IMPLEMENTASI K-MEAN CLUSTERING PADA ESTIMASI TINGKAT OBESITAS BERDASARKAN KONDISI FISIK (BERAT DAN TINGGI BADAN)

PROYEK UAS PEMBELAJARAN MESIN KELAS KARYAWAN (YBM)
Dosen Pengampu: Dr. Dra. DWINA KUSWARDANI, M.Kom

Disusun Oleh :

WAHYU JANUAR ALFIAN
NIM : 202231506

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TELEMATIKA ENERGI
INSTITUT TEKNOLOGI PLN
2025

Abstrak

Penelitian ini mengimplementasikan algoritma K-Mean Clustering untuk mengelompokkan tingkat obesitas berdasarkan kondisi fisik, yakni berat dan tinggi badan. Tinggi dan berat badan menjadi parameter kunci karena keduanya berhubungan langsung dengan massa tubuh, sehingga memungkinkan estimasi tingkat obesitas secara sederhana. Tantangan kesehatan global yang memerlukan pendekatan komprehensif untuk pencegahan dan pengelolaannya. Dengan mengadopsi gaya hidup sehat dan meningkatkan kesadaran akan dampaknya, risiko penyakit terkait obesitas dapat diminimalkan. Upaya bersama dari individu, komunitas, dan pemerintah sangat penting untuk menghadapi masalah ini. Data yang digunakan dianalisis untuk mengidentifikasi pola pengelompokan yang relevan guna memberikan estimasi tingkat obesitas. Teknik dalam analisis data untuk mengelompokkan objek berdasarkan kemiripan. Tujuan utamanya adalah untuk menemukan pola tersembunyi atau struktur alami dalam data. Agglomerative adalah pendekatan *bottom-up* dalam hierarchical clustering. Hasilnya berupa dendrogram (diagram pohon) yang menunjukkan hubungan antar kluster. Algoritma ini sederhana tetapi memiliki biaya komputasi tinggi untuk dataset besar.Dengan demikian, penelitian ini dapat memberikan kontribusi dalam pemantauan kesehatan masyarakat melalui pendekatan berbasis data.

Kata kunci: Obesitas, K-Mean, Clustering, Agglomerative, Hierarchical clustering

I. PENDAHULUAN

1.1. Latar Belakang

Obesitas merupakan masalah kesehatan global yang dapat meningkatkan risiko penyakit kronis seperti diabetes, penyakit jantung, dan hipertensi. Faktor-faktor seperti perubahan gaya hidup, pola makan tidak sehat, dan kurangnya aktivitas fisik telah menyebabkan peningkatan prevalensi obesitas. Identifikasi tingkat obesitas secara cepat dan akurat menjadi penting untuk mencegah komplikasi lebih lanjut. Dengan kemajuan teknologi, metode berbasis data seperti K-Mean Clustering menawarkan solusi inovatif untuk mengelompokkan individu berdasarkan tingkat obesitas dan memberikan wawasan untuk intervensi kesehatan yang lebih efektif. Obesitas didefinisikan sebagai kondisi kelebihan lemak tubuh yang berpotensi mengganggu kesehatan. Penilaian obesitas sering dilakukan menggunakan **Indeks Massa Tubuh (IMT)**, yang dihitung dengan rumus:

$$IMT = fracberat(kg)tinggi(m)^2$$

Kategori obesitas berdasarkan IMT menurut WHO:

- **< 18.5:** Berat badan kurang.
- **18.5–24.9:** Berat badan normal.
- **25–29.9:** Berat badan berlebih.
- **\u226530:** Obesitas.

1.2.Tujuan

- Mengimplementasikan algoritma K-Mean Clustering untuk estimasi tingkat obesitas.
- Menganalisis pola pengelompokan individu berdasarkan berat dan tinggi badan untuk menghasilkan kluster tingkat obesitas yang representatif.
- Menilai kinerja algoritma K-Mean dalam mengelompokkan data kesehatan.

1.3. Manfaat

- Membantu dalam pencegahan dan pemantauan obesitas secara lebih efektif.
- Memberikan wawasan baru terkait pengelompokan data kesehatan masyarakat untuk mendukung kebijakan kesehatan.
- Mengembangkan pendekatan analitik yang dapat diterapkan pada masalah kesehatan lainnya.

1.4. Batasan Penelitian

Penelitian ini hanya menggunakan parameter berat dan tinggi badan dalam analisis. Faktor lain seperti usia, jenis kelamin, aktivitas fisik, dan pola makan tidak diperhitungkan. Selain itu, penelitian ini menggunakan dataset statis tanpa mempertimbangkan data waktu nyata.

II. STUDI LITERATUR

2.2 K-Mean

K-Mean adalah algoritma clustering yang membagi data ke dalam sejumlah kluster berdasarkan kemiripan tertentu. Algoritma ini bekerja dengan menentukan pusat kluster (centroid) secara iteratif hingga konvergensi tercapai. Proses dimulai dengan inisialisasi centroid, penghitungan jarak antara data dan centroid, serta pembaruan centroid hingga data dikelompokkan secara optimal.

2.3 *Clustering*

Clustering adalah teknik analisis data yang mengelompokkan objek-objek serupa ke dalam grup tertentu. Metode ini digunakan untuk memahami struktur data yang kompleks dan menemukan pola tersembunyi di dalamnya. Clustering diterapkan dalam berbagai bidang, termasuk kesehatan, pemasaran, dan kecerdasan buatan.

2.4 Agglomerative Hierarchical Clustering

Metode clustering hirarkis yang dimulai dengan setiap data sebagai kluster tunggal. Kluster-kluster kemudian digabung secara bertahap berdasarkan kemiripan hingga membentuk satu kluster besar. Pendekatan ini sering menggunakan metrik jarak seperti Euclidean atau Manhattan untuk menentukan tingkat kemiripan antar kluster.

2.5 Divisive Clustering

Metode ini dimulai dengan satu kluster besar yang kemudian dipecah menjadi kluster-kluster lebih kecil berdasarkan perbedaan tertentu. Proses ini berlanjut hingga setiap data menjadi kluster tunggal atau sesuai dengan jumlah kluster yang diinginkan. Pendekatan ini lebih jarang digunakan dibandingkan agglomerative karena kompleksitasnya yang lebih tinggi.

2.6 Perbedaan Utama Agglomerative dan Divisive

Agglomerative dimulai dari bawah ke atas (bottom-up), sedangkan Divisive dimulai dari atas ke bawah (top-down). Perbedaan pendekatan ini menghasilkan struktur kluster yang berbeda. Agglomerative lebih sederhana dalam implementasi, sementara divisive lebih fleksibel tetapi membutuhkan lebih banyak komputasi.

2.7. Sklearn

Sklearn dalam konteks clustering mengacu pada proses data normalisasi atau standarisasisebelum diterapkan ke algoritma clustering, seperti K-Mean, Hierarchical Clustering, atau DBSCAN. Scaler memastikan bahwa fitur-fitur dalam data memiliki skala yang seimbang sehingga algoritma dapat bekerja secara optimal.

- **Rumus:**

  $x' =$

  $fracx - textmean(x)textstd(x)$

III. METODOLOGI PENELITIAN

3.1 Identifikasi Masalah

Masalah obesitas yang meningkat secara global menuntut pendekatan analisis data yang efektif untuk mengidentifikasi tingkat obesitas. Penggunaan metode clustering dapat membantu mengelompokkan individu berdasarkan tingkat risiko obesitas sehingga intervensi dapat dilakukan lebih tepat sasaran.

3.2 Studi Literatur

Penelitian ini mengacu pada literatur terkait K-Mean Clustering, metode clustering lainnya, serta penerapannya dalam bidang kesehatan. Studi literatur meliputi prinsip dasar algoritma, implementasi teknis, dan studi kasus sebelumnya.

3.3 Pengumpulan Data

Data berupa berat dan tinggi badan diperoleh dari dataset kesehatan atau survei yang relevan. Dataset ini kemudian diproses untuk menghilangkan data tidak valid atau outlier yang dapat memengaruhi hasil analisis.

3.4 Pengolahan Data

Pengolahan data adalah tahap penting sebelum menerapkan algoritma K-Mean Clustering. Langkah-langkah yang dilakukan meliputi:

1. Pemiilihan data
2. Pembagian Dataset
3. Standarisasi Data
4. Memilih Data untuk dijadikan atribut x
5. Ringkasan Statistik
6. Virtualisasi Persebaran Data

3.5 Penerapan *K-Mean Clustering*

Algoritma K-Mean diterapkan untuk mengelompokkan data ke dalam beberapa kluster tingkat obesitas. Analisis dilakukan untuk mengevaluasi hasil clustering.

3.6 Penarikan Kesimpulan

Penelitian ini menyimpulkan efektivitas K-Mean Clustering dalam mengelompokkan tingkat obesitas, serta memberikan rekomendasi untuk penelitian lanjutan dengan memasukkan parameter tambahan. Dari modul sklearn.metrics, yang digunakan untuk mengukur kualitas klaster. Menyimpan label klaster yang dihasilkan oleh model KMeans ke dalam variabel labels. Setiap titik data diberikan label yang sesuai dengan klaster tempatnya berada.

IV. HASIL PENGUJIAN

Tabel dataset Obesistas, Pemilihan data set untuk atribut X dan tabel ringkasan statistik

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesd
0	Female	21.000000	1.620000	64.000000	yes	no	2.0	3.0	Sometimes	no	2.000000	no	0.000000	1.000000	no	Public_Transportation	Normal_We
1	Female	21.000000	1.520000	56.000000	yes	no	3.0	3.0	Sometimes	yes	3.000000	yes	3.000000	0.000000	Sometimes	Public_Transportation	Normal_We
2	Male	23.000000	1.800000	77.000000	yes	no	2.0	3.0	Sometimes	no	2.000000	no	2.000000	1.000000	Frequently	Public_Transportation	Normal_We
3	Male	27.000000	1.800000	87.000000	no	no	3.0	3.0	Sometimes	no	2.000000	no	2.000000	0.000000	Frequently	Walking	Overweight_Lev
4	Male	22.000000	1.780000	89.800000	no	no	2.0	1.0	Sometimes	no	2.000000	no	0.000000	0.000000	Sometimes	Public_Transportation	Overweight_Leve
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2106	Female	20.976842	1.710730	131.408528	yes	yes	3.0	3.0	Sometimes	no	1.728139	no	1.676269	0.906247	Sometimes	Public_Transportation	Obesity_Type
2107	Female	21.982942	1.748584	133.742943	yes	yes	3.0	3.0	Sometimes	no	2.005130	no	1.341390	0.599270	Sometimes	Public_Transportation	Obesity_Type
2108	Female	22.524036	1.752206	133.689352	yes	yes	3.0	3.0	Sometimes	no	2.054193	no	1.414209	0.646288	Sometimes	Public_Transportation	Obesity_Type
2109	Female	24.361936	1.739450	133.346641	yes	yes	3.0	3.0	Sometimes	no	2.852339	no	1.139107	0.586035	Sometimes	Public_Transportation	Obesity_Type
2110	Female	23.664709	1.738836	133.472641	yes	yes	3.0	3.0	Sometimes	no	2.863513	no	1.026452	0.714137	Sometimes	Public_Transportation	Obesity_Type

2111 rows × 17 columns

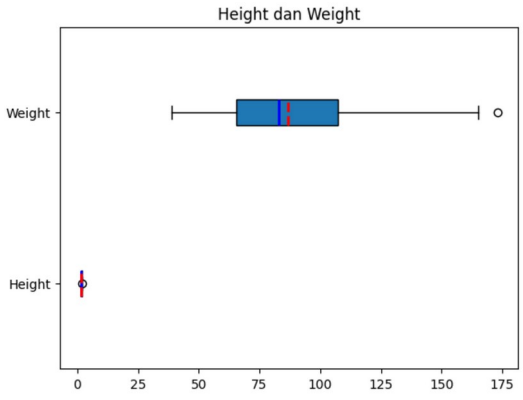
	Height	Weight
0	1.620000	64.000000
1	1.520000	56.000000
2	1.800000	77.000000
3	1.800000	87.000000
4	1.780000	89.800000
...	...	...
495	1.800000	60.000000
496	1.720000	53.000000
497	1.560000	45.000000
498	1.686306	104.572712
499	1.683124	126.673780

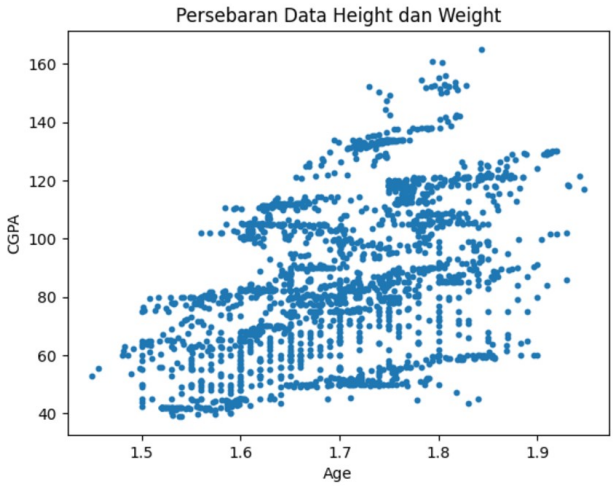
500 rows × 2 columns

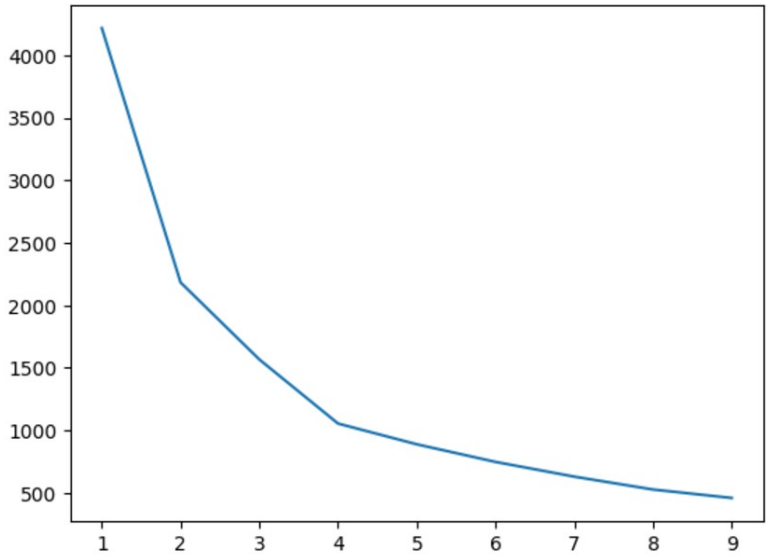
```
#MELIHAT RINGKASAN STATISTIK
data.describe()
```

	Height	Weight
count	2111.000000	2111.000000
mean	1.701677	86.586058
std	0.093305	26.191172
min	1.450000	39.000000
25%	1.630000	65.473343
50%	1.700499	83.000000
75%	1.768464	107.430682
max	1.980000	173.000000

Visualisasi Data







sklearn cluster import KMeans

KMeans

```
KMeans(n_clusters=3, random_state=0)
```

Model K-Mean Cluster

Evaluasi Model

```
#evaluasi model
from sklearn.metrics import davies_bouldin_score
labels = kmeans.labels_
davies_bouldin_score(x_scaled, labels)
```

```
0.9931907991994219
```

V. ANALISIS ATAS HASIL PENGUJIAN

5.1 Data

Tabel 1. Merupakan tabel data Obesitas bertujuan untuk menganalisis atau memprediksi tingkat obesitas seseorang berdasarkan atribut fisik, kebiasaan, dan gaya hidup. Tinggi badan individu (dalam meter), Berat badan individu (dalam kilogram).

Ringkasan statistik deskriptif untuk dua kolom dalam dataset, yaitu *Height* (tinggi badan) dan *Weight* (berat badan), yang dihasilkan menggunakan metode `data.describe()` dari pustaka pandas di Python. Berikut adalah penjelasan rinci:

Statistik Deskriptif:

1. **count**: Jumlah data yang tersedia untuk masing-masing kolom.
    - Kedua kolom memiliki 2111 data yang valid.
2. **mean**: Rata-rata nilai.
    - Rata-rata tinggi badan adalah **1.7017 meter**.
    - Rata-rata berat badan adalah **86.5861 kilogram**.
3. **std**: Standar deviasi, yang menunjukkan seberapa tersebar data dari rata-rata.
    - Tinggi badan memiliki standar deviasi **0.0933 meter**, artinya data tinggi badan cenderung homogen.
    - Berat badan memiliki standar deviasi **26.1912 kilogram**, menunjukkan variasi yang lebih besar dibanding tinggi badan.
4. **min**: Nilai minimum.
    - Tinggi badan terendah adalah **1.45 meter**.
    - Berat badan terendah adalah **39 kilogram**.
5. **25% (1st quartile)**: Nilai di bawah 25% data.
    - 25% data memiliki tinggi badan kurang dari **1.63 meter**.
    - 25% data memiliki berat badan kurang dari **65.47 kilogram**.
6. **50% (median)**: Nilai tengah atau median.
    - Median tinggi badan adalah **1.70 meter**.
    - Median berat badan adalah **83 kilogram**.
7. **75% (3rd quartile)**: Nilai di bawah 75% data.
    - 75% data memiliki tinggi badan kurang dari **1.768 meter**.
    - 75% data memiliki berat badan kurang dari **107.43 kilogram**.
8. **max**: Nilai maksimum.
    - Tinggi badan maksimum adalah **1.98 meter**.
    - Berat badan maksimum adalah **173 kilogram**.

5.2 Visualisasi *Data*

1. **Gambar Kiri: Boxplot untuk Height dan Weight**
    - **Deskripsi**: Boxplot adalah representasi visual dari distribusi data untuk dua variabel, yaitu *Height* (tinggi badan) dan *Weight* (berat badan).
    - **Komponen Boxplot**:
        - **Garis tengah (Median)**: Garis di dalam kotak menunjukkan nilai tengah dari data.
        - **Kotak (Interquartile Range - IQR)**: Kotak menunjukkan rentang data antara kuartil pertama (Q1) dan kuartil ketiga (Q3).
        - **Garis horizontal (Whiskers)**: Menunjukkan data yang berada di luar IQR tetapi masih dalam rentang wajar.
        - **Titik individu (Outlier)**: Titik di luar garis whisker adalah *outliers* (data yang sangat berbeda dari nilai rata-rata).
    - **Interpretasi**:
        - Data *Weight* memiliki nilai *outliers*, yang mungkin berat badan sangat tinggi.
        - Data *Height* lebih terdistribusi dengan baik, tanpa *outliers* signifikan.

2. **Gambar Kanan: Scatter Plot untuk Persebaran Height dan Weight**
    - **Deskripsi**: Scatter plot menampilkan hubungan antara dua variabel kontinu, yaitu *Height* (sumbu x) dan *Weight* (sumbu y).
    - **Pengamatan**:
        - Persebaran data menunjukkan adanya hubungan positif antara *Height* dan *Weight*, yaitu semakin tinggi seseorang, berat badannya cenderung lebih besar.
        - Data tersebar secara bervariasi, menunjukkan keragaman pada sampel dataset.
    - **Potensi Analisis**:
        - Scatter plot ini berguna untuk mengidentifikasi pola hubungan antar variabel sebelum melakukan clustering.
        - Pola-pola ini bisa menjadi dasar untuk pembentukan cluster berdasarkan kelompok berat badan (normal, overweight, obesitas).

5.2 Visualisasi Model

Grafik ini menunjukkan metode **Elbow** yang digunakan untuk menentukan jumlah cluster (n_clusters) yang optimal dalam algoritma K-Means. Sumbu X menunjukkan jumlah cluster, sedangkan sumbu Y menunjukkan *within-cluster sum of squares* (WCSS), yaitu total jarak antara data dalam satu cluster dengan centroid cluster tersebut.

- Garis grafik menunjukkan penurunan WCSS seiring dengan bertambahnya jumlah cluster.
- Titik "siku" atau "elbow" pada grafik sering digunakan untuk menentukan jumlah cluster yang optimal. Dalam kasus ini, titik elbow terlihat berada di **n_clusters = 3**, karena setelah itu penurunan WCSS menjadi tidak signifikan.

Gambar Kedua (KMeans Model)

Ini adalah model **K-Means** yang telah ditentukan menggunakan pustaka `sklearn.cluster`. Model ini mengatur jumlah cluster (`n_clusters=3`) dan menetapkan parameter `random_state=0` untuk memastikan hasil clustering yang konsisten setiap kali model dijalankan. Model ini akan mengelompokkan data menjadi 3 cluster berdasarkan pola data.

5.3 Hasil Clustering K-Means pada datase

Gambar ini menunjukkan hasil clustering K-Means pada dataset yang berkaitan dengan obesitas, dengan variabel-variabel seperti Weight (berat badan) dan komponen lain yang tidak disebutkan secara langsung di sumbu. Berikut penjelasannya:

1. **Visualisasi Cluster**:
    - Titik-titik dalam grafik mewakili data individu.
    - Warna yang berbeda (kuning, ungu, biru kehijauan) menunjukkan kelompok atau cluster yang telah diidentifikasi oleh model K-Means.
    - Skema warna yang ditampilkan dalam legenda mencerminkan label cluster atau nilai kategoris, dengan nilai dari 0 hingga 2 (karena model tampaknya menggunakan 3 cluster).
2. **Centroid (Titik Merah)**:
    - Titik merah di dalam setiap cluster adalah **centroid**, yaitu pusat dari setiap kelompok data.
    - Centroid mewakili nilai rata-rata dari semua data dalam cluster tertentu pada dimensi yang relevan.
3. **Hubungan Variabel**:
    - Sumbu X dan Y (dengan nama **Weight** pada Y) menunjukkan dimensi data yang digunakan untuk memvisualisasikan clustering, mungkin setelah dilakukan normalisasi atau pengurangan dimensi.
4. **Interpretasi Kasus Obesitas**:
    - Cluster ini dapat mencerminkan pengelompokan individu berdasarkan berat badan, mungkin dengan variabel tambahan (seperti tinggi badan, BMI, usia, atau pola makan).
    - Contoh interpretasi:
        - Cluster kuning mungkin adalah individu dengan **berat badan rendah atau kategori tertentu**.
        - Cluster ungu mungkin menunjukkan individu dengan kategori **sedang atau normal**.
        - Cluster biru kehijauan mungkin adalah individu dengan **kategori berat badan tinggi atau obesitas**.

5.4 Evaluasi Model

Kode di atas menunjukkan proses evaluasi sebuah model clustering menggunakan metrik Davies-Bouldin Score.

Pemecahan Kode:

1. **#evaluasi model:**

- Ini adalah komentar (diawali dengan #) yang menjelaskan tujuan dari kode tersebut, yaitu untuk mengevaluasi sebuah model.

2. **from sklearn.metrics import davies_bouldin_score:**

    - **Import:** Mengimpor fungsi davies_bouldin_score dari library scikit-learn. Fungsi ini digunakan untuk menghitung skor Davies-Bouldin, sebuah metrik yang digunakan untuk mengevaluasi kualitas clustering.
    - **Davies-Bouldin Score:** Semakin rendah nilai Davies-Bouldin Score, semakin baik kualitas clustering. Skor ini mengukur seberapa jauh cluster yang berdekatan dan seberapa kompak setiap cluster.

3. **labels = kmeans.labels_:**

    - **Menyimpan label:** Menyimpan label-label cluster yang dihasilkan oleh algoritma k-means ke dalam variabel labels. Label ini menunjukkan kelompok mana setiap data point termasuk.

4. **davies_bouldin_score(x_scaled, labels):**

    - **Menghitung skor:** Memanggil fungsi davies_bouldin_score dengan dua argumen:
        - x_scaled: Data yang telah di-scaling (dinormalisasi) yang digunakan untuk clustering.
        - labels: Label-label cluster yang telah diperoleh dari algoritma k-means.
    - **Hasil:** Fungsi ini akan mengembalikan nilai Davies-Bouldin Score, yang menunjukkan kualitas clustering dari model k-means.

5. **0.9931907991994219:**

    - **Nilai skor:** Nilai yang dihasilkan dari perhitungan Davies-Bouldin Score. Dalam kasus ini, nilainya cukup tinggi (mendekati 1), yang mengindikasikan bahwa kualitas clustering yang dihasilkan oleh model k-means kurang baik. Cluster yang dihasilkan mungkin tumpang tindih atau tidak kompak.

VI. KESIMPULAN:

Pada dua grafik ini memberikan wawasan tentang distribusi dan hubungan variabel yang berguna untuk analisis lebih lanjut, misalnya untuk *K-Mean Clustering*. Kedua grafik ini memberikan wawasan tentang distribusi dan hubungan variabel yang berguna untuk analisis lebih lanjut, misalnya untuk *K-Mean Clustering*.

Menggambarkan pembagian data menjadi 3 cluster menggunakan **K-Means Clustering**. Visualisasi ini mempermudah analisis untuk memahami bagaimana individu dikelompokkan berdasarkan atribut tertentu, dalam konteks ini terkait obesitas.

Pada Kmeans model gambar pertama digunakan untuk menentukan jumlah cluster optimal dengan metode elbow. Gambar kedua menunjukkan implementasi model K-Means setelah jumlah cluster dipilih (n_clusters=3).

Berdasarkan nilai Davies-Bouldin Score yang diperoleh, dapat disimpulkan bahwa model k-means yang digunakan dalam contoh ini menghasilkan clustering yang kurang baik. Perlu dilakukan penyetelan lebih lanjut pada parameter model atau mungkin perlu mencoba algoritma clustering yang berbeda untuk mendapatkan hasil yang lebih optimal.

VII. DAFTAR PUSTAKA

https://p2ptm.kemkes.go.id/uploads N2VaaXIxZGZwWFpEL1VlRFdQQ3ZRZz09/2018/02/FactSheet_Obesitas_Kit_Informasi_Obesitas.pdf

https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

Septiyanti1, Seniwati. (2020), Obesitas dan Obesitas Sentral pada Masyarakat Usia Dewasa di Daerah Perkotaan Indonesia. https://media.neliti.com/media/publications/332464-obesity-and-central-obesity-in-indonesia-ce4fc999.pdf

Geovani, Dite, Umari, Zainal, Ramadini, Suci. (2024), Cluster Analysis Of Obesity Risk Levels Using K-Means And Dbscan Methods.

Li li, Qifa song, Xi Yang, (2020), Categorization of β-cell capacity in patients with obesity via OGTT using K-means clustering https://ec.bioscientifica.com/view/journals/ec/9/2/EC-19-0476.xml

Muhammed Gulam Ahamad, Mohammed Faisal Ahmed, Mohammed Yousuf Uddin, (2016), Clustering as Data Mining Technique in Risk Factors Analysis of Diabetes, Hypertension and Obesity https://ej-eng.org/index.php/ejeng/article/view/202

Link Github : https://github.com/wahyupyan/202231506_WahyuJanuarAlfian_UAS-PM