
Segmentation Of Educational Quality In Indonesian Provinces Based On K-Means Clustering

Baiq Jasmin Sabhira Safwa V.R¹, Zakiy Suryahadi Tectona², Edy Widodo^{3*}

^{1,2,3} Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia, Jl.
Kaliurang km 14.5 Yogyakarta 55584, Indonesia

¹ baiqjasminsabhira@gmail.com, ² zakiy270205@gmail.com, ³ edywidodo@uii.ac.id

*Corresponding author

ABSTRAK

Kualitas pendidikan di Indonesia masih menunjukkan adanya ketimpangan antar provinsi yang mencerminkan perbedaan capaian dan akses pendidikan. Penelitian ini bertujuan untuk melakukan segmentasi kualitas pendidikan provinsi di Indonesia berdasarkan kemiripan karakteristik pendidikan menggunakan metode K-Means Clustering. Data yang digunakan merupakan data pendidikan tingkat provinsi yang telah melalui tahap pemeriksaan outlier dan standarisasi untuk memastikan kesetaraan skala antar variabel. Analisis K-Means Clustering dilakukan dengan membentuk tiga kluster yang merepresentasikan kelompok provinsi dengan kualitas pendidikan rendah, menengah, dan tinggi. Hasil pengelompokan menunjukkan bahwa sebagian besar provinsi berada pada kluster kualitas pendidikan menengah, sementara sebagian kecil provinsi masih berada pada kluster kualitas pendidikan rendah. Temuan ini menunjukkan bahwa metode K-Means Clustering mampu memberikan gambaran yang jelas mengenai pola segmentasi dan ketimpangan kualitas pendidikan antar provinsi di Indonesia, serta dapat menjadi dasar dalam mendukung perumusan kebijakan pendidikan yang lebih tepat sasaran dan berorientasi pada pemerataan kualitas pendidikan.

Kata Kunci: Pendidikan; Kualitas; K-Means; Klasterisasi; Provinsi

ABSTRACT

The quality of education in Indonesia still exhibits disparities among provinces, reflecting differences in educational attainment and access. This study aims to segment the quality of education across Indonesian provinces based on the similarity of educational characteristics using the K-Means Clustering method. The data used consist of provincial-level education data that have undergone outlier detection and standardization to ensure comparability across variables. K-Means Clustering analysis was performed by forming three clusters representing provinces with low, medium, and high levels of educational quality. The clustering results indicate that most provinces fall into the medium education quality cluster, while a smaller number of provinces remain in the low education quality cluster. These findings demonstrate that the K-Means Clustering method is able to provide a clear representation of segmentation patterns and disparities in educational quality across Indonesian provinces and can serve as a basis for supporting more targeted and equity-oriented education policy formulation.

Keywords: Education; Quality; K-Means; Clustering; Provinces

1. PENDAHULUAN

Pendidikan merupakan pilar utama dalam pembangunan sumber daya manusia dan pertumbuhan sosial ekonomi suatu negara. Dalam konteks Indonesia, kualitas pendidikan antarprovinsi masih menunjukkan ketimpangan yang nyata. Data Statistik Pendidikan Indonesia 2024 yang diterbitkan oleh Badan Pusat Statistik menunjukkan adanya perbedaan capaian indikator pendidikan, seperti Rata-rata Lama Sekolah (RLS), Harapan Lama Sekolah (HLS), Angka Partisipasi Murni jenjang SD, SMP, dan SMA (APM SD, APM SMP, APM SMA), Angka Partisipasi Sekolah (APS), serta Angka Melek Huruf (AMH) di antara 38 provinsi di Indonesia [1].

Selama ini, analisis kualitas pendidikan antarwilayah di Indonesia umumnya dilakukan menggunakan indikator statistik pendidikan yang disusun dan dipublikasikan oleh pemerintah, seperti angka partisipasi pendidikan, angka melek huruf, dan statistik pendidikan lainnya yang

tercantum dalam publikasi Indonesia *Educational Statistics in Brief* 2022/2023 oleh Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi [2]. Pendekatan ini memberikan gambaran umum kondisi pendidikan secara cepat dan mudah dipahami, namun bersifat parsial karena belum mampu menangkap keterkaitan antarindikator pendidikan secara simultan.

Meskipun laporan statistik nasional menunjukkan adanya disparitas pendidikan antarprovinsi, penelitian yang secara khusus melakukan segmentasi kualitas pendidikan antarwilayah menggunakan pendekatan multivariat masih relatif terbatas. Beberapa studi menunjukkan bahwa perbedaan akses fasilitas pendidikan, kualitas pengajaran, dan kondisi sosial ekonomi memengaruhi capaian pendidikan antarprovinsi, namun belum banyak penelitian yang mengelompokkan provinsi berdasarkan kemiripan karakteristik pendidikan secara menyeluruh melalui metode *clustering* [3]. Kondisi ini menunjukkan adanya research gap dalam pemetaan kualitas pendidikan yang komprehensif.

Untuk mengatasi keterbatasan tersebut, diperlukan pendekatan statistik multivariat yang mampu mengelompokkan provinsi berdasarkan kemiripan karakteristik pendidikan. Salah satu metode yang relevan adalah *K-Means Clustering*, yang efektif dalam mengidentifikasi kelompok homogen berdasarkan pola data multivariat. Penelitian sebelumnya menunjukkan bahwa metode K-Means dapat digunakan untuk mengklasifikasikan kualitas pendidikan wilayah berdasarkan beberapa indikator secara simultan [4]. Oleh karena itu, penelitian ini bertujuan untuk melakukan segmentasi kualitas pendidikan provinsi di Indonesia menggunakan metode K-Means Clustering berdasarkan indikator RLS, HLS, APM SD, APM SMP, APM SMA, APS, dan AMH, sehingga diharapkan dapat menghasilkan pemetaan kualitas pendidikan yang lebih informatif sebagai dasar perumusan kebijakan pendidikan yang lebih tepat sasaran.

2. METODE

2.1 Data Dan Variabel Penelitian

Penelitian ini menggunakan data sekunder yang bersumber dari publikasi Badan Pusat Statistik (BPS) 2024. Data yang digunakan mencakup indikator-indikator kualitas pendidikan di seluruh provinsi Indonesia yang tersedia secara lengkap pada periode pengamatan. Data yang digunakan dalam penelitian ini merupakan data *cross-section* dengan unit observasi berupa 38 provinsi di Indonesia pada tahun 2024. Seluruh variabel yang digunakan merupakan data rasio yang bersifat kontinu.

Populasi yang digunakan dalam penelitian ini adalah seluruh provinsi di Indonesia yang memiliki data mengenai indikator kualitas pendidikan, yaitu Rata-Rata Lama Sekolah (RLS), Harapan Lama Sekolah (HLS), Angka Partisipasi Murni (APM) SD, APM SMP, APM SMA, Angka Partisipasi Sekolah (APS), dan Angka Melek Huruf (AMH). Penelitian ini menggunakan data pada tahun 2024.

Variabel yang digunakan dalam penelitian ini terdiri dari tujuh indikator kualitas pendidikan. Rincian lengkap mengenai definisi operasional, dan skala data masing-masing variabel disajikan pada **Tabel 1** berikut.

Table 1. Definisi Operasional Variabel

Variabel	Keterangan	Definisi Operasional	Skala Data
RLS	Rata-Rata Lama Sekolah	Rata-rata lama pendidikan penduduk	Rasio
HLS	Harapan Lama Sekolah	Proyeksi lama pendidikan yang diharapkan ditempuh anak usia sekolah	Rasio

Variabel	Keterangan	Definisi Operasional	Skala Data
APM SD/Sederajat	Angka Partisipasi Murni SD/Sederajat	Persentase penduduk usia 7-12 tahun bersekolah di SMA/ sederajat	Rasio
APM SMP/Sederajat	Angka Partisipasi Murni SMP/Sederajat	Persentase penduduk usia 13-15 tahun bersekolah di SMA/ sederajat	Rasio
APM SMA/Sederajat	Angka Partisipasi Murni SMA/Sederajat	Persentase penduduk usia 16-18 tahun bersekolah di SMA/ sederajat	Rasio
APS	Angka Partisipas Sekolah	Persentase Penduduk usia sekolah yang masih bersekolah	Rasio
AMH	Angka Melek Huruf	Presentase penduduk usia 15-24 tahun ke atas yang dapat membaca	Rasio

2.2 Metode Analisis Statistik

Metode analisis yang digunakan dalam penelitian ini adalah *K-Means clustering*. *K-Means* merupakan algoritma *clustering non-heirarki* yang termasuk dalam metode partisi (*partitioning method*), yang bertujuan untuk mengelompokkan data ke dalam sejumlah *cluster* berdasarkan tingkat kemiripan karakteristiknya [5]. Metode ini digunakan untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan indikator kualitas pendidikan yang dimiliki.

Proses *K-Means clustering* diawali dengan pembentukan partisi cluster di awal, yang kemudian dilakukan proses iteratif untuk memperbaiki pembagian cluster hingga tidak terjadi perubahan yang signifikan atau hingga tercapai kondisi konvergen [6]. *K-Means clustering* membagi data berdasarkan jarak antara data dengan pusat cluster (*centroid*). Algoritma ini menggunakan fungsi jarak untuk mengukur kemiripan antar data. Jarak tersebut dihitung menggunakan jarak Euclidean, dan proses ini terus diulang hingga posisi *centroid* stabil [7]. Setelah itu, setiap data dikelompokkan ke dalam cluster yang memiliki jarak terdekat dengan *centroid*nya.

2.2.1 Analisis Deskriptif

Tahap pra-pemrosesan data dilakukan untuk memastikan kualitas data sebelum analisis lebih lanjut. Pada tahap ini dilakukan pemeriksaan terhadap kelengkapan data (*missing value*) serta eksplorasi distribusi data menggunakan statistik deskriptif untuk setiap variabel. Statistik deskriptif yang digunakan meliputi nilai minimum, maksimum, rata-rata (*mean*), dan standar deviasi untuk memberikan gambaran awal mengenai sebaran dan karakteristik data.

Deteksi *outlier* dilakukan menggunakan visualisasi *boxplot* untuk mengidentifikasi pengamatan dengan nilai ekstrem yang berpotensi memengaruhi hasil analisis *clustering*. *Boxplot* merupakan salah satu cara dalam statistik deskriptif untuk menggambarkan data secara grafis yang efektif untuk mendeteksi *outlier* [8]. Metode ini menampilkan distribusi data berdasarkan lima ringkasan statistik yaitu nilai minimum, kuartil pertama (Q1), median (Q2), kuartil tiga (Q3), dan nilai maksimum.

Nilai yang berada di luar batas *Whisker Boxplot*, yaitu $Q1 - 1.5 \times IQR$ atau di atas $Q3 + 1.5 \times IQR$, dianggap sebagai *outlier* potensial. Namun dalam konteks *clustering*, *outlier* tidak selalu dihapus karena dapat mempresentasikan karakteristik unik dari kelompok tertentu yang justru penting untuk diidentifikasi.

Selanjutnya, data distandarisasi menggunakan metode *Z-Score* untuk mengatasi perbedaan skala dan satuan antar variabel sehingga setiap variabel memiliki kontribusi yang seimbang dalam proses pengelompokan. Standarisasi ini mengubah data sehingga rata-rata 0 dan standar deviasi 1. Berikut adalah rumus *Z-Score*

$$Z_i = \frac{x_i - \bar{x}}{s} \quad (1)$$

dengan Z_i menyatakan nilai hasil standarisasi untuk observasi ke- i , x_i merupakan nilai asli variabel pada observasi ke- i , $\bar{x}$ menyatakan nilai rata-rata variabel, dan s merupakan standar deviasi variabel.

Standarisasi *Z-Score* sangat penting dalam *K-Means clustering* karena algoritma ini menggunakan jarak Euclidean yang sensitif terhadap perbedaan skala [9]. Tanpa standarisasi, variabel dengan nilai numerik yang lebih besar akan mendominasi perhitungan jarak dan menghasilkan clustering yang bias. Setelah dilakukan standarisasi, setiap variabel memiliki pengaruh yang seimbang dalam pembentukan cluster, sehingga hasil pengelompokan lebih akurat dan tidak didominasi oleh variabel tertentu.

2.2.2 Penentuan Jumlah Cluster

Penentuan jumlah cluster optimal dalam penelitian ini dilakukan menggunakan Elbow Method dan Silhouette Method. Elbow Method digunakan dengan menganalisis perubahan nilai Within-Cluster Sum of Squares (WCSS) pada beberapa kemungkinan jumlah cluster, di mana jumlah cluster optimal ditunjukkan oleh titik siku (elbow point) pada grafik yang merepresentasikan penurunan WCSS yang mulai melambat [10]. Selanjutnya, Silhouette Method digunakan untuk mengevaluasi kualitas hasil pengelompokan dengan mengukur tingkat kedekatan objek terhadap cluster-nya dibandingkan dengan cluster lain, di mana nilai koefisien silhouette yang lebih tinggi menunjukkan struktur cluster yang lebih baik. Jumlah cluster optimal ditentukan berdasarkan kombinasi hasil kedua metode tersebut untuk memperoleh pengelompokan yang optimal dan mudah diinterpretasikan.

2.2.3 Pembentukan *K-Means Cluster*

Analisis *K-Means Clustering* dilakukan dengan menggunakan jumlah klaster optimal yang telah ditentukan sebelumnya. Proses pengelompokan dilakukan berdasarkan skor komponen utama hasil PCA. Algoritma *K-Means* bekerja secara iteratif dengan memperbarui pusat klaster (centroid) hingga tidak terjadi perubahan signifikan pada keanggotaan klaster atau hingga tercapai kondisi konvergen.

3. HASIL DAN PEMBAHASAN

3.1 Analisis Deskriptif

Analisis Deskriptif merupakan statistik yang mempelajari cara pengumpulan dan penyajian data agar mudah dipahami. Pada penelitian ini, analisis deskriptif digunakan untuk menggambarkan karakteristik dan ringkasan data indikator kualitas pendidikan provinsi di Indonesia tahun 2024, yang meliputi Rata-rata Lama Sekolah (RLS), Angka Partisipasi Murni (APM) pada jenjang SD, SMP, dan SMA/ sederajat, Harapan Lama Sekolah (HLS), Angka Partisipasi Sekolah (APS) usia 7–18 tahun, serta Angka Melek Huruf (AMH) usia 15–24 tahun. Analisis ini bertujuan untuk memberikan gambaran awal mengenai kondisi dan variasi indikator pendidikan antar provinsi sebelum dilakukan proses pengelompokan menggunakan metode *K-Means Clustering*.

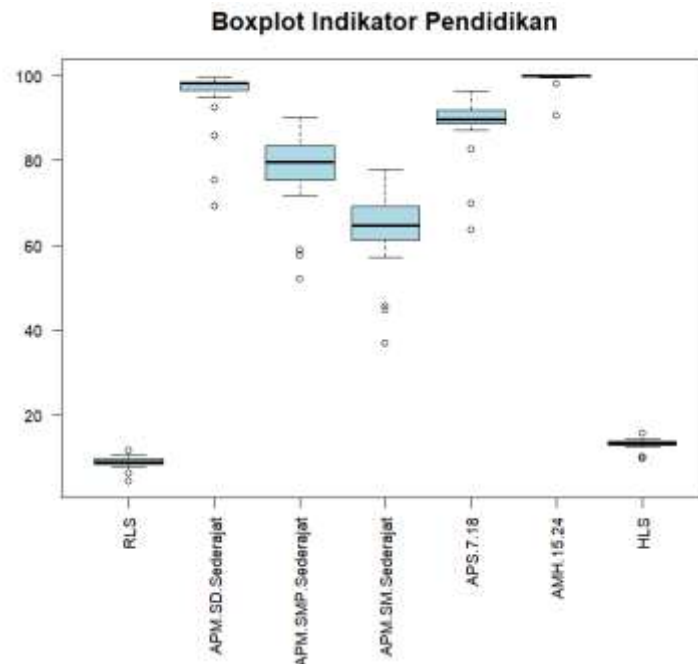
Tabel 2. Analisis Deskriptif

	Minimum	Maximum	Mean
RLS	4.210	11.490	8.841
APM SD/Sederajat	52.03	90.14	78.14
APM SMP/Sederajat	36.84	77.86	64.08
APM SMA/Sederajat	69.03	99.43	95.94
HLS	9.63	15.70	13.21
APS 7-18	63.57	96.30	89.14
AMH 15-24	90.46	99.97	99.32

Pada **Tabel 2**. Indikator pendidikan antar provinsi di Indonesia tahun 2024 menunjukkan adanya variasi yang cukup nyata. Nilai Rata-rata Lama Sekolah (RLS) berada pada rentang 4,21 hingga 11,49 tahun dengan nilai rata-rata sebesar 8,84 tahun, yang mengindikasikan perbedaan capaian pendidikan penduduk antar wilayah. Angka Partisipasi Murni (APM) pada jenjang pendidikan dasar relatif tinggi dengan rata-rata 95,94 persen, namun masih menunjukkan perbedaan antar provinsi, sementara APM pada jenjang SMP dan SMA/ sederajat memiliki nilai rata-rata yang lebih rendah, masing-masing sebesar 78,14 persen dan 64,08 persen, serta rentang nilai yang lebih lebar. Hal ini mencerminkan adanya penurunan tingkat partisipasi pendidikan seiring meningkatnya jenjang pendidikan. Indikator Harapan Lama Sekolah (HLS) memiliki nilai rata-rata sebesar 13,21 tahun dengan rentang antara 9,63 hingga 15,70 tahun, yang menunjukkan perbedaan peluang pendidikan jangka panjang antar provinsi. Sementara itu, Angka Partisipasi Sekolah (APS) usia 7–18 tahun memiliki nilai rata-rata sebesar 89,14 persen, dan Angka Melek Huruf (AMH) usia 15–24 tahun menunjukkan nilai yang sangat tinggi dan relatif merata dengan rata-rata 99,32 persen, menandakan bahwa tingkat literasi dasar pemuda di Indonesia secara umum telah tercapai dengan baik.

3.2 Pengecekan Data *Outlier*

Pemeriksaan data merupakan tahapan awal yang penting dalam analisis deskriptif untuk mengidentifikasi keberadaan nilai pencilan (*outlier*) pada data yang akan digunakan dalam analisis kluster. Keberadaan outlier berpotensi memengaruhi hasil pengelompokan, mengingat metode *K-Means Clustering* bersifat sensitif terhadap nilai ekstrem. Oleh karena itu, pada penelitian ini dilakukan pemeriksaan outlier menggunakan metode boxplot untuk menggambarkan sebaran data serta mendeteksi adanya nilai pencilan antar provinsi sebelum dilakukan analisis kluster lebih lanjut.



Gambar 1. Pengecekan Data *Outlier*

Berdasarkan **Gambar 1**. Hasil visualisasi boxplot menunjukkan bahwa indikator pendidikan antar provinsi di Indonesia memiliki tingkat sebaran yang berbeda-beda, dengan sebagian indikator menunjukkan distribusi yang relatif homogen dan sebagian lainnya memperlihatkan variasi yang cukup besar. Keberadaan beberapa nilai pencilan (*outlier*) mengindikasikan adanya provinsi dengan kondisi pendidikan yang jauh berbeda dibandingkan mayoritas wilayah lainnya, yang mencerminkan ketimpangan kualitas pendidikan antar provinsi. Nilai pencilan tersebut tetap dipertahankan dalam analisis karena merepresentasikan kondisi riil di lapangan, namun pengaruhnya terhadap proses pengelompokan diminimalkan melalui standarisasi data sebelum penerapan metode *K-Means Clustering*, sehingga hasil segmentasi yang diperoleh tetap objektif dan mencerminkan perbedaan karakteristik pendidikan antar wilayah secara proporsional.

3.3 Standarisasi Data

Sebelum dilakukan analisis kluster menggunakan metode *K-Means Clustering*, data indikator pendidikan terlebih dahulu distandarisasi untuk menghilangkan pengaruh perbedaan skala antar variabel. Proses standarisasi dilakukan dengan mentransformasikan setiap variabel ke dalam bentuk nilai baku (*z-score*) sehingga memiliki rata-rata mendekati nol dan simpangan baku sebesar satu. Tahapan ini bertujuan untuk memastikan bahwa seluruh indikator pendidikan memiliki kontribusi yang seimbang dalam perhitungan jarak antar provinsi pada proses pengelompokan. Keberhasilan proses standarisasi dikonfirmasi melalui hasil validasi yang menunjukkan bahwa nilai rata-rata dan simpangan baku masing-masing variabel telah memenuhi kriteria standarisasi sebagaimana ditampilkan pada output analisis.

Tabel 3. Standarisasi Data

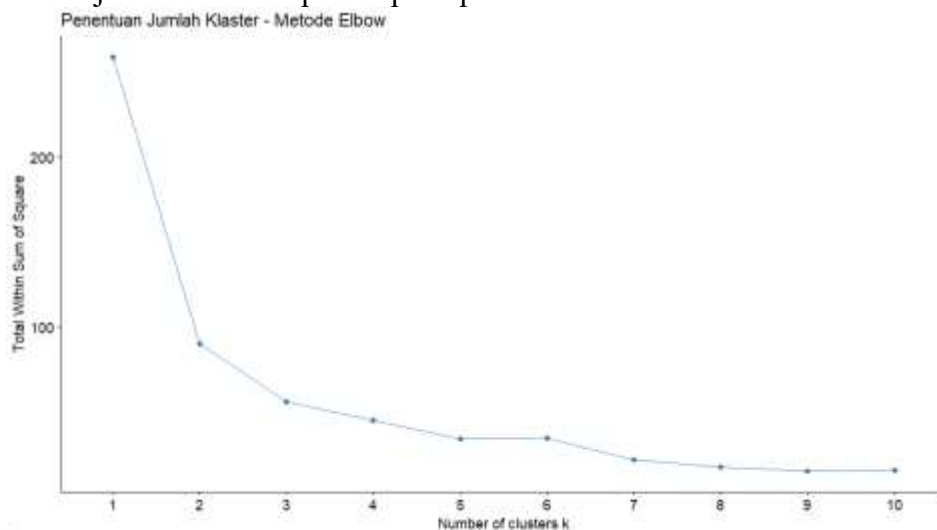
	Mean	Simpangan Baku
RLS	≈ 0	1
APM SD/Sederajat	≈ 0	1
APM SMP/Sederajat	≈ 0	1
APM SMA/Sederajat	≈ 0	1

	Mean	Simpangan Baku
HLS	≈ 0	1
APS 7-18	≈ 0	1
AMH 15-24	≈ 0	1

Seluruh indikator pendidikan yang digunakan dalam penelitian ini telah berhasil distandarisasi dengan baik, yang ditunjukkan oleh nilai rata-rata masing-masing variabel yang mendekati nol dan nilai simpangan baku sebesar satu. Hasil ini mengindikasikan bahwa perbedaan skala dan satuan pengukuran antar indikator telah dieliminasi, sehingga setiap variabel memiliki kontribusi yang seimbang dalam perhitungan jarak pada analisis *K-Means Clustering*. Dengan demikian, proses pengelompokan provinsi yang dilakukan selanjutnya dapat merepresentasikan perbedaan karakteristik kualitas pendidikan secara objektif tanpa dipengaruhi oleh dominasi variabel tertentu, yang mana terdapat pada **Tabel 3**.

3.4 Analisis K-Means Clustering

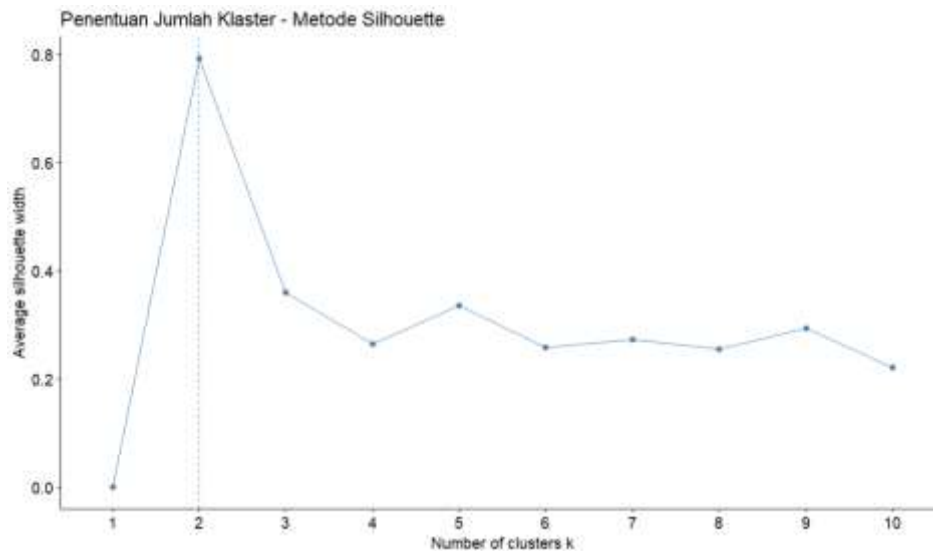
Penentuan jumlah kluster merupakan tahapan penting dalam analisis kluster karena berpengaruh terhadap kualitas serta kemudahan interpretasi hasil pengelompokan. Pada penelitian ini, jumlah kluster optimal ditentukan menggunakan Metode Elbow dengan mengamati pola perubahan nilai *Total Within Sum of Squares* (WSS) seiring dengan bertambahnya jumlah kluster yang dibentuk. Metode ini digunakan untuk mengidentifikasi titik di mana penurunan nilai WSS mulai melambat, sehingga penambahan kluster selanjutnya tidak memberikan peningkatan kualitas pengelompokan yang signifikan. Grafik hubungan antara jumlah kluster dan nilai WSS yang ditampilkan pada gambar berikut digunakan sebagai dasar dalam menentukan jumlah kluster optimal pada penelitian ini.



Gambar 2. Grafik Elbow

Grafik Metode Elbow yang terdapat pada **Gambar 2**, terlihat penurunan nilai *Total Within Sum of Squares* (WSS) terjadi sangat tajam ketika jumlah kluster meningkat dari satu ke dua kluster, kemudian penurunan mulai melandai setelah jumlah kluster tersebut. Pola ini menunjukkan adanya titik siku (*elbow*) pada jumlah kluster tertentu, di mana penambahan kluster selanjutnya tidak lagi memberikan penurunan WSS yang signifikan. Kondisi tersebut mengindikasikan bahwa struktur data mulai terbentuk secara optimal pada jumlah kluster tersebut, sehingga jumlah kluster yang berada di sekitar titik siku dianggap paling representatif untuk menggambarkan pengelompokan provinsi berdasarkan karakteristik kualitas pendidikan.

Selain menggunakan Metode Elbow, penentuan jumlah kluster optimal pada penelitian ini juga didukung oleh Metode Silhouette. Metode ini mengevaluasi kualitas pengelompokan dengan mengukur tingkat kesamaan suatu objek terhadap kluster tempatnya berada dibandingkan dengan kluster lainnya. Nilai *average silhouette width* yang lebih tinggi menunjukkan bahwa objek-objek dalam kluster memiliki tingkat kekompakan dan pemisahan antar kluster yang lebih baik. Grafik berikut menyajikan hubungan antara jumlah kluster dan nilai rata-rata silhouette yang digunakan sebagai dasar tambahan dalam menentukan jumlah kluster optimal.



Gambar 3. Grafik Silhouette

Berdasarkan **Gambar 3.** Grafik Metode Silhouette, nilai *average silhouette width* tertinggi diperoleh pada jumlah kluster sebanyak dua, yang menunjukkan bahwa pembagian data ke dalam dua kluster menghasilkan tingkat kekompakan dalam kluster dan pemisahan antar kluster yang paling baik dibandingkan jumlah kluster lainnya. Nilai silhouette yang relatif tinggi mengindikasikan bahwa objek-objek dalam kluster memiliki kemiripan yang tinggi satu sama lain serta perbedaan yang cukup jelas dengan objek pada kluster lain. Dengan demikian, secara teknis dan statistik, metode Silhouette mengonfirmasi bahwa struktur data indikator pendidikan provinsi di Indonesia paling optimal ketika dikelompokkan ke dalam dua kluster.

Meskipun hasil Metode Elbow dan Silhouette menunjukkan bahwa dua kluster merupakan jumlah kluster yang paling optimal secara statistik, penelitian ini tetap menggunakan tiga kluster dengan mempertimbangkan aspek substantif dan interpretabilitas hasil. Penggunaan tiga kluster memungkinkan segmentasi kualitas pendidikan yang lebih rinci, yaitu tidak hanya membedakan provinsi ke dalam kelompok dengan kualitas pendidikan relatif tinggi dan rendah, tetapi juga mengidentifikasi kelompok provinsi dengan karakteristik pendidikan menengah. Pendekatan ini dinilai lebih relevan dalam konteks analisis kebijakan, karena memberikan informasi yang lebih komprehensif mengenai variasi kualitas pendidikan antar provinsi. Oleh karena itu, pemilihan tiga kluster dalam penelitian ini bertujuan untuk memperoleh hasil segmentasi yang lebih informatif dan mudah diinterpretasikan, meskipun secara teknis dua kluster menunjukkan nilai evaluasi tertinggi.

Setelah jumlah kluster ditetapkan sebanyak tiga berdasarkan pertimbangan statistik dan substantif, langkah selanjutnya adalah melakukan pembentukan kluster menggunakan metode *K-Means Clustering*. Proses pengelompokan dilakukan terhadap data yang telah distandarisasi untuk memastikan setiap indikator memiliki kontribusi yang seimbang dalam perhitungan jarak. Pembentukan kluster ini menghasilkan pengelompokan provinsi ke dalam tiga kluster yang berbeda, lengkap dengan informasi ukuran kluster dan ukuran keragaman dalam dan antar

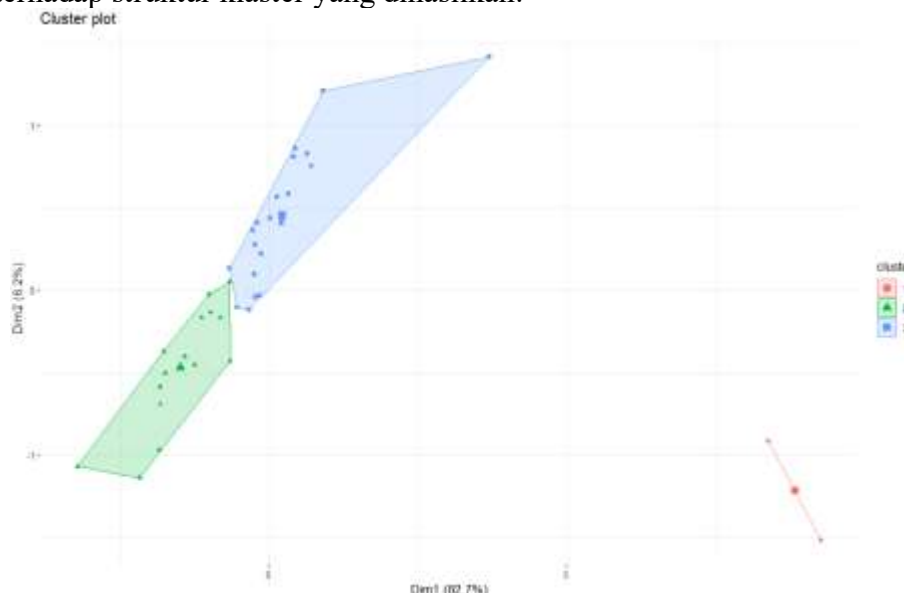
klaster. Hasil lengkap dari proses pengelompokan menggunakan metode K-Means disajikan pada output berikut dan menjadi dasar dalam analisis karakteristik masing-masing klaster.

Tabel 4. Pembentukan *Cluster*

Klaster	Jumlah Provinsi	Kualitas Klaster (Between SS/Total SS)
Klaster 1	2	78,3%
Klaster 2	15	78,3%
Klaster 3	21	78,3%
Total	38	78,3%

Dari **Tabel 4.** diatas. Distribusi klaster, hasil analisis *K-Means Clustering* membagi 38 provinsi di Indonesia ke dalam tiga klaster dengan jumlah anggota yang berbeda, yaitu Klaster 1 sebanyak 2 provinsi, Klaster 2 sebanyak 15 provinsi, dan Klaster 3 sebanyak 21 provinsi. Perbedaan jumlah anggota ini menunjukkan adanya variasi karakteristik kualitas pendidikan antar provinsi. Nilai kualitas klaster yang diukur melalui rasio *between-cluster sum of squares* terhadap *total sum of squares* sebesar 78,3% mengindikasikan bahwa sebagian besar variasi data dapat dijelaskan oleh perbedaan antar klaster, sehingga hasil pengelompokan yang diperoleh tergolong baik dan mampu memisahkan provinsi berdasarkan karakteristik kualitas pendidikan secara cukup jelas.

Untuk mendukung interpretasi hasil pengelompokan, visualisasi klaster digunakan sebagai alat bantu guna menggambarkan pola pemisahan antar klaster yang terbentuk. Visualisasi ini bertujuan untuk menunjukkan sejauh mana objek dalam klaster memiliki kedekatan satu sama lain serta bagaimana jarak antar klaster terbentuk berdasarkan indikator pendidikan yang dianalisis. Gambar berikut menyajikan visualisasi hasil *K-Means Clustering* yang memperlihatkan distribusi provinsi dalam masing-masing klaster, sehingga memudahkan pemahaman terhadap struktur klaster yang dihasilkan.



Gambar 4. Hasil Klasterisasi Dengan *K-Means*

Hasil pengelompokan menggunakan metode K-Means dengan tiga klaster, data terbagi ke dalam Klaster 1 sebanyak 2 anggota, Klaster 2 sebanyak 15 anggota, dan Klaster 3 sebanyak 21 anggota, yang menunjukkan adanya perbedaan karakteristik antar kelompok wilayah. Klaster 1 memiliki jumlah anggota paling sedikit dan merepresentasikan provinsi dengan karakteristik yang paling berbeda, sedangkan Klaster 2 dan Klaster 3 memiliki jumlah anggota yang lebih banyak dan relatif lebih homogen. Untuk memberikan gambaran yang lebih rinci

mengenai komposisi setiap klaster, tabel berikut menyajikan daftar provinsi yang termasuk dalam masing-masing klaster beserta jumlah anggotanya, yang terdapat pada **Gambar 4**.

Tabel 5. Hasil Klaster Provinsi-provinsi Indonesia

Klaster	Provinsi	Jumlah Anggota
1	Papua Pegunungan, Papua Tengah	2
2	Aceh, Bali, Bengkulu, DI Yogyakarta, DKI Jakarta, Kalimantan Timur, Kalimantan Utara, Kepulauan Riau, Maluku, Maluku Utara, Nusa Tenggara Barat, Riau, Sulawesi Tenggara, Sumatera Barat, Sumatera Utara	15
3	Banten, Gorontalo, Jambi, Jawa Barat, Jawa Tengah, Jawa Timur, Kalimantan Barat, Kalimantan Selatan, Kalimantan Tengah, Kepulauan Bangka Belitung, Lampung, Nusa Tenggara Timur, Papua, Papua Barat, Papua Barat Daya, Papua Selatan, Sulawesi Barat, Sulawesi Selatan, Sulawesi Tengah, Sulawesi Utara, Sumatera Selatan	21

Seperti pada **Tabel 5**, terlihat bahwa jumlah anggota paling banyak terdapat pada Klaster 3 dengan 21 provinsi, sedangkan Klaster 1 merupakan klaster dengan jumlah anggota paling sedikit, yaitu 2 provinsi. Setelah diperoleh hasil klasterisasi tersebut, analisis kemudian dilanjutkan dengan profilisasi klaster menggunakan nilai rata-rata variabel independen pada masing-masing klaster, yang dapat dilihat pada **Tabel 6**.

Tabel 6. Profilisasi Klaster

Variabel	Klaster 1	Klaster 2	Klaster 3
RLS	5,17	9,68	8,59
APM SD/Sederajat	72,22	97,99	96,73
APM SMP/Sederajat	56,16	83,34	76,34
APM SMA/Sederajat	41,22	70,70	61,53
HLS	66,65	92,60	88,81
APS 7-18	90,52	99,90	99,75
AMH 15-24	9,8	13,88	13,05

Berdasarkan hasil profilisasi pada **Tabel 6**. Klaster 1 dicirikan oleh nilai yang relatif paling rendah pada hampir seluruh indikator pendidikan dibandingkan dengan klaster lainnya. Klaster ini memiliki rata-rata Rata-rata Lama Sekolah (RLS) dan Harapan Lama Sekolah (HLS) yang paling rendah, yang menunjukkan capaian pendidikan penduduk serta peluang pendidikan jangka panjang yang masih terbatas. Selain itu, tingkat partisipasi pendidikan pada seluruh jenjang, baik pendidikan dasar, menengah pertama, maupun menengah atas, juga berada pada tingkat yang rendah. Nilai Angka Partisipasi Sekolah (APS) usia 7–18 tahun dan Angka Melek Huruf (AMH) usia 15–24 tahun yang relatif lebih rendah dibandingkan klaster lain mengindikasikan bahwa akses dan keberlanjutan pendidikan di wilayah ini masih menghadapi tantangan yang cukup besar. Secara keseluruhan, Klaster 1 merepresentasikan provinsi dengan kualitas pendidikan yang rendah dan memerlukan perhatian serta intervensi kebijakan yang lebih intensif.

Pada Klaster 2, indikator pendidikan menunjukkan nilai yang paling tinggi di antara seluruh klaster. Klaster ini memiliki rata-rata RLS dan HLS tertinggi, yang mencerminkan capaian pendidikan penduduk yang lebih baik serta peluang pendidikan yang lebih panjang. Tingkat partisipasi pendidikan pada seluruh jenjang juga berada pada kategori sangat tinggi,

yang ditunjukkan oleh nilai APM dan APS yang mendekati maksimum. Selain itu, nilai AMH yang hampir mencapai sempurna menunjukkan bahwa tingkat literasi pemuda di klaster ini sudah sangat baik. Secara keseluruhan, Klaster 2 merepresentasikan provinsi dengan kualitas pendidikan yang tinggi, akses pendidikan yang luas, serta sistem pendidikan yang relatif mapan dan berkelanjutan.

Sementara itu, Klaster 3 menunjukkan karakteristik kualitas pendidikan yang berada pada tingkat menengah. Nilai rata-rata indikator pendidikan pada klaster ini umumnya lebih tinggi dibandingkan Klaster 1, namun masih berada di bawah Klaster 2. Hal ini menunjukkan bahwa provinsi-provinsi dalam klaster ini telah memiliki akses dan partisipasi pendidikan yang cukup baik, khususnya pada pendidikan dasar dan menengah, tetapi masih terdapat ruang untuk peningkatan, terutama pada jenjang pendidikan menengah atas. Dengan demikian, Klaster 3 merepresentasikan wilayah dengan kualitas pendidikan yang cukup baik dan memiliki potensi besar untuk ditingkatkan melalui kebijakan pendidikan yang lebih terarah.

Secara keseluruhan, hasil profilisasi klaster ini menunjukkan bahwa metode *K-Means Clustering* mampu mengelompokkan provinsi di Indonesia ke dalam tiga segmen kualitas pendidikan, yaitu rendah, menengah, dan tinggi. Segmentasi ini memberikan gambaran yang jelas mengenai ketimpangan kualitas pendidikan antar provinsi serta dapat menjadi dasar dalam perumusan kebijakan pendidikan yang lebih tepat sasaran sesuai dengan karakteristik masing-masing klaster.

4. KESIMPULAN

Berdasarkan hasil pengelompokan provinsi menggunakan metode *K-Means clustering*, pemerintah pusat dan pemerintah daerah disarankan untuk memfokuskan intervensi kebijakan pendidikan pada provinsi yang tergolong dalam klaster dengan kondisi fasilitas pendidikan rendah. Intervensi tersebut dapat diarahkan pada peningkatan rasio satuan pendidikan terhadap jumlah penduduk usia sekolah, pemenuhan fasilitas dasar dan penunjang yang masih terbatas, serta rehabilitasi ruang kelas dengan tingkat kerusakan tinggi pada jenjang SD, SMP, dan SMA.

Selain itu, komponen utama hasil PCA yang paling dominan dalam menjelaskan keragaman data dapat digunakan sebagai indikator prioritas dalam perencanaan dan evaluasi pembangunan fasilitas pendidikan antarprovinsi. Dengan demikian, alokasi anggaran dan program peningkatan infrastruktur pendidikan dapat difokuskan pada variabel-variabel yang secara statistik memiliki kontribusi terbesar terhadap ketimpangan fasilitas pendidikan.

Untuk pengembangan penelitian selanjutnya, disarankan untuk melakukan analisis lanjutan terhadap dinamika perubahan klaster antar tahun guna mengevaluasi efektivitas kebijakan pendidikan, serta membandingkan hasil analisis metode *K-Means* dengan metode klasterisasi lain seperti *Hierarchical Clustering* atau DBSCAN untuk menguji kestabilan struktur klaster yang terbentuk.

5. REFERENSI

- [1] BPS, "Statistika Pendidikan Indonesia 2024," Badan Pusat Statistika , Jakarta, 2024.
- [2] Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, "Indonesia Educational Statistics in Brief 2022/2023," Pusat Data dan Teknologi Informasi, Kemendikbudristek, Jakarta, 2023.

- [3] A. Z. Kamalia dan I. Nawangsih, "Identifikasi Pola Tingkat Kesenjangan Ketuntasan Pendidikan di Indonesia dengan Menggunakan Metode K-Medoids Clustering," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK)*, pp. 312-330, 2025.
- [4] Z. Mustakim dan R. Kamal, "K-means clustering for classifying the quality management of secondary education in Indonesia," *Jurnal Cakrawala Pendidikan*, p. 40(3), 2021.
- [5] N. N. Afidah dan M. , "Penerapan Metode Clustering dengan Algoritma K-Means," *PRISMA, Prosiding Seminar Nasional Matematika* , p. 731, 2023.
- [6] DQLab, "Algoritma K-Means Clustering pada Data Science," 5 Januari 2025. [Online]. Available: <https://dqlab.id/algoritma-k-means-clustering-pada-data-science>.
- [7] IBM, "What is k-means clustering?," 2025. [Online]. Available: <https://www.ibm.com/think/topics/k-means-clustering>.
- [8] Junaidi, "Deskripsi Data Melalui Box-Plot," June 2015. [Online]. Available: https://www.researchgate.net/publication/278022464_Deskripsi_Data_Melalui_Box-Plot.
- [9] R. G. Whendasromo dan J. Joseph, "Analisis Penerapan Normalisasi Data Dengan Menggunakan Z-Score Pada Kinerja Algoritma K-NN," *Jurikom (Jurnal Riset Komputer)*, pp. 872-876, 2022.
- [10] M. Syakur, B. Kurnia dan A. R. Syahputra, "Integration K-means clustering method and elbow method for identification of the best customer profile cluster," *IOP Conference Series: Materials Science and Engineering*, no. 1, p. 336, 2018.