PENGELOMPOKAN PROVINSI DI INDONESIA BERDASARKAN PENGELUARAN PER KAPITA TAHUN 2023 DENGAN METODE *K-MEANS* DAN *K-MEDOIDS*

Sitti Ulfa Nur Fadhilah[1], Yudi Setyawan[2*], Rokhana Dwi Bekti[3]
[1,2,3] Program Studi Statistika Universitas AKPRINND Indonesia, Penulis Koresponden*
e-mail :[1]nurfadillahulfa9@gmail.com,[2]setyawan@akprind.ac.id, [3]rokhana@akprind.ac.id

ABSTRACT

Based on National Socio-economic Survey data for March 2023 at the national level, the average monthly per capita expenditure on food and non-food consumption increased by 9.35 percent since March 2022. This increase is thought to be triggered by several factors such as the value of inflation, increasing economic growth, decreasing open unemployment rates and the impact of social protection programs that the government has implemented. The purpose of this study is to determine the results of grouping per capita expenditure in Indonesia in 2023 based on food and non-food groups and provide an overview so that it can be a reference for the government to estimate market needs in each province. From the results of the analysis, it is known that the total number of variables used was twenty variables and there were five variables that experienced multicollinearity so that PCA analysis was carried out to change the dimensionality of the data without losing important information from the data. In this analysis, a comparison of K-Means and K-Medoids analysis is carried out using Euclidean distance and Manhattan distance. The best method is determined based on the results of validation tests, namely the Silhouette Index, Davies-Bouldin Index (DBI) and also the Connectivity Index. Based on the validation test results, the best method is the K-Medoids method using Manhattan distance because it has the smallest DBI and Connectivity indices compared to other methods. Clustering based on this best method is divided into 2 clusters where cluster 1 consists of 21 provinces and cluster 2 comprises of 13 provinces.

Keywords: *Euclidean distance, K-Means, K-Medoids, Manhattan distance, PCA*

INTISARI

Berdasarkan data Survei Sosial-Ekonomi Nasional (Susenas) Maret 2023 pada level nasional, rata-rata pengeluaran per kapita sebulan untuk konsumsi makanan dan bukan makanan naik 9,35 persen sejak Maret tahun 2022. Kenaikan ini diduga dipicu oleh beberapa faktor seperti nilai inflasi, meningkatnya pertumbuhan ekonomi, menurunnya tingkat pengangguran terbuka serta dampak dari program-program perlindungan sosial yang telah pemerintah jalankan. Tujuan dari penelitian ini untuk mengetahui hasil pengelompokan pengeluaran per kapita di Indonesia pada tahun 2023 berdasarkan kelompok makanan dan non makanan serta memberikan gambaran sehingga dapat menjadi acuan bagi pemerintah untuk mengestimasi kebutuhan pasar di masing-masing provinsi. Dari hasil analisis diketahui bahwa jumlah total variabel yang digunakan sebanyak dua puluh variabel dan terdapat lima variabel yang mengalami multikolinieritas sehingga dilakukan analisis PCA untuk mengubah dimensi dari data tanpa menghilangkan informasi penting dari data tersebut. Dalam penelitian ini dilakukan perbandingan analisis klaster *K-Means* dan *K-Medoids* dengan menggunakan jarak *Euclid* dan jarak *Manhattan*. Penentuan metode terbaik menggunakan beberapa uji validasi yakni *Silhouette Index*, *Davies-Bouldin Index* (DBI) dan *Connectivity Index*. Uji validasi menunjukkan bahwa metode terbaik adalah metode *K-Medoids* dengan jarak *Manhattan* karena memiliki nilai DBI dan *Connectivity Index* yang paling kecil dibandingkan metode lainnya. Pengelompokan berdasarkan metode terbaik menghasilkan 2 klaster dimana klaster 1 memuat 21 provinsi dan klaster 2 memuat 13 provinsi.

Kata kunci: *K-Means, K-Medoids, jarak Euclid, jarak Manhattan, PCA*

1. PENDAHULUAN

Dalam Survei Sosial-Ekonomi Nasional (Susenas), pengeluaran untuk konsumsi makanan dihitung selama seminggu yang lalu, sedangkan untuk bukan makanan dihitung selama sebulan dan 12 bulan yang lalu. Baik konsumsi makanan maupun bukan makanan selanjutnya dikonversikan ke dalam pengeluaran rata-rata sebulan. Angka-angka konsumsi/pengeluaran rata-rata per kapita yang disajikan dalam publikasi Susenas diperoleh dari hasil bagi jumlah konsumsi seluruh rumah tangga (baik mengkonsumsi makanan maupun tidak) terhadap jumlah penduduk. Berdasarkan data Susenas di bulan Maret 2023 (Samudro, 2023), pada level nasional, rata-rata

pengeluaran per kapita sebulan untuk konsumsi makanan dan bukan makanan sebesar 1.451.870 rupiah atau naik 9,35 persen dari Maret tahun 2022. Data Susenas dapat digunakan untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan pengeluaran per kapita sehingga provinsi-provinsi dalam kelompok yang sama memiliki tingkat kemiripan yang tinggi, sedang provinsi-provinsi dari kelompok yang berbeda memiliki tingkat perbedaan yang tinggi. Pengelompokan ini akan membantu pemerintah dalam pengambilan kebijakan yang diperlukan.

Analisis klaster adalah suatu metode dalam analisis multivariat yang digunakan untuk menempatkan suatu objek ke dalam beberapa kelompok atau klaster berdasarkan karakteristik yang dimilikinya. Klaster-klaster yang terbentuk memiliki karakteristik objek identik, sedangkan antar klaster memiliki karakteristik yang berbeda (Nabilah, Perdana, & Sulistianingsih, 2024).

Algoritma pengelompokan dapat diklasifikasikan menjadi empat jenis, yaitu pengelompokan non-hirarki (partisi), pengelompokan hirarki, pengelompokan berdasarkan densitas, dan pengelompokan berdasarkan *grid*. Metode pengelompokan hirarki digunakan untuk mengelompokkan pengamatan secara terstruktur berdasarkan sifat kemiripannya, dan kelompok yang diinginkan belum diketahui banyaknya. Sedangkan metode pengelompokan non-hirarki, digunakan untuk mengelompokkan objek-objek pengamatan menjadi $\boldsymbol{k}$ kelompok.

Beberapa metode klaster non-hierarki antara lain adalah *X-Means Clustering*, *K-Means Clustering*, *C-Means Clustering*, *K-Medoids*, *Fuzzy K-Means*, *Fuzzy C-Means* dan lain-lain. Dalam penelitian ini dilakukan pengelompokan dengan metode *K-Means* dan *K-Medoids* dengan menggunakan jarak Euclid dan Manhattan. Alasan menggunakan metode *K-Means* dan *K-Medoids* adalah mengetahui algoritma pengelompokan wilayah paling efektif guna membantu pemerintah dalam pengambilan kebijakan yang diperlukan.

Dalam prakteknya, algoritma pengklasteran menemui masalah ketika dihadapkan dengan data dimensi tinggi antara lain masalah multikolinieritas. Untuk itu diperlukan proses optimasi agar kinerja algoritma tetap stabil. Salah satu caranya adalah dengan melakukan proses *PCA* (*Principal Component Analysis*) untuk mereduksi dimensi, yakni proses pengurangan jumlah dimensi tanpa menghilangkan informasi penting dari suatu data. Setelah hasil pengklasteran diperoleh perlu dilakukan validasi. Beberapa metode validasi hasil pengklasteran antara lain adalah *Silhouette Indeks (SI), Davies-Bouldin Index* (*DBI*), dan *Connectivity Index*.

Penelitian yang telah dilakukan sebelumnya meliputi studi perbandingan algoritma *K-Means* dan *K-Medoids* untuk pengelompokan data obat. Penelitian ini menganalisis tentang referensi untuk perencanaan obat yang akan datang di puskesmas Karangsambung. Hasil penelitian menunjukkan bahwa algoritma *K-Means* menghasilkan *Silhouette Coefficient* sebesar 0,627 yang lebih tinggi dibandingkan *K-Medoids* sebesar 0,536 sehingga *K-Means* lebih berkualitas dibandingkan *K-Medoids* (Farissa, Mayasari, & Umaidah, 2021).

Penelitian tentang perbandingan algoritma *K-Means* dan *K-Medoids* pada pengelompokan armada kendaraan truk berdasarkan produktivitas juga dilakukan oleh (Supriyadi, Triayudi, & Sholihati, 2021). Hasil validasi *Davies Bouldin-Index (DBI)* diperoleh metode *K-Means* sebesar 0,67 dan *K-Medoids* sebesar 1,78. Berdasarkan hasil tersebut, algoritma *K-Means* dipilih untuk diimplementasikan pada pembuatan aplikasi *clustering* armada kendaraan berbasis web. Pengujian terhadap hasil *clustering* dengan *tool Rapidminer* dan manual menghasilkan kesesuaian sebesar 97%.

Hingga saat ini belum ada penelitian terkait pengeluaran per kapita yang membandingkan metode *K-Means* dan *K-Medoids*. Oleh karena itu, penelitian ini bertujuan untuk mengetahui algoritma pengelompokan wilayah paling efektif guna membantu pemerintah dalam pengambilan kebijakan yang diperlukan.

2. METODE PENELITIAN

Pada penelitian ini digunakan metode *K-Means* dan *K-Medoids clustering* untuk mengelompokkan 34 Provinsi di Indonesia berdasarkan rata-rata pengeluaran per kapita (rata-rata per orang) per bulan (dalam rupah) tahun 2023. Penelitian menggunakan data sekunder berupa data pengeluaran per kapita per bulan (dalam rupiah) di Indonesia berdasarkan kelompok makanan dan bukan makanan yang diambil dari hasil Susenas bulan Maret 2023.
Variabel yang digunakan dalam penelitian ini berjumlah 20 variabel dalam satuan rupiah meliputi padi-padian, umbi-umbian ikan/udang/cumi-cumi/kepiting, daging, telur dan susu, sayur-sayuran, kacang-kacangan, buah-buahan, minyak dan kelapa, bahan minuman, bumbu-bumbuan, konsumsi lainnya, makanan dan minuman jadi serta rokok, perumahan dan fasilitas rumah tangga, aneka barang dan jasa, pakaian, alas kaki dan penutup kepala, barang tahan lama, pajak pungutan dan asuransi serta keperluan pesta dan upacara/kenduri.
Dari data yang ada dilakukan pendeteksian outlier serta penanganannya apabila diperlukan. Selanjutnya dilakukan uji multikolinieritas untuk mengetahui adanya korelasi antar variabel. Penanganan multikolinieritas dilakukan

dengan analisis komponen utama (*principal component analysis*) sehingga menjamin tidak terjadinya korelasi antar variabel. Langkah berikutnya adalah uji kecukupan sampel menggunakan uji *Kaiser-Mayer-Olkin.* Apabila data sudah lolos semua uji tersebut, barulah dilakukan analisis klaster.
Analisis klaster diawali dengan penentuan jumlah klaster terbaik dengan metode Elbow dan Silhouette. Selanjutnya dilakukan *clustering* dengan metode *K-Means* dan *K-Medoids* menggunakan jarak Euclid dan Manhattan. Untuk memvalidasi hasilnya *digunakan Silhouette Index, Davies-Bouldin Indeks,* dan *Connectivity Index*. Langkah terakhir adalah *profiling* masing-masing klaster agar dapat dimanfaatkan pemangku kepentingan untuk mengambil kebijakan yang sesuai.

2.1 Deteksi *outlier*

Outlier merupakan nilai yang berada pada titik paling ekstrim dari suatu kumpulan data. Beberapa *outlier* merupakan perwakilan dari nilai variansi dalam populasi. *Outlier* tidak selalu merupakan data yang salah sehingga perlu kehati-hatian dalam proses penanganannya. *Outlier* juga dapat disebabkan oleh entri data yang salah, malfungsi peralatan, atau kesalahan pengukuran lainnya, (Bhandari, 2021). Pengecekan *outlier* data dapat dilakukan dengan melihat nilai standar (*Z score*) dari data. Suatu data dianggap sebagai *outlier* jika nilai *Z score* lebih kecil dari -3,00 atau lebih besar dari 3,00. Data awal X yang memiliki mean μ dan simpangan baku σ ditransformasikan dengan menggunakan rumus sebagai berikut: (Ghozali, 2018).

$$Z = \frac{X-\mu}{\sigma} \qquad \text{..... (1)}$$

2.2 Analisis komponen utama (PCA)

Analisis Komponen Utama (*Principal Component Analysis/PCA)* merupakan algoritma yang berguna untuk menentukan komponen dari data yang lebih kompleks. Algoritma *PCA* bekerja untuk mengubah variabel independen awal menjadi variabel independen baru yang tak berkorelasi. Variabel baru ini berisi nilai variabel komponen utama (*PC*), (Ramadhayani & Lusiana, 2022). PCA akan membentuk sekumpulan dimensi baru yang kemudian diurutkan berdasarkan varian datanya. PCA menghasilkan komponen utama melalui dekomposisi *eigen value dan eigen vector* dari matriks kovarians (Jamal, Handayani, Septiandri, & Ripmiatin, 2018).
Langkah-langkah algoritma PCA adalah sebagai berikut:

1. Menghitung *mean* dari data pada tiap dimensi.
2. Menghitung *covariance matrix*.
3. Menghitung *eigen vector* (υ) dan *eigen value* (λ) dari *covariance matrix*.
4. Mengurutkan *eigen value* dari terbesar ke terkecil. *Principal Component* (PC) adalah deretan *eigen vector* sesuai dengan urutan *eigen value* pada tahap 3.
5. Menghasilkan dataset baru.

2.3 Analisis klaster

Analisis klaster adalah kegiatan mencari kesamaan dalam data dan menempatkan data yang sama ke dalam kelompok-kelompok. Tujuan utama dari analisis klaster adalah untuk mengidentifikasi pola atau struktur yang tersembunyi dalam data dan mengelompokkan data ke dalam kelompok-kelompok yang homogen, dengan objek-objek dalam satu kelompok memiliki kemiripan yang tinggi satu sama lain, sementara berbeda dengan objek-objek di kelompok lain.

2.4 Jarak Euclid

Jarak Euclid merupakan perhitungan jarak antara dua titik dalam bidang Euclid untuk mencari hubungan antara sudut dan jarak. Jarak Euclid dapat digunakan untuk menghitung jarak antara titik *centroid* dengan masing-masing objek (Cahaya, 2023). Rumus jarak Euclid adalah:

$$d_{euc}\left(x_i, x_j\right) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \qquad \text{..... (2)}$$

dengan,

$d_{euc}\left(x_i x_j\right)$: jarak Euclid antara x_i dan x_j
p : dimensi data
x_{ik} : objek ke-*i* pada variabel ke-*k*
x_{jk} : objek ke-*j* pada variabel ke-*k*

2.5 Jarak Manhattan

Jarak Manhattan merupakan salah satu pengukuran yang paling banyak digunakan meliputi penggantian perbedaan kuadrat dengan menjumlahkan perbedaan absolut dari variabel-variabel. Prosedur ini disebut *block absolute* atau

P - ISSN: 1979-911X
E - ISSN: 2541-528X

lebih dikenal dengan *city block distance* digunakan untuk menghitung perbedaan absolut (mutlak) antara koordinat sepasang objek. Rumus jarak Manhattan antara x_i dan x_j diberikan oleh:

$$d_{man}\left(x_i, x_j\right) = \sum_{k=1}^{p} |x_{ik} - x_{jk}| \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

2.6 Metode *elbow*

Metode *Elbow* merupakan salah satu metode untuk menentukan jumlah klaster yang tepat melalui persentase hasil perbandingan antara jumlah klaster yang akan membentuk siku pada suatu titik. Untuk mendapatkan perbandingannya adalah dengan menghitung *Sum of Square Error* (SSE) dari masing-masing nilai klaster. Karena semakin besar jumlah nilai klaster K, maka nilai SSE akan semakin kecil. Rumus SSE diberikan oleh:

$$SSE = \sum_{k=1}^{K} \sum_{x_i} d^2(x_i, c_k) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

dengan,
K: banyak klaster
x_i: data objek ke-*i* dalam klaster ke-*k*
c_k: pusat klaster ke-*k*

2.7 Metode *silhouette*

Metode *Silhouette* digunakan untuk melihat kualitas dan kekuatan klaster, yakni seberapa baik atau buruk suatu objek ditempatkan dalam suatu klaster. Metode ini merupakan gabungan dari metode separasi dan kohesi. Nilai *silhouette* diperoleh dari hasil pengukuran kedekatan suatu titik data dengan titik lain dalam klaster yang sama (kohesi) dibandingkan dengan seberapa jauh suatu titik data dari titik lain dalam klaster lain (pemisah). Hasil perhitungan *Silhouette Indeks (SI)* memiliki rentang -1 hingga 1. Rumus *SI* diberikan oleh:

$$SI = \frac{1}{n} \sum_{i=1}^{n} S(i) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

dengan,
S(i), yakni Silhouette Coefficient data ke-i diberikan oleh:

$$S(i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

dengan,
$a(x_i)$: Rata-rata jarak antara observasi ke-i dengan observasi lain dalam satu klaster
$b(x_i)$: Rata-rata jarak antara observasi ke-i dengan observasi lain pada klaster terdekat.

2.8 *K-Means Clustering*

K-Means Clustering merupakan metode pengklasteran paling sederhana karena dapat mengelompokkan data yang berjumlah cukup besar dalam waktu yang cepat dan efisien. *K-Means* menjauhkan data yang berada pada suatu klaster dengan data pada klaster lainnya. Hal ini dilakukan agar variasi antar data yang berada dalam satu klaster yang sama lebih minim dan variasi dengan data yang berada pada klaster lainnya lebih maksimal (Syarief, 2018). Berikut langkah-langkah melakukan pengklasteran dengan metode *K-Means* yaitu:

1. Menentukan jumlah klaster k.
2. Menentukan nilai pusat awal setiap klaster (centroid) secara acak.
3. Menghitung jarak setiap data ke *centroid* dengan menggunakan jarak *Euclid* dan *Manhattan.*
4. Mengklasifikasikan data dengan jarak terkecil kepada kelompok terdekatnya.
5. Memperbarui centroid dari titik tengah klaster bersangkutan menggunakan rumus:

   $$\mu_k = \frac{1}{n_k} \sum_{t=1}^{n_k} x_t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

   dengan n_k adalah jumlah data dalam klaster k.
6. Lakukan kembali langkah ketiga sampai kelima hingga tidak ada anggota klaster yang berpindah.
7. Apabila langkah enam terpenuhi, *centroid* (μ) pada iterasi terakhir digunakan sebagai parameter untuk pengklasifikasian data.

2.9 *K-Medoids*

K-Medoids adalah penyempurnaan dari algoritma *K-Means* pada pengklasteran yang lebih efektif dalam menangani dataset yang mengandung *outliers*. Algoritma *K-Medoids* menetapkan pusat klaster berdasarkan representasi objek klaster yang disebut sebagai medoid. Medoid adalah objek klaster yang terletak paling sentral, dengan total jarak minimum ke titik lainnya, (Orisa & Faisol, 2021).

Keunggulan dari metode ini adalah kemampuannya untuk menangani kelemahan metode *K-Means* yang sensitif terhadap *outlier*. Selain itu, kelebihan lainnya adalah hasil dari proses pengklasteran tidak tergantung pada urutan masukan data dalam dataset.

Algoritma *K-Medoids Clustering* adalah sebagai berikut:

1. Tentukan k (jumlah klaster) yang diinginkan
2. Pilih secara acak medoid awal sebanyak k dari n data
3. Hitung jarak masing-masing objek ke medoid sementara menggunakan rumus jarak *Euclid* dan *Manhattan.* kemudian tandai jarak terdekat objek ke medoid dan hitung totalnya
4. Lakukan iterasi medoid.
5. Hitung total simpangan (S)
6. Ulangi langkah 3 sampai 5 dan hentikan jika sudah tidak terjadi perubahan anggota medoid.

2.10 *Silhouette Index*

Silhouette Index merupakan metode untuk memvalidasi kebaikan sebuah data klaster tunggal atau bahkan keseluruhan klaster. Metode ini digunakan untuk memvalidasi klaster yang menggabungkan nilai kohesi dan separasi. Persamaan *Silhouette Index* dapat ditulis seperti "Persamaan (5) dan (6)", (Nicolaus, Sulistianingsih, & Perdana, 2016). Nilai *Silhouette Index* memiliki rentang -1 hingga 1. Kriteria subjektif pengelompokan berdasarkan *Silhouette Coefficient* (SC). Pada dasarnya nilai *Silhouette Coefficient* diperoleh dengan membandingkan jarak data pada klaster yang sama dengan jarak data di klaster lainnya (Syarief, 2018).

2.11 *Davies-Bouldin Index (DBI)*

Pengukuran *Davies-Bouldin Index* (DBI) digunakan untuk mengevaluasi klaster. Validasi internal yang dilakukannya adalah seberapa baik pengklasteran sudah dilakukan dengan menghitung kuantitas dan fitur turunan dari himpunan data. Semakin kecil nilai DBI (non-negatif), semakin baik klaster yang diperoleh. Rumus DBI dapat ditulis sebagai berikut (Norris S., 2018):

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} \left(\frac{s_i + s_j}{d_{ij}}\right) \quad \text{..... (8)}$$

dengan,

k : jumlah klaster

s_i: rerata jarak klaster ke-*i* dengan centroidnya

s_j: rerata jarak klaster ke-*j* dengan centroidnya

d_{ij}:jarak antara centroid klaster ke-i dengan centroid klaster ke-j.

2.12 *Connectivity Index*

Perhitungan *Connectivity Index* dimulai dengan menetapkan nilai konektivitas awal sebesar 0. Untuk setiap pola dalam dataset, ditentukan sejumlah tetangga terdekat dan diperiksa apakah tetangga tersebut berada dalam klaster yang sama. Jika data ke-*i* tidak berada dalam klaster yang sama, maka *connectivity* bernilai satu. Namun jika data ke-*i* berada pada klaster yang sama, maka *connectivity* bernilai nol.

$$Conn = \frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{L} X_{i,nn_i}(j) \quad \text{..... (9)}$$

dengan:

$Conn$: *connectivity index*

N : banyak pengamatan

L : jumlah tetangga terdekat yang diperiksa untuk setiap objek.

$nn_i(j)$:pengamatan tetangga terdekat dari objek j ke objek di-i

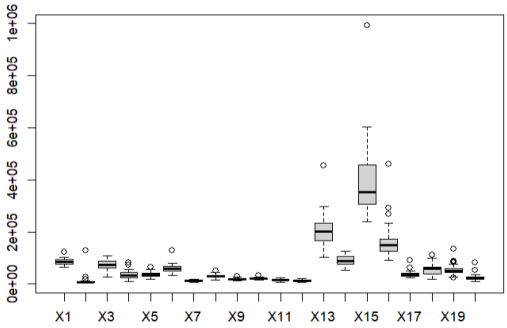
$X_{i,nn_i(j)}$:pengamatan tetangga terdekat (*nearest neighbor*) dari data ke-j ke data ke-i, jika dalam satu klaster bernilai 0 (nol) dan jika berbeda, bernilai 1.

Semakin kecil nilai *Indeks Connectivity* menunjukkan banyak klaster yang terbentuk lebih baik atau optimal, (Halim & Widodo, 2017).

3. HASIL DAN PEMBAHASAN

Hasil deteksi *outlier* dengan *boxplot* adalah seperti berikut.

P - ISSN: 1979-911X
E - ISSN: 2541-528X

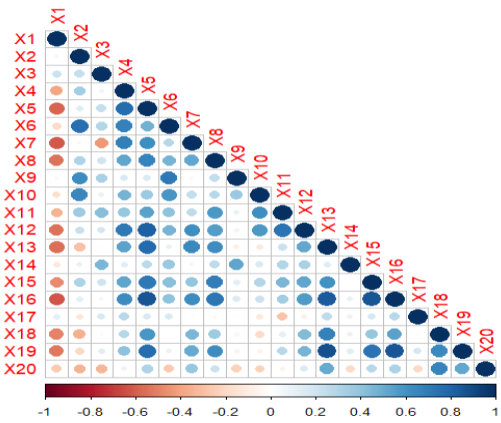


Gambar 1. Hasil Deteksi *Outlier* Menggunakan *Boxplot*

Secara umum data *outlier* harus ditangani yaitu dikeluarkan atau tetap digunakan dengan melakukan transformasi data karena dapat membuat analisis menjadi bias. Dalam penelitian ini penanganan dilakukan dengan transformasi. Hal ini karena jika dikeluarkan dapat menghilangkan informasi penting mengenai hasil analisis pengelompokan.

3.1 Uji Multikolinieritas

Berdasarkan uji multikolinieritas, terdapat beberapa variabel yang memiliki nilai korelasi > 0,8 sehingga terjadi multikolinieritas. Permasalahan multikolinieritas ini ditangani menggunakan *Principal Component Analysis* (PCA). Namun sebelumnya perlu dilakukan pemeriksaan nilai MSA atau *Measure of Sampling Adequacy*. Pemeriksaan ini dilakukan untuk mengetahui variabel apa saja yang layak dilakukan analisis komponen utama.



Gambar 2. Hasil Korelasi Antar Vriabel Bebas

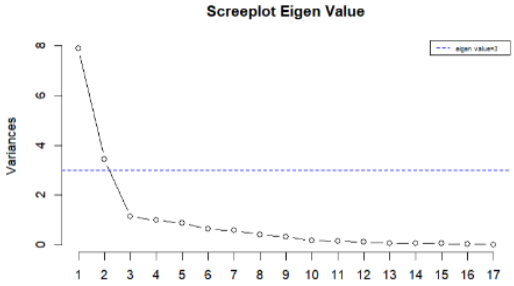
3.2 Uji Kecukupan Sampel

Uji kecukupan data dilakukan dengan uji KMO (*Kaiser-Mayer Olkin*). Nilai uji KMO antara 0 sampai 1. Jika nilai KMO ≥ 0,5 maka sampel dapat dikatakan mewakili populasi dan layak untuk dilakukan analisis lanjutan. KMO atau *Overall* MSA yang diperoleh 0,63 ≥ 0,5 sehingga data yang digunakan layak untuk dianalisis lebih lanjut. Dari perhitungan MSA masing-masing variabel diketahui bahwa variabel Ikan/ Udang/ Cumi-cumi/ Kepiting (X3), Minyak Dan Kelapa (X9), dan variabel Pakaian, Alas Kaki dan Penutup Kepala (X17) tidak layak diikutkan dalam analisis PCA karena memiliki nilai MSA < 0,5, sehingga variabel tersebut tidak diikutkan dalam analisis PCA.

3.3 Analisis Komponen Utama (PCA)

PCA merupakan metode yang mampu mereduksi dimensi data yang besar dan saling berkorelasi menjadi dimensi data yang lebih kecil dan tidak saling berkorelasi tanpa kehilangan banyak informasi. dengan menggunakan konsep nilai *eigen* dan vektor *eigen.* Penentuan menggunakan *screeplot eigen value* seperti pada gambar berikut.

P - ISSN: 1979-911X
E - ISSN: 2541-528X

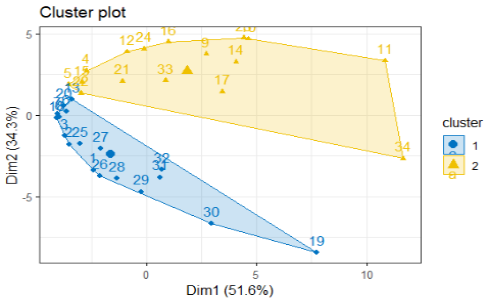


Gambar 3. *Screeplot eigen value*

Berdasarkan Gambar 3 diketahui bahwa titik yang membentuk siku berada pada variansi ke-3, sehingga komponen 1, 2 dan 3 dapat menjelaskan keragaman dan digunakan untuk analisis klaster. Hasil perhitungan ketiga komponen PC1, PC2 dan PC3 digunakan untuk analisis klaster dengan metode *K-Means* dan *K-Medoids*.

3.4 *K-Means* Menggunakan Jarak Euclid

Hasil pengklasteran dengan menggunakan jarak Euclid pada *K-Means* terlihat seperti gambar berikut.



Gambar 4. *Plot Clustering K-Means k=2*

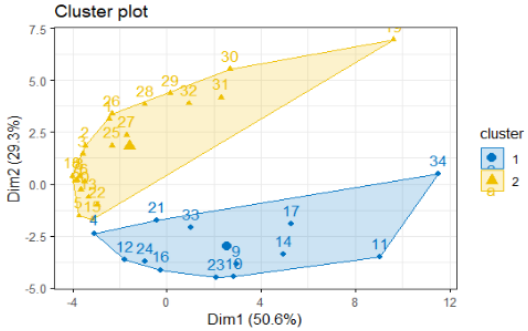
Gambar 6 menunjukkan klaster pertama dilambangkan dengan warna merah dengan jumlah anggota sebanyak 18 anggota dan klaster kedua dilambangkan dengan warna biru dengan jumlah anggota sebanyak 16 anggota. Hasil uji validasi dalam "Tabel 2" diperoleh *Silhouette Index* (SI) sebesar 0,42, *Davies-Bouldin Index* (DBI) sebesar 1,179, dan *Connectivity Index* (CI) sebesar 9,347.

Tabel 1. Hasil Uji Validitas

Jumlah Klaster	SI	DBI	CI
$k = 2$	0,42	1,179	9,347

3.5 *K-Means* Menggunakan Jarak *Manhattan*

Hasil *K-Means clustering* dengan menggunakan jarak Manhattan diberikan oleh gambar berikut:



Gambar 5. Plot *K-Means Clustering* dengan *k=2*

Gambar 9 menunjukkan klaster pertama dilambangkan dengan warna merah dengan jumlah anggota sebanyak 13 anggota dan klaster kedua dilambangkan dengan warna biru dengan jumlah anggota sebanyak 21 anggota. Hasil uji validitas *Silhouette Index* (SI) bernilai 0,38 , *Davies-Bouldin Index* (DBI) bernilai 1,179, dan *Connectivity Index* (CI) bernilai 10,931 sebagaimana ditunjukkan dalam Tabel 3.

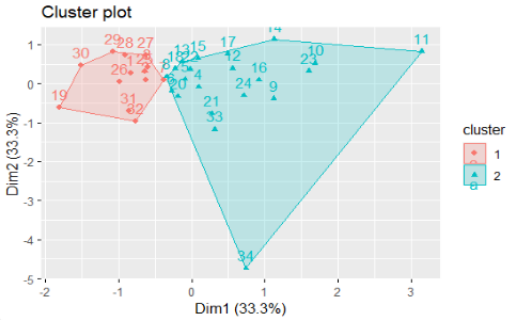
P - ISSN: 1979-911X
E - ISSN: 2541-528X

Tabel 2. Hasil Uji Validitas *K-Means* dengan Jarak Manhattan

Jumlah Klaster	SI	DBI	CI
$k = 2$	0,38	1,179	10,931

3.6 K-Medoids dengan Jarak Euclid

Hasil pengklasteran dengan menggunakan jarak Euclid pada *K-Medoids* diberikan oleh gambar berikut:



Gambar 6. *Plot Clustering K-Medoids k=2*

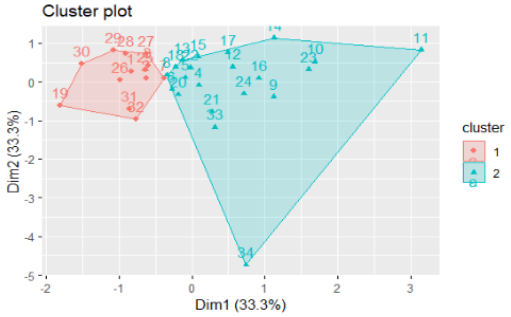
Gambar 10 menunjukkan klaster pertama dilambangkan dengan warna merah dengan jumlah anggota sebanyak 22 anggota dan klaster kedua dilambangkan dengan warna biru dengan jumlah anggota sebanyak 12 anggota. Hasil uji validitas *Silhouette Index* (SI) bernilai 0,42, *Davies-Bouldin Index* (DBI) bernilai 1,179, *Connectivity Index* (CI) bernilai 9,347 sebagaimana terlihat dalam Tabel 4.

Tabel 3. Hasil Uji Validitas *K-Medoids* Jarak *Euclid*

Jumlah Klaster	SI	DBI	CI
$k = 2$	0,42	1,179	9,347

3.7 *K-Medoids* dengan Jarak Manhattan

Pengklasteran dengan menggunakan jarak Manhattan pada *K-Medoids*:



Gambar 7. *Plot Clustering K-Medoids k=2*

Gambar 11 menunjukkan klaster pertama dilambangkan dengan warna merah dengan jumlah anggota sebanyak 21 anggota dan klaster kedua dilambangkan dengan warna biru dengan jumlah anggota sebanyak 13 anggota. Hasil uji validitas *Silhouette Index* (SI) bernilai 0,3, *Davies-Bouldin Index* (DBI) bernilai 1,149, dan *Connectivity Index* (CI) bernilai 5,963 seperti terlihat pada Tabel 5.

Tabel 4. Hasil Uji Validitas *K-Medoids* Jarak *Manhattan*

Jumlah Klaster	SI	DBI	CI
$k = 2$	0,3	1,149	5,963

3.8 Perbandingan *K-Means* dan *K-Medoids*

Perbandingan ini dilakukan untuk mengetahui metode mana yang paling efektif digunakan untuk menganalisis data komponen pengeluaran per kapita di Indonesia tahun 2023. Dari Tabel 6 diketahui bahwa dengan jarak Euclid, pada uji validasi *Silhouette Index,* DBI dan juga *Connectivity Index* tidak dapat disimpulkan metode mana yang paling terbaik karena menghasilkan nilai yang sama persis.

Tabel 5. Hasil Perbandingan *K-Means dan K-Medoids* dengan Jarak Euclid

Metode	jumlah kelompok	SI	DBI	CI
K-Means	$k = 2$	0,42	1,179	9,347
K-Medoids	$k = 2$	0,42	1,179	9,347

Sementara itu Tabel 7 menunjukkan hasil uji validasi pengklasteran dengan jarak Manhattan. Berdasarkan nilai *Silhouette Index, DBI* dan *Connectivity Indeks*, metode *K-Medoids* memiliki nilai DBI dan juga CI lebih kecil dibandingkan dengan metode K-Means, sehingga metode *K-Medoids* lebih baik untuk digunakan.

Tabel 6. Hasil Perbandingan ***K-Means dan K-Medoids*** dengan Jarak Manhattan

Metode	jumlah kelompok	SI	DBI	CI
K-Means	$k = 2$	0,38	1,179	10,931
K-Medoids	$k = 2$	0,3	1,149	5,963

Salah satu alasan bahwa metode *K-Medoids* memberikan hasil yang lebih baik dalam penelitian ini karena data yang digunakan memiliki *outlier* dan sifat dari *K-Medoids* lebih tahan terhadap *outlier* dibandingkan dengan metode *K-Medoids.*

3.9 *Profiling*

Tahap *profiling* meliputi penggambaran karakteristik masing-masing klaster untuk menjelaskan bagaimana setiap variabel relevan pada tiap dimensi. Karakteristik pada klaster dilihat berdasarkan nilai rata-rata setiap variabel pada masing-masing klaster. Namun dikarenakan metode yang lebih baik adalah *K-Medoids* maka *profiling* klaster dilakukan dengan melihat karakteristik berdasarkan nilai median setiap variabel pada masing-masing klaster.

Tabel 7. *Profiling* Klaster Terbaik

	PC1	PC2	PC3
K=1	-1,495	-0,143	-0,243
K=2	2,302	0,175	-0,011

Berdasarkan Tabel 8 dapat diketahui bahwa pada klaster 2 memiliki nilai median PC1, PC2 dan PC3 yang lebih tinggi dibandingkan klaster 1.

Tabel 8. Hasil *Profiling* Klaster Terbaik dengan Data Asli

	x1	x2	x3	x4	x5	x6	x7
k=1	89896	6603	70700	26851	32053	57778	9992
k=2	79887	8516	89540	46950	43493	68279	14180
	x8	x9	x10	x11	x12	x13	x14
k=1	29262	18506	19688	13776	10860	177275	85078
k=2	38010	19403	21617	18878	16908	266887	91207
	x15	x16	x17	x18	x19	x20	
k=1	326626	137402	33405	54497	48969	22073	
k=2	486471	197226	39157	67973	68711	23938	

Berdasarkan Tabel 9 dapat diketahui bahwa pada klaster 1 memiliki nilai median yang lebih tinggi dibandingkan klaster 2 pada variabel padi-padian (x1). Sedangkan pada klaster 2 memiliki nilai median yang lebih tinggi dibandingkan klaster 1 pada variabel umbi-umbian (x2), ikan/udang/cumi-cumi/ kerang (x3), daging (x4), telur dan susu (x5), sayur-sayuran (x6), kacang-kacangan (x7), buah-buahan (x8), minyak dan kelapa (x9), bahan minuman (x10), bumbu-bumbuan (x11), konsumsi lainnya (x12), makanan dan minuman jadi (x13), rokok (x14), perumahan dan fasilitas rumah tangga (x15), aneka barang dan jasa (x16), pakaian, alas kaki dan tutup kepala (x17), barang tahan lama(x18), pajak, pungutan dan asuransi (x19) dan keperluan pesta/kenduri (x20) yang lebih tinggi dibandingkan klaster.

4. KESIMPULAN

Berdasarkan hasil analisis dan pembahasan yang telah dilakukan maka dapat disimpulkan bahwa:

a. Hasil analisis diketahui bahwa rata-rata pengeluaran per kapita paling besar adalah variabel perumahan dan fasilitas rumah tangga (x15) pada provinsi DKI Jakarta sebesar 992.514 rupiah, sedangkan rata-rata pengeluaran per kapita per bulan (dalam rupiah) paling kecil yaitu variabel pajak, pungutan dan asuransi (x19) pada provinsi Sulawesi Barat sebesar 4.759 rupiah. Dari hasil analisis didapatkan bahwa terdapat beberapa variabel yang mengalami multikolinieritas sehingga dilakukan uji PCA. Karena ada 3 variabel yang memiliki nilai MSA < 0,5 maka variabel tersebut tidak disertakan dalam analisis PCA. Selanjutnya nilai dari PCA digunakan sebagai data yang dianalisis dalam proses pengklasteran.
b. *K-Means Clustering* menggunakan jarak Euclid dan Manhattan menghasilkan jumlah klaster optimal sebesar 2. Untuk jarak Euclid, klaster 1 berjumlah 18 provinsi dan klaster 2 berjumlah 16 provinsi. Sedangkan untuk jarak Manhattan, klaster 1 berjumlah 13 provinsi dan klaster 2 berjumlah 21 provinsi.
c. *K-Medoids Clustering* menggunakan jarak Euclid dan Manhattan juga menghasilkan k optimal 2. Klaster 1 memiliki median pada variabel padi-padian (x1) yang lebih tinggi dibandingkan klaster 2. Klaster 2 memiliki median pada variabel umbi-umbian (x2), ikan/udang/cumi-cumi/kerang (x3), daging (x4), telur dan susu (x5), sayur-sayuran (x6), kacang-kacangan (x7), buah-buahan (x8), minyak dan kelapa (x9), bahan minuman (x10), bumbu-bumbuan (x11), konsumsi lainnya (x12), makanan dan minuman jadi (x13), rokok (x14), perumahan dan fasilitas rumah tangga (x15), aneka barang dan jasa (x16), pakaian, alas kaki dan tutup kepala (x17), barang tahan lama (x18), pajak, pungutan dan asuransi (x19) dan keperluan pesta/ kenduri (x20) yang lebih tinggi dibandingkan klaster 1.
d. Berdasarkan hasil uji validitas, pengklasteran dengan jarak Manhattan lebih baik dibandingkan dengan pengklasteran menggunakan jarak Euclid.
e. Perbandingan metode *K-Means* dan *K-Medoids* berdasarkan hasil *Silhouette Index, DBI* dan juga *Connectivity Index* dapat diketahui bahwa metode yang paling baik digunakan dalam analisis ini adalah metode *K-Medoids* dengan menggunakan jarak Manhattan karena memiliki nilai DBI dan juga CI paling kecil dibandingkan dengan metode lainnya. Dengan demikian metode *K-Medoids* adalah metode yang lebih. Hal ini karena data yang digunakan memiliki *outlier* dan sifat dari *K-Medoids* lebih tahan terhadap *outlier* dibandingkan dengan metode *K-Medoids.*

Berdasarkan hasil pembahasan dan kesimpulan yang telah dilakukan maka peneliti dapat memberikan saran bahwa metode yang paling baik digunakan adalah metode *K-Medoids* karena lebih tahan terhadap *outlier.* Namun berdasarkan hasil uji validasi *silhouette index* menunjukkan hasil < 0,5 yang artinya hasil masih kurang baik. Oleh karena itu perlu dicari metode pengklasteran lain yang lebih tahan terhadap *outlier.*

UCAPAN TERIMA KASIH

Terima kasih penulis ucapkan kepada Badan Pusat Statistik (BPS) dan Universitas AKPRIND Indonesia terhadap dukungan data dan fasilitas yang telah diberikan selama penelitian sehingga penelitian ini dapat diselesaikan dengan baik.

DAFTAR PUSTAKA

Bhandari, P. (2021). *How to Find Outliers, 4 Ways with Examples & Explanation*. Diakses 20 Mei 2024 dari https://www.scribbr.com/statistics/outliers/

Cahaya, Y. M. (2023). Perbandingan Metode Perhitungan Jarak Euclidean dengan Perhitungan Jarak Manhattan pada K-Means Clustering dalam Menentukan Penyebaran Covid di Kota Bekasi. *Jurnal Matematika Terapan*, 43-55.

Farissa, R. A., Mayasari, R., & Umaidah, Y. (2021). Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Data Obat dengan Silhouette Coefficient . *JAIC*, Vol.5, No.2, Desember 2021, pp. 109~116.

Ghozali, I. (2018). *Aplikasi Analisis Multivariate dengan Program IBM SPSS 25*. Universitas Diponegoro.

Halim, N. N., & Widodo, E. (2017). Pengklasteran Dampak Gempa Bumi di Indonesia Menggunakan Kohonen Self Organizing Maps. Prosiding *SI MaNIs*, Vol.1, No.1, Juli 2017, Hal. 188-194.

Jamal, A., Handayani, A., Septiandri, A. A., & Ripmiatin, E. (2018). Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction. *Lontar Komputer*, Vol. 9, No. 3.

Nabilah, N. A., Perdana, H., & Sulistianingsih, E. (2024). Pengelompokan Provinsi Di Indonesia Berdasarkan Indikator Kesejahteraan Masyarakat dengan Algoritma K-Means++. *Bimaster*, Volume 13, No. 3 (2024), hal 419 – 426.

Nicolaus, Sulistianingsih, E., & Perdana, H. (2016). Penentuan Jumlah Cluster Optimal pada Median Linkage dengan Indeks Validitas Silhouette. *Bimaster*, 97-102.

Norris, S.(2018) *Assessment Metric for Clustering Algorithm*.

Orisa, M., & Faisol, A. (2021). Analisis Algoritma Partitioning Around Medoid untuk Penentuan Klasterisasi. *J-TIT*, Vol 8, No 2.

Ramadhayani, A. N., & Lusiana, V. (2022). Klasifikasi Jenis Kucing Menggunakan Algoritma Principal Component Analysis dan K-Nearest Neighbor. *Jurnal Informasi dan Komputer*, Vol: 10 No:2.2022.

Samudro, A.B.P. (2023). *Pengeluaran untuk Konsumsi Penduduk Indonesia Per Provinsi Maret 2023*, Badan Pusat Statistik RI.

Supriyadi, A., Triayudi, A., & Sholihati, I. D. (2021). Perbandingan Algoritma K-Means dan K-Medoids pada Pengelompokan Armada Kendaraan Truk Berdasarkan Produktivitas. *JIPI*, Volume 06, Nomor 02, Desember 2021 : 229 – 240.

Syarif, R. M. (2018). Perbandingan Algoritme K-Means dengan Algoritme Fuzzy C Means (FCM) dalam Clustering Moda Transportasi Berbasis GPS. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*.