RESEARCH ARTICLE • OPEN ACCESS

Penerapan Principal Component Analysis untuk Reduksi Variabel pada Algoritma K-Means Clustering

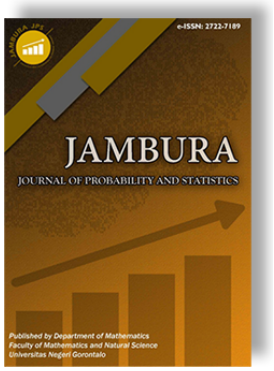
Istina Alya Rosyada, Dina Tri Utari

Volume 5, Issue 1, Pages 6–13, May 2024

Received 30 January 2023, Revised 11 November 2023, Accepted 08 May 2024, Published Online 4 June 2024

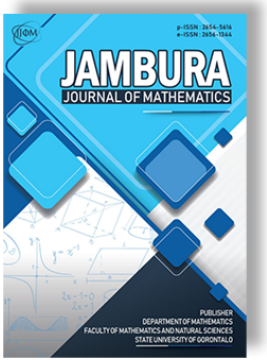
To Cite this Article : I. A. Rosyada, and D. T Utari,“ Penerapan Principal Component Analysis untuk Reduksi Variabel pada Algoritma K-Means Clustering ”, *Jambura J. Probab. Stat.*, vol. 5, no. 1, pp. 6–13, 2024, *https://doi.org/10.34312/jjps.v5i1.18733*

JOURNAL INFO • JAMBURA JOURNAL OF PROBABILITY AND STATISTICS



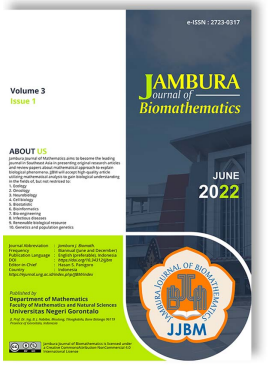
Homepage	:	https://ejurnal.ung.ac.id/index.php/jps/index
Journal Abbreviation	:	Jambura J. Probab. Stat.
Frequency	:	Biannual (May and November)
Publication Language	:	English (preferable), Indonesia
DOI	:	https://doi.org/10.37905/jjps
Online ISSN	:	2722-7189
Editor-in-Chief	:	Ismail Djakaria
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/jps/oai
Google Scholar ID	:	kWdujzMAAAAJ
Email	:	redaksi.jjps@ung.ac.id

JAMBURA JOURNAL • FIND OUR OTHER JOURNALS

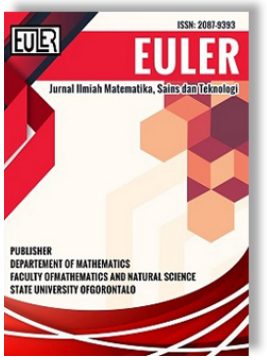


Jambura Journal of Mathematics

Jambura Journal of Mathematics Education



Jambura Journal of Biomathematics



EULER : Jurnal Ilmiah Matematika, Sains, dan Teknologi

Jambura Journal of Probability and Statistics, Volume 5, Issue 1, Pages 6–13, May 2024 https://doi.org/10.34312/jjps.v5i1.18733

Research Article

Check for updates

Penerapan Principal Component Analysis untuk Reduksi Variabel pada Algoritma K-Means Clustering

Istina Alya Rosyada[1,*], **Dina Tri Utari**[1],

[1]*Jurusan Statistika, Fakultas MIPA, Universitas Islam Indonesia*

ARTICLE HISTORY
Received 30 January 2023
Revised 11 November 2023
Accepted 08 May 2024
Published 4 June 2024

KATA KUNCI
clustering
K-Means
Principal Component Analysis
kesejahteraan masyarakat.

KEYWORDS
clustering
K-Means
Principal Component Analysis
public welfare.

ABSTRAK. *K-Means Clustering merupakan algoritma clustering yang banyak digunakan, namun memiliki kelemahan yaitu performa data clustering menurun jika variabel dari data yang diolah sangat banyak. Masalah variabel yang kompleks pada K-Means dapat diatasi dengan mengkombinasikan metode pengurangan variabel Principal Component Analysis (PCA). Penelitian ini menggunakan tujuh variabel indikator kesejahteraan masyarakat Provinsi Jawa Barat tahun 2021 yang bertujuan untuk mengukur tingkat kesejahteraan kabupaten/kota. Hasil analisis tersebut memperoleh dua komponen utama berdasarkan eigenvalues. Pengelompokkan dari analisis cluster dengan metode K-Means dengan PCA terbentuk tiga cluster terbaik dimana jumlah anggota masing-masing cluster terdiri 12, 8, dan 7 kabupaten/kota.*

ABSTRACT. *K-Means clustering is a widely used clustering algorithm. However, it has the disadvantage that the performance of clustering data decreases if the variables of the processed data are immense. The complex variables problem in K-Means can be overcome by combining the Principal Component Analysis (PCA) variable reduction method. This study uses seven indicator variables for the welfare of the people of West Java Province in 2021 to measure the welfare level of districts/cities. The results of the analysis obtained two principal components based on eigenvalues. Clustering from cluster analysis with the K-Means with variable reduction using PCA formed the three best clusters where the number of members of each cluster consisted of 12, 8, and 7 districts/cities.*

1. PENDAHULUAN

Algoritma *K-Means clustering* adalah teknik pembelajaran tanpa pengawasan yang banyak digunakan untuk mempartisi data ke dalam kelompok yang berbeda berdasarkan kesamaannya. Namun, saat menangani data dengan variabel yang sangat kompleks, permasalahan ini dapat menimbulkan tantangan terhadap efektivitas dan efisiensi algoritma *K-Means*. Data seperti ini sering mengalami masalah karena adanya fitur yang tidak relevan atau redundan, yang dapat menyebabkan penurunan kinerja pengelompokan dan peningkatan kompleksitas komputasi [1].

Reduksi variabel menggunakan *Principal Component Analysis* (PCA) merupakan langkah *preprocessing* yang penting dalam algoritma *K-Means clustering* [2]. Dalam konteks reduksi variabel, PCA membantu mengidentifikasi variabel yang paling berkontribusi terhadap varians total dalam data. Dengan mengidentifikasi komponen-komponen utama yang paling signifikan, dengan menggunakan PCA memungkinkan untuk mengeliminasi variabel-variabel yang redundan dalam analisis [3]. Dengan menerapkan PCA, variabel dapat dikurangi dan memungkinkan algoritma *K-Means* untuk fokus pada fitur yang paling informatif [2]. Hal ini mengarah pada peningkatan kinerja *clustering* karena fitur yang tidak relevan dan redundan dihilangkan, sehingga memungkinkan algoritma menangkap struktur dasar data secara lebih efektif.

Selain meningkatkan kinerja *clustering*, reduksi variabel menggunakan PCA juga meningkatkan efisiensi komputasi dengan mengurangi kompleksitas komputasi algoritma *K-Means* [4]. Hal ini membuat PCA cocok untuk menangani kumpulan data berskala besar. Selain itu, PCA memberikan interpretabilitas dan visualisasi dengan memproyeksikan data ke ruang berdimensi lebih rendah, memfasilitasi analisis dan pemahaman yang lebih mudah tentang *cluster* dan hubungannya [5].

Lebih lanjut, PCA membantu meningkatkan ketahanan algoritma *K-Means* terhadap *noise* dan redundansi dengan menghapus fitur yang tidak relevan dan meningkatkan rasio *signal-to-noise* [6][7]. Secara keseluruhan, penerapan PCA untuk pengurangan dimensi dalam algoritma *K-Means clustering* sangat penting untuk meningkatkan kinerja *clustering*, efisiensi komputasi, interpretabilitas, dan ketahanan terhadap *noise* dan redundansi [8][9][10][11][12][13].

PCA juga dapat mengatasi multikolinearitas karena mengubah variabel asli yang saling berkorelasi satu dengan yang lainnya menjadi satu set variabel baru yang lebih kecil dan saling bebas. Berdasarkan penelitian [14] pemakaian PCA pada *K-Means* untuk mengurangi variabel besar yang berkisar antara enam sampai dengan sepuluh variabel.

*Corresponding Author.

Email : *dina.t.utari@uii.ac.id* (D. T. Utari)
Homepage : https://ejurnal.ung.ac.id/index.php/jps/index / E-ISSN : 2722-7189

Berdasarkan latar belakang tersebut, penelitian ini akan membahas mengenai implementasi PCA untuk mereduksi variabel pada *dataset* indikator kesejahteraan masyarakat Provinsi Jawa Barat tahun 2021. Kesejahteraan masyarakat sendiri merupakan konsep yang kompleks dan multidimensi yang mencakup berbagai faktor seperti pendapatan, pendidikan, kesehatan, dan standar hidup. Penggunaan *K-means*, yang merupakan algoritma pengelompokan yang populer, dapat membantu mengidentifikasi berbagai kelompok atau kelompok dalam suatu populasi berdasarkan indikator kesejahteraan ini. Hal ini dapat memberikan wawasan berharga bagi para pembuat kebijakan untuk memahami distribusi kesejahteraan di antara berbagai kelompok dalam masyarakat.

2. METODE PENELITIAN

Pada bagian ini akan dijelaskan tentang data yang digunakan, metode dan tahapan analisis yang dilakukan.

2.1. Data dan Variabel Penelitian

Data yang digunakan dalam penelitian ini merupakan data indikator kesejahteraan masyarakat Provinsi Jawa Barat tahun 2021 yang bersumber dari *website* BPS Provinsi Jawa Barat [15]. Adapun variabel yang digunakan terdiri dari tujuh variabel yang dijelaskan pada tabel 1 berikut.

Tabel 1. Variabel Penelitian

Simbol	Variabel	Satuan
X_1	Rata-rata lama sekolah (RLS)	Tahun
X_2	Daya beli (didekati dengan pengeluaran per kapita yang disesuaikan)	Ribu Rupiah/orang/Tahun
X_3	Tingkat Partisipasi Angkatan Kerja (TPAK)	Persen (10%)
X_4	Pesentase penduduk miskin	Persen (10%)
X_5	Persentase kepemilikan rumah sendiri	Persen (10%)
X_6	Persentase penduduk yang memiliki jaminan kesehatan	Persen (10%)
X_7	Tingkat Pengangguran Terbuka (TPT)	Persen (10%)

2.2. Tahapan Penelitian

Tahapan penelitian yang dilakukan pada penelitian ini dijelaskan melalui diagram alir pada gambar 1.

1. Pra-proses data yaitu dilakukan pengumpulan data mencangkup objek yang relevan dari sumber data yang mendasarinya. Kemudian dilakukan pembersihan data dan menginput data.
2. Uji asumsi *cluster* untuk memastikan sampel sudah representative menggunakan *Kaiser-Mayer-Olkin* (KMO) dan *Measure of Sampling* (MSA), kemudian uji multikolinearitas dilakukan untuk mengetahui apakah variabel yang digunakan dalam penelitian saling berkorelasi atau tidak.
3. PCA digunakan untuk mereduksi variabel dan mengatasi jika data terjadi gejala multikolinearitas.
4. *K-Means clustering* digunakan untuk membagi data ke dalam beberapa kelompok.
5. Langkah terakhir setelah dihasilakan *cluster* yaitu dilakukan profilisasi pada masing-masing *cluster* untuk mengetahui kateristik dari masing-masing *cluster* dengan cara mencari rata-rata masing-masing variabel pada tiap *cluster*.

2.3. Principal Componenet Analysis (PCA)

Principal Component Analysis atau analisis komponen utama merupakan suatu teknik statistik untuk mengubah dari sebagian besar variabel asli yang digunakan yang saling berkorelasi satu dengan yang lainnya menjadi satu set variabel baru yang lebih kecil dan saling bebas (tidak berkorelasi lagi). Diperkenalkan oleh Pearson (1901) dan Hotelling (1933) yang menggambarkan variasi dari data *multivariate* untuk suatu himpunan variabel-variabel tak berkorelasi [16]. Analisis komponen utama berguna untuk mereduksi data, sehingga lebih mudah untuk menginterpretasikan data-data tersebut [17].

Adapun langkah-langkah yang dilakukan untuk menentukan komponen utama adalah sebagai berikut:

1. Melakukan standarisasi data agar memiliki skala nilai yang setara menggunakan *Z-score*.
2. Hitung matriks kovarians untuk menentukan korelasi setiap variabel.
3. Menghitung *eigenvalue* menggunakan persamaan

$$|\lambda I - R| = 0 \tag{1}$$

dan *eigenvector* dengan rumus

$$(R\vec{v} = \lambda\vec{v}) \tag{2}$$

4. Menentukan jumlah komponen utama yang mungkin terbentuk dengan mempertimbangkan *eigenvalue* yang lebih besar atau sama dengan 1.
5. Menghitung transformasi data set baru hasil reduksi dengan PCA dengan persamaan:

$$PC_{at} = \overrightarrow{v_{1a}}Z_1 + \overrightarrow{v_{2a}}Z_2 + \cdots + \overrightarrow{v_{ap}}Z_p \tag{3}$$

6. Membentuk komponen matriks korelasi yang menunjukkan besarnya korelasi variabel terhadap skor komponen yang terbentuk dengan persamaan:

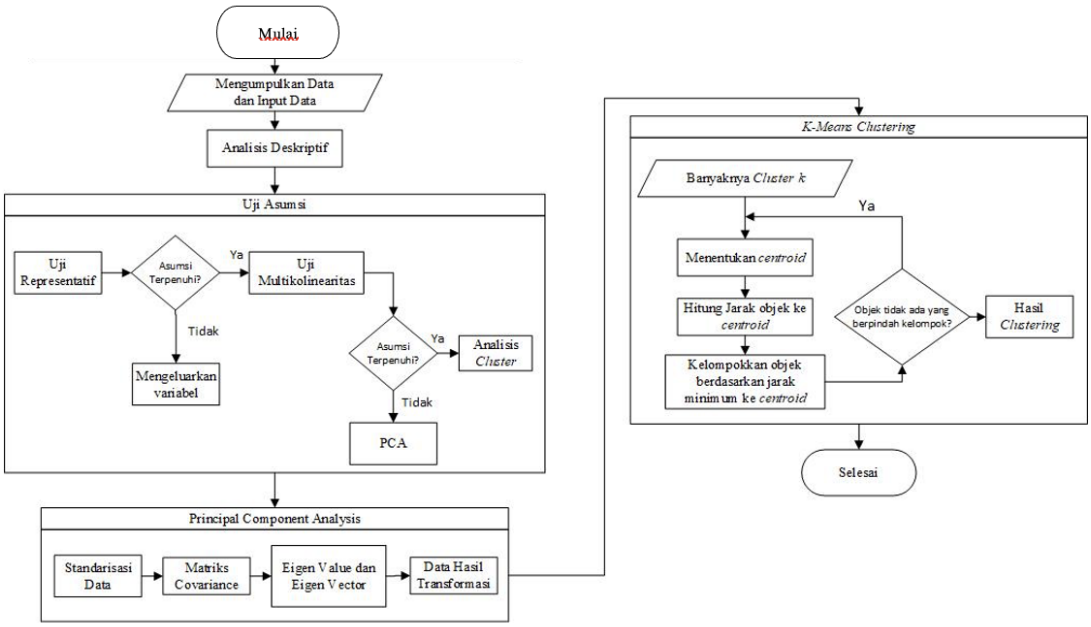
$$rx_p, PC_t = \overrightarrow{v_{1a}}\sqrt{\lambda_t} \tag{4}$$

2.4. Asumsi Cluster

Sebelum melakukan pengelompokkan, diperlukan asumsi-asumsi yang harus dipenuhi terlebih dahulu agar *cluster* yang dihasilkan representatif. Adapun asumsi yang ada pada analisis *cluster* sebagai berikut [18][4]:

1. *Kaiser-Mayer-Olkin* (KMO) KMO merupakan suatu indeks yang digunakan untuk mengukur kecukupan sampling secara menyeluruh dan mengukur kecukupan *sampling*. Ketepatan dalam analisis apabila nilai KMO diantara 0,5 hingga 1,0 dan sebaliknya jika nilai KMO kurang dari 0,5 maka dapat dikatakan bahwa analisis komponen utama tidak tepat digunakan.

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{ij}^{2} + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} \tag{5}$$



Gambar 1. Diagram Alir Penelitian

2. *Measure of Sampling* (MSA) Penilaian kelayakan setiap variabel untuk analisis komponen utama digunakan kriteria MSA. Variabel yang mempunyai ukuran kecukupan sampling kecil (< 0,5) sebaiknya dikeluarkan dari analisis.

$$MSA_i = \frac{\sum_{i=1}^{p} r_{ij}^2}{\sum_{i=1}^{p} r_{ij}^2 + \sum_{j=1}^{p} a_{ij}^2} \tag{6}$$

3. Uji Korelasi Antar Variabel (Multikolinearitas) Uji multikolinearitas juga diperlukan untuk mengetahui adanya hubungan linier antara 2 variabel atau lebih. Gejala multikolinearitas dapat dideteksi dengan beberapa cara seperti berikut:
   a. Pengujian multikolinearitas dapat dilakukan dengan menghitung koefisien korelasinya. Jika nilai koefisien korelasi antara dua variabel independen melibihi 0,8 maka diduga model mengandung unsur multikolinearitas.
   b. Menghitung nilai *tolerance* atau *Variance Inflation Factors* (VIF), yang didefinisikan dengan persamaan berikut [19]. Dua buah variabel dikatakan mengandung multikolinieritas ketika nilai VIF lebih dari 10 [19].

$$VIF_j = \frac{1}{1 - R_j^2} \tag{7}$$

2.5. Algoritma K-Means

K-Means clustering adalah teknik pengelompokan yang berupaya untuk mempartisi N individu dalam sebuah dataset multivariat kedalam kelompok (k kelompok). Pendekatan paling umum dalam implementasi *K-Means* adalah mempartisi dari N individu ke dalam kelompok k yang meminimalkan *within-group sum of square* seluruh variabel [20].

Tujuan dari analisis *cluster* adalah mengelompokkan data observasi ke dalam kelompok hingga anggota kelompok di dalamnya bersifat homogen, sedangkan antar kelompok bersifat heterogen. Proses dalam *K-Means clustering* adalah sebagai berikut:

1. Menentukan besarnya k, k = jumlah *cluster*.
2. Inisiasi secara random nilai *centroid* (pusat *cluster*) sejumlah k.
3. Menghitung jarak dari setiap data terhadap masing-masing *centroid cluster* dengan menggunakan jarak *euclidean*. Jarak *euclidean* antara objek i dan j dirumuskan dengan:

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{8}$$

4. Mengelompokkan data berdasarkan jarak terdekat antara setiap data terhadap *centroid*.
5. Menetukan *centroid* baru dengan menghitung rata-rata seluruh data pada pusat *cluster* yang sama.

$$C_{kj} = \frac{x_1 + x_{2j} + \cdots + x_{aj}}{a}, j = 1, 2, 3, \cdots, p \tag{9}$$

6. Ulangi langkah 3 dan terus melakukan iterasi hingga *centroid cluster* tetap dan anggota *cluster* tidak berpindah (konvergen).

2.6. Penentuan Jumlah Cluster

Penentuan jumlah *cluster* yang optimal bersifat subyektif dan tergantung pada metode yang digunakan untuk mengukur kesamaan dan parameter yang digunakan untuk mempartisi. Ada beberapa metode yang dapat digunakan untuk menentukan jumlah *cluster* optimal pada proses *clustering*. Dalam penelitian ini, metode yang akan digunakan adalah *silhouette, elbow, dan gap statistics* [21].

a). Metode *Silhoutte*
Silhouette coefficient digunakan untuk melihat kualitas dan kekuatan *cluster*, seberapa seberapa baik suatu objek ditem-

patkan dalam suatu *cluster*. Metode ini merupakan gabungan dari dua metode yaitu metode *cohesion* yang berfungsi untuk mengukur seberapa dekat relasi antara objek dalam sebuah *cluster*, dan metode *separation* yang berfungsi untuk mengukur seberapa jauh sebuah *cluster* terpisah dengan *cluster* lain [22]. Tahapan perhitungan *Silhouette coefficient* adalah sebagai berikut:

1. Hitung rata-rata jarak objek dengan semua objek lain yang berada di dalam satu *cluster*.

$$a(i) = \frac{1}{[A]-1} \sum J \in_{A, j \neq i} d(i,j) \tag{10}$$

2. Hitung rata-rata jarak objek dengan semua objek lain yang berada pada *cluster* lain, kemudian ambil nilai paling minimum.

$$a(i,C) = \frac{1}{[A]-1} \sum j \in C \quad d(i,j) \tag{11}$$

3. Hitung nilai *silhouette coefficient*.

$$b(i) = min \quad C \neq A \quad d(i,j) \tag{12}$$

Nilai hasil *silhouette coefficient* terletak pada kisaran nilai -1 hingga 1. Semakin *silhouette coefficient* mendekati nilai 1, maka semakin baik pengelompokan data dalam satu cluster. Sebaliknya jika *silhouette coefficient* mendekati nilai -1, maka semakin buruk pengelompokan data di dalam satu cluster.

b). Metode *Elbow*

Menurut [23], metode *elbow* adalah metode untuk menentukan jumlah *cluster* yang tepat melalui persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik. Jika nilai *cluster* pertama dengan nilai *cluster* kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar maka jumlah nilai *cluster* tersebut yang tepat.

Pada metode ini nilai *cluster* terbaik akan diambil dari nilai *Sum of Square Error* (SSE) yang mengalami penurunan yang signifikan dan berbentuk siku, untuk menghitung SSE digunakan rumus:

$$SEE = \sum_{K=1}^{K} \sum_{x_i} |x_i - C_k|^2 \tag{13}$$

c). Metode *Gap Statistics*

Metode *Gap Statistics* membandingkan total dalam variasi *intra-cluster* untuk nilai k yang berbeda dengan nilai yang diharapkan di bawah distribusi referensi nol data. Perkiraan *cluster* optimal akan menjadi nilai yang memaksimalkan statistik *gap* (yaitu yang menghasilkan statistik *gap* terbesar). Ini berarti struktur pengelompokan jauh dari distribusi titik acak yang seragam [24]. Jarak antara objek berpasangan di dalam *cluster* dirumuskan sebagai:

$$D_r = \sum_{i,i'} \in c_r \quad dii' \tag{14}$$

dengan d adalah jarak kuadrat *euclidien*. Jumlah kuadrat di dalam *cluster* dirumuskan sebagai berikut:

$$W_k = \sum_{r=1}^{k} 1 \frac{1}{2n_r} Dr \tag{15}$$

$$Gap_n(k) = E_n^*\{log(W_k)\} - log\,(W_k) \tag{16}$$

Nilai *gap* merupakan hasil estimasi jumlah *cluster* optimum dengan menggunakan pendekatan standarisasi W_k. Dengan E_n^* adalah ekspektasi dari distribusi jumlah sampel. Kriteria jumlah *cluster* optimal merupakan jumlah *cluster* yang memiliki nilai *gap statistics* tertinggi atau jika nilai *gap* selalu naik maka jumlah *cluster* optimum adalah nilai yang mengindikasi kenaikan *gap* minimum [25].

d). *Package* Nbclust di R

Selain menggunakan pendekatan metode *elbow, silhoute*, dan *gap statistics*, k optimum juga bisa didapat dengan bantuan R *package* yaitu menggunakan NbClust. Package NbClust menyediakan 30 indeks untuk menentukan jumlah *cluster* dan mengusulkan kepada pengguna skema pengelompokkan yang diperoleh dengan memvariasikan semua kombinasi jumlah *cluster*, ukuran jarak, dan metode pengelompokkan.

Pada [26] indeks untuk menentukan jumlah *cluster* pada *package* NbClust antara lain sebagai berikut:

1. *Duda Index*

[27] mengusulkan kriteria rasio seperti persamaan berikut:

$$Duda = \frac{Je(2)}{Je(1)} = \frac{W_k + W_l}{W_m} \tag{17}$$

Jumlah cluster yang optimal adalah nilai dari q terkecil, dimana z adalah skor standar normal.

$$Duda \geq 1 - \frac{2}{\pi p} - z\sqrt{\frac{2(1-\frac{8}{n^2 p})}{n_m P}} = crite \quad Value_{Duda} \tag{18}$$

2. *C-Index*

C-Index diulas oleh [28], yang dihitung menggunakan persamaan berikut. Nilai indeks minimum digunakan untuk menunjukkan jumlah *cluster* yang optimal.

$$Cindex = \frac{S_W - S_{min}}{S_{max} - S_{min}}, S_{min} \neq S_{max}, Cindex \in (0,1) \tag{19}$$

dengan:

S_{min}: jumlah dari N_w jarak terkecil antara semua pasangan titik dalam seluruh kumpulan data

S_{max}: jumlah dari N_w jarak terbesar antara semua pasangan titik dalam seluruh kumpulan data

3. *Gamma Index*

Indeks ini merupakan adaptasi dari [29] untuk digunakan dalam situasi *clustering*. Perbandingan dibuat antara semua *within cluster dissimilarities* dan semua *between cluster dissimilarities*. Nilai maksimum indeks diambil untuk mewakili jumlah *cluster* yang optimal.

$$Gamma = \frac{s(+) - s(-)}{s(+) + s(-)} \tag{20}$$

$s(+)$: Jumlah perbandingan yang sesuai
$s(-)$: Jumlah perbandingan sumbang

4. *Beale Index*
[30] mengusulkan penggunaan *F-ratio* untuk menguji hipotesis antara q_1 dan q_2 *cluster* dalam data $(q_2 > q_1)$. Jumlah *cluster* yang optimal diperoleh dengan membandingkan F dengan $F_{p,(nm-2)p}$. Tingkat signifikasi yang digunakan untuk menolak hipotesis adalah 10%.

$$Beale = F = \frac{\left(\frac{V_{kl}}{W_k+W_l}\right)}{\left(\left(\frac{n_m-1}{n_m-2}\right)2^{\frac{2}{p}}-1\right)} \quad (21)$$

dengan $V_{kl} = W_m - W_k - W_l$
Terdapat dua cara dalam mempertimbangkan jumlah *cluster* optimal pada NbClust yaitu yang pertama didasarkan pada aturan mayoritas yang tersedia dalam *package* NbClust. Misalkan dari 30 indeks, 20 indeks menyarankan 4 sebagai jumlah *cluster* optimal. Pilihan kedua mempertimbangkan hanya indeks yang berkinerja terbaik dalam studi simulasi. Sebagi contoh, digunakan empat *top performer* dalam studi [31] yaitu indeks Duda, C, Gamma, dan Beale sebagai mempertimbangkan jumlah *cluster*.

3. HASIL DAN PEMBAHASAN

3.1. Uji Asumsi

Tahapan sebelum melakukan analisis *cluster* yaitu melakukan uji asumsi, uji asumsi tersebut adalah uji kecukupan data dan kelayakan data serta uji multikolinearitas.

1) Uji Kecukupan Data dan Uji Kelayakan Data
KMO merupakan metode yang digunakan untuk menguji kecukupan data, sedangkan uji kelayakan setiap variabel digunakan kriteria MSA. Uji KMO merupakan indeks perbandingan jarak antara nilai koefisien korelasi terhadap korelasi parsial, dengan pendugaan awal jumlah sampel tidak memenuhi kriteria untuk analisa lanjut.

Tabel 2. Hasil Uji KMO dan MSA

Variabel	Nilai KMO	Nilai MSA
x_1		0,767
x_2		0,812
x_3		0,635
x_4	0,761	0,736
x_5		0,766
x_6		0,946
x_7		0,640

Berdasarkan Tabel 2 nilai KMO yang diperoleh lebih besar dari 0,5, begitu pula nilai MSA yang didapat dari setiap variabel. Oleh karena itu, diperoleh kesimpulan bahwa sampel untuk setiap kabupaten/kota mewakili populasi dan layak untuk dilakukan analisis *cluster*.

2) Uji Multikolinearitas
Hasil dari matriks korelasi menunjukkan bahwa nilai korelasi tertinggi antar variabel terjadi pada variabel ke-1 (RLS) dan ke-5 (persentase kepemilikan rumah sendiri) maka dapat disimpulkan bahwa data terjadi gejala multikolinearitas antar variabel.
Selanjutnya dilakukan pengecekan multikolinearitas juga menggunakan nilai VIF, variabel dependen yang digunakan merupakan jumlahan dari seluruh variabel masing-masing provinsi. Menurut [19], jika nilai VIF melebihi 5 atau 10 maka menunjukkan bahwa terjadi gejala multikolinearitas antar variabel. Berdasarkan hasil nilai VIF, terdapat nilai VIF yang melebihi 5 yaitu variabel RLS, sehingga disimpulkan bahwa data terjadi gejala multikolinearitas. Dengan demikian asumsi multikolinearitas tidak terpenuhi dan akan diatasi dengan metode PCA.

3) Analisis PCA
Pada penelitian ini akan digunakan metode PCA terhadap *dataset* asli, variabel yang memiliki korelasi tinggi akan diubah menjadi variabel baru yang lebih kecil dan saling independen (tidak berkorelasi lagi).
Variabel data indikator kesejahteraan masyarakat memiliki satuan dan rentang yang berbeda, maka data perlu distandarisasi menggunakan *Z-score* agar data memiliki bobot yang sama dalam pembentukan komponen utama. Setelah didapatkan hasil standarisasi data, selanjutnya menghitung matriks kovarians yang hasilnya digunakan untuk menghitung *eigenvalue* dan *eigenvector*.
Banyaknya komponen utama ditentukan berdasarkan *eigenvalues* yang nilainya lebih dari 1. Berdasarkan tabel 3 *eigenvalues* yang bernilai lebih dari 1 terdapat pada komponen utama ke-1 dan ke-2, sehingga banyaknya komponen utama yang terbentuk sebanyak 2 faktor. *Proportion of variance* menunjukan kemampuan tiap faktor dalam menjelaskan variabilitas keseluruhan data. Kemampuan masing-masing komponen utama ke-1 dan ke-2 mampu menjelaskan variansi dari 7 variabel sebesar 55,32% dan 24,175%. Serta *cumulative proportion* yaitu kumulatif atau penjumlahan dari nilai *proportion of variance*, yang berarti bahwa variansi dari 7 variabel mampu dijelaskan oleh komponen ke-1 dan ke-2 sebesar 78,762%. Artinya, pembentukan 2 komponen utama sudah cukup baik untuk digunakan untuk analisis selanjutnya karena memiliki nilai *cumulative propotion* di atas 50%.
Selanjutnya, kombinasi linear pada pembentukan skor baru untuk setiap komponen utama menggunakan nilai *eigenvector* dapat dilihat pada Tabel 4.
Berdasarkan 4 nilai kombinasi linier untuk komponen utama ke-1 dan ke-2 adalah sebagai berikut:

$$y_1 = -0,438x_1 - 0,413x_2 + 0,295x_3 + 0,355x_4 + 0,426x_5 - 0,382x_6 - 0,31x_7 \quad (22)$$

$$y_2 = 0,211x_1 + 0,282x_2 + 0,621 - 0,365x_4 - 0,009x_5 + 0,066x_6 - 0,594x_7 \quad (23)$$

Setelah didapatkan persamaan komponen utamanya, maka dapat dicari data baru hasil pembentukan komponen utama yang dihasilkan dengan cara mensubtitusikan data yang sudah distandarisasi ke dalam Persamaan 22 dan persamaan 23.
Langkah berikutnya adalah melakukan pengujian ulang multikolinearitas untuk membuktikan apakah data sudah tidak

Tabel 3. Hasil Analisis PCA

	F1	F2	F3	F4	F5	F6	F7
Eigen value	4,307	1,206	0,528	0,446	0,273	0,145	0,094
Variability (%)	61,531	17,231	7,543	6,376	3,899	2,074	1,347
Cumulative (%)	61,531	78,762	86,305	92,68	96,579	98,653	100

Tabel 4. *Output Eigenvector*

	PC1	PC2
x_1	-0,438	0,211
x_2	-0,413	0,282
x_3	0,295	0,621
x_4	0,355	-0,365
x_5	0,426	-0,009
x_6	-0,382	0,066
x_7	-0,310	-0,594

terdapat multikolinearitas. Diperoleh hasil bahwa data hasil PCA sudah bebas dari multikolinearitas.

3.2. *K-Means Clustering*

3.2. *Penentuan Jumlah Cluster*

Tahap pertama algoritma dalam K-Means clustering yaitu menentukan jumlah *cluster (k)*. Penentuan k optimum pada penelitian ini menggunakan pendekatan metode *elbow*, *silhoutte*, dan *gap statistics*.

Gambar 2(a) merupakan hasil pendekatan *Elbow* untuk mendapatkan k optimal, berdasarkan plot dapat dilihat bahwa patahan gradien terbesar saat jumlah cluster sebesar 3. Plot metode *silhoutte* (b) menunjukkan bahwa jumlah *cluster* optimal adalah dua, tetapi jumlah ini terlalu sedikit sehingga digunakan nilai terbesar selanjutnya yaitu 3 *cluster*. Sedangkan plot *gap statistics* (c) menunjukkan jumlah *cluster* optimal sebesar satu, namun jumlah ini tidak sesuai dengan penelitian sehingga digunakan nilai terbesar setelahnya yaitu dipilih 3 *cluster*.

Penelitian ini juga menggunakan *package* NbClust pada R untuk menentukan jumlah *cluster* optimum. Hasil yang diperoleh dari total semua indeks, 6 indeks mengusulkan 2 sebagai jumlah *cluster* terbaik, 7 indeks mengusulkan 3 sebagai jumlah *cluster* terbaik, 2 indeks mengusulkan 4 sebagai jumlah *cluster* terbaik, 1 indeks mengusulkan 5 sebagai jumlah *cluster* terbaik, dan 7 indeks mengusulkan 6 sebagai jumlah *cluster* terbaik. Oleh karena itu, disimpulkan bahwa dari sebagian besar indeks dapat disimpulkan bahwa jumlah *cluster* terbaik sebesar 3 *cluster*.

Selanjutnya, SSE digunakan sebagai metode pengujian terhadap *cluster* untuk mengetahui persentase performa dari *cluster* yang digunakan dari masing-masing hasil *cluster*.

Tabel 5. Nilai SSE

Cluster	SSE	Selisih
2	0,60	0,60
3	0,75	0,15
4	0,82	0,07
5	0,87	0,05

Berdasarkan pada tabel 5 nilai SSE yang mengalami penurunan paling besar adalah pada $k = 3$ atau *cluster* 3 dengan jumlah data yang sama. Oleh karena itu, untuk kasus ini jumlah *cluster* yang ideal adalah k = 3 dan dijadikan *default cluster* untuk menentukan karakteristik dari data-data tersebut. Persentase nilai SSE untuk $k = 3$ sebesar 75% yang menunjukkan bahwa nilai $k = 3$ sudah cukup optimal.

3.2. *Pembentukan Cluster*

Hasil penentuan k dari kombinasi metode PCA dan *K-Means* membentuk 3 *cluster* yaitu kelompok dengan kesejahteraan rendah, sedang, dan tinggi. Gambar 3 merupakan hasil *output* pengelompokkan wilayah berdasarkan *cluster* menggunakan warna dan bentuk yang berbeda.

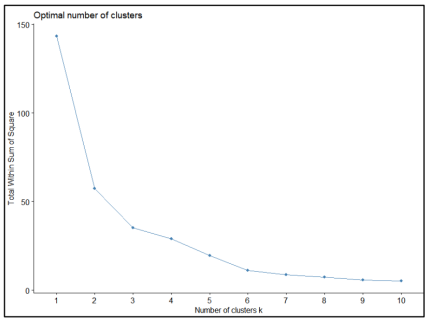
Pada gambar 3 warna merah menunjukkan kelompok wilayah terklasifikasi ke dalam *cluster* 1, warna hijau menunjukkan wilayah terklasifikasi dalam *cluster* 2, serta warna biru dalam *cluster* 3. Jumlah anggota masing-masing *cluster* terdiri 12, 8, dan 7 kabupaten/kota.

Berdasarkan Tabel 6, dapat diketahui bahwa kabupaten/kota yang termasuk dalam *cluster* 1 mempunyai kateristik seluruh variabel dominan sedang dibandingkan dengan *cluster* lainnya. Sedangkan, pada *cluster* 2 mempunyai kateristik dominan tinggi pada variabel x_3 (TPAK), x_4 (Pesentase Penduduk Miskin), dan x_5 (Persentase Kepemilikan Rumah Sendiri) serta variabel tersisa memiliki kateristik dominan rendah. Adapun kabupaten/kota yang termasuk dalam *cluster* 3 dimana variabel x_1 (RLS), x_2 (Daya beli), x_6 (Persentase Penduduk yang memiliki jaminan kesehatan), dan x_7 (TPT) memiliki kateristik dominan tinggi, sedangkan variabel tersisa memiliki variabel rendah.

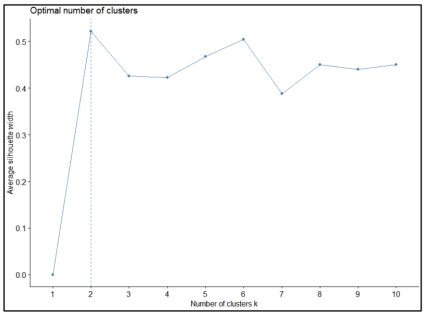
3.3. *Analisis Hasil Clustering*

Hasil profilisasi selanjutnya digunakan untuk mengidentifikasi akar permasalahan kesejahteraan di Provinsi Jawa Barat sehingga dapat diperoleh solusi yang sesuai untuk masing-masing *cluster*.

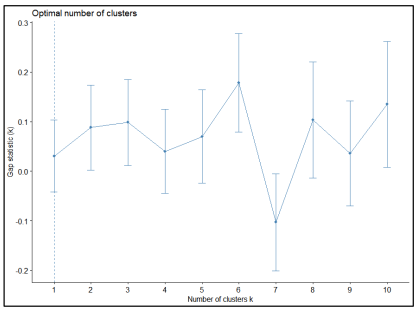
Berdasarkan gambar4, pada kepala diagram ujung kanan merupakan permasalahan kesejahteraan masyarakat Provinsi Jawa Barat tahun 2021. Identifikasi kategori garis besar masalah tersebut terbagi menjadi 3 kelompok, yaitu wilayah pada *cluster* 1, 2, dan 3. *Cluster* 1 terdiri dari 12 Kabupaten/Kota yaitu Bogor, Sukabumi, Bandung, Kuningan, Cirebon, Sumedang, Purwakarta, Bekasi, Bandung Barat, Kota Tasikmalaya, dan Kota Banjar. Sedangkan *cluster* 2 dan 3 masing-masing terdiri 8 dan 7 Kabupaten/Kota. Anggota *cluster* 2 adalah Cianjur, Garut, Tasikmalaya, Ciamis, Majalengka, Indramayu, Subang, dan Pangandaraan, sedangkan anggota *cluster* 3 adalah Kota Bogor, Kota Sukabumi, Kota Bandung, Kota Cirebon, Kota Bekasi, Kota Depok, dan Kota Cimahi. Selanjutnya dijabarkan penjelasan penyebab masalah menggunakan panah kecil. Terlihat wilayah yang berada pada *cluster* 2 perlu penanganan tindak lanjut terlebih dahulu karena memiliki memiliki nilai RLS, daya beli, dan penduduk yang



(a) Metode *Elbow*



(b) Metode *Silhouette*

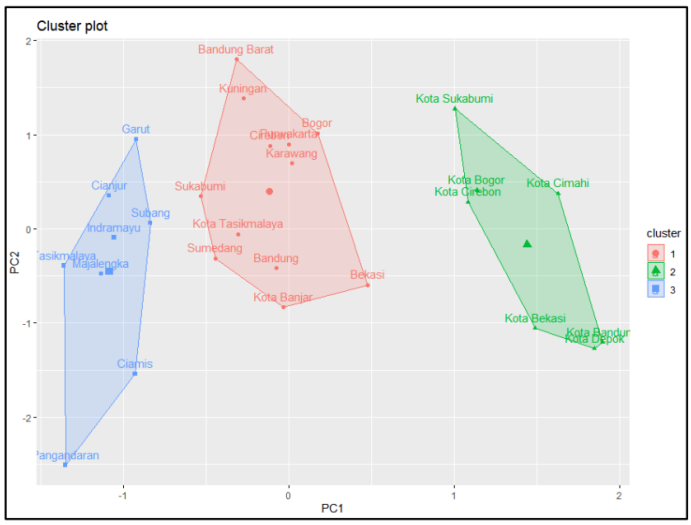


(c) Metode *Gap Statistics*

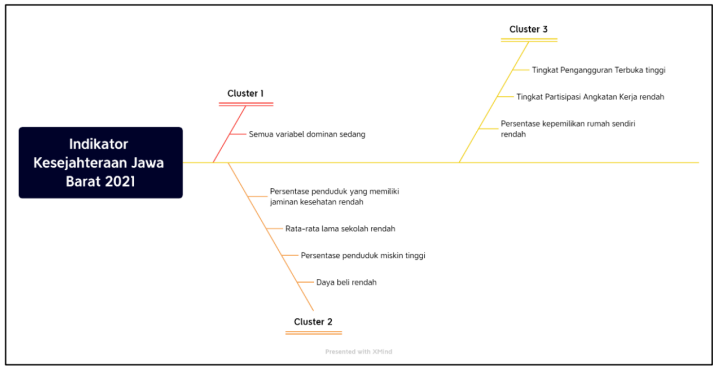
Gambar 2. Penentuan k optimum

Tabel 6. Tabel 6. Profilisasi Hasil Cluster

Cluster	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	8,30	10.21,08	64,27	9,47	83,03	57,82	9,94
2	7,36	9.052,63	68,60	10,75	89,57	46,97	7,03
3	10,76	13.543,71	62,28	6,08	61,20	69,67	11,18



Gambar 3. Visualisasi Hasil *Clustering*



Gambar 4. *Fishbone Diagram*

memiliki jaminan kesehatan rendah tetapi pesentase penduduk miskin tinggi. *Cluster* 2 perlu perhatian lebih karena masalah satu dengan yang lainnya saling berkaitan, penyebab persentase penduduk tinggi salah satu faktornya adalah tingkat pendapatan rumah tangga rendah atau tidak ada pendapatan sama sekali. Hal tersebut menyebabkan daya beli masyarakat rendah, pemerintah perlu menjaga stabilitas harga/daya beli masyarakat karena konsumsi penyumbang utama pertumbuhan ekonomi di suatu daerah. Penyebab lainnya yaitu tidak memiliki pendidikan atau ketrampilan rendah atau tidak sekolah baik Pendidikan formal maupun informal. RLS dapat digunakan untuk mengetahui kualitas Pendidikan masyarakat dalam suatu wilayah, jika nilai RLS rendah menunjukkan kualitas sumber daya manusia di wilayah tersebut juga rendah. Kemudian, karena rendahnya tingkat kesehatan dan tidak cukup memiliki akses fasilitas kesehatan juga menjadi faktor terjadinya kemiskinan.

Penanggulangan kemiskinan merupakan inti tugas dari pemerintah daerah mengeluarkaan aneka kebijakan untuk yang dapat meningkatkan kesejahteraan warganya. Salah satunya dimulai dengan pembenahan basis data informasi tentang kemiskinan serta keterpaduan penanganan di pusat, provinsi, dan kabupaten/kota sehingga upaya penanganan bisa tepat sasaran, efektif, dan efisien. Memanfaatkan angaran untuk mengadakan pelatihan, peminjaman modal, atau gerakan-gerakan konkrit kepada masyarakat. Kemudian, berdayakan kepala desa atau PKK, RT, atau RW sebagai unit terkecit dengan mengoptimalkan dana desa. Selain itu, untuk meningkatkan RLS bisa dengan memberikan kesempatan bagi penduduk yang putus sekolah untuk mengambil paket A, B, atau C. Langkah lain adalah dengan mengupayakan mengajarkan pendidikan dasar di setiap wilayahnya seperti membaca dan menulis. Sosialisasi program BPJS dan memperluas kepesertaan jaminan kesehatan juga diperlukan untuk meningkatkan tingkat kesehatan.

4. KESIMPULAN

Dalam penelitian ini, terdapat uji asumsi yang tidak terpenuhi yaitu data harus bebas dari multikolinearitas, maka untuk mengatasinya digunakan analisis PCA untuk mereduksinya. Faktor yang digunakan sebanyak dua berdasarkan *eigenvalues*

yang lebih dari satu. Hasil pengelompokkan dari analisis *cluster* dengan metode *K-Means* terbentuk 3 *cluster* terbaik dimana jumlah anggota masing-masing terdiri 12, 8, dan 7 kabupaten/kota. Berdasarkan profilisasi data, kabupaten/kota yang termasuk dalam wilayah pada *cluster* 2 perlu penanganan tindak lanjut yang didahulukan daripada wilayah lainnya karena memiliki nilai RLS, daya beli, dan penduduk yang memiliki jaminan kesehatan rendah sehingga menyebabkan pesentase penduduk miskinnya tinggi.

References

[1] L. Zhang, "A feature selection algorithm integrating maximum classification information and minimum interaction feature dependency information," *Computational Intelligence and Neuroscience*, 2021.

[2] J. Shlens, "A tutorial on principal component analysis," *http://arxiv.org/abs/1404.1100*, 2014.

[3] I. Jolliffe, "Principal components analysis," *Wiley StatsRef: Statistics Reference Online*, 2014.

[4] Z. John Lu, "The elements of statistical learning: data mining, inference, and prediction," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 173, no. 3, 2010.

[5] A. Deshpande and K. Varadarajan, "Sampling-based dimension reduction for subspace approximation," in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, 2007. doi: 10.1145/1250790.1250884 pp. 641–650.

[6] J. Wang, C. Xia, Y. Wu, X. Tian, K. Zhang, and Z. Wang, "Rapid detection of carbapenem-resistant klebsiella pneumoniae using machine learning and maldi-tof ms platform," *Infection and Drug Resistance*, vol. 15, pp. 3703–3710, 2022.

[7] J. Yang, Y. K. Wang, X. Yao, and C. T. Lin, "Adaptive initialization method for k-means algorithm," *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[8] A. Deshpande and K. Varadarajan, "Sampling-based dimension reduction for subspace approximation," in *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, 2007. doi: 10.1145/1250790.1250884 pp. 641–650.

[9] J. Duo, P. Zhang, and L. Hao, "A k-means text clustering algorithm based on subject feature vector," *Journal of Web Engineering*, vol. 20, no. 6, pp. 1935–1946, 2021.

[10] K. Katahira, "Evaluating the predictive performance of subtyping: A criterion for cluster mean-based prediction," *Statistics in Medicine*, vol. 42, no. 7, pp. 1045–1065, 2023.

[11] R. Lakshmi and S. Baskar, "Dic-doc-k-means: Dissimilarity-based initial centroid selection for document clustering using k-means for improving the effectiveness of text document clustering," *Journal of Information Science*, vol. 45, no. 6, pp. 818–832, 2019.

[12] K. Shanthi and D. S. .M, "Performance analysis of improved k-means & k-means in cluster generation," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 3, no. 9, pp. 11 878–11 884, 2014.

[13] J. Yang, Y.-K. Wang, X. Yao, and C.-T. Lin, "Adaptive initialization method for k-means algorithm," *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[14] A. L. Yusniyanti, F. Virgantari, and Y. E. Faridhan, "Comparison of average linkage and k-means methods in clustering indonesia's provinces based on welfare indicators," *Journal of Physics: Conference Series*, vol. 1863, no. 1, 2021.

[15] BPS Jabar, *Badan Pusat Statistik Provinsi Jawa Barat*. Badan Pusat Statistik Provinsi Jawa Barat, 2021.

[16] I. T. Jolliffe, *Principal Component Analysis*. Springer Science & Business Media, 2013.

[17] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. Prentice Education, Inc., 2007.

[18] J. F. Hair, R. E. Anderson, R. L. Tatham, and W. C. Black, *Multivariate Data Analysis 5th Edition*, 5th ed. Prentice-Hall, Inc., 1998.

[19] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis Solutions Manual to Accompany*. John Wiley & Sons, 2013.

[20] B. Everitt and T. Hothorn, *An introduction to applied multivariate analysis with R*. Springer Science & Business Media, 2011.

[21] C. L. Clayman, S. M. Srinivasan, and R. S. Sangwan, "K-means clustering and principal components analysis of microarray data of l1000 landmark genes," *Procedia Computer Science*, vol. 168, pp. 97–104, 2020.

[22] T. M. Kodinariva and P. R. Makwana, "Review on determining number of cluster in k-means clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 6, 2013.

[23] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis perbandingan metode elbow dan silhouette pada algoritma clustering k-medoids dalam pengelompokan produk kerajinan bali," *MATRIX: Jurnal Manajemen Teknologi dan Informatika*, vol. 9, no. 3, 2019.

[24] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society*, vol. 63, no. 2, pp. 411–423, 2001.

[25] R. Silvi, "Analisis cluster dengan data outlier menggunakan centroid linkage dan k-means clustering untuk pengelompokan indikator hiv/aids di indonesia," *Jurnal Matematika MANTIK*, vol. 4, no. 1, pp. 22–31, 2018.

[26] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, "Nbclust: An r package for determining the relevant number of clusters in a data set," *Journal of Statistical Software*, vol. 61, no. 6, pp. 1–36, 2014.

[27] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification and scene analysis*. Wiley New York, 1973, vol. 3.

[28] L. J. Hubert and J. R. Levin, "A general statistical framework for assessing categorical clustering in free recall," *Psychological Bulletin*, vol. 83, no. 6, pp. 1072–1080, 1976.

[29] L. A. Goodman, W. H. Kruskal, L. A. Goodman, and W. H. Kruskal, *Measures of association for cross classifications*. Springer, 1979.

[30] E. M. L. Beale, *Cluster analysis*. Scientific Control Systems, 1969.

[31] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.