PERBANDINGAN ALGORITMA K-MEANS DAN HIERARCHICAL CLUSTERING UNTUK PENGELOMPOKKAN DATA PENGUNJUNG MALL

Angga Pangestika

Sekolah Tinggi Manajemen Informatika dan Komputer Jakarta STI&K
Jalan BRI No.17, Radio Dalam Kebayoran Baru, Jakarta Selatan 12140, Indonesia
angga.pangestika@gmail.com

ABSTRAK

Penelitian ini membahas perbandingan algoritma K-Means dan Hierarchical Clustering dalam segmentasi pelanggan pusat perbelanjaan. Dengan menggunakan data pelanggan yang mencakup usia, jenis kelamin, pendapatan tahunan, skor pengeluaran, dan frekuensi kunjungan, studi ini mengevaluasi efektivitas kedua metode dalam mengelompokkan pelanggan. K-Means menggunakan Elbow Method dan Silhouette Score untuk menentukan jumlah klaster optimal, sementara Hierarchical Clustering menggunakan Ward linkage dan dendrogram.. Kedua metode berhasil mengidentifikasi lima segmen pelanggan yang berbeda. K-Means unggul dalam kecepatan dan kualitas klasterisasi, sedangkan Hierarchical Clustering menawarkan analisis yang lebih mendalam melalui struktur hierarkis. Untuk kebutuhan analisis real-time, K-Means lebih direkomendasikan, sementara Hierarchical Clustering lebih cocok digunakan dalam analisis eksploratif. Perbandingan kinerja algoritma clustering K-Means dan Hierarchical Clustering dianalisis dengan memanfaatkan empat metrik evaluasi utama, yaitu Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, serta waktu pemrosesan komputasi. Tools yang digunakan adalah menggunakan Python pustaka library Pandas, Numpy, Scikit-Learn, Matplotlib, Scipy dan Seaborn.

Kata Kunci: *K-Means, Hierarchical, Clustering, Elbow, Silhoutte*

PENDAHULUAN

Industri ritel, khususnya pusat perbelanjaan (mall), menghadapi persaingan yang semakin ketat dalam era digitalisasi saat ini. Pemahaman mendalam terhadap perilaku pelanggan menjadi kunci sukses dalam mengembangkan strategi pemasaran yang efektif dan meningkatkan kepuasan pelanggan [1]. Segmentasi pelanggan merupakan salah satu pendekatan fundamental dalam marketing analytics yang memungkinkan perusahaan untuk mengidentifikasi kelompok-kelompok pelanggan dengan karakteristik serupa.

Data mining, sebagai bagian dari business intelligence, telah menjadi alat yang sangat berharga dalam mengekstrak informasi bermakna dari volume data yang besar [2]. Clustering analysis, sebagai salah satu teknik unsupervised learning, memungkinkan identifikasi pola-pola tersembunyi dalam data pelanggan tanpa memerlukan label atau target variable yang telah ditentukan sebelumnya.

Algoritma clustering yang paling umum digunakan dalam industri adalah K-Means dan Hierarchical Clustering. K-Means dikenal dengan efisiensi komputasionalnya dan kemudahan implementasi, sementara Hierarchical Clustering memberikan insight struktur hierarkis yang lebih mendalam [3]. Namun, pemilihan algoritma yang tepat untuk konteks spesifik data pengunjung mall masih memerlukan evaluasi empiris yang komprehensif.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan quantitative comparative study dengan experimental design untuk membandingkan performa algoritma K-Means dan Hierarchical Clustering. Metodologi yang digunakan mengikuti framework Knowledge Discovery in

Databases (KDD) yang terdiri dari tahapan: data selection, preprocessing, transformation, data mining, dan interpretation/evaluation[4].

Perbandingan kedua algoritma clustering dilakukan melalui evaluasi kuantitatif dan kualitatif yang komprehensif. Secara kuantitatif menggunakan tiga metrik utama untuk mengukur kualitas clustering: (1) Silhouette Score yang mengukur seberapa baik data point cocok dalam clusternya (nilai 0-1, semakin tinggi semakin baik), (2) Davies-Bouldin Index yang mengukur ratio antara within-cluster scatter dan between-cluster separation (semakin rendah semakin baik), dan (3) Calinski-Harabasz Index yang mengevaluasi rasio between-cluster variance terhadap within-cluster variance (semakin tinggi semakin baik). Selain itu, penelitian ini membandingkan efisiensi komputasional dengan mengukur waktu eksekusi dan penggunaan memori, serta stabilitas algoritma dengan menjalankan multiple runs untuk menguji konsistensi hasil. Proses dimulai dengan preprocessing data (standardisasi), penentuan jumlah cluster optimal menggunakan Elbow Method untuk K-Means dan Dendrogram Analysis untuk Hierarchical Clustering, kemudian implementasi kedua algoritma pada dataset yang sama dengan parameter optimal.

Evaluasi kualitatif dilakukan melalui analisis profil cluster dan interpretabilitas bisnis. Setiap cluster dianalisis karakteristiknya berdasarkan mean dan standard deviation dari variabel-variabel seperti usia, pendapatan, dan spending score untuk menentukan apakah segmen yang dihasilkan meaningful dan actionable untuk strategi bisnis. Penelitian ini juga menggunakan visualisasi PCA (Principal Component Analysis) untuk mereduksi dimensi data menjadi 2D dan membandingkan secara visual bagaimana kedua algoritma mengelompokkan data.

Implementasi code perbandingan algoritma clustering menggunakan beberapa library Python utama: Pandas untuk manipulasi data dan analisis profil cluster melalui fungsi read_csv(), groupby(), dan agg(); NumPy untuk operasi matematika dan array manipulation seperti bincount() dan argmax(); Matplotlib dan Seaborn untuk visualisasi data dalam bentuk scatter plots, histogram, pie charts, dan Elbow curves; Scikit-learn sebagai library machine learning utama yang menyediakan algoritma KMeans dan AgglomerativeClustering, preprocessing dengan StandardScaler untuk normalisasi data, dimensionality reduction dengan PCA untuk visualisasi 2D, serta tiga metrik evaluasi clustering yaitu silhouette_score(), davies_bouldin_score(), dan calinski_harabasz_score(); SciPy khususnya modul scipy.cluster.hierarchy untuk membuat linkage matrix dan visualisasi dendrogram dalam hierarchical clustering; serta time module untuk mengukur efisiensi komputasional kedua algoritma. Semua tools ini terintegrasi dalam workflow yang sistematis: pandas untuk data loading dan preprocessing, StandardScaler untuk normalisasi fitur, KMeans dan AgglomerativeClustering untuk clustering implementation, metrik scikit-learn untuk evaluasi kuantitatif, PCA dan matplotlib untuk visualisasi perbandingan, dan pandas kembali untuk analisis profil cluster dan interpretasi bisnis, menghasilkan analisis komprehensif yang mencakup aspek teknis (metrics) maupun praktis (business insights).

K-Means adalah algoritma partitioning clustering yang membagi dataset menjadi k cluster tetap, di mana setiap data point dikelompokkan ke cluster dengan centroid (pusat) terdekat. Algoritma ini bekerja secara iteratif untuk meminimalkan variansi intra-cluster (jarak dalam cluster) sambil memaksimalkan perbedaan antar cluster. Setiap cluster direpresentasikan oleh centroid, yang merupakan rata-rata (mean) dari semua data point dalam cluster tersebut. Pengelompokan didasarkan pada metrik jarak, biasanya Euclidean distance[5].

Asumsi Utama :

- Data bersifat Euclidean (jarak dihitung menggunakan metrik seperti Euclidean distance).
- Cluster berbentuk bulat (spherical) dengan ukuran serupa.
- Tidak ada asumsi ketat tentang bentuk cluster, tetapi performa terbaik pada cluster globular dan terpisah dengan baik.

K-Means meminimalkan within-cluster sum of squares (WCSS), yang setara dengan meminimalkan deviasi kuadrat berpasangan dalam cluster. Ini juga terkait dengan memaksimalkan between-cluster sum of squares (BCSS) berdasarkan hukum variansi total [6].

Algoritma standar K-Means (Lloyd's algorithm) menggunakan teknik iteratif refinement dengan langkah-langkah berikut:

1. **Inisialisasi**: Tentukan jumlah cluster **k** dan pilih **k** centroid secara acak (atau menggunakan metode seperti K-Means++ untuk menghindari inisialisasi buruk). Centroid awal bisa dari data point acak.
2. **Langkah Assignment**: Assign setiap data point ke cluster dengan centroid terdekat, menggunakan jarak Euclidean kuadrat terkecil. Secara matematis, untuk setiap titik (x_p), assign ke cluster $(S_i^{(t)})$ di mana:

$$S_i^{(t)} = \{x_p : |x_p - m_i^{(t)}|^2 \leq |x_p - m_j^{(t)}|^2 \; \forall j, 1 \leq j \leq k\}, \quad (1)$$

dengan $(m_i^{(t)})$ adalah centroid cluster **i** pada iterasi **t**.

3. **Langkah Update**: Hitung ulang centroid sebagai mean dari data point dalam cluster masing-masing:

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j, \quad (2)$$

di mana $(\left|S_i^{(t)}\right|)$ adalah ukuran cluster $(S_i^{(t)})$.

4. **Iterasi**: Ulangi langkah 2-3 hingga centroid stabil (tidak berubah signifikan), assignment tidak berubah, **atau** mencapai iterasi maksimum. Algoritma konvergen ketika WCSS stabil, meskipun mungkin tidak mencapai optimum global.

Objektif utama adalah meminimalkan WCSS:

$$\arg\min_S \sum_{i=1}^{k} \sum_{x \in S_i} |x - \mu_i|^2 = \arg\min_S \sum_{i=1}^{k} |S_i| Var\, S_i, \quad (3)$$

di mana $(\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x)$ adalah mean cluster (S_i), $(|S_i|)$ adalah ukuran cluster, dan $(|\cdot|)$ adalah norma L2. Ini setara dengan meminimalkan deviasi kuadrat berpasangan dalam cluster:

$$\arg\min_S \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{x,y \in S_i} |x - y|^2 \quad (4),$$

menggunakan identitas:

$$|S_i| \sum_{x \in S_i} |x - \mu_i|^2 = \frac{1}{2} \sum_{x,y \in S_i} |x - y|^2. \quad (5)$$

K-Means meminimalkan squared Euclidean distances, bukan regular Euclidean distances (yang lebih sulit, seperti Weber problem).

Parameter Utama:

- Jumlah cluster **k** (harus ditentukan manual, bisa dioptimalkan dengan Elbow Method atau Silhouette Score).
- Metrik jarak (Euclidean default).
- Metode inisialisasi (random atau K-Means++)[5].

Hierarchical clustering mengorganisir data menjadi struktur pohon di mana cluster bersarang, memungkinkan analisis pada berbagai tingkat granularitas. Cluster dibentuk berdasarkan kemiripan (jarak) antar data point atau cluster, menggunakan metrik jarak (seperti Euclidean, Manhattan) dan linkage criteria

untuk menentukan bagaimana jarak antar cluster dihitung. Dendrogram adalah representasi visual utama, di mana sumbu vertikal menunjukkan jarak atau dissimilarity, dan pemotongan horizontal pada ketinggian tertentu menghasilkan partisi cluster. Metode ini hanya memerlukan matriks jarak, bukan data mentah, sehingga fleksibel untuk berbagai tipe data [5].

Hierarchical clustering memiliki dua strategi utama:

1. **Agglomerative (Bottom-Up)**: Mulai dengan setiap data point sebagai cluster tunggal (total **n** cluster, di mana **n** adalah jumlah data) dan secara iteratif menggabungkan dua cluster paling mirip hingga tersisa satu cluster besar atau mencapai kriteria stopping. Ini lebih umum karena kesederhanaan dan efisiensinya untuk dataset kecil hingga sedang [7].
2. **Divisive (Top-Down)**: Mulai dengan semua data point dalam satu cluster besar dan secara rekursif memecahnya menjadi cluster lebih kecil, sering menggunakan heuristik untuk memaksimalkan jarak antar cluster hasil. Kurang umum tetapi berguna untuk mengidentifikasi cluster besar dan distinct terlebih dahulu [7].

Kedua pendekatan menggunakan metode greedy untuk merge dan split, dan pilihan strategi dapat memengaruhi hasil clustering.

Hierarchical clustering bergantung pada metrik jarak (misalnya, Euclidean) dan kriteria linkage untuk mengukur dissimilarity. Kompleksitas waktu untuk agglomerative standar adalah $(\mathcal{O}(n^3))$ dengan memori $(\Omega(n^2))$, meskipun metode efisien seperti SLINK (single-linkage, $(\mathcal{O}(n^2))$ dan CLINK (complete-linkage, $(\mathcal{O}(n^2))$ ada untuk kasus spesifik. Menggunakan heap dapat mengurangi runtime ke $(\mathcal{O}(n^2 \log n))$ dengan biaya memori lebih tinggi. Divisive dengan pencarian exhaustive adalah $(\mathcal{O}(2^n))$, tetapi heuristik seperti K-Means sering digunakan untuk mempercepat split [3]. Persamaan Ward, misalnya, meminimalkan peningkatan sum of squares:$(\sum_{x\in A\cup B}|x-\mu_{A\cup B}|^2-\sum_{x\in A}|x-\mu_A|^2-\sum_{x\in B}|x-\mu_B|^2)$ (6).

Metrik evaluasi klasterisasi yang digunakan adalah:

1. **Silhouette Score** yang bertujuan mengukur **cohesion** (kekompakan intra-cluster) dan **separation** (pemisahan antar-cluster) [8].
   Rumus:
   $s(i)=\frac{b(i)-a(i)}{\max(a(i),b(i))}$ (7)
   - $a(i)a(i)$: Rata-rata jarak dari titik data ii ke semua titik lain dalam **cluster yang sama**.
   - $b(i)b(i)$: Rata-rata jarak dari titik data ii ke titik-titik di **cluster terdekat yang berbeda**.

   Nilai Silhouette Score berada dalam rentang **[-1, 1]**:
   - **Mendekati 1** → Titik terklaster dengan baik (cohesive & well-separated).
   - **Mendekati 0** → Titik berada di batas antara dua cluster.
   - **Negatif** → Titik lebih dekat ke cluster lain daripada ke clusternya sendiri (buruk).

2. **Davies-Bouldin Index (DBI)** yang bertujuan Mengevaluasi trade-off antara **kompaknya**
   cluster dan **jarak antar cluster** (separation)[8].
   Rumus:
   $DBI=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{S_i+S_j}{M_{ij}}\right)$ (8)
   - Si: Rata-rata jarak antar titik dalam cluster ii ke centroid-nya (kompaksi).
   - M_{ij}: Jarak antar centroid cluster ii dan jj.

   Interpretasi:
   - **Nilai lebih kecil** → Cluster yang lebih kompak dan lebih terpisah → lebih baik.

- Nilai optimal mendekati **0**.

3. **Calinski-Harabasz Index (CH Index)** yang bertujuan Mengukur rasio antara **variansi antar-cluster (inter)** terhadap **variansi dalam cluster (intra**[8]**).**
Rumus:
$$CH = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \cdot \frac{n-k}{k-1} \quad (9)$$
- $\mathrm{Tr}(B_k)$: Trace dari matrix variansi antar-cluster.
- $\mathrm{Tr}(W_k)$: Trace dari matrix variansi dalam-cluster.
- nn: Jumlah total sampel.
- kk: Jumlah cluster.

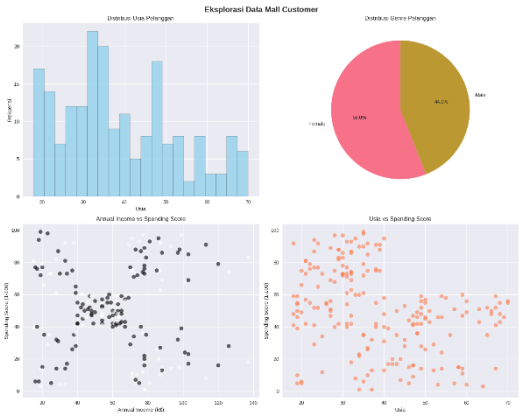
Interpretasi:
- **Nilai lebih tinggi** → Cluster lebih terpisah dan lebih kompak → lebih baik.

4. **Computational Time** tujuannya adalah Mengukur **efisiensi algoritma klasterisasi**, terutama pada **skala besar atau real-time application.**
Parameter yang memengaruhi waktu komputasi:
- **Dimensi data (feature space)**
- **Jumlah data (sample size)**
- **Jumlah cluster**
- **Algoritma yang digunakan** (misal: K-Means, DBSCAN, Agglomerative, dll.)
- **Kompleksitas waktu algoritma**: Misalnya:
  - K-Means → $O(n \cdot k \cdot t \cdot d)O(n \cdot k \cdot t \cdot d)$
  - Agglomerative → $O(n3)O(n^3)$ (dalam implementasi naïf)

Interpretasi:
- **Lebih cepat** → Lebih efisien, terutama untuk penggunaan industri.

HASIL DAN PEMBAHASAN

Hasil dari penelitian ini dirangkum dalam tabel-tabel dan gambar-gambar berikut ini :



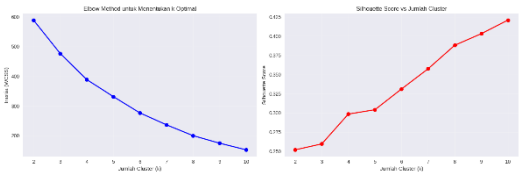
Gambar 1. *Exploratory Data Analysis*

Gambar 1 menunjukkan eksplorasi data pelanggan mall melalui beberapa visualisasi yang berbeda, yang mencakup distribusi usia pelanggan, distribusi gender pelanggan, serta hubungan antara usia dan skor pengeluaran, serta pendapatan tahunan dan skor pengeluaran.

1. **Distribusi Usia Pelanggan:** Visualisasi pertama adalah histogram yang menunjukkan distribusi usia pelanggan. Sumbu x mewakili usia (dalam rentang tertentu), sedangkan sumbu y menunjukkan frekuensi. Histogram ini menunjukkan bahwa usia pelanggan tersebar dengan puncak frekuensi yang signifikan di kisaran usia 30 hingga 40 tahun, dengan beberapa puncak lain di sekitar usia 20 dan 50 tahun. Ini menunjukkan bahwa mayoritas pelanggan mall berada di kelompok usia dewasa muda hingga paruh baya.
2. **Distribusi Gender Pelanggan:** Visualisasi kedua adalah diagram lingkaran yang menggambarkan distribusi gender pelanggan. Sebanyak 56% pelanggan adalah perempuan (ditunjukkan dengan warna merah muda), sementara 44% adalah laki-laki (ditunjukkan dengan warna kuning kecokelatan). Hal ini menunjukkan bahwa perempuan sedikit lebih dominan sebagai pelanggan mall dibandingkan laki-laki.

3. **Annual Income vs Spending Score:** Visualisasi ketiga adalah scatter plot yang menunjukkan hubungan antara pendapatan tahunan (sumbu x, dalam ribuan dolar) dan skor pengeluaran (sumbu y, skala 0-100). Data menunjukkan adanya klasterisasi yang jelas. Pelanggan dengan pendapatan tahunan menengah (sekitar 40-60 ribu dolar) memiliki skor pengeluaran yang bervariasi, dengan konsentrasi tinggi di skor 40-60. Ada juga kelompok kecil pelanggan dengan pendapatan tinggi (100-140 ribu dolar) yang memiliki skor pengeluaran rendah, serta kelompok lain dengan pendapatan rendah hingga menengah yang memiliki skor pengeluaran tinggi (sekitar 80-100).
4. **Usia vs Spending Score:** Visualisasi keempat adalah scatter plot lain yang menunjukkan hubungan antara usia (sumbu x) dan skor pengeluaran (sumbu y). Data ini menunjukkan distribusi yang lebih merata, dengan konsentrasi titik di kisaran usia 20-50 tahun dan skor pengeluaran 40-60. Ada juga beberapa titik outlier, seperti pelanggan dengan usia lebih dari 60 tahun yang memiliki skor pengeluaran tinggi, serta pelanggan muda (di bawah 20 tahun) dengan skor pengeluaran rendah.

Secara keseluruhan, eksplorasi data ini memberikan gambaran bahwa pelanggan mall didominasi oleh kelompok usia 30-40 tahun, dengan perempuan sedikit lebih banyak daripada laki-laki. Pola pengeluaran tampaknya tidak sepenuhnya bergantung pada pendapatan tahunan, dengan beberapa kelompok pelanggan menunjukkan perilaku pengeluaran yang tinggi meskipun pendapatan mereka rendah hingga menengah.



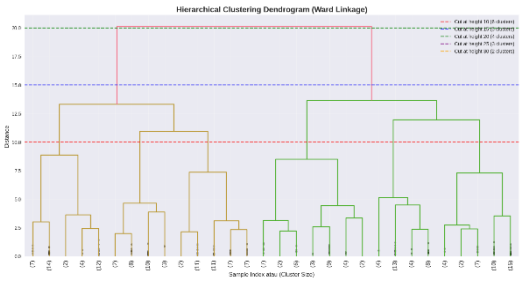
Gambar 2. *Grafik Elbow dan Silhoutte Score*

Gambar 2 menampilkan dua grafik yang digunakan untuk menentukan jumlah klaster optimal dalam analisis klaster, yaitu metode Elbow dan Silhouette Score. Kedua grafik memiliki sumbu x yang mewakili jumlah klaster (k) dari 2 hingga 10.

1. **Elbow Method untuk Menentukan k Optimal:** Grafik pertama menunjukkan metode Elbow dengan sumbu y yang mengukur inertia (Within-Cluster Sum of Squares, WCSS) dalam skala 200 hingga 600. Kurva berwarna biru menurun secara tajam dari k=2 (sekitar 600) hingga k=5 (sekitar 300), lalu melandai secara bertahap hingga k=10. Titik "siku" (elbow) tampaknya terjadi di sekitar k=5, di mana penurunan inertia mulai melambat. Ini menunjukkan bahwa k=5 bisa menjadi jumlah klaster optimal, karena menawarkan keseimbangan antara kompleksitas model dan variasi dalam data yang dijelaskan.
2. **Silhouette Score vs Jumlah Cluster:** Grafik kedua menampilkan Silhouette Score dengan sumbu y yang berkisar dari 0.250 hingga 0.425. Kurva berwarna merah menunjukkan peningkatan skor seiring bertambahnya jumlah klaster, dengan kenaikan signifikan dari k=2 (sekitar 0.250) hingga k=5 (sekitar 0.300), diikuti oleh kenaikan lebih lambat hingga k=10 (sekitar 0.425). Skor Silhouette yang lebih tinggi menunjukkan kohesi dalam klaster yang lebih baik dan pemisahan antar klaster yang lebih jelas. Meskipun skor terus meningkat, peningkatan setelah k=5 relatif kecil, yang

mendukung k=5 sebagai pilihan yang wajar.

Secara keseluruhan, kedua metode ini konsisten menyarankan bahwa jumlah klaster optimal kemungkinan besar adalah 5, berdasarkan titik siku pada metode Elbow dan stabilitas relatif skor Silhouette setelah k=5. Jumlah cluster optimal berdasarkan Silhouette Score = 10, Silhouette Score tertinggi = 0.421.



Gambar 3. *Dendogram*

Dendrogram yang ditampilkan merupakan visualisasi dari proses hierarchical clustering menggunakan metode Ward Linkage, yang merupakan salah satu pendekatan agglomerative clustering di mana klaster digabungkan secara bertahap berdasarkan minimisasi varians dalam klaster (dikenal sebagai Ward's method). Dendrogram ini menggambarkan struktur hierarki penggabungan sampel data, di mana sumbu vertikal (y-axis) mewakili "Distance" atau jarak penggabungan (dalam skala Euclidean atau metrik serupa, dengan nilai dari 0 hingga sekitar 20), sedangkan sumbu horizontal (x-axis) menampilkan "Sample Index atau (Cluster Size)", yang menunjukkan sampel individu atau klaster awal beserta ukurannya, seperti C1(4), C2(3), C3(2), dan seterusnya hingga C10(3). Setiap cabang vertikal mewakili penggabungan dua klaster atau sampel pada tingkat jarak tertentu, membentuk struktur pohon yang semakin menyatu ke atas.

Proses clustering dimulai dari bawah, di mana setiap sampel dianggap sebagai klaster tunggal (leaf nodes), dan secara iteratif digabungkan menjadi klaster yang lebih besar hingga semua sampel berada dalam satu klaster tunggal di puncak. Ketinggian cabang menunjukkan jarak di mana penggabungan terjadi—jarak yang lebih rendah berarti kesamaan yang lebih tinggi antar sampel. Dendrogram ini menyoroti beberapa garis potong (cut lines) horizontal berwarna yang menandai kemungkinan jumlah klaster optimal pada berbagai ketinggian:

- **Garis merah** pada ketinggian 10, menghasilkan 6 klaster.
- **Garis biru** pada ketinggian 15, menghasilkan 5 klaster.
- **Garis ungu** pada ketinggian 20, menghasilkan 4 klaster.
- **Garis hijau** pada ketinggian 25, menghasilkan 3 klaster.
- **Garis kuning** pada ketinggian 30, menghasilkan 2 klaster.

Struktur dendrogram menunjukkan penggabungan yang relatif cepat pada jarak rendah (di bawah 5), di mana klaster kecil terbentuk dari sampel yang sangat mirip, diikuti oleh penggabungan yang lebih lambat pada jarak lebih tinggi, menandakan adanya kelompok yang lebih heterogen. Ada beberapa cabang utama yang terlihat: satu kelompok besar di sebelah kiri (mencakup C1 hingga C4 atau sekitarnya), kelompok tengah (sekitar C5 hingga C7), dan kelompok kanan (C8 hingga C10), yang menyatu pada jarak sekitar 15-20.

Analisis dendrogram ini berguna untuk menentukan jumlah klaster optimal dalam data, terutama dalam konteks eksplorasi data pelanggan mall seperti yang dibahas sebelumnya (misalnya, berdasarkan usia, gender, pendapatan tahunan, dan skor pengeluaran). Metode hierarchical clustering dengan Ward Linkage dipilih karena efektif dalam mendeteksi klaster dengan ukuran dan bentuk yang bervariasi, sambil meminimalkan distorsi varians.

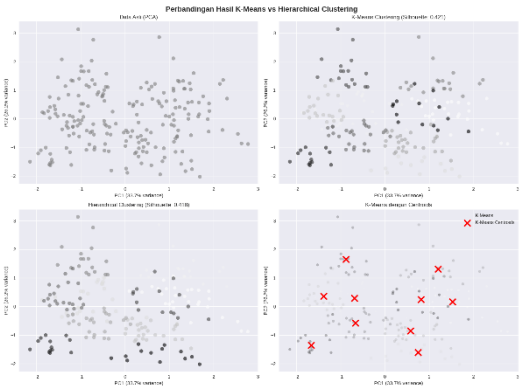
1. **Penentuan Jumlah Klaster Optimal**:
   - Dendrogram sering dianalisis dengan mencari "gap" atau jeda

signifikan dalam ketinggian penggabungan, di mana pemotongan (cut) dilakukan untuk mendapatkan klaster yang kohesif. Di sini, jeda terlihat jelas sekitar ketinggian 10-15, di mana penggabungan besar terjadi setelah klaster kecil terbentuk. Garis biru pada ketinggian 15 (5 klaster) tampak sebagai titik optimal karena berada di tengah jeda ini—di bawahnya, klaster mungkin terlalu terfragmentasi (misalnya, 6 klaster pada ketinggian 10 mungkin over-segmentasi), sementara di atasnya (seperti 4 klaster pada ketinggian 20) mungkin menggabungkan kelompok yang seharusnya terpisah.

- Ini konsisten dengan metode lain seperti Elbow Method (yang menunjukkan "siku" di k=5) dan Silhouette Score (yang stabil atau optimal sekitar k=5), sebagaimana terlihat pada visualisasi sebelumnya. Pemilihan 5 klaster menawarkan keseimbangan antara granularitas (membedakan pola perilaku pelanggan yang berbeda) dan kesederhanaan (menghindari overfitting).

2. **Struktur Klaster dan Interpretasi**:
   - **Klaster Kecil (Leaf Nodes)**: Pada jarak rendah (0-5), sampel digabungkan menjadi klaster kecil dengan ukuran 2-4 (seperti C1(4), C2(3)). Ini menunjukkan adanya sub-kelompok alami dalam data, mungkin mewakili segmen pelanggan dengan karakteristik serupa, seperti usia muda dengan skor pengeluaran tinggi atau pendapatan menengah dengan preferensi gender tertentu.
   - **Penggabungan Utama**: Pada jarak 5-10, klaster ini menyatu menjadi 6 kelompok lebih besar, yang bisa mencerminkan pola seperti: (1) Pelanggan muda berpenghasilan rendah tapi boros, (2) Pelanggan paruh baya berpenghasilan tinggi tapi hemat, dll. (berdasarkan scatter plot sebelumnya yang menunjukkan klasterisasi pada annual income vs spending score).
   - **Heterogenitas**: Jarak penggabungan yang meningkat tajam di atas 15 menandakan bahwa menggabungkan menjadi kurang dari 5 klaster akan menyatukan kelompok yang kurang mirip, potensial mengurangi nilai analitik (misalnya, untuk strategi marketing targeted).

Secara keseluruhan, dendrogram ini memperkuat rekomendasi untuk menggunakan 5 klaster sebagai basis segmentasi pelanggan mall, yang dapat diterapkan untuk strategi bisnis seperti personalisasi promosi berdasarkan usia dan pengeluaran.



Gambar 4. *Perbandingan Hasil K-Means vs Hierarchical Clustering*

Gambar 4 terdiri dari empat panel yang masing-masing menggambarkan tahap dan hasil analisis. Panel pertama

menampilkan data asli pasca-PCA (Principal Component Analysis), di mana PC1 (menjelaskan 33.7% variansi) dan PC2 (menjelaskan 20.2% variansi) digunakan sebagai sumbu, memberikan representasi dua dimensi yang menangkap sekitar 53.9% total variansi data. Titik-titik dalam panel ini tersebar tanpa label klaster, menunjukkan distribusi awal yang menampilkan pola klaster alami, seperti konsentrasi di kuadran kiri bawah (kemungkinan pelanggan dengan pendapatan rendah tetapi pengeluaran tinggi) dan kanan atas (pendapatan tinggi dengan pengeluaran rendah).

Panel kedua menggambarkan hasil K-Means Clustering dengan Silhouette Score sebesar 0.421, menunjukkan kualitas klaster yang moderat hingga baik (nilai mendekati 1 mengindikasikan klaster yang sangat kohesif dan terpisah). Variasi intensitas titik abu-abu mengindikasikan kepadatan intra-klaster, dengan lima kelompok utama yang teridentifikasi berdasarkan iterasi pengoptimalan centroid.

Panel ketiga menampilkan hasil Hierarchical Clustering dengan Silhouette Score 0.418, yang menunjukkan performa serupa, dengan struktur klaster yang konsisten dengan K-Means, mengindikasikan bahwa kedua metode menghasilkan segmentasi yang komparabel.

Panel keempat menyajikan hasil K-Means dengan penanda centroid (simbol silang merah), yang mengkonfirmasi adanya lima klaster utama. Posisi centroid ini mencerminkan pusat-pusat klaster yang dapat diinterpretasikan sebagai segmen pelanggan, seperti pelanggan hemat berpenghasilan rendah (kiri bawah), pelanggan hemat berpenghasilan tinggi (kanan bawah), pelanggan standar berpenghasilan dan pengeluaran sedang (tengah), pelanggan boros berpenghasilan sedang (tengah atas), dan pelanggan boros berpenghasilan tinggi (kanan atas). Konsistensi jumlah klaster (k=5) dengan analisis sebelumnya, seperti Elbow Method dan dendrogram, memperkuat validitas pilihan parameter ini.

Secara keseluruhan, visualisasi ini mengilustrasikan bahwa kedua algoritma clustering menghasilkan segmentasi pelanggan yang serupa pada dataset Mall Customers, dengan Silhouette Score yang mendekati 0.42, menunjukkan kohesi intra-klaster yang memadai dan separasi inter-klaster yang cukup baik. Perbedaan kecil dalam skor Silhouette (0.003) menegaskan bahwa kedua metode memiliki kinerja yang setara dalam konteks data ini, dengan K-Means menawarkan efisiensi komputasi dan Hierarchical Clustering memberikan wawasan hierarkis yang lebih mendalam. Implikasi praktisnya mencakup potensi pengembangan strategi pemasaran berbasis segmen, seperti promosi khusus untuk klaster boros atau penawaran eksklusif untuk segmen premium.

Analisis perbandingan performa algoritma *clustering* K-Means dan Hierarchical Clustering dilakukan dengan menggunakan empat metrik evaluasi utama, yaitu **Silhouette Score**, **Davies-Bouldin Index**, **Calinski-Harabasz Index**, dan **waktu komputasi**. Berdasarkan hasil evaluasi, K-Means menghasilkan **Silhouette Score sebesar 0.4208**, sedikit lebih tinggi dibandingkan Hierarchical Clustering yang memperoleh nilai **0.4176**. Nilai ini menunjukkan bahwa K-Means memiliki kemampuan yang lebih baik dalam membentuk klaster yang kohesif dan terpisah secara internal.

Sebaliknya, Hierarchical Clustering mencatatkan performa yang lebih baik pada **Davies-Bouldin Index**, dengan nilai **0.8247** dibandingkan **0.8331** pada K-Means. Karena indeks ini mengukur kompaksi dan pemisahan klaster (dengan nilai lebih kecil lebih baik), maka hasil ini mengindikasikan bahwa Hierarchical Clustering membentuk klaster yang lebih rapat dan terpisah satu sama lain.

Selanjutnya, pada metrik **Calinski-Harabasz Index**, K-Means kembali menunjukkan keunggulan dengan skor **89.9782**, lebih tinggi dibandingkan Hierarchical Clustering yang memperoleh

87.6113. Nilai indeks ini yang lebih tinggi menunjukkan bahwa K-Means memiliki variansi antar-klaster yang lebih besar relatif terhadap variansi intra-klaster, sehingga segmentasi yang dihasilkan lebih terstruktur.

Dari sisi efisiensi, **Hierarchical Clustering menunjukkan waktu komputasi yang jauh lebih cepat**, yakni hanya **0.0033 detik**, dibandingkan dengan **0.0325 detik** pada K-Means. Ini menandakan bahwa untuk kasus dengan dataset yang tidak terlalu besar, metode Hierarchical dapat menjadi alternatif efisien secara waktu.

Secara keseluruhan, kedua algoritma memiliki kelebihan masing-masing: **K-Means unggul dalam kualitas segmentasi**, sementara **Hierarchical Clustering lebih efisien dalam komputasi dan memberikan struktur klaster yang lebih mendalam.**

Profil Cluster K-Means

Algoritma K-Means menghasilkan **10 cluster** dengan karakteristik yang beragam. Beberapa temuan menonjol antara lain:

- **Cluster 3 dan 7** memiliki profil pelanggan dengan **pendapatan tinggi** (86.05 dan 87.11) dan **pengeluaran tinggi** (81.67 dan 82.67), namun keduanya **didominasi oleh jenis kelamin berbeda**: Cluster 3 berisi 100% perempuan, sedangkan Cluster 7 sepenuhnya laki-laki.
- **Cluster 8** menampilkan kelompok pelanggan **muda (rata-rata umur 25.46)** dengan **pendapatan rendah (25.69)** namun **pengeluaran tinggi (80.54)** dan 100% perempuan, yang dapat dikategorikan sebagai konsumen dengan kecenderungan gaya hidup konsumtif.
- Sebaliknya, **Cluster 5 dan 9** menunjukkan karakteristik dengan **pengeluaran rendah** (14.21 dan 20.64), meskipun pendapatan mereka relatif tinggi (85.89 dan 93.29), menunjukkan pola konsumsi yang hemat atau konservatif.

Secara umum, K-Means mampu membedakan segmen berdasarkan kombinasi umur, penghasilan, dan perilaku pengeluaran secara tajam, termasuk memisahkan segmen berdasar jenis kelamin secara eksplisit dalam beberapa cluster.

Profil Cluster Hierarchical

Hierarchical Clustering juga membentuk **10 cluster** yang memperlihatkan variasi karakteristik yang mirip, namun dengan pembagian yang sedikit berbeda:

- **Cluster 5 dan 6** memperlihatkan profil dengan **pendapatan dan pengeluaran sangat tinggi** (Income_Mean > 86 dan Spending_Mean > 81), dengan perbedaan pada jenis kelamin: Cluster 5 terdiri dari 100% laki-laki, sementara Cluster 6 dari 100% perempuan.
- **Cluster 8** (umur 25.46, income 25.69, spending 80.54) dan **Cluster 2** (umur 24.57, income 39.22, spending 59.65) menunjukkan bahwa algoritma ini juga mengenali kelompok muda berpendapatan rendah namun memiliki pola konsumsi tinggi.
- Sementara itu, **Cluster 1 dan 9** mencerminkan kelompok dengan **pendapatan tinggi namun pengeluaran rendah** (misalnya Cluster 1: income 86.39, spending 11.67), mirip dengan yang ditemukan dalam segmentasi K-Means.

Hierarchical Clustering secara umum mampu menangkap pola yang serupa dengan K-Means, meskipun terdapat perbedaan dalam pembagian klaster dan variansi antar kelompok. Kecenderungan algoritma ini menghasilkan pembagian yang lebih halus terlihat dari nilai standar deviasi yang sedikit lebih tinggi pada beberapa variabel.

Interpretasi Cluster dan Rekomendasi Bisnis

Dalam analisis segmentasi pelanggan, baik metode **K-Means** maupun **Hierarchical Clustering** berhasil mengidentifikasi berbagai kelompok pelanggan berdasarkan usia, pendapatan, kebiasaan belanja (spending score), dan jenis kelamin. Dari hasil clustering ini, kita dapat menyusun strategi pemasaran yang lebih tepat sasaran dan berbasis data.

Segmen Moderate Customers

Kelompok pelanggan terbesar berada pada kategori **Moderate Customers** yang muncul secara konsisten di berbagai cluster, baik di K-Means maupun Hierarchical. Kelompok ini memiliki karakteristik rata-rata usia menengah, pendapatan sedang, dan tingkat belanja yang moderat. Rekomendasi strategi untuk segmen ini adalah pendekatan **balanced**, dengan menawarkan berbagai produk **mid-range** yang dapat menjangkau kebutuhan umum mereka.

Budget Conscious & Young Spenders

Segmen **Budget Conscious** memiliki pendapatan dan spending score yang rendah. Namun, kelompok ini—yang didominasi oleh perempuan—masih memiliki potensi jika ditawarkan **produk dengan harga terjangkau dan penawaran bernilai (value deals)**. Sementara itu, **Young Spenders** adalah kelompok muda dengan penghasilan rendah, namun memiliki kecenderungan belanja tinggi. Mereka sangat responsif terhadap **promosi menarik dan produk affordable**, sehingga strategi pemasaran kreatif sangat efektif untuk mereka.

Premium Customers

Segmen **Premium Customers** sangat menjanjikan karena memiliki pendapatan dan spending score yang tinggi. Mereka tersebar di beberapa cluster dan cenderung berusia produktif. Kelompok ini harus menjadi target utama untuk penawaran **produk premium, layanan eksklusif, serta program loyalitas VIP**, guna meningkatkan retensi dan kepuasan pelanggan.

High Income, Low Spenders

Menariknya, beberapa cluster menampilkan karakteristik pelanggan dengan **pendapatan tinggi namun spending score rendah**. Ini menandakan adanya **potensi pasar yang belum tergarap secara optimal**. Strategi yang disarankan adalah menciptakan engagement yang lebih kuat, edukasi nilai produk, dan pendekatan personal untuk mendorong tingkat pembelanjaan mereka.

INTERPRETASI CLUSTER DAN REKOMENDASI BISNIS

INTERPRETASI CLUSTER K-MEANS

Cluster 0

- **Rata-rata usia:** 58.9 tahun.
- **Rata-rata pendapatan:** $49K.
- **Rata-rata spending score:** 39.9.
- **Persentase perempuan:** 0.0%.
- **Interpretasi:** *Moderate Customers* — Segmen menengah.
- **Rekomendasi:** Strategi seimbang. dengan variasi produk mid-range.

Cluster 1

- **Rata-rata usia:** 25.2 tahun.
- **Rata-rata pendapatan:** $41K.
- **Rata-rata spending score:** 60.9.
- **Persentase perempuan:** 0.0%.
- **Interpretasi:** *Moderate Customers* — Segmen menengah.
- **Rekomendasi:** Strategi seimbang. dengan variasi produk mid-range.

Cluster 2

- **Rata-rata usia:** 41.2 tahun.
- **Rata-rata pendapatan:** $26K.
- **Rata-rata spending score:** 20.1.
- **Persentase perempuan:** 92.9%.

- **Interpretasi:** *Budget Conscious* — Segmen ekonomis.
- **Rekomendasi:** Tawarkan value deals dan produk dengan harga rendah.

Cluster 3
- **Rata-rata usia:** 32.2 tahun.
- **Rata-rata pendapatan:** $86K.
- **Rata-rata spending score:** 81.7.
- **Persentase perempuan:** 100.0%.
- **Interpretasi:** *Premium Customers* — Pendapatan & pengeluaran tinggi.
- **Rekomendasi:** Target produk premium dan layanan VIP.

Cluster 4
- **Rata-rata usia:** 54.1 tahun.
- **Rata-rata pendapatan:** $54K.
- **Rata-rata spending score:** 49.0.
- **Persentase perempuan:** 100.0%.
- **Interpretasi:** *Moderate Customers* — Segmen menengah.
- **Rekomendasi:** Strategi seimbang dengan variasi produk mid-range.

Cluster 5
- **Rata-rata usia:** 38.5 tahun.
- **Rata-rata pendapatan:** $86K.
- **Rata-rata spending score:** 14.2.
- **Persentase perempuan:** 0.0%.
- **Interpretasi:** *High Income, Low Spenders* — Potensi besar.
- **Rekomendasi:** Strategi khusus untuk meningkatkan engagement dan spending.

Cluster 6
- **Rata-rata usia:** 28.0 tahun
- **Rata-rata pendapatan:** $57K
- **Rata-rata spending score:** 47.1
- **Persentase perempuan:** 100.0%
- **Interpretasi:** *Moderate Customers* — Segmen menengah.
- **Rekomendasi:** Strategi seimbang dengan variasi produk mid-range.

Cluster 7
- **Rata-rata usia:** 33.3 tahun.
- **Rata-rata pendapatan:** $87K.
- **Rata-rata spending score:** 82.7.
- **Persentase perempuan:** 0.0%.
- **Interpretasi:** *Premium Customers* — Pendapatan & pengeluaran tinggi.
- **Rekomendasi:** Target produk premium dan layanan VIP.

Cluster 8
- **Rata-rata usia:** 25.5 tahun.
- **Rata-rata pendapatan:** $26K.
- **Rata-rata spending score:** 80.5.
- **Persentase perempuan:** 100.0%.
- **Interpretasi:** *Young Spenders* — Pendapatan rendah, belanja tinggi
- **Rekomendasi:** Fokus pada produk terjangkau dan promosi menarik.

Cluster 9
- **Rata-rata usia:** 43.8 tahun.
- **Rata-rata pendapatan:** $93K.
- **Rata-rata spending score:** 20.6.
- **Persentase perempuan:** 100.0%.
- **Interpretasi:** *High Income, Low Spenders* — Potensi besar.
- **Rekomendasi:** Strategi khusus untuk meningkatkan engagement dan spending.

INTERPRETASI CLUSTER HIERARCHICAL

Cluster 0
- **Rata-rata usia:** 56.5 tahun.
- **Rata-rata pendapatan:** $50K.
- **Rata-rata spending score:** 41.3.
- **Persentase perempuan:** 0.0%.
- **Interpretasi:** *Moderate Customers* — Segmen menengah.
- **Rekomendasi:** Strategi seimbang dengan variasi produk mid-range.

Cluster 1
- **Rata-rata usia:** 38.8 tahun.
- **Rata-rata pendapatan:** $86K.
- **Rata-rata spending score:** 11.7.

- **Persentase perempuan:** 0.0%.
- **Interpretasi:** *High Income, Low Spenders* — Potensi besar.
- **Rekomendasi:** Strategi khusus untuk meningkatkan engagement dan spending.

Cluster 2

- **Rata-rata usia:** 24.6 tahun.
- **Rata-rata pendapatan:** $39K.
- **Rata-rata spending score:** 59.6.
- **Persentase perempuan:** 0.0%.
- **Interpretasi:** *Moderate Customers* — Segmen menengah.
- **Rekomendasi:** Strategi seimbang dengan variasi produk mid-range.

Cluster 3

- **Rata-rata usia:** 54.1 tahun.
- **Rata-rata pendapatan:** $53K.
- **Rata-rata spending score:** 49.5.
- **Persentase perempuan:** 100.0%.
- **Interpretasi:** *Moderate Customers* — Segmen menengah.
- **Rekomendasi:** Strategi seimbang dengan variasi produk mid-range.

Cluster 4

- **Rata-rata usia:** 28.0 tahun.
- **Rata-rata pendapatan:** $57K.
- **Rata-rata spending score:** 47.1.
- **Persentase perempuan:** 100.0%.
- **Interpretasi:** *Moderate Customers* — Segmen menengah.
- **Rekomendasi:** Strategi seimbang dengan variasi produk mid-range.

Cluster 5

- **Rata-rata usia:** 33.3 tahun.
- **Rata-rata pendapatan:** $87K.
- **Rata-rata spending score:** 82.7.
- **Persentase perempuan:** 0.0%.
- **Interpretasi:** *Premium Customers* — Pendapatan & pengeluaran tinggi.
- **Rekomendasi:** Target produk premium dan layanan VIP

Cluster 6

- **Rata-rata usia:** 32.2 tahun.
- **Rata-rata pendapatan:** $86K.
- **Rata-rata spending score:** 81.7.
- **Persentase perempuan:** 100.0%.
- **Interpretasi:** *Premium Customers* — Pendapatan & pengeluaran tinggi.
- **Rekomendasi:** Target produk premium dan layanan VIP.

Cluster 7

- **Rata-rata usia:** 44.6 tahun.
- **Rata-rata pendapatan:** $92K.
- **Rata-rata spending score:** 21.6.
- **Persentase perempuan:** 100.0%.
- **Interpretasi:** *High Income, Low Spenders* — Potensi besar.
- **Rekomendasi:** Strategi khusus untuk meningkatkan engagement dan spending.

Cluster 8

- **Rata-rata usia:** 25.5 tahun.
- **Rata-rata pendapatan:** $26K.
- **Rata-rata spending score:** 80.5.
- **Persentase perempuan:** 100.0%.
- **Interpretasi:** *Young Spenders* — Pendapatan rendah, belanja tinggi.
- **Rekomendasi:** Fokus pada produk terjangkau dan promosi menarik.

Cluster 9

- **Rata-rata usia:** 41.5 tahun.
- **Rata-rata pendapatan:** $27K.
- **Rata-rata spending score:** 20.7.
- **Persentase perempuan:** 100.0%.
- **Interpretasi:** *Budget Conscious* — Segmen ekonomis.
- **Rekomendasi:** Tawarkan value deals dan produk dengan harga rendah.

PENUTUP

Analisis perbandingan hasil clustering antara metode K-Means dan Hierarchical Clustering menunjukkan performa yang relatif seimbang dengan

keunggulan masing-masing. Dari sisi kualitas pemisahan dan kekompakan cluster, K-Means memiliki **Silhouette Score yang sedikit lebih tinggi** (0.4208) dibandingkan Hierarchical (0.4176), yang mengindikasikan bahwa K-Means sedikit lebih baik dalam membentuk cluster yang jelas terpisah dan kompak. Namun, dari nilai **Davies-Bouldin Index**, Hierarchical Clustering menunjukkan keunggulan dengan nilai yang lebih rendah (0.8247 dibandingkan 0.8331), yang berarti cluster yang dihasilkan lebih kompak dan terpisah secara relatif. Sementara itu, **Calinski-Harabasz Index** lebih tinggi pada K-Means (89.9782 vs. 87.6113), menunjukkan bahwa variasi antar cluster lebih besar relatif terhadap variasi dalam cluster, yang biasanya mengindikasikan kualitas segmentasi yang lebih baik. Di sisi efisiensi, **Hierarchical Clustering unggul secara signifikan dalam waktu komputasi**, hanya membutuhkan 0.0033 detik dibandingkan 0.0325 detik pada K-Means. Dengan demikian, pemilihan metode tergantung pada prioritas antara kualitas klasterisasi atau kecepatan eksekusi.

DAFTAR PUSTAKA

[1] V. Kumar and W. Reinartz, "Creating Enduring Customer Value," J Mark, vol. 80, no. 6, pp. 36–68, Nov. 2016, doi: 10.1509/jm.15.0414.

[2] J. Han, M. Kamber, and J. Pei, "Data mining: Concepts and," Techniques, Waltham: Morgan Kaufmann Publishers, 2012.

[3] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview, II," Wiley Interdiscip Rev Data Min Knowl Discov, vol. 7, no. 6, p. e1219, 2017.

[4] M. N. Sari, "Perbandingan Algoritma Clustering DBSCAN Dan K-Means Dalam Pengelompokan Siswa Terbaik," UIN Sumatera Utara, 2024.

[5] R. N. Arvi, "Analisis Segmentasi Pasar Berdasarkan Penjualan Produk Menggunakan Metode Clustering K-Means, K-Medoids Dan Agglomerative Hierarchical Clustering," Universitas Komputer Indonesia, 2024.

[6] M. A. W. Saputra, "Optimasi hasil evaluasi Clustering melalui kombinasi Algoritma Dynamic k-Means dan k-Means Binary Search Centroid," Universitas Islam Negeri Maulana Malik Ibrahim, 2023.

[7] S. Suhirman and H. Wintolo, "System for Determining Public Health Level Using the Agglomerative Hierarchical Clustering Method," Compiler, vol. 8, no. 1, pp. 95–104, 2019.

[8] R. L. Rohmah, D. C. Rini, and W. D. Utami, "Zonasi Daerah Terdampak Bencana Angin Puting Beliung Menggunakan K-means Clustering dengan Analisis Silhouette Coefficient, Davies Bouldin Index dan Purity," Davies Bouldin Index dan Purity (Doctoral dissertation, UIN Sunan Ampel Surabaya), 2019.