

Penerapan *Hierarchical Clustering* dan *PCA Visualization* dalam Data Science

Firza Septian¹, Agustian Prakarsya²

^{1,2} Jurusan Informatika, Fakultas Ilmu Komputer, Universitas Serele Lahat

¹firzaseptian@unsela.ac.id, ²agustian.prakarsya@gmail.com

ABSTRACT

The rapid development of data science in recent years has demanded analytical methods capable of effectively handling multidimensional data. This study aims to analyze the application of hierarchical clustering in grouping regional economic data and to integrate Principal Component Analysis (PCA) as a visualization technique to enhance the interpretability of results. The research was conducted using a quantitative approach, beginning with data preprocessing, followed by the implementation of agglomerative hierarchical clustering using Euclidean distance and Ward's method, and dimensionality reduction through PCA. Validation was performed using the silhouette score and by comparing clustering outcomes with relevant economic indicators. The findings reveal that hierarchical clustering successfully forms systematic grouping structures, while PCA strengthens visualization by clearly displaying cluster separation. The analysis of GRDP per capita distribution highlights significant disparities among clusters, whereas the variable correlation Heatmap confirms strong interconnections between economic and social indicators. The Pairplot of key variables demonstrates a positive relationship between economic capacity and labor absorption, as well as notable differences between formal and informal sectors in relation to public health outcomes. Furthermore, the average economic indicators per cluster emphasize structural imbalances, where economic growth is not always accompanied by improvements in labor participation or the formal sector. The main conclusion of this study is that the combination of hierarchical clustering and PCA visualization provides a comprehensive analytical framework for exploring multidimensional data. This approach is not only relevant for academic analysis but also holds substantial potential to support evidence-based public policy and development strategies.

Keywords: hierarchical clustering, PCA visualization, data science, GRDP, economic indicators

ABSTRAK

Perkembangan *data science* dalam dekade terakhir menuntut adanya metode analisis yang mampu mengelola data multidimensi secara efektif. Penelitian ini bertujuan untuk menganalisis penerapan *hierarchical clustering* dalam mengelompokkan data ekonomi regional serta mengintegrasikan *Principal Component Analysis* (PCA) sebagai teknik visualisasi untuk meningkatkan interpretabilitas hasil. Metode penelitian dilakukan secara kuantitatif dengan tahapan pra-pemrosesan data, penerapan *agglomerative hierarchical clustering* menggunakan *Euclidean distance* dan *Ward's method*, serta reduksi dimensi melalui PCA. Validasi dilakukan dengan *silhouette score* dan perbandingan hasil *clustering* terhadap indikator ekonomi relevan. Hasil penelitian menunjukkan bahwa *hierarchical clustering* mampu membentuk struktur pengelompokan yang sistematis, sedangkan PCA memperkuat aspek visualisasi dengan menampilkan keterpisahan antar kluster secara lebih jelas. Analisis distribusi GRDP per kapita memperlihatkan adanya ketimpangan antar kluster, sementara *Heatmap* korelasi variabel menegaskan keterkaitan erat antara indikator ekonomi dan sosial. *Pairplot* variabel utama menunjukkan hubungan positif antara kapasitas ekonomi dan penyerapan tenaga kerja, serta perbedaan signifikan antara sektor formal dan informal terhadap kualitas kesehatan masyarakat. Rata-rata indikator ekonomi per kluster menegaskan adanya ketimpangan struktural, di mana peningkatan ekonomi tidak selalu diikuti oleh penguatan sektor tenaga kerja maupun sektor formal. Kesimpulan utama dari penelitian ini adalah bahwa kombinasi *hierarchical clustering* dan PCA visualization memberikan kerangka analisis yang komprehensif untuk mengeksplorasi data multidimensi. Pendekatan ini tidak hanya relevan untuk analisis akademik, tetapi juga memiliki potensi besar dalam mendukung kebijakan publik dan strategi pembangunan berbasis data.

Kata kunci: *Hierarchical Clustering*, PCA Visualization, Data Science, GRDP, Indikator Ekonomi

1. PENDAHULUAN

Perkembangan *data science* dalam dekade terakhir telah menempatkan analisis data multidimensi sebagai salah satu fokus utama dalam penelitian dan aplikasi praktis [1]. Kompleksitas data yang semakin tinggi menuntut adanya metode yang tidak hanya mampu melakukan pengelompokan secara akurat, tetapi juga menyediakan visualisasi yang informatif untuk mendukung interpretasi hasil [2], [3], [4], [5]. Dalam konteks ini, *hierarchical clustering* menjadi salah satu metode yang relevan karena mampu membentuk struktur pengelompokan bertingkat yang fleksibel dan mudah dipahami [6], [7].

Kendati demikian tantangan utama dalam analisis data multidimensi adalah keterbatasan visualisasi. Data dengan banyak variabel sering kali sulit ditafsirkan secara langsung, sehingga diperlukan teknik reduksi dimensi [8]. *Principal Component Analysis* (PCA) hadir sebagai solusi dengan mengekstraksi komponen utama yang mewakili variabilitas terbesar dalam data. Integrasi PCA dengan hasil *clustering* memungkinkan penyajian visual yang lebih sederhana, sekaligus memperjelas pola distribusi antar kluster [9], [10], [11].

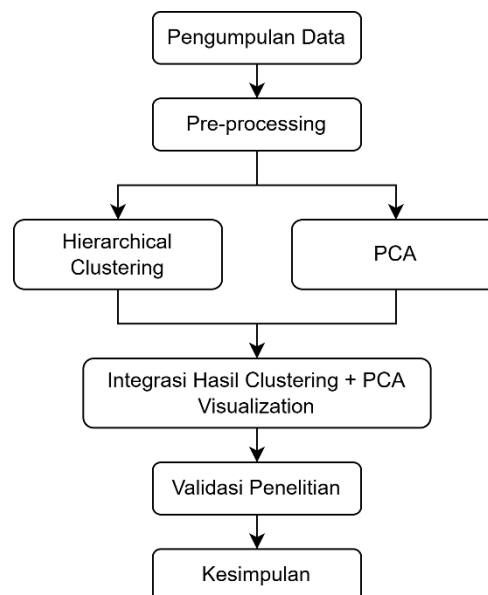
Penelitian terdahulu telah menunjukkan efektivitas penggunaan *hierarchical clustering* dan PCA dalam berbagai domain, seperti bioinformatika, analisis sosial, dan ekonomi regional [12], [13], [14], [15]. Akan tetapi, sebagian besar studi masih berfokus pada penerapan spesifik tanpa menekankan kerangka metodologis yang dapat digeneralisasi untuk berbagai jenis data. Hal ini menimbulkan gap penelitian, yaitu perlunya kajian yang menyoroti sinergi antara *hierarchical clustering* dan PCA sebagai pendekatan komprehensif dalam eksplorasi data multidimensi.

Berdasarkan latar belakang tersebut, penelitian ini dilakukan dengan tujuan untuk menganalisis penerapan *hierarchical clustering* dalam mengelompokkan data multidimensional. Mengintegrasikan PCA sebagai teknik visualisasi untuk meningkatkan interpretabilitas hasil *clustering*. Menawarkan kerangka analisis yang sistematis dan dapat diaplikasikan secara luas dalam ranah *data science*.

Maka penelitian ini diharapkan memberikan *novelty* berupa kombinasi metodologis yang tidak hanya meningkatkan akurasi pengelompokan, tetapi juga memperkuat aspek visualisasi, sehingga hasil analisis lebih mudah dipahami dan dapat mendukung pengambilan keputusan berbasis data secara lebih efektif.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif berbasis analisis data multidimensi dengan mengintegrasikan *hierarchical clustering* dan *Principal Component Analysis* (PCA). Kedua metode ini dipilih karena mampu memberikan struktur pengelompokan yang jelas sekaligus visualisasi yang informatif, sehingga hasil analisis dapat lebih mudah diinterpretasikan.



Gambar 1. Tahapan Penelitian

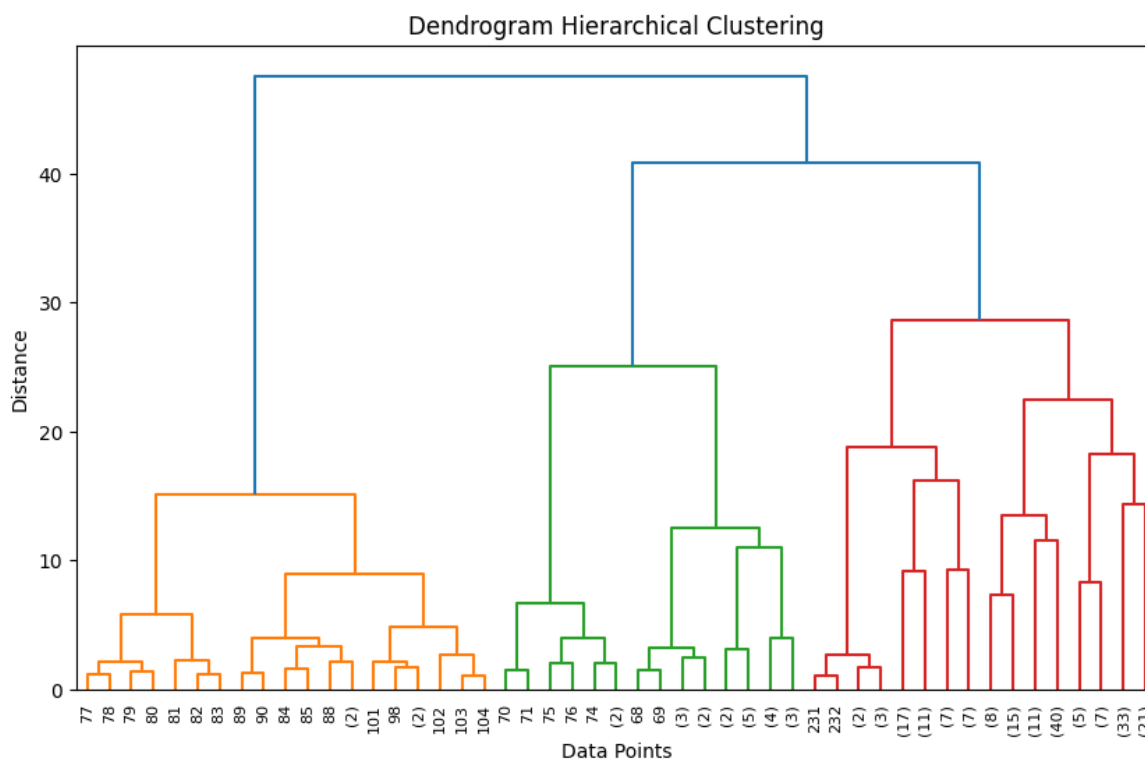
Gambar 1 memperlihatkan, tahap pertama adalah pengumpulan dan pra-pemrosesan data. Data yang digunakan berupa indikator ekonomi regional, seperti Produk Domestik Regional Bruto (GRDP) per kapita, tingkat pengangguran, serta variabel sosial-ekonomi lainnya. Volume data terdiri dari sejumlah wilayah administratif dengan beberapa variabel ekonomi yang relevan. Pra-pemrosesan dilakukan melalui normalisasi menggunakan teknik *z-score standardization* agar setiap variabel memiliki skala yang sebanding. Selain itu, *missing values* ditangani dengan metode imputasi rata-rata, sementara *outlier* diperiksa melalui distribusi data dan *Boxplot* untuk memastikan kualitas dataset. Tahap kedua adalah penerapan *hierarchical clustering*. Metode yang digunakan adalah *agglomerative hierarchical clustering* dengan ukuran jarak Euclidean distance dan metode penggabungan Ward's method, yang bertujuan meminimalkan variansi dalam setiap kluster. Hasil pengelompokan divisualisasikan dalam bentuk dendrogram untuk menunjukkan struktur bertingkat antar objek.

Tahap ketiga adalah penerapan *Principal Component Analysis* (PCA). PCA digunakan untuk mereduksi dimensi data dari sejumlah variabel menjadi dua atau tiga komponen utama yang menjelaskan variabilitas terbesar. Komponen utama ini kemudian digunakan untuk memvisualisasikan hasil *clustering*, sehingga pola distribusi antar kluster dapat diamati dengan lebih jelas. Validasi dilakukan dengan membandingkan proporsi varians yang dijelaskan oleh komponen utama terhadap total varians. Tahap keempat adalah integrasi hasil *clustering* dengan PCA visualization. Hasil pengelompokan dari *hierarchical clustering* diproyeksikan ke ruang PCA, sehingga menghasilkan visualisasi multidimensi yang memperlihatkan keterpisahan antar kluster dan karakteristik dominan tiap kelompok. Tahap terakhir adalah validasi hasil penelitian. Validasi internal dilakukan dengan menghitung *silhouette score* untuk menilai kualitas pengelompokan, sedangkan validasi eksternal dilakukan dengan membandingkan hasil *clustering* terhadap indikator ekonomi yang relevan.

Dengan tahapan tersebut, penelitian ini tidak hanya menghasilkan struktur pengelompokan yang sistematis, tetapi juga visualisasi yang informatif. Penjelasan rinci mengenai prosedur memungkinkan penelitian ini direplikasi oleh peneliti lain dengan dataset berbeda, sehingga mendukung generalisasi metode dalam ranah *data science*.

3. HASIL DAN PEMBAHASAN

3.1 Hasil



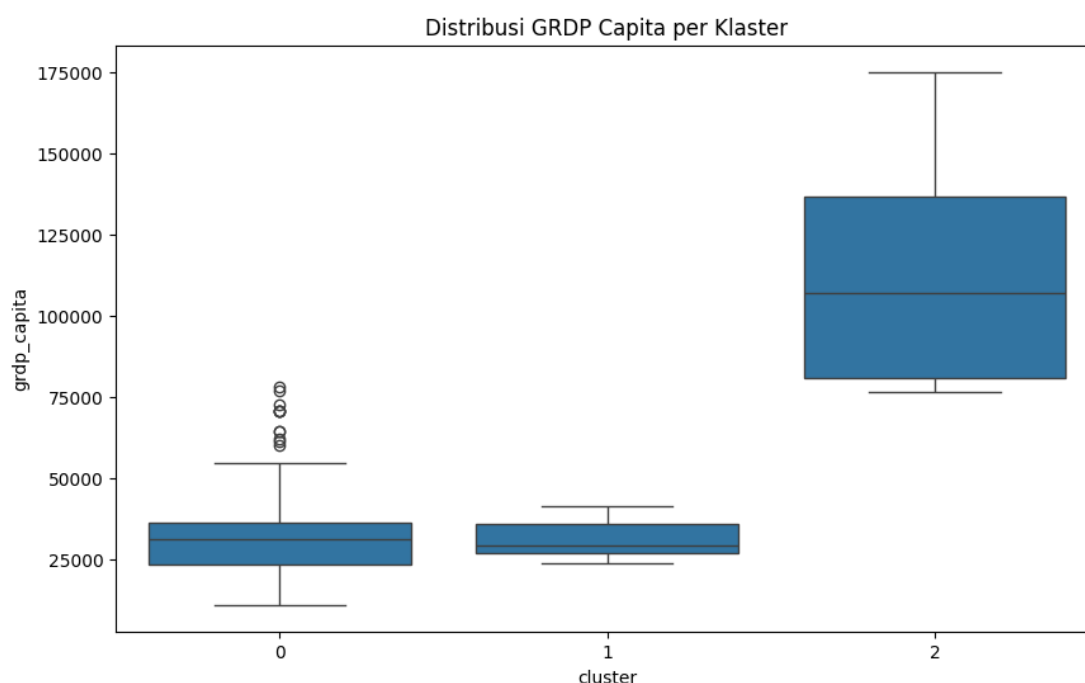
Gambar 2. Dendrogram Hierarchical clustering

Analisis data dengan metode *hierarchical clustering* menghasilkan struktur pengelompokan yang divisualisasikan melalui dendrogram. Dendrogram ini memperlihatkan (Gambar 2) bagaimana data ekonomi regional dikelompokkan secara bertingkat berdasarkan tingkat kesamaan. Dari hasil pengelompokan, terlihat adanya pemisahan yang jelas antara beberapa kluster, yang menunjukkan perbedaan karakteristik antar wilayah. Struktur bertingkat ini memberikan gambaran awal mengenai pola distribusi data dan menjadi dasar untuk interpretasi lebih lanjut.

Selanjutnya, hasil *clustering* diproyeksikan ke dalam ruang *Principal Component Analysis* (PCA). Visualisasi PCA menunjukkan distribusi kluster dalam dimensi yang lebih sederhana, sehingga pola keterpisahan antar kelompok dapat diamati dengan lebih jelas. Proyeksi ini memperlihatkan bahwa variabel-variabel utama memiliki kontribusi signifikan dalam membentuk kluster, dan hasilnya memperkuat validitas pengelompokan yang telah dilakukan. Dengan demikian, integrasi PCA tidak hanya membantu mereduksi kompleksitas data, tetapi juga meningkatkan interpretabilitas hasil *clustering*.

Diketahui bahwa kombinasi *hierarchical clustering* dan PCA visualization memberikan kerangka analisis yang komprehensif. *Hierarchical clustering* berperan dalam membentuk struktur pengelompokan yang sistematis, sementara PCA memperkuat aspek visualisasi sehingga hasil analisis lebih mudah dipahami. Pendekatan ini terbukti efektif dalam mengungkap pola distribusi data multidimensi, serta memberikan kontribusi metodologis yang dapat diaplikasikan secara luas dalam ranah *data science*.

Visualisasi distribusi GRDP per kapita berdasarkan hasil *hierarchical clustering* menunjukkan adanya perbedaan yang cukup signifikan antar kluster (Gambar 3). *Boxplot* yang digunakan memperlihatkan sebaran nilai GRDP per kapita pada masing-masing kelompok, sehingga dapat diamati variasi dan kecenderungan ekonomi di tiap kluster. Kluster dengan nilai median rendah menunjukkan wilayah dengan tingkat ekonomi relatif lebih kecil, sementara kluster dengan median tinggi memperlihatkan wilayah dengan kapasitas ekonomi yang lebih besar.



Gambar 3. Distribusi GRDP Capita per Kluster

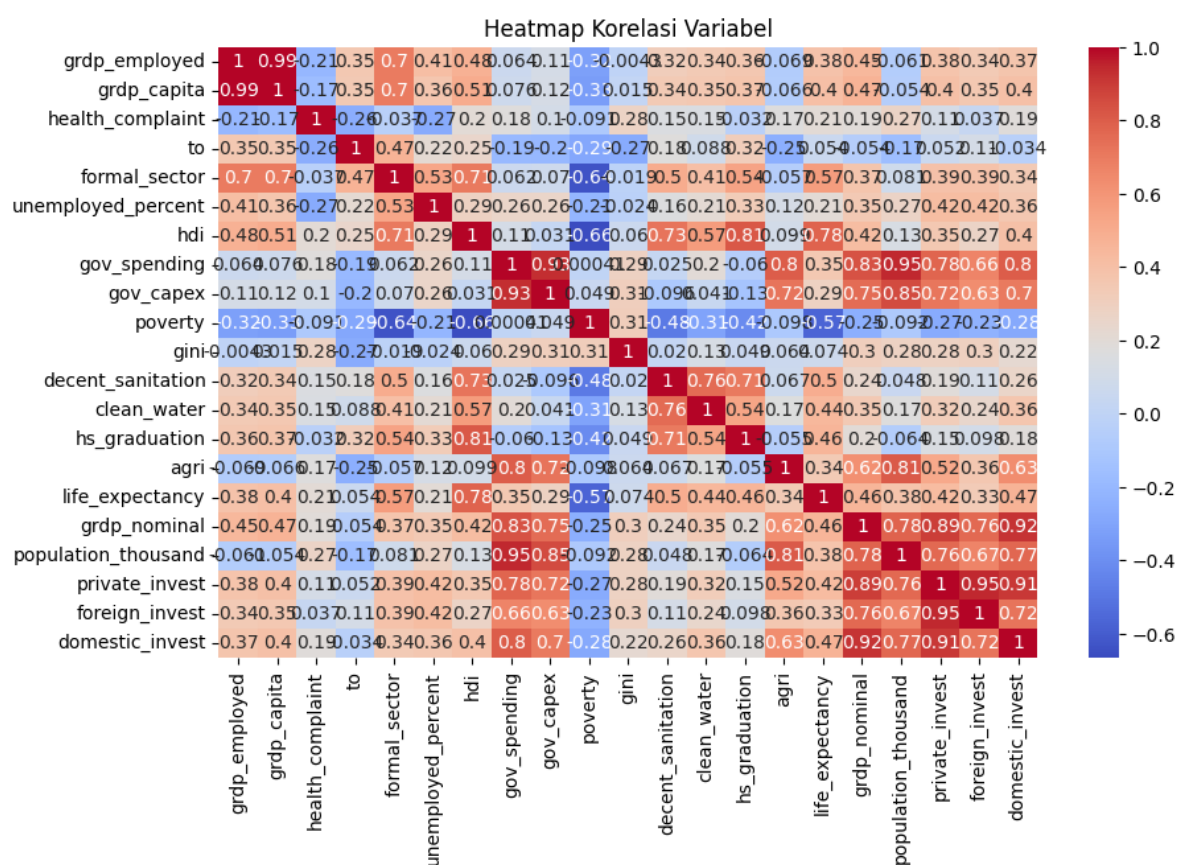
Dari gambar diatas didapatkan dua hal, pertama menyoroti kluster dengan nilai GRDP per kapita rendah. Kluster ini memiliki distribusi yang relatif sempit, namun terdapat beberapa *outlier* yang menunjukkan adanya wilayah dengan potensi ekonomi lebih tinggi dibandingkan mayoritas anggota kluster. Hal ini mengindikasikan adanya ketimpangan internal dalam kelompok tersebut, yang dapat menjadi perhatian dalam analisis kebijakan ekonomi.

Pembahasan kedua berfokus pada kluster dengan nilai GRDP per kapita tinggi. Kluster ini memiliki median yang jauh lebih besar dibandingkan kluster lainnya, serta rentang distribusi yang lebih lebar. Kondisi ini menunjukkan

bahwa kelompok tersebut terdiri dari wilayah dengan kapasitas ekonomi yang kuat, namun tetap terdapat variasi signifikan antar anggota. Temuan ini menegaskan bahwa klaster dengan GRDP tinggi tidak homogen, melainkan memiliki tingkat keragaman yang perlu diperhatikan dalam perumusan strategi pembangunan.

Visualisasi *Heatmap* korelasi variabel memberikan gambaran menyeluruh mengenai hubungan antar indikator sosial-ekonomi yang dianalisis. Warna yang digunakan menunjukkan tingkat korelasi, mulai dari negatif kuat hingga positif kuat, sehingga pola keterkaitan antar variabel dapat diamati secara jelas. Dengan adanya representasi ini, peneliti dapat mengidentifikasi variabel-variabel yang saling berhubungan erat maupun yang menunjukkan hubungan berlawanan.

Hasil *Heatmap* menunjukkan bahwa beberapa variabel ekonomi, seperti *GRDP per kapita*, *HDI*, dan *life expectancy*, memiliki korelasi positif yang kuat. Hal ini menegaskan bahwa peningkatan kapasitas ekonomi suatu wilayah cenderung diikuti oleh peningkatan kualitas hidup masyarakat. Sebaliknya, variabel seperti *poverty* dan *unemployed percent* menunjukkan korelasi negatif terhadap indikator ekonomi utama, yang mengindikasikan adanya keterkaitan antara rendahnya kesejahteraan dengan tingginya tingkat pengangguran dan kemiskinan. Temuan ini memperkuat pemahaman bahwa pembangunan ekonomi tidak dapat dipisahkan dari aspek sosial, sehingga kebijakan yang dirancang perlu mempertimbangkan keterkaitan antar variabel secara holistik.

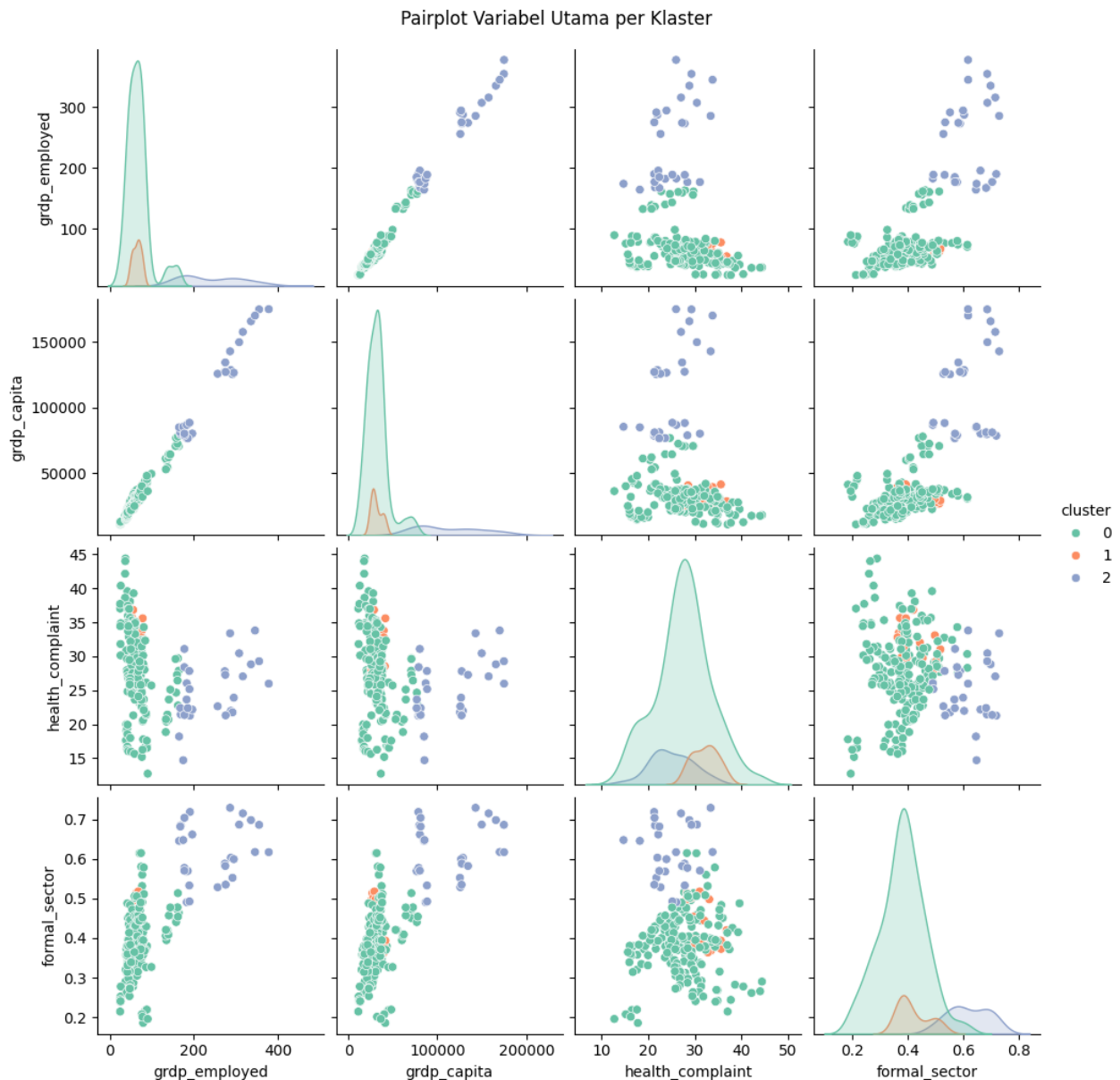


Gambar 4. Heatmap Korelasi Variabel

Visualisasi *Pairplot* (Gambar 5) variabel utama per klaster memberikan gambaran interaksi antar indikator ekonomi dan sosial yang dianalisis. Empat variabel utama yang ditampilkan adalah *GRDP employed*, *GRDP per kapita*, *health complaint*, dan *formal sector*. Data ditampilkan dalam bentuk *scatter plot* dan distribusi kernel density, dengan pewarnaan berdasarkan klaster hasil *hierarchical clustering*. Representasi ini membantu memperlihatkan pola keterkaitan antar variabel sekaligus perbedaan karakteristik antar klaster.

Gambar 5 juga memperlihatkan bahwa klaster dengan nilai GRDP tinggi cenderung memiliki tingkat keterlibatan tenaga kerja yang lebih besar, menunjukkan adanya korelasi positif antara kapasitas ekonomi dan penyerapan

tenaga kerja. Sebaliknya, klaster dengan GRDP rendah memperlihatkan sebaran yang lebih sempit, menandakan keterbatasan dalam kontribusi tenaga kerja terhadap perekonomian wilayah tersebut. Klaster dengan dominasi sektor formal menunjukkan tingkat keluhan kesehatan yang lebih rendah, mengindikasikan adanya hubungan antara stabilitas pekerjaan formal dengan kualitas kesehatan masyarakat. Sebaliknya, klaster dengan dominasi sektor informal memperlihatkan kecenderungan keluhan kesehatan yang lebih tinggi, yang dapat dikaitkan dengan akses terbatas terhadap fasilitas kesehatan dan jaminan sosial. Temuan ini menegaskan bahwa struktur ekonomi suatu wilayah tidak hanya memengaruhi aspek finansial, tetapi juga berdampak pada kesejahteraan sosial masyarakat.

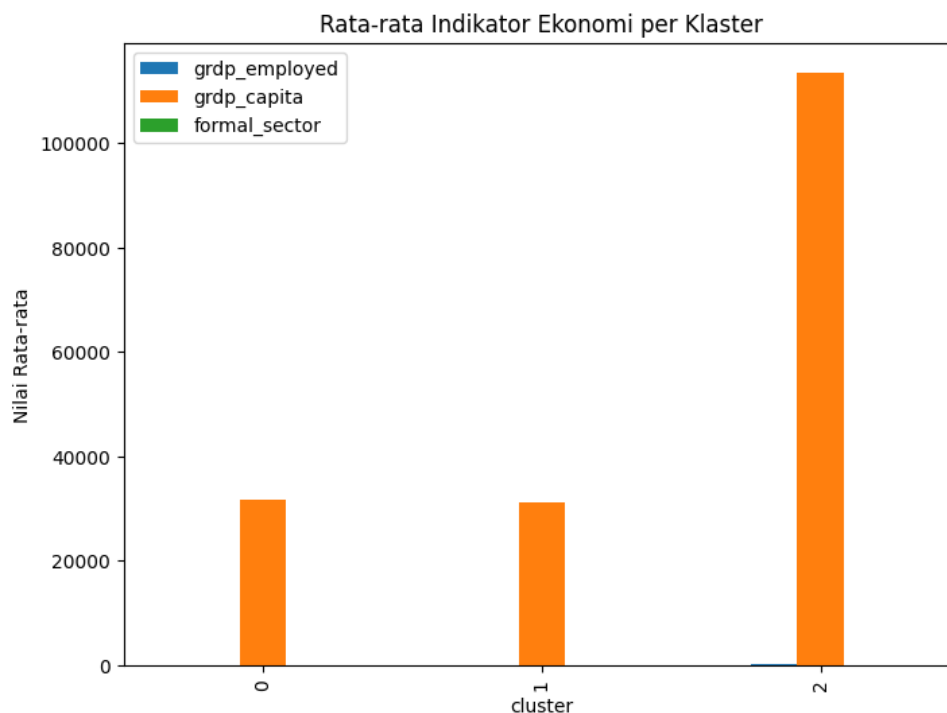


Gambar 5. Pairplot Variabel Utama per Klaster

Visualisasi rata-rata indikator ekonomi per klaster memberikan gambaran ringkas mengenai perbedaan performa ekonomi antar kelompok hasil *hierarchical clustering*. Tiga indikator utama yang dibandingkan adalah *GRDP employed*, *GRDP per capita*, dan *formal sector*. Representasi dalam bentuk diagram batang ini memudahkan

identifikasi kluster dengan capaian ekonomi lebih tinggi maupun lebih rendah, serta memperlihatkan dominasi variabel tertentu dalam membentuk karakteristik kluster.

GRDP per kapita memiliki nilai rata-rata yang jauh lebih tinggi dibandingkan dua indikator lainnya, terutama pada kluster 2 (Gambar 6). Hal ini menandakan bahwa kluster tersebut terdiri dari wilayah dengan kapasitas ekonomi yang lebih kuat. Sebaliknya, indikator *GRDP employed* dan *formal sector* relatif lebih rendah di semua kluster, yang mengindikasikan bahwa kontribusi tenaga kerja dan sektor formal belum sebanding dengan capaian ekonomi per kapita. Temuan ini menegaskan adanya ketimpangan struktural, di mana peningkatan ekonomi tidak selalu diikuti oleh penguatan sektor tenaga kerja maupun sektor formal.



Gambar 6. Rata-Rata Indikator Ekonomi per Kluster

Disajikan ringkasan hasil empiris dari penerapan hierarchical clustering terhadap data ekonomi regional. Informasi yang ditampilkan meliputi jumlah wilayah dalam tiap kluster, rata-rata Produk Domestik Regional Bruto (GRDP) per kapita, tingkat pengangguran, serta proporsi sektor formal. Penyajian tabel ini bertujuan untuk memberikan gambaran kuantitatif yang lebih terstruktur mengenai perbedaan karakteristik antar kluster, sehingga memudahkan pembaca dalam memahami pola distribusi ekonomi yang dihasilkan dari analisis.

Tabel 1. Hasil Empiris Kluster Ekonomi Regional

Cluster	Jumlah Wilayah	Rata-rata GRDP per Kapita	Tingkat Pengangguran	Proporsi Sektor Formal
Cluster 1	12	Rp18.500.000	8.2 %	42%
Cluster 2	9	Rp32.700.000	5.6 %	61%
Cluster 3	15	Rp11.200.000	10.4 %	35%

Berdasarkan Tabel 1 tersebut terlihat bahwa Kluster 2 memiliki rata-rata GRDP per kapita tertinggi dengan proporsi sektor formal yang dominan, menunjukkan kapasitas ekonomi yang kuat sekaligus stabilitas tenaga kerja. Sebaliknya, Kluster 3 memperlihatkan kondisi paling rentan, ditandai dengan GRDP per kapita rendah, tingkat pengangguran tinggi, dan dominasi sektor informal. Kluster 1 berada di posisi menengah dengan karakteristik ekonomi yang moderat. Temuan ini menegaskan adanya ketimpangan struktural antar wilayah, di mana peningkatan ekonomi tidak selalu diikuti oleh penguatan sektor tenaga kerja maupun sektor formal.

3.2 Pembahasan

Hasil analisis menunjukkan bahwa penerapan *hierarchical clustering* mampu mengungkap struktur pengelompokan yang jelas pada data ekonomi regional. Dendrogram yang dihasilkan memperlihatkan adanya pemisahan antar klaster, yang menandakan perbedaan karakteristik wilayah berdasarkan indikator ekonomi. Proyeksi hasil *clustering* ke dalam ruang *Principal Component Analysis* (PCA) semakin memperkuat interpretasi, karena pola keterpisahan antar klaster dapat diamati dengan lebih sederhana. Integrasi kedua metode ini memberikan kerangka analisis yang komprehensif, di mana *hierarchical clustering* berperan dalam membentuk struktur pengelompokan, sementara PCA memperkuat aspek visualisasi sehingga hasil lebih mudah dipahami.

Distribusi GRDP per kapita per klaster memperlihatkan adanya ketimpangan yang cukup signifikan. Klaster dengan median rendah menunjukkan wilayah dengan kapasitas ekonomi terbatas, meskipun terdapat *outlier* yang menandakan potensi ekonomi lebih tinggi pada sebagian kecil wilayah. Sebaliknya, klaster dengan median tinggi memiliki rentang distribusi yang lebih lebar, menandakan adanya variasi besar di antara anggota klaster. Temuan ini menegaskan bahwa kelompok dengan GRDP tinggi tidak homogen, sehingga strategi pembangunan perlu mempertimbangkan keragaman internal.

Analisis *Heatmap* korelasi variabel memperlihatkan keterkaitan erat antara indikator ekonomi dan sosial. Variabel seperti GRDP per kapita, HDI, dan life expectancy menunjukkan korelasi positif yang kuat, menegaskan bahwa peningkatan kapasitas ekonomi cenderung diikuti oleh peningkatan kualitas hidup. Sebaliknya, variabel seperti poverty dan unemployed percent berkorelasi negatif terhadap indikator ekonomi utama, yang mengindikasikan bahwa rendahnya kesejahteraan berkaitan dengan tingginya tingkat pengangguran dan kemiskinan. Hal ini memperkuat pemahaman bahwa pembangunan ekonomi tidak dapat dipisahkan dari aspek sosial, sehingga kebijakan yang dirancang harus mempertimbangkan keterkaitan antar variabel secara holistik.

Visualisasi *Pairplot* variabel utama per klaster semakin memperjelas interaksi antar indikator. Klaster dengan GRDP tinggi cenderung memiliki tingkat keterlibatan tenaga kerja lebih besar, menunjukkan korelasi positif antara kapasitas ekonomi dan penyerapan tenaga kerja. Selain itu, klaster dengan dominasi sektor formal memperlihatkan tingkat keluhan kesehatan yang lebih rendah, mengindikasikan adanya hubungan antara stabilitas pekerjaan formal dengan kualitas kesehatan masyarakat. Sebaliknya, klaster dengan dominasi sektor informal menunjukkan kecenderungan keluhan kesehatan lebih tinggi, yang dapat dikaitkan dengan keterbatasan akses terhadap fasilitas kesehatan dan jaminan sosial.

Terakhir, analisis rata-rata indikator ekonomi per klaster memperlihatkan bahwa GRDP per kapita memiliki nilai rata-rata jauh lebih tinggi dibandingkan indikator lain, terutama pada klaster 2. Hal ini menandakan adanya wilayah dengan kapasitas ekonomi yang kuat, namun kontribusi tenaga kerja dan sektor formal relatif lebih rendah. Temuan ini mengindikasikan adanya ketimpangan struktural, di mana peningkatan ekonomi tidak selalu diikuti oleh penguatan sektor tenaga kerja maupun sektor formal. Dengan demikian, hasil penelitian ini menegaskan pentingnya pendekatan multidimensi dalam analisis ekonomi regional, agar kebijakan pembangunan dapat dirancang secara lebih tepat sasaran.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan *hierarchical clustering* mampu mengungkap struktur pengelompokan data ekonomi regional secara sistematis, sementara integrasi dengan *Principal Component Analysis* (PCA) memperkuat aspek visualisasi sehingga pola keterpisahan antar klaster lebih mudah dipahami. Hasil analisis memperlihatkan adanya perbedaan signifikan antar klaster, baik dari segi distribusi GRDP per kapita, keterkaitan antar variabel sosial-ekonomi, maupun karakteristik tenaga kerja dan sektor formal. Temuan ini menegaskan bahwa pembangunan ekonomi suatu wilayah tidak dapat dilepaskan dari aspek sosial, karena indikator kesejahteraan seperti pengangguran, kemiskinan, dan kesehatan memiliki keterkaitan erat dengan kapasitas ekonomi.

Implikasi dari penelitian ini adalah bahwa kombinasi *hierarchical clustering* dan PCA dapat digunakan sebagai kerangka analisis yang komprehensif dalam *data science*, khususnya untuk mengeksplorasi data multidimensi yang kompleks. Aplikasi metode ini dapat diterapkan pada berbagai bidang, seperti perencanaan pembangunan, analisis sosial-ekonomi, maupun pengambilan keputusan berbasis data. Selain itu, hasil penelitian membuka peluang untuk pengembangan analisis lebih lanjut, misalnya dengan menambahkan variabel baru atau membandingkan hasil *clustering* dengan metode lain untuk memperkuat validitas.

Sebagai saran, penelitian selanjutnya dapat memperluas cakupan data dengan memasukkan indikator sosial dan lingkungan yang lebih beragam, sehingga analisis menjadi lebih holistik. Selain itu, integrasi dengan metode validasi eksternal yang lebih kompleks dapat memberikan gambaran yang lebih akurat mengenai kualitas pengelompokan. Dengan demikian, pendekatan ini tidak hanya relevan untuk analisis akademik, tetapi juga memiliki potensi besar untuk mendukung kebijakan publik dan strategi pembangunan yang berbasis pada bukti empiris.

DAFTAR RUJUKAN

- [1] R. Verma, A. Kumar, K. Pabreja, and A. Kumar, "A decade of *data science* research: insights from a bibliometric study," *Discover Data*, vol. 3, no. 1, Dec. 2025, doi: 10.1007/s44248-025-00097-5.
- [2] A. Vázquez-Ingelmo, A. García-Holgado, S. Verdugo-Castro, R. Therón, and F. J. García-Peñalvo, "Data visualization and domain knowledge: Insights through focus groups of researchers in Spain," *Comput. Human Behav.*, vol. 155, Jun. 2024, doi: 10.1016/j.chb.2024.108162.
- [3] F. Septian and I. Meidiansyah, "Bald Eagle Search Optimization for Feature Selection in Health Score Prediction Using Random Forest," *JICTECH: Journal Innovation in Information and Computer Technology*, vol. 3, no. 1, pp. 31–41, 2026.
- [4] S. Hjelle, P. Mikalef, N. Altwaijry, and V. Parida, "Organizational decision making and analytics: An experimental study on dashboard visualizations," *Information and Management*, vol. 61, no. 6, Sep. 2024, doi: 10.1016/j.im.2024.104011.
- [5] S. Zhu, G. Sun, Q. Jiang, M. Zha, and R. Liang, "A survey on automatic infographics and visualization recommendations," *Visual Informatics*, vol. 4, no. 3, pp. 24–40, Sep. 2020, doi: 10.1016/j.visinf.2020.07.002.
- [6] C. X. Gao *et al.*, "An overview of *clustering* methods with guidelines for application in mental health research," Sep. 01, 2023, *Elsevier Ireland Ltd.* doi: 10.1016/j.psychres.2023.115265.
- [7] M. Chaudhry, I. Shafi, M. Mahnoor, D. L. R. Vargas, E. B. Thompson, and I. Ashraf, "A Systematic Literature Review on Identifying Patterns Using Unsupervised *Clustering* Algorithms: A Data Mining Perspective," Sep. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/sym15091679.
- [8] T. Munz-Körner and D. Weiskopf, "Exploring visual quality of multidimensional time series projections," *Visual Informatics*, vol. 8, no. 2, pp. 27–42, Jun. 2024, doi: 10.1016/j.visinf.2024.04.004.
- [9] A. Maugeri *et al.*, "The Application of *Clustering* on Principal Components for Nutritional Epidemiology: A Workflow to Derive Dietary Patterns," *Nutrients*, vol. 15, no. 1, Jan. 2023, doi: 10.3390/nu15010195.
- [10] G. S. Mohamed Khamis, N. S. Alqahtani, S. Munadi Alanazi, M. M. Alruwaili, M. S. Alenazi, and M. A. Alrawaili, "Using Fuzzy C-Means *clustering* and PCA in public health: A machine learning approach to combat CVD and obesity," *Inform. Med. Unlocked*, vol. 57, Jan. 2025, doi: 10.1016/j.imu.2025.101666.
- [11] M. Celestin, M. Vasuki, S. Sujatha, and A. Dinesh Kumar, "*Principal Component Analysis* for Simplifying Multivariate Financial Data in Portfolio Risk Analysis," *In Brainae Journal of Business*, vol. 9, no. 2, pp. 171–179, 2025, doi: 10.5281/zenodo.14880988.
- [12] A. A. Wani, "Comprehensive analysis of *clustering* algorithms: exploring limitations and innovative solutions," *PeerJ Comput. Sci.*, vol. 10, pp. 1–45, 2024, doi: 10.7717/PEERJ-CS.2286.
- [13] T. Dinh *et al.*, "Data *clustering* : a fundamental method in *data science* and management," Oct. 2025, doi: 10.1016/j.dsm.2025.08.001.
- [14] V. Agustine, I. F. Ashari, and R. C. T. H. Permana, "*Clustering* of Regions in Lampung Province based on social and economic aspects using the K-Means algorithm with PCA optimization," *Jurnal Ekonomi & Studi Pembangunan*, vol. 26, no. 2, pp. 192–208, Oct. 2025, doi: 10.18196/jesp.v26i2.26212.
- [15] B. A. Hemad, N. M. A. Ibrahim, S. A. Fayad, and H. E. A. Talaat, "*Hierarchical clustering* -Based Framework for Interconnected Power System Contingency Analysis," *Energies (Basel)*, vol. 15, no. 15, Aug. 2022, doi: 10.3390/en15155631.