

**IMPLEMENTASI METODE DBSCAN PADA
PENGELOMPOKAN KABUPATEN/KOTA DI PULAU
KALIMANTAN BERDASARKAN INDIKATOR
KESEJAHTERAAN RAKYAT**

Risman^{1*}, Syaripuddin¹, Suyitno¹

¹Jurusan Matematika. Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas
Mulawarman, Indonesia

Corresponding author: rismanstat604@gmail.com

Abstrak. Data Mining (*knowledge discovery*) adalah teknik atau proses untuk mendapatkan informasi dari gudang basis data yang besar. Berdasarkan fungsionalitasnya, salah satu tugas data mining adalah melakukan pengelompokan data, dimana metode yang digunakan dalam penelitian ini adalah metode *Density Based Spatial Clustering Algorithm with Noise* (DBSCAN). Metode pengelompokan DBSCAN diterapkan pada data indikator kesejahteraan rakyat. Tujuan penelitian ini adalah mengetahui berapa cluster yang terbentuk pada pengelompokan menggunakan metode DBSCAN dan mengetahui kombinasi parameter yang optimal berdasarkan nilai Davies Bouldin Index (DBI). Berdasarkan hasil analisis, disimpulkan bahwa kombinasi parameter yang optimal adalah $\varepsilon = 2$ dan MinPts = 2 dengan nilai DBI yang menghasilkan 5 cluster yaitu cluster 0 (titik-titik outlier) beranggotakan 33 Kabupaten/Kota, cluster-1 beranggotakan 2 Kabupaten dengan karakteristik TPT, pengakses internet dan membeli/menerima raskin paling tinggi, cluster-2 beranggotakan 3 kabupaten dengan karakteristik angka pengeluaran per kapita, status kepemilikan rumah sendiri dan korban kejahatan paling tinggi, cluster-3 beranggotakan 2 Kabupaten dengan karakteristik angka Imunisasi, AMH dan penerangan utama listrik PLN paling tinggi dan cluster-4 beranggotakan 2 Kabupaten dengan angka AHH, merokok, jumlah penduduk miskin dan akses air bersih paling tinggi.

Kata Kunci: DBSCAN, DBI, indikator kesejahteraan rakyat.

1 PENDAHULUAN

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi dan pengetahuan dari *database* besar [3]. Salah satu tugas *data mining* berdasarkan fungsionalitasnya adalah melakukan pengelompokan (*clustering*). Analisis kelompok (*cluster analysis*) adalah pengelompokan data (objek) yang didasarkan hanya pada informasi yang ditemukan dalam data yang menggambarkan objek tersebut dan hubungan antar data [4]. Tujuan analisis kelompok adalah agar objek-objek yang bergabung dalam sebuah kelompok merupakan objek-objek yang mirip (atau berhubungan) satu sama lain dan berbeda (atau tidak berhubungan) dengan objek dalam kelompok yang lain [2]. Salah satu metode dalam analisis *cluster* (pengelompokan) adalah metode berbasis kepadatan (*density-based method*) [3].

Density-based method merupakan metode yang dikembangkan berdasarkan *density* (kepadatan) tertentu. Metode ini membentuk *cluster* berdasarkan kepadatan yang cukup tinggi ke dalam *cluster-cluster* dan untuk objek yang tidak masuk ke *cluster* manapun dianggap sebagai *noise*. Sedangkan Salah satu metode berbasis kepadatan (*density-based method*) adalah metode *Density-Based Spatial Clustering Algorithm with Noise* (DBSCAN). Metode ini merupakan sebuah metode *clustering* yang membangun area berdasarkan kepadatan yang terkoneksi (*density-connected*). Setiap objek dari sebuah radius area (*cluster*) harus mengandung setidaknya sejumlah minimum data yang disimbolkan dengan ϵ (radius) dan *MinPts* (titik minimum) [3].

DBSCAN dapat digunakan untuk mengelompokkan berbagai macam aspek kehidupan, salah satunya aspek sosial yaitu kesejahteraan rakyat. Istilah kesejahteraan erat kaitannya dengan tujuan Negara Indonesia yaitu untuk menjamin dan memajukan kesejahteraan umum itu sendiri, dimana bebas dari kemiskinan merupakan salah satu indikator kesejahteraan. Hidup bebas dari kemiskinan atau menikmati kehidupan yang layak merupakan hak asasi setiap warga negara yang menjadi tugas pemerintah yang diwujudkan dalam pembangunan nasional. Diketahui bahwa beberapa Provinsi di Pulau Kalimantan tahun ke tahun selalu mengalami kenaikan jumlah penduduk miskin, tercatat di Badan Pusat Statistik (BPS) Provinsi Kalimantan Timur pada Maret 2017 sebanyak 220070 jiwa (6.19 %). Sedangkan pada september 2016 sebanyak 211240 (6.00 %), yang berarti bahwa jumlah penduduk miskin di Provinsi Kalimantan Timur mengalami kenaikan sebanyak 8930 jiwa (naik 0.19 %). Badan pusat statistik mengamati kesejahteraan rakyat dari berbagai aspek spesifik, yaitu kependudukan, kesehatan, pendidikan, ketenagakerjaan, konsumsi rumah tangga dan kemiskinan, serta perumahan.

Berdasarkan uraian di atas, maka dilakukan penelitian untuk mengelompokkan Kabupaten/Kota di Pulau Kalimantan Berdasarkan Indikator Kesejahteraan Rakyat dengan menerapkan metode DBSCAN. Adapun tujuan dalam penelitian ini adalah untuk mengetahui berapa cluster yang terbentuk pada pengelompokan menggunakan metode DBSCAN dan mengetahui kombinasi parameter yang optimal berdasarkan nilai Davies Bouldin Index (DBI).

2 SUMBER DATA DAN METODOLOGI

2.1 Sumber Data

Data Penelitian ini adalah data sekunder yaitu data Indikator Kesejahteraan Rakyat di 42 Kabupaten/Kota yang ada di Pulau Kalimantan tahun 2017. Variabel dalam data penelitian ini yaitu sebanyak 16 variabel yang tersedia di 4 BPS di Pulau Kalimantan yaitu BPS Provinsi Kalbar, BPS Provinsi Kalsel, BPS Provinsi Kaltim dan BPS Provinsi Kaltara.

Dalam Penelitian ini digunakan 16 variabel yaitu Kepadatan Penduduk (X_1), Angka Harapan Hidup (AHH) (X_2), persentase rumah tangga yang menerima jaminan kesehatan (X_3), persentase penduduk usia 5 tahun ke atas yang merokok (X_4), Balita yang menerima imunisasi lengkap (X_5), Angka Melek Huruf (AMH) (X_6), Tingkat Partisipasi Angkatan Kerja (TPAK) (X_7), Tingkat Pengangguran Terbuka (TPT) (X_8), Pengeluaran Per Kapita (X_9), jumlah penduduk miskin (X_{10}), persentase rumah tangga yang memiliki status Kepemilikan Rumah Sendiri (X_{11}), persentase rumah Tangga yang menggunakan listrik PLN (X_{12}), persentase rumah tangga yang memiliki akses air bersih (X_{13}), persentase rumah tangga yang mengakses internet (X_{14}), persentase penduduk yang menjadi korban kejahatan (X_{15}) dan persentase rumah tangga yang membeli/menerima raskin (X_{16}).

2.2 Teknik Analisis Data

Langkah-langkah analisis data dalam penelitian ini adalah sebagai berikut:

1. Menginput data.
2. Menormalisasi data pengamatan
3. Mendeteksi non multikolinieritas data dengan cara melihat nilai koefisien korelasi antar variabel penelitian. Jika terdapat koefisien korelasi bernilai di atas 0,8 antar variabel penelitian yang berbeda, maka dilakukan reduksi variabel, jika tidak maka dilanjutkan ke langkah berikutnya.
4. Pengelompokkan data dengan menerapkan metode DBSCAN. Tahapannya secara rinci adalah sebagai berikut:
 - a. Inisialisasi parameter ϵ dan $MinPts$.
 - b. Menentukan titik p .
 - c. Menghitung semua jarak titik terhadap p .
 - d. Menentukan titik-titik yang *density-reachable* terhadap p yang berada dalam radius ϵ . Jika banyaknya titik yang *density-reachable* terhadap p lebih dari atau sama dengan $MinPts$ maka titik p adalah *core point* sehingga *cluster* terbentuk dan dilanjutkan ke titik lain disekitarnya yang merupakan titik yang *density-reachable* terhadap p . Jika p adalah *border point* dan tidak ada titik yang *density-reachable* terhadap p , maka proses dilanjutkan ke titik yang lain.
 - e. Ulangi langkah b sampai dengan d untuk kombinasi parameter ϵ dan $MinPts$ yang berbeda.
 - f. Menghitung nilai DBI dari setiap kombinasi parameter ϵ dan $MinPts$ yang berbeda.
 - g. Menentukan kombinasi parameter ϵ dan $MinPts$ yang optimal berdasarkan nilai DBI terkecil.

- Hasil pengelompokkan dianalisis lebih lanjut dengan melakukan interpretasi kelompoknya berdasarkan karakteristik pada setiap kelompok yang terbentuk..

3 HASIL DAN PEMBAHASAN

3.1 Normalisasi Data

Perbedaan nilai yang cukup besar antar nilai variable dapat menyebabkan bias dalam analisis *cluster*, maka data asli perlu standarisasi atau normalisasi. Rentang nilai yang besar akan menyebabkan perhitungan jarak menjadi tidak valid [6]. Cara menentukan nilai normalisasi adalah sebagai berikut [2]

$$z_{pa} = \frac{x_{pa} - \bar{x}_a}{S_a}, p = 1, 2, \dots, n \text{ dan } a = 1, 2, \dots, \ell, \quad (1)$$

Dari hasil perhitungan menggunakan Persamaan (1), didapatkan hasil sebagai berikut:

Tabel 1: Hasil Normalisasi Data

No	z_1	z_2	z_3	...	z_{16}
1	-0,274	-0,665	-0,970	...	0,566
2	-0,296	0,528	-1,017	...	-0,309
3	-0,301	0,303	-0,566	...	1,052
⋮	⋮	⋮	⋮	⋮	⋮
42	0,272	0,727	0,761	...	-0,946

3.2 Pendeteksian Multikolinieritas

Pendeteksian multikolinieritas yaitu dengan melihat nilai koefisien korelasi antar variabel menggunakan Persamaan (2) [1].

$$r_{x_a, x_b} = \frac{n \left(\sum_{p=1}^n x_{pa} x_{pb} \right) - \left(\sum_{p=1}^n x_{pa} \right) \cdot \left(\sum_{p=1}^n x_{pb} \right)}{\sqrt{n \left(\sum_{p=1}^n x_{pa}^2 \right) - \left(\sum_{p=1}^n x_{pa} \right)^2} \cdot \sqrt{n \left(\sum_{p=1}^n x_{pb}^2 \right) - \left(\sum_{p=1}^n x_{pb} \right)^2}}, p = 1, 2, \dots, n \quad (2)$$

Nilai koefisien korelasi antar variabel diperoleh dengan *software R Studio*. Berdasarkan *output software R Studio* diperoleh bahwa nilai mutlak dari koefisien korelasi antar variabel penelitian yang tertinggi adalah 0,781 dimana nilai tersebut masih dibawah 0,8 yang artinya tidak ada multikolinieritas antar variabel dalam penelitian dan dapat dilanjutkan ke proses pengelompokkan dengan metode K-Means dan DBSCAN.

3.3 Metode DBSCAN

Penentuan parameter ϵ dan *MinPts* pada metode DBSCAN bersifat *trial and error*. Pada penelitian ini nilai parameter ϵ dicoba dari 2 sampai dengan 8 (beda = 0.5) dan nilai parameter *MinPts* dicoba dari 2 sampai dengan 8 (beda = 1). Sehingga untuk keseluruhan kombinasi parameter adalah sebanyak 91 kombinasi.

Selanjutnya dilakukan perhitungan jarak seluruh data pengamatan menggunakan Persamaan (3) [2]

$$d(p, q) = \|\mathbf{x}_p - \mathbf{x}_q\|_2 = \sqrt{\sum_{a=1}^{\ell} (x_{pa} - x_{qa})^2}, p, q=1, 2, \dots, n \quad (3)$$

dan menentukan keanggotaan *cluster* berdasarkan hasil pengelompokan 91 kombinasi parameter.

Langkah selanjutnya, dilakukan penentuan parameter optimal dengan melihat nilai DBI. Penentuan parameter ϵ dan *MinPts* yang optimal pada metode DBSCAN dapat diketahui dengan melihat nilai DBI terkecil yang dihasilkan dari masing-masing kombinasi parameter ϵ dan *MinPts*. Sebagai contoh kombinasi parameter pertama yaitu $\epsilon = 2$ dan *MinPts* = 2. Nilai $\epsilon = 2$ dapat diartikan bahwa jarak antara titik p dengan titik q disekitarnya yang *density-reachable* terhadap p masih berada dalam radius $\epsilon = 2$ dan banyaknya titik q lebih dari atau sama dengan 2. Misalkan p adalah z_1 . Jika jarak antara titik z_1 dengan titik lainnya memenuhi $\epsilon = 2$ dan *MinPts* ≥ 2 maka titik z_1 menjadi titik pusat *cluster*. Jika banyaknya titik yang *density-reachable* terhadap titik z_1 adalah 0 atau 1 dengan radius $\epsilon = 2$ maka titik z_1 adalah titik *outlier* dan dilanjutkan untuk titik lainnya sampai dengan z_n .

3.4 Validasi Cluster

Validasi *cluster* yang digunakan adalah *Davies Bouldin Index* (DBI), dimana pengelempokan terbaik adalah pengelompokan yang menghasilkan nilai DBI yang mendekati 0 (non-negatif ≥ 0) [3]. Setelah dilakukan pengelompokan menggunakan metode DBSCAN, kemudian dilakukan validasi *cluster* dengan melihat nilai DBI untuk masing-masing kombinasi parameter sebagai berikut

Tabel 2: Nilai DBI untuk Seluruh Kombinasi Parameter ϵ dan *MinPts*

ϵ	<i>MinPts</i>						
	2	3	4	5	6	7	8
2	0,605	NA	NA	NA	NA	NA	NA
2.5	0.847	0.841	NA	NA	NA	NA	NA
3	1.162	1.162	1.166	1.192	0.992	NA	NA
3.5	1.358	1.358	1.358	1.358	1.363	1.337	1.251
4	NA	NA	NA	NA	1,358	1,358	1,358
4.5	NA	NA	NA	NA	NA	NA	NA
5	NA	NA	NA	NA	NA	NA	NA
5.5	NA	NA	NA	NA	NA	NA	NA
6	NA	NA	NA	NA	NA	NA	NA
6.5	NA	NA	NA	NA	NA	NA	NA
7	NA	NA	NA	NA	NA	NA	NA
7.5	NA	NA	NA	NA	NA	NA	NA
8	NA	NA	NA	NA	NA	NA	NA

Berdasarkan Tabel 2 dapat disimpulkan bahwa kombinasi parameter ϵ dan $MinPts$ yang paling optimal adalah $\epsilon = 2$ dan $MinPts = 2$ yang menghasilkan nilai DBI terkecil yaitu sebesar 0.605. Jika nilai ϵ yang digunakan di atas 4 ($\epsilon = 4$) nilai DBI yang dihasilkan adalah *NA (Not Available)*. Adapun hasil pengelompokkannya dapat dilihat pada Tabel 3.

Tabel 3: Hasil Pengelompokkan dengan Metode DBSCAN

<i>Cluster</i>	Kode Desa/Kelurahan	Nama Desa/Kelurahan
0 (<i>Outlier</i>)	1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 21, 22, 24, 25, 26, 27, 29, 30, 32, 33, 34, 35, 36, 37, 39, 40, 41 dan 42	Sambas, Bengkayang, Landak, Sanggau, Ketapang, Sintang, Kapuas Hulu, Sekadau, Melawi, Kayong Utara, Pontianak, Singkawang, Tanah Laut, Kotabaru, Hulu Sungai Tengah, Hulu Sungai Utara, Tanah Bambu, Balangan, Banjarmasin, Banjar Baru, Kutai Barat, Kutai Kartanegara, Berau, Penajam Paser Utara, Mahakam Ulu, Balikpapan, Samarinda, Bontang, Malinau, Bulungan, Tanah Tidung, Nunukan dan Tarakan
1	4 dan 12	Mempawah dan Kubu Raya
<i>Cluster</i>	Kode Desa/Kelurahan	Nama Desa/Kelurahan
2	17, 18 dan 20	Banjar, Barito Kuala dan Hulu Sungai Selatan
3	19 dan 23	Tapin dan Tabalong
4	28 dan 31	Paser dan Kutai Timur

4 KESIMPULAN

Dari hasil analisis *cluster* menggunakan metode DBSCAN didapatkan bahwa kombinasi parameter yang optimal adalah $\epsilon = 2$ dan $MinPts = 2$ dengan nilai DBI sebesar 0.605 yang menghasil 5 *cluster*. *Cluster* 0 (titik-titik *outlier*) beranggotakan 33 Kabupaten/Kota dimana data pengamatan yang masuk *cluster* ini merupakan data yang memiliki karakteristik sendiri dan tidak bisa di kategorikan ke *cluster* tertentu dan disebut sebagai *noise* atau *outlier*, *cluster* 1 beranggotakan 2 Kabupaten dengan karakteristik TPT, pengakses internet dan membeli/menerima raskin paling tinggi serta, *cluster* 2 beranggotakan 3 kabupaten dengan karakteristik angka pengeluaran per kapita, status kepemilikan rumah sendiri dan korban kejahatan paling tinggi serta, *cluster* 3 beranggotakan 2 Kabupaten dengan karakteristik angka Imunisasi, AMH dan penerangan utama listrik PLN paling tinggi serta dan *cluster* 4 beranggotakan 2 Kabupaten dengan angka AHH, merokok, jumlah penduduk miskin dan akses air bersih paling tinggi serta.

DAFTAR PUSTAKA

- [1] Gujarati, D & Dawn C. Porter. (2010). *Basic Econometrica. Fifth Edition*. New York : Mc Graw Hill.
- [7] Prasetyo, E. (2012). *Data Mining: Konsep dan Aplikasi menggunakan Matlab*. Yogyakarta: Andi Offset.

- [8] Suyanto. (2017). *Data Mining untuk Klasifikasi dan Klusterisasi Data*. Bandung: Informatika.
- [9] Tan, P. (2006). *Introduction to Data Mining*. Boston: Pearson Education.
- [10] Turban, Efraim dkk. (2005). *Decision Support Systems and Intelligent Systems*. Yogyakarta: ANDI.
- [11] Yulianto, S & Hidayatullah, K. H. (2014). Analisis Kluster untuk Pengelompokan Kabupaten/Kota di Provinsi Jateng berdasarkan Indikator Kesejahteraan Rakyat. *Jurnal Akd. Statistika Muhammadiyah Semarang*, 2(1), 58-59.