Metode *Hierarchical Density-Based Spatial Clustering of Application with Noise* (HDBSCAN) Pada Wilayah Desa/Kelurahan Tertinggal di Kabupaten Kutai Kartanegara (Studi Kasus : Data Hasil Pendataan Potensi Desa (PODES) Tahun 2018)

Hierarchical Density-Based Spatial Clustering of Application with Noise (HDBSCAN) *Method to Village or Political District in Kutai Kartanegara Regency* (*Case Study* : *Village Potential Data Collection Results* (*PODES*) *in 2018*)

Nanda Anggun Wahyuni[1], Memi Nor Hayati[2], dan Nanda Arista Rizki[3]
[1]Laboratorium Statistika Komputasi FMIPA Universitas Mulawarman
[2]Laboratorium Statistika Terapan FMIPA Universitas Mulawarman
[3]Program Studi Pendidikan Matematika FKIP Universitas Mulawarman
Email : nandanggunw@gmail.com

Abstract

The underdeveloped areas are generally the districts which are relatively underdeveloped compared to other regions on a national scale. Determination of underdeveloped villages is often done in order to determine the distribution of government assistance so that assistance can be distributed appropriately. The identification is based on facilities, infrastructure, access, social, population and economy provided in the Village Potential data (PODES). The concept of grouping based on regional or spatial is done to find out certain characteristics in an area. HDBSCAN is a grouping concept with a parameter called M_{pts}. The purpose of this study is to know the number of clusters formed in the grouping of underdeveloped villages / urban areas in Kutai Kartanegara Regency using the HDBSCAN method. The Mpts parameters that is used in this study is from 2 to 6. Based on the results of the analysis, the clusters formed in the grouping of underdeveloped villages / urban areas in Kutai Kartanegara Regency using the HDBSCAN method, were 3 clusters. Cluster 0 consists of 19 villages / urban areas , cluster 1 consists of 4 villages / urban areas and cluster 2 consists of 61 villages / urban areas. Based on the analysis, villages / urban areas included in cluster 1 could be the main target of the government in providing assistance and development of regional facilities / infrastructure.

Keywords: Spatial, HDBSCAN, Village Potential

Pendahuluan

Data mining adalah proses menggali informasi baru dari sejumlah besar data yang dapat berguna dalam proses pengambilan keputusan. Data spasial adalah gambaran nyata suatu wilayah yang terdapat di permukaan bumi. Proses *data mining* pada sejumlah besar data spasial dikenal sebagai *spatial data mining* (Matheus dkk, 1993).

Spatial data mining adalah bagian dari *data mining* yang merupakan proses menemukan pola tertentu yang sebelumnya tidak dikenal tetapi secara potensial dapat berguna bagi *dataset* spasial yang besar. Sebagian besar penelitian terbaru pada data spasial menggunakan teknik *clustering* dikarenakan data tersebut memberikan informasi tentang karakteristik suatu wilayah yang dibutuhkan dalam *clustering*.

Clustering merupakan pengelompokkan *record*, pengamatan atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. (Suyanto, 2017). Berdasarkan kategori kemiripan, *clustering* terbagi menjadi dua, yaitu metode hierarki (*hierarchical clustering methods*) dan metode non hierarki (*nonhierarchical clustering methods*). Hasil *clustering* dengan metode hierarki secara umum membentuk diagram pohon (*tree diagram*) dan *dendogram* yang menggambarkan pengelompokkan objek berdasarkan jarak. Contoh metode hierarki adalah metode *single linkage*, *average linkage*, *complete linkage*, dan *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN). Kategori kemiripan selanjutnya adalah metode non hierarki. Metode ini disebut juga metode partisi (*partitional methods*).

Kemiskinan merupakan fokus perhatian dalam menyusun strategi pembangunan hampir di semua negara. Strategi pembangunan diprioritaskan bagi daerah-daerah dengan jumlah penduduk miskin terbesar. Kabupaten Kutai Kartanegara tercatat memiliki penduduk miskin terbanyak di Provinsi Kalimatan Timur dengan jumlah penduduk miskin sebanyak 56.570 jiwa pada tahun 2017.

Data mining

Data mining merupakan sebuah langkah dalam proses *Knowledge Discovery in Database* (KDD) yang terdiri dari penerapan analisis data dan penemuan algoritma yang menghasilkan enumerasi tertentu terhadap pola pada data. *Data mining* ditujukan untuk mengekstrak (mengambil intisari) pengetahuan dari sekumpulan data

sehingga didapatkan struktur yang dapat dimengerti manusia serta meliputi basis data dan managemen data, prapemrosesan data, pertimbangan kompleksitas, pascapemrosesan terhadap struktur yang ditemukan, visualisasi, dan *online updating* (Suyanto, 2017).

Fungsi *Data Mining*

Berdasarkan fungsionalitasnya, tugas-tugas *data mining* terbagi ke dalam enam kelompok, yaitu :

1. *Classification* adalah generalisasi struktur yang diketahui untuk diaplikasikan pada data-data baru.
2. *Clustering* adalah mengelompokkan data yang tidak diketahui label kelasnya ke dalam sejumlah kelompok tertentu sesuai dengan ukuran kemiripan datanya.
3. *Regression* menemukan suatu fungsi yang dapat memodelkan sesuatu dengan galat (kesalahan prediksi) seminimal mungkin.
4. *Anomaly detection* digunakan untuk mengidentifikasi data yang tidak umum, bisa berupa *outlier* (pencilan) dan perubahan (deviasi) yang mungkin sangat penting dan perlu invertigasi lebih lanjut.
5. *Association rule learning* digunakan untuk mencari relasi antar variabel.
6. *Summarization* adalah menyediakan representasi data yang lebih sederhana, meliputi visualisasi dan pembuatan laporan.

Operasi Data Mining

Data mining adalah teknik yang relatif cepat dan mudah untuk menemukan pengetahuan, pola dan/atau relasi antar data secara otomatis. Menurut Silitonga (2016), ada enam tahapan proses *data mining*, tahapan proses *data mining* adalah sebagai berikut :

1. Pembersihan data (*data cleaning*)
2. Integrasi data (*data integration*)
3. Transformasi data (*data transformation*)
4. Penggalian data (*data mining*)
5. Evaluasi pola (*pattern evaluation*)
6. Penyajian pola (*knowledge presentation*)

Analisis *Cluster*

Asumsi dalam analisis kelompok yaitu sampel yang diambil harus mewakili populasi (representatif) dan tidak adanya variabel penelitian yang memiliki hubungan linier yang besar dengan variabel lainnya (nonmultikolinieritas) (Gujarati, 2003).

$$r_{x_k x_l} = \frac{n\left(\sum_{p=1}^{n} x_{kp}x_{kl}\right) - \left(\sum_{p=1}^{n} x_{kp}\right)\cdot\left(\sum_{p=1}^{n} x_{kp}x_{kl}\right)}{\sqrt{n\left(\sum_{p=1}^{n} x_{kp}\right)^2 - \left(\sum_{p=1}^{n} x_{kp}\right)^2}\sqrt{n\left(\sum_{p=1}^{n} x_{lp}\right)^2 - \left(\sum_{p=1}^{n} x_{lp}\right)^2}} \quad (1)$$

dimana

$r_{x_k x_l}$: nilai koefisien korelasi antara variabel x ke-k dan ke-l
n : jumlah data

Standarisasi Data

Jika rentang nilai antar variabel memiliki perbedaan skala yang cukup besar yang dapat menyebabkan bias dalam analisis *cluster* maka data asli perlu standarisasi atau normalisasi. Normalisasi data dapat dilakukan dengan cara semua dimensi atau sub-variabel penyusun atau item ditransformasi ke dalam sata standar atau data Z (nilai rata-rata sama dengan nol, variansi sama dengan satu dan data tanpa satuan/relatif) (Sartono, 2003).

$$Z_{p,k} = \frac{x_{pk} - \bar{x}_k}{s_k} \quad (2)$$

dimana

x_{pk} : data ke-p variabel ke-k
$\bar{x}_k$: rata-rata variabel ke-k
$Z_{p,k}$: normalisasi data untuk data ke-p variabel ke-k
s_k : standar deviasi variabel ke-k

Algoritma HDBSCAN

Hierarchical DBSCAN atau HDBSCAN adalah konsep dan algoritmik peningkatan pada algoritma *Ordering Points to Identify the Clustering Structure* (OPTICS) dengan parameter input tunggal disebut M_{pts}. Metode hierarki yang diproduksi oleh HDBSCAN dapat digunakan sebagai basis untuk lainnya seperti tugas pasca-pemrosesan. Salah satu tugas ini adalah membuat sebuah metode hierarki yang ringkas (Syed, 2015).

Adapun tahapan-tahapan dari algoritma HDBSCAN adalah sebagai berikut:

1. Inisialisasi parameter M_{pts} = 2, 3, 4, 5, dan 6.
2. Menentukan titik x_p yang merupakan calon *core point* secara acak.
3. Menghitung semua jarak titik terhadap x_p dengan menggunakan Persamaan (3) sebagai berikut:

$$d_{p,q} = \|x_p - x_q\| = \sqrt{\sum_{k=1}^{n} |x_{pk}x_{qk}|^2} \quad (3)$$

4. Menentukan titik-titik yang *density-reachable* terhadap x_p yang berada dalam $d_{core}(x_p)$. Jika banyaknya titik yang *density-reachable* terhadap x_p lebih dari atau sama dengan M_{pts} maka titik x_p menjadi *core point* sehingga *cluster* terbentuk dan dilanjutkan ke titik lain disekitarnya yang merupakan titik yang *density-reachable* terhadap x_p. Jika titik yang *density-reachable* juga merupakan *core point* maka *cluster* yang terbentuk merupakan *cluster* yang sama dengan *cluster* untuk *core point* sebelumnya. Namun jika titik yang

density-reachable tersebut adalah *border point* maka proses dilanjutkan ke titik selain yang berada dalam *density-reachable*.

5. Menghitung nilai *SC* dari setiap kombinasi parameter M_{pts} yang diberikan dengan menggunakan Persamaan (9).
6. Ulangi langkah 2 sampai dengan 4 untuk titik-titik yang belum memiliki *cluster*.
7. Ekstrak HDBSCAN sebagai *dendogram*.

Validasi Data Hasil *Clustering*

Salah satu metode evaluasi yang dapat digunakan untuk melihat kualitas dan kekuatan *cluster* adalah metode *Silhouette Coefficient*. Metode ini merupakan metode validasi *cluster* yang menggabungkan metode *cohesion* dan *separation*.

1. Menghitung rata-rata jarak dari suatu data ke-p dengan semua data yang berada pada satu *cluster* yang sama dengan menggunakan Persamaan (4).

$$a_p = \frac{1}{n-1} d_{p,q}, q \neq p \quad (4)$$

dimana $p = 0, 1, 2, \ldots, n$.

2. Menghitung rata-rata jarak suatu data ke-p dengan semua data yang berada pada *cluster* yang berbeda dengan menggunakan Persamaan (5) dan diambil nilai terkecilnya.

$$b_p = min\{d_p(p)\}, q \neq p \quad (5)$$

dengan rumus jarak suatu data ke-p dengan semua data pada *cluster* yang berbeda adalah

$$d_p = \frac{1}{n_i} \sum_{q=1}^{n_i} d_{p,q} \quad (6)$$

3. Menghitung nilai *Silhouette Coefficient*

$$SC_1(p) = \frac{b_p - a_p}{max\{a_p, b_p\}}, p = 1,2,\ldots,n \quad (7)$$

Nilai *SC* dari sebuah *cluster* ($SC_2(i)$) didapatkan dengan menghitung rata-rata nilai $SC_1(p)$ semua data yang bergabung dalam *cluster* tersebut dengan menggunakan Persamaan berikut

$$SC_2(i) = \frac{1}{n_i} \sum_{x_p \epsilon C_i} SC_1(p) \quad (8)$$

Setelah itu nilai *SC* global didapatkan dengan menghitung rata-rata nilai $SC_2(i)$ dari semua *cluster* dengan menggunakan Persamaan berikut

$$SC = \frac{\sum_{i=0}^{C} n_i SC_2(p)}{\sum_{i=0}^{C} n_i} \quad (9)$$

dimana

a_p : Rata-rata jarak data ke-p dengan semua data pada satu *cluster* yang sama
b_p : Rata-rata jarak data ke-p dengan semua data pada *cluster* yang berbeda
$SC_1(p)$: Nilai *Silhouette Coefficient* pada data ke-p
$SC_2(i)$: Nilai *Silhouette Coefficient* pada *cluster* ke-i
SC : Nilai *Silhouette Coefficient* global
x_p : Data pengamatan ke-p
C_i : *Cluster* ke-i
n_i : Jumlah data dalam *cluster* ke-i
c : Jumlah *cluster*

Nilai *Silhouette Coefficient* berdasarkan Kaufman dan Rousseuw (1990) yaitu

Tabel 1 Nilai *Silhouette Coefficient*

No.	Rentang Nilai *SC*	Keterangan
1.	$0,7 < SC \leq 1$	*Strong Structure*
2.	$0,5 < SC \leq 0,7$	*Medium Structure*
3.	$0,25 < SC \leq 0,5$	*Weak Structure*
4.	$SC \leq 0,25$	*No Structure*

Desa tertinggal

Penentuan desa tertinggal seringkali dilakukan dalam rangka menetapkan penyaluran bantuan pemerintah agar bantuan tersebut dapat disalurkan dengan tepat. Penetapan status desa tertinggal diharapkan menjadi identifikasi daerah kantong kemiskinan. Daerah tertinggal umumnya adalah daerah yang kondisinya relatif kurang berkembang dibandingkan daerah lain dalam skala nasional. Hal tersebut yang dicerminkan oleh empat faktor yang diduga menjadi penyebab kemajuan atau ketertinggalan suatu desa menurut Bappenas (2006) dalam Agusta (2007) yaitu, faktor alam/lingkungan, faktor kelembagaan, faktor sarana/prasarana, dan akses serta faktor sosial ekonomi penduduk.

Hasil Penelitian dan Pembahasan

Penelitian ini menggunakan data PODES di 84 desa/kelurahan tertinggal yang ada di Kabupaten Kutai Kartanegara pada Tahun 2018. Variabel penelitian adalah 14 variabel yang tersedia di BPS Provinsi Kalimantan Timur, yaitu:

X_1 : Kepadatan penduduk
X_2 : Ketersediaan sarana pendidikan/sekolah
X_3 : Ketersediaan tenaga kesehatan
X_4 : Ketersediaan sarana kesehatan
X_5 : Jumlah berlangganan telepon kabel
X_6 : Jumlah toko/warung kelontong
X_7 : Jumlah kedai makan/minum
X_8 : Jumlah restoran/rumah makan
X_9 : Jumlah tempat ibadah
X_{10} : Jumlah pengguna listrik PLN
X_{11} : Jumlah keluarga tinggal di bantaran tepi sungai
X_{12} : Jumlah keluarga tinggal di permukiman kumuh
X_{13} : Jumlah penderita gizi buruk selama 3 tahun terakhir
X_{14} : Jumlah penerima kartu JAMKESMAS/JAMKESDA

Hasil statistik deskriptif data PODES di 84 desa/kelurahan tertinggal yang ada di Kabupaten Kutai Kartanegara pada Tahun 2018 dapat dilihat pada Tabel 2.

Tabel 2 Statistika Deskriptif

Variabel	*n*	Minimum	Maksimum	Rata-rata
X_1	84	0,1	886,5	57,608
X_2	84	0	6	1,738
X_3	84	0	11	2,167
X_4	84	0	3	0,083
X_5	84	0	175	3,155
X_6	84	0	200	18,1
X_7	84	0	35	2,917
X_8	84	0	2	0,048
X_9	84	1	14	4,012
X_{10}	84	0	1221	327,70
X_{11}	84	0	359	26,19
X_{12}	84	0	65	1,833
X_{13}	84	0	3	0,083
X_{14}	84	4	967	250,82

Berdasarkan Tabel 2 dari kolom *n* terlihat bahwa seluruh variabel penelitian memiliki jumlah amatan yang sama yakni 84 data pengamatan. Kolom rata-rata menunjukkan nilai rata-rata untuk setiap variabel penelitian. Kolom maksimum dan minimum menunjukkan nilai maksimum dan minimum untuk setiap variabel penelitian. Sebagai contoh variabel kepadatan penduduk (X_1) terlihat bahwa rata-rata kepadatan penduduk di 84 desa/kelurahan tertinggal di Kabupaten Kutai Kartanegara adalah 58 jiwa/km^2. Kepadatan tertinggi dan terendah dari 84 desa/kelurahan tertinggal di Kabupaten Kutai Kartanegara yakni sebesar 887 jiwa/km^2 dan 1 jiwa/km^2. Demikian seterusnya untuk data yang lain.

Selanjutnya menentukan jumlah *cluster* yang terbentuk berdasarkan nilai M_{pts} dengan kombinasi nilai 2 sampai 6 dan melihat masing-masing nilai *SC* pada masing-masing nilai M_{pts}. Dari Tabel 3 dapat dilihat bahwa *cluster* 0 menyatakan *outlier*, nomor *cluster* menyatakan *cluster* yang diperoleh sesuai dengan nilai M_{pts} yang ditentukan dan jumlah anggota menyatakan banyaknya anggota *cluster* sesuai dengan nomor *cluster*. Nilai parameter M_{pts} terbaik adalah 4 karena memiliki nilai *SC* terbesar yaitu 0,184 dan terbentuk 2 *cluster*.

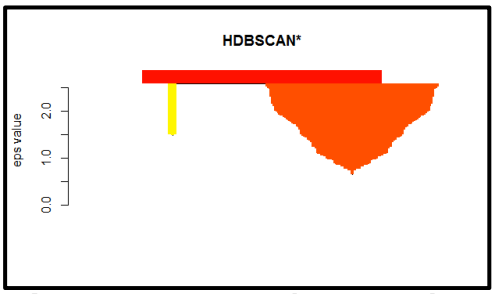
Tabel 4 menunjukkan bahwa desa/kelurahan yang rata-rata jumlah keluarga menggunakan listrik PLN di *cluster* 1 jauh lebih sedikit dibandingkan pada *cluster* 2. Hal ini dapat dilihat dari jumlah rata-rata pengguna listrik PLN pada *cluster* 1 sebanyak 25 keluarga tiap desa/kelurahan, sedangkan jumlah rata-rata pengguna listrik PLN pada *cluster* 2 sebanyak 250 keluarga tiap desa/kelurahan.

Tabel 3 Nilai *SC* dan *Cluster* yang Terbentuk

Nilai M_{pts}	Nilai *SC*	Nomor *Cluster*	Jumlah Anggota
2	0,027	0	31
		1	3
		2	2
		3	2
		4	2
		5	2
		6	3
		7	2
		8	2
		9	2
		10	2
		11	2
		12	3
		13	2
		14	4
		15	10
		16	3
		17	2
		18	2
		19	3
3	0,113	0	25
		1	4
		2	3
		3	52
4	0,184	0	19
		1	4
		2	61
5	NA	0	84
6	NA	0	84

Tabel 4 Nilai Rata-rata Variabel untuk Masing-masing *Cluster*

Variabel	*Cluster* ke-*i*		
	0	1	2
X_1	121,801	2,213	41,246
X_2	3	1	2
X_3	4	1	2
X_4	1	0	0
X_5	14	0	1
X_6	40	4	13
X_7	1	0	0
X_8	6	2	4
X_9	1	0	0
X_{10}	644	25	250
X_{11}	58	0	19
X_{12}	8	0	0
X_{13}	1	1	0
X_{14}	282	87	253



Gambar 1 Diagram pohon (*Dendogram*) menggunakan metode HDBSCAN

Diagram pohon (*dendogram*) seperti pada Gambar 1 terbentuk dengan parameter optimal M_{pts} = 4 dengan rentang nilai *epsilon* yaitu 0,6 sampai dengan 2,5.

Kesimpulan

Berdasarkan hasil penelitian dan pembahasan, maka kesimpulan yang dapat diambil adalah sebagai berikut:

1. *Cluster* yang terbentuk pada pengelompokkan desa/kelurahan tertinggal di Kabupaten Kutai Kartanegara dengan menggunakan metode HDBSCAN adalah sebanyak 2 *cluster*. *Cluster* 1 beranggotakan 4 desa/kelurahan sedangkan *cluster* 2 beranggotakan 61 desa/kelurahan.
2. Hasil pengelompokkan menggunakan metode HDBSCAN menunjukkan bahwa desa/kelurahan yang rata-rata jumlah keluarga menggunakan listrik PLN di *cluster* 1 jauh lebih sedikit dibandingkan pada *cluster* 2. Hal ini dapat dilihat dari jumlah rata-rata pengguna listrik PLN pada *cluster* 1 sebanyak 25 keluarga tiap desa/kelurahan, sedangkan jumlah rata-rata pengguna listrik PLN pada *cluster* 2 sebanyak 250 keluarga tiap desa/kelurahan. Sehingga desa/kelurahan yang masuk dalam *cluster* 1 bisa menjadi sasaran utama pemerintah dalam memberikan bantuan dan pembangunan sarana/prasarana daerah.

Daftar Pustaka

Agusta, I. (2007). Desa Tertinggal di Indonesia. *Jurnal Transdisiplin Sosiologi, Komunikasi dan Ekologi Manusia*, 1(2),233-235.

Badan Pusat Statistik. (2005). *Identifikasi dan Penentuan Desa Tertinggal Tahun 2005*. Jakarta: Badan Pusat Statistik.

Badan Pusat Statistik. (2014). *Pedoman Pencacah PODES 2014*. Jakarta: Badan Pusat Statistik.

Badan Pusat Statistik. (2018). Diakses pada 19 Januari 2019, dari www.bps.go.id: http://www.bps.go.id/dynamictable/2017/08/03/1260/jumlah-penduduk-miskin-menurut-kabupaten-kota-2015---2017.html.

Faqih, A. (2010). *Kependudukan: Teori, Fakta dan Masalah*. Yogyakarta: Dee Publish.

Fayyad, U., Piatetsky-shapiro, G,. & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *Proceeding of the Second International Conference on Knowledge Discovery*, Portland, 82-84.

Han, J., & Kamber, M. (2006). *Data Mining: Concept and Techniques*. San Fransisco: Morgan Kauffman Publisher.

Id, I. D., Mahdiyah, E. (2017). Modifikasi DBSCAN (*Density-Based Spatial Clustering of Application With Noise*) pada Objek 3 Dimensi. *Jurnal Komputer Politeknik Caltex Riau*, 3(1),41-45.

Irabawati, N. (2016). Perbandingan Metode *C-Means* dan *Fuzzy C-Means* (FCM) dalam pengelompokkan Wilayah Desa/Kelurahan di Kabupaten Kutai Kartanegara. *Jurnal EKSPONENSIAL*, 7(1),2-5.

Larose, D.T. (2005). *Discovering Knowledge in Data: Introduction to data mining*. New York: John Willey & Sons.

Mardhiyyah, R. (2014). *Clustering* Dataset Titik Panas dengan Algoritma DBSCAN Menggunakan Web *Framework Shiny* pada Bahasa Pemrograman R. *Jurnal Ilmu Komputer Institut Pertanian Bogor*, 1(2), 1-12