Jurnal Ilmiah Ilmu Komputer Vol. 7, No. 1, April 2021 (P) ISSN 2442-451X
Fakultas Ilmu Komputer (O) ISSN 2503-3832
Universitas AL Asyariah Mandar

KOMPARASI ALGORITMA K-MEANS, K-MEDOID, AGGLOMEARTIVE CLUSTERING TERHADAP GENRE SPOTIFY

Ana Rohmah Zaidah[1], Chandra Indira Septiarani[2], Mar'atus Sholikhatun Nisa[3], Ahmad Yusuf [4], Noor Wahyudi[5]

[1,2,3,4,5]Prrogram Studi Sistem informasi, UIN Sunan Ampel Surabaya, Indonesia
[1]anarohmahzaidah77@gmail.com, [2]cisrani@gmail.com, [3]MaratusSN@gmail.com,
[4]ahmadyusuf@uinsby.ac.id, [5]n.wahyudi@uinsby.ac.id.

ABSTRAK

Perkembangan teknologi di berbagai bidang kehidupan manusia berdampak pada kebutuhan manusia yang semakin kompleks seperti contohnya muncul layanan streaming musik yang bisa didengarkan dengan mudah dari berbagai platform. Platform streaming musik yang banyak digunakan saat ini salah satunya spotify.com, menurut berita yang dirilis kompas.com spotify yang mengalami kenaikan 130 juta pelanggan baru sejak pandemi, sebagai platform yang memiliki pengguna yang banyak perlu penelitian lebih lanjut tentang konten streaming musik untuk menawarkan user experience yang lebih baik dan upaya meningkatkan persaingan dengan platform streaming lain. Penelitian ini menggunakan data publik Global Top 50 yang dianalisis menggunakan algoritma k-means, k-medoid, agglomative clustering Berdasarkan hasil penelitian metode cluster terbaik yang digunakan untuk clustering lagu spotify adalah agglomerative hierarchical clustering metode Average Linked dengan 3 atau 4 klaster. Klaster paling direkomendasikan adalah sebesar 3 klaster karena klaster ke 4 hanya berisi 1 anggota saja. Pada pengelompokkan 3 klaster, klaster ke-1 berisi 2833 anggota, klaster ke-2 berisi 145 anggota, dan klaster ke-3 berisi 21 anggota.

Kata Kunci— Komparasi, Data Mining, K-Means, K-Medoid, Agglomerative clustering, Sportify

ABSTRACT

Technological developments in various fields of human life have an impact on increasingly complex human needs, for example, the emergence of music streaming services that can be heard easily from various platforms. The music streaming platform that is widely used today is spotify.com, according to news released by Kompas.com Spotify which has seen an increase of 130 million new subscribers since the pandemic, as a platform that has many users needs further research on music streaming content to offer users better experience and efforts to increase competition from other streaming platforms. This study uses Global Top 50 public data which is analyzed using the k-means, k-medoid, agglomerative clustering algorithm. The most recommended cluster is 3 clusters because the 4th cluster contains only 1 member. In the 3-cluster grouping, the 1st cluster contains 2833 members, the 2nd cluster contains 145 members, and the 3rd cluster contains 21 members.

Keywords— Comparison, Data Mining, K-Means, K-Medoid, Agglomerative clustering, Sportify

(P) ISSN 2442-451X
(O) ISSN 2503-3832

1. PENDAHULUAN

Teknologi berkembang dengan cepat merubah gaya hidup manusia, seperti contohnya muncul layanan streaming musik yang bisa didengarkan dengan mudah dari berbagai platform yang menawar konten musik seperti mp3, WMA[1]. Platform streaming musik yang banyak digunakan saat ini salah satunya spotify.com, menurut berita yang dirilis kompas.com spotify yang mengalami kenaikan 130 juta pelanggan baru sejak pandemi[2]. Spotify sebagai platform yang memiliki pengguna yang banyak tentu perlu sebuah penelitian lebih lanjut tentang streaming musik yang ditawarkan untuk menawarkan user experience yang lebih baik dan upaya meningkatkan persaingan dengan platform streaming lain melalui analisis data mining.

Pada penelitian ini menggunakan data publik Global Top 50 yang akan dikelompokkan berdasarkan genrenya karena menurut penelitian masyarakat cenderung ingin mendengarkan lagu sesuai jenis genre favorit[3]. Menggunakan perbandingan algoritma K Means, K-medoid,agglomerative yang selain hasilnya tujuan pengelompokan jenis genre musik akan melakukan perbandingan akurasi masing - masing algoritma untuk tujuan pengelompokan atau clustering. Pada penelitian sebelumnya algoritma.

Pada penelitian sebelumnya k-means sudah digunakan untuk proses pengelompokan pelanggan dengan analisis RFM yang hasilnya K Means memiliki kemampuan yang baik dalam identifikasi pelanggan[4],adapun penelitian tentang k-medoid sebelumnya tentang pengelompokan penyebaran covid 19 di indonesia menunjukkan kemampuan clustering yang baik terutama akurasi pada 3 cluster wilayah[5] sedangkan penelitian selanjutnya tentang pengelompokan genre cerpen di kompas menggunakan agglomerative clustering menghasilkan akurasi 47 % dengan perbandingan algoritma clustering yang mendapat akurasi 37%[6].

2. TINJAUAN PUSTAKA

2.1 K-Means

k-mean merupakan algoritma yang cara kerjanya membagi data numeric menjadi beberapa cluster[7]. Algoritma K-Means memiliki proses kerja yaitu [8] :

1. Menentukan k untuk cluster yang dianalisis
2. Menentukan k sebagai titik pusat secara acak
3. Menghitung jarak dari setiap k dari centroid dari cluster menggunakan rumus Euclidean distance
4. Mengalokasikan objek ke dalam centroid terdekat
5. proses iterasi, lalu menentukan tempat centroid yang baru
6. Mengulangi langkah no 3 jika tempat centroid yang baru tidak sama.

2.2 K-Medoids

K-Medoids dan bisa disebut algoritma PAM (Partitioning Around Medoid) adalah algoritma yang memiliki kemiripan dengan K-Means karena keduanya tujuannya membagi datas menjadi kelompok/cluster. Algoritma K-Means dan algoritma K-Medoids memiliki perbedaan pada penentuan pusat cluster, yang mana algoritma K-Means ditentukan dari nilai rata-rata (means) pada setiap cluster, sedangkan K-Medoids menggunakan objek data sebagai perwakilan (medoids)[9]. Berikut langkah - langkah yang digunakan dalam perhitungan algoritma K-Medoids[10]:

1. Menentukan pusat cluster sebanyak k
2. Mengalokasikan data untuk cluster terdekat dengan rumus ukuran jarak euclidean distance
3. Memilih acak objek pada setiap cluster yang digunakan untuk medoid baru
4. Menghitung jarak masing – masing objek berada pada setiap cluster.
5. kandidat medoids baru dengan menggunakan rumus euclidian distance
6. Menghitung total simpangan (S) menggunakan menghitung nilai total distance baru sampai total distance lama. Jika S<0, maka tukar objek dengan hasil data clustering untuk membentuk sekumpulan k objek baru untuk medoids.
7. Ulangi urutan 3 sampai 5 hingga tidak ada perubahan medoids, sampai
8. mendapatkan cluster beserta masing – masing anggota cluster.

2.3 Agglomerative Clustering

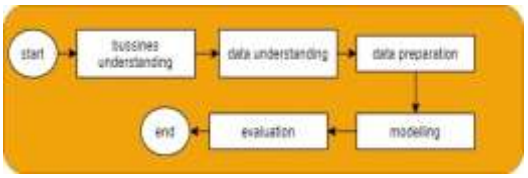
Algoritma agglomerative clustering merupakan algoritma pengelompokkan hierarki dengan pendekatan bawah-atas (bottom up)[11]. Alur proses dari penggunaan algoritma agglomerative clustering sebagai berikut[12]:

1. Menetapkan setiap objek ke cluster yang berbeda.
2. Mengevaluasi semua jarak berpasangan antara jarak cluster metrik
3. Membuat matriks jarak menggunakan nilai jarak.
4. Mencari pasangan cluster dengan jarak terpendek dan hapus pasangan cluster ini dari
5. matriks, lalu gabungkan.
6. Evaluasi semua jarak dari cluster baru ini ke semua lainnya cluster, dan memperbarui
7. matriks.
8. Ulangi sampai matriks jarak menjadi satu elemen

3. METODE YANG DIUSULKAN

Penelitian ini menggunakan bahasa R dengan analisis metode data mining Cross Standard Process For Data Mining (CRISP-DM) yang terdiri dari 6 fase, yaitu : business understanding, data understanding, data preparation, modelling evaluation, dan deployment (Gambar 1). Berikut penjelasan tiap fase pada metode CIRSP-DM[13]:

(P) ISSN 2442-451X
(O) ISSN 2503-3832



Gambar 1. Metode CIRSP-DM

3.1 Business Understanding Phase

Pada proses analisis bisnis dengan tujuan menentukan bisnis yang ingin dilakukan dengan mempertimbangkan formula stategi serta permasalahan data mining yang bisa terselesaikan.

3.2 Data Understanding Phase

Pada proses pemahaman data melakukan proses menentukan data yang diigin dianalisis bisa berasal dari database, scarping kemudian melakukan pemahaman data serta melakukan memeriksa dan evaluasi kualitas data.

3.3 Data Preparation Phase

Pada proses pemahaman data, dilakukannya menyiapkan data , kemudian memilih kasus dan variabel yang akan dianalisis dan menyiapakan data untuk dimodelkan dan pengolahan data yang meliputi proses data reduction, data transformation, uji missing value, uji multikolinearitas, validasi silhouette conffecient

3.5 Modelling Phase

Pada fase ini melakukan implementasi algoritam yang digunakan menggunakan data yang telah disiapkan untuk kebutuah analisis. Pada fase pemodelan menggunakan metode clustering.Metode clustering adalah proses mengidentifikasi pengelompokan atau cluster alami dalam data multidimensi berdasarkan beberapa ukuran kesamaan[14]

3.6 Evaluation Phase

Pada tahap evaluasi peneliti melakukan evaluasi pada model yang digunakan menentukan model yang digunakan memiliki akurasi serta menjawab tujuan bisnis.

3.7 Deployment Phase

Pada fase penyebaran, peneliti menggunakan model yang dihasilkan dan dipresentasikan.

4. HASIL PENELITIAN

3.1 Business Understanding

Sportify sebagai startup layanan streaming music yang memiliki banyak pengguna yang hampir menyentuh 130 juta, akan tetapi layanan ini memiliki banyak pesaing dan kemungkinan besar jumlah pengguna bisa menurun jika tidak meningkatkan layanan sehingga dengan tujuan bisnis bisa bersaing dengan kompetitor, sportify bisa melakukan sebuah analisis pasar berdasarkan genre musik dengan menggunakan data mining.

3.2 Data Understanding

Penelitian ini menggunakan data yang berisi kumpulan lagu pada Spotify dan diambil dari Kaggle. Data yang digunakan sebanyak 5999 data. Data tersebut terdiri dari beberapa atribut, antara lain: key yaitu kunci lagu, loudness yaitu kekerasan lagu, mode, speechiness yaitu tingkat speech lagu, acousticness yaitu tingkat akustik lagu, instrumentalness yaitu tingkat instrumen lagu, liveness yaitu tingkat hidup lagu, valence yaitu valensi lagu, tempo yaitu tempo lagu, type yaitu tipe lagu, id yaitu kode unik lagu, uri yaitu identifikator lagu, track_href yaitu identitas track lagu, analysis url yaitu url lagu, duration_ms yaitu lama lagu, time_signal, genre yaitu genre lagu, song_name yaitu judul lagu, unnamed, dan title.

3.3 Preprocessing

Preprocessing data dilakukan melalui dua tahap, yaitu data reduction dan data transformation

3.3.1 Data Reduction

Data reduction dilakukan dengan menghilangkan atribut type, uri, track_href, analysis_url, genre, song_name, unnamed:0, dan title. Sedangkan atribut yang digunakan pada proses selanjutnya ditunjukkan pada Tabel 1.

Tabel. 1 Hasil reduksi data

danceabilit	energy	key	loudness	mode	speechines	acousticnes	instrumentalness	liveness	valence	tempo	id	time_sig
0.831	0.814	2	-7.364	1	0.42	0.0598	0.0134	0.0556	0.389	156.985	FW8gD9q3	4
0.719	0.493	8	-7.23	1	0.0794	0.401	0	0.118	0.124	115.08	t5VmnJ7x0	4

3.1.2 Data Transformation

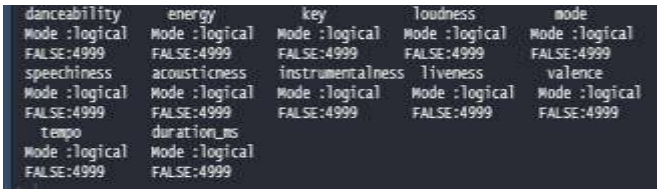
Data transformation dilakukan dengan men-generate id baru dan merubah posisi atribut id, menjadi posisi pertama. Ini dilakukan untuk memudahkan dalam proses analisis. Hasil transformasi data ditunjukkan pada Tabel 2.

Tabel. 2 Hasil transformasi data

id	danceability	energy	key	loudness	mode	speechines	acousticnes	instrumentalness	liveness	valence	tempo	duration_m	time_signat
1	0.831	0.814	2	-7.364	1	0.42	0.0598	0.0134	0.0556	0.389	156.985	124539	4
2	0.719	0.493	8	-7.23	1	0.0794	0.401	0	0.118	0.124	115.08	224427	4

3.2 Uji Missing Value

Setelah dilakukan uji missing value, tidak ada data missing pada semua atribut sebagaiman ditunjukkan pada Gambar 2.



Gambar 2. Hasil uji Missing Value

3.3 Uji Multikolinearitas

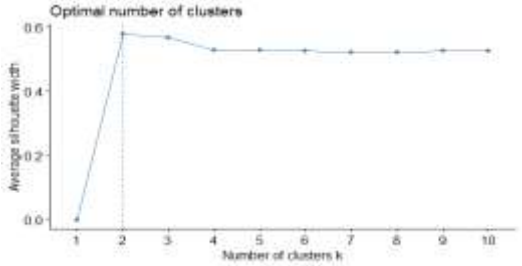
Multikolinearitas terjadi apabila nilai koefisien korelasi antar variabel dalam matriks korelasi bernilai dari 0.8

(P) ISSN 2442-451X
(O) ISSN 2503-3832

hingga 1.0 ($0.8 \leq\leq 1.0$). Hasil uji multikolinearitas ditunjukkan pada Tabel 3.

Tabel 3. Hasil uji multikolinearitas

Berdasarkan hasil uji multikolinearitas diketahui bahwa nilai koefisien terbesar adalah 0.722354 yaitu atribut korelasi antara loudness dan energy. Ini menandakan bahwa multikolinearitas tidak terjadi, sehingga data ini dapat digunakan pada proses selanjutnya.



3.4 Validasi Silhouette Coefficient

Validasi dilakukan untuk menentukkan nilai k yang tepat. Hasil validasi ditunjukkan pada pada Gambar 3.

Gambar 3. Validasi silhoutte coefficient

Pada gambar 2 ditunjukkan bahwa usulan cluster terbaik berdasarkan metode silhouette coefficient adalah 2. Tetapi untuk membuat cluster lebih valid, penelitian akan menggunakan 3, 4, 5, 6, 7, 8, 9, dan 10 cluster sebagai perbandingan pada masing-masing metode, dimana grafik cenderung sejajar. Ini dimungkinkan tingkat kevalidannya tidak jauh berbeda.

3.5 Clustering

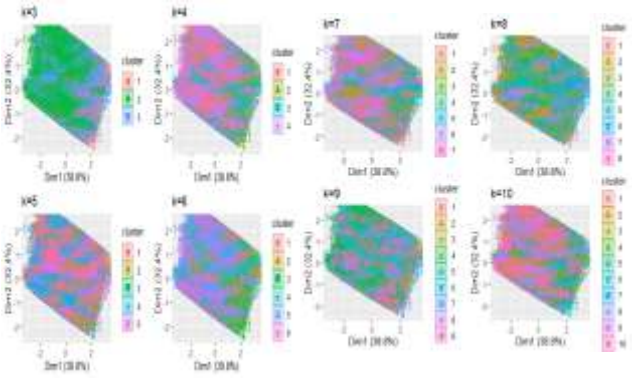
Clustering dilakukan dengan menggunakan 3 metode, k-Means, k-Medoid, dan agglomerative clustering.

3.5.1 k-Means

Dilakukan 8 kali clustering, dengan jumlah klaster 3, 4, 5, 6, 7, 8, 9, dan 10. Jumlah anggota klaster pada setiap proses klaster ditunjukkan pada Tabel 4. Sedangkan sebaran klaster menggunakan metode k-Means ditunjukkan pada Gambar 4

Tabel 4. Jumlah anggota klaster

k	Cluster ke-									
	1	2	3	4	5	6	7	8	9	10
3	865	2293	2041							
4	1852	452	1365	1530						
5	1498	767	190	1122	1422					
6	77	956	456	1201	1356	953				
7	403	571	718	73	1029	1146	1059			
8	296	905	541	990	837	913	457	60		
9	647	30	365	800	912	890	159	405	791	
10	707	296	30	497	706	150	321	734	797	752



Gambar 4. Sebaran cluster k-Means

3.5.2 k-Medoid

Dilakukan 8 kali clustering, dengan jumlah klaster 3, 4, 5, 6, 7, 8, 9, dan 10. Jumlah anggota klaster pada setiap proses klaster ditunjukkan pada Tabel 5. Sedangkan sebaran klaster menggunakan metode k-Medoid ditunjukkan pada Gambar 9

Tabel 5. Jumlah anggota klaster k-Medoid

k	Cluster ke-									
	1	2	3	4	5	6	7	8	9	10
3	2142	1831	1026							
4	1459	1404	1403	731						
5	1119	964	1253	1125	538					
6	851	712	1093	929	954	461				
7	817	837	557	931	862	619	376			
8	793	609	510	911	819	772	387	198		
9	718	683	480	746	656	692	512	330	176	
10	708	472	476	717	609	579	584	311	375	166

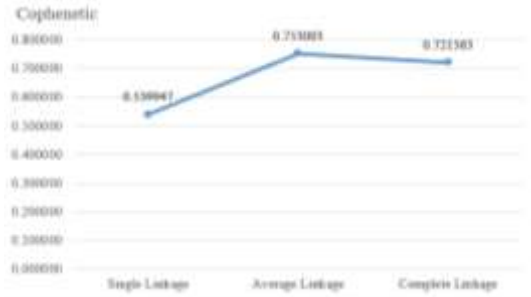
Gambar 5. Sebaran cluster k-Medoid

3.5.3 Agglomerative Clustering

Sebelum dilakukan proses clustering, terlebih dahulu dilakukan proses penentuan metode agglomerative hierarchical clustering yang tepat. Ini dilakukan dengan menghitung koefisien korelasi pada setiap metode.

1.Pengujian Koefisien Korelasi cophenetic

Berdasarkan hasil pengujian koefisien korelasi Cophenetic, metode agglomerative clustering tertinggi adalah Average Linked sebagaimana yang terdapat pada Gambar 5. Oleh karena itu clustering juga akan dilakukan menggunakan metode Average Linked



Gambar 6. Hasil pengujian Koefisien Korelasi *Cophenetic*

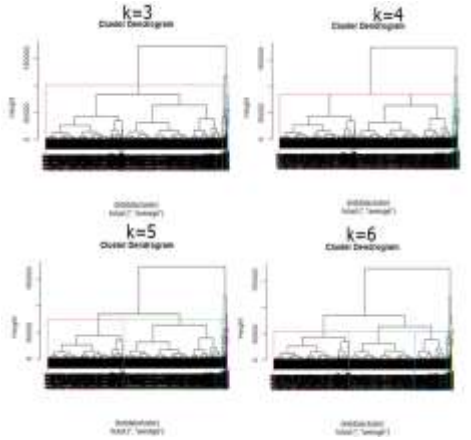
(P) ISSN 2442-451X
(O) ISSN 2503-3832

2. Agglomerative Clustering

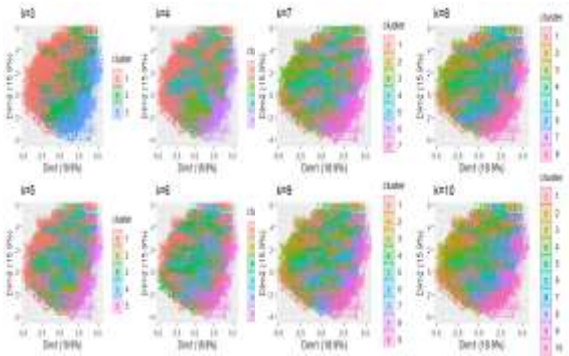
Dilakukan 8 kali clustering, dengan jumlah klaster 3, 4, 5, 6, 7, 8, 9, dan 10. Jumlah anggota klaster pada setiap proses klaster ditunjukkan pada Tabel 6. Sedangkan sebaran klaster ditunjukkan pada Gambar 6 dan 7.

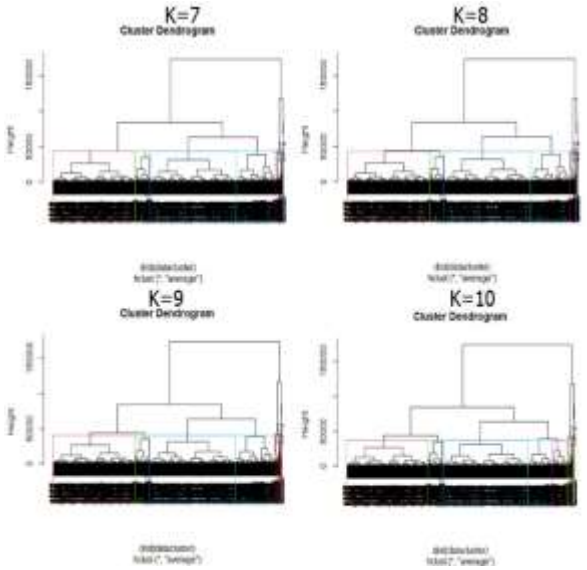
Tabel 6. Jumlah anggota klaster tiap hirarki

k	Cluster ke-									
	1	2	3	4	5	6	7	8	9	10
3	4833	145	21							
4	4833	145	20	1						
5	2115	2718	145	20	1					
6	2115	1862	856	145	20	1				
7	1784	1862	331	856	145	20	1			
8	1784	1862	331	856	145	9	11	1		
9	1784	1862	331	856	48	97	9	11	9	
10	1784	1862	331	538	318	48	97	9	11	1



Gambar 7. hasil clustering Average Linkage dengan jumlah cluster 3,4,5, dan 6





Gambar 8. hasil clustering Average Linkage dengan jumlah cluster 7,8,9, dan 10

3.6 Validasi Dunn Index

Indeks dunn digunakan untuk memvalidasi hasil cluster pada penelitian ini.Semakin besar nilai indeks, semakin valid cluster tersebut. Berdasarkan hasil validasi, nilai indeks terbaik menggunakan agglomerative hierarchical clustering metode Average Linked sebesar 0,00498995 dengan 3 atau 4 cluster jumlah. Hasil validasi ditunjukkan pada Tabel 7.

Tabel 7. Hasil Validasi Index Dunn

Metode	k							
	3	4	5	6	7	8	9	10
K-Means	0.00018007	0.00018995	0.00024458	0.00030568	0.00054814	0.00033074	0.00038810	0.00052067
K-Medoid	0.00011766	0.00020800	0.00008604	0.00016299	0.00007647	0.00026448	0.00026881	0.00017353
Hierarchical Clustering	0.00498995	0.00498995	0.00013180	0.00007457	0.00000000	0.00000000	0.00000000	0.00000000

3.7 Deployment

Berdasarkan hasil penelitian metode cluster terbaik yang digunakan untuk clustering lagu spotify adalah agglomerative hierarchical clustering metode Average Linked dengan 3 atau 4 klaster. Klaster paling direkomendasikan adalah sebesar 3 klaster karena klaster ke 4 hanya berisi 1 anggota saja. Pada pengelompokkan 3 klaster, klaster ke-1 berisi 2833 anggota, klaster ke-2 berisi 145 anggota, dan klaster ke-3 berisi 21 anggota.

5. KESIMPULAN

Penelitian komparasi 3 algoritma data mining yaitu kmeans,kmedoid,agglomerative hierarchical menghasilkan kesimpulan bahwa metode algoritma yang paling baik untuk proses pengelompokan genre adalah agglomerative hierarchical dengan metode average linked dengan klaster terbaik 3 atau 4 Klaster paling direkomendasikan adalah sebesar 3 klaster karena klaster ke 4 hanya berisi 1 anggota saja. Pada pengelompokkan 3 klaster, klaster ke-1 berisi 2833 anggota, klaster ke-2 berisi 145 anggota, dan klaster ke-3 berisi 21 anggota.

(P) ISSN 2442-451X
(O) ISSN 2503-3832

Daftar Pustaka

[1] M. M. A. Amirah, A. W. Widodo, and C. Dewi, 'Pengelompokan Lagu Berdasarkan Emosi Menggunakan Algoritma Fuzzy C-Means', *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN*, vol. 2548, p. 964X, 2017.

[2] K. C. Media, 'Komentar Berita : Jumlah Pelanggan Spotify Tembus 130 Juta di Tengah Pandemi Covid-19', *KOMPAS.com*. https://tekno.kompas.com/komentar/2020/05/02/14020097/jumlah-pelanggan-spotify-tembus-130-juta-di-tengah-pandemi-covid-19 (accessed Jan. 26, 2021).

[3] S. Y. M. Netti and I. Irwansyah, 'Spotify: Aplikasi Music Streaming untuk Generasi Milenial', *Jurnal Komunikasi*, vol. 10, no. 1, pp. 1–16, 2018.

[4] A. HUSNA, 'SEGMENTASI PELANGGAN MENGGUNAKAN MODEL RFM DAN TEORI ROUGH SET UNTUK MEMAHAMI KARAKTERISTIK PELANGGAN (STUDI KASUS: PT. ABBOTT INDONESIA, Tbk CABANG MALANG)', *2015*.

[5] S. Sindi, W. R. O. Ningse, I. A. Sihombing, F. I. R. Zer, and D. Hartama, 'Analisis algoritma k-medoids clustering dalam pengelompokan penyebaran covid-19 di indonesia', *JurTI (Jurnal Teknologi Informasi)*, vol. 4, no. 1, pp. 166–173, 2020.

[6] Z. Arifin, S. Santosa, and M. A. Soeleman, 'Klasterisasi Genre Cerpen Kompas Menggunakan Agglomerative Hierarchical Clustering-Single Linkage', *Jurnal Cyberku*, vol. 13, no. 2, pp. 2–2, 2017.

[7] W. M. P. Dhuhita, 'Clustering Menggunakan Metode K-mean Untuk Menentukan Status Gizi Balita', *Jurnal Informatika Darmajaya*, vol. 15, no. 2, pp. 160–174, 2015.

[8] G. Gustientiedina, M. H. Adiya, and Y. Desnelita, 'Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan', *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 5, no. 1, pp. 17–24, 2019.

[9] B. Wira, A. E. Budianto, and A. S. Wiguna, 'Implementasi Metode K-Medoids Clustering Untuk Mengetahui Pola Pemilihan Program Studi Mahasiswa Baru Tahun 2018 Di Universitas Kanjuruhan Malang', *Rainstek: Jurnal Terapan Sains & Teknologi*, vol. 1, no. 3, pp. 53–68, 2019.

[10] F. Hardiyanti, H. S. Tambunan, and I. S. Saragih, 'PENERAPAN METODE K-MEDOIDS CLUSTERING PADA PENANGANAN KASUS DIARE DI INDONESIA', *KOMIK (Konferensi Nasional Teknologi Informasi dan Komputer)*, vol. 3, no. 1, 2019.

[11] A. T. R. Dani, S. Wahyuningsih, and N. A. Rizki, 'Penerapan Hierarchical Clustering Metode Agglomerative pada Data Runtun Waktu', *Jambura Journal of Mathematics*, vol. 1, no. 2, pp. 64–78, 2019.

[12] M. Reddy, V. Makara, and R. Satish, 'Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering', *Int J of Comp Science Trands and Tech (IJCST)*, vol. 5, no. 5, pp. 5–11, 2017.

[13] I. Purnama, R. Saputra, and A. Wibowo, 'Implementasi Data Mining Menggunakan Crisp-Dm Pada Sistem Informasi Eksekutif Dinas Kelautan Dan Perikanan Provinsi Jawa Tengah', 2014.

[14] T. S. Madhulatha, 'An overview on clustering methods', *arXiv preprint arXiv:1205.1117*, 2012.