PENGELOMPOKAN NEGARA BERDASARKAN INDIKATOR PEMBANGUNAN GLOBAL MENGGUNAKAN METODE PCA DAN CLUSTERING K-MEANS TAHUN 2000-2020

Mika Valentino [1] Yosia Sipahutar [2] Muhammad Farhan [3]

[1] [2] [3] [4] [5] Teknik Informatika Universitas Tarumanagara
Jl. Letjen S. Parman St No.1, Jakarta 11440 Indonesia
email : [1] mika.535230050@stu.untar.ac.id, [2] yosia.535230117@stu.untar.ac.id,
[3] muhammad.535230146@stu.untar.ac.id

ABSTRAK

Penelitian ini bertujuan untuk mengelompokkan negara-negara di dunia berdasarkan indikator pembangunan global dengan menggunakan metode Principal Component Analysis (PCA) dan K-Means Clustering. Data yang digunakan bersumber dari Global Development Indicators tahun 2000–2020 yang mencakup dimensi ekonomi, sosial, dan digital. Tahapan analisis diawali dengan pra-pemrosesan data, termasuk interpolasi dan imputasi nilai hilang, standarisasi menggunakan Z-score, serta transformasi arah untuk variabel berdampak negatif. PCA diterapkan untuk mereduksi kompleksitas data, menghasilkan dua komponen utama (PC1 dan PC2) yang menjelaskan lebih dari 80% variansi data. Selanjutnya, K-Means digunakan untuk melakukan clustering, dengan evaluasi melalui silhouette coefficient yang menunjukkan nilai optimal sebesar 0,71 pada empat klaster. Hasil analisis menunjukkan bahwa kombinasi PCA dan K-Means efektif dalam mengidentifikasi pola pembangunan global dan mengelompokkan negara berdasarkan karakteristik pembangunan yang serupa. Temuan ini dapat digunakan sebagai dasar dalam penyusunan kebijakan pembangunan dan strategi kerja sama internasional yang lebih terarah.

Key words

Principal Component Analysis, *K-Means Clustering*, pembangunan global, indikator pembangunan

1. Pendahuluan

1.1. Latar Belakang

Pembangunan global adalah konsep multidimensional yang mencakup transformasi ekonomi, sosial, lingkungan, dan institusional sebagai segala upaya dalam meningkatkan kesejahteraan. Indikator-indikator pembangunan menjadi patokan penting dalam mengukur tingkat kemajuan antarnegara[1]. Produk Domestik Bruto (PDB) per kapita, sebagai indikator ekonomi, pada awalnya menjadi petunjuk penting dalam mengukur kesejahteraan ekonomi dan daya beli masyarakat [2]. Namun, konsep pembangunan telah berkembang dari hanya menekankan pada pertumbuhan ekonomi menjadi turut mempertimbangkan aspek kehidupan manusia [3]. Oleh karena itu, pembangunan global tidak hanya diukur oleh indikator ekonomi seperti PDB per kapita, melainkan juga mencakup dimensi-dimensi lainnya seperti tata kelola pemerintahan, kesehatan, lingkungan, dan pendidikan [4]. Dengan demikian, pembangunan global adalah proses kompleks dan menyeluruh yang memerlukan pendekatan holistik.

Indikator pembangunan global adalah instrumen penting dalam mengukur kemajuan negara-negara di dunia. Indikator-indikator tersebut tidak hanya lagi berfokus pada aspek ekonomi saja, melainkan juga mencakup dimensi sosial dan lingkungan demi memberikan gambaran menyeluruh tentang kesejahteraan masyarakat [1]. Meskipun demikian, pengukuran pembangunan global masih mengalami tantangan dalam standarisasi data lintas negara, sehingga kerangka kerja yang lebih konsisten sangat diharapkan [6]. Oleh karena itu, indikator pembangunan global yang menyeluruh sangat penting terutama dalam proses perancangan kebijakan berkelanjutan.

1.2. Tujuan dan Kegunaan

Tujuan dari analisis ini adalah membangun model pengelompokan negara-negara di dunia secara komprehensif berdasarkan indikator pembangunan multidimensional menggunakan pendekatan analisis multivariat . Secara spesifik, tujuan yang ingin dicapai antara lain:

1. Mengelompokkan negara-negara di dunia berdasarkan kemiripan indikator pembangunan global menggunakan kombinasi metode *Principal Component Analysis* (PCA) dan *clustering analysis*.
2. Mengidentifikasi pola atau kelompok negara dengan karakteristik pembangunan yang serupa melalui analisis komponen utama untuk mengungkap dimensi-dimensi tersembunyi dalam data pembangunan global periode 2000-2020.

3. Menganalisis jarak dan kemiripan antar kelompok hasil *clustering* untuk memahami jarak pembangunan antar negara dan mengidentifikasi faktor-faktor yang membedakan antar kelompok negara.

Kegunaan bersifat jangka panjang dalam penerapan hasil analisis ini adalah untuk mendukung pengambilan keputusan dalam proses perumusan kebijakan pembangunan global dan kerja sama internasional yang lebih efektif. Berikut adalah beberapa manfaat spesifik dari analisis ini antara lain:

1. Mendukung proses pengambilan keputusan secara strategis dalam alokasi bantuan pembangunan internasional, program kerja sama internasional, dan transfer teknologi antar negara dengan mempertimbangkan kemiripan profil pembangunan.
2. Menyediakan model klasifikasi yang dapat digunakan oleh peneliti dan praktisi pembangunan dalam menganalisis tren pembangunan global, dan merancang strategi pembangunan yang disesuaikan dengan karakteristik spesifik setiap kelompok negara.

2. Landasan Teori

2.1. Konsep Pembangunan Global

Berdasarkan kerangka kerja pembangunan, terdapat beberapa dimensi yang menjadi kunci dalam pembangunan global. Dimensi-dimensi ini saling berkaitan dan berpengaruh satu sama lain untuk membentuk tingkat pembangunan negara secara keseluruhan. Dengan pendekatan multidimensional terhadap pembangunan, dapat dilakukan analisis yang lebih komprehensif terhadap faktor-faktor yang mempengaruhi kemajuan bangsa [2]. Dalam konteks ini, terdapat beberapa pilar utama yang perlu dikaji secara mendalam, yaitu tata kelola pemerintahan, pendidikan-kesehatan, sosial-ekonomi, dan infrastruktur.

2.1.1 Tata Kelola Pemerintahan

Kualitas tata kelola pemerintahan berperan penting dalam mempengaruhi efektivitas pembangunan ekonomi suatu negara [3]. Tata kelola pemerintahan yang baik akan menyediakan lingkungan yang kondusif bagi investasi dan inovasi untuk mendorong pertumbuhan ekonomi berkelanjutan [4]. Transparansi, akuntabilitas, dan efisiensi menjadi poin-poin penting dalam penyelenggaraan pemerintahan yang baik [5]. Stabilitas politik dan kepastian hukum dari tata kelola pemerintahan yang baik memberikan kepercayaan bagi investor domestik maupun asing untuk melakukan investasi jangka panjang. Pada akhirnya, tata kelola pemerintahan yang baik akan mendorong pertumbuhan ekonomi berkelanjutan dan meningkatkan kesejahteraan masyarakat.

2.1.2 Pendidikan Dan Kesehatan

Faktor pendidikan dan kesehatan menjadi pilar utama dalam pembangunan manusia. Partisipasi masyarakat dalam jenjang pendidikan tingkat menengah mencerminkan kemudahan masyarakat dalam mengakses pendidikan. Hal ini menjadi tolok ukur untuk menilai seberapa baik kualitas pendidikan formal di suatu negara [6]. Rasio antara belanja pendidikan dan kesehatan menunjukkan seberapa seimbang investasi pemerintah dalam membangun sumber daya manusia yang berkualitas. Dalam sektor kesehatan, harapan hidup dan angka kematian anak menjadi indikator utama dalam menilai kesejahteraan dan kualitas layanan kesehatan publik. Angka kematian anak dipengaruhi oleh kualitas layanan kesehatan dasar, akses terhadap air bersih, sanitasi, dan gizi masyarakat [7]. Kedua faktor kesehatan tersebut dipengaruhi oleh besarnya belanja kesehatan yang menunjukkan besarnya komitmen pemerintah dalam meningkatkan kualitas sektor kesehatan.

2.1.3 Sosial-Ekonomi

Faktor sosial-ekonomi seperti Indeks Pembangunan Manusia (HDI) dan Indeks Peluang Ekonomi memberikan gambaran tentang tingkat pembangunan suatu negara. HDI mengukur tingkat pembangunan manusia secara keseluruhan dengan mempertimbangkan hal-hal seperti standar hidup, kesehatan, dan pendidikan [8]. Dengan memaksimalkan kedua indeks tersebut, akan diperoleh pertumbuhan ekonomi yang menguntungkan semua orang, bukan hanya kelompok tertentu.

2.1.4 Ekonomi Digital

Kesiapan digital menjadi kunci untuk daya saing ekonomi dan pembangunan negara di era transformasi teknologi global. Kesiapan ini bergantung pada infrastruktur digital, yang mencakup akses masyarakat terhadap teknologi informasi dan komunikasi [14]. Jumlah langganan telepon yang tersedia untuk setiap seratus orang menunjukkan tingkat adopsi teknologi komunikasi penting. Digitalisasi membutuhkan akses listrik sebagai sumber energi untuk semua perangkat dan infrastruktur digital. Penggunaan energi per kapita mencerminkan kemampuan suatu negara dalam mendukung infrastruktur jaringan digital yang menggunakan konsumsi energi yang besar.

Kemampuan ekonomi dan teknologi suatu negara menentukan kapasitas negara dalam mengadopsi dan memanfaatkan teknologi digital. Kemampuan masyarakat dalam mengakses segala fasilitas digital dipengaruhi oleh besarnya GDP per kapita [15]. Pertumbuhan ekonomi riil menunjukkan kemampuan negara untuk menyediakan sumber daya untuk mengembangkan dan membangun infrastruktur digital. Indeks Konektivitas Digital menggambarkan kualitas dari fasilitas internet yang disediakan, baik itu dari kecepatan maupun jangkauan konektivitasnya, konektivitas digital yang baik dapat membantu kegiatan ekonomi dan sosial.

2.1.5 Infrastruktur

Infrastruktur, yang mencakup infrastruktur digital dan infrastruktur dasar seperti akses air bersih, merupakan pilar fundamental dalam pembangunan global di era transformasi teknologi. Kesiapan digital menjadi kunci untuk daya saing ekonomi dan pembangunan negara di era transformasi teknologi global [16]. Infrastruktur digital mencakup berbagai komponen yang saling terkait, mulai dari akses masyarakat terhadap teknologi informasi dan komunikasi hingga ketersediaan energi yang mendukung operasional sistem digital [17]. Jumlah langganan telepon yang tersedia untuk setiap seratus orang menunjukkan tingkat penetrasi teknologi komunikasi, yang menjadi indikator penting tingkat adopsi teknologi dalam masyarakat. Selain itu, digitalisasi membutuhkan akses listrik yang stabil sebagai sumber energi untuk semua perangkat dan infrastruktur digital, sehingga penggunaan energi per kapita mencerminkan kemampuan suatu negara dalam mendukung infrastruktur jaringan digital yang memerlukan konsumsi energi yang besar. Di sisi lain, akses terhadap air bersih merupakan infrastruktur dasar yang tidak dapat dipisahkan dari pembangunan berkelanjutan, karena ketersediaan air bersih mempengaruhi kesehatan masyarakat, produktivitas ekonomi, dan bahkan mendukung operasional infrastruktur teknologi [18].

Kemampuan ekonomi dan teknologi suatu negara menentukan kapasitas negara dalam mengadopsi dan memanfaatkan teknologi digital secara optimal. Kemampuan masyarakat dalam mengakses segala fasilitas digital dipengaruhi oleh besarnya GDP per kapita, yang mencerminkan daya beli dan investasi pemerintah dalam infrastruktur teknologi [19]. Pertumbuhan ekonomi riil menunjukkan kemampuan negara untuk menyediakan sumber daya yang diperlukan demi mengembangkan dan membangun infrastruktur digital yang berkelanjutan. Indeks Konektivitas Digital menggambarkan kualitas dari fasilitas internet yang disediakan, baik dari aspek kecepatan maupun jangkauan konektivitasnya, di mana konektivitas digital yang baik dapat memfasilitasi kegiatan ekonomi dan sosial serta mendorong inovasi dalam berbagai sektor pembangunan [20]. Sementara itu, persentase populasi dengan akses terhadap air bersih yang dikelola secara aman menjadi indikator yang menunjukkan kualitas infrastruktur dasar suatu negara, hal ini karena akses air bersih berkorelasi langsung dengan tingkat kesehatan masyarakat, produktivitas tenaga kerja, dan stabilitas sosial sebagai pendukung pembangunan ekonomi secara keseluruhan [21].

2.2. Metode Analisis Data

2.2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) adalah teknik reduksi dimensionalitas yang bertujuan untuk mengubah data ke dalam sistem koordinat baru sehingga dimensi-dimensi yang menangkap variasi terbesar dalam data dapat dengan mudah diidentifikasi [21]. PCA menciptakan variabel-variabel baru yang tidak saling berkorelasi dan secara berturut-turut memaksimalkan varians, di mana komponen utama pertama menangkap varians terbesar, komponen kedua menangkap varians terbesar berikutnya yang ortogonal terhadap komponen pertama, dan seterusnya [22].

Lebih lanjut, PCA merupakan metode statistik yang digunakan untuk mereduksi dimensi data dengan mengubah variabel asli menjadi sejumlah komponen utama yang tidak berkorelasi. PCA bertujuan untuk mengatasi permasalahan kompleksitas akibat banyaknya variabel yang digunakan dalam analisis, tanpa harus kehilangan informasi penting dari variabel-variabel tersebut [23]. PCA bekerja dengan cara mentransformasikan variabel yang saling berkorelasi menjadi variabel baru yang tidak berkorelasi, atau disebut sebagai *principal components*. Bahwa PCA dapat mengubah data asli menjadi kumpulan variabel baru yang lebih kecil dan tidak berkorelasi, namun tetap merepresentasikan informasi dari data awal [24].

Tujuan utama PCA adalah mendapatkan sebanyak mungkin varians dalam data dengan jumlah komponen yang lebih sedikit. Dalam penerapannya, komponen utama ditentukan berdasarkan nilai *eigen* yang diperoleh dari matriks korelasi, di mana komponen dengan *eigenvalue* lebih besar atau sama dengan satu dipilih untuk mewakili data [23]. Dengan mengurangi jumlah dimensi, PCA membantu dalam visualisasi data dan mengurangi kompleksitas model. PCA sering digunakan sebagai langkah pra-pemrosesan sebelum analisis lebih lanjut seperti *clustering* atau regresi. Dalam konteks analisis indikator pembangunan, PCA memungkinkan identifikasi dimensi-dimensi utama yang menjelaskan sebagian besar variasi dalam data [23]. Data yang telah direduksi dengan PCA kemudian digunakan dalam proses *clustering* [24].

1. Standarisasi Data (*Z-Score Normalization*)
   Sebelum menghitung PCA, data harus di standarisasi terlebih dahulu jika skala antar variabel berbeda, agar setiap variabel memiliki rata-rata 0 dan standar deviasi
   $$z_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j}$$
   x_{ij} nilai data pada observasi ke-i dan variabel ke-j
   $\overline{x_j}$ rata-rata variabel ke-j
   s_j standar deviasi variabel ke-j
2. Hitung Matriks *Kovarians* atau Korelasi
   Setelah di standarisasi, maka gunakan matriks korelasi. Jika tidak di standarisasi, gunakan matriks *kovarians*
   $$\mathrm{Cov}(X) = \frac{1}{n-1} X^T X$$
   X matriks data setelah di kurangi oleh rata-rata
   n jumlah observasi
3. Hitung Nilai *Eigen* dan Vektor *Eigen*

Dari matriks yang sudah di hitung dengan *Kovarians* atau Korelasi selanjutnya hitunglah *eigenvalue* (λ) dan *eigenvector* (v) dengan menyelesaikan:

$$\mathrm{Cov}(X) \cdot v = \lambda \cdot v$$
$$\det(\mathrm{Cov}(X) - \lambda I) = 0$$

Nilai *Eigen* menunjukkan seberapa besar varians yang di jelaskan setiap komponen
Vektor Eigen menunjukkan kombinasi dari variabel dari komponen tersebut

4. Pilih *Principal Components* (PC)
   Mengurutkan nilai *eigen* dari yang terbesar ke yang terkecil. Lalu PC (*Principal Components*) dengan nilai *eigen* ≥ 1 (Dalam kaidah *Kaiser*) Menjelaskan hanya 70–90% varians kumulatif.
5. Transformasi Data ke Ruang Baru
   Membuat Transformasi Dari data asli ke dalam ruang yang baru

$$Z_{\text{baru}} = Z \cdot V$$

Z data hasil dari standarisasi
V matriks dari vektor eigen

2.2.2 Clustering K-Means

Clustering analysis adalah teknik *unsupervised learning* yang bertujuan untuk mengelompokkan objek-objek ke dalam cluster-cluster di mana objek dalam satu *cluster* memiliki kesamaan yang tinggi sementara objek antar *cluster* berbeda secara signifikan [25]. Metode *clustering* yang umum digunakan meliputi *K-means clustering* yang meminimalkan varians dalam *cluster* (*squared Euclidean distance*s), *hierarchical clustering* yang membangun hierarki *cluster* secara bertahap, dan *density-based clustering*, masing-masing dengan kelebihan dan kelemahan tersendiri tergantung pada karakteristik data dan tujuan analisis [26]. Dalam konteks penelitian pembangunan, *clustering* memungkinkan identifikasi kelompok negara dengan profil pembangunan yang serupa, sehingga memfasilitasi analisis komparatif antar negara.

Lebih lanjut, *clustering* adalah metode penambangan data yang bertujuan untuk mengelompokkan data ke dalam beberapa grup atau klaster berdasarkan kesamaan karakteristiknya [27]. Setiap kelompok yang terbentuk memiliki tingkat kesamaan antar anggota di dalam klaster yang tinggi, tetapi berbeda secara signifikan dengan anggota klaster lainnya. Metode *clustering* terbagi menjadi beberapa pendekatan, seperti metode hierarkis (*hierarchical*), metode berbasis densitas (*density-based*), dan metode partisi (*partitioning*), di mana *K-means* adalah salah satu yang paling umum digunakan [28].

Algoritma *K-means*, sebagai salah satu metode *clustering* partisi, bekerja dengan membagi data ke dalam sejumlah klaster tertentu berdasarkan jarak antar titik data terhadap pusat klaster (*centroid*) yang dihitung secara iteratif [27]. Proses ini dimulai dengan *inisialisasi* sejumlah k pusat klaster secara acak. Kemudian, setiap titik data akan ditetapkan ke klaster terdekat dengan pusat klaster yang ada. Setelah semua titik data dialokasikan, posisi pusat klaster akan diperbarui dengan menghitung rata-rata dari semua titik data yang termasuk dalam klaster tersebut. Tahapan penugasan titik data dan pembaruan pusat klaster ini akan terus berulang hingga pusat klaster tidak lagi banyak berubah atau mencapai kriteria konvergensi tertentu.

Salah satu masalah utama dalam *clustering* adalah menentukan jumlah klaster optimal (k) secara otomatis tanpa perlu inisialisasi manual [29]. Untuk menjawab masalah ini, beberapa metode telah dikembangkan guna meningkatkan efisiensi proses penentuan klaster. Metode *Elbow*, misalnya, memanfaatkan nilai WCSS (*Within-Cluster Sum of Squares*) untuk mendeteksi "titik siku" atau perubahan signifikan dari penurunan WCSS sebagai indikator jumlah klaster terbaik [27]. WCSS mengukur total variasi di dalam klaster, di mana nilai WCSS yang lebih kecil menunjukkan klaster yang lebih padat dan homogen.

Di sisi lain, untuk mengatasi ketergantungan pada inisialisasi dan penentuan parameter awal, dengan algoritma *U-k-means* (*Unsupervised K-means*). Algoritma ini menerapkan pendekatan berbasis entropi yang memungkinkan proses *clustering* berjalan tanpa parameter awal dan secara otomatis menyaring jumlah klaster selama proses iterasi berlangsung. Hal ini menjadikan *U-k-means* sebagai metode yang lebih "murni" dalam konteks *unsupervised learning*, karena mampu menemukan jumlah klaster optimal tanpa intervensi pengguna pada tahap awal.

1. Menentukan Jumlah Klaster (k)
   Menentukan nilai dari *Cluster* di lambangkan dengan (k). Nilai k dapat di tentukan dengan manual atau dengan menggunakan metode *Elbow*
2. Inisialisasi Titik Pusat Awal (*Centroid*)
   Titik k titik pusat awal (*centroid*) diambil secara acak dari himpunan yang ada di data.

$$C = \{C_1, C_2, \dots, C_k\}$$

3. Menghitung Jarak Tiap Titik ke Setiap *Centroid*
   Euclidian Distance Berfungsi untuk mengukur jarak antara data di titik x_i dan gunakan centroid c_j

$$d(x_i, c_j) = \sqrt{\sum_{m=1}^{n} (x_{im} - c_{jm})^2}$$

   x_i data ke-i
   $c_j$centroid klaster ke-j
   n jumlah dimensi fitur
4. Ulangi Langkah 3–5 (Iterasi)
   Ulangi proses iterasi terus dari menghitung jarak, mengelompokkan data dan menghitung *centroid* baru sampai :
   - Posisi *centroid* tetap atau tidak berubah.
   - Jumlah Perulangan mencapai maksimum
   - Anggota klaster tidak ada perubahan

5. Evaluasi Hasil dengan WCSS (*Within-Cluster Sum of Squares*)
   Untuk mengukur kualitas dari klaster. Dapat menggunakan rumus WCSS

$$WCSS = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left|\left| x_i - c_j \right|\right|^2$$

   Semakin kecil nilai dari WCSS, menandakan semakin baik hasil klasterisasi karena titik-titik lebih dekat ke centroid-nya

2.3. *Dataset* Global Development Indicators

Dataset Global Development Indicators terdiri dari 5.556 observasi yang mencakup berbagai negara dan wilayah selama periode 2000 hingga 2020. Setiap baris mewakili kondisi suatu negara pada tahun tertentu. Merujuk pada **Tabel 1** *Dataset* ini memuat 22 variabel yang mencerminkan berbagai indikator di bidang pembangunan ekonomi, kesiapan digital, lingkungan, dan sosial.

2.3.1 Stuktur Dataset

Tabel 1 *Dataset Global Development Indicators*

Variabel	Tipe	Deskripsi Singkat
country_name	*Categorical*	Nama negara atau wilayah
year	*Numerik (Int)*	Tahun pengamatan (2000–2020)
region	*Categorical*	Wilayah geografis
income_group	*Categorical*	Kelompok pendapatan negara
gdp_usd	*Numerik (Float)*	Total Produk Domestik Bruto (US$)
population	*Numerik (Float)*	Total populasi
gdp_per_capita	*Numerik (Float)*	PDB per kapita (US$)
inflation_rate	*Numerik (Float)*	Tingkat inflasi (%)
unemployment_rate	*Numerik (Float)*	Tingkat pengangguran (%)
health_expenditure_pct_gdp	*Numerik (Float)*	Pengeluaran kesehatan (% dari PDB)
life_expectancy	*Numerik (Float)*	Harapan hidup rata-rata (tahun)
human_development_index	*Numerik (Float)*	Indeks Pembangunan Manusia (0–1)
climate_vulnerability_index	*Numerik (Float)*	Indeks kerentanan terhadap perubahan iklim
digital_readiness_score	*Numerik (Float)*	Skor kesiapan digital
governance_quality_index	*Numerik (Float)*	Indeks kualitas tata kelola
global_resilience_score	*Numerik (Float)*	Skor ketahanan global
global_development_resilience_index	*Numerik (Float)*	Indeks ketahanan pembangunan global
education_health_ratio	*Numerik (Float)*	Rasio pengeluaran pendidikan terhadap kesehatan
is_pandemic_period	*Numerik (Int)*	Indikator periode pandemi (1 jika pandemi)
years_since_2000	*Numerik (Int)*	Jumlah tahun sejak 2000
years_since_century	*Numerik (Int)*	Jumlah tahun sejak tahun 2000 (abad ke-21)

2.3.2 Missing Value

Dalam *dataset* Global Development *Indicators*, terdapat beberapa variabel yang memiliki nilai hilang (*missing values)* dalam jumlah yang cukup signifikan. Variabel dengan tingkat kehilangan data tertinggi adalah *education_health_ratio*, dengan banyak baris yang tidak memiliki nilai. Selain itu, variabel seperti *income_group*, *unemployment_rate*, health_expenditure_pct_gdp, dan *life_expectancy* juga menunjukkan persentase *missing values* yang relatif tinggi. Hal ini perlu menjadi perhatian dalam proses pra-pemrosesan data, terutama untuk analisis yang sensitif terhadap kelengkapan data seperti PCA atau *clustering*.

3. Penerapan PCA dan Klastering dengan Matlab

3.1. Persiapan Pengujian

Dalam melakukan analisis Data Global Development *Full Analysis* (2000-2020) menggunakan penerapan *Principal Component Analysis* dan *Klastering*, digunakan perangkat lunak, perangkat keras, serta data set. Dilakukan pengolahan data dari data set, analisis statistik, serta visualisasi dari hasil yang diperoleh. Berikut adalah penjelasan dari masing-masing hal yang digunakan:

1. Perangkat Lunak (Software): MATLAB
   Program utama yang digunakan dalam analisis ini adalah MATLAB (Matrix Laboratory). Program ini merupakan sebuah bahasa pemrograman dan lingkungan komputasi numerik yang populer dalam bidang teknik, matematika, dan sains data. MATLAB menjadi pilihan karena memiliki

berbagai fungsi statistik dan visualisasi yang baik dan lengkap. Penggunaan program ini memberikan kemudahan dalam melakukan pemodelan regresi, penanganan data, dan plotting. Beberapa fitur andalan MATLAB yang digunakan, antara lain:

- readtable() – Digunakan untuk membaca data dari file CSV ke dalam bentuk tabel yang memudahkan akses berdasarkan nama kolom.
- unique() – Digunakan untuk mengambil daftar unik dari nama negara yang terdapat dalam data set.
- isnan(), ismissing(), dan logika indexing – Digunakan untuk mengecek dan membersihkan data yang kosong (missing values), termasuk kolom yang seluruhnya kosong.
- matlab.lang.makeValidName() – Digunakan untuk memastikan bahwa nama variabel (misalnya nama negara) sesuai dengan aturan penamaan variabel di MATLAB.
- interp1() – Digunakan untuk melakukan interpolasi nilai yang hilang (missing value) menggunakan metode linear, pchip (kuadratik), dan spline (kubik) sesuai stabilitas data.
- sortrows() – Digunakan untuk mengurutkan data berdasarkan tahun sebelum interpolasi dilakukan.
- zscore() – Digunakan untuk menormalisasi data sebelum dilakukan Principal Component Analysis (PCA) atau klasterisasi.
- pca() – Digunakan untuk melakukan Principal Component Analysis, dalam hal ini digunakan untuk membuat indeks lingkungan, pembangunan, dan kesiapan digital.
- pareto() – Digunakan untuk memvisualisasikan kontribusi masing-masing komponen utama terhadap total variansi.
- plot(), subplot(), title(), xlabel(), ylabel(), grid on – Digunakan untuk menampilkan grafik garis dari indeks yang telah dihitung untuk visualisasi perbandingan antar negara.
- kmeans() – Digunakan untuk melakukan klasterisasi (pengelompokan negara) berdasarkan indeks-indeks utama yang telah dibentuk.
- silhouette() – Digunakan untuk mengevaluasi kualitas hasil klasterisasi dengan melihat nilai silhouette dari tiap observasi.
- scatter3() – Digunakan untuk membuat visualisasi klaster dalam ruang 3 dimensi berdasarkan tiga indeks utama.
- parallelcoords() – Digunakan untuk membuat plot koordinat paralel yang menampilkan pola nilai antar klaster.
- save() – Digunakan untuk menyimpan hasil perhitungan indeks dan hasil klasterisasi ke dalam file MAT agar dapat diakses ulang.
- function – Penggunaan fungsi modular (seperti create_index) memberikan struktur yang lebih bersih dan memungkinkan penggunaan ulang kode secara efisien.

Versi MATLAB yang digunakan dalam proyek ini adalah MATLAB R2022a.

2. Perangkat Keras (Hardware): Laptop MSI GF63 Thin 11UC

   Untuk menjalankan MATLAB dan melakukan analisis data dengan ukuran sedang seperti dataset Global Development Full Analysis, digunakan perangkat keras Laptop MSI GF63 Thin 11UC dengan spesifikasi sebagai berikut:
   - Model: MSI GF63 Thin 11UC
   - Sistem Operasi: Windows 11 Home Single Language 64-bit (Build 22631)
   - Prosesor: 11th Gen Intel® Core™ i7-11800H @ 2.30GHz, 16 threads
   - RAM: 16 GB
   - GPU: NVIDIA GeForce RTX 3050 Laptop GPU
   - Total Memori: 12 GB
   - Memori Tampilan (Display): 3.97 GB
   - Memori Bersama (Shared): 8.04 GB
   - DirectX: Versi 12
   - Pabrikan Sistem: Micro-Star International Co., Ltd.
   - Model BIOS: E16R6IMS.110

   Spesifikasi tersebut sudah memadai untuk menjalankan MATLAB dengan lancar dan cepat, terutama untuk kebutuhan analisis statistik dan pembuatan visualisasi data yang dapat dilakukan tanpa mengalami kendala. Hal ini dikarenakan spesifikasi perangkat yang digunakan sudah sesuai dengan spek yang direkomendasikan untuk menjalankan perangkat lunak MATLAB.

3.2. Pre-processing Dengan Matlab

Pre-processing adalah proses untuk mempersiapkan data mentah agar siap untuk di analisis. Dalam konteks ini, data mentah yang perlu dipersiapkan adalah dataset *Global_Development_Indicators_2010_2020.csv*. Untuk memulai pre-processing, dataset dibaca dahulu dalam matlab menggunakan perintah readtable untuk membaca dataset dengan format .csv. Dengan menggunakan perintah ini, data dalam dataset dapat di akses oleh program MATLAB. Proses ini dapat dilihat pada **Gambar 1**.

```
% === 1. Load and filter dataset ===
data = readtable('Global_Development_Indicators_2000_2020.csv');
```

Gambar 1 Membaca Dataset

Selanjutnya dilakukan pembersihan data untuk menghilangkan variabel-variabel diwakili oleh kolom yang tidak diperlukan. Dari 47 variabel yang ada di dalam dataset *Global_Development_Indicators_2010_2020.csv*, hanya diperlukan 10 variabel saja, oleh karena itu 37 variabel lainnya akan dibuang. Setelah pembersihan dataset, dilakukan pemisahan dataset menjadi tabel-tabel berdasarkan negara masing-masing. Berikut adalah kode pada **Gambar 2** dan **Gambar 3** untuk mewujudkan hal tersebut.

```
% Kolom yang relevan
columns_to_keep = {
    'country_code', ...
    'gdp_per_capita', 'inflation_rate', ...
    'life_expectancy', 'school_enrollment_secondary', ...
    'unemployment_rate', 'electricity_access_pct', ...
    'mobile_subscriptions_per_100', 'internet_usage_pct', 'human_development_index'
};

filtered_data = data(:, [{'country_name', 'year'}, columns_to_keep]);
```

Gambar 2 Filtrasi Variabel-Variabel Dataset

```
% === 2. Pisahkan per negara dan filter data kosong ===
countries = unique(filtered_data.country_name);
country_groups = struct();

for i = 1:length(countries)
    cname = countries{i};
    tmp_table = filtered_data(strcmp(filtered_data.country_name, cname), :);
    tmp_table.country_name = [];
    tmp_table.country_code = [];

    has_fully_missing = false;
    for j = 1:width(tmp_table)
        col_data = tmp_table{:, j};
        if isnumeric(col_data)
            if all(isnan(col_data))
                has_fully_missing = true;
                break;
            end
        end
    end

    if ~has_fully_missing
        valid_name = matlab.lang.makeValidName(cname);
        country_groups.(valid_name) = tmp_table;
    end
end
```

Gambar 3 Pemisahan Dataset

Variabel-variabel yang diperlukan pada *dataset* dapat dibagi menjadi tiga jenis. Jenis pertama adalah identitas yang ditunjukkan oleh *country_name* sebagai identitas negara. Variabel jenis kedua variabel waktu, yakni *year* yang menjadi indikator waktu untuk menggambarkan perubahan negara dari tahun ke tahun. Variabel terakhir, yakni variabel yang akan di analisis adalah variabel-variabel indikator pembangunan negara. **Tabel 2**, **Tabel 3**, dan **Tabel 4** menunjukan variabel-variabel tersebut:

A. Tabel Identitas Negara

Tabel 2 Menunjukan Identitas Suatu Negara

Nama variabel	Deskripsi
country_name	Nama negara sebagai identitas pengamatan

B. Tabel Waktu

Tabel 3 Menunjukkan berapa tahun pengamatan.

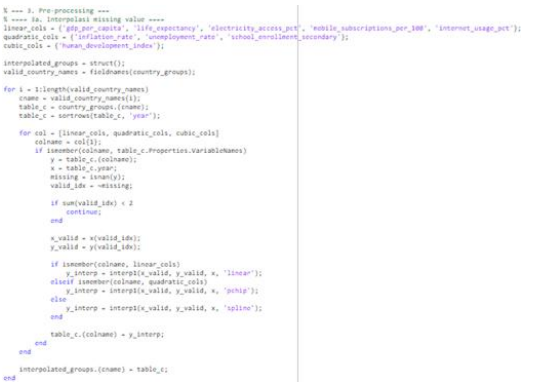
Nama variabel	Deskripsi
year	Tahun pengamatan (2000-2020)

C. Tabel Indikator Pembangunan Negara

Tabel 4 Indikator yang mempengaruhi pembangunan negara

Nama variabel	Deskripsi
gdp_per_capita	Produk domestik bruto per kapita (dalam USD)
inflation_rate	Laju inflasi tahunan (dalam persentase)
life_expectancy	Harapan hidup saat lahir (dalam tahun)
school_enrollment_secondary	Tingkat partisipasi sekolah menengah (% dari kelompok usia terkait)
unemployment_rate	Tingkat pengangguran (% dari angkatan kerja)
electricity_rate	Persentase populasi dengan akses terhadap listrik
mobile_subscriptions_per_100	Jumlah langganan telepon seluler per 100 penduduk
internet_usage_pct	Persentase populasi yang menggunakan internet
human_development_index	Indeks pembangunan manusia (HDI)
governance_quality_index	Indeks kualitas tata kelola pemerintahan
digital_connectivity_index	Indeks konektivitas digital (akses, penggunaan, dan infrastruktur digital)

Setelah melakukan penanganan terhadap variabel-variabel di dalam dataset, selanjutnya perlu dilakukan penanganan juga pada nilai-nilai kosong (NaN (*not a number*)) atau *missing value.* Untuk menangani *missing value* pada dataset, diterapkan metode interpolasi dan imputasi. Interpolasi diterapkan dengan metode interpolasi *linier*, *quadratic*, dan *cubic*, jenis interpolasi yang digunakan disesuaikan berdasarkan tingkat fluktuatif dari setiap variabel. Untuk menggunakan ketiga jenis interpolasi tersebut, digunakan fungsi bawaan *Matlab interpl* dengan metode linear, phchip(*quadratic*), dan spline(*cubic*). Untuk imputasi dilakukan dengan mengisi *missing values* yang tersisah dengan median dari kolom setiap negara. *Imputasi* dilaksanakan sesudah interpolasi. Berikut adalah kode pada **Gambar 4** dan **Gambar 5** dengan referensi dari [9] dan [10] untuk interpolasi dan *imputasi*:



Gambar 4 Interpolasi

```
% ==== 3b. Imputasi NaN yang tersisa dengan median per kolom per negara ====
for i = 1:length(valid_country_names)
    cname = valid_country_names{i};
    tbl = interpolated_groups.(cname);

    for j = 1:width(tbl)
        col_data = tbl{:, j};
        if isnumeric(col_data)
            if any(isnan(col_data))
                median_val = median(col_data, 'omitnan');
                col_data(isnan(col_data)) = median_val;
                tbl{:, j} = col_data;
            end
        end
    end

    interpolated_groups.(cname) = tbl;
end
```

Gambar 5 Imputasi

Untuk tahap terakhir dalam *pre-processing*, perlu dilakukan balik arah pada variabel *inflation_rate* (tingkat inflasi) dan *unemployment_rate* (tingkat pengangguran). Dua variabel ini tentu merupakan variabel yang memiliki interpretasi negatif dalam pembangunan global suatu negara. Semakin besar angka dari kedua nilai tersebut, semakin berdampak negatif pada pembangunan global dari negara terkait. Untuk memberikan interpretasi negatif tersebut dalam proses penelitian ini, maka dilakukan *invers* pada setiap nilai pada dua variabel ini dengan mengubah seluruh nilai kolom menjadi nilai maksimum dari kolom dikurang oleh nilai bawaan pada baris tersebut. Berikut kode pada **Gambar 6** untuk melakukan *invers*.

```
% ==== 3c. Balik arah variabel dengan interpretasi negatif ====
negatively_oriented_vars = {'inflation_rate', 'unemployment_rate'};

for i = 1:length(valid_country_names)
    cname = valid_country_names{i};
    tbl = interpolated_groups.(cname);

    for v = 1:length(negatively_oriented_vars)
        varname = negatively_oriented_vars{v};
        if ismember(varname, tbl.Properties.VariableNames)
            col_data = tbl.(varname);
            % Lakukan invers sederhana: maksimum - nilai
            % Ini akan mengubah arah tanpa mengubah skala relatif
            max_val = max(col_data); % bisa juga median atau konstanta tetap
            tbl.(varname) = max_val - col_data;
        end
    end

    interpolated_groups.(cname) = tbl;
end
```

Gambar 6 Balik Arah Variabel

3.3. PCA Dengan Matlab

PCA (*Principal Component Analysis*) bertujuan untuk mengurangi dimensi demi mengurangi kompleksitas dari variabel yang berjumlah besar. Proses PCA dengan Matlab meliputi, perhitungan rata-rata, standarisasi data, pembuangan outlier, pelaksanaan PCA, dan interpretasi PC1 dan PC2 dari PCA. Metode PCA yang di adaptasikan pada Matlab ini mengambil referensi dari [32]. Berikut adalah langkah-langkah detail dari PCA menggunakan Matlab:

1. Perhitungan Rata-Rata

   Langkah pertama sebelum pelaksanaan PCA adalah perhitungan rata-rata dari setiap kolom. Kolom dari setiap tabel negara yang mewakili satu tahun dihitung rata-ratanya. Rata-rata ini akan digunakan untuk standarisasi data. Untuk menghitung rata-rata, digunakan fungsi mean. Berikut kode pada **Gambar 7** untuk menjalankan tujuan ini:

```
% === 4. Hitung rata-rata tahunan per negara ===
country_means = [];
country_labels = {};

for i = 1:length(valid_country_names)
    cname = valid_country_names{i};
    tbl = interpolated_groups.(cname);
    numeric_vars = tbl(:, ~strcmp(tbl.Properties.VariableNames, 'year'));
    mean_vals = mean(table2array(numeric_vars), 'omitnan');
    country_means = [country_means; mean_vals];
    country_labels{end+1} = cname;
end
```

Gambar 7 Perhitungan Rata-Rata Tahunan per Kolom Negara

2. Standarisasi Data

   Standarisasi data adalah proses mengubah data ke format yang lebih konsisten, agar setiap variabel memiliki *mean* 0 dan standar deviasi 1. Standarisasi dilakukan dengan mengurangkan setiap nilai indikator dengan rata-rata seluruh negara pada indikator tersebut, lalu dibagi dengan standar deviasinya. Proses ini memastikan bahwa semua indikator memiliki skala yang setara (*mean* 0 dan standar deviasi 1), sehingga tidak ada variabel yang mendominasi dalam analisis PCA. Proses ini perlu dilakukan sebelum PCA karena PCA bersifat sensitif terhadap skala variabel yang tidak merata. Hasil standarisasi data berupa *Z-score*. Berikut kode pada **Gambar 8** untuk mewujudkan hal ini:

```
% === 5. Standarisasi data (z-score) ===
X = country_means;
[n, p] = size(X);
Xmean = mean(X);
Xstd = std(X);
Z = (X - Xmean) ./ Xstd;
```

Gambar 8 Standarisasi Data

3. Pembuangan *Outlier*

   Outlier adalah data dengan nilai yang jauh berbeda signifikan dari mayoritas data lainnya. Dalam konteks ini, berupa negara yang

memiliki nilai indikator pembangunan yang ekstrem (sangat besar atau sangat kecil). *Outlier* dapat membuat hasil analisis menjadi tidak proporsional dengan menggeser pusat klaster secara tidak wajar. Oleh karena demikian, menggunakan referensi dari [33], dilakukan penanganan terhadap *outlier* dengan kode pada **Gambar 9** berikut di Matlab.

```
% === 6. Buang Outlier berdasarkan z-score multivariat (threshold Euclidean) ===
z_thresh = 5.2;
z_dist = sqrt(sum(Z.^2, 2)); % Euclidean distance dari pusat (0)

% Negara yang bukan outlier
not_outlier_idx = z_dist < z_thresh;

% Filter data
Z = Z(not_outlier_idx, :);

% Simpan nama negara yang dianggap outlier
outlier_idx = ~not_outlier_idx;
outlier_countries = country_labels(outlier_idx);

% Menampilkan negara-negara outlier yang dibuang
disp('Negara-negara yang dianggap outlier dan dibuang:');
disp(outlier_countries');

country_labels = country_labels(not_outlier_idx);
```

Gambar 9 Penangan Outlier

Z-score yang diperoleh dari standarisasi data digunakan untuk menghitung jarak *Euclidean* setiap negara terhadap pusat distribusi standar (pusat koordinat dalam ruang *z-score*). *Z-threshold* yang diwakili oleh variabel *z_tresh* adalah nilai ambang batas tertinggi yang ditetapkan untuk memisahkan negara-negara normal dan *outlier*. Negara dengan jarak ke pusat distribusi standar yang lebih besar dari *treshold* tersebut di anggap sebagai *outlier* dan akan disaring keluar agar tidak ikut dalam analisis selanjutnya. Sebagai referensi, nama-nama negara *outlier* ditampilkan di terminal *Matlab*.

4. PCA

PCA bertujuan untuk mereduksi dimensi data indikator pembangunan negara namun dengan tetap mempertahankan sebagian besar informasi (varians). Proses ini dilakukan dengan menghitung matriks kovarians dari data yang telah ter standarisasi dengan *Z-score*, untuk memahami hubungan antar variabel. Selanjutnya, dilakukan dekomposisi eigencomposition terhadap matriks kovarians untuk memperoleh eigenvalue dan *eigenvector*. *Eigenvalue* merupakan ukuran varians yang dijelaskan oleh masing-masing komponen utama, sedangkan *eigenvector* menentukan arah komponen tersebut di ruang fitur. Kemudian dilakukan pengurutan eigenvalue dari besar ke kecil, serta ditentukan jumlah komponen utama minimum yang diperlukan untuk menjelaskan 80% total varians, hanya komponen-komponen ini yang akan dipertahankan. Akhirnya, data Z yang diproyeksikan ke ruang komponen utama baru dan menghasilkan representasi baru W untuk digunakan pada analisis selanjutnya. Berikut kode pada **Gambar 10** untuk mewujudkan hal ini:

```
% === 6. PCA ===
S = cov(Z);
[V, D] = eig(S);
[sorted_eigenvals, idx] = sort(diag(D), 'descend');
V_sorted = V(:, idx);

total_variance = sum(sorted_eigenvals);
prop_variance = sorted_eigenvals / total_variance;
cum_prop = cumsum(prop_variance);
komponen = find(cum_prop >= 0.8, 1);
V_utama = V_sorted(:, 1:komponen);
W = Z * V_utama;
```

Gambar 10 PCA

5. Interpretasi PCA

Interpretasi pada PC1 (Principal component 1) dan PC2 (Principal Component 2) digunakan untuk mengetahui seberapa besar kontribusi dari masing-masing variabel terhadap dua komponen PCA yaitu PC1 dan PC2. PCA melakukan reduksi dimensi dengan membentuk sejumlah komponen utama (*principal components*) dari kumpulan variabel asli. Untuk mengetahui makna dari komponen-komponen baru ini, maka perlu diketahui variabel-variabel yang paling dominan dalam membentuk masing-masing komponen utama. Proses ini dilakukan dengan menggunakan interpretasi loading faktor. Berikut kode pada **Gambar 11** untuk mewujudkan tujuan ini:

```
% Interpretasi PC1 dan PC2 pada PCA
variable_names = {
    'gdp_per_capita', ...
    'inflation_rate', ...
    'life_expectancy', ...
    'school_enrollment_secondary', ...
    'unemployment_rate', ...
    'electricity_access_pct', ...
    'mobile_subscriptions_per_100', ...
    'internet_usage_pct', ...
    'human_development_index'
};

% Menggunakan kolom pertama dan kedua dari matrix eigenvector
loading_PC = V_sorted(:, 1:2);

fprintf('\n=== Kontribusi Variabel terhadap PC1 dan PC2 ===\n');
fprintf('%-35s %10s %10s\n', 'Variabel', 'Loading PC1', 'Loading PC2');
fprintf('------------------------------------------------------------\n');
for i = 1:length(variable_names)
    fprintf('%-35s %+10.4f %+10.4f\n', ...
        variable_names{i}, loading_PC(i,1), loading_PC(i,2));
end
```

Gambar 11 Interpertrasi PCA

Kode mengambil dua komponen utama dari matriks *eigenvector* yang mewakili arah dengan variasi data terbesar. Nilai-nilai dalam dua kolom ini disebut *loading*, yang menunjukkan seberapa besar kontribusi setiap variabel asli terhadap masing-masing komponen. Kemudian, program mencetak tabel kontribusi variabel terhadap PC1 dan PC2. Tanda positif atau negatif dari *loading* menunjukkan arah pengaruh variabel. Misalnya, jika *gdp_per_capita*, *internet_usage_pct*, dan *electricity_access_pct* memiliki *loading* tinggi pada PC1, maka PC1 dapat diartikan sebagai dimensi pembangunan ekonomi. Sementara jika *life_expectancy* dan *human_development_index* dominan pada PC2, maka PC2 mencerminkan dimensi sosial atau kesejahteraan.

3.4. *Clustering* Dengan Matlab

Clustering dilakukan sebagai upaya untuk mengelompokkan negara-negara berdasarkan karakteristik pembangunan global. Proses clustering mencakup *Sillhoutte method* dan perhitungan *K-means*. Sebagai ilustrasi hasil, dilakukan juga visualisasi terhadap perhitungan *Sillhoutte*, serta hasil clustering. Pada tahap terakhir anggota-anggota negara dalam setiap *cluster* akan dimasukan disimpan di dalam excel dan rata-rata PC1, PC2, dan nilai *silhoutte* setiap klaster dihitung. Untuk melakukan metode *clustering*, di ambil referensi dari [34]. Berikut adalah langkah-langkah untuk *clustering*:

1. *Silhoutte* Method

   Metode *Sillhoutte* digunakan untuk menentukan jumlah klaster yang optimal dalam membagi data negara berdasarkan indikator pembangunan. Proses ini dilaksanakan dengan mencoba berbagai nilai K (jumlah klaster) dari 2 hingga 10, lalu untuk setiap jumlah klaster tersebut, dihitung rata-rata koefisien *silloutte*-nya. Koefisien ini menjadi indikator seberapa baiknya pengelompokan negara-negara. Nilai tertinggi dari grafik rata-rata *sillhoutte* mengindikasikan jumlah klaster yang memberikan pemisahan paling baik dalam data, nilai ini akan menentukan variabel *k_opt* sebagai jumlah klaster paling optimal. Berikut kode pada **Gambar 12** untuk mewujudkan ini:

```
% === 8. Silhouette Method untuk menentukan jumlah klaster ===
maxK = 10;
silhouette_avg = zeros(maxK,1);

figure;
for k = 2:maxK
    idx_temp = kmeans(W, k, 'Replicates', 10);
    s = silhouette(W, idx_temp);
    silhouette_avg(k) = mean(s);
end

plot(2:maxK, silhouette_avg(2:end), '-o', 'LineWidth', 2);
xlabel('Jumlah Klaster (k)');
ylabel('Rata-rata Silhouette Coefficient');
title('Silhouette Method untuk Menentukan Jumlah Klaster Optimal');
grid on;
```

Gambar 12 Metode *Silhoutte*

   Dilakukan loop untuk K = 2 sampai K = 10 untuk menentukan jumlah klaster paling optimal. Hasil setiap rata-rata *silhoutte* untuk setiap jumlah klaster disimpan dalam array *silhoutte_avg*. Klasterisasi dilakukan dengan menggunakan fungsi *K-Means* dengan 10 replikasi (W komponen yang di klasterisasi dan k sebagai jumlah klaster yang mau dibuat). Untuk menghitung nilai *silhoutte*, digunakan fungsi *silhoutte*, lalu hasil dari fungsi tersebut dirata-ratakan untuk menjadi rata-rata *silhoutte*.

2. Visualisasi *Plot Silhoutte*

   Setelah jumlah klaster paling optimal ditentukan, dilakukan visualisasi plot *sillhoutte* untuk memberikan gambaran detail tentang kualitas dari pemisahan klaster. Plot yang dihasilkan ini menampilkan nilai *silhoutte* kepada masing-masing negara dalam setiap klaster. Ketebalan dari plot ini menggambarkan jumlah negara yang digolongkan dalam klaster tersebut, dan *sillhoutte coeficient* dari plot menggambarkan kecocokan negara dalam klaster terkait (semakin panjang maka semakin cocok dengan klasternya). Untuk memberikan informasi tambahan, dihitung juga rata-rata dari nilai *sillhoutte* dan ditampilkan pada plot. Berikut kode pada **Gambar 13** untuk melakukan visualisasi plot *sillhoutte*:

```
% === 9. Visualisasi plot silhouette untuk k terbaik ===
[~, k_opt] = max(silhouette_avg(2:end));
k_opt = k_opt + 1; % Jumlah kluster optimal berdasarkan sillhoutte method
[idx_opt, ~] = kmeans(W, k_opt, 'Replicates', 10);

figure;
[silh_vals, h] = silhouette(W, idx_opt);
avg_silhouette = mean(silh_vals);

silhouette(W, idx_opt);
title(sprintf('Silhouette Plot untuk k = %d', k_opt));
xlabel('Silhouette Coefficient');
ylabel('Cluster');

hold on;
xline(avg_silhouette, '--r', ...
    sprintf('Rata-rata = %.2f', avg_silhouette), ...
    'LabelOrientation', 'horizontal', ...
    'LabelHorizontalAlignment', 'left', ...
    'FontSize', 10);
hold off;
```

Gambar 13 Visualisasi *Plot Silhoutte*

   Silhoutte_avg yang merupakan *array* dari proses perhitungan rata-rata setiap nilai *silhoutte* untuk setiap jumlah klaster di ambil nilai maksimumnya. Kemudian untuk memperoleh jendela gambar baru, digunakan *figure* yang akan menampilkan *plotting* hasil klasterisasi. Lalu koefisien siluet untuk setiap titik data berdasarkan klasterisasi dihitung untuk memvisualisasikan plot *sillhoutte*. Untuk menampilkan rata-rata serta garis putus-putus posisi rata-rata tersebut dalam plot, digunakan *xline* yang menampilkan garis putus-putus pada posisi variabel *avg_silhoutte*.

3. *Clustering* dan Visualisasi PCA

   Tahap terakhir dalam proses *clustering* adalah menggunakan *K-means* dengan jumlah klaster optimal untuk memvisualisasikan hasil clustering pada ruang dua dimensi. Visualisasi ini menampilkan pola pengelompokan antar negara, di mana setiap klaster diwakili dengan warna yang sama dan klaster lain dengan warna yang berbeda. Untuk memberikan konteks yang lengkap, setiap label-label nama negara diberikan sehingga interpretasi hasil klasterisasi dapat dilakukan dengan baik. Agar setiap anggota klaster dapat dilihat dengan jelas,

dilakukan juga visualisasi tersendiri untuk masing-masing klaster. Berikut kode pada **Gambar 14** untuk *clustering* dan visualisasi PCA:

```
% === 10. Clustering dan Visualisasi PCA ===
[idx, C] = kmeans(W, k_opt, 'Replicates', 10);

figure;
gscatter(W(:,1), W(:,2), idx);
xlabel('PC1'); ylabel('PC2');
title(sprintf('Clustering Negara berdasarkan PCA (k = %d)', k_opt));
legend('Location', 'best');
text(W(:,1)+0.1, W(:,2), country_labels, 'FontSize', 8);

% Visualisasi PCA per klaster

for k = 1:k_opt
    figure;
    cluster_idx = find(idx == k);

    scatter(W(cluster_idx,1), W(cluster_idx,2), 50, 'filled');
    text(W(cluster_idx,1)+0.1, W(cluster_idx,2), country_labels(cluster_idx), 'FontSize', 8);

    xlabel('PC1'); ylabel('PC2');
    title(sprintf('PCA Negara pada Klaster %d', k));
    grid on;
end
```

Gambar 14 *Clustering* dan Visualisasi Hasil PCA Semua Klaster

Visualisasi hasil klasterisasi negara dilakukan berdasarkan hasil *Principal Component Analysis* (PCA), khususnya dua komponen utama yaitu PC1 dan PC2. Pertama, dijalankan algoritma *K-means clustering* pada matriks W, yaitu representasi data negara dalam ruang PCA. Jumlah klaster ditentukan oleh *k_opt*, yang sebelumnya telah diperoleh dari metode silhouette. Kode *[idx, C] = kmeans(W, k_opt, 'Replicates', 10);* melakukan proses pengelompokan ini, di mana idx menyimpan label klaster untuk setiap negara dan C berisi koordinat pusat masing-masing klaster.

Selanjutnya, kode memvisualisasikan seluruh hasil klasterisasi dalam satu plot dua dimensi menggunakan fungsi gscatter(group scatter), yang menampilkan distribusi negara pada sumbu PC1 dan PC2, dengan pewarnaan berbeda untuk tiap klaster. Label nama negara ditambahkan menggunakan fungsi text, sehingga pengguna dapat mengidentifikasi negara mana yang berada di titik tertentu pada ruang PCA. Judul dan label sumbu ditambahkan untuk memperjelas makna visualisasi, serta legenda diletakkan secara otomatis pada lokasi terbaik.

Bagian berikutnya dari kode membuat visualisasi secara terpisah untuk tiap klaster agar dapat dianalisis lebih detail. Ini dilakukan dengan perulangan for dari 1 hingga *k_opt*, di mana untuk setiap klaster, negara-negara yang termasuk dalam klaster tersebut diambil menggunakan *find(idx == k)*. Kemudian, hanya data negara dalam klaster tersebut yang diplot dengan scatter, dan diberi label nama menggunakan text. Visualisasi ini memberi gambaran lebih spesifik mengenai karakteristik dan persebaran negara dalam tiap klaster di ruang PCA.

4. Penyimpanan anggota setiap klaster pada *file excel*

Untuk mempermudah proses analisis, dilakukan penyimpanan anggota setiap klaster. Dengan ini, analisis dapat ditinjau ulang dengan mudah secara manual, serta bisa dilakukan visualisasi lanjut. Anggota setiap cluster disimpan dalam *cell array*. Jumlah klaster disesuaikan oleh jumlah klaster optimal (*k_opt*). Setiap klaster akan disimpan dalam sheet masing-masing pada *file excel*. File akan disimpan dengan nama *Hasil_Klastering.xlsx*. Berikut kode pada **Gambar 15** yang memproses hal ini:

```
% === 11. Simpan anggota tiap klaster dalam tabel ===
cluster_tables = cell(k_opt, 1);
for k = 1:k_opt
    members = country_labels(idx == k)';
    T = table(members', 'VariableNames', {sprintf('Cluster_%d', k)});
    cluster_tables{k} = T;
end

% === Simpan hasil clustering dalam excel ===
output_filename = 'Hasil_Klastering.xlsx';

for k = 1:k_opt
    members = country_labels(idx == k)';
    T = table(members', 'VariableNames', {'Country'});
    writetable(T, output_filename, 'Sheet', sprintf('Cluster_%d', k));
end
```

Gambar 15 Penyimpanan anggota-anggota cluster dalam *excel*

5. Perhitungan rata-rata PC1, PC2, dan nilai *silhoutte* dari setiap klaster

Untuk memahami karakteristik setiap klaster dengan baik, dilakukan perhitungan terhadap rata-rata PC1, PC2, dan nilai silhoutte pada anggota-anggota setiap kluster. Dilakukan loop sebanyak jumlah klaster optimal (*k_opt*) terhadap setiap kluster. Untuk menentukan rata-rata digunakan fungsi mean. diambil indeks negara-negara yang tergolong dalam klaster tersebut dengan *idx == k*. Dengan indeks tersebut, program menghitung rata-rata nilai PC1 dan PC2 dari matriks W, yang merupakan hasil proyeksi data ke komponen utama PCA. Nilai-nilai ini memberi indikasi posisi rata-rata klaster dalam ruang dua dimensi utama yang menjelaskan sebagian besar variasi data. Selain itu, kode juga menghitung rata-rata nilai silhouette untuk anggota klaster. Berikut kode pada **Gambar 16** untuk melaksanakan hal ini dalam Matlab:

```
% === 12. Menghitung rata-rata PC1, PC2 dan silhouette tiap klaster ===
fprintf('\nKarakteristik Setiap Klaster:\n');
for k = 1:k_opt
    cluster_idx = find(idx == k);

    mean_PC1 = mean(W(cluster_idx, 1));
    mean_PC2 = mean(W(cluster_idx, 2));
    mean_silh = mean(silh_vals(cluster_idx));

    fprintf('Klaster %d:\n', k);
    fprintf('  Rata-rata PC1: %.4f\n', mean_PC1);
    fprintf('  Rata-rata PC2: %.4f\n', mean_PC2);
    fprintf('  Rata-rata Silhouette: %.4f\n\n', mean_silh);
end
```

Gambar 16 Perhitungan rata-rata PC1, PC2, nilai sillhoutte setiap klaster

4. Analisis Penerapan PCA dan Klastering

4.1. Skenario Pengujian

Pengujian ini dilakukan untuk menganalisis pola pembangunan global negara-negara dunia berdasarkan indikator indeks pembangunan global. Data set yang digunakan mencakup data pembangunan global dari tahun 2000 hingga 2020 dari berbagai negara dengan latar belakang ekonomi dan sosial yang beragam. Data diproses dalam beberapa tahap, termasuk pembersihan dan interpolasi data, normalisasi (*Z-Score*), reduksi dimensi melalui *Principal Component Analysis* (PCA), dan pengelompokan menggunakan algoritma *K-Means*. Dengan teknik PCA memungkinkan pengurangan dimensi indikator pembangunan yang awalnya tinggi menjadi beberapa komponen penting tanpa kehilangan informasi penting, sehingga proses *clustering* menjadi lebih efisien dan terarah.

Langkah awal dilakukan dengan mengidentifikasi nilai-nilai yang hilang, yang kemudian diisi menggunakan metode interpolasi *linear*, *pchip*, atau *spline* tergantung pada karakteristik variabel. Setelah semua nilai lengkap, dilakukan pembalikan arah untuk variabel yang memiliki pengaruh negatif seperti tingkat pengangguran dan inflasi. Langkah ini penting agar seluruh indikator memiliki arah interpretasi yang konsisten: semakin tinggi nilai, semakin baik pembangunan. Selanjutnya, nilai-nilai indikator distandarkan dan dimasukkan ke dalam PCA untuk mereduksi dimensi. Dari hasil PCA, data dikonversi ke dalam bentuk proyeksi yang hanya memuat komponen utama yang mewakili ≥ 80% variansi total. Proyeksi inilah yang digunakan untuk tahap akhir, yaitu *clustering* dengan *K-Means*. Analisis *silhouette coefficient* dilakukan untuk mendapatkan jumlah klaster terbaik.

Secara khusus, pengujian ini dilakukan dengan tiga tujuan utama berikut:

1. Mengelompokkan negara-negara berdasarkan kemiripan indikator pembangunan.
2. Menerapkan PCA untuk mereduksi dimensi data tanpa kehilangan informasi penting.
3. Mengevaluasi efektivitas metode *clustering* dengan *Silhouette Coefficient*.

4.2. Pengujian

Pengujian ini bertujuan untuk mengetahui seberapa efektif penggunaan metode PCA dan *K-Means* dalam mengelompokkan negara-negara berdasarkan kesamaan indikator pembangunan. Selain itu, juga akan mengetahui seberapa besar kontribusi PCA dalam menyederhanakan kompleksitas data tanpa menghilangkan informasi penting. Pengujian dan analisis dilakukan dengan membandingkan hasil visualisasi, struktur klaster, dan nilai evaluasi *silhouette coefficient* untuk memastikan secara menyeluruh untuk mengevaluasi kualitas pemisahan antar kelompok. Setiap hasil yang diperoleh dianalisis keberhasilan model dalam menyajikan pola global pembangunan negara dan memisahkan karakteristik pembangunan yang signifikan dalam empat klaster yang telah terbentuk.

4.3. Interpretasi Komponen PCA

Principal Component Analysis (PCA) adalah metode untuk mereduksi dimensi yang mentransformasi data berdimensi tinggi menjadi komponen utama yang saling *orthogonal* untuk menyederhanakan analisis. Komponen utama pertama (PC1) merupakan arah yang menjelaskan varians maksimum dalam data, sementara komponen kedua (PC2) menjelaskan varians maksimum kedua yang tegak lurus terhadap PC1. PC1 dan PC2 terbentuk dari kombinasi linear variabel asli dengan koefisien *loading* yang menunjukkan kontribusi relatif setiap variabel terhadap pembentukan komponen tersebut. Interpretasi nilai *loading* memungkinkan pemahaman terhadap makna substantif setiap komponen, seperti dalam konteks pembangunan ekonomi atau indeks kualitas hidup.

```
=== Kontribusi Variabel terhadap PC1 dan PC2 ===
Variabel                          Loading PC1 Loading PC2
---------------------------------------------------------------
gdp_per_capita                      +0.3211     -0.4547
inflation_rate                      -0.0293     +0.0497
life_expectancy                     +0.4280     -0.0622
school_enrollment_secondary         +0.0293     +0.0162
unemployment_rate                   +0.1722     +0.8476
electricity_access_pct              +0.4036     +0.1506
mobile_subscriptions_per_100        +0.3987     +0.0772
internet_usage_pct                  +0.4204     -0.1946
human_development_index             +0.4287     +0.0417
```

Gambar 17 Hasil Interpretasi Nilai *Loading* untuk Setiap Variabel

Hasil interpretasi nilai loading setiap variabel pada **Gambar 16** menunjukan besar kontribusi setiap variabel terhadap setiap komponen PCA. Analisis PCA mengidentifikasi kontribusi variabel terhadap PC1 dan PC2 berdasarkan nilai *loading* masing-masing komponen. PC1 didominasi variabel pembangunan yang memiliki loading tinggi pada variabel *human_development_index* (+0.4287), *life_expectancy* (+0.4280), *internet_usage_pct* (+0.4204), dan *electricity_access_pct* (+0.4036). *Loading* positif ini menunjukkan PC1 sebagai dimensi pembangunan manusia dan infrastruktur, di mana skor tinggi mencerminkan kesejahteraan dan akses teknologi yang superior.

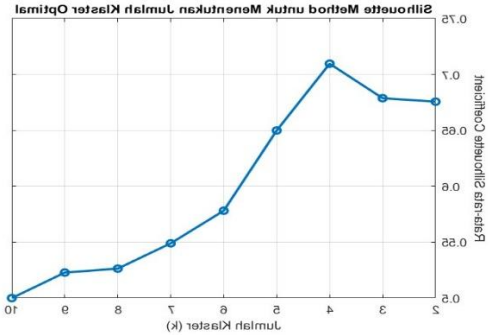
PC2 menampilkan pola berbeda dengan dominasi *unemployment_rate* (+0.8476) sebagai variabel utama. Berbeda dengan PC1, variabel

gdp_per_capita dan *internet_usage_pct* justru berkorelasi negatif dengan PC2, mengindikasikan hubungan terbalik terhadap pengangguran. PC2 merepresentasikan dimensi tekanan ekonomi dan instabilitas ketenagakerjaan, di mana nilai tinggi mencerminkan tingkat pengangguran elevasi tanpa korelasi langsung dengan level pembangunan.

Dengan hasil interpretasi ini, pemetaan negara dalam ruang dua dimensi PCA dapat memiliki makna yang lebih dalam. Skor tinggi PC1 mengindikasikan pembangunan ekonomi-sosial optimal, sedangkan skor tinggi PC2 menunjukkan tantangan struktural ketenagakerjaan. Oleh karena itu, analisis ini membantu mengelompokkan negara tidak hanya berdasarkan rata-rata nilai indikator, tetapi juga berdasarkan pola hubungan antar indikator yang mendasari kondisi pembangunan dan sosial-ekonomi. Interpretasi ini penting untuk memberikan konteks pada hasil klasterisasi.

4.4. Analisis Model Awal

Pada tahap awal proses kelompokan, penentuan jumlah klaster yang ideal sangat penting karena akan menentukan seberapa baik struktur data dapat diwakili dalam kelompok-kelompok yang relevan. Dalam penelitian ini, metode *Silhouette Coefficient* digunakan untuk menghitung sejauh mana setiap negara termasuk dalam klaster yang sesuai dibandingkan dengan klaster terdekat lainnya. Nilai koefisien pada kisaran nilai 0,5 hingga 1 menunjukkan klasifikasi yang sangat baik, sedangkan nilai positif dibawah itu menunjukan klasifikasi yang kurang baik, dan nilai koefisien negatif menunjukkan kemungkinan kesalahan penempatan. Proses pengujian dilakukan untuk berbagai nilai *k* (jumlah klaster), dimulai dari 2 hingga 10, dengan mempertimbangkan bahwa nilai terlalu kecil akan menghasilkan klaster yang terlalu umum, sementara nilai terlalu besar akan menghasilkan klaster yang terlalu spesifik dan tidak stabil.

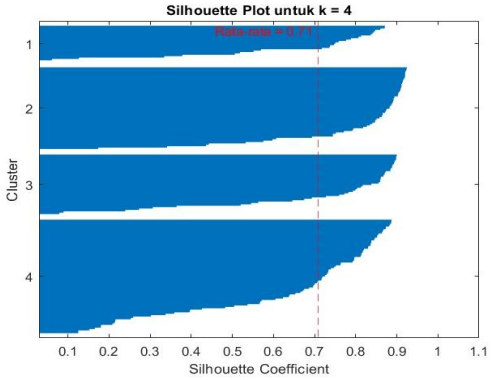


Gambar 18 *Silhouette Method* untuk Menentukan Jumlah Klaster Optimal

Berdasarkan hasil perhitungan dan visualisasi pada **Gambar 17**, nilai rata-rata *silhouette* tertinggi diperoleh pada *k = 4*, yang mengindikasikan bahwa pembagian negara ke dalam empat kelompok merupakan solusi paling optimal untuk mewakili struktur dan keragaman data indikator pembangunan global yang dianalisis.

4.5. Hasil Clustering Dengan Metode K-means

Setelah jumlah klaster optimal ditentukan menggunakan metode *Silhouette*, langkah selanjutnya adalah mengevaluasi hasil pengelompokan secara lebih mendalam untuk memastikan bahwa negara-negara telah diklasifikasikan secara tepat dan memiliki keterkaitan yang logis berdasarkan indikator pembangunan yang digunakan. Evaluasi ini dilakukan dengan menggunakan *Silhouette Plot*, yang memberikan gambaran visual seberapa baik setiap negara cocok berada dalam klaster yang ditentukan dibandingkan dengan klaster lainnya. Dengan analisis ini, dapat diketahui apakah proses clustering sudah efektif memisahkan kelompok negara yang serupa atau masih terdapat *overlap* antar klaster.

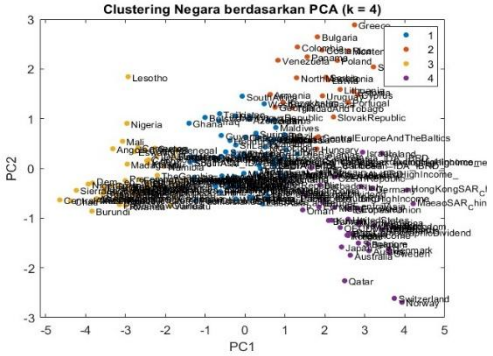


Gambar 19 Visualisasi *Silhouette* Plot

Seperti ditampilkan pada **Gambar 18**, di mana masing-masing batang menggambarkan tingkat kecocokan sebuah negara terhadap klaster yang diberikan berdasarkan nilai *silhouette* individualnya. Dari hasil pengamatan, negara-negara memiliki nilai rata-rata nilai *sillhoutte* 0,71, ini merupakan nilai yang besar sehingga negara-negara dalam *cluster* memiliki tingkat kecocokan yang baik terhadap klaster masing-masing. Selain itu, setiap *clustering* terlihat tanpa ada *cluster* yang tipis, hal ini mengartikan bahwa tidak ada *outlier* dan negara-negara terkelompok jumlah per klaster dengan baik. Pengamatan ini menguatkan bahwa struktur keseluruhan klaster sudah cukup baik dan secara statistik dapat diterima sebagai dasar untuk analisis lebih lanjut. Visualisasi dan Interpretasi Hasil

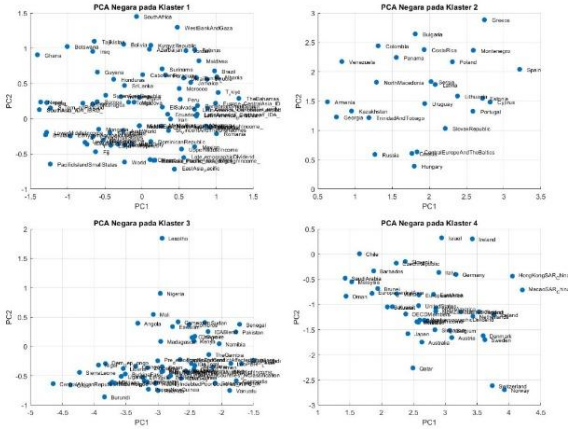
Langkah akhir dalam proses analisis ini adalah melakukan visualisasi hasil *clustering* dalam ruang dua dimensi menggunakan dua komponen utama dari PCA, untuk memperoleh gambaran lebih jelas mengenai hubungan antar negara dalam masing-masing klaster. Visualisasi ini penting karena tidak hanya memberikan konfirmasi visual terhadap hasil

clustering, tetapi juga memperlihatkan bagaimana kelompok-kelompok negara tersebar dalam ruang indikator pembangunan, serta membantu mengidentifikasi negara-negara yang berada di luar pola umum (*outlier*).



Gambar 20 Visualisasi *Clustering* Negara berdasarkan PCA (k = 4)

Pada Gambar 19 menunjukkan hasil proyeksi tersebut, di mana masing-masing titik pada mewakili satu negara dan warna menunjukkan klaster yang terbentuk dari proses *K-Means.* Terlihat jelas bahwa empat klaster dapat dibedakan dengan cukup baik meskipun terdapat beberapa titik yang dekat antar klaster.



Gambar 21 Visualisasi PCA Negara pada Setiap Klaster

Pada **Gambar 20** di atas ditunjukan visualisasi distribusi negara pada masing-masing klaster (1–4) dalam ruang dua dimensi berdasarkan dua komponen utama hasil PCA (PC1 dan PC2). Setiap *subgrafik* merepresentasikan klaster yang terbentuk dari algoritma *K-Means*, dan label negara digunakan untuk mengidentifikasi tiap titik data.

Tabel 5 Hasil Evaluasi Klaster (k = 4)

Klaster	Rata-rata PC1	Rata-rata PC2	Rata-rata Silhouette
1	-0.0271	0.1750	0.6438
2	1.8650	1.6426	0.6416
3	-2.8559	-0.2288	0.8159
4	2.7459	-1.0328	0.7302

Tabel 6 Evaluasi Kateristik Setiap Klaster

Klaster	Kateristik Umum
1	Negara dengan tingkat pembangunan menengah dan seimbang. Tidak menonjol secara ekstrem, cenderung moderat dalam aspek ekonomi maupun sosial.
2	Negara dengan keunggulan sosial dan kelembagaan, memiliki kualitas hidup dan tata kelola yang baik, meskipun tidak dominan dalam kekuatan ekonomi.
3	Negara-negara dengan tingkat pembangunan sangat rendah, tertinggal dalam banyak aspek, terutama kondisi sosial dan ekonomi.
4	Negara maju secara ekonomi dan teknologi, tetapi mungkin menghadapi tantangan sosial atau ketimpangan dalam aspek keadilan sosial.

Informasi pada **Tabel 5** dan **Tabel 6** menunjukkan ringkasan hasil *clustering* empat klaster yang terbentuk berdasarkan algoritma *K-Means* setelah dilakukan reduksi dimensi dengan *Principal Component Analysis* (PCA). Setiap klaster terdiri dari sejumlah negara dengan pola indikator pembangunan yang serupa, baik dalam aspek ekonomi, sosial, maupun kelembagaan. Nilai rata-rata PC1 dan PC2 memberikan gambaran posisi relatif tiap klaster dalam ruang komponen utama, sedangkan nilai rata-rata *silhouette* digunakan untuk menilai kualitas pemisahan antar klaster.

Rata-rata nilai *silhouette* yang cukup tinggi pada keempat klaster mengindikasikan bahwa hasil pengelompokan memiliki konsistensi internal yang baik dan pemisahan yang memadai antar kelompok. Klaster 3, dengan nilai *silhouette* tertinggi, menunjukkan bahwa negara-negara dalam kelompok ini sangat konsisten dalam karakteristik keterbelakangannya, sementara klaster lain mencerminkan keragaman negara dari tingkat pembangunan menengah hingga negara dengan keunggulan kelembagaan. Secara keseluruhan, tabel ini menegaskan bahwa kombinasi metode PCA dan *K-Means* berhasil mengidentifikasi struktur pembangunan global yang jelas dan terkelompok secara statistik, serta dapat memberikan pemahaman yang mendalam mengenai posisi relatif masing-masing negara dalam konteks pembangunan multidimensi.

5. Kesimpulan dan Saran

5.1. Kesimpulan

Berdasarkan hasil penelitian dengan menggunakan metode *Principal Component Analysis* dan *clustering* pada *dataset Global Development Indicators* tahun 2000-2020, dapat disimpulkan bahwa penerapan PCA dapat dengan efektif mereduksi kompleksitas data indikator pembangunan yang semula memiliki banyak variabel menjadi sejumlah variabel yang mampu mewakili lebih dari

80% total variansi. Proses reduksi menyederhanakan proses analisis tetapi tetap mempertahankan esensi informasi penting yang terkandung dalam data asli. Hal ini terbukti dari hasil proyeksi dua dimensi PCA dari komponen utama yang mampu membedakan secara jelas karakteristik pembangunan negara-negara.

Algoritma *K-mean clustering* yang diterapkan dengan hasil proyeksi PCA berhasil mengelompokan negara-negara dalam klaster-klaster pembangunan. Diperoleh *silhuoette coefficient* maksimum dengan nilai yang cukup tinggi yakni 0,7 pada pengelompokan dengan 4 *cluster*. Hasil *cluster* dengan pembagian menjadi empat klaster menampilkan kumpulan klaster dengan anggota negara yang merata dengan baik serta rata-rata *sillhoutte value* yang tinggi hingga mencapai 0,71. Hal ini menunjukkan bahwa klaster-klaster yang diperoleh bersifat konsisten dan kompak antar anggota klaster dengan pemisahan yang baik.

Secara keseluruhan, pendekatan ini berhasil menyajikan gambaran global yang baik mengenai pembangunan negara-negara di dunia. Penggunaan PCA dalam penelitian ini meningkatkan efisiensi dan kejelasan visualisasi serta mendukung akurasi dalam mengidentifikasi pola-pola pembangunan global. Temuan ini menegaskan bahwa metode PCA sangat bermanfaat dan efektif dalam analisis data multivariat yang kompleks.

5.2. Saran

Untuk pengembangan lebih lanjut, terdapat beberapa saran pengembangan. Pertama, metode *clustering* alternatif lainnya seperti *Hierarchical Clustering* atau DBSCAN dapat menjadi perbandingan untuk menguji konsistensi hasil pengelompokan yang diperoleh dari *K-means*. Selain itu, pendekatan temporal melalui *analisis time-series clustering* dapat digunakan untuk mengamati dinamika perubahan pembangunan dari waktu ke waktu untuk mengidentifikasi tren pertumbuhan dan pergeseran posisi negara dalam klaster secara lebih mendalam.

REFERENSI

[1] A. Reza Hariyadi, "Dinamika Kebijakan Perencanaan Pembangunan Nasional Indonesia," *JDKP Jurnal Desentralisasi dan Kebijakan Publik*, vol. 2, no. 2, hlm. 259–276, Sep 2021, doi: 10.30656/jdkp.v2i2.3887.

[2] Sultan, H. C. Rahayu, dan Purwiyanta, "Analisis Pengaruh Kesejahteraan Masyarakat Terhadap Pertumbuhan Ekonomi di Indonesia," *Jurnal Informatika Ekonomi Bisnis*, hlm. 75–83, Mar 2023, doi: 10.37034/infeb.v5i1.198.

[3] United Nations, "The Sustainable Development Goals Report," 2024.

[4] A. Verma, O. Angelini, dan T. Di Matteo, "A new set of cluster driven composite development indicators," *EPJ Data Sci*, vol. 9, no. 1, Des 2020, doi: 10.1140/epjds/s13688-020-00225-y.

[5] E. Emawati Chotim, "PEMBANGUNAN BERKELANJUTAN DENGAN DIMENSI EKONOMI, EKOLOGI, DAN SOSIAL DI INDONESIA," vol. 4, no. 1, 2020.

[6] M. Nilashi, O. Keng Boon, G. Tan, B. Lin, dan R. Abumalloh, "Critical Data Challenges in Measuring the Performance of Sustainable Development Goals: Solutions and the Role of Big-Data Analytics," *Harv Data Sci Rev*, vol. 5, no. 3, Jul 2023, doi: 10.1162/99608f92.545db2cf.

[7] Y. Xiong dkk., "Assessing Sustainable Development through Multidimensional Framework of Synergies and Trade-offs," *Ecosystem Health and Sustainability*, vol. 11, 2025, doi: 10.34133/ehs.0351.

[8] Y. Linawati, H. Suzantia, dan M. G. Wibowo, "Dampak Tata Kelola Pemerintahan Terhadap Pertumbuhan Ekonomi dan Indeks Pembangunan Manusia: Studi Kasus Negara Berkembang OKI," *TEMALI : Jurnal Pembangunan Sosial*, vol. 4, no. 2, hlm. 133–144, Sep 2021, doi: 10.15575/jt.v4i2.12547.

[9] Haris Faozan, "TATA KELOLA PEMERINTAHAN DAERAH YANG BAIK DAN PERTUMBUHAN EKONOMI LOKAL YANG MENIMBULKAN PEMBANGUNAN DAERAH," vol. VII, Jun. 2022.

[10] U. U. Suhardi, U. Pribadi, dan Z. Losi, "The Effects of Good Governance Principles: Accountability, Transparency, and Participation on Public Trust in Village Funds Management," *International Journal of Social Science and Business*, vol. 7, no. 4, hlm. 1050–1060, Nov 2023, doi: 10.23887/ijssb.v7i4.57648.

[11] D. I. Ginting dan I. Lubis, "Pengaruh Angka Harapan Hidup dan Harapan Lama Sekolah Terhadap Indeks Pembangunan Manusia," Des 2023.

[12] Indah Budiastutik, Elly Trisnawati, Giska Hediyanti, Nurul Amaliyah, dan Resky Nanda Pranaka, "Hubungan Penyakit Menular, Sumber Air Bersih, Praktik Kebersihan, dan Sanitasi dengan Kejadian Stunting: Studi Kasus Kontrol di Kabupaten Sambas," *Amerta Nutrition*, Agu 2024.

[13] Dhiaulhaq Luqyana Nizhamul, Vedelya Istighfarah, dan Novri Dwi Damayanti, "Faktor-Faktor Pengaruh Indeks Pembangunan Manusia (IPM) di Hongkong dan Singapura," Nov 2023, doi: 10.25157/je.v11i2.11808.

[14] N. Khairani dan T. Sendjaja, "Akselerasi Transformasi Digital sebagai Katalisator Pertumbuhan Ekonomi: Studi Komparatif Kebijakan Singapura dan Indonesia," Des 2024, vol. 5, no. 12, hlm. 2094.

[15] RHAINA NAWANG WULAN, "Pengaruh Pertumbuhan Bisnis Digital terhadap Pertumbuhan Ekonomi di Indonesia dengan Penyerapan Tenaga Kerja sebagai Variabel Intervening," Des 2021.

[16] Mohammad Rudy Salahuddin, "Pengembangan Ekonomi Digital Indonesia 2030," 2024.

[17] Abdul Karim, "Transformasi Digital dalam Menunjang Pertumbuhan Ekonomi," Jan 2025.

[18] I. Yasmin, "Infrastruktur Air Bersih yang Ada di Indonesia," Sep 2023. [Daring]. Tersedia pada: https://www.researchgate.net/publication/373977763

[19] A. A. Nasution, T. Neyatri Bandrang, D. M. Widiniarsih, M. Syaiful, dan A. R. Munir, "Peran Ekonomi Digital terhadap Ketahanan dan Pertumbuhan Ekonomi di Indonesia," *Universitas Muhammadiyah Pringsewu*, vol. 5, no. 8, hlm. 4224, 2024.

[20] NIZAR FUADI, "Penguasaan Kedaulatan Konektivitas Digital Guna Mendukung Akselerasi Ekonomi dalam Rangka Meningkatkan Ketahanan Nasional," Agu 2023.

[21] S. Rathod, S. Banda, dan N. Bandumula, "Statistical Procedures for Analyzing Agricultural Data using R," 2023. [Daring]. Tersedia pada: https://www.researchgate.net/publication/391930391

[22] J. Qian dkk., "Structured illumination microscopy based on principal component analysis," *eLight*, vol. 3, no. 1, Des 2023, doi: 10.1186/s43593-022-00035-x.

[23] D. Hediyati dan I. M. Suartana, "Penerapan Principal Component Analysis (PCA) untuk Reduksi Dimensi pada Proses Clustering Data Produksi Pertanian di Kabupaten Bojonegoro," Surabaya, Feb 2021.

[24] M. Yafi, R. Goejantoro, A. Tri, dan R. Dani, "K-Medoids Algorithm Clustering with Principal Component Analysis (PCA) (Case Study: Districts/Cities on the Borneo Island Based on Poverty Indicators)," Feb 2023, doi: 10.14710/JSUNIMUS.11.2.2023.31-43.

[25] R. R. Muhima, M. Kurniawan, S. R. Wardhana, A. Yudhana, Sunardi, dan M. Adhimukti, "An Improved Clustering Based on K-Means for Hotspots Data," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 2, hlm. 1109–1117, Agu 2023, doi: 10.11591/ijeecs.v31.i2.pp1109-1117.

[26] S. Surono, K. W. Goh, C. W. Onn, dan F. Marestiani, "Developing an Optimized Recurrent Neural Network Model for Air Quality Prediction Using K-Means Clustering and PCA Dimension Reduction," *International Journal of Innovative Research and Scientific Studies*, vol. 6, no. 2, hlm. 330–343, 2023, doi: 10.53894/ijirss.v6i2.1427.

[27] M. Cui, "Introduction to the K-Means Clustering Algorithm Based on the Elbow Method," Jan 2020, doi: 10.23977/accaf.2020.010102.

[28] E. S. Dalmaijer, C. L. Nord, dan D. E. Astle, "Statistical Power for Cluster Analysis," *BMC Bioinformatics*, vol. 23, no. 1, Des 2022, doi: 10.1186/s12859-022-04675-1.

[29] K. P. Sinaga dan M. S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, hlm. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[30] "Reconstruct missing data," MATLAB & Simulink Example, https://www.mathworks.com/help/signal/ug/reconstruct-missing-data.html (accessed Aug. 7, 2025).

[31] MathWorks, "Data Preprocessing Techniques," Juni 2024, MathWorks. [Daring]. Tersedia pada: https://www.mathworks.com/discovery/data-preprocessing.html

[32] "Principal Components of a Data Set," 2021. [Daring]. Tersedia pada: https://www.mathworks.com/help/stats/pca.html

[33] "Reconstruct missing data," MATLAB & Simulink Example, https://www.mathworks.com/help/signal/ug/reconstruct-missing-data.html (accessed Aug. 7, 2025).

[34] "Train a K-Means Clustering Algorithm," 2024. [Daring]. Tersedia pada: https://www.mathworks.com/help/stats/kmeans.html

Mika Valentino, saat ini sebagai Mahasiswa program studi Teknik Informatika Universitas Tarumanagara.

Yosia Sipahutar, saat ini sebagai Mahasiswa program studi Teknik Informatika Universitas Tarumanagara.

Muhammad Farhan, saat ini sebagai Mahasiswa program studi Teknik Informatika Universitas Tarumanagara.