EXPERT JURNAL MANAJEMEN SISTEM INFORMASI DAN TEKNOLOGI

Implementasi Algoritma K-Means dan C-Means untuk Clustering Angka Kemiskinan

Chairul Habibi [1], **Reni Nursyanti** [2*]
[1] Sistem Informasi, Fakultas Teknologi dan Informatika, Universitas Informatika dan Bisnis Indonesia, Bandung
[2] Informatika, Fakultas Teknologi dan Informatika, Universitas Informatika dan Bisnis Indonesia, Bandung
Jawa Barat, Indonesia
[1] habibi.crl@gmail.com, [2*] reninursyanti@gmail.com

ABSTRACT – Poverty is one of the problems that must be faced by developing countries, including Indonesia and especially the province of West Java. This problem is exacerbated by the Covid-19 pandemic. Poverty can also have other consequences, such as increased crime and death. To facilitate government programs and support, it is necessary to group cities/districts according to the poverty level. The analysis was carried out using the K-Means and Fuzzy C-Means algorithms with the Silhouette method to obtain the optimal number of clusters using RStudio tools. The purpose of this study is to compare which algorithm is based on the Davis-Bouldin Index validation test. Three of the five data generated, the K-Means and C-Means algorithms give the same results. Only poverty data and education data give different results. Based on the results of the Davies-Bouldin Index validation test, the fuzzy c-means and k-means algorithms show that the k-means algorithm is better at clustering with an average of 4.084271. Meanwhile, fuzzy c-means has an average validation score of 4.111375. The smaller the Davies-Bouldin Index value or the closer to 0 shows how good the cluster is.

Keywords: Fuzzy; C-Means; K-Means; Poverty.

ABSTRAK – Kemiskinan merupakan salah satu masalah yang harus dihadapi oleh negara-negara berkembang, termasuk Indonesia dan khususnya provinsi Jawa Barat. Masalah ini diperparah dengan adanya pandemic Covid-19. Kemiskinan juga dapat memiliki konsekuensi lain, seperti meningkatnya kejahatan dan kematian. Untuk memfasilitasi program dan dukungan pemerintah, perlu dilakukan pengelompokan kota atau kabupaten menurut tingkat kemiskinan. Analisis dilakukan menggunakan algoritma K-Means dan Fuzzy C-Means dengan metode Silhouette untuk mendapatkan jumlah cluster yang optimal menggunakan tools RStudio. Tujuan dari penelitian ini adalah untuk membandingkan algoritma mana yang didasarkan pada uji validasi Indeks Davis-Bouldin. Tiga dari lima data yang dihasilkan, algoritma K-Means dan C-Means memberikan hasil yang sama. Hanya data kemiskinan dan data pendidikan yang memberikan hasil berbeda. Berdasarkan hasil uji validasi Davies-Bouldin Index, algoritma fuzzy c-means dan k-means menunjukkan bahwa algoritma k-means lebih baik dalam clustering dengan rata-rata 4.084271. Sedangkan fuzzy c-means memiliki rata-rata skor validasi sebesar 4,111375. Semakin kecil nilai Davies-Bouldin Index atau semakin mendekati nilai 0 menunjukkan seberapa baik cluster yang diperoleh.

Kata Kunci: Fuzzy; C-Means; K-Means; Angka Kemiskinan.

1. PENDAHULUAN

Kemiskinan merupakan masalah sosial yang dapat mempengaruhi individu dan masyarakat secara keseluruhan. Kemiskinan dapat menimbulkan dampak lain, seperti meningkatnya kriminalitas di suatu daerah dan dapat menjadi salah satu penyebab terjadinya kejahatan [1]. Hal ini karena masyarakat miskin cenderung mencari kebutuhan pokoknya dengan cara apapun [2]. Kesulitan dalam akses kesehatan ini dapat menyebabkan peningkatan angka kematian di suatu populasi, terutama di kalangan masyarakat yang hidup dalam kemiskinan. [3]. Berdasarkan data Badan Pusat Statistik (BPS) Jawa Barat, besarnya angka kemiskinan di Jawa Barat adalah 4.195.34 jiwa [4]. Peran pemerintah dalam mengategorikan wilayah-wilayah yang mengalami penurunan ekonomi adalah sebagai faktor yang dipertimbangkan dalam menentukan program yang kemudian dapat meningkatkan perekonomian masyarakat Jawa Barat. Setiap wilayah memiliki data tingkat kemiskinan yang berbeda sehingga untuk menghitung angka kemiskinan, BPS menggunakan pendekatan kebutuhan primer masyarakat dalam memenuhi kebutuhan dasarnya dalam hal pangan [5]. Banyaknya jumlah kota dan kabupaten di daerah Jawa Barat ialah 18 kabupaten dan 9 kota sehingga untuk menganalisis dan mengumpulkan data penting membutuhkan waktu yang relatif lama. Pada penelitian ini mengunakan metode Data Mining dimana suatu proses pengumpulan informasi dan data yang penting dalam jumlah besar atau *big data* [6]. Dalam Data Mining terdapat beberapa algoritma yang digunakan untuk mengelompokan data salah satunya adalah algoritma K-Means. Algoritma K-Means adalah algoritma *clustering* yang paling sederhana dibanding algoritma *clustering* lainnya. Algoritma ini mempunyai kelebihan mudah diterapkan dan dijalankan, relatif cepat,

mudah untuk diadaptasi, dan paling banyak dipraktekkan dalam tugas Data Mining [7]. Selain metode K-Means, ada pendekatan lain untuk mengelompokkan data yaitu logika Fuzzy. Dalam teori Fuzzy, keanggotaan sebuah data tidak diberi nilai secara tegas dengan nilai 1 (menjadi anggota) dan 0 (tidak menjadi anggota), melainkan dengan suatu nilai derajat keanggotaan yang jangkauan nilainya 0 sampai 1 [8]-[10]. Dalam pengelompokan data, algoritma Fuzzy C-Means adalah salah satu metode yang digunakan dalam logika Fuzzy [11]. Dalam sebuah penelitian menjelaskan bahwa metode Fuzzy C-Means adalah metode yang lebih baik untuk melakukan pengelompokan pada data *user knowledge modeling* dikarenakan nilai validitasnya lebih mendekati nilai 1 [12].

Tujuan dari dilaksanakannya penelitian ini adalah untuk mengkategorikan tingkat kemiskinan menjadi beberapa kelompok yang memenuhi kriteria tinggi, cukup tinggi, cukup rendah, dan rendah serta menentukan metode terbaik antara K-Means atau Fuzzy C-Means untuk pengelompokan tingkat kemiskinan di Jawa Barat.

2. DASAR TEORI

Perbandingan K-Means dan Fuzzy C-Means dalam Pengelompokan Daerah Penyebaran Covid-19 Indonesia. Algoritma clustering dapat diterapkan untuk pengelompokan daerah penyebaran Covid-19 pada 34 provinsi di Indonesia berbantuan software RStudio dimana pada penelitian ini melakukan perbandingan dua algoritma yaitu algoritma K-Means dan Fuzzy C-Means. Diperoleh pengelompokan daerah penyebaran Covid-19 pada 34 Provinsi Indonesia lebih baik dilakukan menggunakan algoritma K-Means [13].

Implementasi Data Mining dalam Penentuan Tingkat Kemiskinan Menggunakan Fuzzy C-Means. Sistem aplikasi yang dibuat telah berhasil melakukan klaster 100 data sampel rumah tangga miskin di Kabupaten Bone Bolango Provinsi Gorontalo dalam 3 kategori kemiskinan, dengan persentase setiap kategori adalah 50% sangat miskin, 34% hampir miskin, dan 16% sangat miskin. Adapun saran pengembangannya adalah perlu penggunaan algoritma *clustering* yang berbeda dalam mengelompokkan kategori dan indikator kemiskinan untuk memperoleh hasil yang lebih bervariasi [14].

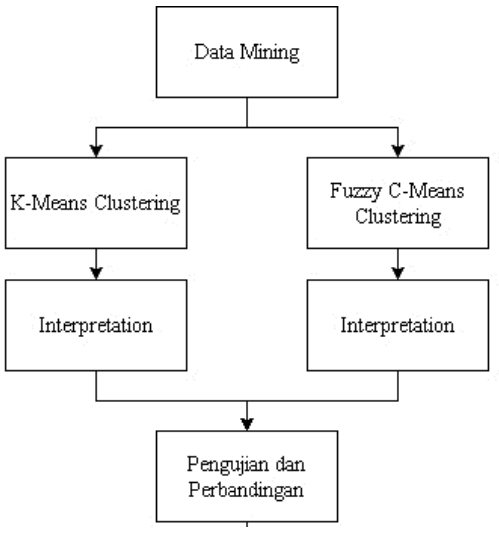
Perbandingan Algoritma K-Means dengan Algoritma Fuzzy C-Means untuk Clustering Tingkat Kedisiplinan Kinerja Karyawan. Hasil *cluster* dari data presensi karyawan menggunakan metode K-Means dan Fuzzy C-Means berbeda. Hal ini dapat dilihat dari jumlah cluster yang diperoleh dari kedua metode tersebut. Dilihat dari hasil validasi, Fuzzy C-Means dominan menghasilkan metode yang lebih baik, dengan nilai validasinya adalah 0.758 dikarenakan nilai validasinya lebih mendekati nilai 1, dibandingkan dengan metode K-Means dengan nilai validasinya adalah 0.528 [15].

Perbandingan Fuzzy C-Means dan K-Means untuk Mengelompokkan Tingkat Buta Huruf Berdasarkan Provinsi di Indonesia. Berdasarkan hasil uji validitas algoritma Fuzzy C-Means serta K-Means menunjukkan bahwa algoritma Fuzzy C-Means lebih baik dalam melakukan clustering dengan nilai validitas 0.8036. Sedangkan pada algoritma K-Means memiliki nilai validitas 0.7894. Validitas yang mendekati nilai 1 memiliki kualitas cluster yang lebih baik [16].

3. METODOLOGI

Metode penelitian ini adalah kuantitatif dimana proses penelitian menggunakan data numerik dan pemodelan Data Mining. Adapun tahapan *clustering* yang di lakukan terdapat pada Gambar 1.

Pada tahap *clustering* Algoritma K-Means maupun C-Means dilakukan dengan menggunakan software RStudio dengan menggunakan data kemiskinan Kabupaten/Kota tahun 2021 yang diperoleh dari Badan Pusat Statistik (BPS) seperti pada Tabel 1 dimana pada table tersebut P1 adalah rata-rata selisih pengeluaran per kapita penduduk miskin dengan garis kemiskinann dan P2 adalah rata-rata dari kuadrat selisih pengeluaran per kapita penduduk miskin dengan garis kemiskinan.



Gambar 1. Tahapan *Clustering*

DOI http://dx.doi.org/10.36448/expert.v13i1.3107
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 13 No. 1
June 30, 2023 – Hal. 51

Tabel 1. Penduduk Miskin, P1, P2, dan Garis Kemiskinan

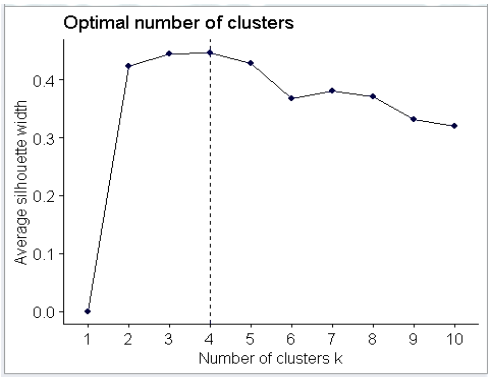
No	Kabupaten/Kota	Jumlah Penduduk Miskin	Persentase Penduduk Miskin	P1	P2	Garis Kemiskinan (Rp/Kap/Bulan)
1	Bogor	491.24	8.13	1.36	0.35	418.483
2	Sukabumi	194.35	7.70	1.04	0.23	342.094
3	Cianjur	260.02	11.18	1.83	0.46	387.631
4	Bandung	269.18	7.15	1.23	0.34	378.819
5	Garut	281.36	10.65	1.40	0.29	320.050
6	Tasikmalaya	200.59	11.15	1.56	0.35	333.909
7	Ciamis	96.60	7.97	0.88	0.15	389.676
8	Kuningan	143.35	13.10	2.01	0.46	358.069
9	Cirebon	271.02	12.30	1.94	0.47	404.635
10	Majalengka	151.14	12.33	2.44	0.77	466.813
11	Sumedang	126.28	10.71	1.71	0.46	360.054
12	Indramayu	228.59	13.04	2.46	0.66	481.754
13	Subang	158.97	10.03	1.92	0.50	360.691
14	Purwakarta	84.27	8.83	1.31	0.30	387.754
15	Karawang	210.78	8.95	1.27	0.29	494.201
16	Bekasi	202.73	5.21	0.91	0.26	549.875
17	Bandung Barat	190.77	11.30	1.62	0.32	374.470
18	Pangandaran	39.07	9.65	1.25	0.27	394.101
19	Kota Bogor	80.09	7.24	1.10	0.27	571.425
20	Kota Sukabumi	27.19	8.25	1.39	0.35	567.734
21	Kota Bandung	112.50	4.37	0.78	0.24	515.396
22	Kota Cirebon	31.98	10.03	2.22	0.68	467.248
23	Kota Bekasi	144.12	4.74	0.66	0.16	692.885
24	Kota Depok	63.86	2.58	0.34	0.07	705.084
25	Kota Cimahi	32.48	5.35	0.92	0.21	522.281
26	Kota Tasikmalaya	89.46	13.13	2.41	0.69	480.341
27	Kota Banjar	13.37	7.11	1.18	0.28	357.210

4. HASIL DAN PEMBHANASAN

Tahapan K-Means *Clustering* (1) – Perancangan algoritma K-Means *Clustering* diimplementasikan pada aplikasi RStudio dalam beberapa tahapan, yaitu:

a. Menginputkan data pada RStudio menggunakan fungsi *read.delim* yang artinya data diinput melalui *clipboard* lalu dijalankan fungsi *read.delim*.
b. Proses *preprocessing* menggunakan atribut jumlah presentase penduduk miskin dan atribut bekerja, lalu data ditransformasi menggunakan skala agar memiliki parameter atau ukuran yang sama untuk *clustering*.
c. Menentukan jumlah *cluster* menggunakan Metode Silhouette karena hasil dari metode ini lebih jelas. Jumlah cluster data miskin pada Gambar 2.
d. Menentukan K-Means *Clustering,* pada proses ini data yang sudah di *preporcessing* akan dieksekusi menjadi beberapa *cluster* yang didasarkan dari hasil Metode Silhouette.
e. Hasil dari K-Means *Clustering* pada data kemiskinan akan menunjukkan tabel *cluster, cluster plot,* dan labelisasi wilayah berdasarkan tingkat kemiskinannya.

Dari hasil *cluster* yang telah diurutkan bahwa *Cluster 2* adalah wilayah dengan tingkat kemiskinan terendah, Gambar 3. *Cluster 1* adalah wilayah dengan kemiskinan cukup rendah. *Cluster 3* adalah wilayah dengan kemiskinan cukup tinggi dan *Cluster 4* adalah wilayah dengan tingkat kemiskinan tertinggi, Tabel 2.



Gambar 2. Jumlah Cluster Data Kemiskinan

Gambar 3. Cluster Plot K-Means Clustering Data Kemiskinan

DOI http://dx.doi.org/10.36448/expert.v13i1.3107
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 13 No. 1
June 30, 2023 – Hal. 52

Tabel 2. Tabel Anggota Cluster K-Means Data Kemiskinan

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Karawang	Bekasi	Bogor	Cianjur
Kota Bogor	Kota Bandung	Sukabumi	Kuningan
Kota Sukabumi	Kota Bekasi	Bandung	Cirebon
	Kota Depok	Garut	Majalengka
	Kota Cimahi	Tasikmalaya	Sumedang
		Ciamis	Indramayu
		Purwakarta	Subang
		Bandung Barat	Kota Cirebon
		Pangandaran	Kota Tasikmalaya
		Kota Banjar	

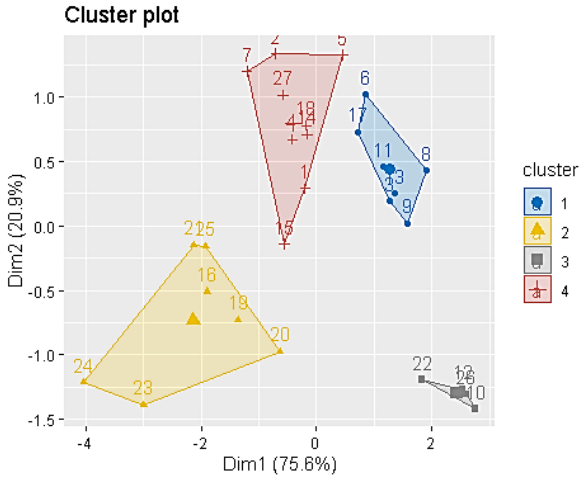
Tahapan Fuzzy C-Means *Clustering* (2) – Perancangan alur algoritma Fuzzy C-Means Clustering yang diimplementasikan pada aplikasi RStudio dengan beberapa tahapan yaitu:

a. Data yang sudah di-*preporcessing* akan dieksekusi menjadi beberapa cluster yang didasarkan dari hasil Metode Silhouette. Semua proses perhitungan clustering dilakukan secara otomatis oleh RStudio.
b. Hasil dari Fuzzy C-Means *Clustering* pada data kemiskinan akan menunjukkan *table cluster, cluster plot* dan labelisasi wilayah berdasarkan tingkat kemiskinannya.

Pemetaan *cluster* dari data kemiskinan pada Gambar 4 terdapat 4 *clusters* (sebelah kanan gambar), *cluster* belum diurutkan dari yang terendah sampai yang tertinggi berdasarkan tingkat kemiskinannya. Jadi perlu dilakukan pengurutan lagi agar tidak terjadi kesalahan saat proses pelabelan.

Berdasarkan Tabel 3, *Cluster 2* adalah wilayah dengan tingkat kemiskinan terendah. *Cluster 4* adalah wilayah dengan kemiskinan cukup rendah. *Cluster 1* adalah wilayah dengan kemiskinan cukup tinggi dan *Cluster 3* adalah wilayah dengan tingkat kemiskinan tertinggi.

Berdasarkan dari hasil yang didapatkan, maka hasil-hasil dari clustering setiap data dapat dibandingkan dalam Tabel 4.



Gambar 4. Cluster Plot Fuzzy C-Means Clustering Data Kemiskinan

Tabel 3. Tabel Anggota Fuzzy C-Means Data Kemiskinan

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cianjur	Bekasi	Majalengka	Bogor
Tasikmalaya	Kota Bogor	Indramayu	Sukabumi
Kuningan	Kota Sukabumi	Kota Cirebon	Bandung
Cirebon	Kota Bandung	Kota Tasikmalaya	Garut
Sumedang	Kota Bekasi		Ciamis
Subang	Kota Depok		Purwakarta
Bandung Barat	Kota Cimahi		Karawang
			Pangandaran
			Kota Banjar

Tabel 4. Tabel Perbandingan Data Kemiskinan

No	Kabupaten/Kota	K-Means	C-Means
1	Bogor		
2	Sukabumi		
3	Cianjur		
4	Bandung		
5	Garut		
6	Tasikmalaya		
7	Ciamis		
8	Kuningan		
9	Cirebon		

No	Kabupaten/Kota	K-Means	C-Means
10	Majalengka		
11	Sumedang		
12	Indramayu		
13	Subang		
14	Purwakarta		
15	Karawang		
16	Bekasi		
17	Bandung Barat		
18	Pangandaran		

DOI http://dx.doi.org/10.36448/expert.v13i1.3107
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 13 No. 1
June 30, 2023 – Hal. 53

No	Kabupaten/Kota	K-Means	C-Means
19	Kota Bogor		
20	Kota Sukabumi		
21	Kota Bandung		
22	Kota Cirebon		
23	Kota Bekasi		
24	Kota Depok		
25	Kota Cimahi		
26	Kota Tasikmalaya		
27	Kota Banjar		

Warna:
Tingkat kemiskinan rendah
Tingkat kemiskinan cukup rendah
Tingkat kemiskinan cukup tinggi
Tingkat kemiskinan tinggi

Berdasarkan hasil uji validasi Tabel 5 bahwa dengan Davies-Bouldin Index algoritma Fuzzy C-Means dan K-Means menunjukkan bahwa algoritma K-Means lebih baik dalam melakukan clustering dengan rata-rata 4.084271. Sedangkan pada Fuzzy C-Means memiliki rata-rata nilai validasi 4.111375. Semakin kecil nilai Davies-Bouldin Index atau semakin mendekati nilai 0 menunjukkan seberapa baik *cluster* yang diperoleh.

Tabel 5. *Davies-Bouldin Index*

Data	DBI	
	K-Means	C-Means
1. Kemiskinan	2.736858	3.119746
2. Tamat Pendidikan	7.942894	8.305394
3. Status Bekerja	4.221959	3.603853
4. Pengeluaran Perkapita	1.385064	1.385064
5. Fasilitas Perumahan	4.13458	4.142818
Rata-rata	4.084271	4.111375

5. KESIMPULAN

Proses *clustering* algoritma K-Means dan Fuzzy C-Means memiliki tahap yang sama. Data diinput menggunakan fungsi *read.delim ("clipboard")* yang artinya data diinput melalui *history clipboard*. Kemudian data dilakukan *preprocessing* yaitu data *selection* dan *transformation* untuk mengeliminasi atribut yang tidak diperlukan. Jumlah *cluster* ditentukan menggunakan Metode Silhouette yang menghasilkan jumlah *cluster* optimal. Lalu dilakukan proses *clustering* kedua algoritma yang menampilkan akurasi dari proses *clustering*. Pada hasil *clustering* didapatkan 3 dari 5 data yang diuji, algoritma K-Means dan Fuzzy C-Means memiliki hasil yang sama. Hanya data kemiskinan dan data tamat pendidikan yang memiliki hasil berbeda, tetapi tidak berbeda terlalu jauh.

DAFTAR PUSTAKA

[1] Bahauddin, A., Fatmawati, A., & Sari, F. P. (2021). Analisis Clustering Provinsi di Indonesia Berdasarkan Tingkat Kemiskinan Menggunakan Algoritma K-Means. Jurnal Manajemen Informatika dan Sistem Informasi, 4(1), 1-8.

[2] D Widyadhan, RB Hastuti, I Kharisudin. (2021). Perbandingan analisis klaster k-means dan average linkage untuk pengklasteran kemiskinan di Provinsi Jawa Tengah. PRISMA.

[3] PN Safitri, R Aristawidya. (2021). Klasterisasi faktor-faktor kemiskinan di Provinsi Jawa Barat menggunakan k-medoids clustering.

[4] Badan Pusat Statistik. (2021). "Data dan Informasi Kemiskinan Kabupaten/Kota Tahun 2021". Badan Pusat Statistik, Jakarta. 161 hal.

[5] NS Fatonah, TK Pancarani. (2022). Analisa Perbandingan Algoritma Clustering Untuk Pemetaan Status Gizi Balita Di Puskesmas Pasir Jaya.

[6] Tan, P. N., Steinbach, M., & Kumar, V. (2006). Data mining introduction. People's. Posts and Telecommunications Publishing House, Beijing.

[7] B Budiman, R Nursyanti, RYR Alamsyah, I Akbar. (2020). Data mining implementation using naïve Bayes algorithm and decision tree J48 in determining concentration selection, International Journal of Quantitative Research and Modeling 1 (3), 123-134.

[8] EF Yogachi, VM Nasution, G Prakarsa (2020). Design and Development of Fuzzy Logic Application Mamdani Method in Predicting the Number of Covid-19 Positive Cases in West Java, IOP Conference Series: Materials Science and Engineering 1115 (1).

[9] AD Permana, VM Nasution, G Prakarsa (2020). Design and Development of Fuzzy Logic Application Tsukamoto Method in Predicting the Number of Covid-19 Positive Cases in West Java, International Journal of Global Operations Research 1 (2), 85-95.

[10] Erlangga, E., & Dharmawan, Y. Y. (2018). Penentuan Penerima Kinerja Dosen Award melalui Metode Tsukamoto dengan Konsep Logika Fuzzy. Explore: Jurnal Sistem Informasi dan Telematika, 9(2), 331236.

[11] G Prakarsa, VM Nasution (2021). Penerapan Logika Fuzzy Menggunakan Metode Mamdani Pada Prediksi Jumlah Kasus Positif Covid-19. Jurnal Media Informatika Budidarma 5 (4).

[12] Ramadhan, A., Efendi, Z., & Mustakim, M. (2017). Perbandingan K-Means dan Fuzzy C-Means untuk Pengelompokan Data User Knowledge Modeling. In Seminar Nasional Teknologi Informasi Komunikasi dan Industri (pp. 219-226).

[13] ALR Putri, N Dwidayati. Analisa perbandingan k-means dan fuzzy c-means dalam pengelompokan daerah penyebaran COVID-19 Indonesia (2021).

[14] NF Kahar, L Hadjaratie, S Suhada, IR Padiku. (2019). Implementasi Data Mining Dalam Penentuan Tingkat Kemiskinan Menggunakan Fuzzy C-Means. Jambura Journal of Informatics, 1(1), 27-36.

[15] Agustina, N., & Prihandoko, P. (2018). Perbandingan Algoritma K-Means dengan Fuzzy C-Means Untuk Clustering Tingkat Kedisiplinan Kinerja Karyawan. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 2(3), 621-626.

[16] Pamungkas, M. A. (2021). Perbandingan Fuzzy C-Means Dan K-Means untuk Mengelompokkan Tingkat Buta Huruf Berdasarkan Provinsi di Indonesia (Doctoral dissertation, Universitas Muhammadiyah Jember).

DOI http://dx.doi.org/10.36448/expert.v13i1.3107
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 13 No. 1
June 30, 2023 – Hal. 55