

Pengelompokan Algoritma *K-Medoids* Dengan *Principal Component Analysis* (PCA) (Studi Kasus : Kabupaten/Kota di Pulau Kalimantan Berdasarkan Indikator Kemiskinan Tahun 2021)

Muhammad Yafi¹, Rito Goejantoro², Andrea Tri Rian Dani³

^{1,2,3}Program Studi Statistika, Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Mulawarman, Indonesia

Corresponding author: yyafimuhammad@gmail.com

Abstrak. Analisis kluster adalah salah satu teknik yang dikenal dalam *data mining* yaitu sebuah metode yang bertujuan untuk mengelompokkan data (objek) berdasarkan informasi dalam data. Penelitian ini menggunakan pengelompokan non-hierarki yaitu algoritma *K-Medoids* untuk mengelompokkan Kabupaten/kota di Pulau Kalimantan berdasarkan indikator kemiskinan dan metode *Principal Component Analysis* (PCA) untuk mereduksi variabel penelitian. Pada penelitian ini juga dilakukan uji validitas kluster untuk melihat berapa jumlah kluster yang memiliki pengelompokan terbaik menggunakan metode *Silhouette Coefficient* (SC). Berdasarkan hasil analisis diketahui bahwa data direduksi menjadi sebanyak 3 *Principal Component* (PC) dengan menggunakan metode PCA dan didapatkan hasil pengelompokan Kabupaten/kota di Pulau Kalimantan berdasarkan indikator kemiskinan menggunakan algoritma *K-Medoids* yang paling optimal adalah dengan kluster sebanyak 2 dan nilai SC sebesar 0,612. Algoritma *K-Medoids* menghasilkan 2 kluster, di mana kluster 1 beranggotakan 49 Kabupaten/kota dan kluster 2 beranggotakan 7 Kabupaten/kota.

Kata Kunci: *Kemiskinan, Kluster, K-Medoids, PCA, Silhouette Coefficient*

1 PENDAHULUAN

Data Mining adalah proses pengumpulan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang besar (*Big Data*) sehingga dapat diperoleh suatu informasi atau pengetahuan baru. *Data Mining* sering juga disebut sebagai *Knowledge Discovery in Database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan dan penggunaan data untuk menemukan keteraturan suatu pola atau hubungan dalam data yang berukuran besar (*Big Data*) [1].

Salah satu teknik yang dikenal dalam *data mining* adalah analisis kluster. Analisis kluster merupakan metode yang bertujuan untuk mengelompokkan data (objek) yang didasarkan pada informasi data. Analisis kluster dibagi menjadi dua, yaitu pengelompokan hierarki dan pengelompokan non-hierarki. Prinsip dari pengelompokan hierarki adalah dimulai dari satu data tunggal dikelompokkan menjadi satu kelompok, dua kelompok atau lebih kelompok kecil yang dapat bergabung menjadi sebuah kelompok besar dan begitu seterusnya [2].

Kelebihan pengelompokan non-hierarki adalah lebih cepat dalam proses komputasi ketika mengelompokkan objek atau observasi yang jumlahnya besar serta memiliki tingkat efisiensi waktu yang lebih baik dibandingkan pengelompokan hierarki karena pengelompokan non-hierarki menentukan terlebih dahulu jumlah kluster diawal. Kemudian objek-objek tersebut akan didistribusikan ke kluster yang sudah ditentukan berdasarkan kemiripan atau kesamaan karakteristiknya [3].

Seiring perkembangan teknologi informasi, keberagaman informasi dijelaskan menggunakan variabel. Semakin banyak variabel yang digunakan maka semakin besar kemungkinan untuk terjadinya masalah kompleksitas dalam proses pengelompokan. Sehingga, penulis tertarik untuk melakukan reduksi variabel dengan tanpa mengurangi banyaknya informasi dari variabel tersebut. Hal ini dapat dilakukan dengan menggunakan metode *Principal Component Analysis* (PCA). PCA adalah salah satu metode statistik yang secara linier mengubah bentuk suatu variabel menjadi kumpulan-kumpulan variabel yang lebih kecil dan tidak berkorelasi yang dapat mewakili informasi dari variabel aslinya [4].

Prinsip dari PCA adalah menemukan nilai eigen dan vektor eigen yang digunakan untuk menyederhanakan atau mereduksi dimensi. Hal ini dilakukan dengan cara menghilangkan korelasi di antara variabel melalui proses transformasi variabel menjadi suatu variabel baru yang tidak berkorelasi sama sekali atau yang biasa disebut *principal component* (PC). Besarnya vektor eigen dan nilai eigen memiliki peranan yang paling penting karena vektor eigen dan nilai eigen yang sangat kecil akan dibuang sehingga data dapat direduksi tanpa menghilangkan banyak informasi yang penting [5].

Analisis kluster telah banyak digunakan di berbagai bidang seperti kemiskinan, kesehatan, pendidikan, ketenagakerjaan dan lain sebagainya. Salah satunya kemiskinan yang merupakan fenomena yang terjadi hampir di seluruh negara khususnya pada negara yang sedang berkembang termasuk Indonesia. Kemiskinan disebabkan oleh ketidakmampuan seseorang dari segi ekonomi untuk memenuhi kebutuhan dasarnya seperti makanan, pakaian, tempat tinggal, pendidikan dan kesehatan [6]. Tolak ukur kemiskinan bukan hanya berdasarkan pemenuhan kebutuhan pangan dan tingkat pendapatan yang rendah, akan tetapi

juga melihat dari sisi kesehatan, pendidikan dan perlakuan adil di muka hukum dan sebagainya [7].

2 TINJAUAN PUSTAKA

2.1 *Principal Component Analysis*

Principal Component Analysis (PCA) adalah sebuah teknik analisis multivariat dalam statistik yang berfungsi untuk mengurangi dimensi, mendeteksi multikolinieritas, dan mengklasifikasikan. Prinsip dari PCA adalah mentransformasikan variabel awal yang kemungkinan besar saling berkorelasi antar dimensinya menjadi variabel baru yang tidak berkorelasi. PCA memberikan hasil yang baik ketika diterapkan pada variabel yang berkorelasi. Oleh karena itu, PCA merupakan suatu teknik seleksi data multivariabel yang mengubah atau mentransformasi suatu matriks data asli menjadi suatu kumpulan kombinasi homogen yang lebih sedikit (reduksi) namun menyerap sejumlah besar varian dari data awal [8].

Secara umum langkah-langkah reduksi variabel dengan menggunakan metode PCA adalah sebagai berikut:

1. Melakukan standarisasi data

Standarisasi yang digunakan adalah standarisasi Z_{score} , cara menghitung standarisasi Z_{score} dapat menggunakan Persamaan (1).

$$Z_{ap} = \frac{x_{ap} - \bar{x}_p}{S_p} \quad (1)$$

dengan rata-rata untuk setiap variabel dapat menggunakan Persamaan (2).

$$\bar{x}_p = \frac{1}{n} \sum_{a=1}^n x_{ap}, a = 1, 2, 3, \dots, n \text{ dan } p = 1, 2, 3, \dots, j \quad (2)$$

dan deviasi standar ditampilkan pada Persamaan (3).

$$S_p = \sqrt{\frac{1}{n-1} \sum_{a=1}^n (X_{ap} - \bar{X}_p)^2} \quad (3)$$

2. Menghitung matriks varians-kovarians

Formula menghitung matriks varians-kovarians:

$$\sum_z = \begin{bmatrix} S_{11}^2 & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22}^2 & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{p10}^2 \end{bmatrix}$$

Formula perhitungan varians ditampilkan pada Persamaan (4).

$$S_p^2 = \frac{1}{n-1} \sum_{a=1}^n (x_{ap} - \bar{x}_p)^2 \quad (4)$$

Formula perhitungan kovarians ditampilkan pada Persamaan (5).

$$S_{x_p x_{p+1}} = \frac{\sum_{a=1}^n (x_{ap} - \bar{x}_p)(x_{ap+1} - \bar{x}_{p+1})}{n-1} \quad (5)$$

3. Menghitung matriks korelasi

Formula untuk menghitung matriks korelasi ditampilkan pada Persamaan (6).

$$\mathbf{R}_z = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

dengan

$$\rho_{x_p x_{p+1}} = \frac{S_{x_p x_{p+1}}}{S_{x_p} S_{x_{p+1}}} \quad (6)$$

4. Menghitung nilai eigen dan vektor eigen

Menghitung nilai eigen (λ) dan vektor eigen (\vec{v}) dari matriks korelasi (\mathbf{R}_z) dengan formula ditampilkan pada Persamaan (7).

$$|\lambda \mathbf{I} - \mathbf{R}_z| = \mathbf{0} \quad (7)$$

5. Agar dapat menjadi nilai eigen, harus terdapat satu solusi tak nol dari $|\lambda \mathbf{I} - \mathbf{R}_z| = \mathbf{0}$. Hal ini terpenuhi jika dan hanya jika memenuhi formula pada Persamaan (7) berikut.

$$\det|\lambda \mathbf{I} - \mathbf{R}_z| = \mathbf{0} \quad (7)$$

di mana \mathbf{I} adalah matriks identitas \vec{v} adalah vektor eigen.

6. Menentukan jumlah *principal component* (PC) yang mungkin terbentuk dengan melihat kriteria nilai eigen yang lebih besar atau sama dengan 1. Nilai eigen yang kurang dari satu dianggap memiliki kontribusi yang kurang. Metode ini hanya bisa diterapkan pada matriks korelasi.

7. Membentuk komponen matriks korelasi yang menunjukkan besarnya korelasi variabel terhadap skor komponen yang terbentuk menggunakan Formula pada Persamaan (9).

$$r_{x_p, PC_t} = \vec{v}_{at} \sqrt{\lambda_t} \quad (9)$$

8. Membentuk variabel baru hasil reduksi

- Membentuk variabel baru berdasarkan kombinasi linier dengan Formula pada Persamaan (10).

$$PC_{a,t} = \sum_{p=10} \vec{v}_p z_p \quad (10)$$

dengan \vec{v} adalah vektor eigen.

2.2 Analisis Kluster

Analisis kluster merupakan metode yang bertujuan untuk mengelompokkan data (objek) yang didasarkan pada informasi data. Informasi tersebut dapat menjelaskan karakteristik dari objek yang diteliti. Tujuan dari analisis kluster adalah mengelompokkan objek-objek bergabung ke dalam satu kluster, sehingga memiliki kemiripan satu sama lain (Homogenitas) dalam satu kluster. Namun berbeda dengan objek yang berada pada kluster yang lain Prasetyo, 2012 [2]. Analisis kluster terbagi menjadi dua yaitu pengelompokan hierarki dan pengelompokan non-hierarki [8].

2.3 Pengukuran Kemiripan

Terdapat beberapa pengukuran jarak yang dapat digunakan salah satunya adalah jarak Euclid (*Euclidean Distance*). *Euclidean Distance* adalah jarak yang efisien untuk digunakan pada data numerik. *Euclidean Distance* juga dinilai bekerja secara efisien untuk menghitung kemiripan dalam pengelompokan objek-objek berdasarkan kemiripan maupun ketidakmiripan objek-objek tersebut. Mohammed dan Abdulazeez, 2007 [9]. Formula untuk menghitung jarak Euclid adalah sebagai berikut:

$$d(x_u, y_v) = \sqrt{\sum_{u,v}^n (x_u - y_v)^2} \quad u, v = 1, 2, \dots, n \quad (11)$$

2.4 Algoritma K-Medoids

K-Medoids atau *Partitioning Around Medoids* (PAM) merupakan metode berbasis partisi dalam mengelompokkan objek-objek menjadi sejumlah kluster. Objek yang terpilih untuk mewakili sebuah kluster disebut *medoids*. *K-Medoids* adalah pengembangan dari algoritma *K-Means* yang sensitif terhadap *outlier*. Selain dapat mengatasi *outlier*, *K-Medoids* juga termasuk algoritma yang fleksibel karena hampir dapat bekerja pada setiap jenis data matriks dan mampu mengelompokkan objek dalam jumlah yang besar dengan waktu komputasi yang cepat Kaufman dan Rousseeuw, 1990 [10]. Tahapan-tahapan dari algoritma *K-Medoids* dirincikan sebagai berikut:

- Memilih secara acak objek sebanyak k sebagai objek representatif o_{mj} (*medoids*)
- Menghitung *Euclidean distance* untuk setiap objek terhadap masing-masing *medoids* seperti dinyatakan oleh Persamaan (12) sebagai berikut:

$$d(x_{ap}, o_{mp}) = \sqrt{(x_{a1} - o_{m1})^2 + (x_{a2} - o_{m2})^2 + \dots + (x_{aj} - o_{mj})^2} \quad (12)$$

dengan $d(x_{ap}, o_{mp})$ adalah jarak dari data ke- a pada variabel ke- p terhadap *medoids* ke- m , di mana $a = 1, 2, \dots, n$ dan $p = 1, 2, \dots, j$ serta $m = 1, 2, \dots, k$.

3. Menetapkan setiap objek ke gugus yang sesuai dengan *medoids* terdekat dan menghitung fungsi objektif yang merupakan jumlah ketidakmiripan dari semua objek ke *medoids* terdekat berdasarkan jarak antara objek terhadap setiap *medoids* yang paling minimum.
4. Memilih secara acak objek yang tidak representatif o_{hj} (non-*medoids*).
5. Menghitung *Euclidean distance* untuk setiap objek terhadap masing-masing non-*medoids* seperti dinyatakan oleh Persamaan (13) sebagai berikut:

$$d(x_{ap}, o_{hp}) = \sqrt{(x_{a1} - o_{h1})^2 + (x_{a2} - o_{h2})^2 + \dots + (x_{aj} - o_{hj})^2} \quad (13)$$

dengan $d(x_{ap} - o_{hp})$ adalah jarak dari data ke- a pada variabel ke- p terhadap objek non-*medoids* ke- h pada variabel ke- p di mana $a = 1, 2, \dots, n$ dan $p = 1, 2, \dots, j$ serta $h = 1, 2, \dots, n-k$.

6. Menetapkan setiap objek ke gugus yang sesuai dengan non-*medoids* terdekat dan menghitung fungsi objektif yang merupakan jumlah ketidaksamaan dari semua objek ke non-*medoids* terdekat berdasarkan jarak antara objek terhadap setiap *medoids* yang paling minimum.
7. Menghitung selisih dari fungsi objektif dengan cara mengurangkan fungsi objektif non-*medoids* dengan fungsi objektif *medoids*.
8. Mengulangi langkah (4-7) sampai tidak ada lagi perubahan objek representatif. Analisis selesai jika sudah tidak terdapat perubahan objek representatif.

2.5 *Silhouette Coefficient*

Salah satu metode evaluasi yang dapat digunakan untuk melihat kualitas dan kekuatan kluster adalah metode *silhouette coefficient*. Hasil perhitungan nilai SC dapat bervariasi antara -1 hingga 1. SC yang bernilai 1 atau mendekati 1 artinya objek sudah berada dalam kluster yang tepat. Jika nilai SC bernilai 0 maka objek berada di antara dua kluster sehingga objek tersebut tidak diketahui harus dimasukkan ke dalam kluster A atau kluster B. Sedangkan jika SC bernilai -1 artinya terdapat banyak objek yang tidak berada dalam kluster yang tepat [11].

3 DATA

3.1 Populasi dan Sampel Penelitian

Populasi yang digunakan dalam penelitian ini adalah indikator kemiskinan di antaranya (rata-rata lama sekolah, harapan lama sekolah, pengeluaran perkapita, persentase penduduk miskin, persentase rumah tangga dengan akses sanitasi layak, persentase rumah tangga dengan akses air minum/bersih yang layak, angka harapan hidup, kepadatan penduduk, tingkat pengangguran terbuka, dan prevalensi balita stunting) di 56 kabupaten/kota Pulau Kalimantan. Sampel yang digunakan dalam penelitian ini adalah indikator kemiskinan di 56 kabupaten/kota yang ada di Pulau Kalimantan tahun 2021.

3.2 Variabel Penelitian

Variabel yang digunakan dalam penelitian ini adalah indikator kemiskinan yang dirincikan pada Tabel 1.

Tabel 1 Variabel penelitian

Variabel	Satuan	Notasi	Dimensi
Rata-rata lama sekolah	Tahun	X_1	Pendidikan
Harapan lama sekolah	Tahun	X_2	Pendidikan
Pengeluaran per kapita	Rupiah (Rp)	X_3	Standar Hidup Layak
Persentase penduduk miskin	Persentase (%)	X_4	Standar Hidup Layak
Persentase rumah tangga dengan akses sanitasi layak	Persentase (%)	X_5	Standar Hidup Layak
Persentase rumah tangga dengan akses air minum/bersih yang layak	Persentase (%)	X_6	Standar Hidup Layak
Angka harapan hidup	Umur	X_7	Kesehatan
Kepadatan penduduk	Jiwa/Km	X_8	Standar Hidup Layak
Tingkat pengangguran terbuka	Persentase (%)	X_9	Standar Hidup Layak
Prevalensi balita <i>stunting</i>	Persentase (%)	X_{10}	Kesehatan

4 HASIL DAN PEMBAHASAN

4.1 Statistika Deskriptif

Hasil pengolahan statistika deskriptif data indikator kemiskinan di 56 Kabupaten/kota yang ada di Pulau Kalimantan pada tahun 2021 dapat dilihat pada Tabel 2.

Tabel 2 Statistika Deskriptif

Variabel	Banyak data	Minimum	Maksimum	Rata-rata
X_1	56	6,020	11,530	8,408
X_2	56	11,170	15,090	12,780
X_3	56	7.064	16.765	10.751
X_4	56	2,890	12,010	6,298
X_5	56	49,230	97,510	79,710
X_6	56	48,850	99,920	77,150
X_7	56	64,100	74,760	70,760
X_8	56	1,700	9198,890	407,350
X_9	56	2,300	12,380	4,953
X_{10}	56	14,200	35,900	24,370

4.2 Principal Component Analysis

Principal Component Analysis (PCA) adalah suatu teknik handal yang berfungsi mengekstraksi struktur dari suatu *set* data dengan dimensi yang cukup banyak. PCA dapat mengurangi besarnya dimensi dari data yang diobservasi menjadi dimensi yang lebih kecil tanpa kehilangan informasi yang signifikan dalam menggambarkan keseluruhan data. Adapun langkah-langkah dalam metode PCA adalah sebagai berikut:

1. Standarisasi data

Pada penelitian ini standarisasi yang digunakan adalah standarisasi Z_{score} , dengan menggunakan standarisasi Z_{score} data akan berada pada rentang yang sama di mana rata-ratanya bernilai 0 dan deviasi standarnya bernilai 1. Hasil standarisasi data menggunakan Z_{score} dapat dilihat pada Tabel 3 sebagai berikut:

Tabel 3 Data Hasil Standarisasi

Kab/kota	x_1	x_2	...	x_{10}
Paser	0,312	0,546	...	-0,159
Kutai Barat	0,238	0,281	...	-1,784
Kutai Kartanegara	0,672	0,949	...	0,423
Kutai Timur	0,835	0,143	...	0,652
⋮	⋮	⋮	...	⋮
Tarakan	1,285	1,444	...	0,319

2. Menghitung matriks korelasi

Matriks korelasi akan bernilai sama dengan matriks varians-kovarians dari data yang sudah distandarisasi. Sehingga dengan bantuan *software R studio* matriks korelasi yang akan terbentuk adalah sebagai berikut:

$$R_z = \begin{pmatrix} 1 & 0,778 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & -0,032 \\ 0,778 & 1 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & -0,193 \\ 0,701 & 0,608 & \ddots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & -0,077 \\ -0,279 & -0,273 & \dots & \ddots & \dots & \dots & \dots & \dots & \dots & \dots & -0,022 \\ 0,495 & 0,489 & \dots & \dots & \ddots & \dots & \dots & \dots & \dots & \dots & -0,245 \\ 0,376 & 0,547 & \dots & \dots & \dots & \ddots & \dots & \dots & \dots & \dots & -0,106 \\ 0,407 & 0,354 & \dots & \dots & \dots & \dots & \ddots & \dots & \dots & \dots & -0,218 \\ 0,382 & 0,437 & \dots & \dots & \dots & \dots & \dots & \ddots & \dots & \dots & -0,136 \\ 0,486 & 0,616 & \dots & \dots & \dots & \dots & \dots & \dots & \ddots & \dots & -0,358 \\ -0,032 & -0,193 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 1 \end{pmatrix}_{(10 \times 10)}$$

3. Menghitung nilai eigen dan vektor eigen

Menghitung nilai eigen (λ) dan vektor eigen (\vec{v}) dengan menggunakan *software R studio* dengan hasil sebagai berikut:

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \\ \lambda_6 \\ \lambda_7 \\ \lambda_8 \\ \lambda_9 \\ \lambda_{10} \end{bmatrix} = \begin{bmatrix} 4,356 \\ 1,509 \\ 0,990 \\ 0,760 \\ 0,710 \\ 0,596 \\ 0,416 \\ 0,284 \\ 0,238 \\ 0,136 \end{bmatrix}_{(10 \times 1)}$$

dan vektor eigen (\vec{v}) sebagai berikut:

$$\begin{pmatrix} -0,388 & 0,037 & -0,429 & 0,135 & 0,259 & -0,030 & 0,199 & -0,182 & 0,035 & 0,708 \\ -0,411 & -0,010 & -0,159 & 0,069 & 0,059 & 0,105 & 0,600 & 0,066 & 0,384 & -0,522 \\ -0,369 & 0,308 & -0,088 & 0,148 & 0,313 & -0,221 & -0,141 & 0,200 & -0,663 & -0,301 \\ 0,195 & -0,602 & -0,245 & -0,177 & -0,123 & -0,348 & 0,380 & 0,323 & -0,344 & 0,048 \\ -0,320 & -0,180 & -0,060 & 0,377 & -0,415 & -0,568 & -0,372 & 0,001 & 0,285 & -0,055 \\ -0,323 & 0,111 & 0,160 & -0,027 & -0,733 & 0,300 & 0,215 & -0,129 & -0,384 & 0,125 \\ -0,205 & -0,570 & -0,304 & -0,046 & 0,090 & 0,449 & -0,398 & -0,311 & -0,131 & -0,233 \\ -0,298 & 0,011 & 0,185 & -0,769 & 0,074 & -0,375 & -0,029 & -0,365 & 0,047 & -0,036 \\ -0,384 & -0,143 & 0,248 & -0,239 & 0,064 & 0,242 & -0,221 & 0,722 & 0,175 & 0,220 \\ 0,135 & 0,386 & -0,710 & -0,357 & -0,294 & 0,052 & -0,202 & 0,220 & 0,113 & -0,089 \end{pmatrix}$$

5 Menentukan jumlah *principal component* (PC) yang mungkin terbentuk

Pada metode *principal component analysis* (PCA) terdapat beberapa metode untuk melihat berapa banyak PC yang terbentuk salah satunya adalah dengan melihat nilai eigen. Banyaknya PC yang akan digunakan berkaitan dengan besarnya nilai eigen yang dihasilkan, banyaknya nilai eigen yang bernilai ≥ 1 berbanding lurus dengan banyaknya PC yang akan terbentuk. Sehingga pada penelitian ini diperoleh tiga nilai eigen yang memenuhi kriteria tersebut, yaitu $\lambda_1 = 4,356$, $\lambda_2 = 1,509$ dan $\lambda_3 = 0,990$.

6 Membentuk komponen matriks korelasi yang menunjukkan besarnya korelasi variabel terhadap skor komponen

Tabel 4 Grafik Komponen Matriks Korelasi

	\mathbf{r}_{xp,PC_1}	\mathbf{r}_{xp,PC_2}	\mathbf{r}_{xp,PC_3}
x_1	-0,811	-0,047	-0,427
x_2	-0,859	0,012	-0,158
x_3	-0,771	-0,379	-0,088
x_4	-0,408	0,740	-0,244
x_5	0,668	0,221	-0,060
x_6	0,676	-0,137	0,160
x_7	0,429	0,700	-0,303
x_8	0,623	-0,015	0,184
x_9	0,802	0,176	0,247
x_{10}	0,282	-0,475	-0,706

Berdasarkan Tabel 4 dapat diketahui bahwa pada variabel rata-rata lama sekolah (x_1), harapan lama sekolah (x_2), pengeluaran per kapita (x_3), persentase rumah tangga dengan akses sanitasi layak (x_5), persentase rumah tangga dengan akses air minum/bersih yang layak (x_6), kepadatan penduduk (x_8) dan variabel tingkat pengangguran terbuka (x_9) sebagian besar data berdistribusi ke PC_1 . Sedangkan pada variabel persentase penduduk miskin (x_4) dan umur harapan hidup (x_7) sebagian besar berdistribusi ke PC_2 . Kemudian pada variabel prevalensi balita *stunting* (x_{10}) sebagian besar berdistribusi ke PC_3 .

6 Membentuk variabel baru hasil reduksi

Berdasarkan hasil perhitungan, maka diperoleh persamaan untuk *principal component* (PC) yang terbentuk dan diperoleh transformasi data *set* baru yang dapat dilihat pada Tabel 5.

Tabel 5 Variabel Baru Hasil Reduksi Menggunakan PCA

Kab/kot	PC_1	PC_2	PC_3
Paser	-0,029	-1,652	-0,861
Kutai Barat	-0,158	-2,584	0,253
Kutai Kartanegara	-0,835	-0,789	-1,158
Kutai Timur	-0,474	-1,379	-1,579
⋮	⋮	⋮	⋮
Tarakan	-1,911	-0,440	-1,161

4.3 Algoritma *K-Medoids*

Algoritma *K-Medoids* merupakan salah satu pengelompokan non-hierarki di mana, diperlukan penentuan banyak kluster di awal dan juga titik pusat (*medoids*) untuk mengelompokkan objek penelitian.

4.4 Validasi Hasil Pengelompokan dengan Algoritma *K-Medoids*

Pada penelitian ini akan dilakukan perhitungan nilai *Silhouette Coefficient* (SC) untuk mengetahui kualitas dari setiap hasil pengelompokan. Hasil Validasi Kluster Berdasarkan Nilai SC_{global} dapat dilihat pada Tabel 6.

Tabel 6 Perbandingan Hasil Validasi Kluster Berdasarkan Nilai SC_{global}

Banyak Kluster	SC_{global}
2	0,612
3	0,319
4	0,308
5	0,285

Berdasarkan Tabel 6 dapat diketahui bahwa nilai SC_{global} untuk validasi data hasil pengelompokan Kabupaten/kota di Pulau Kalimantan berdasarkan indikator kemiskinan dengan menggunakan algoritma *K-Medoids* memiliki nilai SC_{global} yang berbeda-beda. Nilai SC_{global} yang terbesar adalah pengelompokan pada 2 kluster yaitu 0,612. Oleh karena itu, dapat diputuskan bahwa pengelompokan yang

paling optimal pada pengelompokan Kabupaten/kota di Pulau Kalimantan dengan algoritma *K-Medoids* adalah pada $k=2$.

4.5 Pengelompokan algoritma *K-Medoids* untuk $K=2$

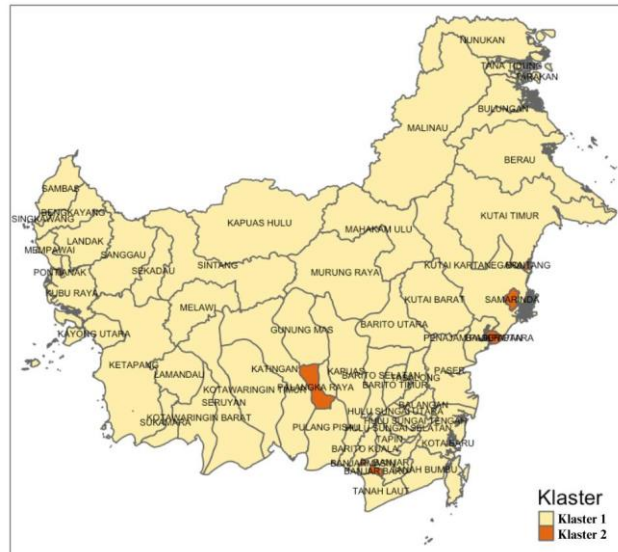
Melakukan pengelompokan menggunakan algoritma *K-Medoids* dengan k sebanyak 2 kluster pada 56 Kabupaten/kota di Pulau Kalimantan berdasarkan indikator kemiskinan. Hasil pengelompokan dapat dilihat pada Tabel 7.

Tabel 7 Anggota Kluster dengan Algoritma *K-Medoids* Untuk $K=2$

No	Kabupaten/kota	Kluster	No	Kabupaten/kota	Kluster
1	Paser	1	29	Tapin	1
2	Kutai Barat	1	30	Hulu Sungai Selatan	1
3	Kutai Kartanegara	1	31	Hulu Sungai Tengah	1
4	Kutai Timur	1	32	Hulu Sungai Utara	1
5	Berau	1	33	Tabalong	1
6	Penajam Paser Utara	1	34	Tanah Bumbu	1
7	Mahakam ulu	1	35	Balangan	1
8	Balikpapan	2	36	Banjarmasin	2
9	Samarinda	2	37	Banjar Baru	2
10	Bontang	2	38	Kotawaringin Barat	1
11	Sambas	1	39	Kotawaringin Timur	1
12	Bengkayang	1	40	Kapuas	1
13	Landak	1	41	Barito Selatan	1
14	Mempawah	1	42	Barito Utara	1
15	Sanggau	1	43	Sukamara	1
16	Ketapang	1	44	Lamandau	1
17	Sintang	1	45	Seruyan	1
18	Kapuas Hulu	1	46	Katingan	1
19	Sekadau	1	47	Pulang Pisau	1
20	Melawi	1	48	Gunung Mas	1
21	Kayong Utara	1	49	Barito Timur	1
22	Kubu Raya	1	50	Murung Raya	1
23	Kota Pontianak	2	51	Palangkaraya	2
24	Kota Singkawang	1	52	Malinau	1
25	Tanah Laut	1	53	Bulungan	1
26	Kota Baru	1	54	Tana Tidung	1
27	Banjar	1	55	Nunukan	1
28	Barito Kuala	1	56	Tarakan	1

4.6 Profilisasi dan Interpretasi Hasil Pengelompokan Terbaik

Pada penelitian ini dilakukan profilisasi dari kluster yang paling optimal dan dengan menggunakan uji validitas *Silhouette Coefficient* didapatkan kluster yang paling optimal adalah kluster sebanyak 2. Profilisasi hasil pengelompokan terbaik dapat dilihat pada Gambar 1 sebagai berikut:



Gambar 1 Peta Profilisasi Hasil Pengelompokan Terbaik

Pada Gambar 1 di atas dapat dilihat Klaster 1 yang berwarna kuning memiliki kecenderungan diisi oleh Kabupaten-kabupaten yang belum semaju perkotaan. Sedangkan pada Klaster 2 yang berwarna jingga cenderung diisi oleh Kota-kota besar yang ada di Pulau Kalimantan. Oleh karena itu dapat disimpulkan bahwa terdapat kesenjangan indikator kemiskinan antara wilayah perkotaan dengan wilayah kabupaten-kabupaten yang belum semaju perkotaan.

5 KESIMPULAN

5.1 Kesimpulan

Berdasarkan hasil penelitian dan pembahasan, maka dapat diambil kesimpulan sebagai berikut:

1. *Principal component* yang terbentuk dari hasil reduksi variabel menggunakan metode PCA dengan memperhatikan nilai eigen yang lebih besar atau sama dengan satu adalah sebanyak 3 PC.
2. Berdasarkan metode *silhouette coefficient* didapatkan bahwa nilai k yang paling optimal dari pengelompokan Kabupaten/kota di Pulau Kalimantan berdasarkan indikator kemiskinan adalah dengan menggunakan 2 klaster ($k=2$) dengan nilai SC_{global} sebesar 0,612.
3. Hasil pengelompokan Kabupaten/kota di Pulau Kalimantan berdasarkan indikator kemiskinan dihasilkan sebanyak 2 klaster. Klaster 1 beranggotakan 49 Kabupaten/kota, sedangkan klaster 2 beranggotakan 7 Kabupaten/kota.

DAFTAR PUSTAKA

- [1] Santoso, B., (2007), *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Yogyakarta: Graha Ilmu.
- [2] Prasetyo, E., (2012), *Data Mining: Konsep Dan Aplikasi Menggunakan MATLAB*, Yogyakarta: Penerbit ANDI.
- [3] Singh, N dan Singh, D., (2012), *Performance evaluation of K-means and Heirarichal clustering in terms of accuracy and running time.*

International journal in computer science and information technology, Vol. 3, Hal. 4119-4121.

- [4] Umar, H. B., (2009), *Principal Component Analysis (PCA) Dan Aplikasinya Dengan SPSS. Jurnal kesehatan*. Vol. 3, No. 2, Hal. 97-101.
- [5] Smith, L. I., (2002), A Tutorial on Principal Component Analysis. *Computer Science Technical Report*, Vol. 1, No. 2, Hal. 1-26.
- [6] Nasution, M. Z., (2019), Penerapan *Principal Component Analysis (PCA)* Dalam Penentuan Faktor Dominan yang Mempengaruhi Prestasi Belajar Siswa. *Jurnal Teknologi Informasi*. Vol. 3, No. 1, Hal. 41–48.
- [7] Pramana, S., Yuniarto, B., Mariyah, S., Santoso, I., Nooraeni, R., (2018). *Data Mining dengan R: Konsep serta Implementasi*. Bogor: IN MEDIA.
- [8] Ghaisani, S. Y., Hikmah, N., Prasetyo, A. H., Widodo, E., (2018), Analisis Cluster Hirarki Untuk Pengelompokan Provinsi Di Indonesia Berdasarkan Indikator Demokrasi Indonesia Tahun 2016. *Konferensi Nasional Penelitian Matematika dan Pembelajarannya (KNPMP) IV*, Surakarta, 2019.
- [9] Mohammed, N. N., dan Abdulazeez, A. M., (2007), Evaluation of partitioning around medoids algorithm with various distances on microarray data," in *IEEE International Conference on Internet of Things (iThings)*, Exeter, UK, pp. 1011-1016. doi: 10.1109/iThings-GreenCom-CPSCoM-SmartData.2017.155
- [10] Kaufman, L. dan Rousseeuw, P. J., (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York.
- [11] Pramesti, D. F., Furqon, M. T., dan Dewi, C., (2017), Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* e-ISSN, 2548, 964X.Y. Vol. 1, No. 9, Hal. 723-732.