

PERBANDINGAN *SINGLE LINKAGE* DAN K-MEDOIDS PADA DATA PENGHASILAN DAN DEMOGRAFI AREA KOMUNITAS CHICAGO

Salsadilla Azizi Firda¹, Haikal Agung Widiyanto², Regina Adelisa³, Etis Sunandi^{4*}

^{1,2}*Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Bengkulu
Jalan W.R. Supratman, Kandang Limun, Bengkulu 38371A, Indonesia*

Korespondensi; Etis Sunandi, Email: esunandi@unib.ac.id

Abstrak

Penelitian ini menganalisis metode *Hierarchical Clustering Single Linkage* dan K-Medoids dalam mengelompokkan data penghasilan dan demografi komunitas di Chicago. Dengan menggunakan data sekunder dari website Kaggle, penelitian ini mengelompokkan 77 area komunitas berdasarkan kesamaan karakteristik sosial dan ekonomi. Hasil analisis dengan metode *Hierarchical Clustering Single Linkage* menunjukkan bahwa terdapat satu area komunitas yang terpisah dalam kluster kedua, sementara sisanya tergabung dalam satu kluster utama. Sebaliknya, metode K-Medoids menghasilkan dua kluster yang lebih seimbang dalam distribusi datanya. Berdasarkan visualisasi clustering, metode K-Medoids dianggap lebih baik karena mampu membagi data secara lebih seimbang. Namun, jika ditinjau dari nilai Silhouette Score dan Dunn Index, metode *Hierarchical Clustering Single Linkage* lebih unggul karena memiliki nilai yang lebih tinggi, mengindikasikan pemisahan kluster yang lebih jelas. Dengan demikian, pemilihan metode terbaik bergantung pada tujuan analisis, di mana K-Medoids lebih sesuai untuk interpretasi distribusi data yang lebih merata, sedangkan *Hierarchical Clustering Single Linkage* lebih optimal dalam pemisahan kluster yang jelas.

Kata Kunci: Clustering, Hierarchical Clustering, K-Medoids, Silhouette Score, Dunn Index, Penghasilan, Demografi.

Abstract

This study analyzes the Hierarchical Clustering Single Linkage and K-Medoids methods in clustering income and demographic data of communities in Chicago. Using secondary data from the Kaggle website, this study clusters 77 community areas based on similarities in social and economic characteristics. The analysis using the Hierarchical Clustering Single Linkage method reveals that one community area is isolated in the second cluster, while the rest are grouped into a single main cluster. In contrast, the K-Medoids method produces two clusters with a more balanced distribution. Based on clustering visualization, the K-Medoids method is considered superior as it provides a more evenly distributed classification. However, when evaluated using the Silhouette Score and Dunn Index, the Hierarchical Clustering Single Linkage method outperforms K-Medoids due to its higher values, indicating clearer cluster separation. Thus, the choice of the best method depends on the analytical objective, where K-Medoids is more suitable for interpreting a more balanced data distribution, while Hierarchical Clustering Single Linkage is optimal for achieving distinct cluster separation.

Keywords: Clustering, Hierarchical Clustering, K-Medoids, Silhouette Score, Dunn Index, Income, Demographics.

Pendahuluan

Pertumbuhan kota yang pesat sering kali diiringi dengan ketimpangan sosial-ekonomi antar wilayah. Chicago, sebagai salah satu kota metropolitan terbesar di Amerika Serikat, memiliki beragam karakteristik demografis dan ekonomi di setiap wilayahnya. Perbedaan tingkat pendapatan, kepadatan penduduk, serta faktor sosial lainnya dapat mencerminkan disparitas yang signifikan dalam akses terhadap layanan publik dan kualitas hidup masyarakat. Ketimpangan ini dapat terlihat dalam berbagai aspek, seperti kesenjangan dalam akses pendidikan, layanan kesehatan, peluang kerja, serta tingkat

kejahatan di berbagai komunitas. Dengan menganalisis penghasilan dan demografi area komunitas Chicago ini, pemerintah, akademisi, dan perencana kebijakan dapat mengidentifikasi wilayah yang membutuhkan perhatian lebih, merancang kebijakan pembangunan yang lebih inklusif, serta meningkatkan akses terhadap layanan publik bagi masyarakat. Selain itu, data penghasilan dan demografi juga memainkan peran penting dalam perencanaan tata kota, analisis pasar tenaga kerja, serta pengembangan strategi ekonomi berbasis komunitas.

Saat ini, konsep data mining semakin dikenal sebagai *tools* penting dalam manajemen informasi karena jumlah informasi yang semakin besar jumlahnya. Data mining sendiri sering disebut sebagai *knowledge discovery in database* (KDD) adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola hubungan dalam set data berukuran besar [1]. Salah satu teknik yang dikenal dalam data mining yaitu clustering. Pengertian clustering keilmuan dalam data mining adalah pengelompokan sejumlah data atau objek ke dalam cluster (*group*) sehingga setiap dalam cluster tersebut akan berisi data yang semirip mungkin dan berbeda dengan objek dalam cluster yang lainnya [2]. 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 KMedoids [3].

Agar dapat menentukan tingkatan prioritas dalam mengidentifikasi wilayah yang membutuhkan perhatian lebih, merancang kebijakan pembangunan yang lebih inklusif, serta meningkatkan akses terhadap layanan publik bagi Masyarakat Chicago, maka perlu dilakukan pengelompokan terhadap data penghasilan dan demografi area komunitas Chicago. Pengelompokan ini dilakukan berdasarkan kemiripan karakteristik wilayahnya. Terdapat beberapa penelitian serupa terkait pengelompokan wilayah berdasarkan indikator tertentu. Thamrin dan Wijayanto (2021) melakukan penelitian untuk menentukan metode clustering terbaik dalam mengelompokkan kabupaten/kota di Pulau Jawa menggunakan metode hard clustering dan soft clustering berdasarkan tingkat kesejahteraan. Berdasarkan penelitian tersebut diperoleh hasil bahwa metode terbaik dalam pengelompokan kabupaten/kota di Pulau Jawa berdasarkan tingkat kesejahteraan yaitu menggunakan Agglomerative Ward Linkage dengan tiga cluster. Selanjutnya, Luthfi dan Wijayanto (2021) juga melakukan penelitian yang serupa.

Penelitian kali ini melakukan pengelompokan pengelompokan terhadap 77 area komunitas di Chicago menggunakan dua metode clustering, yaitu Hierarchical Clustering dan K-Medoids. Hasil dari kedua metode tersebut kemudian dibandingkan untuk menentukan metode yang paling optimal dalam mengelompokkan area komunitas berdasarkan karakteristik sosial, ekonomi, dan demografisnya. Melalui proses clustering ini diharapkan pemerintah dapat menentukan prioritas dan mengambil kebijakan yang tepat mengenai area komunitas di Chicago mana yang perlu ditindak lanjuti terlebih dahulu.

Namun, *research gap* dalam studi ini muncul karena penelitian sebelumnya yang menganalisis pengelompokan wilayah di Chicago masih terbatas, baik dalam hal metode yang digunakan maupun variabel yang dianalisis. Sebagian besar studi terdahulu berfokus pada Spatial Regression atau Hotspot Analysis untuk mendeteksi ketimpangan sosial, sementara penerapan metode non-hierarkis seperti K-Medoids dalam konteks *community area clustering* di Chicago masih jarang dilakukan. Oleh karena itu, penelitian ini memberikan kontribusi baru dengan membandingkan dua pendekatan clustering berbeda untuk memperoleh hasil pengelompokan yang lebih representatif dan dapat mendukung pengambilan kebijakan berbasis data.

Landasan Teori

Clustering

Clustering adalah metode penganalisis data, yang sering dimasukkan sebagai salah satu metode data mining, yang bertujuan adalah untuk mengelompokkan data dengan karakteristik yang sama ke suatu ‘wilayah’ yang sama dan data dengan karakteristik yang berbeda ‘wilayah’ yang lain [4]. Analisis cluster membuat pengelompokan objek berdasarkan jarak antara pasangan objek. Jarak merupakan ukuran yang digunakan untuk mengukur kemiripan dari suatu objek. Pada proses pemilihan ini, dapat dipilih salah satu dari beberapa jarak yang biasa digunakan salah satunya adalah jarak Euclidean. Jarak Euclidean adalah akar dari jumlah kuadrat perbedaan/deviasi di dalam nilai untuk setiap variable. Jarak euclidean antara cluster objek ke-i dan cluster objek ke-g dari p variabel didefinisikan sebagai berikut [5]:

$$d(X_i, X_g) = \left[\sum_{j=1}^p (X_{ij} - X_{gj})^2 \right]^{1/2} \quad (1)$$

Metode Hierarchical Clustering (Single Linkage)

Hierarchical clustering merupakan salah satu metode clustering yang didasarkan pada struktur seperti dendogram, yaitu membagi atau menggabungkan data seperti cabang-cabang pohon dengan mengelompokkan dua atau lebih data yang memiliki kesamaan terdekat [6]. *Hierarchical clustering* dibagi menjadi dua yaitu *Agglomeratif Clustering* dan *Difisive Clustering*. *Agglomeratif Clustering* mengelompokkan data dengan pendekatan bawah atas (*bottom up*), sedangkan *Difisive Clustering* menggunakan pendekatan atas bawah (*top-bottom*) [7]. Salah satu pengelompokan data *Agglomeratif Clustering* adalah *single linkage*. Algoritma pengelompokan *single linkage* diawali dengan memilih jarak terkecil dalam matriks $D = \{d_{ij}\}$, kemudian menggabungkan objek yang bersesuaian misalnya U dan V untuk memperoleh cluster (UV) . Langkah berikutnya adalah mencari jarak antara (UV) dengan cluster lainnya, misalnya W sehingga dapat dituliskan sebagai berikut:

$$d_{(UV)W} = \min(d_{UW}, d_{VW}) \quad (2)$$

Dengan d_{UW} adalah jarrah tetangga terdekat dari cluster U dan W serta d_{VW} adalah jarak tetangga terdekat dari cluster V dan W [8].

Metode K-Medoid

K-Medoids merupakan suatu algoritma yang digunakan untuk menemukan *medoids* didalam sebuah kelompok (cluster) yang merupakan titik pusat dari suatu kelompok (cluster) [9]. Metode K-Medoids merupakan metode clustering yang berkaitan dengan metode KMedoids dan Medoidshift. Algoritma K-Medoids yang biasa dikenal dengan PAM (Partition Around Medoids) adalah algoritma yang mengimplementasikan objek yaitu medoids sebagai perwakilan di setiap cluster [10]. Algoritma K-Medoids tidak menetapkan rata-rata objek dalam suatu cluster sebagai titik referensi, melainkan menggunakan medoid (median), yaitu objek yang terletak paling di tengah cluster [11]. Strategi dasar dari algoritma k-medoids adalah menemukan cluster dari objek k ke n dengan terlebih dahulu menemukan objek asli (medoid) secara acak sebagai perwakilan dari setiap cluster. Setiap objek yang tersisa dikelompokkan dengan medoid yang paling mirip [12].

Silhouette Coefficient

Indeks silhoutte 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 [13].

$$s(i) = \frac{a(i) - b(i)}{\max(a(i), b(i))} \quad (3)$$

Dunn Index

Indeks dunn merupakan rasio jarak terkecil antara observasi yang pada cluster yang berbeda dengan jarak intra cluster terbesar. Indeks dunn berada pada rentang antara 0 sampai tak hingga (∞). Semakin besar nilai indeks dunn pada suatu cluster, semakin baik hasil clustering-nya. Rumus yang digunakan untuk menghitung indeks dunn yaitu [14]:

$$D(C) = \frac{\min_{C_k, C_l \in C, C_k \neq C_l} \left(\min_{i \in C_k, j \in C_l} \text{dist}(i, j) \right)}{\max_{C_m \in C} (\text{diam}(C_m))} \quad (4)$$

dimana $D(C)$ adalah indeks dunn; dan $\text{diam}(C_m)$ adalah jarak maksimum antar observasi dalam cluster C_m .

Bahan dan Metode

Data yang digunakan yaitu data sekunder yang diperoleh dari publikasi website Kaggle :<https://www.kaggle.com/datasets/aniket0712/acs-5-year-data-by-community-area> yaitu data penghasilan dan demografi area komunitas Chicago tahun 2023. Jumlah data yang digunakan berjumlah 78 pengamatan dengan banyak variabel sejumlah 28 variabel sebagai berikut:

Tabel 1 Variabel Data

Variabel	Keterangan	Variabel	Keterangan
Area komunitas (Nama wilayah komunitas di Chicago)	Label	Perempuan usia 25-34 tahun	X_{14}
Pendapatan rendah (< \$25,000 per tahun)	X_1	Perempuan usia 35-49 tahun	X_{15}
Pendapatan menengah bawah (\$25,000-\$49,999 per tahun)	X_2	Perempuan usia 50-64 tahun	X_{16}
Pendapatan menengah (\$50,000-\$74,999 per tahun)	X_3	Perempuan usia ≥ 65 tahun	X_{17}
Pendapatan menengah atas (\$75,000-\$125,000 per tahun)	X_4	Total populasi	X_{18}
Pendapatan tinggi (\geq \$125,000 per tahun)	X_5	Kulit putih (<i>White</i>)	X_{19}
Laki-laki usia 0-17 tahun	X_6	Kulit hitam atau Afrika-Amerika	X_{20}
Laki-laki usia 18-24 tahun	X_7	Penduduk Indian atau Alaska Asli	X_{21}
Laki-laki usia 25-34 tahun	X_8	Penduduk Asia	X_{22}
Laki-laki usia 35-49 tahun	X_9	Penduduk Kepulauan Pasifik	X_{23}
Laki-laki usia 50-64 tahun	X_{10}	Ras lainnya	X_{24}
Laki-laki usia ≥ 65 tahun	X_{11}	Multiras	X_{25}
Perempuan usia 0-17 tahun	X_{12}	Kulit putih bukan Hispanik/Latin	X_{26}

Perempuan usia 18-24 tahun	X_{13}	Hispanik atau Latin	X_{28}
----------------------------	----------	---------------------	----------

Selanjutnya tahapan-tahapan dalam analisis *Hierarchical Clustering (single linkage)* dan K-Medoids pada data penghasilan dan demografi area komunitas Chicago pada penelitian ini yaitu sebagai berikut:

1. Melakukan analisis statistika secara deskriptif pada data penelitian.
2. Menghitung jarak antar observasi dengan menggunakan metrik Euclidean untuk mengukur kedekatan antar wilayah dalam analisis Hierarchical Clustering dan K-Medoids.
3. Menerapkan metode *Hierarchical Clustering* dengan pendekatan *Single Linkage* dengan cara membentuk matriks jarak antar data, membangun dendrogram, dan menentukan jumlah kluster optimal menggunakan metode Silhouette.
4. Menerapkan metode K-Medoids dengan menentukan jumlah kluster k yang optimal menggunakan Silhouette Score, memilih medoids sebagai pusat kluster, mengelompokkan data berdasarkan kedekatan dengan medoids, serta mengoptimalkan pemilihan pusat kluster hingga hasil konvergen.
5. Membuat plot visualisasi hasil metode *Hierarchical Clustering* dengan pendekatan *Single Linkage* dan K-Medoids.
6. Menentukan metode terbaik berdasarkan Silhouette Score dan Dunn Index
7. Menarik Kesimpulan

Hasil dan Pembahasan

Statistik Deskriptif Data

Sebagai tahap awal, dilakukan analisis secara deskriptif terhadap masing-masing variable yang digunakan dalam penelitian. Hal ini bertujuan untuk menggambarkan karakteristik data secara ringkas agar didapatkan pemahaman yang lebih baik mengenai data tersebut.

Tabel 2 Statistik Deskriptif Data

Summary	Minimum	Standar dev	Median	Mean	Maximum
Pendapatan rendah (< \$25,000 per tahun)	31	822,99592	797	987	5506
Pendapatan menengah bawah (\$25,000-\$49,999 per tahun)	104	9930,9139	1082	1266	5084
Pendapatan menengah (\$50,000-\$74,999 per tahun)	36	743,10818	860	1006	3600
Pendapatan menengah atas (\$75,000-\$125,000 per tahun)	61	974,68.595	1295	1516	4588
Pendapatan tinggi (≥\$125,000 per tahun)	34	2565,84199	1609	2567	12236
Laki-laki usia 0-17 tahun	225	2274,59478	3013	3499	12399
Laki-laki usia 18-24 tahun	82	1245,05532	1306	1617	5510
Laki-laki usia 25-34 tahun	20	3545,97462	2029	3357	16967
Laki-laki usia 35-49 tahun	87	2642,58713	2605	3522	11131
Laki-laki usia 50-64 tahun	134	1779,07426	2476	2892	8110
Laki-laki usia ≥ 65 tahun	137	1235,46300	1593	1896	6401
Perempuan usia 0-17 tahun	189	2296,21662	2770	3461	12163
Perempuan usia 18-24 tahun	63	1374,75631	1327	1735	6195

Perempuan usia 25-34 tahun	110	3401,92424	2250	3347	16764
Perempuan usia 35-49 tahun	194	2301,38657	2957	3488	10092
Perempuan usia 50-64 tahun	269	1759,66499	2507	2943	10040
Perempuan usia ≥ 65 tahun	280	1586,58222	2325	2630	8811
Total populasi	2186	2248,359920	29052	34385	100278
Kulit putih (<i>White</i>)	27	15913,45394	10788	14503	71631
Kulit hitam atau Afrika-Amerika	10	12622,49114	5684	9903	73602
Penduduk Indian atau Alaska Asli	0	290,95015	123	254,3	10044
Penduduk Asia	0	3441,44.031	679	2357	16696
Penduduk Kepulauan Pasifik	0	34,42925	0	17,91	171
Ras lainnya	0	5171,59055	1855	3973	25743
Multiras	143	3285,63457	2024	3377	14070
Kulit putih bukan Hispanik/Latin	18	14842,60307	3980	11114	67784
Hispanik atau Latin	48	12238,11978	4707	10170	57511

Pada tabel 1 hasil analisis statistik deskriptif pada data penghasilan dan demografi area komunitas Chicago menunjukkan bahwa distribusi pendapatan dan karakteristik populasi memiliki tingkat variasi yang tinggi. Nilai rata-rata pada beberapa kelompok populasi menunjukkan kecenderungan tertentu, tetapi standar deviasi yang besar pada beberapa kategori mengindikasikan adanya perbedaan yang signifikan dalam distribusi pendapatan dan ukuran populasi. Selain itu, terdapat nilai ekstrem dalam beberapa kategori, terutama dalam jumlah populasi dan distribusi kelompok ras, yang menunjukkan adanya area dengan populasi yang sangat besar atau sangat kecil. Penyebaran data ini menjadi faktor penting dalam penerapan metode *Hierarchical Clustering Single Linkage* dan K-Medoids, karena variasi dalam distribusi penghasilan dan karakteristik demografi dapat memengaruhi pola pengelompokan komunitas di Chicago.

Metode *Hierarchical Clustering*

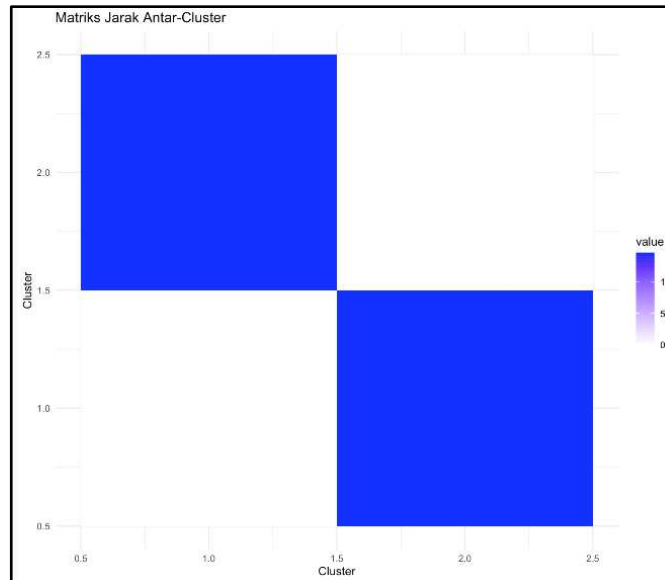
Pada penelitian ini akan dilakukan analisis clustering dengan menerapkan metode *Hierarchical Clustering* dengan pendekatan *Single Linkage*. Dalam melakukan *Hierarchical Clustering* dengan pendekatan *Single Linkage* diperlukan menghitung matriks jarak menggunakan metode Euclidean. Metode ini dipilih karena mampu mengukur kedekatan antar data dalam ruang multidimensi dengan menghitung jarak lurus antara dua titik. Dengan demikian, matriks jarak yang dihasilkan akan memberikan gambaran mengenai seberapa mirip atau berbeda setiap data dalam dataset yang digunakan. Adapun hasil perhitungan matriks jarak pada data kali ini adalah sebagai berikut:

Tabel 3 Matriks Jarak Antar-Cluster

	1	1
1	0.0000	1.466.934
2	1.466.934	0.0000

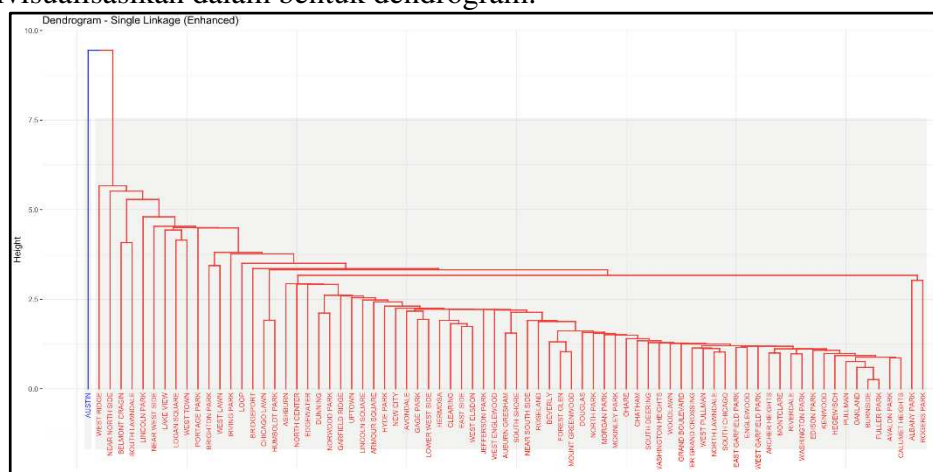
Pada tabel 3 dapat diketahui bahwa setiap data memiliki jarak nol terhadap dirinya sendiri, sebagaimana terlihat pada nilai diagonal utama matriks. Hal ini menunjukkan bahwa tidak ada perbedaan antara suatu data dengan dirinya sendiri. Selain itu, jarak antara data 1 dan data 2 bernilai 1.466.934, yang

mengindikasikan tingkat perbedaan atau ketidaksamaan antara kedua data tersebut. Semakin besar nilai jarak Euclidean, semakin jauh hubungan antara dua data, sehingga dalam kasus ini dapat disimpulkan bahwa data 1 dan data 2 memiliki perbedaan yang cukup signifikan. Selain itu, matriks ini bersifat simetris, yang berarti bahwa jarak antara data 1 ke data 2 sama dengan jarak antara data 2 ke data 1. Hasil matriks jarak ini dapat menjadi dasar dalam proses clustering atau analisis lebih lanjut untuk memahami pola kedekatan antar data dalam dataset.



Gambar 1. Matriks Jarak Antar-Cluster

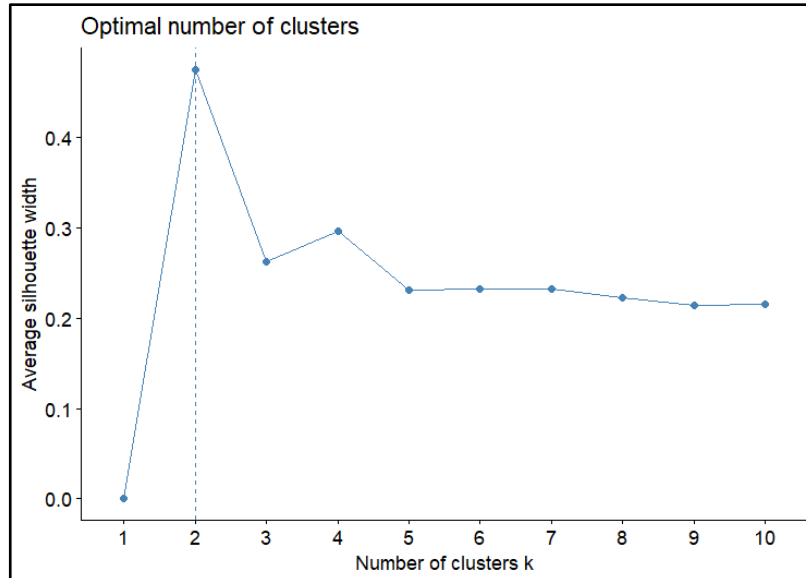
Setelah menghitung matriks jarak antar-cluster, langkah selanjutnya adalah melakukan clustering menggunakan metode hirarki. Dalam penelitian ini, metode yang digunakan adalah *Single Linkage*, yang merupakan salah satu teknik dalam *hierarchical clustering*. Metode ini mengelompokkan data dengan menghubungkan dua kluster terdekat berdasarkan jarak minimum antar titik. Dengan pendekatan ini, proses pembentukan kluster dilakukan secara bertahap hingga seluruh data tergabung dalam satu hierarki yang dapat divisualisasikan dalam bentuk dendrogram.



Gambar 2. Dendrogram Single Linkage

Pada gambar 2 menunjukkan bagaimana data dikelompokkan berdasarkan kesamaannya menggunakan metode *Single Linkage*. Semakin rendah garis penggabungan, semakin mirip objek-objek tersebut. Sebaliknya, garis tinggi menunjukkan bahwa objek lebih berbeda sebelum akhirnya dikelompokkan. Dari grafik ini, kita bisa menentukan jumlah kluster optimal dengan memotong pada titik tertentu, sehingga data terbagi ke dalam kelompok yang lebih seragam.

Setelah model Clustering Hirarki dibuat dengan metode *Single Linkage*, langkah selanjutnya adalah menentukan jumlah klaster yang optimal dan membagi data sesuai hasil analisis. Dendrogram yang dihasilkan kemudian dipotong untuk membentuk dua klaster. Hasil klasterisasi ini ditambahkan ke dalam dataset asli agar setiap data memiliki label klaster. Dengan begitu, data dapat lebih mudah dianalisis dan divisualisasikan untuk memahami karakteristik masing-masing kelompok.



Gambar 3. Cluster Optimal Menggunakan Metode Silhouette

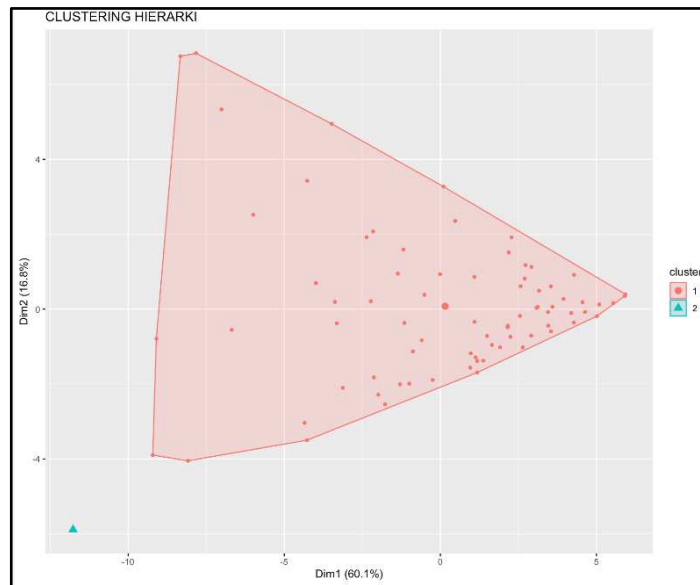
Gambar 3 menunjukkan bahwa jumlah klaster optimal adalah 2, dengan nilai *Silhouette Score* tertinggi. Ini menandakan bahwa data paling baik dipisahkan menjadi dua kelompok, di mana objek dalam klaster lebih mirip satu sama lain dibandingkan dengan klaster lain. Setelah $k = 2$, nilai menurun, menunjukkan bahwa menambah jumlah klaster tidak meningkatkan kualitas pemisahan.

Berdasarkan jumlah cluster optimal silhouette score diperoleh nilai k sebesar dua. Dengan menggunakan jumlah cluster sebanyak dua didapatkan bahwa pada cluster dua hanya terdapat satu buah daerah are komunitas di Chicago, sementara yang lainnya berada pada cluster satu sebagai berikut:

Tabel 4 Hasil *Hierarchical Clustering Single Linkage*

Cluster	Anggota
1	Albany Park, Ashburn, Avondale, Belmont Cragin, Bridgeport, Brighton Park, Chicago Lawn, Dunning, Edgewater, Gage Park, Garfield Ridge, Humboldt Park, Irving Park, Lake View, Lincoln Park, Lincoln Square, Logan Square, Loop, Lower West Side, Near North Side, Near West Side, New City, North Center, Norwood Park, Portage Park, Rogers Park, South Lawndale, Uptown, West Lawn, West Ridge, West Town, Archer Heights, Armour Square, Auburn Gresham, Avalon Park, Beverly, Burnside, Calumet Heights, Chatham, Clearing, Douglas, East Garfield Park, East Side, Edison Park, Englewood, Forest Glen, Fuller Park, Grand Boulevard, Greater Grand Crossing, Hegewisch, Hermosa, Hyde Park, Jefferson Park, Kenwood, Mckinley Park, Montclare, Morgan Park, Mount Greenwood, Near South Side, North Lawndale, North Park, Oakland, Ohare, Pullman, Riverdale, Roseland, South Chicago, South Deering, South Shore, Washington Heights, Washington Park, West Elsdon, West Englewood, West Garfield Park, West Pullman, Woodlawn
2	Austin

Dari hasil *Hierarchical Clustering Single Linkage* Tabel 4 di atas dapat di visualisasikan untuk untuk memahami pola distribusi area komunitas di Chicago adalah sebagai berikut:

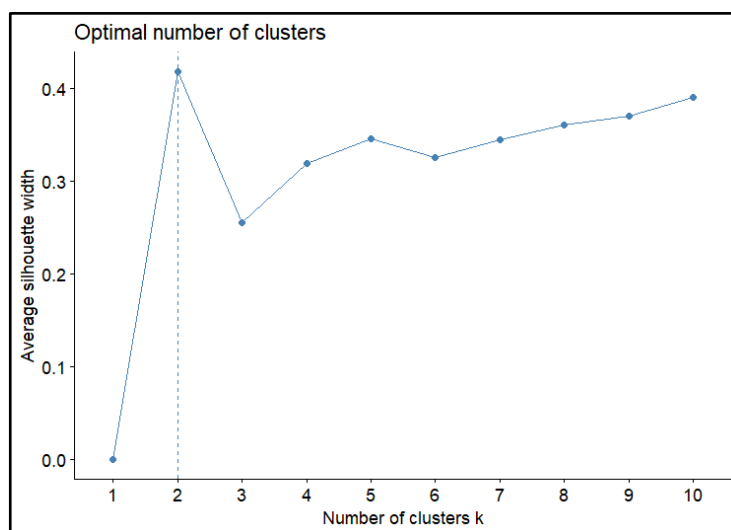


Gambar 4. Visualisasi *Hierarchical Clustering*

Plot hasil *Hierarchical Clustering* di atas menunjukkan dua kluster yang terbentuk. Kluster pertama (warna merah) memiliki distribusi yang lebih luas dan mencakup sebagian besar data, sedangkan kluster kedua (warna biru) hanya terdiri dari satu titik yang jauh terpisah dari kluster utama. Hal ini mengindikasikan bahwa terdapat satu komunitas yang sangat berbeda karakteristiknya dibandingkan dengan yang lain, kemungkinan merupakan *outlier* atau wilayah dengan kondisi sosial-ekonomi yang ekstrem.

Metode K-Medoids

Pada penelitian ini akan dilakukan analisis clustering dengan menerapkan metode K-Medoids. Tahapan pertama yang dilakukan adalah menentukan jumlah cluster optimum menggunakan metode silhouette.



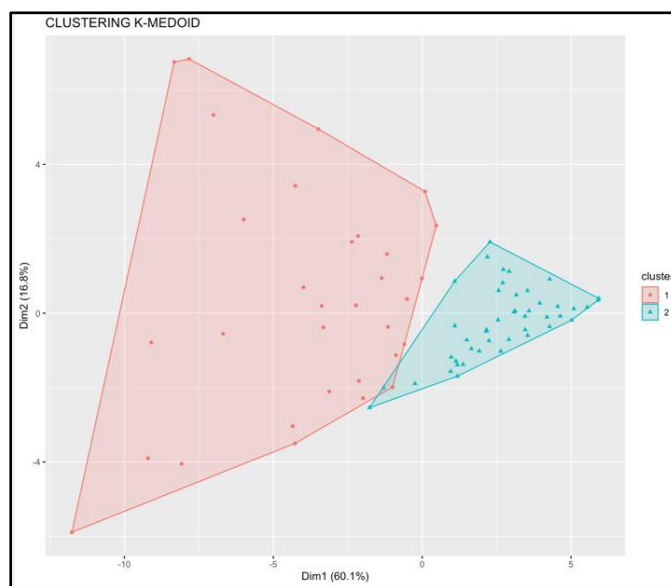
Gambar 5. Cluster Optimal Metode K-Medoids

Berdasarkan plot dari metode silhouette terlihat bahwa jumlah cluster optimal yang terbentuk yaitu sebanyak dua cluster. Dengan menggunakan cluster optimal sebanyak dua cluster, sehingga didapatkan hasil clustering menggunakan metode K-Medoids sebagai berikut:

Tabel 5. Hasil Clustering K-Medoid

Cluster	Anggota
1	Albany Park, Ashburn, Austin, Avondale, Belmont Cragin, Bridgeport, Brighton Park, Chicago Lawn, Dunning, Edgewater, Gage Park, Garfield Ridge, Humboldt Park, Irving Park, Lake View, Lincoln Park, Lincoln Square, Logan Square, Loop, Lower West Side, Near North Side, Near West Side, New City, North Center, Norwood Park, Portage Park, Rogers Park, South Lawndale, Uptown, West Lawn, West Ridge, West Town
2	Archer Heights, Armour Square, Auburn Gresham, Avalon Park, Beverly, Burnside, Calumet Heights, Chatham, Clearing, Douglas, East Garfield Park, East Side, Edison Park, Englewood, Forest Glen, Fuller Park, Grand Boulevard, Greater Grand Crossing, Hegewisch, Hermosa, Hyde Park, Jefferson Park, Kenwood, Mckinley Park, Montclare, Morgan Park, Mount Greenwood, Near South Side, North Lawndale, North Park, Oakland, Ohare, Pullman, Riverdale, Roseland, South Chicago, South Deering, South Shore, Washington Heights, Washington Park, West Elsdon, West Englewood, West Garfield Park, West Pullman, Woodlawn

Berdasarkan hasil clustering K-Medoid pada Tabel 5 di atas dapat dilakukan visualisasi clustering nya sebagai berikut:



Gambar 6. Visualisasi Clustering K-Medoid

Pada gambar 6, hasil clustering menggunakan metode K-Medoids dari visualisasi tersebut, terlihat bahwa klaster 1 memiliki cakupan wilayah yang lebih luas dibandingkan klaster 2, yang memiliki distribusi data yang lebih terkonsentrasi. Klaster 1 mencakup komunitas dengan karakteristik ekonomi dan demografi yang lebih beragam, sementara klaster 2 cenderung lebih homogen dalam aspek-aspek tersebut.

Evaluasi Model Terbaik

Dari hasil kedua metode clustering yaitu *Hierarchical Clustering* serta K-Medoid dapat dilakukan evaluasi model terbaik sebagai berikut:

Tabel 5 Hasil Evaluasi Clustering

Metode	<i>Silhouette Score</i>	<i>Dunn Index</i>
K-Medoids	1.415584	0.1089859
Hierarki	1.987013	0.4981542

Berdasarkan hasil evaluasi clustering Tabel 5 di atas, *Hierarchical Clustering* memiliki kinerja yang lebih baik dibandingkan K-Medoids. Hal ini ditunjukkan oleh *Silhouette Score* yang lebih tinggi (1.987013 vs. 1.415584), mengindikasikan bahwa objek dalam satu cluster lebih homogen dan terpisah dengan baik. Selain itu, *Dunn Index* pada *Hierarchical Clustering* juga lebih besar (0.4981542 vs. 0.1089859), menunjukkan bahwa jarak antar-cluster lebih optimal. Dengan demikian, Hierarchical Clustering lebih efektif dalam membentuk cluster yang terpisah dengan baik dibandingkan K-Medoids.

Kesimpulan

Penelitian ini menganalisis metode *Hierarchical Clustering Single Linkage* dan K-Medoids dalam mengelompokkan data penghasilan dan demografi komunitas di Chicago. Dengan menggunakan metode *Hierarchical Clustering Single Linkage* di dapatkan bahwa terdapat hanya ada satu daerah area komunitas yang berada pada cluster ke dua, dan sisanya berada pada area komunitas yang lain di Chicago. Sementara ketika menggunakan metode K-Medoids di dapatkan bahwa data terbagi lebih jelas menjadi dua cluster yang lebih seimbang. Berdasarkan analisis statistika metode *Hierarchical Clustering Single Linkage* dan K-Medoids dalam mengelompokkan data penghasilan dan demografi komunitas di Chicago dapat disimpulkan bahwa berdasarkan visualisasi clustering menggunakan scatter plot metode yang terbaik untuk digunakan adalah metode K-Medoids, karena pada K-Medoids, data terbagi lebih jelas menjadi dua cluster yang lebih seimbang. Sementara di tinjau dari nilai *Silhouette Score* dan *Dunn Index* diperoleh metode terbaik yaitu metode *Hierarchical Clustering Single Linkage*, sebagaimana dibuktikan oleh nilai yang lebih tinggi.

Referensi

- [1] S. B. "Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis, Graha Ilmu"," vol. 01, no. 01, p. 55511, 2007, doi: 10.1007/978-1-4614-3223.
- [2] T. Alfina, B. Santosa, and A. R. Barakbah, "Analisa Perbandingan Metode Hierarchical Clustering, K-means dan Gabungan Keduanya dalam Cluster Data (Studi kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS)," *J. Tek. ITS*, vol. 1, pp. 521–525, 2023, [Online]. Available: <https://ojs.unikom.ac.id/index.php/komputika%0Ahttp://journal.unhas.ac.id/index.php/ESTIMASI%0Ahttp://ejournal.unp.ac.id/students/index.php/mat/article/view/11552%0Ahttp://ojs.uho.ac.id/index.php/JMIP>
- [3] D. Widyadhana, R. B. Hastuti, I. Kharisudin, and F. Fauzi, "Perbandingan Analisis Klaster K-Means dan Average Linkage untuk Pengklasteran Kemiskinan di Provinsi Jawa Tengah," *Prism. Pros. Semin. Nas. Mat.*, vol. 4, no. 2, pp. 584–594, 2021.
- [4] S. N. Arofah and F. Marisa, "Penerapan Data Mining untuk Mengetahui Minat Siswa pada Pelajaran Matematika menggunakan Metode K-Means Clustering," *JOINTECS (Journal Inf. Technol. Comput. Sci.)*, vol. 3, no. 2, pp. 693–702, 2018, doi: 10.31328/jointecs.v3i2.787.
- [5] D. R. Ningrat, D. Asih, I. Maruddani, and T. Wuryandari, "Analisis Cluster Dengan Algoritma K-Means Dan Fuzzy C-Means Clustering Untuk Pengelompokan Data Obligasi Korporasi," *J. Gaussian*, vol. 5, no. 4, pp. 641–650, 2016, [Online]. Available: <http://ejournal-s1.undip.ac.id/index.php/gaussian>
- [6] S. Singh and S. Srivastava, "Review of Clustering Techniques in Control System," *Procedia Comput. Sci.*, vol. 173, pp. 272–280, 2020, doi: 10.1016/j.procs.2020.06.032.
- [7] S. Oktavia, M. N. Mara, and N. Satyahadewi, "MENGUNAKAN METODE WARD," pp. 93–100, 2013.
- [8] P. N. Pratama Artana, E. Prakarsa Mandyartha, and M. Hanindia Prami S, "Penerapan Data Mining Pada Algoritma Hierarchical Clustering Tentang Pengelolaan Mitra Perjalanan Wisatawan Bali Backpacker," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 7, no. 4, pp. 2903–2909, 2024, doi: 10.36040/jati.v7i4.7284.
- [9] D. Marlina, N. Lina, A. Fernando, and A. Ramadhan, "Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokan Wilayah Sebaran Cacat pada Anak," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 4, no. 2, p. 64, 2018, doi: 10.24014/coreit.v4i2.4498.
- [10] S. B. Faradilla, "Komparasi Analisis K-Medoids Clustering Dan Hierarchical Clustering," (*Studi Kasus Data Kriminalitas di Indones.*

Tahun 2020), 2022.

- [11] R. Mustajab, R. Aristawidya, L. Puspita, and E. Widodo, "Aplikasi Metode K-Medoid pada Pengelompokan Kabupaten/Kota di Jawa Barat berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2020," *J. Stat. dan Apl.*, vol. 5, no. 2, pp. 221–229, 2021.
- [12] S. Defiyanti, M. Jajuli, and N. Rohmawati, "Optimalisasi K-MEDOID dalam Pengklasteran Mahasiswa Pelamar Beasiswa dengan CUBIC CLUSTERING CRITERION," *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 1, pp. 211–218, 2017, doi: 10.25077/teknosi.v3i1.2017.211-218.
- [13] G. R. Suraya and A. W. Wijayanto, "Comparison of Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia according to the Special Index for Handling Stunting," *Indones. J. Stat. Its Appl.*, vol. 6, no. 2, pp. 180–201, 2022, doi: 10.29244/ijsa.v6i2p180-201.
- [14] N. N. Halim and E. Widodo, "Clustering Dampak Gempa Bumi di Indonesia Menggunakan Kohonen Self Organizing Maps," *Pros. SI MaNIS (Seminar Nas. Integr. Mat. dan Nilai Islam.*, vol. 1, no. 1, pp. 188–194, 2017, [Online]. Available: <http://conferences.uin-malang.ac.id/index.php/SIMANIS/article/view/62>