Pengelompokan Wilayah di Jawa Barat Menggunakan *Robust Sparse K-Means* Berdasarkan IPM 2023

Rifkah Nur Fatimah *, Ilham Faishal Mahdy

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

rifkahnurf10@email.com, ilham.faishal@unisba.ac.id

Abstract. The Human Development Index (HDI) is one of the main indicators used to assess the quality of development in a region in terms of health, education, and economy. This study aims to cluster regencies/municipalities in West Java Province based on HDI indicators by applying the Robust Sparse K-Means (RSK-Means) method. This method is an extension of the classical K-Means algorithm, capable of handling the presence of outliers and automatically selecting the most influential variables in the clustering process. The analyzed data consist of 27 regencies/municipalities with 8 HDI indicator variables. The analysis results show that the optimal configuration was obtained with 2 clusters, an L1 penalty value of 1.2, and a trimming α of 0.10. Three regions identified as outliers were trimmed, namely Tasikmalaya Regency, Karawang Regency, and Bekasi City. The Silhouette index value of 0.6 indicates that the formed cluster separation is fairly good. The variables Average Years of Schooling and Life Expectancy were found to be the most influential factors in cluster formation. The first cluster represents regions with lower HDI achievements, while the second cluster reflects regions with higher HDI achievements. These findings demonstrate that the RSK-Means method is effective in clustering regions based on human development indicators and can assist in determining more targeted policy priorities.

Keywords: *Human Development Index, Clustering, Robust Sparse K-Means.*

Abstrak. Indeks Pembangunan Manusia (IPM) merupakan salah satu indikator utama untuk menilai kualitas pembangunan suatu daerah dari segi kesehatan, pendidikan, dan ekonomi. Penelitian ini bertujuan mengelompokkan kabupaten/kota di Provinsi Jawa Barat berdasarkan indikator IPM dengan menerapkan metode *Robust Sparse K-Means* (RSK-*Means*). Metode ini merupakan pengembangan dari K-*Means* klasik yang mampu mengatasi keberadaan pencilan (*outlier*) serta secara otomatis memilih variabel yang paling berpengaruh dalam proses pengelompokan. Data yang dianalisis mencakup 27 kabupaten/kota dengan 8 variabel indikator IPM. Hasil analisis menunjukkan konfigurasi optimal diperoleh dengan 2 klaster, nilai penalti L1 sebesar 1,2, dan trimming α sebesar 0,10. Tiga daerah yang dipangkas karena teridentifikasi sebagai pencilan adalah Kabupaten Tasikmalaya, Kabupaten Karawang, dan Kota Bekasi. Nilai indeks Silhouette sebesar 0,6 mengindikasikan bahwa pemisahan klaster yang terbentuk cukup baik. Variabel Rata-Rata Lama Sekolah dan Umur Harapan Hidup ditemukan sebagai faktor paling berpengaruh dalam pembentukan klaster. Klaster pertama menggambarkan daerah dengan capaian IPM yang lebih rendah, sedangkan klaster kedua mencerminkan daerah dengan IPM yang lebih tinggi. Temuan ini menunjukkan bahwa metode RSK-*Means* efektif dalam mengelompokkan wilayah berdasarkan indikator pembangunan manusia serta dapat membantu dalam penentuan kebijakan yang lebih tepat sasaran.

Kata Kunci: *Indeks Pembangunan Manusia, Klasterisasi, Robust Sparse K-Means.*

*rifkahnurf10@email.com

A. Pendahuluan

Data mining adalah proses mengeksplorasi dan menganalisis data untuk menemukan pola-pola penting yang dapat mendukung pengambilan keputusan (Hendrastuty, 2024). Salah satu cabang dari data mining adalah machine learning yang terbagi menjadi dua pendekatan utama, yaitu supervised learning dan unsupervised learning. Pada pendekatan unsupervised learning, analisis klaster digunakan untuk mengelompokkan data berdasarkan kesamaan karakteristik antar objek. Salah satu algoritma yang banyak digunakan dalam analisis klaster adalah K-*Means*, yang bekerja dengan membagi data ke dalam beberapa klaster berdasarkan jarak terhadap pusat klaster (Kristiawan & Widjaja, 2021).

Meskipun dikenal luas karena kesederhanaan dan efisiensinya, metode K-*Means* memiliki beberapa kelemahan, seperti ketidakmampuan dalam menangani pencilan serta tidak adanya mekanisme seleksi variabel (Nugroho et al., 2024). Untuk mengatasi kendala tersebut, dikembangkanlah metode *Robust Sparse K-Means* (RSK-*Means*). RSK-*Means* merupakan teknik klasterisasi non-hierarki yang menggabungkan dua keunggulan utama sekaligus, yaitu *sparse clustering* dan *robust clustering*. *Sparse clustering* memungkinkan algoritma secara otomatis memilih variabel-variabel yang paling relevan dalam pembentukan klaster, sehingga hasil klasterisasi menjadi lebih informatif dan mudah diinterpretasikan, khususnya pada data berdimensi tinggi. Di sisi lain, *robust clustering* membuat metode ini lebih tahan terhadap pengaruh pencilan (*outlier*) yang sering mengganggu keakuratan hasil pengelompokan. Dengan memadukan kedua pendekatan tersebut, RSK-*Means* mampu menghasilkan klaster yang lebih stabil dan akurat (Kondo et al., 2011).

Penelitian yang dilakukan oleh Muhajir dan Rosadi (2023) membandingkan kinerja metode *Sparse* K-*Means*, *Trimmed* K-*Means*, dan RSK-*Means* dalam konteks pengelompokan kebijakan Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM). Penelitian ini menunjukkan bahwa *Sparse* K-*Means* efektif dalam memilih variabel penting, sementara *Trimmed K-Means* unggul dalam mengatasi data pencilan. Namun, RSK-*Means* terbukti memberikan hasil terbaik karena dapat menangani kedua tantangan tersebut secara bersamaan. Metode ini menghasilkan nilai *Classification Error Rate* (CER) yang paling rendah di antara ketiganya, menunjukkan bahwa RSK-*Means* memiliki keunggulan dalam menghasilkan pengelompokan yang lebih akurat (Muhajir & Rosadi, 2023).

Salah satu bidang penerapan yang relevan adalah pengelompokkan wilayah berdasarkan Indeks Pembangunan Manusia (IPM), yang mencakup aspek-aspek penting pembangunan manusia seperti kesehatan, pendidikan, dan pendapatan (Noviyanti, 2022). Provinsi Jawa Barat, sebagai provinsi dengan jumlah penduduk terbanyak di Indonesia, memperlihatkan adanya kesenjangan yang cukup signifikan antarwilayah dalam hal akses layanan kesehatan, mutu pendidikan, dan tingkat pendapatan. Oleh karena itu, diperlukan metode yang tepat untuk mengelompokkan wilayah-wilayah tersebut berdasarkan karakteristik IPM yang dimilikinya (BPS, 2024).

Melalui penerapan metode RSK-*Means*, diharapkan dapat dihasilkan klaster yang lebih informatif mengenai kondisi kabupaten/kota di Jawa Barat. Hasil pengelompokan tersebut dapat dijadikan landasan dalam menyusun kebijakan pembangunan yang lebih tepat sasaran dan efektif sesuai dengan karakteristik masing-masing wilayah. Berdasarkan latar belakang tersebut, penelitian ini diberi judul "Pengelompokan Kabupaten/Kota di Provinsi Jawa Barat Menggunakan *Robust Sparse* K-*Means* Berdasarkan Indikator Indeks Pembangunan Manusia (IPM) Tahun 2023".

B. Metode

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari website resmi Badan Pusat Statistik Provinsi Jawa Barat yang berjumlah 27 data. Data ini berisi 8 variabel terkait indikator Indeks Pembangunan Manusia (IPM) di kabupaten/kota Provinsi Jawa Barat pada tahun 2023, yaitu variabel Tingkat Pengangguran Terbuka (TPT) (X1), Pengeluaran Riil per kapita (X2), Upah Minimum Regional (UMR) (X3), Produk Domestik Regional Bruto per kapita atas dasar harga (PDRB) (X4), Rumah Tangga yang Memiliki Akses Terhadap Sanitasi Layak (X5), Umur Harapan Hidup (UHH) (X6), Harapan Lama Sekolah (HLS) (X7), dan Rata-Rata Lama Sekolah (RLS) (X8).

Penelitian ini bertujuan untuk mengelompokan daerah di Provinsi Jawa Barat berdasarkan indikator Indeks Pembangunan Manusia (IPM) menggunakan metode *Robust Sparse K-Means* (RSK-*Means*), mengetahui variabel yang paling berkontribusi dalam pembentukan klaster, serta memahami karakteristik klaster yang terbentuk dari hasil pengelompokan, dengan langkah sebagai berikut

Standarisasi Data

Standarisasi memastikan bahwa setiap variabel memiliki kontribusi yang setara dalam proses pembentukan klaster. Proses ini dilakukan dengan menghitung nilai *Z-score* menggunakan rumus berikut (Pearson, 1901):

$$Z_i = \frac{X_i - \mu}{\sigma} \quad \text{... (1)}$$

Pemeriksaan Multikolinearitas

langkah berikutnya adalah melakukan pemeriksaan multikolinearitas antar variabel dengan menghitung nilai VIF (*Variance Inflation Factor*) untuk masing-masing variabel menggunakan rumus berikut (Widarjono, 2010):

$$VIF = \frac{1}{1 - R_j^2} \quad \text{... (2)}$$

Keterangan:

VIF = nilai Variance Inflation Factor (VIF)

R_j^2 = koefisien determinasi antara variabel *j* terhadap semua variabel X yang lain.

Jika hasilnya menunjukkan bahwa tidak ada multikolinearitas atau nilai VIF < 10 maka perhitungan jarak antar objek dapat menggunakan jarak *Euclidean* dengan rumus berikut (Nooraeni & Nurfalah, 2022):

$$d(i, i') = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - y_{(i')j}\right)^2} \quad \text{... (3)}$$

Namun, jika terdeteksi adanya multikolinearitas, maka jarak antar objek dihitung menggunakan jarak *Mahalanobis* seperti dalam rumus berikut (Mahalanobis, 1936):

$$d_M(i, i') = \sqrt{\sum_{j=1}^{p} \sum_{j'=1}^{p} (x_{ij} - y_{i'j})(S^{-1})_{j,j'}(x_{ij'} - y_{i',j'})} \quad \text{... (4)}$$

Penentuan Jumlah Klaster dan Proporsi Data Dipangkas

Tahap selanjutnya adalah penentuan jumlah klaster yang akan diuji, yakni sebanyak 2, 3, 4, dan 5 klaster, serta penetapan proporsi data yang akan dipangkas sebesar α = 0,05 dan α = 0,10. Dalam proses ini juga ditentukan bobot awal variabel dengan asumsi $\boldsymbol{w_j = \frac{1}{\sqrt{p}}}$, dimana *p* adalah jumlah variabel. Selanjutnya, dilakukan penentuan pusat klaster awal secara acak.

Modifikasi Jarak *Euclidean* Berbobot

Pengelompokan selanjutnya dilakukan dengan modifikasi jarak *Euclidean* berbobot, dituliskan dalam rumus berikut (Kondo et al., 2011):

$$d(i, i') = \sqrt{\sum_{j=1}^{p} w_j^2 \left(x_{ij} - \mu_{kj}\right)^2} \quad \text{... (5)}$$

Trimmed K-Means

Setelah jarak dihitung, tahap selanjutnya adalah melakukan proses *trimmed k-means* dengan bobot, yaitu tahap pemangkasan (*trimming*). Objek dengan jarak terbesar sebanyak [αN] ke pusat klaster akan dipangkas sementara dari proses klasterisasi, membentuk himpunan $\boldsymbol{O_w}$ (pemangkasan pada data berbobot) dan $\boldsymbol{O_e}$ (pemangkasan pada data tidak berbobot). Pusat klaster untuk data yang tersisa dihitung kembali dengan rumus berikut (Kondo et al., 2011):

$$tm\,(O)_k = \frac{1}{|C_k \backslash O|} \sum_{i \in C_k \backslash o} x_i \in \mathbb{R}^p \quad \text{... (6)}$$

Keterangan:

$tm\,(O)_k$	=	Pusat klaster setelah melakukan *trimmed* k-*means*, yaitu setelah menghilangkan *outlier*.
O	=	himpunan data yang dianggap *outlier* dan tidak digunakan dalam perhitungan pusat klaster.
C_k	=	himpunan objek yang masuk dalam klaster ke-*k.*
$C_k \backslash O$	=	himpunan anggota klaster C_k yang tidak termasuk *outlier*.
x_i	=	Titik data ke-*i* dalam ruang berdimensi *p* ($\mathbb{R}^p$)

Pemangkasan ini diulang dan dilakukan pembaruan pusat klaster hingga hasilnya konvergen. Kedua himpunan pemangkasan kemudian digabungkan membentuk trimmed set $O = O_w \cup O_e$.

Pembaharuan Bobot Variabel

Pembaharuan bobot variabel untuk memperhitungkan kontribusi masing-masing variabel terhadap variasi antar klaster menggunakan rumus berikut:

$$w_j = \frac{\left[\sum_{k=1}^{K} n_k(\mu_{kj}-\bar{x}_j)^2\right]^{\frac{1}{2}}}{(\sum_{l=1}^{p}[\sum_{k}^{K} n_k(\mu_{kl}-\bar{x}_l)^2])^{\frac{1}{2}}} \quad \text{... (7)}$$

Keterangan:

w_j = bobot untuk variabel ke-*j*.
K = Jumlah klaster.
p = Jumlah variabel.
n_k = Jumlah data dalam klaster ke-*k.*
μ_{kj} = Rata-rata nilai variabel ke-*j* dalam klaster ke-*k.*
$\bar{x}_j$ = Rata-rata variabel ke-*j* dalam klaster ke-*k.*

Setelah memperoleh nilai $\boldsymbol{w_j}$ untuk masing-masing variabel, dilakukan proses *thresholding* dan normalisasi bobot untuk menghasilkan bobot akhir $\boldsymbol{w_j}$ yang bersifat *sparse* menggunakan rumus berikut:

$$w_j^{sparse} = \frac{(\sqrt{BCSS_j}-\Delta)_+}{\sum(\sqrt{BCSS_j}-\Delta)_+} \cdot L1 \quad \text{... (8)}$$

Keterangan:

$\sqrt{BCSS_j}$ = Akar dari variasi antar-klaster untuk variabel ke-*j*.
Δ = Nilai *threshold* yang ditentukan secara iteratif agar total bobot memenuhi syarat sparsity,
$L1$ = parameter yang mengontrol jumlah total bobot (dan tingkat *sparsity*).

Dengan Persamaan (8) bobot w_j akan memiliki nilai nol untuk variabel yang kurang berkontribusi.

Iterasi Bobot Variabel

Kriteria konvergensi bobot menggunakan rumus berikut:

$$\frac{\sum_{j=1}^{p}\left|w_j^{(r)}-w_j^{(r-1)}\right|}{\sum_{j=1}^{p}\left|w_j^{(r-1)}\right|} \leq \delta \quad \text{... (9)}$$

Keterangan:

$w_j^{(r)}$ = Bobot variabel ke-*j* pada iterasi ke-*r.*
$w_j^{(r-1)}$ =Bobot variabel ke-*j* pada iterasi sebelumnya.

Evaluasi Klaster

Evaluasi klaster menggunakan *Silhouette Indeks*. Indeks ini mengukur derajat kepercayaan yang mempunyai kriteria klaster dikatakan terbentuk baik ketika nilai indeks mendekati 1. Dengan rumus sebagai berikut (Afira, N., & Wahyu Wijayanto, 2021):

$$SI_k = \frac{1}{m_k}\sum_{i=1}^{m_k} SI_i^k \quad \text{... (10)}$$

Keterangan:

SI = Rata-rata indeks *Silhouette* dari *dataset*.
SI_k = Rata-rata indeks *Silhouette* klaster ke-*k.*
m_k = Jumlah observasi dalam klaster ke-*k*.

Dimana diketahui kriteria nilai indeks *Silhouette* disajikan pada Tabel 1.

Tabel 1. Kriteria Nilai indeks *Silhouette*

Nilai Indeks *Silhouette*	Kriteria Pengelompokan
0,71 – 1,00	Struktur Klaster Kuat

0,51 – 0,70	Struktur Klaster Baik

C. Hasil Penelitian dan Pembahasan

Eksplorasi Data Penelitian

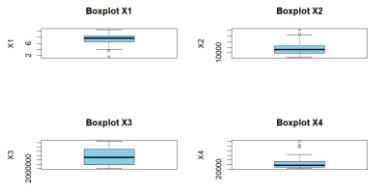
Eksplorasi data dilakukan untuk memberikan pemahaman awal terkait karakteristik data yang digunakan dalam penelitian ini. Data yang dianalisis mencakup delapan variabel yang berhubungan dengan Indeks Pembangunan Manusia (IPM) pada kabupaten/kota di Provinsi Jawa Barat.

Tabel 2. Statistika Deskriptif Variabel Indikator Indeks Pembangunan Manusia

Variabel	Rata-Rata	Standar Deviasi	Minimum	Maksimum
TPT	7,19	2,02	1,52	10,52
Pengeluaran	11.525,148	2.368,139	8.562	18.236
UMR	3.294.994,889	1.082.032,923	1.998.119	5.176.179
PDRB	50.009,074	31.808,218	22.826	140.144
Sanitasi	75,07	16,01	44,76	98,52
UHH	74,93	0,39	73,87	75,86
HLS	12,86	0,81	11,19	14,29
RLS	8,87	1,46	6,94	11,66

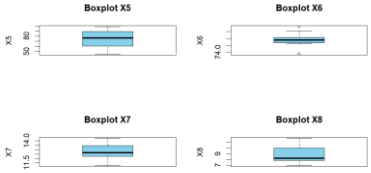
Tabel 2 menyajikan statistik deskriptif dari masing-masing variabel indikator Indeks Pembangunan Manusia (IPM) yang digunakan dalam penelitian. Informasi yang ditampilkan mencakup nilai rata-rata, standar deviasi, nilai minimum, dan maksimum dari tiap variabel. Data ini memberikan gambaran umum mengenai sebaran dan karakteristik awal dari masing-masing indikator sebelum dilakukan proses analisis klaster lebih lanjut.

Setelah analisis deskriptif, kemudian mendeteksi *outlier* karena keberadaannya dapat memengaruhi hasil pengelompokan, keberadaan *outlier* dapat memengaruhi perhitungan pusat klaster dan bobot variabel, sehingga perlu dikenali sejak awal, terutama pada metode *Robust Sparse K-Means* (RSK-*Means*) yang dirancang untuk menangani pencilan. Deteksi dilakukan menggunakan boxplot sebagai alat visual yang efektif untuk mengidentifikasi nilai ekstrem dan memahami distribusi data yang dapat dilihat pada Gambar 1.



Gambar 1. Contoh Tahapan Riset

Berdasarkan boxplot variabel X1 hingga X4 pada Gambar 1 terlihat bahwa beberapa variabel mengandung data pencilan, yang ditandai dengan titik-titik di luar batas garis *whisker*. Pada variabel X1, terdapat dua pencilan di sisi bawah dan Variabel X2 juga menunjukkan beberapa *outlier* di sisi atas, sementara X4 bahkan memiliki lebih banyak *outlier* di bagian atas. Variabel X3 tidak menunjukkan pencilan, tetapi memiliki sebaran nilai yang lebar. Selanjutnya, boxplot untuk variabel X5 hingga X8 ditampilkan pada Gambar 2.



Gambar 2. Contoh Tahapan Riset

Boxplot variabel X5 hingga X8 pada Gambar 2 menunjukkan bahwa sebagian besar variabel memiliki distribusi yang simetris tanpa pencilan ekstrem, kecuali variabel X6 yang memiliki beberapa outlier di bagian atas dan bawah. Perbedaan skala antar variabel tetap perlu diperhatikan dalam analisis klaster. Berdasarkan hasil pada Gambar 1 dan 2 keberadaan pencilan mendukung pemilihan metode *Robust Sparse K-Means* (RSK-*Means*) yang lebih sesuai karena mampu menangani outlier dan melakukan seleksi variabel.

Pemeriksaan Multikolinearitas

Multikolinearitas dapat memengaruhi proses pembobotan dan hasil klasterisasi karena adanya duplikasi informasi antarvariabel. Oleh sebab itu, pemeriksaan multikolinearitas perlu dilakukan untuk memastikan setiap variabel memberikan informasi yang bersifat unik. Hasil pemeriksaan multikolinearitas berdasarkan nilai VIF disajikan pada Tabel 3.

Tabel 3. Hasil Pemeriksaan Multikolinearitas

Variabel	VIF	Keterangan
X1	1,82	Tidak Terdapat Multikolinearitas
X2	5,28	Tidak Terdapat Multikolinearitas
X3	2,26	Tidak Terdapat Multikolinearitas
X4	1,96	Tidak Terdapat Multikolinearitas
X5	1,46	Tidak Terdapat Multikolinearitas
X6	4,3	Tidak Terdapat Multikolinearitas
X7	3,57	Tidak Terdapat Multikolinearitas
X8	5,73	Tidak Terdapat Multikolinearitas

Berdasarkan Tabel 3, seluruh variabel indikator Indeks Pembangunan Manusia tidak menunjukkan adanya multikolinearitas dan tidak terdapat hubungan yang kuat antarvariabel karena nilai VIF < 10. Oleh karena itu, semua variabel dinyatakan layak digunakan dalam proses analisis. Dengan tidak adanya multikolinearitas, perhitungan jarak dapat dilakukan menggunakan jarak Euclidean.

Hasil Seleksi Variabel

Analisis dilakukan terhadap seluruh data kabupaten/kota di Provinsi Jawa Barat menggunakan bantuan perangkat lunak *RStudio*. Analisis dilakukan dengan menguji berbagai parameter, di antaranya jumlah klaster sebanyak 2, 3, 4, dan 5 klaster, serta nilai *trimming* (α) sebesar 0,05 dan 0,10. Selain itu, digunakan juga berbagai nilai parameter *sparsity* (L1) untuk mengevaluasi pengaruhnya terhadap seleksi variabel dan hasil pengelompokan. Hasil seleksi variabel untuk semua kombinasi dimana variabel yang terpilih merupakan variabel yang memiliki bobot tidak nol, atau variabel yang memilliki bobot lebih besar dari nol.

Tabel 4 Hasil Seleksi Variabel Dengan Metode RSK-*Means* Pada 2 Klaster

L1	2 Klaster			
	α = 0,05		α = 0,10	
	Jumlah Variabel Terpilih	Variabel Terpilih	Jumlah Variabel Terpilih	Variabel Terpilih
1,05	3	X2,X6,X8	2	X6,X8
1,10	3	X2,X6,X8	2	X6,X8
1,20	3	X6,X7,X8	2	X6,X8
1,30	3	X6,X7,X8	3	X6,X7,X8
1,40	3	X6,X7,X8	4	X2,X6,X7,X8
1,50	4	X2,X6,X7,X8	4	X2,X4,X6,X7
2	4	X2,X3,X4,X6	4	X2,X3,X4,X5
4	8	X1-X8	8	X1-X8

Tabel 4 menunjukkan hasil seleksi variabel menggunakan metode *Robust Sparse K-Means* (RSK-*Means*) pada pengelompokan 2 klaster dengan beragam kombinasi nilai parameter penalty L1

dan tingkat trimming α= 0,05 dan 0,10. Pada setiap kombinasi, ditampilkan jumlah variabel yang terpilih dan variabel yang memiliki bobot tidak nol.

Tabel 5. Hasil Seleksi Variabel Dengan Metode RSK-*Means* Pada 3 Klaster

L1	3 Klaster			
	α = 0,05		α = 0,10	
	Jumlah Variabel Terpilih	Variabel Terpilih	Jumlah Variabel Terpilih	Variabel Terpilih
1,05	2	X1,X3	2	X2,X3
1,10	2	X1,X8	3	X2,X3,X4
1,20	3	X1,X3,X8	4	X1,X3,X4,X8
1,30	2	X1,X8	2	X7,X8
1,40	3	X1,X2,X3	3	X1,X6,X8
1,50	3	X1,X6,X8	4	X2,X6,X7,X8
2	4	X2,X3,X4,X6	5	X2,X3,X4,X6,X8
4	8	X1-X8	8	X1-X8

Tabel 5 menampilkan hasil seleksi variabel menggunakan metode RSK-*Means* untuk 3 klaster dengan variasi parameter L1 dan α.

Tabel 6. Hasil Seleksi Variabel Dengan Metode RSK-*Means* Pada 4 Klaster

L1	4 Klaster			
	α = 0,05		α = 0,10	
	Jumlah Variabel Terpilih	Variabel Terpilih	Jumlah Variabel Terpilih	Variabel Terpilih
1,05	2	X2,X8	2	X6,X8
1,10	2	X2,X8	2	X7,X8
1,20	2	X3,X8	2	X7,X8
1,30	2	X3,X5	3	X2,X3,X8
1,40	3	X2,X7,X8	3	X2,X3,X8
1,50	4	X2,X3,X6,X8	4	X3,X6,X7,X8
2	6	X2,X3,X4,X6,X7,X8	5	X2,X3,X6,X7,X8
4	8	X1-X8	8	X1-X8

Tabel 6 menampilkan hasil seleksi variabel menggunakan metode RSK-*Means* untuk 4 klaster dengan variasi parameter L1 dan α.

Tabel 7. Hasil Seleksi Variabel Dengan Metode RSK-*Means* Pada 4 Klaster

L1	5 Klaster			
	α = 0,05		α = 0,10	
	Jumlah Variabel Terpilih	Variabel Terpilih	Jumlah Variabel Terpilih	Variabel Terpilih
1,05	2	X2,X8	2	X2,X8
1,10	2	X3,X8	2	X7,X8
1,20	2	X1,X8	2	X3,X8
1,30	3	X3,X6,X8	2	X5,X8
1,40	3	X1,X3,X8	4	X3,X6,X7,X8
1,50	3	X1,X3,X8	4	X1,X2,X7,X8
2	5	X3,X5,X6,X7,X8	6	X1,X2,X3,X6,X7,X8
4	8	X1-X8	8	X1-X8

Tabel 7 menampilkan hasil seleksi variabel menggunakan metode RSK-*Means* untuk 5 klaster dengan variasi parameter L1 dan α.

Validasi Klaster

Validasi ini bertujuan untuk mengukur sejauh mana hasil pengelompokan yang dihasilkan telah optimal, dengan menggunakan indeks Silhouette yang bernilai antara -1 hingga 1. Dalam penelitian ini, validasi klaster dihitung menggunakan *silhouette coefficient* untuk mengevaluasi kualitas klasterisasi.

Tabel 8 Hasil indeks *Silhoutte*

L1	2 Klaster		3 Klaster		4 Klaster		5 Klaster	
	Silhoutte α = 0,05	**Silhoutte α = 0,10**	**Silhoutte α = 0,05**	**Silhoutte α = 0,10**	**Silhoutte α = 0,05**	**Silhoutte α = 0,10**	**Silhoutte α = 0,05**	**Silhoutte α = 0,10**
1,05	0,509	0,585	0,289	0,354	0,366	0,267	0,260	0,260
1,1	0,509	0,585	0,288	0,246	0,366	0,520	0,390	0,264
1,2	0,525	0,585	0,256	0,226	0,476	0,520	0,423	0,390
1,3	0,525	0,525	0,521	0,528	0,453	0,288	0,350	0,436
1,4	0,525	0,484	0,413	0,430	0,365	0,279	0,390	0,347
1,5	0,484	0,435	0,430	0,414	0,259	0,317	0,422	0,381
2	0,381	0,335	0,411	0,401	0,307	0,353	0,280	0,333
4	0,355	0,355	0,225	0,254	0,214	0,214	0,230	0,173

Tabel 8 memperlihatkan bahwa kombinasi terbaik diperoleh pada jumlah 2 klaster dengan parameter L1 = 1,2 dan α = 0,10, yang menghasilkan nilai indeks Silhouette tertinggi sebesar 0,585. Kombinasi ini dipilih karena tidak hanya memberikan evaluasi klaster yang paling baik, tetapi juga menghasilkan distribusi bobot variabel yang seimbang serta seleksi variabel yang optimal, sehingga mencerminkan struktur klaster yang baik.

Interpretasi Klaster

Hasil terbaik diperoleh pada 2 klaster dengan parameter L1 = 1,2 dan α = 0,10, setelah dilakukan proses pemangkasan pada data ke-6 (Kabupaten Tasikmalaya), ke-15 (Kabupaten Karawang), dan ke-23 (Kota Bekasi), dengan variabel Umur Harapan Hitup dan Rata-Rata Lama Sekolah terpilih sebagai variabel terbaik untuk pengelompokan. berikut anggota dari masing-masing klaster disajikan pada tabel 9.

Tabel 9. Anggota Klaster

Klaster	Anggota Klaster	Banyaknya Anggota
1	Kabupaten Bogor, Kabupaten Sukabumi, Kabupaten Cianjur, Kabupaten Bandung, Kabupaten Garut, Kabupaten Tasikmalaya, Kabupaten Ciamis, Kabupaten Kuningan, Kabupaten Cirebon, Kabupaten Majalengka, Kabupaten Sumedang, Kabupaten Indramayu, Kabupaten Subang, Kabupaten Purwakarta, Kabupaten Karawang, Kabupaten Bandung Barat, Kabupaten Pangandaran, Kota Banjar	18
2	Kabupaten Bekasi, Kota Bogor, Kota Sukabumi, Kota Bandung, Kota Cirebon, Kota Bekasi, Kota Depok, Kota Cimahi, Kota Tasikmalaya	9

Seperti yang terlihat pada Tabel 9 anggota klaster 1 ada 18 anggota dan untuk klaster 2 memiliki 9 anggota dari berbagai Kabupaten/Kota di Provinsi Jawa Barat. Adapun hasil karakteristik pengelompokkan klaster dengan 2 variabel X6 yaitu Umur Harapan Hitup dan X8 yaitu Rata-Rata Lama Sekolah disajikan dalam Tabel 10.

Tabel 10. Rata-Rata Variabel Setiap Klaster

	Umur Harapan Hidup (UHH) (bobot= 0,226)	**Rata-Rata Lama Sekolah (RLS) (bobot= 0,974)**
Klaster 1	74,73	7,96
Klaster 2	75,32	10,68

Berdasarkan Tabel 10, hasil pengelompokan dua klaster menggunakan metode RSK-*Means* memperlihatkan adanya perbedaan karakteristik, di mana Klaster 2 memiliki nilai UHH dan RLS yang lebih tinggi dibandingkan Klaster 1. Bobot variabel juga menunjukkan bahwa RLS (0,974) memberikan kontribusi yang jauh lebih besar dalam pembentukan klaster dibandingkan UHH (0,226). Hasil pengelompokan ini menunjukkan bahwa Klaster 2 menggambarkan daerah dengan IPM lebih tinggi, yang ditandai dengan kualitas pendidikan dan umur harapan hidup yang lebih baik. Sebaliknya, Klaster 1 mencerminkan daerah dengan IPM yang lebih rendah. Temuan ini menegaskan bahwa variabel pendidikan memiliki pengaruh yang lebih dominan dibandingkan variabel kesehatan dalam membedakan tingkat pembangunan manusia di berbagai wilayah di Provinsi Jawa Barat.

D. Kesimpulan

Berdasarkan hasil analisis pengelompokan menggunakan metode *Robust Sparse* K-*Means* (RSK-*Means*) terhadap data indikator Indeks Pembangunan Manusia (IPM) di Provinsi Jawa Barat menunjukkan bahwa kombinasi parameter terbaik diperoleh pada L1 = 1,2 dan $\alpha = 0{,}10$ dengan dua klaster, serta pemangkasan pada tiga daerah yang dianggap pencilan, yaitu Kabupaten Tasikmalaya, Kabupaten Karawang, dan Kota Bekasi. Klaster 1 terdiri dari 18 daerah, sementara Klaster 2 terdiri dari 9 daerah. Validitas pengelompokan ini didukung oleh nilai indeks Silhouette sebesar 0,6 yang menunjukkan struktur klaster yang baik. Variabel yang paling berkontribusi dalam proses pengelompokan adalah Rata-Rata Lama Sekolah (RLS) dan Umur Harapan Hidup (UHH), di mana RLS memiliki bobot tertinggi. Klaster 2 ditandai dengan nilai rata-rata UHH dan RLS yang lebih tinggi, sehingga dapat diartikan sebagai kelompok daerah dengan capaian IPM yang lebih baik, khususnya karena didukung oleh kualitas pendidikan yang lebih tinggi.

Ucapan Terimakasih

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah memberikan bantuan dan dukungan selama penyusunan penelitian ini, khususnya kepada dosen pembimbing, keluarga, serta instansi terkait yang telah menyediakan data dan informasi yang diperlukan.

Daftar Pustaka

Afira, N., & Wahyu Wijayanto, A. (2021). Analisis Cluster Kemiskinan Provinsi di Indonesia Tahun 2019 dengan Metode Partitioningdan Hierarki. *Komputika: Jurnal Komputer*.

BPS, B. P. S. P. J. B. (2024). *Indeks Pembangunan Manusia Provinsi Jawa Barat 2023*. *8*.

Fitriani, A. N., & Suliadi. (2021). Selang Kepercayaan Koefisien Korelasi Berdasarkan Empirical Likelihood dan Penerapannya pada Data Rata-Rata Lama Sekolah dan Penduduk Miskin Kota/Kabupaten di Indonesia. *Jurnal Riset Statistika*, *1*(1), 51–56. https://doi.org/10.29313/jrs.v1i1.146

Hasna, & Achmad, A. I. (2022). Metode Regresi Probit Biner untuk Pemodelan Faktor-Faktor yang Mempengaruhi Diagnosis Penyakit Jantung. *Jurnal Riset Statistika*, 28–34. https://doi.org/10.29313/jrs.vi.721

Hendrastuty, N. (2024). Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa. *Jurnal Ilmiah Informatika Dan Ilmu Komputer (Jima-Ilkom)*, *3*(1), 46–56.

Kondo, Y., Salibian-Barrera, M., & Zamar, R. (2011). *A robust and sparse K-means clustering algorithm*.

Kristiawan, K., & Widjaja, A. (2021). Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel. *Jurnal Teknik Informatika Dan Sistem Informasi*, *7*(1), 35–46. https://doi.org/10.28932/jutisi.v7i1.3182

Mahalanobis, P. C. (1936). On the Generalised Distance in Statistics. In *Proceedings of the National Institute of Sciences of India* (Vol. 2, Issue 1, pp. 49–55).

Muhajir, M., & Rosadi, D. (2023). Robust clustering using sparse K means on text analitycs of PPKM policy in Indonesia. *AIP Conference Proceedings*, *2720*(January). https://doi.org/10.1063/5.0137220

Nooraeni, R., & Nurfalah, G. (2022). Kajian Penerapan Jarak Euclidean, Manhattan, Minkowski, dan Chebyshev pada Algoritma Clustering K-Prototype. *Sains, Aplikasi, Komputasi Dan Teknologi Informasi*, *4*(2), 72–82. https://doi.org/10.30872/jsakti.v4i2.9241

Noviyanti, D. (2022). Faktor-Faktor Yang Memengaruhi Indeks Pembangunan Manusia Provinsi Jawa Barat Tahun 2020. *Jurnal Sosial Humaniora Sigli*, *5*(2), 117–124. https://doi.org/10.47647/jsh.v5i2.909

Nugroho, D. W., Bramhatchi, F., Wulandari, S. P., & Eka, A. (2024). *Pengelompokan Indikator Kesejahteraan Masyarakat Berdasarkan Kabupaten / Kota di Jawa Tengah Tahun 2023 Menggunakan Analisis Cluster Institut Teknologi Sepuluh Nopember , Indonesia pendidikan , kualitas pekerjaan , serta pertumbuhan ekonomi masyarakat sua*.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11), 559-572.

Widarjono, A. (2010). Analisis Statistika Multivariat Terapan. *UPP STIM YKPN, Yogyakarta*.

Zahara, A., & Kurniati, E. (2024). Penerapan Ekstrim Fungsi untuk Menentukan Profit Maksimum dalam Pasar Persaingan Sempurna untuk Short-Run dan Long-Run. *DataMath: Journal of Statistics and Mathematics*, *2*(1), 33–40.