Terbit pada laman web jurnal : **http://ojsamik.amikmitragama.ac.id**

JURNAL JARINGAN SISTEM INFORMASI ROBOTIK (JSR)
Vol. 9 No. 1 TAHUN 2025 E - ISSN : 2579-373X

KOMBINASI HYBRID K-MEANS UNTUK KLASTERISASI MULTIVARIAT DALAM ANALISIS STUNTING

Devi Sartika[1], Febie Elfaladonna[2*], Ayu Octarina[2,*]
[1,2,3] D4 Manajemen Informatika, Politeknik Negeri Sriwijaya, Palembang, Indonesia
Email: [1]devi_sartika_mi@polsri.ac.id, [2,*]febie_elfaladonna_mi@polsri.ac.id, , [3]ayu_octarina_mi@polsri.ac.id,

Abstrak
Penelitian ini mengintegrasikan metode Principal Component Analysis (PCA) dan K-Means clustering untuk melakukan klasterisasi multivariat pada dataset stunting. PCA diterapkan untuk mengurangi dimensi data yang kompleks, dengan mempertahankan sekitar 69,65% dari variansi total melalui 17 komponen utama, sehingga analisis menjadi lebih efisien tanpa kehilangan informasi penting. Setelah reduksi dimensi, algoritma K-Means digunakan untuk mengelompokkan individu berdasarkan kesamaan karakteristik, dengan jumlah cluster optimal ditentukan menggunakan analisis Within-Cluster Sum of Squares (WCSS). Hasil klasterisasi membagi data menjadi dua cluster yang memiliki perbedaan karakteristik, yang mencerminkan variasi dalam faktor-faktor yang mempengaruhi stunting, seperti status gizi, akses terhadap layanan kesehatan, dan faktor sosial ekonomi. Penerapan PCA dan K-Means clustering memberikan pemahaman yang lebih jelas mengenai pola dan distribusi faktor penyebab stunting, serta mendukung analisis stunting lebih lanjut.

Kata Kunci: Principal Component Analysis, K-Means Clustering, Stunting, Klasterisasi Multivariat, Analisis Data.

Abstract
This study integrates Principal Component Analysis (PCA) and K-Means clustering methods to perform multivariate clustering on stunting datasets. PCA was applied to reduce the dimensionality of complex data, retaining about 69.65% of the total variance through 17 principal components, making the analysis more efficient without losing important information. After dimensionality reduction, the K-Means algorithm was used to cluster individuals based on similar characteristics, with the optimal number of clusters determined using Within-Cluster Sum of Squares (WCSS) analysis. The clustering results divided the data into two clusters with different characteristics, reflecting variations in factors that influence stunting, such as nutritional status, access to health services, and socioeconomic factors. The application of PCA and K-Means clustering provides a clearer understanding of the pattern and distribution of factors that cause stunting, and supports further analysis of stunting.

Keywords: Principal Component Analysis, K-Means Clustering, Stunting, Multivariate Clustering, Data Analysis.

1. PENDAHULUAN

Stunting atau keterlambatan pertumbuhan, merupakan masalah gizi kronis yang banyak dialami oleh anak-anak di berbagai negara, termasuk Indonesia. Kondisi ini tampak ketika seorang anak memiliki tinggi badan yang lebih rendah dari standar normal untuk usianya. Anak yang mengalami stunting tidak hanya menunjukkan hambatan dalam pertumbuhan fisik dibandingkan dengan anak seusianya, tetapi juga berisiko mengalami gangguan pada perkembangan kognitif serta kemampuan belajarnya [1].
Menurut Survei Kesehatan Indonesia pada tahun 2023 yang diterbitkan oleh Kementerian Kesehatan, prevalensi stunting di Indonesia saat ini tercatat pada angka 21,5 persen. Angka ini hanya turun sebesar 0,1 persen dibandingkan data Survei Status Gizi Balita Indonesia tahun 2022, yang mencatat prevalensi stunting sebesar 21,6 persen. Capaian ini masih cukup jauh dari target penurunan stunting sebesar 14 persen yang ditetapkan untuk tahun 2024 [2].
Ada banyak pihak yang dilibatkan dalam penanganan stunting di Indonesia termasuk salah satunya adalah puskesmas. Puskesmas berperan sebagai pelopor dalam upaya penurunan kasus stunting [3]. Hal ini dikarenakan puskesmas menjadi jembatan komunikasi antara pemerintah dengan masyarakat.

Puskesmas XYZ, yang terletak di Sumatera Selatan, merupakan salah satu fasilitas kesehatan utama di wilayah tersebut. Namun, Puskesmas ini menghadapi tantangan signifikan dalam mengelompokkan status gizi balita berdasarkan berbagai parameter yang berkaitan dengan penyebab stunting. Kendala ini diperburuk oleh kurangnya ketelitian kader posyandu dalam melakukan penimbangan balita serta minimnya tindak lanjut berupa analisis mendalam. Akibatnya, proses pemetaan status gizi balita menjadi tidak optimal, sehingga menyulitkan identifikasi kelompok risiko tinggi dan intervensi yang tepat.

Penelitian ini menawarkan solusi melalui pendekatan Hybrid K-Means untuk klasterisasi multivariat dalam analisis stunting. Pendekatan ini memadukan algoritma K-Means dengan metode lain untuk meningkatkan akurasi dan efektivitas klasterisasi, memungkinkan pengelompokan status gizi balita secara lebih akurat berdasarkan berbagai parameter sosial, ekonomi, kesehatan, dan lingkungan.

Salah satu metode yang bisa dikombinasikan dengan Kmeans Clustering untuk pemecahan permasalahan di atas adalah PCA [4]. Prinsip PCA yaitu mengubah bentuk sekumpulan variabel asli menjadi kumpulan variabel yang lebih kecil. Dalam analisis Principal Component Analysis (PCA), langkah pertama yang perlu dilakukan adalah membentuk matriks korelasi. Matriks ini bertujuan untuk mengidentifikasi sejauh mana hubungan atau keterkaitan antara variabel-variabel yang dianalisis [5].

Penerapan PCA dalam algoritma K-Means memungkinkan pengurangan dimensi data tanpa kehilangan informasi penting yang terkandung di dalamnya seperti pada penelitian yang dilakukan oleh [6] , teknik clustering K-Means dan PCA digunakan untuk mengelompokkan tingkat pendidikan masyarakat di Kabupaten Semarang berdasarkan variabel jenis kelamin, umur, dan status individu dalam keluarga. Hasilnya menunjukkan bahwa penerapan PCA pada algoritma K-Means dapat mengurangi dimensi data tanpa mengurangi informasi penting. Setelah penerapan PCA, diperoleh dua komponen utama yang secara kumulatif mewakili 70% dari total variabilitas data. Setelah dilakukan reduksi dimensi dengan PCA, analisis cluster dilanjutkan menggunakan algoritma K-Means, yang menghasilkan empat kelompok dengan karakteristik berbeda untuk setiap cluster.

Terdapat beberapa penelitian lain yang terkait dengan Kmeans dan PCA, yang meneliti tentang klasifikasi Kabupaten/Kota di Pulau Kalimantan Berdasarkan Indikator Tingkat Pengangguran Terbuka. Penelitian ini menyimpulkan Jumlah komponen utama (principal component) yang terbentuk dari hasil reduksi variabel menggunakan metode PCA dengan kriteria nilai eigen yanglebih besar atau sama dengan satuadalah sebanyak dua komponen [7].

Penelitian selanjutnya menyatakan untuk mengatasi asumsi yang tidak terpenuhi, maka data harus bebas dari multikolinearitas. Dengan analisis PCA guna mereduksi multikolinearitas tersebut. Dua faktor digunakan berdasarkan eigenvalue yang lebih besar dari satu. Hasil pengelompokkan menggunakan metode K-Means menghasilkan 3 cluster terbaik, dengan jumlah anggota masing-masing 12, 8, dan 7 kabupaten/ kota. Berdasarkan profil data, kabupaten/ kota yang masuk dalam cluster 2 memerlukan penanganan prioritas karena memiliki nilai RLS, daya beli, dan proporsi penduduk dengan jaminan kesehatan yang rendah, sehingga menyebabkan tingginya persentase penduduk miskin di wilayah tersebut [8]. Selanjutnya adalah penelitian yang dilakukan terhadap uji simulasi data dalam skala besar, yang memerlukan metode klasifikasi yang efektif, salah satunya adalah Radial Basis Function Neural Network (RBFNN). Dalam pelatihan data RBFNN, digunakan struktur khusus yang melibatkan dimensi tinggi pada hidden layer. Namun, struktur ini sering menimbulkan masalah karena ukuran hidden layer yang terlalu besar, sehingga diperlukan pendekatan penyederhanaan jaringan seperti PCA dan K-Means Clustering. PCA digunakan untuk mereduksi dimensi input pada RBFNN, sementara K-Means Clustering digunakan untuk menentukan inisialisasi pusat awal RBFNN. Hasil eksperimen dengan metode PCA menunjukkan bahwa komponen utama pertama dan kedua masing-masing mewakili 55,2288% dan 27,3108% dari total variabilitas, dengan kedua komponen utama tersebut secara kumulatif menyumbang 82,5396%. Hasil percobaan iterasi menunjukkan bahwa akurasi rata-rata pada proses training dan testing di PC-2 Klas-3 mencapai lebih dari 90%, dengan tingkat kesalahan klasifikasi di bawah 10% [9].

Penelitian berikutnya bertujuan untuk mengembangkan sistem rekomendasi film dengan menggabungkan metode Collaborative Filtering, PCA, dan K-Means. PCA diterapkan pada data untuk mempercepat proses clustering. Rata-rata kompleksitas waktu yang diperoleh adalah 1,061282. Proses clustering digunakan untuk menentukan karakteristik pengguna berdasarkan kesamaan dengan pengguna lainnya. Hasil pengujian menggunakan Silhouette Coefficient dan metode Elbow menunjukkan bahwa nilai k terbaik adalah 3. Rekomendasi yang dihasilkan kemudian dihitung menggunakan Mean Reciprocal Rank (MRR) untuk menilai tingkat akurasi rekomendasi. Rata-rata nilai MRR yang diperoleh adalah 0,44533417402269865, yang menunjukkan bahwa rekomendasi yang dihasilkan kurang tepat [10].

Pada dataset yang didapatkan dari mitra, jumlah keseluruhan variabel adalah 27 variabel yaitu : BB Lahir, TB Lahir, Usia Saat Ukur, Berat, Tinggi,

Cara Ukur, LiLA, BB/U, TB/U, BB/TB, ASI Ekslusif, Kesesuaian Menu MPASI, Air Bersih, Kecacingan, Jamban Sehat, Imunisasi, Merokok (Keluarga), Kelahiran, Penyakit Penyerta, Ketercukupan Mikronutrien, Pekerjaan Kepala Keluarga, Penghasilan Total keluarga, Pendidikan Ayah, Pendidikan Ibu, Tinggi Ayah (cm), Tinggi Ibu (cm). Variabel-variabel tersebut mencakup berbagai aspek sosial, ekonomi, lingkungan, dan kesehatan yang saling berhubungan dan berpotensi mempengaruhi prevalensi stunting.

Untuk mengatasi kompleksitas data yang besar dan keterkaitan antar variabel, penerapan pendekatan Hybrid K-Means yang dipadukan dengan *Principal Component Analysis* (PCA) dapat menjadi solusi yang efektif. Dengan menggunakan PCA untuk mereduksi dimensi data, informasi penting dapat tetap dipertahankan, namun dengan representasi yang lebih sederhana dan lebih mudah dikelola. Pendekatan ini memungkinkan analisis lebih terfokus pada pola-pola utama yang mengungkap faktor-faktor risiko stunting yang paling signifikan, serta memberikan pemahaman yang lebih mendalam dalam merancang kebijakan intervensi yang lebih tepat dan terarah. Kombinasi hybrid K-Means dan PCA juga dapat meningkatkan akurasi dalam proses klasterisasi data, yang pada gilirannya akan memperbaiki pemetaan status gizi dan risiko stunting secara lebih efektif.

Prinsip pengelompokan data (*clustering*) dengan prinsip utama membentuk k prototipe atau pusat massa (*centroid*) yang mewakili rata-rata dari kumpulan data berdimensi n. Dalam metode ini, nilai k harus ditentukan sebelumnya (*a priori*). Algoritma K-Means dimulai dengan menetapkan *prototipe cluster* awal, yang kemudian diperbarui secara iteratif hingga mencapai kondisi konvergen, yaitu saat perubahan pada *prototipe cluster* tidak lagi signifikan. Proses perubahan ini diukur melalui fungsi objektif J, yang biasanya didefinisikan sebagai jumlah atau rata-rata jarak setiap data ke pusat massa kelompoknya [11]. Langkah-langkah dalam melakukan klasterisasi menggunakan metode K-Means adalah sebagai berikut [12]:

1. Menentukan nilai k sebagai jumlah klaster yang akan dibentuk.
2. Memilih nilai awal secara acak atau random untuk centroid awal sebanyak k, kemudian menghitung jarak setiap data input ke masing-masing centroid dengan menggunakan rumus Euclidean Distance berikut:

$$d(xi, \mu j) = \sqrt{\sum (xi - \mu j)^2} \quad (1)$$

Keterangan:
xi : data kriteria
μj : centroid pada cluster ke-js

3. Mengelompokkan setiap data ke klaster berdasarkan jarak terdekatnya dengan centroid.
4. Memperbarui centroid dengan menghitung nilai baru, yang diperoleh dari rata-rata data dalam klaster terkait, menggunakan rumus berikut:

$$\mu j(t+1) = \frac{1}{Nsj} \sum_{j \in sj} xj \quad (2)$$

Keterangan:
μj(t+1) : centroid baru pada iterasi (t+1)
Nsj : data pada cluster Sj

5. Jika data dalam setiap klaster belum stabil, ulangi langkah 2 hingga 5 hingga tidak ada lagi perubahan dalam keanggotaan setiap klaster.

Clustering menggunakan pendekatan unsupervised learning, di mana tidak memerlukan fase pembelajaran (learning) dan tidak melibatkan pelabelan pada setiap kelompok [13]. *Principal Component Analysis* (PCA) adalah teknik statistik *multivariat* yang secara *linear* mengubah kumpulan variabel asli menjadi sejumlah variabel baru berukuran lebih kecil yang tidak saling berkorelasi. Variabel-variabel baru ini tetap mampu merepresentasikan informasi dari variabel asli. Tujuan utama PCA adalah menjelaskan sebanyak mungkin varian dari data asli menggunakan sesedikit mungkin komponen utama, yang disebut faktor [14] . Metode PCA membentuk serangkaian dimensi baru yang diurutkan berdasarkan besar varian data yang diwakilinya. PCA menghasilkan komponen utama (principal components) yang diperoleh melalui dekomposisi nilai eigen (eigenvalue) dan vektor eigen (eigenvector) dari matriks kovariansi [15]. Tahapan algoritma PCA dapat dijelaskan sebagai berikut:

1. Perhitungan rata-rata (*mean*) tiap dimensi menggunakan persamaan:

$$X = \frac{1}{n} \sum_{i=1}^{n} Xi \quad (3)$$

Keterangan:
n : jumlah data sampel
Xi : data sampel

2. Menghitung *covariance matrix* (Cx) menggunakan persamaan:

$$X = \frac{1}{n} \sum_{i=1}^{n} Xi \quad (4)$$

Keterangan:
n : jumlah data sampel
Xi : data sampel

3. Menghitung *eigenvector* (vm) dan *eigenvalue* (λm) dari *covarince matrix* menggunakan persamaan:

$$C_x v_m = \lambda_m v_m \quad (5)$$

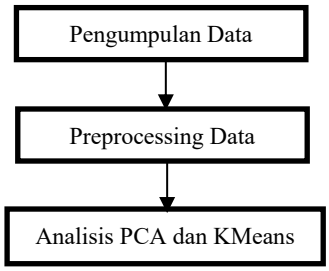
Keterangan:
λm : *eigenvalue*
vm : *eigenvector*
Cx : *matriks kovarian*

4. Urutkan eigenvalue dalam urutan menurun (descending). Principal Component (PC) adalah kumpulan eigenvector yang diurutkan berdasarkan nilai eigenvalue dari langkah sebelumnya.
5. Membentuk dataset baru berdasarkan hasil transformasi tersebut.

Berdasarkan latar belakang ditas, diharapkan hasil penelitian ini dapat mendukung Puskesmas XYZ dalam meningkatkan kualitas pemetaan stunting, memperbaiki manajemen data, dan memperkuat intervensi kesehatan masyarakat di wilayahnya.

2. METODOLOGI PENELITIAN

Analisis faktor-faktor yang mempengaruhi prevalensi stunting dengan menerapkan kombinasi Hybrid K-Means dalam klasterisasi multivariat, yang dipadukan dengan Principal Component Analysis (PCA) digunakan untuk mengurangi dimensi data. Penelitian ini mencakup serangkaian langkah yang terstruktur untuk memastikan hasil analisis yang valid dan dapat diandalkan. Adapun tahapan penelitian dapat dilihat pada gambar berikut:



Gambar 1. Tahapan Penelitian

2.1 Pengumpulan Data

Berikut ini merupakan potongan dataset yang didapatkan dari Puskesmas XYZ dengan total variabel 27 dan jumlah baris data 122 yang mencakup faktor-faktor sosial, ekonomi, lingkungan, dan kesehatan yang dapat memengaruhi status gizi balita dan risiko stunting. Variabel tersebut meliputi BB Lahir, TB Lahir, Usia Saat Ukur, Berat, Tinggi, LiLA, BB/U, TB/U, BB/TB, ASI Eksklusif, Kesesuaian Menu MPASI, Air Bersih, Kecacingan, Jamban Sehat, Imunisasi, Merokok (Keluarga), Kelahiran, Penyakit Penyerta, Ketercukupan Mikronutrien, Pekerjaan Kepala Keluarga, Penghasilan Total Keluarga, Pendidikan Ayah, Pendidikan Ibu, Tinggi Ayah, dan Tinggi Ibu.

JK	BB Lahir	TB Lahir	Usia Saat Ukur	Berat	Tinggi	Cara Ukur	LiLA	BB/U	TB/U	BB/TB
0	3	51	4.1	9.1	84	0	11	0	0	0
1	4	50	2.1	9.8	80	0	17	1	1	1
1	3.8	51	4.5	13.1	94.1	0	16	1	1	1
1	3	49	1.9	8.7	77	1	16	1	1	1
1	2.5	46	3.1	11.7	88.5	0	16	2	1	1
1	2.8	50	4.3	13.1	95.6	0	17	2	1	1
0	4	51	3.10	12.3	90.7	0	17	2	1	1
0	2.6	49	3.0	11.2	87	0	17	2	1	1
0	2.3	48	2.9	12.2	84.5	0	18	2	1	1
1	3.4	51	4.10	12.7	97.6	0	16	1	1	1

Show 10 per page 1 2 10 13

Gambar 2. Dataset Balita di Puskesmas XYZ

2.2 Preprocessing Data

Preprocessing data adalah langkah krusial dalam analisis data mining yang bertujuan untuk membersihkan, mengubah format, dan mempersiapkan data sehingga lebih siap dan tepat untuk dianalisis [16]. Penelitian ini menggunakan beberapa tahapan yang diperlukan dalam preprocessing data, seperti:

1. Normalisasi Data: Menerapkan teknik seperti min-max scaling atau z-score *standardization* untuk menyamakan skala semua variabel, sehingga tidak ada satu variabel pun yang lebih dominan dalam proses klasterisasi. Pada dataset, proses normalisasi hanya menggunakan min-max scaling dengan hasil dari proses normalisasi seperti berikut:

JK	BB Lahir	TB Lahir	Usia Saat Ukur	Berat
0.0	0.409090909090906	0.7272727272727275	0.8333333333333331	0.32954545454545436
1.0	0.8636363636363636	0.6363636363636367	0.4166666666666667	0.409090909090917
1.0	0.7727272727272726	0.7272727272727275	0.9166666666666665	0.7840909090909
1.0	0.409090909090906	0.5454545454545459	0.37499999999999994	0.28409090909090895
1.0	0.18181818181818177	0.2727272727272725	0.6249999999999999	0.6249999999999998
1.0	0.3181818181818181	0.6363636363636367	0.8749999999999999	0.7840909090909
0.0	0.8636363636363636	0.7272727272727275	0.6249999999999999	0.6931818181818181
0.0	0.2272727272727718	0.5454545454545459	0.6041666666666666	0.5681818181818181
0.0	0.090909090909072	0.45454545454545414	0.5833333333333333	0.6818181818181817
1.0	0.5909090909090909	0.7272727272727275	0.8333333333333331	0.7386363636363635

Show 10 per page

Gambar 3. Hasil normalisasi min-max scaling

Persamaan yang digunakan untuk proses normalisasi dengan min-max scaling adalah:

$$\text{Normalized Value} = \frac{\text{Value} - \text{Min}}{\text{Max} - \text{Min}} \quad (6)$$

Keterangan:
Value : nilai asli dari data.
Min dan Max : nilai minimum dan maksimum dalam kolom tersebut.
Hasil di atas menunjukkan bahwa nilai-nilai telah dinormalisasi ke dalam rentang [0, 1], sehingga tidak ada atribut yang memiliki skala lebih besar atau lebih kecil yang dapat mempengaruhi atau mendominasi proses analisis.

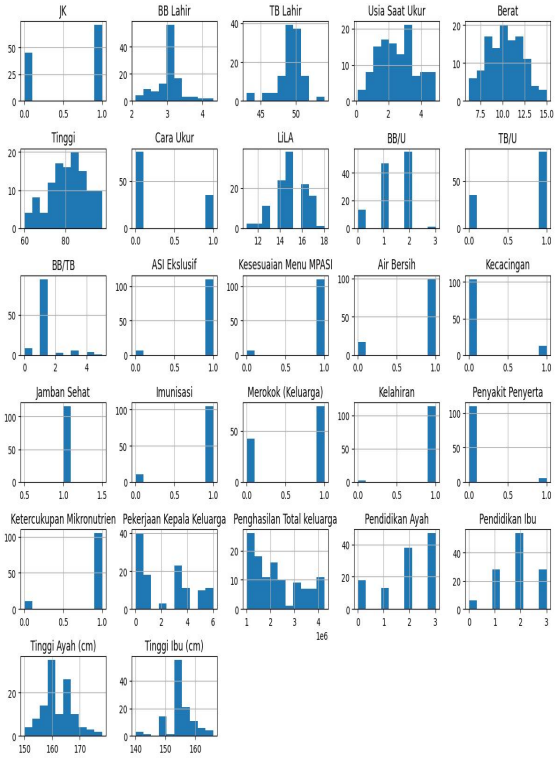
2. Deteksi Outlier : Menemukan dan menangani outlier yang bisa memengaruhi hasil klasterisasi, terutama karena algoritma K-Means sangat peka terhadap outlier. Deteksi outlier menggunakan

isolation forest yang merupakan algoritma efektif dalam mendeteksi outlier pada data multivariat dan dapat memberikan hasil yang lebih baik ketika diterapkan pada data yang lebih kompleks serta termasuk unsupervised learning dan nonparametrik yang berbasiskan pohon keputusan (*decision trees*) [17]. Hasil menunjukan terdapat 7 baris outlier yaitu baris ke 0, 47, 50, 55, 70, 72, dan 119. Baris-baris ini dianggap memiliki karakteristik yang sangat berbeda (dalam konteks multivariat) dibandingkan dengan sebagian besar data lainnya dalam dataset. Hal ini berarti, titik data tersebut "terpisah" atau "terisolasi" dari mayoritas data lainnya dalam ruang fitur multivariat.

Hal yang dilakukan setelah menemukan outlier yaitu menghapus baris yang terindikasi outliers, sehingga total baris dataset saat ini adalah 116 data dengan 27 variabel.

2.3 Pemeriksaan Distribusi Data

Distribusi data pada setiap fitur dilakukan untuk memastikan bahwa data tidak mengalami distorsi setelah penghapusan outlier. Penelitian ini menggunakan visualisasi untuk memeriksa distribusi data secara lebih jelas seperti gambar berikut:



Gambar 4. Visualisasi distribusi data

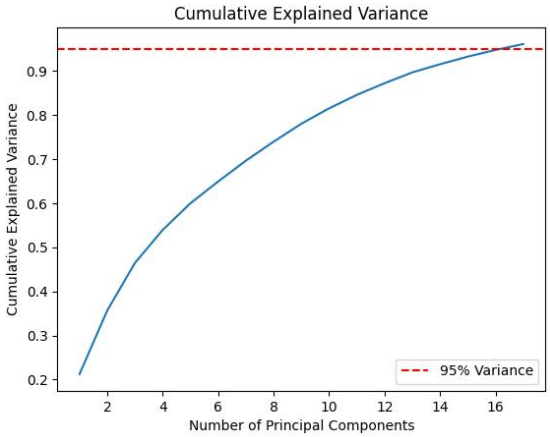
3. HASIL DAN PEMBAHASAN

Setelah melakukan *preprocessing data*, tahapan berikutnya yaitu menganalisis *Principal Component Analysis* (PCA) dan K-means *clustering*.

3.1 Principal Component Analysis (PCA)

Menggunakan PCA untuk mereduksi dimensi data adalah teknik yang bertujuan untuk mengurangi jumlah fitur sambil mempertahankan sebanyak mungkin informasi yang terkandung dalam variansi data. Metode ini sangat bermanfaat, terutama ketika data memiliki banyak variabel atau fitur yang saling berkorelasi. Dengan demikian, PCA dapat meningkatkan efisiensi pengolahan data sekaligus membantu meningkatkan akurasi model. Seringkali informasi tersebut dikelompokkan menjadi variabel baru yang lebih sederhana berdasarkan kemiripan informasi yang diperoleh dari data awal [18]. Berikut ini hasil dari perhitungan variansi kumulatif oleh PCA menunjukkan seberapa banyak informasi (variansi) yang dipertahankan seiring dengan penambahan komponen utama satu per satu.

Jumlah komponen utama untuk menjelaskan 95% variansi sebanyak 17. Rasio variansi yang dijelaskan oleh setiap komponen yaitu [0.21216677, 0.14425921 0.10809877, 0.07468302, 0.06049184, 0.04942359, 0.04725046, 0.04302692, 0.04072877, 0.03490417 0.03047046, 0.02651119, 0.02449155, 0.01853654, 0.01724846, 0.01518458, 0.01333107]



Gambar 5. Variansi Kumulatif

Hasil analisis PCA menunjukkan rasio variansi yang dijelaskan oleh setiap komponen utama, yang mencerminkan sejauh mana setiap komponen membawa informasi (variansi) dalam dataset. Komponen pertama (PC1) menjelaskan 21,22% dari total variansi, menjadikannya yang paling berpengaruh. Komponen kedua (PC2) menjelaskan 14,43%, dan komponen ketiga (PC3) menjelaskan 10,81% dari variansi. Komponen keempat hingga ketujuh masing-masing menyumbang 7,47%, 6,05%, dan 4,94%. Komponen-komponen setelah PC7 memberikan kontribusi yang semakin kecil, dengan masing-masing menjelaskan hampir 1% dari variansi. Secara keseluruhan, tujuh komponen pertama telah menjelaskan sekitar 69,65% dari variansi data. Komponen utama merupakan representasi yang lebih sederhana dari data yang masih mempertahankan informasi penting di dalamnya. Dengan demikian, komponen-komponen

dalam 69,65% PCA adalah hasil kombinasi linier dari variabel-variabel yang ada dan bukan merupakan variabel asli dari dataset.

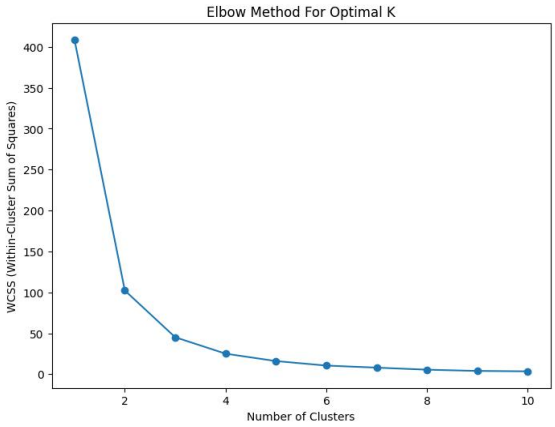
3.2 KMEANS Clustering

Berikut ini merupakan tahapan Kmeans clutering yang dilakukan setelah mendapatkan nilai dari *principal component analysis* (PCA):

1. Menentukan jumlah cluster yang optimal.

Menentukan jumlah cluster yang optimal (K) merupakan salah satu langkah krusial dalam algoritma K-means. Untuk menemukan nilai K yang paling sesuai, berbagai metode dapat digunakan, salah satunya adalah Elbow Method, yang menghubungkan jumlah cluster (K) dengan nilai within-cluster sum of squares (WCSS). Pada metode ini, titik "elbow" dicari di mana penurunan WCSS mulai melambat. Selain itu, Silhouette Score juga dapat diterapkan untuk mengevaluasi seberapa baik setiap titik data terkelompok dalam cluster yang terbentuk, dengan nilai silhouette yang lebih tinggi menunjukkan pemisahan cluster yang lebih jelas dan lebih baik [19]. Karena pada penelitian ini proses PCA sudah dilakukan, maka penggunaan kombinasi Elbow Method dan Silhouette Score dapat dilakukan untuk menentukan jumlah cluster terbaik. Elbow Method terlebih dahulu untuk menemukan rentang jumlah cluster, dan selanjutnya adalah silhouette score untuk mengevaluasi klasterisasi yang dihasilkan oleh K-Means pada berbagai nilai K, dan pilih K yang memiliki skor Silhouette terbaik. Penggabungan kedua metode ini akan memberikan hasil yang lebih efektif dalam menentukan jumlah cluster yang optimal.

Berikut ini merupakan grafik terbentuknya cluster berdasarkan elbow method:

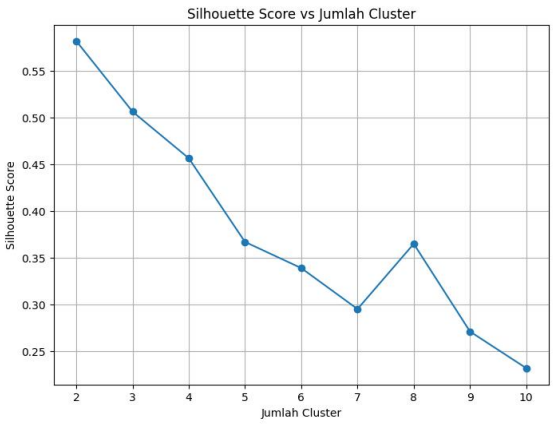


Gambar 6. Elbow method untuk pencarian nilai cluster optimal

Jika grafik WCSS (*Within-Cluster Sum of Squares*) menunjukkan penurunan yang signifikan hingga mencapai nilai maksimum 400 dengan jumlah cluster (K) mencapai 10, maka hal ini dapat diartikan bahwa WCSS akan cenderung menurun seiring dengan bertambahnya jumlah cluster, karena pembagian data yang lebih banyak akan mengurangi variasi dalam setiap cluster. Penurunan tajam di awal grafik mengindikasikan bahwa menambah jumlah cluster memberikan dampak besar dalam mengurangi keragaman antar cluster, yang umumnya terjadi ketika jumlah cluster yang sedikit tidak mampu mewakili distribusi data dengan efektif. Titik elbow, di mana penurunan WCSS mulai melambat, memberikan indikasi jumlah cluster yang optimal. Pada grafik di atas didapatkan bahwa penurunan WCSS melambat pada K=3 atau K=4, maka jumlah cluster yang optimal kemungkinan ada pada titik tersebut, karena menambah lebih banyak cluster setelah titik itu tidak memberikan pengurangan WCSS yang berarti. Nilai tersebut dianggap sebagai jumlah cluster yang optimal, sementara penurunan yang terus berlangsung hingga K=10 mungkin memerlukan evaluasi lebih lanjut atau penggunaan metrik lain, seperti Silhouette Score, untuk memastikan bahwa jumlah cluster tersebut menghasilkan pemisahan data yang lebih baik. Berikut adalah hasil analisis cluster menggunakan Silhouette Score:

Tabel 1. Silhouette Score

Jumlah Cluster	Silhouette Score
2	0.5815668500896095
3	0.5060464848367403
4	0.45615390235609293
5	0.3665934819881008
6	0.3386226146353719
7	0.2947892034423691
8	0.36459465158424065
9	0.2707394494275095
10	0.23152480222423052



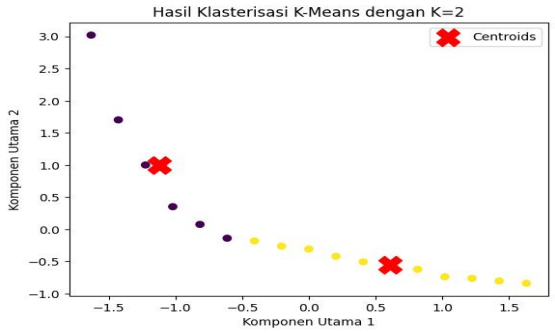
Gambar 7. Silhouette score untuk pencarian nilai cluster optimal

Berdasarkan hasil yang diperoleh dari Silhouette Score, jumlah cluster yang optimal untuk dataset ini adalah 2. Silhouette Score digunakan untuk menilai kualitas pemisahan antar cluster, dengan nilai berkisar antara -1 hingga +1. Nilai yang mendekati +1 menunjukkan pemisahan antar cluster yang baik,

nilai mendekati 0 menunjukkan data yang berada di perbatasan antara dua cluster, dan nilai mendekati -1 mengindikasikan kemungkinan kesalahan dalam pengelompokan. Hasil analisis menunjukkan bahwa Cluster 2 memiliki Silhouette Score tertinggi (0.5815), yang mencerminkan pemisahan antar cluster yang baik. Namun, setelah jumlah cluster meningkat dari 3 hingga 10, nilai Silhouette Score secara bertahap menurun, dengan Cluster 10 mencapai nilai terendah (0.2315), yang menunjukkan pemisahan cluster yang kurang baik dan kemungkinan terjadinya overfitting. Dengan demikian, Cluster 2 adalah pilihan yang lebih baik, dan disarankan untuk memilih K=2 sebagai jumlah cluster optimal dalam penelitian ini.

2. Menerapkan Kmeans dengan nilai K=2

Berikut ini adalah grafik kmeans clustering dengan jumlah cluster=2 dengan cluster labels untuk setiap data point adalah : [0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1].



Gambar 8. Hasil klasterisasi dengan k=2

Hasil dari algoritma K-Means ini menunjukkan pengelompokan data ke dalam dua cluster berdasarkan kedekatannya, dengan setiap angka (0 atau 1) merepresentasikan cluster tempat data tersebut berada. Sebagai contoh, jika label untuk data pertama adalah 0, maka data tersebut termasuk dalam cluster 0. Berdasarkan distribusi data yang diperoleh, sebagian besar data dikelompokkan ke dalam cluster dengan label 1, sementara sisanya berada dalam cluster dengan label 0. Hal ini dapat mengindikasikan bahwa cluster dengan label 1 lebih besar atau lebih padat dibandingkan cluster dengan label 0. Visualisasi di atas hanya memungkinkan untuk dua dimensi, sehingga dua komponen yang dipilih untuk divisualisasikan adalah dua komponen utama pertama dari hasil PCA yang sudah dilakukan.

3. Melakukan analisis cluster lebih lanjut

Untuk memahami karakteristik masing-masing cluster, dapat dilakukan analisis deskriptif pada fitur-fitur utama di setiap cluster guna mengidentifikasi apakah terdapat pola atau perbedaan yang membedakan satu cluster dengan cluster lainnya sepeti gambar berikut:

	Principal Component							\
	count	mean	std	min	25%	50%	75%	max
Cluster								
0	6.0	3.5	1.870829	1.0	2.25	3.5	4.75	6.0
1	11.0	12.0	3.316625	7.0	9.50	12.0	14.50	17.0

	Explained Variance Ratio		...		\
	count	mean	...	75%	max
Cluster			...		
0	6.0	0.108187	...	0.135219	0.212167
1	11.0	0.028335	...	0.037816	0.047250

	Cumulative Variance					
	count	mean	std	min	25%	50%
Cluster						
0	6.0	0.470191	0.163159	0.212167	0.383451	0.501866
1	11.0	0.854598	0.087838	0.696374	0.797581	0.872015

	75%	max
Cluster		
0	0.584577	0.649123
1	0.923667	0.960807

Gambar 8. Hasil analisis cluster

4. KESIMPULAN

Penerapan Principal Component Analysis (PCA) berhasil mereduksi dimensi data dengan mempertahankan sekitar 69,65% dari variansi total menggunakan 17 komponen utama, sehingga membuat analisis menjadi lebih efisien tanpa mengorbankan informasi penting. Dalam pemilihan jumlah cluster optimal, algoritma K-Means yang didukung oleh analisis WCSS menunjukkan bahwa K=2 adalah jumlah cluster yang paling tepat, berdasarkan grafik elbow yang memperlihatkan penurunan tajam pada awalnya dan pelambatan penurunan setelah K=2. Klasterisasi membagi data menjadi dua cluster dengan karakteristik yang berbeda, di mana Cluster 0 cenderung memiliki nilai yang lebih rendah dibandingkan dengan Cluster 1. Analisis deskriptif pada fitur utama setiap cluster mengungkapkan pola yang membedakan keduanya, memberikan pemahaman yang lebih dalam tentang distribusi data. Secara keseluruhan, penerapan PCA dan K-Means clustering berhasil menyederhanakan data dan mengelompokkan informasi, memberikan wawasan yang lebih jelas mengenai pola dalam dataset yang kompleks, serta membuka peluang untuk analisis dan aplikasi lebih lanjut. Penerapan PCA dan K-Means clustering pada dataset stunting berfungsi untuk mereduksi dimensi data yang kompleks dan mengelompokkan individu berdasarkan karakteristik serupa. Hasil klasterisasi mengungkapkan kelompok-kelompok dengan ciri khas yang berbeda, yang mencerminkan perbedaan dalam faktor-faktor yang mempengaruhi stunting, seperti status gizi, akses terhadap layanan kesehatan, dan faktor sosial ekonomi. Analisis ini memberikan wawasan lebih dalam tentang distribusi stunting dan dapat membantu merancang intervensi yang lebih tepat sasaran, seperti program gizi yang terarah atau edukasi kesehatan yang disesuaikan dengan kondisi setiap kelompok.

5. REFERENCES

[1] Kemenkes, "Membentengi Anak Dari Stunting," *Edisi 167,* pp. 1-56, Juni 2024.

[2] Kemenkes, "Cegah Stunting dengan ABCDE," 16 Februari 2023.

[3] M. E. Rahmuniyati, "Peran Puskesmas Dalam Upaya Mengurangi Kasus Stunting Melalui Program Sanitasi Total Berbasis Masyarakat (Stbm)," In *Seminar Nasional Unriyo* , Yogyakarta, 2020.

[4] M. Wangge, "Penerapan Metode Principal Component Analysis (PCA) Terhadap Faktor-faktor yang Mempengaruhi Lamanya Penyelesaian Skripsi Mahasiswa Program Studi Pendidikan Matematika FKIP UNDANA," *Jurnal Cendekia: Jurnal Pendidikan Matematika,* vol. 5, no. 2, pp. 974-988, 2021.

[5] S. Manullang, D. Aryani and H. Rusydah, "Analisis Principal Component Analysis (PCA) dalam Penentuan Faktor Kepuasan Pengunjung terhadap Layanan Perpustakaan Digilib," *Edumatic: Jurnal Pendidikan Informatika,* vol. 7, no. 1, pp. 123-130, 2023.

[6] S. Dewi And M. A. Ineke Pakereng, "Implementasi Principal Component Analysis Pada K-Means Untuk Klasterisasi Tingkat Pendidikan Penduduk Kabupaten Semarang," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika),* vol. 8, no. 4, pp. 1186-1195, 2023.

[7] Rais. M, Goejantoro, and Prangga Surya, "Optimalisasi K-MeansCluster dengan Principal Com ponent Analysis pada Pengelompokan Kabupaten/Kota di Pulau Kalimantan Berdasark an Indik ator Tingk at Pengangguran Terbuka," *Jurnal Eksponensial*, vol. 12, no. 2, pp. 129-136. 2085-7829, 2021.

[8] Rosyada and Dina Tri Utari, " Penerapan Principal Component Analysis untuk Reduksi Variabel pada Algoritma K-Means Clustering," *Jambura Journal of Probability and Statistics*, vol. 5, no. 1, pp. 6-13. 2722-7189, 2024.

[9] Hayqal Hazmi Qastari, Oni Soesanto, and Yuana Sukmawaty, "K-Means Clustering dan Principal Component Analysis (PCA) Dalam Radial Basis Function Neural Network (RBFNN) Untuk Klasifikasi Data Multivariat," *Journal of Mathematics: Theory and Applications,* vol. 4, no.1, pp. 1-7, 2685-9653, 2022.

[10] M. Billah, M. Aidil Zartesya, and Desta Sandya Prasvita, "Penerapan Collaborative Filtering, PCA dan K-Means dalam Pembangunan Sistem Rekomendasi Film," *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, pp. 579-587, 20 April 2021.

[11] B. A. Nugroho, A. Izzah, K. Eliyen and R. Widyastuti, "Metode Hybrid-DPSO dan Clustering untuk Rekomendasi Rute Perjalanan Wisata Berbasis Mobile Android," *JEPIN (Jurnal Edukasi dan Penelitian Informatika),* vol. 10, no. 2, pp. 201-2018, 2024.

[12] H. Priyatman, F. Sajid and D. Haldivany, "Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa," *JEPIN (Jurnal Edukasi dan Penelitian Informatika),* vol. 5, no. 1, pp. 62-66, 2019.

[13] R. K. Dinata, Safwandi, N. Hasdyna and N. Azizah, "Analisis K-

Means Clustering pada Data Sepeda Motor," *Informatics Journal,* vol. 5, no. 1, pp. 10-17, 2020.

[14] Nurdiansyah, Muliadi, R. Herteno, D. Kartini And I. Budiman, "Implementasi Metode Principal Component Analysis (Pca) Dan Modified K-Nearest Neighbor Pada Klasifikasi Citra Daun Tanaman Herbal," *Jurnal MNEMONIC,* vol. 7, no. 1, pp. 1-9, 2024.

[15] D. Hediyati and . I. M. Suartana, "Penerapan Principal Component Analysis(PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi Pertanian Di Kabupaten Bojonegoro," *JIEET (Journal Information Engineering and Educational Technology),* vol. 5, no. 2, pp. 49-54, 2021.

[16] A. A. Aryasatya Daniswara and I. K. Dwi Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *JINACS (Journal of Informatics and Computer Science),* vol. 5, no. 1, pp. 97-100, 2023.

[17] A. Zulfikar, F. A. Rahmani and N. Azizah, "Deteksi Anomali Menggunakan Isolation Forest Belanja Barang Persediaan Konsumsi Pada Satuan Kerja Kepolisian Republik Indonesia," *Jurnal Manajemen Perbendaharaan,* vol. 4, no. 1, pp. 1-15, 2023.

[18] D. R. Puspita Sari, "Metode Principal Component Analysis(Pca) Sebagai

Penanganan Asumsi Multikolinearitas (Studi Kasus: Data Produksi Tapioka)," *Parameter (Jurnal Matematika, Statistika Dan Terapannya),* vol. 2, no. 2, pp. 115-124, 2023.

[19] F. Dikarya and S. Muharni, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokanuniversitas Terbaik Di Dunia," *Jurnal Informatika,* vol. 22, no. 2, pp. 124-131, 2022.