ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 12, Nomor 2, Tahun 2023, Halaman 178 - 188
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

ANALISIS *k-MEDOIDS* DENGAN VALIDASI INDEKS PADA IPM DAERAH 3T DI INDONESIA

Maria Dafrosa Doi [1], Agus Rusgiyono [2], Triastuti Wuryandari [3]
[1,2,3] Jurusan Statistika Fakultas Sains dan Matematika Universitas Diponegoro
***e-mail:** dafrosadoi@gmail.com

DOI: 10.14710/j.gauss.12.2.178-188

Article Info:
Received: 2022-12-06
Accepted: 2023-02-15
Available Online: 2023-07-28

Keywords:
Human Development Index, k-Medoids, Euclidean Distance, Manhattan Distance, Cluster Validation

Abstract: Human development is a development paradigm that places humans as the main target of all development activities, namely controling over resources, improving health and improving education. The Human Development Index (HDI) in Indonesia varies in each district, especially in the 3T areas. The 3T area is an area that is classified as underdeveloped, remote and outermost in terms of economy, health, education and infrastructure. The k-Medoids method is a partitional clustering method for grouping several objects into clusters. This clustering algorithm uses the medoid as the center of the cluster, so it is robust to data containing outliers. This study aims to classify the 3T regions in Indonesia based on the Human Development Index to find out which areas require more attention from the government in optimizing the Human Development Index numbers. The size of object similarity is calculated by using the Euclidean distance and Manhattan distance, for the selection of the best number of clusters, internal cluster validation, such as Calinski – Harabasz index, Gamma Index, and Silhouette index. The result of this study showed that the best cluster were four by using Euclidean distance measurement, having Calinski – Harabasz index score of 37.15764, Gamma index score of 0.7821181, and Silhouette index score of 0.3354435.

1. PENDAHULUAN

Indonesia adalah negara berkembang dimana penentuan pembangunan merupakan salah satu landasan perubahan. Pembangunan adalah proses perubahan yang disengaja dengan tujuan untuk mengubah kehidupan masyarakat menjadi lebih baik. Manusia adalah insan yang bisa membantu kemajuan, pergeseran ini dimulai dengan perkembangan manusia. Paradigma pembangunan yang dikenal dengan "pembangunan manusia" memandang manusia sebagai tujuan utama dari semua upaya pembangunan, termasuk menguasai sumber daya alam, meningkatkan kesehatan, dan meningkatkan standar pendidikan. *United Nations Development Programme* awalnya menetapkan konsep pengukuran pembangunan manusia pada tahun 1990. Sebuah konsep baru dalam pengukuran manusia, diperkenalkan oleh UNDP adalah Indeks Pembangunan Manusia yang bisa disingkat dengan IPM, (BPS, 2021).

Pertumbuhan IPM Indonesia menurut Badan Pusat Statistik (BPS) konsisten lebih besar dari 0,7%, bahkan sebelum pandemi COVID-19. IPM terus meningkat selama pandemi, naik dari 71,92 pada 2019 menjadi 71,94 pada 2020, kemudian naik lagi menjadi 72,29 pada 2021. Kondisi pembangunan manusia kabupaten dan kota di Indonsia sangat bervariasi pada tahun 2021. Kota Yogyakarta, memiliki IPM terbesar dengan capaian indeks 87,18. Kabupaten Nduga memiliki capaian IPM terendah, dengan indeks 32,84. Terdapat 250 kabupaten dan kota berstatus IPM "sedang" pada tahun 2021, 22 kabupaten/kota berstatus "rendah" (4,28%), 204 kabupaten/kota berstatus "tinggi" (36,69%), dan 38 kabupaten/kota berstatus " status IPM sangat tinggi" (7,39%). (BPS, 2021). Presiden Joko Widodo menetapkan identifikasi daerah tertinggal periode 2020 hingga 2024 dalam

Peraturan Presiden (Perpes) Nomor 63 Tahun 2020. Di Indonesia, terdapat 62 daerah 3T atau sering disebut daerah tertinggal yang tersebar di beberapa provinsi. Daerah 3T merupakan daerah yang tergolong tertinggal, terpencil, dan terluar segi ekonomi, kesehatan, infrastruktur, dan pendidikan.

Bidang statistik menggunakan analisis klaster sebagai salah satu metode analisisnya untuk mengklasifikasikan objek. Metode *k-Means* dan metode *k-Medoids*, merupakan pendekatan yang digunakan dalam analisis klaster. Han dan Kamber (2006) menegaskan bahwa pendekatan *k-Medoids* lebih unggul dari metode *k-Means* karena dapat digunakan untuk mengelompokkan objek yang mengandung *outlier*. Pengukuran jarak dapat digunakan untuk mengukur seberapa dekat sifat dua item yang mirip satu sama lain. Semakin dekat objek dengan pusat cluster jarak yang diperoleh semakin kecil, dan semakin besar jarak yang diperoleh semakin jauh letak objek dengan pusat klaster. Jarak Manhattan dan jarak Euclidean adalah satuan jarak yang digunakan.

Validasi klaster merupakan faktor yang harus diperhitungkan dalam analisis klaster untuk mengidentifikasi klaster yang terbaik. Validasi internal adalah metode validasi cluster yang digunakan. Untuk menetapkan jumlah klaster/kelompok yang terbaik, penulis akan menggunakan pendekatan *k-Medoids* untuk mengklasifikasikan daerah 3T Indonesia berdasarkan Indeks Pembangunan Manusia (IPM) dan validasi indeks internal dengan menggunakan Calinski-Harabasz *index*, *Gamma index*, dan *Silhouette index* di RStudio 4.2.1.

2. TINJAUAN PUSTAKA

Pembangunan manusia memandang manusia bukan hanya sebagai input, melainkan tujuan akhir dari pembangunan tersebut. Pembangunan ini dengan tujuan adalah untuk menciptakan kondisi dimana masyarakat dapat hidup lama, sehat, dan produktif (UNDP, 1990). IPM adalah metrik penting untuk melihat sisi pembangunan. Setiap indikator komponen IPM datap digunakan untuk menilai kemajuan suatu masyarakat atau penduduk dalam meningkatkan kualitas hidup. Beberapa indikator komponen IPM yang digunakan pada penelitian ini yaitu: Usia Harapan hidup (UHH), Harapan Lama Sekolah (HLS), Rata-rata Lama Sekolah (RLS), serta Pendapatan per kapita yang disesuaikan (PPD).

Untuk menghindari masalah yang disebabkan oleh penggunaan satuan data yang berbeda, variabel penelitian harus dibakukan jika memungkinkan. Untuk mengonversi setiap peubah ke nilai standarnya (Z *score*), bagi nilai rata-rata dengan nilai standar deviasi atau simpangan baku. Rumus standarisasi untuk setiap variabel menurut Walpole dan Mayes (1995) sebagai berikut:

$$Z_i = \frac{x_i - \bar{x}}{s}, \; i = 1, 2, 3, \dots, n \qquad (2.1)$$

dengan, Z_i = variabel standarisasi, x_i = data ke-*i*, $\bar{x}$ = nilai mean semua data untuk masing-masing variabel, s = standar deviasi.

Data *outlier* adalah data yang menyimpang dari pola khusus dan tersebar jauh dari pusat data. *Outlier*, menurut Sembiring (1995), adalah titik data yang menyimpang dari pola tipikal model. Data yang *outlier* mengungkapkan anomali yang tidak sesuai dengan data lainnya. Hair *et al*., (2010) menyatakan bahwa keberadaan *outlier* dapat mengakibatkan hasil analisis yang tidak akurat sehingga tidak mencerminkan kondisi populasi secara akurat. Dengan mengukur jarak antara setiap nilai pengamatan dan pusat data menggunakan kuadrat

jarak Mahalanobis. Berikut adalah persamaan yang dapat digunakan untuk menentukan kuadrat jarak Mahalanobis antara pusat data dan objek ke-*i*:

$$d_{MD}^2(i) = (\boldsymbol{x_i} - \bar{\boldsymbol{x}})^{\boldsymbol{T}} \boldsymbol{\Sigma^{-1}} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}), i = 1,2, \dots, n \qquad (2.2)$$

dengan, $d_{MD}^2(i)$ = kuadrat jarak Mahanabis objek ke-*i* dengan pusat data, $\boldsymbol{x_i}$ = vektor data objek variabel ke-*i*, $\bar{\boldsymbol{x}}$ = vektor rata-rata setiap variat, dan $\boldsymbol{\Sigma^{-1}}$ = matriks kovarians variat. Kuadrat jarak mahalanobis (d_{MD}^2) kemudian dievaluasi dengan *distribusi chi-kuadrat* $(\chi_{\alpha,p}^2)$di mana *p* adalah jumlah variabel observasi. Apabila suatu data memiliki nilai (d_{MD}^2) > *chi-kuadrat* ($\chi_{\alpha,p}^2$) maka data tersebut diidentifikasi sebagai data *outlier* .

Teknik analisis multivariat seperti analisis klaster sebagian besar digunakan untuk mengkategorikan hal-hal menurut sifatnya. Teknik mengklasifikasikan objek ke dalam kelompok-kelompok yang pada dasarnya sama (homogen) disebut analisis klaster. Objek di setiap klaster memiliki kecenderungan untuk menyerupai satu sama lain dan berbeda dari hal-hal di klaster lain. Selain itu, tidak ada tumpang tindih atau interaksi karena setiap objek hanya dimiliki oleh satu kelompok (Supranto, 2004).

Ada dua asumsi yang perlu dipenuhi dalam analisis klaster, menurut Hair *et al* ., (2010) adalah:

1. Kecukupan Sampel (Sampel Representatif)

Sampel yang akurat dapat mewakili seluruh populasi dikatakan sebagai sampel representatif. *Kaiser-Mayer-Olkin* (KMO) dapat digunakan untuk menguji sampel yang representatif. Dengan uji KMO ini, kecukupan sampel dari setiap indikator serta ukuran sampel secara keseluruhan akan dievaluasi. Jika nilai KMO berada di antara 0,5 dan 1, maka sampel dianggap representatif atau dapat mencerminkan populasi. Berikut adalah rumus KMO (Widarjono, 2010):

$$KMO = \frac{\sum_{s=1}^{p} \sum_{t=1,k\neq j}^{p} r_{st}^2}{\sum_{s=1}^{p} \sum_{t=1}^{p} r_{st}^2 + \sum_{s,t=1}^{p} \sum_{r=1}^{p} \rho_{s(tu)}^2} \qquad (2.2)$$

dengan, p = banyaknya variabel, n = banyaknya objek, r_{st} = koefisien korelasi antara variabel ke $-s$ dan variabel ke $-t$, $\rho_{s(tu)}^2$ = koefisien korelasi parsial antara variabel ke $-s$, ke $-t$ dan ke $-u$

2. Tanpa Multikolinearitas (Non-Multikolinearitas)

Multikolinearitas menurut Gujarati (2009), didefinisikan sebagai adanya hubungan linier yang sempurna antara beberapa atau seluruh peubah penelitian. Untuk mengetahui multikolinearitas adalah dengan menggunakan nilai *Variance Inflation Factor* (VIF), yang bisa dihitung dengan rumus:

$$VIF_j = \frac{1}{1 - R_j^2} \qquad (2.4)$$

dengan : R_j^2 = koefisien determinasi variabel ke- j . Dalam kasus pemodelan regresi linier, R_j^2diperoleh dengan meregresikan variabel independen j dengan variabel dependen. Sedangkan pada analisis klaster R_j^2 dapat diperoleh dengan meregresikan variabel ke- j

dengan variabel lainnya. Jika nilainya $VIF \geq 10$, maka dapat dikatakan terjadi multikolinieritas pada suatu data.

Jarak Euclidean dan jarak Manhattan digunakan dalam penelitian ini untuk membandingkan hasil klaster dengan berbagai jarak.

1. Jarak Euclidean

Jarak Euclidean antar objek dihitung dengan mengambil akar kuadrat dari penjumlahan kuadrat selisih nilai setiap peubah pada item (Andreberg, 1973). Rumus jarak Euclidean adalah sebagai berikut:

$$d_{(x_i.C_k)} = \sqrt{\sum_{j=1}^{p}(x_{ij} - C_{kj})^2}, i = 1,2,3 \dots, n;\ j = 1,2,3,..,n \quad (2.5)$$

$$k = 1,2,3, \dots, p$$

dengan, x_{ij} = objek pada pengamatan ke $- i$ pada variabel ke $- j$, C_{kj} = pusat klaster ke $- k$ pada variabel ke $- j$, p = banyaknya variabel yang diteliti.

2. Jarak Manhattan

Jarak Manhattan mengukur jarak objek dengan menghitung jumlah mutlak perbedaan objek pada setiap variabel (Andreberg, 1973). Jarak manhattan dirumuskan sebagai berikut:

$$d_{(x_i.C_k)} = \sum_{j=1}^{p}\left|x_{ij} - C_{kj}\right|,\ i = 1,2,3, \dots, n\ ; j = 1,2,3, \dots, n \quad (2.6)$$

$$k = 1,2,3, \dots, p$$

dengan, x_{ij} = objek pada pengamatan ke $- i$ pada variabel ke $- j$, C_{kj} = pusat klaster ke $- k$ pada variabel ke $- j$, p = banyaknya variabel yang diteliti.

Leonard Kauffman dan Peter J.Rousseuw menemukan pendekatan clustering partisi yang dikenal sebagai *k-Medoids*, dengan mengurangi jarak antara titik yang ditunjuk sebagai cluster dan titik yang ditunjuk sebagai pusat cluster. Metode *k-Medoids* menggunakan objek data *medoids* sebagai pusat data, sedangkan metode *k-means* menggunakan nilai *means* sebagai pusat cluster. (Kaur, et al., 2014).

k-Medoids merupakan algoritma clustering yang lebih tahan terhadap outlier dibandingkan dengan *k-Means* (Aggawal and Reddy, 2014). Mirip dengan *k-means* , Tujuan *k-Medoids* adalah untuk menemukan solusi pengelompokan *k*-klaster diantara semua objek data di dalam sebuah kelompok data. Algoritme *k-Medoids* memilih titik data aktual sebagai prototipe dan lebih kuat terhadap noise dan *outlier* dalam data. Untuk mengelompokkan objek ke dalam cluster, *algoritma k-medoids* menggunakan clustering partisi. Algoritma ini mewakili cluster dalam satu set objek menggunakan objek. Medoids adalah objek yang akan dipilih untuk mewakili cluster. Kedekatan *objek medoid* dan *non medoid* akan dihitung untuk membentuk cluster.

Algoritma dari *k-Medoids* menurut Han dan Kamber (2006), adalah sebagai berikut:

1. Tetapkan *k* sebagai jumlah cluster yang akan dibentuk

2. Memilih *k* objek secara acak dalam kumpulan *n* objek sebagai *medoid* awal
3. Alokasikan jarak antara objek non *medoid* dengan *medoid* awal dengan ukuran jarak euclidean dan jarak manhatttan, kemudian tandai jarak ke *medoid* terdekat, jumlahkan jarak terdekatnya
4. Pilih objek non *medoid* secara acak sebagai bakal calon *medoid* baru
5. Hitung jarak setiap objek non *medoid* dengan bakal calon *medoid* baru, lalu menandai jarak *medoid* terdekat, dan hitung total jarak terdekatnya
6. Hitung total jarak $S_{\text{total}}(\text{jarak})$, dengan $S_{\text{total}}(\text{jarak})$ =total jarak di *medoid* baru - total jarak di *medoid* lama.
7. Jika hasil $S_{\text{total}}(\text{jarak}) < 0$, maka jadikan kandidat *medoid* baru menjadi *medoid* baru, lalu lakukan iterasi dengan mengulangi langkah (4) sampai (7). Jika didapat $S_{\text{total}}(\text{jarak}) > 0$, maka iterasi berakhir.

Untuk menganalisis jumlah klaster yang dibuat dari metode klasterisasi yang digunakan untuk menghasilkan pengelompokan yang sesuai dengan data kajian, maka harus dilakukan validasi klaster. Setiap cluster yang dibuat memiliki ukuran yang berbeda untuk karakteristiknya, seperti nilai indeks validasi klaster (Brock, et al., 2008). Hasil klaster divalidasi menggunakan dua kriteria indeks dalam analisis klaster: validasi indeks internal dan validasi indeks eksternal. Validasi klaster internal mengevaluasi klaster menggunakan informasi internal yang terkandung dalam data penelitian, sedangkan validasi eksternal membandingkan hasil analisis klaster dengan hasil yang diketahui secara eksternal, seperti label kelas yang diberikan dari luar data yang digunakan (Nerukar et al., 2019). Karena belum diketahui label eksternal yang digunakan sebagai acuan validasi klaster eksternal pada penelitian ini, maka validasi internal digunakan untuk mengevaluasi hasil klaster yang terbentuk.Dalam penelitian ini, validasi klaster internal dilakukan dengan menggunakan Calinski-Harabasz *index, Gamma index,* dan *Silhouette index.*

1. Calinski-Harabasz *index*

Calinski-Harabasz *index* (CH *index*), menurut Baarsch dan Celebi (2012), menawarkan evaluasi hasil klaster berdasarkan perbandingan nilai jumlah kuadrat antar klaster (SSB) sebagai pemisahan dan nilai jumlah kuadrat dalam klaster (SSW) sebagai kekompakan dikalikan dengan faktor normalisasi yaitu selisih jumlah data dengan jumlah cluster dibagi banyaknya klaster dikurangi satu. Nilai klaster Calinski-Harabasz *index* harus lebih tinggi untuk hasil klaster yang lebih baik. Persamaan berikut dapat digunakan untuk mendapatkan nilai validitas Calinski-Harabasz *index*:

$$CH = \frac{tr\ (SSB)}{tr\ (SSW)} \times \frac{n-k}{k-1} \tag{2.7}$$

di mana *k* = banyaknya cluster, n = banyaknya objek yang diteliti, dan *trace (tr)* adalah jumlah elemen di sepanjang diagonal utama matriks.

2. *Gamma index*

Gamma index adalah validasi klaster internal yang diusulkan oleh Baker dan Hubert pada tahun 1975. Indeks ini merupakan gagasan dari Godman dan Kruskal pada tahun 1954. Persamaan berikut dapat digunakan untuk menentukan *Gamma index*:

$$G = \frac{S^+ - S^-}{S^+ + S^-} \tag{2.8}$$

dengan, S^+= banyaknya pasangan objek yang *concordant* , S^-= banyaknya pasangan benda yang *discordant* . Pasangan objek dikatakan *concordant* jika memenuhi *d(q, r) < d(s, t)* , objek *q* dan *r* berada pada cluster yang sama, sedangkan objek *s* dan *t* berada pada cluster yang berbeda atau *d(q, r) > d(s, t)*, objek *q* dan *r* berada dalam klaster yang berbeda, sedangkan objek *s* dan *t* berada dalam klaster yang sama. Pasangan objek dikatakan *discordant* jika memenuhi *d(q, r) < d(s, t)* , objek *q* dan *r* berada pada cluster yang berbeda, sedangkan objek *s* dan *t* berada dalam cluster yang serupa atau *d(q, r) > d(s, t*), objek *q* dan *r* berada dalam klaster yang sama, sedangkan objek *s* dan *t* berada dalam klaster yang berbeda. Nilai *Gamma index* berkisar antara rentang -1 sampai dengan 1, dengan nilai indeks maksimum yang diambil untuk mewakili jumlah cluster terbaik (Milligan and Cooper, (1985).

3. *Silhouette index*

Setiap titik dalam pengumpulan data dirata-ratakan dengan menggunakan indeks silhouette (Rousseeuw, 1987). Nilai pada masing-masing titik dihitung dengan cara mengurangkan nilai separasi dan kekompakan, kemudian membagi hasilnya dengan selisih maksimum antara keduanya. Rumus berikut digunakan untuk menentukan koefisien Silhouette:

$$SI = \frac{1}{k}\sum_{j=1}^{k} SI_j \qquad (2.9)$$

dimana, $SI_j = \frac{1}{m_j}\sum_{i=1}^{m_j} SI_i^j$,$SI_i^j = \frac{b_i^j - a_i^j}{max\left\{a_i^j; b_i^j\right\}}$ $a_i^j = \frac{1}{m_j - 1}\sum_{r=1, r\neq i}^{m_j} d(x_i^j, x_r^j)$,

$b_i^j min\left\{\frac{1}{m_l}\sum_{r=1, r\neq i}^{m_j} d\,(x_i^j, x_r^l)\right\}, l \neq j, l = 1, \ldots, k$, *SI= Indeks Silhouette* global, SI_j= *Indeks Silhouette* di cluster *j* , a_i^j= Rata-rata objek *i* dan semua objek lain pada satu cluster *j* ,

b_i^j= nilai mean minimum objek *i* dan seluruh data dari cluster lain *j* . Nilai *Silhouette Coefficient (SC)* dapat digunakan untuk menentukan apakah hasil cluster baik atau tidak . Nilai *SC* dihitung menggunakan persamaan berikut:

$$SC = max_x SI(k) \qquad (2.10)$$

dengan, *SC= Koefisien Silhouette* , $SI(k)$= *Indeks Silhouette dengan k* cluster . Semakin besar nilai *Silhouette Coefficient (SC)* , maka cluster tersebut semakin baik.

3. METODOLOGI PENELITIAN

Data yang digunakan dalam penelitian ini adalah data sekunder yang bersumber dari publikasi Badan Pusat Statistik tahun 2021. Indikator/variabel penelitiannya meliputi: UHH (X1), HLS (X2), RLS (X3), dan PPD (X4).

Pada penelitian ini data diolah dengan menggunakan RStudio 4.2.1 dan *Microsoft Excel*. Data penelitian dianalisis dengan tahapan sebagai berikut:

1. Data harus distandarisasi
2. Pendeteksian data *outlier* dengan kuadrat jarak mahalanobis
3. Melakukan uji non-multikolinearitas dengan nilai *Variance Inflation Factor (VIF)*.
4. Tentukan *k* sebagai banyaknya klaster yang akan dibentuk, *k* yang digunakan adalah *k* = 3, 4, 5, 6, dan 7
5. Mengunakan *algoritma k-Medoids* untuk melakukan analisis klaster
6. Validasi hasil klaster menggunakan indeks internal yaitu:

a. Menghitung nilai CH *index* , *Gamma index* , dan *Silhouette index* pada masing-masing *k* klaster
b. Membandingkan nilai CH *index*, *Gamma index* , dan *Silhouette index*. Nilai maksimum dari CH *index*, *Gamma index* , dan *Silhouette index* menunjukan hasil klaster terbaik

7. Melakukan interprestasi karakteristik daerah berdasrakan hasil pengklasteran yang terbaik

4. HASIL DAN PEMBAHASAN

Standarisasi dilakukan untuk membakukan satuan data penelitian. Hal ini dilakukan karena dalam data penelitian ini terdapat data dengan satuan yang berbeda, dimana variabel X_1 (dalam tahun), X_2 (dalam tahun), X_3 (dalam tahun), dan X_4 (dalam rupiah). Tabel 1 di bawah ini menunjukkan hasil standardisasi data:

Tabel 1 . Hasil Standardisasi Data

Objek ke-	X1	X2	X3	X4
1	1,5140	0,5591	-0,5263	0,2012
2	1,2441	0,3166	-0,3155	0,2318
3	1,4668	0,6441	0,0410	-0,3577
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
62	0,1146	-0,7254	-1,7262	-1,3438

Deteksi data outlier dengan kuadrat jarak mahalanobis, suatu objek teridentifikasi sebagai outlier jika kuadrat jarak mahalanobis (d_{MD}^2) > *chi-kuadrat* ($\chi_{\alpha,p}^2$). Penelitian ini menggunakan *p* sebanyak 4 variabel dengan nilai α yang dipakai adalah 0,05. Diperoleh nilai $\chi_{0,05;4}^2$ = 9,487729. Berdasarkan hasil pendeteksian kuadrat jarak mahanlanobis tiap objek dan membandingkannya dengan nilai *chi-squared* 9,487729 diketahui terdapat 2 kabupaten/kota yang *outlier* yaitu Kabupaten Nduga dan Kabupaten Intan Jaya. Kedua data outlier dalam hal ini tetap dimasukkan dalam analisis selanjutnya, karena metode *k-Medoid* merupakan pengelompokan data yang *robust* terhadap *outlier* sehingga tidak akan mempengaruhi hasil.

Analisis klaster memiliki dua asumsi yang harus dipenuhi, yaitu: asumsi kecukupan sampel (sampel representatif) dan asumsi non-multikolinearitas. Uji kecukupan sampel pada penelitian ini, karena data yang digunakan adalah data IPM Daerah 3T pada tahun 2021 yang termasuk data populasi. Perhitungan nilai *Variance Inflation Factor* (VIF) untuk setiap peubah yang digunakan dalam penelitian ini digunakan untuk menguji asumsi non multikolinearitas. Hasil dari VIF ditunjukkan pada Tabel 2 di bawah ini:

Tabel 2 . Nilai *Variance Inflation Factor* dari 4 Variabel

Variabel	VIF
X_1	1,061372
X_2	1,849360
X_3	2,561896
X_4	1,895932

Berdasarkan Tabel 2 terlihat bahwa dari variabel X_1 sampai dengan X_4 tidak terdapat variabel yang memiliki nilai VIF > 10, sehingga disimpulkan asumsi non multikolinearitas terpenuhi.

Analisis clustering menggunakan metode *k-Medoids* dengan k = 3, 4, 5, 6 dan 7. Diolah dengan *software* Rstudio 4.2.1, objek *medoid* dan jumlah anggota pada setiap cluster disajikan pada Tabel 3 berikut :

Tabel 3 . Hasil clustering dengan k = 3, 4, 5, 6 dan 7

k	Jarak	Klaster ke-	Jumlah Anggota	*Medoid*
3	Euclidean	1	27	Objek ke-35
		2	21	Objek ke-7
		3	14	Objek ke-50
	Manhattan	1	32	Objek ke-12
		2	16	Objek ke-27
		3	14	Objek ke-50
4	Euclidean	1	28	Objek ke-12
		2	17	Objek ke-7
		3	5	Objek ke-37
		4	12	Objek ke-50
	Manhattan	1	29	Objek ke-12
		2	15	Objek ke-27
		3	5	Objek ke-37
		4	13	Objek ke-50
⋮	⋮	⋮	⋮	⋮
7	Euclidean	1	9	Objek ke-18
		2	8	Objek ke-35
		3	16	Objek ke-31
		4	11	Objek ke-27
		5	5	Objek ke-37
		6	12	Objek ke-50
		7	1	Objek ke-55
	Manhattan	1	9	Objek ke-18
		2	12	Objek ke-35
		3	10	Objek ke-51
		4	13	Objek ke-27
		5	5	Objek ke-37
		6	12	Objek ke-50
		7	1	Objek ke-55

Nilai dari masing-masing indeks validasi dengan k = 3, 4, 5, 6, dan 7 dapat digunakan untuk menentukan jumlah cluster terbaik, pada Tabel 4 berikut:

Tabel 4 . Nilai Indeks Validasi Jumlah Cluster

k	Jarak	Indeks Validasi		
		Calinski-Harabazs index	Gamma index	Silhouette index
3	Euclidean	36,7833	0,6506	0,2869
	Manhattan	31,9834	0,6622	0,2794
4	**Euclidean**	**37,1576**	**0,7821**	**0,3354**

	Manhattan	31,9841	0,7643	0,3147
5	Euclidean	32,5158	0,7308	0,2715
	Manhattan	32,2467	0,7496	0,2947
6	Euclidean	30,1575	0,7248	0,2720
	Manhattan	29,6892	0,7171	0,2653
7	Euclidean	34,4094	0,7815	0,2862
	Manhattan	33,7762	0,7750	0,2779

Berdasarkan Tabel 4 terlihat bahwa nilai Calinski-Harabazs *index* , *Gamma index* , dan *Silhouette index* terbesar berada pada jumlah cluster k = 4, dengan pengukuran jarak Euclidean. Profilisasi klaster terbaik k = 4 disajikan pada Tabel 5 berikut:

Tabel 5 . Hasil Pengelompokan Terbaik *(k = 4)*

k	Objek Medoid	Jumlah Anggota	Daerah
1	Objek ke-12	28	Kepulauan Mentawai, Lombok Utara, Sumba Barat, Kupang, Timor Tengah Selatan, Belu, Alor, Rote Ndao, Sumba Barat Daya, Manggarai Timur, Sabu Raijua, Donggala, Tojo Una-Una, Nias, Nias Selatan, Nias Utara, Nias Barat, Sula Kepulauan, Sorong Selatan, Sorong, Maybrat, Manokwari Selatan, Pegunungan Arfak, Paniai, Mappi, Waropen, dan Supiori
2	Objek ke-7	17	Musi Rawas Utara, Pesisir Barat, Sumba Timur, Lembata, Sumba Tengah, Sigi, Kepulauan Aru, Seram Barat, Seram Timur, Maluku Barat Daya, Buru Selatan, Pulau Taliabu, Teluk Wondama, Teluk Bintuni, Nabire, Boven Digoel, dan Keerom
3	Objek ke-37	5	Malaka, Tambrauw, Jayawijaya, Asmat, dan Mamberamo Raya
4	Objek ke-50	12	Puncak Jaya, Yahukimo, Pegunungan Bintang, Tolikara, Nduga, Lanny Jaya, Mamberamo Tengah, Yalimo, Puncak, Dogiyai, Intan Jaya, dan Deiyai

Interpretasi dilakukan untuk melihat nilai rata-rata/*centroid* masing-masing variabel dari cluster terbaik, ditunjukkan pada Tabel 6:

Tabel 6 . Nilai *Centroid* Masing-masing Variabel pada setiap Klaster Terbaik *(k = 4)*

Variabel	Klaster 1	Klaster 2	Klaster 3	Klaster 4
UHH/X_1 (Tahun)	67,1	63,5	59,5	64,6
HLS/X_2 (Tahun)	12,7	12,4	11,8	7,51
RLS/X_3 (Tahun)	7,09	8,45	5,72	3,66
PPD/X_4 (Tahun)	6659	8334	5610	4896

Berdasarkan Tabel 5 dan Tabel 6 diperoleh informasi sebagai berikut:

a. Klaster 1

Klaster 1 memiliki dua variabel atau indikator IPM yaitu X3 dan X4 yang rata-ratanya lebih rendah dari klaster 2. Klaster 1 mempunyai nilai *centroid* yang lebih tinggi untuk variabel X3 dan X4 dibandingkan klaster 3 dan 4, serta rata-rata yang lebih tinggi untuk variabel X1 dan X2 dibandingkan kelompok 2, 3, dan 4.

b. Klaster 2

Pada variabel X1, klaster ini mempunyai nilai *centroid* lebih rendah dari klaster 1 dan 4, serta variabel X2 pada klaster 1. Klaster 2 mengungguli klaster 3 pada variabel X1, klaster 3 dan klaster 4 pada variabel X2, serta klaster 1, 3, dan 4 pada variabel X3 dan X4.

c. Klaster 3

Anggota klaster ini terdiir dari 5 kabupaten yang tesebar di beberapa provinsi, yaitu: Malaka, Tambrauw, Jayawijaya, Asmat, dan Mamberamo Raya. Pada variabel X1, klaster ini mempunyai nilai *centroid*/rata-rata lebih rendah pada klaster 1, 2, dan 4 serta pada variabel X2, X3, X4 pada klaster 1 dan 2. Klaster 3 juga mempunyai nilai rata-rata lebih tinggi dari klaster 4 pada variabel X2, X3, dan X4.

d. Klaster 4

Pada variabel X2, X3, X4, dan X1, klaster ini mempunyai nilai centroid yang lebih rendah dibandingkan dengan klaster 1, 2, dan 3. Klaster 4 mengungguli klaster 2 dan 3 pada variabel X1. Anggota klaster 4 terdiri dari 12 kabupaten, yaitu: Puncak Jaya, Yahukimo, Pegunungan Bintang, Tolikara, Nduga, Lanny Jaya, Mamberamo Tengah, Yalimo, Puncak, Dogiyai, Intan Jaya, dan Deiyai.

5. KESIMPULAN

Pengelompokan menggunakan metode *k-Medoids* dengan pengukuran jarak euclidean dan jarak manhattan untuk k = 3, 4, 5, 6, dan 7 diperoleh klaster yang terbaik pada k = 4 dengan jarak euclidean dimana nilai CH *index* = 37,1576, nilai *Gamma index* = 0,7821, dan nilai Silhouette *index* = 0,3354. Hasil pengelompokan metode ini menunjukkan bahwa jarak pengukuran berpengaruh terhadap hasil pengelompokan. Profilisasi hasil analisis klaster menunjukan bahwa pada klaster 3 indikator IPM yang paling rendah yaitu, Umur Harapan Hidup berada pada kabupaten Malaka, Tambrauw, Jayawijaya, Asmat, dan Mamberamo Raya. Untuk klaster 4 indikator yang paling rendah yaitu, Harapan Lama Sekolah, Rata-rata Lama Sekolah dan Pengeluaran per kapita disesuaikan berada pada kabupaten Puncak Jaya, Yahukimo, Pegunungan Bintang, Tolikara, Nduga, Lanny Jaya, Mamberamo Tengah, Yalimo, Puncak, Dogiyai, Intan Jaya, dan Deiyai. Klaster 1 dan klaster 2 didominasi indikator IPM dengan rata-rata tinggi diantara klaster yang lain.

Untuk menentukan clustering mana yang lebih tahan terhadap data *outlier*, penelitian lebih lanjut dapat dilakukan dengan menggunakan clustering dan perbandingan pendekatan clustering *k-Medoids* dan *k-Means*. Penelitian selanjutnya bisa menggunakan metode klastering *k-Prototypes* dengan menambahkan variabel penelitian dan bisa juga menggunakan pengukuran jarak yang lainnya seperti jarak *g*ower.

DAFTAR PUSTAKA

Aggarawal, C. C.,& Reddy,K. C. 2014. *Data Clustering Algoritmas and Applications*. New York: CRC Press.

Anderberg, M. 1973. *Cluster Analysis for Application*. New York: Academic Press.

Baarsch, J., & Celebi, M. E. 2012. Investigastion of Internal Validity Measures for K-Means Clustering. *International Multiconference of Engineers and Computer Scientist 1*. Los Angeles: Louisiana Board of Regents. 14–16.

Brock, G., Vasyl, P., Susmita, D., & Somnath, D. 2008. CValid: An R Package for Cluster Validation. *Journal of Statistical Software*. Vol. 25, No.4, 1–22.

[BPS] Badan Pusat Statistik. 2021. *Indeks Pembangunan Manusia 2021* . Jakarta: Badan Pusat Statistik.

Gujarati, D. 2009. *Dasar-Dasar Ekonometrika*. Jakarta: Erlangga.

Hair, J. F., Anderson, R. E., Thatham, R. L., & Black, W. C. 2010. *Multivariate Data Analysis Seventh Edition*. New Jersey: Pearson Education Inc.

Han, J., & Kamber, M. 2006. *Data Mining: Concepts and Techniques*. San Fransisco: Elsevier Inc.

Kaufman, L., & Rousseeuw, P. J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics)*. New York: Wiley.

Kaur, N. K., Kakur, U., & Singh, D. 2014. K-Medoids Clustering Algorithm. *International Journal of Computer Application and Technology (IJCAT)*. Vol.1, No.1.

Milligan, G., Cooper, M. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*. Vol.50, No. 2, 159-179.

Nerukar, P., Pavate, A., Shah, M., & Jacob, S. 2019. Performance of Internal Cluster Validations Measures for Evolutionary Clustering. *Computing, Communication and Signal Processing*. Singapore: Springer Nature Singapore Pte Ltd. 305–312.

[PERPES] Peraturan Presiden. 2020. *Peraturan Presiden tentang Penetapan Daerah Tertinggal Tahun 2020-2024 Nomor 63 Tahun 2020*. Tersedia: https://peraturan.bpk.go.id/Home/Details/136563/perpres-no-63-tahun-2020 (diakses pada tanggal 6 Juni 2022)

Rousseeuw, P. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Jornal of Computational and Applied Mathematics*. Vol.20, 53–65.

Sembiring, R. 1995. *Analisis Regresi*. Bandung: Penerbit ITB.

Supranto, J. 2004. *Analisis Multivariat : Arti dan Interpretasi*. Jakarta: PT. Rineka Cipta.

Walpole, R.E., dan Myers, R.H. 1995. *Ilmu Peluang dan Statistika untuk Insinyur dan Ilmuwan Edisi ke-4*. Bandung : Penerbit ITB

Widarjono, A. 2010. Analisis Statistika Multivariat Terapan. Yogyakarta: UPP STIM YKPN.