Jurnal Rekursif, Volume 12 Nomor 2, November 2024, ISSN 2303-0755, e-ISSN 2777-0427
www.ejournal.unib.ac.id/index.php/rekursif

Analisis Klaster Negara Berdasarkan Indikator Sosial-Ekonomi Menggunakan Fuzzy C-Means dan K-Means

Davi Sulaiman, Rafi Afrian Al-Haritz, Fahim Ahmad Saputra, Ade Irawan

Informatika, Fakultas Teknik, Universitas Bengkulu, Jl. Wr. Supratman Kandang Limun, Bengkulu Bengkulu 38371 A Telp: (0736) 344087, 22105 - 227

[1]davisulaiman1@gmail.com
[2]rafiafrian003@gmail.com
[3]fahimokensyah@gmail.com
[4]otusade11@gmail.com

Abstrak: Penelitian ini menganalisis pengelompokan negara berdasarkan indikator sosial-ekonomi menggunakan algoritma *Fuzzy C-Means* dan *K-Means*. Setiap negara memiliki karakterisitik sosial-ekonomi yang unik, meliputi aspek seperti angka kematian anak, ekspor, impor, pendapatan per kapira, inflasi, angka harapan hidup, dan produk domestik bruto. Dataset mencakup 167 negara dengan indikator 10 utama. Evaluasi efektivitas klasterisasi dilakukan menggunakan *Sum of Squared Errors (SSE)* dan *Silhoutte Score*. Hasil menunjukkan metode *StandardScaler* menghasilkan *Sum of Squared Errors* terbaik pada klaster ke-10 sebesar 416.04, sedangkan *MinMaxScaler* memiliki *Sum of Squared Errors* terendah pada klaster ke-10 sebesar 10.30. Untuk pengurangan dimensi dengan *StandardScaler + PCA*, nilai *Sum of Squared Errors* mencapai 5771653036.34 pada klaster ke-10. Berdasarkan *Silhouttte Score*, jumlah klaster optimal adalah 2 dengan skor 0.41, menunjukkan *K-Means* efektif untuk memisahkan data dalam dua kelompok utama berdasarkan karakteristik sosial-ekonomi.
Kata Kunci: Klasterisasi, *Fuzzy C-Means*, *K-Means*, Indikator sosial-ekonomi, *Sum of Squared Errors (SSE)*, *Silhouttte Score*.

Abstract: This research analyzes the clustering of countries based on socio-economic indicators using Fuzzy C-Means and K-Means algorithms. Each country has unique socio-economic characteristics, including aspects such as child mortality, exports, imports, per capita income, inflation, life expectancy, and gross domestic product. The dataset covers 167 countries with 10 main indicators. Evaluation of clustering effectiveness was conducted using Sum of Squared Errors (SSE) and Silhoutte Score. Results show the StandardScaler method produces the best Sum of Squared Errors at the 10th cluster of 416.04, while MinMaxScaler has the lowest Sum of Squared Errors at the 10th cluster of 10.30. For dimension reduction with StandardScaler + PCA, the Sum of Squared Errors value reached 5771653036.34 in the 10th cluster. Based on the Silhouttte Score, the optimal number of clusters is 2 with a score of 0.41, indicating K-Means is effective for separating data into two main groups based on socio-economic characteristics.

Keywords: Clustering, Fuzzy C-Means, K-Means, Socio-economic indicators, Sum of Squared Errors (SSE), Silhouttte Score.

I. PENDAHULUAN

Setiap negara memiliki karakteristik sosial-ekonomi yang unik, mencakup berbagai aspek seperti kematian anak, ekspor, impor, kesehatan, *Gross Domestic Product*, inflasi, pemasukan, angka harapan hidup, *total fertility rate*.(Suryadi, 2015) Perbedaan ini tidak hanya memengaruhi kebijakan pemerintah dalam Upaya meningkatkan kesejahteraan masyrakat, tetapi juga menciptakan tantangan yang kompleks dalam pengelolaan sumber daya dan perencanaan pembangunan. (Sanusi et al., 2020a).Dengan semakin kompleksnya data sosial-ekonomi yang tersedia metode analisis data menjadi sangat penting untuk membantu memahami pola yang ada dan mengelompokkan negara-negara berdasarkan kesamaan karakteristik sosial-ekonomi.(Qaadan et al., 2024)

Metode klastering sering digunakan dalam analisis data sosial-ekonomi yntuk mengelompokkan

data menjadi beberapa kelompok homogen .(Putriana et al., 2016). Dalam konteks negara, analisis klaster membantu mengidentifikasi pola perbedaan dan kesamaan antar indikator sosial-ekonomi. Seperti yang dilaporkan Perserikatan Bangsa-Bangsa (UNDP). Data menunjukkan bahwa negara maju seperti Norwegia dan Swiss memiliki Indeks Pembangunan Manusia (IPM) tinggi (0.94-0.96), semenata negara berkembang seperti Chad dan Sudan Selatan memiliki IPM rendah (masing-masing 0.394 dan 0.385). Ketidakmerataan distiribusi sumber daya ini menciptakan kebutuhan pendekatan analitik untuk mengelompokkan negara dan mendukung kebijakan pembangunan yang lebih adil (Hussain et al., 2023)

Penelitian sebelumnya telah menujukkan keunggulan metode *Fuzzy C-Means (FCM)* dan *K-Means* dalam analisis klasterisasi. Misalnya, (Putriana et al., 2016) menggunakan *K-Means* untuk mengelompokkan kabupaten di Jawa Tengah berdasarkan indikator kemiskinan yang berhasil menghasilkan klaster efektif dalam memetakan tingkat kemiskinan rendah, sedang, dan tinggi. *FCM*, di sisi lain, menawarkan fleksibilitas dengan memungkinkan setiap data menjadi anggota lebih dari satu klaster melalui derajat kenaggotaan (Sanusi et al., 2020). Menurut (Singh et al., 2023) mengonfirmasi bahwa *FCM* memberikan hasil pengelopokan yang lebih akuran dibandingkan metode lain, terutama dalam konteks data sosial-ekonomi yang kompleks.

Dalam penelitian ini, *FCM* dan *K-Means* dipilih karena keunggulan masing-masing. *K-Menas* merupakan metode yang sederhana, cepat, dan efektif untuk klasterisasi dengan batas yang jelas (Atiqah et al., 2018). Sebaliknya, FCM unggul dalam fleksibilitas dan akurasi karena mampu menangkap tumpang tindih antar-klaster (Sun et al., 2024). Kombinasi kedua algoritma ini memungkinkan peneliti untuk memanfaatkan keunggulan masing-masing sekaligus memvalidasi hasil pengelompokan.

Klaster negara berdasarkan indikator sosial-ekonomi membutuhkan algoritma yang akurat sehingga peneliti memilih dua metode algoritma yang popular yaitu Fuzzy C-Means dan K-Means. Fuzzy C-Means (FCM) adalah algoritma klastering yang memungkinkan setiap data point untuk menjadi anggota dari lebih dari satu kluster dengan derajat keanggotaan yang berbeda. Menurut (Saatchi, 2024) FCM menggabungkan prinsip fuzzy dengan metode K-Means, di mana setiap data memiliki nilai keanggotaan yang berkisar antara 0 hingga 1 untuk setiap kluster yang ada. Hal ini berbeda dengan metode K-Means yang bersifat keras, di mana setiap data hanya dapat menjadi anggota dari satu kluster. (Zhou et al., 2024)

Dengan mengaplikasikan metode ini, peneliti yakin terhadap penelitian ini untuk memberikan wawasan baru tentang bagaimana negara-negara dapat dikelompokan berdasarkan indikator sosial-ekonomi mereka. Pendekatan ini dharapkan dapat membantuk dalam pengambilan keputusan kebijakan yang lebih baik di tingkat globar(Anwar, 2023).

II. METODE DAN BAHAN

Dataset yang digunakan dalam penelitian ini adalah data negara dalam format CSV berjudul "*Country-data.csv*". Dataset ini mencakup berbagai indikator sosial, ekonomi, dan kesehatan dari berbagai negara. Setiap baris dalam dataset ini mewakili satu negara, sementara setiap kolom merepresentasikan indikator tertentu seperti angka kematian anak (*child_mort*), ekspor barang dan jasa sebagai persentase dari PDB (*exports*), pengeluaran kesehatan sebagai persentase dari PDB (*health*), impor barang dan jasa sebagai persentase dari PDB (imports), pendapatan per kapita dalam dolar AS (*Income*), inflasi tahunan (*Inflation*), harapan hidup

dalam tahun (*life_expec*), total fertilitas atau jumlah anak per wanita (*total_fer*), dan produk domestik bruto per kapita dalam dolar AS (*gdpp*). Dataset ini diimpor menggunakan pustaka *pandas* dalam *Python* dan terdiri dari sejumlah baris dan kolom yang mewakili data dari berbagai negara. Metode shape pada DataFrame digunakan untuk mengetahui jumlah data, yang menunjukkan bahwa dataset ini terdiri dari nnn baris dan mmm kolom, di mana nnn adalah jumlah negara dan mmm adalah jumlah indikator. Data ini digunakan untuk mengkaji berbagai studi kasus yang melibatkan analisis terhadap indikator-indikator yang tercantum, serta digunakan sebagai parameter dalam berbagai metode analisis dan pemodelan statistik.(Rahakbauw et al., 2017)

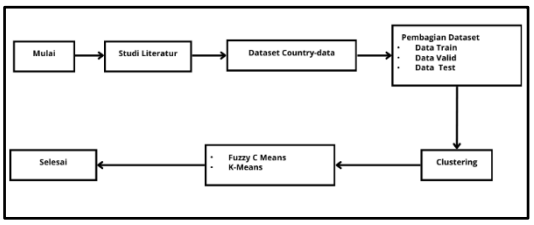
Penelitian ini bertujuan untuk menemukan cara optimal dalam mengelompokkan negara-negara berdasarkan indikator sosial-ekonomi, yang kemudian dapat digunakan sebagai acuan dalam perbandingan serta analisis pola ekonomi dan sosial global. Selain itu, penelitian ini juga bertujuan untuk mengidentifikasi algoritma yang paling efektif dan efisien antara Fuzzy C-Means dan K-Means dalam mengelompokkan negara berdasarkan indikator sosial-ekonomi tersebut.(Aisah et al., 2022)

Dataset yang digunakan adalah *"Country-data.csv"* yang mencakup berbagai indikator sosial, ekonomi, dan kesehatan dari 167 negara. Setiap baris dalam dataset merepresentasikan satu negara, sementara kolom-kolomnya terdiri dari indikator seperti angka kematian anak, ekspor sebagai persentase dari PDB, pengeluaran kesehatan, impor, pendapatan per kapita, inflasi, harapan hidup, total fertilitas, dan PDB per kapita. Dataset ini berjumlah 167 baris (negara) dan 10 kolom (indikator).

Evaluasi hasil klastering dilakukan menggunakan dua algoritma, yaitu *Fuzzy C-Means* dan *K-Means*. Penghitungan ini melibatkan metrik evaluasi seperti *Sum of Squared Errors* (SSE) untuk melihat seberapa baik negara-negara terkelompok dalam klaster, serta waktu komputasi untuk menilai efisiensi masing-masing algoritma.

Setelah klastering, dilakukan analisis hasil pengelompokan negara berdasarkan indikator sosial-ekonomi yang terbentuk dari kedua algoritma. Analisis ini menggambarkan pola dan tren dalam data berdasarkan klaster yang terbentuk. Diskusi dengan para ahli melalui wawancara atau diskusi kelompok terfokus juga dilakukan untuk mendapatkan wawasan tambahan mengenai hasil klastering dan implikasi kebijakan.

Penelitian ini menghasilkan pengelompokan negara berdasarkan indikator sosial-ekonomi yang dapat digunakan sebagai acuan untuk kebijakan dan perencanaan di bidang sosial-ekonomi. Selain itu, penelitian ini juga mengidentifikasi algoritma klastering (antara *Fuzzy C-Means* dan *K-Means*) yang paling optimal dalam hal akurasi dan efisiensi waktu, sehingga memberikan rekomendasi terkait metode yang paling efektif untuk pengelompokan negara berdasarkan indikator sosial-ekonomi.



Gambar 1. Strategi Penelitian

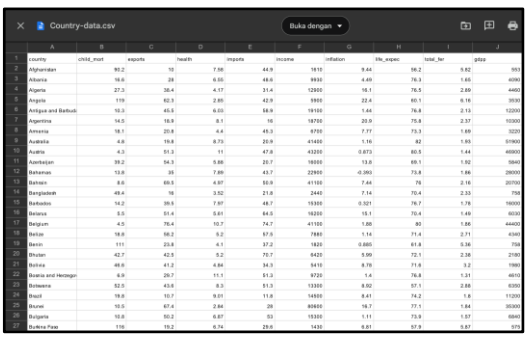
2.1.Studi Literatur

Pada tahap studi literatur, peneliti melakukan pencarian sumber teori yang valid yang mendukung penelitian ini. Sumber data dan teori ini ditemukan melalui jurnal penelitian sebelumnya, jurnal yang relevan untuk mendukung teori yang dibuat, buku-buku, dan artikel di situs

web yang mendukung pemahaman tentang penelitian ini. Tahap ini memiliki tujuan untuk digunakan sebagai referensi dan memperkuat penelitian yang sedang dilakukan.

2.2.Dataset

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari situs web Kaggle. Dataset ini berisi informasi mengenai 167 negara dengan berbagai indikator sosial dan ekonomi. Kolom pertama adalah country, yang mencatat nama negara. Kolom *child_mort* menunjukkan tingkat kematian anak per 1.000 kelahiran. Kolom *exports*, *health*, dan *imports* masing-masing menggambarkan persentase ekspor, pengeluaran untuk kesehatan, dan impor terhadap produk domestik bruto (PDB). Kolom *income* mencatat pendapatan rata-rata per kapita dalam mata uang tertentu, sementara *inflation* menunjukkan tingkat inflasi tahunan. Harapan hidup rata-rata suatu negara tercatat dalam kolom *life_expec*, sedangkan kolom total_fer menunjukkan angka kelahiran rata-rata per wanita. Akhirnya, kolom gdpp menyajikan pendapatan domestik bruto per kapita. Dataset ini memberikan gambaran menyeluruh tentang kesejahteraan sosial, ekonomi, dan kesehatan dari berbagai negara, yang dapat digunakan untuk analisis lebih lanjut seperti pengelompokan, korelasi antar variabel, atau studi perbandingan antar negara.



Gambar 2. Dataset yang digunakan

2.3. Modeling

Modeling adalah tahap dalam penelitian ini di mana menggunakan model clustering, yaitu proses pengelompokkan data tanpa label menggunakan algoritma *K-Means* dan *Fuzzy C-Means* (FCM). Sebelum modeling, data diproses dengan teknik scaling, seperti *Min-Max Scaling* dan *Standard Scaling*, untuk memastikan semua fitur berada pada skala yang sama. Setelah itu, dilakukan dimensionality reduction menggunakan *Principal Component Analysis* (PCA) untuk mengurangi jumlah fitur, yang membantu meningkatkan efisiensi algoritma *clustering*.

Pada proses *clustering*, *K-Means* digunakan untuk mengelompokkan data ke dalam sejumlah cluster dengan pendekatan berbasis jarak antara data dan pusat *cluster* (*centroid*). Hasilnya adalah pembagian data ke *cluster* tertentu secara eksklusif. Selain itu, digunakan juga *Fuzzy C-Means* (FCM), yang merupakan varian dari *K-Means*, tetapi dengan fleksibilitas lebih tinggi karena memungkinkan data memiliki derajat keanggotaan ke lebih dari satu *cluster*. Model-model ini cocok untuk analisis eksplorasi data, terutama ketika struktur label data tidak diketahui sebelumnya.

2.4. Evaluasi Model

Pada tahap evaluasi, model menggunakan metrik clustering seperti *Silhouette Score* atau Inertia untuk menilai kualitas pengelompokan data. *K-Means* dievaluasi berdasarkan kedekatan data dengan centroid dan pemisahan antar cluster, sedangkan *Fuzzy C-Means* (FCM) dinilai dari derajat keanggotaan tiap data pada. Visualisasi hasil clustering, seperti scatter *cluster* plot setelah reduksi dimensi dengan PCA, digunakan untuk memvalidasi hasil pengelompokan secara intuitif. Evaluasi ini memastikan bahwa *cluster* yang terbentuk memiliki koherensi internal yang baik dan pemisahan antar *cluster* yang jelas.

Pada tahap evaluasi, Sum of Squared Errors

(SSE) digunakan untuk mengevaluasi kualitas klasterisasi. SSE menunjukkan seberapa dekat data dalam sebuah klaster dengan pusat atau centroidnya; semakin rendah nilai SSE, semakin baik gambaran data klaster tersebut. Untuk algoritma K-Means, rumus SSE adalah sebagai berikut:

$$SEE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} ||x_j^{(i)} - c_i||^2$$

Di mana k adalah jumlah klaster, n_i adalah jumlah data pada klaster ke- i, $x_j^{(i)}$ adalah data ke-j pada klaster ke-i, dan c_i adalah centroid klaster ke-i.

Untuk algoritma *Fuzzy C-Means*, perhitungan SSE sedikit berbeda karena mempertimbangkan derajat keanggotaan (u_{ij})dari setiap data terhadap kluster. Dalam hal ini, setiap data tidak hanya menjadi anggota satu klaster, tetapi dapat memiliki hubungan dengan lebih dari satu klaster berdasarkan derajat keanggotaannya. Rumus SSE untuk *Fuzzy C-Means* adalah $SEE = \sum_{i=1}^{k} \sum_{j=1}^{N} u_{ij}^{m} \cdot ||x_j - c_i||^2$, di mana N adalah jumlah total data, u_{ij} adalah derajat keanggotaan data ke-j pada klaster ke-i, dan m adalah parameter fuzzy yang biasanya bernilai 2. Pendekatan ini lebih fleksibel dibandingkan K-Means karena mampu menangkap tumpang tindih antar-klaster.

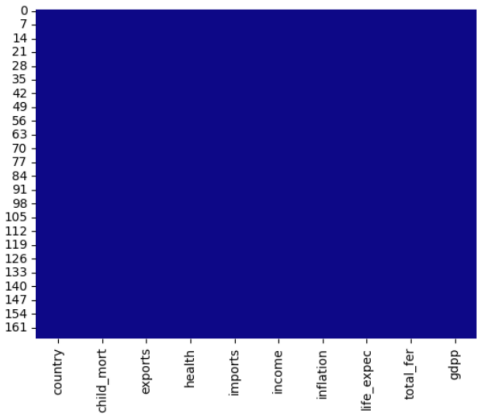
Untuk evaluasi model, metrik SSE sangat penting karena memberikan gambaran tentang seberapa dekat data dalam sebuah klaster. Nilai SSE yang rendah menunjukkan bahwa data dalam klaster memiliki penyebaran yang rendah dan berada di dekat centroid. Oleh karena itu, penilaian ini membantu menjamin kualitas klasterisasi yang dihasilkan oleh algoritma *K-Means* dan C-Means yang tidak jelas.

III. HASIL DAN PEMBAHASAN

Dilakukan pengolahan data terhadap dataset negara yang mengandung berbagai indikator sosial-ekonomi, seperti angka kematian anak (*child_mort*), ekspor (*exports*), kesehatan (*health*), impor (*imports*), pendapatan (*income*), inflasi (*inflation*), harapan hidup (*life_expec*), fertilitas total (total_fer), dan Produk Domestik Bruto per kapita (gdpp). Berdasarkan total data yang tersedia dalam dataset ini, yaitu sejumlah 202 negara, data akan diproses untuk diidentifikasi nilai *cluster*-nya melalui dua metode utama, yaitu *K-Means* dan *Fuzzy C-Means*.

Data diperiksa lebih rinci untuk memahami struktur dan karakteristik dasar dari dataset. Langkah pertama adalah menampilkan bentuk data untuk memastikan ukuran dan cakupan informasi yang dimiliki. Pada dataset ini, terlihat bahwa data menunjukkan adanya 202 baris data negara dengan 10 kolom indikator sosial-ekonomi. Langkah selanjutnya adalah memastikan bahwa tidak terdapat nilai kosong (*null*) atau data duplikat yang dapat memengaruhi hasil analisis.

Hasil yang diperoleh menunjukkan bahwa tidak ada data yang memiliki nilai kosong dan data duplikat dalam dataset. Visualisasi *heatmap* juga digunakan untuk memastikan tidak adanya nilai kosong (*null*) pada dataset.

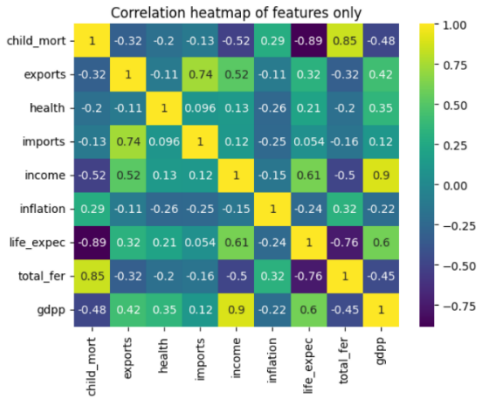


Gambar 2. Visualisasi *Heatmap*

Dilakukan pemisahan antara fitur kategorikal dan numerik untuk memudahkan pemrosesan lebih lanjut. Fitur kategorikal, yang berisi informasi non-numerik seperti nama negara, akan disimpan terpisah dari fitur numerik yang mewakili indikator sosial-ekonomi.

Dengan pemisahan ini, fitur numerik dapat diproses untuk analisis klaster tanpa melibatkan fitur kategorikal, yang hanya digunakan sebagai referensi untuk mengidentifikasi negara dalam setiap klaster. Langkah ini juga membantu memastikan bahwa algoritma klaster hanya berfokus pada karakteristik sosial-ekonomi setiap negara, sesuai tujuan analisis.

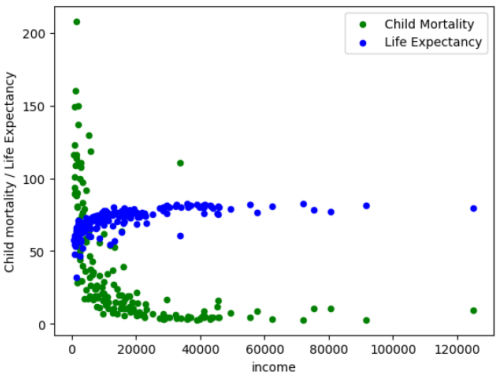
Dengan menggunakan heatmap korelasi, dapat terlihat bagaimana setiap fitur numerik saling berkaitan. Heatmap ini disusun berdasarkan koefisien korelasi Pearson, yang menunjukkan kekuatan dan arah hubungan antara setiap pasangan indikator.



Gambar 3. Visualisasi *Heatmap* Korelasi

Warna yang lebih terang menunjukkan korelasi yang lebih kuat, sedangkan warna yang lebih gelap menunjukkan korelasi yang lebih lemah. Misalnya, terdapat korelasi negatif yang kuat antara *child_mort* dan *life_expec* (-0.89), yang menunjukkan bahwa negara dengan angka kematian anak yang tinggi cenderung memiliki harapan hidup yang lebih rendah. Sebaliknya, *income* dan gdpp memiliki korelasi positif yang cukup tinggi (0.9), menunjukkan hubungan langsung antara pendapatan dan Produk Domestik Bruto per kapita.

Untuk memperoleh pemahaman lebih dalam tentang hubungan antara pendapatan dan indikator sosial-ekonomi lainnya, dilakukan visualisasi dengan grafik sebar (scatter plot) yang membandingkan fitur income dengan child_mort dan life_expec. Visualisasi ini memperlihatkan bagaimana tingkat pendapatan berhubungan dengan angka kematian anak dan harapan hidup di setiap negara.

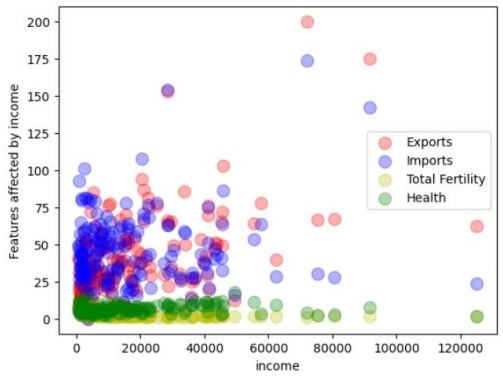


Gambar 3. Visualisasi Grafik Sebar

Dalam grafik ini, titik hijau mewakili hubungan antara pendapatan (*income*) dan angka kematian anak (*child_mort*), sementara titik biru menunjukkan hubungan antara pendapatan dan harapan hidup (*life_expec*). Dari visualisasi ini, dapat dilihat bahwa negara dengan pendapatan lebih tinggi cenderung memiliki angka kematian anak yang lebih rendah dan harapan hidup yang lebih tinggi, sesuai dengan tren global yang menunjukkan hubungan positif antara pendapatan dan kualitas hidup.

Untuk memahami lebih lanjut pengaruh pendapatan (*income*) terhadap berbagai indikator sosial-ekonomi lainnya, dibuat visualisasi *scatter plot* yang menghubungkan income dengan *exports*, *imports*, total_fer, dan *health*. Setiap

indikator diwakili oleh warna dan label berbeda, serta ditampilkan dengan tingkat transparansi (alpha) untuk memudahkan interpretasi.

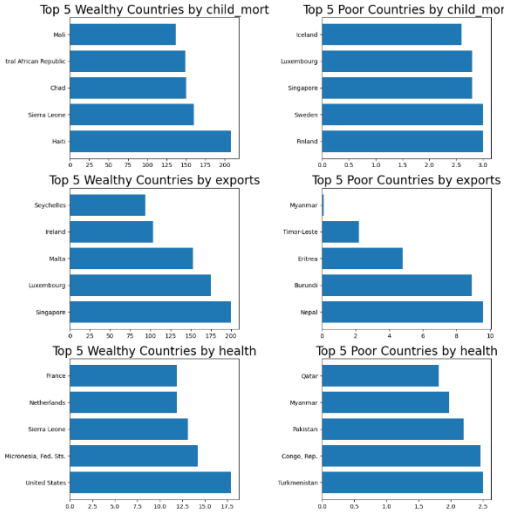


Gambar 4. Visualisasi *Scatter Plot*

- Titik merah mewakili hubungan antara income dan exports (ekspor).
- Titik biru menunjukkan hubungan income dengan imports (impor).
- Titik kuning memperlihatkan hubungan antara income dan total_fer (fertilitas total).
- Titik hijau menggambarkan hubungan antara income dan health (kesehatan).

Dari grafik ini, dapat dilihat pola-pola umum yang menunjukkan bahwa pendapatan suatu negara sering kali berhubungan dengan angka ekspor, impor, fertilitas, dan pengeluaran kesehatan. Negara dengan pendapatan lebih tinggi cenderung memiliki indikator yang berbeda dibandingkan negara dengan pendapatan lebih rendah, seperti dalam aspek fertilitas dan pengeluaran untuk kesehatan.

Untuk memahami perbedaan antar-negara berdasarkan berbagai indikator sosial-ekonomi, dibuat visualisasi perbandingan antara lima negara dengan nilai tertinggi (terkaya) dan lima negara dengan nilai terendah (termiskin) pada setiap fitur numerik. Visualisasi ini memperlihatkan perbedaan signifikan dalam setiap indikator yang mencerminkan kondisi sosial-ekonomi di berbagai negara.



Gambar 5. Visualisasi Grafik Batang

- Grafik batang horizontal di sebelah kiri menunjukkan lima negara teratas dengan nilai tertinggi untuk setiap fitur, seperti pendapatan (*income*), harapan hidup (*life_expec*), dan indikator lainnya.
- Grafik batang horizontal di sebelah kanan menunjukkan lima negara terbawah dengan nilai terendah untuk masing-masing indikator.

Dengan cara ini, kita dapat melihat perbedaan antara negara-negara kaya dan miskin dalam hal indikator sosial-ekonomi, yang menjadi penting dalam analisis klasterisasi untuk mengidentifikasi pola-pola pengelompokan.

Pada tahap pra-pemrosesan data, dilakukan modifikasi pada dataset untuk memfokuskan analisis secara eksklusif pada nilai indikator sosial-ekonomi. Nama negara (*'country'*) dihapus dari dataframe, sehingga dataset hanya berisi kolom-kolom indikator sosial-ekonomi yang akan digunakan dalam proses klastering. Langkah ini memastikan bahwa hanya data numerik dari setiap indikator yang terlibat dalam penghitungan jarak klaster.

Agar setiap indikator sosial-ekonomi berada dalam skala yang sama, data pada masing-masing kolom di-normalisasi menggunakan metode *Min-Max Scaling.* Normalisasi ini penting untuk menghindari dominasi dari satu atribut yang memiliki rentang nilai lebih besar dibandingkan atribut lain, yang dapat mempengaruhi hasil klastering. Dengan *Min-Max Scaling*, semua nilai indikator diubah ke dalam rentang 0 hingga 1, sehingga perbandingan antar indikator menjadi lebih seimbang.

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0.426485	0.049482	0.358608	0.257765	0.008047	0.126144	0.475345	0.736593	0.003073
1	0.068160	0.139531	0.294593	0.279037	0.074933	0.080399	0.871795	0.078864	0.036833
2	0.120253	0.191559	0.146675	0.180149	0.098809	0.187691	0.875740	0.274448	0.040365
3	0.566699	0.311125	0.064636	0.246266	0.042535	0.245911	0.552268	0.790221	0.031488
4	0.037488	0.227079	0.262275	0.338255	0.148652	0.052213	0.881657	0.154574	0.114242
...	...	...	...	...	...	...	...	...	...
162	0.129503	0.232582	0.213797	0.302609	0.018820	0.063118	0.609467	0.370662	0.026143
163	0.070594	0.142032	0.192666	0.100809	0.127750	0.463081	0.854043	0.208202	0.126650
164	0.100779	0.359651	0.312617	0.460715	0.031200	0.150725	0.808679	0.126183	0.010299
165	0.261441	0.149536	0.209447	0.197397	0.031120	0.257000	0.698225	0.555205	0.010299
166	0.391918	0.184556	0.253574	0.177275	0.021473	0.168284	0.392505	0.670347	0.011731

Gambar 6. Metode *Min-Max Scaling*

Data juga ditransformasikan menggunakan metode *Standard Scaling* untuk memastikan distribusi setiap indikator memiliki rata-rata nol dan standar deviasi satu. Standardisasi ini membantu menyeimbangkan data dengan menghilangkan pengaruh skala asli masing-masing indikator, yang penting terutama saat data beragam secara signifikan dalam rentang nilai.

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	1.291532	-1.138280	0.279088	-0.082455	-0.808245	0.157336	-1.619092	1.902882	-0.679180
1	-0.538949	-0.479658	-0.097016	0.070837	-0.375369	-0.312347	0.647866	-0.859973	-0.485623
2	-0.272833	-0.099122	-0.966073	-0.641762	-0.220844	0.789274	0.670423	-0.038404	-0.465376
3	2.007808	0.775381	-1.448071	-0.165315	-0.585043	1.387054	-1.179234	2.128151	-0.516268
4	-0.695634	0.160668	-0.286894	0.497568	0.101732	-0.601749	0.704258	-0.541946	-0.041817
...	...	...	...	...	...	...	...	...	...
162	-0.225578	0.200917	-0.571711	0.240700	-0.738527	-0.489784	-0.852161	0.365754	-0.546913
163	-0.526514	-0.461363	-0.695862	-1.213499	-0.033542	3.616865	0.546361	-0.316678	0.029323
164	-0.372315	1.130305	0.008877	1.380030	-0.658404	0.409732	0.286958	-0.661206	-0.637754
165	0.448417	-0.406478	-0.597272	-0.517472	-0.658924	1.500916	-0.344633	1.140944	-0.637754
166	1.114951	-0.150348	-0.338015	-0.662477	-0.721358	0.590015	-2.092785	1.624609	-0.629546

Gambar 7. Metode *Standard Scaling*

Untuk mengurangi kompleksitas data tanpa kehilangan informasi yang signifikan, dilakukan *Principal Component Analysis* (PCA) pada data yang telah dinormalisasi dengan *Min-Max Scaling*. PCA membantu dalam menemukan pola dalam data dengan mereduksi jumlah dimensi dan mempertahankan variabilitas utama. Dalam analisis ini, PCA digunakan untuk mengidentifikasi komponen utama yang dapat menjelaskan variasi data terbesar.

Setelah menerapkan PCA pada data yang telah dinormalisasi, langkah selanjutnya adalah melakukan analisis yang sama pada dataset yang telah terstandarisasi. Dengan menggunakan PCA pada data ini, kita dapat mengevaluasi seberapa baik komponen utama yang dihasilkan dapat menjelaskan variansi dalam data yang sudah berada dalam skala standar. Tujuan dari langkah ini adalah untuk melihat apakah ada perbedaan signifikan dalam variansi yang dijelaskan oleh komponen utama ketika menggunakan data terstandarisasi dibandingkan dengan data normalisasi.

Setelah melakukan PCA dan memvisualisasikan hasilnya, langkah berikutnya adalah memilih komponen utama yang paling relevan untuk analisis lebih lanjut. Dalam kasus ini, kita akan menggunakan lima komponen utama yang dihasilkan dari analisis PCA, dengan mengabaikan komponen yang kurang signifikan. Hal ini penting untuk memastikan bahwa analisis klaster yang dilakukan nantinya akan lebih akurat dan fokus pada informasi yang paling berpengaruh.

	PC1	PC2	PC3	PC4	PC5
0	-2.913025	0.095621	-0.718118	1.005255	-0.158310
1	0.429911	-0.588156	-0.333486	-1.161059	0.174677
2	-0.285225	-0.455174	1.221505	-0.868115	0.156475
3	-2.932423	1.695555	1.525044	0.839625	-0.273209
4	1.033576	0.136659	-0.225721	-0.847063	-0.193007
...	...	...	...	...	...
162	-0.820631	0.639570	-0.389923	-0.706595	-0.395748
163	-0.551036	-1.233886	3.101350	-0.115311	2.082581
164	0.498524	1.390744	-0.238526	-1.074098	1.176081
165	-1.887451	-0.109453	1.109752	0.056257	0.618365
166	-2.864064	0.485998	0.223167	0.816364	-0.274068

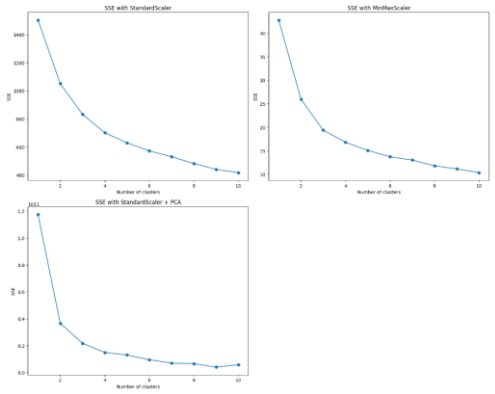
Gambar 8. *Principal Component Analysis* (PCA)

Untuk lebih memahami kontribusi masing-masing komponen utama yang dihasilkan dari PCA, perlu mengevaluasi rasio variansi yang dijelaskan oleh setiap komponen serta total variansi kumulatif yang dijelaskan. Analisis ini memberikan wawasan tentang seberapa baik komponen utama tersebut menggambarkan data, yang penting untuk keputusan terkait jumlah komponen yang akan digunakan dalam analisis klaster.

	PC	explained_variance_ratio	Cumulative_variance
0	PC1	0.459517	0.459517
1	PC2	0.171816	0.631334
2	PC3	0.130043	0.761376
3	PC4	0.110532	0.871908
4	PC5	0.073402	0.945310
5	PC6	0.024842	0.970152
6	PC7	0.012604	0.982757
7	PC8	0.009813	0.992569
8	PC9	0.007431	1.000000

Gambar 9. Variansi Kumulatif

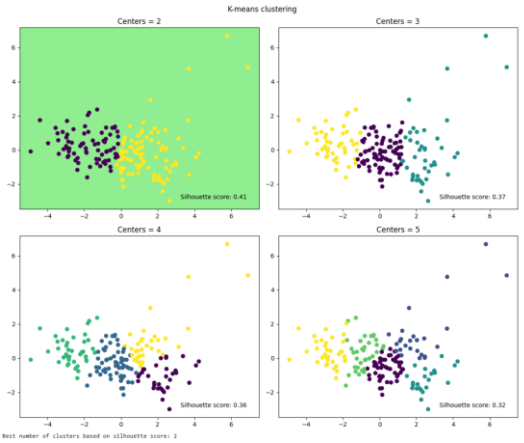
Melakukan evaluasi jumlah kluster yang optimal untuk analisis kluster menggunakan algoritma *K-Means* dengan memplot *nilai Sum of Squared Errors* (SSE) untuk berbagai jumlah kluster. SSE adalah metrik yang digunakan untuk mengukur seberapa baik model klustering memisahkan data. Semakin rendah nilai SSE, semakin baik kluster tersebut dalam menggambarkan data.



Gambar 10. Evaluasi Kluster Menggunakan K-Means dan Sum of Squared Errors (SSE)

Dengan melakukan analisis ini, dapat mengidentifikasi jumlah kluster yang optimal berdasarkan nilai SSE. Dalam mencari titik di mana penurunan SSE mulai melambat, dikenal sebagai "*elbow point*". Hasil ini memberikan informasi penting untuk pemilihan jumlah kluster yang tepat dalam analisis kluster.

Pada tahap ini, penerapkan metode *Fuzzy C-Means* (FCM) untuk melakukan analisis kluster dan mengevaluasi hasilnya dengan menggunakan *Silhouette Score*. Metode FCM adalah varian dari *K-Means* yang memperbolehkan data menjadi bagian dari beberapa kluster dengan derajat keanggotaan yang berbeda, menjadikannya lebih fleksibel dalam situasi di mana batas antar kluster tidak jelas.



Gambar 11. Analisis Kluster Menggunakan *Fuzzy C-Means*

dan Evaluasi *Silhouette Score*

Analisis ini memungkinkan untuk tidak hanya menentukan jumlah kluster yang optimal tetapi juga memahami bagaimana data terbagi dalam kluster yang berbeda dengan menggunakan pendekatan yang lebih fleksibel daripada *K-Means*. Nilai *Silhouette Score* memberikan indikator yang jelas tentang kualitas pemisahan kluster, membantu dalam pengambilan keputusan lebih lanjut untuk analisis kluster.

Penerapkan algoritma *K-Means* untuk melakukan klustering pada dataset yang telah dinormalisasi, terstandarisasi, dan dataset asli yang telah diproses menggunakan PCA. Proses ini memungkinkan untuk mengidentifikasi pola dan struktur dalam data berdasarkan fitur-fitur yang ada.

Selanjutnya melakukan pemeriksaan terhadap dataset untuk memastikan bahwa semua kolom yang ada berisi data numerik. Proses ini penting, terutama dalam konteks analisis kluster dan teknik statistik lainnya, yang biasanya memerlukan data numerik untuk pengolahan dan perhitungan.

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553	2
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090	1
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460	1
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530	2
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200	1
...	...	...	...	...	...	...	...	...	...	...	...
162	Vanuatu	29.2	46.6	5.25	52.7	2950	2.62	63.0	3.50	2970	1
163	Venezuela	17.1	28.5	4.91	17.6	16500	45.90	75.4	2.47	13500	1
164	Vietnam	23.3	72.0	6.84	80.2	4490	12.10	73.1	1.95	1310	1
165	Yemen	56.3	30.0	5.18	34.4	4480	23.60	67.5	4.67	1310	2
166	Zambia	83.1	37.0	5.89	30.9	3280	14.00	52.0	5.40	1460	2

Gambar 12. Pemeriksaan Kolom Non-Numerik dalam Dataset

Melakukan pemilihan kolom dari dataset yang akan digunakan dalam analisis lebih lanjut. Dengan mengeluarkan kolom yang tidak relevan, kita dapat fokus pada fitur-fitur yang berkontribusi terhadap analisis kluster.

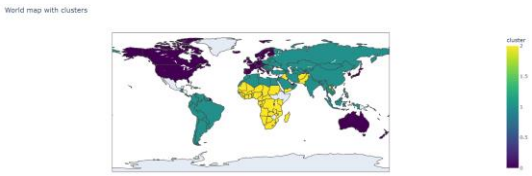
Di tahap selanjutnya, menyusun tabel pivot yang bertujuan untuk meringkas data berdasarkan kluster yang telah terbentuk sebelumnya. Tabel pivot adalah alat yang sangat berguna untuk melihat ringkasan statistik dari berbagai fitur dalam dataset.

cluster	child_mort	exports	gdpp	health	imports	income	inflation	life_expec	total_fer
0	5.000000	58.738889	42494.444444	8.807778	51.491667	45672.222222	2.671250	80.127778	1.752778
1	21.927381	40.243917	6486.452381	6.200952	47.473404	12305.595238	7.600905	72.814286	2.307500
2	92.961702	29.151277	1922.382979	6.388511	42.323404	3942.404255	12.019681	59.187234	5.008085

Gambar 12. Tabel Pivot

Setelah membangun tabel pivot untuk meringkas data, langkah selanjutnya adalah mengidentifikasi negara-negara yang termasuk dalam setiap kluster. Proses ini sangat penting untuk memahami karakteristik geografis dan sosial-ekonomi dari masing-masing kluster yang telah terbentuk.

Setelah mengidentifikasi negara-negara dalam setiap kluster, langkah berikutnya adalah memvisualisasikan hasil klustering dalam bentuk peta. Visualisasi ini akan membantu memahami distribusi geografis negara-negara berdasarkan kluster yang terbentuk dari analisis sebelumnya.



Gambar 13. Visualisasi Peta Dunia dengan Kluster

Dari visualisasi peta dunia di atas, dapat dilihat bahwa negara-negara dengan karakteristik sosial-ekonomi yang mirip cenderung berada dalam klaster yang sama, ditunjukkan dengan warna yang serupa. Hal ini menunjukkan adanya pola distribusi dan kesamaan dalam indikator sosial-ekonomi antar negara di setiap klaster.

IV. KESIMPULAN

Penelitian ini menganalisis pengelompokan negara berdasarkan indikator sosial-ekonomi seperti angka kematian anak, pendapatan per kapita, harapan hidup, dan inflasi menggunakan

algoritma *Fuzzy C-Means* (FCM) dan *K-Means*, Nilai *SSE* terendah diperoleh dengan *MinMaxScaler* pada klaster ke-10 sebesar 10.30, sementara *Standard Scaler* menghasilkan *SSE* sebesar 416.04 pada klaster ke-10. Untuk data yang direduksi menggunakan *PCA*, *SSE* mencapai 5771653036.34 pada klaster ke-10. Berdasarkan *Silhoutte Score*, jumlah klaster optimal adalah 2 dengan skor 0.41, yang mengindikasikan pemisahan klaster yang baik. *FCM* dipilih karena memungkinkan setiap negara menjadi bagian dari kelompok yang sesuai dengan kemiripan karakterisitiknya, semenatara *K-Means* hanya menetapkan satu kelompok per negara. Hasil penelitian menunjukkan bahwa *FCM* menghasilkan kelompok yang lebih akuran dan fleksibel dalam menggambarkan pola sosial-ekonomi antar negara. Dengan evaluasi menggunakan metrik *Sum of Squared Errors (SSE)* dan *Silhoutte Score*, penelitian ini mendukung bahwa *FCM* lebih efektif dalam mengelompokkan negara berdasarkan indikator sosial-ekonomi.

REFERENSI

Aisah, S. N., Nurcahyani, A., & Rini, D. C. (2022). Implementasi Fuzzy C–Means Clustering (Fcm) Pada Pemetaan Daerah Potensi Transmigrasi Di Jawa Timur. *Jurnal Teknik Informatika UNIKA Santo Thomas*, *07*, 33–40. https://doi.org/10.54367/jtiust.v7i1.1841

Anwar, A. N. (2023). Implementasi Fuzzy C-Mean (Fcm). *Jurnal Ilmu Komputer JIK*, *VI*(01).

Atiqah, N., Hamzah, B., Tun, U., & Onn Malaysia, H. (2018). *Malaysia Household Incomes Classification Prediction With K-Means Clustering and Fuzzy Inference System*. *August*.

Hussain, I., Sinaga, K. P., & Yang, M. S. (2023). Unsupervised Multiview Fuzzy C-Means Clustering Algorithm. *Electronics (Switzerland)*, *12*(21). https://doi.org/10.3390/electronics12214467

Putriana, U., Setyawan, Y., & Noeryanti. (2016). Metode Cluster Analysis Untuk Pengelompokan Kabupaten/Kota Di Provinsi Jawa Tengah Berdasarkan Variabel Yang Mempengaruhi Kemiskinan Pada Tahun 2013. *Jurnal Statistika Industri Dan Komputasi*, *1*(1), 38–52.

Qaadan, S., Alshare, A., & Popp, A. (2024). *Prediction of Lithium-Ion Battery Health Using GRU-BPP*. 1–23.

Rahakbauw, D. L., Ilwaru, V. Y. I., & Hahury, M. H. (2017). Implementasi Fuzzy C-Means Clustering Dalam Penentuan Beasiswa. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *11*(1), 1–12. https://doi.org/10.30598/barekengvol11iss1pp1-12

Saatchi, R. (2024). Fuzzy Logic Concepts, Developments and Implementation. *Information (Switzerland)*, *15*(10). https://doi.org/10.3390/info15100656

Sanusi, W., Zaky, A., & Afni, B. N. (2020). Analisis Fuzzy C-Means dan Penerapannya Dalam Pengelompokan Kabupaten/Kota di Provinsi Sulawesi Selatan Berdasarkan Faktor-faktor Penyebab Gizi Buruk. *Journal of Mathematics, Computations, and Statistics*, *2*(1), 47. https://doi.org/10.35580/jmathcos.v2i1.12458

Singh, P., Shamseldin, A. Y., Melville, B. W., & Wotherspoon, L. (2023). Development of statistical downscaling model based on Volterra series realization, principal components and ridge regression. *Modeling Earth Systems and Environment*, *9*(3), 3361–3380. https://doi.org/10.1007/s40808-022-01649-3

Sun, C., Shao, Q., Zhou, Z., & Zhang, J. (2024). An Enhanced FCM Clustering Method Based on Multi-Strategy Tuna Swarm Optimization. *Mathematics*, *12*(3), 0–16. https://doi.org/10.3390/math12030453

Suryadi, A. (2015). Sistem Pengenalan Wajah Menggunakan Metode Principal Component Analysis (PCA) Dengan Algoritma Fuzzy C-Means (FCM). *Mosharafa: Jurnal Pendidikan Matematika*, *4*(2), 58–65. https://doi.org/10.31980/mosharafa.v4i2.329

Zhou, W., Zhong, L., Kang, W., & Xu, Y. (2024). *Modal Parameter Identification of Electric Spindles Based on Covariance-Driven Stochastic Subspace*.