KOMPUTA : Jurnal Ilmiah Komputer dan Informatika
Vol. 14, No. 1, April 2025, ISSN : 2089-9033, e- ISSN : 2715-7849
DOI: 10.34010/komputa.v14i1. 15249

Klasterisasi Negara Dunia Berdasarkan Data Sosioekonomi dan Demografi Tahun 2023 dengan PCA dan K-Means

Dhea Romantika Marpaung[1], Erin Gunawan[2], Farrell Rio Fa[3] Albert Christianto[4]

[1,2,3,4] Program Studi Teknik Informatika, Universitas Mikroskil
E-mail : dhea.r.marpaung@gmail.com[1]

Abstrak

Perkembangan sosial, ekonomi, dan demografi merupakan indikator penting untuk menilai kemajuan suatu negara. Faktor-faktor ini mencerminkan kondisi kehidupan masyarakat, kualitas ekonomi, dan dinamika populasi yang dapat mempengaruhi kebijakan dan perencanaan pembangunan. Oleh karena itu, untuk memahami kondisi negara secara lebih mendalam, penting untuk mengelompokkan negara-negara berdasarkan karakteristik serupa dalam berbagai aspek tersebut. Tujuan penelitian ini adalah untuk mengidentifikasi klaster negara-negara di dunia berdasarkan analisis data sosio-ekonomi dan demografi tahun 2023 menggunakan metode *Principal Component Analysis* (PCA) dan *K-Means Clustering*. Analisis ini dilakukan dengan memeriksa hubungan antara GDP, tingkat kelahiran, tingkat kematian, populasi, dan emisi CO2. Hasil analisis mengungkapkan terdapat tiga klaster dengan *silhouette score*. *Cluster* 0 menunjukkan GDP yang tinggi dengan angka kematian bayi rendah dan emisi CO2 yang terkendali. *Cluster* 1 menunjukkan GDP yang lebih rendah, angka kematian bayi tinggi, dan tantangan dalam sektor kesehatan dan ekonomi. *Cluster* 2, yang mencakup negara-negara seperti China, India, dan AS, memiliki GDP tinggi tetapi menghadapi masalah emisi CO2 yang tinggi. Hasil ini mengindikasikan perlunya kebijakan yang terintegrasi untuk meningkatkan kesejahteraan global dengan memperhatikan faktor ekonomi, kesehatan, dan lingkungan secara berkelanjutan.
Kata kunci : Klasterisasi negara, Analisis komponen utama, K-Means

World Country Clustering Based on Socioeconomic and Demographic Data of 2023 Using PCA and K-Means

Abstract

The development of social, economic, and demographic factors is an important indicator for assessing the progress of a country. These factors reflect the quality of life, economic conditions, and population dynamics that can influence policies and development planning. Therefore, to better understand a country's conditions, it is important to cluster countries based on similar characteristics in these various aspects. The purpose of this study is to identify clusters of countries worldwide based on the analysis of socio-economic and demographic data for 2023 using Principal Component Analysis (PCA) and K-Means Clustering methods. This analysis examines the relationship between GDP, birth rate, death rate, population, and CO2 emissions. The results reveal three clusters with distinct characteristics. Cluster 0 shows high GDP with low infant mortality and controlled CO2 emissions. Cluster 1 shows lower GDP, high infant mortality, and challenges in the health and economic sectors. Cluster 2, which includes countries like China, India, and the US, has high GDP but faces high CO2 emission issues. These findings indicate the need for integrated policies to improve global well-being by considering economic, health, and environmental factors in a sustainable manner.
Keywords : *Country clustering, Principal component analysis, K-Means*

1. Pendahuluan

Dalam era globalisasi, perkembangan sosial, ekonomi, dan demografi merupakan aspek penting yang mencerminkan kemajuan suatu negara. Untuk memahami dinamika ini, analisis terhadap indikator sosio-ekonomi dan demografi, seperti PDB, tingkat pengangguran, tingkat kelahiran, dan distribusi populasi, menjadi krusial [1]. Klasterisasi adalah proses mengelompokkan titik data berdasarkan kesamaan yang ada di antara mereka, dengan tujuan untuk mengidentifikasi dan mengecualikan data yang tidak biasa atau

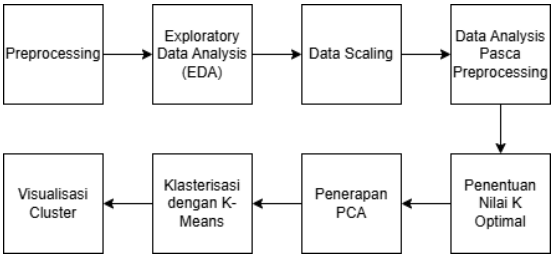
Naskah dikirim: 16-01-2025; direvisi 16-03-2025; diterima: 19-03-2025

anomali. Metode ini penting untuk memastikan bahwa setiap kelompok yang terbentuk lebih representatif dan bebas dari anomali. Algoritma klasterisasi mengandalkan hubungan dan parameter yang ditentukan oleh pengguna, yang berfungsi sebagai panduan dalam menentukan cara data dikelompokkan. Dengan cara ini, klasterisasi membantu mengungkap pola dan struktur tersembunyi dalam data yang kompleks [2]. Melalui pendekatan klasterisasi, negara-negara dapat dikelompokkan berdasarkan karakteristik yang serupa, sehingga pola tersembunyi dalam data dapat diungkap dan perbandingan antar negara dapat dilakukan dengan lebih jelas. Sosio-ekonomi, yang mempelajari hubungan antara perilaku sosial dan kegiatan ekonomi dalam masyarakat, membantu memahami bagaimana aspek-aspek seperti pendapatan dan pekerjaan memengaruhi dinamika negara-negara dalam klaster tersebut [3]. Selain itu, analisis demografi, yang berfokus pada karakteristik populasi seperti usia, tingkat kelahiran, dan harapan hidup, memberikan wawasan lebih dalam tentang struktur sosial dalam suatu negara. Dengan mengaitkan data demografi dengan indikator sosio-ekonomi, klasterisasi memungkinkan identifikasi negara-negara dengan pola perkembangan populasi yang serupa, sehingga dapat dipahami lebih baik faktor-faktor yang mempengaruhi perbedaan atau kemiripan antar negara [4]**.**

Sejalan dengan ini, pada penelitian sebelumnya, algoritma K-Means digunakan untuk mengelompokkan wilayah berdasarkan jumlah kasus Covid-19 di beberapa daerah di Indonesia, seperti yang terlihat dalam studi tentang prioritas vaksin di Provinsi Sumatera Utara. Di sini, K-Means *clustering* digunakan untuk mengelompokkan area berdasarkan tingkat infeksi, kematian, dan pemulihan. Hasil analisis menunjukkan kluster dengan kategori C1 (Tinggi), C2 (Sedang), dan C3 (Rendah), memberikan gambaran yang lebih jelas tentang distribusi kasus dan membantu mengidentifikasi area prioritas untuk vaksinasi [5]**.** Algoritma K-Means juga diterapkan pada klasterisasi ketimpangan pembangunan pada 41 daerah di Jawa Barat[6]. Dengan menggunakan pendekatan serupa, penelitian ini bertujuan untuk mengidentifikasi klaster negara-negara di dunia berdasarkan data *Global Country Information* 2023 pada Google Datasets, yaitu data sosio-ekonomi dan demografi tahun 2023 yang berisi 195 baris (195 negara) dan 35 fitur. Penelitian dilakukan menggunakan metode PCA, yang merupakan alat standar dalam analisis data modern dan digunakan oleh hampir semua disiplin ilmu. PCA memiliki banyak aplikasi, seperti pengurangan dimensi, kompresi data, ekstraksi fitur, dan visualisasi data [7], dan K-Means *clustering* yang merupakan metode pengelompokan unsupervised learning yang klasik, berdasarkan jarak. Metode ini membagi sampel dengan nilai eigen yang mirip untuk membentuk beberapa *cluster*. Jarak antara dua objek dianggap dekat, sehingga tingkat kesamaannya lebih tinggi. Algoritma ini secara acak memilih K objek data dari N objek data sebagai pusat pengelompokan awal, kemudian menghitung jarak antara objek data yang tersisa dan K pusat pengelompokan. Objek data yang tersisa kemudian dikelompokkan ke dalam *cluster* yang sesuai berdasarkan jarak terkecil. Selanjutnya, rata-rata pasangan dari semua objek data dalam pengelompokan sampel baru dihitung dan dianggap sebagai pusat pengelompokan baru. Proses ini diulang hingga jumlah deviasi kuadrat rata-rata diminimalkan [8]. Dengan menggunakan data terbaru tahun 2023, penelitian ini bertujuan untuk mengidentifikasi klaster negara-negara di dunia berdasarkan indikator sosio-ekonomi dan demografi. Hasil dari analisis ini diharapkan mampu memberikan gambaran yang jelas mengenai pola kemiripan antar negara, mengidentifikasi kelompok negara dengan karakteristik serupa, serta menjadi dasar dalam merumuskan strategi yang tepat. Selain itu, wawasan yang diperoleh dapat menjadi acuan bagi para pembuat kebijakan, peneliti, dan organisasi internasional dalam menangani tantangan global serta meningkatkan kesejahteraan masyarakat secara berkelanjutan.

2. Metodologi

Penelitian ini dilakukan melalui serangkaian tahapan yang terstruktur untuk memastikan proses klasterisasi data menghasilkan hasil yang representatif. Gambar 1 menunjukkan tahapan-tahapan yang dilakukan untuk melakukan klasterisasi.



Gambar 1. Tahapan klaterisasi.

a. Preprocessing

Tahapan pertama adalah tahap *preprocessing* untuk meningkatkan kualitas data. Tahap *preprocessing* data merupakan tahap penting dalam analisis data yang bertujuan untuk membersihkan, mengubah format, dan menyiapkan data agar lebih mudah dan tepat dalam analisis [9]. Proses yang dilakukan mencakup pembersihan data (*data cleaning*): menghapus fitur yang mengandung lebih dari 10 baris data kosong, menghapus baris data kosong, menghapus fitur yang tidak relevan. Selain itu, atribut kategorikal diubah menjadi numerik melalui proses encoding agar dapat dimengerti oleh model *machine learning*.

b. Exploratory Data Analysis (EDA)

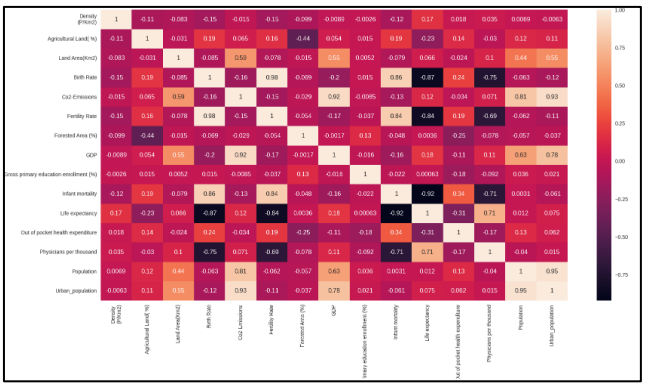
Tahapan Exploratory Data Analysis (EDA) dilakukan dengan mengidentifikasi pola dan memperoleh wawasan dari data. Proses ini biasanya melibatkan penyajian ringkasan karakteristik utama dari data melalui visualisasi. EDA adalah tahapan yang sangat penting untuk memahami data secara mendalam sebelum menerapkan model machine learning [10]. Eksplorasi data dilakukan guna memahami karakteristik data yang akan digunakan. Proses ini mencakup analisis statistik dan visualisasi untuk mengidentifikasi pola, distribusi variabel, serta potensi ketidakseimbangan kelas dalam data.

c. Data Scaling

Selanjutnya, dilakukan *data scaling*, yaitu proses penyesuaian nilai-nilai fitur dalam dataset agar berada dalam skala yang seragam. Tujuan dari *data scaling* adalah untuk mengatasi isu yang timbul akibat perbedaan skala di antara fitur-fitur tersebut. Tanpa melakukan data scaling, algoritma pembelajaran mesin mungkin mengalami kesulitan dalam menemukan solusi optimal, karena fitur dengan skala yang berbeda akan memberikan bobot yang tidak seimbang dalam model [11]. Pada penelitian ini dilakukan standarisasi terhadap seluruh data bertipe numerik untuk memastikan setiap fitur memiliki kontribusi yang setara dalam analisis. Ketidakhadiran standar yang seragam dapat menyebabkan hasil yang berbeda ketika metrik yang sama digunakan dalam lingkungan yang berbeda, sehingga menyulitkan untuk membandingkan model dan mengurangi kepercayaan terhadap hasil [12]. Proses standarisasi dilakukan menggunakan *standard scaler*, di mana hasilnya adalah data dengan nilai rata-rata (*mean*) 0 dan standar deviasi 1.

d. Data Analysis Pasca Preprocessing

Setelah tahap *preprocessing*, langkah berikutnya adalah melakukan analisis *pasca-preprocessing*, di mana pada tahap ini dilakukan analisis lanjutan untuk lebih memahami hubungan antar fitur dengan memvisualisasikan *correlation matrix* dalam bentuk *heatmap* seperti pada gambar 2. Matrix korelasi ini mengandung berbagai informasi spasial dan temporal yang luas tentang sistem dinamis yang mendasarinya. Dalam konteks ini, memprediksi matrix korelasi dari informasi *time series parsial* yang diperoleh dari beberapa node secara acak dapat menggambarkan dinamika spasial dan temporal dari keseluruhan sistem yang mendasari data. Visualisasi *correlation matrix* memungkinkan untuk mengidentifikasi hubungan linier antar fitur, dan hal ini dapat membantu untuk memahami struktur dasar data. Sebagai contoh, dalam analisis yang lebih luas, teknik ini digunakan untuk memprediksi struktur jaringan yang mendasari, seperti menginferensi hubungan neuron dari data *spike* atau mengidentifikasi ketergantungan kausal antar gen dari data ekspresi. Dalam hal ini, meskipun hanya menggunakan data time series yang terbatas dari sebagian kecil node, model tersebut dapat memberikan prediksi yang baik untuk keseluruhan matrix korelasi [13].



Gambar 2. *Correlation Matrix* dalam bentuk diagram *Heatmap*.

e. Penentuan Nilai K Optimal

Tahap selanjutnya adalah penentuan nilai k optimal menggunakan metode Elbow. Metode ini merupakan salah satu cara untuk menentukan jumlah *cluster* yang optimal dengan mengamati persentase setiap *cluster* yang membentuk siku pada titik tertentu. Proses ini dilakukan untuk mencari nilai k yang paling optimal, yang berguna untuk penentuan *cluster* semakin akurat menggunakan Optimasi Elbow [14]. Metode ini biasanya ditampilkan dalam bentuk grafik untuk mempermudah identifikasi siku yang terbentuk. Tujuan dari metode Elbow adalah memilih nilai k yang kecil sambil mempertahankan nilai withinss yang rendah. Nilai k dalam kombinasi siku dengan k-means ditunjukkan melalui grafik yang menggambarkan hubungan antara *cluster* dan penurunan error. Jumlah *cluster* k yang diperoleh dari pengujian menggunakan k-means dievaluasi dengan teknik SSE. SSE (Sum of Square Error) adalah rumus yang digunakan untuk mengukur perbedaan antara data yang telah dianalisis sebelumnya [15]. Metode ini menentukan jumlah *cluster* yang optimal dengan menganalisis variasi inertia dalam data dan memplot hubungan antara jumlah *cluster* dan nilai inertia. Nilai k optimal ditentukan pada titik di mana terjadi perubahan yang signifikan dalam kurva plot (*elbow point*).

f. Penerapan PCA

Setelah menentukan nilai k optimal menggunakan metode Elbow, tahap selanjutnya adalah menerapkan *Principal Component Analysis* (PCA). PCA digunakan untuk mereduksi dimensi data sebelum melakukan klasterisasi menggunakan K-Means. PCA bertujuan untuk menyederhanakan data berdimensi tinggi menjadi bentuk yang lebih sederhana, tanpa kehilangan informasi penting, sehingga mempercepat proses klasterisasi dan meningkatkan efektivitas model [7][16].

PCA bekerja dengan mengidentifikasi komponen utama yang menjelaskan sebagian besar variansi dalam data dengan menghitung hubungan antar fitur melalui matriks kovarians. Kemudian, PCA menghasilkan basis/*axis* baru berdasarkan *eigenvalues* dan *eigenvectors*, yang menyusun dimensi baru yang lebih informatif dan bebas dari korelasi antar fitur. Basis yang paling bermakna (memiliki varians terbesar) akan menjadi PC1, basis paling bermakna berikutnya menjadi PC2, dan seterusnya. Diharapkan bahwa basis baru ini dapat mengungkapkan struktur tersembunyi dalam set data dan menyaring kebisingan/*noise*.

g. K-Means

Pada tahap berikutnya, algoritma K-Means digunakan untuk membagi data yang telah melalui tahapan sebelumnya ke dalam *cluster* berdasarkan nilai k yang telah ditentukan. Klasterisasi ini bertujuan untuk mengelompokkan data berdasarkan kesamaan antar titik data, sehingga pola dalam data dapat diidentifikasi dengan baik. Setelah proses klasterisasi selesai, hasil klaster divisualisasikan dalam data yang telah direduksi dimensinya pada tahapan sebelumnya, yaitu dengan PC1 dan PC2, serta PC1, PC2, dan PC3. Teknik ini sangat penting dalam analisis sistem di mana reduksi dimensi dapat digunakan untuk membantu mengatasi kompleksitas komputasi dengan mendekati sistem berdimensi tinggi menggunakan representasi berdimensi lebih rendah, tanpa kehilangan karakteristik utama sistem [17].

Tahapan dalam K-Means, pertama-tama, menentukan nilai *k*, yaitu jumlah cluster yang ingin dibentuk. Lalu, dipilih *k* objek dari *n* objek data sebagai *centroid*/pusat pengelompokan. Kemudian, hitung jarak antara setiap objek data ke masing-masing centroid menggunakan Euclidean Distance. Setelah ini, cluster dari setiap data ialah centroid yang terdekat dengannya. Nilai rata-rata dari setiap data pada cluster yang terbentuk akan dihitung untuk mendapatkan centroid baru. Proses perhitungan centroid dan clustering dilakukan hingga jumlah deviasi kuadrat rata-rata *(sum of squared errors)* minimal. Pada penelitian ini, nilai *k* berasal dari tahapan penentuan nilai *k* optimal sebelumnya.

h. Visualisasi Cluster

Tahap terakhir adalah visualisasi *cluster* yang telah terbentuk. Visualisasi ini memberikan gambaran bagaimana data terbagi ke dalam kelompok-kelompok yang berbeda dan membantu mengidentifikasi pola atau anomali yang mungkin ada. Dalam hal ini, grafik membantu menjelaskan ketegangan antara observasi dan model, serta sensitivitas observasi terhadap parameter tertentu. Selain itu, visualisasi ini juga mengilustrasikan bagaimana beberapa parameter memiliki dampak yang berbeda terhadap data di setiap *cluster* [18].

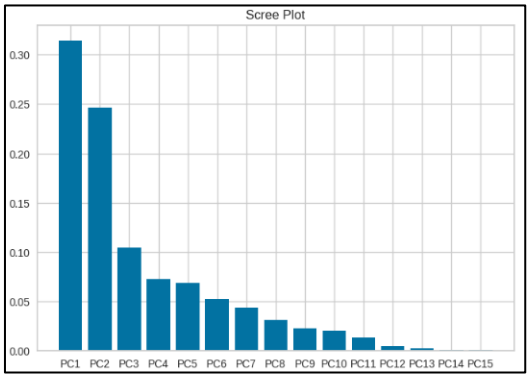
3. HASIL DAN PEMBAHASAN

3.1 Principal Component Analysis

Data hasil preprocessing terdiri dari 180 baris dengan 16 fitur, salah satunya berupa nama negara untuk identifikasi dan analisis hasil. Data input untuk model *clustering* hanya berupa data numerik, yaitu 180 baris dan 15 fitur. Karena jumlah fitur masih banyak, PCA diterapkan untuk mereduksi dimensi menggunakan objek PCA dari scikit-learn. PCA mereduksi dimensi data sambil mempertahankan sebanyak mungkin informasi dari dataset awal dengan memutar dan memproyeksikan data sepanjang arah peningkatan varians. PCA mencari kombinasi linear terbaik dari variabel asli sehingga varians di sepanjang variabel baru maksimal. Fitur dengan varians maksimum menjadi komponen utama, dengan jumlah maksimum komponen setara dengan jumlah fitur atau sampel, mana yang lebih kecil. Untuk membatasi jumlah komponen yang digunakan, dilakukan visualisasi total variasi yang dijelaskan oleh tiap komponen dari rentang 1 hingga jumlah maksimum komponen. Total variasi yang dijelaskan oleh PC1, PC2, PC3, hingga PC15 dapat diakses melalui atribut `explained_variance_ratio_` dari objek PCA.

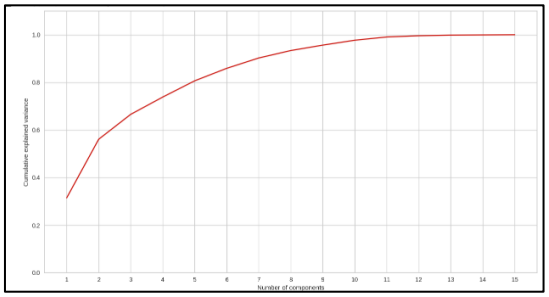
```
[ ]  # cek explained ratio
     pca.explained_variance_ratio_

⇒  array([0.31418148, 0.24651161, 0.1046095 , 0.07263952, 0.06838594,
          0.05248224, 0.04345092, 0.03130452, 0.02278299, 0.02059814,
          0.01367252, 0.00516376, 0.00254689, 0.00086286, 0.0008071 ])
```

Gambar 3. Nilai variasi yang dijelaskan oleh komponen PC.



Gambar 4. Visualisasi Scree Plot variasi tiap komponen PC.

Artinya, PC1 memiliki explained variance sebesar 31% dan PC2 memiliki *explained variance* sebesar 24%. Jika ditotal, menjadi 56%., yang berarti jika kita menggunakan kedua principal component, maka hanya bisa menjelaskan sebesar 56% variasi pada data. Total variasi yang dijelaskan juga dapat direpresentasikan dalam visualisasi *cumulative explained variance*.



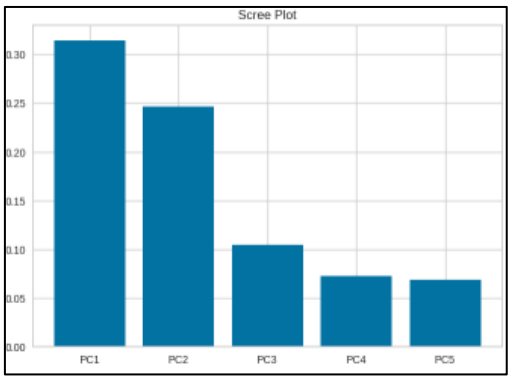
Gambar 5. Visualisasi Total Variasi Kumulatif yang dijelaskan oleh komponen PC.

Sebanyak 80% varians (0.80632806) dijelaskan oleh 5 principal components, sementara 93% varians (0.93356573) dijelaskan oleh 8 principal components. Rule of thumb adalah mempertahankan setidaknya 80% varians, tetapi komponen 2 atau 3 sering dipilih untuk memudahkan visualisasi dalam dua atau tiga dimensi. Dalam penelitian ini, digunakan 5 principal components untuk menjelaskan 80% varians, dengan hasil *clustering* divisualisasikan menggunakan PC1 dan PC2, serta PC1, PC2, dan PC3 dari 5 komponen

tersebut. Implementasi PCA dibatasi pada 5 komponen dengan random_state yang diatur agar hasil konsisten setiap kali dijalankan.

```
[ ] N_COMPONENTS = 5
    pca = PCA(n_components = N_COMPONENTS, random_state=SEED)
    X_pca = pca.fit_transform(df_scaled)
    X_pca.shape

→ (179, 5)
```

Gambar 6. Membuat PCA dengan Jumlah Komponen=5.



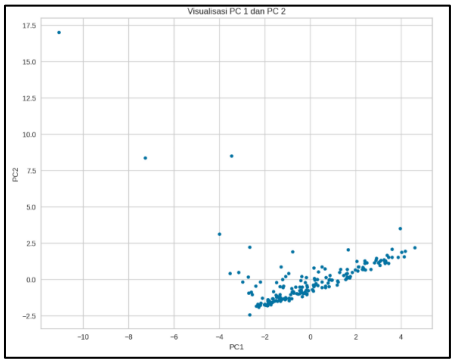
Gambar 7. Scree Plot untuk PCA dengan 5 Principal Component.

Loading scores dapat dianalisis untuk melihat proporsi fitur-fitur yang membentuk masing-masing *principal component*. Pada objek PCA di scikit-learn, loading scores diakses melalui atribut `components_`. *Loading scores* untuk 5 *principal component* ditampilkan pada gambar 8. Nilai 0.407 pada indeks 0 kolom *Birth Rate* menunjukkan bahwa fitur *Birth Rate* memiliki proporsi 0.407 terhadap PC1. Fitur *Birth Rate*, *Fertility Rate, Infant Mortality, dan Life Expectancy* memiliki nilai tertinggi pada indeks 0, menunjukkan pentingnya dalam membentuk PC1. Sementara itu, fitur *Co2-emissions, Population*, dan GDP memiliki nilai tertinggi pada indeks 1, sehingga berkontribusi paling besar pada PC2. Oleh karena itu, fitur *Birth Rate, GDP, Infant Mortality, Population, dan Co2-emissions* akan dieksplorasi lebih lanjut dalam analisis hasil.

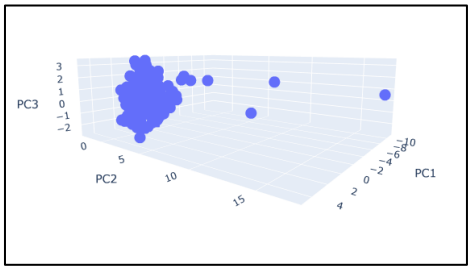
	Density\n(P/Km2)	Agricultural Land(%)	Land Area(Km2)	Birth Rate	Co2-Emissions	Fertility Rate	Forested Area (%)	GDP	Gross primary education enrollment (%)	Infant mortality	Life expectancy	Out of pocket health expenditure	Physicians per thousand	Population	Urban_population
0	-0.061	0.075	-0.172	0.407	-0.249	0.394	-0.024	-0.252	-0.006	0.390	-0.393	0.138	-0.328	-0.177	-0.222
1	-0.061	0.129	0.292	0.196	0.418	0.194	-0.059	0.363	-0.001	0.215	-0.212	0.115	-0.192	0.419	0.437
2	-0.112	-0.512	0.090	0.057	0.044	0.066	0.660	0.065	0.281	0.025	-0.046	-0.381	-0.201	-0.021	0.010
3	-0.530	0.505	0.019	0.023	-0.018	0.006	-0.123	0.014	0.537	-0.051	-0.009	-0.379	0.091	-0.061	-0.039
4	0.718	0.082	-0.182	0.010	-0.001	-0.036	-0.132	-0.029	0.610	0.003	0.023	-0.066	-0.179	0.110	0.061

Gambar 8. *Loading Scores* untuk 5 Principal Components.

Visualisasi PC1 dan PC2 dalam dua dimensi tampak pada gambar 9. Visualisasi PC1, PC2 dan PC3 tampak pada gambar 10. Setelah klasterisasi, akan divisualisasikan Kembali kedua plotting ini.



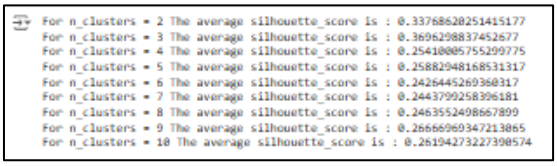
Gambar 9. Hasil PC1 dan PC2.



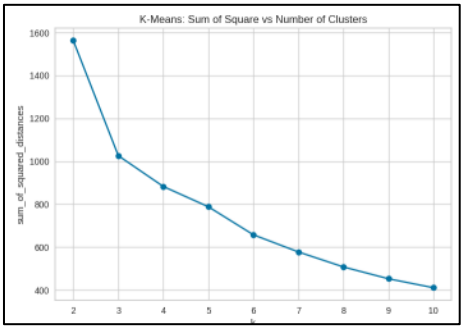
Gambar 10. Hasil PC1, PC2 dan PC3.

3.2 K-Means

Tahap ini bertujuan menentukan jumlah klaster optimal dalam K-Means *clustering* dengan evaluasi metrik Inertia (SSE) dan Silhouette Score. Algoritma dijalankan untuk k = 2 hingga k = 10, dengan max_iter = 100 dan n_init = 5. Hasil evaluasi menunjukkan Silhouette *Score* tertinggi pada k = 3 (nilai 0.3693), yang menunjukkan pemisahan klaster optimal. Inertia juga diukur untuk memastikan penurunan jarak kuadrat antara titik data dan pusat klaster. Kombinasi kedua metrik ini menunjukkan k = 3 sebagai jumlah klaster optimal, yang akan divalidasi melalui visualisasi seperti Elbow Method dan distribusi Silhouette *Score*.

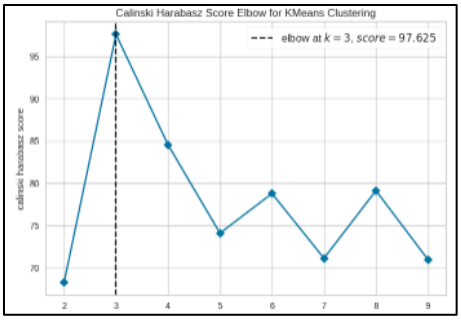


Gambar 11. Penentuan Jumlah Klaster Optimal K-Means.



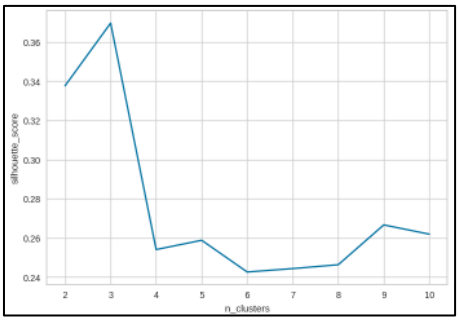
Gambar 12. Visualisasi Elbow Method.

Visualisasi Elbow Method menunjukkan bahwa SSD menurun seiring bertambahnya K, namun laju penurunan melambat pada titik tertentu, membentuk kurva yang merata. Titik siku ini menunjukkan jumlah klaster optimal, yaitu sekitar k = 3 atau k = 4.



Gambar 13. Visualisasi Skor Calinski-Harabasz untuk Klasterisasi K-Means.

Visualisasi ini menunjukkan Skor Calinski-Harabasz untuk klasterisasi K-Means dengan berbagai K. Skor ini mencerminkan pemisahan antar klaster, yang meningkat seiring bertambahnya K. Namun, setelah titik tertentu, peningkatan melambat, menandakan penambahan klaster tidak signifikan. Skor tertinggi tercapai pada K = 3, menunjukkan tiga klaster sebagai pilihan optimal.

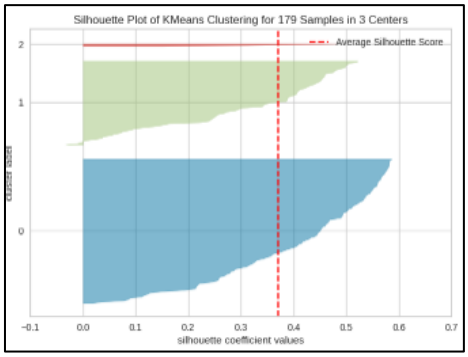


Gambar 14. Visualisasi Skor Silhouette untuk Klasterisasi K-Means.

Visualisasi ini menggambarkan Skor Silhouette untuk klasterisasi dengan berbagai K. Skor ini mengukur seberapa baik pemisahan dan kohesi antar klaster, dengan nilai antara -1 hingga 1. Nilai lebih tinggi menunjukkan pemisahan yang lebih baik. Grafik menunjukkan bahwa skor Silhouette lebih tinggi pada K yang lebih rendah, mengindikasikan bahwa jumlah klaster optimal mungkin berada pada kisaran K yang lebih kecil.

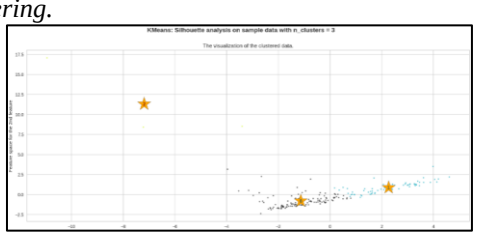
3.3 Evaluasi

Evaluasi dilakukan dengan nilai Silhouette Score dan visualisasi Silhouette Plot, visualisasi PCA, serta visualisasi distribusi box plot dari variabel-variabel PCA. Evaluasi dilakukan untuk memahami hasil cluster yang terbentuk.



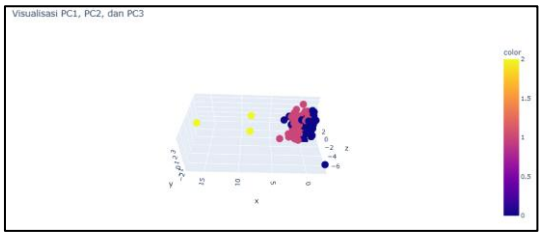
Gambar 15. Evaluasi model K-Means menggunakan Silhouette Score.

Visualisasi pada gambar 15 menunjukkan hasil evaluasi model K-Means menggunakan Silhouette *Score*, yang mengukur pemisahan titik data dalam klaster. Nilai mendekati +1 menunjukkan pemisahan baik, sementara nilai mendekati -1 menunjukkan kesalahan pengelompokan. Pada plot, setiap baris mewakili satu klaster, dengan panjang bar menunjukkan nilai Silhouette, dan garis merah putus-putus menandakan *Average* Silhouette Score. Jika sebagian besar nilai Silhouette berada di atas garis rata-rata, model K-Means dengan 3 klaster dianggap berhasil. Skor Silhouette 0.3693 atau sekitar 37%, menunjukkan klaster ini tergolong *fair clustering.*



Gambar 16. Visualisasi hasil K-Means dengan PC1 dan PC2.

Visualisasi pada gambar 16 menampilkan pemetaan dua dimensi dari data yang telah direduksi menggunakan Principal Component Analysis (PCA). Dua komponen utama (PC1 dan PC2) digunakan sebagai sumbu untuk menggambarkan struktur data. Setiap titik merepresentasikan sampel, dengan warna titik menunjukkan keanggotaan klaster yang telah ditentukan. Titik berwarna kuning berfungsi sebagai pusat klaster atau centroid, menunjukkan lokasi sentral masing-masing kelompok.



Gambar 17. Visualisasi Hasil K-Means dengan PC1, PC2, dan PC3.

Visualisasi pada gambar 17 menampilkan pemetaan tiga dimensi dari data yang telah direduksi menggunakan *Principal Component Analysis* (PCA). Tiga komponen utama (PC1, PC2, dan PC3) digunakan sebagai sumbu untuk menggambarkan struktur data. Setiap titik mewakili sampel data, dengan warna menunjukkan keanggotaan klaster yang dihasilkan oleh model K-Means. Titik-titik berwarna menggambarkan distribusi data dalam berbagai klaster, sementara gradien warna di sisi plot memberikan indikasi kategori atau label yang berbeda.

Cluster ID	GDP	Birth Rate	Infant Mortality	Population	CO2 Emissions
0	406550367464.872	14.080	9.266	23032212.624	129716.734
1	53732463105.044	30.982	41.010	28928179.426	27923.559
2	14649566666666.666	13.453	14.300	1030790759.000	5769004.000

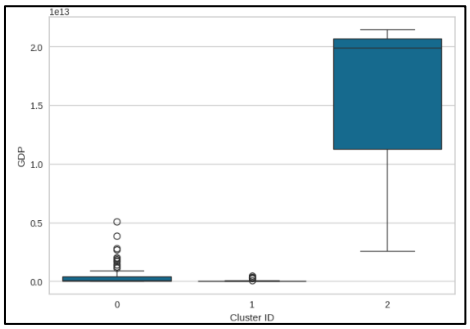
Gambar 18. Analisis variabel yang mendefiniskan PC1 dan PC2.

Visualisasi ini menyajikan analisis data pada dataframe yang telah digabungkan, menampilkan informasi terkait klaster yang dihasilkan dari model K-Means. DataFrame ini mencakup variabel-variabel penting yang mendefinisikan dua komponen utama (PC1 dan PC2) dari analisis PCA. Kolom yang ditampilkan meliputi *Cluster* ID, GDP (kesehatan ekonomi), tingkat kelahiran per 1000 populasi per tahun (informasi demografis), angka kematian bayi per 1000 kelahiran sebelum mencapai umur 1 tahun (kualitas layanan kesehatan), jumlah populasi (konteks skala), serta emisi CO2 dalam skala ton (aktivitas industri).

```
Cluster 0
['Albania' 'Algeria' 'Antigua and Barbuda' 'Argentina' 'Armenia'
 'Australia' 'Austria' 'Azerbaijan' 'The Bahamas' 'Bahrain' 'Barbados'
 'Belarus' 'Belgium' 'Belize' 'Bhutan' 'Bolivia' 'Brazil' 'Brunei'
 'Bulgaria' 'Cape Verde' 'Canada' 'Chile' 'Colombia' 'Costa Rica'
 'Croatia' 'Cyprus' 'Czech Republic' 'Denmark' 'Dominica'
 'Dominican Republic' 'Ecuador' 'El Salvador' 'Estonia' 'Fiji' 'Finland'
 'France' 'Georgia' 'Germany' 'Greece' 'Grenada' 'Guyana' 'Honduras'
 'Hungary' 'Iceland' 'Indonesia' 'Iran' 'Republic of Ireland' 'Israel'
 'Italy' 'Jamaica' 'Japan' 'Jordan' 'Kazakhstan' 'Kuwait' 'Latvia'
 'Lebanon' 'Libya' 'Lithuania' 'Luxembourg' 'Malaysia' 'Maldives' 'Malta'
 'Mauritius' 'Mexico' 'Moldova' 'Mongolia' 'Montenegro' 'Morocco'
 'Netherlands' 'New Zealand' 'Nicaragua' 'Norway' 'Oman' 'Palau' 'Panama'
 'Paraguay' 'Peru' 'Poland' 'Portugal' 'Qatar' 'Romania' 'Russia'
 'Saint Kitts and Nevis' 'Saint Lucia' 'Saint Vincent and the Grenadines'
 'Samoa' 'Saudi Arabia' 'Serbia' 'Seychelles' 'Singapore' 'Slovakia'
 'Slovenia' 'South Korea' 'Spain' 'Sri Lanka' 'Suriname' 'Sweden'
 'Switzerland' 'Thailand' 'Trinidad and Tobago' 'Tunisia' 'Turkey'
 'Ukraine' 'United Arab Emirates' 'United Kingdom' 'Uruguay' 'Uzbekistan'
 'Venezuela' 'Vietnam']
Cluster 1
['Afghanistan' 'Angola' 'Bangladesh' 'Benin' 'Botswana' 'Burkina Faso'
 'Burundi' 'Ivory Coast' 'Cambodia' 'Cameroon' 'Central African Republic'
 'Chad' 'Comoros' 'Republic of the Congo'
 'Democratic Republic of the Congo' 'Djibouti' 'Egypt' 'Equatorial Guinea'
 'Eritrea' 'Ethiopia' 'Gabon' 'The Gambia' 'Ghana' 'Guatemala' 'Guinea'
 'Guinea-Bissau' 'Haiti' 'Iraq' 'Kenya' 'Kiribati' 'Kyrgyzstan' 'Laos'
 'Lesotho' 'Liberia' 'Madagascar' 'Malawi' 'Mali' 'Marshall Islands'
 'Mauritania' 'Federated States of Micronesia' 'Mozambique' 'Myanmar'
 'Namibia' 'Nepal' 'Niger' 'Nigeria' 'Pakistan' 'Papua New Guinea'
 'Philippines' 'Rwanda' 'S����������' 'Senegal' 'Sierra Leone'
 'Solomon Islands' 'South Africa' 'Sudan' 'Syria' 'Tajikistan' 'Tanzania'
 'East Timor' 'Togo' 'Tonga' 'Turkmenistan' 'Uganda' 'Vanuatu' 'Yemen'
 'Zambia' 'Zimbabwe']
Cluster 2
['China' 'India' 'United States']
```

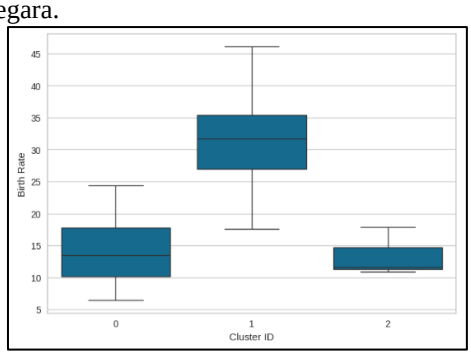
Gambar 19. Distribusi negara berdasarkan *cluster*.

Gambar 4.21 menampilkan hasil pengelompokan negara berdasarkan klaster yang dihasilkan dari analisis data, dengan tiga klaster yang mengidentifikasi negara berdasarkan kesamaan tertentu, seperti faktor ekonomi atau demografi.



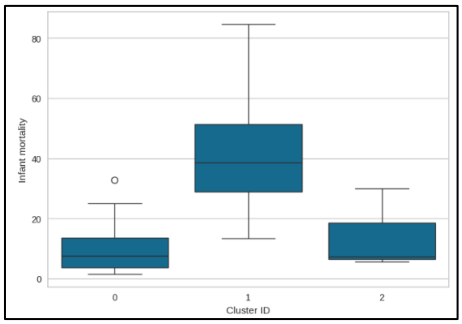
Gambar 20. Distribusi GDP dari setiap *cluster*.

Visualisasi melalui *Box plot* pada Gambar 20 menunjukkan distribusi data dan *outlier* tingkat kesejahteraan ekonomi antar negara.



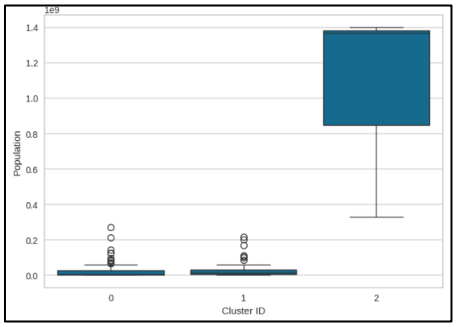
Gambar 21. Distribusi Tingkat kelahiran dari setiap *cluster*.

Visualisasi melalui *Box plot* pada Gambar 21 menunjukkan distribusi data dan outlier tingkat kelahiran antar negara.



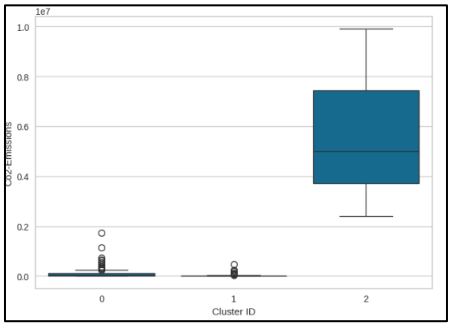
Gambar 22. Distribusi Tingkat kematian bayi dari setiap *cluster*.

Visualisasi melalui *Box plot* pada Gambar 22 menunjukkan distribusi data dan *outlier* tingkat kematian bayi antar negara.



Gambar 23. Distribusi Tingkat populasi dari setiap *cluster*.

Visualisasi melalui *Box plot* pada Gambar 23 menunjukkan distribusi data dan *outlier* tingkat populasi antar negara.



Gambar 24. Distribusi Tingkat emisi CO2 dari setiap *cluster*.

Visualisasi melalui Box plot pada Gambar 24 menunjukkan distribusi data dan *outlier* tingkat emisi CO2 antar negara.

4. Kesimpulan

Dalam analisis klasterisasi negara berdasarkan data sosioekonomi dan demografi tahun 2023 menggunakan pca dan K-Means, ditemukan tiga kluster dengan karakteristik berbeda. *Cluster* 0 menunjukkan karakteristik positif dengan GDP sedang hingga tinggi, populasi dan angka kelahiran sedang, serta angka kematian bayi rendah, mencerminkan kesehatan dan kebijakan lingkungan yang baik. *Cluster* 1 mengalami tantangan kesehatan dan ekonomi, dengan angka kematian bayi tinggi dan GDP rendah, meskipun memiliki angka kelahiran tinggi. *Cluster* 2, yang mencakup negara-negara seperti china, india, dan as, memiliki GDP tinggi dan angka kematian bayi rendah, tetapi menghadapi tantangan emisi Co2 tinggi. Analisis ini menunjukkan perlunya kebijakan terintegrasi dan berkelanjutan untuk meningkatkan kesehatan, ekonomi, dan lingkungan global. Kinerja K-Means yang didukung PCA menunjukkan hasil klasterisasi yang baik, tetapi interpretasi dapat ditingkatkan dengan memilih fitur yang lebih fokus pada domain tertentu, seperti ekonomi.

Daftar Pustaka

[1] B. S. Lal, “Demographic and Socio-Economic Development Evidence from G7 Countries,” *Studies in Social Science & Humanities*, vol. 2, no. 8, pp. 17–26, Aug. 2023, doi: 10.56397/sssh.2023.08.03.

[2] N. Mohammad, “A Computational Theory and Semi-Supervised Algorithm for Clustering,” Jun. 2023, [Online]. Available: http://arxiv.org/abs/2306.06974

[3] E. Tarver, “What Is Social Economics, and How Does It Impact Society?,” investopedia.com. Accessed: Jan. 11, 2025. [Online]. Available: https://www.investopedia.com/terms/s/social-economics.asp#toc-what-is-social-economics

[4] A. Hayes, “Demographics: How to Collect, Analyze, and Use Demographic Data,” investopedia.com. Accessed: Jan. 11, 2025. [Online]. Available: https://www.investopedia.com/terms/d/demographics.asp

[5] R. Kurniawan, M. S. Hasibuan, and R. Hasibuan, "Klasterisasi Wilayah Prioritas Vaksin Menggunakan Algoritma K-Means Clustering," *Media Online*, vol. 4, no. 3, pp. 1585–1592, 2023, doi: 10.30865/klik.v4i3.1334.

[6] T. A. Munandar and D. Handayani, "K-Means Cluster Algorithm for Grouping Inequality in Regional Development," *International Journal of Information Technology and Computer Science Applications*, vol. 1, no. 1, 2023, doi: 10.58776/ijitcsa.v1i1.20.

[7] T. Kurita, "Principal Component Analysis (PCA)," in *Computer Vision*, Cham: Springer International Publishing, 2020, pp. 1–4. doi: 10.1007/978-3-030-03243-2_649-1.

[8] X. Lin and J. Xu, "Road network partitioning method based on canopy-kmeans clustering algorithm," *Archives of Transport*, vol. 54, no. 2, pp. 95–106, 2020, doi: 10.5604/01.3001.0014.2970.

[9] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *Journal of Informatics and Computer Science*, vol. 05, 2023.

[10] C. M. P. Santosa, E. Sumirat, and O. Y. Sudrajad, "An Exploratory Data Analysis (EDA) Approach for Analyzing Financial Statements in Pharmaceutical Companies Using Machine Learning," *International Journal of Current Science Research and Review*, vol. 07, no. 07, Jul. 2024, doi: 10.47191/ijcsrr/V7-i7-12.

[11] Baihaqiyazid, "Data Scaling," Medium.com. Accessed: Jan. 11, 2025. [Online]. Available: https://medium.com/@baihaqiyazid16/data-scaling-3669f475790a

[12] M. R. Salmanpour *et al.*, "Machine Learning Evaluation Metric Discrepancies across Programming Languages and Their Components: Need for Standardization," Nov. 2024.

[13] N. Easaw, W. S. Lee, P. S. Lohiya, S. Jalan, and P. Pradhan, "Estimation of Correlation Matrices from Limited time series Data using Machine Learning," Sep. 2022, [Online]. Available: http://arxiv.org/abs/2209.01198

[14] A. Muqoddam, "Pengelompokan Produksi tambak Garam dengan Metode Cluster K-Means dan Optimasi Cluster Menggunakan Elbow (studi kasus: dinas kelautan Kabupaten Bangkalan)," *Jurnal METHODIKA*, Mar. 2023.

[15] N. A. Maori, "Metode Elbow Dalam Optimasi Jumlah Cluster Pada K-Means Clustering," *Jurnal SIMETRIS*, vol. 14, 2023.

[16] G. Li and Y. Qin, "An Exploration of the Application of Principal Component Analysis in Big Data Processing," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, 2024, doi: 10.2478/amns-2024-0664.

[17] M. Redmann, "Dimension reduction for large-scale stochastic systems with non-zero initial states and controlled diffusion," Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.00581

[18] U. Laa and G. Valencia, "Clustering and visualization tools to study high dimensional parameter spaces: B anomalies example," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2304.00151