E-ISSN : 2988-1986
https://ejournal.cahayailmubangsa.institute/index.php/kohesi

Kohesi: Jurnal Multidisiplin Saintek
Volume 8 No. 4 Tahun 2025

ANALISIS KLASTER KABUPATEN/KOTA DI PROVINSI JAWA TIMUR TAHUN 2023 BERDASARKAN ANGKA HARAPAN HIDUP DAN KEMISKINAN MENGGUNAKAN METODE K-MEANS DAN PCA

Alivia Salma Namira1, Genesis Bunga2, Erna Novita Anggie3, Nahda Hayu Hemalina4, Alfan Rizaldy Pratama5
Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Email: 23083010032@student.upnjatim.ac.id, 23083010037@student.upnjatim.ac.id, 23083010038@student.upnjatim.ac.id, 23083010066@student.upnjatim.ac.id, alfan.fasilkom@upnjatim.ac.id

ABSTRACT

This research aims to cluster regencies/cities in East Java Province based on the variables of life expectancy and poverty rate. The data used for clustering was sourced from the East Java Provincial Statistics Agency in 2023. The method used in this study is a combination of Principal Component Analysis (PCA) for dimensionality reduction and clustering for the clustering process. Evaluation of the clustering results was conducted using the Sum of Square Error (SSE) and Silhouette Score to determine the optimal number of clusters. The analysis results indicate that the optimal number of clusters is k = 2, with an SSE value difference of 7.97. Cluster 0 consists of 11 districts/cities that have the characteristics of high life expectancy and low poverty levels. Meanwhile, cluster 1 consists of 27 districts/cities with the opposite characteristics, namely lower life expectancy and higher poverty levels.

Keywords: *regional clustering; life expectancy; poverty; K-Means Clustering; Principal Component Analysis (PCA).*

ABSTRAK

Penelitian ini bertujuan untuk mengelompokkan kabupaten/kota di Provinsi Jawa Timur berdasarkan variabel angka harapan hidup dan tingkat kemiskinan. Data yang digunakan untuk klasterisasi dari Badan Pusat Statistik Provinsi Jawa Timur tahun 2023. Metode yang digunakan dalam penelitian ini adalah kombinasi antara Principal Component Analysis (PCA) untuk reduksi dimensi dan klasterisasi clustering untuk proses klasterisasi. Evaluasi terhadap hasil klasterisasi dilakukan menggunakan Sum of Square Error (SSE) dan Silhouette Score guna menentukan jumlah klaster yang optimal. Hasil analisis menunjukkan bahwa jumlah klaster optimal adalah k = 2, dengan selisih nilai SSE sebesar 7,97. Klaster 0 terdiri dari 11 kabupaten/kota yang memiliki karakteristik angka harapan hidup tinggi dan tingkat kemiskinan rendah. Sementara itu, klaster 1 terdiri dari 27 kabupaten/kota dengan karakteristik sebaliknya, yaitu angka harapan hidup lebih rendah dan tingkat kemiskinan lebih tinggi.

Article History
Received: Juni 2025
Reviewed: Juni 2025
Published: Juni 2025

Plagiarism Checker No 235
Prefix DOI : 10.8734/Kohesi.v1i2.365
Copyright : Author
Publish by : Kohesi

Kata kunci: klasterisasi wilayah; angka harapan hidup; kemiskinan; K-Means Clustering; Principal Component Analysis (PCA).

1. PENDAHULUAN

Pembangunan manusia merupakan tujuan utama dari kebijakan publik yang tercermin melalui angka harapan hidup dan kemiskinan. Di Indonesia, khususnya di Jawa Timur, terdapat ketimpangan sosial-ekonomi antar kabupaten/kota, yakni beberapa wilayah memiliki angka harapan hidup tinggi namun disertai tingkat kemiskinan yang masih tinggi, dan sebaliknya. Ketimpangan ini menghambat pembangunan yang merata dan berdampak pada kesejahteraan masyarakat secara keseluruhan. Oleh karena itu, pemahaman mendalam mengenai pola ketimpangan ini sangat mendesak untuk diwujudkan. Untuk memahami pola ketimpangan ini secara lebih mendalam, metode clustering seperti K-Means telah digunakan dalam penelitian bidang serupa. Contohnya, Fadilah et al. menerapkan K-Means dan evaluasi elbow pada indikator kemiskinan distrik di Papua, menghasilkan dua klaster yang membedakan wilayah dengan kemiskinan tinggi dan rendah [1]. Selain itu, Nugraha et al. menggunakan K-Means untuk mengelompokkan provinsi di Indonesia berdasarkan angka harapan hidup dan berhasil menemukan tiga klaster dengan karakteristik tinggi, menengah dan, rendah [2].

Namun, penelitian-penelitian sebelumnya umumnya berfokus pada skala nasional atau menggunakan indikator yang lebih terbatas. Penelitian ini bertujuan untuk mengisi gap tersebut dengan mengelompokkan 38 kabupaten/kota di Jawa Timur menggunakan data tahun 2023, yang merupakan data terkini, berdasarkan empat indikator utama: angka harapan hidup, jumlah penduduk miskin, persentase penduduk miskin, dan garis kemiskinan. Metode yang dipilih adalah K-Means Clustering, dengan jumlah klaster ditentukan melalui metode elbow dan validasi menggunakan Silhouette Score. Untuk mempermudah visualisasi dan meningkatkan interpretabilitas, dilakukan reduksi dimensi menggunakan Principal Component Analysis (PCA). Hasil dari penelitian ini diharapkan memberikan segmentasi wilayah yang jelas berdasarkan kondisi kesejahteraan, sehingga dapat menjadi dasar dalam perumusan kebijakan daerah dan penyaluran program bantuan sosial yang lebih tepat sasaran. Secara konkret, hasil klasterisasi ini dapat digunakan untuk mengidentifikasi wilayah yang memerlukan prioritas dalam program peningkatan kesehatan, pemberdayaan ekonomi, atau intervensi sosial lainnya. Penelitian ini memiliki batasan, antara lain ketergantungan pada data yang tersedia dari BPS dan asumsi yang digunakan dalam metode K-Means dan PCA. Penelitian lebih lanjut dapat dilakukan dengan mempertimbangkan faktor-faktor lain yang mempengaruhi kesejahteraan atau menggunakan metode analisis yang lebih kompleks.

ristik tinggi, menengah dan, rendah [2].

Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi yang signifikan dalam upaya mengatasi ketimpangan sosial-ekonomi di Jawa Timur dan meningkatkan kesejahteraan masyarakat secara merata.

2. TINJAUAN PUSTAKA

Beberapa penelitian terdahulu telah menerapkan metode klasterisasi pada data sosial dan kesehatan di berbagai daerah Indonesia.

Dimas Reza Nugraha et al. (2024) menggunakan algoritma K-Means untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan angka harapan hidup. Mereka menemukan tiga klaster yaitu tinggi, sedang, dan rendah dengan Jawa Timur berada pada klaster harapan hidup tinggi. Namun, penelitian ini mencatat adanya dua provinsi (Jawa Timur dan Gorontalo) dengan nilai silhouette rendah, yang kemudian menyebabkan hasil klaster kurang optimal.

Penelitian lain oleh Sri Pingit Wulandari et al. (2025) berfokus pada analisis faktor kesehatan masyarakat pada skala provisi menggunakan Principal Component Analysis (PCA). Mereka mereduksi variabel menjadi dua faktor utama, sosio-ekonomi dan keamanan pangan yang mampu menjelaskan sekitar 76% variansi data.

Di tingkat internasional, sebuah studi yang dipublikasikan baru-baru ini menggunakan kombinasi PCA dan K-Means untuk mengelompokkan negara berdasarkan indikator sosial dan ekonomi, seperti pendapatan, angka kematian anak-anak, dan harapan hidup. Mereka memanfaatkan PCA untuk mereduksi dimensi, lalu K-Means untuk membentuk klaster wilayah yang dapat membantu lembaga kemanusiaan dalam mendistribusikan bantuan secara lebih tepat.

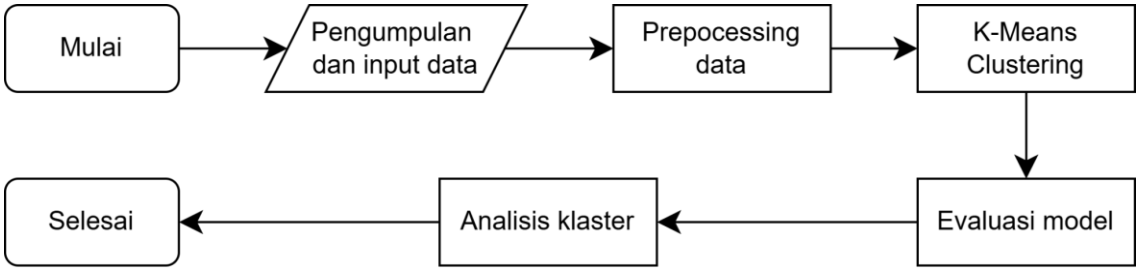
Selain itu, penelitian oleh Indah Fahmiyah et al. (2023) mengelompokkan provinsi di Indonesia berdasarkan indikator HDI (termasuk angka harapan hidup) menggunakan K-Means. Mereka mendapatkan empat klaster dengan metode Elbow dan indeks *Calinski-Harabasz*, serta menekankan pentingnya normalisasi data dalam proses klasterisasi.

Dari berbagai penelitian sebelumnya, terlihat bahwa metode K-Means dan PCA banyak digunakan dalam pengelompokkan wilayah berdasarkan indikator sosial. Nugraha et al. (2024) menggunakan K-Means untuk analisis harapan hidup di tingkat provinsi tanpa reduksi dimensi. Wulandari et al. (2025) menerapkan PCA untuk faktor kesehatan, namun tidak dikombinasikan dengan klasterisasi. Sementara itu, Fahmiyah et al. (2023) mengelompokkan wilayah berdasarkan HDI secara nasional, bukan fokus regional.

Kebaruan dari penelitian ini terletak pada penerapan gabungan PCA dan K-Means secara spesifik terhadap dua indikator sosial, yaitu angka harapan hidup dan persentase kemiskinan, di tingkat kabupaten/kota di Jawa Timur. Pendekatan ini memanfaatkan data terbaru dan menyederhanakan variabel untuk memperoleh klaster yang lebih jelas dan relevan secara kebijakan.

3. METODE PENELITIAN

Kerangka kerja sistematis yang dikenal sebagai metodologi penelitian digunakan untuk mengumpulkan, menganalisis, dan menginterpretasikan data dalam upaya menguji hipotesis atau menjawab pertanyaan penelitian. Metodologi ini menggabungkan berbagai langkah dan prosedur untuk memastikan bahwa penelitian dilakukan secara sistematis, sah, dan dapat diandalkan [5]. Ini termasuk memilih metode penelitian yang tepat, mengumpulkan data, melakukan analisis statistik, dan menginterpretasikan hasil yang diperoleh. Penelitian ini menggunakan metode kuantitatif untuk menganalisis data Angka Harapan Hidup (AHH) dan kemiskinan di Jawa Timur. Berikut adalah alur tahapan penelitian :



Gambar 1. Alur tahapan penelitian

3.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini bersumber dari Badan Pusat Statistik Provinsi Jawa Timur. Data yang dikumpulkan meliputi data Angka Harapan Hidup (AHH) per kabupaten/kota dan data kemiskinan (jumlah penduduk miskin, garis kemiskinan, dan persentase penduduk miskin) pada tahun 2023.

3.2. Data Pre-processing

Tahap preprocessing untuk memastikan bahwa data yang akan dianalisis bersih, terstandarisasi, dan siap digunakan. Langkah pertama dalam preprocessing adalah Data Cleaning. Pada langkah ini, dilakukan normalisasi pada kolom “Kabupaten/Kota” karena ketidaksesuaian format seperti perbedaan kapitalisasi huruf serta spasi tambahan yang tidak konsisten, yang berpotensi menimbulkan ketidakcocokan saat proses penggabungan data. Menyesuaikan nama-nama kolom yang tidak relevan dan menghapus semua baris data dengan nilai kosong (NaN) untuk memastikan bahwa data yang akan dianalisis lengkap dan tidak ada nilai yang hilang yang dapat mengganggu hasil analisis. Data bersih ini kemudian digunakan sebagai dasar untuk tahapan *preprocessing* berikutnya[8]. Langkah selanjutnya adalah memilih dan menyiapkan fitur (variabel) yang akan digunakan dalam proses klasterisasi. Empat variabel numerik digunakan sebagai fitur utama penelitian ini. Selanjutnya, data distandarisasi untuk memastikan bahwa semua fitur dataset memiliki skala yang sama. Ini dilakukan untuk memastikan bahwa fitur dengan skala yang berbeda dapat mempengaruhi hasil analisis clustering. Teknik StandardScaler dari pustaka Sklearn digunakan untuk standarisasi, yang mengubah setiap fitur menjadi memiliki mean 0 dan standar deviasi 1.

3.3. Metode K-Means

K-Means membagi data ke dalam kkk cluster yang telah ditentukan sebelumnya, yang menjadikannya salah satu algoritma *clustering* yang paling populer dan paling sederhana untuk digunakan[9]. Tujuan dari algoritma ini adalah untuk meminimalkan perbedaan antara setiap cluster[9]. Untuk melakukan ini, data dikelompokkan ke dalam cluster dengan centroid

terdekat. Metode K-Means digunakan untuk mengelompokkan kabupaten/kota di Jawa Timur berdasarkan data Angka Harapan Hidup (AHH) dan kemiskinan. Implementasi K-Means dilakukan dengan menguji beberapa jumlah klaster untuk menentukan klaster optimal.

3.4. Evaluasi dan Analisis Model

Pada penelitian ini, penelitian ini menerapkan dua metode evaluasi, yaitu *Elbow* dan *Silhouette Score*. Metode Elbow untuk penentuan klaster optimal dengan persentase perbandingan antara klaster sebelum dan sesudah yang mengalami penurunan terbesar pada grafik di suatu titik yang membentuk suatu siku. Dari hasil persentase yang dihitung dijadikan pembanding antara penjumlahan jumlah klaster. *Sum of Square Error* untuk evaluasi jumlah klaster dari hasil pengujian K-Means. SSE dihitung menggunakan rumus persamaan berikut :

$$\text{SSE} = \sum_{k=1}^{k} \sum_{xi \epsilon Sk} \left|\left| x_i - C_k \right|\right|_2^2$$

Dimana :
k = jumlah nilai kluster
xi = data ke-i
Ck = centroid pada kluster

Untuk memperkuat keputusan ini, metode Silhouette Score juga digunakan untuk mengevaluasi tingkat kemiripan data antara kluster mereka sendiri dan kluster lain. Nilai Silhouette antara -1 dan 1; nilai yang lebih tinggi menunjukkan kelompokkan data yang lebih baik [13]. Setelah didapat klaster terbaik menggunakan metode *Elbow*, kemudian menentukan nama kategori pada masing-masing label atau klaster. Dalam penentuan kategori diperlukan analisis perbandingan nilai rata-rata antar klaster. Hasil klasterisasi kemudian divisualisasikan menggunakan *Principal Component Analysis (PCA*) untuk mempermudah interpretasi hasil.

3.5. Analisis Klaster

Pada penelitian ini dilakukan analisis mendalam terhadap masing-masing klaster guna mengidentifikasi perbedaan karakteristik sosial ekonomi antar kelompok. Untuk mendukung interpretasi visual dari hasil klasterisasi, digunakan pendekatan Principal Component Analysis (PCA) sebagai metode reduksi dimensi. PCA diterapkan pada data yang telah distandarisasi guna mengekstrak dua komponen utama—disebut PCA1 dan PCA2—yang memuat proporsi terbesar dari variasi data asli. Kedua komponen ini kemudian digunakan sebagai sumbu dalam visualisasi dua dimensi yang mampu menyederhanakan representasi data tanpa kehilangan informasi penting.

Di samping visualisasi, analisis kuantitatif juga dilakukan untuk menggambarkan karakteristik internal masing-masing klaster. Analisis ini mencakup perhitungan nilai rata-rata (mean), median, dan simpangan baku (standar deviasi) dari seluruh fitur utama seperti Angka Harapan Hidup, garis kemiskinan, jumlah penduduk miskin, serta persentase kemiskinan. Tujuannya adalah untuk menemukan pola dominan dan membandingkan kondisi sosial ekonomi antar klaster. Selain itu, dilakukan juga konversi kembali nilai *centroid* dari hasil klasterisasi

ke skala semula melalui proses *inverse transform* terhadap parameter standardisasi awal. Proses ini memungkinkan hasil akhir interpretasi dapat disajikan dalam satuan yang dapat dipahami secara substantif, sehingga mendukung penarikan kesimpulan terhadap kondisi khas dari masing-masing kelompok wilayah yang terbentuk dalam klasterisasi.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini bersumber dari Badan Pusat Statistik (BPS) Provinsi Jawa Timur, dengan tahun rujukan data yaitu tahun 2023. Variabel yang digunakan mencakup empat indikator utama yang merepresentasikan kesejahteraan sosial masyarakat pada tingkat kabupaten/kota, yaitu: angka harapan hidup saat lahir (tahun),jumlah penduduk miskin (dalam ribuan jiwa), persentase penduduk miskin terhadap jumlah penduduk, dan garis kemiskinan (dalam rupiah per kapita per bulan). Data ini mencakup seluruh 38 kabupaten/kota di Jawa Timur yang menjadi unit observasi dalam penelitian. Tujuan pemilihan variabel ini adalah untuk menangkap kompleksitas dimensi kesehatan dan kemiskinan secara simultan guna dianalisis melalui pendekatan unsupervised learning.

Seluruh data diunduh melalui situs resmi BPS (https://jatim.bps.go.id/), di mana setiap variabel tersedia dalam bentuk tabel terpisah berdasarkan topik statistik. Oleh karena itu, proses awal pengumpulan data dilakukan dengan menggabungkan keempat tabel indikator ke dalam satu set data utama berdasarkan nama kabupaten/kota. Kemudian, dilakukan pembersihan data (*data cleaning*) untuk menghindari adanya duplikasi, *missing values,* maupun kesalahan entri.

4.2. Hasil Klasterisasi *Clustering K-Means*

Penelitian ini menggunakan bahasa pemrograman python dalam mengimplementasikan metode K-Means *Clustering* guna mengelompokkan data kesejahteraan wilayah kabupaten/kota di Provinsi Jawa Timur. Proses klasterisasi dilakukan berdasarkan sejumlah variabel sosial ekonomi seperti angka harapan hidup, jumlah penduduk miskin, dan persentase kemiskinan.

Implementasikan pengelompokan data menggunakan persamaan jarak euclidean dengan uji cluster uji k=1 hingga k=10. Berikut merupakan hasil klaster pada uji k=2

Table 1. Hasil Klasterisasi tahun 2023

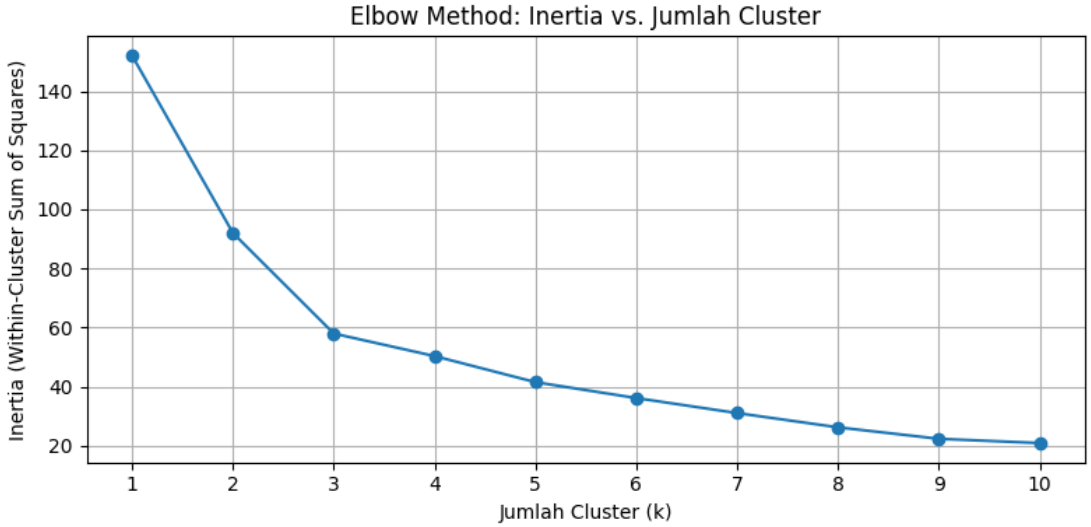
Kabupaten/Kota	2023
Kabupaten Bangkalan	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Banyuwangi	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Blitar	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Bojonegoro	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Bondowoso	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Gresik	Klaster 0 (Harapan Hidup Tinggi

	dan Kemiskinan Rendah)
Kabupaten Jember	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Jombang	Klaster 0 (Harapan Hidup Tinggi dan Kemiskinan Rendah)
Kabupaten Kediri	Klaster 0 (Harapan Hidup Tinggi dan Kemiskinan Rendah)
Kabupaten Lamongan	Klaster 0 (Harapan Hidup Tinggi dan Kemiskinan Rendah)
Kabupaten Lumajang	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Madiun	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Magetan	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Malang	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Mojokerto	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Nganjuk	Klaster 0 (Harapan Hidup Tinggi dan Kemiskinan Rendah)
Kabupaten Ngawi	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Pacitan	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Pamekasan	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Pasuruan	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Ponorogo	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Probolinggo	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Sampang	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Sidoarjo	Klaster 0 (Harapan Hidup Tinggi dan Kemiskinan Rendah)
Kabupaten Situbondo	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Sumenep	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Trenggalek	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)

Kabupaten Tuban	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kabupaten Tulungagung	Klaster 0 (Harapan Hidup Tinggi dan Kemiskinan Rendah)
Kota Batu	Klaster 0 (Harapan Hidup Tinggi dan Kemiskinan Rendah)
Kota Blitar	Klaster 0 (Harapan Hidup Tinggi dan Kemiskinan Rendah)
Kota Kediri	Klaster 0 (Harapan Hidup Tinggi dan Kemiskinan Rendah)
Kota Madiun	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kota Malang	Klaster 0 (Harapan Hidup Tinggi dan Kemiskinan Rendah)
Kota Mojokerto	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kota Pasuruan	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kota Probolinggo	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)
Kota Surabaya	Klaster 1 (Kemiskinan Lebih Tinggi dan Harapan Hidup Lebih Rendah)

4.3. Evaluasi dan Analisis Hasil

Dalam proses klasterisasi, penelitian ini menerapkan dua metode untuk menentukan jumlah klaster yang paling tepat, yakni metode *Elbow* dan *Silhouette Score*, dengan tujuan memperoleh hasil yang optimal baik dari segi efisiensi pembentukan klaster maupun kualitas pemisahan antar kelompok data. Metode Elbow digunakan dengan cara membandingkan nilai *Sum of Squared Error* (SSE) pada beberapa variasi jumlah klaster, yaitu dari k=2 hingga k=6.



Gambar 2. Grafik Elbow pada 38 data

Berikut merupakan hasil pengujian metode elbow pada 38 data. Hasil penggambaran grafik menunjukkan bahwa nilai SSE menurun seiring bertambahnya jumlah klaster. Namun, setelah titik tertentu, penurunannya menjadi tidak signifikan dan grafik membentuk sudut tajam menyerupai "siku" (*elbow*). Titik siku inilah yang mengindikasikan nilai kk terbaik, karena menandakan titik efisiensi di mana penambahan jumlah klaster tidak lagi memberikan pengurangan error yang berarti. Dalam kasus ini, titik optimal tersebut terjadi pada k=2, sehingga dipilih sebagai jumlah klaster yang ideal berdasarkan pendekatan *Elbow*.

Berikut pada Tabel 2. merupakan hasil pengujian perhitungan nilai SSE dan selisih nilai SSE antar klaster sebelum dan sesudah pada 38 data.

Tabel 2. Hasil Perhitungan Metode Elbow

Kluster	Nilai SEE	Selisih Nilai SEE
K2	87.60	30.29
K3	57.31	8.55
K4	48.76	7.97
K5	40.79	5.60
K6	35.19	4.71
K7	30.48	4.22
K8	26.27	3.98
K9	22.29	2.15
K10	20.14	-

Pada Tabel 2 ditampilkan hasil uji coba metode *Elbow* terhadap 38 data, yang menunjukkan nilai SSE untuk setiap jumlah klaster dari k = 2 sebesar 87.60, k = 3 sebesar 57,31, k = 4 sebesar 48,76, k = 5 sebesar 40,79, dan k = 6 sebesar 35,19. Sementara itu, nilai SSE pada k = 7 sebesar 30,48, k = 8 sebesar 26,27, k = 9 sebesar 22,29, dan k = 10 sebesar 20,14.

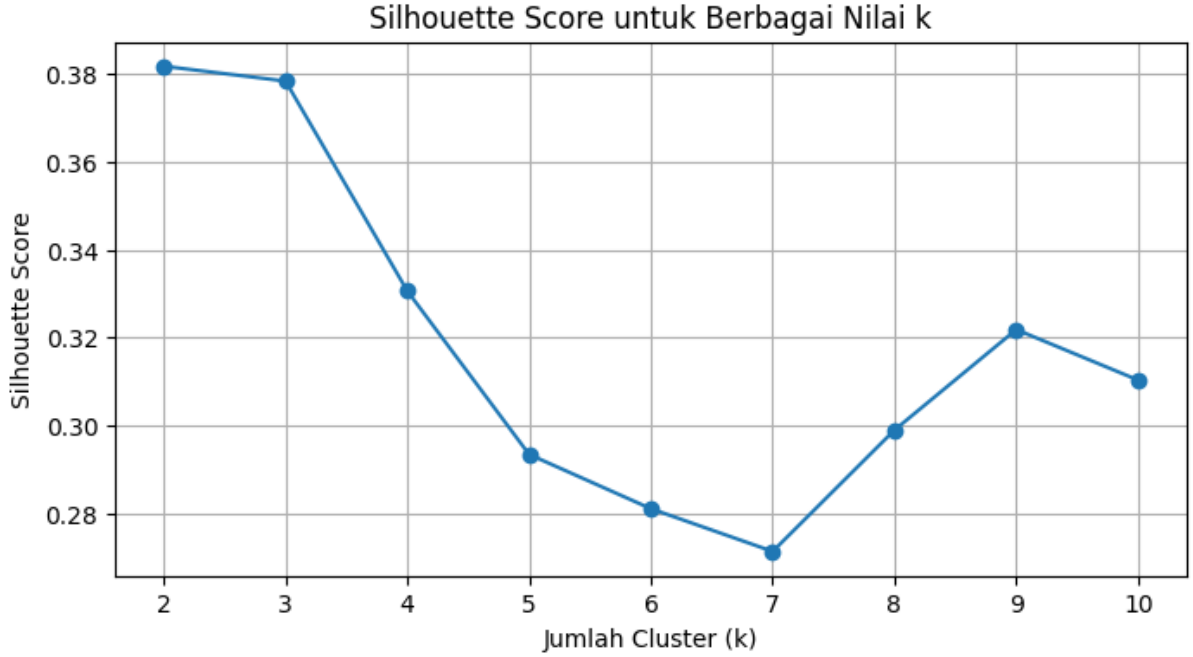
Dari hasil perhitungan selisih nilai SSE antar klaster, penurunan nilai terbesar terjadi dari k = 2 ke k = 3, yaitu sebesar 30,29. Setelah titik tersebut, penurunan nilai SSE relatif kecil dan cenderung melandai. Oleh karena itu, jumlah klaster yang dianggap optimal berdasarkan metode *Elbow* adalah 2 klaster, karena pada titik ini terjadi penurunan SSE paling signifikan dibandingkan nilai k selanjutnya. Berikut merupakan hasil jumlah data masing-masing klaster:

Tabel 3. Jumlah Data Kabupaten/Kota Setiap Kluster

Kluster	Total Data
0	11

1	27

Sebagai bentuk validasi tambahan, digunakan pula metode *Silhouette Score*, yang menilai kualitas pembentukan klaster dengan mempertimbangkan kedekatan suatu objek dengan anggota klasternya sendiri dibandingkan dengan objek dari klaster lain. Nilai Silhouette Score berada dalam rentang -1 hingga 1, dengan nilai yang lebih tinggi menunjukkan pemisahan klaster yang semakin jelas. Perhitungan dilakukan pada jumlah klaster dari k=2 hingga k=10. Hasil evaluasi menunjukkan bahwa nilai *Silhouette Score* tertinggi juga diperoleh pada k=2, yang menunjukkan bahwa pembagian klaster pada jumlah tersebut menghasilkan struktur klaster yang paling optimal dan paling terpisah secara internal antar kelompok data. Berikut merupakan grafik metode *Silhouette Score* pada 38 data sebagai berikut:



Gambar 3. Grafik *Silhouette Score* pada 38 data

Dengan mempertimbangkan hasil kedua metode di atas, dapat disimpulkan bahwa jumlah klaster optimal untuk penelitian ini adalah dua klaster. Keputusan ini didasarkan pada pertimbangan kuantitatif berupa penurunan nilai SSE yang signifikan hingga titik k=2 (0.38), serta pertimbangan kualitatif berupa nilai *Silhouette Score* yang paling tinggi pada jumlah klaster yang sama. Oleh karena itu, pembagian data ke dalam dua klaster dipandang paling representatif dalam menggambarkan karakteristik kelompok yang berbeda secara konsisten dan akurat.

4.4. Hasil Principal Component Analysis (PCA)

Proses pengurangan dimensi juga dikenal sebagai pengurangan dimensi—dilakukan melalui metode *Principal Component Analysis* (PCA). Metode ini memungkinkan representasi data berdimensi tinggi yang mencakup empat fitur numerik utama ke dalam dua dimensi utama, yang paling banyak mengandung informasi tentang variasi dalam data awal. Dua komponen utama disebut PCA1 dan PCA2 dihasilkan dari transformasi PCA. Keduanya digunakan sebagai sumbu visualisasi dua dimensi. Hasil nilai *explained variance ratio* pada tabel 4 sebagai berikut:

Tabel 4. Jumlah Data Kabupaten/Kota Setiap Kluster

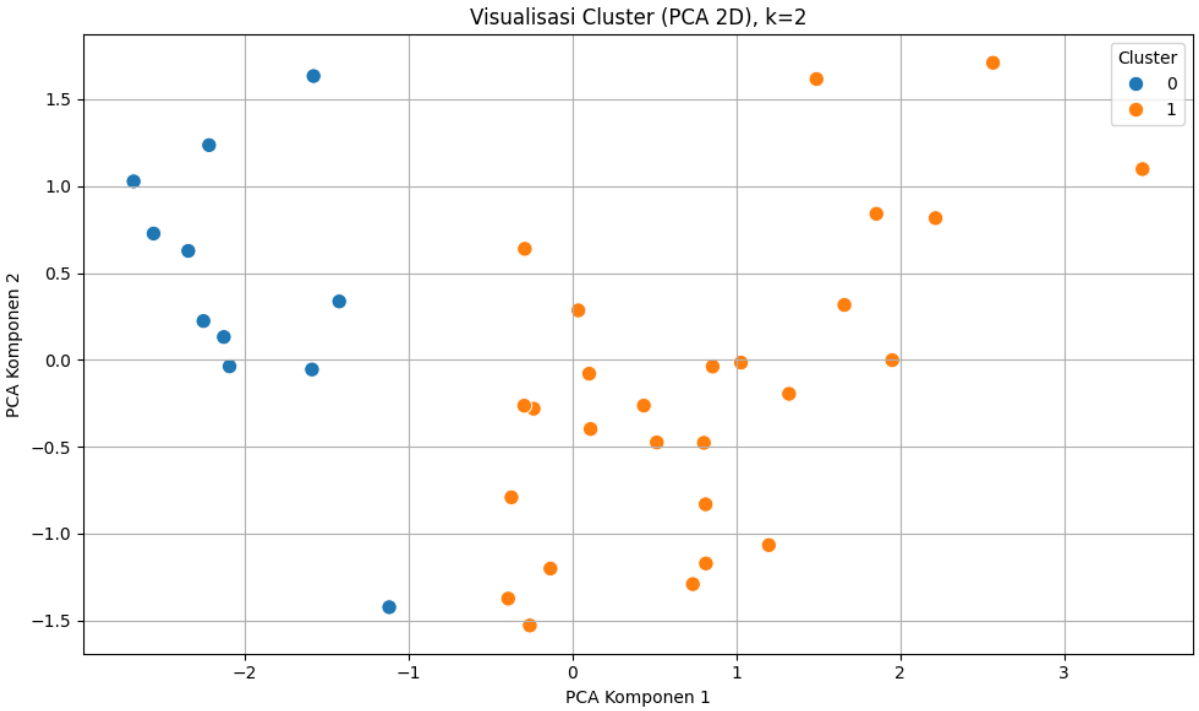
Komponen Utama	Variansi yang Dijelaskan (%)	Variansi Kumulatif (%)
PCA1	59,73%	59,73%
PCA2	19,14%	78,88%

Tabel tersebut menunjukkan bahwa dua komponen utama hasil PCA, yaitu PCA1 dan PCA2, secara kumulatif mampu menjelaskan 78,88% dari total variasi yang terdapat dalam data asli. Komponen pertama (PCA1) berkontribusi paling besar dengan menjelaskan sekitar 59,73%, sementara PCA2 menambahkan 19,14% informasi tambahan. Berdasarkan nilai tersebut, penggunaan dua dimensi ini dianggap cukup representatif dalam menggambarkan struktur data dan pemisahan antar klaster secara visual.

Tabel 5. Rata-rata Nilai Fitur pada Seluruh Klaster

Klaster	Angka Harapan Hidup (Tahun)	Garis Kemiskinan	Jumlah Penduduk Miskin (Ribu)	Persentase Penduduk Miskin
PCA1	78.81	591.147	40.58	5.62
PCA2	71.85	445.807	136.61	12.20

Berdasarkan Tabel 1, Klaster 0 memiliki rata-rata Angka Harapan Hidup lebih tinggi dan tingkat kemiskinan lebih rendah dibandingkan Klaster 1. Hal ini memperjelas pemisahan karakteristik sosial ekonomi antara dua klaster.



Gambar 4. Hasil Visualisasi PCA

Berdasarkan evaluasi menggunakan *Silhouette Score*, jumlah klaster optimal adalah 2, dengan nilai tertinggi dicapai saat k=2. Hasil ini diperkuat oleh visualisasi menggunakan *Principal Component Analysis* (PCA), yang menunjukkan bahwa klaster 0 memiliki penyebaran yang lebih kompak dan homogen, sementara klaster 1 tampak lebih menyebar, mencerminkan adanya keragaman karakteristik sosial ekonomi dalam klaster tersebut. Meskipun klaster 1 lebih bervariasi, kedua klaster tetap terlihat terpisah secara jelas, menandakan bahwa proses klasterisasi telah berhasil memetakan struktur data dengan cukup baik dan representatif terhadap variasi yang ada dalam data asli.

4.5. Hasil Analisis Statistik Klaster

Selanjutnya dilakukan analisis terhadap kontribusi (*loadings*) setiap variabel terhadap dua komponen utama, yaitu PC1 dan PC2. Nilai loadings menggambarkan seberapa besar pengaruh masing-masing fitur dalam menentukan arah dan struktur komponen utama. Fitur dengan nilai absolut yang lebih tinggi dianggap lebih dominan dalam membentuk dimensi tersebut. Tabel 2 berikut menyajikan kontribusi masing-masing fitur terhadap dua komponen utama PCA yang digunakan dalam visualisasi klaster:

Tabel 5. Rata-rata Nilai Fitur pada Seluruh Klaster

Klaster	PCA1	PCA2
Angka Harapan Hidup (Tahun)	-0,464	-0,611
Garis Kemiskinan	-0,419	0,788
Jumlah Penduduk Miskin (Ribu)	0,522	0,002
Persentase Penduduk Miskin	0,581	0,079

Tabel 5 menunjukkan kontribusi fitur terhadap dua komponen utama hasil PCA. Fitur yang paling dominan terhadap PC1 adalah *Persentase Penduduk Miskin* (0,581), sedangkan PC2 didominasi oleh *Garis Kemiskinan* (0,788). Nilai ini menunjukkan bahwa dimensi pertama (PC1) cenderung mencerminkan tingkat kemiskinan secara langsung, sedangkan PC2 lebih mengarah pada perbedaan daya beli masyarakat.

KESIMPULAN

Berdasarkan hasil analisis klasterisasi menggunakan metode K-Means pada data Angka Harapan Hidup (AHH) dan kemiskinan di Jawa Timur, terdapat dua klaster yang optimal berdasarkan *Elbow*. Klaster 0 memiliki karakteristik AHH yang lebih tinggi dan tingkat

kemiskinan yang lebih rendah, menunjukkan bahwa daerah-daerah dalam klaster ini memiliki kondisi kesehatan dan ekonomi yang lebih baik. Sementara itu, klaster 1 memiliki karakteristik AHH yang lebih tinggi dan tingkat kemiskinan yang lebih rendah

Hasil klasterisasi kemudian dianalisis secara deskriptif dan visual. Klaster pertama (Klaster 0) mewakili daerah dengan angka harapan hidup yang tinggi dan kemiskinan yang rendah, sedangkan klaster kedua (Klaster 1) mewakili daerah dengan angka harapan hidup yang lebih rendah dan kemiskinan yang lebih tinggi. Dua komponen utama dapat menjelaskan 78,88% variasi dalam data, dengan pemisahan klaster yang cukup jelas dalam representasi dua dimensi, menurut analisis dimensi menggunakan *Principal Component Analysis* (PCA). Selain itu, kontribusi fitur PCA menunjukkan bahwa dimensi pertama terkait erat dengan indikator kemiskinan, sedangkan dimensi kedua lebih banyak menggambarkan variasi garis kemiskinan dan harapan hidup.

Keseluruhan hasil ini menunjukkan bahwa pendekatan klasterisasi yang digunakan mampu memberikan pemetaan sosial ekonomi yang menarik dan dapat digunakan sebagai referensi awal untuk pengambilan kebijakan pembangunan daerah yang berbasis data. Hasil visualisasi menggunakan *Analysis of Principal Components* (PCA) menunjukkan bahwa komponen utama yang menjelaskan variasi data dapat membedakan kedua klaster. Hal ini menunjukkan bahwa pola-pola dapat ditemukan dalam data AHH dan kemiskinan di Jawa Timur dengan menggunakan teknik K-Means dan PCA. Penelitian ini dapat membantu pemerintah dan pihak-pihak yang terlibat dalam menentukan lokasi yang membutuhkan perhatian khusus untuk meningkatkan AHH dan mengurangi kemiskinan, menyusun strategi pembangunan yang lebih tepat sasaran, mengurangi kesenjangan antarwilayah, dan meningkatkan kualitas hidup masyarakat secara merata.

DAFTAR PUSTAKA

[1] Y. C. Fadilah, A. Sani, and A. Andrianingsih, “Applying K-Means Clustering for Grouping Papua’s Districts Based on Poverty Indicators Analysis,” *J. Ilm. Pemgetah. Teknol. Komput.*, vol.10, no. 3, 2023, doi: 10.33480/jitk.v10i3.5865.

[2] D. R. Nugraha, A. T. Zy, and A. Supriyadi, “The Use of K-Means Algorithm Clustering in Grouping Life Expectancy (Case Study: Provinces in Indonesia),” *J. Comput. Netw. Archit. High Perform. Comput.*, vo 6, no. 3, 2021, doi: 10.47709/cnahpc.v6i3.4171.

[3] Sri Wahyuni, Agustia Hananto, Baenil Huda, Tukino Tukino, “Identifying Regional Patterns of Poverty in Indonesia: a Clustering Approach Using K-Means,” *Int.J. Comput. Inform. Sci*, vol. -, no.-, 2025. Studi ini mengelompokkan 38 provinsi di Indonesia berdasarkan tingkat kemiskinan menggunakan K-Means (silhouette score0.613).

[4] Gustriza Erda, Chairani Gunawan, Zulya Erda, “Grouping of Poverty in Indonesia Using K-Means with Silhouette Coefficient,” *Parameter: Journal of Statistics*, vol. 3, no.1, 2023, doi:10.22487/27765660.2023.v3.i1.16435

[5] Resti Wahyuni, "K-Means Clustering for Grouping Indonesia Underdeveloped Regions in 2020 Based on Poverty Indicators," *Parameter: Journal of Statistics*, vol.2, no.1, 2021, doi:10.22487/27765660.2021.vs.i1.15675

[6] A. Bahauddin, A. Fatmawati, FP. Sari, "Analisis Clustering Provinsi Di Indonesia Berdasarkan Tingkat Kemiskinan Menggunakan Algoritma K-Means," *Jurnal Manajemen Informatika dan Sistem Informasi*, vol.4, no.1, pp. 1-8, 2021

[7] Albert V. Dian Sano, Hendro Nindito, "Application of K-Means Algorithm for Cluster Analysis on Poverty of Provinces in Indonesia," ComTech (BINUS Univ.), 2014

[8] K. Setiawan, Kastum, Pratama Y.P., "K-Means Clustering Analysis of Poverty Data in Cilacap District," *Int'l J. Software Eng. & Computer Sci. (IJSECS*), vol.5, no.1, 2025

[9] J. Ramadhani, Y.S. Anugraha, A. Fauzan, R. Rahmaddeni, L.Efrizoni, "Perbandingan Algoritma K-Means Clustering dan K-Medoids dalam Mengelompokkan Tingkat Kemiskinan di Provinsi Riau," JSR: Jaringan Sistem Informasi Robotik, vol.8, no.1, 2024

[10] K.H. Izzuddin, A.W.Wijayanto, "Pemodelan Clustering Ward K-Means, DIANA, dan PAM dengan PCA untuk Karakterisasi Kemiskinan Indonesia Tahun 2021," *Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi,* vol.4, no.1, 2025

[11] "Implementation of K-Means dan RapidMiner untuk Clustering Data Kemiskinan Provinsi Banten," oleh Deny Haryadi dkk., *Jurnal Informatika & Komunikasi (JICT),* vol.6, no.1, 2024, DOI:10.33480/jitk.v9i2.4381

[12] Albert Alamsyah, T.T.Gustyana, A.D.Fajaryanto, D.Septiafani, "Open Data Analytical Model for Human Development Index Optimization to Support Government Policy," arXiv, 2018 (menggunakan ANN dan K-Means untuk HDI di Indonesia)

[13] M. Emre Celebi, Hassan A. Kingravi, Patricio A. Vela, "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm," arXiv, 2012

[14] Mehrdad Ghadiri, Samira Samadi, Santosh Vempala, "Socially Fair k-Means Clustering," arXiv, 2020

[15] "Simulation of the K-Means Clustering Algorithm With The Elbow Method in Making Clusters Of Provincial Poverty Levels in Indonesia" (Jurnal Mantik, UNISA), menentukan 5 cluster kemiskinan provinsi Indonesia

[16] A. Novita, I. Ernawati, dan N. Chamidah, "Klasterisasi Provinsi di Indonesia Berdasarkan Produktivitas Komoditas Pangan Menggunakan Algoritma K-Means," Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA), Jakarta, Indonesia, 20 Agustus 2022.