

PENGELOMPOKAN PROVINSI DI INDONESIA BERDASARKAN INDIKATOR PENDIDIKAN MENGGUNAKAN METODE K-MEANS DAN HIERARCHICAL CLUSTERING

Nia Sukmawijaya^{1*}, Emeylia Safitri²

^{1,2}Program Studi Statistika, Fakultas Sains dan Teknologi, Universitas Terbuka, Tangerang Selatan, Indonesia

*Penulis korespondensi: niasukmawijaya9@gmail.com

ABSTRAK

Pendidikan merupakan salah satu indikator penting dalam pembangunan sumber daya manusia dan kemajuan suatu daerah. Namun, kualitas pendidikan di Indonesia masih menunjukkan adanya ketimpangan antarprovinsi, baik dalam hal akses ke pendidikan maupun pencapaian berbagai indikator pendidikan. Kondisi itu menunjukkan bahwa perlu adanya pemetaan tentang karakteristik pendidikan di berbagai daerah agar kebijakan pendidikan bisa dibuat dengan lebih tepat. Penelitian ini bertujuan mengelompokkan provinsi-provinsi di Indonesia berdasarkan indikator-indikator pendidikan dengan menggunakan dua metode, yaitu *K-Means* dan *hierarchical clustering*. Data yang digunakan berasal dari publikasi Badan Pusat Statistik (BPS) tahun 2025, yang mencakup berbagai indikator terkait pendidikan, seperti rata-rata lama sekolah, harapan lama sekolah, angka partisipasi dalam pendidikan, tingkat kemampuan membaca dan menulis, serta indikator pendidikan lainnya. Penelitian ini menunjukkan bahwa metode *K-Means* dengan nilai $k=3$ menghasilkan *silhouette score* sebesar 0,2438. Pembagian kluster terdiri dari 24 provinsi dalam kluster tinggi, 12 provinsi dalam kluster sedang, serta 2 provinsi dalam kluster rendah, yaitu Papua Tengah dan Papua Pegunungan. Sementara itu, metode *Hierarchical Clustering Average Linkage* dengan nilai $k=2$ menghasilkan *silhouette score* sebesar 0,7536. Pembagian kluster terdiri dari 36 provinsi yang berada di kluster tinggi, serta 2 provinsi yang berada di kluster rendah, yaitu Papua Tengah dan Papua Pegunungan. Berdasarkan nilai *silhouette score*, *hierarchical clustering* dengan *average linkage* menghasilkan kualitas pengelompokan yang lebih baik dibandingkan *K-Means*. Namun, *K-Means* memberikan segmentasi yang lebih detail melalui pembentukan tiga kategori kluster, sehingga lebih mampu menggambarkan variasi karakteristik pendidikan antarprovinsi. Oleh karena itu, kedua metode memiliki keunggulan masing-masing, baik dari aspek kualitas pengelompokan maupun interpretasi karakteristik kluster.

Kata kunci: pendidikan, *clustering*, *k-means*, *hierarchical*, provinsi, Indonesia.

1 PENDAHULUAN

Menurut Ki Hajar Dewantara, pendidikan adalah proses pembentukan potensi dan karakter manusia agar dapat bertumbuh dengan maksimal (Sugiarta et al., 2019). Oleh sebab itu, kualitas pendidikan sering kali menjadi tolak ukur utama untuk menilai kemajuan sebuah wilayah. Hal tersebut membuat sektor pendidikan terus menjadi perhatian penting dalam pembangunan di Indonesia.

Walaupun pemerintah telah menjalankan berbagai program untuk meningkatkan kualitas pendidikan, kenyataannya kondisi pendidikan di Indonesia masih belum merata di setiap wilayah. Masih terdapat perbedaan kualitas pendidikan antarprovinsi, baik dari segi akses pendidikan maupun capaian indikator pendidikannya (Putri, et al., 2025). Pemerintah melalui Badan Pusat Statistik (BPS) secara rutin mempublikasikan data pendidikan yang berasal dari Survei Sosial Ekonomi Nasional (SUSENAS). Data tersebut mencakup berbagai

indikator pendidikan, pada penelitian ini digunakan beberapa indikator pendidikan meliputi rata-rata lama sekolah, harapan lama sekolah, angka partisipasi sekolah tiga jenjang, angka partisipasi kasar perguruan tinggi, angka partisipasi murni tiga jenjang, angka kesiapan sekolah, angka anak tidak sekolah umur 13-15 tahun, angka melek huruf, rasio guru dan murid SD, dan tingkat penyelesaian pendidikan untuk tiga jenjang. Karena setiap provinsi memiliki karakteristik pendidikan yang berbeda, maka diperlukan analisis untuk mengetahui pola kesamaan antarprovinsi berdasarkan tingkat pendidikannya.

Salah satu metode yang bisa digunakan untuk melihat kesamaan karakteristik antarwilayah adalah analisis *clustering*. *Clustering* adalah metode yang digunakan untuk mengelompokkan data berdasarkan tingkat kesamaan karakteristik yang dimiliki (Ferreira & de Souza, 2017). Dengan cara ini, data yang memiliki sifat atau ciri hampir sama akan dikelompokkan ke dalam satu kelompok yang sama. Dalam penelitian ini, metode *K-Means* dan *hierarchical clustering* digunakan karena kedua metode tersebut memiliki cara kerja yang berbeda dalam membagi data ke dalam kelompok-kelompok. Metode *K-Means* bekerja dengan membagi data ke dalam beberapa klaster, dengan cara mengukur jarak setiap data terhadap pusat klaster dan mengelompokkan data ke klaster yang paling dekat. Sementara itu, metode *hierarchical clustering* melakukan pengelompokan data secara bertahap hingga membentuk struktur hierarki yang biasanya ditampilkan dalam bentuk dendrogram (Siregar, et al., 2023). Perbedaan proses pada kedua metode tersebut memungkinkan munculnya hasil pengelompokan yang berbeda sehingga menarik untuk dibandingkan.

Penelitian sebelumnya oleh Hanifah dan Primandari (2023) menggunakan metode *K-Means* untuk mengelompokkan kabupaten/kota berdasarkan indikator pendidikan dan menunjukkan adanya perbedaan kualitas pendidikan antarwilayah. Penelitian lain dilakukan oleh Syafiq et al. (2026) membandingkan metode *K-Means* dan *Hierarchical Clustering* pada pengelompokan kabupaten/kota di Jawa Tengah berdasarkan indikator pendidikan. Hasilnya menunjukkan bahwa *Hierarchical Clustering* memberikan kualitas cluster yang lebih baik. Selain itu, penerapan metode *K-Means Clustering* untuk klasifikasi kualitas pendidikan dan segmentasi wilayah di Indonesia juga telah dilakukan oleh Mustakim dan Kamal (2021), Safwa et al. (2026) dan Sari et al. (2025) yang membuktikan efektivitas metode ini dalam menemukan pola karakteristik pendidikan antarwilayah. Sementara itu, segmentasi dengan metode *Hierarchical Clustering* telah dilakukan oleh Dewi & Ahadiyah (2022) dan Muflihan et al. (2022) untuk menganalisis pengelompokan berdasarkan indikator pendidikan dan raport mutu sekolah antardaerah dengan memberikan visualisasi hubungan antardaerah melalui dendrogram. Penelitian ini bertujuan untuk mengelompokkan provinsi di Indonesia berdasarkan indikator pendidikan sehingga dapat diketahui kelompok provinsi yang memiliki karakteristik pendidikan yang relatif sama.

2 METODE

2.1 Data

Penelitian ini mencakup 38 Provinsi di Indonesia dengan tahun 2025. Data yang digunakan adalah data sekunder yang diambil dari Badan Pusat Statistik (BPS) dengan enam belas indikator pendidikan yang disajikan pada Tabel 1 (Badan Pusat Statistik, 2025). Pemilihan indikator-indikator tersebut didasarkan pada indikator yang mampu menggambarkan kondisi pendidikan, baik dari aspek akses, partisipasi, maupun capaian pendidikan masyarakat.

Tabel 1. Variabel Penelitian

Indikator	Satuan
Rata-rata Lama Sekolah (RLS)	Tahun
Harapan Lama Sekolah (HLS)	Tahun
Angka Partisipasi Sekolah Umur 7 – 12 Tahun	Persen
Angka Partisipasi Sekolah Umur 13 – 15 Tahun	Persen
Angka Partisipasi Sekolah Umur 16 – 18 Tahun	Persen
Angka Partisipasi Kasar Perguruan Tinggi	Persen
Angka Partisipasi Murni SD	Persen
Angka Partisipasi Murni SMP	Persen
Angka Partisipasi Murni SMA/SMK	Persen
Angka Kesiapan Sekolah	Persen
Angka Anak Tidak Sekolah Umur 13 – 15 Tahun	Persen
Angka Melek Huruf	Persen
Rasio Guru dan Murid SD	Orang
Tingkat Penyelesaian Pendidikan SD	Persen
Tingkat Penyelesaian Pendidikan SMP	Persen
Tingkat Penyelesaian Pendidikan SMA/SMK	Persen

2.2 Pra-pemrosesan Data

Tahap pra-pemrosesan data meliputi *data cleaning* dan standardisasi data, sehingga dapat dipastikan data yang digunakan untuk penelitian telah akurat guna mendapatkan hasil klusterisasi yang tepat (Gori et al., 2024). Karena data bersumber dari BPS, sehingga telah dipastikan tidak ada *missing values* ataupun data duplikat. Lalu standardisasi data pada setiap indikator yang digunakan. Standarisasi dilakukan karena setiap variabel memiliki satuan dan rentang nilai yang berbeda, sehingga diperlukan penyetaraan skala agar tidak terjadi dominasi variabel tertentu dalam proses pengelompokan. Metode standarisasi yang digunakan adalah *z-score standardization*.

2.3 Validasi Jumlah Klaster

Validasi jumlah klaster dilakukan untuk menentukan jumlah klaster optimal. Pada penelitian ini, penentuan jumlah klaster menggunakan metode *elbow method* dan *silhouette coefficient*. Metode *Elbow* digunakan dengan melihat penurunan nilai *Within Cluster Sum of Squares* (WCSS). Jumlah klaster optimal ditentukan pada titik yang membentuk pola siku (*elbow*), yaitu ketika penurunan nilai WCSS mulai melambat. Sedangkan *silhouette coefficient* untuk mengukur kualitas hasil *clustering*. Nilai *silhouette* yang semakin mendekati 1 menunjukkan bahwa pengelompokan data semakin baik. Berikut rumus perhitungan *elbow method* dan *silhouette score*:

$$\text{elbow method} = \sum_{k=1}^K \sum_{x_i \in C_k} |x_i - \mu_k|^2 \quad (1)$$

Keterangan:

K	= jumlah klaster
k	= indeks klaster ke- k
x_i	= data ke- i
C_k	= himpunan data yang tergabung dalam klaster ke- k
μ_k	= <i>centroid</i> klaster ke- k

$||x_i - \mu_k||^2$ = jarak Euclidean kuadrat antara data ke- i dan *centroid* klaster ke- k

$$silhouette\ score = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \quad (2)$$

Keterangan:

- $a(i)$ = rata-rata jarak data ke anggota dalam klaster yang sama
- $b(i)$ = rata-rata jarak data ke klaster terdekat lainnya
- $\max[a(i), b(i)]$ = nilai maksimum antara $a(i)$ dan $b(i)$

2.4 Analisis Klaster

2.4.1 K-Means Clustering

K-Means clustering termasuk dalam teknik klasterisasi non-hierarki. Metode ini bertujuan memetakan data ke dalam jumlah klaster dengan meminimalkan jumlah kuadrat jarak antara data dengan *centroid cluster*. Proses pengelompokan dilakukan secara iteratif hingga tidak terjadi lagi perubahan anggota klaster atau centroid telah konvergen (Prasetyo, 2014). Secara matematis dirumuskan sebagai berikut:

$$J = \sum_{j=1}^k \sum_{i=1}^n ||x_i - C_j||^2 \quad (3)$$

Keterangan:

- J = fungsi objektif
- k = jumlah klaster
- x_i = data ke- i
- C_j = *centroid* cluster ke- j

Langkah-langkah metode *K-Means* adalah sebagai berikut:

- a. Menentukan jumlah klaster (k).
- b. Menentukan *centroid* awal secara acak.
- c. Menghitung jarak setiap data terhadap *centroid*.
- d. Mengelompokkan data ke klaster dengan jarak terdekat.
- e. Menghitung ulang *centroid* berdasarkan rata-rata anggota klaster.
- f. Mengulangi proses hingga *centroid* stabil atau iterasi berhenti.

2.4.2 Hierarchical Clustering

Hierarchical clustering merupakan metode pengelompokan data yang membentuk struktur bertingkat dalam bentuk dendrogram. Metode ini digunakan untuk mengetahui hubungan kedekatan antarobjek berdasarkan tingkat kemiripan karakteristik data. Algoritma ini memiliki banyak cara, salah satunya adalah *Average Linkage Method*. Secara matematis dirumuskan sebagai berikut :

$$D(A, B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(x_i, x_j) \quad (4)$$

Keterangan:

- $D(A, B)$ = jarak antara klaster (A) dan (B)
- n_A = jumlah anggota klaster (A)
- n_B = jumlah anggota klaster (B)
- $d(x_i, x_j)$ = jarak Euclidean kuadrat antara data x_i dan x_j

Metode ini menggunakan pendekatan bottom-up, yakni menggabungkan setiap data secara

Bertahap menjadi satu kluster hingga seluruh data berhasil dikelompokkan (Justitia et al., 2021). Langkah – langkah metode *Hierarchical Clustering Average Linkage* adalah:

- Menghitung matriks jarak antar seluruh data.
- Menggabungkan dua data atau kluster dengan jarak terdekat.
- Menghitung kembali jarak antarkluster menggunakan *Average Linkage*.
- Mengulangi proses hingga seluruh data menjadi satu kluster.
- Menampilkan hasil pengelompokan dalam bentuk dendrogram.

2.5 Interpretasi Hasil

Interpretasi hasil dilakukan untuk memahami karakteristik dari setiap kluster yang terbentuk. Interpretasi dilakukan dengan menganalisis nilai rata-rata (*centroid*) pada masing - masing kluster sehingga dapat diketahui pola dan perbedaan karakteristik antar kelompok. Pada penelitian ini, hasil klustering digunakan untuk mengidentifikasi kondisi pendidikan antar provinsi di Indonesia berdasarkan indikator pendidikan. Selain itu, visualisasi seperti dendrogram dan grafik *clustering* digunakan untuk mempermudah proses interpretasi dan memahami hubungan antar kluster yang terbentuk.

3 HASIL DAN PEMBAHASAN

3.1 Statistik Deskriptif

Pada tahap sebelumnya telah dijelaskan bahwa penelitian ini menggunakan data sekunder dari BPS yang mencakup enam belas indikator pendidikan tahun 2025 dari 38 provinsi di Indonesia. Tabel 2 menyajikan statistik deskriptif yang mencakup nilai rata-rata (*mean*), standar deviasi, nilai minimum, dan nilai maksimum dari variabel indikator pendidikan yang digunakan dalam penelitian ini.

Tabel 2. Statistik deskriptif variabel penelitian

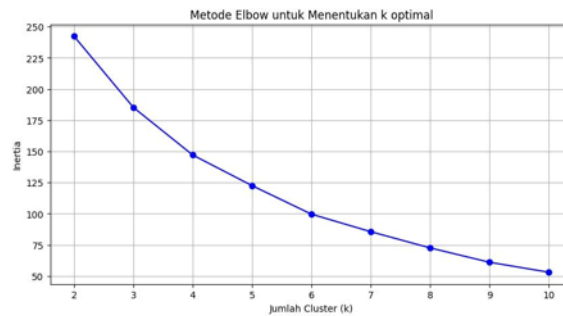
Variabel	Mean	Standar Deviasi	Nilai Minimum	Nilai Maksimum
Rata-rata Lama Sekolah (RLS)	9,0295	123,993	4,30	11,59
Harapan Lama Sekolah (HLS)	132,621	102,826	9,64	15,78
Angka Partisipasi Sekolah Umur 7-12 Tahun	980,574	424,883	80,13	99,74
Angka Partisipasi Sekolah Umur 13-15 Tahun	946,592	635,862	67,86	99,60
Angka Partisipasi Sekolah Umur 16-18 Tahun	767,453	895,066	44,61	93,03
Angka Partisipasi Kasar Perguruan Tinggi	335,621	1.017,913	13,34	74,70
Angka Partisipasi Murni SD	950,921	549,059	72,55	98,77
Angka Partisipasi Murni SMP	773,397	745,669	54,33	88,67
Angka Partisipasi Murni SMA/SMK	644,239	901,429	35,59	78,42
Angka Kesiapan Sekolah	740,550	1.589,349	29,77	96,48
Angka Anak Tidak Sekolah Umur 13-15Tahun	7,3655	672,040	2,22	35,14
Angka Melek Huruf	965,826	539,508	68,32	99,80
Rasio Guru dan Murid SD	130,000	292,242	9,00	21,00
Tingkat Penyelesaian Pendidikan SD	952,161	575,676	72,10	99,02
Tingkat Penyelesaian Pendidikan SMP	877,603	915,933	55,49	97,06
Tingkat Penyelesaian Pendidikan SMA/SMK	706,587	1.267,953	28,30	92,55

3.2 K – Means Clustering

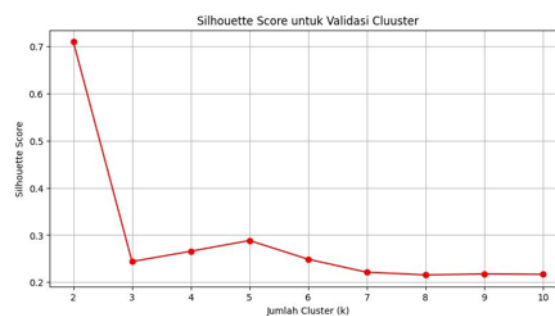
3.2.1 Penentuan nilai k

Penentuan jumlah kluster optimal dilakukan menggunakan metode *elbow* dan *silhouette score*. Metode *elbow* digunakan untuk melihat hubungan antara jumlah kluster (*k*) dengan nilai

inertia. Nilai inertia menunjukkan total variasi data di dalam *cluster*. Semakin kecil nilai inertia, maka anggota dalam klaster semakin homogen.



Gambar 1. Elbow Method



Gambar 2. Silhouette Score

Evaluasi berdasarkan metode *elbow*, terlihat bahwa penurunan inertia cukup tajam terjadi dari $k=2$ hingga $k=3$. Setelah itu, penurunan inertia mulai melandai sehingga penambahan jumlah *cluster* tidak lagi memberikan perubahan yang signifikan. Kondisi tersebut menunjukkan adanya titik siku (*elbow point*) pada $k=3$. Selain menggunakan *elbow method*, evaluasi kualitas klaster juga dilakukan menggunakan *silhouette score*. Nilai *silhouette score* pada metode *K-Means* diperoleh sebesar 0,2438 untuk $k=3$. Oleh karena itu, penelitian ini menggunakan 3 klaster karena dianggap mampu memberikan pengelompokan yang lebih informatif dan mudah diinterpretasikan ke dalam kategori tinggi, sedang, dan rendah.

3.2.2 Pengelompokkan dengan *K - Means*

Setelah ditentukan $k=3$ sebagai jumlah klaster optimal, dilakukan proses klasterisasi menggunakan metode *K-Means*. Sebaran provinsi per klaster dapat dilihat pada Tabel 3.

Tabel 3. Hasil Klaster dengan metode *K-Means*

Klaster	Jumlah Anggota	Provinsi
1	24	Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Bengkulu, Lampung, Kepulauan Riau, DKI Jakarta, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, Banten, Bali, Nusa Tenggara Barat, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Maluku, dan Maluku Utara.
2	2	Papua Tengah dan Papua Pegunungan.

3 12 Sumatera Selatan, Kepulauan Bangka Belitung, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Gorontalo, Sulawesi Barat, Papua Barat, Papua Barat Daya, Papua, dan Papua Selatan.

Tabel 4. *Centroid K-Means*

Indikator	Klaster 1 (Tinggi)	Klaster 2 (Rendah)	Klaster 3 (Sedang)
Rata-Rata Lama Sekolah (RLS)	9,54	5,22	8,64
Harapan Lama Sekolah (HLS)	13,64	9,81	13,07
Angka Partisipasi Sekolah Umur 7-12 Tahun	99,40	81,18	98,19
Angka Partisipasi Sekolah Umur 13-15 Tahun	96,93	69,47	94,32
Angka Partisipasi Sekolah Umur 16-18 Tahun	80,18	47,70	74,71
Angka Partisipasi Kasar Perguruan Tinggi	37,50	13,48	29,03
Angka Partisipasi Murni SD	96,97	73,70	94,90
Angka Partisipasi Murni SMP	80,89	55,24	73,92
Angka Partisipasi Murni SMA/SMK	69,03	37,73	59,67
Angka Kesiapan Sekolah	80,23	34,79	68,29
Angka Anak Tidak Sekolah Umur 13-15 Tahun	4,74	32,79	8,38
Angka Melek Huruf	97,73	77,07	97,54
Rasio Guru dan Murid SD	13,58	12,50	11,92
Tingkat Penyelesaian Pendidikan SD	97,74	75,86	93,40
Tingkat Penyelesaian Pendidikan SMP	92,29	55,94	84,00
Tingkat Penyelesaian Pendidikan SMA/SMK	76,73	31,33	65,08

Berdasarkan nilai *centroid* pada Tabel 4, setiap klaster memiliki karakteristik pendidikan yang berbeda. Klaster 1 memiliki nilai rata-rata tertinggi pada hampir seluruh indikator pendidikan, seperti Rata-Rata Lama Sekolah (RLS), Harapan Lama Sekolah (HLS), Angka Partisipasi Sekolah (APS), Angka Partisipasi Murni (APM), Angka Melek Huruf (AMH), serta tingkat penyelesaian pendidikan. Selain itu, Angka Anak Tidak Sekolah (AATS) pada klaster ini paling rendah yaitu sebesar 4,74%. Kondisi tersebut menunjukkan bahwa provinsi-provinsi dalam klaster 1 memiliki kualitas pendidikan yang relatif baik dengan tingkat partisipasi pendidikan yang tinggi. Sebaliknya, klaster 2 memiliki nilai terendah pada hampir seluruh indikator pendidikan. Nilai RLS, HLS, APS, APM, dan AMH pada klaster ini jauh lebih rendah dibanding klaster lainnya, sedangkan angka anak tidak sekolah sangat tinggi yaitu sebesar 32,79%. Hal ini menunjukkan bahwa provinsi pada klaster 2 masih memiliki kondisi pendidikan yang rendah dan akses pendidikan yang belum optimal.

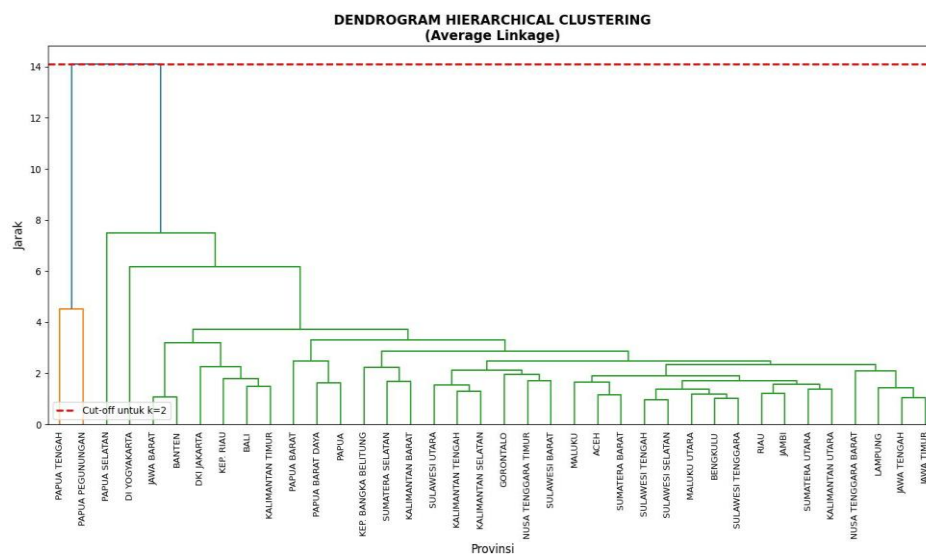
Sementara itu, klaster 3 memiliki nilai indikator pendidikan yang berada di antara klaster 1 dan klaster 2. Sebagian besar indikator pendidikan pada klaster ini menunjukkan kondisi yang cukup baik, namun masih lebih rendah dibanding klaster 1. Nilai angka anak tidak sekolah juga masih lebih tinggi dibanding klaster 1, meskipun jauh lebih rendah dibanding klaster 2. Oleh karena itu, klaster 3 dapat dikategorikan sebagai kelompok provinsi dengan kondisi pendidikan sedang atau berkembang. Secara umum, hasil *centroid* menunjukkan adanya perbedaan karakteristik pendidikan antar klaster.

Sementara itu, kluster 3 yang ditunjukkan dengan warna hijau terdiri dari 12 provinsi. Posisi kluster ini berada di tengah antara kluster 1 dan kluster 2, yang menunjukkan bahwa tingkat pendidikan di kluster ini berada dalam kategori sedang atau sedang berkembang.

3.3 Hasil *Hierarchical Clustering*

3.3.1 Penentuan Jumlah Kluster

Pada metode *Hierarchical Clustering*, penentuan jumlah kluster dilakukan menggunakan dendrogram. Dendrogram merupakan diagram berbentuk pohon yang menunjukkan proses penggabungan data berdasarkan tingkat kemiripan antarobjek. Metode yang digunakan dalam penelitian ini adalah *average linkage*. Pada metode ini, jarak antar kluster dihitung berdasarkan rata-rata jarak antar anggota kluster.



Gambar 4. Dendrogram

Berdasarkan dendrogram, di tahap akhir terdapat jarak penggabungan yang cukup besar sebelum semua data akhirnya bergabung menjadi satu kelompok. Kondisi ini menunjukkan bahwa data dari kedua kelompok terpisah dengan jelas, sehingga diputuskan untuk menggunakan 2 kluster sebagai jumlah kluster yang paling optimal. Hasil evaluasi menggunakan *silhouette score* sebesar 0,7536 menunjukkan bahwa struktur kluster yang terbentuk memiliki kualitas yang baik, dimana anggota dalam satu kluster memiliki tingkat homogenitas yang tinggi dan heterogenitas yang cukup jelas antar kluster.

3.3.2 Hasil Pengelompokan

Berdasarkan pemotongan dendrogram pada tiga kluster, diperoleh hasil pengelompokan berikut:

Tabel 5. Hasil Kluster dengan Metode *Hierarchical Clustering*

Kluster	Jumlah Anggota	Provinsi
1	2	Papua Tengah dan Papua Pegunungan.
2	36	Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Bengkulu, Lampung, Sumatera Selatan, Kepulauan Bangka Belitung, Kepulauan Riau, DKI Jakarta, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, Banten, Bali, Nusa Tenggara Barat,

Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Maluku, Maluku Utara, Gorontalo, Sulawesi Barat, Papua Barat, Papua Barat Daya, Papua Selatan, Papua.

Tabel 6. Profiling Hasil Klaster *Hierarchical Clustering*

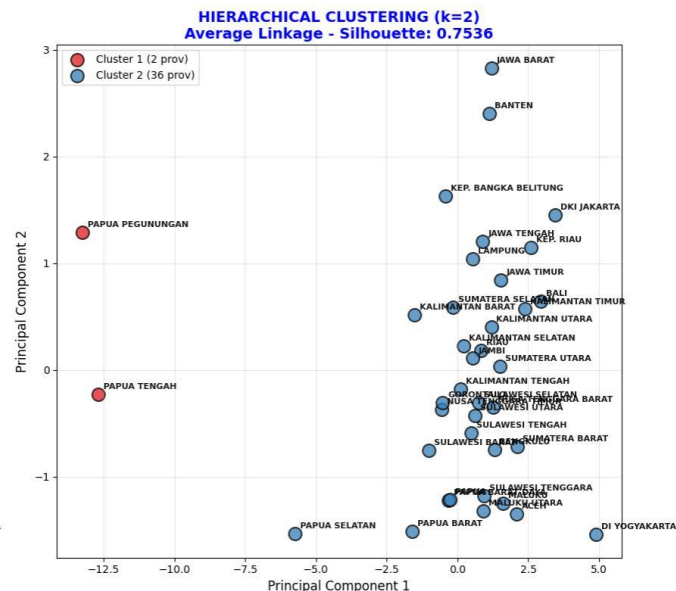
Indikator	Klaster 1 (Rendah)	Klaster 2 (Tinggi)
Rata – Rata Lama Sekolah (RLS)	5,21	9,24
Harapan Lama Sekolah (HLS)	9,81	13,45
Angka Partisipasi Sekolah Umur 7-12 Tahun	81,19	98,99
Angka Partisipasi Sekolah Umur 13-15 Tahun	69,47	96,06
Angka Partisipasi Sekolah Umur 16-18 Tahun	47,70	78,36
Angka Partisipasi Kasar Perguruan Tinggi	13,48	34,68
Angka Partisipasi Murni SD	73,70	96,28
Angka Partisipasi Murni SMP	55,24	78,57
Angka Partisipasi Murni SMA/SMK	37,73	65,91
Angka Kesiapan Sekolah	34,52	76,25
Angka Anak Tidak Sekolah Umur 13-15 Tahun	32,79	5,95
Angka Melek Huruf	77,07	97,67
Rasio Guru dan Murid SD	12,50	13,03
Tingkat Penyelesaian Pendidikan SD	75,86	96,29
Tingkat Penyelesaian Pendidikan SMP	55,94	89,53
Tingkat Penyelesaian Pendidikan SMA/SMK	31,34	72,84

Berdasarkan hasil profiling *Hierarchical Clustering*, setiap klaster memiliki karakteristik pendidikan yang berbeda. Klaster 1 menunjukkan nilai yang lebih rendah pada hampir seluruh indikator pendidikan dibandingkan klaster 2. Hal ini terlihat dari rendahnya nilai Rata-Rata Lama Sekolah (RLS) sebesar 5,21 tahun, Harapan Lama Sekolah (HLS) sebesar 9,81 tahun, serta rendahnya Angka Partisipasi Sekolah (APS), Angka Partisipasi Murni (APM), Angka Kesiapan Sekolah, Angka Melek Huruf (AMH), dan tingkat penyelesaian pendidikan. Selain itu, angka anak tidak sekolah usia 13–15 tahun pada klaster ini tergolong tinggi, yaitu 32,79%. Kondisi tersebut menunjukkan bahwa Klaster 1 merupakan kelompok provinsi dengan tingkat pendidikan yang relatif rendah atau masih tertinggal.

Sebaliknya, klaster 2 memiliki nilai yang lebih tinggi pada hampir seluruh indikator pendidikan. Nilai RLS mencapai 9,24 tahun dan HLS sebesar 13,45 tahun, disertai tingginya APS, APM, Angka Kesiapan Sekolah, AMH, serta tingkat penyelesaian pendidikan pada semua jenjang. Selain itu, angka anak tidak sekolah usia 13–15 tahun relatif rendah, yaitu hanya 5,95%. Meskipun rasio guru dan murid SD pada kedua klaster tidak menunjukkan perbedaan yang signifikan, secara keseluruhan Klaster 2 menunjukkan kualitas pendidikan yang lebih baik dibandingkan Klaster 1. Oleh karena itu, Klaster 2 dapat dikategorikan sebagai kelompok provinsi dengan kondisi pendidikan tinggi atau maju.

3.3.3 Visualisasi PCA Berdasarkan *Hierarchical Clustering*

Berdasarkan visualisasi PCA hasil *Hierarchical Clustering* dengan metode *average linkage* dan jumlah cluster sebanyak 2, terlihat bahwa provinsi-provinsi di Indonesia berhasil dikelompokkan ke dalam dua kelompok utama berdasarkan kemiripan indikator pendidikan.



Gambar 5. PCA *Hierarchical Clustering*

Berdasarkan Gambar 5, visualisasi PCA dari metode *Hierarchical Clustering* dengan jumlah klaster sebanyak 2 menunjukkan bahwa sebagian besar provinsi di Indonesia berada dalam klaster 2 yang terdiri dari 36 provinsi. Titik-titik pada klaster ini terpusat di area tengah hingga sebelah kanan grafik, yang menunjukkan bahwa provinsi-provinsi tersebut memiliki karakteristik pendidikan yang hampir sama. Kedekatan posisi antarprovinsi pada grafik menunjukkan tingkat kesamaan yang cukup besar berdasarkan indikator pendidikan yang digunakan dalam penelitian.

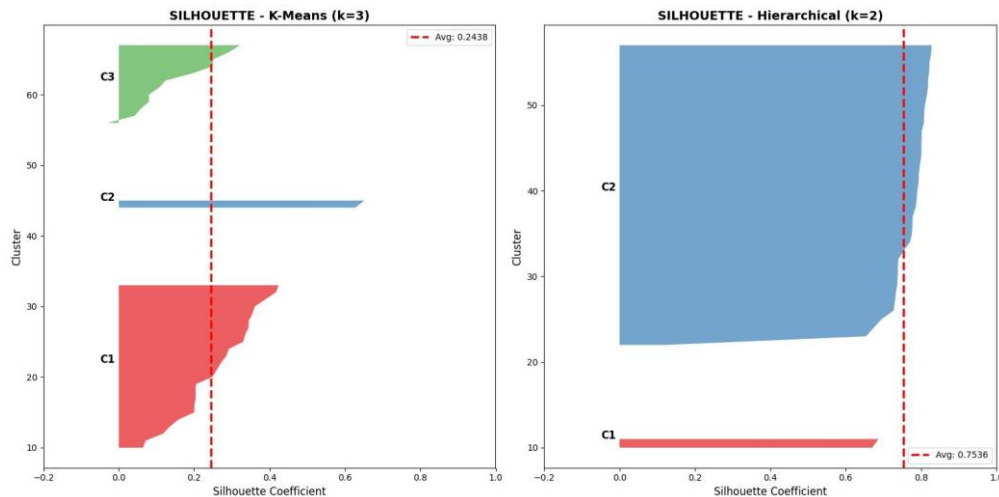
Sementara itu, klaster 1 hanya terdiri dari dua provinsi, yaitu Papua Pegunungan dan Papua Tengah. Dua provinsi tersebut tampak berada cukup jauh di bagian kiri grafik dan terpisah dari kelompok provinsi lainnya. Menunjukkan adanya perbedaan karakteristik pendidikan yang signifikan dibandingkan dengan mayoritas provinsi di Indonesia. Oleh karena itu, klaster 1 dapat dianggap sebagai kelompok dengan kondisi pendidikan yang relatif berbeda serta memerlukan perhatian lebih dalam berbagai indikator pendidikan.

Di klaster 2, beberapa provinsi seperti Jawa Barat, Banten, DKI Jakarta, Jawa Tengah, dan Kepulauan Bangka Belitung berada di posisi yang lebih tinggi pada komponen utama kedua. Ini menunjukkan bahwa beberapa provinsi memiliki capaian lebih baik dalam beberapa indikator pendidikan dibandingkan sebagian besar provinsi lainnya dalam satu klaster. Di sisi lain, ada beberapa provinsi yang berada di bagian bawah grafik, tetapi tetap termasuk dalam klaster 2 karena secara keseluruhan karakteristik pendidikannya masih memiliki kemiripan dengan anggota klaster tersebut.

3.4 Perbandingan *K-Means* dan *Hierarchical Clustering*

3.4.1 Nilai *Silhouette Score*

Perbandingan kedua metode dilakukan menggunakan *silhouette score*.



Gambar 6. *Silhouette Metode K-Means dan Hierarchical Clustering*

Tabel 7. *Nilai Silhouette*

Metode	<i>Silhouette Score</i>
<i>K-Means Clustering</i>	0,2438
<i>Hierarchical Clustering (Average Linkage)</i>	0,7536

Berdasarkan Tabel 7, metode *hierarchical clustering* dengan *average linkage* memiliki nilai *silhouette score* yang lebih tinggi sebesar 0,7536 dibanding metode *K-Means* yang hanya 0,2438. Hal tersebut menunjukkan bahwa kualitas pengelompokan yang dihasilkan oleh *hierarchical clustering* lebih baik. Nilai *silhouette score* yang tinggi mengindikasikan bahwa objek dalam suatu kluster memiliki tingkat kemiripan yang tinggi dengan anggota kluster yang sama dan memiliki perbedaan yang cukup jelas dengan kluster lainnya.

Berdasarkan evaluasi menggunakan *silhouette score*, metode *hierarchical clustering* (*average linkage*) dengan $k=2$ memberikan hasil pengelompokan yang lebih optimal dibandingkan metode *K-Means* dengan $k=3$. Meskipun demikian, metode *K-Means* tetap mampu memberikan segmentasi yang lebih rinci karena menghasilkan tiga kelompok provinsi, sehingga dapat memberikan gambaran yang lebih detail mengenai variasi karakteristik pendidikan antarprovinsi di Indonesia.

4 KESIMPULAN

Berdasarkan hasil penelitian menunjukkan adanya perbedaan karakteristik pendidikan antarprovinsi, yang terlihat dari munculnya beberapa kelompok dengan tingkat kemiripan tertentu. Metode *K-Means* menghasilkan tiga kluster dengan *silhouette score* sebesar 0,2438. Pengelompokan ini memberikan gambaran yang lebih rinci mengenai kondisi pendidikan antarprovinsi karena mampu membedakan provinsi ke dalam kategori tinggi, sedang, dan rendah. Sementara itu, metode *hierarchical clustering* dengan *average linkage* menghasilkan dua kelompok dan *silhouette score* yang diperoleh adalah 0,7536. Nilai tersebut menunjukkan bahwa kualitas pengelompokan yang dihasilkan lebih baik karena memiliki tingkat kesamaan yang tinggi di dalam kluster serta pemisahan yang lebih jelas antar kluster. Penelitian ini juga menunjukkan bahwa Papua Tengah dan Papua Pegunungan secara konsisten berada pada kelompok yang berbeda dari sebagian besar provinsi lainnya. Secara umum, metode *hierarchical clustering* dengan *average linkage* memberikan hasil yang lebih baik berdasarkan nilai *silhouette score*. Namun, *K-Means* mampu memberikan segmentasi yang lebih detail sehingga variasi karakteristik pendidikan antarprovinsi dapat terlihat dengan lebih jelas.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada seluruh pihak yang telah memberikan dukungan, bantuan, dan masukan selama proses pelaksanaan penelitian hingga penyusunan artikel ini.

DAFTAR PUSTAKA

- Badan Pusat Statistik. (2025). Statistik pendidikan 2025. <https://www.bps.go.id/id/publication/2025/11/21/d048070f37740b0e04d99350/statistik-pendidikan-2025.html>
- Dewi, A. F., & Ahadiyah, K. (2022). Agglomerative hierarchy clustering pada penentuan kelompok kabupaten/kota di Jawa Timur berdasarkan indikator pendidikan. *Zeta – Math Journal*, 7(2), 57–63. <https://doi.org/10.31102/zeta.2022.7.2.57-63>
- Ferreira, L., & de Souza, A. (2017). Clustering analysis for educational data mining: A review of applications and methods. *International Journal of Advanced Computer Science and Applications*, 8(6), 1–8. <https://doi.org/10.14569/IJACSA.2017.080601>
- Gori, T., Sunyoto, A., & Al Fatta, H. (2024). Preprocessing data dan klasifikasi untuk prediksi kinerja akademik siswa. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 11(1), 215–224. <https://doi.org/10.25126/jtiik.20241118074>
- Hanifah, N., & Primandari, A. H. (2023). Penerapan metode K-Means dalam pengelompokan kabupaten/kota berdasarkan indikator pendidikan di Indonesia. *Jurnal Informatika dan Sistem Informasi*, 4(2), 115–124. <https://doi.org/10.33005/jursistek.v4i2.178>
- Justitia, R. P., Hidayat, N., & Santoso, E. (2021). Implementasi metode agglomerative hierarchical clustering pada segmentasi pelanggan barbershop (Studi kasus: RichDjoe Barbershop Malang). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 5(3), 1048–1054. <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/8730>
- Muflihan, Y., Retnawati, H., & Kistian, A. (2022). Analisis cluster dengan metode hierarki untuk pengelompokan sekolah menengah atas berdasarkan raport mutu sekolah di Kabupaten Nagan Raya. *Measurement in Educational Research*, 2(1), 22–33. <https://doi.org/10.33292/meter.v2i1.155>
- Mustakim, Z., & Kamal, R. (2021). K-Means clustering for classifying the quality management of secondary education in Indonesia. *Jurnal Cakrawala Pendidikan*, 40(3), 663–675. <https://doi.org/10.21831/cp.v40i3.40150>
- Prasetyo, E. (2014). Data mining: Mengolah data menjadi informasi menggunakan MATLAB. Andi.
- Putri, F. A., Sani, M. R., Azura, S. M., Gustiawati, L., & Suryani, M. (2025). Evaluasi sistem pendidikan nasional dalam mewujudkan Generasi Emas Indonesia. *Jurnal Pendidikan Tambusai*, 9(2), 25551–25556. <https://doi.org/10.31004/jptam.v9i2.30841>
- Safwa, B. J. S., Tectona, Z. S., & Widodo, E. (2026). Segmentation of educational quality in Indonesian provinces based on K-Means clustering. *Jurnal Sintak*, 4(2), 71–81. <https://doi.org/10.62375/jsintak.v4i2.794>
- Sari, N. N., Adzima, K. R., Sahila, S., & Khotimah, T. H. (2025). K-Means clustering to classify Indonesian provinces based on school participation and socio-economic indicators. *Jurnal Riset Rumpun Matematika dan Ilmu Pengetahuan Alam*, 4(2), 45–56. <https://doi.org/10.55606/jurrimipa.v4i2.6657>
- Siregar, M. R., Lubis, N. F., & Hasibuan, Y. S. (2023). Perbandingan metode K-Means dan hierarchical clustering dalam pengelompokan data pendidikan. *Jurnal Teknologi*

Informasi dan Ilmu Komputer, 10(4), 721–730.
<https://doi.org/10.25126/jtiik.20231041234>

- Sugiarta, I. M., Mardana, I. B. P., Adiarta, A., & Artanayasa, I. W. (2019). Filsafat pendidikan Ki Hajar Dewantara (Tokoh Timur). *Jurnal Filsafat Indonesia*, 2(3), 124–136.
<https://doi.org/10.23887/jfi.v2i3.22187>
- Syafiq, M., Millati, N. F., Idris, M. A., Fitrianto, A., Alifviansyah, K., & Erfiani. (2026). Clustering of Central Java districts based on educational indicators: A comparison of K-Means and hierarchical methods. *Journal of Mathematics, Computations and Statistics*, 9(1), 192–205. <https://doi.org/10.35580/xen35m31>