Analisis *Silhouette Coefficient* pada 6 Perhitungan Jarak *K-Means Clustering*

Silhouette Coefficient Analysis in 6 Measuring Distances of K-Means Clustering

Rahmatina Hidayati[1], Anis Zubair[2], Aditya Hidayat Pratama[3], Luthfi Indana[4]
[1,2,3,4]Program Studi Sistem Informasi, Fakultas Teknologi Informasi,
Universitas Merdeka Malang
E-mail: [1]rahmatina.hidayati@unmer.ac.id, [2]anis.zubair.2@unmer.ac.id,
[3]aditya.hidayat@unmer.ac.id, [4]luthfi.indana@unmer.ac.id

Abstrak

Clustering merupakan proses pengelompokan sekumpulan data ke dalam klaster yang memiliki kemiripan. Kemiripan dalam satau klaster ditentukan dengan perhitungan jarak. Untuk melihat perfoma beberapa perhitungan jarak, dalam penelitian ini penulis menguji pada 6 data yang memiliki atribut berbeda, yakni 2, 3, 4, dan 6 atribut. Dari hasil uji perbandingan rumus jarak pada *K-Means clustering* menggunakan *Silhouette coefficient* dapat disimpulkan bahwa: 1) *Chebyshev distance* memiliki performa yang stabil baik untuk data dengan sedikit atribut maupun banyak. 2) *Average distance* memiliki hasil *Silhouette coefficient* paling tinggi dibandingkan dengan pengukuran jarak lain untuk data yang memiliki *outliers* seperti data 3. 3) *Mean Character Difference* mendapatkan hasil yang baik hanya untuk data dengan sedikit atribut. 4) *Euclidean distance, Manhattan distance,* dan *Minkowski distance* menghasilkan nilai baik untuk data yang memiliki sedikt atribut, sedangkan untuk data yang banyak atribut mendapatkan nilai cukup yang mendekati 0,5.

Kata kunci*: K-Means, Clustering, Silhouette Coefficient*

Abstract

Clustering is the process of grouping a set of data into clusters that have similarities. The similarity in a cluster is determined by calculating the distance. To see the performance of some distance calculations, in this study the authors tested 6 data which had different attributes, namely 2, 3, 4, and 6 attributes. From the results of the comparison test for the distance formula on K-Means clustering using the Silhouette coefficient, it can be concluded that: 1) Chebyshev distance has stable performance both for data with few and many attributes. 2) Average distance has the highest Silhouette coefficient result compared to other distance measurements for data that has outliers such as data 3. 3) Mean Character Difference gets good results only for data with few attributes. 4) Euclidean distance, Manhattan distance, and Minkowski distance produce good values for data that have few attributes, while data with many attributes get a sufficient value that is close to 0.5.

Keywords: *K-Means, Clustering, Silhouette Coefficient*

1. PENDAHULUAN

Clustering merupakan proses pengelompokan sekumpulan data ke dalam klaster yang memiliki kemiripan, di mana kesamaan dalam satu klaster dimaksimalkan, sedangkan kesamaan di luar klaster diminimalkan dengan perhitungan jarak. Salah satu metode yang digunakan untuk menemukan klaster dalam data adalah *K-Means clustering* yang mana k mempresentasikan jumlah klaster[1].

Penelitian [2] membandingkan *Euclidean distance, Manhattan distance,* dan *Minkowski distance* dalam pengukuran jarak *K-Means*. Berdasarkan uji homogenitas klaster dengan *Silhouette coefficient, Minkowski distance* mendapatkan nilai tertinggi, sedangkan *Euclidean*

distance memperoleh akurasi terbaik dengan *Chi-squre*. Penelitian [3] menguji *Euclidean distance*, *Manhattan distance*, dan *Chebyshev distance* dalam pengelompokan *member* di salah satu tempat fitenes menggunakan algoritma *K-Means*. Hasilnya, *Chebyshev distance* memiliki nilai *Silhouette coefficient* yang lebih baik dibandingkan *Euclidean distance* dan *Manhattan distance*.

Berdasarkan uji *Silhouette coefficient* pada penelitian [4], *Euclidean distance* memperoleh nilai tertinggi sebesar 0,64018 dibandingkan *Manhattan distance*, dan *Cosine distance.* Pada penelitian [5] *Euclidean distance* memiliki kinerja lebih baik dibandingkan dengan *Manhattan distance* berdasarkan evaluasi *Davies Bouldin Index (DBI)*.

Penelitian [6] menerapkan *K-Means clustering* pada data Ujian Nasional SMP tahun 2018/2019. Persamaan jarak yang digunakan adalah *Euclidean distance*. Hasil evaluasi dengan *Silhouette* mendapatkan nilai 0,464 yang berarti struktur pengelompokannya lemah. Sedangkan pada penelitian [7], penggunaan *Euclidean distance* untuk *text mining* dokumen skripsi menghasilkan *Silhouette* tertinggi 0,12 yang menandakan pengelompokannya tidak terstruktur. Hal ini berbeda dengan hasil pebelitian [8] di mana penggunaan *Euclidean distance* dalam *K-Means clustering* untuk pengelompokan pelanggan potensial berstruktur kuat dengan nilai *Silhouette* 0,85.

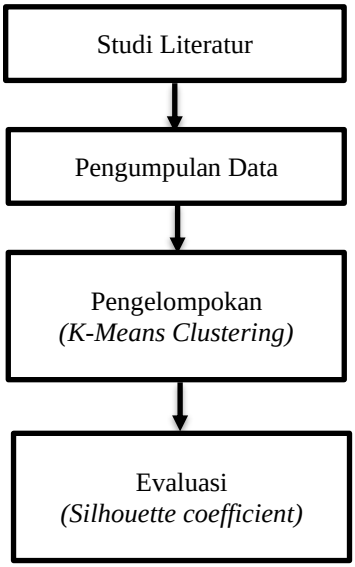
Penelitian [9] menguji *Euclidean distance* pada data evaluasi pembelajaran. Hasil terbaik didapat dengan nilai k=5 yang memperoleh *Silhouette coefficient* > 0,5. Pada penelitian [10], performa *Euclidean distance* lebih optimal dibandingkan dengan *Manhattan distance* untuk data penentuan promosi.

Berdasarkan studi literatur dari beberapa penelitian sebelumnya, persamaan perhitungan jarak yang sering digunakan adalah *Euclidean distance.* Analisis yang dihasilkan hanya berupa pemilihan jumlah k optimal. Sedangkan untuk sebagian penelitian yang lain melakukan perbandingan *Euclidean distance, Manhattan distance, Chebyshev distance,* dan *Minkowski distance*. Perbandingan rumus jarak tersebut dilakukan hanya untuk satu data.

Penelitian [11] menguji 3 perhitungan jarak *K-Means clustering* pada 3 data yang memiliki atribut berbeda. Pada pengujian data dengan 4 atribut, *Euclidean distance* memperoleh nilai *Purity* terbaik. Sedangkan pada data dengan 2 dan 6 atribut, *Carberra Purity* mendapatkan hasil terbaik dibandingkan *Euclidean Purity* dan *City Block Purity*.

Pada penelitian ini, penulis juga akan menguji beberapa perhitungan jarak pada 6 data yang memiliki atribut berbeda, yakni 2, 3, 4, dan 6 atribut. Persamaan jarak tersebut antara lain *Euclidean distance, Manhattan distance (City Block), Chebyshev distance, Minkowski distance*, *Average distance*, dan *Mean Character Difference*. Untuk mengevaluasi hasil pengelompokan, penulis menggunakan rumus *Silhouette coefficient*.

2. METODE PENELITIAN



Gambar 1 Tahapan Penelitian Analisis *Silhouette Coefficient* pada 6 Perhitungan Jarak *K-Means Clustering*

2.1 Pengumpulan data

Dalam penelitian ini data yang digunakan terdiri dari 6 data yang memiliki jumlah atribut berbeda. Data tersebut diperoleh dari Badan Pusat Statisik Indonesa, Badan Pusat Statistik Jawa Timur, dan situs resmi Provinsi Jawa Timur. Rincian dari data tersebut ditampilkan pada Tabel 1.

Tabel 1 Rincian Data Penelitian

No.	Data	Atribut	Sumber
1.	Presentase Rumah Tangga Mengakses Internet dalam 3 Bulan Terakhir	Tahun 2018-2019	hlm 449 [12]
2.	Produksi Tanaman Sayuran Semusim Menurut Jenis Tanaman (ton)	Tahun 2018-2019	hlm 281 [12]
3.	Angka Partisipaso Sekolah (APS) menurut Kota/Kab. Dan Kelompok Usia Sekolah Tahun 2019	Usia 7-12 Tahun, usia 13-15 tahun, usia 16-18 tahun	hlm 39 [13]
4.	Resiko Penularan Covid-19 di Provinsi Jawa Timur	Pasien positif yang masih aktif, pasien meninggal, dan suspek, *probable*	[14]
5.	Jumlah Des/Kel yang Memiliki Fasilitas Sekolah Menurut Provinsi dan Tingkat Pendidikan Tahun 2019	SD, SMP, SMA, SMK	hlm 198-199 [12]
6.	Jumlah Des/Kel yang Memiliki Sarana Kesehatan Menurut Provinsi Tahun 2019	Rumah Sakit, Rumah Sakit Bersalin, Poliklinik, Puskesmas, Puskesmas Pembantu, Apotek	hlm 200-201 [12]

2.2 Metode

A. K-Means Clustering

Clustering merupakan proses pengelompokan sekumpulan data ke dalam klaster yang memiliki kemiripan. Salah satu metode yang digunakan untuk menemukan klaster dalam data adalah *K-Means clustering* yang mana k mempresentasikan jumlah klaster [1]. Suatu data dikelompokkan ke dalam satu klaster berdasarkan kemiripan atribut yang dimiliki. Kemiripan ini bisa diketahui dengan menerapkan pengukuran jarak. Metode perhitungan jarak dalam penelitian ini antara lain [15]:

1) Euclidean Distance

Euclidean distance merupakan perhitungan jarak yang paling umum digunakan. Untuk 2 titik data x dan y dalam d-dimensi data, perhitungan jarak menggunaka *Euclidean distance* dirumuskan dengan Persamaan (1):

$$d_{euc}(x,y) = \sqrt{\sum_{j=1}^{d}(x_j - y_j)^2} \quad (1)$$

di mana,
$x_j, y_j =$ nilai dari j atribut

2) Manhattan Distance

Manhattan distance yang disebut juga *city block distance* didefinisikan sebagai jumlah jarak dari semua atribut.

$$d_{man}(x,y) = \sum_{j=1}^{d} |x_j - y_j| \quad (2)$$

3) Chebyshev Distance

Chebyshev disebut juga *Maximum distance* karena dalam penerapannya mencari nilai maksimum dari selisih jarak titik data x dan y di dimensi ruang d. Jarak maksimum tersebut ditulis seperti Persamaan (3):

$$d_{Che}(x,y) = \max_{1 \le k \le d} |x_j - y_j| \quad (3)$$

4) Minkowski Distance

Euclidean distance, *Manhattan distance*, dan *Chebyshev distance* merupakan tiga kasus khusus dari *Minkowski distance* yang didefinisikan seperti Persamaan (4):

$$d_{minw}(x,y) = \left(\sum_{j=1}^{d} |x_j - y_j|^r\right)^{\frac{1}{r}}, r \ge 1 \quad (4)$$

Jika nilai r = 2, 1, dan maka akan didapat *Euclidean distance*, *Manhattan distance*, dan *Chebyshev distance*. Dalam penelitian ini, menggunakan nilai r = 1,5.

5) Average Distance

Average distance merupakan modifikasi dari *Euclidean distance* untuk memperbaiki keurangan, di mana dua titik data yang tidak mempunyai nilai atribut biasanya memiliki jarak lebih kecil dibandingkan pasangan titik data yang berisi nilai atribut serupa.

$$d_{Ave}(x,y) = \frac{1}{d}\sqrt{\sum_{j=1}^{d}(x_j - y_j)^2} \quad (5)$$

6) Mean Character Difference

Mean character difference merupakan rata-rata dari *Manhattan distance*.

$$d_{mean}(x,y) = \frac{1}{d}\sum_{j=1}^{d} |x_j - y_j| \quad (6)$$

B. Silhouette Coefficient

Untuk melihat kualitas hasil pengelompokan masing-masing perhitungan jarak, maka perlu dilakukan uji homogenitas. Pengujian dilakukan setelah mencapai konvergensi 0 di mana hasil pengelompokan terakhir sama dengan pengelompokan sebelumnya. Dengan kata lain, tidak ada data yang berpindah klaster. Pengujian dihitung menggunakan persamaan *Silhouette coefficient*. Langkah dalam menghitung *Silhouette coefficient* di mulai dengan mencari jarak ratak-rata data ke-*i* dengan semua data di klaster yang sama, di sini kita asumsikan data ke-*i* berada di klaster A. Rumus dari *a(i)* ditulis dalam Persamaan (7) [16].

$$a(i) := \frac{1}{|A|-1} \sum_{j \in A, j \neq 1} d(i,j) \qquad (7)$$

di mana,
A = banyaknya data di klaster A

Selanjutnya menghitung nilai *b(i)* yang merupakan nilai minimum dari jarak rata-rata data ke-*i* dengan semua data di klaster berbeda. Sekarang, mari asumsikan klaster berbeda selain A dengan klaster C. Maka, perhitungan jarak rata-rata data ke-*i* dengan semua data di klaster C ditulis sebagai berikut:

$$d(i,C) := \frac{1}{|C|} \sum_{j \in C} d(i,j) \qquad (8)$$

di mana C = banyaknya data di klaster C

Setelah menghitung *d(i,C)* untuk semua klaster C ≠ A, selanjutnya memilih nilai jarak paling minimum sebagai nilai *b(i).*

$$b(i) := \min_{C \neq A} d(i,j) \qquad (9)$$

Jika klaster B memiliki nilai jarak minimum, maka *d(i,B) = b(i)* yang disebut sebagai tetangga dari data ke-*i* dan merupakan klaster terbaik kedua untuk data ke-*i* setelah klaster A. Setelah *a(i)* dan *b(i)* diketahui, maka proses terakhir menghitung *Silhouette coefficient* [16].

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i) - b(i)\}} \qquad (10)$$

Nilai *s(i)* berada antara -1 dan 1, di mana setiap nilai diiinterpretasi sebagai berikut:
s(i) ≈ 1 => data ke-*i* digolongkan dengan baik (dalam A)
s(i) ≈ 0 => data ke-*i* berada di tengah antara dua klaster (A dan B)
s(i) ≈ -1 => data ke-*i* digolongkan dengan lemah (dekat ke klaster B daripada A)

Penafsiran nilai *Silhouette coefficient* ditunjukkan dalam Tabel 2 [16].

Tabel 2 Interpretasi Nilai *Silhouete Coefficient*

Silhouette Coefficient	Interpretasi
0.71 - 1.00	Struktur yang dihailkan kuat
0.51 – 0.70	Struktur yang dihasilkan baik
0.26 – 0.50	Struktur yang dihasilkan lemah
≤ 0.25	Tidak tersruktur

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan 6 data dengan jumlah atribut berbeda yakni 2, 3, 4, dan 6 atribut. Salah satu data yang memiliki 4 atribut adalah data jumlah desa/kelurahan yang memiliki fasilitas sekolah menurut provinsi dan tingkat pendidikan tahun 2019 yang ditampilkan pada Tabel 3.

Tabel 3 Jumlah Des/Kel yang Memiliki Fasilitas Sekolah

No.	Provinsi	SD	SMP	SMA	SMK
1	Aceh	3475	1405	716	190
2	Sumatera Utara	5039	2280	1094	678
3	Sumatera Barat	1243	793	410	173
4	Riau	1816	1187	588	254
5	Jambi	1471	770	359	157
6	Sumatera Selata	2985	1333	639	233
7	Bengkulu	1187	477	186	100
8	Lampung	2532	1429	654	397
9	Kepulauan Bang	388	206	80	49
10	Kepulauan Riau	396	244	121	65
11	DKI Jakarta	264	257	215	211
12	Jawa Barat	5948	4187	2022	1989
13	Jawa Tengah	8443	3607	1266	1212
14	DI Yogyakarta	437	316	144	147
15	Jawa Timur	8443	4762	2420	1586
16	Banten	1541	1235	689	519
17	Bali	709	321	144	124
18	Nusa Tenggara B	1129	875	548	269
19	Nusa Tenggara T	3202	1613	575	280
20	Kalimantan Bara	2052	1117	403	171
21	Kalimantan Teng	1553	792	277	117
22	Kalimantan Selat	1867	769	305	104
23	Kalimantan Timu	984	541	234	153
24	Kalimantan Utar	316	154	59	30
25	Sulawesi Utara	1556	725	239	172
26	Sulawesi Tengah	1925	945	336	168
27	Sulawesi Selatan	2947	1739	770	337
28	Sulawesi Tengga	1871	887	386	156
29	Gorontalo	655	361	107	52
30	Sulawesi Barat	630	404	156	111
31	Maluku	1089	597	259	97
32	Maluku Utara	1093	629	292	140
33	Papua Barat	928	290	126	48
34	Papua	2147	612	236	126

Langkah-langkah K-Means Clustering:

(1) Tentukan jumlah k
Pada penelitian ini, k untuk seluruh data ditetapkan sama dengan 3.

(2) Tentukan nilai *centroid* awal
Nilai *centroid* didapat berdasarkan distribusi data. Tabel 4 menampilkan *centroid* awal dari data fasilitas sekolah di mana klaster 1 = nilai maksimu, klaster 2 = nilai rata-rata, dan klaster 3 = nilai minimum.

Tabel 4 Nilai *Centroid* Awal

	SD	**SMP**	**SMA**	**SMK**
c1	8443	4762	2420	1989
c2	2125,324	1113,5	501,6176	312,2059
c3	264	154	59	30

Perhitungan jarak titik data ke-1 dengan *centroid* klaster 1 menggunakan Persamaan (1) sampai Persamaan (6) dinyatakan sebagai berikut:

$$d_{euc}(x,y) = \sqrt{(3475-8443)^2 + (1405-4762)^2 + (716-2420)^2 + (190-1989)^2}$$
$$= 6487,72$$

$$d_{man}(x,y) = |3475-8443| + |1405-4762| + |716-2420| + |190-1989|$$
$$= 11828$$

$$d_{Che}(x,y) = \max|3475-8443| + \max|1405-4762| + \max|716-2420| + max|190-1989|$$
$$= 4968$$

$$d_{minw}(x,y) = (\,|3475-8443|^{1,5} + |1405-4762|^{1,5} + |716-2420|^{1,5} + |190-1989|^{1,5})^{1/1,5}$$
$$= 7818,369$$

$$d_{Ave}(x,y) = \frac{1}{34}\sqrt{(3475-8443)^2 + (1405-4762)^2 + (716-2420)^2 + (190-1989)^2}$$
$$= 1052,447$$

$$d_{mean}(x,y) = \frac{1}{34}(|3475-8443| + |1405-4762| + |716-2420| + |190-1989|)$$
$$= 311,263$$

Lakukan perhitungan jarak titik data ke-1 dengan nilai *centroid* klaster 2 dan juga klaster 3. Serta lakukan perhitungan serupa untuk titik data ke-2 sampai titik data ke-34. Setelah mengetahui jarak data dengan *centroid* semua klaster, selanjutnya mengelompokkan data ke dalam klaster yang memiliki hasil jarak minimum.

(3) *Update* nilai *centroid*

Nilai *centroid* baru merupakan rata-rata nilai data yang berada di tiap klaster. Pada hasil perhitungan *Euclidean distance*, Provinsi Jawa Barat, Jawa Tengah, dan Timur berada dalam klaster 1. Jadi, perhitungan *centroid* baru klaster 1 dilakukan dengan rumus berikut:

$$c1_{SD} = \frac{1}{3}(5948 + 8443 + 8443) = 7611{,}333$$
$$c1_{SMP} = \frac{1}{3}(4178 + 3607 + 4762) = 4185{,}333$$
$$c1_{SMA} = \frac{1}{3}(2022 + 1266 + 2420) = 1902{,}667$$
$$c1_{SMK} = \frac{1}{3}(1989 + 1212 + 1586) = 1595{,}667$$

Lakukan perhitungan yang sama untuk *centroid* kedua pada klaster 2 dan 3. Setelah mendapatkan nilai *centroid* baru, selanjutnya menghitung jarak kembali. Proses ini dilakukan berulang sampai mencapai konvergensi 0. Tabel 5 menampilkan jumlah data masing-masing klaster pada pengelompokan terakhir.

Tabel 5 Jumlah data Masing-Masing Klaster

Measure Distance	**C1**	**C2**	**C3**
Euclidean	3	13	18
Manhattan	3	9	22
Chebyshev	4	14	16
Minkowski	3	12	19
Average	3	13	18
Mean Character Difference	3	20	11

(4) Uji Homogenitas

Untuk melihat kemiripan data yang tergabung ke dalam satu kelompok, maka perlu dilakukan uji homogenitas. Pengujian ini menggunakan rumus *Silhouette coefficient.* Hasil perhitungan nilai *(ai), b(i),* dan *Silhouette coefficient (s)* masing-masing data ditunjukkan di Tabel 6.

Tabel 6 Nilai *a(i), b(i),* dan *s(i) Euclidean Distance*

a(1)	1358,78	**b(1)**	2792,73	**s(1)**	0,5135
a(2)	2991,21	**b(2)**	3619,69	**s(2)**	0,1736
a(3)	644,96	**b(3)**	1472,11	**s(3)**	0,5619
a(4)	1048,72	**b(4)**	1240,08	**s(4)**	0,1543
a(5)	744,19	**b(5)**	1291,30	**s(5)**	0,4237
a(6)	1056,24	**b(6)**	2305,37	**s(6)**	0,5418
a(7)	544,66	**b(7)**	1688,09	**s(7)**	0,6774
a(8)	977,17	**b(8)**	1967,49	**s(8)**	0,5033
a(9)	672,95	**b(9)**	2507,73	**s(9)**	0,7317
a(10)	646,81	**b(10)**	2475,57	**s(10)**	0,7387
a(11)	755,59	**b(11)**	2565,05	**s(11)**	0,7054
a(12)	2701,90	**b(12)**	5025,14	**s(12)**	0,4623
a(13)	2228,24	**b(13)**	6433,75	**s(13)**	0,6537
a(14)	607,77	**b(14)**	2400,40	**s(14)**	0,7468
a(15)	2148,65	**b(15)**	7206,02	**s(15)**	0,7018

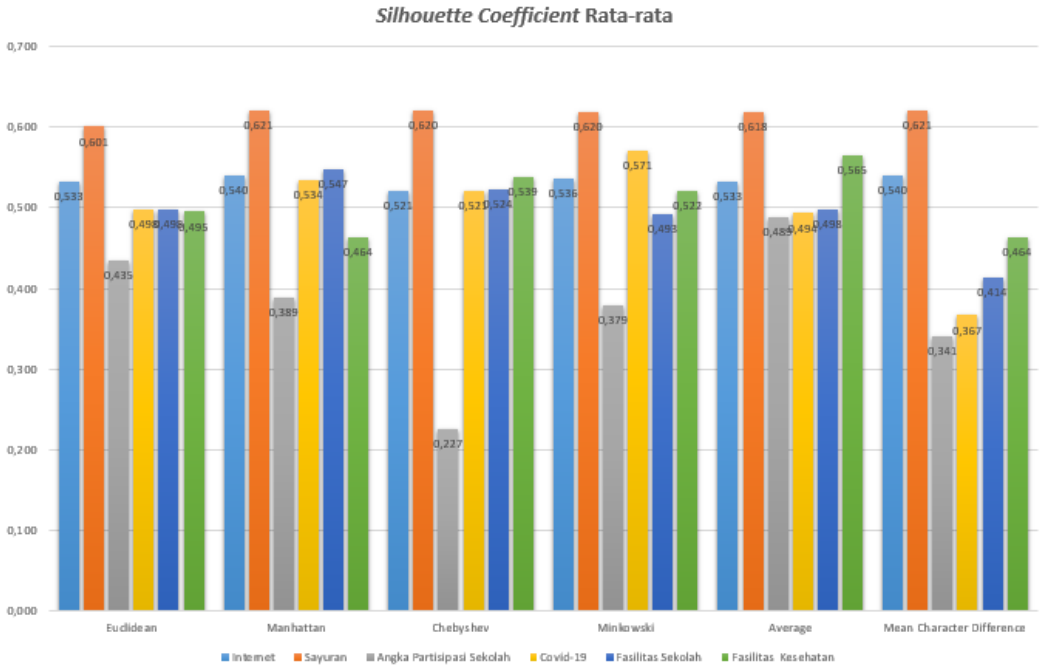
a(16)	1306,97	**b(16)**	1202,43	**s(16)**	-0,0800
a(17)	505,86	**b(17)**	2163,15	**s(17)**	0,7661
a(18)	728,33	**b(18)**	1548,70	**s(18)**	0,5297
a(19)	1201,54	**b(19)**	2604,78	**s(19)**	0,5387
a(20)	947,40	**b(20)**	1344,39	**s(20)**	0,2953
a(21)	803,63	**b(21)**	1237,18	**s(21)**	0,3504
a(22)	1112,22	**b(22)**	1039,55	**s(22)**	-0,0653
a(23)	482,66	**b(23)**	1814,32	**s(23)**	0,7340
a(24)	752,71	**b(24)**	2598,28	**s(24)**	0,7103
a(25)	788,54	**b(25)**	1265,78	**s(25)**	0,3770
a(26)	1004,24	**b(26)**	1150,35	**s(26)**	0,1270
a(27)	1164,48	**b(27)**	2483,86	**s(27)**	0,5312
a(28)	1035,48	**b(28)**	1085,64	**s(28)**	0,0462
a(29)	518,08	**b(29)**	2205,90	**s(29)**	0,7651
a(30)	512,05	**b(30)**	2195,17	**s(30)**	0,7667
a(31)	504,41	**b(31)**	1701,83	**s(31)**	0,7036
a(32)	513,66	**b(32)**	1676,45	**s(32)**	0,6936
a(33)	529,04	**b(33)**	2007,31	**s(33)**	0,7364
a(34)	1153,74	**b(34)**	1287,85	**s(34)**	0,1041

$$s_{rata-rata} = \frac{1}{34}(16{,}9203) = 0{,}4976$$

Tabel 7 menampilkan hasil keseluruhan *Silhouette coefficient* rata-rata semua data pada masing-masing jarak, sedangkan Gambar 2 menunjukkan grafik perbandingannya.

Tabel 7 Hasil *Silhouette Coefficient* Rata-rata

Data / **Jarak**	***Euclidean***	***Manhattan***	***Chebyshev***	***Minkowski***	***Average***	***Mean Character Difference***
Internet	0,533	0,540	0,521	0,536	0,533	0,540
Sayuran	0,601	0,621	0,620	0,620	0,618	0,621
Angka Partisipasi Sekolah	0,435	0,389	0,227	0,379	0,489	0,341
Covid-19	0,498	0,534	0,521	0,571	0,494	0,367
Fasilitas Sekolah	0,498	0,547	0,524	0,493	0,498	0,414
Fasilitas Kesehatan	0,495	0,464	0,539	0,522	0,565	0,464



Gambar 2 Grafik perbandingan

4. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Dari hasil uji perbandingan rumus jarak *K-Means clustering* pada 6 data yang memiliki atribut berbeda menggunakan *Silhouette coefficient*, dapat disimpulkan bahwa:

1) *Chebyshev distance* memiliki performa yang stabil baik untuk data dengan sedikit atribut maupun banyak.
2) *Average distance* memiliki hasil *Silhouette coefficient* paling tinggi dibandingkan dengan pengukuran jarak lain untuk data yang memiliki *outliers* seperti data 3.
3) *Mean Character Difference* mendapatkan hasil yang baik hanya untuk data dengan sedikit atribut.
4) *Euclidean distance*, *Manhattan distance*, dan *Minkowski distance* menghasilkan nilai baik untuk data yang memiliki sedikt atribut, sedangkan untuk data yang banyak atribut mendapatkan nilai cukup yang mendekati 0,5.

4.2 Saran

Pada penelitian selanjutnya, pengujian hasil pengelompokan menggunakan selain *Silhouette coefficient*.

DAFTAR PUSTAKA

[1] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Ed.2, Morgan Kaufmann, San Francisco, 2006.

[2] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma *K-Means* Clustering berbasis Chi-Square," *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, vol.04, no.01, pp. 20-24, 2019, doi: 10.30591/jpit.v4i1.1253.

[3] M. Anggara, H. Sujiani, and H. Nasution, "Pemilihan Distance Measure Pada K-Means Clustering Untuk Pengelompokkan *Member* Di Alvaro Fitness," *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, vol. 1, no. 1, pp. 1-6, 2016, [Online]. Available: https://jurnal.untan.ac.id/index.php/justin/article/view/13119/11875.

[4] I G. Harsemadi, "Perbandingan Distance Measure Pada K-Means Clustering untuk Pengelompokkan Musik Terhadap Suasana Hati," *Seminar Nasional Teknologi Informasi dan Multimedia*, pp. 13-18, 2018, [Online]. Available: https://ojs.amikom.ac.id/index.php/semnasteknomedia/article/view/2050.

[5] W. Gie, and D. Jollyta, "Perbandingan Euclidean dan Manhattan Untuk Optimasi Cluster Menggunakan Davies Bouldin Index: Status Covid-19 Wilayah Riau," *Prosiding Seminar Nasional Riset Dan Information Science (SENARIS)*, vol. 2, pp.187-191, 2020, doi: http://dx.doi.org/10.30645/senaris.v2i0.160.

[6] A. Aditya, I. Jovian, and B. N. Sari, "Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019," *Jurnal Media Informatika Budidarma*, vol. 4, no. 1, pp. 51-58, 2020, doi: 10.30865/mib.v4i1.1784.

[7] D. A. C. Rachman, R. Goejantoro, and F. D. T. Amijaya, "Implementasi Text Mining Pengelompokkan Dokumen Skripsi Menggunakan Metode K-Means Clustering," *Jurnal EKSPONENSIAL*, vol. 11, no. 2, pp. 167-174, 2020, [Online]. Available: http://jurnal.fmipa.unmul.ac.id/index.php/exponensial/article/view/660.

[8] A. P. Tiratana, B. Mulyawan, and M. D. Lauro, "Pembuatan Aplikasi Customer Relationship Management Berbasis Web Menggunakan Metode K-Means," *Jurnal Ilmu Komputer dan Sistem Informasi*, vol 7, no. 2, pp. 179-184, 2019, [Online]. Available: https://journal.untar.ac.id/index.php/jiksi/article/view/7372.

[9] C. C. B. Pradana, "Pengelompokan Data Evaluasi Pembelajaran Menggunakan Algoritma K-Means++ Clustering," Skripsi, Program Studi Teknik Informatika, Univ. Sanata Dharma, Yogyakarta, 2019.

[10] R. I. Fajriah, H. Sutisna, and B. K. Simpony, "Perbandingan Distance Space Manhattan Dengan Euclidean Pada K-Means Clustering Dalam Menentukan Promosi," *Indonesian Journal on Computer and Information Technology (IJCIT)*, vol. 4, no. 1, pp. 36-49, 2019, doi: https://doi.org/10.31294/ijcit.v4i1.4630.

[11] F. A. Sebayang, M. S. Lydia, and B. B. Nasution, "Optimization on Purity K-Means Using Variant Distance Measure," *IEEE 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT)*, pp. 143-147, 2020, doi: 10.1109/MECnIT48290.2020.9166600.

[12] BPS-Statistics Indonesi, "Statistik Indonesia," 2020.

[13] BPS Provinsi Jawa Timur, "Statistik Pendidikan Provinsi Jawa Timur," 2019.

[14] http://covid19.go.id diakses Februari 2021.

[15] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics, Alexandria, VA, 2007.

[16] A. Struyf, M. Hubert, and P. J. Rousseeuw, "Clustering in an Object-Oriented Environment," *Journal of Statictical Software*, vol 1, Issue 4, pp. 1-30, 1997.