
Implementasi Algoritma *Hierarchical Clustering* dan *Non-Hierarchical Clustering* untuk Pengelompokan Pengguna Media Sosial

Zulkifli Alamtaha^{1*}, Ismail Djakaria², Nisky Imansyah Yahya³
^{1,2,3}Jurusan Matematika, Fakultas MIPA,
Universitas Negeri Gorontalo, Bone Bolango, 96554, Indonesia
*Corresponding author, email: zulalamtaha27@gmail.com

Abstract

Social media is a means to interact with other people through sentences, pictures and videos online. Excessive use of social media has a negative impact on mental health. The grouping process in this study was carried out to see the level of social media use in Bone Bolango Regency. Before grouping, data pre-processing is carried out and the optimal number of clusters is determined using the Silhouette index. The optimal cluster results obtained are two clusters for all methods. After that, grouping is done using Hierarchical Clustering and Non-Hierarchical Clustering Algorithms. The Hierarchical Clustering algorithm consists of two methods, namely the single linkage method and the complete linkage method. The Non-Hierarchical Clustering Algorithm consists of two methods, namely the K-Means and K-Medoids methods. The next step is to determine the best method using the Davies-Bouldin Index (DBI). The smaller the DBI value, the better the method used. The smallest DBI value is obtained in the complete linkage method. The grouping results for cluster 1 consisted of 70 respondents and cluster 2 consisted of 80 respondents.

Keywords: Davies-Bouldin Index, Hierarchical Clustering, Non-Hierarchical Clustering, Silhouette Index, Social Media

Abstrak

Media sosial merupakan sarana untuk berinteraksi dengan orang lain melalui kalimat, gambar dan video secara *online*. Penggunaan media sosial yang berlebihan berdampak negatif pada kesehatan mental. Proses pengelompokan pada penelitian ini dilakukan untuk melihat tingkat penggunaan media sosial di Kabupaten Bone Bolango. Sebelum melakukan pengelompokan, terlebih dahulu dilakukan *pre-processing* data dan penentuan jumlah *cluster* optimal menggunakan indeks *silhouette*. Hasil *cluster* optimal yang diperoleh adalah dua *cluster* untuk semua metode. Setelah itu dilakukan pengelompokan menggunakan Algoritma *Hierarchical Clustering* dan *Non-Hierarchical Clustering*. Algoritma *Hierarchical Clustering* terdiri dari dua metode yaitu metode *single linkage* dan *complete linkage*. Algoritma *Non-Hierarchical Clustering* terdiri dari dua metode yaitu metode K-Means dan K-Medoids. Tahap berikutnya adalah penentuan metode terbaik menggunakan *Davies-Bouldin Index (DBI)*. Semakin kecil nilai *DBI*, maka semakin baik metode yang digunakan. Nilai *DBI* terkecil diperoleh pada metode *complete linkage*. Hasil pengelompokan untuk *cluster* 1 terdiri dari 70 responden dan *cluster* 2 terdiri dari 80 responden.

Kata Kunci: Davies-Bouldin Index, Hierarchical Clustering, Media Sosial, Non-Hierarchical Clustering, Silhouette Index.

1. Pendahuluan

Pembelajaran mesin (*machine learning*) adalah riset ilmu tentang kombinasi konsep dari sejumlah bidang keilmuan, seperti kecerdasan buatan, statistika, matematika dan teori informasi. Tujuan spesifik dari penelitian *machine learning* yaitu untuk pengembangan algoritma pembelajaran yang cepat dan akurat dalam menghasilkan deskripsi dan prediksi tentang data. Dalam hal analisis data, *machine learning* adalah pendekatan yang digunakan untuk menghasilkan model deskriptif dan prediktif [1]. Salah satu jenis *machine learning* adalah *unsupervised learning*.

Algoritma *unsupervised learning* merupakan algoritma yang mempunyai kegunaan dalam mengelompokkan dan mengklasifikasikan data. Pada algoritma ini tidak dilakukan proses *training* pada *dataset*, karena algoritma ini bersifat deskriptif, bukan prediktif, maka dari itu hanya membutuhkan pembelajaran dari data itu sendiri. Dalam algoritma *unsupervised learning* tersedia beberapa *input* sampel namun ada juga beberapa *output* yang tidak diinginkan. Karena terdapat beberapa *output* yang tidak diinginkan, klasifikasi dilakukan supaya algoritma dapat membedakan dengan benar data yang tersedia. Tugas tersebut merupakan fungsi dalam penggambaran struktur yang ada pada data, terutama pada data yang tidak memiliki label [2]. Salah satu analisis yang sering digunakan dalam *unsupervised learning* adalah analisis *cluster*.

Analisis *cluster* adalah model multivariat yang dibuat dengan tujuan mengklasifikasikan data menurut karakteristik data tersebut. Ada dua metode dalam analisis *cluster*, yaitu metode hierarki dan non-hierarki. Metode hierarki adalah metode yang dirancang untuk dekomposisi berhierarki (tingkatan) dari kumpulan data menurut karakteristik pada data tersebut. Sedangkan metode non-hierarki adalah metode yang digunakan untuk mengelompokkan data dan jumlah *cluster* yang akan dibuat dapat ditetapkan sebelumnya. Ada berbagai jenis analisis *cluster* yang digunakan dalam metode hierarki, yaitu *single linkage* dan *complete linkage*. Sedangkan jenis analisis *cluster* yang digunakan dalam metode non-hierarki, yaitu metode K-Means dan K-Medoids [3]. Studi kasus yang diangkat dalam penelitian ini adalah tentang penggunaan media sosial.

Media sosial sudah menjadi hal yang cukup penting di era sekarang. Kebutuhan tentang sebuah informasi yang cepat dan tepat untuk kepentingan maupun kepuasan pribadi yang terkadang melupakan satu hal yaitu durasi yang mereka gunakan untuk mengakses media sosial tersebut. Kehadiran media sosial tentu membawa dampak positif maupun dampak negatif. Dampak positif yang diperoleh salah satunya adalah pertukaran informasi yang cepat dan tepat. Sedangkan dampak negatif yang diperoleh adalah tentang durasi penggunaan media sosial yang berlebihan. Dampak negatif yang dimaksud yaitu gangguan tidur, tingkat konsentrasi menurun, merusak mata, gangguan kecemasan dan kecanduan media sosial [4].

Seorang psikoterapis dari California School of Professional Psychology, Philip Cushman, menyampaikan hasil penelitiannya bahwa kita harus membatasi penggunaan

media sosial dari setengah jam sampai satu jam atau tidak lebih dari dua jam per hari. Hal ini mirip dengan penelitian yang dilakukan di University of Pittsburgh. Studi ini memverifikasi bahwa Orang yang terlalu aktif di media sosial setiap hari memiliki risiko depresi hingga tiga kali lebih besar dibandingkan mereka yang jarang menggunakan media sosial saat ini. Berdasarkan penelitian ini, terlalu sering menggunakan media sosial dapat berdampak negatif pada kesehatan mental, dan orang yang terlalu aktif di media sosial setiap hari memiliki risiko depresi tiga kali lebih besar jika dibandingkan dengan orang yang jarang menggunakannya [5].

Penelitian ini dilakukan untuk melihat kelompok yang penggunaan media sosialnya masih dalam batas wajar dan kelompok yang penggunaan media sosialnya sudah berlebihan. Dari hal tersebut, penulis dapat mengetahui karakteristik dari masing-masing kelompok yang diperoleh. Dari hasil pengelompokan yang diperoleh juga dapat menjadi cerminan dan peringatan bagi penulis maupun pembaca agar tidak berlebihan dalam menggunakan media sosial di era sekarang.

2. Material dan Metode

2.1 Data

Data yang digunakan pada penelitian ini merupakan data primer tentang durasi penggunaan setiap pengguna media sosial. Sampel yang digunakan yaitu 150 pengguna media sosial di Kabupaten Bone Bolango. Variabel yang digunakan sebanyak 8 Variabel. Variabel-variabel yang akan digunakan dalam penelitian ini sebagai berikut:

1. Usia (X_1)
2. Durasi Penggunaan WhatsApp (X_2)
3. Durasi Penggunaan Telegram (X_3)
4. Durasi Penggunaan YouTube (X_4)
5. Durasi Penggunaan Facebook (X_5)
6. Durasi Penggunaan Instagram (X_6)
7. Durasi Penggunaan TikTok (X_7)
8. Durasi Penggunaan Twitter (X_8)

2.2 Pre-Processing Data

Pre-processing data adalah proses mengubah data ke dalam format yang sederhana, efektif dan sesuai dengan kebutuhan peneliti. Indikator yang dapat digunakan sebagai referensi adalah kelengkapan data, hasil yang akurat dan waktu komputasi yang singkat [6]. Terdapat beberapa bentuk transformasi data yang dapat digunakan dalam analisis *cluster*. Diantaranya adalah dengan mengubah bentuk data ke dalam skala nominal atau ordinal menyesuaikan dengan data yang digunakan serta pembobotan data yang jelas [7].

2.3 Indeks *Silhouette*

Indeks *silhouette* adalah nilai yang digunakan untuk menentukan nilai k yang paling optimal. Dengan *cluster* disebut terbentuk dengan baik jika nilai indeks mendekati 1. Sedangkan *cluster* disebut terbentuk kurang baik jika nilai indeks mendekati -1 [8].

$$s(i) = \frac{a(i) - b(i)}{\max(a(i), b(i))} \quad (1)$$

2.4 Uji Asumsi Multikolinearitas

Uji asumsi multikolinearitas memiliki tujuan untuk menguji apakah data mempunyai korelasi yang tinggi dari variabel yang diikutsertakan atau tidak. Variabel analisis *cluster* tersebut disebut baik jika antar variabel bebas tidak terjadi multikolinearitas. Salah satu cara dalam menganalisis terjadinya multikolinearitas bisa dianalisis dari nilai *Variance Inflation Factor (VIF)* [9].

$$VIF = \frac{1}{Tolerance} = \frac{1}{1 - R_p^2} \quad (2)$$

2.5 *Euclidean Distance*

Jarak antar objek ke- i dan objek ke- k disimbolkan dengan d_{ik} pada variabel ke- p . Perhitungan jarak *euclid* dari dua objek sebanyak p variabel menggunakan persamaan sebagai berikut [9].

$$d_{ik} = \sqrt{\sum_{p=1}^c |x_{ip} - x_{kp}|} \quad (3)$$

2.6 *Hierarchical Clustering*

Single linkage adalah jarak antara dua *cluster* yang diukur menurut jarak minimum objek dalam suatu *cluster* dan objek pada *cluster* lainnya [10].

$$d(uv)w = \min (d_{uw}d_{vw}) \quad (4)$$

Complete linkage adalah jarak antara dua *cluster* yang diukur menurut jarak maksimum objek dalam suatu *cluster* dan objek pada *cluster* lainnya.

$$d(uv)w = \max (d_{uw}d_{vw}) \quad (5)$$

2.7 *Non-Hierarchical Clustering*

K-Means adalah sebuah algoritma yang memiliki *centroid* terdekat di setiap *cluster*. Dalam menentukan *centroid* awal ditentukan secara acak dan juga menentukan partisi objek sejumlah k *cluster*, selanjutnya dihitung *centroid* ke- i variabel ke- p .

$$v(ip) = \frac{\sum_{i=1}^n x_{ip}}{n_{pj}} \quad (6)$$

K-Medoids atau nama lainnya *Partitioning Around Medoids (PAM)* adalah jenis lain dari metode non-hierarki selain K-Means. Perhitungan metode K-Medoids tidak berpatokan pada nilai rata-rata (mean) semua data dari setiap *cluster* melainkan berpatokan pada titik data secara acak, dan mempunyai tujuan untuk mengurangi *outlier* atau sensitivitas dari *cluster* yang dihasilkan dalam data. Algoritma K-Means maupun K-Medoids membuat *cluster* dari jarak minimal antara setiap titik data berada dalam satu *cluster* dan titik data pada pusat *cluster*. Dalam pemilihan titik data sebagai pusat (medoid) pada algoritma K-Medoids dilakukan secara acak pada titik data dalam masing-masing *cluster*, bukan berdasarkan nilai rata-rata (*mean*) seperti K-Means [11].

2.8 Davies-Bouldin Index

Davies-Bouldin Index atau *DBI* memaksimalkan jarak inter *cluster* dan meminimalkan jarak intra *cluster*. Semakin kecil nilai *DBI*, maka semakin baik metode yang digunakan. Rumus untuk menghitung nilai *DBI* terdapat pada persamaan berikut [12].

$$DBI = \frac{1}{K} \sum_{i=1}^k \max_{i \neq j} (R_{ij}) \quad (7)$$

3. Hasil dan Diskusi

3.1 Pre-Processing Data

Data awal memuat data untuk X_1 yaitu usia dalam satuan tahun, dan data X_2 - X_8 yaitu durasi penggunaan media sosial dalam satuan menit. Jumlah data yang digunakan dalam penelitian ini sebanyak 150 data. Berikut ini 10 data awal dari 150 data yang digunakan.

Tabel 1. Data Awal

No.	Nama	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1.	Toni	47	49	17	181	193	8	3	0
2.	Yati	39	26	25	93	99	26	78	0
3.	Agus	32	69	53	79	30	80	186	40
4.	Kiwi	42	75	18	73	62	27	74	0
5.	Roni	44	64	14	148	176	15	87	0
6.	Weti	41	62	19	183	139	20	81	21
7.	Maman	38	189	29	44	34	86	186	0
8.	Santi	37	97	6	28	33	41	195	0
9.	Dedi	33	47	27	133	146	28	66	0
10.	Debi	27	155	75	88	56	90	199	10

Sumber: Data olah, 2022

Tabel 1 merupakan tabel yang memuat data awal sebelum ditransformasikan. Selanjutnya data pada Tabel 1 tersebut ditransformasikan menurut kriteria pembobotan. Untuk Variabel X_1 terbagi atas empat kategori sebagai berikut.

1. Usia 17-25 Tahun = 1 (Masa Remaja Akhir)
2. Usia 26-35 Tahun = 2 (Masa Dewasa Awal)
3. Usia 36-45 Tahun = 3 (Masa Dewasa Akhir)
4. Usia 46-55 Tahun = 4 (Masa Lansia Awal)

Untuk Variabel X_2 - X_8 terbagi atas empat kategori sebagai berikut.

1. Durasi < 1 Jam = 1 (Singkat)
2. Durasi 1-2 Jam = 2 (Sedang)
3. Durasi 3-4 Jam = 3 (Lama)
4. Durasi 5-6 Jam = 4 (Sangat Lama)

Jumlah data yang digunakan dalam penelitian ini sebanyak 150 data. Berikut ini 10 data transformasi dari 150 data yang digunakan.

Tabel 2. Data Transformasi

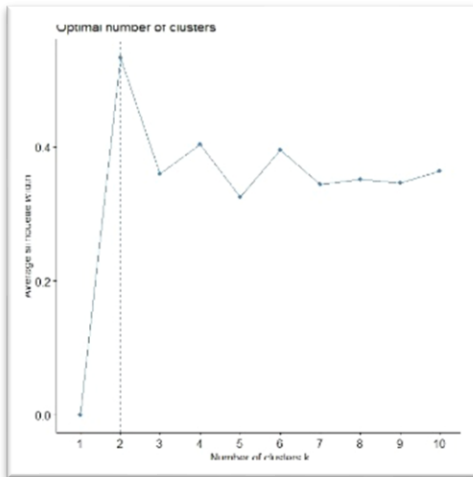
No.	Nama	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1.	Toni	4	1	1	3	3	1	1	1
2.	Yati	3	1	1	2	2	1	2	1
3.	Agus	2	2	1	2	1	2	3	1
4.	Kiwi	3	2	1	2	2	1	2	1
5.	Roni	3	2	1	2	2	1	2	1
6.	Weti	3	2	1	3	2	1	2	1
7.	Maman	3	3	1	1	1	2	3	1
8.	Santi	3	2	1	1	1	1	3	1
9.	Dedi	2	1	1	2	2	1	2	1
10.	Debi	2	2	2	2	1	2	3	1

Sumber: Data olah, 2022

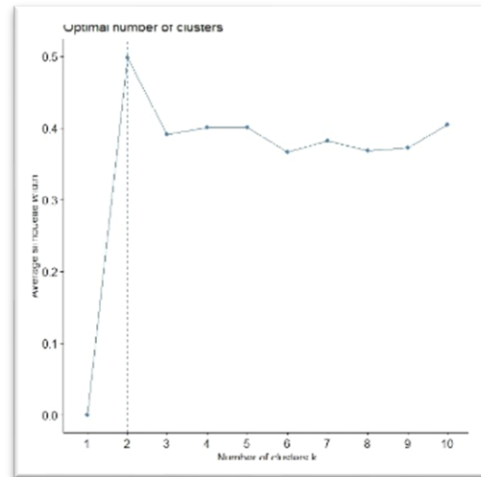
Tabel 2 merupakan tabel yang memuat data yang sudah ditransformasikan menurut kriteria pembobotannya. Setelah dilakukan transformasi data, tahap selanjutnya adalah penentuan jumlah k cluster optimal dan uji asumsi multikolinearitas.

3.2 Penentuan Jumlah K Cluster Optimal

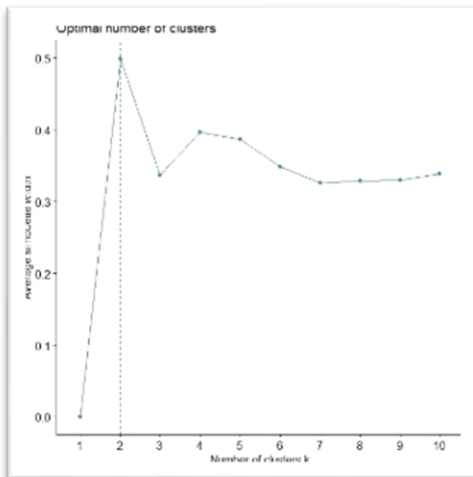
Menentukan jumlah k cluster optimal pada penelitian ini menggunakan grafik indeks *silhouette*. Berikut grafik indeks *silhouette* untuk metode *single linkage*, *complete linkage*, K-Means dan K-Medoids.



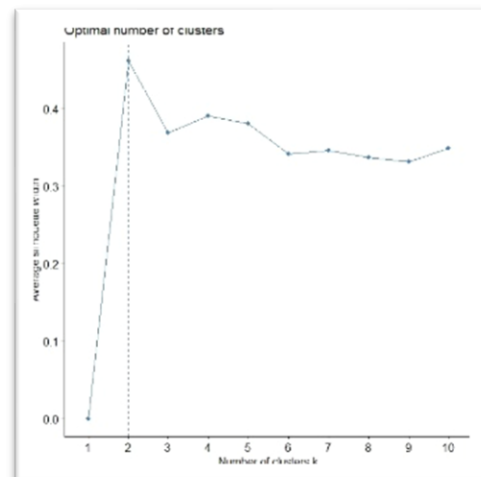
Gambar 1. Single Linkage



Gambar 2. Complete Linkage



Gambar 3. K-Means



Gambar 4. K-Medoids

Gambar 1-4 merupakan gambar yang memuat grafik indeks *silhouette* dari masing-masing metode yang digunakan. Dari grafik indeks *silhouette* di atas, menunjukkan jumlah *cluster* optimal untuk empat metode yang digunakan sebanyak dua *cluster*.

3.3 Uji Asumsi Multikolinearitas

Uji asumsi multikolinearitas digunakan untuk menguji apakah data mempunyai korelasi yang tinggi dari variabel yang diikutsertakan atau tidak.

Tabel 3. Nilai *VIF*

Variabel	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈
X ₁	1							
X ₂	2,2589	1						
X ₃	1,1980	1,1323	1					
X ₄	1,0549	1,0305	1,0613	1				
X ₅	2,3262	1,5496	1,1359	1,1777	1			
X ₆	2,4497	2,3713	1,1625	1,0353	1,5495	1		
X ₇	4,0407	2,1219	1,2689	1,1402	2,7631	1,9521	1	
X ₈	1,2770	1,2374	1,0346	1,0014	1,0562	1,2760	1,1518	1

Sumber: Data olah, 2022

Dari nilai *VIF* pada tabel di atas tidak terdapat nilai *VIF* > 10. Maka dapat disimpulkan, data yang digunakan tidak terjadi multikolinearitas.

3.4 Hasil Cluster

Setelah dilakukan proses pembentukan *cluster*, diperoleh hasil *cluster* untuk masing-masing metode sebagai berikut.

Tabel 4. Hasil *Cluster*

Metode	Cluster 1	Cluster 2
Single Linkage	132 Sampel	18 Sampel
Complete Linkage	70 Sampel	80 Sampel
K-Means	71 Sampel	79 Sampel
K-Medoids	89 Sampel	61 Sampel

Sumber: Data olah, 2022

Setelah diperoleh hasil *cluster* dari masing-masing metode, langkah selanjutnya adalah menentukan metode mana yang terbaik dari empat metode yang digunakan.

3.5 Penentuan Metode Terbaik

Menentukan metode terbaik menggunakan nilai *Davies-Bouldin Indeks (DBI)*. Semakin kecil nilai *DBI*, maka semakin baik metode yang digunakan.

Tabel 5. Nilai *DBI*

Single Linkage	Complete Linkage	K-Means	K-Medoids
0,75469	0,42595	0,43057	0,54621

Sumber: Data olah, 2022

Dari nilai *DBI* pada tabel di atas, nilai *DBI* terkecil ada pada metode *complete linkage*. Maka dapat disimpulkan metode *complete linkage* adalah metode terbaik.

3.6 Interpretasi Cluster

Dari empat metode yang telah di uji, nilai DBI terkecil ada pada metode *complete linkage*, sehingga metode *complete linkage* dipilih sebagai metode terbaik dalam penelitian ini. Hasil *clustering* dari metode *complete linkage* akan diambil sebagai dasar pengambilan kesimpulan terhadap hasil *cluster* yang terbentuk. Jumlah optimal *cluster* yang terbentuk adalah dua *cluster*. Sehingga dapat di interpretasikan sebagai berikut.

1. Berikut ini adalah rata-rata usia untuk X_1 (tahun) dan durasi penggunaan media sosial X_2 - X_8 (menit) untuk setiap *cluster*.

Tabel 6. Rata-rata Tiap Cluster

Cluster	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	44,86	48,27	16,77	131,57	163,99	15,37	58,46	34,36
2	25,48	197,28	45,91	64,50	48,18	61,35	229,99	37,24

Sumber: Data olah, 2022

2. Pada metode *complete linkage* terbentuk dua metode, yaitu:
Cluster 1 (Cluster Old) terdiri dari 70 subjek adalah *cluster* yang durasi penggunaan media sosialnya masih dalam batas wajar. Media sosial yang paling sering digunakan oleh *Cluster Old* adalah Facebook dan Youtube. *Cluster Old* didominasi oleh kelompok umur dewasa.
Cluster 2 (Cluster Milenial) terdiri dari 80 subjek adalah *cluster* yang durasi penggunaan media sosialnya sudah berlebihan. Media sosial yang paling sering digunakan oleh *Cluster Milenial* adalah TikTok dan WhatsApp. *Cluster Milenial* didominasi oleh kelompok umur remaja.
3. Pembagian 70 subjek pada *Cluster Old* dan 80 subjek pada *Cluster Milenial* berdasarkan persamaan dari beberapa subjek pengamatan yang terdiri dari Usia dan Durasi Penggunaan Media Sosial seperti WhatsApp, Telegram, YouTube, Facebook, Instagram, TikTok dan Twitter.

4. Kesimpulan

Berdasarkan hasil penelitian dapat disimpulkan bahwa:

1. Penentuan jumlah *cluster* optimal menggunakan algoritma *hierarchical clustering* dan *non-hierarchical clustering* untuk pengelompokkan pengguna media sosial adalah dua *cluster*. Dari empat metode yang digunakan yaitu metode *single linkage*, *complete linkage*, K-Means dan K-Medoids, semuanya menunjukkan jumlah *cluster* optimal yang didapatkan adalah dua *cluster*. Penentuan jumlah *cluster* optimal menggunakan grafik *silhouette index*.
2. Penentuan metode terbaik yang akan digunakan untuk pengelompokkan pengguna media sosial menggunakan uji validasi *cluster Davies-Bouldin Index (DBI)*. Semakin kecil nilai DBI yang diperoleh, maka semakin baik metode yang digunakan. Dari empat metode yang digunakan, nilai DBI terkecil yang diperoleh

adalah nilai DBI metode *complete linkage* yaitu sebesar 0,42595 untuk *cluster* optimal sejumlah dua *cluster*.

3. Hasil pengelompokkan pengguna media sosial menggunakan metode *complete linkage* terbentuk jumlah *cluster* optimal sejumlah dua *cluster*. Adapun hasil pengelompokkan pada *Cluster 1 (Cluster Old)* terdiri dari 70 subjek adalah *cluster* yang durasi penggunaan media sosialnya masih dalam batas wajar. Media sosial yang paling sering digunakan oleh *Cluster Old* adalah Facebook dan Youtube. *Cluster Old* didominasi oleh kelompok umur dewasa. Sedangkan *Cluster 2 (Cluster Milenial)* terdiri dari 80 subjek adalah *cluster* yang durasi penggunaan media sosialnya sudah berlebihan. Media sosial yang paling sering digunakan oleh *Cluster Milenial* adalah TikTok dan WhatsApp. *Cluster Milenial* didominasi oleh kelompok umur remaja.

Daftar Pustaka

- [1] Wiranda, N., Purba, H. S., dan Sukmawati, R. A. Survei penggunaan tensorflow pada machine learning untuk identifikasi ikan kawasan lahan basah. *Indonesian Journal of Electronics and Instrumentation Systems*, 10:179. doi:10.22146/ijeis.58315. 10 2020.
- [2] Roihan, A., Sunarya, P. A., dan Rafika, A. S. Pemanfaatan machine learning dalam berbagai bidang: Review paper. *Indonesian Journal on Computer and Information Technology*, 5:75–82. 2019.
- [3] Widyadhana, D., Hastuti, R. B., Kharisudin, I., dan Fauzi, F. Perbandingan analisis klaster k-means dan average linkage untuk pengklasteran kemiskinan di provinsi jawa tengah. *Prosiding Seminar Nasional Matematika*, 4:584–594. 2021.
- [4] Rahmawati, H. N., Iqom, M. K. B., dan Hermanto. Hubungan durasi penggunaan media sosial dengan motivasi belajar remaja. *Jurnal Keperawatan Jiwa*, 5:77–81. 2017.
- [5] Sosiady, M., Djamil, N., dan Ermansyah, E. The effects of internet addiction disorder on students' learning motivation. *Al-Ishlah: Jurnal Pendidikan*, 14:3449–3460. doi:10.35445/alishlah.v14i3.1712. 7 2022.
- [6] Saifullah, Zarlis, M., dan Sembiring, R. W. Analisa terhadap perbandingan algoritma decision tree dengan algoritma random tree untuk pre-processing data. *Jurnal Sains Komputer dan Informatika (J-SAKTI)*. 2017.
- [7] Aziz, A. R., Warsito, B., dan Prahutama, A. Pengaruh transformasi data pada metode learning vector quantization terhadap akurasi klasifikasi diagnosis penyakit jantung. 10. 2021.
- [8] N. N. Halim dan E. Widodo. Clustering dampak gempa bumi di Indonesia menggunakan kohonen self organizing maps. *Seminar Nasional Integrasi Matematika dan Nilai-nilai Islami*, 1:14-16. 2017.

- [9] Khoirunnisa, H., Ruchjana, B. N., Irianingsih, I., dan Suhandi, B. Perbandingan penerapan metode agglomerative dengan metode kmeans pada data curah hujan di wilayah bogor. 2020.
- [10] C. Suhaeni, A. Kurnia, dan R. Ristiyanti. Perbandingan hasil pengelompokkan menggunakan analisis cluster berhirarki, kmeans cluster, dan cluster ensemble (studi kasus data indikator pelayanan kesehatan ibu hamil), *Jurnal Media Infotama*. 14:1. 2018.
- [11] Supriyadi, A., Triayudi, A., dan Sholihati, I. D. Perbandingan algoritma k-means dengan k-medoids pada pengelompokan armada kendaraan truk berdasarkan produktivitas. 6:229–240. 2021.
- [12] B. J. D. Sitompul, O. S. Sitompul, dan P. Sihombing. Enhancement clustering evaluation result of davies-bouldin index with determining initial centroid of k-means algorithm. *Journal of Physics*. 1235:1. 2019.