PERBANDINGAN K-MEANS DAN HIERARCHICAL CLUSTERING PADA DATA TEBU DI INDONESIA

Darren Amadeus Kurniawan [1] Teny Handhayani [2]

[1] Teknik Informatika, FTI Universitas Tarumanagara
Jl. Letjen S. Parman St No.1, RT.6/RW.16, Tomang, Grogol petamburan, Kota Jakarta Barat, Jakarta 11440 Indonesia
email : darren.535230165@stu.untar.ac.id
[2] Teknik Informatika, FTI Universitas Tarumanagara
Jl. Letjen S. Parman St No.1, RT.6/RW.16, Tomang, Grogol petamburan, Kota Jakarta Barat, Jakarta 11440 Indonesia
email : tenyh@fti.untar.ac.id

ABSTRAK

Penelitian ini bertujuan untuk membandingkan dua algoritma pembelajaran mesin tanpa pengawasan, yaitu K-Means dan Hierarchical Clustering dalam mengelompokkan data tebu dari seluruh Indonesia selama 2010 sampai 2024. Dataset yang digunakan dalam penelitian ini berisi tiga fitur: luas panen, volume produksi, dan produktivitas. Dataset yang digunakan dalam penelitian ini, telah diterapkan melalui langkah-langkah praproses seperti normalisasi dan penanganan nilai yang hilang untuk memastikan konsistensi model algoritma. Perbandingan itu sendiri berfokus pada kualitas dan interpretabilitas klaster yang terbentuk. Kedua algoritma diuji pada dataset yang sama untuk memastikan kualitas dan interpretabilitas. Perbandingan dilakukan dengan menggunakan metrik silhouette score dan davies-bouldin indeks yang berfungsi mengukur kualitas dan jarak antar klaster. Hasil yang didapatkan dari eksperimen yang dilakukan adalah algoritma model K-Means mendapatkan hasil terbaik, terutama pada jumlah K = 2 menggunakan silhouette score dengan nilai rata-rata di atas 0.8 untuk seluruh fitur, pada davies-bouldin indeks nilai terendah didapatkan pada jumlah klaster yang berbeda tergantung pada fitur yang dipilih. Pada algoritma model hierarchical clustering, algoritma berjalan dengan cukup baik menggunakan silhouette score dan davies-bouldin indeks, tetapi nilai yang didapatkan cukup tidak stabil terutama pada fitur produktivitas karena cukup banyaknya nilai 0 yang ada pada kolom tahun 2024. Berdasarkan penjabaran yang telah dilakukan, algoritma model K-Means dapat dikatakan lebih unggul dan stabil untuk hal klasterisasi dataset yang digunakan dibandingkan dengan model algoritma hierarchical clustering yang dapat dikatakan kurang optimal dan stabil dalam eksperimen yang dilakukan dengan dataset ini. Penelitian ini memberikan informasi yang memungkinkan berkontribusi sebagai acuan dalam pengambilan keputusan ke depannya tentang bagaimana nasib pengembangan dan perencanaan produksi tebu di Indonesia.

Kata Kunci

Algoritma, Dataset, Davies-bouldin Index, K-Means, Klaster, Silhouette,

ABSTRACT

This study aims to compares of two unsupervised algorithms of machine learning, those algorithms are K-Means and Hierarchical Clustering in grouping sugarcane data from all across Indonesia over multiple years. The dataset that used in this study contains three features: harvested area, production volume, and productivity. Dataset that used in this study, has been apllied by preprocessing steps such as normalization and missing values handling to ensure algorithms model consistency. The comparison itself focuses on the quality and interpretability of clusters formed. Both algorithms are tested on the sama dataset to ensure the quality and interpretability. Comparisons were made using the silhouette score and Davies-Bouldin index metrics which function to measure the quality and distance between clusters. The results obtained from the experiments carried out were that the K-Means model algorithm got the best results, especially at the number of K = 2 using the silhouette score with an average value above 0.8 for all features, in the Davies-Bouldin index the lowest value was obtained at a different number of clusters depending on the selected feature. In the hierarchical clustering model algorithm, the algorithm runs quite well using the silhouette score and Davies-Bouldin index, but the values obtained are quite unstable, especially in the productivity feature because there are quite a lot of 0 values in the 2024 column. Based on the explanation that has been done, the K-Means model algorithm can be said to be superior and more stable for the clustering of the dataset used compared to the hierarchical clustering algorithm model which can be said to be less than optimal and stable in the experiments carried out with this dataset. This study provides information that can contribute as a reference in

future decision making about the fate of sugarcane production development and planning in Indonesia.

Key words

Algorithms, Clusterin, Dataset, Davies-bouldin Index, K-Means, Silhouette,

1. Pendahuluan

Manusia membutuhkan rasa manis untuk sebagian besar masakan dan makanan, rasa manis yang didapatkan secara alami dan cukup mudah ditemukan adalah tebu. Tebu adalah salah satu tumbuhan yang merupakan bahan baku pembuatan gula[1]. Tebu juga menjadi salah satu tumbuhan yang cukup penting dalam pertumbuhan perekonomian Indonesia dalam bidang perkebunan[2]. Pentingnya tebu dalam perekonomian di Indonesia dapat dilihat selama lima tahun terakhir yang dimana tebu mencatat pertumbuhan rata-rata 3,5% per tahun yang mencapai 2,27 ton pada tahun 2023[3]. Meningkatnya pertumbuhan tebu di Indonesia ini dapat kita analisis berdasarkan tiga indikator utama yaitu luas area, produksi, dan produktivitas tebu di dalam masing-masing provinsi dan kabupaten yang ada di Indonesia. Tiga indikator utama yang telah disebutkan sebelumnya bisa di dapatkan dari website Badan Data dan Sistem Informasi Pertanian (BDSP). Badan Data dan Sistem Informasi Pertanian (BDSP) sendiri adalah lembaga pemerintah yang yang mengelola data pertanian, perkebunan, dan holtikultura yang ada di seluruh daerah di Indonesia[4].

Dataset yang digunakan juga diambil dari website Badan Data dan Sistem Informasi Pertanian (BDSP) yang telah disebutkan sebelumnya. Di dalam website Badan Data dan Sistem Informasi Pertanian (BDSP) disediakan data-data yang dapat dicari sesuai rentang waktu per tahun, luas area panen, total produksi, dan produktivitas suatu tumbuhan pertanian atau perkebunan, dimana di penelitian ini data yang diambil adalah tebu dalam rentang waktu dari tahun 2010 sampai 2024 yang tersebar di seluruh provinsi di Indonesia.

Clustering merupakan alat yang berguna dalam ilmu *data science*. *Clustering* merupakan metode untuk menemukan struktur klaster dalam kumpulan data yang dicirikan oleh kesamaan terbesar dalam klaster yang sama dan perbedaan terbesar antara klaster yang berbeda [5]. Metode *clustering* biasanya memerlukan fitur yang dibuat secara manual atau dipelajari untuk setiap dataset dan tugas yang ingin dibuat [6]. Salah satu metode clustering yang nanti akan digunakan dalam penelitian ini adalah K-Means dan *hierachical clustering*. Teknik *clustering* adalah teknik yang digunakan secara luas dan cukup bertambah popular diikuti dengan perkembangan zaman yang dituntut untuk mengolah data yang bertambah besar[7]. Dengan penjelasan teknik *clustering* yang sudah dijelaskan, diharapkan *clustering* dapat dengan mudah dan cocok untuk mengolah dataset yang dipakai ini.

K-Means merupakan salah satu model algoritma *unsupervised learning* yang cukup sederhana yang dapat mengatasi *clustering* yang cukup terkenal. Algoritma ini mengikuti cara sederhana dan mudah untuk mengklasifikasikan dataset melalui sejumlah klister (diasumsikan k = klaster) yang ditetapkan sebelumnya atau singkatnya K-means adalah prosedur iterative yang membagi N objek menjadi K klaster yang terpisah[8]. Model algoritma lain yang digunakan untuk mengolah dan menganalisis dataset tebu ini adalah *Hierarchical Clustering*. *Hierarchical Clustering* sendiri adalah metode analisis yang bertujuan untuk membangun hierarki klaster di mana data atau kumpulan data yang dikelompokkan [9]. Dalam *Hierarchical Clustering* juga terdapat 2 metode yaitu *Agglomerative (bottom-up)* dan *Devisive (top-down)* [10].

Dalam penelitian ini yang menggunakan model algoritma *clustering* K-Means dan Hierarchical Clustering, hasil evaluasi akan menggunakan metrik silhouette score dan Davies-Bouldin Indeks (DBI). Digunakan silhouette score untuk metrik evaluasi yang pertama untuk menunjukkan kualitas pengelompokan, dengan skor yang lebih tinggi mencerminkan pengelompokan yang lebih optimal, semakin mendekati nilai 1, maka semakin ter klasterisasi [11]. Sementara Davies-Bouldin Indeks (DBI) adalah rasio antara jarak-jarak dalam klaster dan antar klaster untuk setiap klaster terhadap klaster yang terdekat. Davies Bouldin Index (DBI) juga merupakan ukuran untuk mengevaluasi kinerja *clustering*. Davies Bouldin Index (DBI) memiliki korelasi yang positif terhadap kasus *within-class* dan korelasi negatif terhadap kasus *between-class*[12].

Tujuan dari penelitian ini sendiri adalah untuk membandingkan kedua model algoritma yaitu K-Means dan *Hierarchical Clustering* dalam mengelompokkan dataset luas area, volume produksi, dan produktivitas tebu di seluruh provinsi yang ada di Indonesia. Tujuan lain dari penelitian ini adalah untuk mengindetifikasikan pola klasifikasi sehingga mendapat gambaran tentang luas area, volume produksi, dan produktivitas tebu yang ada di provinsi-provinsi di Indonesia. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam perencanaan dan pengambilan keputusan dalam peningkatan produksi dan hasil panen tebu di Indonesia.

2. Metode Penelitian

Metode penelitian yang digunakan dalam penelitian ini dilakukan secara eksperimen masing-masing model algoritma dan studi literatur. Penelitian ini juga dilakukan dengan cara kedua model algoritma dijalankan selama 9 kali dengan klaster yang berbeda. 9 klaster yang dimaksud berbeda adalah karena setiap menjalankan kode, klaster (K) akan berubah dari K = 2 sampai K = 10.

Pengolahan dan penelitian data digunakan dengan bahasa pemrograman python.merupakan Bahasa pemrograman tingkat tinggi, lintas platform, dan terinterpretasi yang berfokus pada keterbacaan kode. Tersedia sejumlah besar pustaka berkualitas tinggi dan dukungan untuk segala jenis komputasi ilmiah terjamin. Karakteristik ini menjadikan bahasa pemrograman menjadikan bahasa pemrograman Python alat yang tepat untuk banyak proyek penelitian dan industri yang

penyelidikannya bisa agak rumit[13] Dalam proses koding menggunakan bahasa python, akan digunakan beberapa library untuk memudahkan jalannya penelitian. Beberapa library yang dimaksud adalah sklearn untuk meng*import* KMeans, AgglomerativeClustering, silhouette_score, dan davies_bouldin_score, matplotlib untuk membuat plot, numpy, serta pandas. Numpy dan pandas merupakan dua *library* utama dan yang paling sering digunakan. Numpy merupakan *library* pemrograman array utama untuk bahasa Python[14]. Pandas merupakan library python untuk membaca dan memproses tabel [15]. Sklearn sendiri adalah library Python yang dikembangkan untuk memberikan kemudahan dalam pengkodean machine learning dalam bahasa pemrograman Python library ini dirancang di atas modul NumPy (Numerical Python) dan SciPy (Sientific Python), sehingga perhitungan di dalamnya menjadi lebih efisien[16].

2.1 K-Means

K-means merupakan sebuah algoritma *unsupervised learning* yang digunakan dalam pengelompokan data dalam dataset yang tidak memiliki label kedalam sebuah klaster-klaster yang berbeda[17]. Alasan utama mengapa algoritma model K-Means cukup populer adalah karena algoritma model K-Means ini memiliki kecermatan yang cukup baik karena lebih terukur dan efisien untuk pengolahan data yang memiliki ukuran cukup besar, alasan lainnya adalah algoritma model K-Means ini tidak terpengaruh urutan data yang ada dalam dataset[18]. Penjelasan perhitungan dasar algoritma model K-Means adalah sebagai berikut

$$f(x) = \sum_{n=1}^{k} \sum_{n=1}^{n} |X_i - C_j|^2$$

k adalah jumlah klaster, n adalah jumlah kasus dan C adalah jumlah centroid dan X adalah titik data yang jarak Euclidean dari centroid dihitung. K berarti algoritma memiliki fase inisialisasi dan iterasi. Pada fase pertama, titik data ditetapkan secara acak ke dalam k klaster, kemudian pada fase iterasi algoritma menghitung jarak antara setiap titik data ke setiap pusat klaster[19].

Untuk memperharui centroid selama proses iterasi, dapat menggunakan rumus di bawah ini.

$$V_{ij} = \frac{1}{Ni} \sum_{k=0}^{Ni} Xkj$$

Keterangan dari rumus di atas, adalah sebagai berikut:

- Vij = Centroidrata-rata Clusterke-I variabel ke-j
- Ni = Jumlah anggota Clusterke-i
- i, k = Indeks Cluster
- j = Indeks variable
- Xkj = nilai ke-k 3 variable ke-j untuk Cluster

Pada Metode yang digunakan dalam algoritma model K-Means, data akan di kelompokkan menjadi beberapa bagian kelompok dan setiap bagian kelompok memiliki ciri-ciri yang mirip satu sama lain, tapi berbeda dengan kelompok lainnya. Hal tersebut mempunyai tujuan untuk memperkecil kesempatan perbedaan data dengan klister lain. Berikut beberapa istilah dalam algoritma model K-Means.

- *Cluster*, suatu kelompok atau grup
- Centroid, titik pusat untuk menentukan *Euclidean distance*
- Iterasi, pengulangan suatu proses yang akan berhenti ketika hasil iterasi telah konvergen [20]

Tujuan pengelompokan adalah untuk meminimalkan fungsi objektif yang ditetapkan dalam proses pengelompokan, secara umum pengelompokan dimaksudkan untuk meminimalkan variasi dalam suatu kelompok dan memaksimalkan variasi antar kelompok[21].

2.2 Hierarchical Clustering

Hierarchical clustering merupakan merupakan pilihan model algoritma alami untuk tujuan pengelompokan, di mana jumlah klaster tidak diketahui dan di mana semua contoh ditetapkan ke klaster yang paling relevan[22]. Metode *hierarchical clustering* dapat dibagi lagi menjadi cabang-cabang aglomeratif dan cabang-cabang divisif. hierarchical clustering aglomeratif,adalah yang menggunakan strategi *bottom-up*, pertama-tama mengklasifikasikan setiap objek sebagai suatu kelompok dan kemudian menggabungkan kelompok-kelompok ini menjadi kelompok-kelompok yang lebih besar. Sebaliknya, *hierarchical clustering* hierarkis divisif, yang menggunakan strategi *top-down*, pertama-tama mengklasifikasikan semua objek dalam suatu kelompok dan kemudian secara bertahap membaginya menjadi kelompok-kelompok yang lebih kecil. Pengelompokan divisif umumnya lebih kompleks daripada pengelompokan aglomeratif karena pengelompokan ini membagi data hingga setiap kelompok berisi satu item data.[23].

Model algoritma hierarchical clustering yang dipakai dalam penelitian ini adalah *hierarchical clustering* aglomeratif. Digunakannya model hierarchical clustering aglomeratif bukannya tanpa alasan, hierarchical clustering aglomeratif adalah model yang paling umum digunakan karena kemudahan implementasi dan kemampuan untuk menangkap pola data yang terjadi secara alami [24]. Aglomeratif dalam adalah yang berarti adanya tahap dimana algoritma nanti akan mempunyai proses mempelajari data-data yang memiliki kemiripan akan digabung menjadi kelompok baru dan nanti nya seluruh data akan terbentuk menjadi kelompok baru[25]

hierarchical clustering aglomeratif memiliki beberapa rumus dasar untuk menghitung pengelompokannya. Rumus-rumus tersebut adalah

$$d_{AB} = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$

Rumus pertama untuk melakukan *hierarchical clustering* aglomeratif adalah rumus Euclidean *distance* atau jarak Euclidean. Rumus Euclidean *distance* berfungsi sebagai metode pengukuran jarak antar dua titik data[26].

$$d(U,V) = \frac{1}{n_u \times n_v} \sum d(U,V); d(U,V) \in D$$

Rumus kedua untuk melakukan *hierarchical clustering* aglomeratif adalah rumus *average linkage* yang berfungsi untuk mengukur jarak antar klister.[26]

3. Hasil Percobaan

Penelitian dilakukan berdasarkan dataset yang diambil dari website https://bdsp2.pertanian.go.id/bdsp/id/lokasi. Dalam website tersebut diambil dataset yang berisi luas area, volume produksi, dan produktivitas tanaman tebu yang ada di seluruh Indonesia dalam rentang waktu 14 tahun, yaitu dari tahun 2010 sampai 2024. Di dalam dataset ini luas panen memiliki satuan hektar, volume produksi memiliki satuan ton, dan produktivitas memiliki satuan kg/ha. Dengan masing-masing variabel mempunyai perwakilan berupa satuan, diharapakan banyak atau luas yang ada dalam dataset penelitian ini dapat terwakili dengan satuan tersebut. Dataset yang dipakai tidak memiliki missing value, tetapi data yang diambil untuk dimasukkan ke dalam dataset hanyalah provinsi atau kabupaten yang memiliki lebih banyak nilai luas area, volume produksi, dan produktivitas selama rentang waktu 14 tahun, sebagai contoh suatu kabupaten dalam rentang waktu 14 tahun memiliki 9 tahun nilai yang bukan 0, maka data tersebut akan diambil dan dimasukkan ke dalam dataset. Sebaliknya, kabupaten atau provinsi yang memiliki nilai 0 lebih dari 50% tidak akan diambil dan dimasukkan ke dalam dataset yang akan diteliti.

Tabel 1 Deskripsi dataset tebu

	Sheet	Kolom	Jumlah data
Dataset Tebu	Luas area	17	116
	Produksi	17	88
	Produktivitas	17	90

Dari tabel di atas, dapat dilihat dari dataset tebu yang digunakan ada 3 sheet yang berisi masing-masing variabel yang digunakan yaitu luas area, produksi (volume produksi), dan produktivitas. Jumlah kolom yang terdapat dalam dataset berjumlah 17 kolom, sementara untuk jumlah data dari masing-masing sheet adalah luas area berjumlah 116 data, sheet produksi berjumlah 88 data, dan produktivitas 90 data.

Setelah dataset selesai disiapkan, dilanjutkan dengan proses koding, dalam proses koding mengetahui skor rata-rata silhouette dan rata-rata davies-boulding indeks akan dilakukan eksperimen selama 9 kali dari setiap masing-masing model algoritma yang digunakan yaitu K-Means dan *hierarchical clustering* aglomeratif.

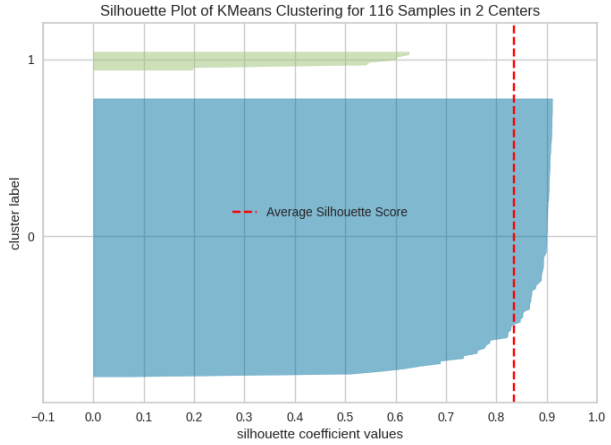
Masuk ke dalam hasil klastering K-Means untuk sheet luas area yang dilakukan selama 9 kali dengan jumlah klaster dari 2 sampai 10

Tabel 2 hasil klasterisasi K-Means luas area

Jumlah klaster	Rata-rata silhouette	Rata-rata davies-bouldin index
2	0.8341	0.4307
3	0.5695	0.6727
4	0.5750	0.7534
5	0.5831	0.5916
6	0.5790	0.4543
7	0.5931	0.3492
8	0.5851	0.3375
9	0.5089	0.4743
10	0.5294	0.4597

Berdasarkan hasil klasterisasi yang menggunakan algoritma model K-Means untuk dataset pada sheet luas area, didapatkan kesimpulan bahwa pemilihan jumlah klaster pada proses klasterisasi memberikan pengaruh yang cukup besar terhadap hasil nilai rata-rata silhouette dan nilai rata-rata davies-boulding indeks. Untuk klasterisasi terbaik yang menggunakan rata-rata silhouette didapatkan saat jumlah klaster nya adalah 2, dengan nilai 0.8341, yang dimana dalam metode menggunakan rata-rata silhouette jika skor semakin mendekati nilai 1 berarti titik sangat cocok dengan klasternya atau data dalam masing-masing klaster terpisah secara jelas dengan klaster lain. Untuk model algoritma yang menggunakan rata-rata davies-bouldin indeks mencapai nilai terendah pada saat jumlah klaster 8, dengan nilai 0.3375, yang dimana dalam metode menggunakan rata-rata davies-bouldin indeks, jika hasil nilai semakin kecil dibawah 1 maka titik sangat cocok dengan klasternya atau data dalam masing-masing klaster terpisah secara jelas dengan klaster lain.

Berikut akan ditampilkan gambar plot dari hasil algoritma yang menggunakan jumlah klaster terbaik.



Gambar 1 Plot K-Means luas area K=2

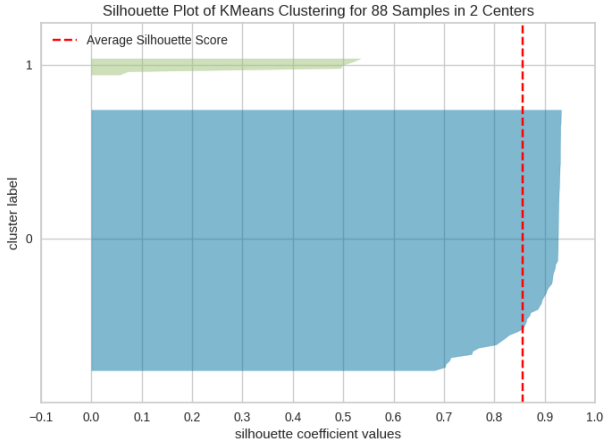
Di atas adalah tampilan plot silhouette K = 2 pada sheet luas area

Tabel 3 hasil klasterisasi K-Means produksi

Jumlah klaster	Rata-rata silhouette	Rata-rata davies-bouldin index
2	0.8571	0.4627
3	0.6227	0.6533
4	0.6122	0.6768
5	0.6124	0.6115
6	0.6075	0.5576
7	0.6022	0.3456
8	0.5510	0.5081
9	0.5250	0.5377
10	0.4648	0.6585

Berdasarkan hasil klasterisasi yang menggunakan algoritma model K-Means untuk dataset pada sheet produksi, didapatkan kesimpulan bahwa pemilihan jumlah klaster pada proses klasterisasi memberikan pengaruh yang cukup besar terhadap hasil nilai rata-rata silhouette dan nilai rata-rata davies-boulding indeks. Untuk klasterisasi terbaik yang menggunakan rata-rata silhouette didapatkan saat jumlah klaster nya adalah 2, dengan nilai 0.8571, yang dimana dalam metode menggunakan rata-rata silhouette jika skor semakin mendekati nilai 1 berarti titik sangat cocok dengan klasternya atau data dalam masing-masing klaster terpisah secara jelas dengan klaster lain. . Untuk model algoritma yang menggunakan rata-rata davies-bouldin indeks mencapai nilai terendah pada saat jumlah klaster 7, dengan nilai 0.3456, yang dimana dalam metode menggunakan rata-rata davies-bouldin indeks, jika hasil nilai semakin kecil dibawah 1 maka titik sangat cocok dengan klasternya atau data dalam masing-masing klaster terpisah secara jelas dengan klaster lain.

Berikut akan ditampilkan gambar plot dari hasil algoritma yang menggunakan jumlah klaster terbaik.



Gambar 2 Plot K-Means produksi K=2

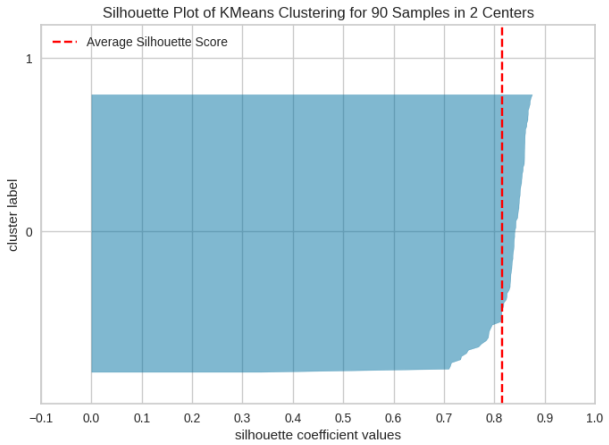
Di atas adalah tampilan plot silhouette K = 2 pada sheet produksi

Tabel 4 hasil klasterisasi K-Means produktivitas

Jumlah klaster	Rata-rata silhouette	Rata-rata davies-bouldin index
2	0.8160	0.1217
3	0.3037	0.8692
4	0.3181	0.6941
5	0.2907	0.9208
6	0.3019	1.037
7	0.1876	1.1485
8	0.2723	1.1136
9	0.2718	1.0322
10	0.2529	0.9547

Berdasarkan hasil klasterisasi yang menggunakan algoritma model K-Means untuk dataset pada sheet produktivitas, didapatkan kesimpulan bahwa pemilihan jumlah klaster pada proses klasterisasi memberikan pengaruh yang cukup besar terhadap hasil nilai rata-rata silhouette dan nilai rata-rata davies-boulding indeks. Untuk klasterisasi terbaik yang menggunakan rata-rata silhouette didapatkan saat jumlah klaster nya adalah 2, dengan nilai 0.8160, yang dimana dalam metode menggunakan rata-rata silhouette jika skor semakin mendekati nilai 1 berarti titik sangat cocok dengan klasternya atau data dalam masing-masing klaster terpisah secara jelas dengan klaster lain. . Untuk model algoritma yang menggunakan rata-rata davies-bouldin indeks mencapai nilai terendah pada saat jumlah klaster 7, dengan nilai 0.1217, yang dimana dalam metode menggunakan rata-rata davies-bouldin indeks, jika hasil nilai semakin kecil dibawah 1 maka titik sangat cocok dengan klasternya atau data dalam masing-masing klaster terpisah secara jelas dengan klaster lain.

Berikut akan ditampilkan gambar plot dari hasil algoritma yang menggunakan jumlah klaster terbaik



Gambar 3 Plot K-Means produktivitas K=2

Terjadinya penurunan nilai rata-rata silhouette dan nilai rata-rata davies-bouldin indeks karena adanya data pada tahun 2024 yang memiliki nilai 0 sehingga membuat proses klastering menjadi tidak seimbang.

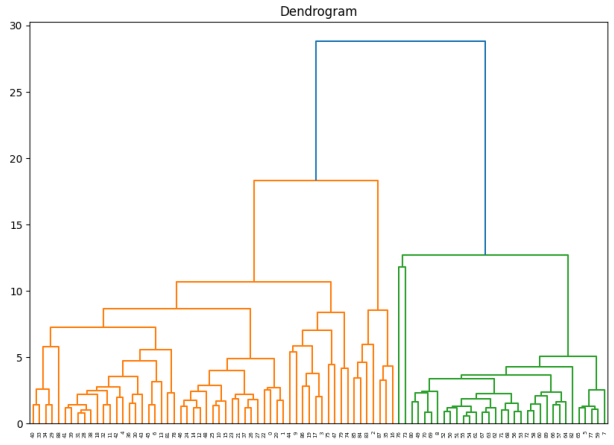
Hasil kesimpulan yang dapat ditarik dari seluruh eksperimen yang dilakukan dengan dataset tebu dengan algoritma model K-Means adalah setelah melakukan eksperimen selama 9 kali dengan jumlah klaster yang berjumlah k = 9, performa terbaik terjadi pada saat jumlah klaster yang dipakai adalah k = 2 dengan skor rata-rata silhouette semua sheet diatas 0.8, sementara untuk rata-rata davies-bouldin indeks pada sheet luas area mendapatkan hasil terbaik pada k = 8 dengan nilai 0.3375, untuk sheet produksi mendapatkan hasil terbaik pada K = 7 dengan nilai 0.3456, dan untuk sheet produktivitas mendapatkan hasil terbaik pada K = 2 dengan nilai 0.1217.

Berikut akan ditampilkan nilai rata-rata dari hasil seluruh eksperimen dengan algoritma model K-Means dari seluruh sheet yang ada di dalam dataset..

Tabel 5 Nilai rata-rata masing-masing sheets

Sheet	Nilai rata-rata silhouette	Nilai rata-rata davies-bouldin index
Luas Area	0,5952	0,5026
Produksi	0,6061	0,5568
Produktivitas	0,3350	0,8769

Masuk ke dalam hasil klastering hierarchical clustering aglomeratif untuk sheet luas area yang dilakukan selama 9 kali dengan jumlah klaster dari 2 sampai 10.



Gambar 4 Dendogram hierarchical clustering

Gambar di atas adalah tampilan dendogram untuk hierachical clustering, dimana dendogram memberikan informasi bahwa data membetuk 3 klaster dan semakin ke kanan maka semakin kecil klaster atau grupnya.

Tabel 6 hasil klasterisasi hierarchical clustering aglomeratif luas area

Jumlah klaster	Rata-rata silhouette	Rata-rata davies-bouldin index
2	0.8223	0.4964
3	0.5621	0.6658
4	0.5674	0.8122
5	0.5807	0.5703
6	0.5940	0.4807
7	0.5661	0.5758
8	0.4453	0.6816
9	0.4782	0.7578
10	0.4732	0.6408

Berdasarkan hasil klasterisasi yang menggunakan algoritma model *hierarchical clustering* aglomeratif untuk dataset pada sheet luas area, didapatkan kesimpulan bahwa pemilihan jumlah klaster pada proses klasterisasi memberikan pengaruh yang cukup besar terhadap hasil nilai rata-rata silhouette dan nilai rata-rata davies-boulding indeks. Untuk klasterisasi terbaik yang menggunakan rata-rata silhouette didapatkan saat jumlah klaster nya adalah 2, dengan nilai 0.8223, yang dimana dalam metode menggunakan rata-rata silhouette jika skor semakin mendekati nilai 1 berarti titik sangat cocok dengan klasternya atau data dalam masing-masing klaster terpisah secara jelas dengan klaster lain. . Untuk model algoritma yang menggunakan rata-rata davies-bouldin indeks mencapai nilai terendah pada saat jumlah klaster 6, dengan nilai 0.4807, yang dimana dalam metode menggunakan rata-rata davies-bouldin indeks, jika hasil nilai semakin kecil dibawah 1 maka titik sangat cocok dengan klasternya atau data dalam masing-masing klaster terpisah secara jelas dengan klaster lain

Tabel 7 hasil klasterisasi hierarchical clustering aglomeratif produksi

Jumlah klaster	Rata-rata silhouette	Rata-rata davies-bouldin index
2	0.8468	0.5161
3	0.8243	0.7980
4	0.5860	0.8461
5	0.5863	0.6656
6	0.5847	0.4681
7	0.5844	0.3634
8	0.5708	0.3988
9	0.5664	0.6556
10	0.5719	0.6205

Berdasarkan hasil klasterisasi yang menggunakan algoritma model *hierarchical clustering* aglomeratif untuk dataset pada sheet produksi, didapatkan kesimpulan bahwa pemilihan jumlah klaster pada proses klasterisasi memberikan pengaruh yang cukup besar terhadap hasil nilai rata-rata silhouette dan nilai rata-rata davies-boulding indeks. Untuk klasterisasi terbaik yang menggunakan rata-rata silhouette didapatkan saat jumlah klaster nya adalah 2, dengan nilai 0.8468, yang dimana dalam metode menggunakan rata-rata silhouette jika skor semakin mendekati nilai 1 berarti titik sangat cocok dengan klasternya atau data dalam masing-masing klaster terpisah secara jelas dengan klaster lain. . Untuk model algoritma yang menggunakan rata-rata davies-bouldin indeks mencapai nilai terendah pada saat jumlah klaster 7, dengan nilai 0.3634, yang dimana dalam metode menggunakan rata-rata davies-bouldin indeks, jika hasil nilai semakin kecil dibawah 1 maka titik sangat cocok dengan klasternya atau data dalam masing-masing klaster terpisah secara jelas dengan klaster lain

Tabel 8 hasil klasterisasi hierarchical clustering aglomeratif produktivitas

Jumlah klaster	Rata-rata silhouette	Rata-rata davies-bouldin index
2	0.3236	1.0712
3	0.3249	1.1203
4	0.3234	1.1189
5	0.3288	0.7606
6	0.3112	1.1552
7	0.2113	1.2547
8	0.2210	1.2093
9	0.2310	1.1741
10	0.2353	1.1699

Berdasarkan hasil klasterisasi yang menggunakan algoritma model *hierarchical clustering* aglomeratif untuk dataset pada sheet produktivitas, didapatkan kesimpulan bahwa pemilihan jumlah klaster pada proses klasterisasi memberikan pengaruh yang tidak terlalu besar terhadap hasil nilai rata-rata silhouette dan nilai rata-rata davies-boulding indeks. Untuk klasterisasi terbaik yang menggunakan rata-rata silhouette didapatkan saat jumlah klaster nya adalah 3, dengan nilai 0.3249, yang dimana dalam metode menggunakan rata-rata silhouette jika skor semakin menjauhi nilai 1 berarti titik terlalu cocok dengan klasternya atau data dalam masing-masing klaster terpisah secara kurang jelas dengan klaster lain. . Untuk model algoritma yang menggunakan rata-rata davies-bouldin indeks mencapai nilai terendah pada saat jumlah klaster 5, dengan nilai 0.7606, yang dimana dalam metode menggunakan rata-rata davies-bouldin indeks, jika hasil nilai semakin besar dibawah 1 maka titik kurang cocok dengan klasternya atau data dalam masing-masing klaster terpisah secara kurang jelas dengan klaster lain

Terjadinya penurunan nilai rata-rata silhouette dan nilai rata-rata davies-bouldin indeks karena adanya data pada tahun 2024 yang memiliki nilai 0 sehingga membuat proses klastering menjadi tidak seimbang.

Hasil kesimpulan yang dapat ditarik dari seluruh eksperimen yang dilakukan dengan dataset tebu dengan algoritma model *hierarchical clustering* aglomeratif adalah setelah melakukan eksperimen selama 9 kali dengan jumlah klaster yang berjumlah k = 9, performa terbaik terjadi pada saat jumlah klaster yang dipakai adalah k = 2 dengan skor rata-rata terbaik silhouette semua sheet diatas 0.8 kecuali pada sheet produktivitas yang mendapatkan rata-rata 0.2 sampai 0.3 saja, sementara untuk rata-rata davies-bouldin indeks pada sheet luas area mendapatkan hasil terbaik pada k = 2 dengan nilai 0.4964, untuk sheet produksi mendapatkan hasil terbaik pada K = 7 dengan nilai 0.3634, dan untuk sheet produktivitas mendapatkan hasil terbaik pada K = 5

Gambar 4 Dendogram hierarchical clustering

dengan nilai 0.7606.

Berikut akan ditampilkan nilai rata-rata dari hasil seluruh eksperimen dengan algoritma model K-Means dari seluruh sheet yang ada di dalam dataset..

Tabel 9 Nilai rata-rata masing-masing sheets

Sheet	Nilai rata-rata silhouette	Nilai rata-rata davies-bouldin index
Luas Area	0.5654	0,6312
Produksi	0.6357	0.5924
Produktivitas	0.2784	1.1592

4. Kesimpulan

Kesimpulan yang bisa didapatkan dari hasil penelitan penggunaan algoritma model K-Means dan model *hierarchical clustering* yang menggunakan dataset luas area, produksi, dan produktivitas tebu di Indonesia, didapatkan :

- K-Means memiliki performa yang lebih stabil dibandingkan model *hierarchical clustering* yang dalam penelitian

menggunakan parameter metriks skor silhouette dan davies-bouldin index.

- Rata-rata nilai silhouette yang tertinggi didapatkan saat penggunaan jumlah klister = 2.
- Nilai davies-bouldin index yang terbaik/terkecil didapatkan tergantung pada pemilihan penggunaan sheet pada dataset.
- Sheet produktivitas pada dataset menjadi fitur atau variabel yang memiliki nilai silhouette dan davies-bouldin index terendah, hal tersebut dimungkinkan disebabkan oleh adanya nilai 0 di tahun 2024 yang membuat dataset tidak seimbang saat proses klastering.
- Dataset bekerja lebih baik dan optimal dengan menggunakan algortma model karena K-Means bekerja lebih baik dalam proses klastering atau pengelompokkan yang dilakukan.

Dapat disimpulkan bahwa algoritma model K-Means bekerja lebih optimal dengan dataset yang digunakan dibanding dengan menggunakan algoritma model hierarchical clustering. Peneliitian ini juga terdapat keterbatasan, yaitu banyaknya nilai 0 pada dataset di tahun 2024 pada sheet produktivitas. Penelitian ke depannya atau selanjutnya akan bisa mendapatkan hasil yang lebih baik dengan model algoritma dan metode lain terutama dalam menghadapi keterbatasan dalam penelitian ini.

REFERENSI

[1] Nurhajijah, "PENGARUH PEMBERIAN NIPAGIN TERHADAP PERKEMBANGAN PENGGEREK BATANG TEBU BERGARIS (Chilo sacchariphagus)," *Jurnal Ilmu Pertanian*, vol. 1, no. 1, pp. 1–5, Jun. 2022.

[2] Agus Supriono *et al.*, "Review Peraturan Daerah Provinsi Jawa Timur Nomor 17 Tahun 2012 tentang Peningkatan Rendemen dan Hablur Tanaman Tebu," *Jurnal Pangan*, vol. 32, no. 3, pp. 1–18, Dec. 2023.

[3] Agnes Verawaty Silalahi, "KEBIJAKAN PENGEMBANGAN TEBU MENUJU SWASEMBADA GULA KONSUMSI," *Journal of Agricultural Development Planning*, vol. 1, no. 1, pp. 1–12, Dec. 2024.

[4] K. R. Ardiansyah and N. Palasara, "Perancangan Sistem Informasi Monitoring Capaian Kinerja Pegawai Di Badan Pusat Statistik Kabupaten Tasikmalaya," 2022. [Online]. Available: http://jurnal.bsi.ac.id/index.php/simpatik

[5] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[6] M. Allaoui, M. L. Kherfi, and A. Cheriet, "Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, 2020, pp. 317–325. doi: 10.1007/978-3-030-51935-3_34.

[7] E. A. Saputra and Y. Nataliani, "Analisis Pengelompokan Data Nilai Siswa untuk Menentukan Siswa Berprestasi Menggunakan Metode Clustering K-Means," *Journal of Information Systems and Informatics*, vol. 3, no. 3, 2021, [Online]. Available: http://journal-isi.org/index.php/isi

[8] E. U. Oti, M. O. Olusola, F. C. Eze, and S. U. Enogwe, "Comprehensive Review of K-Means Clustering Algorithms," *International Journal of Advances in Scientific Research and Engineering*, vol. 07, no. 08, pp. 64–69, 2021, doi: 10.31695/ijasre.2021.34050.

[9] A. Caggiano, F. Napolitano, and R. Teti, "Hierarchical cluster analysis for pattern recognition of process conditions in die sinking EDM process monitoring," in *Procedia CIRP*, Elsevier B.V., 2021, pp. 514–519. doi: 10.1016/j.procir.2021.03.071.

[10] J. Homepage, K. Pratama Simanjuntak, and U. Khaira, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Hotspot Clustering in Jambi Province Using Agglomerative Hierarchical Clustering Algorithm Pengelompokkan Titik Api di Provinsi Jambi dengan Algoritma Agglomerative Hierarchical Clustering," vol. 1, pp. 7–16, 2021.

[11] D. Arizki Kuswardana, D. Arman Prasetya, and I. Gede Susrama Mas Diyasa, "Comparison of Elbow and Silhouette Methods in Optimizing K-Prototype Clustering for Customer Transactions," *Jurnal Ilmiah Edutic : Pendidikan dan Informatika*, vol. 12, no. 1, pp. 43–48, 2025, doi: 10.21107/edutic.v12i1.29744.

[12] D. A. Tarigan, "Optimization of the K-Means Clustering Algorithm Using Davies Bouldin Index in Iris Data Classification," *Media Online)*, vol. 4, no. 1, pp. 545–552, 2023, doi: 10.30865/klik.v4i1.964.

[13] J. Blank and K. Deb, "Pymoo: Multi-Objective Optimization in Python," *IEEE Access*, vol. 8, pp. 89497–89509, 2020, doi: 10.1109/ACCESS.2020.2990567.

[14] C. R. Harris *et al.*, "Array programming with NumPy," Sep. 17, 2020, *Nature Research*. doi: 10.1038/s41586-020-2649-2.

[15] Polina Lemenkova, "Python libraries matplotlib, seaborn and pandas for visualization geospatial datasets generated by QGIS," *HAL Open Science*, vol. 64, no. 1, pp. 13–32, 2020.

[16] M. N. Fahmi, "Implementasi Mechine Learning menggunakan Python Library : Scikit-Learn (Supervised dan Unsupervised Learning)," *Sains Data Jurnal Studi Matematika dan Teknologi*, vol. 1, no. 2, pp. 87–96, Dec. 2023, doi: 10.52620/sainsdata.v1i2.31.

[17] A. Yudhistira and R. Andika, "Pengelompokan Data Nilai Siswa Menggunakan Metode K-Means Clustering," *Journal of Artificial Intelligence and Technology Information (JAITI)*, vol. 1, no. 1, pp. 20–28, Mar. 2023, doi: 10.58602/jaiti.v1i1.22.

[18] A. A. Simangunsong, I. Gunawan, Z. M. Nasution, and G. Artikel, "Pengelompokkan Hasil Produksi Tanaman Perkebunan Berdasarkan Provinsi Menggunakan Metode K-Means Clustering Production of Plantation Crops by Province Using the K-Means Method Article Info ABSTRAK," *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, vol. 1, no. 4, pp. 2828–9099, 2022, doi: 10.55123/jomlai.v1i4.1661.

[19] S. S. Yassin and Pooja, "Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach," *SN Appl Sci*, vol. 2, no. 9, Sep. 2020, doi: 10.1007/s42452-020-3125-1.

[20] F N Dhewani, D Amelia, D N Alifah, B N Sari, and M Jajuli, "Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM," *Jurnal Teknologi dan Informasi (JATI)*, vol. 12, no. 1, pp. 1–14, Mar. 2022.

[21] A A Aldino, D Darwis, A T Prastowo, and C Sujana, "Implementation of K-Means Algorithm for Clustering Corn Planting Feasibility Area in South Lampung Regency," *J Phys Conf Ser*, vol. 1, pp. 2–10, 2020.

[22] Christopher Briggs, Zhong Fan, and Peter Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-IID data," *School of Computing & Mathematics*, vol. 2, pp. 1–9, May 2020.

[23] F. Wang *et al.*, "An Automatic Hierarchical Clustering Method for the LiDAR Point Cloud Segmentation of Buildings via Shape Classification and Outliers Reassignment," *Remote Sens (Basel)*, vol. 15, no. 9, May 2023, doi: 10.3390/rs15092432.

[24] I. Maflahah, D. F. Asfan, S. J. Utomo, A. S. Fathor, and R. A. Firmansyah, "Agglomeration Agricultural Zone Analysis Based on the Hybrid Hierarchical Clustering Method in Madura, Indonesia," *International Journal of Industrial Engineering and Production Research*, vol. 35, no. 4, pp. 75–90, Dec. 2024, doi: 10.22068/ijiepr.35.4.2110.

[25] Afdhah Nur Riadhoh, Galuh Eka Puspita, Inas Rafidah, and Edy Widodo, "Pengelompokan Kabupaten/Kota Berdasarkan Produksi Tanaman Pangan Sumatera Utara Tahun 2020 Menggunakan Pengelompokan Hirarki Aglomeratif," *KUBIK: Jurnal Publikasi Ilmiah Matematika*, vol. 6, no. 2, pp. 1–10, Nov. 2021.

[26] S. Dalimunthe and A. Hanafiah, "Implementation of Agglomerative Hierarchical Clustering Based on The Classification of Food Ingredients Content of Nutritional Substances," *IT Journal Research and Development*, vol. 6, no. 1, pp. 60–69, Aug. 2021, doi: 10.25299/itjrd.2021.6872.