JURNAL SISFOTENIKA
Vol. 13 No. 1, Januari 2023
p-ISSN: 2087-7897 ; e-ISSN : 2460-5344
DOI : 10.30700/jst.v13i1.1337

K-Means untuk Klasterisasi Daerah Rawan Penyakit Demam Berdarah

K-Means for Clustering of Dengue Fever Prone Areas

Novianti Puspitasari*[1], Andre Ardin Maulana[2], Rosmasari[3], Faza Alameka[4]
[1,2,3]Program Studi Informatika, Universitas Mulawarman, Samarinda
[4]Program Studi Sistem Informasi, Universitas Mulia, Samarinda
e-mail: *[1]novia.ftik.unmul@gmail.com, [2]andreardinmaulanaa@gmail.com,
[3]rosmasari.unmul@gmail.com [4]faza.alameka@gmail.com

Abstrak

Jumlah penderita penyakit demam berdarah mengalami peningkatan dan berada pada level cukup tinggi. Tingginya jumlah kasus demam berdarah terkadang tidak diimbangi dengan ketersediaan informasi yang dimiliki oleh dinas maupun pemerintah daerah tentang daerah-daerah penyebaran demam berdarah. Oleh karena itu, clustering daerah rawan demam berdarah perlu dilakukan untuk memberikan informasi bagi pihak yang berkepentingan agar pemerintah dapat melakukan tindakan penanganan yang tepat berdasarkan tingkat penyebarannya. Algoritma clustering yang digunakan adalah K-Means. Metode perhitungan jarak yang digunakan adalah Euclidean Distance, Manhattan Distance dan Minkowski Distance. Dari ketiga metode jarak tersebut dilakukan perhitungan akurasi menggunakan Sum of Squared Error (SSE) untuk mengetahui perhitungan jarak yang ideal dan untuk melihat jumlah cluster yang optimal. Berdasarkan data jumlah kasus demam berdarah di Kota Samarinda selama lima tahun dari berbagai wilayah diperoleh hasil bahwa cluster C1 merupakan tingkat kerawanan tinggi, C2 untuk tingkat kerawanan sedang dan C3 untuk tingkat kerawanan rendah. Lebih lanjut, tiga cluster merupakan merupakan jumlah ideal untuk melakukan clustering karena memiliki nilai SSE lebih kecil. Metode pengukuran jarak yang ideal adalah Minkowski Distance karena nilai selisih SSE Minkowski Distance paling rendah diantara metode jaraknya lainnya yaitu sebesar 13.0803.

Kata kunci*—demam berdarah, daerah, clustering, K-Means, SSE*

Abstract

The number of dengue fever sufferers has increased at a reasonably high level. The high number of cases of dengue fever is sometimes different from the availability of information owned by the department or local government about the areas where dengue fever is spread. Therefore, clustering areas prone to dengue fever needs to be carried out to provide information for interested parties so that the government can take appropriate handling measures based on the level of its spread. The clustering algorithm used is K-Means. The calculation methods used are Euclidean Distance, Manhattan Distance and Minkowski Distance. Accuracy calculations of the three distance methods are carried out using the Sum of Squared Error (SSE) to determine the ideal distance calculation. In addition, SSE is also used to see the optimal number of clusters. Based on data on the number of cases of dengue fever in Samarinda City for five years from various regions, the results show that cluster C1 is a high vulnerability level, C2 is a medium vulnerability level, and C3 is a low vulnerability level. Furthermore, three clusters are the ideal number for clustering because it has a smaller SSE value. The perfect distance measurement method is the Minkowski Distance because the

Minkowski Distance SSE difference is the lowest among the other distance methods, which is 13.0803.
Keywords—*dengue fever, area, clustering, K-Means, SSE*

1. PENDAHULUAN

Kementerian Kesehatan Republik Indonesia mencatat pada tahun 2021 terdapat 71.856 kasus penderita Demam berdarah *Dengue* (DBD) di seluruh wilayah Indonesia dengan jumlah penderita meninggal sebanyak 696. Pada tahun 2022 jumlah kasus demam berdarah cenderung meningkat dan bersifat fluktuatif, namun masih termasuk ke dalam kategori yang cukup tinggi [1]. DBD adalah salah satu masalah kesehatan masyarakat di Indonesia yang disebabkan oleh lingkungan, dimana penyebarannya semakin meluas sehingga jumlah penderitanya pun cenderung meningkat. Virus *dengue* ditularkan kepada manusia melalui gigitan nyamuk *Aedes Aegypti* dan *Aedes Albopictus* yang telah terinfeksi [2]. Kedua jenis nyamuk ini terdapat hampir di seluruh pelosok dunia termasuk di Indonesia. Salah satu provinsi di Indonesia yang memiliki jumlah penderita penyakit DBD tertinggi adalah provinsi Kalimantan Timur, khususnya di Kota Samarinda. Berdasarkan data profil kesehatan di Kota Samarinda pada tahun 2016 yang dilaporkan melalui Sistem Informasi Daerah (SIKDA), jumlah penderita DBD adalah 2814 kasus dengan jumlah kematian sebanyak 18 orang. Angka kesakitan (*Incidance Rate*/IR) sebesar 290.6 per 100000 penduduk dan angka kematian (*Case Fatality Rate*/CFR) sebesar 0.6%. Dari laporan tersebut, terlihat bahwa angka kesakitan DBD di kota Samarinda tergolong tinggi. Lebih lanjut, berdasarkan data dari website Badan Pusat Statistik Provinsi Kalimantan Timur tahun 2019, jumlah penderita DBD sebanyak 1554 kasus [3].

Jumlah penderita yang cukup besar membuat penyakit demam berdarah menjadi permasalahan yang serius bagi masyarakat dan pemerintah. Informasi terkait daerah rawan atau daerah yang memiliki jumlah penyebaran penyakit DBD sangat besar, tentunya dibutuhkan oleh pemerintah daerah. Informasi tersebut dapat digunakan oleh pemerintah untuk memberikan penanganan yang tepat dalam mencegah penyakit demam berdarah sehingga jumlah penderita demam berdarah mengalami penurunan di masa yang akan datang. Salah satu cara untuk memperoleh informasi tentang daerah rawan demam berdarah adalah melakukan pengelompokan terhadap daerah-daerah rawan demam berdarah. Hal ini bertujuan untuk membantu dinas maupun pemerintah dalam memantau daerah yang rawan demam berdarah dan mencegah penyebarannya agar tidak semakin menyebar luas ke daerah yang belum terdampak. Pengelompokan daerah rawan penyakit dilakukan menggunakan metode *Clustering*.

Pengelompokkan atau *clustering* telah banyak dilakukan menggunakan berbagai metode dan algoritma, diantaranya adalah metode *Fuzzy C-Means* untuk *clustering* rekredensialing fasilitas kesehatan dengan menghasilkan nilai PCI 0.50002 dan PEI 0.99998 yang berarti tingkat akurasi dan nilai keanggotaan dari *cluster* cukup baik [4]. Selanjutnya, *clustering* data menggunakan metode PAM atau *K-Medoids* untuk transaksi bongkar muat di Provinsi Riau yang menghasilkan waktu rata-rata 1 menit 38 detik untuk pengolahan data pada *RapidMiner* dengan nilai DBI sebesar 0.119 [5]. Metode berikutnya yaitu *clustering* menggunakan *subtractive fuzzy c-means* (SFCM) untuk pengelompokkan rumah tangga miskin dengan presentasi nilai akurasi yang dihasilkan lebih besar dibandingkan *K-Means* dengan persentase sebesar 94.8%. Namun, penelitian membuktikan bahwa dari sisi waktu pemrosesan, *K-Means* lebih cepat bila dibandingkan dengan SFCM dengan durasi sekitar 9 detik [6]. Selain itu masih banyak lagi penelitian tentang *clustering* yang menggunakan berbagai metode dengan pengukuran akurasi *cluster* yang berbeda-beda [7]–[12].

Dari sekian banyak metode *clustering* yang telah diimplementasikan, metode *K-Means* merupakan metode yang efektif dan optimal untuk mendapatkan hasil pengelompokkan yang lebih baik dibandingkan dengan metode lainnya [13]–[17]. *Clustering* dengan metode *K-Means* ini sangat cocok jika dipakai untuk memetakan atau mengelompokkan sebuah tempat karena

metode ini dapat mengelompokkan sebuah data dalam satu *cluster*. Hal tersebut menjadi dasar penelitian ini untuk menerapkan *K-Means* pada pengelompokan daerah rawan penyakit demam berdarah di Kota Samarinda. Harapannya, hasil dari penelitian ini membantu pemerintah dalam mencegah, mengendalikan dan memberantas penyebaran penyakit demam berdarah.

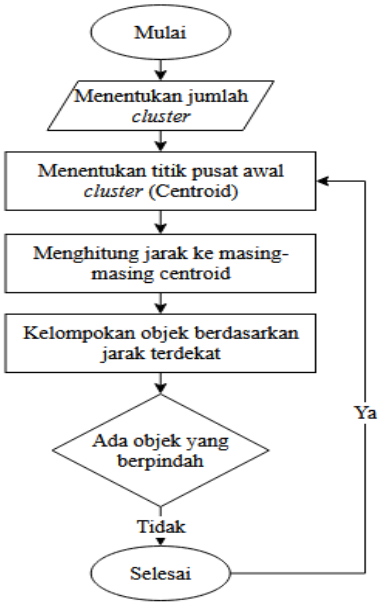
2. METODE PENELITIAN

2. 1 Clustering

Clustering merupakan suatu proses pengelompokkan *record*, *observasi*, atau mengelompokkan kelas yang memiliki kesamaan objek [18]. Tujuan utama dari metode *clustering* adalah mengelompokkan sejumlah data atau objek ke dalam *cluster* (*group*) sehingga dalam setiap *cluster* dapat berisi data yang semirip mungkin. Analisis klaster bertujuan menemukan kelompok objek sedemikian rupa sehingga objek-objek dalam grup akan sama atau terkait satu sama lain dan berbeda dari (atau tidak terkait) objek-objek dalam grup lain [19].

2. 2 K-Means

K-Means adalah metode *clustering* berbasis jarak yang membagi data ke-*n* dalam *cluster* dan algoritma ini bekerja pada atribut numerik. Metode *K-Means* termasuk dalam *partitioning clustering* yang memisahkan data ke *k* daerah bagian yang terpisah [20]. *K-Means* termasuk dalam metode non-*hierarchical* yang mempartisipasi data ke dalam bentuk satu atau lebih *cluster* [21]. Algoritma ini secara acak akan memilih pola *k* sebagai titik awal *centroid*. Jumlah iterasi untuk mencapai *cluster centroid* akan dipengaruhi oleh kandidat *cluster centroid* awal yang ditentukan secara *random* jika posisi *centroid* baru tidak berubah [22]. Gambar 1 merupakan *flowchart* Algoritma *K-Means* [19].



Gambar 1. *Flowchart* Algoritma *K-Means*

Gambar 1 merupakan tahapan algoritma K-Means, yang pertama adalah menentukan jumlah *k* (*cluster*) yang menjadi acuan pengelompokan data. Kedua, menentukan titik pusat (*centroid*) dari masing-masing *cluster* yang ada secara *random*. Tahap ketiga adalah menghitung jarak terdekat masing-masing data ke pusat *centroid*. Jarak paling dekat antara satu data dengan satu *cluster* tertentu akan menentukan suatu data masuk dalam *cluster* mana. Perhitungan jarak

semua data ke setiap titik pusat *cluster* dapat menggunakan metode perhitungan jarak [23]. Langkah keempat yaitu melakukan perhitungan kembali pusat *cluster* dengan keanggotaan *cluster* yang sekarang. Pusat *cluster* adalah rata-rata dari semua data/objek dalam *cluster* tertentu [24]. Perhitungan kembali pusat *cluster* dapat menggunakan Persamaan (1).

$$C(i) = \frac{x_1 + x_2 + x_{...} + x_{...}}{\sum x} \quad (1)$$

Di mana:
x_1 = Nilai data *record* ke-1, dst
$\sum x$ = Jumlah data *record*

Mengulangi langkah ketiga dan keempat hingga tidak ada data *record* yang berpindah *cluster* atau konvergen.

2.3 Metode Pengukuran Jarak

Pengukuran jarak memegang peran yang sangat penting dalam menentukan kemiripan atau keteraturan di antara data dan item [10]. Pada penelitian ini metode pengukuran jarak yang digunakan yaitu *Euclidean Distance*, *Manhattan Distance* dan *Minkowski Distance*. Tiga metode ini merupakan metode pengukuran jarak yang umum digunakan [23].

2.3.1 Euclidean Distance

Euclidean distance merupakan salah satu metode perhitungan jarak yang digunakan untuk mengukur jarak dari 2 (dua) buah titik dalam *Euclidean space* (meliputi bidang *euclidean* dua dimensi, tiga dimensi, atau bahkan lebih) [10]. Rumus pengukuran jarak menggunakan metode *Euclidean Distance* dapat dilihat pada Persamaan (2).

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

Di mana:
d = Jarak antara x dan y
x = Data pusat *cluster*
y = Data pada atribut
i = Setiap data
n = Jumlah data
x_i = Data pada pusat *cluster* ke i
y_i = Data pada setiap data ke i

2.3.2 Manhattan Distance

Manhattan distance digunakan untuk menghitung perbedaan absolut (mutlak) antara koordinat sepasang objek [10]. Rumus metode *Manhattan Distance* adalah pada Persamaan (3).

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

Di mana:
d = Jarak antara x dan y
x = Data pusat *cluster*
y = Data pada atribut
i = Setiap data
n = Jumlah data

2.3.3 Minkowski Distance

Minkowski distance merupakan sebuah metrik dalam ruang vektor di mana suatu norma didefinisikan (*normed vector space*) sekaligus dianggap sebagai generalisasi dari *Euclidean distance* dan *Manhattan distance*. Pada pengukuran jarak objek menggunakan *minkowski distance* biasanya nilai p yang digunakan adalah 1 atau 2 [10]. Rumus perhitungan jarak menggunakan metode *Minkowski Distance* dapat dilihat pada Persamaan (4).

$$d(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \quad (4)$$

Di mana:
d = Jarak antara x dan y
x = Data pusat *cluster*
y = Data pada atribut
i = Setiap data
n = Jumlah data
x_i = Data pada pusat *cluster* ke i
y_i = Data pada setiap data ke i
p = *Power*

2.4 Metode Normalisasi Z-Score

Normalisasi adalah proses penskalaan nilai atribut dari data sehingga bisa terletak pada rentang tertentu [25]. Normalisasi *Z-Score* merupakan metode normalisasi berdasarkan *mean* (nilai rata-rata) dan *standard deviation* (deviasi standar) dari data. Metode ini sangat berguna jika tidak diketahui nilai aktual minimum dan maksimum dari data. Pencarian nilai normalisasi *z-score* menggunakan Persamaan (5).

$$Z_i = \frac{X_i - \overline{X}}{S} \quad (5)$$

Di mana:
Z_i = Data Hasil Normalisasi
X_i = Data Asli
$\overline{X}$ = Nilai Rata-rata Baku Rujukan
S = Nilai Simpang Baku Rujukan

2.5 Data Penelitian

Tahapan penelitian yang harus dilakukan untuk mendapatkan data penelitian adalah:

1. Mengidentifikasi masalah yang berkaitan dengan kurangnya informasi terkait daerah rawan atau daerah yang memiliki jumlah penyebaran penyakit DBD sangat besar di Kota Samarinda.
2. Menentukan variabel yang akan digunakan dalam penelitian ini. Variabel yang digunakan adalah jumlah kasus DBD dari seluruh kecamatan di Kota Samarinda.
3. Melakukan observasi lapangan dengan mengunjungi Dinas Kesehatan Kota Samarinda untuk melakukan pengambilan data secara langsung.
4. Mengolah data penelitian yang telah didapatkan menggunakan *Ms. Excel* dan menganalisis data menggunakan metode CRISP-DM.

Hasil dari observasi lapangan dan pengolahan data, diperoleh data jumlah kasus demam berdarah selama lima tahun dari tahun 2015 sampai 2018 yang tersebar di berbagai puskesmas kecamatan di Kota Samarinda. Data jumlah kasus demam berdarah tersebut menjadi data penelitian yang ditampilkan oleh Tabel 1.

Tabel 1. Data Jumlah Kasus DBD Kota Samarinda

No	Puskesmas	Tahun			
		2015	2016	2017	2018
1	Palaran	81	241	17	55
2	Bantuas	4	13	11	2
3	Bukuan	20	53	0	14
4	Mangkupalas	58	99	42	30
5	Baqa	93	83	16	43
6	Harapan Baru	165	128	14	34
7	Trauma Center	35	63	22	12
8	Loa Bakung	104	107	50	45
9	Karang Asam	97	133	11	47
10	Wonorejo	51	143	40	34
11	Juanda	46	109	20	93
12	Air Putih	83	149	41	101
13	Segiri	77	133	18	54
14	Pasundan	116	134	19	59
15	Sidomulyo	151	349	40	79
16	Sungai Kapih	24	53	7	5
17	Sambutan	45	109	8	46
18	Makroman	7	44	4	8
19	Bengkuring	48	95	37	71
20	Sempaja	63	114	11	75
21	Sungai Siring	6	39	4	32
22	Lempake	17	74	15	72
23	Remaja	60	123	15	62
24	Temindung	90	226	36	99

Tabel 1 menampilkan data set yang digunakan dalam penelitian ini. Data set penelitian berjumlah 24 data yang berasal dari Dinas Kesehatan Kota Samarinda. Data set ini memiliki variabel tahun yang di dalamnya berisi jumlah kasus DBD dari setiap puskesmas yang berada di setiap kecamatan di Kota Samarinda.

2.6 Pengukuran Akurasi K-Means

Pada metode *K-Means,* pengelompokkan data dilakukan dengan cara mengelompokkan atribut data ke dalam sejumlah *cluster* berdasarkan kemiripan data atribut. Kemudian, kemiripan suatu kelompok data diukur menggunakan suatu metode pengukuran jarak. Metode pengukuran jarak yang digunakan adalah *Sum of Squared Error* (SSE). Metode ini akan memberikan informasi *error* jarak data ke *centroid*. Jika semakin kecil nilai SSE, maka semakin akurat hasil *clustering* yang dibuat. Berikut rumus untuk menghitung SSE pada Persamaan (6) [23].

$$SSE = \sum_{k=1}^{k} \sum_{p \in C_i} d = (p, m_i)^2 \quad (6)$$

Di mana:
$p \in C_i$ = Tiap data poin pada *cluster i*;
m_i = *Centroid* dari *cluster i*;
d = Jarak/ *distances/ variance* terdekat pada masing-masing *cluster* i.

3. HASIL DAN PEMBAHASAN

Pengelompokkan daerah rawan demam berdarah menggunakan *K-Means* sudah dilakukan menggunakan tiga metode pengukuran jarak. Pada proses awal, dilakukan normalisasi terlebih dahulu dengan tujuan mencari rentang nilai *maximum* dan *minimum* dari data jumlah kasus demam berdarah menggunakan Persamaan (5). Data jumlah kasus demam berdarah yang sudah dinormalisasi ditampilkan oleh Tabel 2.

Tabel 2. Normalisasi Jumlah Kasus DBD

No	Puskesmas	Tahun			
		2015	2016	2017	2018
1	Palaran	0.3207	3.1339	-0.8046	-0.1364
2	Bantuas	-1.0332	-0.8749	-0.9101	-1.0683

No	Puskesmas	Tahun			
		2015	2016	2017	2018
3	Bukuan	-0.7518	-0.1716	-1.1035	-0.8573
4	Mangkupalas	-0.0837	0.6372	-0.3650	-0.5760
5	Baqa	0.5317	0.3559	-0.8222	-0.3474
6	Harapan Baru	1.7976	1.1471	-0.8573	-0.5057
7	Trauma Center	-0.4881	0.0042	-0.7167	-0.8925
8	Loa Bakung	0.7251	0.7779	-0.2244	-0.3123
9	Karang Asam	0.6020	1.2350	-0.9101	-0.2771
10	Wonorejo	-0.2068	1.4108	-0.4002	-0.5057
11	Juanda	-0.2947	0.8130	-0.7518	0.5317
12	Air Putih	0.3559	1.5163	-0.3826	0.6724
13	Segiri	0.2504	1.2350	-0.7870	-0.1540
14	Pasundan	0.9361	1.2526	-0.7694	-0.0661
15	Sidomulyo	1.5515	5.0329	-0.4002	0.2855
16	Sungai Kapih	-0.6815	-0.1716	-0.9804	-1.0156
17	Sambutan	-0.3123	0.8130	-0.9628	-0.2947
18	Makroman	-0.9804	-0.3299	-1.0332	-0.9628
19	Bengkuring	-0.2595	0.5669	-0.4529	0.1449
20	Sempaja	0.0042	0.9009	-0.9101	0.2152
21	Sungai Siring	-0.9980	-0.4178	-1.0332	-0.5409
22	Lempake	-0.8046	0.1976	-0.8398	0.1625
23	Remaja	-0.0485	1.0592	-0.8398	-0.0134
24	Temindung	0.4789	2.8702	-0.4705	0.6372

Selanjutnya, jumlah *cluster* telah ditentukan sebanyak K=3, yaitu *cluster* daerah rawan kategori tinggi (C1), sedang (C2) dan rendah (C3). Setelah menentukan jumlah *cluster*, proses berikutnya adalah menentukan *centroid* dari tiap-tiap *cluster*. Pengambilan *centroid* dilakukan dengan mengurutkan data yang dinormalisasi dari terbesar sampai terkecil kemudian mengambil nilai data yang pertama, data yang berada di tengah dan data yang terakhir untuk dijadikan titik pusat awal atau *centroid* awal. Dalam penelitian ini, *centroid* awal yang diambil dapat dilihat pada Tabel 3.

Tabel 3. Nilai *Centroid* Awal

Centroid	*Centroid*			
	2015	2016	2017	2018
C1	1.7976	5.0329	-0.2244	0.6724
C2	-0.0837	0.8130	-0.8222	-0.2947
C3	-1.0332	-0.8749	-1.1035	-1.0683

Selanjutnya adalah menghitung jarak data *centroid* setiap *cluster* dengan menggunakan perhitungan jarak yaitu *Euclidean Distance, Manhattan Distance* dan *Minkowski Distance*.

3.1 Perhitungan Euclidean Distance

Dari hasil perhitungan *K-Means* menggunakan metode jarak *Euclidian Distance* menggunakan persamaan (2) diperoleh hasil perhitungan keseluruhan data seperti terlihat pada Tabel 4.

Tabel 4. Hasil *Centroid Euclidean Distance*

No	Puskesmas	Jarak Data			
		C1	C2	C3	*CLUSTER*
1	Palaran	2.3372	1.9075	3.6260	C2
2	Bantuas	6.6087	2.7001	0.7370	C3
3	Bukuan	5.8474	1.9747	0.2237	C3
4	Mangkupalas	4.7686	0.9268	1.3022	C2
5	Baqa	4.8470	0.9632	1.5376	C2
6	Harapan Baru	3.9994	1.5758	2.9783	C2
7	Trauma Center	5.5620	1.6801	0.5013	C3
8	Loa Bakung	4.3791	0.8054	2.0369	C2
9	Karang Asam	3.9877	0.4423	2.1088	C2
10	Wonorejo	4.1033	0.7469	1.8688	C2
11	Juanda	4.6260	0.9426	1.7500	C2
12	Air Putih	3.7344	0.8395	2.6111	C2

13	Segiri	4.0571	0.1575	1.9297	C2
14	Pasundan	3.8638	0.6444	2.4146	C2
15	Sidomulyo	0.0000	4.0250	5.9076	C1
16	Sungai Kapih	5.8397	1.9854	0.3211	C3
17	Sambutan	4.6834	0.8314	1.2608	C2
18	Makroman	6.0934	2.2417	0.2993	C3
19	Bengkuring	4.8216	0.9197	1.4172	C2
20	Sempaja	4.4421	0.5763	1.7089	C2
21	Sungai Siring	6.1069	2.1842	0.3262	C3
22	Lempake	5.3981	1.5397	1.0132	C3
23	Remaja	4.3166	0.4276	1.6887	C2
24	Temindung	2.4406	1.8003	3.6816	C2

Dari proses perhitungan yang telah dilakukan selama tiga kali iterasi, terlihat bahwa *Cluster* 1 (C1) beranggotakan 1 data, *Cluster* 2 (C2) beranggotakan 16 data dan *Cluster* 3 (C3) beranggotakan 7 data.

3.2 Perhitungan Manhattan Distance

Perhitungan *K-Means* menggunakan metode jarak *Manhattan Distance* dilakukan dengan Persamaan (3) yang menghasilkan jarak data ke *centroid* seperti pada Tabel 5.

Tabel 5. Hasil *Centroid Manhattan Distance*

No	Puskesmas	Jarak Data			
		C1	C2	C3	*CLUSTER*
1	Palaran	2.4792	2.1931	5.2698	C2
2	Bantuas	8.8793	4.5070	1.2006	C3
3	Bukuan	7.8770	3.5048	0.4245	C3
4	Mangkupalas	5.5210	1.6434	2.3686	C2
5	Baqa	5.2748	1.3902	2.4741	C2
6	Harapan Baru	4.9759	2.1052	4.3379	C2
7	Trauma Center	7.0858	2.7136	0.9696	C3
8	Loa Bakung	4.4484	1.4453	3.7225	C2
9	Karang Asam	4.3429	0.8217	3.4060	C2
10	Wonorejo	4.7649	1.4617	3.0544	C2
11	Juanda	4.8352	1.6024	3.0544	C2
12	Air Putih	3.3583	1.5414	4.9181	C2
13	Segiri	4.4484	0.2989	3.3005	C2
14	Pasundan	3.6396	0.9061	4.1093	C2
15	Sidomulyo	1.8286	5.8492	9.2259	C1
16	Sungai Kapih	7.8419	3.4696	0.5300	C3
17	Sambutan	5.7495	1.3773	2.0346	C2
18	Makroman	8.2990	3.9268	0.5501	C3
19	Bengkuring	4.9935	1.5871	2.7555	C2
20	Sempaja	4.7825	1.0573	2.9664	C2
21	Sungai Siring	7.9825	3.6103	0.6305	C3
22	Lempake	4.8352	0.6529	2.9137	C2
23	Remaja	1.8286	2.8953	6.2720	C1
24	Temindung	5.7495	1.3773	2.0346	C2

Dari hasil perhitungan algoritma *K-Means* menggunakan metode *Manhattan Distance* yang berhenti pada iterasi ke-3, diperoleh *Cluster* 1 (C1) beranggotakan 2 data, *Cluster* 2 (C2) beranggotakan 15 data dan *Cluster* 3 (C3) beranggotakan 7 data.

3.3 Perhitungan Minkowski Distance

Metode jarak *Minkowski Distance* menggunakan persamaan (2) untuk mencari nilai jarak data ke *centroid*. Tabel 6 menampilkan hasil keseluruhan jarak data ke *centroid* menggunakan *Minkowski Distance.*

Tabel 6. Hasil *Centroid Minkowski Distance*

No	Puskesmas	Jarak Data			
		C1	C2	C3	*CLUSTER*
1	Palaran	0.6976	2.1540	3.4348	C1
2	Bantuas	4.6851	2.1238	0.6597	C3
3	Bukuan	3.9638	1.4654	0.1860	C3

No	Puskesmas	Jarak Data			
		C1	C2	C3	*CLUSTER*
4	Mangkupalas	3.0861	0.6071	1.0918	C2
5	Baqa	3.3310	0.6506	1.4018	C2
6	Harapan Baru	2.6087	1.5229	2.7449	C2
7	Trauma Center	3.7616	1.2338	0.4098	C3
8	Loa Bakung	2.9101	0.5777	1.7359	C2
9	Karang Asam	2.4554	0.4050	1.8430	C2
10	Wonorejo	2.3572	0.6569	1.7103	C2
11	Juanda	2.9170	0.7720	1.5056	C2
12	Air Putih	2.1735	0.8669	2.1648	C2
13	Segiri	2.4571	0.2629	1.6771	C2
14	Pasundan	2.4291	0.6668	2.0900	C2
15	Sidomulyo	1.4323	4.0952	5.4533	C1
16	Sungai Kapih	3.9664	1.4755	0.2896	C3
17	Sambutan	2.9278	0.6292	1.1257	C2
18	Makroman	4.1575	1.7044	0.2545	C3
19	Bengkuring	3.1508	0.6375	1.1536	C2
20	Sempaja	2.8002	0.4105	1.4360	C2
21	Sungai Siring	4.2179	1.7070	0.2655	C3
22	Lempake	3.5889	1.2161	0.9380	C3
23	Remaja	2.6496	0.3507	1.4575	C2
24	Temindung	0.8485	1.9291	3.2816	C1

Hasil perhitungan metode *K-Means* dengan *Minkowski Distance* selama tiga iterasi, diperoleh *Cluster* 1 (C1) beranggotakan 3 data, *Cluster* 2 (C2) beranggotakan 14 data dan *Cluster* 3 (C3) beranggotakan 7 data.

3.4 Pengukuran Akurasi K-Means

Tahap berikutnya adalah melakukan pengujian akurasi ketiga metode jarak menggunakan metode *Sum of Squared Error* (SSE). SSE memberikan informasi *error* jarak data ke *centroid*. SSE digunakan untuk menguji hasil *clustering* yang dibuat sebanyak tiga *cluster* dan hasil *clustering* sebanyak dua *cluster*. Hasil perhitungan SSE dapat dilihat pada Tabel 7.

Tabel 7. Hasil SSE

Cluster	*Euclidean*	*Manhattan*	*Minkowski*
K=2	72.7043	122.0097	63.4208
K=3	57.7040	92.3633	50.3405

Dari Tabel 7, diperoleh informasi bahwa nilai SSE 3 *cluster* menunjukkan nilai yang lebih kecil dibandingkan nilai SSE sebanyak 2 *cluster*. Hal ini menunjukkan jumlah 3 *cluster* adalah yang paling ideal dan optimal untuk pengklasteran daerah rawan demam berdarah. Selain itu, nilai SSE yang dimiliki oleh metode jarak *Minkowski Distance* lebih kecil dibandingkan dengan dua metode jarak lainnya. Hal yang sama juga terlihat di nilai selisih *Minkowski Distance* antara 2 *cluster* dan 3 *cluster* sebesar 13.0803 yang artinya perhitungan jarak *Minkowski* adalah perhitungan jarak yang paling ideal untuk klusterisasi daerah rawan penyakit demam berdarah di kota Samarinda, karena memiliki selisih nilai lebih sedikit dibandingkan dua metode jarak lainnya yaitu, *Manhattan distance* dan *Euclidean distance*.

Setelah memperoleh hasil perhitungan jarak dan jumlah *cluster* yang ideal, maka model tersebut diterapkan untuk klusterisasi daerah rawan penyakit demam berdarah. Proses perhitungan *cluster*ing berhenti di iterasi ke-3. Hasil *cluster*ing algoritma *K-Means* pada iterasi terakhir ditunjukkan oleh Tabel 8.

Tabel 8. Hasil *Clustering*

No	Puskesmas	Jarak *Centroid*			Hasil Cluster		
		C1	C2	C3	C1	C2	C3
1	Palaran	0.6976	2.1540	3.4348	*		
2	Bantuas	4.6851	2.1238	0.6597			*
3	Bukuan	3.9638	1.4654	0.1860			*
4	Mangkupalas	3.0861	0.6071	1.0918		*	

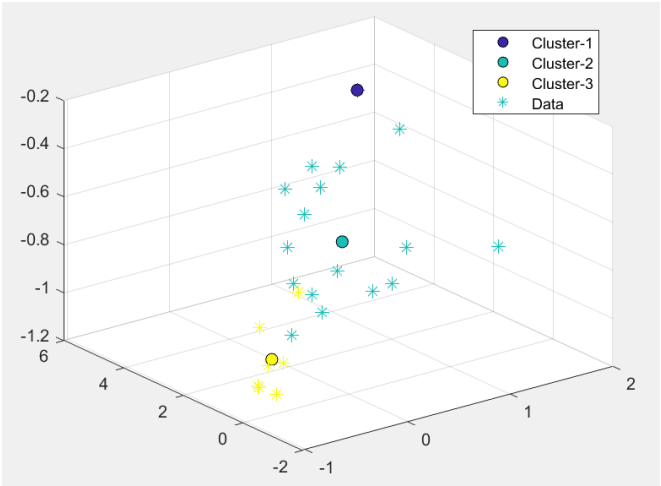
No	Puskesmas	Jarak *Centroid*			Hasil Cluster		
		C1	C2	C3	C1	C2	C3
5	Baqa	3.3310	0.6506	1.4018		*	
6	Harapan Baru	2.6087	1.5229	2.7449		*	
7	Trauma Center	3.7616	1.2338	0.4098			*
8	Loa Bakung	2.9101	0.5777	1.7359		*	
9	Karang Asam	2.4554	0.4050	1.8430		*	
10	Wonorejo	2.3572	0.6569	1.7103		*	
11	Juanda	2.9170	0.7720	1.5056		*	
12	Air Putih	2.1735	0.8669	2.1648		*	
13	Segiri	2.4571	0.2629	1.6771		*	
14	Pasundan	2.4291	0.6668	2.0900		*	
15	Sidomulyo	1.4323	4.0952	5.4533	*		
16	Sungai Kapih	3.9664	1.4755	0.2896			*
17	Sambutan	2.9278	0.6292	1.1257		*	
18	Makroman	4.1575	1.7044	0.2545			*
19	Bengkuring	3.1508	0.6375	1.1536		*	
20	Sempaja	2.8002	0.4105	1.4360		*	
21	Sungai Siring	4.2179	1.7070	0.2655			*
22	Lempake	3.5889	1.2161	0.9380			*
23	Remaja	2.6496	0.3507	1.4575		*	
24	Temindung	0.8485	1.9291	3.2816	*		

Berdasarkan hasil klusterisasi pada Tabel 8, terlihat bahwa wilayah palaran, sidomulyo dan temindung termasuk ke dalam *Cluster* 1 (C1). Daerah yang termasuk ke dalam *Cluster* 2 (C2) adalah Mangkupalas, Baqa, Harapan Baru, Loa Bakung, Karang Asam, Wonorejo, Juanda, Air Putih, Segiri, Pasundan, Sambutan, Bengkuring, Sempaja, dan Remaja. Daerah Bantuas, Bakuan, Trauma Center, Sungai Kapih, Makroman, Sungai Siring dan Lempake merupakan daerah yang termasuk ke dalam *Cluster* 3 (C3). Hasil klusterisasi juga menghasilkan *centroid* baru yang telah terbentuk. Tabel 9 menampilkan *centroid* dari iterasi terakhir.

Tabel 9. Nilai *Centroid* Akhir

Centroid	*Centroid*			
	2015	2016	2017	2018
C1	0.7837	3.6790	-0.5584	0.2621
C2	0.2855	0.9801	-0.6740	-0.1063
C3	-0.8197	-0.2520	-0.9453	-0.7393

Nilai *centroid* terakhir yang ditampilkan oleh Tabel 9 menunjukkan bahwa C3 merupakan nilai *centroid* terendah dari dua *cluster* lainnya. Selanjutnya, C1 memiliki nilai *centroid* paling tinggi dari dua *cluster* lainnya. Hasil perhitungan metode *K-Means* juga ditampilkan dalam bentuk grafik dengan alat bantu *Matlab* seperti yang terlihat pada Gambar 2.



Gambar 2. Grafik Hasil *Clustering K-Means*

Gambar 2 menampilkan data-data yang telah dikelompokkan dalam bentuk grafik plot. Titik berwarna biru tua merupakan *cluster* C1 dengan posisi diatas *cluster* C2 dan C3. *Cluster* C2 merupakan plot berwarna toska dengan posisi diantara *cluster* C1 dan C3, sedangkan titik plot berwarna kuning merupakan *cluster* tiga C3 yang posisinya terletak dibawah C2 dan C3. Korelasi yang muncul antara grafik dan hasil perhitungan adalah *cluster* C1 pada grafik merupakan *cluster* C1 pada proses perhitungan yang termasuk ke *cluster* tinggi. Lebih lanjut, *cluster* C2 pada grafik merupakan *cluster* yang sama pada proses perhitungan yaitu C2 dengan kategori sedang dan *cluster* C3 merupakan C3 di proses perhitungan yang termasuk *cluster* rendah. Berdasarkan grafik dan hasil perhitungan yang telah dilakukan, cluster C1 merupakan daerah rawan demam berdarah dengan kategori tinggi, C2 termasuk dalam kategori sedang dan C3 daerah kategori rendah. Hasil tersebut memberikan informasi bahwa daerah rawan demam berdarah yang termasuk kategori tinggi terdiri dari Palaran, Sidomulyo dan Temindung. Daerah Mangkupalas, Baqa, Harapan Baru, Loa Bakung, Karang Asam, Wonorejo, Juanda, Air Putih, Segiri, Pasundan, Sambutan, Bengkuring, Sempaja, dan Remaja termasuk ke dalam daerah rawan demam berdarah kategori sedang. Sedangkan Bantuas, Bakuan, Trauma Center, Sungai Kapih, Makroman, Sungai Siring dan Lempake merupakan wilayah rawan demam berdarah kategori rendah. Informasi ini dapat menjadi acuan bagi pemerintah kota untuk lebih memperhatikan dan memprioritaskan daerah Palaran, Sidomulyo dan Temindung dalam pencegahan dan penanganan kasus Demam Berdarah.

4. KESIMPULAN

Klusterisasi daerah rawan penyakit demam berdarah di Kota Samarinda menggunakan Algoritma *K-Means* menghasilkan kelompok tingkat kerawanan tinggi, sedang dan rendah. Perhitungan jarak menggunakan *Minkowski Distance* merupakan perhitungan jarak ideal berdasarkan hasil uji metode SSE karena menghasilkan nilai selisih SSE yang lebih kecil dibandingkan 2 pengukuran jarak lainnya. *Cluster* yang optimal untuk studi kasus ini sebanyak 3 *cluster* menggunakan metode SSE. Hal ini menandakan bahwa kategori ideal untuk mengelompokkan daerah penyebaran demam berdarah adalah sebanyak tiga kategori.

DAFTAR PUSTAKA

[1] D. Nuroniyah, S. Nitalana, P. J. N. Wati, S. I. M. Sari, and N. G. Prihartanti, "Penguatan Kapasitas Kader Resik (Remaja Sadar Jentik) Sebagai Pencegahan Primer Demam Berdarah Dengue di Daerah Rawan Banjir," *SNHRP*, pp. 1420–1425, 2022.

[2] N. A. Pascawati, S. Sahid, S. Sukismanto, and H. Yuningrum, "Faktor yang Berhubungan dengan Pola Pengelompokkan Kasus Demam Berdarah Dengue (DBD) di Temanggung, Jawa Tengah," *Balaba J. Litbang Pengendali. Penyakit Bersumber Binatang Banjarnegara*, vol. 18, no. 1, pp. 65–78, 2022, doi: 10.22435/blb.v18i1.5957.

[3] E. A. Idris and F. Zulaikha, "Hubungan Jenis Kelamin Terhadap Kejadian DHF pada Anak di TK RA AL Kamal 4 di Wilayah Bukuan Kota Samarinda," *Borneo Student Res.*, vol. 2, no. 3, pp. 1592–1598, 2021.

[4] V. Herlinda, D. Darwis, and D. Dartono, "Analisis Clustering Untuk Recredesialing Fasilitas Kesehatan Menggunakan Metode Fuzzy C-Means," *J. Teknol. dan Sist. Inf.*, vol. 2, no. 2, pp. 94–99, 2021.

[5] F. Hardiyanti, H. S. Tambunan, and I. S. Saragih, "Penerapan Metode K-Medoids Clustering Pada Penanganan Kasus Diare Di Indonesia," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 3, no. 1, pp. 598–603, 2019, doi: 10.30865/komik.v3i1.1666.

[6] W. Widayani and H. Harliana, "Perbandingan Algoritma K-Means dan SFCM Pada Pengelompokkan Rumah Tangga Miskin," *J. Sains dan Inform.*, vol. 6, no. 1, pp. 1–9,

2020, doi: 10.34128/jsi.v6i1.200.
[7] B. E. Adiana, I. Soesanti, and A. E. Permanasari, "Analisis Segmentasi Pelanggan Menggunakan Kombinasi RFM Model dan Teknik Clustering," *J. Terap. Teknol. Inf.*, vol. 2, no. 1, pp. 23–32, 2018, [Online]. Available: internal-pdf://100.141.155.94/76-Article Text-384-1-10-20180426.pdf.
[8] A. Saputra, B. Mulyawan, and T. Sutrisno, "Rekomendasi Lokasi Wisata Kuliner di Jakarta Menggunakan Metode K-Means Clustering dan Simple Additive Weighting," *J. Ilmu Komput. dan Sist. Inf.*, vol. 7, no. 1, pp. 14–21, 2019.
[9] T. Syahputra, J. Halim, and E. P. Sintho, "Penerapan Data Mining Dalam Menentukan Pilihan Jurusan Bidang Studi Sma Menggunakan Metode Clustering Dengan Teknik Single Linkage," *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. IV, no. 2, pp. 205–208, 2018.
[10] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," *J. Inform. J. Pengemb. IT*, vol. 4, no. 1, pp. 20–24, 2019, doi: 10.30591/jpit.v4i1.1253.
[11] S. Witanto, D. E. Ratnawati, and S. Anam, "Pengelompokan Fungsi Aktif Senyawa Data SMILES (Simplified Molecular Input Line Entry System) Menggunakan Metode K-Means Dengan Inisialisasi Pusat Klaster Menggunakan Metode Heuristic O (N LogN)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 1, pp. 702–707, 2019.
[12] Haviluddin *et al.*, "A Performance Comparison of Euclidean, Manhattan and Minkowski Distances in K-Means Clustering," in *2020 6th International Conference on Science in Information Technology: Embracing Industry 4.0: Towards Innovation in Disaster Management, ICSITech 2020*, 2020, pp. 184–188, doi: 10.1109/ICSITech49800.2020.9392053.
[13] R. D. Ramadhani and D. J. Ak, "Evaluasi K-Means dan K-Medoids pada Dataset Kecil," in *Seminar Nasional Informatika dan Aplikasinya*, 2017, no. September, pp. 20–24.
[14] N. L. Anggreini and S. Tresnawati, "Komparasi Algoritma K-Means Dan K-Medoids Untuk Menangani Strategi Promosi Di Politeknik TEDC Bandung," *J. TEDC*, vol. 14, no. 2, pp. 120–127, 2020.
[15] F. Harahap, "Perbandingan Algoritma K-Means dan K-Medoids Untuk Clustering Kelas Siswa Tunagrahita," *TIN Terap. Inform. Nusant.*, vol. 2, no. 4, pp. 191–197, 2021.
[16] Y. H. Susanti and E. Widodo, "Perbandingan K-Means dan K-Medoids Clustering terhadap Kelayakan Puskesmas di DIY Tahun 2015," in *Prosiding SI MaNIs (Seminar Nasional Integrasi Matematika dan Nilai Islami)*, 2017, vol. 1, no. 1, pp. 116–122.
[17] R. Adha, N. Nurhaliza, U. Soleha, P. Studi, S. Informasi, and F. Sains, "Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia," *SITEKIN J. Sains, Teknol. dan Ind.*, vol. 18, no. 2, pp. 206–211, 2021.
[18] Athifaturrofifah, R. Goejantoro, and D. Yuniarti, "Perbandingan Pengelompokan K-Means dan K-Medoids Pada Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Studi Kasus : Data Titik Panas Di Indonesia Pada 28 April 2018)," *J. EKSPONENSIAL*, vol. 10, no. 2, pp. 143–152, 2019.
[19] H. Sulastri and A. I. Gufroni, "Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia," *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 2, pp. 299–305, 2017.
[20] M. Y. Rizki, S. Maysaroh, and A. P. Windarto, "Implementasi K-Means Clushtering dalam Mengelompokkan Minat Membaca Penduduk Menurut Wilayah," *JUST IT J. Sist. Informasi, Teknol. Inf. dan Komput.*, vol. 11, no. 2, pp. 41–49, 2021, doi: 10.24853/justit.11.2.41-49.
[21] N. T. Hartanti, "Measuring Students ' Level of Understanding in Programming Courses with the K-Means Clustering Algorithm," *SISFOTENIKA*, vol. 12, no. 1, pp. 62–73, 2022.
[22] N. Hasanah, M. Ugiarto, and N. Puspitasari, "Sistem Pengelompokan Curah Hujan Menggunakan Metode K-Means di Wilayah Kalimantan Timur," in *Prosiding Seminar*

Nasional Ilmu Komputer dan Teknologi Informasi, 2017, vol. 2, no. 2, pp. 122–126.

[23] A. Septiarini, I. A. Thaher, and N. Puspitasari, "Pengelompokan Kualitas Kinerja Pegawai Menggunakan Metode K-Means Clustering," *Komputika J. Sist. Komput.*, vol. 11, no. 2, pp. 131–141, 2022, doi: 10.34010/komputika.v11i2.5518.

[24] S. Butsianto and N. T. Mayangwulan, "Penerapan Data Mining untuk Prediksi Penjualan Mobil Menggunakan Metode K-Means Clustering," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 3, no. 3, pp. 187–201, 2020, doi: 10.32672/jnkti.v3i3.2428.

[25] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.