J-COSINE (Journal of Computer Science and Informatics Engineering)
Vol. 7, No. 1, June 2023 E-ISSN:2541-0806
Accredited Sinta-3 by RISTEKDIKTI Decree No. 28/E/KPT/2019 P-ISSN:2540-8895

Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan Menggunakan Metode K-Means Clustering

Grouping Provinces in Indonesia Based on Education Indicators Using the K-Means Clustering Method

Mindi Richia Putri*, Gibran Satya Nugraha, Ramaditia Dwiyansaputra
Dept Informatics Engineering, University of Mataram
Jl. Majapahit 62, Mataram, Lombok NTB, INDONESIA
Email: mindirichiaputri@gmail.com, gibransn@unram.ac.id, rama@unram.ac.id

Penulis Korespondensi

Abstract **The education level of the Indonesian people has improved but has not yet reached the entire population. The gap in educational attainment between socioeconomic classes is still an issue and gets worse as education levels rise. When comparing areas, the education difference is still substantial. Additionally, there hasn't been an optimal flow of quality education across areas. Accurate and complete information is needed as a reference in planning and determining the right strategy in facing development challenges in the education sector. This information is expected to explain the current condition and situation of education development in Indonesia. This study aims to group provinces in Indonesia based on educational indicators using the K-Means method. The data and parameters used are based on a portrait of education statistics in Indonesia in 2020. This study shows that clustering produces the best cluster quality at K=3 with Silhouette Coefficient (SC) value = 0.6308 based on parameters that have been previously selected.**

Key words**: Educational Indicators; Provinces; Grouping; K-Means; Silhouette Coefficient**

I. PENDAHULUAN

Pendidikan memiliki peranan penting dalam meningkatkan kualitas SDM (Sumber Daya Manusia). Pendidikan menjadi salah satu bagian dari arah pembangunan SDM yang bertujuan untuk menciptakan SDM yang produktif, dinamis, terampil, dan memiliki pengetahuan dan kemampuan di bidang ilmu pengetahuan dan teknologi yang didukung dengan kerjasama talenta global maupun industri. Hal tersebut merupakan salah satu bagian agenda pembangunan nasional tahun 2020-2024 yakni meningkatkan SDM yang berdaya saing dan berkualitas. Peningkatan daya saing serta kualitas SDM diupayakan dapat membentuk generasi yang cerdas, sehat, terampil, adaptif, inovatif, serta berkarakter [1].

Menurut Pasal 3 Undang-Undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional, terdapat pernyataan bahwa tujuan pendidikan nasional adalah untuk mengembangkan kemampuan individu dan membentuk karakter serta peradaban bangsa yang memiliki martabat. Hal ini dilakukan dengan maksud untuk meningkatkan kecerdasan kehidupan bangsa. Pendidikan tersebut bertujuan agar peserta didik dapat mengembangkan potensi mereka sehingga menjadi individu yang beriman dan taqwa kepada Tuhan Yang Maha Esa, memiliki akhlak yang mulia, sehat, berpengetahuan, kompeten, kreatif, mandiri, serta menjadi warga negara yang demokratis dan bertanggung jawab [2].

Namun, pendidikan di Indonesia saat ini masih belum merata dan terdapat perbedaan kualitas pendidikan yang signifikan antara daerah satu dengan yang lain. Faktanya, beberapa daerah di Indonesia masih menghadapi tantangan dalam mencapai standar pendidikan yang memadai. Permasalahan dalam sistem pendidikan Indonesia mencakup kesenjangan dalam sarana dan prasarana pendidikan antara daerah perkotaan dan pedesaan, kekurangan dalam manajemen pendidikan, persepsi masyarakat yang menganggap rendah pentingnya pendidikan, kekurangan tenaga pengajar berkualitas, biaya pendidikan yang tinggi, dan sebagainya. Beberapa hal tersebut menjadi faktor kualitas pendidikan di Indonesia tergolong rendah [3]. Untuk mengetahui dan mengevaluasi kualitas pendidikan, diperlukan indikator pendidikan yang dapat menggambarkan kualitas pendidikan di Indonesia.

Berdasarkan data Potret Statistik Pendidikan Indonesia tahun 2020 yang diperoleh dari *website* Badan Pusat Statistik (BPS) didapatkan indikator pendidikan yang dapat menggambarkan kondisi, situasi serta capaian pembangunan dalam bidang pendidikan pada setiap provinsi di Indonesia, yaitu meliputi Jumlah Sekolah, Jumlah Peserta Didik, Jumlah Rombel, Jumlah Ruang Kelas, Jumlah Perpustakaan, Jumlah Guru, APK (Angka Partisipasi Kasar), APM (Angka Partisipasi Murni), APS (Angka Partisipasi Sekolah), AKS (Angka Kesiapan Sekolah), AMH (Angka Melek Huruf), Angka Mengulang, Angka Bertahan, Angka Melanjutkan, Angka Putus Sekolah, Angka Tidak Bersekolah, dan Tingkat

Penyelesaian Sekolah [1]. Beberapa indikator tersebut memiliki kategori didalamnya seperti Peserta Didik (SD, SMP, SMA, SMK) yang berada di dalam indikator Jumlah Sekolah, Jumlah Peserta Didik, Jumlah Guru, Jumlah Perpustakaan dan lain-lain. Banyaknya jumlah indikator pedidikan dan kategorinya membuat pengukuran kualitas pendidikan di Indonesia menjadi sulit dilakukan. Oleh sebab itu, dengan provinsi di Indonesia sebagai objek dan indikator pendidikan sebagai parameter, maka akan dilakukan pengelompokan provinsi di Indonesia berdasarkan indikator pendidikan.

Pada penelitian ini data dianalisis dengan metode analisis *cluster* yaitu K-Means Clustering yang mana termasuk dalam metode non hirarki. K-Means menggunakan pendekatan yang berbeda dengan metode lain seperti Fuzzy C-Means, Agglomerative Hierarchical Clustering, dan K-Medoids. K-Means hanya memungkinkan suatu data menjadi bagian dari satu *cluster*, sedangkan metode lain seperti Fuzzy C-Means memungkinkan suatu data tidak hanya menjadi bagian dari satu *cluster* saja [4]. Dengan adanya penelitian ini, dapat diketahui pola kesamaan data indikator pendidikan dari provinsi-provinsi yang berada dalam *cluster* yang sama. Selain itu, diharapkan dapat menjadi salah satu referensi serta evaluasi untuk pembangunan dan pengembangan pada bidang pendidikan.

II. TINJAUAN PUSTAKA

A. Penelitian Terkait

Beberapa penelitian terdahulu telah melakukan pengelompokan menggunakan K-Means. Penelitian [5] membahas mengenai perbandingan hasil klasterisasi dengan K-Means dan Fuzzy C-Means. Penelitian ini menunjukkan K-Means lebih baik dalam mengelompokkan daerah penyebaran Covid-19 di Indonesia. Penelitian [6] membahas mengenai pengelompokkan daerah rawan bencana kebakaran menggunakan K-Means dan mendapatkan hasil kualitas *cluster* yang sangat baik (*strong stucture*). Penelitian [7] membahas mengenai pengelompokan data kemiskinan provinsi Jawa Barat dengan algoritma K-Means, yang mana menghasilkan nilai Silhouette Coefficient sebesar 0.576.

Penelitian lainnya yang membandingkan metode K-Means adalah penelitian [8] dan [9]. Penelitian [8] membahas mengenai perbandingan algoritma *clustering* untuk pengelompokan data persediaan produk. Penelitian ini menunjukkan bahwa algoritma *clustering* yang terbaik adalah K-Means, yang mana menghasilkan Silhouette Coefficient senilai 0.52. Penelitian [9] membahas mengenai perbandingan metode Partitioning dan Hierarchical Clustering dalam pengelompokan berdasarkan IPM (Indeks Pembangunan Manusia) Indonesia pada Tahun 2019. Penelitian ini menunjukkan bahwa K-Means merupakan metode terbaik dengan Silhouette Coefficient senilai 0.6291.

Berdasarkan penelitian-penelitian tersebut terlihat bahwa K-Means mampu mengelompokkan data dengan dimensi yang besar. Pengujian kualitas *cluster* dapat dilakukan dengan menggunakan Silhouette Coefficient. Dengan demikian, penulis akan melakukan pengelompokan provinsi di Indonesia berdasarkan indikator pendidikan menggunakan metode K-Means Clustering.

B. Teori Penunjang

B.1 Pendidikan

Pendidikan merupakan suatu proses yang bertujuan untuk memaksimalkan kemampuan individu. Asal usul kata "pendidikan" dapat ditelusuri ke kata "padegogik" dalam bahasa Yunani yang berarti ilmu mengarahkan anak-anak. Menurut definisi dari Kamus Besar Bahasa Indonesia (KBBI), pendidikan didefinisikan sebagai kegiatan mendidik, yaitu upaya untuk memelihara dan memberikan pengajaran mengenai akhlak dan kecerdasan pikiran. Maka, dapat disimpulkan bahwa pendidikan merupakan upaya untuk membimbing manusia melalui pemberian ajaran tentang moral dan kecerdasan pikiran, dengan tujuan untuk mengembangkan dan mengoptimalkan potensi yang dimilikinya [10].

B.2 Indikator Pendidikan

Berikut adalah beberapa indikator atau parameter yang dapat menggambarkan kualitas pendidikan di Indonesia [1]. Seluruh parameter ini akan dinilai apakah termasuk kategori *cost* atau *benefit*. Parameter yang termasuk kategori *cost* apabila semakin kecil nilainya maka semakin bagus, contohnya Angka Mengulang yang dimana semakin kecil nilai Angka Mengulang di suatu provinsi menunjukkan kualitas pendidikan di daerah tersebut cukup baik, sedangkan parameter yang termasuk kategori *benefit* apabila semakin besar nilainya maka semakin bagus, contohnya Angka Bertahan di suatu provinsi yang dimana apabila Angka Bertahan semakin tinggi kualitas pendidikan di daerah tersebut cukup bagus. Detail parameter dengan kategori *cost* atau *benefit* ditunjukkan pada Tabel I.

TABEL I. KATEGORI PARAMETER

No	Parameter	Parameter (*cost*/*benefit*)
1	Jumlah Sekolah	*benefit*
2	Jumlah Peserta Didik	*benefit*
3	Jumlah Rombongan Belajar	*benefit*
4	Jumlah Ruang Kelas	*benefit*
5	Jumlah Perpustakaan	*benefit*
6	Jumlah Guru	*benefit*
7	APK (Angka Partisipasi Kasar)	*benefit*
8	APM (Angka Partisipasi Murni)	*benefit*
9	APS (Angka Partisipasi Sekolah)	*benefit*
10	AKS (Angka Kesiapan Sekolah)	*benefit*
11	AMH (Angka Melek Huruf)	*benefit*
12	Angka Mengulang	*cost*
13	Angka Bertahan	*benefit*
14	Angka Melanjutkan	*benefit*
15	Angka Putus Sekolah	*cost*
16	Angka Tidak Bersekolah	*cost*
17	Tingkat Penyelesaian Sekolah	*benefit*

B.3 Clustering

Clustering merupakan salah satu metode *unsupervised learning*, yang mana *dataset* akan dipartisi menjadi kelompok atau *cluster* yang berbeda berdasarkan ukuran kesamaan tertentu [11]. Metode ini akan mengelompokkan objek-objek ke dalam cluster berdasarkan karakteristik yang memiliki tingkat kemiripan yang signifikan jika berada dalam satu kluster, dan memiliki perbedaan yang cukup besar jika objek berada dalam kluster yang berbeda [12].

Pengukuran kesamaan yang digunakan dalam penelitian ini adalah Euclidean Distance. Euclidean Distance merupakan salah satu *distance metric* yang sering digunakan karena dapat mendukung hasil perhitungan *cluster* dan sangat mudah dipahami. Euclidean Distance menghasilkan perhitungan jarak terkecil antara dua titik yang diperhitungkan. Proses perhitungan dilakukan selama nilai pengelompokan yang konvergen belum ditemukan. Dalam beberapa kasus, perhitungan tersebut membutuhkan banyak iterasi, mulai dari puluhan hingga ratusan iterasi, sebelum nilai konvergen ditemukan. Adapun Euclidean Distance dihitung menggunakan Persamaan (1)[13].

$$D_{(a,b)} = \sqrt{(x_{1a} - x_{1b})^2 + \cdots + (x_{ka} - x_{kb})^2} \quad (1)$$

Dimana:

$D_{(a,b)}$: Jarak data ke a ke *centroid* b
x_{ka} : Data ke a pada parameter data ke k
x_{kb} : *Centroid* ke b pada parameter ke k

B.4 K-Means

K-Means adalah metode pengelompokan yang mengelompokkan data berdasarkan kedekatan data dengan *centroid* (titip pusat kluster). Pengelompokan dengan K-Means bertujuan untuk meminimalkan kemiripan karakteristik data dalam *cluster* yang berbeda dan memaksimalkan kemiripan karakteristik data dalam *cluster* yang sama. Ukuran kemiripan karakteristik yang digunakan ialah *distance metric*. Dengan demikian, pemaksimalan kemiripan karakteristik data dapat diperoleh berdasarkan jarak terkecil antara data dengan titik pusat *cluster* [14]. Adapun tahapan dalam metode K-Means adalah sebagai berikut [15]:

a. Menentukan jumlah *cluster*
b. Menentukan *centroid cluster k* secara *random*.
c. Menghitung jarak masing-masing data ke *centroid cluster* menggunakan Euclidean Distance, jarak terkecil antara data dengan titik pusat *cluster* menentukan data keanggotaan data tersebut.
d. Menghitung *centroid* baru dengan keanggotaan *cluster* yang didapatkan sebelumnya, dengan mencari *mean* (rata-rata) dari semua data dalam *cluster* menggunakan dengan Persamaan (2).

$$c = \frac{\sum_{i=1}^{n} x_i}{n} \quad (2)$$

Dimana:

x_i : Data ke i
n : *Jumlah* data yang menjadi anggota *cluster*

e. Mengulang dari langkah ke-3 hingga nilai titik pusat *cluster* tidak berubah.

B.5 Normaalisasi

Normalisasi bertujuan untuk mengubah nilai numerik dalam *dataset* dengan tujuan membentuk data dengan *range* nilai yang sama [16]. Dalam proses pengelompokan, normalisasi diperlukan untuk yang terkait dengan metrik jarak, seperti *Euclidean Distance*, karena metrik tersebut sangat sensitif terhadap perbedaan skala atau ukuran pada parameter Perbedaan rentang nilai dari suatu parameter menyebabkan salah satu parameter akan mendominasi parameter lainnya. Normalisasi mencegah hal tersebut dengan menyamakan skala data antara satu parameter dengan parameter lainnya tanpa mengubah makna perbedaan dalam *range* nilai. Berikut adalah beberapa metode untuk melakukan normalisasi data [12]:

1. Min-Max Normalization

   Min-Max Normalization melakukan normalisasi dengan mengubah nilai data asli menjadi nilai dengan *range* 0 hingga 1. Min-Max Normalization untuk parameter *benefit* terdapat pada Persamaan (3). Sedangkan, untuk parameter *cost* terdapat pada persamaan (4).

$$r_{ij} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}} \quad (3)$$

$$r_{ij} = \frac{x_j^{max} - x_{ij}}{x_j^{max} - x_j^{min}} \quad (4)$$

   Dimana *r*, *x*, *i*, *j*, *min* dan *max* secara berurutan ialah hasil normalisasi, nilai awal, baris, kolom, nilai paling kecil dan nilai paling besar.

2. Z-Score Normalization

   Z-Score Normalization melakukan normalisasi pada sebuah parameter A berdasarkan *mean* dan standar deviasi dari parameter tersebut. v melambangkan nilai A yang akan dinormalisasikan pada v' menggunakan rumus pada Persamaan (5).

$$v' = \frac{v - \overline{A}}{\sigma_A} \quad (5)$$

   Dimana $\overline{A}$ dan σ_A secara berurutan ialah *mean* dan standar deviasi dari parameter A. Metode ini akan sangat tepat apabila nilai minimum maupun maksimum tidak diketahui secara aktual.

3. Normalization by Decimal Scaling

   Normalisasi ini melakukan normalisasi dengan mengubah titik desimal dari nilai parameter A. Nilai di titik desimal bergantung dari nilai absolut

maksimum dari A. v melambangkan A akan dinormalisasikan terhadap v' menggunakan rumus pada Persamaan (6).

$$v' = \frac{v}{10^j} \quad (6)$$

Dimana j ialah nilai integer terkecil yang menyebabkan Max($|v'|$) < 1.

B.6 Principle Component Analysis (PCA)

PCA adalah suatu metode yang digunakan untuk mengurangi dimensi dalam data dengan mengubah himpunan dimensi yang saling berkorelasi menjadi tidak berkorelasi. Tujuan dari metode ini adalah untuk menghasilkan nilai komponen utama (PC) yang merupakan hasil dari penggabungan linear nilai-nilai asli sebelum dilakukan pengurangan dimensi.

Langkah pertama dalam metode PCA adalah menghitung data $X^*_{i,j}$ dengan dimensi $m\,x\,n$, dimana m ialah jumlah sampel dan n ialah jumlah atributnya. Data $X^*_{i,j}$ didapatkan menggunakan rumus pada Persamaan 7.

$$X^*_{i,j} = X_{i,j} - \bar{X} \quad (7)$$

Selanjutnya, menghitung nilai kovarian dari matriks $X_{i,j}$, C_x menunjukkan nilai dari kovarian yang dihitung. C_x ialah matriks kovariansi dari $j\,x\,j$, m ialah jumlah dari sampel. Rumus untuk mencari C_x menggunakan Persamaan (8).

$$C_x = \frac{1}{m-1} . X^{*T}_{i,j} . X^*_{i,j} \quad (8)$$

Kemudian, mencari nilai eigen menggunakan Persamaan (9).

$$|C_x - \lambda I| = 0 \text{ dan} \quad (C_x - \lambda I).v = 0 \quad (9)$$

Dimana I menunjukkan matriks identitas, λ menunjukkan nilai eigen dan v menunjukkan vektor eigen. Variabel baru ditentukan oleh Vektor eigen. Jumlah variabel baru bergantung pada persentasi kontribusi kumulatif variansi (V_r), yang dihitung menggunakan Persamaan (10).

$$V_r = \frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{D} \lambda_j} . 100\% \quad (10)$$

Dimana D menunjukkan jumlah atribut awal, dan r menunjukkan jumlah komponen yang dipilih [17].

B.7 Silhouette Coefficient

Silhouette Coefficient (SC) merupakan salah satu metode untuk evaluasi *cluster*. SC digunakan untuk menguji kualitas suatu *cluster* dan menggambarkan derajat kepemilikan setiap objek pada suatu *cluster*. SC terbentuk dari gabungan metode kohesi dan separasi. Kohesi digunakan untuk mengetahui keterkaitan antar objek dalam *cluster*. Sedangkan Separasi digunakan untuk mendapatkan jarak objek dengan objek lainnya yang berada pada *cluster* yang berbeda. Berikut tahapan dalam menghitung nilai SC:

a. Menghitung nilai *mean* jarak data i dengan semua data dengan *cluster* yang sama menggunakan Persamaan (11).

$$a_i = \frac{1}{[A]-1} \sum j \quad \in_{A,j} \neq d(i,j) \quad (11)$$

b. Menghitung *mean* data i dengan seluruh data yang berada pada *cluster* yang lain, kemudian mengambil nilai rata-rata terkecil yang diinisialisasi sebagai b_i. Adapun nilai b_i diperoleh menggunakan Persamaan (12).

$$b_i = \min\,(D(i,C)) \quad (12)$$

c. Menghitung nilai SC menggunakan Persamaan (13):

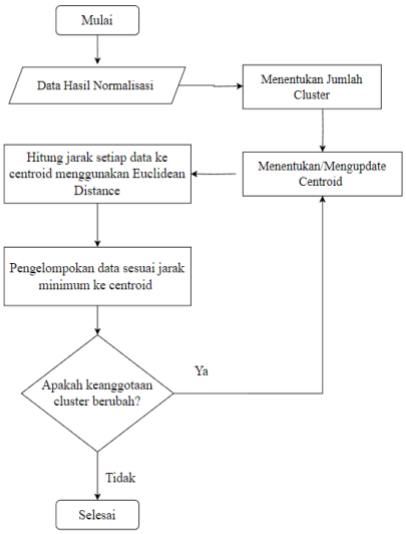
$$S_i = \frac{b_i - a_i}{\max\,(a_i, b_i)} \quad (13)$$

Hasil nilai SC yang diperoleh berada pada *range* -1 hingga 1. Nilai SC bernilai positif dimana (a_i)<(b_i) dan a_i mendekati nilai 0 menunjukkan suatu *clustering* dapat dikatakan baik. SC bernilai 1 ketika $a_i = 0$, yang mana menunjukkan bahwa *cluster* 1 berada pada *cluster* yang sesuai. SC bernilai 0 menunjukkan objek i berada pada *cluster* dengan struktur yang tidak jelas. Lalu, SC bernilai -1 menunjukkan objek i lebih tepat apabila dimasukkan pada *cluster* lain. Berikut nilai SC menurut Rousseau (1987) [6] :

a. 0.7<SC≤1, menunjukkan *cluster* memiliki struktur yang kuat.
b. 0.5< SC ≤0.7, menunjukkan *cluster* memiliki struktur standar.
c. 0.25<SC≤0.5, menunjukkan *cluster* memiliki struktur yang lemah.
d. SC≤0.25, menunjukkan *cluster* tidak memiliki struktur.

III. METODOLOGI

Dalam penelitian ini, metode yang digunakan adalah K-Means Clustering. Gambar 1 menunjukkan langkah-langkah dalam proses pengelompokan menggunakan K-Means Clustering. Langkah pertama adalah menentukan jumlah kluster yang diinginkan dari data yang telah dinormalisasi. Selajutnya, dilakukan proses penentuan nilai centroid yang akan digunakan. Nilai awal *centroid* ditentukan secara acak (*random*). Setelah nilai *centroid* ditentukan, langkah tahapan berikutnya adalah menghitung jarak masing-masing data ke *centroid* dengan menggunakan metode *euclidian distance*. Pengelompokan data dilakukan dengan mengelompokkan data dengan *centroid* terdekatnya. Selanjutnya, nilai *centroid* akan diperbaharui secara terus menerus sesuai dengan pengelompokan data yang terbentuk hingga keanggotaan kluster tidak berubah.



Gambar 1. Algoritma K-Means Clustering

IV. Hasil dan Pembahasan

A. Akuisis Data

Data bersumber dari data potret statistik pendidikan Indonesia tahun 2020 yang bersumber dari *website* https://www.bps.go.id/publication/2020/11/27/347c85541c34e7dae54395a3/statistik-pendidikan-2020.html dan https://statistik.data.kemdikbud.go.id. *Dataset* tersebut terdiri dari data indikator pendidikan di 34 provinsi di Indonesia. Pengelompokan dilakukan menggunakan 62 parameter. Setelah melakukan akuisisi data, selanjutnya *dataset* dinormalisasi menggunakan Min-Max Normalization. Rincian parameter yang digunakan ditunjukkan pada Tabel II.

TABEL II. Rincian Parameter

No	Parameter (Jumlah Sub Kategori)	Sub Kategori
1	Jumlah Sekolah (4)	SD
		SMP
		SMA
		SMK
2	Jumlah Peserta Didik (4)	SD
		SMP
		SMA
		SMK
3	Jumlah Rombongan Belajar (4)	SD
		SMP
		SMA
		SMK
4	Jumlah Ruang Kelas (4)	SD
		SMP
		SMA
		SMK
5	Jumlah Perpustakaan (4)	SD
		SMP
		SMA
		SMK
6	Jumlah Guru (4)	SD
		SMP
		SMA
		SMK
7	Angka Partisipasi Kasar (7)	PAUD Anak Usia 3-5 Tahun
		PAUD Anak Usia 3-5 Tahun
		SD
		SMP
		SM
		PT 19-24
		PT 19-23
8	Angka Partisipasi Murni (7)	PAUD Anak Usia 3-5 Tahun
		PAUD Anak Usia 3-5 Tahun
		SD
		SMP
		SM
		PT 19-24
		PT 19-23
9	Angka Partisipasi Sekolah (5)	Umur 7-12
		Umur 13-15
		Umur 16-18
		Umur 19-24
		Umur 19-23
10	Angka Kesiapan Sekolah (1)	Total
11	Angka Melek Huruf (3)	15-24 Tahun
		15-59 Tahun
		15 Tahun ke Atas
12	Angka Mengulang (3)	SD/sederajat
		SMP/sederajat
		SM/sederajat
13	Angka Bertahan (1)	Perkotaan+Pedesaan
14	Angka Melanjutkan (2)	Melanjutkan ke SMP/sederajat
		Melanjutkan ke SM/sederajar
15	Angka Putus Sekolah (3)	SD/sederajat
		SMP/sederajat
		SM/sederajat
16	Angka Tidak Bersekolah (3)	7-12 Tahun
		3-15 Tahun
		16-18 Tahun
17	Tingkat Penyelesaian Sekolah (3)	SD/sederajat
		SMP/sederajat
		SM/sederajat
Total Parameter = 62 Parameter		

B. Pengelompokan dengan K-Means

Pada penelitian ini pengelompokan dilakukan menggunakan 62 parameter dengan nilai K yang dicoba adalah K = 2 hingga K = 7. *Distance metric* yang digunakan pada adalah Euclidean. Hasil pengelompokan ditunjukkan pada Tabel III.

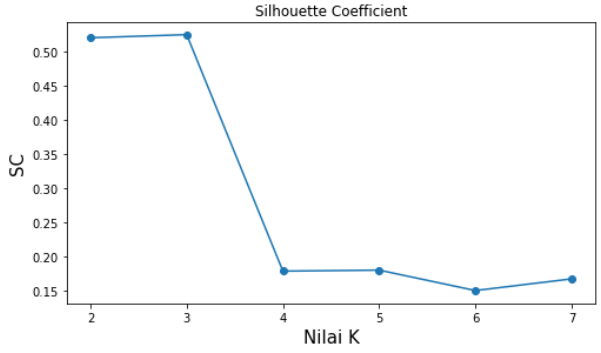
TABEL III. Hasil Pengelompokan

Provinsi	K=2	K=3	K=4	K=5	K=6	K=7
Aceh	0	0	0	0	0	6
Sumatera Utara	0	0	1	2	1	3
Sumatera Barat	0	0	0	0	0	6
Riau	0	0	1	2	2	3
Jambi	0	0	0	0	1	0
Sumatera Selatan	0	0	1	2	1	3
Bengkulu	0	0	0	0	0	6
Lampung	0	0	1	2	2	3
Kep. Bangka Belitung	0	0	1	2	2	2

Provinsi	K=2	K=3	K=4	K=5	K=6	K=7
Kep. Riau	0	0	0	0	0	6
DKI Jakarta	0	0	1	2	2	2
Jawa Barat	1	1	2	1	3	1
Jawa Tengah	1	1	2	1	3	1
DI Yogyakarta	0	0	0	3	5	4
Jawa Timur	1	1	2	1	3	1
Banten	0	0	1	2	1	0
Bali	0	0	0	0	0	6
Nusa Tenggara Barat	0	0	0	0	2	6
Nusa Tenggara Timur	0	0	1	2	1	2
Kalimantan Barat	0	0	1	2	1	2
Kalimantan Tengah	0	0	1	2	2	2
Kalimantan Selatan	0	0	1	2	2	2
Kalimantan Timur	0	0	0	0	0	6
Kalimantan Utara	0	0	0	0	2	6
Sulawesi Utara	0	0	0	0	2	6
Sulawesi Tengah	0	0	0	0	2	2
Sulawesi Selatan	0	0	1	2	1	2
Sulawesi Tenggara	0	0	1	2	2	2
Gorontalo	0	0	1	2	2	2
Sulawesi Barat	0	0	1	0	2	2
Maluku	0	0	0	0	0	6
Maluku Utara	0	0	0	0	0	6
Papua Barat	0	0	0	0	0	6
Papua	0	2	3	4	4	5

Ket: 0 = *cluster* ke-1, 1 = *cluster* ke-2, 2 = *cluster* ke-3, 3 = *cluster* ke-4, 4 = *cluster* ke-5, 5 = *cluster* ke-6, 6 = *cluster* ke-7

C. Evaluasi dan Analisis



Gambar 2. Evaluasi menggunakan Silhouette Coefficient

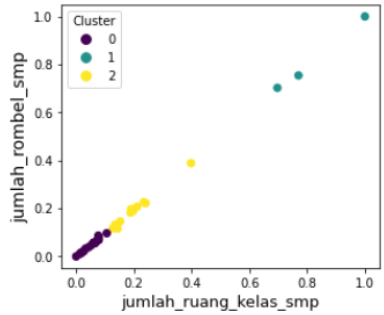
Gambar 2 menunjukkan hasil evaluasi pengelompokan menggunakan Silhouette Coefficient. Silhouette Coefficient menunjukkan nilai K optimal berada pada K=3 dengan nilai SC=0.524016 (struktur standar).

D. Pengujian

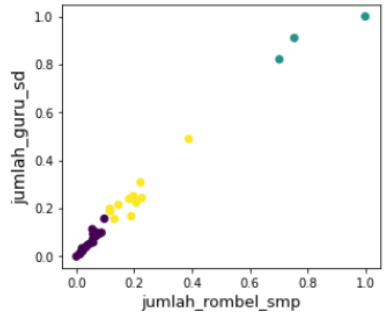
Pengujian dilakukan untuk mengetahui pengaruh seleksi fitur/parameter menggunakan PCA terhadap kualitas *cluster* yang terbentuk. Pemilihan parameter dilakukan dengan bantuan *software* Weka dengan variasi koefisien parameter 0.207, ≥0.206, ≥0.205, ≥0.204 dan ≥0.203 dengan proporsi 0.6318. Berdasarkan hasil Silhouette Coefficient yang dihasilkan sebelumnya, diketahui bahwa nilai K optimal adalah 3. Sehingga, pengujian ini dilakukan dengan nilai K=3. Hasil pengujian ditunjukkan pada Tabel IV.

TABEL IV. HASIL PENGUJIAN

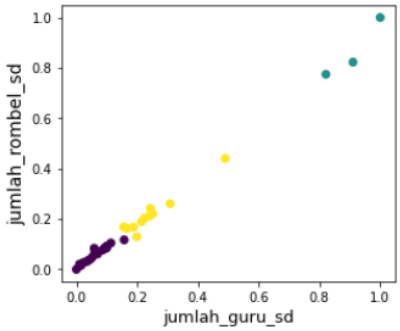
No	Parameter Terpilih	Koefisien Parameter	Silhouette Coefficient
1	Jumlah Ruang Kelas SMP, Jumlah Rombel SMP, Jumlah Guru SD	0.207	0.5928
2	Jumlah Ruang Kelas SMP, Jumlah Rombel SMP, Jumlah Guru SD, Jumlah Rombel SD, Jumlah Guru SMP, Jumlah Peserta Didik SMP	≥0.206	0.6308
3	Jumlah Ruang Kelas SMP, Jumlah Rombel SMP, Jumlah Guru SD, Jumlah Rombel SD, Jumlah Guru SMP, Jumlah Peserta Didik SMP, Jumlah Ruang Kelas SD, Jumlah Perpustakaan SMK	≥0.205	0.6161
4	Jumlah Ruang Kelas SMP, Jumlah Rombel SMP, Jumlah Guru SD, Jumlah Rombel SD, Jumlah Guru SMP, Jumlah Peserta Didik SMP, Jumlah Ruang Kelas SD, Jumlah Perpustakaan SMK, Jumlah Peserta Didik SMA, Jumlah Peserta Didik SD, Jumlah Rombel SMK, Jumlah Ruang Kelas SMK, Jumlah Ruang Kelas SMA, Jumlah Sekolah SMK, Jumlah Sekolah SMP, Jumlah Peserta Didik SMK	≥0.204	0.5900
5	Jumlah Ruang Kelas SMP, Jumlah Rombel SMP, Jumlah Guru SD, Jumlah Rombel SD, Jumlah Guru SMP, Jumlah Peserta Didik SMP, Jumlah Ruang Kelas SD, Jumlah Perpustakaan SMK, Jumlah Peserta Didik SMA, Jumlah Peserta Didik SD, Jumlah Rombel SMK, Jumlah Ruang Kelas SMK, Jumlah Ruang Kelas SMA, Jumlah Sekolah SMK, Jumlah Sekolah SMP, Jumlah Peserta Didik SMK, Jumlah Rombel SMA, Jumlah Sekolah SD	≥0.203	0.5928



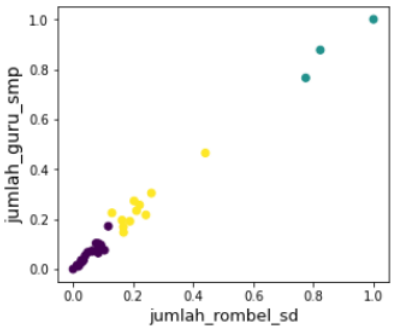
Gambar 3. Visualisasi hasil pengelompokan dengan seleksi fitur untuk parameter Jumlah Ruang Kelas SMP dan Jumlah Rombel SMP



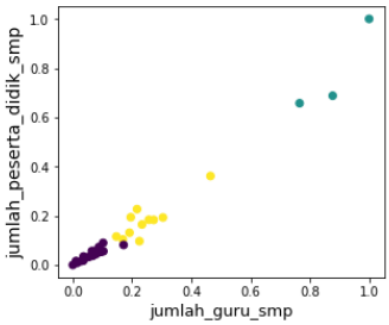
Gambar 4. Visualisasi hasil pengelompokan dengan seleksi fitur untuk parameter Jumlah Rombel SMP dan Jumlah Guru SD



Gambar 5. Visualisasi hasil pengelompokan dengan seleksi fitur untuk parameter Jumlah Guru SD dan Jumlah Rombel SD



Gambar 6. Visualisasi hasil pengelompokan dengan seleksi fitur untuk parameter Jumlah Rombel SD dan Jumlah Guru SMP



Gambar 7. Visualisasi hasil pengelompokan dengan seleksi fitur untuk parameter Jumlah Guru SMP dan Jumlah Peserta Didik SMP

TABEL IV menunjukkan parameter terpilih dengan koefisien ≥0.206 mendapatkan nilai Silhouette Coefficient terbesar yaitu 0.6308 yang mana memiliki struktur *cluster* standar. Visualisasi hasil pengelompokan menggunakan PCA ditunjukkan pada Gambar 3, Gambar 4, Gambar 5, Gambar 6, dan Gambar 7. Hasil tersebut menunjukkan pengelompokan menggunakan PCA dapat membedakan *cluster* 0, *cluster* 1, dan *cluster* 2 dengan cukup baik dan mendapatkan Silhouette Coefficient senilai 0.6308. Hasil tersebut lebih besar dibandingkan dengan hasil Silhouette Coefficient untuk K=3 dengan menggunakan seluruh parameter. Hal tersebut menunjukkan bahwa penggunaan seleksi fitur pada penelitian ini dapat meningkatkan hasil Silhouette Coefficient dan mendapatkan kualitas *cluster* yang lebih baik. Parameter/fitur yang terpilih merupakan fitur yang merepresentasikan data dengan tegas. Hal tersebut membuat seleksi fitur dengan PCA mampu meningkatkan nilai Silhouette Coefficient.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, metode K-Means Clustering dapat mengelompokkan provinsi di Indonesia berdasarkan indikator pendidikan dengan cukup baik yang diukur dengan metode Silhouette Coefficient dengan struktur *cluster* standar dengan nilai 0.6308 dengan parameter yang sudah di seleksi sebelumnya. Selanjutnya, penentuan jumlah *cluster* terbaik dilakukan dengan melakukan percobaan menggunakan nilai K = 2 hingga K = 7. Berdasarkan evaluasi menggunakan metode Silhouette Coefficient, *cluster* dengan kualitas terbaik terbentuk ketika nilai K=3. Selain itu, penggunaan seleksi fitur/parameter menggunakan metode PCA menghasilkan kualitas *cluster* yang lebih baik dibandingkan tanpa menggunakan seleksi fitur.

B. Saran

Berdasarkan penelitian yang telah dilakukan nilai Silhouette Coefficient yang didapatkan masih tergolong struktur standar sehingga untuk penelitian ataupun pengembangan selanjutnya dapat menggunakan metode *clustering* lain seperti K-Medoid, DBSCAN, Agglomerative Clustering, ataupun yang lainnya tanpa mengurangi nilai kualitas *cluster* yang sudah didapatkan sebelumnya. Selain itu, walaupun sudah menggunakan

PCA sebagai metode seleksi fitur, namun hasil Silhouette Coefficient masih termasuk struktur standar, sehingga perlu untuk dicoba metode seleksi fitur yang lain untuk memperoleh hasil yang lebih akurat dan tepat dalam mempengaruhi kualitas pendidikan.

REFERENCES

[1] R. Agustina, S. W. Nugroho, N. P. Sulistyowati, L. Annisa, and R. Putrianti, Potret Pendidikan Indonesia 2020. Jakarta: Badan Pusat Statistik, 2020.

[2] P. R. Indonesia, "Undang-Undang Republik Indonesia," 2003. https://pusdiklat.perpusnas.go.id

[3] L. E. Wahyudi et al., "Mengukur Kualitas Pendidikan di Indonesia," Ma'arif J. Educ. Madrasah Innov. Aswaja Stud., vol. 1, no. 1, pp. 18–22, 2022, [Online]. Available: https://jurnal.maarifnumalang.id/ (diunduh 10 Februari 2022)

[4] A. Chusyairi, U. B. Insani, P. Ramadar, and N. Saputra, "Perbandingan Algoritma Fuzzy C-Means Dan K- Means Clustering Dalam Pengelompokan Data Puskesmas," Conf. Inf. Technol. Inf. Syst. Electr. Eng., 2019.

[5] A. L. R. Putri and N. Dwidayati, "Analisa Perbandingan K-Means Dan Fuzzy C-Means Dalam Pengelompokan Daerah Penyebaran Covid-19 Indonesia," UNNES J. Math., vol. 10, no. 2, pp. 4–7, 2021, [Online]. Available: http://journal.unnes.ac.id/sju/index.php/ujme

[6] F. N. Dhewayani, D. Amelia, D. N. Alifah, B. N. Sari, and M. Jajuli, "Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM," J. Teknol. dan Inf., vol. 12, no. 1, pp. 64–77, 2022, doi: 10.34010/jati.v12i1.6674.

[7] N. N. Fransiska, D. S. Anggraeni, and U. Enri, "Pengelompokkan Data Kemiskinan Provinsi Jawa Barat Menggunakan Algoritma K-Means dengan Silhouette Coefficient," Temat. (Jurnal Teknol. Inf. Komunikasi), vol. 9, no. 1, pp. 29–35, 2022, [Online]. Available: https://doi.org/10.38204/tematik.v9i1.921

[8] C. Nisa1 and W. Yustanti2, "Studi Perbandingan Algoritma Klastering Dalam Pengelompokan Persediaan Produk (Studi Kasus : Subdirektorat Perencanaan Sarana Prasarana Dan Logistik PTN X)," Jeisbi, vol. 02, no. 03, p. 2021, 2021.

[9] A. M. Sikana and A. W. Wijayanto, "Analisis Perbandingan Pengelompokan Indeks Pembangunan Manusia Indonesia Tahun 2019 dengan Metode Partitioning dan Hierarchical Clustering," J. Ilmu Komput., vol. 14, no. 2, p. 66, 2021, doi: 10.24843/jik.2021.v14.i02.p01.

[10] Nurkholis, "Pendidikan dalam Upaya Memajukan Teknologi," J. Kependidikan, vol. 1, no. 1, pp. 24–44, 2013.

[11] R. A. Haraty, M. Dimishkieh, and M. Masud, "An Enhanced K-means Clustering Algorithm For Pattern Discovery In Healthcare Data," Int. J. Distrib. Sens. Networks, vol. 2015, p. 11, [Online]. Available: https://journals.sagepub.com/doi/10.1155/2015/615740

[12] G. S. Nugraha, Hairani, and R. F. P. Ardi, "Aplikasi Pemetaan Kualitas Pendidikan Di Indonesia Menggunakan Metode K-Means," J. Matrik, vol. 17, no. 2, pp. 13–23, 2018.

[13] D. Jollyta, H. Mawengkang, M. Siddik, and S. Efendi, Teknik Evaluasi Cluster Solusi Menggunakan Python Dan Rapidminer. Pekan Baru: Penerbit Deepublish, 2021.

[14] N. Purba, P. Poningsih, and H. S. Tambunan, "Penerapan Algoritma K-Means Clustering Pada Penyebaran Penyakit Infeksi Saluran Pernapasan Akut (ISPA) di Provinsi Riau," J. Inf. Syst. Res., vol. 2, no. 3, pp. 220–226, 2021, [Online]. Available: http://ejurnal.seminar-id.com/index.php/josh/article/view/736

[15] S. Sukamto, I. D. Id, and T. R. Angraini, "Penentuan Daerah Rawan Titik Api di Provinsi Riau Menggunakan Clustering Algoritma K-Means," JUITA J. Inform., vol. 6, no. 2, p. 137, 2018, doi: 10.30595/juita.v6i2.3172.

[16] Ahmad Harmain, P. Paiman, H. Kurniawan, K. Kusrini, and Dina Maulina, "Normalisasi Data Untuk Efisiensi K-Means Pada Pengelompokan Wilayah Berpotensi Kebakaran Hutan Dan Lahan Berdasarkan Sebaran Titik Panas," Tek. Teknol. Inf. dan Multimed., vol. 2, no. 2, pp. 83–89, 2022, doi: 10.46764/teknimedia.v2i2.49.

[17] R. Pujianto, Adiwijaya, and A. A. Rahmawati, "Analisis Ekstraksi Fitur Principle Component Analysis pada Klasifikasi Microarray Data Menggunakan Classification And Regression Trees," eProceeding Eng., vol. 6, no. 1, pp. 2368–2379, 2019.