Seminar Nasional Sains Data 2025 (SENADA 2025)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Optimasi Segmentasi Tingkat Biaya Hidup Provinsi di Indonesia Berdasarkan Harga Komoditas dan Layanan

Muhammad Erlangga Kurniawan[1], Aditya Putra Ananta[2], Muhammad Cahya Raka Anugrah[3], Shindi Shella May Wara[4], Wahyu Syaifullah Jauharis Saputra[5]

[1, 2, 3, 4, 5]*Sains Data, UPN "Veteran" Jawa Timur*

[1]23083010045@student.upnjatim.ac.id
[2]23083010048@student.upnjatim.ac.id
[3]23083010070@student.upnjatim.ac.id
[4]shindi.shella.fasilkom@upnjatim.ac.id
[5]wahyu.s.j.saputra.if@upnjatim.ac.id

Abstract: *The cost of living in Indonesia varies significantly across provinces, influenced by staple commodity prices (rice, chicken eggs, beef, red chili, cooking oil), electricity tariffs, provincial minimum wage (UMP), poverty line, and construction cost index (IKK). This study clusters provinces based on cost-of-living characteristics using two approaches: K-Means and Hierarchical Clustering (Ward linkage), selecting 3 clusters. Data were standardized using Z-score and reduced via Principal Component Analysis (PCA). Analysis revealed PCA improved clustering quality, with Ward achieving the highest Silhouette Score (0.44). However, K-Means proved more stable (Silhouette Score 0.41) with no outliers. K-Means identified three distinct clusters: Cluster 2 (low cost-of-living regions like Java, Bali, and Sulawesi), Cluster 0 (medium cost-of-living areas in Eastern Indonesia, including Papua and Maluku), and Cluster 1 (a single high cost-of-living outlier province). Given its stability, K-Means was selected as the preferred method. These findings provide a robust basis for targeted regional economic policies tailored to each cluster's characteristics.*
Keywords: *K-Means, cluster, PCA, silhouette score, ward*
Abstrak: Tingkat biaya hidup di Indonesia menunjukkan variasi signifikan antarprovinsi, dipengaruhi oleh harga komoditas pokok (beras, telur ayam, daging sapi, cabai merah, minyak goreng), tarif listrik, upah minimum provinsi (UMP), garis kemiskinan, dan indeks kemahalan konstruksi (IKK). Penelitian ini bertujuan mengelompokkan provinsi berdasarkan kesamaan karakteristik biaya hidup menggunakan dua pendekatan *clustering*: *K-Means* dan *Hierarchical Clustering* dengan *Ward linkage* dengan klaster yang terpilih sejumlah 3 klaster. Data distandardisasi menggunakan *Z-Score* dan direduksi dimensinya dengan *Principal Component Analysis* (PCA). Hasil analisis menunjukkan bahwa PCA meningkatkan kualitas *clustering* dengan *Silhouette Score* tertinggi 0.44 pada metode *Ward*. Di sisi lain, *K-Means* menghasilkan *Silhouette Score* 0.41 dengan distribusi klaster yang lebih stabil dan tanpa *outlier*. Karakteristik pada metode *K-Means clustering* menunjukkan tiga pola klaster berbeda. Klaster 0 mencakup wilayah dengan biaya hidup menengah terutama di Indonesia bagian Timur seperti Papua dan Maluku, klaster 2 mencakup wilayah dengan dengan biaya hidup rendah seperti Jawa, Bali, dan Sulawesi, sementara klaster 1 mencakup provinsi *outlier* dengan biaya hidup yang sangat tinggi. Berdasarkan pertimbangan ini, penelitian memilih metode *K-Means* sebagai pendekatan *clustering* yang lebih andal untuk analisis segmentasi biaya hidup. Temuan ini memberikan dasar penting bagi penyusunan kebijakan publik yang lebih tepat sasaran sesuai karakteristik ekonomi masing-masing wilayah.
Kata kunci: *K-Means*, *cluster*, PCA, *silhouette score*, *ward*

I. PENDAHULUAN

Tingkat biaya hidup di Indonesia menunjukkan adanya perbedaan yang cukup tinggi antar provinsi yang dipengaruhi oleh beberapa variabel ekonomi seperti harga komoditas pokok, tarif listrik, hingga standar upah minimum. Perbedaan tersebut tidak hanya dipengaruhi oleh pendapatan masyarakat, tetapi juga oleh volatilitas harga komoditas pokok dan tingkat inflasi. Perbedaan ini bukan hanya mencerminkan kondisi ekonomi daerah, tetapi juga berimplikasi langsung terhadap kesejahteraan masyarakat, daya beli, dan efektivitas kebijakan fiskal nasional maupun daerah. Studi oleh Zaman et al. menunjukkan bahwa kenaikan harga pangan dapat meningkatkan pendapatan petani dalam jangka pendek, meskipun tekanan inflasi tetap berpotensi dalam menurunkan daya beli dan kesejahteraan mereka secara keseluruhan [1]. Hal

ini menunjukkan bahwa tidak semua peningkatan harga pangan memberikan keuntungan bersih bagi pelaku ekonomi di sektor bawah.

Hingga kini, pengukuran dan pemetaan tingkat biaya hidup masih cenderung menggunakan pendekatan umum atau rata-rata nasional, tanpa adanya pertimbangan karakteristik khusus masing-masing provinsi. Perbedaan struktur konsumsi, akses terhadap komoditas, serta daya beli masyarakat di tiap provinsi dapat menciptakan profil biaya hidup yang unik. Menurut penelitian oleh Auliasari et al., fluktuasi harga bahan pangan seperti beras, gula, dan telur dapat bervariasi cukup signifikan antar provinsi, dengan wilayah timur Indonesia memiliki tingkat harga komoditas tertinggi jika dibandingkan dengan wilayah barat [2]. Menurut Rizal dan Fitria, sistem perdagangan bahan pangan saat ini semakin terbuka yang menyebabkan produk pangan dalam negeri sulit dikendalikan akibat transmisi dari situasi dan kondisi harga keseluruhan [3]. Hal ini dapat disebabkan oleh faktor-faktor seperti tingginya permintaan barang lokal, tingkat biaya hidup, maupun ketersediaan infrastruktur yang memengaruhi distribusi penjualan bahan pangan. Situasi tersebut mempertegas bahwa diperlukan adanya pendekatan segmentasi biaya hidup yang lebih akurat dan terperinci. Dengan memetakan provinsi-provinsi berdasarkan indikator harga komoditas dan layanan esensial, diharapkan dapat diperoleh pemahaman yang lebih mendalam terhadap karakteristik ekonomi di masing-masing daerah. Pendekatan pengelompokkan berbasis data seperti *K-Means* maupun *hierarchical clustering* lainnya telah terbukti mampu mengidentifikasi pola kemiripan karakteristik wilayah dalam konteks sosial dan ekonomi. Penelitian yang dilakukan oleh Anis et al. menegaskan bahwa segmentasi wilayah berbasis indikator sosial ekonomi sangatlah penting untuk mendukung kebijakan berbasis data [4].

Dalam penelitian ini, variabel yang digunakan mencerminkan aspek-aspek utama yang memengaruhi tingkat biaya hidup di setiap provinsi di Indonesia. Variabel-variabel tersebut mencakup harga komoditas pokok yang menjadi kebutuhan dasar masyarakat, antara lain harga beras, telur ayam, daging sapi, cabai merah, dan minyak goreng. Selain itu, turut dipertimbangkan variabel layanan esensial seperti tarif listrik dan upah minimum provinsi (UMP), yang merefleksikan daya beli masyarakat. Garis kemiskinan digunakan sebagai indikator kesejahteraan minimum di suatu wilayah, sedangkan indeks kemahalan konstruksi (IKK) merepresentasikan tingkat biaya pembangunan infrastruktur di masing-masing provinsi. Kesembilan variabel ini dipilih karena dinilai mampu memberikan gambaran komprehensif terhadap perbedaan struktur biaya hidup antarwilayah dan sekaligus relevan untuk dianalisis dalam proses segmentasi wilayah berbasis metode *clustering*.

Beberapa penelitian terdahulu telah banyak menerapkan metode *clustering* untuk segmentasi wilayah berdasarkan indikator ekonomi. Albertus dan Maria menggunakan metode *K-Means* untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan indikator sosial ekonomi seperti pengeluaran per kapita, tingkat pendidikan, dan akses layanan kesehatan [5]. Penelitian tersebut menunjukkan bahwa pendekatan pengelompokan wilayah dapat membantu dalam memahami karakteristik kesejahteraan masyarakat secara regional sebagai dasar dalam perumusan kebijakan pembangunan daerah. Sementara itu, Aditya et al. menerapkan metode *K-Means* untuk mengelompokkan kabupaten/kota di Provinsi Jawa Timur berdasarkan indikator produktivitas komoditas pangan [6]. Hasil penelitian mereka menunjukkan bahwa pendekatan segmentasi wilayah mampu mengidentifikasi ketimpangan ekonomi secara efektif. Penelitian lain mengatakan bahwa metode *K-Means* lebih baik dibandingkan metode *Ward* dalam membuat segmentasi wilayah berdasarkan beberapa kasus, seperti yang dilakukan oleh Nahya et al. di mana metode *K-Means* menghasilkan pengelompokkan klaster yang lebih baik

saat ditinjau menggunakan nilai *Silhouette* sebesar 0.48 dan membentuk dua klaster, sedangkan metode *Ward* menghasilkan nilai *Silhouette* sebesar 0.47 dalam mengelompokkan daerah berisiko *stunting* [7]. Metode *K-Means* juga digunakan oleh Shindi et al. yang menggabungkan dengan SVM untuk meningkatkan efektivitas klasifikasi pengolahan citra digital [8]. Penelitian lain yang dilakukan oleh Sri et al. mengatakan bahwa metode *K-Means clustering* menghasil rasio Sw (simpangan baku dalam klaster) dan Sb (simpangan baku antar klaster) yang lebih besar dibandingkan metode *Ward* [9].

Merujuk dari hasil penelitian tersebut, kami melakukan pendekatan *clustering* dengan metode *K-Means* sebagai landasan penting dalam penelitian ini yang bertujuan untuk memetakan provinsi di Indonesia berdasarkan indikator yang merepresentasikan tingkat biaya hidup. Untuk mengoptimalkan hasil segmentasi, kami menerapkan metode *Principal Component Analysis* (PCA) untuk mereduksi dimensi data serta meningkatkan kualitas pemisahan klaster. Pendekatan ini diharapkan mampu memberikan gambaran yang lebih komprehensif terhadap pola biaya hidup antarprovinsi dan dapat dimanfaatkan untuk penyusunan kebijakan yang lebih adaptif terhadap kondisi ekonomi lokal.

II. METODE PENELITIAN

II.1. Standardisasi

Proses normalisasi data menggunakan algoritma *Z-Score* dilakukan sebelum pengelompokan untuk menyamakan skala variabel numerik, mencegah dominasi salah satu variabel akibat perbedaan skala atau unit, baik dalam pendekatan dengan maupun tanpa reduksi dimensionalitas (PCA). Tahap ini krusial karena variasi skala berdampak signifikan pada algoritma berbasis jarak, seperti *K-Means* dan hierarki, di mana penelitian Ismail et al. menunjukkan bahwa *Z-Score* menghasilkan klaster lebih baik dibanding *Min-Max* atau *Decimal Scaling* [10]. Pada pendekatan non-PCA, standardisasi memastikan bobot variabel setara, sedangkan sebelum PCA, standardisasi mencegah komponen utama didominasi variabel bervariansi tinggi. Dengan demikian, standardisasi menjadi langkah penting untuk memastikan hasil klasterisasi akurat dan merepresentasikan struktur alami data. *Z-Score* merupakan suatu transformasi linear yang mengonversi data mentah ke dalam skala relatif terhadap distribusi normal standar, di mana nilai yang dihasilkan mengkuantifikasi penyimpangan suatu observasi dari *mean* populasi dalam satuan standar deviasi dan dapat ditulis seperti persamaan berikut.

$$z = \frac{x-\mu}{\sigma} \qquad (1)$$

II.2. Metode klasterisasi: *K-Means* dan Hierarki

Penelitian ini mengimplementasikan dua pendekatan *clustering*, yaitu *K-Means* (non-hierarki) dan *Agglomerative Hierarchical Clustering* - AHC (hierarki), untuk memperoleh hasil pengelompokan yang komprehensif melalui metodologi berbeda, sebagaimana digunakan dalam penelitian sejenis di bidang fasilitas kesehatan [11]. *K-Means* ini efisien dalam pengunaannya, tetapi sensitif terhadap inisialisasi *centroid* awal [12]. Algoritma metode ini dimulai dengan memilih jumlah klaster K dan kemudian secara acak menetapkan pusat klaster, setiap titik data dihubungkan ke pusat terdekat menggunakan jarak *Euclidean*, kemudian pusat diperbarui sebagai lokasi rata-rata data dalam setiap klaster, dan iterasi berlanjut hingga konvergensi atau batas iterasi yang telah ditetapkan tercapai. Algoritma *K-Means* dapat dilihat pada persamaan berikut.

$$J = \sum_{i=1}^{k} \quad \sum_{x \in Ci} \quad \left|\left|x - \mu_i\right|\right|^2 \qquad (2)$$

Berbeda dengan klasterisasi non-hierarki, AHC membangun struktur hierarki *bottom-up* dengan menggabungkan klaster paling similar secara progresif hingga membentuk satu klaster utama, menggunakan *Ward linkage* yang meminimalkan varians intra-klaster dan optimal untuk *dataset* kecil seperti yang dibuktikan Apriliana et al. [13], dengan keunggulan tidak memerlukan penentuan jumlah klaster awal karena dapat ditentukan melalui analisis dendrogram [14]. Persamaan (3) merupakan rumus dari *Ward linkage*, yakni penggabungan dua klaster berdasarkan kenaikan terkecil dalam total varians (jumlah kuadrat jarak) jika dua klaster digabung. Metode ini menggunakan jarak *Euclidean* kuadrat antar *centroid* untuk mengukur seberapa besar penyebaran data akan bertambah setelah penggabungan.

$$D(A,B) = \frac{|A||B|}{|A|+|B|} \cdot \left|\left|\mu_a - \mu_b\right|\right|^2 \qquad (3)$$

II.3. Elbow Method

Penentuan jumlah klaster optimal dalam *K-Means* dilakukan menggunakan Metode *Elbow* dengan menganalisis pola penurunan *Sum of Squared Errors* (SSE) seperti pada persamaan (2), yang mengkuantifikasi total jarak kuadrat antara titik data dan *centroid* klaster terhadap peningkatan jumlah klaster. Titik optimal ditandai dengan perlambatan signifikan penurunan SSE yang membentuk pola siku pada grafik [15]. Metode ini kemudian diaplikasikan secara konsisten baik pada analisis tanpa PCA maupun dengan PCA, serta menjadi acuan penentuan jumlah klaster dalam *Agglomerative Hierarchical Clustering* (AHC) untuk memastikan komparasi yang objektif antara kedua pendekatan *clustering* [16]. Sebagaimana yang telah diterapkan dalam penelitian serupa oleh Nadia dan Evanita [17] sehingga memungkinkan evaluasi kinerja yang konsisten antar metode.

II.4. PCA sebagai teknik reduksi dimensi

Principal Component Analysis (PCA) merupakan teknik reduksi dimensi yang mentransformasikan data kompleks menjadi komponen utama dengan mempertahankan informasi esensial, di mana penerapannya dalam *clustering* berfungsi untuk meminimalkan *noise* dan redundansi variabel sehingga memperjelas batas antar klaster. PCA diaplikasikan pada data terstandardisasi dengan mempertahankan komponen utama yang menjelaskan 70-80% varians total, kemudian hasil reduksi dimensi ini dijadikan input untuk algoritma *K-Means* dan *Agglomerative Hierarchical Clustering* (AHC) dengan perbandingan hasil terhadap *clustering* data tanpa reduksi untuk mengevaluasi efektivitasnya. Studi oleh Zang et al. membuktikan bahwa integrasi PCA dengan *K-Means* menghasilkan pemisahan klaster lebih stabil pada data dimensi tinggi [18], sementara penelitian Rifqi et al. menunjukkan peningkatan signifikan dalam efektivitas *clustering* hierarki [19]. Selain meningkatkan kualitas klaster, PCA juga memberikan manfaat komputasional seperti pencegahan *overfitting* dan percepatan proses analisis, khususnya pada *dataset* multivariat. Hasil reduksi dimensi dengan PCA didapatkan dengan mengalikan data asli dengan matriks eigen vektor, yang tercantum pada persamaan (4).

$$Z = XW \qquad (4)$$

II.5. *Silhouette Score* dan Plot

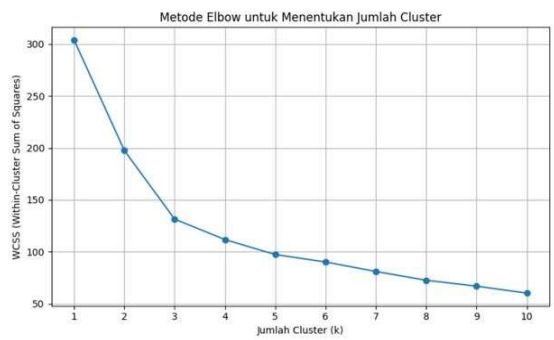
Evaluasi kualitas klasterisasi dilakukan menggunakan *Silhouette Score* yang mengukur seberapa baik suatu objek cocok dengan klaster asalnya dibandingkan klaster lain, dengan rentang nilai -1 hingga 1 dimana nilai mendekati 1 menunjukkan pengelompokan optimal, nilai sekitar 0 mengindikasikan posisi di batas klaster, dan nilai negatif menandakan kesalahan pengelompokan. Analisis ini dilengkapi dengan *Silhouette Plot* untuk visualisasi distribusi kualitas pengelompokan tiap klaster, memungkinkan identifikasi klaster tidak stabil, terlalu kecil, atau heterogen — di mana klaster ideal ditunjukkan oleh plot yang lebar dan seragam, sementara ketidakseimbangan dan nilai negatif mengungkap kelemahan hasil klasterisasi. Metode evaluasi ini telah terbukti efektif dalam berbagai penelitian, termasuk studi Hasan yang mendemonstrasikan kehandalan *Silhouette Score* dalam menilai performa klasterisasi data berdimensi tinggi [20]. Persamaan (5) adalah rumus perhitungan skor *silhouette* yang dimulai dengan menghitung setiap titik data, menghitung rata-rata jarak antara titik tersebut dan semua titik lain dalam kluster yang sama (a). Lalu, dihitung pula rata-rata jarak antara titik tersebut dan semua titik dalam kluster terdekat yang berbeda (b). Selisih antara b dan a, yang kemudian dibagi dengan nilai maksimum dari keduanya, akan menghasilkan skor *silhouette* untuk titik tersebut.

$$s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \qquad (5)$$

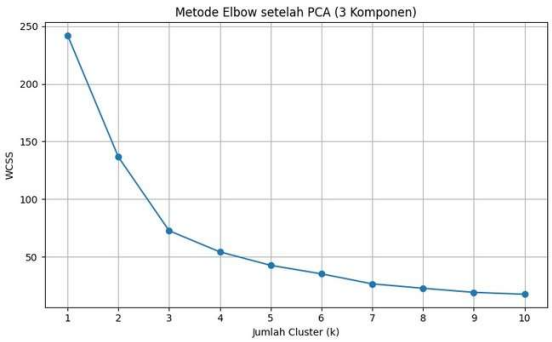
III. HASIL DAN PEMBAHASAN

Data yang kami gunakan diambil dari web resmi BPS (Badan Pusat Statistik) Indonesia. Jumlah data yang dianalisis terdapat 38 baris, mengikuti banyaknya jumlah ibukota setiap provinsi di Indonesia. Data ini melibatkan delapan variabel antara lain harga komoditas meliputi harga beras, daging sapi, telur, minyak, dan cabai merah. Selain itu, terdapat tiga variabel berupa layanan esensial seperti garis kemiskinan, tarif listrik, dan UMP (upah minimum provinsi). Kedelapan variabel tersebut dinilai relevan untuk menganalisis taraf biaya hidup tiap daerah di Indonesia sebagaimana yang telah disampaikan di bagian pendahuluan.

Berdasarkan hasil analisis yang telah dilakukan menggunakan metode-metode pada bagian metode penelitian, selanjutnya akan dibahas secara singkat, padat, dan jelas hasil-hasil analisisnya. Pada klasterisasi, penting untuk mengetahui berapa jumlah klaster yang optimal sebelum memulai analisis. Pada analisis ini, kami menggunakan *elbow method* (metode siku) sebanyak dua kali untuk membandingkan pengaruh antara sebelum dan sesudah menerapkan PCA (Principal Component Analysis). Hasil *elbow method* terlampir sebagai berikut.

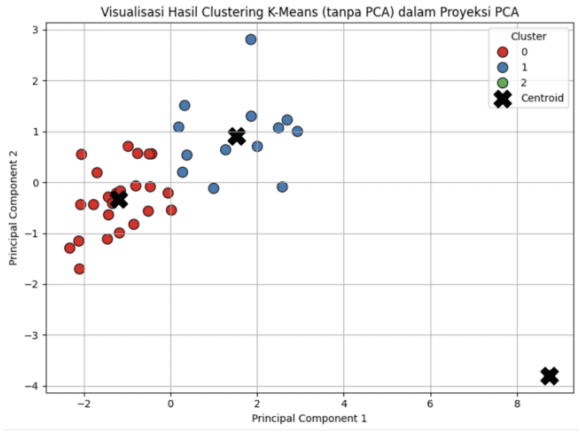


Gambar 1a. *Elbow Method* Sebelum PCA

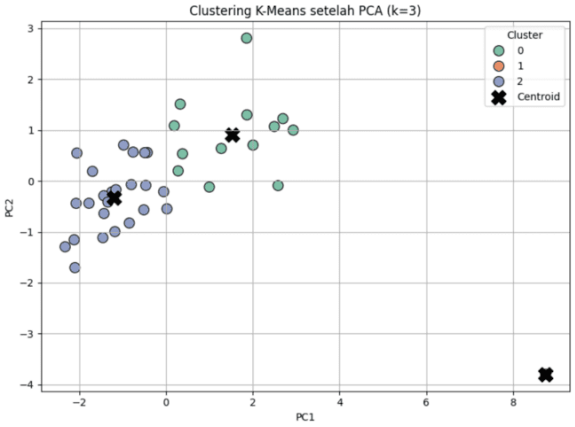


Gambar 1b. *Elbow Method* Setelah PCA

Berdasarkan dua gambar di atas, sekilas tampak tidak ada yang berbeda antara sebelum dan sesudah menerapkan PCA pada *elbow method*. Namun, jika diteliti lebih lanjut, tampak nilai SSE menurun pada *elbow method* yang disisipkan PCA di dalamnya (gambar 1b). Hal ini juga membuktikan dengan mereduksi jumlah dimensi data, menyebabkan nilai SSE/WCSS menjadi lebih efisien sehingga membantu algoritma *K-Means* untuk mengelompokkan data secara lebih padat. Dari hasil *elbow method*, didapatlah jumlah klaster yang optimal adalah tiga. Setelah mendapatkan jumlah klaster yang optimal, selanjutnya kita dapat melakukan klasterisasi menggunakan metode yang telah ditentukan, yakni *K-Means*. Hasil visualisasi klasterisasi yang ditunjukkan pada Gambar 2.



Gambar 2a. *Scatter Plot K-means* Sebelum PCA

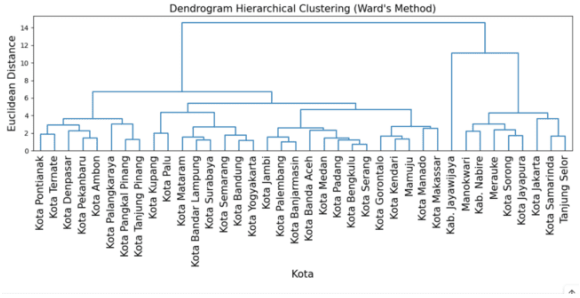


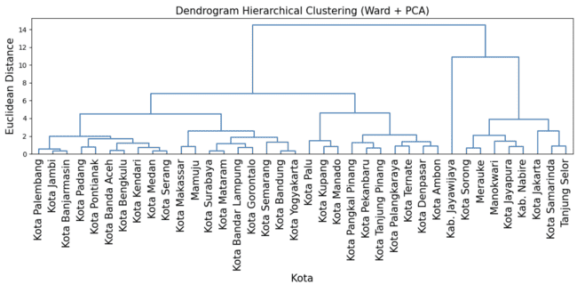
Gambar 2b. *Scatter Plot K-means* Setelah PCA

Dari kedua tersebut, tampak tidak ada perbedaan antara sebelum dan sesudah penerapan metode PCA pada algoritma *K-Means*. *Centroid* pada *K-Means clustering* dipilih berdasarkan titik acak murni hasil algoritma *K-Means* itu sendiri. Lalu, terdapat satu *centroid* yang terletak sangat jauh dibandingkan *centroid-centroid* lainnya, hal tersebut dikarenakan terdapat satu data pencilan dengan karakteristik yang paling berbeda daripada data-data lainnya. Oleh karena itu, terbentuklah sebuah *centroid* yang terletak sangat jauh daripada *centroid* lainnya.

Meskipun penerapan PCA pada *elbow method* sebelumnya untuk menentukan jumlah klaster yang optimal memberikan perbedaan, hal ini mencerminkan bahwasanya PCA memengaruhi representasi antar varians fitur, yang pada gilirannya berpengaruh terhadap evaluasi inersia dalam proses pembentukan klaster. PCA bekerja dengan cara mereduksi kompleksitas data tanpa menghilangkan varians utama data, sehingga memungkinkannya untuk menghasilkan struktur spasial yang berpotensi mengubah inersia antar jumlah klaster. Namun, stabilitas hasil klaster menunjukkan jika PCA tidak secara drastis mengubah struktur dasar klaster, meskipun persepsi nilai optimal klaster telah diubah. Hasil klasterisasi yang tidak memberikan perubahan bisa saja terjadi karena sifat *robust* dari *K-Means* itu sendiri yang menyebabkannya tidak terlalu terpengaruh meskipun data diputar, digeser, ataupun diskalakan secara proporsional [21]. Sebagai alternatif dari pembahasan visualisasi klasterisasi *K-Means*, kita dapat menggunakan klasterisasi hierarki untuk melihat hubungan antar datanya. Gambar 3 merupakan visualisasi hasil pemodelan data menggunakan metode *Ward*.

E-ISSN 2808-5841
P-ISSN 2808-7283



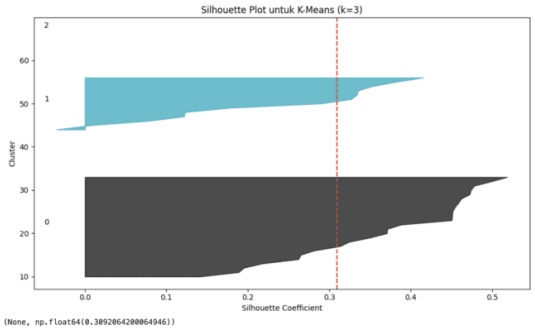


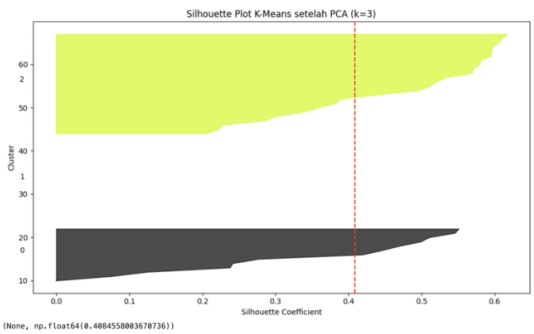
Gambar 3a. Dendrogram Metode *Ward* Sebelum PCA

Gambar 3b. Dendrogram Metode *Ward* Setelah PCA

Visualisasi dendogram di atas menunjukkan hasil pengelompokkan/penggabungan data berdasarkan kemiripan karakteristiknya. Semua label pada sumbu horizontal mewakili kota-kota di Indonesia. Sebelum penerapan PCA, tampak pemisahan klaster cenderung terjadi pada jarak *euclidean* yang cukup tinggi, menandakan varians yang masih tinggi. Sementara itu, setelah penerapan PCA, data berhasil dipisahkan saat jarak *euclidean* yang masih rendah. Hal ini menandakan PCA membantu menyederhanakan struktur data dengan menggunakan beberapa komponen utama (tidak semua variabel) tanpa mengurangi informasi penting yang ada pada data. Meskipun puncak dari kedua dendogram sama-sama berada di jarak sekitar 14, tetapi pengelompokkan data setelah penerapan PCA jauh lebih cepat dikarenakan varians yang telah disesuaikan oleh PCA itu sendiri.

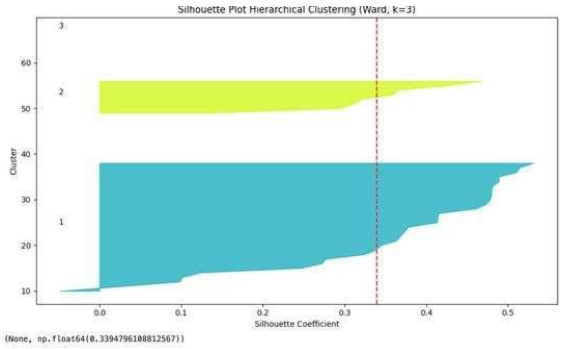
Untuk mengetahui tingkat akurasi metode-metode klasterisasi yang telah digunakan sebelumnya, kita bisa memanfaatkan *silhouette score* beserta menampilkan *plot*-nya untuk mengetahui apakah hasil klasterisasi dari metode-metode yang digunakan benar secara proporsional atau terdapat kesalahan dalam pengklasteran. Gambar 4a, 4b, 4c, dan 4d merupakan nilai *silhouette* beserta visualisasinya dalam bentuk *plot* perbandingan antara *silhouette score* sebelum dan sesudah PCA untuk kedua metode *clustering*.



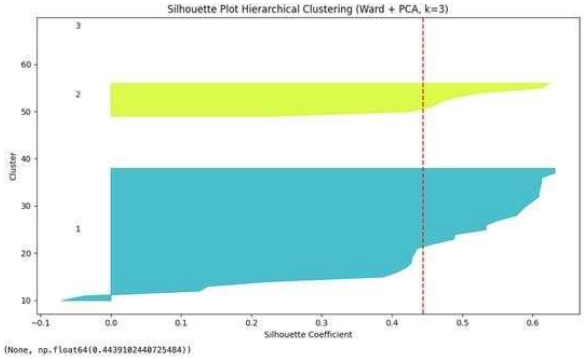


Gambar 4a. *Silhouette Plot K-means* Sebelum PCA

Gambar 4b. *Silhouette Plot K-means* Sebelum PCA



Gambar 4c. *Silhouette Plot* Metode *Ward* Sebelum PCA



Gambar 4d. *Silhouette Plot* Metode *Ward* Setelah PCA

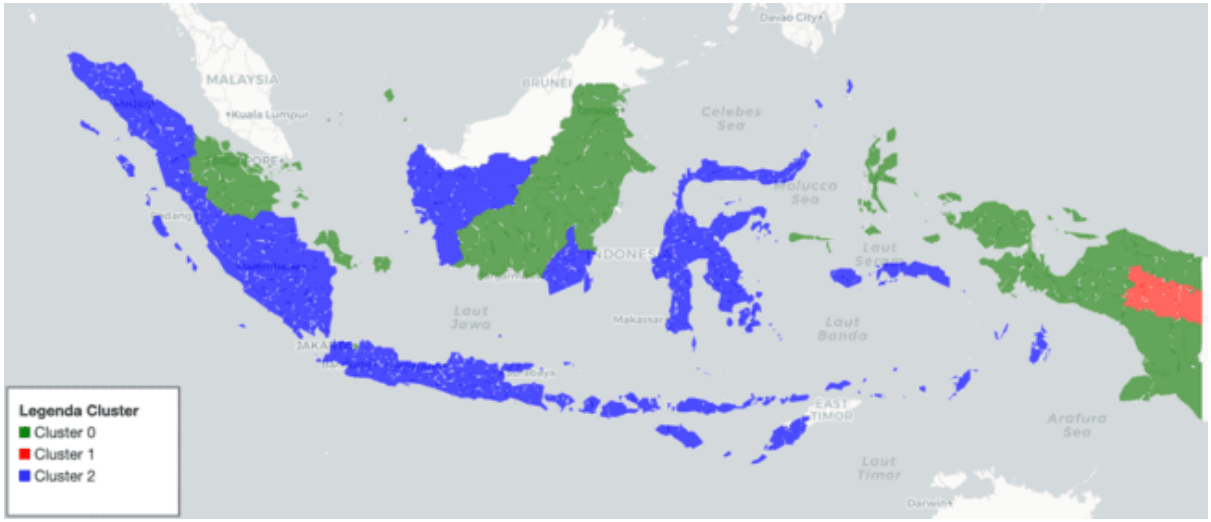
Untuk mengetahui kualitas hasil klasterisasi pada masing-masing metode, dilakukan evaluasi menggunakan *silhouette score* baik sebelum maupun setelah penerapan PCA. Nilai ini memberikan gambaran seberapa baik observasi cocok dengan klaster tempatnya tergabung dibandingkan dengan klaster lainnya. Tabel 1 merupakan hasil perbandingan *silhouette score* dari metode *K-Means* dan *Ward Clustering*.

Tabel 1. Tabel Perbandingan *Silhouette Score* Kedua Metode Clustering

Nama Metode	***Silhouette Score* Sebelum PCA**	***Silhouette Score* Setelah PCA**
K-Means	0.31	**0.41**
Ward	0.34	**0.44**

Berdasarkan hasil dari visualisasi *silhouette plot* dan tabel perbandingan di atas, terlihat adanya perbedaan yang signifikan antara hasil klasterisasi sebelum dan sesudah penerapan PCA pada masing-masing metode. Pada metode *K-Means clustering*, nilai *silhouette score* meningkat dari 0.31 sebelum PCA menjadi 0.41 setelah penerapan PCA. Visualisasi *silhouette plot* juga menunjukkan distribusi yang lebih seimbang setelah di PCA yang di mana tidak ada nilai koefisien negatif (outlier). Hal ini menandakan bahwa metode *K-Means clustering* memiliki sifat *robust* terhadap transformasi (reduksi) dimensi dan tetap menghasilkan klaster yang stabil.

Sebaliknya, pada metode *hierarchical clustering* dengan pendekatan *Ward*, penerapan PCA memberikan dampak yang lebih signifikan. Metode *Ward* sebelum penerapan PCA menghasilkan *silhouette score* sebesar 0.34, tetapi terdapat sejumlah observasi dengan nilai negatif. Setelah PCA diterapkan, sama halnya dengan metode *K-Means clustering*, *silhouette score* meningkat menjadi sebesar 0.44. Meskipun telah dilakukan praproses PCA, nilai observasi negatif masih terdapat di visualisasi *silhouette plot*. Hal ini menunjukkan bahwa metode *clustering* hierarki sangat sensitif terhadap struktur spasial data, dan penerapan PCA berperan penting dalam meningkatkan kualitas representasi jarak antar titik. Untuk visualisasi lebih jelasnya kami menggunakan persebaran peta di Indonesia pada Gambar 5.



Gambar 5. Visualisasi Peta *Clustering K-Means* Setelah PCA

Sebuah provinsi tentu memiliki taraf biaya hidup yang berbeda dengan provinsi lainnya. Seiring berjalannya waktu, dinamika biaya hidup dipengaruhi oleh perkembangan teknologi, upah minimum tiap daerah, maupun gaya hidup memengaruhi biaya hidup seseorang. Biaya hidup juga dipengaruhi oleh konsumsi rumah tangga, seperti barang elektronik. Selain itu, upah minimum provinsi (UMP) sering dijadikan indikator tak langsung dalam menentukan standar kebutuhan hidup layak [22]. Meskipun demikian, biaya hidup relatif ditentukan berdasarkan jumlah perbandingan antara pendapatan dan pengeluaran. Hal ini dikatakan lebih akurat karena dapat diukur sebagai tingkat kemampuan masyarakat untuk membeli barang dan jasa sebagaimana yang dijelaskan melalui kerangka *Purchasing Power Parity* (PPP) [23].

Berdasarkan visualisasi peta *clustering* di atas, tampak klaster 2 tersebar di seluruh Pulau Jawa, Bali, dan Sulawesi. Selain itu, juga menyebar di ¾ bagian Pulau Sumatra, sebagian Pulau Kalimantan, dan beberapa kepulauan kecil di sekitar Sulawesi dan Papua. Daerah pada klaster 2 ini cenderung padat penduduk, maju dari segi perekonomian, dan terhubung dengan infrastruktur nasional. Pada klaster 2, harga komoditas dan layanan esensialnya cenderung rendah, serta garis kemiskinannya relatif rendah. Selain itu, didukung dengan UMP yang tinggi, menandakan bahwa daerah yang terklaster dalam klaster 2 merupakan daerah dengan biaya hidup yang relatif rendah.

Klaster 0 tersebar di wilayah timur Indonesia seperti Papua, Maluku, dan Nusa Tenggara, serta beberapa wilayah di Pulau Sumatra, Kalimantan, dan Jawa. Wilayah yang tergolong pada klaster 0 merupakan daerah yang cenderung kurang padat dari segi kependudukan serta kurang dalam pembangunan dan ekonomi. Selain itu, harga komoditas dan layanan esensial pada daerah-daerah klaster 0 cenderung lebih tinggi daripada klaster 2, begitu juga dengan garis kemiskinannya. Namun, menariknya adalah Kota Jakarta yang merupakan ibukota dari Indonesia justru masuk ke dalam klaster 0. Setelah dianalisis lebih lanjut, ternyata terdapat perbandingan yang cukup kontras antara garis kemiskinan dan UMP di sana. Mengingat Jakarta merupakan ibukota dari Indonesia, menyebabkan biaya hidup di sana tergolong cukup tinggi, sekali pun didukung dengan UMP yang tinggi. Sementara itu, mayoritas daerah-daerah pada klaster 0 cenderung memiliki UMP yang lebih rendah daripada daerah-daerah pada klaster 2, menandakan bahwa klaster 0 merupakan daerah dengan biaya hidup menengah.

Pada klaster 1, hanya terdapat satu daerah di dalamnya. Hal ini mengindikasikan adanya *outlier* pada *dataset* karena yang memiliki karakteristik paling berbeda dari data lain. Selain itu, daerah yang tergolong pada klaster 1 ini menandakan wilayah dengan nilai ekstrim dalam variabel tertentu, seperti harga komoditas yang tinggi. Tanpa didukung UMP yang tinggi, hal ini berimbas pada daerah pada klaster 1 memiliki garis kemiskinan yang tinggi. Dengan demikian, daerah pada klaster 1 memiliki biaya hidup yang relatif tinggi.

IV. KESIMPULAN

Berdasarkan hasil analisis yang dilakukan menggunakan metode *K-Means* dan *hierarchical clustering* dengan pendekatan *Ward*, baik sebelum maupun sesudah penerapan PCA, diperoleh bahwa metode *Ward* setelah PCA memberikan performa yang terbaik. Hal ini dapat dibuktikan dengan *silhouette score* tertinggi sebesar 0.44, akan tetapi setelah dilakukan visualisasi *silhouette plot*, terdapat nilai observasi negatif *(*outlier). Sementara itu, pada metode *K-Means* menunjukkan hasil *silhouette score* yang lebih rendah sebesar 0.41 setelah penerapan PCA. Meskipun memiliki *silhouette score* yang lebih rendah, hasil visualisasi *silhouette plot* menunjukkan tidak adanya nilai observasi negatif (outlier) yang menandakan bahwa seluruh observasi relatif cocok dengan klasternya masing-masing.

Nilai observasi negatif yang terdapat pada metode *Ward* mengindikasikan bahwa terdapat sebagian data yang secara spasial lebih dekat ke klaster lain dibandingkan klaster tempatnya tergabung. Fenomena ini bisa terjadi karena algoritma *Ward* yang sangat bergantung pada jarak dan varian antar kelompok, serta adanya potensi struktur data yang *overlap* atau distribusi yang tidak sepenuhnya terpisah antar klaster. Dengan mempertimbangkan hal tersebut, metode *Ward* memiliki *silhouette score* yang lebih tinggi, meskipun *clustering* dengan metode *K-Means* memiliki distribusi yang lebih seimbang dan lebih aman untuk dasar pengambilan kebijakan segmentasi wilayah.

Sebagai saran untuk penelitian selanjutnya, analisis ini dapat diperluas dengan menambahkan variabel yang relevan seperti tingkat pengangguran atau tingkat pendidikan serta metode lain yang tergolong pada tingkat lanjut seperti DBSCAN. Dengan itu, hasil analisis yang didapatkan akan lebih lengkap Melalui pendekatan yang komprehensif, hasil analisis diharapkan dapat menjadi dasar penyusunan kebijakan publik yang lebih akurat, khususnya kebijakan berbasis wilayah yang mempertimbangkan kondisi sosial ekonomi secara menyeluruh.

V. DAFTAR PUSTAKA

[1] M. H. Zaman, D. Wahyuningsih, R. Yuwono, and Y. Nugroho, 'The Response of Farmer Welfares Amidst Food Prices Shock and Inflation in the Province of East Java', 2024, doi: 10.56472/25835238/IRJEMS-V3I12P129.

[2] K. Auliasari, F. Qurrotuna, M. Orisa, N. Nurina, and P. M. Mirenty, 'Analisis Fluktuasi Harga Pangan Antar Provinsi di Indonesia Menggunakan Pendekatan Data Mining dan Big Data', *Digital Transformation Technology*, vol. 4, no. 2, pp. 1184–1191, Jan. 2025, doi: 10.47709/digitech.v4i2.5217.

[3] R. Bahtiar and F. D. Raswatie, 'Analisis Fluktuasi Harga Pangan di Kota Bogor', *Indonesian Journal of Agriculture Resource and Environmental Economics*, vol. 1, no. 2, pp. 70–81, Oct. 2023, doi: 10.29244/ijaree.v1i2.42020.

[4] Y. Anis *et al.*, 'Unsupervised Clustering Untuk Pengolahan Data Kemiskinan di Jawa Tengah Dengan Menggunakan Metode Self-Organizing Maps', *Journal of Information System Research*, vol. 6, no. 2, p. 828, 2025, doi: 10.47065/josh.v6i2.6439.

[5] Aditya Novita, Iin Ernawati, and Nurul Chamidah, 'KLASTERISASI PROVINSI DI INDONESIA BERDASARKAN PRODUKTIVITAS KOMODITAS PANGAN MENGGUNAKAN ALGORITMA K-MEANS', 2022.

[6] A. Eka Putra Haryanto, M. Ulfa Yanuar, D. Statistika Bisnis, and F. Vokasi, 'Metode K-Means Clustering untuk Pengelompokan Kabupaten/Kota dalam Upaya Pengendalian Tingkat Inflasi di Pulau Jawa dan Sumatera K-Means Clustering Method for District/City Grouping in Effort to Control Inflation Rates in Java and Sumatera', pp. 29–42, 2022, doi: 10.21787/govstat.1.1.2022.29-42.

[7] N. Nur, M. Iqram, and N. Inayah, 'Perbandingan K-Means dan Hierarchical Clustering dalam Pengelompokan Daerah Beresiko Stunting', vol. 8, no. 2, p. 2023, 2023.

[8] Shindi Sheila May Wara, Andri Fauzan Adziima, Muhammad Nasrudin, and Alfan Rizaldy Pratama, *2024 IEEE 10th Information Technology International Seminar*. IEEE, 2024.

[9] S. P. Lestari, E. D. Supandi, and P. P. Rahayu, 'Pengklasteran Kabupaten/Kota di Jawa Tengah berdasarkan Tenaga Kesehatan dengan Menggunakan Metode Ward dan K-Means', *Jurnal Fourier*, vol. 7, no. 2, pp. 103–109, Oct. 2018, doi: 10.14421/fourier.2018.72.103-109.

[10] I. Bin Mohamad and D. Usman, "Standardization and its effects on K-means clustering algorithm," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, no. 17, pp. 3299–3303, 2013, doi: 10.19026/rjaset.6.3638.

[11] A. Azzahra and A. W. Wijayanto, "Comparison of Agglomerative Hierarchical and K-Means in Grouping Provinces Based on Maternal Health Services," *SISTEMASI*, vol. 11, no. 2, p. 481, May 2022, doi: 10.32520/stmsi.v11i2.1829.

[12] M. S. Premkumar and S. H. Ganesh, "A median based external initial centroid selection method for k-means clustering," in 2017 World Congress on Computing and Communication Technologies (WCCCT), Tiruchirappalli, India, 2017, pp. 143-146, doi: 10.1109/WCCCT.2016.42.

[13] T. Apriliana and E. Widodo, "Analisis Cluster Hierarki untuk Pengelompokan Provinsi di Indonesia berdasarkan Jumlah Base Transceiver Station dan Kekuatan Sinyal," 2023.

[14] N. I. Boyko and O. A. Tkachyk, "Hierarchical clustering algorithm for dendrogram construction and cluster counting," *Informatics and mathematical methods in simulation*, vol. 13, no. 1–2, pp. 5–15, Apr. 2023, doi: 10.15276/imms.v13.no1-2.5.

[15] D. Fitrianingsih and M. D. Kartikasari, "Penerapan K-Means Clustering dengan Metode Elbow untuk Mengelompokkan Kabupaten/Kota Berdasarkan Faktor-Faktor yang Mempengaruhi Indeks Pembangunan Manusia di Provinsi Jawa Barat," *Emerging Statistics and Data Science Journal*, vol. 2, no. 2, 2024.

[16] A. Az-Zahra and A. W. Wijayanto, "Tinjauan Kesejahteraan di Daerah Perbatasan Republik Indonesia Tahun 2021: Penerapan Analisis Klaster K-Means dan Hierarki," *Jurnal Sistem dan Teknologi Informasi (JustIN)*, vol. 12, no. 1, p. 55, Jan. 2024, doi: 10.26418/justin.v12i1.69040.

[17] N. A. Maori, "METODE ELBOW DALAM OPTIMASI JUMLAH CLUSTER PADA K-MEANS CLUSTERING," *Jurnal SIMETRIS*, vol. 14, 2023.

[18] C. Zhang, J. Ou, W. He, H. Huang, G. Cheng, and Y. Gu, "Optimisation research on K-means clustering algorithm based on principal component analysis and percentile improvement," in 2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 2024, pp. 148–153, doi: 10.1109/ICAICA63239.2024.10823007.

[19] R. H. Bhahari and K. Kusnawi, "Clustering Analysis of Socio-Economic Districts/Cities In East Java Province Using PCA And Hierarchical Clustering Methods," *sinkron*, vol. 8, no. 4, pp. 2242–2251, Oct. 2024, doi: 10.33395/sinkron.v8i4.14078.

[20] Y. Hasan, "Pengukuran Silhouette Score dan Davies-Bouldin Index pada Hasil Cluster K-Means dan Dbscan," 2024.

[21] L. A. García-Escudero and A. Mayo-Iscar, "Robust clustering based on trimming Statistical Learning and Exploratory Methods of the Data Sciences > Clus-tering and Classification Statistical and Graphical Methods of Data Analysis > Robust Methods," 2024, doi: 10.13039/501100011033/FEDER.

[22] Y. Adella, S. Program, S. E. Pembangunan, F. Ekonomi, and D. Bisnis, "PENGARUH UPAH MINIMUM TINGKAT PENGANGGURAN TERBUKA DAN JUMLAH PENDUDUK TERHADAP KEMISKINAN DI PROVINSI JAWA TENGAH."

[23] V. Bénard et al., "Tackling Europe's cost of living crisis," SSRN Electronic Journal, 2024, doi: 10.2139/ssrn.4723161.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih yang sebesar-besarnya kepada Tim SENADA 2025 yang telah meluangkan waktu dan tenaga dalam menyusun serta menyediakan *template* jurnal ini. Ketersediaan *template* yang rapi dan terstruktur sangat membantu dalam proses penulisan, penyusunan, serta penyuntingan artikel ilmiah ini. Dukungan dan kontribusi Tim SENADA dalam menyediakan sarana penulisan yang efektif dan efisien sangat kami hargai, dan diharapkan dapat terus memberikan manfaat bagi peneliti dan akademisi di masa mendatang.