 http://dx.doi.org/10.29313/.v0i0.28974

Penerapan Algoritme K-*Prototypes* untuk Pengelompokkan Desa-Desa di Provinsi Jawa Barat Berdasarkan Indikator Indeks Desa Membangun Tahun 2020

Mayang Ganmanah*, Abdul Kudus

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

*mayangganmanah99@gmail.com, akudus69@unisba.ac.id

Abstract. The grouping of villages in West Java needs to be done as material for planning and evaluating government program targets. The aim is to determine the status of village development based on the indicators of the developing village index and so that the policy programs carried out by the government are more focused according to the characteristics of the results of village grouping. Thus, it is necessary to conduct an analysis related to the clustering of the indicators of the developing village index. The analysis that can be used to group a village based on its characteristics is cluster analysis. The K-Prototypes algorithm is a cluster analysis method for large data with mixed-type data. A total of 3698 villages from 18 regencies and 9 cities were used as observation units. The grouping of villages in West Java Province based on the index of developing villages in 2020 using the K-prototypes algorithm produces 5 clusters. In the 1st cluster there are 54 villages, the 2nd cluster there are 1273 villages, the 3rd cluster there are 1272 villages In the 4th cluster there are 464 villages and in the 5th cluster there are 635 villages. Where the 5th cluster is the best cluster among the other 5 clusters.

Keywords: *Clustering*, K-*Prototypes*, Village Index Build.

Abstrak. Pengelompokkan desa di Jawa Barat perlu dilakukan sebagai bahan perencanaan dan evaluasi sasaran program pemerintah. Tujuannya untuk mengetahui penetapan status perkembangan desa berdasarkan indikator indeks desa membangun dan agar program kebijakan yang dilakukan pemerintah lebih terarah sesuai karakteristik hasil dari pengelompokkan desa. Dengan demikian perlu dilakukan analisis terkait klasterisasi terhadap indikator indeks desa membangun. Analisis yang dapat digunakan untuk mengelompokkan suatu desa berdasarkan karakteristik-karakteristik yang dimilikinya adalah analisis *cluster*. Algoritme K-*Prototypes* merupakan salah satu metode analisis *cluster* pada data berukuran besar dengan data bertipe campuran. Sebanyak 3698 desa dari 18 Kabupaten dan 9 Kota digunakan sebagai unit pengamatan. Pengelompokan desa-desa di Provinsi Jawa Barat berdasarkan indikator indeks desa membangun tahun 2020 dengan menggunakan algoritme K-*prototypes* menghasilkan 5 *cluster*. Pada *cluster* ke-1 terdapat 54 desa, *cluster* ke-2 terdapat 1273 desa, *cluster* ke-3 terdapat 1272 desa, *cluster* ke-4 terdapat 464 desa dan pada *cluster* ke-5 terdapat 635 desa. Dimana *cluster* ke-5 merupakan *cluster* terbaik diantara 5 *cluster* lainnya.

Kata Kunci: *Clustering*, K-*Prototypes*, Indeks Desa Membangun.

1. Pendahuluan

Dalam rangka memperkuat upaya pencapaian pembangunan di tingkat desa, dilakukan sebuah pengembangan ukuran yang mampu memotret perkembangan kemandirian desa, yaitu IDM atau Indeks Desa Membangun (Kementerian Desa, PDT dan Transmigrasi, 2019). IDM merupakan sebuah indeks komposit yang dibentuk berdasarkan tiga indeks, yaitu Indeks Ketahanan Sosial (IKS), Indeks Ketahanan Ekonomi (IKE), dan Indeks Ketahanan Ekologi/Lingkungan (IKL).

Pengelompokkan desa di Jawa Barat perlu dilakukan sebagai bahan perencanaan dan evaluasi sasaran program pemerintah. Tujuannya untuk mengetahui penetapan status perkembangan desa berdasarkan indikator indeks desa membangun dan agar program kebijakan yang dilakukan pemerintah lebih terarah sesuai karakteristik hasil dari pengelompokkan desa. Dengan demikian perlu dilakukan analisis terkait klasterisasi desa-desa di Jawa Barat berdasarkan indikator indeks desa membangun. Analisis yang digunakan adalah analisis data *mining*, data *mining* merupakan proses mengeksplorasi dan menganalisis data dalam jumlah besar untuk menemukan pola yang berarti. Analisis *cluster* adalah salah satu jenis data *mining* yang sering digunakan untuk mengelompokkan data berdasarkan variabel yang dimilikinya. Analisis *cluster* digunakan untuk mengelompokkan desa-desa berdasarkan variabel yang dimilikinya.

Terdapat dua metode dalam analisis *cluster* yaitu analisis berhierarki dan analisis tidak berhierarki. Metode berhierarki digunakan ketika banyaknya *cluster* yang akan dibentuk tidak diketahui sebelumnya dan kurang efisien untuk data berukuran sangat besar. Sedangkan metode tak berhierarki digunakan jika *cluster* yang akan dibentuk sudah diketahui sebelumnya Dalam penelitian ini data yang digunakan adalah data bertipe campuran yaitu data numerik dan data kategorik, maka algoritme yang digunakan dalam analisis ini adalah Algoritme K-*prototypes*.. Metode ini menggabungkan ukuran jarak pada algoritme K-*means* dan algoritme K-*modes* (Huang Z. , 1997). Algoritme ini memiliki keunggulan yaitu algoritmenya yang tidak terlalu kompleks, bisa mengatasi data yang bertipe campuran yaitu numerik dan kategorik, sangat efektif untuk menangani data yang sangat besar Kelemahan dari algoritme K-*Prototypes* adalah sensitif terhadap penentuan inisialisasi pusat *cluster* sehingga menyebabkan solusi lokal optimum.

Berdasarkan penjelasan latar belakang di atas, maka identifikasi masalah dalam penelitian ini yaitu "Bagaimana hasil pengelompokkan desa di Provinsi Jawa Barat berdasarkan indikator indeks desa membangun tahun 2020 dengan menggunakan algoritme K-*prototypes*? ". Selanjutnya, tujuan dalam penelitian ini yaitu untuk mengelompokkan desa-desa di Provinsi Jawa Barat berdasarkan indikator indeks desa membangun tahun 2020 dengan menggunakan algoritme K-*prototypes*.

2. Metodologi

Data yang digunakan untuk penelitian ini adalah data sekunder yang didapatkan dari website idm.kemendesa.go.id, dimana pengambilan sampel dilakukan dengan pengambilan data pada seluruh desa di Jawa Barat tahun 2020 yang dipublikasikan oleh Kementerian Desa Pembangunan Daerah Tertinggal dan Transmigrasi. Unit penelitian yang digunakan yaitu desa-desa di Provinsi Jawa Barat yang terdiri dari 3698 desa dari 18 kabupaten dan 9 kota. Semua variabel bebasnya berjumlah 54 variabel.

Metode Analisis Data

Analisis Cluster

Analisis cluster merupakan suatu teknik multivariat untuk mengelompokkan individu atau objek ke dalam cluster-cluster tersebut dengan sedemikian rupa. Sehingga objek-objek yang ada di dalam cluster yang sama atau mirip satu sama lain itu yang nantinya dibandingkan dengan objek-objek atau cluster lainnya. (Hajarisman, 2019)

Ukuran Kemiripan (Similarity Measure)

Ukuran kemiripan atau *similarity measure* merupakan proses pengukuran suatu objek terhadap objek yang menjadi acuan. Semakin meningkat jarak diantara dua objek maka akan semakin berbeda dua objek tersebut. Sebaliknya semakin kecil jarak diantara dua objek tersebut akan semakin mirip objeknya (Rencher, 2007). Ukuran Jarak yang akan digunakan pada penelitian yaitu ukuran jarak tipe data campuran K-*Prototypes*.

Algoritme K-*Prototypes* adalah gabungan dari algoritme k-*means* dan k-*modes* yang digunakan untuk mengklasterisasikan objek-objek bertipe campuran. Parameter γ digunakan untuk menghindari salah satu jenis atribut dan sebagai penyeimbang perbedaan skala peubah campuran yaitu peubah numerik dan peubah kategorik. Ukuran jarak untuk tipe data campuran yang digunakan pada algoritme K- *Prototypes* adalah sebagai berikut (Huang Z. , 1998) :

$$d_{ij} = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2 + \gamma \sum_{k=p+1}^{p+m} \delta(x_{ik}, x_{jk}) \quad ...(1)$$

Dimana :
d_{ij} : Ukuran jarak antara objek *i* dan *j*
$\sum_{k=1}^{p}(x_{ik} - x_{jk})^2$: Ukuran jarak untuk data tipe numerik
γ: Penyeimbang jarak
$\sum_{k=p+1}^{p+m} \delta(x_{ik}, x_{jk})$: Ukuran jarak untuk data tipe kategorik

Algoritme *K-Prototypes*

Algoritme K-*Prototypes* adalah salah satu metode yang diperkenalkan pertama kali untuk menerapkan analisis cluster pada data yang berukuran besar dengan data bertipe campuran. Metode ini menggabungkan ukuran jarak pada algoritme K-*means* dan algoritme K-*modes* (Huang Z. , 1997). Menurut (Gan, MA, & Wu, 2007) Tahapan-tahapan algoritme K-*prototype* yaitu yang pertama menentukan banyaknya kelompok (C) yang akan dibentuk. Dimana C0 = 2 dan Ci = i + 2 dengan i = 0, 1, 2,…. , yang kedua yaitu menentukan C *prototype* awal yaitu Z1, Z2,…., ZC sebagai pusat kelompok di masing-masing *cluster*. Selanjutnya melakukan perhitungan jarak semua observasi pada dataset terhadap cluster awal. Tahapan ketiga mengalokasikan semua observasi ke dalam *cluster* yang memiliki jarak terdekat dengan objek yang diukur. Setelah dialokasikan lakukan perhitungan titik pusat *cluster* yang baru dan tahapan terakhir yaitu merealokasikan semua data observasi pada dataset terhadap *prototype* yang baru. Apabila titik pusat klaster tidak berubah dan sudah tidak ada lagi perpindahan objek maka proses algoritme berhenti.

Evaluasi Hasil *Cluster*

Kinerja hasil pengklasteran untuk peubah bertipe numerik bisa diketahui melalui rasio diantara nilai simpangan baku di dalam kelompok (Sw) dan simpangan baku antar kelompok (SB) Rumus Sw dan SB untuk data bertipe numerik adalah sebagai berikut (Bunkers, Miller, & DeGaetano, 1996) :

$$S_w = \frac{1}{C}\sum_{i=1}^{C} S_i \quad ...(2)$$

$$S_B = \left[\frac{1}{C-1}\sum_{i=1}^{C}(\bar{x}_i - \bar{x})^2\right]^{1/2} \quad ...(3)$$

Variabel kategorik diukur dengan cara berbeda pada evaluasi hasklasterisasinya. Menurut Okada (1999) serta Kader dan Perry (2007) rumus S_w dan S_B untuk peubah bertipe kategorik adalah sebagai berikut :

$$Sw = [\frac{1}{n-C}\left[\frac{n}{2} - \frac{1}{2}\sum_{i=1}^{C}\frac{1}{n_i}\sum_{j=1}^{k} n^2{}_{ji}\right]]^{1/2} \quad ...(4)$$

$$S_B = [\frac{1}{C-1}\left[\frac{1}{2}\left(\sum_{i=1}^{c}\frac{1}{n_i}\sum_{j=1}^{k} n^2{}_{ji}\right) - \frac{1}{2n}\sum_{j=1}^{k} n^2{}_{j}\right]]^{1/2} \quad ...(5)$$

Menurut (Hair, Black, Babin, & Anderson, 2010) suatu *Cluster* dikatakan optimal jika keragaman antar kelompoknya semakin besar dan keragaman di dalam kelompoknya semakin kecil. Hal tersebut menunjukan bahwa didalam suatu kelompok semakin homogen dan antar kelompok semakin heterogen. Hasil Penklasterisasi yang baik jika diperoleh nilai rasio terkecil antara S_w dan S_B .

3. Pembahasan dan Diskusi

Pen*clusteran* dengan Algoritme K-*Prototypes*

Dalam algoritme K-*prototypes* yang pertama kali dilakukan yaitu penentuan jumlah *cluster* awal sebelum semua proses dijalankan. Pada penelitian ini, banyak nya *cluster* yang akan digunakan dimulai dari k=2. Data yang diolah menggunakan algoritme ini adalah data dengan 39 peubah kategorik dan 15 peubah numerik yang telah di standarisasikan. Penentuan jumlah *cluster* yang berbeda akan menghasilkan kesimpulan *cluster* yang berbeda juga.

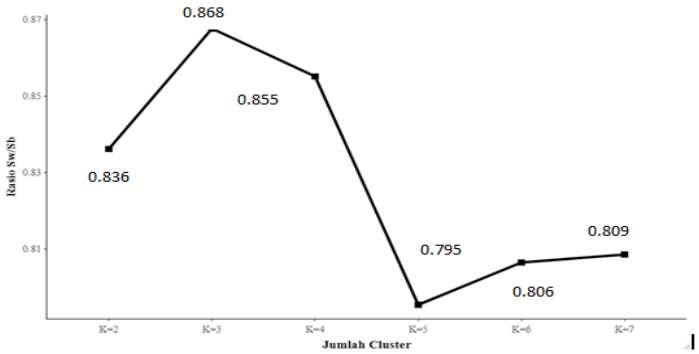
Berdasarkan hasil pengolahan untuk pengukuran jarak , diperoleh koefisien penimbang (γ) sebesar 3.579. Langkah akhir dari algoritme ini ketika titik pusat *cluster* tidak mengalami

perubahan. Distribusi desa yang terbentuk dari hasil akhir algoritme K-Prototypes adalah sebagai berikut:

Tabel 1. Sebaran Hasil Pengelompokan dengan Algoritme *K-Prototypes*

Algoritme *k-prototypes*	Jumlah Desa Pada Setiap Cluster						
	1	2	3	4	5	6	7
k = 2	1115	2583					
k = 3	814	1491	1393				
k = 4	701	1258	1271	468			
k = 5	54	1273	1272	464	635		
k = 6	56	1328	894	365	647	408	
k = 7	54	865	587	295	568	327	1002

Agar semua menggambarkan kondisi sebenarnya maka sangat diperlukan pemilihan jumlah *cluster* optimal. Dimana *cluster* yang optimal adalah *cluster* yang memiliki nilai rasio (S_w/S_b) semakin kecil, karena artinya semakin kecil juga keragaman di dalam *cluster* dan semakin besar keragaman antar *cluster* maka hasilnya akan semakin baik.



Gambar 1. Rasio S_w/S_b Hasil Pengelompokan

Berdasarkan gambar 1 nilai rasio S_w/S_b terkecil ditunjukan pada saat K=5 karena sangat terlihat jelas pada K=5 terjadi penurunan yang sangat drastis maka jumlah *cluster* yang ditetapkan sebagai *cluster* yang optimal adalah K=5. Hasil penclusteran ini memiliki nilai rasio keragaman dalam kelompok terhadap keragaman antar kelompok terkecil yaitu sebesar 0.795.

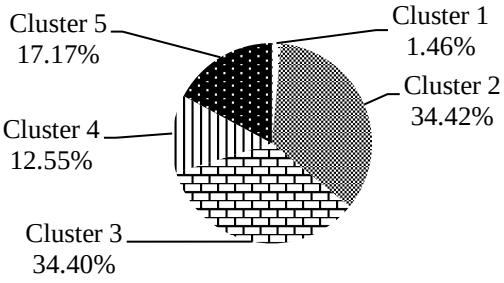
Menentukan Banyaknya Anggota *Cluster*

Penetuan jumlah *cluster* optimum dengan jumlah *cluster* k=5 menunjukan bahwa *cluster* 1, 2 dan 3 memiliki jumlah desa masing-masing sebesar 54 desa, 1273 desa, dan 1272 desa dengan persentase sebesar 1.46%, 34.42% dan 34.40%.

Tabel 2. Sebaran Hasil Pengelompokan

	***Cluster* 1**	***Cluster* 2**	***Cluster* 3**	***Cluster* 4**	***Cluster* 5**
Jumlah Desa	54	1273	1272	464	635
Persentase Desa(%)	1.46%	34.42%	34.40%	12.55%	17.17%

Cluster 4 memiliki jumlah anggota *cluster* sebesar 464 desa atau 12.55% dari seluruh desa- desa di Provinsi Jawa Barat. *Cluster* 5 memiliki jumlah anggota *cluster* sebanyak 635 desa atau 17.17% dari seluruh desa. Apabila digambarkan dalam bentuk *pie chart* adalah sebagai berikut :



Gambar 2. *Pie Chart* Sebaran Hasil Pengelompokan

4. Kesimpulan

Berdasarkan hasil penelitian menggunakan algoritme K-*prototypes* yang telah dibahas sebelumnya maka kesimpulan yang didapatkan adalah sebagai berikut:

1. Pengelompokan desa- desa di Provinsi Jawa Barat berdasarkan indikator indeks desa membangun menggunakan algoritme K-*prototypes* menghasilkan 5 *cluster*. Dimana penentuan hasil *cluster* tersebut diperoleh dari evaluasi hasil *cluster* yang optimal berdasarkan nilai rasio keragaman di dalam *cluster* terhadap keragaman antar *cluster* (Sw/Sb) terkecil yaitu, sebesar 0.795
2. Masing- masing *cluster* memiliki karakteristik yang berbeda. *Cluster* 1 merupakan wilayah dengan pencapaian pembangunan paling rendah dari ke 5 *cluster* lainnya. *Cluster* 1 terdiri dari 54 desa. *Cluster* 2 merupakan wilayah dengan pencapaian pembangunan lebih baik dari *cluster* 1. Yang terdiri dari 1273 desa. *Cluster 3* merupakan wilayah dengan pencapaian pembangunan lebih baik dari *cluster* 1 dan 2. *Cluster* 3 terdiri dari 1272 desa. *Cluster* 4 terdiri dari 464 desa yang merupakan wilayah pencapaian paling baik diantara *cluster* 1, *cluster* 2 dan *cluster* 3. *Cluster* 5 terdiri dari 635 desa yang merupakan wilayah paling baik karena dilihat dari dimensi kesehatan, dimensi pendidikan, dimensi permukiman, dimensi sosial, dimensi ekonomi dan dimensi lingkungan nilai nya lebih baik daripada ke lima *cluster* lainnya.

Acknowledge

Puji dan syukur penulis panjatkan kehadirat Allah SWT karena berkat rahmat dan hidayah-Nya penulis dapat menyelesaikan penelitian ini. Terimakasih kepada ayah, ibu, adik –adik dan kang Iqbal serta keluarga yang selalu mendo'akan dan memberi dukungan baik moral maupun materi kepada penulis. Kepada Bapak Abdul Kudus, S.Si., M.Si., Ph.D yang telah memberi bimbingan kepada penulis hingga penelitian ini selesai. Dosen-dosen Program Studi Statistika Universitas Islam Bandung yang telah banyak memberikan ilmu pengetahuan. Sahabat dan teman-teman serta Semua pihak yang telah hingga penelitian ini selesai.

Daftar Pustaka

[1] Bunkers, M., Miller, J., & DeGaetano, A. (1996). Definition of climate regions in the northen plains using an objective *cluster* modification technique. *Journal of Climate*, 130-146

[2] Gan, G., MA, C., & Wu, J. (2007). *Data Clustering Theory, Algorithms, sand Applications.* Virginia (US): American Statistical Association (ASA).

[3] Hair, J., Black, W., Babin, J., & Anderson, R. (2010). *Multivariate Data Analysis "7th ed"*. New Jersey(US): Pearson Prentice Hall.

[4] Hajarisman, N. (2019). *Statistika Multivariat: Analisis Klaster.* Bandung: NH Press.

[5] Huang, Z. (1998). Extensions to the k-Means Algorithm for *Cluster*ing Large data Sets with Categorical Values . *Data Mining and Knowledge Discovery*, 2833-304.

[6] Kader, G., & Perry, M. (2007). Variability for categorical variables. *Journal of Statistics*, 15(2): 1-16.

[7] Kementerian Desa, PDT dan Transmigrasi. (2019). *Status IDM Indeks Desa Membangun Provinsi Kabupaten Kecamatan Tahun 2019*. Jakarta Selatan: Direktorat Jenderal Pembangunan dan Pemberdayaan Masyarakat Desa.

[8]Okada, T. (1999). Sum of Squares Decompositionfor Categorical Data. *Kwansei Gakuin Studies in Computer Science*, 14:1-6.

[9] Rencher, A. C. (2007). *Methods of Multivariate Analysis Second Edition.* Canada (US): Johm Wiley & Sons.

[10] Yulianto Anggi Priliani, Darwis Sutawanir. (2021). *Penerapan Metode K-Nearest Neighbors (kNN) pada Bearing*. Jurnal Riset Statistika, 1(1), 10-18.