Perbandingan Algoritma Klasterisasi dengan *Principal Component Analysis* pada Indikator Sosial Ekonomi Kesehatan Jawa Timur

Comparison of Clustering Algorithms with Principal Component Analysis on Socioeconomic and Health Indicators in East Java

Uswatun Hasanah[1], Monica Rahma Fauziah[2], Anwar Fitrianto[3], Erfiani[4], L.M. Risman Dwi Jumansyah[5]
[1,2,3,4,5]Statistika dan Sains Data, IPB University
E-mail: [1]21uswatun@apps.ipb.ac.id, [2]007monica@apps.ipb.ac.id, [3]anwarstat@gmail.com, [4]erfiani@apps.ipb.ac.id, [5]rismandwijumansyah@apps.ipb.ac.id

Abstrak

K-Means dan *K-Medoids* digunakan untuk menilai indikator sosial ekonomi dan kesehatan di Provinsi Jawa Timur tahun 2023 melalui metode klasterisasi. Dengan menggunakan *Principal Component Analysis* (PCA) untuk mereduksi dimensi variabel, penelitian ini mengelompokkan wilayah berdasarkan karakteristik sosial ekonomi dan kesehatan. Data yang dianalisis termasuk angka harapan hidup, tingkat kemiskinan, pengangguran, dan akses ke layanan kesehatan. Kebaruan penelitian ini terletak pada kombinasi unik antara PCA dan *K-Medoids* untuk menghasilkan klaster yang lebih akurat dan *robust* terhadap *outlier,* dibandingkan metode yang biasanya hanya menggunakan satu teknik klasterisasi atau tidak melibatkan reduksi dimensi. Hasil penelitian menunjukkan bahwa *K-Medoids* dengan PCA menghasilkan klaster yang lebih koheren dan terpisah daripada *K-Means*, terutama dalam menangani *outlier*. Menurut metode *Elbow* dan *Silhouette*, empat hingga lima klaster adalah pilihan terbaik. PCA meningkatkan akurasi dan efisiensi klasterisasi dengan mengurangi kompleksitas data, yang menghasilkan klaster yang lebih baik Diharapkan temuan ini akan membantu pemerintah membuat kebijakan yang lebih baik untuk mengatasi ketimpangan kesehatan dan sosial ekonomi di Jawa Timur.

Kata kunci: Klasterisasi, *Outlier, Principal Component Analy*sis (PCA)

Abstract

K-Means and K-Medoids were used to evaluate socioeconomic and health indicators in East Java Province in 2023 through clustering methods. Using Principal Component Analysis (PCA) for dimensionality reduction, this study grouped regions based on socioeconomic and health characteristics, including life expectancy, poverty rate, unemployment, and access to healthcare services. The novelty of this research lies in the unique combination of PCA and K-Medoids to produce more accurate and robust clusters against outliers, compared to methods that typically use only one clustering technique or do not involve dimensionality reduction. The results indicate that K-Medoids with PCA produced more coherent and well-separated clusters than K-Means, especially in handling outliers. According to the Elbow and Silhouette methods, four to five clusters were the best choices. By reducing data complexity, PCA improved clustering accuracy and efficiency, resulting in better-structured and more representative clusters. These findings are expected to assist the government in formulating more targeted policies to address health and socioeconomic disparities in East Java more effectively.

Keywords: Clustering, Outlier, Principal Component Analysis (PCA)

1. PENDAHULUAN

Dalam perkembangan statistika modern, klasterisasi menjadi salah satu teknik penting untuk menangani data yang kompleks dan multidimensi. Klasterisasi adalah teknik pengelompokan objek atau data berdasarkan kesamaan karakteristik. Teknik ini membantu dalam mengidentifikasi pola-pola tersembunyi dalam *dataset* besar dan kompleks, seperti data yang mencakup berbagai indikator sosial ekonomi dan kesehatan.

K-Means adalah algoritma klasterisasi populer karena kesederhanaan dan kecepatannya dalam mengelompokkan data berdasarkan *centroid* kluster. Namun, kelemahan utamanya adalah sensitivitas terhadap outlier, yang dapat menggeser centroid dan membuat hasil klasterisasi kurang representatif [1]. Untuk mengatasi kelemahan tersebut, *K-Medoids* hadir sebagai alternatif yang lebih kuat dalam menangani *outlier*. *K-Medoids* mirip dengan *K-Means*, tetapi alih-alih menggunakan rata-rata (*centroid*), *K-Medoids* memilih titik data aktual sebagai pusat kluster (*medoid*). Pendekatan ini membuat *K-Medoids* lebih tahan terhadap *outlier* karena *medoid* dipilih dari titik data yang ada, bukan nilai rata-rata yang bisa dipengaruhi oleh *outlier* [2]. Penelitian [3] dengan judul "*Analysis of K-Means and K-Medoids Algorithm For Big Data*" mengevaluasi algoritma *K-Means* dan *K-Medoids* pada *dataset* transaksi besar, menemukan bahwa *K-Medoids* lebih unggul dalam menangani *outlier* dan lebih efisien dalam hal waktu eksekusi dibandingkan *K-Means.*

Tantangan utama dalam klasterisasi data sosial ekonomi dan kesehatan adalah banyaknya variabel yang berkorelasi tinggi. *Principal Component Analysis* (PCA) membantu mengatasi ini dengan mereduksi dimensi data, menyederhanakan variabel berkorelasi tinggi menjadi beberapa komponen utama. PCA meningkatkan efisiensi dan akurasi *K-Means* dan *K-Medoids* dengan mengurangi *redundansi* variabel [4]. Dalam analisis multidimensi dengan banyak variabel sosial ekonomi dan kesehatan, PCA menyederhanakan data sehingga *K-Means* dan *K-Medoids* dapat bekerja lebih efektif [5]. Penelitan [5] mengusulkan model prediksi diabetes menggunakan PCA untuk reduksi dimensi dan algoritma *K-Means*, serta regresi logistik. Hasilnya menunjukkan bahwa PCA meningkatkan akurasi *K-Means* dan model regresi logistik dibandingkan tanpa PCA.

Di Jawa Timur, ada perbedaan yang signifikan dalam indikator sosial ekonomi dan kesehatan, yang menunjukkan adanya ketidakmerataan yang memerlukan penyelidikan lebih lanjut. Ketidakmerataan ini dapat memengaruhi kebijakan pembangunan daerah dan kualitas hidup masyarakat. Akibatnya, penelitian ini bertujuan untuk memperoleh pemahaman yang lebih mendalam tentang distribusi ketimpangan sosial ekonomi dan kesehatan dengan mengidentifikasi klaster wilayah di Jawa Timur dengan menggunakan pendekatan klasterisasi berbasis *K-Means* dan *K-Medoids* [6].

Data yang digunakan mencakup berbagai indikator sosial ekonomi dan kesehatan dari kabupaten/kota di Jawa Timur. Penelitian ini akan mengelompokkan wilayah berdasarkan karakteristik yang serupa dengan memanfaatkan Teknik *Principal Component Analysis* (PCA) dan algoritma *K-Means* serta *K-Medoids*. Hasil penelitian diharapkan dapat membantu pemerintah daerah dalam merancang kebijakan pembangunan yang lebih efektif dan adil, guna mengurangi ketimpangan sosial ekonomi dan mengoptimalkan kualitas hidup masyarakat di Jawa Timur [7].

2. METODE PENELITIAN

2.1. *Principal Component Analysis* (PCA)

Analisis komponen utama (PCA) merupakan salah satu teknik dalam analisis data multivariat yang bertujuan untuk menyederhanakan data dengan banyak variabel. Pada dasarnya, PCA digunakan untuk mengurangi jumlah variabel dalam dataset yang kompleks, dengan tetap mempertahankan informasi penting. Teknik ini memungkinkan data berdimensi tinggi menjadi lebih mudah dianalisis tanpa kehilangan nilai-nilai utama yang terkandung di dalamnya [8]. Seringkali, informasi tersebut disederhanakan menjadi variabel-variabel baru yang lebih ringkas berdasarkan kesamaan informasi yang diperoleh dari data asli. Hingga saat ini, metode PCA terus

berkembang dan diaplikasikan dalam berbagai bidang. Beberapa manfaat dari penerapan metode PCA di antaranya adalah:

1. Membentuk variabel baru yang mewakili data multivariat secara lebih sederhana,
2. Mengurangi jumlah variabel independen menjadi lebih sedikit tanpa kehilangan informasi penting dari data asli,
3. Menghilangkan variabel asli yang memberikan sedikit kontribusi informasi.

2.2 Analisis *Cluster*

Klasterisasi merupakan metode analisis multivariat yang bertujuan untuk mengelompokkan objek (atau responden) dimana objek dalam suatu cluster memiliki kemiripan (homogen) secara simultan dan perbedaan ukuran antar cluster jauh lebih besar. Berbeda dengan klasifikasi, *cluster* tidak memiliki variable target saat mengelompokkan proses *clustering*. Selain itu, metode klasterisasi juga memiliki kemampuan unik untuk mengidentifikasi pola kesamaan dalam kelompok data penelitian. Kualitas hasil pengelompokan seringkali bergantung pada metode yang digunakan dan terdapat banyak metode lain yang digunakan untuk pengukuran kemiripan antar objek yang akan dibandingkan. Metode *clustering* secara umum dapat dikelompokkan sebagai berikut:

a. Metode partisi yaitu metode untuk membuat suatu partisi pada basis data yang disebut sebagai jumlah *cluster* (*k*). Setiap *cluster* harus memiliki 8 minimal satu objek untuk setiap partisi yang terbentuk, dan objek wajib dimiliki tepat satu *cluster*.
b. Metode hirarki ialah jenis klasifikasi terstruktur dan stabil yang dibentuk berdasarkan dendogram menggunakan beberapa kriteria.
c. Metode berdasarkan *grid* yaitu metode yang melakukan pendekatan menurut struktur *multiple level granularity*.
d. Metode dengan model dengan hipotesis untuk masing-masing *cluster* serta ide awalnya digunakan untuk menentukan model yang pas untuk setiap *cluster* [9].

2.3 *K-Means Clustering*

K-Means clustering adalah metode yang membagi N individu dalam *dataset* multivariat menjadi k kelompok. Pendekatan umum dalam *K-Means* adalah mempartisi N individu ke dalam k kelompok dengan meminimalkan jumlah kuadrat jarak antar anggota dalam kelompok (*within-group sum of squares*) untuk semua variabel. Tujuan utamanya adalah membentuk kelompok yang homogen di dalamnya dan heterogen antar kelompok. Tahapan dalam *K-Means clustering* meliputi:

1. Menentukan nilai k, di mana k adalah jumlah klaster yang diinginkan.
2. Menginisiasi secara acak posisi centroid (titik pusat klaster) sebanyak k buah.
3. Menghitung jarak dari setiap data ke masing-masing centroid klaster menggunakan jarak *Euclidean.* Jarak *Euclidean* antara objek i dan j dirumuskan sebagai:

$$d(x,y) = \sqrt{\sum_{k=1}^{n}\left(x_{ik} - y_{jk}\right)^2} \quad (1)$$

4. Mengelompokkan data ke dalam klaster berdasarkan jarak terdekat dari setiap data ke centroid yang ada.
5. Menentukan centroid baru dengan menghitung rata-rata posisi semua data dalam klaster yang sama.

$$v = \sqrt{\frac{\sum_{i=1}^{n} x_i}{n}}\ ; n = 1,2,3,\dots,n \quad (2)$$

6. Ulangi langkah 3 dan teruskan iterasi hingga centroid klaster stabil dan anggota klaster tidak mengalami perpindahan (mencapai konvergensi). [8]

2.4 K-Medoids Clustering

Algoritma *K-Medoids*, atau yang dikenal sebagai PAM (*Partitioning Around Medoid*), dikembangkan oleh Leonard Kaufman dan Peter J. Rousseeuw. Algoritma ini mirip *dengan K-*

Means karena keduanya bersifat partisional, yaitu membagi *dataset* menjadi beberapa kelompok. Perbedaan utama antara *K-Means* dan *K-Medoids* terletak pada penentuan pusat klaster: *K-Means* menggunakan rata-rata nilai dalam setiap klaster sebagai pusat, sedangkan *K-Medoids* memilih objek data sebagai pusat klaster. Langkah-langkah dalam algoritma *K-Medoids* adalah sebagai berikut.

a. Inisialisasi pusat klaster sebanyak k (sesuai jumlah klaster yang diinginkan).
b. Hitung jarak setiap objek ke klaster terdekat menggunakan rumus jarak Euclidean. Perhitungan jarak Euclidean menggunakan persamaan (2).
c. Setelah menghitung jarak Euclidean, pilih secara acak pusat klaster baru pada setiap objek sebagai kandidat non-medoid.
d. Hitung jarak antara setiap objek dalam klaster dengan kandidat non-medoid yang ada.
e. Hitung total simpangan (S) dengan membandingkan total jarak baru dan total jarak lama. Jika S < 0, maka tukar objek dengan data klaster non-medoid untuk membentuk k objek baru sebagai medoid.
f. Ulangi langkah c–e hingga medoid tidak mengalami perubahan, sehingga terbentuk klaster beserta anggota klaster masing-masing. [10]

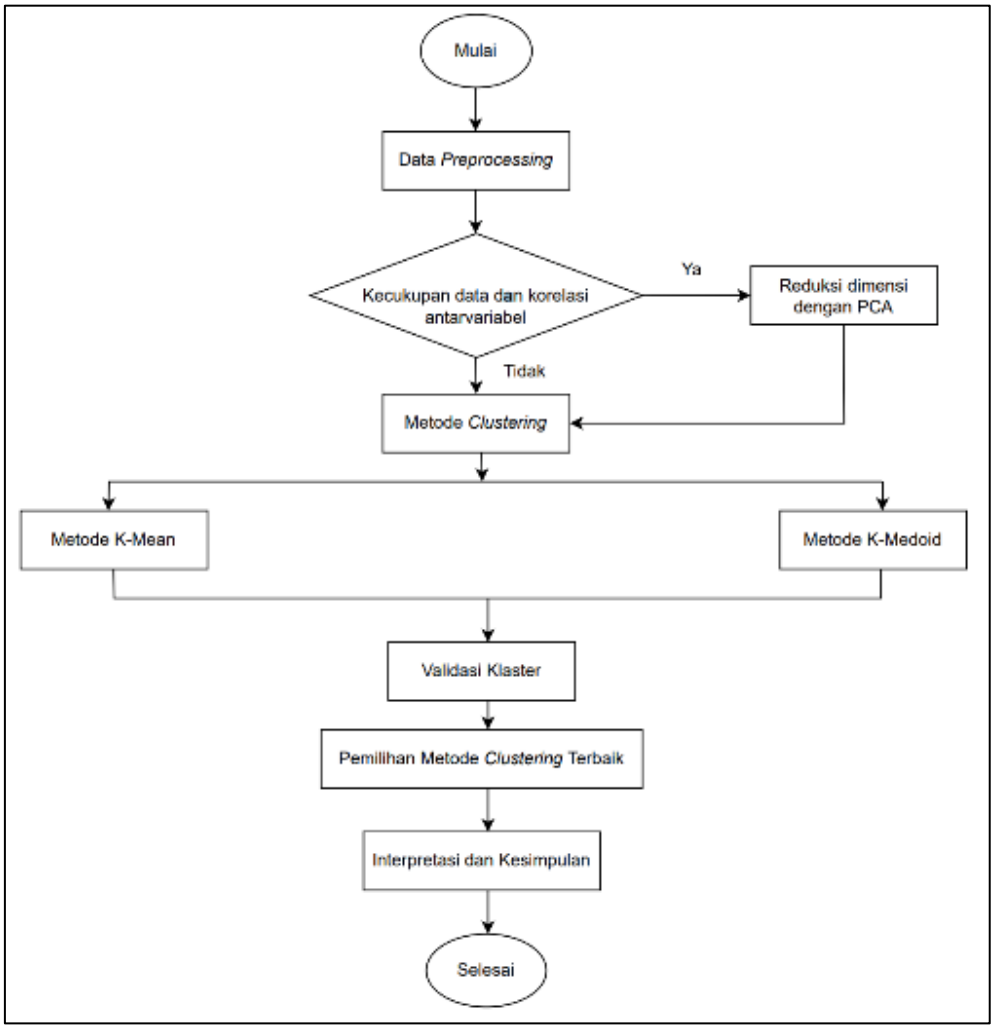
2.5 Alur Penelitian

Langkah awal metode penelitian dimulai dengan melakukan pengumpulan data sosial ekonomi dan kesehatan dari kabupaten/kota di Jawa Timur. Data yang sudah diproses kemudian direduksi dimensinya menggunakan *Principal Component Analysis* (PCA) untuk menyederhanakan variabel berkorelasi. Tahap berikutnya adalah klasterisasi dengan algoritma klasterisasi *K-Means* untuk membentuk klaster dan *K-Medoids* untuk menangani *outlier*. Evaluasi klaster dilakukan dengan *silhouette score* dan *Calinski-Harabasz Index* untuk menilai kualitas klaster. Hasil klasterisasi kemudian diinterpretasikan dan divisualisasikan, serta disimpulkan guna mengatasi ketimpangan sosial ekonomi dan kesehatan di Jawa Timur.Penelitian ini menggunakan data indikator kesehatan dan sosial ekonomi dari Jawa Timur tahun 2023, yang mencakup 29 kabupaten dan 9 kota dengan sumber data dari website BPS Jatim https://jatim.bps.go.id/id. Tabel 1 menampilkan kumpulan variabel yang digunakan dalam penelitian ini.

Tabel 1 Variabel Penelitian

Variabel	Penjelas	Keterangan
X1	Persentase Wanita yang Kawin dan Menggunakan KB	Numerik
X2	Angka Harapan Hidup (Tahun)	Numerik
X3	Rata-Rata Usia Perkawinan	Numerik
X4	Jumlah Kasus Penyakit - HIV/AIDS	Numerik
X5	Persentase Rumah Tangga yang memiliki Fasilitas Buang Air Besar Pribadi	Numerik
X6	Kepadatan Penduduk per km persegi (km^2)	Numerik
X7	Persentase Balita di Jawa Timur dengan Pemberian Imunisasi Lengkap 2023	Numerik
X8	Tingkat Pengangguran Terbuka (TPT) Per Kabupaten/Kota (Persen)	Numerik
X9	Rumah Tangga yang Memiliki Sanitasi Layak	Numerik
X10	Persentase Penduduk dengan kemampuan Membaca dan Menulis, Dapat Baca Tulis Huruf/Latin, Arab, atau Lainnya	Numerik
X11	Rata-rata Lama Sekolah (Tahun)	Numerik
X12	Rata-rata Jumlah Rokok per Minggu yang dikonsumsi dalam 1 Bulan Terakhir	Numerik
X13	Persentase Wanita yang Pernah Menikah dan Melahirkan Hidup	Numerik
X14	Jumlah Penduduk	Numerik
X15	Laju Pertumbuhan Penduduk per Tahun	Numerik
X16	Produk Domestik Regional Bruto (PDRB) atau Produk Domestik Bruto (PDB) atas Harga Berlaku (dalam miliar rupiah)	Numerik
X17	Jumlah Penduduk Miskin	Numerik
X18	Jumlah Puskesmas	Numerik
X19	Persentase Penduduk dengan BPJS Kesehatan	Numerik

Proses penelitian ini melibatkan beberapa tahap, mulai dari *preprocessing* data hingga interpretasi dan kesimpulan, sebagaimana ditampilkan pada Gambar 1.



Gambar 1 Alur Penelitian

2.5 Tahapan Penelitian:

1. Pengumpulan Data
   Langkah awal adalah mengumpulkan data sosial ekonomi dan kesehatan dari kabupaten/kota di Jawa Timur. Data ini diperoleh dari sumber seperti BPS Jawa Timur.
2. Preprocessing Data
   Pada langkah ini, data yang telah dikumpulkan akan diproses lebih lanjut. *Preprocessing* mencakup langkah-langkah berikut:
   a. Membersihkan data dengan menghapus data yang tidak lengkap, inkonsisten, atau kurang relevan.
   b. Melakukan standarisasi data untuk memastikan setiap variabel berada dalam skala yang sama.
3. Reduksi Dimensi dengan PCA (*Principal Component Analysis)*
   PCA digunakan untuk mereduksi variabel dengan korelasi tinggi menjadi sejumlah komponen utama yang mewakili variasi terbesar dalam data. Tahap ini menghasilkan dataset yang lebih ringkas namun tetap mempertahankan informasi penting.
4. Klasterisasi dengan *K-Means* dan *K-Medoids*

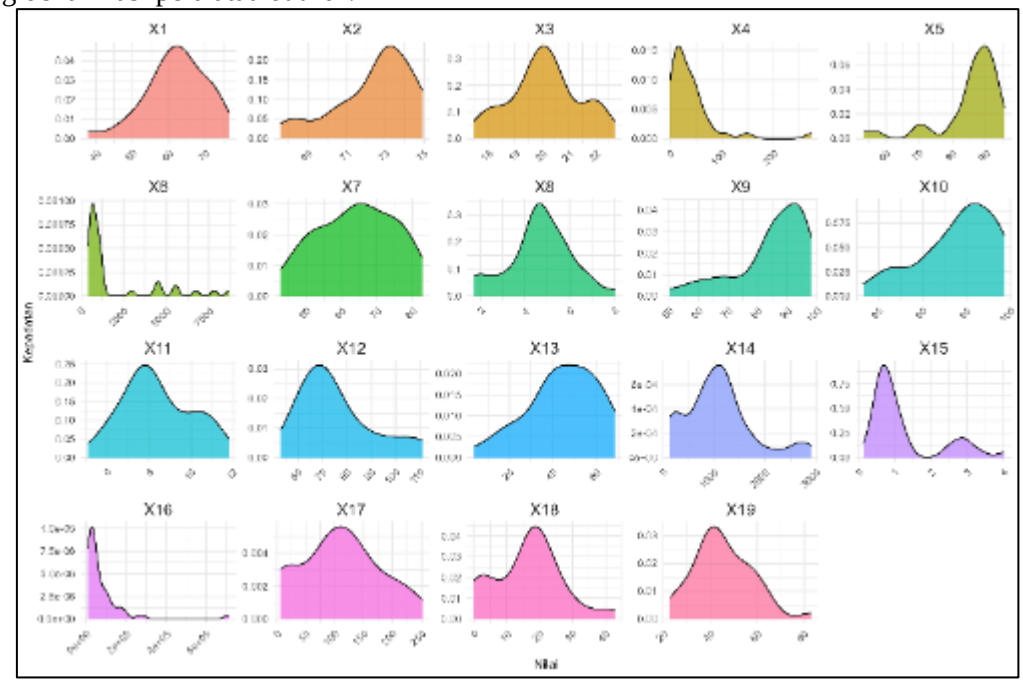
Setelah PCA diterapkan, data yang telah direduksi dimensinya akan diproses untuk klasterisasi. Algoritma *K-Means* dan *K-Medoids* akan diterapkan untuk mengelompokkan kabupaten/kota berdasarkan indikator sosial ekonomi dan kesehatan.

5. Evaluasi Klaster
   Setelah klasterisasi dilakukan, evaluasi hasil klasterisasi akan dilakukan untuk menentukan kualitas dari kluster yang terbentuk dengan menggunakan *silhouette score* dan *Calinski-Harabasz Index*.
6. Interpretasi Hasil dan Visualisasi
   Hasil klasterisasi akan diinterpretasikan untuk memberikan pemahaman yang lebih dalam terkait ketimpangan sosial ekonomi dan kesehatan antar kabupaten/kota di Jawa Timur. Visualisasi kluster digunakan untuk menampilkan hasil secara lebih jelas dengan menggunakan plot *scatter*, peta tematik, dan dendrogram.
7. Penyusunan Kesimpulan dan Saran
   Langkah akhir adalah menyusun kesimpulan berdasarkan hasil penelitian.. Berdasarkan klaster yang terbentuk, kesimpulan yang diambil akan memberikan gambaran mengenai kondisi sosial ekonomi dan kesehatan antar wilayah di Jawa Timur. Saran diberikan untuk mendukung analisis lebih lanjut serta pengembangan metode yang lebih efektif dalam mengelompokkan data sosial ekonomi dan kesehatan, guna memahami ketimpangan yang ada dengan lebih mendalam.

3. HASIL DAN PEMBAHASAN

3.1 Eksplorasi Data

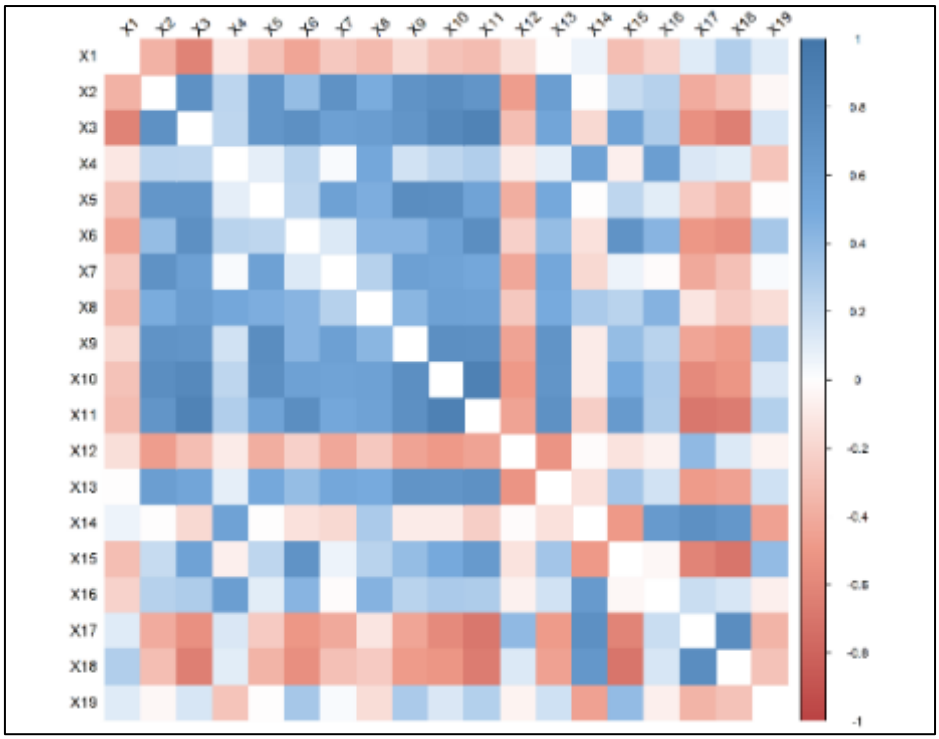
Eksplorasi data dengan plot densitas menggambarkan distribusi indikator sosial ekonomi dan kesehatan di Jawa Timur 2023, memberikan wawasan tentang persebaran data serta mengidentifikasi pola atau *outlier*.



Gambar 2 Plot Densitas untuk Indikator Sosial Ekonomi dan Kesehatan Jawa Timur 2023

Plot densitas indikator sosial ekonomi dan kesehatan di Jawa Timur 2023 menunjukkan distribusi data pada variabel X6 (Kepadatan Penduduk), menunjukkan distribusi data yang agak terpusat di kisaran nilai yang lebih tinggi, dengan skewness ke kiri. Hal ini mungkin mengindikasikan bahwa sebagian besar wilayah memiliki kepadatan penduduk yang sangat

tinggi, dengan beberapa kabupaten/kota yang lebih jarang penduduknya. Sementara itu, X2 (Angka Harapan Hidup), memperlihatkan distribusi yang lebih mendekati normal, dengan puncak yang berada di sekitar nilai sedang. Kombinasi kedua distribusi ini bisa memberikan wawasan menarik, seperti bagaimana kepadatan penduduk yang tinggi cenderung berasosiasi dengan variabilitas dalam harapan hidup, yang berpotensi membuka diskusi tentang akses ke layanan kesehatan dan kualitas hidup di wilayah-wilayah dengan berbagai tingkat kepadatan.



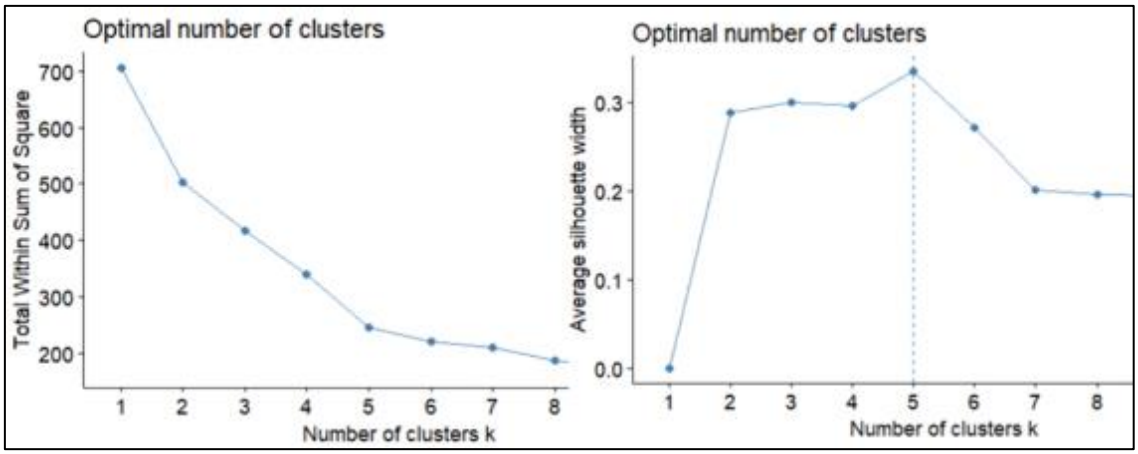
Gambar 3 Korelasi Matriks Indikator Sosial Ekonomi dan Kesehatan Jawa Timur 2023

Hasil korelasi matriks terlihat adanya korelasi tinggi antar variabel sosial ekonomi di Jawa Timur, yang dapat memicu multikolinieritas dalam analisis statistik. Misalnya, X16 (PDRB) dan X14 (Jumlah Penduduk) memiliki korelasi positif yang kuat, sementara X17 (Jumlah Penduduk Miskin) berkorelasi negatif dengan X19 (Persentase Penduduk dengan BPJS Kesehatan). Multikolinieritas ini dapat memengaruhi akurasi model, sehingga diperlukan teknik seperti PCA untuk mereduksi dimensi data dan mengatasi masalah ini.

3.2 Klasterisasi Tanpa *Principal Component Analysis* (PCA)

3.2.1. Metode Klasterisasi *K-Means*

Algoritma *K-Means* merupakan sebuah algoritma klaster *nonhierarki* yang membutuhkan penentuan jumlah klaster terlebih dahulu. Oleh karena itu, metode *Elbow* dan *Silhouette* digunakan untuk mengidentifikasi jumlah klaster yang paling sesuai, sebagaimana ditunjukkan pada Gambar 4.



(a) (b)

Gambar 4 (a) Metode *elbow* (b) Metode *Silhouette*

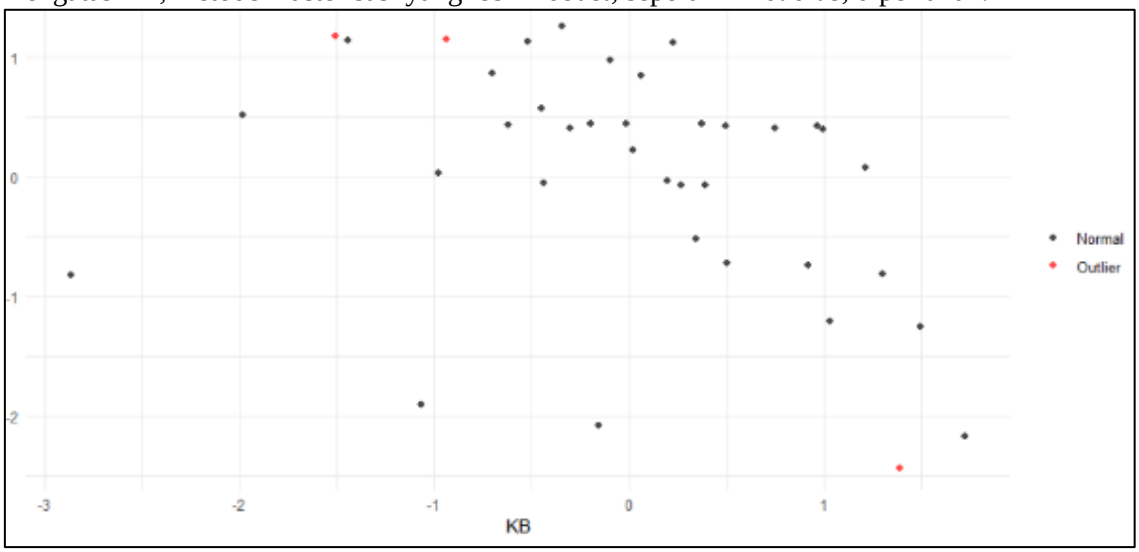
Hasil penentuan jumlah kluster dengan metode *Elbow*, TWSS mulai landai setelah klaster kelima, menunjukkan bahwa jumlah klaster optimal berkisar antara 2 hingga 5. Metode *Silhouette* menunjukkan puncak pada klaster kelima, mengindikasikan 5 klaster memberikan hasil terbaik dengan struktur klaster yang lebih terpisah dan kohesif.

Gambar 5 Visualisasi Hasil *K-Means* Clustering

Hasil klasterisasi dengan metode *K-Means* wilayah Jawa Timur tahun 2023 terbagi menjadi lima klaster dengan karakteristik sosial ekonomi dan kesehatan yang berbeda. Klaster pertama (merah) mencakup Jember dan sekitarnya, menunjukkan tantangan sosial ekonomi yang lebih besar. Klaster kedua (hijau) terdiri dari wilayah Madura seperti Pamekasan dan Sampang, yang memerlukan peningkatan sosial ekonomi. Klaster ketiga (*cyan*) mencakup kota-kota seperti Malang dan Pasuruan dengan perkembangan sosial ekonomi lebih baik. Klaster keempat (biru) meliputi wilayah tengah dan barat seperti Pacitan dan Ngawi, dengan indikator yang cukup

konsisten. Klaster kelima (ungu) terdiri dari Sidoarjo dan Surabaya, yang menampilkan kondisi sosial ekonomi dan kesehatan paling maju.

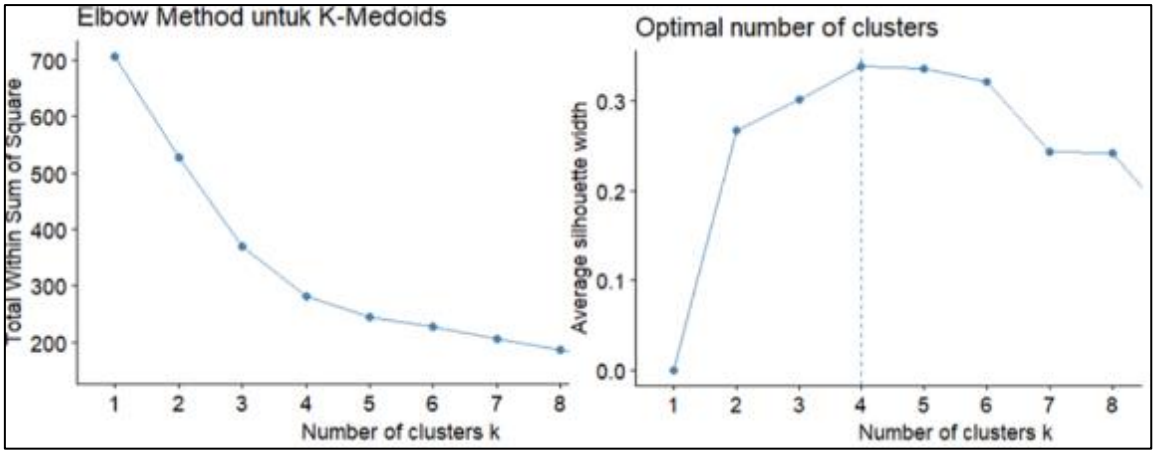
Hasil klasterisasi *K-Means* kurang *robust* terhadap *outlier*, karena algoritma ini menggunakan *centroid* sebagai rata-rata data dalam klaster. *Outlier* dapat menggeser posisi *centroid*, sehingga hasil klasterisasi menjadi kurang representatif. Hasil klusterisasi dengan *K-Means* menunjukkan *outlier* di wilayah Bondowoso dan Sidoarjo, yang dapat memengaruhi hasil *K-Means* karena *centroid* sensitif terhadap *outlier* seperti yang ditunjukkan pada Gambar 6. Untuk mengatasi ini, metode klasterisasi yang lebih *robust*, seperti *K-Medoids*, diperlukan.



Gambar 6 Scatter Plot *Outlier*

3.2.2. Metode Klasterisasi K-Medoids

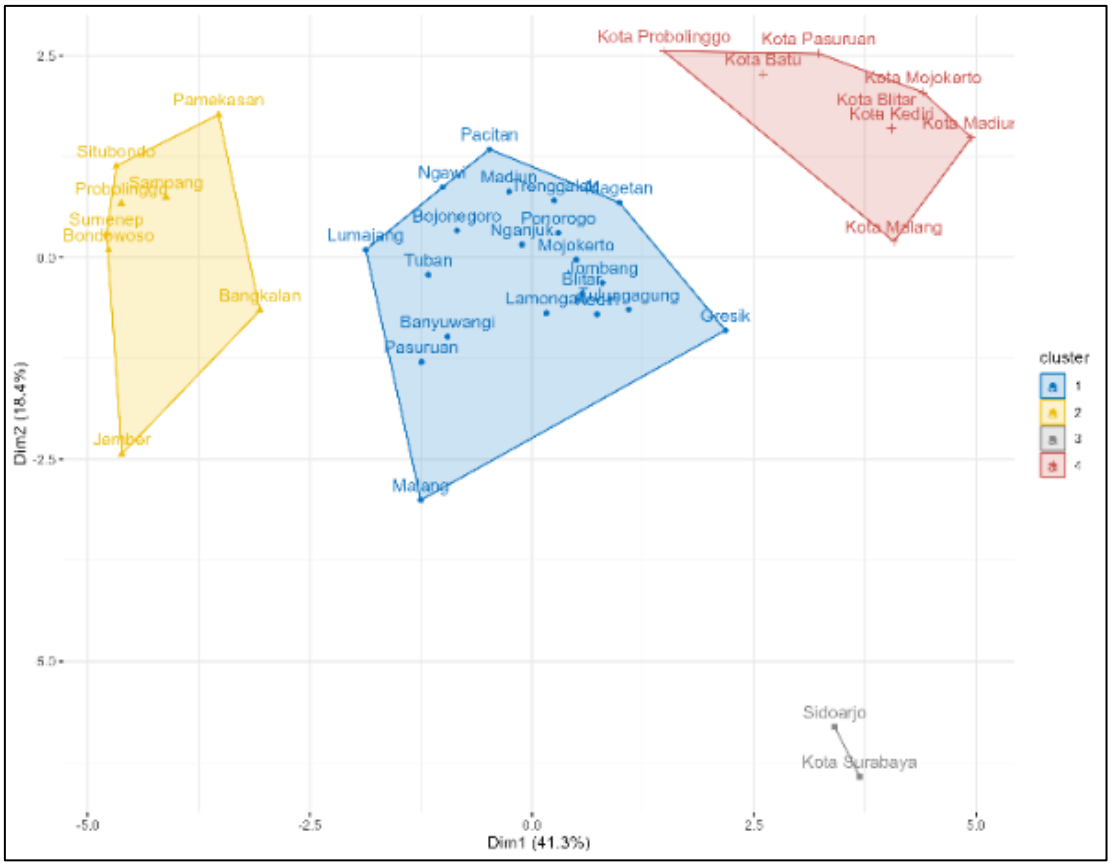
K-Medoids menetapkan medoid sebagai pusat klaster, sehingga lebih *robust* terhadap *outlier* dan mampu menghasilkan klaster yang lebih akurat. Metode *Elbow* dan *Silhouette* digunakan untuk mengidentifikasi jumlah klaster optimal, sebagaimana diperlihatkan Gambar 7.



(a) (b)

Gambar 7 (a) Metode *Elbow* (b) Metode *Silhouette*

Hasil identifikasi jumlah kluster menggunakan metode *Silhouette* dan *Elbow*, jumlah klaster optimal adalah 4. Grafik *Silhouette* menunjukkan nilai tertinggi pada klaster ke-4, menunjukkan pemisahan klaster yang baik. Hasil *Elbow* juga mendukung, dengan penurunan signifikan pada TWSS setelah klaster ke-4.



Gambar 8 Visualisasi Hasil *K-Medoids* Clustering

Hasil klasterisasi dengan *K-Medoids* menunjukkan bahwa wilayah Jawa Timur terbagi dalam 4 klaster berdasarkan indikator sosial ekonomi dan kesehatan. Klaster 1 (biru) mencakup wilayah seperti Banyuwangi dan Gresik dengan indikator sosial ekonomi yang relatif baik. Klaster 2 (kuning) meliputi wilayah Madura seperti Bangkalan dan Sampang yang memerlukan peningkatan. Klaster 3 (abu-abu) terdiri dari kota-kota besar seperti Surabaya dan Sidoarjo yang lebih maju. Klaster 4 (merah) mencakup kota-kota seperti Blitar dan Pasuruan dengan karakteristik berbeda dari klaster lainnya.

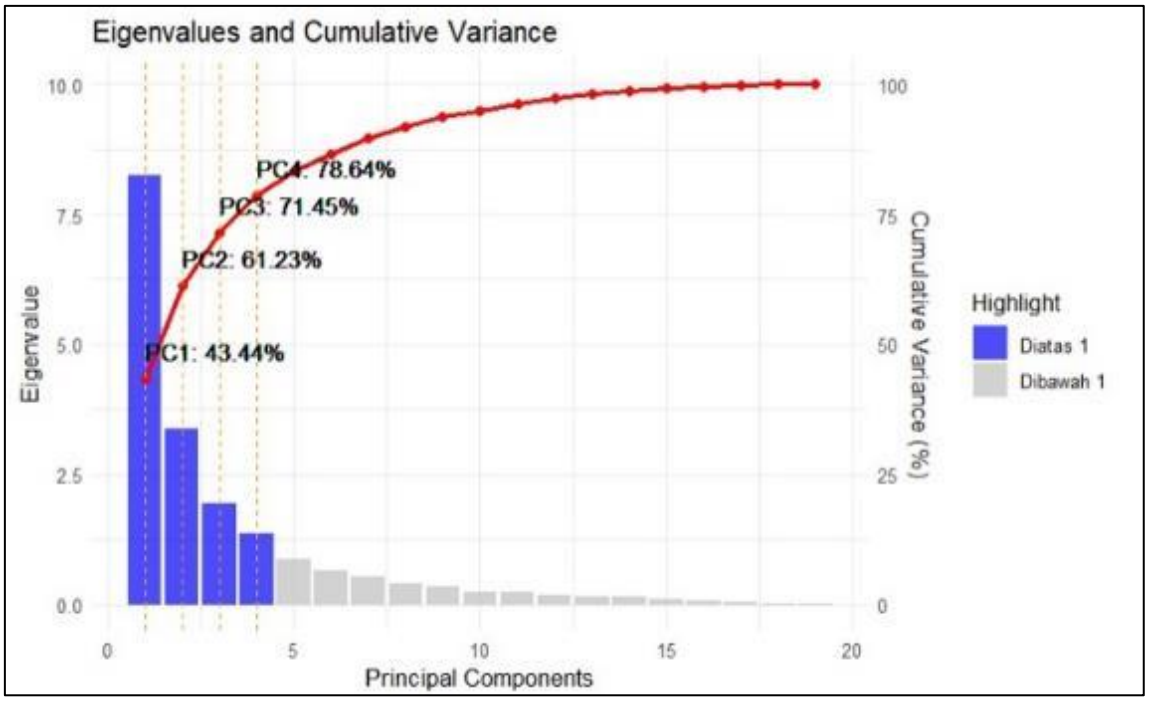
3.3 Reduksi Dimensi Principal Component Analysis

Untuk memastikan data siap untuk algorima klasterisasi, kecukupan sampel harus diperiksa dengan uji KMO serta korelasi antar variabel menggunakan uji *Bartlett.*

Tabel 2 Nilai KMO dan Uji *Barlett*

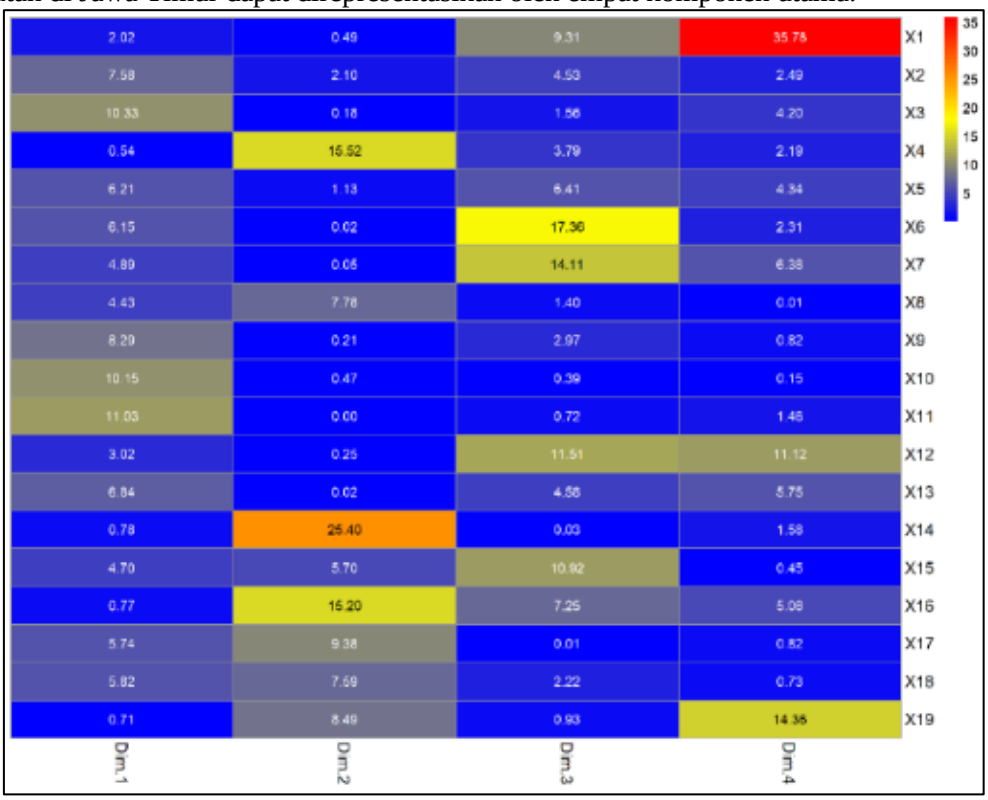
Uji		Nilai
Kaiser-Meyer-Olkin (KMO)		0.75
Bartlett's Test of Sphericity	Approx. Chi-Square	701.3841
	Degrees of freedom	171
	Sig.	0.000

Tabel 2 memperlihatkan bahwa data cukup secara statistik dan reliabel, dengan nilai KMO 0,75 yang nilai tersebut lebih besar dari 0,6. Hubungan signifikan antara variabel ditunjukkan oleh *p-value* dari uji *Bartlett* yang lebih rendah dari taraf nyata $\alpha = 0{,}05$. sehingga langkah-langkah untuk mengurangi dimensi harus dilakukan. Jumlah komponen utama yang digunakan dihitung dengan acuan nilai *eigen* yang lebih besar dari 1.



Gambar 9 Nilai *Eigenvalue* dan Varian Kumulatif

Nilai *eigenvalue* menunjukkan hasil PCA dengan empat komponen utama yang signifikan karena nilai *eigenvalue* lebih dari 1. PC1 menjelaskan 43,44% variansi, diikuti oleh PC2 (17,79%), PC3 (10,22%), dan PC4 (7,19%), dengan total 78,64% variansi. Hal ini menunjukkan bahwa sebagian besar informasi penting dalam indikator sosial ekonomi dan kesehatan di Jawa Timur dapat direpresentasikan oleh empat komponen utama.



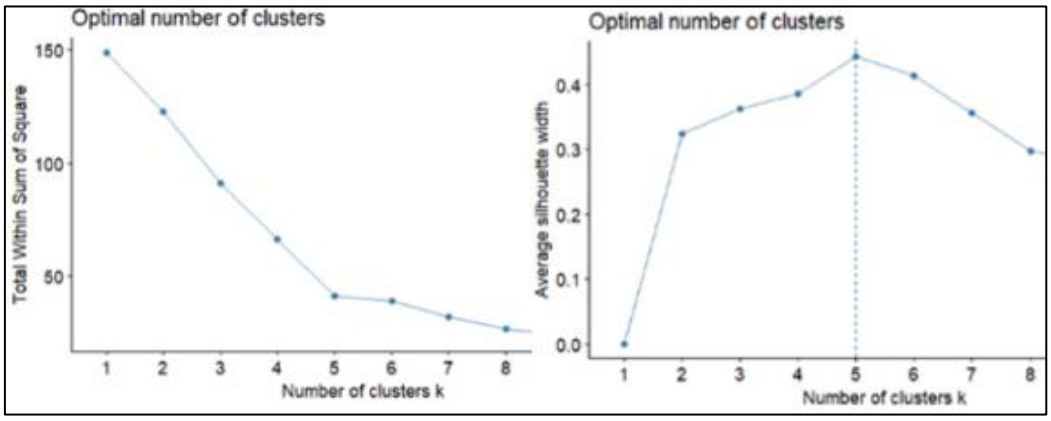
Gambar 10 Kontribusi Variabel Terhadap Komponen Utama

Gambar 10 menunjukkan kontribusi variabel terhadap empat komponen utama PCA. X1 (Persentase Wanita yang Kawin dan Menggunakan KB) mendominasi Dimensi 4 dengan kontribusi tertinggi (35.78), sementara X6 (Kepadatan Penduduk) memberikan kontribusi kuat di Dimensi 2 dan 3, menunjukkan pentingnya faktor demografi dalam variasi data. X14 (Jumlah Penduduk) dan X19 (Persentase Penduduk dengan BPJS Kesehatan) juga menunjukkan pengaruh besar dalam membentuk pola klaster sosial ekonomi. Analisis ini menyoroti variabel-variabel kunci yang sangat berpengaruh dalam memisahkan klaster sosial ekonomi di Jawa Timur.

3.4 Klasterisasi Tanpa Principal Component Analysis (PCA)

3.4.1. Metode Klasterisasi *K-Means*

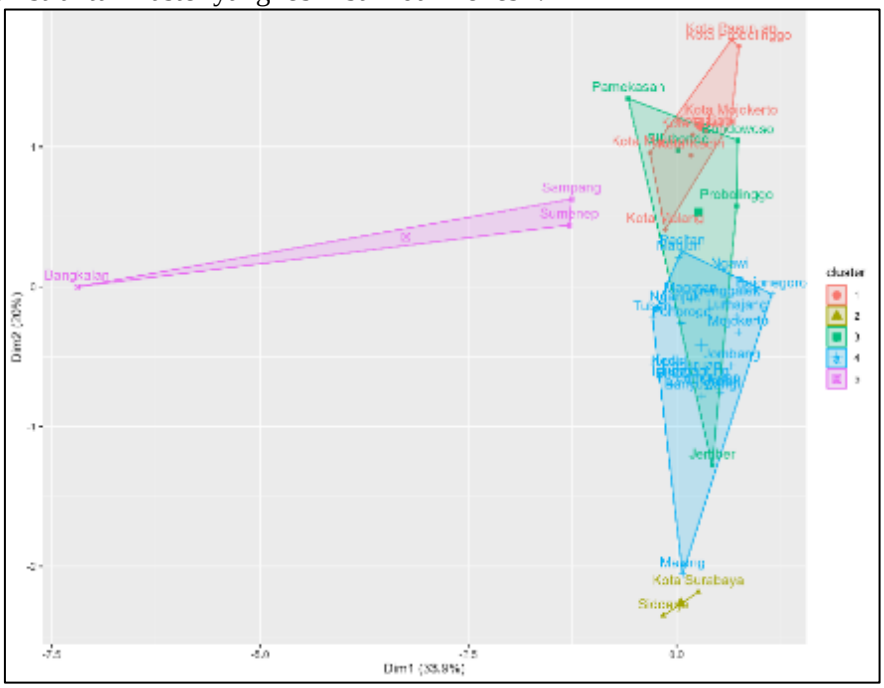
Penentuan jumlah klaster optimal pada algoritma *K-Means* dilakukan menggunakan dua metode, yaitu *Elbow* dan *Silhouette* yang dihasilkan oleh Gambar 11.



(a) (b)

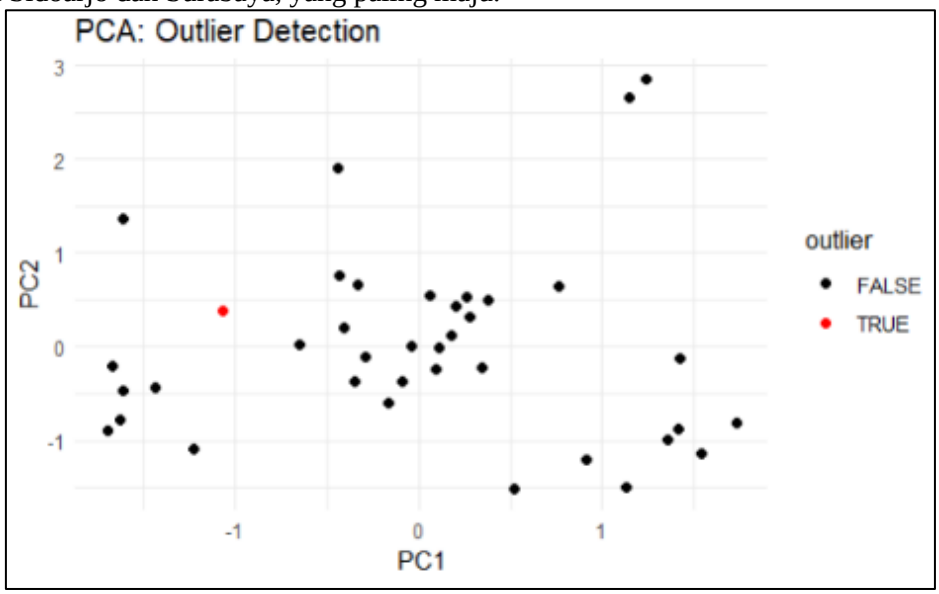
Gambar 11 (a) Metode *Elbow* (b) Metode *Silhouette*

Metode *Elbow* menunjukkan bahwa setelah klaster kelima, penambahan klaster tidak mengurangi TWSS secara signifikan, mengindikasikan jumlah optimal antara 2 hingga 5 klaster. Metode *Silhouette* juga menunjukkan nilai tertinggi pada klaster kelima, yang berarti 5 klaster memberikan struktur klaster yang lebih baik dan kohesif.



Gambar 12 Visualisasi Hasil *K-Means* dengan PCA

Hasil klasterisasi *K-Means* wilayah Jawa Timur terbagi menjadi lima klaster. Klaster 1 (merah) mencakup Kota Pasuruan dan Mojokerto dengan tantangan sosial ekonomi. Klaster 2 (hijau) meliputi wilayah Madura seperti Pamekasan dan Bangkalan yang masih memerlukan peningkatan. Klaster 3 (*cyan*) mencakup Jember hingga Probolinggo dengan kondisi lebih baik. Klaster 4 (biru) terdiri dari Bojonegoro dan Ngawi, dengan variasi indikator. Klaster 5 (ungu) meliputi Sidoarjo dan Surabaya, yang paling maju.

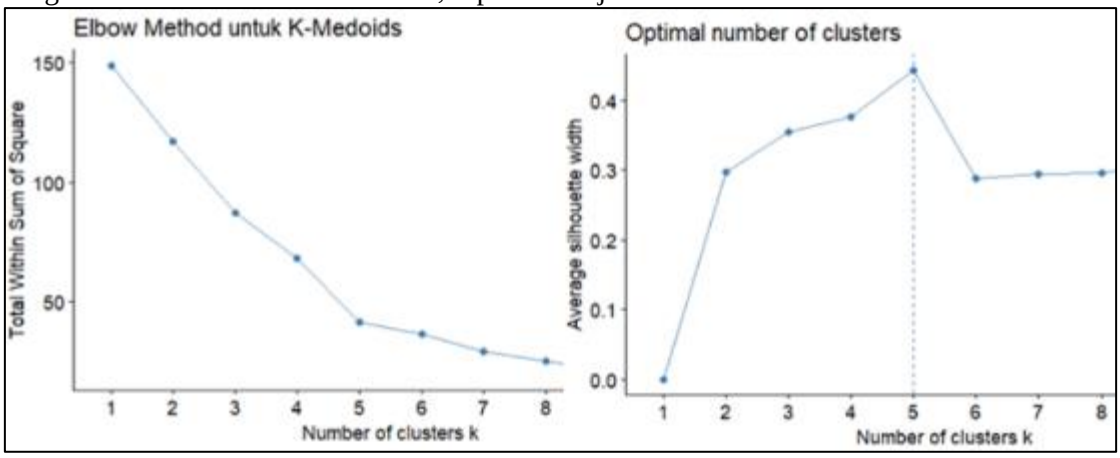


Gambar 13 Mendeteksi Outlier

Kabupaten Bangkalan teridentifikasi sebagai *outlier* dalam PCA pada Dim.4 dengan nilai -3.6, menunjukkan perbedaan signifikan dalam karakteristik sosial ekonomi dan kesehatan. Ini menekankan pentingnya metode *robust* seperti *K-Medoids* untuk menangani *outlier*.

3.4.2. Metode Klasterisasi *K-Medoids*

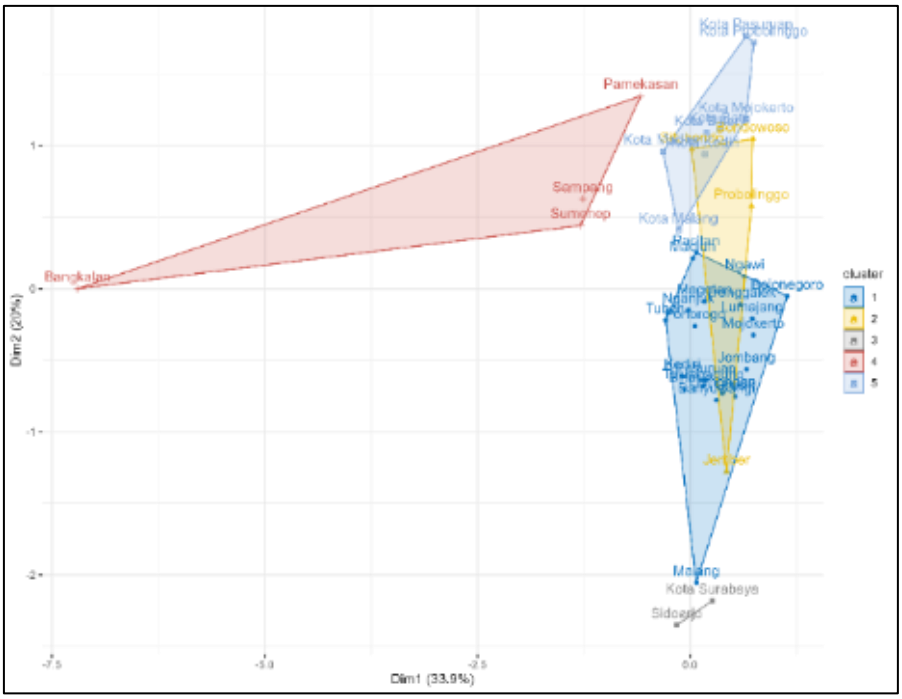
K-Medoids menetapkan medoid sebagai pusat klaster, sehingga lebih *robust* terhadap *outlier* dan mampu menghasilkan klaster yang lebih akurat. Jumlah klaster terbaik ditentukan dengan metode *Elbow* dan *Silhouette*, seperti ditunjukkan oleh Gambar 14.



(a) (b)

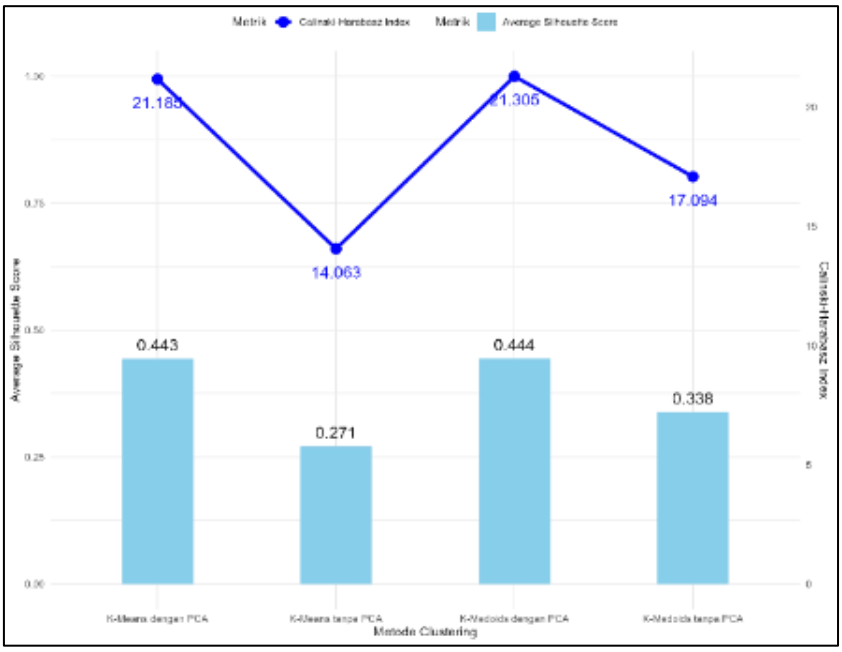
Gambar 14 (a) Metode *Elbow* (b) Metode *Silhouette*

Hasil identifikasi jumlah klaster dengan *Silhouette* dan *Elbow*, jumlah klaster optimal adalah 4. *Silhouette* menunjukkan nilai tertinggi pada klaster ke-4, menandakan pemisahan klaster yang baik. Hasil *Elbow* juga mendukung dengan penurunan signifikan pada TWSS setelah klaster ke-4.



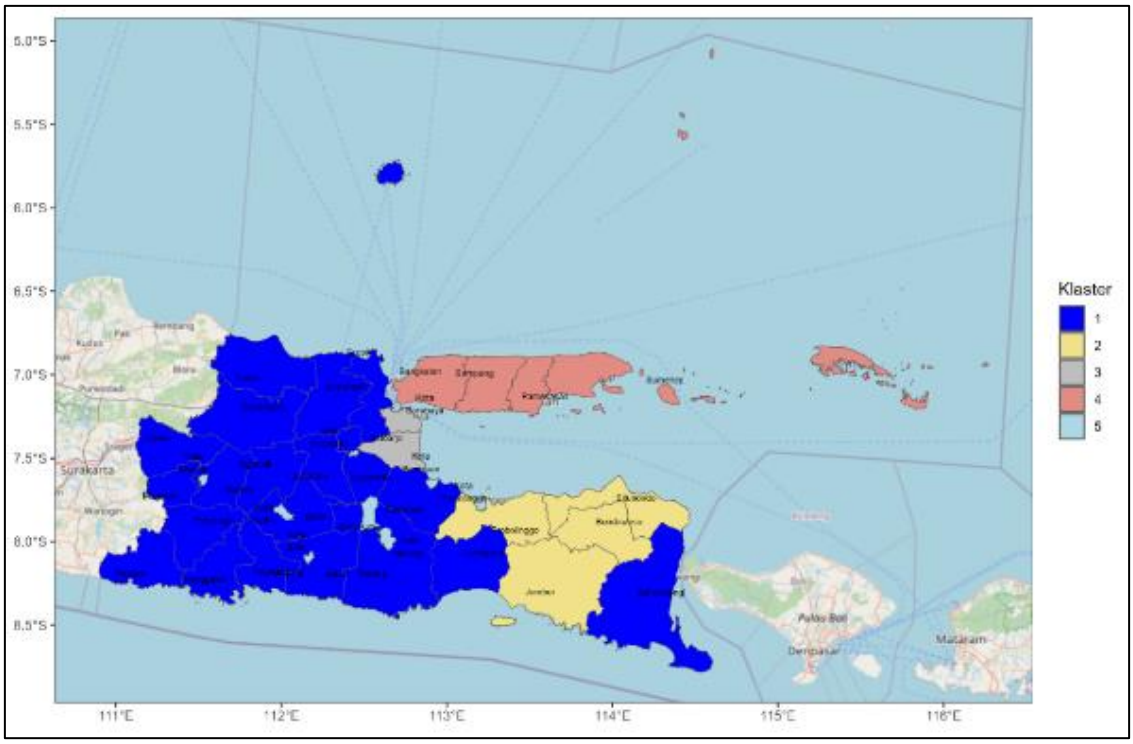
Gambar 15 Visualisasi Hasil *K-Medoids Clustering*

Wilayah Jawa Timur tahun 2023 terbagi dalam lima klaster sosial ekonomi dan kesehatan berdasarkan metode *K-Medoids*. Klaster 1 (merah) meliputi Bangkalan dan Pamekasan, Klaster 2 (kuning) termasuk Probolinggo, Klaster 3 (biru muda) mencakup Kota Pasuruan, Klaster 4 (abu-abu) meliputi Kota Malang dan Surabaya, dan Klaster 5 (biru tua) terdiri dari wilayah lainnya dengan karakteristik serupa.



Gambar 16 Perbandingan Hasil Algoritma Klasterisasi dengan PCA

Perbandingan hasil menunjukkan bahwa *K-Medoids* dengan PCA memiliki *Average Silhouette Score* tertinggi (0.444), menunjukkan klaster yang lebih terpisah dan kohesif. *K-Means* tanpa PCA memiliki performa terendah. Berdasarkan *Calinski-Harabasz* Index, *K-Medoids* dengan PCA mencetak nilai tertinggi (21.305), menunjukkan klaster yang paling optimal. Secara keseluruhan, penggunaan PCA meningkatkan kualitas klasterisasi, terutama pada *K-Means*.



Gambar 17 Peta Jawa Timur dengan Metode *K-medoids* dengan PCA

Penelitian ini memberikan kontribusi signifikan dalam bidang klasterisasi data sosial ekonomi dan kesehatan dengan mengombinasikan *Principal Component Analysis* (PCA) dan *K-Medoids* secara simultan, yang jarang diaplikasikan dalam penelitian sebelumnya. Metode ini terbukti efektif dalam menghadapi data berdimensi tinggi dengan *outlier*. Penggunaan PCA dalam penelitian ini berfungsi untuk mengurangi multikolinieritas antar variabel, sehingga mengurangi kompleksitas dan meningkatkan efisiensi analisis klasterisasi yang terbukti melalui nilai *Average Silhouette Score* tertinggi (0,444) dan *Calinski-Harabasz Index* (21,305) menunjukkan klaster yang lebih terpisah dan koheren, serta memperbaiki performa klasterisasi pada data berdimensi tinggi [11].

Secara khusus, penelitian ini menunjukkan bahwa penggunaan *K-Medoids* dan PCA secara bersamaan lebih *robust* dalam menangani *outlier* dibandingkan dengan *K-Means*, yang sensitif terhadap *outlier*. Hasil ini menegaskan keunggulan *K-Medoids* dalam aplikasi data sosial ekonomi yang mengandung nilai ekstrem, memberikan klaster yang lebih representatif dan stabil. Selain itu, temuan ini relevan bagi pemangku kebijakan karena memungkinkan analisis yang lebih mendalam mengenai perbedaan sosial ekonomi dan kesehatan di berbagai wilayah, serta dapat menjadi dasar untuk pengembangan kebijakan yang lebih tepat sasaran dalam mengurangi ketimpangan di Jawa Timur.

Kebaruan lainnya terletak pada aplikasi metode ini dalam konteks wilayah spesifik, yaitu Provinsi Jawa Timur, dengan fokus pada indikator sosial ekonomi dan kesehatan yang jarang dijadikan objek penelitian dalam kombinasi metode yang digunakan. Pendekatan ini memungkinkan analisis yang lebih detail dan akurat terhadap ketimpangan di berbagai wilayah, yang diharapkan mampu memberikan dasar kebijakan yang lebih spesifik. Dengan membuktikan bahwa PCA dan *K-Medoids* dapat secara signifikan meningkatkan ketepatan dalam klasterisasi data berdimensi tinggi, penelitian ini memperluas cakupan studi klasterisasi menjadi lebih aplikatif dan relevan bagi pemerintah daerah dalam upaya pemerataan sosial ekonomi dan peningkatan akses layanan kesehatan.

Berbeda dengan penelitian lain yang umumnya mengandalkan satu metode klasterisasi, penelitian ini memberikan inovasi dengan penerapan dua metode yang saling melengkapi: PCA untuk mengurangi multikolinieritas dan *K-Medoids* untuk mengatasi *outlier*. Penelitian ini berbeda dari studi yang hanya mengandalkan *K-Means* atau tidak menggunakan reduksi dimensi,

karena sensitivitas *K-Means* terhadap *outlier* dan multikolinieritas terbukti dapat mengurangi akurasi klasterisasi. Penggunaan PCA dalam penelitian ini terbukti efektif dalam menangani masalah multikolinieritas yang sering diabaikan pada studi sebelumnya, memberikan kebaruan dalam aspek metodologi dan aplikasi di lapangan. Studi terbaru oleh [7] juga menegaskan pentingnya PCA dalam mengurangi multikolinearitas dan mempermudah interpretasi hasil. *Calinski-Harabasz Index* mendukung hasil penelitian ini dengan *K-Medoids* menggunakan PCA memiliki indeks tertinggi (21.305), menunjukkan separasi dan kepadatan optimal [12]. Sebaliknya, *K-Means* tanpa PCA memperlihatkan performa yang lebih rendah menunjukkan bahwa metode ini kurang efektif tanpa pengurangan dimensi [5]. Hal ini menunjukkan bahwa penggunaan PCA mampu meningkatkan efisiensi klasterisasi, terutama dalam konteks data dengan banyak variabel yang saling berhubungan.

Studi lain juga mendukung pentingnya PCA dalam klasterisasi untuk mengurangi *noise* dan mempercepat proses tanpa mengorbankan informasi penting [6]. Secara keseluruhan, hasil penelitian ini menegaskan peran penting PCA dalam meningkatkan kualitas klasterisasi. Kombinasi PCA dan *K-Medoids* menghasilkan klaster yang lebih baik dalam hal separasi dan kepadatan dan lebih unggul dari *K-Means*. Penelitian oleh [13]. juga memperkuat temuan ini, di mana PCA digunakan untuk mereduksi dimensi sebelum menerapkan K-Means, yang menghasilkan klaster yang lebih akurat dalam data berdimensi tinggi Demikian pula, [14] membuktikan bahwa K-Medoids dengan PCA menghasilkan klaster yang lebih konsisten, khususnya dalam menangani outlier pada data kompleks [15] menggunakan kombinasi PCA dan K-Medoids untuk segmentasi pelanggan dan menemukan bahwa metode ini dapat mengatasi masalah multikolinearitas serta meningkatkan akurasi klasterisasi. Studi oleh [16] mengonfirmasi bahwa PCA yang terintegrasi dengan algoritma klasterisasi seperti K-Medoids dapat mengelola kompleksitas big data dengan lebih baik, menyoroti kemampuan PCA dalam meningkatkan efisiensi klasterisasi. Selain itu, [17] juga menunjukkan bahwa kombinasi PCA dan K-Medoids memberikan klaster yang lebih kohesif dan efektif dalam menangani noise, mendukung metodologi penelitian ini untuk menghasilkan klaster yang terpisah dan stabil pada data sosial ekonomi

Secara keseluruhan, hasil penelitian ini menegaskan peran penting PCA dalam meningkatkan kualitas klasterisasi. Kombinasi PCA dan K-Medoids menghasilkan klaster yang lebih baik dalam hal separasi dan kepadatan, serta lebih unggul dari K-Means. Temuan ini konsisten dengan penelitian terbaru yang menunjukkan bahwa PCA dapat membantu meningkatkan akurasi metode klasterisasi dalam data yang kompleks [18].

4. KESIMPULAN DAN SARAN

Metode klasterisasi *K-Means* dan *K-Medoids* yang didukung dengan *Principal Component Analysis (PCA)* diterapkan dalam penelitian ini untuk menganalisis indikator sosial ekonomi dan kesehatan di Jawa Timur tahun 2023. Hasil penelitian menghasilkan bahwa *K-Medoids* dengan PCA memberikan klasterisasi yang lebih akurat dan *robust* dibandingkan *K-Means*, terutama dalam menangani *outlier*. Jumlah klaster optimal ditentukan melalui metode *Elbow* dan *Silhouette*, yang menunjukkan bahwa empat hingga lima klaster merupakan pilihan terbaik untuk menggambarkan variasi antar kabupaten/kota. Penelitian ini berhasil mereduksi dimensi data dengan menggunakan PCA, sehingga proses klasterisasi dapat berjalan lebih efisien dan akurat. PCA juga membantu algoritma *K-Medoids* untuk menghasilkan klaster yang lebih terpisah dan koheren, dengan menangani *outlier* secara efektif. Diharapkan temuan penelitian ini dapat memberikan pandangan yang membantu dalam merumuskan kebijakan yang lebih efektif untuk mengurangi ketimpangan sosial ekonomi dan kesehatan di berbagai wilayah di Jawa Timur. Saran penelitian ini meliputi perluasan cakupan wilayah dan waktu, penggunaan metode klasterisasi lain seperti DBSCAN untuk data dengan banyak *outlier*, serta penambahan variabel lingkungan dan infrastruktur. Evaluasi bersama pemangku kebijakan diperlukan agar hasil klasterisasi dapat diimplementasikan secara efektif.

DAFTAR PUSTAKA

[1] R. Sonawane and H. D.Patil, "Prediction of Heart Disease by Optimized Distance and Density-Based Clustering," *Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, 2022, doi: 10.1109/ICAIS53314.2022.9742885.

[2] A. Arifa, C. Rindu, R. Astriratma, and A. Zaidiah, "K-Means Algorithm Implementation for Project Health Clustering," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, 2023, doi:10.29207/resti.v7i5.5181.

[3] P. Arora, Deepali, and S. Varshney, "Analysis of K-Means and K-Medoids Algorithm For Big Data," *Procedia Computer Science*, vol. 78, 2016, doi:10.1016/j.procs.2016.02.095.

[4] E. K. Korir, "Comparative clustering and visualization of socioeconomic and health indicators: A case of Kenya," *Socio-Economic Planning Sciences*, vol. 95, 2024, doi: 10.1016/j.seps.2024.101961.

[5] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-Means techniques," *Informatics in Medicine Unlocked*, 2019, doi: 10.1016/j.imu.2019.100179.

[6] S. Suhirman and H. Wintolo, "System for Determining Public Health Level Using the Agglomerative Hierarchical Clustering Method," *Compiler*, vol. 8, no. 1, pp. 1-8, 2019, doi: 10.28989/compiler.v8i1.425.

[7] R. Saha, M. T. Tariq, M. Hadi, and Y. Xiao, "Pattern Recognition Using Clustering Analysis to Support Transportation System Management, Operations, and Modeling," *Journal of Advanced Transportation*, pp. 1-12, 2019, doi: 10.1155/2019/1628417.

[8] I. A. Rosyada dan D. T. Utari, "Penerapan Principal Component Analysis untuk Reduksi Variabel pada Algoritma K-Means Clustering," *JAMBURA*, vol. 5, 2024.

[9] D. R. P. Sari, "Metode Principal Component Analysis (PCA) sebagai Penanganan Asumsi Multikolinearitas," *Parameter*, vol. 02, 2023.

[10] B. Wira, A. E. Budianto, dan A. S. Wiguna, "Implementasi Metode K Medoids Clustering untuk Mengetahui Pola Pemilihan Program Studi Mahasiswa Baru Tahun 2018 di Universitas Kanjuruhan Malang," *Rainstek*, vol. 1, 2019.

[11] M. Zhu and A. Ghodsi, "Automatic dimensionality selection from the scree plot via the use of profile likelihood," *Computational Statistics & Data Analysis*, vol. 51, 2018, doi: 10.1016/j.csda.2005.09.010.

[12] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," *arXiv preprint arXiv*, 2020, doi: 10.48550/arXiv.1811.12808.

[13] T. Jiang, S. Li, and Z. Li, "An enhanced K-means algorithm based on PCA for dimensionality reduction in high-dimensional data," *IEEE Access*, 2016.

[14] S. Sharma and S. Dey, "Performance analysis of clustering techniques for identifying patterns in higher-dimensional datasets," *Expert Systems with Applications*, 2019.

[15] S. Ahmed and M. Nadeem, "Application of PCA and K-Medoids clustering for customer segmentation," *International Journal of Advanced Computer Science and Applications*, 2018.

[16] A. Verma and P. Verma, "Dimensionality reduction using PCA integrated with clustering algorithms for big data," *Big Data Research*, 2020.

[17] J. Wang, Y. Zhang, and Z. Xu, "Exploring high-dimensional data using PCA and robust clustering methods," *IEEE Transactions on Cybernetics*, 2021.

[18] C. Boutsidis, M. W. Mahoney, and P. Drineas, "An Improved Approximation Algorithm for the Column Subset Selection Problem," *SIAM Journal on Computing*, 2015, doi: 10.1137/1.9781611973068.105.