p-ISSN 2085-7829
e-ISSN 2798-3455

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

Nurlia Sucianti Putri[1], Memi Nor Hayati[2a)], dan Rito Goejantoro[3]
[1, 2, 3]Program Studi Statistika, Jurusan Matematika, FMIPA, Universitas Mulawarman, Indonesia

[a)]Corresponding author: meminorhayati@fmipa.unmul.ac.id

ABSTRACT

Cluster analysis is used to group several objects based on similarities within the group. There are many methods included in cluster analysis, including k-means. K-means is a non-hierarchical cluster analysis method. The assumption that needs to be considered in cluster analysis is that there is no strong correlation between research variables. An alternative that can be done to deal with variables that are strongly correlated is to use Principal Component Analysis (PCA). This research aims to group districts/cities in Kalimantan based on education indicators in 2022 using k-means with PCA optimization, as well as finding out the optimal cluster based on the smallest Davies Bouldin Index (DBI) value. Based on the results of the analysis, from 11 research variables two main components were formed. From these two main components, new data transformations are produced which are then used in grouping districts/cities in Kalimantan based on education indicators using the k-means methods. The analysis results, it was found that the optimal cluster with k-means grouping was 5 clusters with a DBI value of 0.835. Cluster 1 has 8 regencies/cities, cluster 2 has 16 regencies/cities, cluster 6 has 5 regencies/cities, cluster 4 has 21 regencies/cities, and cluster 5 has 5 regencies/cities.

Keywords*: dbi, education indicators, k-means, pca*

1. Pendahuluan

Analisis klaster adalah suatu metode dalam statistika yang digunakan untuk mengidentifikasi kelompok-kelompok alami dalam suatu kumpulan data, di mana objek-objek dalam satu kelompok memiliki kemiripan yang lebih besar dibandingkan dengan objek-objek di kelompok lainnya (Ulinnuh & Veriani, 2020). Salah satu metode analisis klaster adalah *k-means*. Metode *k-means* metode yang tergolong sederhana dan mudah dalam penerapannya, sehingga sering digunakan dalam berbagai aplikasi. Namun, di balik kemudahan penerapannya metode *k-means* dengan data dimensi tinggi tidak mampu memberikan performa yang baik dikarenakan masalah yang muncul seperti menurunnya nilai akurasi, sehingga untuk mengatasi masalah pada data dimensi tinggi yaitu dilakukan reduksi dimensi. Adapun metode yang dapat digunakan untuk mereduksi dimensi salah satunya *Principal Component Analysis* (Suyanto, 2017).

Principal Component Analysis (PCA) merupakan teknik transformasi linier yang mengubah sekumpulan variabel asal menjadi sekumpulan variabel baru yang tidak berkorelasi. Variabel baru ini, disebut komponen utama, dipilih sedemikian rupa sehingga mampu menjelaskan variabilitas data asli secara maksimal. Dengan kata lain, PCA bertujuan mereduksi dimensi data sambil mempertahankan informasi yang paling relevan (Wangge, 2021).

Pada analisis klaster, perlu dilakukan validasi untuk mengukur kualitas dari hasil analisis klaster yang diperoleh. Salah satu metode evaluasi klaster yaitu *Davies-Bouldin Index* (DBI). DBI dapat mengukur kualitas hasil klasterisasi dengan memaksimalkan jarak antar klaster dan meminimalkan jarak antar objek dalam satu klaster (Badruttamam, dkk., 2020). Selain itu, DBI merupakan metode evaluasi klaster yang familiar dikarenakan sudah banyak literatur yang menggunakan metode dalam mengevaluasi hasil klaster.

Penelitian ini mengacu pada penelitian sebelumnya. Sopyan, dkk. (2022), melakukan penelitian untuk mengelompokkan kasus tingkat perceraian di Kuningan menggunakan algoritma *k-means*. Pengelompokan dilakukan dengan beberapa variasi jumlah klaster yaitu 2,3,4 dan 5. Kemudian dari hasil pengelompokan tersebut dibandingkan berdasarkan nilai DBI dan diperoleh bahwa hasil klaster optimal yaitu sebanyak empat klaster dengan nilai DBI sebesar 0,535. Penelitian lainnya yang dilakukan oleh Amrullah, dkk (2022) dengan metode klaster dan validitas serupa untuk mengelompokan faktor penunjang pendidikan di Kabupaten Karawang menghasilkan hasil klaster optimal sebanyak 2 klaster dengan nilai DBI sebesar 0,408.

Menurut Bashori dan Aprima (2019), pendidikan merupakan salah satu kunci dalam menentukan kualitas hidup manusia. Pendidikan dianggap sebagai jembatan bagi manusia untuk dapat mengembangkan potensi diri, oleh karena itu masyarakat diharapkan dapat berpartisipasi dalam pendidikan. Menurut Badan Pusat Statistik (2023), menunjukkan bahwa Angka Partisipasi Sekolah (APS) di Indonesia sepanjang tahun 2020 sampai 2022 masih berada di bawah 100%. Terlebih yang harus menjadi perhatian saat ini adalah APS pada jenjang pendidikan SMA yang hanya mencapai angka 73,15% di tahun 2022.

DOI 10.30872/eksponensial.v15i2.1373

p-ISSN 2085-7829
e-ISSN 2798-3455

Kalimantan merupakan salah satu provinsi terbesar di Indonesia, jika dilihat APS pada kelompok usia 16-18 tahun di masing-masing provinsi mengalami fluktuatif dalam tiga tahun terakhir (2020-2022). Pada tahun 2021, Kalimantan Barat mengalami kenaikan sebesar 0,42% tetapi pada tahun 2022 mengalami penurunan sebesar 0,66%. Kalimantan Selatan mengalami penurunan sebesar 0,07% di tahun 2021 namun, di tahun selanjutnya mengalami kenaikan sebesar 0,57%. Sedangkan Kalimantan Tengah mengalami penurunan selama dua tahun berturut-turut yaitu 2021 dan 2022 sebesar 0,22% dan 0,38%. Adapun Kalimantan Timur mengalami kenaikan sebesar 0,13% di tahun 2021 tetapi sayangnya, pada tahun 2022 mengalami penurunan sebesar 0,58%. Sama halnya, Kalimantan Utara juga mengalami kenaikan sebesar 0,4% pada tahun 2021 kemudian di tahun 2022 nilai APS cenderung konstan dari tahun sebelumnya. Dari semua provinsi di Kalimantan, wilayah Kalimantan Barat memiliki nilai APS paling rendah dibandingkan wilayah lainnya yaitu hanya berada pada rata-rata 69,02%.

Berdasarkan uraian tersebut, maka pada penelitian ini akan dilakukan pengelompokan Kabupaten/Kota di Kalimantan berdasarkan indikator pendidikan menggunakan metode *K-Means* dan *Fuzzy C-Means* dengan Optimasi *Principal Component Analysis* (PCA) yang bertujuan untuk memperoleh hasil pengelompokan yang optimal berdasarkan nilai *Davies-Bouldin Index* (DBI) terkecil.

2. Tinjauan Pustaka

2.1 Analisis Klaster

Analisis klaster merupakan metode yang digolongkan sebagai *unsupervised learning*. Tujuan utama dari analisis klaster adalah mengelompokkan sejumlah objek ke dalam klaster, sehingga setiap klaster berisi objek yang semirip mungkin (Santosa & Umam, 2018). Asumsi yang harus dipenuhi untuk dapat menggunakan analisis klaster, yaitu tidak terjadi korelasi yang tinggi antar variabel atau non-multikolinieritas (Nawari, 2010). Besaran yang dapat digunakan untuk mendeteksi adanya multikolinieritas adalah *Variance Inflation Factor* (VIF). Adapun persamaan VIF (Gujarati, 2003) adalah sebagai berikut:

$$VIF = \frac{1}{1-R_j^2}; j = 1, 2, 3, ..., p \tag{1}$$

di mana R_j^2 merupakan koefisien determinasi dari variabel X_j yang diregresikan dengan variabel lainnya

2.2 *K-Means*

Algoritma *K-Means* digunakan untuk membagi data menjadi beberapa kelompok dengan sistem partisi tanpa menggunakan label kelas (Wanto, dkk., 2020). Metode ini bekerja dengan meminimalkan jumlah kuadrat jarak antara objek dan masing-masing *centroid* atau pusat klaster (Irwansyah & Faisal, 2015). Tahapan metode *K-Means* menurut Prasetyo (2012) adalah sebagai berikut:

1. Menentukan jumlah klaster (C)
2. Menentukan pusat klaster (v_{cj}) awal secara acak
3. Menghitung jarak Euclid untuk setiap objek terhadap pusat klaster

$$d(x_i, v_c) = \sqrt{\sum_{j=1}^{p} (x_{ij} - v_{cj})^2} \tag{2}$$

di mana:

$d(x_i, v_c)$: jarak Euclid objek pengamatan ke-i dengan pusat klaster ke-c
v_{cj} : pusat klaster ke-c pada variabel ke-j
x_{ij} : objek pengamatan ke-i variabel ke-j

4. Mengalokasikan masing-masing objek ke klaster yang objeknya paling mirip berdasarkan jarak terdekat antara objek terhadap setiap pusat klaster.
5. Memperbarui pusat klaster dengan menghitung nilai rata-rata dari setiap objek untuk setiap klaster.

$$v_{cj}^t = \sum_{i=1}^{n_c} \frac{x_{ijc}}{n_c} \tag{3}$$

di mana:

v_{cj}^t : pusat klaster baru ke-c variabel ke-j pada iterasi ke-t
x_{ijc} : objek pengamatan ke-i variabel ke-j pusat klaster ke-c
n_c : banyaknya objek pada klaster ke-c

6. Mengulangi langkah 3,4, dan 5 sampai tidak ada lagi anggota pada suatu klaster yang berpindah klaster.
7. Analisis selesai jika tidak ada lagi perubahan anggota pada setiap klaster.

2.3 *Principal Component Analysis*

Principal Component Analysis **(**PCA) adalah salah satu metode analisis multivariat yang dapat digunakan dalam mereduksi sejumlah variabel asal menjadi beberapa variabel baru yang bersifat orthogonal

DOI 10.30872/eksponensial.v15i2.1373

p-ISSN 2085-7829
e-ISSN 2798-3455

dengan tetap mempertahankan varians dari variabel asalnya. *Principal component* bersifat saling orthogonal (saling bebas) yang artinya setiap variabel saling independen dan tidak saling berkorelasi (Nugroho, 2008). Tahapan analisis yang dilakukan pada PCA menurut Prasetyo (2012) adalah sebagai berikut:

1. Melakukan normalisasi pada data pengamatan menggunakan persamaan berikut.

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j} \tag{4}$$

di mana:

Z_{ij} : hasil normalisasi objek pengamatan ke-i variabel ke-j
x_{ij} : objek pengamatan ke-i variabel ke-j
$\bar{x}_j$: rata-rata nilai dari variabel ke-j
s_j : simpangan baku variabel ke-j

dengan:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}; i = 1,2,3,...,n;\ j = 1,2,3,...,p \tag{5}$$

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \tag{6}$$

2. Menghitung koefisien korelasi dengan persamaan berikut.

$$r_{a,b} = \frac{n\left(\sum_{i=1}^{n} z_{ia}z_{ib}\right) - \left(\sum_{i=1}^{n} z_{ia}\right)\left(\sum_{i=1}^{n} z_{ib}\right)}{\sqrt{n\left(\sum_{i=1}^{n} z_{ia}^2\right) - \left(\sum_{i=1}^{n} z_{ia}\right)^2} \cdot \sqrt{n\left(\sum_{i=1}^{n} z_{ib}^2\right) - \left(\sum_{i=1}^{n} z_{ib}\right)^2}}; a = b; 1,2,3,..., p \tag{7}$$

di mana:

$r_{a,b}$: koefisien korelasi data hasil normalisasi variabel ke-a dan ke-b

Menurut Sugiyono (2013), sebagai bahan penafsiran terhadap koefisien korelasi maka dapat berpedoman pada ketentuan pada **Tabel 1** berikut:

Tabel 1. Tingkat Hubungan Koefisien Korelasi

Interval Koefisien	Tingkat Hubungan
0,000 – 0,199	Sangat Lemah
0,200 – 0,399	Lemah
0,400 – 0,599	Sedang
0,600 – 0,799	Kuat
0,800 – 1,000	Sangat Kuat

3. Membuat matriks korelasi berdasarkan koefisien korelasi.

$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix} \tag{8}$$

4. Menentukan nilai eigen berdasarkan persamaan berikut.

$$\det(\lambda\mathbf{I} - \mathbf{R}) = 0 \tag{9}$$

dan vektor eigen sesuai persamaan:

$$\mathbf{R}\vec{v} = \lambda\vec{v} \tag{10}$$

di mana:

λ adalah nilai eigen, $\vec{v}$ vektor eigen, dan **I** merupakan matriks identitas dengan persamaan berikut.

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{11}$$

5. Menentukan banyaknya komponen utama yang terbentuk berdasarkan kriteria $\lambda \geq 1$.
6. Menghitung komponen matriks korelasi yang menandakan besarnya korelasi variabel terhadap skor komponen yang terbentuk menggunakan persamaan berikut.

DOI 10.30872/eksponensial.v15i2.1373

p-ISSN 2085-7829
e-ISSN 2798-3455

$$r_{X_j, PC_t} = \overset{\mathrm{r}}{v}_{it}\sqrt{\lambda_t} \tag{12}$$

di mana:

PC_t : komponen utama ke-t
X_j : variabel ke-j
$\vec{v}_{it}$: vektor eigen ke-i komponen utama ke-t
λ_t : nilai eigen komponen utama ke-t

7. Menghitung transformasi himpunan data baru hasil reduksi dengan menggunakan PCA sesuai persamaan berikut.

$$PC_{it} = \vec{v}_{i1}Z_{i1} + \vec{v}_{i2}Z_{i2} + ... + \vec{v}_{ip}Z_{ip} \tag{13}$$

2.4 *Davies-Bouldin Index*

Menurut Rahmayanti, dkk. (2022) DBI merupakan metode validitas internal dalam melakukan evaluasi pada hasil klasterisasi. DBI menentukan seberapa baik klasterisasi yang dilakukan dengan menghitung kuantitas dan fitur turunan dari himpunan data. Adapun tahapan perhitungan DBI adalah sebagai berikut:

1. Menghitung nilai *Sum of Square Within* (SSW) dengan formulasi sebagai berikut:

$$SSW_c = \sqrt{\frac{1}{n_c}\sum_{i=1}^{n_c}\sum_{j=1}^{p}(x_{ij} - v_{cj})^2} \tag{14}$$

Keterangan:

SSW_c : *sum of square within* klaster ke-c
n_c : banyaknya objek pada klaster ke-c
v_{cj} : pusat klaster ke-c variabel ke-j
x_{ij} : objek pengamatan ke-i variabel ke-j

2. Menghitung nilai *Sum of Square Between* (SSB) dengan mengukur jarak antar pusat klaster v_{cj} dan v_{ej} menggunakan persamaan berikut:

$$SSB_{ce} = \sqrt{\sum_{j=1}^{p}\left(v_{cj}, v_{ej}\right)^2} \tag{15}$$

Keterangan:

SSB_{ce} : *sum of square between* pusat klaster v_c dan v_e

di mana $c, e = 1,2,3,\ldots,C$.

3. Menghitung ukuran rasio kebaikan perbandingan antara klaster ke-c dan klaster-e (R_{ce}) dengan rumus berikut:

$$R_{ce} = \frac{SSW_c + SSW_e}{SSB_{ce}} \tag{16}$$

Keterangan:

R_{ce} : rasio perbandingan klaster ke-c dengan klaster ke-e

4. Menghitung nilai DBI dengan menggunakan formula berikut:

$$DBI = \frac{1}{C}\sum_{c=1}^{C} max(R_{ce}) \tag{17}$$

Hasil klaster terbaik adalah klaster dengan nilai DBI paling kecil.

3. Bahan dan Metode

3.1 Sumber Data

Data yang digunakan pada penelitian ini bersumber dari website https://www.bps.go.id/ Badan Pusat Statistik (BPS) dari lima provinsi di Kalimantan tahun 2022.

3.2 Variabel Penelitian

Variabel yang diteliti merupakan indikator pendidikan di Kalimantan dengan rincian sebagai berikut.

Tabel 2. Variabel Penelitian

Notasi	Variabel	Notasi	Variabel
X_1	Jumlah SD/Sederajat	X_7	Jumlah Murid SD/Sederajat
X_2	Jumlah SMP/Sederajat	X_8	Jumlah Murid SMP/Sederajat
X_3	Jumlah SMA/Sederajat	X_9	Jumlah Murid SMA/Sederajat
X_4	Jumlah Guru SD/Sederajat	X_{10}	Rata-rata Lama Sekolah
X_5	Jumlah Guru SMP/Sederajat	X_{11}	Harapan Lama Sekolah
X_6	Jumlah Guru SMA/Sederajat		

DOI 10.30872/eksponensial.v15i2.1373

p-ISSN 2085-7829
e-ISSN 2798-3455

3.3 Teknik Pengumpulan Data

Teknik pengumpulan data yang dilakukan pada penelitian ini adalah dengan cara mengambil data yang sudah ada (data sekunder) berdasarkan 56 kabupaten/kota di Kalimantan pada tahun 2022.

3.4 Teknik Analisis Data

Adapun tahapan yang akan dilakukan untuk menganalisis data penelitian menggunakan bantuan *software* R adalah sebagai berikut:

1. Melakukan analisis statistika deskriptif pada data penelitian menggunakan *spatial mapping* dan grafik antar dua variabel yang saling berkaitan.
2. Melakukan pendeteksian multikolinieritas pada variabel penelitian menggunakan persamaan (1). Jika terjadi multikolinieritas maka lakukan reduksi dimensi dengan metode PCA.
3. Melakukan analisis PCA dengan tahapan sebagai berikut:
   a. Melakukan normalisasi pada objek pengamatan menggunakan persamaan (4).
   b. Menghitung koefisien korelasi antar variabel dengan persamaan (7).
   c. Membuat matriks korelasi berdasarkan koefisien korelasi sesuai dengan persamaan (8).
   d. Menentukan nilai eigen berdasarkan persamaan (9) dan vektor eigen sesuai persamaan (10).
   e. Menentukan banyaknya komponen utama yang terbentuk dengan melihat nilai eigen yang lebih besar atau sama dengan 1.
   f. Membentuk komponen matriks korelasi yang menandakan besarnya korelasi variabel terhadap skor komponen yang terbentuk menggunakan persamaan (12).
   g. Menghitung transformasi himpunan data baru hasil reduksi menggunakan metode PCA sesuai persamaan (13).
4. Melakukan pengelompokan menggunakan metode *K-Means* dengan langkah-langkah berikut:
   a. Menentukan jumlah klaster (C) yang akan digunakan yaitu 2,3,4, dan 5.
   b. Menentukan pusat klaster (v_{cj}) awal secara acak menggunakan *software* R.
   c. Menghitung jarak Euclid untuk setiap objek terhadap pusat klaster dengan persamaan (2).
   d. Mengalokasikan masing-masing objek ke klaster yang objeknya paling mirip berdasarkan jarak terdekat antara objek terhadap setiap pusat klaster.
   e. Memperbarui pusat klaster dengan menghitung nilai rata-rata dari setiap objek untuk setiap klaster sesuai persamaan (3).
   f. Mengulangi langkah c, d, dan e sampai tidak ada anggota pada suatu klaster yang berpindah.
   g. Mengulangi langkah c, d, e, dan f untuk jumlah klaster yang berbeda.
5. Melakukan pemilihan klaster optimal dari hasil pengelompokan metode *k-means* menggunakan nilai DBI berdasarkan persamaan (17). Klaster optimal ditentukan berdasarkan nilai DBI terkecil.
6. Menginterpretasi hasil pengelompokan *k-means* dari klaster optimal berdasarkan nilai DBI terkecil.

4. Hasil dan Pembahasan

4.1 Statistika Deskriptif

Analisis statistika deskriptif merupakan penyajian data yang berfungsi untuk melihat ringkasan atau gambaran data, sehingga memberikan informasi yang lebih mudah dipahami. Hasil analisis statistika deskriptif dapat dilihat pada **Tabel 3** berikut:

Tabel 3. Statistika Deskriptif

Variabel	Minimum	Maksimum	Rata-Rata	Simpangan Baku
X_1	31,000	564,000	247,300	136,616
X_2	11,000	252,000	85,460	47,645
X_3	5,000	152,000	44,460	27,371
X_4	365,000	6.303,000	2.593,000	1.279,646
X_5	194,000	2.820,000	1.189,000	611,726
X_6	99,000	2.790,000	937,000	553,258
X_7	3.510,000	92.562,000	34.082,000	21.433,810
X_8	1.282,000	42.434,000	14.597,000	9.661,422
X_9	1.021,000	43.264,000	12.480,000	8.915,635
X_{10}	6,210	11,550	8,494	1,193
X_{11}	11,180	15,100	12,830	0,850

4.2 Pendeteksian Multikolinieritas

Pendeteksian multikolinieritas sebagai syarat agar dapat menggunakan metode analisis klaster. Besaran yang dapat digunakan untuk mendeteksi multikolinieritas adalah nilai VIF. Berikut ini merupakan hasil hitung nilai VIF pada setiap variabel yang dapat dilihat pada **Tabel 4** sebagai berikut.

Tabel 4. Nilai VIF

DOI 10.30872/eksponensial.v15i2.1373

p-ISSN 2085-7829
e-ISSN 2798-3455

X_1	X_2	X_3	X_4	X_5	X_6
23,006	38,645	30,065	81,029	94,372	93,485
X_7	X_8	X_9	X_{10}	X_{11}	
128,056	242,747	185,826	4,887	8,666	

Dari **Tabel 4** dapat dilihat bahwa dari variabel X_1 hingga X_9 masing-masing memiliki nilai nilai VIF lebih besar dari 10. Hal ini dapat diartikan bahwa terjadi multikolinieritas antar variabel dan proses klasterisasi tidak dapat dilanjutkan karena variabel yang mengalami multikolinieritas akan menyebabkan hasil klaster yang tidak optimal. Adapun alternatif yang dapat dilakukan yaitu mereduksi variabel dengan metode PCA.

4.3 *Principal Component Analysis*

Pada tahapan pendeteksian multikolinieritas, telah diketahui bahwa terdapat 9 dari 11 variabel yang masih saling berkorelasi, sehingga untuk dapat melanjutkan ke proses pengelompokan tanpa harus menghilangkan variabel yang berkorelasi dapat diterapkan metode PCA. Tahapan analisis dalam metode PCA sebagai berikut.

a. Melakukan normalisasi data

Normalisasi data dilakukan guna menghasilkan rentang nilai yang sama dari semua variabel. Normalisasi data dihitung dengan menggunakan persamaan (4). Hasil perhitungan dapat dilihat pada **Tabel 5** berikut.

Tabel 5. Data Hasil Normalisasi Menggunakan *Z-Score*

Objek	Z_1	Z_2	Z_3	⋯	Z_{11}
1	1,527	1,585	1,079	⋯	-0,153
2	0,268	0,158	-0,053	⋯	-0,811
3	1,630	0,683	0,714	⋯	-0,459
⋮	⋮	⋮	⋮	⋱	⋮
56	-1,261	-1,080	-0,602	⋯	1,424

b. Menghitung koefisien korelasi

Perhitungan koefisien korelasi antar variabel berguna untuk melihat keterkaitan hubungan antar variabel. Nilai ini dihitung dengan menggunakan persamaan (7). Hasil perhitungan dapat dilihat pada **Tabel 6**.

Tabel 6. Koefisien Korelasi

	Z_1	Z_2	Z_3	⋯	Z_{11}
Z_1	1,000	0,898	0,687	⋯	-0,104
Z_2	0,898	1,000	0,882	⋯	0,116
Z_3	0,687	0,882	1,000	⋯	0,486
⋮	⋮	⋮	⋮	⋱	⋮
Z_{11}	-0,104	0,116	0,486	⋯	1,000

Berdasarkan hasil perhitungan di atas, diperoleh nilai koefisien korelasi antara $\boldsymbol{X_1}$ dan $\boldsymbol{X_2}$ sebesar 0,898. Nilai tersebut mengartikan bahwa antar variabel tersebut memiliki korelasi yang kuat.

c. Membuat matriks korelasi

Setelah diperoleh nilai koefisien korelasi, maka dapat dibentuk matrik korelasi sebagai berikut:

$$\mathbf{R} = \begin{pmatrix} 1 & 0,898 & 0,687 & \cdots & -0,104 \\ 0,898 & 1 & 0,882 & \cdots & 0,116 \\ 0,687 & 0,882 & 1 & \cdots & 0,486 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -0,104 & 0,116 & 0,486 & \cdots & 1 \end{pmatrix}$$

d. Menentukan nilai eigen dan vektor eigen

Perhitungan nilai eigen didasarkan pada persamaan (9) dan vektor eigen dengan persamaan (10). Hasil perhitungan nilai eigen dapat dilihat pada **Tabel 7**.

Tabel 7. Nilai Eigen

Notasi	Nilai Eigen	Notasi	Nilai Eigen
λ_1	8,031	λ_7	0,035
λ_2	2,260	λ_8	0,021
λ_3	0,254	λ_9	0,007
λ_4	0,227	λ_{10}	0,004
λ_5	0,108	λ_{11}	0,002
λ_6	0,051		

Dapat dilihat berdasarkan nilai eigen tersebut, diperoleh sebanyak 11 nilai eigen. Kemudian pembentukan komponen utama dipilih berdasarkan kriteria $\lambda \geq 1$. Sehingga dari delapan nilai eigen yang diperoleh hanya dua nilai eigen yang memenuhi kriteria, yaitu $\lambda_1 = 8{,}301$ dan $\lambda_2 = 2{,}260$. Selanjutnya, diperoleh vektor eigen sebagai berikut:

DOI 10.30872/eksponensial.v15i2.1373

p-ISSN 2085-7829
e-ISSN 2798-3455

$$\vec{v} = \begin{pmatrix} -0,267 & 0,389 & \cdots & 0,030 \\ -0,310 & 0,260 & \cdots & -0,062 \\ \vdots & \vdots & \ddots & \vdots \\ -0,165 & -0,548 & \cdots & -0,065 \end{pmatrix}$$

DOI 10.30872/eksponensial.v15i2.1373

p-ISSN 2085-7829
e-ISSN 2798-3455

e. Menghitung komponen matriks korelasi
Komponen korelasi menunjukkan besarnya suatu korelasi variabel terhadap skor komponen yang terbentuk. Perhitungan komponen korelasi menggunakan persamaan (12) dan hasilnya dapat dilihat pada **Tabel 8** sebagai berikut:

Tabel 8. Komponen Matriks Korelasi

Variabel	PC_1	PC_2	Variabel	PC_1	PC_2
X_1	-0,757	0,585	X_7	-0,977	-0,024
X_2	-0,879	0,390	X_8	-0,974	-0,106
X_3	-0,953	-0,005	X_9	-0,939	-0,254
X_4	-0,946	0,230	X_{10}	-0,075	-0,938
X_5	-0,989	0,077	X_{11}	-0,467	-0,824
X_6	-0,947	-0,268			

Korelasi variabel X_1 terhadap PC_1 sebesar -0,757 dan terhadap PC_2 sebesar 0,585. Karena korelasi variabel X_1 terhadap PC_1 lebih besar maka variabel X_1 tergabung dengan PC_1. Begitupun seterusnya untuk variabel yang lain.

f. Membentuk persamaan *principal component*
Berdasarkan hasil vektor eigen yang terbentuk, maka persamaan (13) dapat dituliskan sebagai berikut:

$$PC_{i,1} = -0,267Z_{i,1} - 0,310Z_{i,2} - 0,336Z_{i,3} - 0,334Z_{i,4} - 0,349Z_{i,5} - 0,334Z_{i,6} - 0,345Z_{i,7} - 0,344Z_{i,8} - 0,331Z_{i,9} - 0,027Z_{i,10} - 0,165Z_{i,11}$$

$$PC_{i,2} = 0,389Z_{i,1} + 0,260Z_{i,2} - 0,003Z_{i,3} + 0,153Z_{i,4} + 0,051Z_{i,5} - 0,178Z_{i,6} - 0,016Z_{i,7} - 0,071Z_{i,8} - 0,169Z_{i,9} - 0,624Z_{i,10} - 0,548Z_{i,11}$$

g. Menghitung transformasi himpunan data baru
Perhitungan transformasi himpunan data baru dengan PCA secara lengkap dapat dilihat pada **Tabel 9**.

Tabel 9. Hasil Transformasi Himpunan Data Baru

No.	PC_1	PC_2
1.	-4,295	1,787
2.	0,294	1,399
3.	-1,830	1,675
⋮	⋮	⋮
56.	1,556	-2,425

4.4 Analisis Pengelompokan dengan Metode *K-Means*

Analisis pengelompokan akan dilakukan berdasarkan hasil transformasi himpunan data baru pada perhitungan PCA dengan langkah-langkah berikut:

a. Menentukan jumlah klaster (C)
Pada langkah pertama, menentukan banyaknya klaster (C) yang akan diuji yaitu sebanyak 2, 3, 4, dan 5. Sebagai contoh, perhitungan pada penelitian ini dilakukan dengan menggunakan $C = 5$.

b. Menentukan pusat klaster (v_{cj}) awal secara acak
Pusat klaster awal dilakukan secara *trial* dan *error*. Pusat klaster awal dapat dilihat pada **Tabel 10** berikut:

Tabel 10. Pusat Klaster Awal untuk $C = 5$

x_i	Wilayah	Pusat klaster (v_{cj})	Variabel	
			PC_1	PC_2
1	Kabupaten Sambas	1	-4,295	1,787
2	Kabupaten Bengkayang	2	0,294	1,399
12	Kabupaten Kubu Raya	3	-7,046	1,554
9	Kabupaten Sekadau	4	1,746	1,315
28	Kota Palangka Raya	5	-0,554	-3,646

c. Menghitung jarak Euclid untuk setiap objek terhadap pusat klaster
Perhitungan jarak menggunakan jarak Euclid sesuai persamaan (2). Hasil perhitungan dapat dilihat pada **Tabel 11** berikut:

Tabel 11. Hasil Perhitungan Jarak Euclid terhadap Pusat Klaster Awal

Data ke-i	Jarak Euclid Data terhadap Pusat Klaster				
	v_1	v_2	v_3	v_4	v_5
1	0	4,606	2,760	6,060	6,597
2	4,606	0	7,341	1,455	5,116
3	2,468	2,142	5,217	3,595	5,472
⋮	⋮	⋮	⋮	⋮	⋮

DOI 10.30872/eksponensial.v15i2.1373

p-ISSN 2085-7829
e-ISSN 2798-3455

56	7,210	4,026	9,478	3,744	2,438

d. Mengalokasikan masing-masing objek ke klaster terdekat
Alokasi pusat klaster ditentukan berdasarkan jarak Euclid yang paling kecil. Hasil alokasi data dapat dilihat pada **Tabel 12** berikut:

Tabel 12. Hasil Alokasi Klaster

Data ke-i	Jarak Euclid Data terhadap Pusat Klaster					Alokasi Klaster
	v_1	v_2	v_3	v_4	v_5	
1	0	4,606	2,760	6,060	6,597	1
2	4,606	0	7,341	1,455	5,116	2
3	2,468	2,142	5,217	3,595	5,472	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
56	7,210	4,026	9,478	3,744	2,438	5

e. Memperbarui pusat klaster
Pusat klaster baru dihitung berdasarkan keanggotan klaster pada **Tabel 12** dengan menggunakan persamaan (3). Hasil perhitungan pusat klaster baru dapat dilihat pada **Tabel 13** berikut:

Tabel 13. Pusat Klaster Baru

Pusat klaster Baru	Variabel	
(v_{cj})	PC_1	PC_2
1	-3,341	1,110
2	-0,154	0,690
3	-7,292	-0,613
4	2,301	0,008
5	-0,854	-2,981

f. Mengulangi langkah 3,4, dan 5 sampai tidak ada objek yang berpindah klaster
Berdasarkan hasil perhitungan, pengelompokan dengan 5 pusat klaster dihentikan pada iterasi ke-5. Adapun hasil pengelompokan *k-means* dengan $C = 5$ dapat dilihat pada **Tabel 14**. sebagai berikut:

Tabel 14. Hasil Pengelompokan Kabupaten/Kota Menggunakan 5 Klaster

Klaster	Banyak anggota	Anggota Klaster	
		Kode	Kabupaten/Kota
1	8	1,3,5,6,7, 16, 17,31	Sambas, Landak, Sanggau, Ketapang, Sintang, Kotawaringin Timur, Kapuas, Banjar
2	16	2,4,8,10,15, 29,30,32,34 35,36,37, 38,42,43,45	Bengkayang, Mempawah, Kapuas Hulu, Melawi, Kotawaringin Barat, Tanah Laut, Kota Baru, Barito Kuala, Hulu Sungai Selatan, Hulu Sungai Tengah, Hulu Sungai Utara, Tabalong, Tanah Bumbu, Paser, Kutai Barat, Kutai Timur
3	6	12,13,40, 44,49,50	Kubu Raya, Pontianak, Banjarmasin, Kutai Kartanegara, Balikpapan, Samarinda
4	21	9,11,14,18, 19,20,21,22, 23,24,25,26, 27,33,39,47, 48,52,53,54,55	Sekadau, Kayong Utara, Singkawang, Barito Selatan, Barito Utara, Sukamara, Lamandau, Seruyan, Katingan, Pulang Pisau, Gunung Mas, Barito Timur, Murung Raya, Tapin, Balangan, Penajam Paser Utara, Mahakam Ulu, Malinau, Bulungan, Tana Tidung, Nunukan
5	5	28,41,46, 51,56	Palangka Raya, Banjar Baru, Berau, Bontang, Tarakan

g. Mengulangi langkah b hingga f untuk jumlah klaster sebanyak 2, 3, dan 4.

4.5 Validasi Klaster dengan *Davies-Bouldin Index*

Validasi klaster guna memastikan hasil pengelompokan sudah optimal dan akurat. Adapun tahapannya adalah sebagai berikut.

a. Menghitung nilai *Sum of Square Within* (SSW)
Langkah awal dalam menghitung nilai DBI yaitu menentukan terlebih dahulu nilai *Sum of Square Within* (SSW) menggunakan persamaan (14). Perhitungan nilai SSW dapat dilihat pada **Tabel 15** berikut:

Tabel 15. Nilai SSW pada 5 Klaster

Klaster ke-c	Nilai SSW_c
1	1,010
2	0,992
3	2,432
4	1,119

DOI 10.30872/eksponensial.v15i2.1373

p-ISSN 2085-7829
e-ISSN 2798-3455

Klaster ke-c	Nilai SSW_c
5	1,317

Sumber: Lampiran 27

b. Menghitung nilai *Sum of Square Between* (SSB)
Nilai SSB dapat dihitung dengan mengukur jarak antar pusat klaster v_c dan v_e menggunakan Persamaan (15). Adapun perhitungan nilai SSB dirangkum pada **Tabel 16** berikut.

Tabel 16. Nilai SSB pada 5 Klaster

SSB	C_1	C_2	C_3	C_4	C_5
C_1	0	3,243	4,444	5,559	5,664
C_2	3,243	0	6,360	2,322	3,232
C_3	4,444	6,360	0	8,364	6,679
C_4	5,559	2,322	8,364	0	3,024
C_5	5,664	3,232	6,679	3,024	0

c. Menghitung ukuran rasio antar klaster
Perhitungan ukuran rasio perbandingan antara klaster ke-c dan klaster-e (R_{ce}) berdasarkan rumus pada persamaan (16). Perhitungan ukuran rasio dapat dilihat pada **Tabel 17** sebagai berikut:

Tabel 17. Nilai SSB pada 5 Klaster

R_{ce}	C_1	C_2	C_3	C_4	C_5
C_1	–	0,618	0,775	0,383	0,411
C_2	0,618	–	0,539	0,909	0,715
C_3	0,775	0,539	–	0,425	0,561
C_4	0,383	0,909	0,425	–	0,805
C_5	0,411	0,715	0,561	0,805	–

d. Menghitung nilai DBI untuk 5 klaster
Setelah diperoleh nilai rasio perbandingan antar klaster, maka nilai DBI dapat dihitung dengan formulasi sesuai pada persamaan (17). Perhitungan nilai DBI untuk $C = 2,3,4$ dapat dilihat pada **Tabel 18** berikut:

Tabel 18. Hasil Perhitungan DBI

Jumlah Klaster	Nilai DBI
2	0,840
3	0,890
4	1,085
5	0,835

Berdasarkan **Tabel 18** dapat dilihat bahwa nilai DBI dari hasil pengelompokan sebanyak 5 klaster memiliki nilai paling kecil dibandingkan jumlah klaster lainnya, yaitu bernilai 0,835. Hal ini mengartikan bahwa pengelompokan dengan metode *k-means* akan menghasilkan klaster optimal jika menggunakan jumlah klaster sebanyak 5 dibandingkan jumlah klaster sebanyak 2, 3, dan 4.

4.6 Interpretasi Hasil Pengelompokan Klaster Optimal

Setelah kelompok terbentuk, langkah selanjutnya yaitu menghitung nilai rata-rata variabel untuk setiap cluster. Hasil perhitungan rata-rata dapat dilihat pada **Tabel 19** sebagai berikut:

Tabel 19. Nilai Rata-rata Variabel untuk Setiap Klaster

Variabel	Klaster				
	1	2	3	4	5
X_1	477,500	267,250	342,833	149,190	112,800
X_2	146,000	88,250	149,833	51,048	47,000
X_3	64,125	41,500	102,833	24,571	36,000
X_4	4.223,625	2.712,000	4.611,000	1.482,333	1.848,200
X_5	1.852,000	1.207,563	2.336,333	656,524	924,400
X_6	1.238,875	885,000	2.215,000	495,667	940,800
X_7	56.061,000	32.509,250	76.970,333	16.006,286	28.405,400
X_8	23.398,875	13.268,125	35.379,833	6.680,333	13.082,400
X_9	18.432,625	10.856,875	33.115,833	5.617,095	12.206,600
X_{10}	7,456	8,044	9,757	8,371	10,592
X_{11}	12,481	12,519	14,302	12,483	14,074

Berdasarkan Tabel 19, diketahui bahwa klaster 1 merupakan wilayah dengan jumlah SD/sederajat paling tinggi, namun memiliki rata-rata lama sekolah dan harapan lama sekolah paling rendah dibandingkan klaster lainnya. Klaster 3 merupakan wilayah dengan jumlah SMP, jumlah SMA, jumlah guru SD hingga SMA, jumlah murid SD hingga SMA, dan harapan lama sekolah paling tinggi dibandingkan klaster lainnya. Klaster 4 merupakan wilayah dengan jumlah SMA, jumlah guru SD hingga SMA, dan jumlah murid SD hingga SMA

DOI 10.30872/eksponensial.v15i2.1373

p-ISSN 2085-7829
e-ISSN 2798-3455

paling rendah dibandingkan klaster lainnya. Klaster 4 merupakan wilayah dengan jumlah SD dan SMP paling rendah, namun memiliki rata-rata lama sekolah paling tinggi dibandingkan klaster lainnya.

Adapun kabupaten/kota yang termasuk pada masing-masing klaster 1 hingga 5 dapat dilihat pada Tabel 20 berikut:

Tabel 20. Hasil Pengelompokan Kabupaten/Kota Menggunakan 5 Klaster

Klaster	Banyak anggota	Kabupaten/Kota
1	8	Sambas, Landak, Sanggau, Ketapang, Sintang, Kotawaringin Timur, Kapuas, Banjar
2	16	Bengkayang, Mempawah, Kapuas Hulu, Melawi, Kotawaringin Barat, Tanah Laut, Kota Baru, Barito Kuala, Hulu Sungai Selatan, Hulu Sungai Tengah, Hulu Sungai Utara, Tabalong, Tanah Bumbu, Paser, Kutai Barat, Kutai Timur
3	6	Kubu Raya, Pontianak, Banjarmasin, Kutai Kartanegara, Balikpapan, Samarinda
4	21	Sekadau, Kayong Utara, Singkawang, Barito Selatan, Barito Utara, Sukamara, Lamandau, Seruyan, Katingan, Pulang Pisau, Gunung Mas, Barito Timur, Murung Raya, Tapin, Balangan, Penajam Paser Utara, Mahakam Ulu, Malinau, Bulungan, Tana Tidung, Nunukan
5	5	Palangka Raya, Banjar Baru, Berau, Bontang, Tarakan

5. Kesimpulan

Berdasarkan hasil penelitian dan pembahasan, maka dapat diperoleh bahwa Klaster optimal yang terbentuk pada pengelompokan Kabupaten/Kota di Kalimantan menurut indikator pendidikan menggunakan metode *k-means* dengan optimasi PCA adalah sebanyak 5 klaster dengan nilai DBI sebesar 0,835. Klaster 1 terdiri dari 8 kabupaten/kota, klaster 2 terdiri dari 16 kabupaten/kota, klaster 3 terdiri dari 6 kabupaten/kota, klaster 4 terdiri dari 21 kabupaten/kota, dan klaster 5 terdiri dari 5 kabupaten/kota.

6. Daftar Pustaka

Amrullah, A., Purnamasari, I., Sari. B. N., Garno, Voutama, A. (2022). Analisis Cluster Faktor Penunjang Pendidikan Menggunakan Algoritma K-Means (Studi Kasus: Kabupaten Karawang). *Jurnal Informatika* & *Rekayasa Elektronika*, 5(2), 244 – 252.

Badan Pusat Statistik. (2023). *Statistik Indonesia 2023.* Jakarta: Badan Pusat Statistik Republik Indonesia.

Badruttamam, A., Sudarno, & Maruddani, D. I. (2020). Penerapan Analisis Klaster K-Modes dengan Validasi Davies Bouldin Index dalam Menentukan Karakteristik Kanal Youtube di Indonesia. *Jurnal GAUSSIAN, 9*(3), 263-272.

Bashori, & Aprima, S. G. (2019). Analisis Kebijakan Program Wajib Belajar 12 Tahun di Provinsi Lampung. *Jurnal Manajemen Pendidikan Islam, 1*(1), 18-28.

Gujarati, D. (2003). *Ekonometrika Dasar.* Jakarta: Erlangga.

Irwansyah, E., & Faisal, M. (2015). *Advanced Clustering: Teori dan Aplikasi.* Yogyakarta: DeePublish.

Nawari. (2010). *Analisis Regresi dengan MS Excel* . Jakarta: Elex Media Komputindo.

Nugroho, S. (2008). *Statistika Multivariat Terapan.* Bengkulu: UNIB Press.

Prasetyo, E. (2012). *Data Mining: Konsep dan Aplikasi Menggunakan Matlab.* Yogyakarta: Andi Offset.

Rahmayanti, A., Juita, R., & Suhendra, C. D. (2022). Penerapan Metode K-Means untuk Clustering Data Anak Berdasarkan Kepemilikan Akta Kelahiran dan KIA. *Jurnal Informatik, 7*(3), 210-219.

Santosa, B., & Umam, A. (2018). *Data Mining dan Big Data Analytics.* Yogyakarta: Penebar Media Pustaka.

Sopyan, Y., Lesmana, A. D., & Juliane, C. (2022). Analisis Algoritma K-Means dan Davies Bouldin Index dalam Mencari Cluster Terbaik Kasus Perceraian di Kabupaten Kuningan. *Building of Informatics, Technology and Science, 4*(3), 1464-1470.

Supranto, J. (2004). *Analisis Multivariat: Arti dan Interpretasi.* Jakarta: Rineka Cipta.

Suyanto. (2017). *Data Mining untuk Klasifikasi dan Klasterisasi Data.* Bandung: Informatika .

Ulinnuha, N., Veriani, R. (2020). Analisis Cluster dalam Pengelompokan Provinsi di Indonesia Berdasarkan Penyakit Menular Menggunakan Metode Complete Linkage, Average Linkage, dan Ward. *Jurnal Nasional Informatika dan Teknologi Jaringan,* 5(1). 101 - 108.

Wangge, M. (2021). Penerapan Metode Principal Component Analysis (PCA) Terhadap Faktor-Faktor yang Memengaruhi Lamanya Penyelesaian Skripsi Mahasiswa Program Studi Pendidikan Matematika FKIP UNDANA. *Jurnal Cendekia: Jurnal Pendidikan Matematika,* 5(1). 974 - 988.

Wanto, A., Siregar, M. N., Windarto, A. P., Hartama, D., Ginantra, L. W., Napitupulu, D., . . . Prianto, C. (2020). *Data Mining: Algoritma dan Implementasi.* Medan: Yayasan Kita Menulis.