

LAPORAN PRAKTIKUM DATA MINING

Afwa Fuadi Nugraha¹⁾, Angelica Noviana²⁾, Asrizal³⁾, Dara Cantika Dewi⁴⁾,
Kanaya Dea Thalita Akhmad⁵⁾, Rizky Andrian Bennovry⁶⁾

IMPLEMENTASI AGGLOMERATIVE CLUSTERING PADA DATA NUMERIK MENGGUNAKAN DATASET OBESITY CLASSIFICATION

Program Studi Sains Data, Fakultas Sains, Institut Teknologi Sumatera

afwa.121450019@student.itera.ac.id¹⁾, angelica.121450064@student.itera.ac.id²⁾,
asrizal.121450010@student.itera.ac.id³⁾, dara.121450127@student.itera.ac.id⁴⁾,
kanaya.121450001@student.itera.ac.id⁵⁾, rizky.121450073@student.itera.ac.id⁶⁾

Abstrak

Kondisi obesitas merupakan masalah kesehatan serius yang dapat memengaruhi kualitas hidup seseorang serta berpotensi memicu berbagai penyakit kronis. Penelitian ini bertujuan untuk mengidentifikasi hubungan antara berat badan, Body Mass Index (BMI), dan atribut lainnya dalam sebuah dataset yang menggambarkan kondisi obesitas. Dataset Obesity Classification digunakan sebagai basis analisis, dengan fitur-fitur seperti usia, jenis kelamin, tinggi, berat badan, BMI, dan label kategori berat badan. Metode analisis yang diterapkan meliputi visualisasi data menggunakan scatter plot serta penerapan teknik clustering seperti K-Means dan Agglomerative Hierarchical Clustering.

Visualisasi data menggambarkan hubungan antara berat badan, BMI, dan usia dalam dataset. Percobaan clustering menunjukkan tantangan dalam menentukan jumlah cluster yang optimal, walaupun penggunaan K-Means dengan 3 cluster menampilkan distribusi yang lebih baik dalam membedakan pola tertentu. Evaluasi menggunakan metrik seperti Elbow method dan silhouette score memberikan gambaran kesulitan dalam menentukan struktur yang jelas dalam data. Rekomendasi penelitian untuk melakukan analisis lanjutan dengan teknik reduksi dimensi atau eksplorasi fitur tambahan diharapkan dapat memberikan pemahaman yang lebih mendalam terhadap pola yang ada. Validasi ulang dengan model clustering yang berbeda juga direkomendasikan untuk memastikan keandalan hasil. Hasil analisis menunjukkan kompleksitas dalam mencari pola terkait obesitas dari dataset yang digunakan, serta menggarisbawahi pentingnya eksplorasi lebih lanjut untuk pemahaman yang lebih komprehensif.

Kata Kunci : Obesitas, Berat Badan, BMI, Clustering (KMeans, Agglomerative Hierarchical Clustering)

I. Pendahuluan

I.1 Latar Belakang

Obesitas adalah sebuah status kondisi gangguan kesehatan tubuh yang diakibatkan oleh adanya penumpukan lemak, dan seringkali bisa diukur dari berat badan yang di atas normalnya. Asupan kalori yang berlebihan apabila dibandingkan dengan aktivitas pembakaran kalori mengakibatkan kalori yang berlebihan itu menumpuk dalam bentuk lemak.

Keadaan obesitas dalam jangka panjang akan berdampak pada kualitas hidup yang bersangkutan dalam aktivitas hariannya karena munculnya ketidaknyamanan atau nyeri badan. Selain itu, obesitas juga dapat memicu munculnya penyakit kronis, di antaranya diabetes melitus, jantung koroner, stroke iskemik dan hemoragik. Di sisi lain, adanya obesitas juga dapat meningkatkan jumlah biaya pengeluaran.

Obesitas menjadi salah satu masalah kesehatan yang tidak biasa lagi, dikarenakan angka kasus tingkat obesitas semakin meningkat, dan dapat diidap oleh setiap orang tanpa melihat usia. Penelitian yang dilakukan dengan melakukan pemantauan berkala untuk kelebihan berat badan dan obesitas pada semua populasi di dunia menunjukkan bahwa penderita obesitas terdapat di banyak negara, seperti di negara Peru sebanyak 19,7 % [4], Kolombia sebanyak 22,3% [5], termasuk di Indonesia sebanyak 6,9%. Banyak hal yang dapat dilakukan untuk mencegah munculnya obesitas, misalnya melalui edukasi di masyarakat mengenai gejala, dan akibat kalau mengidap obesitas. Salah satu cara edukasi adalah dengan menyediakan sistem sederhana yang dapat digunakan oleh masyarakat guna melakukan cek secara mandiri mengenai kondisi kesehatannya. Apabila dari hasil pengecekan secara mandiri ditemukan gejala adanya penyakit seperti obesitas, maka masyarakat bisa melanjutkan pemeriksaan ke dokter.

I.2 Rumusan Masalah

1. Bagaimana hubungan antara berat badan, BMI, dan atribut lainnya?

I.2 Tujuan

1. Mengidentifikasi hubungan antara berat badan, BMI, dan atribut lainnya dalam dataset yang ada.

II. Metode

I.1 Data

Sumber dataset Obesity Classification diperoleh dari website kaggle yaitu <https://www.kaggle.com/datasets/sujithmandala/obesity-classification-data-set>. Jumlah dataset Obesity Classification yang diperoleh sebanyak 110 entri. Terdapat informasi atribut yaitu 7 fitur kolom. Dapat kami tampilkan beberapa bagian dataset Obesity Classification.

Tabel.1 Data Obesity Classification

ID	Age	Gender	Height	Weight	BMI	Label
1	25	Male	175	80	25.3	Normal Weight
2	30	Female	160	60	22.5	Normal Weight
3	35	Male	180	90	27.3	Overweight
4	40	Female	150	50	20	Underweight
5	45	Male	190	100	31.2	Obese
...
109	26	Female	150	15	5.6	Underweight
110	31	Male	190	20	8.3	Underweight

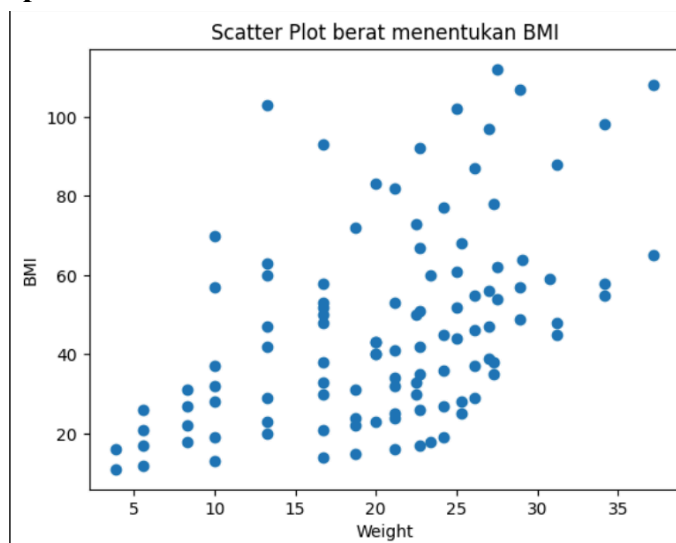
I.2 Agglomerative Hierarchical

Penelitian ini menggunakan metode Agglomerative Hierarchical Clustering. Agglomerative hierarchical clustering merupakan metode pengelompokan hierarki dengan pendekatan bawah atas (bottom up), dengan proses yang dimulai dari data sebagai sebuah kelompok, dilanjutkan dengan mencari kelompok potensial berdasarkan jarak untuk bergabung sebagai suatu kelompok yang lebih besar dan terus berulang sehingga tampak bergerak ke atas (Agglomerative) membentuk jenjang (Hierarki). Hierarchical Clustering adalah teknik clustering yang membentuk hirarki sehingga membentuk struktur pohon.

III. Hasil dan Pembahasan

Visualisasi ini dimulai dengan mengimpor beberapa library yang umum digunakan dalam analisis data dan pembuatan model machine learning, seperti pandas, matplotlib, dan scikit-learn. Selanjutnya, dataset dimuat dari file CSV dengan nama "Obesity Classification.csv". Selanjutnya kita menampilkan beberapa baris pertama dari dataset untuk memastikan pengambilan data berfungsi. Setelah itu, dilakukan pengecekan missing values pada dataset. Kemudian, kita lakukan visualisasi data dengan scatter plot.

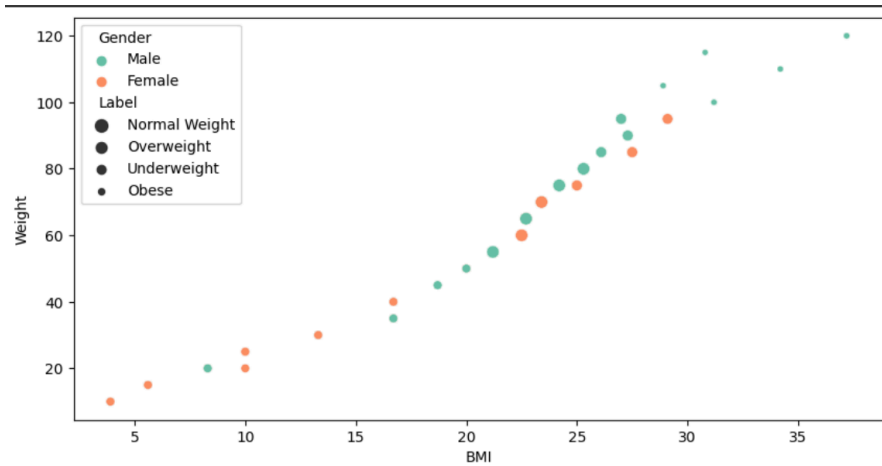
Visualisasi Eksplorasi Data secara linier



Visualisasi data diatas adalah dengan menggunakan scatter plot. Dalam scatter plot tersebut, sumbu x direpresentasikan oleh kolom "BMI" dari dataset, sementara sumbu y direpresentasikan oleh kolom "Age". Scatter plot ini bertujuan untuk menunjukkan hubungan antara berat badan (BMI) dan usia (Age) dalam dataset. Kode `'plt.scatter(X['BMI'], y)'` digunakan untuk membuat scatter plot dengan data berat badan sebagai sumbu x dan usia sebagai sumbu y.

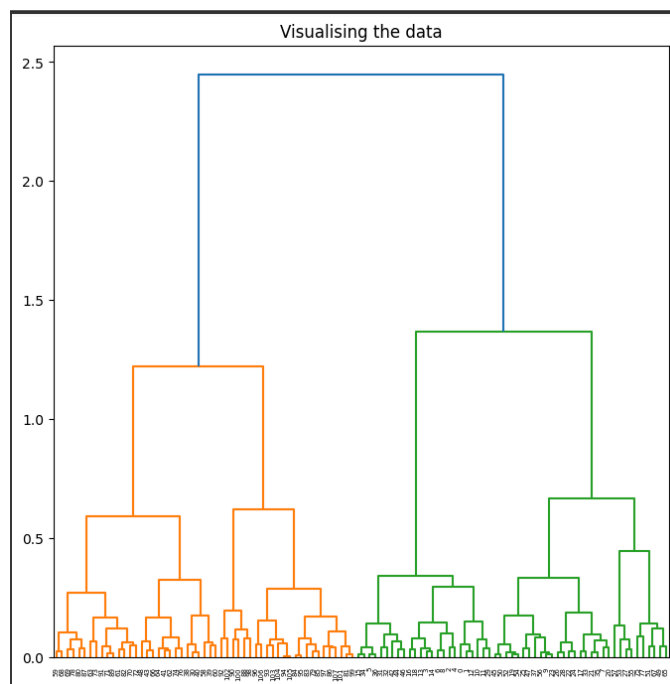
Pada visualisasi ini, kita dapat melihat sebaran titik-titik data di sepanjang sumbu x dan y, yang dapat memberikan gambaran visual tentang hubungan antara berat badan dan usia.

Visualisasi secara kelompok



Scatter plot di atas memanfaatkan library seaborn untuk menggabungkan beberapa variabel. Scatter plot tersebut memiliki sumbu x yang direpresentasikan oleh kolom "BMI", sumbu y oleh kolom "Weight", warna (hue) oleh kolom "Gender", dan ukuran titik (size) oleh kolom "Label". Warna titik pada plot menunjukkan perbedaan jenis kelamin, sedangkan ukuran titik adalah kategori atau label tertentu dalam dataset.

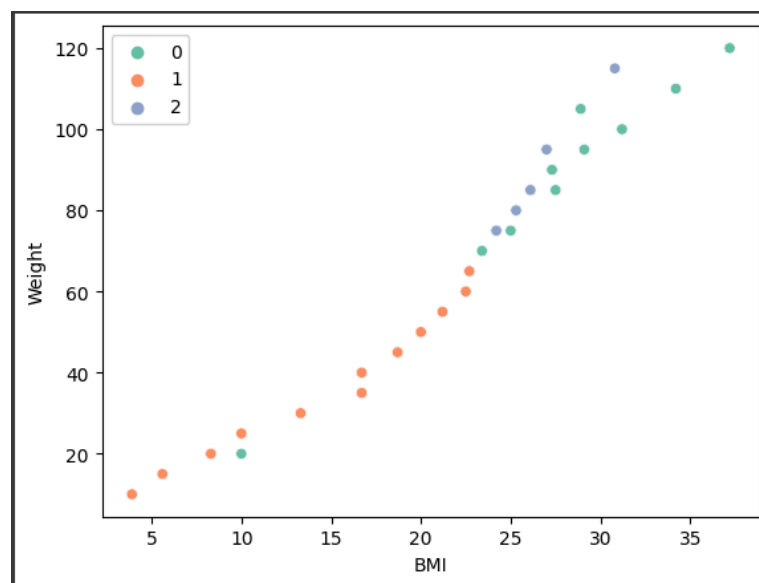
Membangun Model



Dendrogram yang dihasilkan oleh kode tersebut menunjukkan bahwa data dapat dikelompokkan menjadi dua kelompok utama. Kelompok pertama terdiri dari data dengan nilai P1 yang lebih rendah dan nilai P2 yang lebih tinggi. Kelompok kedua terdiri dari data dengan nilai P1 yang lebih tinggi dan nilai P2 yang lebih rendah.

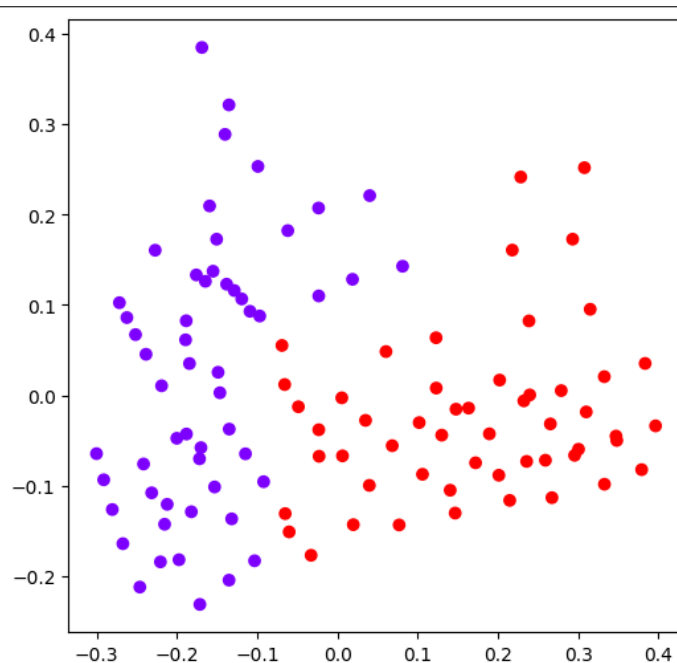
Pemisahan dua kelompok ini cukup jelas, karena jarak antara dua kelompok tersebut relatif besar. Hal ini menunjukkan bahwa data memiliki struktur yang jelas dan dapat dikelompokkan dengan baik menggunakan hierarchical clustering.

Visualisasi Cluster



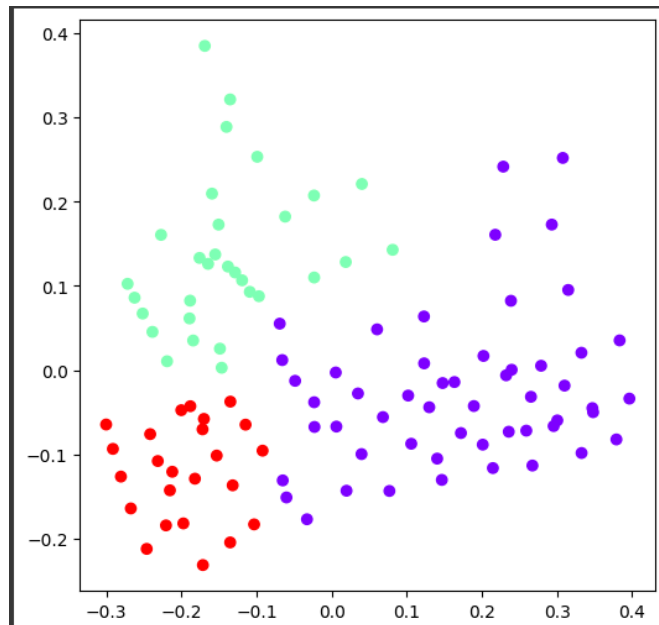
Scatter plot yang dihasilkan oleh kode tersebut menunjukkan bahwa ketiga cluster memiliki distribusi yang berbeda. Cluster 1 terdiri dari data dengan nilai BMI yang lebih rendah dan nilai berat badan yang lebih rendah. Cluster 2 terdiri dari data dengan nilai BMI yang lebih tinggi dan nilai berat badan yang lebih tinggi. Cluster 3 terdiri dari data dengan nilai BMI yang lebih tinggi dan nilai berat badan yang lebih rendah.

Visualisasi Cluster menggunakan metode Agglomerative



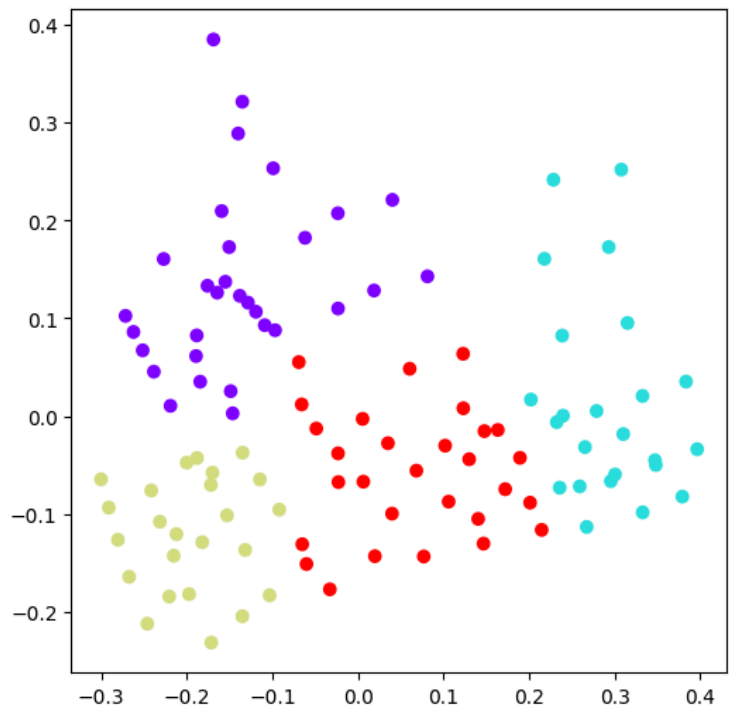
Hasil clustering dapat dilihat pada scatter plot yang dihasilkan oleh kode tersebut. Scatter plot menunjukkan bahwa data dikelompokkan menjadi dua cluster yang berbeda. Cluster pertama terdiri dari data dengan nilai P1 yang lebih rendah dan nilai P2 yang lebih tinggi. Cluster kedua terdiri dari data dengan nilai P1 yang lebih tinggi dan nilai P2 yang lebih rendah.

Pemisahan dua kelompok ini cukup jelas, karena jarak antara dua kelompok tersebut relatif besar. Hal ini menunjukkan bahwa data memiliki struktur yang jelas dan dapat dikelompokkan dengan baik menggunakan agglomerative clustering.



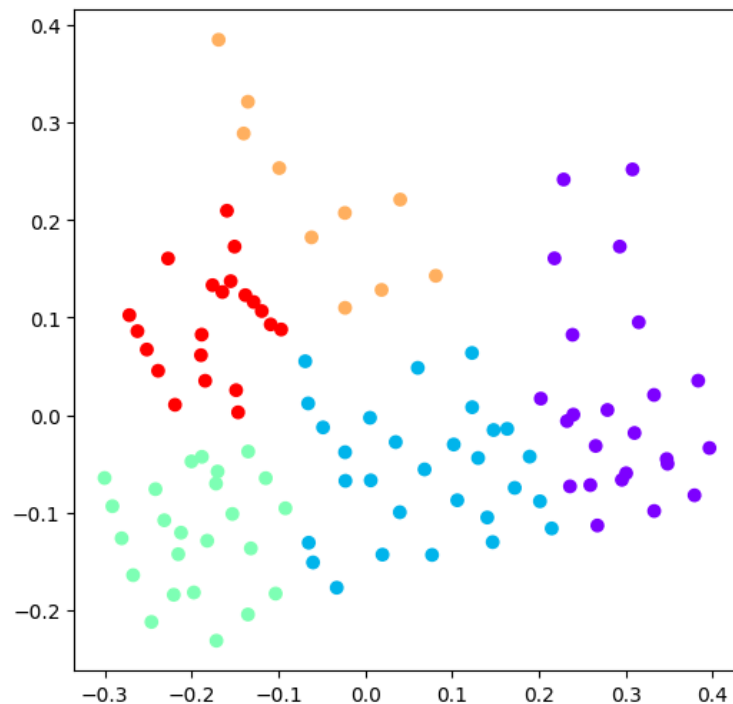
Hasil clustering dapat dilihat pada scatter plot yang dihasilkan oleh kode tersebut. Scatter plot menunjukkan bahwa data dikelompokkan menjadi tiga cluster yang berbeda. Cluster pertama terdiri dari data dengan nilai P1 yang lebih rendah dan nilai P2 yang lebih tinggi. Cluster kedua terdiri dari data dengan nilai P1 yang lebih tinggi dan nilai P2 yang lebih rendah. Cluster ketiga terdiri dari data yang berada di antara dua cluster pertama.

Pemisahan dua kelompok pertama cukup jelas, karena jarak antara dua kelompok tersebut relatif besar. Namun, cluster ketiga tidak begitu jelas, karena jaraknya dengan dua cluster pertama tidak terlalu jauh. Hal ini menunjukkan bahwa data memiliki struktur yang tidak terlalu jelas dan dapat dikelompokkan dengan baik menggunakan agglomerative clustering, tetapi dengan hasil yang tidak terlalu optimal.



Hasil clustering dapat dilihat pada scatter plot yang dihasilkan oleh kode tersebut. Scatter plot menunjukkan bahwa data dikelompokkan menjadi empat cluster yang berbeda. Cluster pertama terdiri dari data dengan nilai P1 yang lebih rendah dan nilai P2 yang lebih tinggi. Cluster kedua terdiri dari data dengan nilai P1 yang lebih tinggi dan nilai P2 yang lebih rendah. Cluster ketiga terdiri dari data yang berada di antara dua cluster pertama. Cluster keempat terdiri dari data yang berada di luar dua cluster pertama.

Pemisahan dua kelompok pertama cukup jelas, karena jarak antara dua kelompok tersebut relatif besar. Namun, cluster ketiga dan cluster keempat tidak begitu jelas, karena jaraknya dengan dua cluster pertama tidak terlalu jauh. Hal ini menunjukkan bahwa data memiliki struktur yang tidak terlalu jelas dan dapat dikelompokkan dengan baik menggunakan agglomerative clustering, tetapi dengan hasil yang tidak terlalu optimal.



Hasil clustering dapat dilihat pada scatter plot yang dihasilkan oleh kode tersebut. Scatter plot menunjukkan bahwa data dikelompokkan menjadi lima cluster yang berbeda. Cluster pertama terdiri dari data dengan nilai P1 yang lebih rendah dan nilai P2 yang lebih tinggi. Cluster kedua terdiri dari data dengan nilai P1 yang lebih tinggi dan nilai P2 yang lebih rendah. Cluster ketiga terdiri dari data yang berada di antara dua cluster pertama. Cluster keempat terdiri dari data yang berada di luar dua cluster pertama. Cluster kelima terdiri dari data yang berada di luar dua cluster pertama.

Pemisahan dua kelompok pertama cukup jelas, karena jarak antara dua kelompok tersebut relatif besar. Namun, cluster ketiga, cluster keempat, dan cluster kelima tidak begitu jelas, karena jaraknya dengan dua cluster pertama tidak terlalu jauh. Hal ini menunjukkan bahwa data memiliki struktur yang tidak terlalu jelas dan dapat dikelompokkan dengan baik menggunakan agglomerative clustering, tetapi dengan hasil yang tidak terlalu optimal.

Evaluasi Model dengan Boxplot

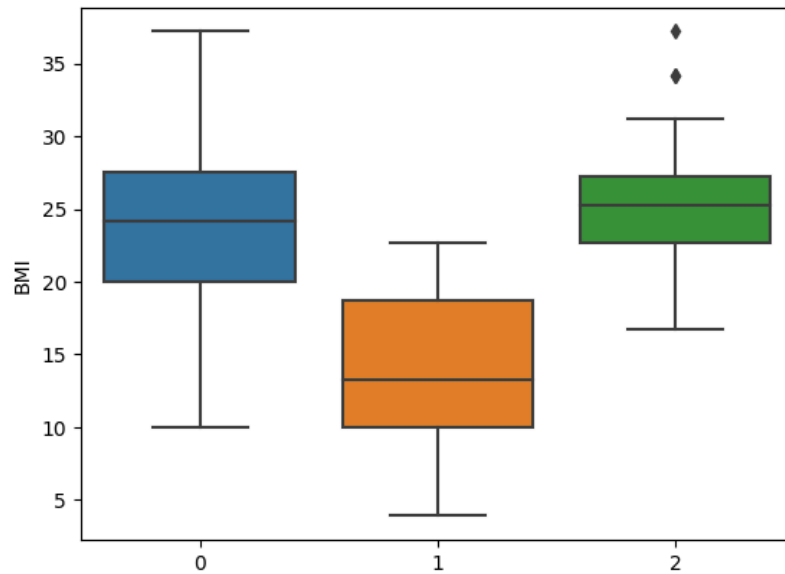
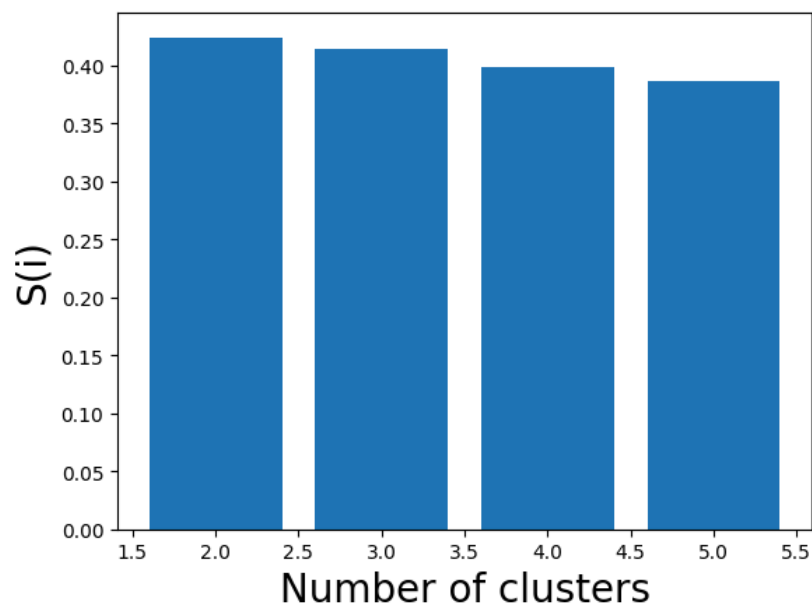


Diagram boxplot yang dihasilkan oleh kode tersebut menunjukkan bahwa ada perbedaan yang signifikan dalam distribusi BMI antara cluster yang berbeda. Pada cluster 1, nilai BMI terdistribusi dengan rentang yang lebih sempit dan nilai median yang lebih rendah. Hal ini menunjukkan bahwa cluster 1 terdiri dari orang-orang dengan BMI yang lebih rendah secara keseluruhan. Pada cluster 2, nilai BMI terdistribusi dengan rentang yang lebih luas dan nilai median yang lebih tinggi. Hal ini menunjukkan bahwa cluster 2 terdiri dari orang-orang dengan BMI yang lebih tinggi secara keseluruhan. Perbedaan distribusi BMI antara cluster yang berbeda ini menunjukkan bahwa K-Means clustering dapat digunakan untuk membedakan kelompok orang dengan BMI yang berbeda.

Evaluasi Agglomerative Hasil Antara Agglomerative 2 dan 3



Bar plot yang dihasilkan oleh kode menunjukkan bahwa silhouette score tertinggi diperoleh pada cluster 3. Hal ini menunjukkan bahwa cluster 3 adalah cluster terbaik untuk data tersebut, karena memiliki pemisahan yang paling jelas antara cluster yang berbeda.

IV. Kesimpulan


Dari analisis yang dilakukan terhadap dataset mengenai berat badan (Weight), BMI (Body Mass Index), dan faktor-faktor terkait lainnya, dapat disimpulkan beberapa hal. Visualisasi awal antara berat badan dan BMI menunjukkan hubungan yang tidak jelas secara langsung. Hasil dari beberapa percobaan clustering menunjukkan bahwa tidak ada jumlah cluster yang secara tegas membagi data menjadi kelompok yang terpisah dengan jelas. Meskipun demikian, penggunaan K-Means dengan 3 cluster menampilkan distribusi yang relatif lebih baik dalam membedakan pola-pola tertentu. Evaluasi menggunakan metrik seperti Elbow method dan silhouette score menunjukkan tantangan dalam menentukan jumlah cluster yang optimal. Rekomendasi untuk melakukan ekspansi analisis dengan teknik reduksi dimensi atau eksplorasi fitur tambahan mungkin dapat membantu menemukan struktur yang lebih jelas dalam data ini. Terakhir, validasi ulang dengan model clustering yang berbeda juga direkomendasikan untuk memastikan hasil yang diperoleh konsisten dan reliabel.

Referensi

[1] Dessy Yulianti, Teguh Hermanto, Meriska Defriani. (2023). Analisis Clustering Donor Darah dengan Metode Agglomerative Hierarchical Clustering, 3(6). <https://doi.org/10.30865/resolusi>

Lampiran

Berikut adalah lampiran pendukung pada laporan:

1. Link dataset:
<https://www.kaggle.com/datasets/sujithmandala/obesity-classification-data-set>
2. Link kode program:  Kel4_Damin_13&14_RB