Penerapan Algoritma K-Medoids Untuk Mengelompokkan Status Obesitas

Retno Wahyusari[a*]
[a] Sekolah Tinggi Teknologi Ronggolawe
[*] retnowahyusari@gmail.com

Diterima : 1-6-2024 - Direview : 5-6-2024 - Diterbitkan : 30-6-2024

Intisari

Obesitas merupakan salah satu masalah kesehatan global yang semakin mendesak untuk ditangani. Organisasi Kesehatan Dunia (WHO) mencatat bahwa angka kejadian obesitas terus meningkat sampai kelipatan tiga sejak tahun 1975, dengan lebih dari 1,9 miliar orang dewasa mengalami kelebihan berat badan, dan di antaranya, lebih dari 650 juta orang terdiagnosis obesitas. Penelitian ini bertujuan untuk menerapkan algoritma K-Medoids dalam pengelompokan data status obesitas. Pengelompokan dilakukan menggunakan RapidMiner dengan berbagai konfigurasi jumlah kluster, dan hasil dievaluasi menggunakan beberapa metrik validasi kluster termasuk Davies-Bouldin Index (DBI). Hasil penelitian menunjukkan bahwa jumlah kluster tiga (k = 3) adalah yang paling optimal dalam mengelompokkan data obesitas, dengan nilai DBI sebesar 0,071, dibandingkan dengan dua kluster (k = 2) yang memiliki nilai DBI sebesar 0,101. Nilai DBI yang lebih rendah menunjukkan bahwa kluster yang terbentuk lebih kompak dan terpisah dengan baik, menandakan performa yang lebih baik dalam pengelompokan data status obesitas.

Kata Kunci: Obesitas, K-Medoids, Pengelompokan, *Davies-Bouldin Index*, RapidMiner.

Abstract

Obesity is a global health problem that is increasingly urgent to be addressed. The World Health Organization (WHO) notes that the incidence of obesity has tripled since 1975, with more than 1.9 billion adults being overweight, and of these, more than 650 million people diagnosed with obesity. This research aims to apply the K-Medoids algorithm in grouping obesity status data. Clustering was performed using RapidMiner with various configurations of the number of clusters, and results were evaluated using several cluster validation metrics including the Davies-Bouldin Index (DBI). The results showed that three clusters (k = 3) were the most optimal in grouping obesity data, with a DBI value of 0.071, compared to two clusters (k = 2) which had a DBI value of 0.101. A lower DBI value indicates that the clusters formed are more compact and well separated, indicating better performance in grouping obesity status data.

Keywords : *Obesity, K-Medoids, Clustering, Davies-Bouldin Index, RapidMiner.*

1. Pendahuluan

Obesitas merupakan salah satu masalah kesehatan global yang semakin mendesak untuk ditangani. Organisasi Kesehatan Dunia (WHO) mencatat bahwa angka kejadian obesitas terus meningkat sampai kelipatan tiga sejak tahun 1975, dengan jumlah orang dewasa yang mengalami kelebihan berat badan lebih besar dari 1,9 dari 650 juta orang terdiagnosis obesitas (WHO, 2023). Dampak obesitas tidak hanya terbatas pada peningkatan risiko penyakit kardiovaskular, diabetes tipe 2, dan berbagai penyakit metabolik, tetapi juga mempengaruhi kualitas hidup dan kesejahteraan psikososial individu yang bersangkutan (Hruby and Hu, 2015).

Pemahaman yang lebih baik tentang berbagai pola obesitas dan faktor risiko terkait diperlukan untuk mendukung upaya pencegahan dan pengobatan yang lebih efektif. Salah satu pendekatan yang dapat diterapkan dalam analisis data kesehatan adalah pengelompokan (*clustering*). Pengelompokan memungkinkan identifikasi kelompok-kelompok dalam data tanpa pengetahuan apriori mengenai kategori yang ada, sehingga dapat mengungkap pola dan karakteristik tersembunyi dalam dataset (Jain, 2010).

Algoritma *K-Medoids*, sebagai varian dari metode pengelompokan berbasis partisi seperti K-Means, menawarkan keunggulan dalam hal *robust* terhadap *outlier* dan kepekaan terhadap pusat *medoid* yang mewakili setiap kluster. Berbeda dengan K-Means yang menggunakan rata-rata (*mean*) sebagai pusat kluster, K-Medoids menggunakan *medoid* yang merupakan titik aktual dari dataset, sehingga lebih tahan terhadap pengaruh data pencilan (Park and Jun, 2009).

Dalam konteks analisis status obesitas, penggunaan algoritma *K-Medoids* dapat membantu dalam mengidentifikasi sub-kelompok individu berdasarkan karakteristik obesitas yang berbeda. Dengan mempertimbangkan variabel-variabel seperti indeks massa tubuh (BMI), lingkar pinggang, riwayat medis, dan pola aktivitas fisik, algoritma ini dapat mengelompokkan individu ke dalam kategori-kategori yang lebih representatif. Pengelompokan ini berpotensi meningkatkan pemahaman tentang variasi dalam status obesitas, yang dapat mengarah pada intervensi yang lebih spesifik dan personalisasi perawatan (Han et al., 2011).

K-Medoids pertama kali diperkenalkan sebagai alternatif yang lebih stabil dan robust terhadap *outlier* dibandingkan *K-Means* (Kaufman and Rousseeuw, 1990). Algoritma ini

meminimalkan total *dissimilarity* antara titik data dalam satu kluster dengan *medoid*, yang dipilih sebagai titik pusat dari dataset. Proses pemilihan *medoid* ini, meskipun lebih komputasi-intensif, menghasilkan kluster yang lebih representatif untuk data yang memiliki distribusi tidak normal atau mengandung *outlier*.

Penggunaan algoritma *K-Medoids* menawarkan beberapa keunggulan penting dibandingkan dengan algoritma pengelompokan lain seperti *K-Means* dan *Gaussian Mixture Models* (GMM). *K-Means*, meskipun populer, memiliki kelemahan dalam hal sensitivitas terhadap titik awal dan kesulitan dalam menangani bentuk kluster yang *non-sferis* atau ukuran yang berbeda. Sementara itu, GMM, meskipun lebih fleksibel dalam menangani distribusi data yang beragam, memerlukan estimasi parameter yang rumit dan rentan terhadap *overfitting* (Bishop, 2006).

Penelitian ini bertujuan untuk menerapkan algoritma K-Medoids dalam pengelompokan data status obesitas. Tujuan utama adalah hasil pengelompokan data status obesitas serta menilai kinerja *K-Medoids* dalam mengelompokkan individu berdasarkan parameter-parameter obesitas.

2. Kerangka Teori

2.1. Obesitas

Obesitas adalah kondisi medis yang ditandai dengan penumpukan lemak tubuh yang berlebihan, yang dapat meningkatkan risiko berbagai masalah kesehatan. Organisasi Kesehatan Dunia (WHO) mendefinisikan obesitas sebagai indeks massa tubuh (BMI) yang sama dengan atau lebih besar dari 30 kg/m². Obesitas bukan hanya masalah estetika, tetapi juga merupakan masalah kesehatan yang serius karena berkaitan dengan peningkatan risiko berbagai penyakit kronis seperti diabetes tipe 2, penyakit jantung, dan beberapa jenis kanker (World Health Organization, 2020).

Menurut data WHO, prevalensi obesitas global telah meningkat tiga kali lipat sejak 1975. Pada tahun 2016, lebih dari 1,9 miliar orang dewasa berusia 18 tahun ke atas mengalami kelebihan berat badan, dengan lebih dari 650 juta di antaranya tergolong obesitas. Tren ini tidak hanya terjadi di negara-negara berpenghasilan tinggi tetapi juga di negara-negara berpenghasilan menengah dan rendah, termasuk Indonesia (World Health Organization, 2020).

2.2. Pustaka Rujukan

Penelitian yang dilakukan oleh Suraya, G. R., & Wijayanto, A. W. (2022) membandingkan berbagai metode pengelompokan dan menemukan bahwa K-Medoids lebih robust terhadap outlier dibandingkan K-Means, namun aplikasi K-Medoids pada data kesehatan khususnya obesitas belum banyak dieksplorasi.

Penelitian yang dilakukan oleh Thamrin, N., & Wijayanto, A. W. (2021) menghasilkan aplikasi K-Medoids dalam mengelompokkan data kesejahteraan, menyoroti efektivitasnya dalam menangani data dengan outlier. Namun, penelitian serupa untuk data obesitas masih terbatas.

Penelitian yang menjadi rujukan berikutnya dilakukan oleh Govender, P., & Sivakumar, V. (2020), peneliti mereview berbagai teknik pengelompokan dan aplikasinya dalam data lingkungan, menunjukkan bahwa K-Medoids jarang digunakan meskipun memiliki keunggulan dalam menangani data dengan distribusi yang kompleks. Kebutuhan aplikasi pada data kesehatan menjadi jelas.

2.3. Algoritma K-Medoids

Algoritma K-Medoids adalah teknik pengelompokan yang mirip dengan K-Means, tetapi lebih robust terhadap outlier. K-Medoids menggunakan titik yang sebenarnya dalam dataset sebagai pusat kluster (medoid), berbeda dengan K-Means yang menggunakan rata-rata (mean) titik-titik dalam kluster. K-Medoids pertama kali diperkenalkan oleh Kaufman dan Rousseeuw pada tahun 1987 dalam buku mereka "Finding Groups in Data: An Introduction to Cluster Analysis" (Kaufman and Rousseeuw, 1990).

Algoritma K-Medoids terdiri dari beberapa langkah utama sebagai berikut:

1. Tentukan k sebagai banyaknya *cluster* yang akan dibentuk.
2. Pilih secara acak k objek untuk dijadikan pusat *cluster* (medoid).
3. Hitung kemiripan benda medoid dengan benda bukan medoid menggunakan jarak Euclidean dengan persamaan 1 dan letakkan benda bukan medoid pada medoid terdekat, lalu hitung jarak totalnya
   $d(p, q) = \sqrt{(p1 - q1)\ 2 + (p2 - q2)\ 2 + (p3 - q3)2}$ (1)
   Keterangan:
   d = jarak obyek
   p = data
   q = centroid / medoid
4. Pilih acak obyek non-medoid untuk tiap-tiap *cluster* dijadikan kandidat medoid baru.
5. Hitung kemiripan benda *medoid* dengan benda *non-medoid* menggunakan jarak *Euclidean* dengan persamaan 1 dan letakkan benda non-medoid pada medoid terdekat, kemudian hitung jarak totalnya.
6. Hitung selisih jarak total (*Stotal distance*), dengan *Stotal distance* = total jarak pada calon *medoid* baru – total jarak pada *medoid* lama.
7. Apabila diperoleh nilai jarak Stotal < 0 maka calon *medoid* yang baru akan menjadi medoid yang baru, dan apabila diperoleh nilai jarak *Stotal* > 0 maka proses iterasi akan terhenti.
8. Ulangi langkah 4 sampai dengann 7 hingga tidak terjadi perubahan *medoid* atau Stotal jarak > 0.

3. Metodologi

Penelitian ini bertujuan untuk menerapkan algoritma K-Medoids dalam mengelompokkan data status obesitas dengan tujuan mengidentifikasi pola-pola obesitas yang lebih representatif. Bagian ini menjelaskan metode penelitian yang digunakan, termasuk desain penelitian, populasi dan sampel, pengumpulan data, serta prosedur analisis data.

Penelitian ini menggunakan desain kuantitatif eksperimental yang bertujuan untuk mengevaluasi efektivitas algoritma K-Medoids dalam pengelompokan data status obesitas. Desain ini dipilih untuk memungkinkan pengujian hipotesis dan analisis data secara sistematis. Langkah yang dilakukan sebagai berikut (Gambar 1) :

1. Pengumpulan Data

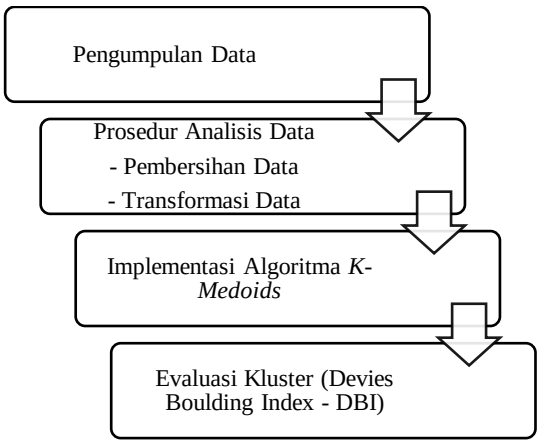
Data diambil dari dataset kaggle pada file *obesity or CVS Risk*. Atribut yang digunakan dalam menentukan status obesitas sebnayak 17 antara lain usia, berat badan, tinggi

badan, jenis kelamin, merokok atau tidak. Jumlah data yang tersedia sebanyak 2111.

2. Prosedur Analisis Data

Proses analisis data dilakukan melalui tahapan utama berupa *preprocessing data*. Proses *preprocessing data* dengan cara pembersihan data (menghapus data yang tidak lengkap atau *anomalous*) dan transformasi data (merubah data polinominal menjadi nominal)

3. Implementasi Algoritma *K-Medoids*: implementasi menggunakan bantuan RapidMminer.
4. Evaluasi Kluster: Menggunakan metode Devies Boulding Index (DBI) untuk mengukur kualitas pengelompokan.



Gambar 1 Metode Penelitian

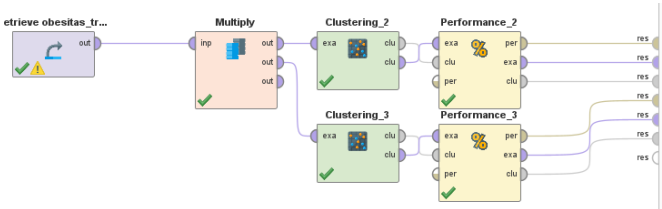
4. Hasil dan Pembahasan

Hasil *preprocessing data* data yang digunakan menjadi 500 data yang digunakandan merubah data polinominal menjadi nominal. Data yang telah dilakukan *preprocessing data* kemudian dilakukan proses pengelompokkan menggunakan algoritma K-Medoids dengan bantuan aplikasi rapid miner. Tabel 1 merupakan hasil dari proses *preprocessing data*.

Tabel 1
Data Obesitas

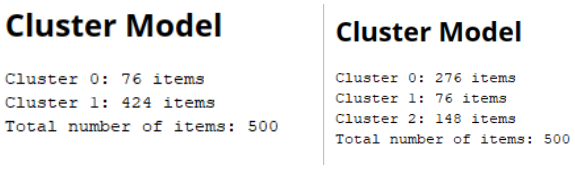
Gender	Age	Height	…	…	…	…	MTRANS
1	21	162	...	...	...	...	1
1	21	152	…	…	…	…	1
2	23	180	...	...	...	...	1
2	27	180	…	…	…	…	2
2	22	178	...	...	...	...	1
2	29	162	…	…	…	…	3
1	23	150	...	...	...	...	4
2	22	164	…	…	…	…	1
2	24	178	...	...	...	...	1

Data yang didapat kemudian dilakukan proses pengelompokkan menggunakan algoritma *K-Medoids* dengan bantuan aplikasi rapidminer. Desain pada rapidminer tersaji pada Gambar 2.



Gambar 2 Desain Pengelompokkan Mengggunakan Algortima K-Medoids

Hasil pengelompokan dengan menggunakan jumlah *cluster* 2 menghasilkan nilai DBI sebesar 0,101 dengan *cluster* normal sebanyak 76 data dan *cluster* obesitas sebesar 424. Sedangkan hasil untuk 3 *cluster* menghasilkan nilai DBI sebesar 0,071 dengan rincian *cluster* normal sebanyak 276 data, *cluster over weight* sebanyak 148 data, dan *cluster* obesitas sebanyak 76 data. Tersaji pada Gambar 3 dan Gambar 4.



Gambar 3 Jumlah Anggota Tiap *Cluster*



Gambar 4 Hasil Nilai Deviies Bouldin Index

Berdasarkan nilai DBI, jumlah *cluster* yang direkomendasikan adalah 3 *cluster*. Hal ini didasarkan pada banyaknya data yang dekat dengan pusat *cluster* dari *cluster* yang dilacak. Perhatikan bahwa jarak adalah jarak antara pusat cluster dan *cluster*. Suatu *cluster* dianggap optimal jika nilai aglomerasinya rendah dan nilai separasinya tinggi. Nilai DBI yang mendekati 0 menunjukkan bahwa cluster tersebut optimal.

5. Simpulan

Hasil penelitian dapat simpulkan bahwa jumlah *cluster* 3 yang paling baik dalam mengelompokkan pada data obesitas sebanyak 500 data dengan 17 atribut dilihat dari nilai DBI yang mendekati nilai 0 yaitu 0,071 dbandingkan dengan menggunakan jumlah *cluster* 2 dengan nilai DBI 0,101. Hasil pengelompokan dengan menggunakan jumlah *cluster* 2 menghasilkan *cluster* normal sebanyak 76 data dan *cluster* obesitas sebesar 424. Sedangkan hasil untuk 3 *cluster* menghasilkan n *cluster* normal sebanyak 276 data, *cluster over weight* sebanyak 148 data, dan *cluster* obesitas sebanyak 76 data.

Daftar Pustaka

Bishop, C.M., 2006. Pattern recognition and machine learning, Information science and statistics. Springer, New York.

Govender, P., Sivakumar, V., 2020. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmospheric Pollution Research 11, 40–56. https://doi.org/10.1016/j.apr.2019.09.009

Han, J., Kamber, M., Pei, J., 2011. Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

Hruby, A., Hu, F.B., 2015. The Epidemiology of Obesity: A Big Picture. PharmacoEconomics 33, 673–689. https://doi.org/10.1007/s40273-014-0243-x

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31, 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis, 1st ed, Wiley Series in Probability and Statistics. Wiley. https://doi.org/10.1002/9780470316801

Park, H.-S., Jun, C.-H., 2009. A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications 36, 3336–3341. https://doi.org/10.1016/j.eswa.2008.01.039

Suraya, G.R., Wijayanto, A.W., 2022. 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: Perbandingan Metode Hierarchical Clustering, K-Means, K-Medoids, dan Fuzzy C-Means dalam Pengelompokan Provinsi di Indonesia Menurut Indeks Khusus Penanganan Stunting. IJSA 6, 180–201. https://doi.org/10.29244/ijsa.v6i2p180-201

Thamrin, N., Wijayanto, A.W., 2021. Comparison of Soft and Hard Clustering: A Case Study on Welfare Level in Cities on Java Island: Analisis cluster dengan menggunakan hard clustering dan soft clustering untuk pengelompokkan tingkat kesejahteraan kabupaten/kota di pulau Jawa. IJSA 5, 141–160. https://doi.org/10.29244/ijsa.v5i1p141-160