ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 13, Nomor 2, Tahun 2024, Halaman 479 - 489
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PENERAPAN ALGORITMA *k-PROTOTYPE* UNTUK PENGELOMPOKAN DESA DI KABUPATEN BEKASI BERDASARKAN INFRASTRUKTUR DIGITAL

Ariani Fitri Purba[1], Mustafid[2], Puspita Kartikasari[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***E-mail** : arianiifp@gmail.com

DOI: 10.14710/j.gauss.13.2.479-489

Article Info:
Received: 2024-01-15
Accepted: 2025-01-24
Available Online: 2025-01-31

Keywords:
Digital Infrastructure; Mixed Type Data; Cluster Analysis; k-Prototype Algorithm; Average Silhouette Width

Abstract: Grouping villages in Bekasi Regency is necessary as a planning and evaluation material for government program targets, especially in digital transformation efforts. The goal is to find out which villages are prioritized based on the characteristics of digital infrastructure. Thus, grouping villages in Bekasi Regency based on digital infrastructure needs to be done to support the success of digital transformation efforts. The analysis that can be used to group a village is cluster analysis. The clustering method used in this research is the k-Prototype algorithm using the value of k = 1, 2, 3, ..., and 10. The k-Prototype algorithm is a clustering method that can handle mixed type data, namely numeric and categorical types and large data. The k-Prototype algorithm has the advantage that the algorithm is less complex and better than hierarchy-based algorithms. Based on the results of analysis, the optimal number of groups formed was four groups with an Average Silhouette Width value of 0,505. Group 3, which consists of 9 villages, is the best group based on digital infrastructure characteristics, while group 1, which consists of 41 villages, has the lowest average cellular phone communication service operator and is dominated by villages that do not have internet access and internet facilities at the village office. In addition, group 1 has the most villages with weak cellular phone signals compared to the other groups.

1. PENDAHULUAN

Pada era industri 4.0, transformasi digital merupakan suatu hal yang tidak dapat dihindarkan. Transformasi digital adalah proses di mana teknologi digital digunakan untuk menciptakan perubahan yang signifikan dalam berbagai aspek (Wahyuddin *et al.*, 2022). Kabupaten Bekasi merupakan salah satu kabupaten di provinsi Jawa Barat yang terdiri dari 23 kecamatan dan 187 desa sudah mulai menerapkan transformasi digital. Penjabat Bupati Kabupaten Bekasi, Dr. H Dani Ramdan dalam diskusi pada 29 Juni 2022 menyatakan bahwa Kabupaten Bekasi sudah mulai mempersiapkan teknologi digital pada administrasi pemerintahan dan layanan publik. Banyak tantangan yang harus dihadapi untuk melakukan transformasi digital. Salah satu cara yang pertama dilakukan yaitu dengan mengelompokkan desa-desa di Kabupaten Bekasi berdasarkan infrastruktur digital sehingga dapat diketahui desa mana yang belum siap menghadapi transformasi digital.

Analisis *cluster* menjadi pilihan untuk mengelompokkan desa di Kabupaten Bekasi berdasarkan infrastruktur digital. Data pada dunia nyata cenderung bersifat variatif dan sering ditemukan beberapa campuran tipe data, yaitu bertipe numerik dan kategorik sehingga diperlukan suatu algoritma pengelompokan yang dapat mengatasi campuran tipe data (Pham *et al.*, 2016). Salah satu metode untuk mengelompokkan campuran tipe data adalah dengan menggunakan algoritma *k-Prototype* (Huang, 1997). Algoritma *k-Prototype* adalah kombinasi dari metode pengelompokan *k-Means* dan *k-Modes*. Kelebihan dari algoritma *k-Prototype* adalah algoritmanya tidak terlalu rumit dan dapat menangani data yang besar.

Selain itu menurut Huang (1997), algoritma *k-Prototype* lebih unggul daripada algoritma berbasis hierarki (Huang, 1997). Nooraeni *et al.* (2019) menyimpulkan bahwa algoritma *k-Prototype* dapat mempertahankan efisiensi algoritma *k-Means* dalam menangani data berukuran besar tetapi menghilangkan keterbatasan penerapan *k-Means* yang hanya dapat digunakan untuk data bertipe numerik.

Jumlah kelompok optimal dari hasil algoritma *k-Prototype* dapat diperiksa menggunakan *Average Silhouette Width*. Berdasarkan penelitian yang dilakukan oleh Aschenbruck dan Szepannek (2020) yang menguji algoritma *k-Prototype* menggunakan delapan metode validasi, diperoleh hasil bahwa *Average Silhouette Width* lebih unggul dalam *run time* tercepat dan penentuan jumlah kelompok lebih stabil dan akurat daripada metode validasi lainnya. Sehingga pada penelitian ini, dilakukan analisis pengelompokan desa di Kabupaten Bekasi berdasarkan infrastruktur digital menggunakan algoritma *k-Prototype* dengan metode validasi *Average Silhouette Width* dalam menentukan jumlah kelompok optimal. Hasil dari pengelompokan ini dapat membantu pemerintah Kabupaten Bekasi dalam menyusun strategi peningkatan infrastruktur digital pada desa yang belum siap diadakannya transformasi digital.

2. TINJAUAN PUSTAKA

Infrastruktur digital adalah pondasi yang mendukung sistem komputasi. Pembangunan infrastruktur digital bertujuan untuk memberikan kemudahan dan meningkatkan efisiensi serta efektivitas waktu dalam menjalankan berbagai kegiatan dan ekonomi digital dengan menggunakan teknologi informasi (Kemenkeu, 2022). Pengelompokan desa berdasarkan infrastruktur digital di Kabupaten Bekasi merupakan salah satu cara yang dapat digunakan untuk mengetahui kesiapan adanya transformasi digital di Kabupaten tersebut.

Analisis *cluster* merupakan teknik statistik dengan mempartisi data ke dalam kelompok-kelompok berdasarkan kesamaan antara objek-objek data. (Everitt *et al.,* 2011). Setiap objek data dalam satu kelompok memiliki kemiripan, tetapi berbeda dengan objek data di kelompok lain. Hal ini berarti setiap kelompok memiliki homogenitas internal yang tinggi dan heterogenitas eksternal yang tinggi. (Hair *et al*., 2010). Analisis *cluster* memiliki dua asumsi (Hair *et al.,* 2010).

a. Sampel Representatif (*Representativeness of the sample*)

Sampel pada analisis *cluster* harus merepresentasikan populasi yang akan diteliti. Uji KMO (Kaiser-Mayer-Olkin) merupakan salah satu pengujian untuk sampel representatif. Jika nilai KMO kurang dari 0,5 maka dapat dikatakan bahwa sampel belum mewakili populasi. Formula nilai KMO tertulis pada Persamaan (1)

$$KMO = \frac{\sum_{p,q=1}^{m} r_{pq}^2}{\sum_{p,q=1}^{m} r_{pq}^2 + \sum_{p,q=1}^{m} \sum_{r=1}^{m} a_{pq}^2} \quad (1)$$

dengan

r_{pq} = koefisien korelasi antara variabel p dan q

$a_{r(pq)}$= koefisien korelasi parsial antara variabel p dan q dengan variabel r

b. Non-Multikolinieritas

Multikolinieritas adalah sebuah situasi yang menunjukkan adanya hubungan kuat antara beberapa variabel independen. Variabel independen alam analisis *cluster* sebaiknya tidak terindikasi multikolinieritas (Hair *et al*., 2010). Dengan mengetahui nilai *Variance Inflating Factor* (VIF) dapat melihat apakah terindikasi multikolinieritas atau tidak. Formula VIF tertulis pada Persamaan (2).

$$VIF = \frac{1}{1-r_{pq}^2} \quad (2)$$

Tetapi, VIF tidak selalu dapat diterapkan pada semua model. Misalnya, pada kumpulan variabel independen yang memiliki kategori untuk variabel kategorik, nilai VIF kurang akurat. Diperlukan analisis berulang untuk memastikan variabel tidak memiliki nilai VIF di atas nilai batas pada setiap tingkat variabel kategorik. Sehingga perlu menghitung VIF dari kategori variabel tersebut. Ini akan rumit apabila variabel kategorik lebih dari satu. Fox dan Monette (1992) menyatakan bahwa *Generalized* VIF atau GVIF dapat digunakan dalam mengatasi hal tersebut. GVIF adalah pengganti yang biasa digunakan untuk VIF standar sebagai faktor inflasi varians. Formula GVIF seperti Persamaan (3) (Fox dan Monette, 1992)

$$GVIF_q = \frac{det\left(\boldsymbol{R}_{\boldsymbol{X_q}}\right) det\left(\boldsymbol{R}_{\boldsymbol{X}_{[-q]}}\right)}{\det(\boldsymbol{R})} \quad (3)$$

dengan R_{X_q} yaitu matriks korelasi X_q, $R_{X_{[-q]}}$ merupakan matriks korelasi $X_{[-q]}$, dan R merupakan matriks korelasi seluruh variabel pada matriks X. Fox dan Monette (1992) menyarankan menggunakan $GVIF^{\frac{1}{2\times df}}$ untuk membuat GVIF sebanding di seluruh dimensi. df diperoleh dari $df = k - 1$ dengan k adalah jumlah kategori dalam variabel kategorik. Kemudian, untuk mendeteksi adanya multikolinieritas yaitu sama dengan aturan VIF, $GVIF^{\frac{1}{2\times df}} < 10{,}00$ menunjukkan bahwa tidak terjadi multikolinieritas.

Ukuran jarak kemiripan yang digunakan pada penelitian ini adalah jarak *Euclidean* untuk data numerik, jarak *Simple Matching* untuk data kategorik, dan jarak untuk data campuran. Jarak *Euclidean* merupakan jarak yang biasa digunakan dalam analisis *cluster*, namun hanya dapat digunakan untuk data bertipe numerik. Jarak *Euclidean* dituliskan pada Persamaan (4)

$$d(x_i, c_y) = \left(\sum_{p=1}^{m}\left(x_{ip} - c_{yp}\right)^2\right)^{1/2} \quad (4)$$

dengan x_{ip} adalah nilai objek ke- i pada variabel ke- p, c_{yp} adalah nilai pusat kelompok y pada variabel p.

Simple Matching adalah ukuran jarak untuk tipe data kategorik yang digunakan dalam metode *k-Modes* (Huang, 1998). *k-Modes* adalah sebuah algoritma *clustering* yang bekerja dengan mencari modus atau nilai dominan dari tiap variabel kategorik dalam sebuah kelompok (Huang, 1998). Persamaan jarak data tipe kategorik dinyatakan seperti Persamaan (5) (Huang, 1997)

$$d(x_i, c_y) = \left(\sum_{q=1}^{m} \delta\left(x_{iq}; c_{yq}\right)\right) \quad (5)$$

dengan $\delta\left(x_{iq}; c_{yq}\right) = \begin{cases} 0, x_{iq} = c_{yq} \\ 1, x_{iq} \neq c_{yq} \end{cases}$

x_{iq} adalah nilai objek ke- i pada variabel q, c_{yq} adalah nilai pusat kelompok ke- y pada variabel q.

Jarak campuran tipe data *(k-Prototype)* adalah ukuran jarak yang dapat digunakan untuk mengelompokkan data campuran yang terdiri dari data numerik dan data kategorik. Pengukuran jarak campuran tipe data diperoleh dari pengukuran jarak *Euclidean* (untuk tipe data numerik) yang dikuadratkan dan dijumlahkan unsur koefisien penimbang gamma (γ) yang dikalikan dengan ukuran jarak tipe data kategorik (Huang, 1998). Persamaan untuk menghitung ukuran jarak campuran tipe data dinyatakan seperti Persamaan (6) (Huang, 1997)

$$d(x_i, c_y)^2 = \sum_{p=1}^{m}(x_{ip} - c_{yp})^2 + \gamma \sum_{q=1}^{m} \delta(x_{iq}; c_{yq}) \quad (6)$$
$$\gamma = \frac{1}{m}\sum_{p=1}^{m} s_p$$

dengan $\sum_{p=1}^{m}(x_{ip} - c_{yp})^2$adalah ukuran jarak tipe data numerik, $\sum_{q=1}^{m} \delta(x_{iq}; c_{yq})$ adalah ukuran jarak untuk data kategorik, γ adalah koefisien penimbang, m adalah jumlah variabel numerik, dan $s_p$adalah simpangan baku variabel numerik.

Algoritma *k-Prototype* adalah metode dengan mengintegrasikan perhitungan jarak kemiripan pada algoritma *k-Means* dan *k-Modes* untuk pengelompokan campuran tipe data (numerik dan kategorik) dengan mengintegrasikan perhitungan jarak kemiripan pada algoritma *k-Means* dan *k-Modes*. Beberapa tahapan dalam algoritma *k-Prototype* adalah sebagai berikut (Huang, 1997).

a. Menentukan banyak kelompok (k) yang akan dibentuk.
b. Inisialisasi awal *prototype* sebagai pusat kelompok sejumlah k sesuai dengan jumlah kelompok yang dibuat.
c. Melakukan perhitungan jarak dengan ukuran jarak campuran tipe data terhadap pusat kelompok yang telah ditentukan.
d. Alokasi objek ke dalam kelompok yang memiliki nilai jarak terdekat.
e. Melakukan perhitungan nilai *centroid* baru menggunakan nilai rata-rata untuk variabel numerik nilai modus untuk variabel kategorik.
f. Realokasi objek ke dalam masing-masing kelompok berdasarkan perhitungan jarak terdekat dengan nilai pusat kelompok yang baru. Tahap c diulang sampai iterasi maksimum tercapai atau tidak ada lagi perubahan objek dalam kelompok.

Algoritma *k-Prototype* memerlukan metode validasi agar jumlah kelompok yang terbentuk optimal. Salah satu metode validasi pada analisis *cluster* yaitu *Average Silhouette Width*. Nilai *Average Silhouette Width* ada di rentang -1 sampai +1. Jika nilai *Average Silhouette Width* mendekati +1, maka jumlah kelompok yang terbentuk semakin optimal. Persamaan *Average Silhouette Width* dituliskan seperti pada Persamaan (7)

$$S(i) = \frac{1}{n}\sum_{i=1}^{n} \frac{b(x_i) - a(x_i)}{max(a(x_i); b(x_i))} \quad (7)$$
$$a(x_i) = \frac{1}{n_y - 1}\sum_{jy=1}^{n_y} d(x_{iy}, x_{jy}) \text{ dan } b(x_i) = min_{y \neq z} \frac{1}{n_z}\sum_{jz=1}^{n_z} d(x_{iy}, x_{jz})$$

dengan
$d(x_{iy}, x_{jy})$ = jarak objek ke-i pada kelompok y dengan objek ke-j pada kelompok y
$d(x_{iy}, x_{jz})$ = jarak objek ke-i pada kelompok y dengan objek ke-j pada kelompok z

3. METODE PENELITIAN

Penelitian ini menggunakan data sekunder berupa data Potensi Desa (PODES) di Kabupaten Bekasi yang tercatat pada tahun 2019 yang diperoleh dari Badan Pusat Statistik Kabupaten Bekasi yang mencatat 187 desa. Variabel yang digunakan terdiri dari 3 variabel numerik dan 4 variabel kategorik.

Tabel 1. Variabel Penelitian

	Variabel	Jenis Variabel	Keterangan
X_1	Persentase jumlah keluarga pengguna listrik	Numerik	Persen

	Variabel	Jenis Variabel	Keterangan
X_2	Jumlah menara telepon seluler atau *Base Transceiver Station* (BTS)	Numerik	Satuan (Benda)
X_3	Jumlah operator layanan komunikasi telepon seluler yang menjangkau desa	Numerik	Satuan (Benda)
X_4	Keberadaan internet untuk fasilitas di desa	Kategorik	-
X_5	Sinyal telepon seluler di wilayah desa	Kategorik	-
X_6	Sinyal internet telepon seluler di sebagian besar wilayah desa	Kategorik	-
X_7	Fasilitas internet di kantor kepala desa	Kategorik	-

Dengan kategori variabel X_4 , X_5, X_6 dan X_7 dapat dilihat pada Tabel 2.

Tabel 2. Kategori Variabel X_4, X_5, X_6 dan X_7

Variabel	Kategori Variabel
X_4	1: Ada
	2: Tidak Ada
X_5	1: Sinyal sangat kuat
	2: Sinyal kuat
	3: Sinyal lemah
X_6	1: 4G/LTE
	2: 3G/H/H+/EVDO
	3: 2,5G/E/GPRS
X_7	1: Berfungsi
	2: Jarang berfungsi
	3: Tidak berfungsi

Penelitian ini menggunakan *software* RStudio versi 4.1.2 dan Microsoft Excel untuk menganalisis data dengan langkah-langkah sebagai berikut.

a. Memasukkan data.
b. Melakukan *pre-processing* data.
c. Melakukan normalisasi pada variabel numerik menggunakan metode *z-score*.
d. Melakukan pelabelan data pada variabel kategorik.
e. Uji asumsi non-multikolinieritas menggunakan nilai *Generalized Variance Inflation Factor* yang dipangkatkan $\frac{1}{2\times df}$, $GVIF^{\frac{1}{2\times df}} = \left[\frac{det\left(\boldsymbol{R}_{\boldsymbol{X_q}}\right)det\left(\boldsymbol{R}_{\boldsymbol{X}_{[-q]}}\right)}{det(\boldsymbol{R})}\right]^{\frac{1}{2\times df}}$.
f. Menentukan nilai k (jumlah kelompok) yaitu k = 2, 3, 4, …, 10.
g. Melakukan pengelompokan desa di Kabupaten Bekasi menggunakan algoritma *k-Prototype* berdasarkan infrastruktur digital.
h. Melakukan validasi *Average Silhouette Width*.
i. Memilih k optimal dengan melihat nilai Silhouette tertinggi yang mendekati +1.
j. Interpretasi hasil pengelompokan dengan algoritma *k-Prototype*.

4. HASIL DAN PEMBAHASAN

Proses *data mining* dimulai dengan mengecek data yang hilang dan duplikasi pada seluruh data atau disebut dengan *pre-processing*. Hasil *Pre-processing* diketahui bahwa data tidak ada yang hilang dan tidak terjadi duplikasi pada data. Selanjutnya, dilakukan normalisasi pada variabel numerik menggunakan metode *z-score*.

Uji asumsi sampel representatif pada penelitian ini tidak dilakukan karena data yang digunakan merupakan data seluruh desa di Kabupaten Bekasi yang merupakan data populasi. Uji asumsi non-multikolinieritas menggunakan nilai $GVIF^{\frac{1}{2\times df}}$ dapat dilihat pada Tabel 3.

Tabel 3. Hasil Uji Asumsi Analisis *Cluster*

Variabel	$GVIF^{\frac{1}{2\times df}}$	Variabel	$GVIF^{\frac{1}{2\times df}}$
X_1	1,024	X_5	1,135
X_2	1,166	X_6	1,141
X_3	1,187	X_7	1,037
X_4	1,110		

Pada uji asumsi non-multikolinieritas, nilai $GVIF^{\frac{1}{2\times df}}$ setiap variabel kurang dari 10. Hal ini menunjukkan bahwa pada setiap variabel independen tidak terjadi multikolinieritas, sehingga asumsi non-multikolinieritas terpenuhi.

Setelah asumsi analisis *cluster* terpenuhi, selanjutnya dilakukan pengelompokan menggunakan algoritma *k-Prototype*. Tahapan perhitungan manual pada proses pengelompokan dengan algoritma *k-Prototype* dijelaskan sebagai berikut.

a. Menentukan k (jumlah kelompok) yang akan digunakan, misalnya nilai k = 2.
b. Memilih objek sesuai dengan jumlah kelompok yang akan digunakan sebagai *centroid* secara acak. Objek data yang menjadi *centroid* awal ditulis pada Tabel 4.

Tabel 4. Pusat Kelompok Awal pada Pengelompokan dengan k = 2

Objek ke-(i)	Pusat Kelompok	Indeks Variabel ke-(p)						
		1	2	3	4	5	6	7
90	c_1	0,145	0,625	0,617	1	2	1	1
180	c_2	0,145	-0,501	-0,902	1	2	1	2

c. Menentukan koefisien gamma (γ). Berdasarkan hasil output RStudio, diperoleh s_1 = 1,000; s_2 = 1,000; s_3 = 1,000; sehingga nilai koefisien gamma (γ) adalah sebagai berikut.

$$\gamma = \frac{1}{3}(1{,}000 + 1{,}000 + 1{,}000) = 1{,}000$$

d. Menghitung jarak semua objek data ke pusat kelompok awal yang telah ditentukan menggunakan jarak *k-Prototype* seperti pada Persamaan 7.

$$d(x_1, c_1)^2 = ((x_{11} - c_{11})^2 + (x_{12} - c_{12})^2 + (x_{13} - c_{13})^2) + \gamma\big(\delta(x_{14}; c_{14}) + \delta(x_{15}; c_{15}) + \delta(x_{16}; c_{16}) + \delta(x_{17}; c_{17})\big)$$

$$= ((0{,}145 - 0{,}145)^2 + (-0{,}051 - 0{,}625)^2 + (-0.902 - 0{,}617)^2) + \gamma\big(\delta(1;1) + \delta(2;2) + \delta(1;1) + \delta(1;1)\big)$$

$$= (0{,}000 + 0{,}457 + 0{,}2307) + 1(0 + 0 + 0 + 0) = 2{,}764$$

$$d(x_1, c_1) = 1{,}663$$

$$d(x_1, c_2)^2 = ((x_{11} - c_{21})^2 + (x_{12} - c_{22})^2 + (x_{13} - c_{23})^2 + (x_{14} - c_{24})^2) + \gamma\big(\delta(x_{15}; c_{25}) + \delta(x_{16}; c_{26}) + \delta(x_{17}; c_{27})\big)$$

$$d(x_1, c_2)^2 = ((0{,}145 - 0.145)^2 + (-0{,}051 - (-0.501))^2 + (-0.902 - (-0.902))^2) + \gamma\big(\delta(1;1) + \delta(2;2) + \delta(1;1) + \delta(1;2)\big)$$

$$= (0{,}000 + 0{,}203 + 0{,}000) + 1(0 + 0 + 0 + 1) = 1{,}203$$

$$d(x_1, c_2) = 1{,}097$$

dan seterusnya sampai objek ke-187 $\left(d(x_{187}, c_1)\right)$ dan $\left(d(x_{187}, c_2)\right)$.

e. Mengalokasikan objek ke kelompok berdasarkan dari pusat kelompok terdekat.
f. Menghitung nilai pusat kelompok baru menggunakan rata-rata untuk variabel numerik dan modus untuk variabel kategorik.
g. Proses pada langkah ke c sampai langkah ke e akan terus berulang hingga objek sudah tidak berpindah dan nilai pusat kelompok pada dua iterasi terakhir bernilai sama (konvergen).

Hasil dari pengelompokan menggunakan algoritma *k-Prototype* setelah mencapai iterasi maksimum dengan k = 2, 3, ..., 10 disajikan pada Tabel 5.

Tabel 5. Hasil Pengelompokan Menggunakan Algoritma *k-Prototype*

k	Kelompok	Jumlah Anggota
2	1	58
	2	129
3	1	42
	2	136
3	3	9
4	1	41
	2	133
	3	9
	4	4
5	1	50
	2	83
5	3	9
	4	4
	5	41
6	1	18
	2	91
	3	28
	4	4
	5	37
	6	9
7	1	28
	2	65
	3	26
	4	15
	5	27
	6	9
	7	17
8	1	26
	2	8
	3	68
	4	21
8	5	20
	6	23
	7	17
	8	4
9	1	17
	2	63
	3	20
9	4	12
	5	24
	6	9
	7	4
	8	15
	9	23
10	1	11
	2	64
	3	15
	4	9
	5	22
	6	9
	7	4
	8	20
	9	15
	10	18

Penentuan jumlah kelompok optimal pada penelitian ini menggunakan *Average Silhouette Width* seperti pada Persamaan 8. Nilai *Average Silhouette* terbesar akan menjadi jumlah kelompok optimal. Tahapan dalam menghitung nilai *Average Silhouette* dijelaskan sebagai berikut.

a. Menghitung nilai $a(x_i)$ yang diperoleh dari menghitung jarak antara objek ke-1 terhadap objek lainnya yang terletak pada kelompok 1.

$$d\left(x_{1,1}; x_{2,1}\right)^2 = ((0{,}145 - 0{,}145)^2 + (-0{,}051 - 0{,}175)^2 + (-0{,}902 - (-1{,}661))^2) + 1\left(\delta(1;1) + \delta(2;3) + \delta(1;2) + \delta(1;1)\right)$$

$$= (0{,}000 + 0{,}051 + 0{,}577) + 1(0 + 1 + 1 + 0) = = 2{,}628$$

$$d\left(x_{1,1}; x_{2,1}\right) = 1{,}621$$

⋮

Perhitungan dilakukan seterusnya hingga objek ke-1 terhadap objek ke-186 pada kelompok ke-1. Kemudian nilai $a(x_i)$ diperoleh sebagai berikut.

$a(x_1) = \frac{1}{58-1}(1{,}621 + 2{,}008 + \cdots + 3{,}457) = 2{,}043$

b. Menghitung nilai $b(x_i)$ yang diperoleh dari nilai minimum rata-rata jarak antara objek ke- i terhadap seluruh objek lainnya pada kelompok yang berbeda.

$$d(x_{1,1}; x_{3,2})^2 = ((0{,}145 - 0{,}145)^2 + (-0{,}051 - (-0{,}051))^2 + (-0{,}902 - 0{,}617)^2) + 1(\delta(1;2) + \delta(2;2) + \delta(1;1) + \delta(1;1))$$

$$= (0{,}000 + 0{,}000 + 2{,}307) + 1(1 + 0 + 0 + 0) = 3{,}307$$

$$d(x_{1,1}; x_{3,2}) = 1{,}819$$

⋮

Perhitungan dilakukan seterusnya hingga objek ke-1 terhadap objek ke-187 pada kelompok ke-2. Kemudian nilai $b(x_i)$ diperoleh sebagai berikut.

$b(x_1) = \frac{1}{129-1}(1{,}819 + 0{,}759 + \cdots + 0{,}883) = 1{,}823$

Perhitungan $a(x_i)$ dan $b(x_i)$ kemudian dilanjutkan untuk setiap objek.

c. Menghitung nilai *Average Silhouette Width* menggunakan Persamaan 8.

$$S(i) = \frac{1}{187}\left[\left(\frac{1{,}823-2{,}043}{max(2{,}043;1{,}823)}\right) + \left(\frac{2{,}797-2{,}433}{max(2{,}433;2{,}797)}\right) + \cdots + \left(\frac{4{,}551-3{,}869}{max(3{,}869;4{,}551)}\right)\right]$$

$$= \frac{1}{187}[-0{,}322 + 0{,}079 + \cdots + 0{,}265] = 0{,}412$$

Sehingga diperoleh nilai *Average Silhouette Width* untuk k = 2 adalah 0,412.

Hasil perhitungan nilai *Average Silhouette Width* untuk k = 2, 3, 4, …, dan 10 menggunakan bantuan RStudio ditulis pada Tabel 6.

Tabel 6. Nilai *Average Silhouette Width* k = 2, 3, 4, …, 10

k	Nilai *Average Silhouette Width*
2	0,412
3	0,456
4	**0,505**
5	0,297
6	0,323
7	0,091
8	0,322
9	0,332
10	0,298

Setelah diketahui jumlah kelompok optimal yaitu k = 4, selanjutnya dilakukan penilaian karakteristik untuk hasil pengelompokan. Karakteristik setiap kelompok berdasarkan variabel numerik dan kategorik terlihat pada Tabel 7.

Tabel 7. Karakteristik Kelompok Berdasarkan Variabel

Kelompok	Rata-Rata setiap Variabel Numerik		
	Persentase jumlah keluarga pengguna listrik	Menara telepon seluler	Operator layanan komunikasi telepon seluler
1	1,000	0,976	2,537

2	1,000	3,090	4,677
3	1,000	16,667	5,111
4	0,900	0,500	2,750

Rata-Rata setiap Variabel Kategorik

Kelompok	Variabel				
	Keberadaan internet untuk fasilitas desa		Sinyal telepon seluler di desa		
	Ada	Tidak Ada	Sinyal Sangat Kuat	Sinyal Kuat	Sinyal Lemah
1	0,341	0,659	0,073	0,634	0,293
2	0,797	0,203	0,158	0,767	0,075
3	1,000	0,000	0,556	0,444	0,000
4	1,000	0,000	0,000	0,750	0,250

Kelompok	Variabel					
	Sinyal internet telepon seluler di desa			Fasilitas internet di kantor kepala desa		
	4G/ LTE	3G/H/H+/E VDO	2,5G/E/G PRS	Berfungsi	Jarang Berfungsi	Tidak Berfungsi
1	0,585	0,366	0,049	0,220	0,122	0,659
2	0,947	0,045	0,008	0,692	0,023	0,286
3	0,889	0,111	0,000	0,889	0,111	0,000
4	0,750	0,250	0,000	0,750	0,000	0,250

Berdasarkan Tabel 7, karakteristik masing-masing kelompok diinterpretasikan sebagai berikut.

a. Kelompok 1
   Kelompok 1 terdiri dari desa dengan rata-rata operator layanan komunikasi telepon selular (X_3) paling rendah. Kelompok 1 didominasi oleh desa yang belum memiliki akses internet (X_5) dan desa yang belum memiliki fasilitas internet di kantor desa (X_7). Selain itu, kelompok 1 merupakan desa dengan sinyal telepon selular yang lemah terbanyak dibandingkan dengan kelompok lainnya.
b. Kelompok 2
   Kelompok 2 terdiri dari desa dengan desa dengan seluruh keluarga sudah menggunakan listrik (X_1). Kelompok 2 didominasi oleh desa dengan sinyal internet telepon selular 4G/LTE tertinggi dibanding kelompok lainnya.
c. Kelompok 3
   Kelompok 3 terdiri dari rata-rata jumlah menara telepon selular atau atau *Base Transceiver Station* (BTS) dan operator layanan komunikasi telepon selular tertinggi dibandingkan kelompok lainnya. Kelompok 3 didominasi oleh desa dengan sinyal telepon selular sangat kuat. Selain itu, desa pada kelompok 3 sudah banyak memiliki fasilitas internet di kantor desa. Secara karakteristik, kelompok 3 merupakan kelompok dengan fasilitas infrastruktur digital tertinggi.
d. Kelompok 4
   Kelompok 4 adalah satu-satunya kelompok dengan persentase keluarga yang menggunakan listrik tidak mencapai 100%. Selain itu, rata-rata jumlah menara telepon selular pada kelompok 4 paling rendah dibandingkan dengan kelompok lain. Namun, kelompok 4 sudah cukup lebih baik dibandingkan kelompok 1. Hal tersebut terbukti dari rata-rata keberadaan internet untuk fasilitas di desa, Sinyal telepon seluler, sinyal internet, dan fasilitas internet di kantor desa jauh lebih tinggi dibandingkan kelompok 1.

Secara umum, keempat kelompok dapat dikategorikan dengan penilaian yang tertulis pada Tabel 8.

Tabel 8. Penilaian setiap Kelompok

Kelompok	Karakteristik
3	Sangat Baik
2	Baik
4	Cukup
1	Kurang

5. KESIMPULAN

Jumlah kelompok optimal yang diperoleh dari hasil validasi menggunakan metode *Average Silhouette Width* adalah empat kelompok ($k = 4$) dengan nilai terbesar yaitu 0,505. Jumlah desa pada masing-masing kelompok adalah 41 desa pada kelompok 1, kelompok 2 sebanyak 113 desa, kelompok 3 sebanyak 9 desa, dan pada kelompok 4 sebanyak 4 desa.

Berdasarkan karakteristik keempat kelompok yang terbentuk, kelompok 1 merupakan kelompok dengan karakteristik infrastruktur digital yang paling kurang dibandingkan dengan kelompok lainnya. Berdasarkan hal tersebut, Pemerintah Kabupaten Bekasi perlu menetapkan kelompok 1 sebagai prioritas utama dalam melakukan evaluasi terhadap pembangunan infrastruktur digital agar transformasi digital berjalan maksimal.

DAFTAR PUSTAKA

Aschenburck, R., dan Szepannek, G. 2020. Cluster Validation for Mixed Type Data. *Archives of Data Science, Series A*. Vol. 6 No. 1, Hal:1-12.

Badan Pusat Statistik Kabupaten Bekasi. 2019. *Hasil Pendataan Potensi Desa (PODES) 2019 Kabupaten Bekasi*. Bekasi: Badan Pusat Statistika Kabupaten Bekasi.

Everitt, B., Landau, S., Leese, M., dan Stahl, D. 2011. *Cluster Analysis (5th Ed)*. London: WILEY.

Fox, J., Monette, G. 1992. *Generalized Collinearity Diagnotics*. Journal of the American Statistical Association Vol. 87 No. 417, Hal: 178-183.

Hair, J., Black, W., Babin, B., dan Anderson, R. 2010. *Multivariate Data Analysis (7th Ed)*. New York: Pearson Prentice Hall.

Huang, Z. 1997. *Clustering Large Data Set with Mixed Numeric and Categorical Values*. In the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining, 21-34.

Huang, Z. 1998. *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*. Data Mining and Knowledge Discovery2. Vol.2 No. 3, Hal:283—304.

Kementrian Keuangan. 2022. *Infrastruktur Digital, Seberapa Penting?*. Tersedia: https://kpbu.kemenkeu.go.id/read/1152-1408/umum/kajian-opini-publik/infrastruktur-digital-seberapa-penting (diakses pada tanggal 25 November 2022).

Munthe, A., Sumertajaya, I., dan Syafitri, U. 2018. Penggerombolan Desa/Kelurahan Berdasarkan Indikator Kemiskinan dengan Menerapkan Algoritma TSC dan k-Prototypes. Indonesia Journal of Statistics and Its Applications. Vol.2, No. 2, Hal: 63-76.

Nooraeni, R., Suprijadi, J., dan Zulhanif. 2019. *k-Prototype* untuk Pengelompokan Data Campuran. *Jurnal Statistika Teori dan Aplikasi: Biomedics, Industry & Business and Social Statistics*. Vol 13 No. 1, Hal: 9-16.

Wahyuddin, S., Rachmat, Z., Amriadi, Kusumawarhani Z., Abbas, S. 2022. *Transformasi Digital Menuju Kelompok Informasi Masyarakat (KIM) di Desa Palangiseng*. Prosiding PEPADU 2022. e-ISSN: 2715-5811 Vol. 4, Hal: 141-144.

Pham, dan Maria, M. 2011. *Random Search with k-Prototypes Algorithm for Clustering Mixed Datasets.* Proceedings of The Royal Society a Mathematical Physical and Engineering Sciences. https://doi:10.1098/rspa.2010.0594.