p-ISSN 2085-7829
e-ISSN 2798-3455

Penerapan Metode *K-Means* Dalam Pengelompokan Kabupaten/Kota Di Kalimantan Berdasarkan Indikator Pendidikan

Application Of the K-Means Method In Regencies/Cities Grouping In Kalimantan Based On Educational Indicators

Gerald Claudio Messakh [a], Memi Nor Hayati[b], dan Sifriyani[c]
Laboratorium Statistika Terapan, FMIPA, Universitas Mulawarman

[a] geraldmessakh@gmail.com
[b] *Corresponding author*: meminorhayati@fmipa.unmul.ac.id
[c] sifriyani@fmipa.unmul.ac.id

ABSTRACT

Cluster analysis is an analysis that aims to classify data based on the similarity of spesific characteristics. Based on the structure, cluster analysis is divided into two, namely hierarchical and non-hierarchical methods. One of the non-hierarchical methods used in this study is K-Means. K-Means is a partition-based non-hierarchical data grouping method. This purpose of this study is to obtain the best results of grouping regencies/cities on the island of Kalimantan based on education indicators using the K-Means method based on the smallest ratio of standard deviation. The education indicators used are Expected length of schooling, Average Length of School, Number of Elementary Schools, Number of Middle Schools, Number of High Schools, Number of Elementary School Teachers, Number of Middle School Teachers, Number of High School Teachers, Number of Elementary Students, Number of Middle School Students, Number of Senior High School Students . Based on the results of the analysis, it can be concluded that the best grouping results based on the smallest ratio of standard deviation is 0.6052 which produces optimal clusters of 2 clusters with the first cluster consisting of 14 Regencies/Cities while the second cluster consists of 42 Regencies/Cities on Kalimantan Island. With most members of cluster 2, it can be said that they are regencies/cities with a smaller average variable compared to regencies/cities that are members of cluster 1, where the average value is 5 of the 6 variables in cluster 2 are smaller than the regencies/cities which are members of cluster 1. So it can be said that the regencies/cities in cluster 1 are more dominant and better than the regencies/cities in cluster 2

Keywords: *Education Indicators, K-Means, Standard Deviation Ratio*

1. Pendahuluan

Analisis *cluster* merupakan metode analisis statistika multivariat yang digunakan untuk mengelompokkan objek pengamatan yang mempunyai kemiripan karakteristik tertentu dan dapat dipisahkan dengan kelompok lainnya sehingga dalam setiap kelompok akan berisi objek yang homogen (Nurjanah dkk, 2014). Analisis *cluster* terbagi menjadi dua yaitu metode hirarki dan non hirarki. Metode non hirarki memiliki kelebihan dapat melakukan analisis dengan jumlah sampel yang lebih besar dibandingkan metode hirarki. (Suyanto, 2017). Salah satu metode yang termasuk ke dalam metode non hirarki adalah *K-means* (Triyanto, 2015).

Metode *K-Means* adalah metode yang membagi objek data ke dalam C *cluster* dengan menetapkan tiap objek ke pusat *cluster* yang terdekat (Yohannes, 2016). Setelah objek masuk pada pusat *cluster* terdekat dan membentuk *cluster* baru, pusat *cluster* baru ditentukan kembali dengan menghitung rata-rata objek pada *cluster* yang sama. Jika masih ada perbedaan dengan pusat *cluster* yang sebelumnya, maka dilakukan perhitungan kembali pusat *cluster* baru (Lathifaturrahmah, 2010).

Metode *K-Means* dapat diaplikasikan dalam berbagai bidang, salah satunya sosial kependudukan. Karti & Irhamah (2013) melakukan pengelompokan kabupaten/kota di Provinsi Jawa Timur berdasarkan indikator pendidikan dengan menggunakan metode K-Means dan FCM. Berdasarkan nilai Internal *Cluster Dispersion Rate* (ICDRate), metode K-Means memiliki nilai ICDRate lebih kecil dibandingkan metode *Fuzzy C-Means*, sehingga K-Means memiliki hasil clustering yang lebih baik.

Pendidikan merupakan motor penggerak untuk memajukan kehidupan manusia di masa depan. Pendidikan terbangun dari beberapa komponen yang saling berkaitan. Proses pendidikan sendiri merupakan suatu kegiatan yang menggerakkan segenap komponen untuk mewujudkan tujuan pendidikan (Direktorat Statistik Kesejahteraan Rakyat, 2021). Pada tahun 2021, Rata-rata lama sekolah (RLS) penduduk usia 15 tahun ke atas di Kalimantan adalah sebesar 8,41 tahun. Angka tersebut lebih kecil 0,56 dibandingkan RLS Indonesia yaitu sebesar 8,97 dimana Kota Palangkaraya memiliki nilai RLS tertinggi dan yang terendah adalah Kabupaten Kayong Utara. Di antara 5 Provinsi di Kalimantan, Provinsi Kalimantan Timur memiliki RLS terbesar yaitu 9,44, sedangkan Provinsi Kalimantan Barat memiliki RLS terkecil yaitu sebesar 7,35. Mengingat pentingnya peran pendidikan dalam kemajuan bangsa maka pengukuran dan perhitungan

p-ISSN 2085-7829
e-ISSN 2798-3455

indikator pendidikan harus dilakukan untuk melihat sejauh mana pemerataan pendidikan. Untuk mengetahui sebaran pendidikan atau ciri-ciri tingkat pendidikan dapat dilakukan melalui pengelompokan dengan menggunakan metode *K-Means*.

Meskipun telah banyak pencapaian dan program yang menghasilkan output yang positif dalam bidang pendidikan, namun masih banyak cela dan tantangan yang perlu diselesaikan dalam pembangunan pendidikan agar target berbagai indikator pendidikan mampu terpenuhi pada akhir tahun 2025. Berdasarkan uraian di atas, pada penelitian ini akan dilakukan pengelompokan Kabupaten/Kota di Kalimantan berdasarkan indikator pendidikan menggunakan metode *K-Means* yang bertujuan untuk mendapatkan hasil pengelompokan yang optimal berdasarkan nilai rasio simpangan baku terkecil..

2. Tinjauan Pustaka

2.1 Analisis *Cluster*

Analisis *cluster* adalah metode analisis statistika multivariat yang bertujuan untuk mengidentifikasi sekelompok objek yang mempunyai kesamaan karakteristik dan dapat dipisahkan dengan kelompok objek lainnya. (Nurjanah dkk, 2014). Suatu *cluster* dianggap baik jika memiliki homogenitas yang tinggi antar anggota dalam suatu kelompok dan memiliki heterogenitas yang tinggi antar kelompok (Santoso, 2015).

Dalam analisis *cluster* rentang nilai yang besar antara nilai variabel, dapat menyebabkan perhitungan jarak menjadi tidak stabil, maka perlu dilakukan proses standarisasi data (Yulianto & Hidayatullah, 2014). Standarisasi data dapat dilakukan dengan menggunakan Algoritma *Min-Max* (Suyanto, 2018):

$$x'_{i,k} = \frac{x_{i,k} - x_{kmin}}{x_{kmax} - x_{kmin}} \tag{1}$$

dimana:

$x'_{i,k}$: nilai baru hasil standarisasi untuk data ke-i variabel ke-k
$x_{i,k}$: data ke-i variabel ke-k
x_{kmin} : data minimum variabel ke-k
x_{kmax} : data maksimum variabel ke-k

Analisis *cluster* memiliki dua asumsi yaitu sampel yang digunakan bersifat *representative* dan non multikolinearitas. Menurut Gujarati (2003), multikolinearitas adalah suatu keadaan dimana terdapat hubungan linier yang kuat antara variabel bebas. Salah satu cara untuk mendeteksi multikolinearitas yaitu dengan melihat nilai *Variance Inflation Factor* (VIF) yang dapat diperoleh menggunakan persamaan berikut :

$$VIF_k = \frac{1}{1 - R_k^2}, k = 1,2, \dots, p \tag{2}$$

dimana ${R_k}^2$ adalah koefisien determinasi antara variabel X_k yang diregresikan terhadap variabel lainnya. Jika nilai VIF lebih besar dari 10, maka mengindikasikan terdapat multikolinearitas. Salah satu cara yang dapat dilakukan untuk mengatasi multikolinearitas yaitu dengan mereduksi variabel (Masnarivan, 2021).

2.2 Metode *K-Means*

Metode *K-Means* merupakan salah satu metode analisis *cluster* non hirarki berbasis partisi yang mengelompokkan data ke dalam kelompok sehingga data yang memiliki kemiripan dan karakteristik yang sama berada di *cluster* yang sama dan data yang memiliki karakteristik berbeda akan dikelompokan ke dalam *cluster* yang lain (Prasetyo, 2012). Dengan menggunakan metode *K-Means*, hasil pengelompokan akan bergantung pada nilai pusat *cluster* awal. Nilai awal yang berbeda dapat menghasilkan kelompok yang berbeda. Beberapa cara untuk memberi nilai awal yaitu mengambil sampel awal dari objek kemudian mencari nilai pusatnya, menentukan nilai awal secara random, atau menggunakan hasil dari kelompok hirarki dengan jumlah yang sesuai (Santoso, 2007).

Menurut Prasetyo (2012), langkah-langkah metode *K-Means* adalah sebagai berikut:

1. Menentukan jumlah *cluster* (C).
2. Menentukan pusat *cluster* (v_{ck}) awal secara acak dari objek pengamatan.
3. Menghitung jarak *euclidean* untuk setiap objek pengamatan terhadap pusat *cluster* menggunakan persamaan berikut :

$$d(x'_i, v_c) = \sqrt{\sum_{k=1}^{p} \left(x'_{i,k} - v_{c,k}\right)^2} \tag{3}$$

dimana:

$d(x_i, v_c)$: Jarak *euclidean* data pengamatan ke-i dengan pusat *cluster* ke-c
$x'_{i,k}$: Nilai baru hasil standarisasi data pengamatan ke-i pada variabel ke-k
$v_{c,k}$: Pusat *cluster* pada *cluster* ke-c pada variabel ke-k

4. Mengalokasikan masing-masing objek ke *cluster* yang objeknya paling mirip, berdasarkan jarak terdekat antara objek terhadap setiap pusat *cluster*.

p-ISSN 2085-7829
e-ISSN 2798-3455

5. Memperbarui pusat *cluster* dengan menggunakan persamaan berikut.

$$v_{c,k}^{t} = \sum_{i=1}^{n_c} \frac{x'_{i,k,c}}{n_c} \tag{4}$$

dimana :

v_{ck}^{t} : pusat *cluster* baru ke-c variabel ke-k pada iterasi ke-t
$x'_{i,k,c}$: standarisasi data ke-i variabel ke-k dalam *cluster* ke-c
n_c : banyaknya data pada *cluster* ke-c

6. Mengulangi langkah 3, 4, dan 5 sampai tidak ada lagi anggota *cluster* yang mengalami perubahan *cluster*.

2.3 Rasio Simpangan Baku

Menurut Santoso (2007), suatu metode pengelompokan yang dapat digunakan untuk membentuk *cluster* dikatakan memiliki kinerja baik jika memiliki nilai simpangan baku dalam *cluster* yang minimum terhadap simpangan baku antar *cluster*. Menurut Michael & Gordon (1981), simpangan baku yaitu nilai statistik yang digunakan untuk menentukan bagaimana sebaran data serta seberapa dekat titik data individu ke rata-rata. Simpangan baku dalam *cluster* (S_w) dapat dihitung menggunakan persamaan berikut (Santoso, 2007):

$$S_w = \frac{1}{C}\sum_{c=1}^{C} S_c \tag{5}$$

$$S_{c,k} = \sqrt{\frac{1}{n_c-1}\sum_{i=1}^{n_c}\left(x'_{i,k} - \bar{x}_{c,k}\right)^2} \tag{6}$$

$$S_c = \frac{1}{p}\sum_{k=1}^{p} S_{c,k} \tag{7}$$

Adapun rumus dari simpangan baku antar *cluster* (S_b) adalah sebagai berikut :

$$\bar{x}_c = \frac{1}{p}\sum_{k=1}^{p} \bar{x}_{c,k} \tag{8}$$

$$S_b = \sqrt{\frac{1}{C-1}\sum_{c=1}^{C}(\bar{x}_c - \bar{x})^2} \tag{9}$$

$$\bar{x} = \frac{1}{C}\sum_{c=1}^{C} \bar{x}_c \tag{10}$$

dimana:

S_w : nilai simpangan baku dalam *cluster*
$S_{c,k}$: simpangan baku *cluster* ke-c variabel ke-k
$\bar{x}_{c,k}$: rata-rata dari *cluster* ke-c variabel ke-k
S_c : simpangan baku *cluster* ke-c
S_b : nilai simpangan baku antar *cluster*
$\bar{x}_c$: nilai rata-rata *cluster* ke-c
$\bar{x}$: nilai rata-rata keseluruhan *cluster*
C : banyaknya *cluster*

2.4 Pendidikan

Menurut Undang-Undang (UU) Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional, pendidikan adalah usaha sadar dan terencana untuk mewujudkan suasana belajar dan proses yang bertujuan agar peserta didik secara aktif mengembangkan potensi jiwa keagamaan, disiplin diri, kepribadian, kecerdasan, akhlak mulia, serta keterampilan yang diperlukan dirinya, masyarakat, dan bangsa dan negara. Sistem pendidikan nasional adalah keseluruhan komponen pendidikan yang satu sama lain saling terintegrasi untuk mencapai tujuan pendidikan nasional. Sebagai suatu sistem, pendidikan bertumpu pada beberapa komponen, seperti peserta didik, bahan/alat ajar, lingkungan pendidikan, dan tujuan pendidikan. (Direktorat Statistik Kesejahteraan Rakyat, 2021).

3. Bahan dan Metode

3.1 Sumber Data

Adapun data yang digunakan dalam penelitian ini bersumber dari *website* Badan Pusat Statistik (BPS) dari lima Provinsi yang ada di Pulau Kalimantan pada tahun 2021.

3.2 Variabel Penelitian

Penentuan variabel indikator pendidikan yang digunakan menggunakan referensi perhitungan IPM aspek pendidikan di BPS dan mengikuti penelitian sebelumnya oleh (Dewi dkk, 2021). Adapun variabel yang digunakan dalam penelitian ini adalah sebagai berikut :

X_1 : Harapan Lama Sekolah
X_2 : Rata-rata Lama Sekolah
X_3 : Jumlah SD/Sederajat

p-ISSN 2085-7829
e-ISSN 2798-3455

X_4 : Jumlah SMP/Sederajat
X_5 : Jumlah SMA/Sederajat
X_6 : Jumlah Guru SD/Sederajat
X_7 : Jumlah Guru SMP/Sederajat
X_8 : Jumlah Guru SMA/Sederajat
X_9 : Jumlah Murid SD/ Sederajat
X_{10} : Jumlah Murid SMP/ Sederajat
X_{11} : Jumlah Murid SMA/ Sederajat

3.3 Teknik Pengumpulan Data

Teknik pengumpulan data yang dilakukan dalam penelitian ini adalah dengan cara mengambil data sekunder yang diperoleh melalui *website* Badan Pusat Statistik (BPS) dari lima Provinsi yang ada di Pulau Kalimantan, yaitu provinsi Kalimantan Barat, Klimantan Selatan, Kalimantan Tengah, Kalimantan Timur dan Kalimantan Utara. Adapun objek penelitian yang digunakan adalah 56 Kabupaten/Kota yang ada di Pulau Kalimantan.

3.4 Teknik Analisis Data

Adapun langkah-langkah analisis data yang dilakukan adalah sebagai berikut :

1. Melakukan standarisasi data dengan menggunakan metode Min-Max.
2. Mendeteksi multikolinearitas data.
3. Melakukan pengelompokan menggunakan metode *K-Means* sebagai berikut :
   a. Menentukan banyaknya *cluster* yang akan digunakan yaitu 2,3,4,5 dan 6
   b. Menentukan pusat *cluster* awal secara acak dari objek pengamatan.
   c. Menghitung jarak *euclidean* untuk setiap objek pengamatan terhadap masing-masing pusat *cluster* menggunakan persamaan (3).
   d. Menetapkan masing-masing objek ke *cluster* yang objeknya paling mirip
   e. Memperbarui pusat *cluster* dengan menghitung nilai rata-rata dari objek untuk setiap *cluster* berdasarkan persamaan (4).
   f. Mengulangi langkah c sampai dengan langkah e sampai tidak ada lagi anggota suatu *cluster* yang mengalami perubahan *cluster*.
   g. Mengulangi langkah b sampai dengan langkah f untuk jumlah *cluster* yang berbeda.
4. Menghitung nilai rasio simpangan baku untuk setiap jumlah *Cluster*.
5. Menginterpretasi hasil pengelompokan terbaik berdasarkan nilai rasio $S_w$terhadap S_b terkecil.

4. Hasil dan Pembahasan

4.1 Statistika Deskriptif

Analisis statistika deskriptif dilakukan untuk melihat gambaran data secara umum. Analisis statistika deskriptif pada data indikator pendidikan berdasarkan 11 variabel di 56 kabupaten/kota yang ada di pulau Kalimantan tahun 2021 menggunakan nilai minimum, maksimum, rata-rata dan simpangan baku. Adapun hasil analisis statistika deskriptif dapat dilihat pada Tabel 1 hingga Tabel 5 sebagai berikut.

Tabel 1. Statistika Deskriptif Provinsi Kalimantan Barat

Variabel	Min	Max	Rata-rata	Simpangan Baku
X_1	11,170	15,010	12,430	0,954
X_2	6,020	10,430	7,304	0,966
X_3	110	560	346,500	149,641
X_4	47	245	119,143	49,847
X_5	21	145	59,429	31,304
X_6	1.124	4.944	2.813,143	1.101,474
X_7	578	2.430	1.407,500	558,397
X_8	337	2.348	1.018,286	513,466
X_9	13.810	75.215	44.749,929	19.439,927
X_{10}	6.244	34.278	19.454,643	9.055,144
X_{11}	5.507	39.535	17.086,214	8.512,803

p-ISSN 2085-7829
e-ISSN 2798-3455

Tabel 2. Statistika Deskriptif Provinsi Kalimantan Selatan

Variabel	Min	Max	Rata-rata	Simpangan Baku
X_1	11,950	14,820	12,727	0,780
X_2	7,460	10,960	8,280	1,068
X_3	101	470	267,077	82,504
X_4	38	132	74,769	26,000
X_5	18	64	38,615	13,077
X_6	1.891	4.826	2.897,154	841,694
X_7	670	2.240	1.287,231	442,053
X_8	491	2.125	996,692	403,928
X_9	15.090	67.145	34.043,846	14.050,418
X_{10}	5.135	29.783	13.909,462	6.185,801
X_{11}	4.933	31.760	12.436,538	6.566,839

Tabel 3. Statistika Deskriptif Provinsi Kalimantan Tengah

Variabel	Min	Max	Rata-rata	Simpangan Baku
X_1	11,760	14,960	12,641	0,740
X_2	7,600	11,530	8,641	0,946
X_3	52	502	208,500	108,702
X_4	19	168	72,500	36,541
X_5	11	62	32,929	14,063
X_6	681	4.755	2.065,786	1.028,822
X_7	340	1.868	886,143	406,983
X_8	262	1.417	675,429	332,811
X_9	6.658	55.436	22.542,143	13.220,882
X_{10}	1.858	23.725	9.519,500	5.784,983
X_{11}	2.258	17.569	7.938,071	4.708,575

Tabel 4. Statistika Deskriptif Provinsi Kalimantan Timur

Variabel	Min	Max	Rata-rata	Simpangan Baku
X_1	12,570	15,090	13,376	0,731
X_2	8,180	10,910	9,442	0,943
X_3	39	510	203,400	124,603
X_4	15	201	83,800	50,898
X_5	14	127	53,600	33,052
X_6	462	6.346	2.806,300	1.640,697
X_7	327	2.861	1.386,700	800,524
X_8	214	2.383	1.137,300	670,020
X_9	4.000	92.197	44.151,000	29.143,783
X_{10}	1.828	42.164	19.227,100	13.520,054
X_{11}	1.305	23.687	10.255,900	6.993,369

Tabel 5. Statistika Deskriptif Provinsi Kalimantan Utara

Variabel	Min	Max	Rata-rata	Simpangan Baku
X_1	12,220	14,030	13,044	0,613
X_2	8,170	9,980	9,108	0,604
X_3	30	152	102,400	45,916
X_4	11	69	41,600	19,946
X_5	4	32	23,400	10,210
X_6	362	1.879	1.298,800	507,984
X_7	193	789	637,600	227,467
X_8	91	697	501,200	225,828
X_9	3.402	27.083	16.687,000	8.678,177
X_{10}	1.228	11.897	7.236,400	3.828,502
X_{11}	1.017	10.639	6.291,600	3.395,776

4.2 Standardisasi Data

Perhitungan standardisasi data dilakukan menggunakan persamaan (1), contoh perhitungan menggunakan data kabupaten/kota yang pertama sebagai berikut :

p-ISSN 2085-7829
e-ISSN 2798-3455

$$x'_{1,1} = \frac{x_{1,1} - x_{1mi}}{x_{1ma} - x_{1min}} = \frac{12{,}63 - 11{,}17}{15{,}09 - 11{,}17} = 0{,}3724$$
$$x'_{1,2} = \frac{x_{1,2} - x_{2mi}}{x_{2max} - x_{2min}} = \frac{6{,}72 - 6{,}02}{11{,}53 - 6{,}02} = 0{,}1270$$
$$\vdots$$
$$x'_{1,11} = \frac{x_{1,11} - x_{11min}}{x_{11max} - x_{11min}} = \frac{23.456 - 1.017}{39.535 - 1.017} = 0{,}5826$$

4.3 Multikolinearitas

Pendeteksian multikolinearitas dilakukan dengan melihat nilai VIF. Berikut ini merupakan hasil perhitungan nilai VIF pada setiap variable yang dapat dilihat pada Tabel 6 sebagai berikut.

Tabel 6. Nilai VIF

Variabel	VIF	Variabel	VIF
X_1	8,6666	X_7	61,3757
X_2	5,1965	X_8	64,5255
X_3	19,8039	X_9	124,9046
X_4	29,2192	X_{10}	125,1102
X_5	28,2840	X_{11}	13,5941
X_6	42,1840		

Berdasarkan Tabel 6, terdapat 9 variabel yang memiliki nilai VIF lebih besar dari 10, sehingga dapat dikatakan bahwa terdapat multikolinearitas antar variabel dan belum dapat dilanjutkan ke dalam proses pengelompokan. Untuk mengatasi multikolinearitas dapat dilakukan dengan mengeluarkan variabel yang memiliki nilai VIF paling besar kemudian meregresikan kembali variabel yang tersisa. Setelah mengeluarkan beberapa variabel dan melakukan perhitungan regresi didapatkan nilai VIF akhir sebagai berikut :

Tabel 7. Nilai VIF setelah mengeluarkan variabel X_{10}, X_8, X_7, X_4 dan X_6

Variabel	VIF	Variabel	VIF
X_1	6,2222	X_5	7,3830
X_2	4,1703	X_9	7,7411
X_3	4,4504	X_{11}	4,7305

Berdasarkan Tabel 7 terlihat bahwa semua variabel yang tersisa memiliki nilai VIF lebih kecil dari 10, sehingga dapat dikatakan bahwa tidak terdapat multikolinearitas antar variabel dan dapat dilanjutkan ke proses pengelompokan dengan menggunakan variabel yang tidak dikeluarkan.

4.4 Pengelompokan *K-Means*

Adapun langkah-langkah pengelompokan menggunakan metode *K-Means* sebagai berikut :

a. Menentukan Banyaknya *Cluster*
Dalam penelitian ini, jumlah *cluster* yang akan digunakan yaitu $C = 2$, 3, 4, 5 dan 6. Sebagai contoh perhitungan pada penelitian ini, dilakukan dengan menggunakan $C = 2$.

b. Memilih Pusat *Cluster* Awal
Pusat *cluster* awal yang terpilih adalah data ke-18 (Kabupaten Barito Kuala) dan data ke-36 (Kabupaten Katingan) sebagai berikut :

Tabel 8. Pusat *Cluster* Awal untuk $C = 2$

Variabel	Pusat *Cluster*		Variabel	Pusat *Cluster*	
	$v_{1,k}$	$v_{2,k}$		$v_{1,k}$	$v_{2,k}$
X_1	0,3138	0,4184	X_5	0,3050	0,2340
X_2	0,2777	0,4828	X_9	0,3466	0,1894
X_3	0,5642	0,3528	X_{11}	0,2648	0,1548

c. Menghitung Jarak Semua Data Pengamatan dengan Pusat *Cluster* Awal
Langkah selanjutnya adalah menghitung jarak semua data pengamatan dengan masing-masing pusat *cluster* awal berdasarkan variabel-variabel yang tidak dikeluarkan pada tahap pendeteksian multikolnearitas. Perhitungan jarak data pengamatan dengan pusat *cluster* awal yang terpilih menggunakan persamaan (3). Adapun contoh perhitungan jarak antara data pengamatan ke-1 (Kabupaten Sambas) terhadap pusat *cluster* awal adalah sebagai berikut :
Jarak *euclidean* data pengamatan terhadap pusat *cluster* 1

$$d(x'_{1,k}, v_{1,k}) = \sqrt{(x'_{1,1} - v_{1,1})^2 + (x'_{1,2} - v_{1,2})^2 + \cdots + (x'_{1,11} - v_{1,11})^2}$$

p-ISSN 2085-7829
e-ISSN 2798-3455

$$= \sqrt{(0{,}3724 - 0{,}3138)^2 + (0{,}1270 - 0{,}2777)^2 + \cdots + (0{,}5826 - 0{,}2648)^2}$$
$$= 0{,}6521$$

Jarak *euclidean* data pengamatan terhadap pusat *cluster* 2

$$d(x'_{2,k}, v_{2,k}) = \sqrt{\left(x'_{1,1} - v_{2,1}\right)^2 + \left(x'_{1,2} - v_{2,2}\right)^2 + \cdots + \left(x'_{1,11} - v_{2,11}\right)^2}$$
$$= \sqrt{(0{,}3724 - 0{,}4184)^2 + (0{,}1270 - 0{,}4828)^2 + \cdots + (0{,}5826 - 0{,}1548)^2}$$
$$= 0{,}9764$$

Perhitungan jarak data pengamatan dengan pusat *cluster* awal dilakukan pada data pengamatan ke-2 hingga data pengamatan ke-56. Adapun keseluruhan hasil perhitungan dapat dilihat pada Tabel 9 berikut :

Tabel 9. Data Hasil Perhitungan Jarak *Eulidean* Terhadap Masing-Masing Pusat *Cluster* Awal

Data Pengamatan	Jarak *Euclidean* Data Pengamatan ke Pusat *Cluster* Awal	
	Pusat *Cluster* 1	Pusat *Cluster* 2
1	0,6521	0,9764
2	0,1812	0,4550
3	0,3721	0,7183
⋮	⋮	⋮
56	0,7910	0,4953

d. Menempatkan Data Pengamatan ke Pusat *Cluster* Terdekat

Langkah selanjutnya adalah menempatkan setiap data pengamatan ke dalam pusat *cluster* terdekat berdasarkan jarak terdekat objek terhadap pusat *cluster*. Hasil alokasi dapat dilihat pada Tabel 10 berikut:

Tabel 10. Hasil Penempatan Setiap Data ke Pusat *Cluster* Terdekat

Data Pengamatan	Jarak *Euclidean* Data Pengamatan ke Pusat *Cluster* Awal		Alokasi *Cluster*
	Pusat *Cluster* 1	Pusat *Cluster* 2	
1	0,6521	0,9764	1
2	0,1812	0,4550	1
3	0,3721	0,7183	1
⋮	⋮	⋮	⋮
56	0,7910	0,4953	2

Berdasarkan Tabel 10 dapat dilihat bahwa jarak *euclidean* untuk data pengamatan ke-1 terhadap pusat *cluster* 1 lebih kecil dibandingkan jarak *euclidean* data pengamatan ke-1 terhadap pusat *cluster* 2, sehingga data pengamatan ke-1 masuk dalam keanggotaan *cluster* 1 demikian seterusnya sampai dengan data pengamatan ke-56. Berdasarkan hasil penempatan diperoleh bahwa *cluster* 1 beranggotakan 24 Kabupaten/Kota dan *cluster* 2 beranggotakan 32 Kabupaten/Kota.

e. Memperbarui Pusat *Cluster*

Pusat *cluster* diperbarui dengan menggunakan rumus rata-rata pada persamaan (4) untuk $k = 1, 2, 3, 5, 9, 11$. Adapun hasil perhitungan pusat *cluster* baru sebagai berikut :

Tabel 11. Pusat *Cluster* Baru

Variabel	Pusat *Cluster*		Variabel	Pusat *Cluster*	
	$v_{1,k}^1$	$v_{2,k}^1$		$v_{1,k}^1$	$v_{2,k}^1$
X_1	0,3833	0,4293	X_5	0,4072	0,1875
X_2	0,3303	0,5106	X_9	0,5197	0,2152
X_3	0,6299	0,2415	X_{11}	0,4190	0,1637

Berdasarkan hasil perhitungan pada Tabel 11, dapat dilihat bahwa terdapat perbedaan pusat *cluster* baru dengan pusat *cluster* sebelumnya, maka pengelompokan dilanjutkan ke iterasi selanjutnya.

f. Mengulangi langkah c, d dan e sampai tidak ada perubahan pusat *cluster* dengan pusat *cluster* sebelumnya

Berdasarkan hasil perhitungan, pengelompokan dihentikan pada iterasi ke-5, dimana tidak terdapat perubahan pada keanggotaan *cluster.* Sehingga pusat *cluster* baru akan sama dengan pusat *custer* lama. Hasil pengelompokan metode *K-Means* dengan $C = 2$ dapat dilihat pada Tabel 12 sebagai berikut :

p-ISSN 2085-7829
e-ISSN 2798-3455

Tabel 12. Hasil Pengelompokan *K-Means* untuk $C = 2$

Cluster	Jumlah Anggota	Anggota *Cluster*
1	14	Sambas, Landak, Sangau, Ketapang, Sintang, Kubu Raya, Pontianak, Banjar, Banjarmasin, Kotawaringin Timur, Kapuas, Kutai Kartanegara, Balikpapan, Samarinda.
2	42	Bengkayang, Mempawah, Kapuas Hulu, Sekadau, Melawi, Kayong Utara, Singkawang, Tanah Laut, Kotabaru, Barito Kuala, Tapin, Hulu Sungai Selatan, Hulu Sungai Tengah, Hulu Sungai Utara, Tabalong, Tanah Bumbu, Balangan, Banjar Baru, Kotawaringin Barat, Barito Selatan, Barito Utara, Sukamara, Lamandau, Seruyan, Katingan, Pulang Pisau, Gunung Mas, Barito Timur, Murung Raya, Palangkaraya, Paser, Kutai Barat, Kutai Timur, Berau, Penajam Paser Utara, Mahakam Ulu, Bontang, Malinau, Bulungan, Tana Tidung, Nunukan, Tarakan.

Berdasarkan Tabel 12, dapat dilihat bahwa hasil akhir pengelompokan Kabupaten/Kota di Pulau Kalimantan berdasarkan indikator pendidikan pada tahun 2021 menggunakan metode *K-Means* dengan $C = 2$ didapatkan *cluster* ke-1 beranggotakan 14 Kabupaten/Kota, sedangkan *cluster* ke-2 beranggotakan 42 Kabupaten/Kota. Selain perhitungan pada $C = 2$, dilakukan juga perhitungan dengan langkah yang sama terhadap banyak *cluster* (C) 3, 4, 5, dan 6.

4.5 Rasio Simpangan Baku

a. Simpangan baku dalam *cluster* (S_w)

Langkah awal sebelum menghitung simpangan baku dalam *cluster* (S_w) adalah menghitung nilai rata-rata dari *cluster* ke-c variabel ke-k $(\bar{X}_{c,k})$ untuk k =1, 2, 3, 5, 9, 11. Hasil perhitungan rata-rata dari *cluster* ke-c variabel ke-k pada $C = 2$ adalah sebagai berikut:

Tabel 13. Nilai Rata-rata dari *cluster* ke-c variabel ke-k

Variabel	$\bar{x}_{1,k}$	$\bar{x}_{2,k}$
X_1	0,5168	0,3738
X_2	0,4281	0,4351
X_3	0,7296	0,3007
X_5	0,5274	0,1998
X_9	0,6942	0,2296
X_{11}	0,5440	0,1828

Setelah mendapatkan rata-rata dari *cluster* ke-c variabel ke-k $(\bar{X}_{c,k})$, selanjutnya menghitung simpangan baku *cluster* ke-1 variabel ke-k $(S_{1,k})$ menggunakan persamaan (6) sehingga diperoleh simpangan baku *cluster* ke-c variabel ke-k pada *cluster* 1 dan 2 untuk $k = 1, 2, 3, 5, 9, 11$ sebagai berikut:

Tabel 14. Simpangan Baku *cluster* ke-c variabel ke-k

Variabel	$S_{1,k}$	$S_{2,k}$
X_1	0,2852	0,0348
X_2	0,2769	0,0418
X_3	0,2295	0,0249
X_5	0,2058	0,0065
X_9	0,1639	0,0140
X_{11}	0,1780	0,0093

Langkah selanjutnya adalah menghitung nilai simpangan baku *cluster* pada *cluster* (c) 1 dan 2 dengan menggunakan persamaan (7) untuk $k = 1, 2, 3, 5, 9, 11$. Sehingga diperoleh nilai simpangan baku *cluster* 1 dan 2 adalah S_1 sebesar 0,2232 dan S_2 sebesar 0,0219. Langkah selanjutnya adalah menghitung nilai simpangan baku dalam *cluster* (S_w) dengan menggunakan persamaan (5) sebagai berikut :

$$S_w = \frac{1}{2}\sum_{c=1}^{2} S_c = \frac{(0{,}2232 + 0{,}0219)}{2} = 0{,}1225$$

b. Simpangan baku antar *cluster* (S_b)

Dalam menghitung nilai S_b terlebih dahulu dilakukan perhitungan nilai rata-rata *cluster* ke-c dengan menggunakan persamaan (8) untuk $k = 1, 2, 3, 5, 9, 11$ sebagai berikut :

p-ISSN 2085-7829
e-ISSN 2798-3455

$$\bar{x}_1 = \frac{1}{6}\sum_k \bar{x}_{1,k}$$
$$= \frac{0{,}5168 + 0{,}4281 + 0{,}7296 + 0{,}5274 + 0{,}6942 + 0{,}5440}{6}$$
$$= 0{,}5733$$
$$\bar{x}_2 = \frac{1}{6}\sum_k \bar{x}_{2,k}$$
$$= \frac{0{,}3738 + 0{,}4351 + 0{,}3007 + 0{,}1998 + 0{,}2296 + 0{,}1828}{6}$$
$$= 0{,}2870$$

Langkah selanjutnya menghitung nilai rata-rata keseluruhan *cluster* dengan menggunakan persamaan (10) kemudian menghitung simpangan baku antar *cluster* (S_b) dengan menggunakan persamaan (9) sebagai berikut :

$$\bar{x} = \frac{1}{2}\sum_{c=1}^{2} \bar{x}_c$$
$$= \frac{0{,}5733 + 0{,}2870}{2}$$
$$= 0{,}4302$$
$$S_b = \sqrt{\frac{1}{2-1}\sum_{c=1}^{2} (\bar{x}_c - \bar{x})^2}$$
$$= \sqrt{\frac{(0{,}5733 - 0{,}4302)^2 + (0{,}2870 - 0{,}4302)^2}{1}}$$
$$= 0{,}2025$$

Berdasarkan hasil perhitungan nilai S_w dan S_b, maka nilai rasio simpangan baku dalam *cluster* dan simpangan baku antar *cluster* adalah sebagai berikut :

$$Rasio = \frac{S_w}{S_b} = \frac{0{,}1225}{0{,}2025} = 0{,}6052$$

Selain perhitungan pada jumlah *cluster* $(C) = 2$, dilakukan juga perhitungan rasio simpangan baku dengan langkah-langkah yang sama untuk jumlah *cluster* $(C) = 3$, 4, 5 dan 6 dan didapatkan hasil secara lengkap dapat dilihat pada Tabel 15 berikut:

Tabel 15. Nilai Rasio Simpangan Baku Metode *K-Means*

Jumlah *Cluster*	Rasio Simpanan Baku
2	0,6052
3	0,8471
4	0,7466
5	0,6788
6	0,6169

Berdasarkan Tabel 15 dapat dilihat bahwa hasil pengelompokan dengan $C = 2$ memiliki hasil pengelompokan yang lebih baik dibandingkan dengan hasil pengelompokan pada $C = 3, 4, 5$ dan 6. Hal ini dapat dilihat dari nilai rasio simpangan baku pada $C = 2$ sebesar 0,6052 lebih kecil dibandingkan pengelompokan dengan $C = 3, 4, 5$ dan 6.

4.6 Interpretasi Hasil Pengelompokan

Setelah kelompok terbentuk, langkah selanjutnya yaitu menghitung nilai rata-rata variabel untuk setiap *cluster*. Hasil perhitungan rata-rata dapat dilihat pada Tabel 16 sebagai berikut:

Tabel 16. Nilai Rata-rata Variabel untuk Masing-masing *Cluster*

Cluster	Jumlah Anggota	Rata-rata Variabel					
		X_1	X_2	X_3	X_5	X_9	X_{11}
1	14	13,20	8,38	417	78	65.045	21.972
2	42	12,64	8,42	189	32	23.785	8.058

Berdasarkan hasil pengelompokan terbaik menggunakan metode *K-Means* dengan 2 *cluster*, didapatkan *cluster* 1 yang beranggotakan 14 Kabupaten/Kota dimana dari 14 anggota *cluster* 1 terdapat 3 anggota berstatus ibu kota provinsi dan terdapat beberapa anggota yang termasuk kota besar yang ada di Pulau Kalimantan seperti Balikpapan. Sedangkan *cluster* 2 beranggotakan 42 Kabupaten/Kota di Pulau Kalimantan dengan sebagian besar anggota *cluster* 2 dapat dikatakan merupakan Kabupaten/Kota

p-ISSN 2085-7829
e-ISSN 2798-3455

dengan rata-rata variabel yang lebih kecil dibandingkan dengan Kabupaten/Kota yang tergabung dalam *cluster* 1. Hal ini dapat dilihat dari Tabel 12 dimana nilai rata-rata variabel harapan lama sekolah (X_1), variabel Jumlah SD/Sederajat (X_3), Jumlah SMA/Sederajat (X_5), variabel Jumlah Murid SD/Sederajat (X_9), dan variabel Jumlah Murid SMA/Sederajat (X_{11}) pada *cluster* 2 lebih kecil dibandingkan dengan Kabupaten/Kota yang menjadi anggota *cluster* 1. Namun pada variabel rata-rata lama sekolah (X_2) untuk *cluster* 2 memiliki nilai yang lebih besar dibandingkan dengan *cluster* 1. Sehingga dapat dikatakan bahwa Kabupaten/kota dalam *cluster* 1 lebih dominan dan lebih baik dibandingkan Kabupaten/Kota dalam *cluster* 2.

5. Kesimpulan

Berdasarkan hasil pengelompokan metode *K-Means* dengan menggunakan jumlah *cluster* (C) = 2 sampai dengan jumlah *cluster* (C) = 6 berdasarkan nilai rasio simpangan baku menunjukkan bahwa metode *K-Means* dengan C = 2 memiliki hasil pengelompokan yang lebih baik digunakan dibandingkan dengan C yang lainnya dengan *cluster* 1 beranggotakan 14 Kabupaten/Kota dan *cluster* 2 beranggotakan 42 Kabupaten/Kota. Temuan penelitian ini dapat dijadikan informasi bagi instansi pemerintah yang berkepentingan untuk mengambil kebijakan terkait indikator pendidikan di Pulau Kalimantan. khususnya kabupaten/kota yang berada di klaster 2 agar dapat dijadikan bahan evaluasi dalam peningkatan taraf pendidikan di Kalimantan.

Referensi

Dewi, L.S., Talakua, M.W., Lesnussa, Y.A., & Matdoan, M.Y. (2021). Analisis Klaster untuk Pengelompokan Kabupaten/Kota di Provinsi Maluku Berdasarkan Indikator Pendidikan dengan Menggunakan Metode Ward. *Jurnal Statistika dan Aplikasinya*, 5(1).

Direktorat Statistik Kesejahteraan Rakyat. (2021). Statistik Pendidikan 2021. Badan Pusat Statistik.

Gujarati, D. (2003). *Ekonometrika Dasar*. Terjemah Sumarno Zein. Jakarta: Erlangga.

Lathifaturrahmah. (2010). *Perbandingan Hasil Penggerombolan Metode K-Means, Fuzzy C-Means, dan TWO Step Cluster*. Bogor: Institut Pertanian Bogor.

Masnarivan, Y. (2021). *Memahami Penyakit Demam Berdarah Dengue di Sumatera Barat.* Yogyakarta: Bintang Pustaka Madani.

Michael, E & Gordon, G. (1981). *Analysis and Adjustment of Survey Measurement*, New York: Van Nostrand Reinhold Company.

Nurjanah, Farmadi, A., & Indriani, F. (2014). Implementasi Metode Fuzzy C-Means Pada Sistem *Clustering* Data Varietas Padi. *Jurnal Ilmu Komputer*, 1(1), 23–32.

Prasetyo, E. (2012). *Data Mining: Konsep dan Aplikasi menggunakan Matlab.*Yogyakarta: Andi Offset.

Santoso, S. (2007). *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: PT. Elex Media Komputindo Gramedia.

Santoso, S. (2010). *Statistika Multivariat Konsep & Aplikasi dengan SPSS*. Jakarta: PT. Elex Media Komputindo.

Santoso, S. (2015). *Menguasai Statistik Multivariat*. Jakarta: PT. Elex Media Komputindo.

Suyanto. (2017). *Data Mining Untuk Klasifikasi dan Klasterisasi Data*. Bandung: Informatika.

Suyanto. (2018). *Machine Learning Tingkat Dasar dan Lanjut*. Bandung: Informatika Bandung.

Triyanto, W.A. (2015). Algoritma K-Medoids untuk Penentuan Strategi Pemasaran Produk. *Jurnal SIMETRIS*, 6(1), 183 – 188.

Vulandari, R.T. (2017). *Data Mining Teori dan Aplikasi Rapid Miner*. Yogyakarta: Gava Media.

Yohannes. (2016). *Analisis Perbandingan Algoritma Fuzzy C-Means dan K-Means*. Palembang: STMIK-Global Informatika MDP.

Yulianto, S. & Hidayatullah, K.H. (2014). Analisis Klaster untuk Pengelompokan Kabupaten/Kota di Provinsi Jateng berdasarkan Indikator Kesejahteraan Rakyat. *Jurnal Akd. Statistika Muhammadiyah Semarang*.