Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 2, Juni 2026
https://doi.org/10.58602/jaiti.v4i2.256
Page 120-131

Penerapan Metode K-Means Clustering dan Principal Component Analysis (PCA) untuk Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan

Clinton Lumbantoruan[1], Dyah Ayu Megawaty[2*]
[1,2]Sistem Informasi, Universitas Teknokrat Indonesia, Bandar Lampung, Indoneisa
[1]Clinton_Lumbantoruan_@teknokrat.ac.id, [2*]dyahayumegawaty@teknokrat.ac.id

Abstrak: Latar belakang studi ini beranjak dari adanya kesenjangan tingkat pendidikan di berbagai wilayah Indonesia yang membutuhkan analisis berbasis data untuk memperoleh gambaran kondisi pendidikan secara lebih sistematis. Penelitian ini bertujuan untuk mengelompokkan provinsi di Indonesia berdasarkan indikator pendidikan menggunakan metode Principal Component Analysis (PCA) dan K-Means Clustering.Penelitian ini bertujuan untuk mengategorikan provinsi-provinsi di Indonesia berdasarkan indikator-indikator pendidikan, dengan harapan memberikan gambaran menyeluruh dan berbasis data mengenai kondisi pendidikan. Latar belakang studi ini beranjak dari adanya kesenjangan tingkat pendidikan di berbagai wilayah yang membutuhkan analisis mendalam dengan pendekatan data mining. Data yang digunakan diperoleh dari Badan Pusat Statistik (BPS) yang mencakup sejumlah indikator pendidikan, seperti rata-rata lama bersekolah, angka partisipasi kasar (APK), angka partisipasi murni (APM), angka partisipasi sekolahan (APS), serta persentase penduduk yang tidak pernah atau belum mengenyam pendidikan. Pendekatan yang diterapkan dalam penelitian ini adalah Principal Component Analysis (PCA) untuk reduksi dimensi dan K-Means Clustering untuk pengelompokan data. Langkah-langkah penelitian mencakup preprocessing data, normalisasi dengan menggunakan StandardScaler, reduksi dimensi melalui PCA, dan clustering dengan K-Means. Temuan dari penelitian ini menunjukkan bahwa dua komponen utama dari PCA mampu menerangkan hingga 79% variasi dalam data, sehingga bermanfaat dalam menyederhanakan dataset yang ada. Proses pengelompokan menghasilkan tiga kelompok provinsi dengan karakteristik pendidikan yang bervariasi, yaitu kategori tinggi, menengah, dan rendah. Penilaian menggunakan silhouette score mengindikasikan bahwa model ini memiliki kualitas pengelompokan yang baik. Temuan dari penelitian ini diharapkan dapat memberikan kontribusi dalam membantu pengambilan kebijakan pendidikan yang lebih tepat serta menjadi landasan untuk penelitian selanjutnya dengan menambahkan variabel yang lebih luas.
Kata Kunci: K-Means; PCA; Clustering; Pendidikan; Data Mining

Abstract: The background of this study stems from the educational disparities across various regions of Indonesia, which require data-driven analysis to obtain a more systematic overview of the state of education. This study aims

Dyah Ayu Megawaty: * Penulis Korespondensi

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 2, Juni 2026
https://doi.org/10.58602/jaiti.v4i2.256
Page 120-131

to cluster Indonesian provinces based on educational indicators using Principal Component Analysis (PCA) and K-Means Clustering. This study aims to categorize provinces in Indonesia based on educational indicators, with the hope of providing a comprehensive, data-driven overview of the state of education. The background of this study stems from the existence of disparities in educational attainment across various regions, which require in-depth analysis using a data mining approach. The data used were obtained from the Central Statistics Agency (BPS) and include a number of educational indicators, such as average years of schooling, gross enrollment rate (GER), net enrollment rate (NER), school enrollment rate (SER), and the percentage of the population that has never attended or has not yet received any education. The approaches applied in this study are Principal Component Analysis (PCA) for dimensionality reduction and K-Means Clustering for data clustering. The research steps include data preprocessing, normalization using StandardScaler, dimensionality reduction via PCA, and clustering with K-Means. The findings of this study indicate that the two principal components of PCA explain up to 79% of the variation in the data, making them useful for simplifying the existing dataset. The clustering process resulted in three groups of provinces with varying educational characteristics: high, medium, and low. An assessment using the silhouette score indicates that this model has good clustering quality. The findings of this study are expected to contribute to more informed educational policy-making and serve as a foundation for future research that incorporates a broader range of variables.
Keywords: K-Means; PCA; Clustering; Education; Data Mining

1. PENDAHULUAN

Pendidikan dianggap sebagai salah satu tolok ukur utama dalam menilai perkembangan suatu negara. Di Indonesia, perbedaan dalam keadaan pendidikan di masing-masing provinsi dipengaruhi oleh berbagai elemen, seperti akses ke fasilitas pendidikan, kondisi ekonomi, dan kebijakan pemerintah daerah. Variasi ini menimbulkan kesenjangan pendidikan yang perlu diteliti secara mendalam agar menjadi pijakan dalam pengambilan keputusan yang tepat. Informasi pendidikan dari Badan Pusat Statistik (BPS) mencakup banyak indikator, termasuk rata-rata lama sekolah, angka partisipasi kasar (APK), angka partisipasi murni (APM), angka partisipasi sekolah (APS), serta persentase penduduk yang tidak atau belum pernah bersekolah[1].

Namun, banyaknya indikator yang digunakan sering kali menambah kompleksitas dalam proses analisis, terutama disebabkan oleh adanya hubungan atau korelasi antara variabel. Oleh karena itu, diperlukan pendekatan yang dapat menyederhanakan data tanpa menghilangkan informasi penting serta mampu mengelompokkan data menurut kesamaan karakteristik. Pendekatan machine learning, khususnya teknik pengelompokan, menjadi alternatif yang tepat untuk mengatasi masalah ini.

Metode K-Means Clustering adalah salah satu algoritma yang banyak diterapkan dalam pengelompokan data karena efisiensinya dalam mengelompokkan data berdasarkan jarak ke centroid. Namun, kinerja K-Means bisa terpengaruh oleh jumlah variabel yang digunakan. Untuk mengatasi masalah ini, Principal Component Analysis (PCA) digunakan sebagai teknik pengurangan dimensi yang dapat mengubah variabel asli menjadi beberapa komponen utama yang lebih representatif. Penelitian ini menerapkan PCA untuk mereduksi data indikator pendidikan sehingga proses pengelompokan menjadi lebih optimal[2].

Beberapa studi sebelumnya menunjukkan bahwa kombinasi PCA dan K-Means dapat meningkatkan kualitas hasil pengelompokan, terutama pada data dengan dimensi tinggi. Dengan penerapan PCA, data dapat direduksi tanpa kehilangan informasi penting, sehingga K-Means dapat beroperasi lebih efektif dalam mengenali pola data[3].

Dyah Ayu Megawaty: * Penulis Korespondensi

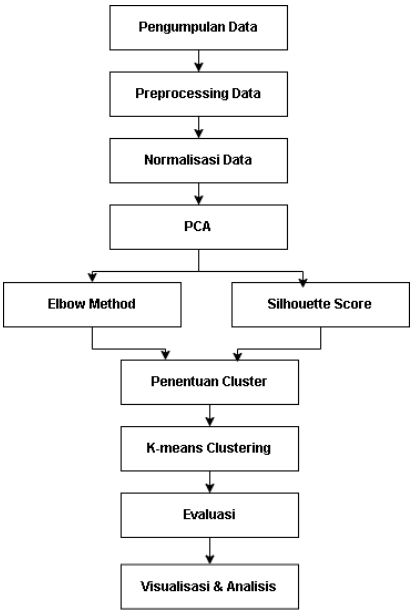
Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 2, Juni 2026
https://doi.org/10.58602/jaiti.v4i2.256
Page 120-131

Penelitian oleh Dewi dan Pakereng menerapkan PCA dan K-Means untuk klasterisasi tingkat pendidikan penduduk Kabupaten Semarang dan menunjukkan bahwa PCA mampu meningkatkan efisiensi clustering[3]. Penelitian Putri dan Hayati menggunakan optimasi PCA pada metode K-Means untuk mengelompokkan kabupaten/kota di Kalimantan berdasarkan indikator pendidikan[4]. Selain itu, Rianti et al. Menerapkan PCA dan clustering dalam analisis mutu pendidikan tinggi dan memperoleh hasil pengelompokan yang cukup baik. Berbeda dengan penelitian sebelumnya, penelitian ini berfokus pada pengelompokan provinsi di Indonesia menggunakan indikator pendidikan yang lebih beragam serta menambahkan evaluasi cluster menggunakan Silhouette Score untuk memperkuat kualitas hasil clustering[1].

Melihat permasalahan tersebut, penelitian ini bertujuan untuk mengelompokkan provinsi di Indonesia berdasarkan indikator pendidikan menggunakan metode PCA dan K-Means Clustering. Di samping itu, penelitian ini ingin memberikan gambaran yang lebih sistematis mengenai kondisi pendidikan antar daerah serta mengidentifikasi kelompok provinsi dengan karakteristik pendidikan yang sejenis. Hasil dari penelitian ini diharapkan mampu berkontribusi dalam mendukung pengambilan keputusan kebijakan pendidikan berbasis data dan menjadi acuan untuk penelitian di masa mendatang[5]. Kontribusi utama penelitian ini terletak pada penerapan kombinasi PCA dan K-Means Clustering untuk memetakan kondisi pendidikan antar provinsi di Indonesia secara lebih sistematis menggunakan indikator pendidikan yang lebih beragam. Penelitian ini juga memberikan evaluasi kualitas cluster menggunakan Silhouette Score sehingga hasil pengelompokan tidak hanya bersifat deskriptif, tetapi juga terukur secara kuantitatif.

2. METODE PENELITIAN

Penelitian ini menggunakan metode Principal Component Analysis (PCA) dan K-Means Clustering untuk mengelompokkan provinsi di Indonesia berdasarkan indikator pendidikan melalui tahapan pengumpulan data, preprocessing, reduksi dimensi, clustering, evaluasi, dan visualisasi hasil.



Gambar 1. Tahapan Penelitian

Langkah-langkah yang diambil dalam penelitian ini mengikuti alur terencana seperti yang dibahas dalam studi-studi sebelumnya, di mana proses dimulai dengan pemrosesan data mentah hingga menghasilkan informasi yang bisa diteliti lebih lanjut dan dapt dilihat

Dyah Ayu Megawaty: * Penulis Korespondensi

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 2, Juni 2026
https://doi.org/10.58602/jaiti.v4i2.256
Page 120-131

pada Gambar 1.

Alur penelitian yang terlihat pada Gambar 1 menggambarkan langkah-langkah yang dilakukan dari tahap pengumpulan data hingga visualisasi dan analisis. Setelah menerapkan reduksi dimensi melalui PCA, langkah selanjutnya adalah menentukan jumlah cluster dengan dua cara, yaitu Metode Elbow dan Skor Siluet. Kedua cara ini dimanfaatkan untuk menemukan jumlah cluster yang maksimal sebelum melanjutkan ke proses pengelompokan dengan algoritma K-Means[6][7].

Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data indikator pendidikan provinsi di Indonesia tahun 2023 yang diperoleh dari Badan Pusat Statistik (BPS). Dataset mencakup beberapa indikator pendidikan, seperti rata-rata lama sekolah, angka partisipasi kasar (APK), angka partisipasi murni (APM), angka partisipasi sekolah (APS), serta persentase penduduk yang tidak atau belum pernah bersekolah. Proses pengumpulan data dilakukan dengan mengunduh dataset dari sumber resmi BPS, kemudian data tersebut disimpan dalam format Comma Separated Values (CSV) agar dapat dengan mudah diproses menggunakan perangkat lunak analisis data. Selanjutnya, dataset yang telah diperoleh diimpor ke dalam lingkungan pemrograman Python melalui Google Colab untuk dilakukan proses pengolahan dan analisis lebih lanjut[8][9].

Pre-processing Data

Tahap preprocessing dilakukan untuk menyiapkan data agar layak digunakan dalam proses analisis. Pada tahap ini dilakukan beberapa langkah, yaitu menghapus data yang tidak valid, seperti baris yang tidak merepresentasikan data sebenarnya, serta menghilangkan kolom yang tidak diperlukan. Selanjutnya, seluruh variabel numerik dikonversi ke dalam format numerik agar dapat diproses secara komputasional[10].

Pengecekan terhadap nilai yang hilang dilakukan pada seluruh variabel. Meskipun tidak ditemukan missing value, proses pengisian nilai menggunakan rata-rata tetap diterapkan sebagai langkah antisipasi. Selain itu, dilakukan deteksi outlier menggunakan visualisasi boxplot untuk mengidentifikasi nilai ekstrem pada data. Outlier yang ditemukan tetap dipertahankan karena masih mencerminkan kondisi nyata dari data pendidikan di setiap provinsi.

Pemisahan Fitur

Data kemudian dipisahkan menjadi dua bagian, yaitu kolom identitas berupa nama provinsi dan variabel numerik sebagai fitur yang digunakan dalam proses analisis. Pemisahan ini bertujuan untuk mempermudah pengolahan data pada tahap berikutnya, khususnya dalam proses normalisasi dan clustering[11].

Normalisasi Data

Normalisasi dilakukan untuk menyamakan skala antar variabel sehingga tidak terjadi dominasi nilai tertentu dalam proses perhitungan jarak. Metode yang digunakan adalah StandardScaler yang mengubah data menjadi memiliki rata-rata nol dan standar deviasi satu. Tahap ini penting karena metode K-Means sangat sensitif terhadap perbedaan skala data[12]. Berikut adalah rumus STANDART SCALER (Normalisasi) yang digunakan :

$$Z = \frac{X - \mu}{\sigma}$$

Dimana:
X = nilai data
μ = rata-rata

Dyah Ayu Megawaty: * Penulis Korespondensi

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 2, Juni 2026
https://doi.org/10.58602/jaiti.v4i2.256
Page 120-131

σ = standar deviasi

Normalisasi ini penting karena metode K-Means menggunakan jarak sebagai dasar pengelompokan, sehingga perbedaan skala dapat menyebabkan bias.

Reduksi Dimensi Menggunakan PCA

Principal Component Analysis (PCA) digunakan untuk mengurangi jumlah variabel dengan tetap mempertahankan informasi utama dalam data. Pada tahap awal dilakukan analisis terhadap cumulative explained variance untuk menentukan jumlah komponen yang optimal. Berdasarkan hasil perhitungan, dua komponen utama dipilih karena mampu menjelaskan sebagian besar variasi data. Kedua komponen tersebut kemudian digunakan sebagai representasi baru dari dataset untuk proses clustering[13][14]. PCA digunakan untuk mengurangi jumlah variabel dengan tetap mempertahankan informasi utama, dengan rumus sebagai berikut :

$$Z = XW$$

Dimana:
X = data asli
W = matriks eigenvector
Z = data hasil transformasi

Penentuan Jumlah Cluster

Penentuan jumlah cluster dilakukan dengan dua pendekatan, yaitu Elbow Method dan Silhouette Score. Elbow Method digunakan untuk melihat perubahan nilai inertia terhadap jumlah cluster, sedangkan Silhouette Score digunakan untuk mengukur kualitas pemisahan antar cluster. Hasil analisis menunjukkan bahwa jumlah cluster optimal berada pada rentang tertentu, dan dalam penelitian ini dipilih tiga cluster untuk menghasilkan pengelompokan yang lebih informatif[15][7].

Clustering Menggunakan K-Means

Proses clustering dilakukan menggunakan algoritma K-Means dengan jumlah cluster yang telah ditentukan sebelumnya. Data yang digunakan sebagai input adalah hasil reduksi PCA. Algoritma akan membagi data ke dalam beberapa kelompok berdasarkan kedekatan terhadap pusat cluster (centroid). Proses iterasi dilakukan hingga posisi centroid stabil dan tidak mengalami perubahan yang signifikan. Metode K-Means diterapkan untuk mengelompokkan informasi berdasarkan seberapa dekatnya dengan pusat data[16][17].

$$J = \sum_{i=1}^{k} \sum x \in ci \parallel x - \mu 2$$

Dimana:
ci = cluster ke-i
$\mu 2$= centroid
x = data

Tujuan metode ini adalah meminimalkan jarak antar data dalam satu cluster.

Evaluasi Model

Evaluasi dilakukan menggunakan Silhouette Score untuk menilai kualitas hasil clustering. Nilai silhouette digunakan untuk mengukur sejauh mana suatu data berada dalam cluster yang tepat dibandingkan dengan cluster lainnya. Nilai yang dihasilkan menunjukkan bahwa model mampu membentuk cluster dengan tingkat pemisahan yang cukup baik. Evaluasi dilakukan menggunakan Silhouette Score[18][19].

Dyah Ayu Megawaty: * Penulis Korespondensi

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 2, Juni 2026
https://doi.org/10.58602/jaiti.v4i2.256
Page 120-131

$$S(i) = \frac{b(i) - a(i)}{\max a(i), b(i)}$$

Dimana:
$a(i)$= jarak dalam cluster
$b(i)$= jarak ke cluster terdekat
Nilai mendekati 1 menunjukkan cluster yang baik.

Visualisasi dan Analisis Hasil

Hasil clustering divisualisasikan menggunakan scatter plot berdasarkan dua komponen utama PCA. Visualisasi ini digunakan untuk melihat pola distribusi data serta pemisahan antar cluster secara visual. Selain itu, dilakukan analisis terhadap rata-rata setiap indikator pada masing-masing cluster untuk mengidentifikasi karakteristik yang membedakan setiap kelompok provinsi[20][21][22].

Hasil clustering divisualisasikan menggunakan scatter plot berdasarkan PC1 dan PC2. Analisis dilakukan dengan melihat:

a. Distribusi cluster
b. Rata-rata tiap indikator
c. Karakteristik masing-masing cluster

Hasil menunjukkan:

a. Cluster tinggi
b. Cluster sedang
c. Cluster rendah

3. HASIL DAN PEMBAHASAN

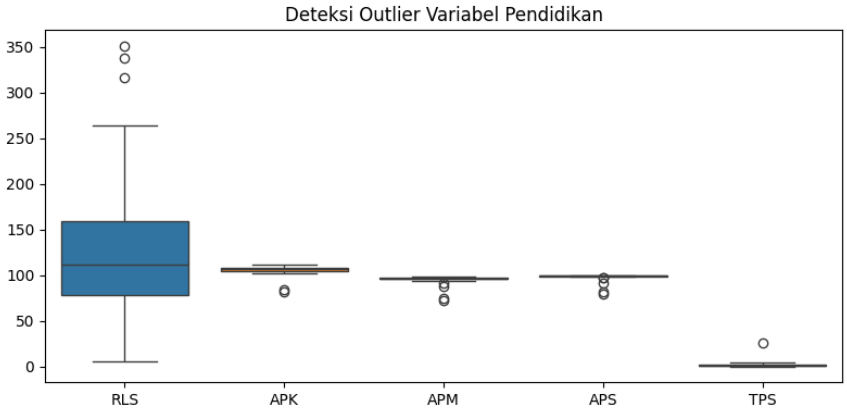
Bab ini menyampaikan hasil dari pengolahan data yang telah dilakukan dengan metode Analisis Komponen Utama (PCA) dan K-Means Clustering. Pembahasan diawali dengan hasil reduksi dimensi melalui PCA yang bertujuan untuk menyederhanakan variabel-variabel, kemudian dilanjutkan dengan penentuan jumlah cluster menggunakan metode Elbow dan Skor Silhouette, serta proses pengelompokan dengan algoritma K-Means. Di samping itu, bab ini juga menyajikan visualisasi dari hasil pengelompokan dan analisis karakteristik setiap cluster untuk memberikan pemahaman tentang kondisi pendidikan di berbagai provinsi di Indonesia dengan cara yang lebih terstruktur.

Pre-Processing Data

Sebelum analisis dilakukan dengan PCA dan K-Means Clustering, langkah prabaca dilakukan untuk memastikan kualitas data yang akan digunakan. Hasil dari prabaca ditunjukkan untuk mengkonfirmasi bahwa data sudah siap untuk dianalisis. Dari hasil pemeriksaan, semua variabel telah berada dalam bentuk numerik dan tidak ada nilai yang hilang. Ini mengindikasikan bahwa dataset yang dipakai sudah utuh dan tidak memerlukan pengisian tambahan.

Dyah Ayu Megawaty: * Penulis Korespondensi

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 2, Juni 2026
https://doi.org/10.58602/jaiti.v4i2.256
Page 120-131

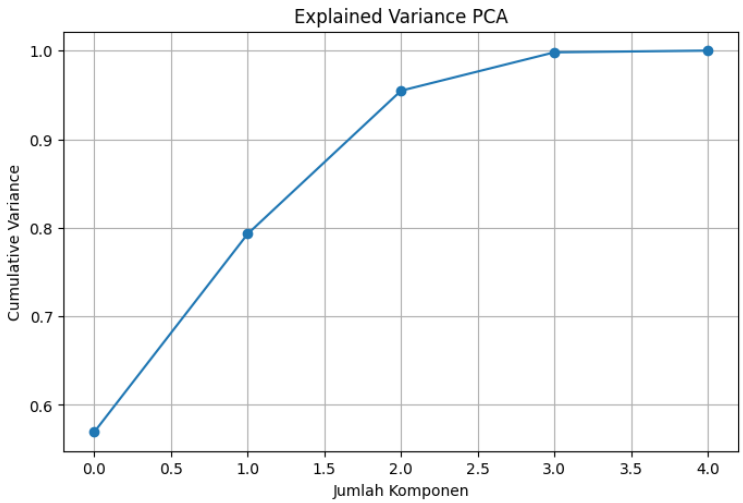


Gambar 3. Deteksi Outlier

Berdasarkan ilustrasi 3, dapat dilihat bahwa ada beberapa nilai unik pada variabel tertentu, terutama pada indikator rata-rata durasi pendidikan. Namun, nilai-nilai ini tetap dipertahankan karena masih mencerminkan keadaan sebenarnya mengenai variasi pendidikan di antara provinsi-provinsi di Indonesia. Maka dari itu, data ini terus digunakan dalam analisis selanjutnya. Hasil dari langkah preprocessing ini mengindikasikan bahwa data kini berada dalam keadaan yang siap untuk dilakukan normalisasi, pengurangan dimensi, dan pengelompokan.

Hasil Reduksi Dimensi (PCA)

Reduksi dimensi dilakukan menggunakan Principal Component Analysis (PCA) untuk menyederhanakan jumlah variabel tanpa menghilangkan informasi penting dalam dataset.



Gambar 4. Explained Variance PCA

Berdasarkan hasil analisis yang dilakukan, diperoleh nilai variasi yang dijelaskan sebagai berikut:

a. Komponen utama pertama (PC1) mencapai 56,95%
b. Komponen utama kedua (PC2) mencapai 22,35%

Total variasi data yang dapat dijelaskan oleh kedua komponen utama tersebut adalah 79,31%. Angka ini menandakan bahwa mayoritas informasi dalam dataset telah berhasil

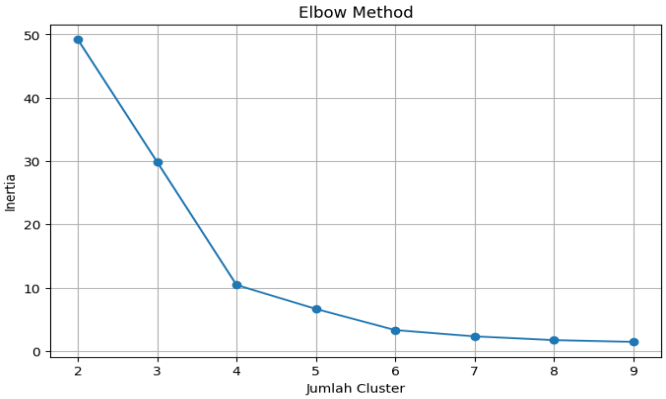
Dyah Ayu Megawaty: * Penulis Korespondensi

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 2, Juni 2026
https://doi.org/10.58602/jaiti.v4i2.256
Page 120-131

direpresentasikan dengan baik melalui kedua komponen utama. Dengan demikian, kedua komponen ini digunakan sebagai dasar dalam proses pengelompokan[23].

Penentuan Jumlah *Cluster*

Penentuan jumlah cluster dilakukan menggunakan metode Elbow dan Silhouette Score.



Gambar 5. *Elbow Method*

Grafik Elbow menggambarkan nilai inersia menunjukkan penurunan yang signifikan sampai pada titik tertentu, setelah itu mulai stabil. Ini menunjukkan bahwa jumlah kelompok yang paling ideal berada di antara 2 sampai 3 kelompok.

Hasil penilaian dengan menggunakan Skor Silhouette menunjukkan:

- k = 2 → 0,85
- k = 3 → 0,54
- k = 4 → 0,52
- k = 5 → 0,52

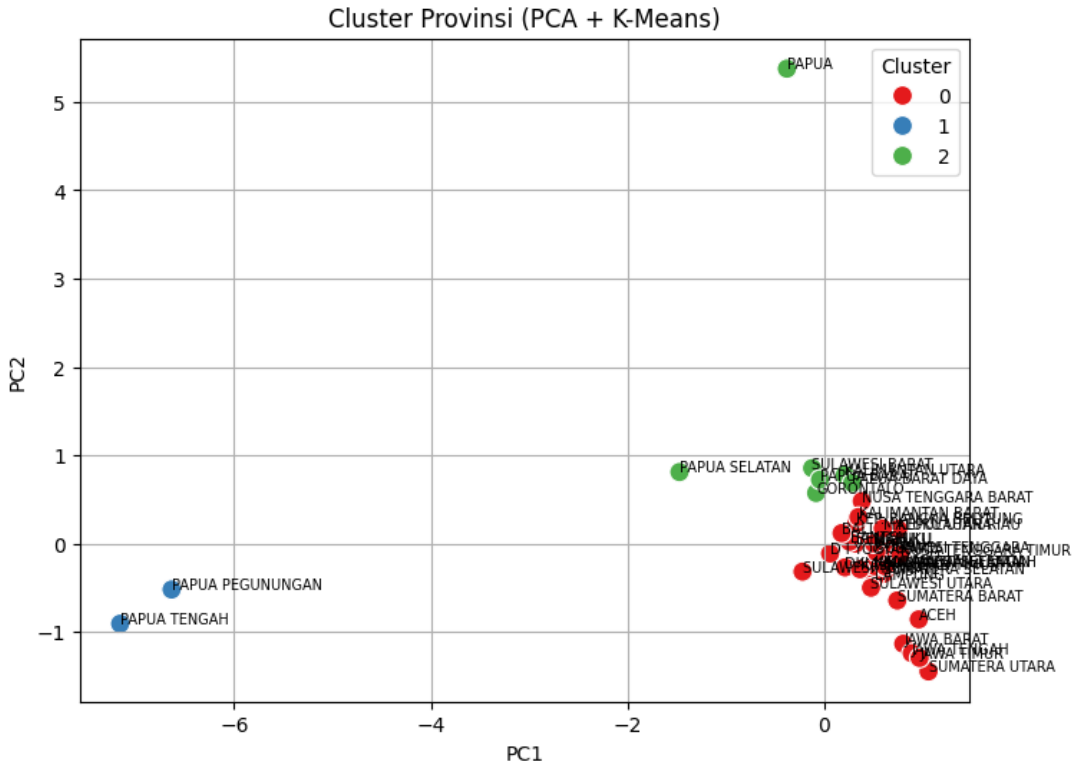
Meskipun nilai silhouette score tertinggi diperoleh pada k=2 sebesar 0,85, penelitian ini menggunakan k=3 dengan pertimbangan interpretatif dan kebutuhan segmentasi data yang lebih detail. Penggunaan tiga cluster memungkinkan data pendidikan provinsi di Indonesia dikategorikan ke dalam kelompok tinggi, sedang, dan rendah, sehingga hasil analisis menjadi lebih informatif dibandingkan hanya dua kelompok. Selain itu, nilai silhouette score sebesar 0,54 pada k=3 masih termasuk dalam kategori cukup baik untuk proses clustering, karena menunjukkan bahwa pemisahan antar cluster masih dapat dibedakan dengan cukup jelas.

Hasil *Clustering K-Means*

Pengelompokan dilakukan melalui algoritma K-Means dengan total tiga kelompok. Data yang dipakai adalah hasil dari reduksi PCA (PC1 dan PC2).

Dyah Ayu Megawaty: * Penulis Korespondensi

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 2, Juni 2026
https://doi.org/10.58602/jaiti.v4i2.256
Page 120-131



Gambar 6. Visualisasi Cluster (PCA + K-Means)

Mengacu pada Gambar 6, dapat dilihat bahwa informasi terbagi menjadi tiga kategori yang cukup nampak. Kebanyakan provinsi tergabung dalam satu kelompok utama, sementara beberapa provinsi lainnya membentuk kelompok lain yang memperlihatkan perbedaan ciri yang nyata.

Analisis Karakteristik Cluster

Analisis dilakukan dengan menghitung rata-rata setiap indikator pendidikan pada masing-masing cluster.

Tabel 1. Rata-rata Indikator Pendidikan Tiap Cluster

Cluster	RLS	APK	APM	APS	TPS
0	156.06	106.23	96.88	99.37	1.62
1	67.26	108.63	93.78	97.42	6.72
2	30.9	83.42	73.7	81.18	2.61

Berdasarkan hasil pengolahan data, diperoleh karakteristik sebagai berikut:

a. Cluster 0 (Kategori Pendidikan Tinggi)
   Rata-rata lama sekolah = 156,06
   APS = 99,37
   Persentase tidak sekolah = 1,61

Cluster ini mencakup sebagian besar provinsi di Indonesia dan menunjukkan kondisi pendidikan yang relatif baik dan merata.

b. Cluster 1 (Kategori Pendidikan Rendah)
   Rata-rata lama sekolah = 30,90
   APS = 81,18

Dyah Ayu Megawaty: * Penulis Korespondensi

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 2, Juni 2026
https://doi.org/10.58602/jaiti.v4i2.256
Page 120-131

Persentase tidak sekolah = 2,60

Cluster ini didominasi oleh wilayah Papua Tengah dan Papua Pegunungan yang memiliki tingkat pendidikan paling rendah dibandingkan wilayah lainnya.

c. Cluster 2 (Kategori Pendidikan Sedang)
Rata-rata lama sekolah ≈ 67,25
APS ≈ 97,42
Persentase tidak sekolah ≈ 6,71

Cluster ini menunjukkan kondisi pendidikan yang berada pada tahap transisi antara rendah dan tinggi.

Interpretasi Hasil

Hasil pengelompokan dengan metode K-Means Clustering yang dilengkapi dengan reduksi dimensi PCA memperlihatkan adanya perbedaan karakteristik yang jelas antara provinsi-provinsi di Indonesia berdasar pada indikator pendidikan. Penerapan PCA yang menggunakan dua komponen utama yang menjelaskan 79,31% variasi data menunjukkan bahwa proses reduksi dimensi berhasil menjaga informasi inti dari dataset, sehingga hasil pengelompokan dapat dipahami secara representatif.

Secara keseluruhan, pengelompokan ini membagi provinsi-provinsi di Indonesia menjadi tiga kelompok utama, yaitu kluster dengan tingkat pendidikan tinggi, sedang, dan rendah. Perbedaan antar kluster terlihat jelas baik dari visualisasi yang berbasis PCA maupun dari rata-rata setiap indikator pendidikan.

Kluster 0 adalah kelompok dengan karakteristik pendidikan yang cenderung tinggi dan mencakup sebagian besar provinsi di Indonesia. Ini tergambar dari rata-rata lama sekolah yang tinggi, angka partisipasi sekolah (APS) yang hampir maksimal, serta persentase penduduk yang tidak atau belum pernah sekolah yang relatif rendah. Situasi ini menunjukkan bahwa provinsi dalam kluster ini memiliki akses pendidikan yang baik dan sistem pendidikan yang cukup merata. Selain itu, tingginya nilai APM dan APK menunjukkan bahwa tingkat partisipasi dalam pendidikan formal di kelompok ini sudah optimal.

Kluster 1 mencerminkan kelompok dengan kondisi pendidikan yang paling rendah dibandingkan kluster lainnya. Wilayah ini sebagian besar terdiri dari Papua Tengah dan Papua Pegunungan. Rata-rata lama sekolah yang sangat rendah serta angka partisipasi sekolah yang jauh di bawah kluster lainnya menunjukkan adanya hambatan dalam akses dan kualitas pendidikan. Selain itu, proporsi penduduk yang tidak atau belum pernah sekolah masih cukup tinggi, yang menandakan adanya tantangan serius dalam pemerataan pendidikan. Faktor-faktor geografis, infrastruktur, dan distribusi tenaga pendidik bisa jadi penyebab utama dari kondisi tersebut.

Kluster 2 berada di kategori menengah atau dalam tahap transisi. Kelompok ini memperlihatkan nilai indikator pendidikan yang cukup baik, tetapi belum mencapai tingkat optimal seperti pada kluster 0. Walaupun angka partisipasi sekolah relatif tinggi, persentase penduduk yang tidak atau belum pernah sekolah masih lebih tinggi dibandingkan kluster 0. Hal ini menunjukkan bahwa meskipun akses pendidikan telah tersedia, masih ada tantangan dalam pemerataan dan keberlanjutan pendidikan di kawasan tersebut.

Hasil penelitian ini juga memperlihatkan adanya perbedaan pendidikan yang cukup signifikan di berbagai daerah di Indonesia. Provinsi yang berada dalam kelompok rendah dan menengah membutuhkan perhatian lebih dalam hal peningkatan akses pendidikan, pengembangan infrastruktur, serta peningkatan mutu tenaga pengajar. Sementara itu, provinsi yang termasuk dalam kelompok tinggi dapat dijadikan sebagai model dalam pengembangan kebijakan pendidikan nasional.

Dyah Ayu Megawaty: * Penulis Korespondensi

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 2, Juni 2026
https://doi.org/10.58602/jaiti.v4i2.256
Page 120-131

Dalam aspek metodologi, gabungan PCA dan K-Means Clustering terbukti sangat berguna dalam mengelompokkan data dengan banyak variabel. PCA berhasil mengurangi kompleksitas data tanpa kehilangan informasi penting, sedangkan K-Means efektif dalam menemukan pola kesamaan antar provinsi. Visualisasi yang didasarkan pada PCA juga memperkuat hasil analisis dengan memperlihatkan pemisahan cluster yang cukup jelas.

Dengan kata lain, hasil dari studi ini tidak hanya memberikan penjelasan mengenai keadaan pendidikan secara deskriptif, tetapi juga dapat mengidentifikasi pola dan kelompok daerah yang memiliki karakteristik yang sama. Ini dapat dijadikan pijakan dalam pembuatan kebijakan pendidikan yang lebih terarah, berbasis data, dan sesuai dengan kebutuhan spesifik setiap wilayah.

4. KESIMPULAN

Penelitian ini berhasil menerapkan metode Principal Component Analysis (PCA) dan K-Means Clustering untuk mengelompokkan provinsi di Indonesia berdasarkan indikator pendidikan tahun 2023 yang diperoleh dari Badan Pusat Statistik (BPS). Hasil reduksi dimensi menggunakan PCA mampu mempertahankan 79,31% variasi data melalui dua komponen utama, sehingga proses clustering dapat dilakukan dengan lebih efisien tanpa menghilangkan informasi penting. Berdasarkan hasil clustering, provinsi di Indonesia terbagi menjadi tiga kelompok utama, yaitu kategori pendidikan tinggi, sedang, dan rendah. Evaluasi menggunakan Silhouette Score menunjukkan bahwa model memiliki kualitas cluster yang cukup baik, dimana nilai silhouette sebesar 0,54 pada k=3 masih mampu menunjukkan pemisahan cluster yang jelas dan lebih informatif dibandingkan penggunaan dua cluster. Penelitian ini memberikan kontribusi dalam pemetaan kondisi pendidikan berbasis data yang dapat digunakan sebagai pendukung pengambilan kebijakan pendidikan yang lebih terarah, meskipun penelitian masih memiliki keterbatasan pada jumlah indikator dan penggunaan data dalam satu periode waktu.

5. REFERENCES

[1] R. Rianti *et al.*, "Penerapan PCA dan Algoritma Clustering untuk Analisis Mutu Perguruan Tinggi di LLDIKTI Wilayah IV," vol. 18, 2024.
[2] I. A. Naya, "Penerapan Algoritma K- means Cliustering Untuk Segmentasi Penyakit Daun Mangga," 2025.
[3] S. Dewi and M. A. I. Pakereng, "Implementasi principal component analysis pada k-means untuk klasterisasi tingkat pendidikan penduduk kabupaten semarang," vol. 8, no. 4, pp. 1186–1195, 2023.
[4] F. A. Statistik, M. Nur, and H. Tri, "Klasterisasi Kabupaten / Kota di Provinsi Kalimantan Barat Berdasarkan Rasio Guru dengan Murid di Tingkat SD , SMA , dan SMA Serta IPM Tahun 2022 Menggunakan Metode K-Means Clustering," vol. 3, no. 1, pp. 42–50, 2023.
[5] N. S. Putri and M. N. Hayati, "Pengelompokan Kabupaten/Kota di Kalimantan Berdasarkan Indikator Pendidikan Menggunakan Metode K-Means dengan Optimasi Principal Component Analysis Grouping Regencies/Cities in Kalimantan Based on Educational Indicators Used K-Means Method with Principal Component Analysis Optimization," vol. 15, no. November, pp. 128–138, 2024, doi: 10.30872/eksponensial.v15i2.1373.
[6] G. W. and T. H. Robert Tibshirani, "Estimating the number of clusters in a data set via the gap statistic."
[7] K. Dbscan, "Analisis Perbandingan Silhouette dengan Elbow pada Algoritma," 2025, doi: 10.47002/metik.v9i1.1027.
[8] Y. Dalva, H. Pehlivan, O. I. Hatipoglu, C. Moran, and A. Dundar, "Image-to-Image Translation with Disentangled Latent Vectors for Face Editing," pp. 1–12.

Dyah Ayu Megawaty: * Penulis Korespondensi

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 2, Juni 2026
https://doi.org/10.58602/jaiti.v4i2.256
Page 120-131

[9] F. Susy, "Unflavored Leptogenesis and Neutrino Masses in," no. 5.
[10] P. Martinez-azcona and A. Kundu, "Stochastic Operator Variance: an observable to diagnose noise and scrambling," pp. 1–14.
[11] L. Angeles, "Estimating Ejecta Masses of Stripped Envelope Supernovae Using Late-Time Light Curves," 2023.
[12] R. Molina, N. Ikeno, and E. Oset, "Reply to 'Comment on Sequential Single-pion Production Explaining the dibaryon d ∗ (2380) peak,'" no. 2380, pp. 1–4, 2023.
[13] R. Str, M. Schaller, K. Worthmann, J. Berberich, and F. Allg, "SafEDMD : A Koopman-based data-driven controller design framework for nonlinear dynamical systems ⋆," no. 2021, 2022.
[14] A. Mandal, "A characterization of orthogonal permutative matrices of".
[15] R. Gong *et al.*, "ARNOLD : A Benchmark for Language-Grounded Task Learning With Continuous States in Realistic 3D Scenes".
[16] M. Jourdan, E. Kaufmann, M. Jourdan, and I. Fr, "Dealing with Unknown Variances in Best-Arm Identification," vol. 201, pp. 1–74, 2023.
[17] N. N. Putriwijaya and N. Yudistira, "Learning-Augmented K-Means Clustering Using Dimensional Reduction," pp. 1–17, 2021.
[18] R. Capuani and O. C. Feb, "a pursuit evasion-like game," pp. 1–26, 2023.
[19] C. St, "JPerceiver: Joint Perception Network for Depth, Pose and Layout Estimation in Driving Scenes," pp. 1–26, 2008.
[20] J. Shah, G. Nambiar, A. V Gorshkov, and V. Galitski, "Quantum spin ice in three-dimensional Rydberg atom arrays," no. 1, pp. 1–33, 2024.
[21] X. Ma, J. Zhang, X. Feng, and C. Zhang, "perturbed convection diffusion problem on Shishkin," pp. 1–33.
[22] N. Ramsey, "Some symbolic dynamics in real quadratic fields with applications to inhomogeneous minima," vol. 1, pp. 1–10.
[23] N. Roth and A. L. Goodwin, "disordered crystals Pre-print , second version May 17th 2023," pp. 1–34, 2023.

Dyah Ayu Megawaty: * Penulis Korespondensi