PERBANDINGAN BISECTING K-MEANS DAN HIERARCHICAL CLUSTERING PADA DATA CABAI RAWIT

Julius Juan [1)] Teny Handhayani [2)]

[1)] Teknik Informatika Universitas Tarumanagara
Jl. Taman S. Parman No.10, Tomang, Jakarta 11440, Indonesia
email : julius.535230078@stu.untar.ac.id
[2)] Teknik Informatika, Universitas Tarumanagara
Jl. Letjen S. Parman No. 1, Jakarta, 11440, Indonesia
email : tenyh@fti.untar.ac.id

ABSTRAK

Pengelompokan data merupakan salah satu pendekatan penting dalam analisis data baik di bidang pertanian maupun lainnya, untuk menemukan sebuah pola tertentu yang bermanfaat dalam menentukan keputusan. Penelitian ini bertujuan untuk membandingkan kinerja dari dua model klasterisasi yaitu Bisecting K-Means dan Hierarchical Clustering menggunakan data cabai rawit. Selain itu, data yang digunakan dibedakan menjadi 3 kelompok yaitu data luas panen, produksi, dan produktivitas dari komoditas cabe rawit Indonesia berdasarkan data yang didapatkan dari situs Basis Data Statistik Pertanian (BDSP) tahun 2010–2024 dengan total 515 sampel data dari berbagai provinsi di Indonesia. Metode yang digunakan untuk menganalisis data pada penelitian ini adalah Klasterisasi atau Clustering yang merupakan salah satu bagian dari Unsupervised Learning. Untuk evaluasinya, dilakukan dengan menguji jumlah klaster mulai dari 2 hingga 10 menggunakan tiga nilai metrik yaitu Silhouette Score, Davies-Bouldin Index, dan waktu komputasi.

Hasil dari penelitian ini, menunjukkan model Bisecting K-Means menghasilkan Silhouette Score tertinggi sebesar 0,9258 dan Davies-Bouldin Index sebesar 0,9698 pada klaster ke-2, namun membutuhkan waktu komputasi 1,6022 detik. Sedangkan model Hierarchical Clustering, menghasilkan Silhouette score terbaik sebesar 0,9100 dan Davies-Bouldin Index sebesar 0,8714 pada klaster ke-3 dengan waktu komputasi yang lebih cepat (0,0265 detik) dibandingkan model Bisecting K-Means. Dengan demikian, model Hierarchical Clustering memiliki keunggulan dari sisi waktu komputasi yang sangat cepat, sementara model Bisecting K-Means cenderung menghasilkan klaster yang lebih terpisah secara struktur dengan nilai evaluasi yang tinggi. Berdasarkan hasil tersebut, walaupun model Bisecting K-Means memiliki nilai evaluasi yang tinggi, namun model Hierarchical Clustering menunjukkan lebih sesuai digunakan dalam konteks analisis spasial dan pertanian berbasis data yang memerlukan hasil cepat dan akurat.

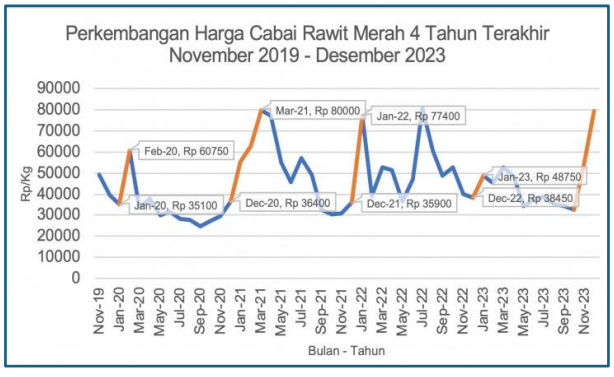
Key words

Bisecting K-Means, Cabe Rawit, Hierarchical Clustering, Klasterisasi, Unsupervised Learning

1. Pendahuluan

1.1 Latar Belakang

Pertanian memiliki peran penting dalam kemajuan ekonomi serta ketahanan pangan di tingkat nasional. Salah satu tanaman hortikultura yang memiliki nilai komersial tinggi dan stabilitas permintaan di pasar domestik adalah cabai rawit. Produksi cabai rawit di Indonesia mengalami fluktuasi yang dipengaruhi oleh berbagai faktor, termasuk perubahan iklim, serangan hama dan penyakit, teknik budidaya, serta aspek distribusi dan sarana pertanian [1], [2]. Ketidakseimbangan dalam distribusi hasil produksi dan produktivitas antar daerah memerlukan pendekatan yang didasarkan pada data untuk mendapatkan pemahaman yang lebih mendalam mengenai pola-pola tersembunyi dalam statistik produksi pertanian. Berikut ini gambar perubahan harga cabai rawit merah dari tahun 2019 hingga 2023 sebagai gambaran bentuk fluktuasi dari tahun ke tahun.



Gambar 1. Grafik Perkembangan Harga Cabai Rawit Merah

(Sumber: pangannews.com)

Penerapan metode klasterisasi dalam pembelajaran mesin telah menjadi salah satu cara yang efisien untuk menganalisis serta menggolongkan data pertanian berdasar karakteristik tertentu [3]. Klasterisasi sebagai salah satu elemen dari pembelajaran tanpa pengawasan (*Unsupervised Learning)*, memfasilitasi proses identifikasi struktur atau pola tanpa adanya label kelas, yang sangat bermanfaat untuk membagi wilayah tanam, memetakan produktivitas, serta merencanakan distribusi hasil panen [4]. Dua metode klasterisasi yang umum digunakan dalam penelitian adalah *Bisecting K-Means* dan *Clustering Hierarchical*. Setiap metode memiliki keunggulan dan kelemahan masing-masing yang penting untuk dibandingkan secara sistematis, terutama dalam konteks data spasial pertanian [5].

Penelitian sebelumnya menunjukkan bahwa kedua metode ini telah digunakan secara luas di berbagai bidang, mulai dari kesehatan, keuangan, hingga pertanian. Namun, masih sedikit studi yang secara khusus membandingkan kinerja Bisecting K-Means dan Clustering Hierarkis pada data jangka panjang komoditas pertanian seperti cabai rawit. Evaluasi dilakukan dengan menggunakan metrik Silhouette Score, Indeks Davies-Bouldin, dan waktu komputasi, untuk menentukan metode mana yang lebih baik dalam mendukung pengambilan keputusan berbasis data di sektor pertanian [4].

1.2 Penelitian Terdahulu

Beberapa penelitian telah menerapkan clustering di sektor pertanian. Misalnya, penggunaan K-Means dan K-Medians untuk segmentasi wilayah produksi cabai merah menunjukkan efektivitas metrik Davies-Bouldin dalam identifikasi zona optimal [6]. Selain itu, pendekatan partisi-rekursif seperti Bisecting K-Means mulai dipertimbangkan dalam analisis spasial, terutama ketika digabungkan dengan inisialisasi berbasis hirarki [7]. Di lain sisi, *Hierarchical Clustering* dengan metode *Ward* adalah teknik yang umum digunakan untuk struktur cluster bertingkat, meskipun memerlukan pertimbangan performa saat data berukuran besar [8]. Metode evaluasi internal seperti Silhouette Score dan Davies-Bouldin Index juga telah diakui sebagai standar dalam validasi kualitas cluster di berbagai domain, termasuk pertanian dan analisis citra satelit [9].

1.3 Rumusan Masalah

Pada penelitian ini, ditemukannya 3 permasalahan penting yang menjadi pembahasan utama yaitu:

1. Model apa yang lebih cocok dan memiliki performa terbaik dalam menganalisis data cabai rawit?
2. Apa kelebihan dan kekurangan dari model *Bisecting K-Means* dan *Hierarchical Clustering*?
3. Bagaimana hasil nilai evaluasi dari setiap model?

1.4 Tujuan

Penelitian ini bertujuan untuk membandingkan performa antara dua model *clustering* yaitu *Bisecting K-Means* dan *Hierarchical Clustering* dalam mengelompokkan data cabe rawit berdasarkan tiga indikator (luas panen, produksi, produktivitas) selama periode 2010–2024, untuk menentukan model yang paling terbaik untuk analisis data pertanian.

1.5 Manfaat

Hasil penelitian ini diharapkan dapat memberikan panduan bagi peneliti dan pembuat kebijakan dalam menentukan metode clustering yang paling sesuai untuk analisis spasial dan agribisnis berbasis data, serta meningkatkan presisi pengambilan keputusan wilayah.

1.6 Kebaruan

Terdapat beberapa perbedaan utama dari penelitian ini dengan yang lainnya yaitu:

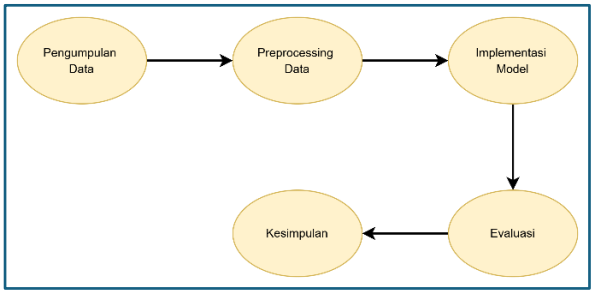
1. Komparasi langsung antara dua pendekatan berbeda, yaitu metode partisi-rekursif *Bisecting K-Means* dan metode hierarkis *Hierarchical Clustering*, yang jarang dibandingkan secara sistematis dalam konteks pertanian.
2. Evaluasi tidak hanya berfokus pada kualitas dan struktur klaster menggunakan metrik Silhouette Score dan Davies-Bouldin Index, tetapi juga memperhitungkan efisiensi waktu pemrosesan, yang penting dalam implementasi nyata untuk pengambilan keputusan cepat.
3. Analisis data jangka panjang mengenai produksi cabe rawit selama 14 tahun, yang memberikan kontribusi baru terhadap literatur mengenai aplikasi *clustering* dalam sektor pertanian Indonesia, khususnya pada komoditas dengan volatilitas tinggi seperti cabai rawit.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan *Machine Learning* dengan metode *Unsupervised Learning. Unsupervised Learning* adalah sebuah mesin atau komputer yang memiliki tujuan untuk mempelajari suatu pola dari kumpulan data tanpa merujuk pada respon tertentu [10]. Metode ini pun memiliki 2 jenis yaitu *clustering* dan *association*. Namun pada penelitian ini, yang akan digunakan adalah *clustering*, yaitu pembelajaran yang bertugas untuk menemukan pola tersembunyi pada suatu data masukan / *input* yang tidak memiliki label dalam bentuk klaster [11]. Setiap klaster berisikan data yang memiliki tingkat kemiripan yang tinggi dan diukur kemiripannya berdasarkan jarak antar

klaster [12]. Klasterisasi memiliki banyak manfaat di berbagai macam bidang seperti pemasaran, pertanian, kesehatan, keuangan, dan lain-lain. Selain itu, klasterisasi (*clustering*) memiliki berbagai macam algoritma seperti *k-means* yang merupakan salah satu algoritma terkenal dan sering kali digunakan [13].

Oleh karena itu, penelitian ini menggunakan algoritma *Bisecting K-Means* dan *Hierarchical Clustering* untuk memberikan hasil yang baru dan berbeda. Hal ini sejalan dengan penelitian terkini oleh Moseley & Wang (2020), yang menunjukkan bahwa *Bisecting K-Means* memiliki koneksi langsung dengan tujuan *Hierarchical Clustering* serta mampu mengoptimalkan struktur hierarkis secara global [14]. Metode tersebut akan digunakan pada bahasa pemrograman *Python,* alasannya karena cukup fleksibel untuk keperluan dalam menganalisis data [15]. Dengan mengetahui metode penelitiannya, diperlukannya tahap-tahap penelitian yang tepat sesuai dengan metode yang akan digunakan sebelum masuk ke pembahasan hasil evaluasi. Dapat dilihat pada gambar diagram alur tahap penelitian dibawah ini.



Gambar 2. Diagram Alir Penelitian

Pada Gambar 2, alur tersebut menerapkan kerangka kerja seperti CRISP-DM karena memiliki tahap yang sesuai untuk digunakan pada penelitian ini yaitu *Business Understanding* (tidak digunakan), *Data Understanding* (Pengumpulan Data), *Data Preparation* (Preprocessing Data), *Modeling* (Implementasi Model), *Evaluation* (Evaluasi), dan *Deployment* (Kesimpulan) [16]. Berikut penjelasan lengkap dari setiap tahap penelitian:

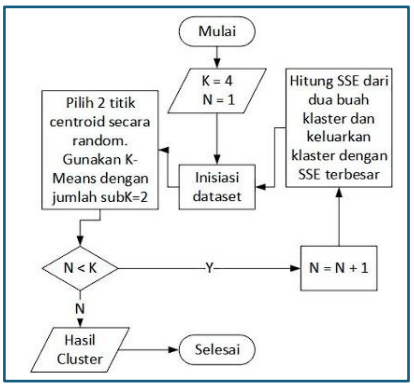
1. Pengumpulan Data: Menggunakan data dari situs BDSP bernama cabai rawit dan berjumlah 515 sampel.
2. Preprocessing Data: Data memiliki 3 indikator yaitu luas panen, produksi, dan produktivitas dari tahun 2010 hingga 2024. Dari 3 indikator tersebut dibagi menjadi 3 *sheet* atau lembar pada *file* excel, ditambah dengan lembar bernama Gabungan yang berisikan 3 indikator tersebut.
3. Implementasi Model:
   a. Bisecting K-Means: Menggunakan 2 parameter yaitu klaster (2-10), dan *n init* (500).
   b. Hierarchical Clustering: Terdapat 3 parameter yang digunakan yaitu klaster (2-10), *metric* (euclidean), dan *linkage* (ward).
4. Evaluasi: Hasil evaluasi yang digunakan untuk membandingkan model adalah Silhouette Score dan Davies-Bouldin Index.
5. Kesimpulan: Ringkasan dari hasil penelitian.

2.1. Bisecting K-Means

Model *Bisecting K-Means* merupakan bagian dari algoritma *K-Means*, yang digunakan pada penelitian ini karena memiliki kemampuan untuk menginisialisasi *centroid* secara acak dan dapat membagi satu *cluster* menjadi dua *sub-cluster* setiap langkah [17], [18]. Selain itu, *Bisecting K-Means* efektif untuk mengatasi suatu kondisi dimana algoritma memasuki kondisi optimal lokal hingga batas-batas tertentu. Berikut ini, terdapat tahap-tahap umum dari model *Bisecting K-Means*.

1. Menentukan *cluster* yang akan dibagi (*split)* [18].
2. Membagi klaster tersebut menjadi dua sub-klaster menggunakan *K-Means* (*bisecting)* [18].
3. Pemisahan beberapa kali (misalnya dua atau lebih iterasi awal), kemudian mengambil hasil pemisahan dengan nilai SSE total terkecil [18].
4. Mengulangi tahap-tahap sebelumnya hingga mencapai jumlah *cluster* yang sudah ditentukan [18].

Untuk pemahaman lebih mudahnya, dapat dilihat contoh diagram alir dari model *Bisecting K-Means* pada gambar berikut ini sebagai gambaran umum seperti apa prosesnya.



Gambar 3. Diagram Alir Bisecting K-Means

Dengan mengetahui diagram alir pada gambar 3, dapat melanjutkan dengan rumus dari model *Bisecting K-Means*.

1. Rumus untuk menemukan nilai SSE:

$$\text{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2 [19] \ (1)$$

2. Rumus pemilihan *cluster* terbaik:

$$C_{\text{best}} = \arg\min_{j} \text{SSE}\left(C_j\right) [19] \ (2)$$

SSE : *Sum of Squared Errors*
Ci : Klaster ke-i
Mi : *centroid dari klaster Ci*
X : *data point*
arg min : Memilih klaster dengan nilai SSE terkecil
Cbest : hasil pemisahan terbaik dari beberapa inisialisasi.

2.2. Hierarchical Clustering

Algoritma *Hierarchical Clustering* merupakan suatu proses pengelompokan data yang dilakukan dengan membuat suatu bagan hirarki atau disebut dengan dendrogram yang bertujuan untuk menampilkan kemiripan antara satu data dengan yang lain [20]. *Hierarchical Clustering* memiliki 2 tipe yaitu *agglomerative* dan *divisive* [21]. Tipe yang digunakan pada penelitian ini adalah *agglomerative*, yaitu penggabungan jumlah *cluster* hingga terbentuknya menjadi satu *cluster* [21]. Berikutnya, terdapat tahap-tahap dari algoritma *Hierarchical Clustering* yaitu:

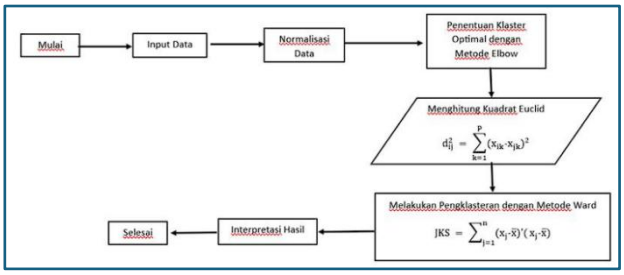
1. Menghitung jarak antar seluruh pasangan objek (*instance*) [22].
2. Mengidentifikasi dua objek atau klaster dengan jarak paling dekat [22].
3. Menggabungkan dua objek atau klaster terdekat menjadi satu klaster baru [22].
4. Menghitung ulang jarak antara klaster baru dengan klaster lainnya (*linkage*) [22].
5. Mengulangi tahap 2-4 hingga seluruh objek tergabung menjadi satu *cluster* besar [22].

Selain tahap dari algoritma *Hierarchical Clustering*, terdapat juga rumus untuk menghitung jarak antar objek atau disebut dengan *Euclidean*.

$$d_{(x,y)} = \sqrt{\sum_{j=1}^{p} \left(x_j - y_j\right)^2} \ [23] \ (3)$$

(x,y) : Jarak euclidean antara objek X dengan objek Y.
P : Banyaknya variabel yang diamati.
xj : Nilai *j* pada objek X.
yj : Nilai *j* pada objek Y [23].

Dengan mengetahui rumus diatas, dapat memahami langkah-langkah dari *Hierarchical Clustering Ward* berbentuk diagram alir yang dapat dilihat di bawah ini [24].



Gambar 4. Diagram Alir Hierarchical Clustering

2.3. Metode Evaluasi

Dari dua model yang telah dibahas sebelumnya, terdapat juga metode evaluasi yang digunakan untuk membandingkan hasil antar kedua model. Berikut ini, terdapat metode-metode evaluasi yang digunakan pada penelitian ini.

1. Silhouette Score: Merupakan nilai yang digunakan untuk mengevaluasi seberapa baik objek-objek dalam sebuah klaster terbentuk. Semakin mendekati nilai 1, semakin baik kualitas klaster tersebut. Sebaliknya, jika nilai Silhouette mendekati 0 atau negatif, maka kualitas klaster dianggap kurang baik atau buruk [25].

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \ (4)$$

- $a(i)$: rata-rata jarak objek i dengan semua titik dalam klaster yang sama.
- $b(i)$: rata-rata jarak objek i dengan semua titik dalam klaster terdekat.
- $s(i)$: nilai Silhouette untuk objek ke-i, berkisar antara -1 sampai 1.

2. Davies-Bouldin Index : Davies-Bouldin Index (DBI) adalah salah satu metrik evaluasi dalam klasterisasi yang digunakan untuk mengukur seberapa baik hasil klaster yang terbentuk. Metrik ini bertujuan untuk meminimalkan jarak antar data dalam satu klaster dan memaksimalkan jarak antar pusat klaster yang berbeda. Semakin kecil nilai DBI, maka semakin baik kualitas klasterisasi yang dihasilkan, karena klaster lebih kompak dan lebih terpisah satu sama lain. Rumus perhitungan DBI dapat dituliskan pada persamaan berikut ini [25].

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) [25] \ (5)$$

- k: jumlah total klaster.
- S_i: rata-rata jarak antara setiap titik dalam klaster i ke pusat klaster i.

- M_{ij}: jarak antara pusat klaster i dan j.
- $\frac{S_i+S_j}{M_{ij}}$: rasio dari gabungan kedekatan internal dua klaster terhadap jarak antar klaster.

3. Waktu komputasi: jumlah waktu pelatihan yang diperlukan untuk menjalankan suatu model.

3. Hasil dan Pembahasan

3.1 Dataset

Penelitian ini menggunakan data sekunder yang bersumber dari situs Basis Data Statistik Pertanian (BDSP). Dataset yang digunakan terdiri dari 515 sampel dan jangka waktu informasi mulai dari tahun 2010 hingga 2024. Data ini dipilih karena memiliki informasi historis yang cukup panjang sehingga memungkinkan analisis tren dan pola distribusi pertanian jangka panjang secara lebih mendalam. Terdapat tiga indikator utama dalam dataset ini, yaitu:

1. Luas Panen (dalam hektar): Menggambarkan seberapa luas lahan yang digunakan untuk budidaya cabai rawit.
2. Produksi (dalam ton): Menunjukkan total hasil panen yang dihasilkan dari lahan tersebut.
3. Produktivitas (dalam kuintal/hektar): Merupakan rasio hasil panen terhadap luas lahan, sebagai indikator efisiensi pertanian.

Ketiga indikator ini akan digunakan pada penelitian ini karena memiliki variabel yang umum digunakan dalam analisis performa sektor tanaman hortikultura, seperti cabai rawit. Data dikumpulkan berdasarkan unit administratif provinsi dan kabupaten/kota yang tersebar di seluruh wilayah Indonesia. Hal ini memungkinkan penerapan pendekatan spasial dalam analisis, untuk melihat perbedaan performa antar daerah dan mengelompokkannya ke dalam klaster.

Namun, terdapat beberapa kendala keterbatasan data dari sisi kelengkapan. Terdapat sejumlah provinsi hasil pemekaran baru seperti Papua Barat Daya, Papua Tengah, Papua Selatan, dan Papua Pegunungan yang belum memiliki data lengkap hingga level kabupaten/kota dalam kurun waktu yang diteliti. Oleh karena itu, provinsi-provinsi tersebut tidak akan digunakan untuk analisis. Sehingga, dari total 38 provinsi di Indonesia, hanya 34 provinsi yang digunakan dalam penelitian ini. Pendekatan ini dilakukan untuk menjaga konsistensi data dan keakuratan hasil analisis klasterisasi.

3.2 Hasil Evaluasi Bisecting K-Means

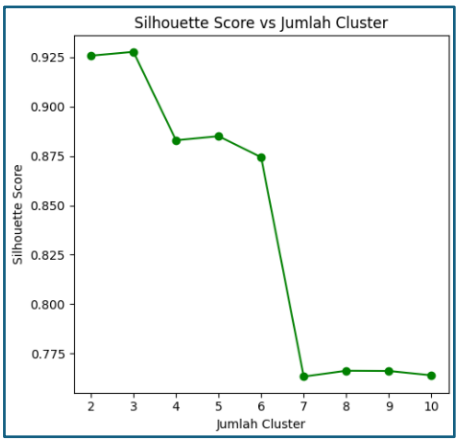
Hasil yang didapatkan dari pelatihan algoritma *Bisecting K-Means* menunjukkan performa yang cukup tinggi yang dimulai dari jumlah *cluster* 2. Berikut ini terdapat hasil evaluasi berbentuk tabel dari *Bisecting K-Means* dengan matrik *Silhouette Score*, *Davies-Bouldin Index*, dan waktu komputasi.

Tabel 1 Hasil Evaluasi Algoritma Bisecting K-Means

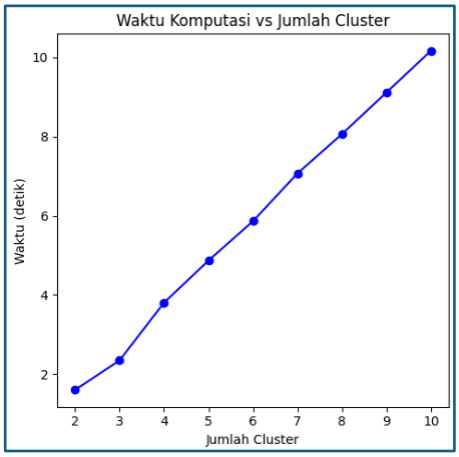
Jumlah Cluster	Silhouette Score	Davies-Bouldin Index	Waktu Komputasi (detik)
2	0,9258	0,9698	1,6022
3	0,9277	0,7561	2,347
4	0,8829	1,058	3,8058
5	0,8850	0,9578	4,8693
6	0,8744	0,7503	5,8644
7	0,7633	0,8613	7,0711
8	0,7662	0,781	8,0643
9	0,7661	0,6368	9,1145
10	0,7639	0,6165	10,169

Hasil klasterisasi menggunakan *Bisecting K-Means* pada tabel 1, menunjukkan pada jumlah klaster ke-2 diperoleh nilai *Silhouette Scor*e tertinggi sebesar 0,9258 dan *Davies-Bouldin Index* sebesar 0,9698, yang membuktikan bahwa pemisahan antar klaster cukup baik tetapi masih kurang optimal. Namun, pada jumlah klaster ke-3 nilai Davies-Bouldin Index menurun cukup drastis menjadi 0,7561, yang menandakan peningkatan kualitas klaster, meskipun waktu komputasinya meningkat menjadi 2,347 detik.

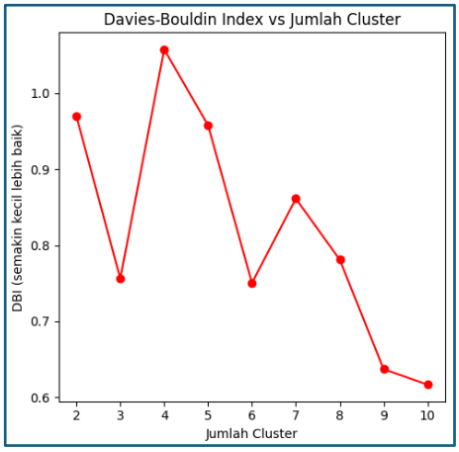
Dengan bertambahnya jumlah klaster, terlihat dari waktu komputasi yang meningkat secara signifikan dan terus-menerus, dari 1,6022 detik pada jumlah 2 klaster menjadi 10,169 detik pada jumlah 10 klaster. Penurunan nilai Davies-Bouldin Index hingga 0,6165 pada klaster ke-10 menunjukkan peningkatan dalam struktur klaster yang semakin kompak dan terpisah. Namun sebaliknya, nilai Silhouette Score mengalami penurunan secara bertahap setelah klaster ke-5, dari 0,8850 menjadi 0,7639, yang mengindikasikan bahwa kualitas pemisahan antar klaster mulai berkurang. Berikut ini, terdapat tiga grafik visualisasi dari hasil evaluasi matrik *Silhouette Score*, *Davies-Bouldin Index*, dan waktu komputasi.



Gambar 5. Grafik Silhouette Score Bisecting K-Means



Gambar 6. Grafik Davies-Bouldin Index Bisecting K-Means

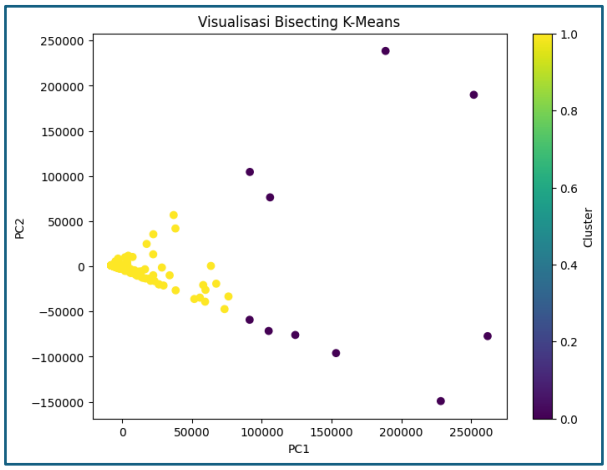


Gambar 7. Grafik Waktu Komputasi Bisecting K-Means

Pada gambar 5, terlihat *Silhouette Score* mencapai titik puncaknya pada jumlah klaster 2 dan 3, dengan skor sekitar 0,925. Ini menunjukkan bahwa pada jumlah klaster tersebut, pemisahan antar klaster yang sangat baik dan koherensi dalam klaster juga tinggi. Namun, setelah jumlah klaster 5, *Silhouette Score* menurun drastis mencapai sekitar 0,76 pada klaster ke-7 hingga 10, menandakan penurunan kualitas pemisahan klaster akibat kelebihan segmentasi data.

Gambar 6 memperlihatkan nilai *Davies-Bouldin Index* yang cenderung menurun seiring bertambahnya jumlah klaster, dengan nilai terendah terletak pada klaster ke-10 sebesar 0,6165. Hal ini menunjukkan bahwa semakin bertambah jumlah klasternya, jarak antar klaster menjadi lebih jauh dan kompak secara internal. Terakhir, gambar 7 menunjukkan waktu komputasi yang secara konsisten meningkat seiring bertambahnya jumlah klaster, mulai dari sekitar 1,6 detik pada 2 klaster hingga lebih dari 10 detik pada 10 klaster. Kenaikan ini diakibatkan karena proses pemisahan dan evaluasi klaster menjadi lebih kompleks saat jumlah klaster bertambah.

Secara keseluruhan, visualisasi dari ketiga grafik ini mengindikasikan bahwa jumlah klaster yang paling optimal pada model *Bisecting K-Means* berada di sekitar jumlah klaster 3 hingga 5, karena merupakan titik keseimbangan antara kualitas klaster dan efisiensi waktu komputasi. Selain itu, terdapat grafik *scatterplot* yang memperlihatkan distribusi data menggunakan PCA dibawah ini.



Gambar 8. Grafik Scatter plot Bisecting K-Means (2 klaster)

Gambar 8 merupakan hasil klasterisasi *Bisecting K-Means* yang memperlihatkan distribusi data pada dua klaster menggunakan proyeksi dua dimensi hasil reduksi *Principal Component Analysis* (PCA). Warna pada titik-titik data menunjukkan keanggotaan klaster, yaitu warna kuning untuk klaster 1 dan warna ungu untuk klaster 0.

Titik-titik berwarna kuning tampak terkumpul di satu area dengan distribusi yang lebih padat dan mendekati titik pusat, menunjukkan bahwa klaster ini mencerminkan kawasan dengan sifat pertanian yang lebih konsisten, kemungkinan merupakan daerah dengan tingkat hasil dan produktivitas yang cukup stabil. Di sisi lain, titik-titik berwarna ungu menyebar lebih luas dengan jarak yang lebih jauh dari pusat distribusi, menandakan adanya *outlier* atau wilayah dengan karakteristik produksi yang ekstrem, baik yang sangat tinggi maupun yang sangat rendah.

Visualisasi ini memperkuat hasil evaluasi yang terdapat pada tabel sebelumnya, di mana jumlah klaster 2 menunjukkan *Silhouette Score* yang tinggi, mengindikasikan bahwa data di setiap klaster cukup terkoherensi di dalamnya dan terpisah dengan baik dari klaster lain. Selain itu, distribusi titik-titik yang tidak tumpang tindih dengan signifikan membuktikan bahwa struktur klaster dari *Bisecting K-Means* berhasil memisahkan data berdasarkan karakteristiknya dengan efisien, terutama untuk dua kelompok besar yang terbentuk.

3.3 Hasil Evaluasi Hierarchical Clustering

Algoritma Hierarchical Clustering menggunakan metode *Ward linkage* untuk mengelompokkan data secara bertingkat berdasarkan minimasi varians antar klaster. Evaluasi dilakukan terhadap hasil klasterisasi dari jumlah

klaster 2 hingga 10 dengan tiga metrik utama: Silhouette Score, Davies-Bouldin Index, dan waktu komputasi. Hasil evaluasi tersebut ditampilkan pada Tabel 2 berikut.

Tabel 2 Hasil Evaluasi Algoritma Hierarchical Clustering

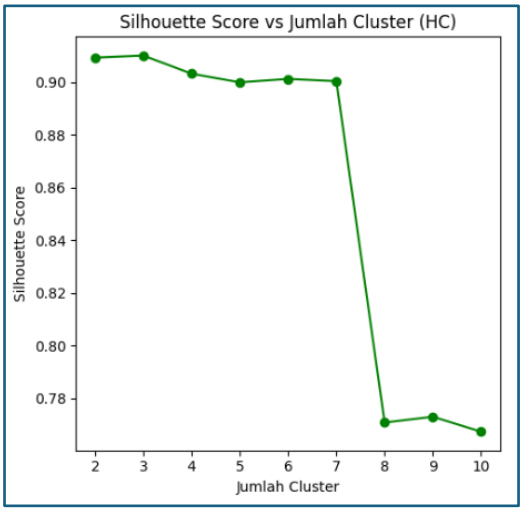
Jumlah Cluster	Silhouette Score	Davies-Bouldin Index	Waktu Komputasi (detik)
2	0,9093	1,0845	0,0321
3	0,9100	0,8714	0,0265
4	0,9032	0,7718	0,0287
5	0,8999	0,6028	0,03
6	0,9012	0,5451	0,0334
7	0,9003	0,3885	0,032
8	0,7709	0,5658	0,0257
9	0,7731	0,6382	0,0278
10	0,7674	0,6653	0,0309

Hasil *clustering* dengan algoritma *Hierarchical Clustering* pada Tabel 2 menunjukkan kualitas klaster yang cukup konsisten dari jumlah klaster 2 hingga 6, di mana *Silhouette Score* berada di atas 0,89. Nilai tertinggi dari matrik *Silhouette Score* terletak pada jumlah klaster 2 sebesar 0,9093, dan tetap stabil hingga jumlah klaster 6 dengan nilai 0,9012, menandakan bahwa data pada klaster-klaster tersebut memiliki kepadatan dan pemisahan yang baik.

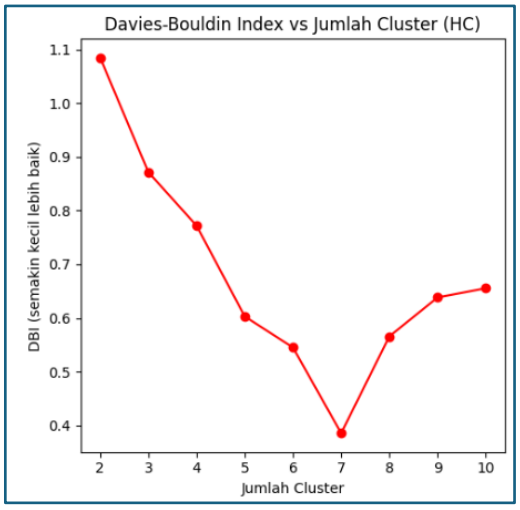
Sementara itu, nilai *Davies-Bouldin Index* (DBI) menunjukkan penurunan hingga mencapai titik terendah pada jumlah klaster 7 (0,3885), yang menandakan struktur klaster yang semakin baik karena nilai DBI yang lebih rendah membuktikan klaster yang lebih kompak dan terpisah secara optimal.

Waktu komputasi pada *Hierarchical Clustering* secara keseluruhan sangat rendah, yang berkisar antara 0,0257 hingga 0,0334 detik. Hal ini menunjukkan salah satu keunggulan dari algoritma ini dalam efisiensi waktu dibandingkan *Bisecting K-Means*, terutama ketika jumlah klaster meningkat.

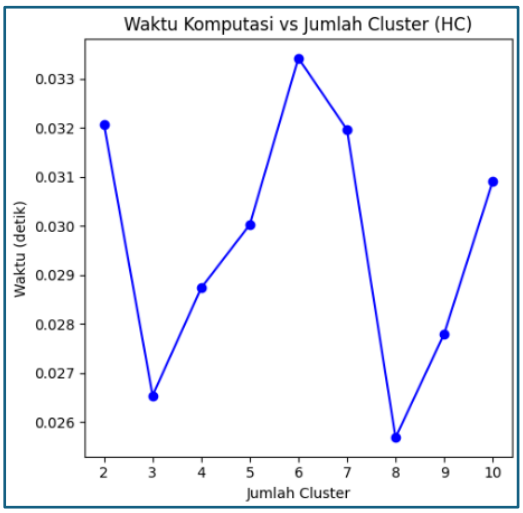
Hasil evaluasi dari ketiga metrik, jumlah klaster 5 hingga 7 memberikan evaluasi terbaik berupa keseimbangan antara kualitas pemisahan klaster yang terbukti dari *Silhouette Score* yang tetap tinggi dan nilai DBI yang sangat rendah, menjadikan interval ini sebagai kandidat optimal untuk jumlah klaster yang digunakan dalam *Hierarchical Clustering*. Sama seperti *Bisecting K-Means*, terdapat tiga grafik yang menampilkan hasil dari *Silhouette Score*, *Davies-Bouldin Index*, dan waktu komputasi dari jumlah klaster 2 hingga 10 pada algoritma *Hierarchical Clustering*.



Gambar 9. Grafik Silhouette Score Hierarchical Clustering



Gambar 10. Grafik Davies-Bouldin Index Hierarchical Clustering

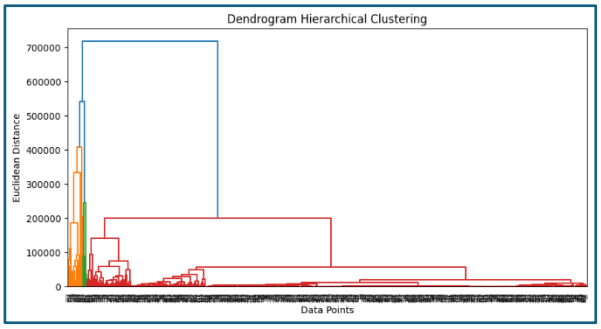


Gambar 11. Grafik Waktu Komputasi Hierarchical Clustering

Pada gambar 9, terlihat *Silhouette Score* tertinggi terdapat pada jumlah klaster 3, yaitu sebesar 0,9012, sedikit lebih tinggi dari jumlah klaster 2 dan 4 walaupun tidak beda jauh. Hal ini menunjukkan bahwa pada jumlah klaster 6, struktur klaster memiliki kekompakan internal dan pemisahan antar klaster yang paling optimal. Setelah jumlah klaster melebihi 7, terjadi penurunan drastis pada *Silhouette Score*, yang menandakan kualitas pemisahan antar klaster mulai menurun secara signifikan.

Gambar 10, menunjukkan nilai *Davies-Bouldin Index* (DBI) yang semakin kecil mulai dari jumlah klaster 2 hingga mencapai titik terendah pada jumlah klaster 7 (0,3885). Artinya, struktur klaster yang paling terpisah dan padat berada pada titik tersebut. Namun, bila digabungkan dengan penurunan *Silhouette Score* pada jumlah klaster 7, maka kualitas pemisahan klaster tidak bisa dinilai hanya dari DBI saja. Selanjutnya, gambar 11 menunjukkan waktu komputasi *Hierarchical Clustering* relatif konstan dan rendah pada semua jumlah klaster, yang berada di kisaran antara 0,0257 hingga 0,0334 detik. Ini membuktikan bahwa model *Hierarchical Clustering* memiliki efisiensi waktu yang tinggi, meskipun kompleksitas data meningkat.

Hasil secara keseluruhan dari tiga grafik yang telah dibahas, menunjukkan bahwa jumlah klaster optimal untuk Hierarchical Clustering adalah 7, karena pada titik ini terjadi keseimbangan antara nilai *Silhouette Score* yang tinggi dan nilai DBI yang rendah. Selain itu, kebutuhan waktu komputasi yang sangat singkat menjadikan model ini efisien untuk digunakan dalam pengolahan data klasterisasi skala kecil hingga menengah. Tiga grafik tersebut merupakan hasil dari evaluasi matrik dari setiap jumlah klaster, terdapat pula grafik dendrogram yang menampilkan hubungan hirarkis antara data poin berdasarkan jarak *Euclidean* dan dengan metode *Ward* yang dapat dilihat pada gambar berikut ini.



Gambar 12. Grafik Dendrogram dari Hierarchical Clustering

Gambar 12 menampilkan dendrogram hasil klasterisasi menggunakan model *Hierarchical Clustering*. Tinggi cabang pada grafik dendrogram menunjukkan jarak antar klaster saat penggabungan terjadi, di mana jarak yang lebih besar menandakan dua klaster yang digabung memiliki perbedaan yang lebih signifikan. Dari dendrogram ini, terlihat bahwa terdapat beberapa klaster utama yang terbentuk sebelum penggabungan besar terjadi pada jarak vertikal yang tinggi.

Namun, karena jumlah data sangat besar yang mewakili seluruh kota di Indonesia, visualisasi pada sumbu X tampak penuh atau terlihat hanya berwarna hitam dan saling tumpang tindih, sehingga identifikasi visual terhadap label masing-masing kota menjadi sulit. Meski demikian, dendrogram ini tetap memberikan informasi penting mengenai titik-titik pemotongan (*cutting point*) yang optimal dalam membagi data ke dalam sejumlah klaster, misalnya 2 atau 3 klaster utama.

3.4 Perbandingan Hasil Evaluasi

Berdasarkan hasil evaluasi, model *Bisecting K-Means* mencatat nilai *Silhouette* tertinggi terletak pada jumlah klaster 3 sebesar 0,9277, sedikit lebih unggul dari model *Hierarchical Clustering* yang mencatat nilai 0,91 pada jumlah klaster yang sama. Namun, nilai *Davies-Bouldin Index* (DBI) terbaik terdapat pada model *Hierarchical Clustering* pada jumlah klaster 7 dengan nilai 0,3885, jauh lebih rendah dibandingkan nilai terbaik yang diperoleh *Bisecting K-Means* pada klaster ke-10, yaitu 0,6165. Dari sisi efisiensi waktu komputasi, *Hierarchical Clustering* secara konsisten memiliki waktu proses yang jauh lebih cepat, berkisar antara 0,0257 hingga 0,0334 detik, sedangkan *Bisecting K-Means* yang memerlukan waktu lebih dari 1 detik hingga di atas 10 detik untuk jumlah klaster 10. Hal ini menunjukkan keunggulan signifikan dari Hierarchical Clustering dalam efisiensi waktu. Dengan demikian, penjelasan sederhananya:

1. *Bisecting K-Means* unggul pada Silhouette Score untuk klaster rendah (k=2 atau 3).
2. *Hierarchical Clustering* unggul pada nilai DBI dan efisiensi waktu komputasi, serta lebih konsisten menghasilkan klaster yang baik pada jumlah klaster 4 hingga 7.

4. Kesimpulan

Hasil secara keseluruhan yang didapatkan pada penelitian ini, dapat dilihat dalam bentuk ringkasan pada poin-poin berikut ini.

1. Perbandingan dari kedua model, *Bisecting K-Means* unggul dalam kualitas pemisahan klaster pada jumlah klaster rendah, tetapi memiliki kekurangan dari segi waktu komputasi yang lebih tinggi. Sementara itu, *Hierarchical Clustering* lebih unggul dalam efisiensi waktu komputasi dan konsistensi pada struktur klaster, terutama untuk jumlah klaster 4 hingga 7.
2. Kelebihan dari model *Bisecting K-Means*:
   - Memiliki *Silhouette Score* yang tinggi untuk klaster kecil (jumlah klaster 2-3).
   - Cocok digunakan untuk mendeteksi struktur klaster yang kompleks.
3. Kekurangan dari model *Bisecting K-Means*:
   - Waktu komputasi yang semakin meningkat seiring bertambahnya jumlah klaster.

- Nilai DBI yang tinggi pada jumlah klaster 2 dan rendah ketika sudah mencapai jumlah klaster 10.

4. Kelebihan dari model *Hierarchical Clustering*:
   - Nilai DBI sangat rendah pada klaster menengah.
   - Waktu komputasi jauh lebih cepat dan stabil.
5. Kekurangan dari model *Hierarchical Clustering*:
   - Visualisasi dendrogram yang kurang informatif ketika menggunakan dataset besar.
6. Pengembangan selanjutnya: Membandingkan metode dengan algoritma klasterisasi lain seperti DBSCAN, *Gaussian Mixture*, atau metode *deep clustering*.

REFERENSI

[1] E. Elwi Jeksen dan D. Sari, "ANALISIS PROSPEK PENINGKATAN PRODUKSI CABAI RAWIT (Capsicum frutescens L.) DI INDONESIA," *PAPER SSRN*, 2022, [Daring]. Tersedia pada: https://ssrn.com/abstract=4285742

[2] H. Xu *dkk.*, "Identification of the pollution sources and hidden clustering patterns for potentially toxic elements in typical peri-urban agricultural soils in southern China," *Environmental Pollution*, vol. 370, hlm. 125904, Apr 2025, doi: 10.1016/j.envpol.2025.125904.

[3] Udhaya Priya J dan Dr. K. Nirmala, "An Ensemble Based Clustering and Classification Framework for Prediction of Agricultural Crop Yield," *Nanotechnol Percept*, hlm. 397–413, Nov 2024, doi: 10.62441/nano-ntp.vi.2787.

[4] S. H. Javadi, A. Guerrero, dan A. M. Mouazen, "Clustering and Smoothing Pipeline for Management Zone Delineation Using Proximal and Remote Sensing," *Sensors*, vol. 22, no. 2, hlm. 645, Jan 2022, doi: 10.3390/s22020645.

[5] D. Ulayya Tsabitah, Y. Angraini, dan I. M. Sumertajaya, "Implementation of Time Series Clustering with DTW to Clustering and Forecasting Rice Prices Each Provinces in Indonesia," *INFERENSI*, vol. 8, no. 1, hlm. 2721–3862, 2025, doi: 10.12962//j27213862.v8i1.21952.

[6] Y. Asriningtias dan J. Aryanto, "K-Means Algorithm with Davies Bouldin Criteria for Clustering Provinces in Indonesia Based on Number of Events and Impacts of Natural Disasters," *International Journal of Engineering Technology and Natural Sciences*, vol. 4, no. 1, hlm. 75–80, Jul 2022, doi: 10.46923/ijets.v4i1.147.

[7] Y. Chen, W. Liu, H. Zhao, S. Cao, S. Fu, dan D. Jiang, "Bisecting K-Means Based Fingerprint Indoor Localization," dalam *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, Springer, 2019, hlm. 1–12. doi: 10.1007/978-3-030-32216-8_1.

[8] S. E. H. Pang, J. W. F. Slik, D. Zurell, dan E. L. Webb, "The clustering of spatially associated species unravels patterns in tropical tree species distributions," *Ecosphere*, vol. 14, no. 6, Jun 2023, doi: 10.1002/ecs2.4589.

[9] L. E. Ekemeyong Awong dan T. Zielinska, "Comparative Analysis of the Clustering Quality in Self-Organizing Maps for Human Posture Classification," *Sensors*, vol. 23, no. 18, hlm. 7925, Sep 2023, doi: 10.3390/s23187925.

[10] D. Valkenborg, A.-J. Rousseau, M. Geubbelmans, dan T. Burzykowski, "Unsupervised learning," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 163, no. 6, hlm. 877–882, Jun 2023, doi: 10.1016/j.ajodo.2023.04.001.

[11] M. Usama *dkk.*, "Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges," *IEEE Access*, vol. 7, hlm. 65579–65615, 2019, doi: 10.1109/ACCESS.2019.2916648.

[12] T. Tendean, W. Purba, dan M. Kom, "Analisis Cluster Provinsi Indonesia Berdasarkan Produksi Bahan Pangan Menggunakan Algoritma K-Means," *Jurnal Sains dan Teknologi)*, vol. 1, no. 2, hlm. 5–11, 2020.

[13] B. Chong, "K-means clustering algorithm: a brief review," *Academic Journal of Computing & Information Science*, vol. 4, no. 5, 2021, doi: 10.25236/AJCIS.2021.040506.

[14] B. Moseley dan Y. Wang, "An Objective for Hierarchical Clustering in Euclidean Space and its Connection to Bisecting K-means," Agu 2020.

[15] B. Nurina Sari dan M. Yamin Nurzaman, "Implementasi K-Means Clustering Dalam," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 10, no. 3, 2023, [Daring]. Tersedia pada: http://jurnal.mdp.ac.id

[16] C. Schröer, F. Kruse, dan J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput Sci*, vol. 181, hlm. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[17] A. Fauzi, D. M. Furqon, R. A. Maulana, N. D. Cahya, dan M. N. Sidiq, "Analisis Pendapatan dan Pengeluaran Film menggunakan Algoritma Bisecting K-Means (Analysis of Film Budget and Profit using the Bisecting K-Means Algorithm)," *Gunung Djati Conference Series*, vol. 3, 2021.

[18] N. Puspitasari, J. A. Widians, dan N. B. Setiawan, "Customer segmentation using bisecting k-means algorithm based on recency, frequency, and monetary (RFM) model," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, hlm. 78–83, Apr 2020, doi: 10.14710/jtsiskom.8.2.2020.78-83.

[19] S. Dwididanti dan D. A. Anggoro, "Analisis Perbandingan Algoritma Bisecting K-Means dan Fuzzy C-Means pada Data Pengguna Kartu Kredit," *Emitor: Jurnal Teknik Elektro*, vol. 22, no. 2, hlm. 110–117, Agu 2022, doi: 10.23917/emitor.v22i2.15677.

[20] N. K. Zuhal, "Study Comparison K-Means Clustering dengan Algoritma Hierarchical Clustering," 2022.

[21] E. Widodo, P. Ermayani, L. N. Laila, dan A. T. Madani, "Pengelompokkan Provinsi di Indonesia Berdasarkan Tingkat Kemiskinan Menggunakan Analisis Hierarchical Agglomerative Clustering," *Seminar Nasional Official Statistics*, vol. 2021, no. 1, hlm. 557–566, Nov 2021, doi: 10.34123/semnasoffstat.v2021i1.971.

[22] I. Rahma, P. P. Arhandi, dan A. T. Firdausi, "PENERAPA METODE HIERARCHICAL CLUSTERING DAN K-MEANS CLUSTERING UNTUK MENGELOMPOKKAN POTENSI LOKASI

PENJUALAN LINKAJA," *Jurnal Informatika Polinema*, vol. 6, no. 1, hlm. 15–22, Jan 2020, doi: 10.33795/jip.v6i1.287.

[23] G. Zahra Nur Fadhilah *dkk.*, "Klasterisasi Pola Gejala Depresi Menggunakan Agglomerative Hierarchical Cluster Analysis Berdasarkan Skor Depresi PHQ-9," *Jurnal Ilmiah Sistem Informasi dan Ilmu Komputer*, vol. 5, no. 2, hlm. 1–16, 2025, doi: 10.55606/juisik.v5i2.993.

[24] Y. Asyfani *dkk.*, "Pengelompokan Kabupaten/Kota di Jawa Tengah Berdasarkan Kepadatan Penduduk Menggunakan Metode Hierarchical Clustering," *Journal Of Data Insights*, vol. 2, no. 1, hlm. 1–8, Jun 2024, doi: 10.26714/jodi.v2i1.158.

[25] Z. Alamtaha, I. Djakaria, N. I. Yahya, J. Matematika, dan F. Mipa, "Implementasi Algoritma Hierarchical Clustering dan Non-Hierarchical Clustering untuk Pengelompokkan Pengguna Media Sosial," *Estimasi: Journal of Statistics and Its Application*, vol. 4, no. 1, hlm. 2721–379, 2023, doi: 10.20956/ejsa.vi.24830.

Julius Juan, Saat ini sebagai Mahasiswa studi Teknik Informatika Universitas Tarumanagara.