JURNAL FASILKOM
Volume 16 No. 1 | April 2026: 105-112

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Analisis Pemerataan Pendidikan di Indonesia Menggunakan Reduksi Dimensi PCA dan Klasterisasi *K-Means*

Erwin Arry Kusuma[1], Adani Dharmawati[2]
[1]Teknik Informatika, STMIK Banjarbaru
[2]Teknik Informatika, Fakultas Teknologi Informasi, Universitas Islam Kalimantan MAB
[1]erwinarry@gmail.com*, [2]adani.dharmawati@gmail.com

Abstract

Educational equity in Indonesia continues to face substantial challenges due to significant disparities in achievement across provinces. This study aims to map these gaps by combining Principal Component Analysis (PCA) for dimensionality reduction and K-Means Clustering for regional grouping. Utilizing 2023 data from the Indonesian Central Bureau of Statistics (BPS) with eight key indicators, the analysis reveals that three principal components effectively capture 91.85% of the data variance. The clustering procedure successfully categorizes provinces into two distinct groups: 36 provinces in the high-achievement cluster and two provinces that lag significantly (Central Papua and Papua Mountains). A Silhouette Score of 0.782 confirms the high validity and consistency of the clustering results. These findings serve as a critical alert for policymakers to implement targeted interventions in underperforming regions to prevent further widening of the educational gap.

Keywords: educational equity, k-means clustering, principal component analysis

Abstrak

Pemerataan pendidikan di Indonesia masih terganjal disparitas capaian antarprovinsi yang cukup lebar. Penelitian ini bertujuan untuk memetakan kesenjangan tersebut dengan mengombinasikan *Principal Component Analysis* (PCA) untuk reduksi dimensi dan *K-Means Clustering* untuk pengelompokan wilayah. Menggunakan data Badan Pusat Statistik (BPS) tahun 2023 dengan delapan indikator utama, hasil analisis menunjukkan bahwa tiga komponen utama mampu menangkap 91,85% varians data secara efektif. Prosedur klasterisasi berhasil mengelompokkan provinsi menjadi dua grup yang berbeda: 36 provinsi masuk dalam klaster capaian tinggi dan dua provinsi (Papua Tengah dan Papua Pegunungan) tertinggal secara signifikan. Skor *Silhouette* sebesar 0,782 mengonfirmasi validitas dan konsistensi hasil klasterisasi yang sangat kuat. Temuan ini menjadi peringatan bagi pembuat kebijakan untuk melakukan intervensi target di wilayah yang tertinggal guna mencegah semakin lebarnya jurang pendidikan nasional.

Kata kunci: pemerataan pendidikan, *k-means clustering, principal component analysis*

1. Pendahuluan

Pemerataan akses dan kualitas pendidikan hingga kini masih menjadi persoalan fundamental yang dihadapi oleh bangsa Indonesia dalam menyongsong era industri 4.0. Sebagai instrumen utama dalam pembangunan nasional, kualitas pendidikan berkorelasi langsung dengan kualitas daya saing sumber daya manusia di kancah global. Namun, kenyataannya peningkatan indeks pembangunan manusia di seluruh pelosok negeri belum berjalan secara merata. Hal ini menjadi tantangan besar dalam upaya meningkatkan indeks pembangunan manusia di seluruh pelosok negeri secara merata. Sebagai negara kepulauan dengan karakteristik geografis yang kompleks, kesenjangan capaian pendidikan antarwilayah seringkali tidak terelakkan [1]. Kondisi topografi yang sulit dan sebaran penduduk yang tidak merata menciptakan hambatan tersendiri dalam distribusi fasilitas pendidikan. Perbedaan infrastruktur dan aksesibilitas antar pulau menjadi faktor utama sulitnya standarisasi mutu pendidikan secara nasional [2]. Akibatnya, daerah-daerah yang berada di zona pusat pertumbuhan ekonomi cenderung memiliki fasilitas yang jauh lebih mapan dibandingkan wilayah terpencil atau daerah perbatasan. Tantangan logistik dan perbedaan ketersediaan sarana pendukung di wilayah terpencil seringkali membuat kualitas fasilitas pendidikan tertinggal jauh dibandingkan dengan wilayah perkotaan. Ketidakseimbangan ini terlihat jelas pada sarana fisik sekolah, ketersediaan tenaga pendidik, hingga akses terhadap teknologi. Berdasarkan data terbaru dari Badan Pusat Statistik (BPS) tahun 2023, berbagai indikator capaian pendidikan masih menunjukkan variasi yang signifikan antarprovinsi [3]. Data tersebut bukan sekadar angka, melainkan cerminan adanya ketimpangan nyata dalam hal daya saing sumber daya manusia di tingkat regional yang jika dibiarkan akan memperlebar jurang ekonomi antarwilayah. Masalah utama yang dihadapi adalah sulitnya melakukan analisis pada data multidimensi yang memiliki rentang nilai berbeda-beda tanpa menimbulkan bias [4]. Sifat data pendidikan yang heterogen dimulai dari angka partisipasi hingga tingkat buta aksara, memerlukan ketelitian dalam pengolahannya. Kumpulan data yang besar dan beragam ini menuntut metode pengolahan yang tepat agar informasi yang dihasilkan benar-benar akurat serta mampu menjadi basis pengambilan keputusan yang valid. Ketidakmampuan dalam membaca pola data

P-ISSN : 2089-3353
E-ISSN : 2808-9162

secara utuh dapat menyebabkan salah sasaran dalam penyaluran bantuan atau program pemerintah.. Disparitas ini jika tidak dipetakan secara akurat akan menghambat efektivitas kebijakan pemerintah dalam melakukan intervensi [5].

Oleh karena itu, diperlukan sebuah pendekatan komputasional yang mampu melakukan pengelompokan wilayah secara objektif dan berbasis data (*data-driven*). Metode statistik konvensional seringkali mengalami keterbatasan ketika harus berhadapan dengan variabel yang sangat banyak dan memiliki korelasi yang rumit. Dalam beberapa tahun terakhir, teknik pengolahan data atau *data mining* telah banyak diandalkan untuk memecahkan masalah segmentasi sosial yang kompleks [6]. Salah satu keunggulan utama dari teknik ini adalah kemampuannya dalam memproses data secara masif tanpa prasangka subjektif dari peneliti. Pendekatan komputasional ini memungkinkan peneliti untuk menemukan pola tersembunyi (*hidden pattern*) dari sekumpulan data yang sangat besar yang mungkin tidak terlihat melalui observasi manual. Dalam ranah klasterisasi, salah satu metode yang paling populer dan teruji efektivitasnya adalah algoritma K-Means [7]. Algoritma ini bekerja dengan cara mempartisi data ke dalam beberapa kelompok berdasarkan kemiripan karakteristiknya. Meskipun dikenal efisien dalam mengolah data dalam jumlah besar, algoritma ini memiliki keterbatasan tertentu yang perlu diwaspadai. Kelemahan utamanya terletak pada cara algoritma menghitung jarak antar objek. Metode ini dipilih karena kemampuannya dalam melakukan pengelompokan secara cepat dan efisien pada berbagai jenis data. Meskipun dikenal efisien, K-Means memiliki sensitivitas yang sangat tinggi terhadap skala fitur data, sehingga memerlukan tahapan pra-pemrosesan yang tepat sebelum dijalankan [8]. Sebagai contoh, variabel dengan satuan jutaan akan sangat mendominasi variabel yang bersatuan persentase dalam perhitungan jarak Euclidean. Tanpa proses normalisasi atau standarisasi yang baik, variabel dengan skala angka yang besar akan mendominasi hasil klasterisasi secara tidak adil dan menghasilkan pengelompokan yang bias. Tantangan teknis lainnya dalam pemetaan pendidikan muncul ketika data yang dianalisis bersifat multidimensi dan memiliki korelasi antarvariabel yang tinggi (multikolinieritas), yang seringkali justru menimbulkan *noise* dalam proses klasterisasi. Dalam data pendidikan, variabel seperti Angka Partisipasi Sekolah dan Harapan Lama Sekolah seringkali saling berkaitan erat, yang jika langsung diproses dapat menimbulkan *noise* dan redundansi informasi. Adanya redundansi informasi antar-indikator dapat menyebabkan hasil pengelompokan menjadi kurang stabil dan sulit untuk diinterpretasikan secara tajam. Untuk mengatasi batasan tersebut, integrasi metode reduksi dimensi menjadi sangat krusial. Penggunaan *Principal Component Analysis* (PCA) menjadi solusi strategis untuk menyederhanakan variabel asli menjadi set variabel baru yang tidak saling berkorelasi tanpa menghilangkan esensi informasi penting di dalamnya [9,10]. Dengan PCA, dimensi data yang kompleks dapat diringkas sehingga proses klasterisasi menjadi lebih ringan dan lebih akurat dalam menangkap variansi data.

Selain pemilihan algoritma inti, keberhasilan sebuah model pemetaan wilayah juga sangat bergantung pada penentuan parameter model yang optimal. Banyak penelitian sebelumnya hanya berfokus pada hasil klasterisasi tanpa mempertimbangkan validitas internal dari klaster yang terbentuk. Ketepatan hasil klasterisasi sangat ditentukan oleh seberapa baik peneliti menentukan jumlah kelompok atau nilai *k* yang paling ideal. Penggunaan metode *Elbow* seringkali menjadi standar industi untuk menentukan jumlah klaster terbaik melalui analisis grafik *Sum of Squared Errors* (SSE) guna meminimalkan subjektivitas peneliti. Dengan metode ini, titik potong siku pada grafik menjadi indikator jumlah klaster paling ideal bagi data yang sedang diuji. Di sisi lain, standarisasi data melalui *StandardScaler* (Z-Score) menjadi langkah wajib untuk memastikan setiap indikator memiliki bobot yang setara dalam perhitungan jarak. Hal ini penting agar perbedaan satuan antar indikator, seperti persentase dan tahun, tidak mengacaukan perhitungan matematis algoritma dan memastikan setiap variabel berkontribusi secara proporsional dalam membentuk profil wilayah. Implementasi gabungan antara PCA dan K-Means ini telah divalidasi dalam berbagai studi sebagai pendekatan yang efisien untuk evaluasi data kompleks. Namun, penelitian ini membawa beberapa nilai kebaruan yang membedakannya dengan studi-studi terdahulu.

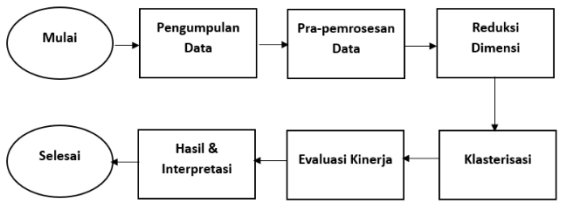
Penelitian ini bertujuan untuk memetakan 38 provinsi di Indonesia berdasarkan delapan indikator pendidikan utama menggunakan kombinasi PCA dan K-Means. Fokus utama penelitian diarahkan pada data tahun 2023 yang sudah mencakup seluruh wilayah Indonesia, termasuk empat provinsi baru hasil pemekaran di wilayah Papua (Papua Tengah, Papua Pegunungan, Papua Selatan, dan Papua Barat Daya). Berbeda dengan studi sebelumnya yang seringkali hanya menggunakan satu metrik validasi, penelitian ini menerapkan evaluasi ganda melalui *Silhouette Score* untuk mengukur kerapatan klaster dan *Davies-Bouldin Index* untuk menilai pemisahan antar klaster. Penggunaan dua metrik evaluasi ini bertujuan untuk memberikan tingkat kepercayaan yang lebih tinggi terhadap validitas hasil pengelompokan. Hasil penelitian ini diharapkan dapat menjadi rujukan strategis bagi pengambil kebijakan dalam menentukan prioritas intervensi pendidikan di wilayah yang terdeteksi tertinggal. Dengan demikian, distribusi sumber daya pendidikan dapat dilakukan secara lebih efektif dan tepat sasaran.

2. Metode Penelitian

Penelitian ini menggunakan kerangka kerja *Knowledge Discovery in Databases* (KDD) yang dilakukan secara sistematis. Penggunaan kerangka KDD bertujuan untuk

P-ISSN : 2089-3353
E-ISSN : 2808-9162

memastikan bahwa proses ekstraksi pengetahuan dari data indikator pendidikan dilakukan melalui tahapan yang terstruktur dan tervalidasi, mulai dari seleksi data hingga tahap interpretasi pola yang ditemukan. Tahapan penelitian secara keseluruhan dapat dilihat pada Gambar 1 berikut.



Gambar 1. Alur Tahapan Penelitian

2.1. Data dan Sumber Data

Objek penelitian ini mencakup 38 provinsi di Indonesia dengan menggunakan data sekunder yang dirilis oleh Badan Pusat Statistik (BPS) pada tahun 2023. Penggunaan data terbaru ini sangat penting untuk memberikan gambaran kondisi riil pendidikan pasca-pandemi di seluruh wilayah Indonesia, termasuk daerah otonomi baru (DOB) di wilayah Papua yang memiliki tantangan geografis unik. Pemilihan variabel ini didasarkan pada standar penilaian pemerataan pendidikan nasional yang telah lazim digunakan dalam studi-studi sektoral [11]. Data sekunder dari BPS digunakan untuk memastikan validitas informasi yang diolah karena telah melalui proses verifikasi data tingkat nasional [12]. Penentuan indikator yang relevan sangat menentukan akurasi hasil pengelompokan yang akan dihasilkan. Indikator yang dipilih dalam penelitian ini mencakup dimensi partisipasi sekolah dan kualitas literasi guna memberikan gambaran komprehensif mengenai daya saing SDM regional. Detail indikator yang digunakan dalam penelitian ini dirangkum dalam Tabel 1.

Tabel 1. Indikator Pendidikan

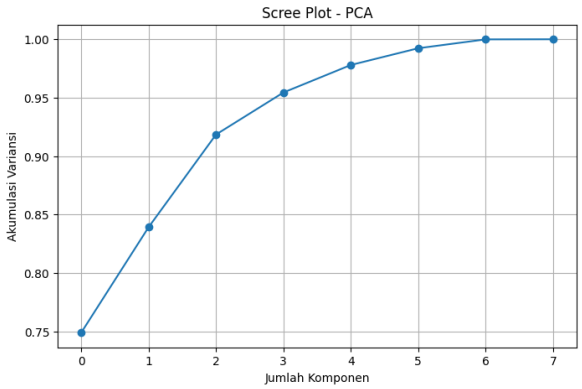
Nomor	Indikator
1	Rata-rata lama sekolah penduduk umur 15+ (tahun)
2	APK SD (%)
3	APK SMP (%)
4	APK SMA (%)
5	Harapan lama sekolah (tahun)
6	Angka buta aksara usia 15+ (%)
7	Angka buta aksara usia 15–44 (%)
8	Angka buta aksara usia 45+ (%)

2.2. Pra-pemrosesan Data

Tahap awal yang dilakukan adalah pembersihan data (*data cleaning*) untuk memastikan tidak terdapat nilai yang hilang (*missing values*). Keberadaan data yang kosong atau tidak konsisten dapat menyebabkan bias yang signifikan pada hasil klasterisasi karena algoritma berbasis jarak sangat sensitif terhadap ketimpangan data. Mengingat variabel yang digunakan memiliki satuan dan rentang nilai yang sangat kontras seperti Angka Partisipasi Kasar (APK) yang bernilai ratusan dibandingkan dengan indikator Rata-rata Lama Sekolah yang berada pada kisaran angka satuan, penelitian ini menggunakan teknik *StandardScaler* untuk mengubah data menjadi nilai *Z-Score* dengan rata-rata 0 dan variansi 1 [13]. Transformasi ini dilakukan dengan mengurangkan nilai asli dengan rata-rata kemudian membaginya dengan standar deviasi. Langkah ini sangat krusial agar variabel dengan skala besar tidak mendominasi perhitungan jarak dalam algoritma klasterisasi secara tidak proporsional [14]. Tanpa standarisasi, indikator dengan rentang nilai ratusan seperti APK akan dianggap lebih penting daripada indikator dengan rentang nilai satuan.

2.3. Reduksi Dimensi dengan PCA

Data yang telah distandarisasi kemudian diproses menggunakan *Principal Component Analysis* (PCA). Tujuan utama dari PCA adalah untuk mereduksi dimensi data tanpa menghilangkan variansi informasi yang signifikan [15]. Penggunaan PCA menjadi solusi strategis untuk mengatasi masalah multikolinieritas, di mana beberapa indikator pendidikan seringkali memiliki korelasi yang sangat kuat satu sama lain. Dengan teknik ini, delapan indikator pendidikan yang awalnya saling memiliki korelasi akan ditransformasi menjadi komponen utama baru (*Principal Components*) yang bersifat independen satu sama lain.



Gambar 2. Grafik Scree Plot PCA

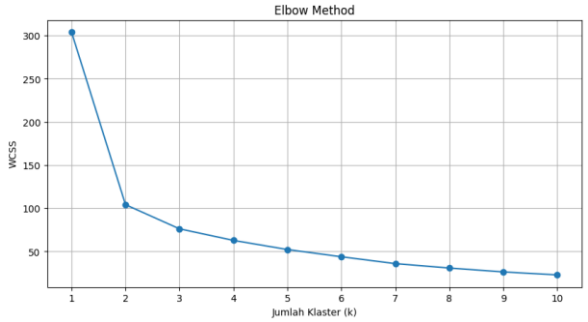
Pemilihan jumlah komponen dilakukan dengan melihat nilai *eigenvalue* di atas satu atau berdasarkan persentase varians kumulatif yang dihasilkan [16] [17]. Hal ini memungkinkan penyederhanaan model tanpa kehilangan karakteristik penting dari dataset asli. Secara matematis, transformasi variabel asal menjadi komponen utama dapat dirumuskan sebagai berikut:

$$det(C - \lambda I) = 0 \qquad (1)$$

P-ISSN : 2089-3353
E-ISSN : 2808-9162

2.4. Klasterisasi K-Means dan Evaluasi

Tahap akhir adalah pengelompokan provinsi menggunakan algoritma K-Means yang diterapkan pada skor hasil PCA. Algoritma ini bekerja dengan meminimalkan jarak antara data dengan pusat klaster (*centroid*) yang dipilih secara iteratif hingga mencapai titik konvergensi.



Gambar 3. Penentuan Jumlah Klaster dengan Metode Elbow

Proses minimalisasi ini dilakukan dengan mengoptimalkan fungsi objektif *Sum of Squared Errors* (SSE) sebagai berikut:

$$J = \sum_{j=1}^{k} \cdot \sum_{i \in S_j} \cdot \left|\left| x_i - c_j \right|\right| \text{^}2 \qquad (2)$$

Untuk menentukan jumlah klaster (k) yang paling representatif, digunakan *Elbow Method* dengan menganalisis titik kelandaian pada kurva *Within-Cluster Sum of Squares* (WCSS) [18].

Pemilihan nilai k yang tepat sangat menentukan interpretasi profil wilayah yang akan terbentuk. Selain metode Elbow, kualitas dari hasil pengelompokan divalidasi menggunakan *Silhouette Coefficient* untuk mengukur derajat kepadatan klaster dan pemisahan antar objek [19]. Terakhir, digunakan *Davies-Bouldin Index* (DBI) sebagai metrik evaluasi tambahan untuk menilai kekompakan di dalam klaster dan pemisahan antar klaster, di mana nilai DBI yang semakin kecil menunjukkan hasil klasterisasi yang semakin baik karena menandakan tingkat tumpang tindih antar kelompok yang rendah [20].

3. Hasil dan Pembahasan

Bagian ini menguraikan temuan eksperimen dari integrasi metode PCA dan K-Means dalam memetakan profil pendidikan di Indonesia berdasarkan data BPS 2023. Analisis dilakukan secara bertahap, mulai dari reduksi dimensi hingga interpretasi mendalam terhadap klaster yang terbentuk. Pendekatan ini dipilih untuk memastikan bahwa pengelompokan tidak hanya didasarkan pada kemiripan angka, tetapi juga pada variansi informasi yang paling menentukan perbedaan kualitas antarwilayah.

3.1. Analisis Reduksi Dimensi (PCA)

Proses awal dimulai dengan mereduksi delapan indikator pendidikan guna mengeliminasi redundansi data. Mengingat banyak indikator pendidikan yang saling berkorelasi seperti Harapan Lama Sekolah dengan Angka Partisipasi Sekolah maka reduksi dimensi menjadi langkah krusial agar model tidak mengalami *overfitting*. Berdasarkan hasil ekstraksi matriks kovarians, terbentuk tiga komponen utama (PC1, PC2, dan PC3) yang memiliki nilai *eigenvalue* di atas 0,6. Meskipun secara teori standar *Kaiser* menyarankan *eigenvalue* di atas 1, namun dalam penelitian ini nilai 0,6 diambil untuk memastikan informasi spesifik mengenai buta aksara kelompok usia tertentu tetap terjaga. Ketiga komponen ini mampu merangkum hingga 91,85% total informasi dari dataset asli.

Tingginya angka kumulatif varians ini menjamin bahwa representasi data dalam dimensi rendah tidak menghilangkan esensi dari kondisi pendidikan riil di tiap provinsi. Hal ini membuktikan bahwa penyusutan dimensi dari delapan variabel menjadi tiga komponen saja sudah cukup untuk menggambarkan dinamika pendidikan nasional tanpa kehilangan detail signifikan. Detail hasil ekstraksi komponen disajikan pada Tabel 2.

Tabel 2. Hasil Ekstraksi Komponen Utama (PCA)

Komponen	Eigenvalue	Varians (%)	Kumulatif (%)
PC1	6,153	74,89	74,89
PC2	0,742	9,03	83,92
PC3	0,651	7,93	91,85

PC1 merepresentasikan kualitas pendidikan umum secara luas, sementara PC2 dan PC3 menangkap variasi spesifik pada angka buta aksara dewasa. PC1 memiliki kontribusi dominan sebesar 74,89%, yang menunjukkan bahwa sebagian besar perbedaan antarprovinsi ditentukan oleh indikator-indikator utama seperti lama sekolah dan APK. Dominasi PC1 ini mengindikasikan bahwa struktur data pendidikan di Indonesia memiliki pola searah, di mana wilayah yang unggul di satu indikator partisipasi cenderung unggul di indikator lainnya. Dengan mempertahankan >90% informasi, analisis klaster selanjutnya dipastikan tetap memiliki akurasi tinggi meski dimensinya telah disederhanakan.

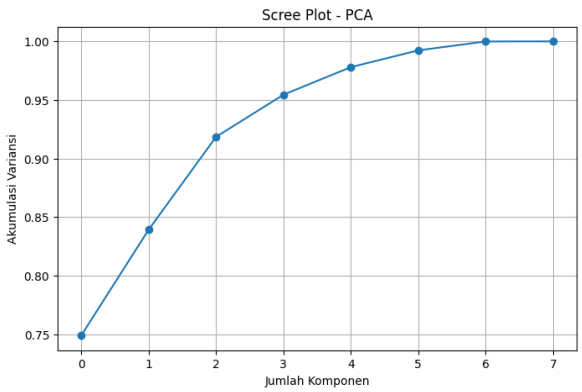
P-ISSN : 2089-3353
E-ISSN : 2808-9162

Tabel 3. Matriks Komponen (Loading Factors) PCA

Indikator Pendidikan	PC1	PC2	PC3
Rata-rata lama sekolah (RLS)	0.921	-0.112	0.085
APK SD	0.885	0.214	-0.102
Indikator Pendidikan	PC1	PC2	PC3
APK SMP	0.912	0.156	-0.042
APK SMA	0.898	0.098	0.115
Harapan lama sekolah (HLS)	0.934	-0.087	0.064
Angka buta aksara usia 15+	-0.712	0.645	0.214
Angka buta aksara usia 15–44	-0.685	0.124	0.756
Angka buta aksara usia 45+	-0.742	0.598	0.187

Berdasarkan Tabel 3, diketahui bahwa PC1 memiliki korelasi yang sangat kuat (>0,8) dengan indikator Harapan Lama Sekolah (0,934) dan Rata-rata Lama Sekolah (0,921). Hal ini menunjukkan bahwa PC1 merupakan representasi dari Dimensi Capaian dan Akses Pendidikan Utama. Sebaliknya, PC2 lebih banyak menangkap variasi pada Angka Buta Aksara 15+ (0,645) dan 45+ (0,598), yang merepresentasikan Masalah Literasi Generasi Tua. Sementara itu, PC3 secara spesifik menangkap variansi Angka Buta Aksara usia 15-44 (0,756) yang menunjukkan hambatan literasi pada Kelompok Usia Produktif.

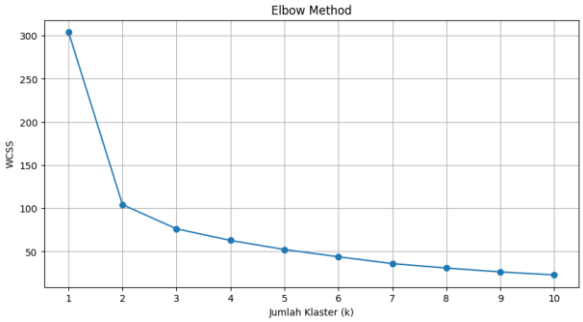
Visualisasi akumulasi varians untuk setiap komponen dapat dilihat melalui grafik *Scree Plot* pada Gambar 4 berikut.



Gambar 4. Grafik Persentase Kumulatif Varians pada Setiap Komponen Utama

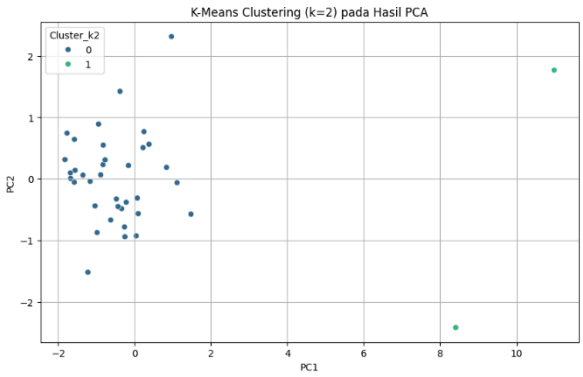
3.2. Evaluasi Model dan Penentuan Nilai k

Untuk mendapatkan jumlah kelompok yang paling objektif, dilakukan pengujian rentang k=2 hingga k=6. Tahapan ini sangat krusial untuk menghindari subjektivitas peneliti dalam menentukan jumlah klaster. Penentuan nilai k yang salah dapat menyebabkan interpretasi wilayah menjadi terlalu sempit atau terlalu luas. Berdasarkan metode *Elbow*, terjadi penurunan drastis pada nilai WCSS saat k=2 dan mulai melandai setelahnya yang terlihat dari gambar 5. Patahnya kurva pada nilai k=2 menandakan bahwa penambahan klaster lebih lanjut tidak lagi memberikan penurunan inersia yang signifikan secara statistik. Dengan kata lain, membagi Indonesia menjadi lebih dari dua kelompok pada dataset 2023 ini justru akan melemahkan struktur kelompok yang terbentuk.



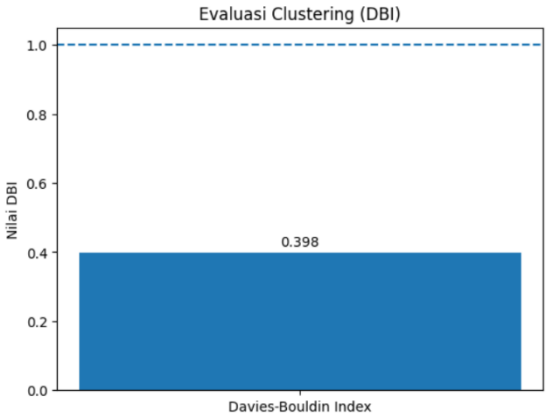
Gambar 5. Evaluasi Parameter Klaster Metode Elbow

Konsistensi model ini divalidasi dengan nilai *Silhouette Score* sebesar 0,782 pada gambar 6 dan *Davies-Bouldin Index* sebesar 0,398 pada gambar 7. Skor Silhouette yang mendekati angka 0,8 merupakan indikator yang sangat kuat bahwa pemisahan antar klaster telah mencapai titik optimal. Hal ini menunjukkan bahwa pengelompokan menjadi dua klaster memiliki struktur yang paling kuat dan pemisahan antar-anggota yang paling optimal dibanding nilai k lainnya.



Gambar 6. Evaluasi Parameter Klaster Nilai Silhouette Coefficient

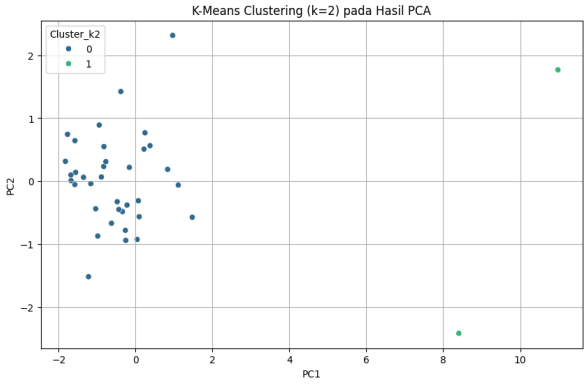
P-ISSN : 2089-3353
E-ISSN : 2808-9162



Gambar 7. Evaluasi Parameter Klaster Nilai Davies-Bouldin Index

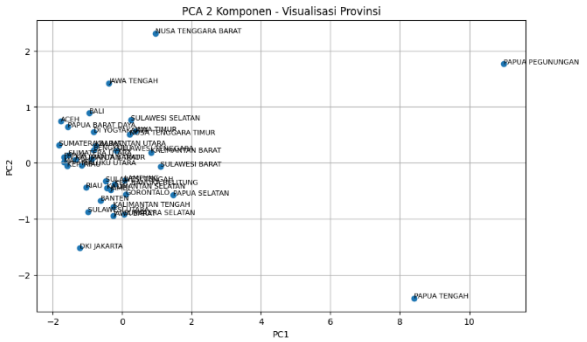
3.3. Klasterisasi

Implementasi algoritma K-Means menyajikan visualisasi sebaran spasial 38 provinsi di Indonesia dalam ruang komponen utama (PC1 dan PC2) setelah penerapan algoritma K-Means (k=2). Plot tersebut menunjukkan adanya asimetri statistik yang tajam, di mana mayoritas wilayah nasional yang tergabung dalam Klaster 0 membentuk densitas tinggi yang mencerminkan homogenitas capaian pendidikan. Sebaliknya, Klaster 1 terisolasi secara ekstrem di sisi kanan sumbu PC1, yang mengonfirmasi status Papua Tengah dan Papua Pegunungan sebagai *outlier* dengan gap karakteristik pendidikan yang sangat lebar dibandingkan wilayah lainnya. Pemisahan yang dominan pada sumbu horizontal ini memvalidasi efektivitas PCA dalam mereduksi dimensi tanpa kehilangan esensi informasi, mengingat PC1 mampu merangkum 74,89% variansi data yang didorong oleh indikator akses dan durasi sekolah, terlihat pada gambar 8 visualisasi centroid di bawah ini.



Gambar 8. Proyeksi Spasial Klaster Provinsi dalam Ruang Komponen Utama

Hasil implementasi K-Means menunjukkan bahwa dari 38 provinsi, mayoritas (36 provinsi) tergabung dalam Klaster 0, sementara Klaster 1 hanya beranggotakan dua provinsi, yaitu Papua Tengah dan Papua Pegunungan. Temuan ini cukup mengejutkan karena secara statistik, Klaster 1 bertindak sebagai *outlier* kelompok. Pengelompokan yang sangat timpang ini mengindikasikan adanya perbedaan karakteristik yang sangat mencolok (outlier) pada kedua provinsi baru tersebut dibandingkan dengan provinsi lainnya di Indonesia. Hal ini terlihat pada gambar 9 di bawah ini.



Gambar 9. Distribusi Frekuensi Jumlah Provinsi Per Klaster

3.4. Analisis Kritis Ketidakseimbangan Klaster

Ketimpangan jumlah anggota antara Klaster 0 (n=36) dan Klaster 1 (n=2) menunjukkan adanya polarisasi data pendidikan yang signifikan di Indonesia. Secara statistik, Klaster 1 tidak dapat dianggap sebagai kelompok variasi biasa, melainkan bertindak sebagai statistical outlier yang memiliki jarak Euclidean sangat lebar dari pusat massa (centroid) nasional. Fenomena ini dibuktikan secara kuantitatif melalui gap capaian rata-rata lama sekolah; di mana Klaster 1 (5,64 tahun) memiliki deviasi negatif yang ekstrem dibandingkan rata-rata Klaster 0 (9,50 tahun). Validitas pemisahan yang asimetris ini didukung oleh *Silhouette Score* sebesar 0,782, yang mengonfirmasi bahwa struktur klaster yang terbentuk memiliki kepadatan internal tinggi dan inter-cluster distance yang lebar. Hal ini menegaskan bahwa ketidakseimbangan jumlah anggota bukan merupakan kelemahan model, melainkan representasi akurat dari adanya anomali pembangunan pendidikan di wilayah DOB Papua.

3.5. Profil Indikator

Pada tabel 4 di bawah ini menunjukkan bahwa klaster 0 memiliki capaian pendidikan lebih baik dengan rata-rata lama sekolah dan harapan lama sekolah lebih tinggi, APK merata tinggi di semua jenjang, serta tingkat buta aksara rendah pada seluruh kelompok usia. Hal ini mencerminkan bahwa sebagian besar wilayah Indonesia telah mencapai standar pelayanan minimal pendidikan. Sebaliknya, klaster 1 ditandai oleh lama sekolah dan APK yang rendah serta tingkat buta aksara yang sangat tinggi, khususnya pada kelompok usia dewasa dan lanjut usia.

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Tabel 4. Profil Rata-rata Indikator Per Klaster

Indikator	Klaster 0	Klaster 1
Rata-rata lama sekolah (th)	9,50	**5,64**
Harapan lama sekolah (th)	13,38	9,80
APK SD (%)	105,96	77,50
APK SMP (%)	91,10	76,75
APK SMA (%)	89,84	58,96
Buta aksara 15+ (%)	2,66	22,47
Buta aksara 15–44 (%)	0,43	17,88
Buta aksara 45+ (%)	6,20	**34,19**

3.6. Pembahasan

Berdasarkan data pada Tabel 4, terlihat adanya kesenjangan pendidikan yang sangat tajam antara Klaster 0 dan Klaster 1. Kesenjangan ini menunjukkan bahwa pemerataan kualitas pendidikan belum sepenuhnya menyentuh wilayah otonomi baru di Papua. Secara statistik, perbedaan ini bukan sekadar variasi angka biasa, melainkan sebuah polarisasi yang memisahkan antara wilayah yang sudah mapan dengan wilayah yang masih berjuang pada level literasi dasar.

Klaster 0 menunjukkan kondisi pendidikan yang relatif stabil dan sesuai standar nasional dengan angka harapan lama sekolah mencapai 13,38 tahun. Ini berarti secara rata-rata, penduduk di klaster ini diharapkan mampu mengenyam pendidikan hingga tingkat Diploma atau awal Perguruan Tinggi. Hal ini menandakan bahwa infrastruktur pendidikan di mayoritas provinsi Indonesia sudah cukup mampu mendukung keberlanjutan studi hingga ke jenjang pasca-menengah.

Sebaliknya, Klaster 1 muncul sebagai anomali statistik dengan angka buta aksara pada usia 45+ mencapai 34,19%. Angka ini sangat mengkhawatirkan karena sepertiga penduduk usia lanjut di wilayah tersebut belum memiliki kemampuan literasi dasar. Ketertinggalan ini menciptakan beban ganda bagi pembangunan daerah, di mana rendahnya kemampuan literasi pada generasi tua dapat menghambat adopsi teknologi dan informasi di tingkat rumah tangga. Selisih rata-rata lama sekolah sebesar 3,86 tahun mengindikasikan bahwa penduduk di wilayah Papua Tengah dan Papua Pegunungan secara umum belum menuntaskan jenjang pendidikan dasar secara kolektif. Jika ditarik perbandingannya secara linear, terdapat jarak waktu hampir empat dekade pembangunan yang harus dikejar oleh Klaster 1 untuk bisa setara dengan pencapaian rata-rata di Klaster 0. Jika rata-rata lama sekolah di Klaster 0 sudah mencapai 9,50 tahun (setara tamat SMP), Klaster 1 hanya berada di angka 5,64 tahun (belum tamat SD). Hal ini menunjukkan bahwa tantangan di wilayah tersebut berkaitan erat dengan akses literasi dasar yang belum tuntas di masa lalu serta kemungkinan adanya hambatan bahasa atau ketersediaan tenaga pendidik yang kronis di wilayah pegunungan.

Temuan ini mengonfirmasi bahwa penggunaan PCA sebelum K-Means sangat membantu dalam menangani *multicollinearity* antar indikator pendidikan di Indonesia. PCA berhasil menyaring informasi esensial dari variabel yang saling tumpang tindih, sehingga proses klasterisasi dapat fokus pada dimensi yang benar-benar membedakan karakteristik antarwilayah. Tanpa PCA, indikator yang saling berkorelasi kuat mungkin akan mengaburkan perbedaan nyata yang ada di lapangan. Secara praktis, pengelompokan ini menunjukkan bahwa masalah utama di Papua Tengah dan Papua Pegunungan (Klaster 1) bukan sekadar pada partisipasi sekolah saat ini, melainkan pada akumulasi buta aksara di kelompok usia dewasa. Data ini memberikan bukti empiris bahwa strategi satu ukuran untuk semua dalam kebijakan pendidikan nasional tidak lagi relevan. Hal ini memberikan masukan bagi pemerintah bahwa program literasi dasar bagi penduduk usia produktif dan lansia harus menjadi prioritas di wilayah tersebut, selain dari pembangunan infrastruktur fisik sekolah. Intervensi kebijakan tidak bisa lagi disamaratakan.

Di sisi lain, bagi wilayah yang tergabung dalam Klaster 0, tantangannya sudah bergeser pada aspek kualitas dan relevansi lulusan dengan dunia kerja. Klaster 0 yang sudah mapan perlu difokuskan pada peningkatan kualitas pendidikan tinggi (X4), mengingat rata-rata capaian pada indikator tersebut masih memiliki ruang pengembangan yang besar dibandingkan dengan jenjang pendidikan dasar. Dengan demikian, strategi pembangunan pendidikan di Indonesia harus dilakukan secara asimetris berdasarkan karakteristik klaster masing-masing wilayah, di mana Klaster 1 memerlukan percepatan literasi dasar, sementara Klaster 0 memerlukan penguatan inovasi di jenjang pendidikan menengah dan tinggi.

4. Kesimpulan

Berdasarkan hasil analisis dan pembahasan, dapat disimpulkan bahwa integrasi metode *Principal Component Analysis* (PCA) dan K-Means *Clustering* berhasil memetakan profil pemerataan pendidikan di Indonesia secara objektif dan terukur. Pendekatan ini terbukti efektif dalam menangani data indikator pendidikan yang kompleks dan multidimensi tanpa mengurangi esensi informasi asli. Reduksi dimensi menggunakan PCA mampu merangkum 91,85% informasi dari delapan indikator menjadi tiga komponen utama, yang terbukti meningkatkan stabilitas proses klasterisasi dengan meminimalisir efek multikolinieritas antar-variabel. Penentuan jumlah kelompok melalui metode *Elbow* menghasilkan dua klaster optimal dengan *Silhouette Score* sebesar 0,782, yang menunjukkan tingkat validitas pengelompokan

P-ISSN : 2089-3353
E-ISSN : 2808-9162

yang sangat kuat dan struktur pemisahan data yang signifikan secara statistik.

Hasil pemetaan menunjukkan adanya disparitas yang tajam, di mana 36 provinsi berada pada klaster capaian tinggi (Klaster 0), sementara Provinsi Papua Tengah dan Papua Pegunungan terisolasi dalam klaster capaian rendah (Klaster 1). Kondisi di Klaster 1 ini ditandai dengan rendahnya angka partisipasi sekolah dan tingginya angka buta aksara pada usia dewasa yang menjadi penghambat utama peningkatan indeks pembangunan manusia di wilayah tersebut. Temuan ini merekomendasikan perlunya kebijakan yang bersifat asimetris dan berbasis karakteristik wilayah, terutama peningkatan program pemberantasan buta aksara dan percepatan lama sekolah di wilayah Papua. Intervensi pemerintah tidak bisa dilakukan secara seragam, melainkan harus fokus pada pemulihan literasi di wilayah tertinggal melalui pendekatan sosial-budaya yang spesifik. Untuk penelitian selanjutnya, disarankan untuk menambahkan variabel ekonomi seperti tingkat kemiskinan atau alokasi anggaran pendidikan daerah guna memberikan perspektif yang lebih komprehensif mengenai faktor penyebab kesenjangan tersebut. Dengan tambahan variabel eksternal, model pemetaan diharapkan dapat mengungkap korelasi antara kapasitas fiskal daerah dengan kualitas output pendidikan secara lebih mendalam sebagai basis perencanaan pembangunan jangka panjang yang lebih inklusif.

Daftar Rujukan

[1] R. Syahputra, "Analisis Ketimpangan Pendidikan antar Wilayah di Indonesia.," *Jurnal Ekonomi dan Kebijakan Publik,* vol. 12, no. 1, pp. 45-58, 2021.

[2] A. Nurahman, "Disparitas Kualitas Sumber Daya Manusia di Wilayah Tertinggal," *Jurnal Sosial dan Humaniora,* vol. 8, no. 2, pp. 112-125, 2022.

[3] Badan Pusat Statistik, Potret Pendidikan Indonesia: Statistik Pendidikan 2023, Jakarta: BPS RI, 2023.

[4] K. d. F. M. Anwar, "Penerapan Principal Component Analysis untuk Reduksi Dimensi pada Data Multivariat," *Jurnal Sains Data,* vol. 4, no. 1, pp. 12-20, 2020.

[5] A. Sudrajat, "Evaluasi Kebijakan Pemerintah dalam Pemerataan Pendidikan Nasional," *Jurnal Administrasi Publik,* vol. 11, no. 2, pp. 101-115, 2020.

[6] J. K. M. d. P. J. Han, Data Mining: Concepts and Techniques. 3rd ed, Waltham: Morgan Kaufmann, 2012.

[7] M. A. Syakur, "Integrasi Algoritma PCA dan K-Means dalam Analisis Klasterisasi," *Jurnal Teknologi Informasi,* vol. 15, no. 3, pp. 201-210, 2021.

[8] S. Raschka, Python Machine Learning, Birmingham: Packt Publishing, 2015.

[9] I. T. d. C. J. Jolliffe, "Principal Component Analysis: A Review and Recent Developments," *Philosophical Transactions of the Royal Society A,* vol. 374, no. 2061, 2016.

[10] J. Shlens, "A Tutorial on Principal Component Analysis," *arXiv preprint arXiv:1404.1100,* 2014.

[11] S. Sauri, "Analisis Multivariat dalam Pemetaan Pendidikan Dasar," *Jurnal Pendidikan Dasar,* vol. 3, no. 2, pp. 15-28, 2019.

[12] A. Suryadi, "Metodologi Penelitian Pendidikan dan Validitas Data Sekunder," *Jurnal Ilmiah Pendidikan,* vol. 7, no. 1, pp. 34-46, 2020.

[13] S. Hardiani, "Pentingnya Standarisasi Data pada Algoritma Berbasis Jarak," *Jurnal Informatika,* vol. 10, no. 2, pp. 88-97, 2021.

[14] L. d. R. P. J. Kaufman, "Finding Groups in Data: An Introduction to Cluster Analysis.," New York, John Wiley & Sons, 2009.

[15] I. T. d. C. J. olliffe, "Principal Component Analysis: A Review and Recent Developments.," *hilosophical Transactions of the Royal Society A,* vol. 374, no. 2061, 2016.

[16] H. d. W. L. J. Abdi, "Principal Component Analysis," *Wiley Interdisciplinary Reviews: Computational Statistics,* vol. 2, no. 4, pp. 433-459, 2010.

[17] J. F. B. W. C. B. B. J. d. A. R. E. Hair, Multivariate Data Analysis. 7th ed, Harlow: Pearson Education, 2014.

[18] P. d. K. A. Bholowalia, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means," *International Journal of Computer Applications,* vol. 105, no. 9, pp. 17-24, 2014.

[19] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics,* vol. 20, pp. 53-65, 1987.

[20] D. L. d. B. D. W. Davies, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 1, no. 2, pp. 224-227, 1979.