

PENGUNAAN METODE *FUZZY C-MEANS* UNTUK PENGELOMPOKAN PROVINSI DI INDONESIA BERDASARKAN INDIKATOR KESEHATAN LINGKUNGAN

WIDYA WIJAYANTI, IZZATI RAHMI HG*, FERRA YANUAR

*Program Studi S1 Matematika,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Andalas,
Kampus UNAND Limau Manis Padang, Indonesia.
email : widyawijaya06@gmail.com, izzatirahmihg@sci.unand.ac.id, ferrayanuar@sci.unand.ac.id*

Diterima 15 Desember 2020 Direvisi 29 Desember 2020 Dipublikasikan 12 Januari 2021

Abstrak. Analisis Klaster (*cluster analysis*) memiliki tujuan untuk mengelompokkan objek-objek berdasarkan karakteristik yang dimiliki. Pengelompokan dengan analisis klaster dibagi menjadi dua metode yaitu metode berhirarki dan metode tak berhirarki. Salah satu metode tak berhirarki adalah Metode *Fuzzy C-Means*. Metode *Fuzzy C-Means* merupakan suatu metode yang mempertimbangkan derajat keanggotaan dan himpunan *fuzzy* sebagai dasar pembobotan. Untuk jumlah klaster optimum didapatkan dengan menggunakan kriteria indeks validitas *Partition Entropy Index (PE)*. Pada penelitian ini Metode *Fuzzy C-Means* digunakan untuk mengelompokkan provinsi-provinsi yang ada di Indonesia berdasarkan indikator kesehatan lingkungan. Indikator kesehatan lingkungan yang digunakan adalah STBM, tatanan kawasan sehat, air minum layak rumah tangga, TPTU yang memenuhi syarat kesehatan, TPM yang memenuhi syarat kesehatan, sanitasi layak rumah tangga dan rumah tangga layak huni. Jumlah klaster optimum diperoleh sebanyak 2 klaster yang memiliki nilai *partition entropy* 0,4633829. Klaster 1 terdiri dari 22 provinsi dan klaster 2 terdiri dari 12 provinsi.

Kata Kunci: *Fuzzy C-Means*, Indikator Kesehatan Lingkungan, *Partition Entropy Index*

1. Pendahuluan

Menurut WHO, kesehatan lingkungan meliputi seluruh faktor fisik, kimia, dan biologi dari luar tubuh manusia dan segala faktor yang dapat mempengaruhi perilaku manusia. [3] Sebagai salah satu negara berkembang dengan jumlah penduduk 266,91 juta jiwa (tahun 2019), masalah kesehatan lingkungan di Indonesia terutama di kota-kota besar menjadi sangat kompleks. Penyakit yang diakibatkan oleh kondisi lingkungan masih terus menerus terjadi di Indonesia.

Indikator yang mempengaruhi kesehatan lingkungan tiap provinsi di Indonesia berbeda-beda. Namun masih ada provinsi yang memiliki kesamaan terhadap indika-

*penulis korespondensi

tor yang mempengaruhi. Sehingga pengelompokan provinsi-provinsi di Indonesia dapat dibentuk menjadi beberapa kluster berdasarkan kesamaan karakteristik yang dimiliki. Salah satu metode yang sering digunakan dalam pengelompokan adalah analisis kluster (*cluster analysis*).

Analisis kluster dibagi menjadi dua metode yaitu metode berhirarki dan metode tak berhirarki. Pada metode tak berhirarki, salah satu metode sederhana yang populer digunakan adalah metode *K-Means cluster* yang masih menggunakan teknik *hard clustering* dimana suatu objek hanya menempati satu kluster saja. Selanjutnya berkembang sebuah metode pengklasteran baru yang mempertimbangkan derajat keanggotaan dan himpunan fuzzy sebagai dasar pembobotannya yang disebut dengan Metode *Fuzzy C-Means*.

2. Landasan Teori

2.1. Analisis Kluster

Analisis kluster adalah teknik peubah ganda yang mempunyai tujuan utama untuk mengelompokkan objek berdasarkan kemiripan karakteristik yang dimilikinya. Karakteristik objek-objek dalam suatu kluster memiliki tingkat kemiripan yang tinggi, sedangkan karakteristik antar objek pada suatu kluster dengan kluster lain memiliki tingkat kemiripan yang rendah. Dengan kata lain, keragaman dalam suatu kluster minimum sedangkan keragaman antar kluster maksimum. [5] Perbedaan antara kluster satu dengan kluster yang lain diukur dengan menggunakan sistem jarak. Ukuran jarak yang digunakan pada penelitian ini adalah jarak Euclidean. Jarak Euclidean dirumuskan sebagai berikut: [2]

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2} \quad (2.1)$$

dimana d_{ij} adalah jarak antara objek ke- i dengan objek ke- j , x_{ik} adalah nilai objek ke- i pada peubah ke- k , x_{jk} adalah nilai objek ke- j pada peubah ke- k dan p adalah banyak variabel yang diamati.

Pengelompokan dengan analisis kluster dibagi menjadi dua metode, yaitu: [2]

1. *Metode Berhirarki* digunakan untuk mengelompokkan objek - objek pengamatan secara terstruktur berdasarkan kemiripan sifat yang dimiliki dan kluster yang diinginkan belum diketahui jumlahnya. Teknik pengelompokan hirarki dilakukan dengan salah satu dari dua cara berikut, yaitu penggabungan metode hirarki (*Agglomerative hierarchical methods*) dan Pemisahan metode hirarki (*Divisive hierarchical methods*) yang dapat ditampilkan dalam bentuk dendrogram. Pada penggabungan metode berhirarki terdapat beberapa metode yang cukup terkenal yaitu *single linkage*, *complete linkage*, *average linkage* dan *metode ward*.
2. *Metode tak berhirarki* digunakan untuk mengelompokkan objek-objek pengamatan dimana banyaknya kluster yang akan dibentuk dapat ditentukan terlebih dahulu sebagai bagian dari prosedur pengklasteran. Salah satu metode pengelompokan tak berhirarki adalah Metode *Fuzzy C-Means*.

2.2. Fuzzy C-Means

Fuzzy C-Means (FCM) adalah suatu teknik pengklasteran data yang mana keberadaan tiap-tiap data dalam suatu klaster ditentukan oleh derajat keanggotaan. Teknik ini pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981. [4]

Berikut adalah algoritma *fuzzy c-means* :

- (1) Input data yang akan diklaster X , berupa matriks berukuran $n \times m$ dimana n adalah jumlah sampel data, m adalah atribut setiap data dan X_{ij} adalah data sampel ke- i ($i = 1, 2, \dots, n$), atribut ke- j ($j = 1, 2, \dots, m$).
- (2) Tentukan jumlah klaster (c), pangkat pembobot ($w > 1$), error terkecil yang diharapkan (ξ), fungsi objektif awal ($P_0 = 0$) dan iterasi awal $t = 1$.
- (3) Bangkitkan bilangan random μ_{ik} , $i = 1, 2, \dots, n$; $k = 1, 2, \dots, c$; sebagai elemen-elemen matriks partisi awal U . Matriks partisi awal (U) pada pengelompokan fuzzy memenuhi kondisi sebagai berikut:

$$\mu_{ik} = [0, 1]; 1 \leq i \leq n; 1 \leq k \leq c$$

$$\sum_{i=1}^n \mu_{ik} = 1; 1 \leq i \leq n$$

$$0 < \sum_{k=1}^c \mu_{ik} < n. 1 \leq k \leq c$$

- (4) Hitung pusat *cluster* ke- k : $V_{kj} = V_{k1}, V_{k2}, \dots, V_{km}$ dengan $k = 1, 2, \dots, c$; dan $j = 1, 2, \dots, m$.

$$V_{kj} = \frac{\sum_{i=1}^n (\mu_{ik})^w (X_{ij})}{\sum_{i=1}^n (\mu_{ik})^w} \quad (2.2)$$

- (5) Hitung fungsi objektif pada iterasi ke- t (P_t).

$$P_t = \sum_{i=1}^n \sum_{k=1}^c ([\sum_{j=1}^m (X_{ij} - V_{kj})^2] (\mu_{ikt})^w) \quad (2.3)$$

- (6) Hitung perubahan matriks partisi

$$\mu_{ik} = \frac{[\sum_{j=1}^m (X_{ij} - V_{kj})^2]^{\frac{-1}{w-1}}}{\sum_{k=1}^c [\sum_{j=1}^m (X_{ij} - V_{kj})^2]^{\frac{-1}{w-1}}} \quad (2.4)$$

- (7) Cek kondisi berhenti, jika $|P_t - P_{t-1}| < \xi$ maka iterasi berhenti, jika tidak maka $t = t + 1$ dan mengulang kembali ke langkah 4. [4]

2.3. Partition Entropy Index (PE)

Partition Entropy Index merupakan suatu indeks validitas yang hanya mengukur kekaburan (*fuzzyness*) dari partisi kelompok. Persamaan indeks PE dapat dituliskan sebagai berikut [7]:

$$PE(c) = -\frac{1}{n} \sum_{i=1}^c \sum_{k=1}^n \mu_{ik} \ln \mu_{ik} \quad (2.5)$$

dengan n adalah banyak objek penelitian, c adalah banyak klaster, dan μ_{ik} adalah nilai keanggotaan objek ke- k dengan pusat kelompok ke- i . Indeks ini memiliki rentang 0 sampai $\ln(c)$. Indeks PE yang semakin kecil menunjukkan pengelompokan yang lebih baik.

3. Metode Penelitian

Data yang digunakan dalam penelitian ini merupakan data yang bersumber dari data Kementerian Kesehatan Republik Indonesia pada Profil Kesehatan 2019 tentang indikator kesehatan lingkungan tahun 2019 dan data Badan Pusat Statistik tentang Perumahan tahun 2019.

Variabel-variabel yang digunakan pada penelitian ini adalah persentase desa/kelurahan melaksanakan Sanitasi Total Berbasis Masyarakat (STBM) (X_1), persentase Kab/Kota melaksanakan tatanan kawasan sehat (X_2), persentase jumlah air minum layak rumah tangga (X_3), Persentase Tempat-Tempat Umum (TTU) yang memenuhi syarat kesehatan (X_4), persentase Tempat Pengolahan Makanan (TPM) yang memenuhi syarat kesehatan (X_5), persentase Sanitasi layak rumah tangga (X_6) dan persentase rumah layak huni (X_7).

Tahap-tahap yang dilakukan adalah :

- (1) Menyiapkan data yang akan digunakan dalam pengklasteran.
- (2) Menghitung statistika deskriptif untuk setiap variabel.
- (3) Memeriksa korelasi antar variabel.
- (4) Menganalisis banyaknya klaster awal yang mungkin dibentuk pada dendrogram Metode Hirarki.
- (5) Menentukan jumlah klaster (c), pangkat pembobot ($w = 2$), galat terkecil yang diharapkan ($\xi = 10^{-5}$), fungsi objektif awal ($P_0 = 0$) dan iterasi awal ($t = 1$).
- (6) Membangkitkan derajat keanggotaan awal.
- (7) Menghitung pusat klaster, menghitung fungsi objektif iterasi $ke - t$, serta memperbaharui nilai derajat keanggotaan.
- (8) Proses iterasi berhenti untuk nilai $|P_t - P_{t-1}| < \xi$, jika tidak demikian maka ulangi langkah ke-7.
- (9) Menentukan jumlah klaster optimum dengan menggunakan indeks validitas *Partition Entropy Index (PE)*
- (10) Menentukan anggota klaster optimum menggunakan derajat keanggotaan berdasarkan nilai tertinggi.
- (11) Interpretasi hasil pengelompokan dan karakteristik klaster.

4. Pembahasan

4.1. Jumlah Klaster Optimum

Jumlah klaster optimum ditentukan dengan melihat indeks validitas. Indeks validitas yang digunakan adalah *Partition Entropy*. Jumlah klaster optimum apabila memiliki nilai indeks *Partition Entropy* yang terkecil.

Berdasarkan Tabel 1 dapat dilihat bahwa jumlah klaster yang optimum diberikan ketika jumlah klaster = 2 dengan nilai *Partition Entropy* lebih kecil

Tabel 1. Penentuan Klater Optimum

Klaster	Iterasi	Fungsi Objektif	PE
2	25	32650.106670	0.4633829
3	64	20939.034505	0.8150949
4	69	15427.129253	1.028335
5	78	12136.867758	1.190519
6	82	10140.043690	1.36586
7	115	8514.666073	1.390845
8	92	7446.525271	1.455882

dibanding klaster lain. Banyak Iterasi yang diperoleh adalah 25 dengan nilai fungsi objektifnya 32650,106670.

4.2. Anggota Klaster Optimum

Anggota klaster Optimum ditentukan dari derajat keanggotaan klaster pada iterasi terakhir. Berikut diberikan derajat keanggotaan dari dua klaster provinsi-provinsi di Indonesia:

Tabel 2. Derajat Keanggotaan Klaster Optimum

No	Provinsi	Klaster 1	Klaster 2
1.	Aceh	0.14138541	0.85861459
2.	Sumatera Utara	0.30100226	0.69899774
3.	Sumatera Barat	0.90353391	0.09646609
4.	Riau	0.96446564	0.03553436
...	...	...	...
31.	Maluku	0.09045430	0.90954570
32.	Maluku Utara	0.12285563	0.87714437
33.	Papua Barat	0.13661872	0.86338128
34.	Papua	0.24336862	0.75663138

Derajat keanggotaan terbesar menunjukkan bahwa kecenderungan tertinggi untuk masuk menjadi anggota klaster tersebut. Sebagai contoh provinsi Aceh, dapat menjadi:

- (a) Anggota klaster ke-1 dengan derajat keanggotaan 0,14138541
- (b) Anggota klaster ke-2 dengan derajat keanggotaan 0,85861459

Berdasarkan uraian, dapat dilihat bahwa nilai derajat keanggotaan terbesar provinsi Aceh terletak pada klaster $ke - 2$, sehingga provinsi Aceh merupakan anggota dari klaster ke-2.

4.3. Hasil Pengelompokan Menggunakan Metode Fuzzy C-Means

Hasil pengelompokan provinsi di Indonesia berdasarkan indikator yang mempengaruhi kesehatan lingkungan dapat dilihat pada tabel berikut:

Tabel 3. Hasil Pengelompokan Provinsi-Provinsi di Indonesia

Klaster	Provinsi
Klaster 1	Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur, DI Yogyakarta, Banten, Bali Kalimantan Timur, Kalimantan Utara, Nusa Tenggara Barat, Kalimantan Selatan, Sulawesi Utara, Sulawesi Selatan, Gorontalo
Klaster 2	Aceh, Sumatera Utara, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan tengah, Sulawesi tengah, Sulawesi Tenggara, Sulawesi Barat Maluku, Maluku Utara, Papua Barat, Papua

Setelah terbentuk klaster, selanjutnya akan diambil rata-rata dari masing-masing variabel indikator kesehatan lingkungan. Dimana faktor-faktor tersebut yaitu Sanitasi Total Berbasis Masyarakat (STBM), kabupaten/kota penyelenggara Tatanan Kawasan Sehat, rumah tangga yang memiliki akses terhadap sumber air minum layak, TTU yang memenuhi syarat kesehatan, TPM yang memenuhi syarat kesehatan, rumah tangga yang layak huni dan sanitasi layak yang dilambangkan dengan ($\bar{X}$).

Selanjutnya untuk masing-masing klaster juga akan diambil rata-rata untuk Sanitasi Total Berbasis Masyarakat (STBM), kabupaten/kota yang menyelenggarakan Tatanan Kawasan Sehat, rumah tangga yang mempunyai akses terhadap sumber air minum layak, TTU yang memenuhi syarat-syarat kesehatan, TPM yang memenuhi syarat kesehatan, rumah tangga yang layak huni dan sanitasi layak yang dilambangkan dengan ($\bar{X}_c$). Setiap variabel didalam klaster diberi tanda, jika $\bar{X}_c > \bar{X}$ maka diberi tanda positif (+), sedangkan jika $\bar{X}_c < \bar{X}$ maka diberi tanda negatif (-). Nilai $\bar{X}$ dan $\bar{X}_c$ untuk masing-masing peubah dapat dilihat pada Lampiran 3. Adapun karakteristik hasil pengelompokan disajikan pada Tabel 4.

Tabel 4. Karakteristik Hasil Pengelompokan Menggunakan Fuzzy C-Means

Klaster	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Klaster 1	+	+	+	+	+	+	+
Klaster 2	-	-	-	-	-	-	-

Berdasarkan Tabel 4, klaster 1 beranggotakan 22 provinsi. Pada klaster 1 dapat dilihat bahwa setiap variabel memiliki nilai rata-rata lebih besar daripada nilai rata-rata di Indonesia. Artinya, provinsi yang menjadi anggota klaster 1 memiliki

tingkat kesehatan lingkungan yang cukup baik. Sedangkan pada kluster 2 yang beranggotakan 12 provinsi, setiap variabel memiliki nilai rata-rata yang lebih kecil dibanding rata-rata di Indonesia. Artinya provinsi yang menjadi anggota kluster 2 memiliki tingkat kesehatan lingkungan yang rendah. Sehingga perlu menjadi prioritas pemerintah dalam melaksanakan program peningkatan kesehatan lingkungan.

Selanjutnya akan dilakukan pengujian nilai rata-rata (uji-t) pada setiap variabel. Pengujian ini dilakukan untuk melihat apakah variabel tersebut memiliki perbedaan yang signifikan di antara masing-masing kluster. Nilai α yang digunakan adalah $\alpha = 0,5$. Jika hasil dari p -value uji-t lebih kecil dari $\alpha = 0,5$ maka terdapat perbedaan yang signifikan antara nilai tengah variabel pada masing-masing kluster. Sebaliknya, jika hasil dari p -value uji-t lebih besar dari $\alpha = 0,5$ maka tidak terdapat perbedaan yang signifikan antara nilai tengah variabel pada masing-masing kluster. Pada Tabel 5 diberikan hasil yang diperoleh pada setiap variabel.

Tabel 5. P -value Pada Hasil Pengujian Nilai Rata-Rata Setiap Variabel

Variabel	Kluster		Keterangan
	Kluster 1	Kluster 2	
X_1	0.005	0.001	terdapat perbedaan yang signifikan
X_2	0.000	0.000	terdapat perbedaan yang signifikan
X_3	0.015	0.026	terdapat perbedaan yang signifikan
X_4	0.596	0.578	tidak terdapat perbedaan yang signifikan
X_5	0.147	0.176	tidak terdapat perbedaan yang signifikan
X_6	0.240	0.225	tidak terdapat perbedaan yang signifikan
X_7	0.000	0.002	terdapat perbedaan yang signifikan

Pada Tabel 5 yang merupakan hasil dari pengujian nilai tengah setiap variabel antara kluster 1 dan kluster 2 dapat dilihat bahwa variabel X_1 (STBM), X_2 (Tatanan Kawasan Sehat), X_3 (Air Minum Layak Rumah Tangga) dan X_7 (rumah tangga layak huni) memiliki perbedaan nilai tengah yang signifikan antara kluster 1 dan kluster 2. Sedangkan pada variabel X_4 (TTU yang Memenuhi Syarat Kesehatan), X_5 (TPM yang Memenuhi Syarat Kesehatan), dan X_6 (Sanitasi Layak Rumah Tangga) tidak memiliki perbedaan nilai tengah yang signifikan antara kluster 1 dan kluster 2.

5. Kesimpulan

Dalam penelitian ini dilakukan pengelompokan provinsi-provinsi di Indonesia berdasarkan indikator kesehatan lingkungan menggunakan metode *Fuzzy C-Means*. Kluster optimum diperoleh dengan menggunakan validitas kluster *Partition Entropy* (PE) yang terbentuk sebanyak dua kluster. Provinsi-provinsi yang menjadi anggota kluster 1 (22 provinsi) adalah Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, DKI Jakarta, Jawa Barat, Kepulauan Bangka Belitung, Kepulauan Riau, Jawa Tengah, Jawa Timur, DI Yogyakarta, Banten, Bali, Kalimantan

Timur, Kalimantan Utara, Nusa Tenggara Barat, Sulawesi Utara, Sulawesi Selatan Kalimantan Selatan dan Gorontalo. Sedangkan provinsi yang menjadi klaster 2 (12 provinsi) adalah Aceh, Sumatera Utara, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan tengah, Sulawesi tengah, Sulawesi Tenggara , Sulawesi Barat, Maluku, Maluku Utara, Papua Barat dan Papua.

Klaster 1 merupakan klaster dengan nilai semua indikator kesehatan lingkungan berada diatas rata-rata indikator secara nasional. Sedangkan klaster 2 merupakan klaster dengan rata-rata setiap indikator kesehatan lingkungan berada dibawah setiap indikator nasional. Dengan demikian, klaster 1 adalah klaster yang lebih baik dibandingkan klaster 2.

6. Ucapan Terima kasih

Penulis mengucapkan terimakasih banyak kepada Ibu Hazmira Yozza M.Si, Bapak Dr. Ahmad Iqbal Baqi dan Bapak Prof, Syafrizal Sy yang telah memberikan masukan dan saran sehingga paper ini dapat diselesaikan dengan baik.

Daftar Pustaka

- [1] Badan Pusat Statistik. 2019. *Perumahan 2019*. Jakarta: Badan Pusat Statistik. <https://www.bps.go.id/indicator/29/1241/1/persentase-rumah-tangga-yang-memiliki-akses-terhadap-hunian-yang-layak-dan-terjangkau-menurut-provinsi.html>.
- [2] Johnson R, Wichern D. 2007. *Applied Multivariate Statistical Analysis, Sixth Edition*. New Jersey : Pearson Education.
- [3] Kemenkes RI. 2020. *Profil Kesehatan Indonesia 2019*. Jakarta : Kemenkes RI. <https://pusdatin.kemkes.go.id/resources/download/pusdatin/profil-kesehatan-indonesia/Profil-Kesehatan-indonesia-2019.pdf>.
- [4] Kusumadewi dan Purnomo .2004. *Aplikasi Logika Fuzzy untuk Pendukung Keputusan*. Yogyakarta : Graha Ilmu.
- [5] Mattjik, A.A. dan I Made Sumertajoyo. 2011. *Sidik Peubah Ganda dengan Menggunakan SAS*. Bogor : IPB Press.
- [6] Mukaka MM. 2012. Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. *Malawi Medical Journal*. 24(3): 69 – 71.
- [7] Wang, W. dan Zhang, Y. On Fuzzy Cluster Validity Indices, *Fuzzy Sets Syztem*. 2007. 158(19): 2095 – 2117