JuTech
Jurnal Ilmiah Teknologi & Informasi

Website: http://ojs.itb-ad.ac.id/index.php/JUTECH
E-ISSN: 2797 - 4111

IMPLEMENTASI TEKNIK CLUSTERING BERBASIS VALIDASI CLUSTER PADA DATA STUNTING

Hardian Oktavianto[1], Nur Qodariyah Fitriyah[2] (*)

[1]*Universitas Muhammadiyah Jember*
[2]*Universitas Muhammadiyah Jember*

Abstract

Kecenderungan prevalensi status gizi balita di Jawa Timur mengalami penurunan dari tahun 2010 hingga tahun 2015. Sementara itu, menurut data Kementerian Kesehatan Republik Indonesia, persentase balita pendek di Jawa Timur pada tahun 2015 adalah sebesar 27,1%. Persentase tersebut menempatkan Provinsi Jawa Timur menjadi wilayah dengan persentase balita pendek tertinggi di Pulau Jawa. Data mining adalah penambangan yang mencirikan proses dalam menemukan informasi berharga dalam sekumpulan data yang sangat besar. Clustering merupakan salah satu cara untuk menganalisis data melalui pengelompokan objek data yang mirip satu sama lain dalam cluster yang sama dan berbeda dengan objek di cluster lain, sehingga setiap objek yang menjadi anggota cluster yang sama akan memiliki karakteristik yang lebih mirip satu sama lain dibandingkan dengan objek dari cluster berbeda. Salah satu metode clustering yang banyak digunakan adalah K-Means, yaitu metode clustering yang termasuk ke dalam jenis non-hierarchical. Penelitian ini melakukan pengelompokan kabupaten di provinsi Jawa Timur berdasarkan faktor – faktor yang mempengaruhi stunting melalui teknik data mining yaitu clustering yang secara khusus menerapkan algoritma K-Means dengan validasi cluster. Hasil dari penelitian ini adalah, bahwa dengan dataset yang ada, dan dengan menerapkan validasi cluster, maka diperoleh jumlah cluster optimal adalah sejumlah 10 cluster. Terdapat beberapa cluster yang hanya mempunyai 1 anggota, yaitu pada cluster 2, 4, dan 9, sedangkan anggota cluster terbanyak adalah pada cluster kelompok 7 yang mempunyai 11 anggota.

Kata Kunci: *Clustering, Stunting, Kmeans*

Juli – Desember 2021, Vol 2 (2) : hlm 61-70

(*) Korespondensi: hardian@unmuhjember.ac.id *(Hardian Oktavianto)*, nurfitriyah@unmuhjember.ac.id *(Nur Qodariyah Fitriyah)*

PENDAHULUAN

Stunting adalah balita dengan status gizi yang berdasarkan panjang atau tinggi badan menurut umurnya bila dibandingkan dengan standar baku WHO-MGRS (Multicentre Growth Reference Study) tahun 2005 , nilai z-score balita tersebut kurang dari -2SD dan masuk kategori sangat pendek jika nilai z-score balita tersebut kurang dari -3SD. Seorang balita dapat dikatakan mengalami kondisi stunting atau balita pendek, apabila hasil perbandingan antara tinggi badan dan nilai standar balita tersebut berada dibawah normal. Masalah stunting ini menggambarkan adanya masalah gizi kronis, dipengaruhi dari kondisi ibu atau calon ibu, masa janin, dan masa bayi atau balita, termasuk penyakit yang diderita selama masa balita. Selain itu, masalah stunting juga dipengaruhi kondisi – kondisi lain yang secara tidak langsung memengaruhi kesehatan (Indonesia, 2016).

Data "Profil Kesehatan Provinsi Jawa Timur Tahun 2015" menunjukan bahwa kecenderungan prevalensi status gizi balita di Jawa Timur mengalami penurunan atau perbaikan dari tahun 2010 hingga tahun 2015. Sementara itu, menurut data Kementerian Kesehatan Republik Indonesia, persentase balita pendek di Jawa Timur pada tahun 2015 adalah sebesar 27,1%. Persentase tersebut menempatkan Provinsi Jawa Timur menjadi wilayah dengan persentase balita pendek tertinggi di Pulau Jawa. Selain itu, disparitas yang sangat tinggi terjadi antar kabupaten/kota di Provinsi Jawa Timur dengan angka prevalensi stunting balita pada tingkatan yang sama dengan angka nasional, yaitu prevalensi stunting balita tinggi sebesar 30-39%. Di sisi lain, Rencana Pembangunan Jangka Menengah Nasional (RPJMN) 2015-2019 menetapkan target prevalensi stunting balita menurun dari angka 32,9% pada tahun 2013 menjadi sebesar 28% pada tahun 2019.

Data mining adalah penambangan yang mencirikan proses dalam menemukan informasi berharga dalam sekumpulan data yang sangat besar. Berbagai metode dalam data mining digunakan untuk menganalisis data demi mendapatkan informasi berharga (Han, Kamber, & Pei, 2011). Clustering merupakan salah satu cara untuk menganalisis data melalui pengelompokan objek data yang mirip satu sama lain dalam cluster yang sama dan berbeda dengan objek di cluster lain (Tan, Steinbach, & Kumar, 2005). Clustering merupakan teknik data mining yang bertujuan untuk menetapkan objek-objek menjadi sekumpulan kelompok terpisah yang disebut dengan cluster sehingga setiap objek yang menjadi anggota cluster yang sama akan memiliki karakteristik yang lebih mirip satu sama lain dibandingkan dengan objek dari cluster berbeda (Gorunescu, 2011).

Salah satu metode clustering yang banyak digunakan adalah K-Means, yaitu metode clustering yang termasuk ke dalam jenis non-hierarchical. Adapun kelemahan pada clustering dengan menggunakan metode K-Means adalah belum adanya ketentuan jumlah cluster yang tepat dari sejumlah k yang diujikan pada data. Tidak ada ukuran terbaik untuk pengelompokan data (Wu, Kumar, Ross Quinlan, & al, 2008). Hal ini menyebabkan informasi dari pengelompokan terkadang tidak sesuai dengan pola yang ada atau yang diharapkan. Cara yang dapat digunakan untuk menentukan cluster terbaik adalah dengan melakukan validasi cluster sehingga dapat diketahui berapa jumlah cluster yang optimal. Saat ini telah banyak metode atau

algoritma yang dihasilkan untuk menentukan jumlah cluster terbaik. Berbagai metode seperti Elbow, Dynamic Cluster Algorithm, Hamming Distance dan Xie Beni Index, merupakan cara yang ditemukan untuk menghasilkan jumlah cluster yang optimal. Metode-metode tersebut tidak berdiri sendiri dalam menentukan jumlah cluster yang tepat (Gosain & Dahiya, 2016). Terdapat pula teknik untuk mengevaluasi cluster yang dihasilkan berdasarkan error tertinggi dan index terkecil yang dihasilkan, seperti Sum of Square Error (SSE) dan Davies Bouldin Index (DBI). Kedua teknik ini sangat membantu dalam memastikan jumlah cluster yang paling optimal dari sejumlah k yang diujikan pada data (Ghosh & Dubey, 2013).

Penelitian ini melakukan pengelompokan kabupaten di provinsi Jawa Timur berdasarkan faktor – faktor yang mempengaruhi stunting melalui teknik data mining yaitu clustering yang secara khusus menerapkan algoritma K-Means dengan validasi cluster, dimana nantinya cluster yang terbentuk digunakan untuk melihat apakah kebijakan dan upaya yang dilakukan telah diterapkan pada daerah yang sesuai sehingga kebijakan untuk mengatasi kondisi stunting ini lebih tepat sasaran. Selain itu, hasil dari pengelompokan di Provinsi Jawa Timur dapat dijadikan rujukan bagi pemerintah untuk mengatasi prevalensi stunting balita pada tingkat nasional.

METODE

Tahapan atau langkah – langkah penelitian yang akan dilakukan secara umum terdiri dari 4 buah tahapan, mulai dari studi literatur, kemudian pengumpulan dan pemrosesan dataset, dilanjutkan dengan implementasi clustering, dan yang terakhir adalah penarikan kesimpulan atau analisis hasil. Tahapan atau langkah – langkah penelitian ini secara umum dapat dilihat pada gambar 1.

Gambar 1 Tahapan Penelitian

Studi literatur merupakan langkah yang dilakukan untuk mempelajari referensi berupa jurnal penelitian, paper, buku-buku referensi yang lain terkait dengan penelitian untuk melengkapi pengetahuan awal, guna memahami teori yang dapat digunakan untuk menunjang penelitian.

Dataset yang digunakan adalah data status gizi berat badan terhadap umur, status gizi tinggi badan terhadap umur, dan status gizi berat badan terhadap tinggi badan. Data tersebut mengambil dari website resmi BPS.

Clustering akan dilakukan menggunakan algoritma K-Means yang kemudian dilakukan validitas cluster menggunakan Davies Bouldin Index. Secara umum prosedur metode K-Means adalah sebagai berikut:

1) Tentukan k sebagai jumlah cluster yang di bentuk. Penentuan banyaknya jumlah cluster k dilakukan dengan beberapa faktor seperti pertimbangan teoritis dan konseptual yang diusulkan untuk menentukan berapa banyak cluster.

2) Bangkitkan k Centroid (titik pusat cluster) awal secara random. Untuk menentukan centroid awal dilakukan secara acak dari beberapa objek yang tersedia sebanyak k cluster, untuk menghitung centroid cluster ke-i berikutnya, menggunakan rumus sebagai berikut:

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \; ; i = 1,2,3, \dots n$$

Dimana:
V = centroid pada cluster
x_i = obyek ke-i
N = banyaknya obyek atau jumlah obyek yang menjadi anggota cluster

3) Hitung jarak setiap objek ke masing-masing centroid dari masing-masing cluster. Kemudian hitung jarak antara objek dengan centroid, dalam penelitian ini menggunakan Euclidian Distance:

$$d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \; ; i = 1,2,3, \dots n$$

Dimana;
xi : objek x ke-i
y: daya y ke-i
n: banyaknya objek

4) Alokasikan masing-masing objek ke dalam centroid yang paling terdekat.
5) Lakukan iterasi, kemudian tentukan posisi centroid baru dengan menggunakan persamaan
6) Ulangi langkah 3 jika posisi centroid baru tidak sama.

Validasi cluster yang terbentuk menggunakan Davies Bouldin Index (DBI), yaitu metode yang diusulkan oleh oleh David L. Davies dan Donald W.Bouldin pada tahun 1979, yang mengukur baik atau tidaknya hasil cluster dilihat dari kuantitas dan kedekatan antar data hasil cluster (Davies & Bouldin, 1979). DBI adalah fungsi rasio dari jumlah antara cluster scatter sampai dengan cluster separation. Pendekatan pengukuran DBI yaitu memaksimalkan jarak inter cluster serta meminimalkan jarak intra cluster (Rose, 2016). Semakin kecil nilai DBI menunjukan skema cluster yang paling optimal (He, Zhang, & Sun, 2014). Rumus DBI adalah:

$$R_{j,k} = \frac{MAE_j + MAE_k}{d(c_j, c_k)}$$

$$DBI = \frac{1}{M} \sum_{j=1}^{M} \max_{j \neq k} R_{j,k}$$

Yang terakhir adalah tahap analisis hasil, yaitu melakukan pendeskrispsian terhadap cluster yang terbentuk, atau umumnya disebut dengan profiling cluster. Hasil yang diharapkan yaitu bisa mendapatkan informasi terkait kasus stunting berdasarkan atribut yang dipakai.

HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam penelitian ini adalah data status gizi berat badan terhadap umur, status gizi tinggi badan terhadap umur, dan status gizi berat badan terhadap tinggi badan. Data tersebut mengambil dari website resmi BPS. Adapun data tersebut merupakan nilai dari masing – masing propinsi di Jawa Timur.

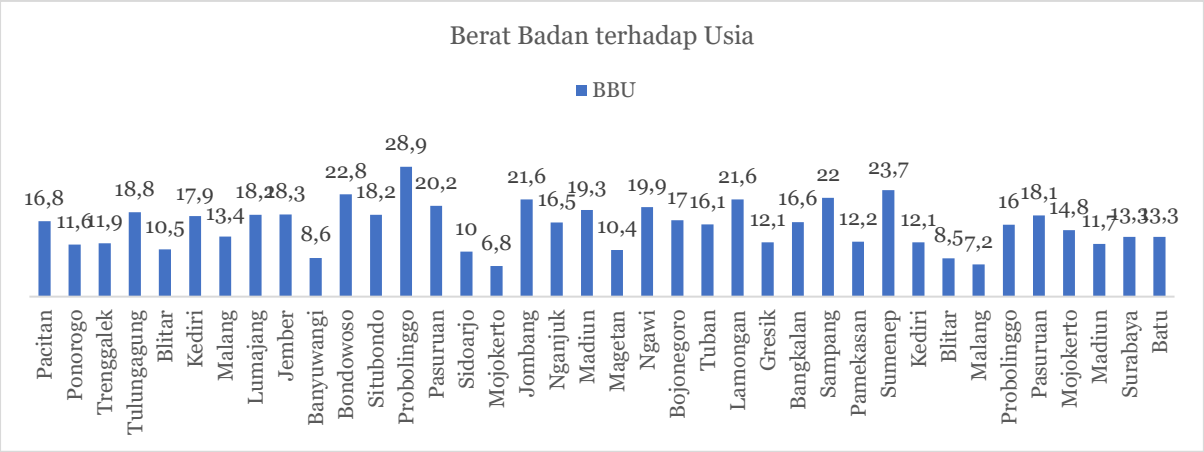
Dataset berasal dari 3 buah data yang berbeda, yang kemudian dijadikan satu sehingga menjadi dataset yang siap untuk dilakukan clustering. Dataset dapat dilihat pada tabel 1. Adapun visualisasi secara grafis dapat dilihat pada gambar 2, gambar 3, dan gambar 4, untuk masing – masing atribut.

Tabel 1. Dataset

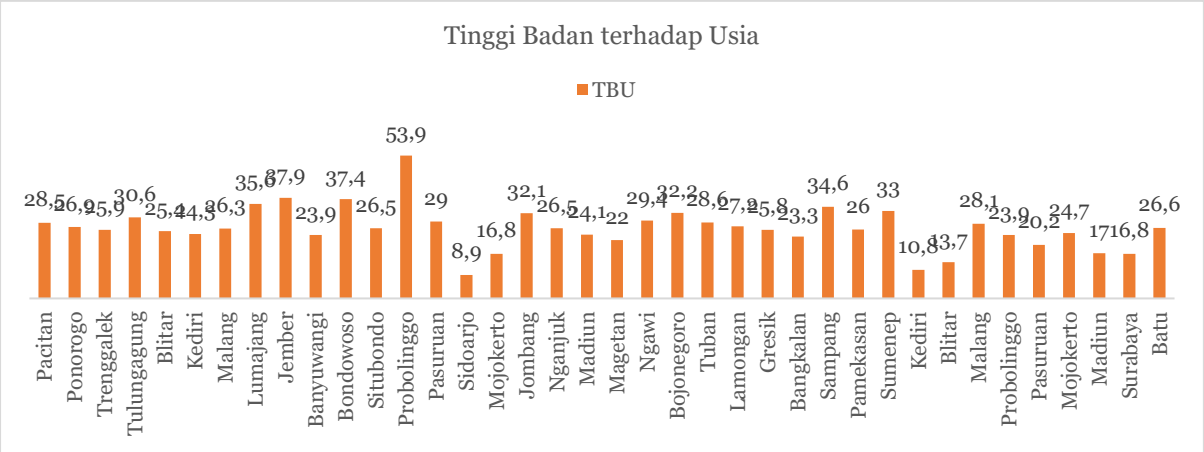
KABKOTA	BBU	TBU	BBTB
Pacitan	16.8	28.5	6.4
Ponorogo	11.6	26.9	9.1
Trenggalek	11.9	25.9	3
Tulungagung	18.8	30.6	7.2
Blitar	10.5	25.4	2.4
Kediri	17.9	24.3	7.5
Malang	13.4	26.3	3.5
Lumajang	18.2	35.6	6.8
Jember	18.3	37.9	6.1
Banyuwangi	8.6	23.9	6
Bondowoso	22.8	37.4	8.5
Situbondo	18.2	26.5	7.9
Probolinggo	28.9	53.9	5.6
Pasuruan	20.2	29	7.2
Sidoarjo	10	8.9	9.1
Mojokerto	6.8	16.8	1.4
Jombang	21.6	32.1	5.1
Nganjuk	16.5	26.5	8.1
Madiun	19.3	24.1	10.9
Magetan	10.4	22	3.5
Ngawi	19.9	29.4	10.3
Bojonegoro	17	32.2	6.4
Tuban	16.1	28.6	7.5
Lamongan	21.6	27.2	5.5
Gresik	12.1	25.8	6
Bangkalan	16.6	23.3	14.6
Sampang	22	34.6	6.8
Pamekasan	12.2	26	3.7
Sumenep	23.7	33	3.5
Kediri	12.1	10.8	21.4

Blitar	8.5	13.7	5.2
Malang	7.2	28.1	2.7
Probolinggo	16	23.9	3.9
Pasuruan	18.1	20.2	7.1
Mojokerto	14.8	24.7	9.4
Madiun	11.7	17	10.7
Surabaya	13.3	16.8	9
Batu	13.3	26.6	0.1

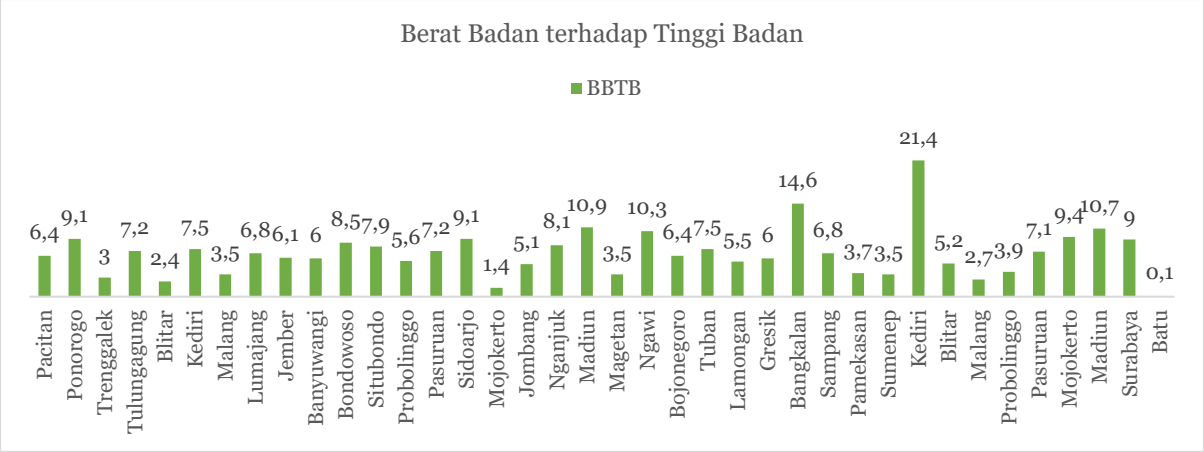
Sumber : BPS



Gambar 2. Atribut Berat Badan terhadap Usia



Gambar 3. Atribut Tinggi Badan terhadap Usia

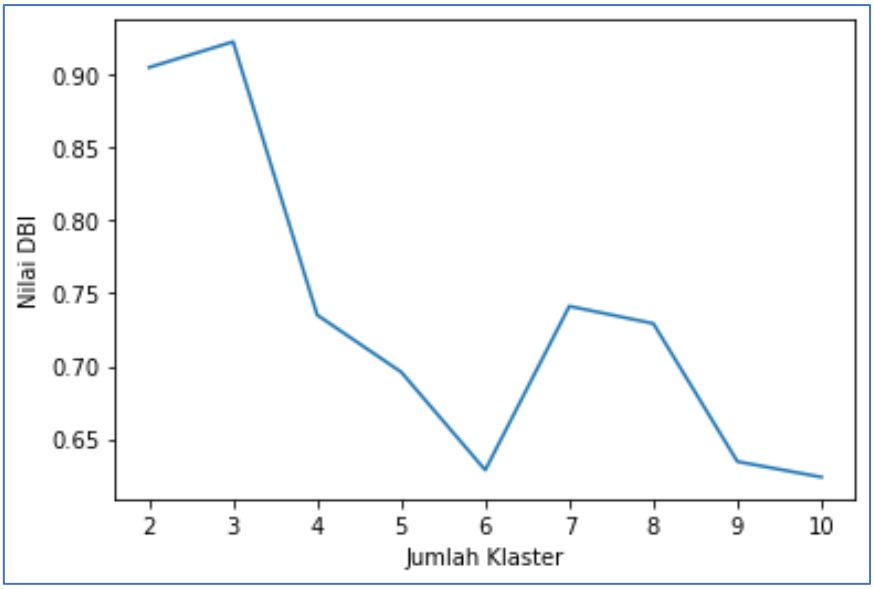


Gambar 4. Atribut Berat Badan terhadap Tinggi Badan

Proses clustering dilakukan dengan bantuan tools python, dengan menggunakan library dari *sklearn* dan *pandas* yang telah disediakan pada python. Clustering dilakukan dengan melakukan percobaan mulai dari 2 cluster hingga 10 cluster, dan dari masing – masing percobaan cluster, langsung dihitung nilai davies bouldin index (DBI) nya. Nilai DBI disajikan pada tabel 2 sedangkan grafik nilai DBI ditunjukkan pada gambar 5.

Tabel 2. Nilai DBI

Jumlah Cluster	Nilai DBI
2	0.904589355
3	0.922163388
4	0.735000606
5	0.695911986
6	0.628770342
7	0.741034956
8	0.729160562
9	0.634587735
10	0.623955184



Gambar 5. Grafik Nilai DBI

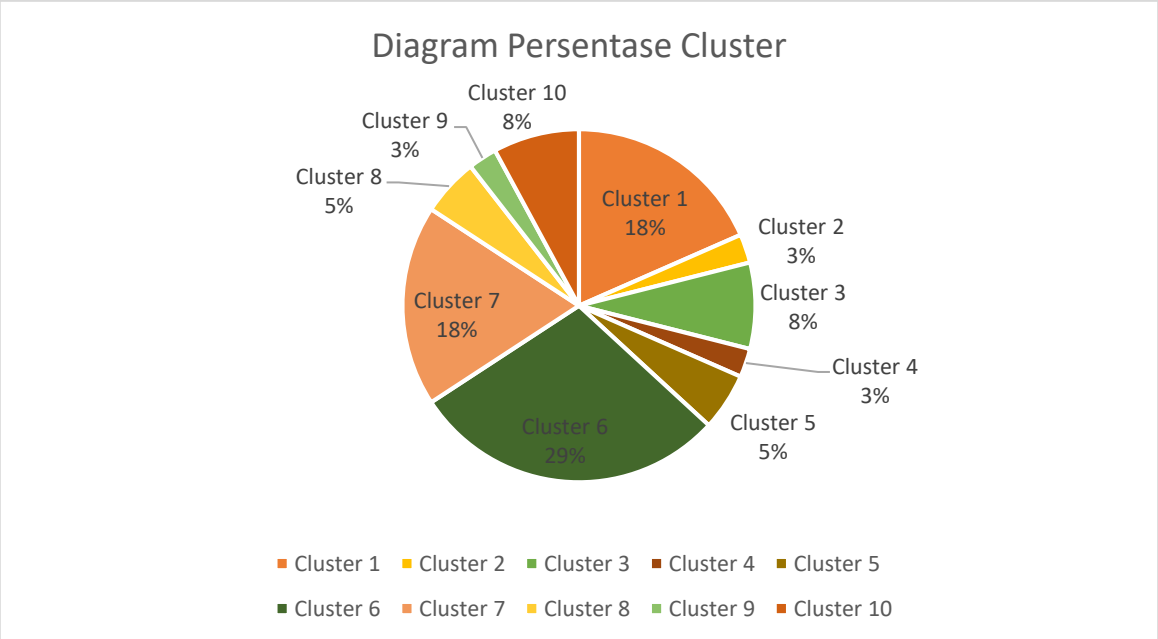
Nilai DBI yang paling rendah menunjukkan bahwa jumlah cluster yang terbentuk tersebut adalah yang paling optimal. Dari grafik dapat kita lihat bahwa nilai DBI paling rendah adalah ketika terbentuk 10 cluster, sehingga cluster optimum adalah 10 cluster dimana nilai DBI dengan pembulatan adalah sebesar 0.624.

Hasil 10 cluster yang terbentuk disajikan pada tabel 3 yang diurutkan berdasarkan cluster yang terbentuk. Dari kelompok yang terbentuk, terdapat beberapa analisis yang menarik, terutama jika dilihat dari sudut pandang jumlah anggota masing – masing cluster. Terdapat beberapa cluster yang hanya mempunyai 1 anggota, yaitu pada cluster 2, 4, dan 9, sedangkan anggota cluster terbanyak adalah pada cluster kelompok 7 yang mempunyai 11 anggota.

Tabel 3. Kelompok Cluster

KABKOTA	BBU	TBU	BBTB	KLASTER
Kediri	17.9	24.3	7.5	1
Situbondo	18.2	26.5	7.9	1
Nganjuk	16.5	26.5	8.1	1
Madiun	19.3	24.1	10.9	1
Bangkalan	16.6	23.3	14.6	1
Pasuruan	18.1	20.2	7.1	1
Mojokerto	14.8	24.7	9.4	1
Sidoarjo	10	8.9	9.1	2
Lumajang	18.2	35.6	6.8	3
Jember	18.3	37.9	6.1	3
Bondowoso	22.8	37.4	8.5	3
Probolinggo	28.9	53.9	5.6	4
Mojokerto	6.8	16.8	1.4	5
Blitar	8.5	13.7	5.2	5
Ponorogo	11.6	26.9	9.1	6
Trenggalek	11.9	25.9	3	6

Blitar	10.5	25.4	2.4	6
Malang	13.4	26.3	3.5	6
Banyuwangi	8.6	23.9	6	6
Magetan	10.4	22	3.5	6
Gresik	12.1	25.8	6	6
Pamekasan	12.2	26	3.7	6
Malang	7.2	28.1	2.7	6
Probolinggo	16	23.9	3.9	6
Batu	13.3	26.6	0.1	6
Pacitan	16.8	28.5	6.4	7
Tulungagung	18.8	30.6	7.2	7
Pasuruan	20.2	29	7.2	7
Ngawi	19.9	29.4	10.3	7
Bojonegoro	17	32.2	6.4	7
Tuban	16.1	28.6	7.5	7
Lamongan	21.6	27.2	5.5	7
Madiun	11.7	17	10.7	8
Surabaya	13.3	16.8	9	8
Kediri	12.1	10.8	21.4	9
Jombang	21.6	32.1	5.1	10
Sampang	22	34.6	6.8	10
Sumenep	23.7	33	3.5	10



Gambar 6. Diagram Persentase Cluster

KESIMPULAN

Kesimpulan dari penelitian ini adalah, bahwa dengan dataset yang ada, dan dengan menerapkan validasi cluster, maka diperoleh jumlah cluster optimal adalah sejumlah

10 cluster. Terdapat beberapa cluster yang hanya mempunyai 1 anggota, yaitu pada cluster 2, 4, dan 9, sedangkan anggota cluster terbanyak adalah pada cluster kelompok 7 yang mempunyai 11 anggota.

DAFTAR PUSTAKA

Davies, D. L.;& Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 224-227.

Ghosh, S.;& Dubey, S. K. (2013). Comparative Analysis of KMeans and Fuzzy C-Means Algorithms. *International Journal of Adavnced Computer Science and Application Vol. 4, No. 4*, 35-39.

Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques*. Heidelberg: Springer Science dan Business Media.

Gosain, A.;& Dahiya, S. (2016). Performance Analysis of Various Fuzzy Clustering Algorithms: A Review. *Procedia Computer Science*, 100-111.

Han, J.;Kamber, M.;& Pei, J. (2011). *Data Mining: Concepts and Techniques, 3rd ed*. Morgan Kaufmann Publishers.

He, Y.;Zhang, K.;& Sun, Z. (2014). A possibilistic fuzzy c-means clustering algorithm based on improved particle swarm optimization. *Journal of Computational Information Systems 10(18)*, 7845-7857.

Indonesia, K. K. (2016). *Situasi Balita Pendek*. Jakarta Selatan: Kementerian Kesehatan RI Pusat Data dan Informasi.

Rose, J. (2016). An efficient association rule based hierarchical algorithm for text clustering. *International Journal of Advanced Engineering Technology 7(4)*, 751-753.

Tan, P.;Steinbach, M.;& Kumar, V. (2005). *Cluster Analysis: Basic Concepts and Algorithms. In: Introduction to Data Mining*. Boston: Addison-Wesley.

Wu, X.;Kumar, V.;Ross Quinlan, J.;& al, e. (2008). Top 10 Algorithms in Data Mining. *Knowledge and Information Systems 14*, 1-37.