ESTIMASI: Journal of Statistics and Its Application
Vol. 7, No. 1, Januari, 2026, Hal. 81-96
DOI: 10.20956/ejsa.v7i1.44767

Segmentasi Wilayah Jawa Timur Berdasarkan Ketersediaan Fasilitas dan Tenaga Kesehatan

Muhammad Erlangga Kurniawan[1], Aditya Putra Ananta[2], Muhammad Cahya Raka Anugrah[3], Aviolla Terza Damaliana[4], Shindi Sheila May Wara[5*]
[12345]Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jawa Timur Surabaya
*** Corresponding author, email: shindi.shella.fasilkom@student.upnjatim.ac.id**

Abstract

Health is a fundamental indicator in measuring societal well-being, where the equitable distribution of healthcare facilities and personnel plays a critical role. This study aims to segment regions in East Java Province based on the availability of healthcare facilities (community health centers, general/special hospitals, pharmacies, integrated health posts, and primary clinics) and healthcare personnel (doctors, midwives, nurses, pharmacists). The methods used include Principal Component Analysis (PCA) for dimensionality reduction, followed by K-Means and Agglomerative Hierarchical Clustering (AHC) algorithms using Average Linkage and Cosine Similarity. The analysis results show that AHC provides more optimal outcomes, with a silhouette score of 0.75, compared to K-Means which only achieved 0.51. The segmentation produced three main clusters: low (Pacitan, Ponorogo, Madura), medium (Bojonegoro, Jember, Banyuwangi), and high (Surabaya, Malang, Sidoarjo). These findings reveal disparities in the distribution of healthcare services in East Java and can serve as a foundation for more targeted policy formulation to improve equitable access to healthcare, particularly in underserved regions.

Keywords*: Health facilities, health workers, Clustering, PCA, AHC.*

Abstrak

Kesehatan merupakan indikator fundamental dalam mengukur kesejahteraan masyarakat, di mana pemerataan fasilitas dan tenaga kesehatan menjadi faktor penting. Penelitian ini bertujuan untuk melakukan segmentasi wilayah di Provinsi Jawa Timur berdasarkan ketersediaan fasilitas kesehatan (puskesmas, rumah sakit umum/khusus, apotek, posyandu, dan klinik pratama) serta tenaga kesehatan (dokter, bidan, perawat, apoteker). Metode yang digunakan meliputi Principal Component Analysis (PCA) untuk reduksi dimensi, dilanjutkan dengan algoritma K-Means dan Agglomerative Hierarchical Clustering (AHC) menggunakan Average Linkage dan Cosine Similarity. Hasil analisis menunjukkan bahwa AHC memberikan hasil yang lebih optimal dengan nilai silhouette score sebesar 0,75 dibandingkan K-Means yang hanya mencapai 0,51. Segmentasi menghasilkan tiga klaster utama, yaitu klaster rendah (Pacitan, Ponorogo, Madura), sedang (Bojonegoro, Jember, Banyuwangi), dan tinggi (Surabaya, Malang, Sidoarjo). Temuan ini mengungkap adanya ketimpangan distribusi layanan kesehatan di Jawa Timur dan dapat menjadi dasar perumusan kebijakan yang lebih terarah untuk meningkatkan pemerataan akses layanan kesehatan, khususnya di wilayah dengan ketersediaan fasilitas yang masih terbatas.

Kata Kunci: Fasilitas kesehatan, tenaga kesehatan, klasterisasi, PCA, AHC.

1. Pendahuluan

Kesehatan merupakan salah satu indikator utama dalam mengukur tingkat kesejahteraan masyarakat. Status kesehatan masyarakat di suatu negara berperan penting dalam meningkatkan kualitas Sumber Daya Manusia, sementara kualitas Sumber Daya Manusia dapat menjadi aspek dalam mengukur kemajuan dari negara tersebut [1]. Pemerataan fasilitas kesehatan menjadi prioritas penting dalam menjaga kesejahteraan masyarakat. Menurut Gadisty dan Susilawati, berbagai macam upaya dan usaha yang diperlukan agar tubuh masyarakat selalu dalam keadaan sehat [2]. Penelitian lain yang dilakukan oleh Anastasia dan Kemal mengatakan bahwa fasilitas kesehatan merupakan hak utama masyarakat yang harus dipenuhi dalam pembangunan kesehatan sehingga terjadi pemerataan terhadap suatu indeks [3]. Oleh karena itu, mudahnya akses fasilitas dan tenaga kesehatan sangat diperlukan di kehidupan masyarakat, khususnya di wilayah dengan keragaman geografis dan demografis yang tinggi seperti Provinsi Jawa Timur. Perbedaan jumlah fasilitas kesehatan antar kabupaten/kota dapat mencerminkan ketimpangan dalam pelayanan kesehatan yang berpengaruh langsung terhadap kualitas hidup masyarakat.

Provinsi Jawa Timur sebagai salah satu provinsi terbesar di Indonesia memiliki tantangan tersendiri dalam menjalankan pemerataan layanan kesehatan. Tidak meratanya fasilitas layanan kesehatan merupakan permasalahan penting yang harus mendapatkan perhatian dari pemerintah setempat. Menurut penelitian dari Henny, masih banyak rumah sakit dan puskesmas yang mengalami kekurangan tenaga kesehatan, terutama di puskesmas [4]. Kurangnya tenaga kesehatan ini tentunya akan berpengaruh terhadap pelayanan yang didapatkan oleh masyarakat yang membutuhkan. Penelitian lain yang dilakukan oleh Rafli, menunjukkan bahwa distribusi dokter di Jawa Timur sangat bervariasi antarwilayah, sedangkan jumlah dokter cenderung lebih banyak tersebar di daerah dengan jumlah penduduk dan fasilitas kesehatan yang tinggi. Ketimpangan ini menyebabkan buruknya akses masyarakat terhadap pelayanan kesehatan yang layak, khususnya di wilayah dengan aksesibilitas rendah dan jumlah fasilitas yang cukup terbatas. Oleh karena itu, untuk mengetahui karakteristik dari persebaran ketersediaan fasilitas dan tenaga kesehatan di Provinsi Jawa Timur, perlu dilakukannya analisis dengan metode *clustering*. Metode *clustering* akan digunakan untuk mengidentifikasi pola distribusi untuk membentuk segmentasi wilayah dengan karakteristik yang serupa.

Pada penelitian ini, kami menggunakan beberapa variabel yang merepresentasikan ketersediaan fasilitas dan tenaga kesehatan di setiap kabupaten/kota di Provinsi Jawa Timur. Variabel fasilitas kesehatan terdiri dari jumlah puskesmas, posyandu, rumah sakit umum, rumah sakit khusus, apotek, dan klinik pratama. Sementara itu, variabel tenaga kesehatan mencakup jumlah dokter, bidan, perawat, dan apoteker. Variabel-variabel tersebut digunakan karena mencerminkan kapasitas dasar suatu wilayah dalam

menyediakan pelayanan kesehatan yang menjangkau masyarakat secara langsung. Kondisi tenaga kesehatan sering kali tidak seimbang (inequality) dengan jumlah fasilitas kesehatan terutama di Jawa Timur. Sebagai contoh, tenaga kesehatan di fasilitas kesehatan seperti puskesmas, berdasarkan rasio jumlah tenaga kesehatan puskesmas di Jawa Timur belum memenuhi kebutuhan penduduk di provinsi tersebut. Seluruh variabel bersifat kuantitatif dan dianalisis menggunakan pendekatan multivariat untuk mengidentifikasi pola segmentasi wilayah berdasarkan layanan kesehatannya. Penelitian ini menggunakan dua pendekatan metode *clustering*, yaitu *K-Means* dan *Hierarchical Clustering* dengan *Average Linkage* berbasis *Cosine Similarity*. Pemilihan kedua metode ini bertujuan untuk membandingkan hasil segmentasi wilayah untuk memperoleh gambaran yang lebih komprehensif terhadap pola distribusi layanan kesehatan di Provinsi Jawa Timur.

Beberapa penelitian terdahulu telah mengkaji distribusi layanan kesehatan menggunakan berbagai pendekatan kuantitatif. Seperti pada penelitian yang dilakukan oleh Wibowo dan Mulyastuti menerapkan metode *K-Means* untuk mengelompokkan wilayah administratif di Provinsi DKI Jakarta berdasarkan jumlah fasilitas kesehatan, seperti rumah sakit dan puskesmas. Hasil *clustering* menghasilkan sejumlah tiga klaster yang digunakan untuk mengidentifikasi daerah dengan kebutuhan pembangunan fasilitas kesehatan yang lebih besar [5]. Pada penelitian lain, Hamami dan Dahlan melakukan segmentasi fasilitas dan tenaga kesehatan di Kota Bandung menggunakan algoritma *K-Means*. Hasil penelitian menunjukkan bahwa distribusi fasilitas dan tenaga kesehatan tidak merata, dengan sebagian besar kecamatan termasuk dalam kategori mengalami kekurangan Sumber Daya Manusia (SDM) [6]. Sementara itu, Yolanda dan Kristiana juga menerapkan metode *clustering* untuk menganalisis persebaran fasilitas kesehatan di Provinsi Jawa Barat. Penelitian tersebut menghasilkan tiga klaster yang menggambarkan wilayah dengan ketersediaan fasilitas (tinggi, sedang, dan rendah). Dukungan terhadap pemilihan metode ini juga ditemukan dalam penelitian Yusniyanti et al., yang membandingkan *Average Linkage* dan *K-Means* untuk segmentasi provinsi di Indonesia berdasarkan indikator kesejahteraan. Hasil penelitian tersebut menyimpulkan bahwa metode *Average Linkage* menghasilkan *silhouette score* yang lebih tinggi, dan berdasarkan rasio varians dibandingkan metode *K-Means* [7].

Meskipun berbagai hasil penelitian sebelumnya telah berhasil mengidentifikasi ketimpangan layanan kesehatan menggunakan metode *clustering*, beberapa penelitian masih terbatas pada satu wilayah kota atau hanya menggunakan satu jenis metode. Penelitian ini dilakukan pada cakupan wilayah Provinsi Jawa Timur dengan mempertimbangkan data fasilitas dan tenaga kesehatan secara bersamaan. Untuk memperoleh hasil segmentasi yang lebih komprehensif, kami akan menggunakan dua metode, yaitu *K-Means* dan *Hierarchical Clustering* dengan pendekatan *Average Linkage*. Hasil segmentasi yang diperoleh diharapkan dapat memberikan gambaran

yang lebih menyeluruh mengenai pola distribusi layanan kesehatan antar kabupaten/kota serta menjadi dasar pertimbangan dalam upaya perumusan kebijakan pemerataan pelayanan kesehatan di wilayah tersebut..

2. Material dan Metode

Kesehatan Provinsi Jawa Timur yang berisi data fasilitas dan tenaga kesehatan tiap kabupaten/kota. Data tersebut kemudian distandarisasi memakai teknik *Z-Score* agar semua variabel memiliki skala yang sebanding. Selanjutnya, dilakukan penyederhanaan dimensi data menggunakan *Principal Component Analysis* (PCA) untuk mengurangi kerumitan data tanpa menghilangkan informasi penting. Output dari PCA kemudian dipakai dalam pengelompokan wilayah dengan algoritma K-Means serta *Hierarchical* berdasarkan kesamaan karakteristik. Pemilihan jumlah klaster optimal dilakukan dengan teknik *Elbow Method*, *Silhouette Score,* dan *Silhouette Plot*. Tahap terakhir, hasil pengelompokan ditampilkan secara visual melalui peta distribusi wilayah agar lebih mudah dipahami dan dianalisis lebih lanjut.

2.1 Standardisasi

Langkah pertama dalam analisis adalah melakukan standardisasi pada semua variabel numerik terkait fasilitas dan tenaga kesehatan. Proses standardisasi ini penting untuk memastikan setiap variabel memberikan kontribusi yang setara dalam analisis, karena skala pengukuran tiap variabel mungkin berbeda. *Z-Score* sering kali dipakai sebagai metode standardisasi sebelum melakukan proses *clustering*, seperti penelitian yang dilakukan oleh Afifah dan Arie [8]. Hasilnya, setiap variabel akan berdistribusi dengan rata-rata nol dan standar deviasi satu, sehingga lebih siap untuk tahap analisis berikutnya seperti PCA dan klasterisasi. Nilai standar *Z-Score* dihitung dengan mengurangkan nilai aktual X_i dari rata-ratanya μ, lalu dibagi dengan standar deviasi σ, seperti persamaan berikut.

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{1}$$

2.2 Uji Asumsi

Sebelum melakukan reduksi dimensi melalui *Principal Component Analysis* (PCA), data dievaluasi awal melalui berbagai uji asumsi statistik untuk memastikan bahwa data tersebut memenuhi prasyarat untuk analisis dengan pendekatan PCA [9]. Penilaian awal ini bertujuan untuk menentukan sejauh mana variabel-variabel kumpulan data menunjukkan korelasi yang cukup, yang memungkinkannya untuk diringkas menjadi beberapa komponen utama tanpa mengorbankan informasi yang signifikan.

Evaluasi awal melibatkan pengukuran Kaiser-Meyer-Olkin (KMO), yang menilai kecukupan sampel berdasarkan kekuatan hubungan parsial antara variabel. Nilai KMO

berkisar dari nol hingga satu, di mana nilai ≥ 0.5 menandakan bahwa data tersebut sesuai untuk analisis PCA. Semakin dekat nilainya dengan satu, semakin kuat struktur korelasi di antara variabel-variabel, sehingga meningkatkan kemungkinan reduksi dimensi yang berhasil dan efisien. Rumus KMO tertulis pada persamaan (2), r_{ij} menggambarkan korelasi sederhana antara variabel ke-i dan ke-j, sedangkan q_{ij} mewakili korelasi parsial antara kedua variabel tersebut. Korelasi parsial diperoleh dengan mengendalikan pengaruh variabel lain dalam dataset, sehingga mencerminkan hubungan langsung antara dua variabel tersebut. Baik r_{ij}^2 maupun q_{ij}^2 merupakan kuadrat dari nilai korelasi sederhana dan korelasi parsial. Seluruh pasangan variabel dihitung nilai kuadrat korelasinya, lalu hasilnya dijumlahkan.

$$KMO = \frac{\Sigma\Sigma r_{ij}^2}{\Sigma\Sigma r_{ij}^2 + \Sigma\Sigma q_{ij}^2} \tag{2}$$

Pengujian pengukuran Kecukupan Sampel (MSA) dilakukan untuk setiap variabel. Nilai MSA yang tinggi (≥ 0.5) menandakan bahwa variabel yang berkorelasi mempertahankan korelasi yang wajar dengan variabel lain dan cocok untuk dimasukkan dalam proses PCA. Variabel yang menghasilkan nilai MSA di bawah ambang batas ini disarankan untuk dihilangkan, karena dapat membahayakan struktur faktor yang ditetapkan. MSA merupakan versi variabel individual dari KMO dan dihitung dengan membandingkan seberapa besar variabel tersebut berkorelasi secara umum dibandingkan secara parsial terhadap variabel lain, yang tertulis pada persamaan (3). Simbol r_{ij}^2 mengacu pada jumlah kuadrat korelasi sederhana antara variabel ke-i dengan seluruh variabel lainnya, sedangkan q_{ij}^2 menunjukkan jumlah kuadrat korelasi parsial antara variabel ke-i dengan variabel lainnya.

$$MSA_i = \frac{\Sigma r_{ij}^2}{\Sigma r_{ij}^2 + \Sigma q_{ij}^2} \tag{3}$$

Selain itu, Uji Bartlett digunakan untuk mengevaluasi H_0 yang menyatakan bahwa matriks korelasi menyerupai matriks identitas, yang menunjukkan kurangnya korelasi di antara variabel. X^2 hitung lebih tinggi dari tabel chi-square mengindikasikan ada korelasi signifikan antar variabel dalam dataset, sehingga PCA dapat dilakukan. Persamaan (4) menunjukkan perhitungan dari Bartlett, $X_{chisquare}^2$ merupakan nilai statistik uji chi-square yang akan digunakan untuk menguji H_0 bahwa matriks korelasi antar variabel sama dengan matriks identitas. Simbol n menunjukkan jumlah observasi atau sampel dalam *dataset*, sementara p menyatakan jumlah variabel yang dianalisis. Simbol $|R|$ adalah determinan dari matriks korelasi antar variabel.

$$X^2 = -(n - 1 - \frac{2p + 5}{6}) \, ln \, |R| \tag{4}$$

2.3 Reduksi Dimensi

Principal Component Analysis (PCA) digunakan untuk mengurangi dimensi data dengan mentransformasikan variabel-variabel asli menjadi komponen utama yang saling ortogonal. Komponen-komponen ini dipilih berdasarkan kriteria nilai eigen *value* yang lebih besar dari 1 dan kemampuan menjelaskan minimal 80% dari varians kumulatif data, sehingga dapat dikatakan sebagian besar informasi penting berhasil dipertahankan. Dengan teknik ini, analisis dapat disederhanakan dengan hanya berfokus pada beberapa komponen utama yang paling signifikan, memungkinkan reduksi dimensi yang efektif tanpa kehilangan esensi informasi dari *dataset* asli. Zang et al. menemukan bahwa kombinasi PCA dan *K-Means* mampu meningkatkan kestabilan dan akurasi pemisahan klaster pada data berdimensi tinggi [10]. Sementara itu, Rifqi et al. melaporkan bahwa penggunaan PCA secara signifikan memperbaiki kinerja metode klasterisasi hierarki [11]. Rumus perhitungan PCA dapat dilihat pada persamaan (5). Dalam rumus ini, **Z** adalah matriks skor komponen utama hasil PCA, yang merepresentasikan data dalam dimensi yang telah direduksi. X merupakan data asli yang telah distandardisasi, berukuran $n \times p$, dengan n observasi dan p variabel. Sementara itu, **W** adalah matriks eigen vektor dari kovarians atau korelasi X, yang menunjukkan arah komponen utama dan merupakan kombinasi linier dari variabel awal.

$$Z = XW \tag{5}$$

2.4 Metode Klasterisasi

Proses segmentasi wilayah dilakukan terhadap data hasil PCA menggunakan dua pendekatan klasterisasi, yaitu *K-Means Clustering* dan *Agglomerative Hierarchical Clustering* (AHC). *K-Means* bekerja dengan membagi data ke dalam sejumlah klaster (k) yang telah ditentukan, di mana setiap titik data dikelompokkan berdasarkan jarak *Euclidean* terdekat ke *centroid*. Proses iteratif ini berlanjut hingga posisi *centroid* stabil. *K-Means* masih menjadi pilihan yang baik karena kemampuannya dalam menghasilkan pembagian klaster yang jelas dan konsisten pada data yang telah direduksi dimensinya, seperti yang dilakukan oleh [12]. Setelah PCA merangkum variabel-variabel penting ke dalam beberapa komponen utama, struktur data menjadi lebih sederhana dan teratur, sehingga memudahkan *K-Means* dalam mengelompokkan data secara efektif. Dalam rumus *K-Means* pada persamaan (8), J adalah total kuadrat jarak antar data dan *centroid* klasternya (within-cluster sum of squares) yang diminimalkan oleh *K-Means*. k menunjukkan jumlah klaster, Ci adalah anggota klaster ke-i, x adalah data individual, dan μ_i adalah *centroid* klaster ke-i. Notasi $\left|\left|x - \mu_i\right|\right|^2$ menyatakan jarak *Euclidean*

kuadrat antara x dan μ_i. Tujuan *K-Means* adalah meminimalkan jarak ini agar tiap data sedekat mungkin ke *centroid*-nya.

$$J = \sum_{i=1}^{k} \sum_{x \in Ci} \left|\left|x - \mu_i\right|\right|^2 \tag{6}$$

Penerapan AHC kali ini menggunakan pendekatan hierarki *bottom-up* dengan *average linkage* dan *cosine distance*, dimulai dari setiap objek sebagai klaster individual yang kemudian digabungkan secara bertahap berdasarkan kemiripan hingga membentuk struktur dendrogram. Dalam analisis ini, *average linkage* digunakan sebagai strategi penggabungan, menentukan jarak antara klaster berdasarkan jarak rata-rata antara semua pasangan titik dari setiap klaster. *Averag*e *linkage* digunakan karena lebih tahan terhadap *outlier* daripada *single* dan *complete linkage* [13]. Metrik jarak yang dipilih harus tahan juga terhadap *outlier*, sehingga metrik *cosine* dirasa cocok karena *cosine* menilai perbedaan arah antara vektor dalam ruang berdimensi tinggi. Sudut lebih dapat diandalkan daripada jarak untuk deteksi *outlier* di dimensi tinggi [14]. Karena jarak kosinus juga berbasis sudut dan mengabaikan magnitudo, ia mewarisi properti ketahanan terhadap perubahan magnitudo yang sering disebabkan oleh *outlier*, terutama di lingkungan data yang kompleks dan berdimensi tinggi. Persamaan (9) merupakan rumus dari AHC *average linkage* dengan *metric cosine*. $D(A, B)$, menyatakan jarak antara dua klaster A dan B menggunakan metode *average linkage*, yaitu rata-rata jarak antar semua pasangan data dari kedua klaster. $|A|$ dan $|B|$ adalah jumlah anggota masing-masing klaster, sedangkan $x \epsilon A$ dan $y \epsilon B$ menunjukkan vektor data dari klaster A dan B. Jarak antar vektor dihitung dengan *cosine distance*, yang didasarkan pada *dot product* dan norma masing-masing vektor. Nilai *cosine distance* kecil menunjukkan arah vektor yang mirip.

$$D(A, B) = \frac{1}{|A||B|} \sum_{x \in A} \sum_{y \in B} d_{cos}(x, y) \tag{7}$$

2.5 Jumlah Klaster Optimal

Penentuan jumlah klaster optimal merupakan aspek kritis dalam algoritma *K-Means* dan *Agglomerative Hierarchical Clustering* (AHC). Untuk menentukannya, terdapat metode evaluasi yang umum digunakan, yaitu *Elbow Method*. *Elbow Method* mengidentifikasi titik optimal melalui grafik yang menunjukkan hubungan antara *Within-Cluster Sum of Squares* (WCSS) dan jumlah klaster, dengan titik siku (elbow) sebagai indikasi jumlah klaster yang ideal [15]. Sementara itu, *Silhouette Score* mengukur seberapa baik suatu objek ditempatkan dalam klasternya sendiri dibandingkan dengan klaster terdekat lainnya, dengan skor berkisar antara -1 hingga 1. Nilai yang mendekati 1 menunjukkan struktur klaster yang terpisah dengan baik [16].

Pendekatan ganda ini memastikan validitas hasil klasterisasi dengan mempertimbangkan baik efisiensi (Elbow Method) maupun kualitas pemisahan (Silhouette Score). Untuk lebih memperjelas hasil evaluasi, digunakan juga *Silhouette Plot* yang menggambarkan nilai *silhouette* untuk masing-masing anggota klaster. *Plot* ini memvisualisasikan seberapa konsisten setiap titik berada di dalam klasternya, sehingga memudahkan dalam menilai kualitas segmentasi yang terbentuk. *Plot* ini dikatakan baik apabila setiap klaster memiliki bentuk batang yang lebar, tinggi, dan tidak tumpang tindih satu sama lain serta tidak ada klaster yang berada pada rentang negatif [17]. Bentuk tersebut mengindikasikan bahwa objek dalam klaster memiliki keseragaman internal yang tinggi dan terpisah secara jelas dari klaster lainnya.

3. Hasil dan Diskusi

Data yang digunakan dalam analisis ini berupa data sekunder yang bersumber dari web resmi Badan Pusat Statistik (BPS) Provinsi Jawa Timur dan Dinas Kesehatan Provinsi Jawa Timur. Dari kedua web tersebut, kami mendapatkan berbagai variabel yang relevan dengan penelitian ini seperti jumlah puskesmas, jumlah rumah sakit umum, jumlah rumah sakit khusus, jumlah dokter, jumlah bidan, dan variabel-variabel lainnya. Selanjutnya, seluruh variabel dilakukan analisis secara eksploratif guna mendapatkan informasi terkait karakteristik *dataset*.

Dalam statistika, sangat penting untuk melakukan analisis eksploratif data (EDA) guna mengetahui karakteristik data dan mempermudah saat menginterpretasikan hasil analisis. Pada hal ini, statistika deskriptif sangat diperlukan, khususnya dalam tahap analisis eksploratif data. Melalui statistika deskriptif, kita bisa terbantu untuk memahami skala dan distribusi setiap variabel serta mengetahui variabel-variabel yang relevan. Berikut tabel hasil statistika deskriptif untuk mengetahui rata rata-rata, standar deviasi, nilai minimum, dan nilai maksimum dari data yang akan dianalisis.

Tabel 1. Hasil statistika deskriptif

Nama Variabel	Mean	Standar Deviasi	Nilai Minimum	Nilai Maksimum
Jumlah Puskesmas	25.56	12.87	3	63
Rumah Sakit Umum	8.37	7.33	2	39
Rumah Sakit Khusus	2.16	3.92	0	21
Dokter	800.34	1433.76	172	8733
Bidan	838.13	407.89	153	1942
Perawat	1890.76	1885.85	529	11933
Jumlah Apotek	55.87	36.41	11	159
Apoteker	345	416.24	42	2575
Jumlah Posyandu	1238.10	711.83	168	2876
Jumlah Klinik	25.61	34.54	0	170

Dari hasil statistika deskriptif di atas, kita dapat mengetahui bahwa terdapat beberapa skala data yang berbeda. Selain itu, beberapa variabel juga tampak menunjukkan perbedaan varians yang sangat besar. Sebagai contoh, jumlah rumah sakit khusus berada di kisaran 0-21 dan sangat berbanding terbalik dengan jumlah perawat, di mana jumlahnya berada di kisaran 529-11.933. Hal ini akan menimbulkan hasil yang bias terhadap analisis *unsupervised* seperti *clustering* jika data tidak distandardisasi. Mengutip informasi dari hasil statistika deskriptif, analisis akan dilanjutkan menggunakan metode *clustering* untuk mengelompokkan berbagai wilayah berdasarkan kemiripan jumlah ketersediaan fasilitas dan tenaga kesehatannya.

Analisis *clustering* merupakan suatu metode statistik yang bertujuan untuk mengelompokkan data berdasarkan kemiripan karakteristik. Seperti metode statistik pada umumnya, *clustering* memiliki beberapa prasyarat yang harus dipenuhi sebelum memulai analisis. Pada analisisnya, terdapat setidaknya tiga syarat yang harus dipenuhi, yaitu uji Bartlett, uji KMO (Kaiser-Meyer-Olkin), dan uji MSA (Measure of Sampling Adequacy).

Uji asumsi digunakan untuk menentukan penerimaan hipotesis nol atau hipotesis alternatif. Ketiga uji asumsi mengharapkan hasil uji untuk menolak H_0 sehingga analisis dapat dilanjutkan karena dianggap memenuhi semua prasyarat. Uji asumsi dilakukan dengan membandingkan hasil uji statistik terhadap tabel chi-square untuk uji Bartlett dan membandingkan uji statistik terhadap ambang batas uji (rule of thumb) untuk uji KMO dan MSA. Pada hal ini, H_0 akan diterima ketika nilai uji statistik kurang dari tabel chi-square atau nilai ambang batas. Sebaliknya, H_0 akan ditolak ketika nilai uji statistik lebih dari tabel chi square atau nilai ambang batas. Selain itu, ketiga uji asumsi juga menggunakan taraf signifikansi 5% sebagai batas toleransi wajar dalam uji statistik. Berikut tabel hasil pengujian prasyarat analisis *clustering*.

Tabel 2. Hasil uji asumsi analisis *clustering*

Nama Uji	Kriteria	Hasil Uji	Keputusan
Bartlett	$> \chi^2$	118.226 > 73.331	Tolak H_0; terdapat korelasi signifikan antar variabel
KMO	> 0.5	0.99 > 0.5	Tolak H_0; ukuran *sampling adequacy* data layak

Nama Uji	Kriteria	Hasil Uji	Keputusan
MSA	> 0.5	0.99 > 0.5	Tolak H_0; data dapat dianalisis lebih lanjut

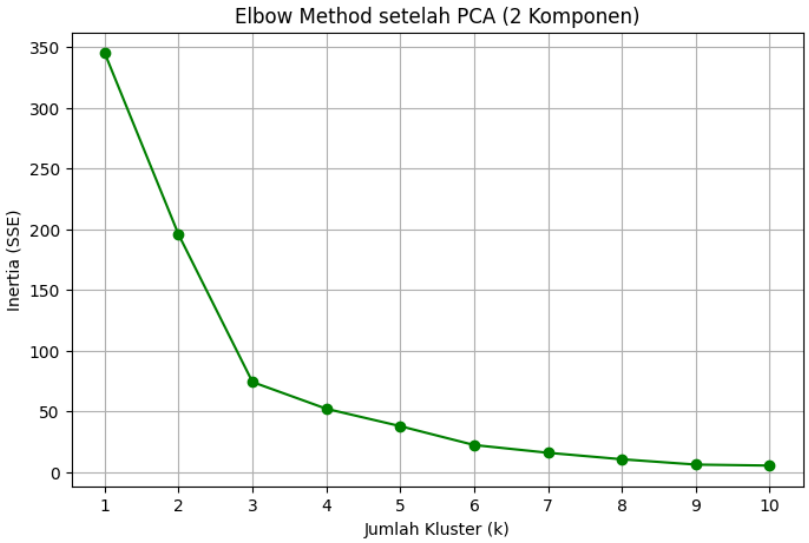
Dengan mengutip hasil dari uji asumsi, semua uji menunjukkan hasil yang sama, yaitu menolak H_0, oleh karena itu analisis dapat dilanjutkan. Dalam melakukan analisis, penting untuk mengetahui cara memaksimalkan informasi melalui data yang dimiliki. Pada *dataset* ini, terdapat sepuluh variabel, tetapi tidak semuanya perlu dipertahankan selama proses analisis. Dengan ini, bukan berarti variabel-variabel lain tidak penting, tetapi keseluruhan informasi yang ada pada data dapat dijelaskan tanpa menggunakan seluruh variabel yang ada. Dengan menggunakan PCA, kita dapat mengetahui jumlah komponen optimal yang mampu menjelaskan informasi keseluruhan pada *dataset* tanpa mengurangi sedikit pun informasi yang ada. Selain itu, hal ini juga membantu dalam proses visualisasi dan interpretasi, serta mengurangi kemungkinan terjadinya *overfitting*. Berikut dua proporsi varians tertinggi dari kesepuluh variabel pada *dataset*.

Tabel 3. Proporsi varians setiap variabel

Komponen	Proporsi varians
1	0.7418
2	0.1660

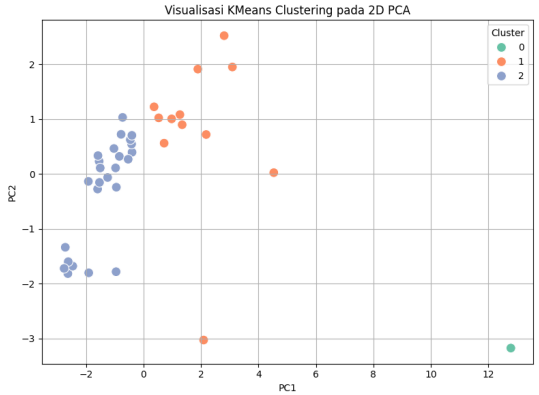
Berdasarkan hasil analisis proporsi varians tiap variabel, dengan menjumlahkan proporsi varians dari dua variabel pertama kita telah mendapatkan proporsi sebesar 0.91. Hal ini menandakan sepuluh variabel yang ada pada *dataset* mampu dijelaskan dengan baik cukup menggunakan dua data.

Pada analisis *clustering*, selain mengetahui jumlah variabel optimal, sangat penting untuk mengetahui jumlah klaster yang paling optimal. Dengan mengetahui jumlah klaster yang optimal, proses analisis akan menjadi lebih mudah. Metode siku (elbow method) sangat cocok digunakan untuk mengetahui jumlah klaster yang optimal. Pada analisis kali ini, kami melakukan PCA sebelum menerapkan *elbow method*. PCA dapat membantu menghindari sumber kesalahan, mendukung efisiensi komputasi, dan membantu saat evaluasi model. Berikut visualisasi *elbow method* setelah penerapan PCA untuk mengetahui jumlah klaster yang optimal.

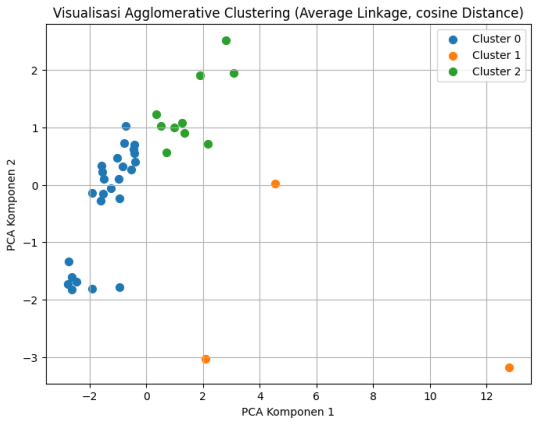


Gambar 1. *Elbow method* dengan penerapan PCA

Berdasarkan visualisasi *elbow method* di atas, diketahui jumlah klaster yang optimal pada *dataset* kami adalah tiga. Jumlah klaster yang optimal akan dijadikan acuan selama analisis *clustering* menggunakan metode yang telah dipilih. Berikut visualisasi hasil *clustering* sebanyak tiga klaster menggunakan algoritma *K-Means* dan *Average Linkage* berbasis *Cosine Similarity*.



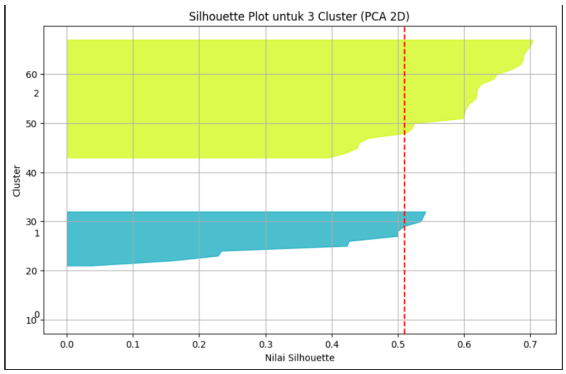
Gambar 2. *K-Means Clustering*



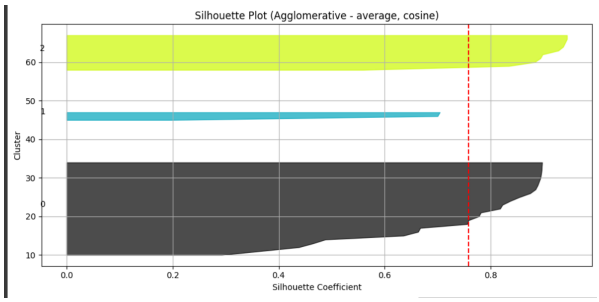
Gambar 3. *Average Linkage* (Cosine)

Dari kedua visualisasi hasil *clustering* di atas, terdapat perbedaan yang cukup signifikan pada pembagian klasternya. Pada *K-Means Clustering*, klasterisasinya dipartisi berdasarkan *centroid*, sedangkan pada AHC dipartisi berdasarkan kemiripan (cosine similarity). Dua metode tersebut sama-sama menghasilkan pencilan (outlier), tetapi algoritma AHC yang menggunakan pendekatan *cosine similarity* tampak lebih baik dalam memvisualisasikan hasil *clustering*. Saat menggunakan *K-Means*, terdapat satu data yang tergolong ekstrim (memiliki karakteristik berbeda), tetapi jika menggunakan AHC terdapat tiga data ekstrim dan berhasil dikelompokkan menjadi

satu. Secara umum, hal ini mungkin terkesan sama saja, tetapi hasilnya akan terlihat saat diuji menggunakan *silhouette score*. Berikut merupakan visualisasi *silhouette plot* dari *silhouette score* berdasarkan hasil klasterisasi masing-masing metode.



Gambar 4. *Silhouette score K-Means*



Gambar 5. *Silhouette score* AHC

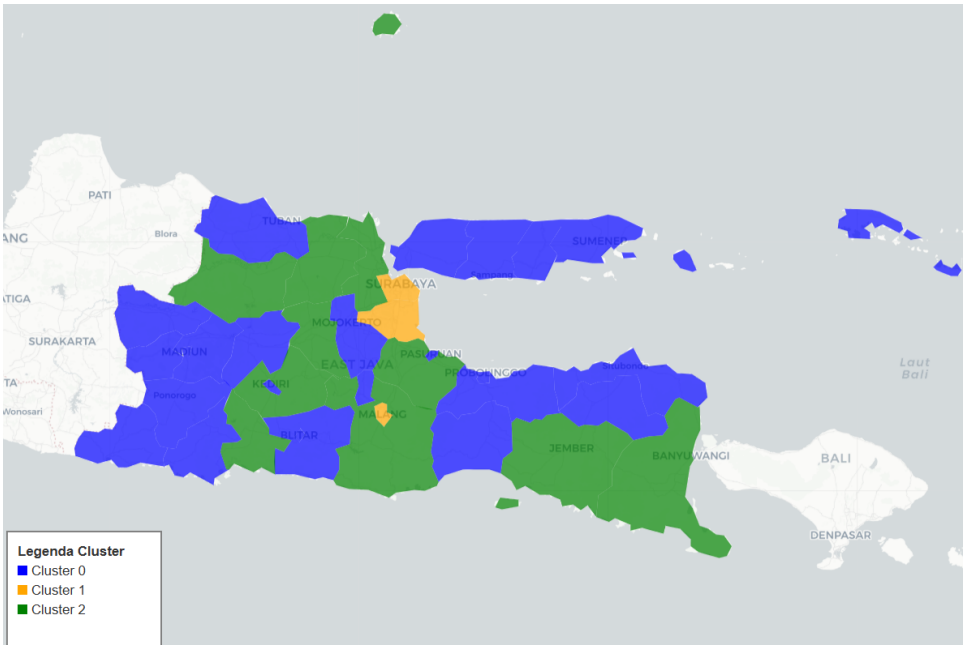
Dari kedua visualisasi *plot silhouette score* di atas, tampak dengan jelas perbedaannya secara signifikan. Pada hasil klasterisasi *K-Means*, terdapat satu data pencilan yang menyebabkannya tidak terbaca oleh *silhouette plot* dikarenakan jumlahnya yang hanya satu dari sekian banyak baris data pada *dataset*. Hal itu memengaruhi hasil *silhouette score* sehingga perbedaan *silhouette score* kedua metode cukup jauh. Sebaliknya, klasterisasi AHC yang dibantu pendekatan *cosine similarity* dinilai cukup baik berdasarkan *silhouette score*-nya. Klasterisasi dengan metode AHC menggunakan metrik *cosine similarity* berhasil mengelompokkan beberapa data yang dinilai ekstrim dengan baik, sehingga seluruh klasternya berhasil tervisualisasikan pada *silhouette plot*. Hal itulah yang menjadi pembeda mutlak dari kedua metode. Berikut hasil *silhouette score* masing-masing metode.

Tabel 4. *Silhouette score* masing-masing metode

Metode	***Silhouette Score***
K-Means	0.51
Average Linkage (Cosine Similarity)	**0.75**

Perbedaan *silhouette score* yang cukup jauh antara kedua metode merupakan pengaruh dari perbedaan algoritma pada masing-masing metode. Pada *K-Means*, terdapat satu pencilan data (outlier) di antara sekian banyak baris data, menyebabkannya gagal dibaca oleh *silhouette plot*. Sementara itu, pada AHC dengan pendekatan *cosine similarity*, data-data yang memiliki karakteristik paling berbeda dapat dikelompokkan dengan lebih baik ke dalam satu klaster, sehingga seluruh klaster data dapat dibaca oleh *silhouette plot*. Hal ini memengaruhi tinggi atau rendahnya *silhouette score* dan menjadi pembeda mutlak antara kedua metode yang digunakan. Selanjutnya, untuk mengetahui

secara lebih jelas hasil klasterisasinya, kami menggunakan peta di Provinsi Jawa Timur. Berikut peta Provinsi Jawa Timur beserta klasternya.



Gambar 5. Visualisasi peta *clustering* Provinsi Jawa Timur

Pada peta *clustering*, klaster nol tersebar rata di Provinsi Jawa Timur dengan meliputi beberapa kabupaten dan kota seperti Pacitan, Ponorogo, Magetan, Trenggalek, Tuban, Madura, Kediri, Mojokerto, Probolinggo dan sekitarnya, serta beberapa pulau di sekitar Pulau Madura. Daerah-daerah pada klaster nol cenderung memiliki jumlah ketersediaan fasilitas dan tenaga kesehatan yang rendah daripada daerah lain. Sebagai contoh, Kota Batu hanya memiliki masing-masing lima puskesmas dan rumah sakit umum yang menjadikannya daerah dengan jumlah puskesmas paling sedikit di Provinsi Jawa Timur. Kekurangan ketersediaan fasilitas dan tenaga kesehatan pada daerah di klaster nol dapat dijadikan pertimbangan oleh pemerintah untuk menambah jumlah fasilitas pelayanan kesehatan yang tersedia.

Pada klaster dua, beberapa kabupaten dan kota seperti Bojonegoro, Tulungagung, Lamongan, Jombang, Jember, Banyuwangi, beberapa daerah di sekitar Malang dan Pasuruan, serta Pulau Bawean terkategori di dalamnya. Daerah-daerah pada klaster ini tergolong sebagai wilayah dengan jumlah ketersediaan fasilitas dan tenaga kesehatan yang cukup. Hal ini dapat dijadikan acuan bagi pemerintah untuk mempertahankan jumlah ketersediaan fasilitas layanan kesehatan. Akan tetapi, lebih baik lagi jika pemerintah mengusahakan penambahan jumlah layanan kesehatan di daerah-daerah pada klaster dua, sehingga akses ke layanan kesehatan akan lebih mudah didapat oleh masyarakat.

Sementara itu, klaster satu diisi oleh daerah-daerah seperti Kota dan Kabupaten Malang, Kota Surabaya, serta Kabupaten Sidoarjo. Wilayah-wilayah tersebut tergolong sebagai pusat perekonomian, pendidikan, dan perindustrian di Provinsi Jawa Timur. Kota Surabaya, sebagai ibukota Provinsi Jawa Timur menempati posisi teratas dalam hal ketersediaan jumlah layanan kesehatan, di mana terdapat 63 puskesmas, 39 rumah sakit umum, dan 21 rumah sakit khusus beserta ribuan hingga belasan ribu tenaga kesehatan. Dengan melimpahnya jumlah layanan kesehatan yang tersedia, diharapkan pemerintah dapat mempertahankan kualitas serta kuantitas fasilitas kesehatan pada daerah-daerah di klaster satu.

4. Kesimpulan

Berdasarkan hasil analisis, dapat disimpulkan bahwa pemilihan pendekatan dalam metode klasterisasi berpengaruh signifikan terhadap kualitas hasil analisis. Metode *K-Means* yang berbasis *centroid* cenderung kurang efektif dalam mendeteksi data pencilan, sebagaimana tercermin dari nilai *silhouette score* sebesar 0.51 dan adanya klaster yang tidak terbaca pada visualisasi *plot*. Sebaliknya, metode *Average Linkage* dengan pendekatan *cosine similarity* mampu mengelompokkan data pencilan dengan lebih baik, menghasilkan *silhouette score* yang lebih tinggi, yaitu 0.75. Hasil ini menunjukkan bahwa metode AHC lebih sesuai untuk data berdimensi tinggi yang mengandung *outlier*. Oleh karena itu, hasil klasterisasi ini dapat digunakan sebagai dasar bagi pemerintah dalam merumuskan kebijakan peningkatan serta pemerataan fasilitas dan tenaga kesehatan secara lebih akurat di Provinsi Jawa Timur.

Daftar Pustaka

[1] Rakasiwi, L. S., Kautsar, A., & Keuangan, K. E. Pengaruh Faktor Demografi dan Sosial Ekonomi terhadap Status Kesehatan Individu di Indonesia. *Jurnal Keuangan dan Ekonomi*, 5, 12220, 2021. https://doi.org/10.31685/kek.V5.2.1008

[2] Mentari, G. B., & Susilawati, S. Faktor-Faktor yang Mempengaruhi Akses Pelayanan Kesehatan di Indonesia. *Jurnal Health Sains*, 3(6), 767–773, 2022. https://doi.org/10.46799/jhs.v4i06.512

[3] Rumahorbo, A. C., Kemal, D., & Sekarwati, A. Penerapan Data Mining dengan Menggunakan Algoritma C4.5 pada Klasifikasi Fasilitas Kesehatan Provinsi di Indonesia. *Jurnal Ilmiah Komputasi dan Sistem Informasi*, 2020. https://doi.org/10.32409/jikstik.19.1.2681

[4] Hidayanti, H. Pemerataan Tenaga Kesehatan di Kabupaten Lamongan. *Cakrawala*, 12(2), 162–177, 2019. https://doi.org/10.32781/cakrawala.v12i2.272

[5] Wibowo, A. S., & Mulyastuti, I. D. Penerapan Algoritma K-Means Clustering pada Jumlah Fasilitas Kesehatan Menurut Pemerintah Provinsi DKI Jakarta. *Badan Pusat Statistik DKI Jakarta*, 2022.

[6] Hamami, F., & Dahlan, A. Penerapan Algoritma K-Means untuk Memetakan Persebaran Fasilitas dan Tenaga Kesehatan di Kota Bandung. *SWADHARMA (JRIS)*, 2024.

[7] Yusniyanti, A. L., Virgantari, F., & Faridhan, Y. E. Comparison of Average Linkage and K-Means Methods in Clustering Indonesia's Provinces Based on Welfare Indicators. *Journal of Physics: Conference Series*, 2021. https://doi.org/10.1088/1742-6596/1863/1/012071

[8] Az-Zahra, A., & Wijayanto, A. W. Tinjauan Kesejahteraan di Daerah Perbatasan Republik Indonesia Tahun 2021: Penerapan Analisis Klaster K-Means dan Hierarki. *Jurnal Sistem dan Teknologi Informasi (JustIN)*, 12(1), 55, 2024. https://doi.org/10.26418/justin.v12i1.69040

[9] Amelia, R. N., Aji, S., Kriswantoro, K., & Sukmasari, H. Proof of Unidimensionality in Cognitive Test Instrument for Evaluation Science Learning. *Journal of Innovation in Educational and Cultural Research*, 6(1), 16–24, 2025. https://doi.org/10.46843/jiecr.v6i1.1897

[10] Zhang, C., Ou, J., He, W., Huang, H., Cheng, G., & Gu, Y. Optimisation Research on K-Means Clustering Algorithm Based on Principal Component Analysis and Percentile Improvement. *Proceedings of the 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 148–153, 2024. https://doi.org/10.1109/ICAICA63239.2024.10823007

[11] Bhahari, R. H., & Kusnawi, K. Clustering Analysis of Socio-Economic Districts/Cities in East Java Province Using PCA and Hierarchical Clustering Methods. *Sinkron*, 8(4), 2242–2251, 2024. https://doi.org/10.33395/sinkron.v8i4.14078

[12] Dewi, S., & Pakereng, M. A. I. Implementasi Principal Component Analysis pada K-Means untuk Klasterisasi Tingkat Pendidikan Penduduk Kabupaten Semarang. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, 8(4), 1186–1195, 2023. https://doi.org/10.29100/jipi.v8i4.4101

[13] Jarman, A. M. Hierarchical Cluster Analysis: Comparison of Single Linkage, Complete Linkage, Average Linkage and Centroid Linkage Method. *ResearchGate Preprint*, n.d. https://doi.org/10.13140/RG.2.2.11388.90240

[14] Kriegel, H.-P., Schubert, M., & Zimek, A. Angle-Based Outlier Detection in High-Dimensional Data. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*, 444–452, 2008

[15] Ekasetya, V. A., & Jananto, A. Klusterisasi Optimal dengan Elbow Method untuk Pengelompokan Data Kecelakaan Lalu Lintas di Kota Semarang. *Dinamika Informatika*, 12(1), 20–28, n.d.

[16] Nugroho, N., & Adhinata, F. D. Penggunaan Metode K-Means dan K-Means++ sebagai Clustering Data Covid-19 di Pulau Jawa. *Teknika*, 11(3), 170–179, 2022. https://doi.org/10.34148/teknika.v11i3.502

[17] Shutaywi, M., & Kachouie, N. N. Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy*, 23(6), 2021. https://doi.org/10.3390/e23060759