal Component Analysis

Abstrak

Pendidikan yang berkualitas sangat dipengaruhi oleh ketersediaan sarana dan prasarana yang memadai. Penelitian ini bertujuan untuk membandingkan kinerja dua algoritma *clustering*, yaitu *K-Means* dan *K-Medoids*, dalam mengelompokkan 497 sekolah negeri di Provinsi Riau yang terdiri dari jenjang SD, SMP, SMA, dan SMK. Data yang dianalisis meliputi jumlah guru, siswa, ruang kelas, laboratorium, akses internet, sanitasi, dan status akreditasi. Data diperoleh dari Dinas Pendidikan dan Badan Pusat Statistik (BPS) Provinsi Riau, kemudian dianalisis melalui *Exploratory Data Analysis* (EDA), *preprocessing*, dan reduksi dimensi dengan *Principal Component Analysis* (PCA). Hasil evaluasi menggunakan *Davies-Bouldin Index* (DBI) dengan k=3 menunjukkan bahwa *K-Medoids* menghasilkan *cluster* yang lebih terpisah dan lebih baik (0,61) dibandingkan *K-Means* (0,80). Keunggulan *K-Medoids* terletak pada ketahanannya terhadap *outlier* dan distribusi data yang tidak merata. Hasil penelitian ini dapat digunakan sebagai acuan dalam perencanaan kebijakan pendidikan yang lebih merata dan tepat sasaran di Provinsi Riau.

Kata Kunci: *Clustering,* Davies-Bouldin Index, *K-Means, K-Medoids,* Principal Component Analysis

1. PENDAHULUAN

Pendidikan memegang peranan yang sangat penting dalam mendorong pembangunan suatu daerah. Pentingnya pendidikan sebagai indikator pembangunan juga tercermin dalam tujuan *Sustainable Development Goals* (SDGs), yaitu "Menjamin kualitas pendidikan yang inklusif dan merata, serta

mendukung kesempatan belajar seumur hidup bagi semua." Salah satu cara untuk mewujudkan hal ini adalah dengan memastikan bahwa fasilitas pendidikan terdistribusi secara merata dan memadai di seluruh wilayah, termasuk daerah yang lebih terpencil dan kurang berkembang [1].

Pendidikan yang berkualitas merupakan modal utama bagi suatu bangsa untuk maju. Dalam pelaksanaannya, pendidikan di Indonesia umumnya dilaksanakan di sekolah yang bertujuan untuk mempersiapkan generasi muda agar dapat menghadapai masa depan yang lebih baik. Namun, meskipun upaya peningkatan kualitas pendidikan Indonesia telah dilakukan, masalah distribusi fasilitas pendidikan yang tidak merata masih menjadi tantangan besar yang harus dihadapi, khususnya di daerah-daerah yang jauh dari pusat- pusat kota besar. Dalam perbandingan dengan negara-negara tetangga, kualitas pendidikan Indonesia masih tertinggal [1].

Ketimpangan dalam distribusi sarana dan prasarana sekolah merupakan salah satu faktor utama yang menyebabkan perbedaan kualitas pendidikan di berbagai wilayah. Sekolah-sekolah di daerah perkotaan cenderung memiliki fasilitas yang lebih lengkap dibandingkan dengan sekolah di daerah terpencil, yang dapat berdampak pada kesenjangan kualitas pembelajaran. Oleh karena itu, diperlukan metode untuk mengelompokkan sekolah berdasarkan ketersediaan fasilitas guna mengidentifikasi daerah yang membutuhkan perhatian khusus dalam perencanaan kebijakan pendidikan [2].

Clustering sekolah berdasarkan ketersediaan sarana dan prasarana dapat menjadi salah satu solusi untuk mengatasi permasalahan ketidakmerataan fasilitas pendidikan. Dengan pengelompokan yang tepat, pemerintah dapat mengidentifikasi daerah-daerah yang membutuhkan perhatian khusus dalam hal penyediaan fasilitas dan sumber daya pendidikan. Hal ini penting untuk memastikan bahwa semua sekolah memiliki akses yang setara terhadap fasilitas yang dapat mendukung kualitas pendidikan [3].

Algoritma *clustering* telah banyak digunakan dalam mengelompokkan objek berdasarkan karakteristik tertentu, seperti DBSCAN, *Gaussian Mixture Model* (GMM), dan *Agglomerative Hierarchical Clustering* (AHC). Dalam penelitian oleh Kurniawan (2023), model GMM menunjukkan performa terbaik dalam mengelompokkan rumah sakit di Jakarta berdasarkan jumlah tenaga medis dan fasilitas tempat tidur, dengan nilai DBI sebesar 0.6457, lebih baik dibandingkan DBSCAN dan AHC [4].

Selain itu, hasil perbandingan antara *K-Means* dan *K-Medoids* pada penelitian yang dilakukan oleh Farahdina (2019) menunjukkan bahwa *K-Medoids* lebih *robust* terhadap *outlier* dibandingkan *K-Means. K-Means* menggunakan *centroid* yang merupakan rata-rata dari objek dalam *cluster*, sehingga dapat terpengaruh oleh nilai ekstrem. Sebaliknya, *K-Medoids* menggunakan objek representatif *(medoids)* sebagai pusat *cluster*, yang tidak terpengaruh oleh *outlier*, sehingga lebih stabil dalam pengelompokan data yang memiliki nilai ekstrem[5]. Temuan ini relevan untuk penelitian ini karena data fasilitas dan infrastruktur sekolah yang digunakan mengandung nilai ekstrem akibat variasi yang signifikan dalam fasilitas dan infrastruktur antar sekolah. Oleh karena itu, penerapan metode *K-Medoids* diharapkan mampu memberikan hasil pengelompokan yang lebih akurat dan stabil.

Penelitian oleh Tusyakdiah Halima (2023) melakukan implementasi *K-Means* dan *K-Medoids* untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan aspek pendidikan pemuda. Hasilnya, K-Means memiliki performa lebih baik dalam hal rasio simpangan baku (0,527941) dibandingkan *K-Medoids* (0,5612719) [6]. Selain itu, Damanik (2019) menerapkan algoritma *K-Medoids* untuk mengelompokkan desa-desa di Indonesia berdasarkan ketersediaan fasilitas sekolah. Penelitian ini menegaskan keunggulan *K-Medoids* dalam menangani data dengan kemungkinan *outlier* dan urutan masukan data yang tidak seragam [7].

Dalam penelitian ini, *K-Means* dan *K-Medoids* dipilih karena sifatnya yang lebih sederhana dan banyak digunakan dalam analisis *clustering.* Untuk mengevaluasi hasil *clustering,* digunakan metode *Davies-Bouldin Index* (DBI) untuk mengukur seberapa baik *cluster* yang terbentuk. DBI mengevaluasi rasio jarak antara *cluster* dan variasi dalam *cluster,* sehingga membantu dalam menilai efektivitas setiap metode *clustering* yang diterapkan. Penelitian ini juga menerapkan *Principal Component Analysis* (PCA) sebagai teknik reduksi dimensi sebelum proses *clustering,* yang bertujuan untuk meningkatkan akurasi hasil *clustering.*

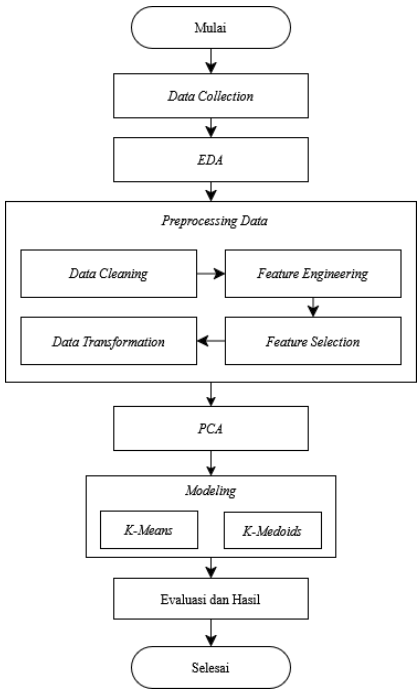
Penelitian ini bertujuan untuk membandingkan kinerja algoritma *K-Means* dan *K-Medoids* dalam pengelompokan sekolah-sekolah di Provinsi Riau menjadi tiga *cluster* (k=3) berdasarkan ketersediaan sarana dan prasarana. Hasil *clustering* ini dapat digunakan oleh pemerintah daerah sebagai dasar perumusan kebijakan pendidikan yang lebih merata dan tepat sasaran, seperti pengalokasian anggaran, prioritas pembangunan infrastruktur pendidikan, serta distribusi tenaga pengajar dan sumber daya lainnya sesuai dengan kebutuhan masing-masing sekolah.

2. METODOLOGI PENELITIAN

Metodologi penelitian ini dibagi dalam enam tahap dapat dilihat pada gambar 1. Adapan penjelasannya taitu:

1. *Data Collection,* Data sekolah di Provinsi Riau dikumpulkan dari Dinas Pendidikan Provinsi Riau dan Badan Pusat Statistik (BPS) Provinsi Riau.

2. EDA, Data yang dikumpulkan kemudian dianalisis menggunakan metode EDA untuk memahami distribusi data dan identifikasi *outlier*.
3. *Data Preprocessing*, Data yang telah dianalisis kemudian diproses untuk menghilangkan *missing value* dan melakukan normalisasi data.
4. *Modeling*, Algoritma *K-Means* dan *K-Medoids* digunakan untuk mengelompokkan sekolah di Provinsi Riau berdasarkan ketersediaan sarana dan prasarana.
5. *PCA* adalah alat analisis data yang digunakan dalam berbagai disiplin ilmu untuk mengidentifikasi struktur tersembunyi dalam kumpulan data[8].
6. Evaluasi dan Hasil, Hasil pengelompokan sekolah dievaluasi menggunakan indeks evaluasi DBI.



Gambar 1. Metode Penelitian

2.1. Data Collection

Data collection didefinisikan sebagai proses mengumpulkan, mengukur, dan merekam data yang akurat dan relevan dari berbagai sumber untuk tujuan analisis dan pengambilan keputusan. Tujuan utama *data collection* adalah untuk mengumpulkan informasi dan data yang akurat dan relevan untuk membuat keputusan bisnis yang tepat, meningkatkan kualitas produk atau jasa, meningkatkan efisiensi operasional, dan meningkatkan kepuasan pelanggan [9].

Dalam melakukan *data collection*, terdapat beberapa metode yang dapat digunakan, antara lain observasi, wawancara, kuesioner, analisis dokumen, dan eksperimen. *Data collection* sangat penting dalam berbagai bidang, termasuk bisnis, kesehatan, pendidikan, dan lain-lain. Dengan mengumpulkan data yang akurat dan relevan, organisasi dapat membuat keputusan yang tepat, meningkatkan kualitas produk atau jasa, dan meningkatkan efisiensi operasional [9].

2.2. *Exploratory Data Analysis (*EDA*)*

EDA adalah proses menganalisis dan menampilkan data bertujuan mendapatkan pemahaman yang lebih baik tentang wawasan dari data[10]. Ada berbagai langkah yang dilakukan saat melakukan EDA, berikut ini adalah langkah-langkah umum yang dapat diambil dalam melakukan analisis EDA data:

1. Memaksimalkan wawasan ke dalam kumpulan data.
2. Mengungkap struktur data.
3. Ekstrak variabel yang penting.
4. Mendeteksi outlier dan anomali.
5. Melakukan uji asumsi.
6. Mengembangkan model.
7. Menentukan faktor yang optimal.

Kebanyakan Teknik EDA adalah berbentuk grafis dengan beberapa Teknik kuantitatif. Peran utama EDA adalah untuk mengeksplorasi data secara terbuka, dan grafik bertujuan memperkuat analisis yang dilakukan[10]. Berikut adalah beberapa jenis teknik grafis sederhana yang banyak digunakan.

1. *Plotting* data mentah seperti *data traces, histograms, bihistograms, probability plots, lag plots, block plots, dan Youden plots.*
2. *Plotting* statistik sederhana seperti *mean plots, standard deviation plots, box plots.*

2.3. *Preprocessing Data*

Preprocessing Data adalah serangkaian langkah yang dilakukan untuk mempersiapkan data mentah menjadi format yang lebih baik untuk dianalisis atau digunakan oleh model pembelajaran mesin [11]. Tujuan utama dari *data preprocessing* adalah meningkatkan kualitas data, mengatasi masalah yang mungkin ada, dan memastikan bahwa data siap digunakan dalam proses analisis atau pelatihan model [12]. *Proses preprocessing* data yang digunakan untuk penelitian ini adalah sebagai berikut.

1. *Feature Engineering*, adalah proses mengubah data mentah menjadi fitur-fitur baru yang lebih informatif untuk analisis atau model prediktif, guna meningkatkan kualitas data dan kinerja model [13].
2. *Feature Selecion,* adalah suatu kegiatan pemodelan atau penganalisaan data yang umumnya dapat dilakukan secara *preprocessing* dan bertujuan untuk memilih fitur yang berpengaruh (fitur optimal) dan mengesampingkan fitur yang tidak berpengaruh[14].
3. *Label Encoding,* adalah teknik dalam pembelajaran mesin untuk mengubah data kategori menjadi numerik [15].
4. *Data Normalization,* adalah proses mengubah data ke dalam rentang nilai yang seragam, biasanya antara 0 dan 1, untuk memperbaiki kualitas data dan meningkatkan efisiensi algoritma [16].
5. *Outlier Handling,* adalah teknik untuk mengidentifikasi dan menghilangkan pengaruh data yang tidak biasa [17][18].
6. *Missing Value Handling,* adalah teknik untuk mengatasi kehilangan data yang dapat mempengaruhi akurasi analisis [19].

2.4. *Principal Component Analysis (*PCA*)*

PCA adalah teknik analisis multivariat yang digunakan untuk mengurangi dimensi data dan mengidentifikasi pola-pola yang terkait dalam data. PCA bekerja dengan mengubah variabel-variabel asli menjadi komponen-komponen baru yang tidak terkait dan memiliki variansi yang maksimum. Dengan demikian, PCA dapat membantu mengidentifikasi struktur data yang tersembunyi, mengurangi kompleksitas data, dan meningkatkan akurasi model prediksi [20][8]. Selain itu, PCA juga dapat digunakan untuk mengurangi *noise* dalam data dan mengidentifikasi *outlier* [21].

Dalam proses PCA, terdapat beberapa langkah yang harus dilakukan, yaitu pengumpulan data, normalisasi data, perhitungan kovariansi, perhitungan *eigenvalue* dan *eigenvector,* seleksi komponen utama, dan transformasi data. Dengan melakukan langkah-langkah tersebut, PCA dapat membantu mengidentifikasi komponen-komponen yang paling berpengaruh dalam data dan mengurangi dimensi data menjadi lebih sederhana [22]. Proses PCA meliputi langkah-langkah berikut:

1. Menentukan *eigenvalue*

$$\det(A - \lambda I) = 0 \tag{1}$$

Dimana A adalah matriks kovariansi dan λ merupakan nilai *eigen.* Nilai ini mencerminkan besarnya variansi data sepanjang suatu arah tertentu.

2. *Menentukan eigenvector*

$$(A - \lambda I)v = 0 \tag{2}$$

Vektor ini menunjukkan arah komponen utama, yang menjadi dasar dalam pembentukan ruang fitur baru.

3. Menghitung proporsi variansi masing-masing komponen:

$$P_i = \frac{\lambda_i}{\Sigma_{j=1}^{n} \lambda_j} \tag{3}$$

P_i menyatakan proporsi variansi dari komponen ke- i, dan λ_i merupakan nilai *eigen* yang bersesuaian.

4. Menghitung kumulatif proporsi variansi:

$$CP_r = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{j=1}^{n} \lambda_j}, dengan\ \lambda_i > \lambda_2 > \cdots > \lambda_n \quad (4)$$

Nilai *eigen* diurutkan dari terbesar ke terkecil untuk membantu menentukan jumlah komponen yang cukup untuk merepresentasikan data.

5. Transformasi akhir ke dalam ruang komponen utama:

$$Z = XW \quad (5)$$

Mentransformasikan data ke ruang dimensi baru, di mana X adalah data terstandarisasi dan W adalah matriks vektor *eigen*.

PCA juga memiliki beberapa kelemahan, seperti tidak dapat menangani data yang tidak berdistribusi normal dan tidak dapat menangani data yang memiliki banyak *outlier* [20]. Oleh karena itu, perlu dilakukan pengecekan dan pengolahan data sebelum melakukan analisis PCA.

2.5. *Clustering*

Clustering atau pengelompokan data merupakan suatu teknik penting dalam analisis data yang bertujuan untuk mengidentifikasi kesamaan dan pola dalam data, serta mengelompokkan data yang serupa ke dalam kelompok-kelompok yang berbeda. Sebagai salah satu metode *unsupervised learning*, *clustering* memungkinkan kita untuk menemukan struktur dan pola dalam kumpulan data yang tidak berlabel. Dalam *clustering*, data dibagi menjadi sejumlah kelompok yang memiliki kesamaan karakteristik yang lebih besar dibandingkan dengan kelompok lainnya [23].

Algoritma *clustering* sangat tergantung pada jenis data, tujuan, dan aplikasi yang spesifik. Dalam analisis *cluster*, algoritma dapat digunakan sebagai alat deskriptif atau eksplorasi untuk memperoleh *insights* dari data. Oleh karena itu, tidak jarang beberapa algoritma *clustering* dicoba pada data yang sama untuk memperoleh hasil yang optimal. Secara umum, metode *clustering* dapat diklasifikasikan ke dalam beberapa kategori, salah satunya adalah metode partisi (*partitioning method*). Metode partisi ini melibatkan penentuan awal jumlah kelompok, diikuti dengan realokasi objek secara iteratif untuk menemukan kembali kelompok-kelompok yang optimal. Dua algoritma yang populer dalam metode partisi ini adalah *K-Means* dan *K-Medoids* [23].

2.6. *K-Means*

K-Means adalah salah satu metode data clustering non-hierarki yang mempartisi data menjadi satu atau lebih cluster berdasarkan karakteristik yang sama. Tujuan dari metode ini adalah untuk mengelompokkan data yang memiliki karakteristik yang sama ke dalam satu *cluster* dan data yang memiliki karakteristik yang berbeda ke dalam *cluster* lainnya [24].

K-Means adalah metode *clustering* berbasis jarak yang membagi data ke dalam sejumlah *cluster*. Algoritma ini hanya bekerja pada atribut *numeric* dan termasuk dalam kategori *partitioning clustering*. *K-Means* memisahkan data ke dalam k daerah bagian yang terpisah dan sangat terkenal karena kemudahan dan kemampuannya untuk mengklaster data yang besar dan data *outlier* dengan sangat cepat [25].

Dalam algoritma *K-Means*, setiap data harus termasuk ke dalam *cluster* tertentu dan dapat berpindah ke *cluster* lainnya pada tahap berikutnya. Algoritma ini merupakan metode non-hierarki yang pada awalnya mengambil sebagian besar komponen populasi untuk dijadikan pusat *cluster* awal. Pusat *cluster* dipilih secara acak dari sekumpulan populasi data [25].

Kemudian, *K-Means* menguji masing-masing komponen di dalam populasi data dan menandai komponen tersebut ke salah satu pusat *cluster* yang telah didefinisikan tergantung dari jarak minimum antar komponen dengan tiap-tiap *cluster*. Posisi pusat *cluster* akan dihitung kembali sampai semua komponen data digolongkan ke dalam tiap-tiap pusat *cluster* dan terakhir akan terbentuk posisi pusat *cluster* yang baru [26].

Kelebihan utama dari algoritma K-Means adalah kesederhanaan dan efisiensinya dalam mengolah data berskala besar. *K-Means* memiliki waktu komputasi yang relatif cepat karena hanya membutuhkan perhitungan rata-rata untuk setiap iterasi. Selain itu, algoritma ini dapat dengan mudah diimplementasikan dan dikombinasikan dengan metode lain dalam proses analisis lanjutan. Namun demikian, *K-Means* memiliki kelemahan penting, yaitu sangat sensitif terhadap outlier dan nilai ekstrim, karena pusat cluster dihitung berdasarkan rata-rata yang mudah terdistorsi oleh data yang ekstrem. *K-Means* juga mensyaratkan pengguna untuk menentukan jumlah cluster (k) secara manual, yang sering kali sulit diketahui tanpa eksplorasi awal terhadap data. Selain itu, algoritma ini kurang optimal ketika digunakan pada data yang memiliki bentuk *cluster* tidak sferis atau distribusi yang tidak merata antar *cluster* [27].

2.7. *K-Medoids*

K-Medoids adalah sebuah metode clustering yang merupakan varian dari metode *K-Means.* Metode ini dikembangkan oleh Kaufman dan Rousseeuw (1987) untuk mengatasi kelemahan metode *K-Means* yang sensitif terhadap *outlier* dan *noise* dalam data. *K-Medoids* berbeda dengan *K-Means* dalam beberapa hal, yaitu representasi *cluster,* kriteria pemilihan *cluster,* dan *robustness* terhadap *outlier* [7].

K-Medoids menggunakan objek data yang sebenarnya sebagai representasi *cluster,* sedangkan *K-Means* menggunakan *mean* (rata-rata) dari objek data dalam *cluster.* Selain itu, *K-Medoids* menggunakan jarak antara objek data dan *medoid cluster* sebagai kriteria pemilihan *cluster,* sedangkan *K-Means* menggunakan jarak antara objek data dan *mean cluster. K-Medoids* juga lebih *robust* terhadap *outlier* dan *noise* dalam data karena menggunakan *medoid* sebagai representasi *cluster* [28].

Algoritma ini juga lebih stabil untuk data yang memiliki distribusi tidak merata atau struktur cluster yang kompleks. Berdasarkan penelitian Marlina dkk, (2018), implementasi *K-Medoids* dalam pengelompokan data sebaran anak cacat di Provinsi Riau menunjukkan kinerja yang lebih baik dibandingkan *K-Means,* dengan nilai validasi *Silhouette Coefficient* sebesar 0.5009, jauh lebih tinggi dibanding K-Means yang hanya sebesar 0.1443 [29]. Hasil ini menunjukkan bahwa K-Medoids mampu membentuk cluster yang lebih konsisten dan terpisah dengan baik.

Namun demikian, algoritma ini memiliki kelemahan dari sisi komputasi, karena membutuhkan perhitungan jarak antar semua pasangan titik dalam *cluster,* sehingga kompleksitas waktunya lebih tinggi dibanding *K-Means.* Proses pemilihan *medoid* baru juga dapat menjadi mahal secara waktu terutama pada *dataset* berskala besar. Selain itu, seperti halnya *K-Means,* algoritma ini juga mengharuskan penentuan jumlah *cluster* k di awal, yang bisa menjadi tantangan dalam tahap eksplorasi awal data.

3. HASIL DAN PEMBAHASAN

3.1. *Data Collection*

Penelitian ini menggunakan data sekolah jenjang Sekolah Dasar (SD), Sekolah Menengah Pertama (SMP), Sekolah Menengah Akhir (SMA), dan Sekolah Menengah Kejuruan (SMK) negeri di Provinsi Riau yang diperoleh dari situs Badan Pusat Statistik (BPS). Data yang dikumpulkan mencakup berbagai aspek pendidikan, seperti jumlah guru, jumlah, pegawai, jumlah siswa, jumlah ruang kelas, jumlah lab, akses air, akses internet, sanitasi, dan akreditasi dengan total data sebanyak 497 baris. Data yang digunakan dapat dilihat pada Tabel 1.

Tabel 1. Data Sekolah Provinsi Riau

No	Jumlah Guru	Jumlah PTK	Jumlah Siswa Laki-laki	Jumlah Siswa Perempuan	Rombongan Belajar	Jumlah Kelas	Lab	Internet	...	Akreditasi
1	40	8	448	413	27	16	1	Ada	...	A
2	23	1	231	213	12	6	0	Ada	...	B
3	9	1	51	51	6	0	0	Tidak	...	A
4	10	1	123	123	8	6	1	Tidak	...	A
5	61	17	573	637	34	5	0	Ada	...	A
...	...	...	...	...	...	...	...	...	...	...
497	11	1	78	76	6	5	0	Tidak	...	B

3.2. *Preprocessing Data*

Dataset yang digunakan dalam penelitian ini menjalani tahapan *preprocessing* yang meliputi *data cleaning* dengan imputasi untuk menangani *missing values*, diikuti *feature engineering* untuk menciptakan atau menggabungkan fitur baru yang lebih relevan. Setelah itu, dilakukan *feature selection* untuk menghapus variabel yang tidak signifikan. Tahap selanjutnya adalah *data transformation,* di mana variabel dengan distribusi miring ditransformasikan menggunakan *log transformation* untuk mendekatkan distribusi data ke bentuk normal. Berikut *dataset* setelah dilakukan *preprocessing data* ditunjukkan pada Tabel 2.

Tabel 2. *Preprocessing Data*

No	Jumlah Guru	Jumlah PTK	Jumlah Siswa	Rombongan Belajar	Jumlah Kelas	Lab	Internet	...	Akreditasi
1	3.713572	2.197225	6.759255	3.332205	3.258097	0.693147	0.693147	...	0.693147
2	3.178054	0.693147	6.098074	2.564949	2.564949	0.000000	0.693147	...	1.098612
3	2.302585	0.693147	4.682131	1.945910	1.945910	0.000000	0.000000	...	0.693147
4	2.397895	0.693147	5.513429	2.197225	2.079442	0.000000	0.000000	...	0.693147
5	4.127134	2.890372	7.099202	3.555348	3.610918	1.791759	0.693147	...	0.693147
...	...	...	...	...	...	...	...	...	...
497	2.484907	0.693147	5.043425	1.945910	1.791759	0.000000	0.000000	...	1.098612

3.3. Principal Component Analysis (PCA)

Pada tahap ini, metode PCA digunakan untuk mereduksi *dataset* menjadi dua komponen utama, yaitu PC1 dan PC2. PCA berguna meningkatkan efisiensi komputasi, mengatasi *curse of dimensionality*, dan menghilangkan redundansi data. Dengan mengubah data ke dalam komponen utama, PCA membantu mempermudah interpretasi serta visualisasi pola dalam *dataset*, terutama dalam *clustering* dan klasifikasi. Selain itu, PCA memungkinkan analisis yang lebih fokus pada fitur paling signifikan, sehingga menghasilkan representasi data yang lebih efektif tanpa kehilangan informasi penting[30]**.** Berikut *dataset* setelah melalui proses PCA yang ditunjukkan pada Tabel 3.

Tabel 3. Hasil PCA

No	PC1	PC2
1	2.521525	0.311275
2	-0.143047	0.058620
3	-1.199501	1.022920
4	-0.891423	0.247006
5	3.957351	-0.796619
...	...	...
497	-1.438864	0.244363

3.4. Modeling

Pada tahap ini, dilakukan *modeling* dari algoritma *K-Means* dan *K-Medoids* untuk melakukan *clustering* dan menetapkan k=3 pada masing-masing algoritma.

3.4.1. *K-Means*

Data yang telah melewati proses *preprocessing* dan reduksi dimensi (PCA) akan diolah menggunakan metode *K-Means clustering.* Algoritma ini bekerja dengan menentukan *centroid* awal secara acak, mengelompokkan data berdasarkan jarak ke *centroid* terdekat, dan memperbarui posisi *centroid* hingga *cluster* stabil.

```
K-Means Model
Input: df_pca
Output: df_pca
  Initialization n_cluster, random_state
    Get kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
  kmeans.fit(df_pca)
```

3.4.2. *K-Medoids*

Modeling K-Medoids clustering juga menggunakan data yang sama pada tahap modeling *K-Means. K-Medoids* menggunakan titik data aktual *(medoid)* sebagai pusat *cluster*, sehingga lebih tahan terhadap *outlier.*

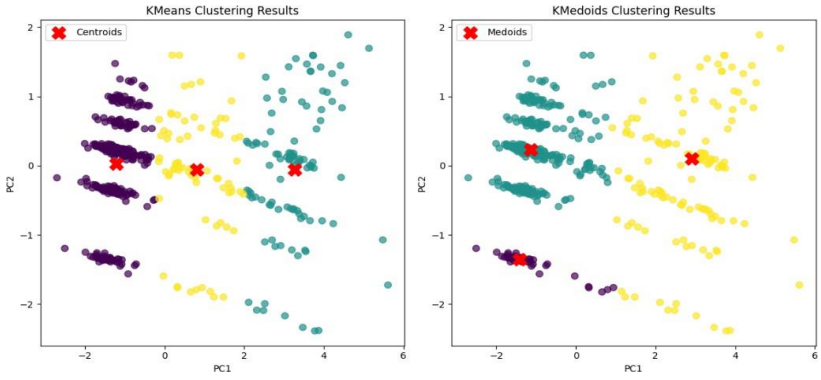
```
K-Medoids Model
Input: df_pca
Output: df_pca
    Initialization n_cluster, random_state
    Get kmedoids = Kmedoids(n_clusters=3, random_state=42)
  kmedoids.fit(df_pca)
```

3.5. Evaluasi

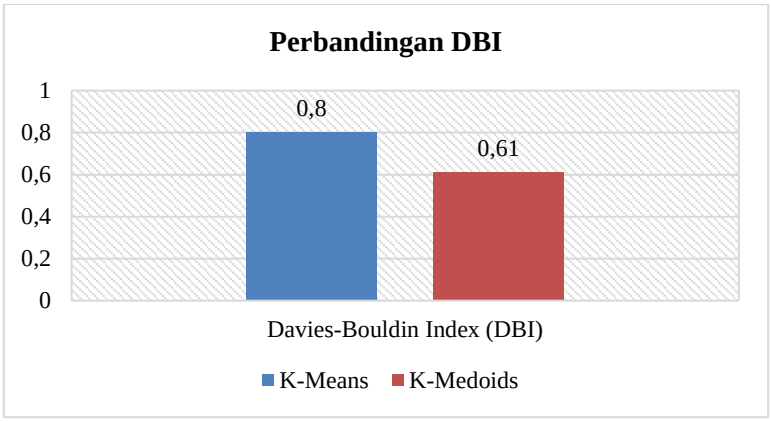
Pada penelitian ini, algoritma *K-Means* dan *K-Medoids* digunakan untuk mengelompokkan data sekolah di Provinsi Riau berdasarkan ketersediaan sarana dan prasarana. Jumlah *cluster* ditentukan sebanyak tiga (k = 3), sehingga hasil *clustering* akan membentuk tiga kelompok. *Plotting* hasil *clustering* dari kedua algoritma ditampilkan pada Gambar 2.

Gambar 2 menunjukkan data yang direduksi menjadi dua dimensi (PC1 dan PC2) menggunakan teknik PCA. *Cluster* yang dihasilkan algoritma *K-Means* menunjukkan lebih banyak tumpang tindih *(overlapping)* dibandingkan dengan algoritma *K-Medoids.* Titik pusat *cluster (centroid)* pada algoritma *K-Means* terbentuk berdasarkan rata-rata posisi seluruh anggota *cluster.* Sementara itu, pada hasil *clustering K-Medoids, medoid* yang terbentuk merupakan titik data yang meminimalkan total jarak terhadap semua titik lain dalam *cluster.*

Selanjutnya melakukan evaluasi hasil *clustering* menggunakan metode DBI. DBI dapat digunakan tanpa memerlukan label data atau *ground truth*, sehingga cocok untuk evaluasi *clustering* pada data yang tidak berlabel[31]. Perbandingan nilai DBI dapat dilihat pada Gambar 3.

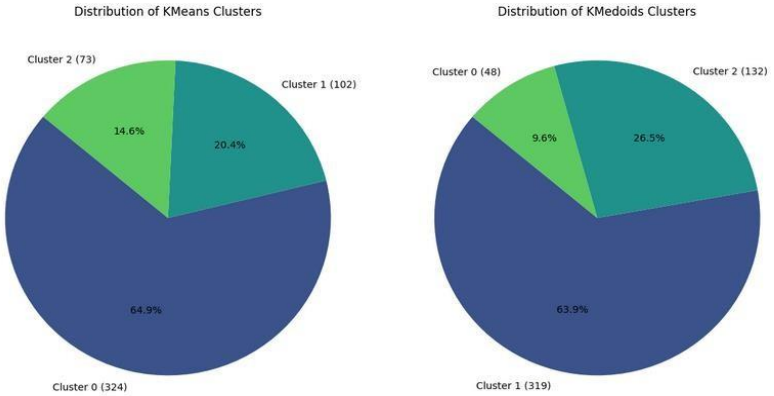


Gambar 2. *Ploting* Hasil *Clustering*



Gambar 3. Perbandingan DBI Algoritma *K-Means* dan *K-Medoids*

K-Medoids memiliki DBI yang lebih rendah (0,61) dibandingkan dengan *K-Means* (0,80), yang menunjukkan bahwa *K-Medoids* memiliki kualitas *clustering* yang lebih baik. Selanjutnya distribusi data dalam tiga *cluster* yang dihasilkan oleh algoritma *K-Means* dan *K-Medoids* ditunjukkan pada Gambar 4.



Gambar 4. Distribusi Data *Cluster* Algoritma *K-Means* dan *K-Medoids*

Distribusi hasil *clustering* menunjukkan bahwa baik *K-Means* maupun *K-Medoids* menghasilkan satu *cluster* dominan dengan persentase terbesar, yaitu *cluster* 0 pada *K-Means* (64.9%) dan *cluster* 1 pada *K-Medoids* (63.9%).

4. DISKUSI

Hasil *clustering* menunjukkan bahwa algoritma *K-Medoids* memiliki nilai *Davies-Bouldin Index* (0,61) yang lebih rendah daripada *K-Means* (0,80), menandakan bahwa *K-Medoids* membentuk *cluster* yang lebih terpisah dan konsisten. Hal ini disebabkan oleh metode pemilihan pusat *cluster* pada *K-Medoids* yang tidak terpengaruh oleh nilai ekstrem, sehingga lebih stabil ketika digunakan pada data pendidikan yang memiliki *outlier*, seperti sekolah dengan fasilitas sangat minim atau sangat lengkap.

Temuan ini sejalan dengan studi sebelumnya (Farahdinna dkk, 2019) dan (Marlina dkk, 2018) yang menunjukkan keunggulan *K-Medoids* dalam mengelola data yang tidak merata dan mengandung *noise.* Dibandingkan *K-Means* yang sensitif terhadap distribusi data yang tidak merata, *K-Medoids* memberikan hasil yang lebih representatif terhadap kondisi nyata sekolah-sekolah di Riau.

Dari hasil *clustering*, diketahui bahwa terdapat satu *cluster* dominan yang merepresentasikan mayoritas sekolah dengan fasilitas dasar yang cukup, hal ini mencerminkan bahwa kondisi rata-rata sekolah di Riau yang tidak kekurangan ekstrem tapi juga belum mencapai level terbaik. sedangkan dua *cluster* lainnya mewakili sekolah dengan infrastruktur sangat baik atau sangat buruk. Namun demikian, hasil *clustering* belum mencakup variabel sosial-ekonomi atau letak geografis sekolah, yang juga berpotensi memengaruhi hasil kebijakan. Oleh karena itu, interpretasi hasil *clustering* harus dikombinasikan dengan kajian kontekstual lapangan untuk implementasi kebijakan yang tepat sasaran.

5. KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma *K-Medoids* memberikan performa lebih baik dibandingkan *K-Means* dalam mengelompokkan sekolah di Provinsi Riau berdasarkan ketersediaan sarana dan prasarana, sebagaimana ditunjukkan oleh nilai DBI yang lebih rendah. Hal ini menegaskan keunggulan *K-Medoids* dalam menangani data dengan *outlier* dan distribusi yang tidak merata. Hasil pengelompokan ini dapat dijadikan sebagai dasar dalam merumuskan kebijakan pendidikan yang lebih merata, seperti alokasi anggaran, distribusi guru, dan pembangunan infrastruktur pendidikan yang berbasis pada kebutuhan nyata sekolah. Kelemahan dari penelitian ini terletak pada keterbatasan variabel yang digunakan, yang hanya mencakup indikator fisik sarana dan prasarana tanpa mempertimbangkan aspek sosial, geografis, dan kualitas pembelajaran. Penelitian selanjutnya disarankan untuk mengintegrasikan lebih banyak dimensi, termasuk data spasial dan faktor-faktor non-fisik lainnya, serta mengeksplorasi algoritma *clustering* lainnya untuk validasi hasil.

REFERENSI

[1] S. A. Nurfatimah, S. Hasna, dan D. Rostika, "Membangun Kualitas Pendidikan di Indonesia dalam Mewujudkan Program Sustainable Development Goals (SDGs)," *Jurnal Basicedu*, vol. 6, no. 4, hlm. 6145–6154, Mei 2022, doi: 10.31004/basicedu.v6i4.3183.

[2] A. Edo dan M. Yasin, "Dampak Kesenjangan Akses Pendidikan dan Faktor Ekonomi Keluarga terhadap Mobilitas Sosial," 2024.

[3] T. Hidayat dan A. Kosasih, "Analisis Peraturan Menteri Pendidikan Dan Kebudayaan Republik Indonesia Nomor 22 Tahun 2016 Tentang Standar Proses Pendidikan Dasar Dan Menengah Serta Implikasinya Dalam Pembelajaran Pai Di Sekolah," 2019.

[4] K. E. Setiawan dan A. Kurniawan, "Jurnal Informatika Terpadu Pengelompokan Rumah Sakit Di Jakarta Menggunakan Model Dbscan, Gaussian Mixture, Dan Hierarchical Clustering," *Jurnal Informatika Terpadu*, vol. 9, no. 2, hlm. 149–156, 2023, [Daring]. Tersedia pada: https://journal.nurulfikri.ac.id/index.php/JIT

[5] F. Farahdinna, I. Nurdiansyah, A. Suryani, dan A. Wibowo, "Perbandingan Algoritma K-Means Dan K-Medoids Dalam Klasterisasi Produk Asuransi Perusahaan Nasional," *Jurnal Ilmiah FIFO*, vol. 11, no. 2, hlm. 208, Nov 2019, doi: 10.22441/fifo.2019.v11i2.010.

[6] Tusyakdiah Halima *dkk.*, "Implementasi Metode K-Means Dan K-Medoids Pada Pengelompokan Provinsi Indonesia Berdasarkan Aspek Pendidikan Pemuda," *Community Services Social Work Bulletin*, 2023, doi: http://dx.doi.org/10.31000/cswb.v3i1.10153.

[7] S. Rahayu Ningsih, I. Sudahri Damanik, A. Perdana Windarto, H. Satria Tambunan, A. Wanto, dan S. A. Tunas Bangsa Pematangsiantar Jl Jenderal Sudirman Blok No, "Prosiding Seminar Nasional Riset Information Science (SENARIS) Analisis K-Medoids Dalam Pengelompokkan Penduduk Buta Huruf Menurut Provinsi," 2019.

[8] T. Kurita, "Principal component analysis (PCA)," *Computer vision: a reference guide (pp. 1013-1016). Cham: Springer International Publishing*, 2021.

[9] S. A. Mazhar, "Methods of Data Collection: A Fundamental Tool of Research," *Journal of Integrated Community Health*, vol. 10, no. 01, hlm. 6–10, Jun 2021, doi: 10.24321/2319.9113.202101.

[10] A. Chandra, "Memahami Data Dengan Exploratory Data Analysis," 2019.

[11] J. Khatib Sulaiman, Z. Azhari, L. Efrizoni, W. Agustin, R. Yanti, dan S. AMIK Riau Pekanbaru, "Opinion Mining menggunakan Algoritma Deep Learning untuk Menganalisis Penggunaan Aplikasi Jamsostek Mobile," *Indonesian Journal of Computer Science Attribution*, vol. 12, no. 2, hlm. 2023–666, 2023.

[12] M. A. Rofiq dan A. Qoiriah, "Pengelompokan Kategori Buku Berdasarkan Judul Menggunakan Algoritma Agglomerative Hierarchical Clustering Dan K-Medoids," *Journal of Informatics and Computer Science*, vol. 02, 2021.

[13] R. Riyaddulloh dan A. Romadhony, "Normalisasi Teks Bahasa Indonesia Berbasis Kamus Slang Studi Kasus: Tweet Produk Gadget Pada Twitter," *Agustus*, vol. 8, no. 4, 2021.
[14] F. F. Firdaus, H. A. Nugroho, dan I. Soesanti, "A Review of Feature Selection and Classification Approaches for Heart Disease Prediction," 2020.
[15] F. Bolikulov, R. Nasimov, A. Rashidov, F. Akhmedov, dan Y. I. Cho, "Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms," *Mathematics*, vol. 12, no. 16, Agu 2024, doi: 10.3390/math12162553.
[16] J. Brownlee, "Machine Learning Resource Guide," 2014.
[17] W. Zhou, H. Zhu, W. Chen, C. Chen, dan J. Xu, "Outlier Handling Strategy of Ensembled-Based Sequential Convolutional Neural Networks for Sleep Stage Classification," *Bioengineering*, vol. 11, no. 12, Des 2024, doi: 10.3390/bioengineering11121226.
[18] A. A. A dan L. R. Nair, "Anovel Study of Silhouette Method to Solve The Issues of Outlier and Improve the Quality of Cluster," *Journal of Data Acquisition and Processing*, vol. 38, 2023.
[19] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, dan O. Tabona, "A survey on missing data in machine learning," *J Big Data*, vol. 8, no. 1, Des 2021, doi: 10.1186/s40537-021-00516-9.
[20] I. T. Jollife dan J. Cadima, "Principal component analysis: A review and recent developments," 13 April 2016, *Royal Society of London*. doi: 10.1098/rsta.2015.0202.
[21] S. Anastassia Amellia Kharis dan A. Haqqi Anna Zili, "Learning Analytics dan Educational Data Mining pada Data Pendidikan," *Jurnal Riset Pembelajaran Matematika Sekolah*, vol. 6, 2022.
[22] H. Chong *dkk.*, "High-resolution mapping of SO2 using airborne observations from the GeoTASO instrument during the KORUS-AQ field study: PCA-based vertical column retrievals," *Remote Sens Environ*, vol. 241, Mei 2020.
[23] T. Soni Madhulatha, "Comparison Between K-means and K-medoids Clustering Algorithms," *International Journal of Advanced Computing (IJAC)*, 2011, Diakses: 10 April 2025. [Daring].
[24] Y. Darmi, A. Setiawan, J. Bali, K. Kampung Bali, K. Teluk Segara, dan K. Bengkulu, "Penerapan Metode Clustering K-Means Dalam Pengelompokan Penjualan Produk," 2016.
[25] M. Benri, H. Metisen, dan S. Latipa, "Analisis Clustering Menggunakan Metode K-Means Dalam Pengelompokkan Penjualan Produk Pada Swalayan Fadhila," 2015.
[26] S. Agustina, D. Yhudo, H. Santoso, N. Marnasusanto, A. Tirtana, dan F. Khusnu, "Clustering Kualitas Beras Berdasarkan Ciri Fisik Menggunakan Metode K-Means," 2012. Diakses: 10 April 2025. [Daring].
[27] T. Avini, Mk. Zumhur Alamin, M. Kom Giandari Maulani, dan Mk. Eza Budi Perkasa, *Fundamental Algorithma*. PT Sada Kurnia Pustaka, 2024.
[28] R. A. Farissa, R. Mayasari, dan Y. Umaidah, "Perbandingan Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Data Obat dengan Silhouette Coefficient," 2021. [Daring]. Tersedia pada: http://jurnal.polibatam.ac.id/index.php/JAIC
[29] D. Marlina, N. Fauzer Putri, A. Fernando, dan A. Ramadhan, "Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak," *Jurnal CoreIT*, vol. 4, no. 2, 2018.
[30] V. Benny Alexsius Pardosi, *Kecerdasan Komputasional*. 2022. Diakses: 11 April 2025. [Daring]. Tersedia pada: https://www.researchgate.net/profile/Ilham-Ilham-17/publication/385125274_Kecerdasan_Komputasional/Links/6717570268ac304149aa4c56/Kecerdasan-Komputasional.pdf
[31] Y. Hasan, "Pengukuran Silhouette Score dan Davies-Bouldin Index pada Hasil Cluster K-Means dan Dbscan," 2024.