PERBANDINGAN BISECTING K-MEANS DAN AGGLOMERATIVE HIERARCHICAL CLUSTERING UNTUK PENGELOMPOKAN DAERAH PENGHASIL KACANG HIJAU

Lisa[1], Teny Handhayani [2]*
[1,2] Program Studi Teknik Informatika, Universitas Tarumanagara, Jakarta, Indonesia
tenyh@fti.untar.ac.id

ABSTRAK

Tujuan dari penelitian ini adalah membandingkan metode klasterisasi, seperti Bisecting K-Means dan Agglomerative Hierarchical Clustering (AHC) untuk mengklaster produksi kacang hijau di Indonesia. Bisecting K-Means dipilih karena dapat mengelompokkan data dengan cepat, efisien, dan merata, sedangkan AHC dipilih karena dapat menangani data multivariat dengan struktur hierarkinya. Selama lima tahun terakhir, permintaan kacang hijau untuk ekspor internasional telah meningkat, menjadikan hal ini sebagai potensial bagi Indonesia untuk meningkatkan perekonomiannya. Kacang hijau sendiri memiliki manfaat yang beragam, seperti harganya yang murah, fleksibilitas budidaya, dan juga meningkatkan hemoglobin. Permasalahan yang diteliti yaitu bagaimana memetakan daerah-daerah penghasil kacang hijau Indonesia. Metode yang digunakan yaitu Bisecting K-Means dan Agglomerative Hierarchical Clustering. Penelitian ini mengelompokkan wilayah di Indonesia berdasarkan besarnya luas panen, produksi, dan produktivitas kacang hijau yang diharapkan dapat membantu meningkatkan distribusi produksinya. Hasil penelitian menemukan bahwa Agglomerative Hierarchical Clustering berkinerja lebih baik dibandingkan Bisecting K-Means dengan skor Silhouette sebesar 0,841 and Davies-Bouldin Index 0,238. Walaupun Bisecting K-Means bekerja 3,6 milidetik lebih cepat, metode ini hanya menghasilkan skor Silhouette sebesar 0,777 and DBI sebesar 0,785. Menggunakan metode AHC, data dikelompokkan menjadi 2 klaster (tinggi dan rendah) yang menunjukkan bahwa hanya 4 wilayah memiliki produksi kacang hijau yang tinggi, yakni Grobogan, Demak, Temanggung, dan Sumbawa. Kontribusi penelitian ini yaitu pemetaan wilayah penghasil kacang hijau di Indonesia untuk identifikasi daerah yang berpotensi untuk industri pertanian kacang hijau.

Keyword : *agglomerative hierarchical clustering, bisecting k-means, kacang hijau*

1. PENDAHULUAN

Di Indonesia, agrikultur memainkan peran yang penting dalam kehidupan masyarakatnya, baik sebagai sumber distribusi pangan dan juga untuk perkembangan ekonomi [1]. Keanekaragaman pangan di Indonesia menguntungkan dalam menjaga ketahanan pangan karena setiap tanaman memiliki keunggulan dan manfaat yang berbeda – beda. Pangan dapat diartikan sebagai produk dari sumber hasil hayati, seperti pertanian, perkebunan, dan peternakan yang diolah maupun tidak diolah yang dapat dikonsumsi oleh manusia [2].

Jenis tanaman kacang – kacangan (leguminosae atau legum) termasuk komoditas yang murah namun kaya akan kandungan gizi, lemak, vitamin, dan serta, serta dapat dilakukan pengolahan pada tahap pertumbuhan muda sampai tua [3]. Pemanfaatan legum dapat dimaksimalkan karena fleksibilitas dalam pengolahannya tersebut. Hal ini membuat kacang – kacangan menjadi jenis pangan yang sangat bermanfaat karena memiliki manfaat yang beragam.

Kacang hijau merupakan jenis tanaman polong – polongan yang memiliki tingkat protein nabati tertinggi ketiga, setelah kedelai dan kacang tanah [3]. Penelitian lain juga menemukan bahwa ternyata kacang hijau juga dapat meningkatkan kadar hemoglobin pada orang yang terkena anemia, yang rentan ditemukan pada remaja putri [4]. Hal lain yang menjadikannya unggul dibandingkan tanaman serupa adalah cepatnya memasuki tahap dewasa, tahan terhadap kekeringan, jarang terkena penyakit, dan kemampuannya untuk berkembang di tanah yang kurang subur [5].

Dalam lima tahun terakhir, permintaan ekspor terhadap kacang hijau terus meningkat sehingga menjadi sebuah potensi untuk mengembangkan perekonomian negara [6]. Permintaan ekspor yang tinggi berarti juga terdapat daya saing yang tinggi dari negara penghasil kacang hijau lain. Namun jumlah hasil produksi kacang hijau di Indonesia pada tahun 2018 mengalami penurunan sebesar 2,7% dari tahun sebelumnya yang disebabkan oleh salinitas tanah atau kadar garam pada tanah yang tinggi dan juga rendahnya minat tanam oleh petani [6][7]. Maka perlu adanya manajemen distribusi untuk memaksimalkan produksi yang lebih merata demi meningkatkan daya saing kacang hijau Indonesia di pasar internasional. Penelitian ini mengusulkan penerapan metode clustering *Bisecting K-Means* dan *Agglomerative Hierarchical Clustering* (AHC) untuk menganalisis pola distribusi kacang hijau di Indonesia.

Clustering adalah metode pembelajaran mesin tanpa pengawasan untuk mengelompokkan data tak berlabel berdasarkan kemiripan fitur karakteristik [8][9]. Teknik ini menghasilkan kategori data yang dapat digunakan untuk mengungkapkan pola yang

ada dari beberapa atribut [10]. Clustering merupakan metode yang banyak digunakan untuk analisis data dalam berbagai studi kasus [11] [12][13][14][15] .

Penelitian ini melakukan perbandingan algoritma Bisecting K-Means dan Agglomerative Hierarchical Clustering dalam pengelompokan distribusi produksi kacang hijau di Indonesia. Kinerja kedua metode akan dibandingkan untuk mencari metode yang paling tepat dan dengan parameter yang optimal untuk masalah pengelompokan wilayah berdasarkan besarnya luas panen, produksi, dan produktivitas kacang hijau. Kemudian analisis dibuat berdasarkan model klaster dengan kinerja terbaik. Model klaster dievaluasi menggunakan Silhouette Score dan Davies-Bouldin Index (DBI). Diharapkan hasil analisis dan pemetaan dari penelitian ini dapat menjadi arahan atau pedoman yang dapat mendukung peningkatan produksi kacang hijau yang juga lebih merata di seluruh Indonesia. Persebaran penanaman yang lebih merata dapat mencegah adanya keterbatasan kacang hijau yang dapat timbul karena perubahan iklim dan sedikitnya wilayah yang membudidayakannya.

Metode *clustering* sudah banyak diterapkan pada data pertanian untuk melihat potensi daerah menghasilkan suatu tanaman. Metode *K-Means* sudah umum digunakan untuk mengelompokkan daerah penghasil pangan. Salsabila, dkk. pada tahun 2025 menggunakan algoritma *K-Means* untuk mengelompokkan daerah penghasil tanaman jahe di Kabupaten Sumenep berdasarkan produktivitas menjadi 2 dengan produktivitas tinggi dan rendah, yaitu *cluster* dengan 104 kecamatan dan *cluster* 4 kecamatan [8]. Data yang digunakan sebanyak 108 sampel yang berisi luas panen dan produksi dari tahun 2020 sampai 2023 dari Badan Pusat Statistik (BPS) Kabupaten Seumenep. Penelitian tersebut berhasil melakukan *clustering* produktivitas tanaman jahe di Kabupaten Sumenep dengan metode *K-Means* secara akurat sesuai dengan parameter yang ditentukan.

Pengelompokan pangan lain ditelaah oleh Nadilla, dkk. dengan melakukan perbandingan metode *K-Means* dan *K-Medoids* untuk mengelompokan produksi padi di Pulau Sumatera. Penelitian yang dilakukan bertujuan untuk melihat pola produksi padi dengan mengelompokkan daerah dengan hasil produksi tertinggi dan terendah. Pada penelitian ini, performa *K-Means* lebih cepat dan efisien, sedangkan *K-Medoids* lebih resisten terhadap data yang strukturnya tidak umum [16]. *K-Medoids* dengan 2 *cluster* merupakan metode terbaik dengan DBI sebesar 0,43, sedangkan *K-Means* dengan 3 cluster memiliki DBI sebesar 0,44. K-Medoids mengelompokkan data menjadi 2 klaster, yaitu klaster pertama dengan produksi tinggi yang berisi 54 wilayah dan klaster kedua dengan produksi rendah yang berisi 34 wilayah.

Azari, dkk. melakukan penelitian guna mengelompokan produksi padi dan beras di Jawa Timur tahun 2023 menggunakan metode AHC. Padi dan beras merupakan makanan pokok di Indonesia sehingga penelitian ini berguna untuk melihat wilayah dengan produksi padi dan beras yang rendah agar dapat dilakukan mitigasi. Algoritma terbaik pada penelitian ini menggunakan *average linkage* berdasarkan korelasi *cophenetic* tertinggi [17]. Nilai *silhouette* tertinggi yang dicapai adalah sebesar 0,6347 dengan menggunakan 4 klaster. Hasilnya diperoleh klaster 1 dengan kategori produksi tinggi sebanyak 4 kabupaten, klaster 2 kategori rendah yang berisi 18 kabupaten, klaster 3 berkategori sedang dengan 7 kabupaten, dan klaster 4 berkategori sangat rendah sebanyak 9 kota.

Adapun pengelompokan tanaman padi dilakukan oleh Siti Fitri berdasarkan wilayah menjadi tingkat penghasil tinggi, sedang, dan rendah menggunakan metode *Self-Organizing Maps* (SOM) [18]. Ditemukan hasil model terbaik ketika menggunakan learning rate pada metode SOM sebesar 0,9 dengan Silhouette tertinggi sebesar 0,6388. Selain itu, melalui penelitian ini ditemukan banyak daerah dengan distribusi padi yang rendah di provinsi Jawa Timur.

Sejauh ini metode K-Means paling sering digunakan untuk pengelompokan produksi tanaman pangan namun belum ditemukan penggunaan metode *Bisecting K-Means*. Sedangkan metode AHC dengan berbagai parameter sudah banyak digunakan untuk mengelompokkan daerah produksi pangan. Dalam penelitian ini model AHC menghasilkan skor Silhouette yang lebih tinggi dibandingkan dengan hasil yang diperoleh Azari, dkk. yaitu sebesar 0,841. Selain itu dibandingkan dengan penelitian yang sudah disebutkan, studi ini mengelompokkan kacang hijau berdasarkan variabel luas panen, produksi, dan produktivitas kabupaten / kota di seluruh Indonesia.

Penelitian pendahulu masih bersifat analisis pada wilayah tertentu [16][17][18], sehingga diperlukan penelitian pemetaan wilayah produksi pertahian yang mencakup wilayah nasional. Penelitian ini bertujuan untuk membandingkan algoritma Bisecting K-Means dan Agglomerative Hierarchical Clustering untuk pemetaan wilayah penghasil kacang hijau di Indonesia. Kontribusi penelitian ini yaitu pemetaan wilayah penghasil kacang hijau di Indonesia untuk identifikasi daerah yang berpotensi untuk industri pertanian kacang hijau.

2. TINJAUAN PUSTAKA

2.1. Bisecting K-Means

Metode *Bisecting K-Means* merupakan turunan dari algoritma *K-Means*. *K-Means* bekerja dengan menetapkan pusat *cluster* dari sampel data sebagai centroid awal, lalu data diklaster berdasarkan pada jarak terdekatnya dengan centroid.

Centroid setiap klaster dihitung ulang setiap iterasi sampai data pada *cluster* tidak berubah lagi [19].

Namun pada *Bisecting K-Means*, data awalnya dianggap menjadi 1 klaster. Kemudian metode *Bisecting K-Means* membaginya menjadi 2 sub-klaster menggunakan metode *K-Means*, lalu proses ini diulang terus menerus [20]. Metode membagi berdasarkan sub-klaster dengan *sum of squared errors* (SSE) tertinggi sehingga mendapat hasil klaster terbaik, yaitu ketika anggota klaster paling selaras [21]. Iterasi berhenti sampai jumlah klaster mencapai jumlah yang telah ditentukan sebelumnya.

2.2. Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering (AHC) termasuk metode *clustering* hierarki, yang bekerja dengan menggabungkan pasangan *cluster* berdasarkan jarak terdekat atau paling mirip sehingga membentuk *cluster* baru [22]. Pada hasil akhir, data akan menjadi *cluster* tunggal yang memiliki identitas berupa tingkatan yang mewakili similitaritas. Jarak pada metode ini dihitung menggunakan rumus *Euclidean* yang ditampilkan pada Persamaan (1) [17][23]:

$$D_{(x,y)} = \sqrt{\sum_{i=1}^{n} \left(x_{ij} - y_{jk}\right)^2} \quad (1)$$

dimana $D_{(x,y)}$ adalah jarak *Euclidean* antara data x dan y, dan n mewakili jumlah sampel data. Variabel x_{ij} adalah nilai data ke-i pada variabel j dan y_{jk} merupakan nilai data ke-j pada variabel k.

Metode AHC memiliki beberapa macam teknik yang dapat digunakan untuk menghubungkan klaster. Adapun algoritma yang sering digunakan adalah [22], [24]:

a. *Single Linkage*
   Metode ini menemukan hubungan antara klaster dengan menentukan jarak terdekat dari anggota klaster. Rumus perhitungan metode *single linkage* dapat dilihat pada persamaan (2):
   $$D_{(x,y)w} = min\{D_{(x,w)} \cdot D_{(y,w)}\} \quad (2)$$
   dimana $D_{(x,w)}$ adalah jarak terdekat antara klaster x dan w, sedangkan $D_{(y,w)}$ mewakili jarak terdekat antara klaster y dan w.
b. *Complete Linkage*
   Metode ini menemukan hubungan antara klaster dengan menentukan jarak terdekat dari anggota klaster. Rumus perhitungan metode *single linkage* dapat dilihat pada persamaan (2):
   $$D_{(x,y)w} = maks\{D_{(x,w)} \cdot D_{(y,w)}\} \quad (3)$$
   dimana $D_{(x,w)}$ adalah jarak terdekat antara klaster x dan w, sedangkan $D_{(y,w)}$ mewakili jarak terdekat antara klaster y dan w.
c. *Average Linkage*
   Metode AHC dengan teknik Average Linkage mengelompokkan data dengan menghitung rata – rata jarak dari kelompok klaster. Perhitungan ini dapat dilakukan dengan persamaan (4):
   $$D_{(x,y)w} = \frac{\Sigma_i \Sigma_k D_{ik}}{N_{x,y} N_w} \quad (4)$$
   dimana $D_{(i,k)}$ adalah jarak antara data i pada klaster (x, w) dan data k pada klaster w. Sedangkan $N_{x,y}$ dan N_w mewakili banyaknya data pada klaster (x, w) dan w.

2.3. Skor *Silhouette* dan *Davies-Bouldin Index*

Skor *Silhouette* digunakan untuk mengevaluasi kinerja model *clustering*. Metode ini memuat perhitungan jarak dengan metode *Euclidean*. *Silhouette* skor diukur dengan menghitung rata – rata jarak dekatnya suatu sampel data dengan sampel data lain di klaster yang sama dan berbeda [25]. Skor ini menghasilkan nilai yang memiliki rentang antara -1 hingga 1 [26]. Klaster dianggap optimal jika nilai skor mendekati 1 dan sampel data lebih dekat dengan sesama anggota sampel di klaster yang sama dibandingkan dengan sampel di klaster lain. Perhitungan skor *Silhoutte* ditampilkan pada persamaan (5) [27]:

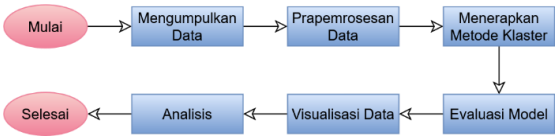
$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (5)$$

dimana $S(i)$ merupakan skor *Silhouette*, $a(i)$ adalah jarak rata – rata antara sampel i dengan seluruh objek data di klaster yang sama, sedangkan $b(i)$ mewakili jarak rata – rata antara sampel i dengan seluruh data di klaster lainnya.

Davies-Bouldin Index (DBI) menilai performa model dalam melakukan pengelompokan data dengan dengan menghitung jarak antara data dengan anggota data lain di klaster yang sama [26]. Maka nilai DBI yang ideal adalah nilai yang mendekati 0 tapi bukan negatif [28]. Klaster yang optimal adalah ketika data memiliki jarak yang dekat dengan sesama anggota di klaster yang sama karena menunjukkan hasil pengelompokan yang baik dan jelas.

3. METODE PENELITIAN

Tahap alur penelitian ini diilustrasikan pada Gambar 1. Tahap pertama adalah mengumpulkan data dari sumber yang dapat dipercaya dan sesuai dengan metode yang digunakan. Pada penelitian ini data dikumpulkan dari situs Basis Data Statistik Pertanian (BDSP). Hasil tahap pertama yaitu dataset yang berisi daftar kota, luas lahan, dan hasil produksi kacang hijau di Indoensia. Kemudian data dilakukan prapemrosesan dengan meniadakan daerah yang memiliki nilai kosong lebih dari 35%, mengisi nilai yang hilang, serta melakukan normalisasi data. Selanjutnya menerapkan metode *Bisecting K-Means* dan AHC pada data dan melakukan evaluasi kinerja model dengan skor *Silhouette* dan *Davies-Bouldin Index*. Visualisasi data dibuat berdasarkan model dengan performa terbaik. Visualisasi dibuat untuk mendukung hasil analisis.



Gambar 1. Tahap alur penerapan metode clustering

Sampel data memiliki banyak nilai kosong, baik karena daerah tidak menghasilkan kacang hijau maupun karena nilai yang dicatat pada hari tertentu, seperti pada hari libur nasional. Terdapat kabupaten / kota dengan nilai kosong sebanyak 100%, namun tetap akan diikutsertakan pada penelitian ini karena daerah tersebut artinya tidak memiliki lahan pertumbuhan kacang hijau. Sedangkan data dengan nilai kosong sebanyak lebih dari 35% tidak akan dimasukkan dalam penelitian.

Sampel data yang memiliki nilai kosong diisi dengan *forward fill* dan *backward fill*. *Forward fill* atau dengan nama lain *Last Observation Carried Forward* (LOCF) melengkapi nilai kosong dengan nilai pada sampel sebelumnya yang terisi [29]. *Backward Fill* atau disebut juga dengan *Next Observation Carried Backward* (NOCB) adalah metode yang pengisian nilai kosong dengan mengambil dari nilai pada sampel setelahnya yang terisi [30]. Dengan menggunakan kedua metode ini, data yang sebelumnya memiliki nilai kosong akan terisi semua sehingga data dapat digunakan. Hasil dari pra-pemrosesan data yaitu dataset yang sudah siap digunakan untuk masukan bagi metode clustering.

Data berisi nilai numerikal sehingga terdapat nilai *outlier* atau nilai yang menonjol dari data lainnya sehingga perlu dilakukan normalisasi. Pada penelitian ini normalisasi menggunakan fungsi *Standard Scaler*. Hasil nilai yang dinormalisasi memiliki rentang dari 1 sampai -1.

Tahapan selanjutnya yaitu menerapkan metode *clustering* dan mengevaluasi hasilnya. Metode clustering yang digunakan yaitu Bisecting K-Means dan AHC. Hasil dari tahap clustering yaitu cluster (kelompok) wilayah. Sedangkan metode evaluasi yaitu Silhouette dan Davies-Bouldin Index. Untuk mencari klaster yang optimal, data dicoba untuk dikelompokkan menjadi 2 sampai 11 klaster. Hasil dari tahapan evaluasi model yaitu nilai Silahouette dan Davies-Bouldin Idnex. Setelah melakukan *clustering*, tahapan dilanjutkan dengan visualisasi hasil *clustering* dan analisis. Hasil dari vaisualisasi yaitu peta wilayah Indonesia berdasarkan cluster dan pola dari setiao cluster.

4. HASIL DAN PEMBAHASAN

Data yang digunakan untuk penelitian ini diperoleh dari situs Basis Data Statistik Pertanian (BDSP) milik Kementerian Pertanian Republik Indonesia. Data terdiri dari variabel luas panen, produksi, dan produktivitas kacang hijau yang mencakup seluruh kabupaten / kota sehingga total sampel menjadi 515 baris. Data memiliki rentang periode 2010 sampai 2024. Variabel luas panen, produksi, dan produktivitas digabung, sehingga jumlah banyaknya kolom menjadi 46.

Data terdapat banyak nilai yang hilang sehingga perlu dilakukan prapemrosesan data. Data dengan nilai kosong yang lebih dari 35% akan tidak diikutsertakan. Namun jika persentase nilai kosong adalah 100% maka sampel akan termasuk pada data karena artinya daerah tersebut tidak menghasilkan kacang hijau. Nilai kosong juga terdapat pada seluruh daerah di tahun 2023 dan 2024, yang artinya data belum tercatat pada periode tersebut sehingga akan dibuang juga. Ditemukan jumlah data yang dibuang sebanyak 191 sampel, sehingga jumlah baris data menjadi 324 dan banyaknya kolom menjadi 40.

Selanjutnya data diterapkan metode normalisasi, yaitu menggunakan perpustakaan fungsi dari *scikit-learn* dengan nama *Standard Scaler*. Hasilnya nilai pada data hanya memiliki rentang antara -1 hingga 1. Hal ini untuk mencegah dampak *outlier* pada performa metode *clustering* yang sensitif.

Metode *clustering Bisecting K-Means* diterapkan pada data secara iteratif menggunakan biji acak dari parameter agar didapat hasil dari metode yang lebih beragam dan merata. Untuk mencari kinerja terbaik dari model, data akan dikelompokkan mulai dari 2 klaster hingga 10 klaster. Gunanya adalah mencari banyaknya klaster yang optimal untuk mengelompokkan data kacang hijau. *Bisecting K-Means* pada penelitian ini menggunakan parameter strategi *bisecting biggest innertia*, algoritma *lloyd*, toleransi 0,0001, dan maksimum iterasi sebesar 300. Hasil *Bisecting K-Means* dengan 2 klaster lebih unggul dibandingkan dengan jumlah klaster lain dengan rata – rata skor *Silhouette* 0.7771, *Davies-Bouldin Index* 0.7853, dan waktu komputasi hanya 4,3 milidetik.

Selanjutnya menerapkan algoritma AHC pada data menggunakan metode pengukuran jarak *Euclidean* dan algoritma *linkage average*. Antara klaster 2 sampai 11, metode ini paling unggul ketika mengelompokkan data menjadi 2 klaster. Dihasilkan rata – rata skor *Sillhouette* 0.8410, DBI sebesar 0.2384, dan waktu komputasi yang selama 7,9 milidetik.

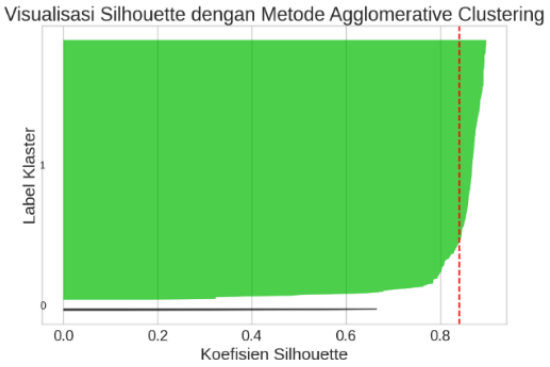
Tabel 1. Perbandingan *Bisecting K-Means* dan AHC

Algoritma	Klaster	*Silhouette*	DBI
Bisecting K-Means	**2**	**0,777**	**0,785**
	3	0,380	1,144
	4	0,217	1,393
	5	0,144	1,565
	6	0,087	1,625
	7	0,077	1,656
	8	0,086	1,589
	9	0,101	1,574
	10	0,096	1,586
Rata – rata		0,218	1,435

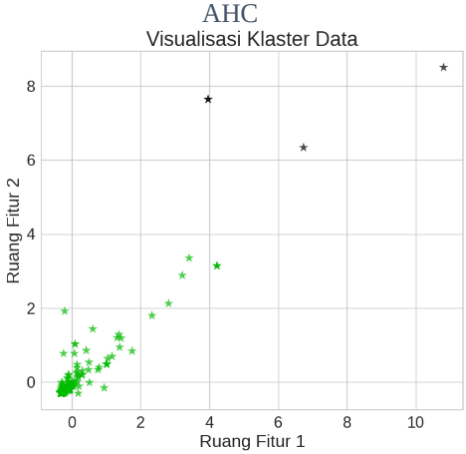
Algoritma	Klaster	*Silhouette*	DBI
AHC	**2**	**0,841**	**0,435**
	3	0,798	0,253
	4	0,698	0,567
	5	0,683	0,517
	6	0,681	0,400
	7	0,631	0,399
	8	0,631	0,345
	9	0,632	0,238
	10	0,564	0,244
Rata – rata		0,684	0,377

Tabel 1 menunjukkan perbandingan algoritma *Bisecting K-Means* dan AHC yang diukur dengan metrik evaluasi skor *Sillhouette* dan *Davies-Bouldin Index*. Jika kedua metode *clustering* dibandingkan berdasarkan skor *Silhouette* dan DBI, AHC menghasilkan kinerja terbaik dibandingkan *Bisecting K*-Means. Metode AHC menghasilkan skor *Silhouette* tertinggi dengan menggunakan 2 klaster, yaitu sebesar 0,841. Sedangkan DBI terendahnya dihasilkan ketika menggunakan 9 klaster, yaitu 0,238. Maka hasil visualisasi dan analisis akan dibuat berdasarkan metode AHC.

Gambar 2 menunjukkan visualisasi *Silhouette* pemetaan klaster menggunakan metode AHC. Nilai 0 adalah daerah dengan luas panen, produksi dan produktivitas kacang hijau yang tinggi di Indonesia, sedangkan angka 1 adalah yang rendah. Dapat diketahui dari gambar *Silhouette* bahwa daerah penghasil kacang hijau yang sangat sedikit.



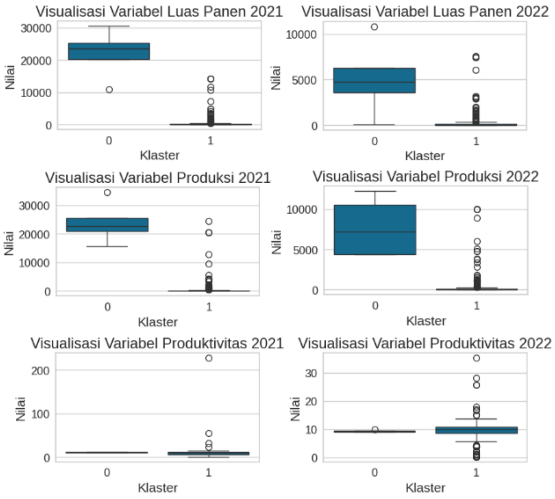
Gambar 2. Visalisasi Silhouette dengan metode AHC



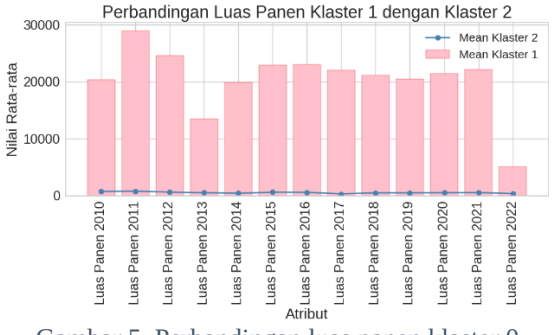
Gambar 3. Visualisasi klaster data

Didukung dengan visualisasi klaster data pada Gambar 3, dapat terlihat bahwa hanya beberapa daerah yang memiliki daya tarik untuk menanam kacang hijau. Hal ini kemungkinan terjadi karena mayoritas petani lebih mengutamakan produksi tanaman lain dibandingkan kacang hijau, yang bisa dilihat berdasarkan

Gambar 4. Gambar 4 menunjukkan pola distribusi data berdasarkan luas panen, produksi, dan produktivitas di setiap kabupaten / kota dari tahun 2020 sampai 2021. Dapat dilihat bahwa produksi kacang hijau di data klaster 0 memiliki keragaman rentang, yaitu mulai dari 5000 sampai 10.000 ton. Diketahui bahwa daerah klaster 0 memiliki luas panen dan produksi yang jauh lebih tinggi dibandingkan klaster 1. Sedangkan produktivitasnya rendah di kedua klaster, terutama pada tahun 2021. Hal ini bisa terjadi karena kacang hijau mampu dipanen dalam tahap hidup muda maupun tua sehingga tidak merata. Namun bisa juga dikarenakan kurangnya daya tarik dalam pengolahan kacang hijau.



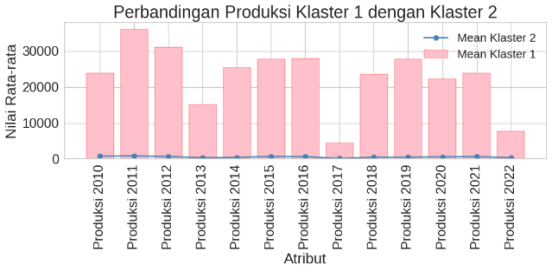
Gambar 4. Visualisasi data klaster



Gambar 5. Perbandingan luas panen klaster 0 dengan klaster 1

Gambar 5 memperlihatkan perbandingan luas panen klaster 0 dan klaster 1. Luas panen pada daerah di klaster 0 jauh lebih tinggi dibandingkan dengan klaster 1, yaitu hampir mencapai 30.000 hektar. Namun dapat terlihat juga bahwa di tahun 2022 luas panen kacang hijau turun drastis dari

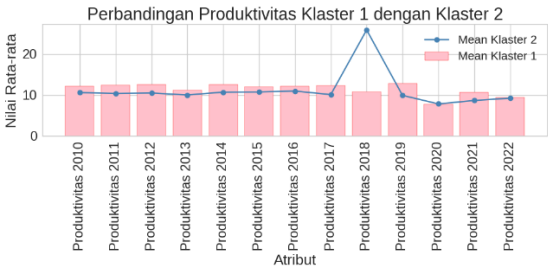
tahun 2021. Gambar 6 menunjukkan perbandingan produksi klaster 0 dan klaster 1. Berdasarkan diagram, jumlah produksi dari klaster 0 semakin lama semakin menurun, walau juga tidak stabil. Hal ini dikarenakan produksi kacang hijau yang kurang merata dan kurang ada di banyak wilayah, walau luas panen sudah besar.



Gambar 6. Perbandingan produksi klaster 0 dengan klaster 1

Kondisi ini menyebabkan produksi kacang hijau menjadi terhambat sehingga terlihat volatil pada diagram.Perbandingan produktivitas kacang hijau pada klaster 0 dan klaster 1 dapat dilihat pada Gambar 7, pada gambar terlihat bahwa produktivitas kacang hijau dari klaster 1 mengalami kenaikan yang pesat di tahun 2018. Kemungkinan peristiwa ini terjadi karena adanya penurunan produksi di tahun 2017 dari klaster 0, yang merupakan daerah sumber produksi kacang hijau pada biasanya. Produktivitas kacang hijau pada klaster 0 dan klaster 1 secara garis besar bersifat stabil dari tahun 2018 sampai 2024.

Dapat dilihat pada Gambar 7 sebelumnya bahwa distribusi data kurang optimal dan butuh dilakukan normalisasi data lebih lanjut. Data dari daerah yang menghasilkan kacang hijau tinggi sangat menonjol dibandingkan dengan daerah yang menghasilkan kacang hijau rendah. Selain itu Bisecting K-Means sangat rentan terhadap overfitting karena dilihat dari Tabel 1 model dengan klaster 3 sampai 11 menghasilkan Silhouette yang rendah, di bawah 0,4, menunjukkan kinerja model yang kurang baik dalam mengelompokkan data.



Gambar 7. Perbandingan produktivitas klaster 0 dengan klaster 1

Gambar 8 menunjukkan visualisasi daerah di Indonesia dengan produksi kacang hijau yang tinggi dan rendah. Wilayah dengan produksi kacang hijau yang rendah diberi warna biru, sedangkan yang tinggi ditandai warna merah. Dari hasil pemetaan, ternyata daerah yang memproduksi kacang hijau tinggi di Indonesia hanya ada di beberapa wilayah saja. Daerah dengan produksi kacang hijau tinggi di provinsi Jawa Tengah hanya terdapat di wilayah Grobongan, Demak, dan Temanggung. Selain itu, kabupaten Sumbawa di Nusa Tenggara Barat juga termasuk wilayah dengan produksi kacang hijau yang tinggi.

Gambar 8. Visualisasi daerah penghasil kacang hijau di Indonesia

Kekurangan metode AHC yiatu komputasi yang lama untuk data besar. Sedangkan kelemahan dari Bisecting K-Means yaitu terpengaruh oleh inisialisasi dari nilai awal pada saat menentukan centroid dan membutuhkan penentuan jumlah cluster. Peningkatan yang dapat dilakukan yaitu mengembangkan metode penentuan centroid untuk Bisecting K-Means. Sedangkan kecepatan performa AHC dapat tingkatkan melalui komputasi paralel untuk data besar. .

5. KESIMPULAN DAN SARAN

Berdasarkan penelitian ini, dapat disimpulkan bahwa bahwa metode Agglomerative Hierarchical Clustering mengungguli metode Bisecting K-Means dari segi skor Silhouette dan DBI. AHC menghasilkan performa terbaik dengan 2 klaster sehingga mendapat skor Silhouette tertinggi, yaitu sebesar 0,841. Sedangkan skor Silhouette tertinggi oleh Bisecting K-Means hanya mencapai 0,777 dengan 2 klaster, walaupun memakan waktu lebih cepat 3,6 milidetik dibandingkan metode AHC. Penelitian ini menunjukkan bahwa wilayah dengan produksi kacang hijau tertinggi hanya terdapat di Jawa Tengah, seperti Grobogan, Demak, dan Temanggung, serta di Nusa Tenggara Barat, yaitu di kabupaten Sumbawa. Peningkatan daya tarik ekspor kacang hijau dapat menjadi kesempatan bagi Indonesia untuk meningkatkan produksi kacang hijau yang lebih banyak dan terdistribusi. Namun kondisi seperti ini membuat potensi daya saing produksi kacang hijau Indonesia di pasar internasional menjadi rendah. Dari hasil visualisasi pengelompokan data, terlihat bahwa data masih terdapat outlier walaupun sudah dilakukan normalisasi dengan Standard Scaler. Hal ini menunjukkan bahwa data kurang seimbang sehingga metode normalisasi kurang dapat menangani adanya *outlier* antara daerah dengan produksi kacang hijau tinggi dan rendah. Selain itu

Bisecting K-Means butuh dilakukan optimasi parameter lebih lanjut karena hanya model dengan 2 klaster yang menghasilkan skor Silhouette melebihi 0,4. Selain itu jarak antara klaster dengan pusat masih relatif jauh jika dilihat dari skor DBI yang tinggi, yaitu sebesar 0,785. Penelitian ini dapat dikembangkan dengan menambahkan komponen meteorologi untuk menganalisis pengaruh perubahan iklim terhadap hasil produksi pertanian.

DAFTAR PUSTAKA

[1] Sari and Indah Retno, "Implementasi Convolutional Neural Networks (Cnn) Untuk Klasifikasi Citra Benih Kacang Hijau Berkualitas," UNIVERSITAS MUHAMMADIYAH SEMARANG, 2021.

[2] S. Amin, "Perlindungan Hukum Bagi Konsumen Muslim terhadap produk Pangan yang Tidak Bersertifikat Halal Menurut Undang-Undang Nomor 33 Tahun 2014 tentang Jaminan Produk Halal," Universitas Islam Sultan Agung Semarang, 2022.

[3] D. Ratnasari, Y. D. Rahmawati, H. Fajarini, and D. Nafisyah, "Potensi Kacang Hijau Sebagai Makanan Alternatif Penyakit Degenaratif," *JAMU: Jurnal Abdi Masyarakat UMUS*, vol. 1, no. 02, 2021.

[4] B. T. Carolin, S. Suprihatin, I. Indirasari, and S. Novelia, "Pemberian Sari Kacang Hijau untuk Meningkatkan Kadar Hemoglobin pada Siswi Anemia," *Journal for Quality in Women's Health*, vol. 4, no. 1, pp. 109–114, 2021.

[5] R. Santosa, "Analisis Daya Saing Kacang Hijau Di Kecamatan Saronggi Kabupaten Sumenep," *JURNAL PERTANIAN CEMARA*, vol. 17, no. 2, pp. 35–49, Dec. 2020, doi: 10.24929/fp.v17i2.1146.

[6] N. E. Ningsih, T. Ekowati, and S. Nurfadillah, "Analisis Daya Saing Kacang Hijau (vigna radiata) Indonesia di Pasar Internasional," *Jurnal Ekonomi Pertanian dan Agribisnis*, vol. 6, no. 4, pp. 1644–1654, 2022.

[7] D. A. A. Elisabeth, S. Sutrisno, S. A. Riyanto, H. Kuntyastuti, and F. Rozi, "Kemampuan Daya Saing Kacang Hijau di Tingkat Usahatani pada Lahan Salin (Studi Kasus di Desa Gesik Harjo, Kecamatan Palang, Kabupaten Tuban)," *Buletin Palawija*, vol. 19, no. 2, pp. 93–102, 2021.

[8] N. Salsabila, K. Aulisari, and H. Z. Zahro, "Penerapan Algoritma K-Means Untuk Klasterisasi Produktivitas Tanaman Jahe," *Infotek: Jurnal Informatika dan Teknologi*, vol. 8, no. 1, pp. 228–238, Jan. 2025, doi: 10.29408/jit.v8i1.28195.

[9] F. Gisolf, Z. J. M. H. Geradts, and M. Worring, "Post-Clustering Merging with Novel Metrics for Multi-Label Image Collections," *Expert Syst Appl*, vol. 288, p. 127875, Sep. 2025, doi: 10.1016/j.eswa.2025.127875.

[10] G. R. Bauer, M. Mahendran, C. Walwyn, and M. Shokoohi, "Latent Variable and Clustering Methods in Intersectionality Research: Systematic Review of Methods Applications," *Soc Psychiatry Psychiatr Epidemiol*, vol. 57, no. 2, pp. 221–237, Feb. 2022, doi: 10.1007/s00127-021-02195-6.

[11] T. Handhayani and Z. Rusdi, "K-Means Using Dynamic Time Warping For Clustering Cities in Java Island According to Meteorological Conditions," 2023. doi: 10.1109/ICIC60109.2023.10381899.

[12] T. Handhayani and I. Lewenusa, "An Intelligent Clustering Approach For Analyzing A Multivariate Time Series Dataset, Case Study COVID-19 Outbreak in Indonesia," in *2023 17th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/TSSA59948.2023.10367007.

[13] T. Handhayani and I. Lewenusa, "An Analysis of Meteorological Data in Sumatra and Nearby using Agglomerative Clustering," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 2, pp. 234–241, Apr. 2024, doi: 10.29207/resti.v8i2.5663.

[14] Y. A. Ishak, T. Handhayani, M. D. L. Sitorus, W. William, J. Pragantha, and I. Lewenusa, "Advanced Clustering Approach for Mapping Regions of Paddy Productivity in Indonesia Using Intelligent K-Means," in *IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS)*, Manado: IEEE, Mar. 2024, pp. 269–274.

[15] William, L. Bayuaji, N. J. Perdana, and T. Handhayani, "Mapping Indonesia's Regions Based on Carbon Emissions Using the K-Means Algorithm," in *2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI)*, IEEE, Jan. 2025, pp. 200–205. doi: 10.1109/ICoCSETI63724.2025.11020099.

[16] S. Nadilla, W. Syahrifa, F. Razi, R. Maulana, and M. Ula, "Analisis Perbandingan Kinerja Algoritma K-Means dan K Medoids untuk Klasterisasi Produksi Padi di Pulau Sumatera," in *Prosiding Seminar Nasional Teknologi dan Teknik Informatika (SENASTIKA)*, 2024.

[17] H. Azari, D. Hartanti, and A. A. Sari, "Pengelompokan Produksi Padi dan Beras Provinsi Jawa Timur dengan Metode Agglomerative Hierarchical Clustering," *Infotek: Jurnal Informatika dan Teknologi*, vol. 7, no. 2, pp. 379–389, Jul. 2024, doi: 10.29408/jit.v7i2.26016.

[18] S. Fitri, "Penetuan Cluster Terbaik pada Pengelompokan Ketahanan Pangan

Menggunakan Metode Self-Organizing Maps (SOM) dan Metode Silhouette Coefficient (Sc): Studi Kasus Provinsi Jawa Timur," Universitas Islam Negeri Maulana Malik Ibrahim, 2022.

[19] U. Govindarajulu and S. Bedi, "K-means for Shared Frailty Models," *BMC Med Res Methodol*, vol. 22, no. 1, p. 11, Dec. 2022, doi: 10.1186/s12874-021-01424-5.

[20] H. Han and J. Oh, "Application of Various Machine Learning Techniques to Predict Obstructive Sleep Apnea Syndrome Severity," *Sci Rep*, vol. 13, no. 1, p. 6379, Apr. 2023, doi: 10.1038/s41598-023-33170-7.

[21] D. Rodríguez-Esparragón, P. Gamba, and J. Marcello, "Automatic Methodology for Forest Fire Mapping with SuperDove Imagery," *Sensors*, vol. 24, no. 16, p. 5084, Aug. 2024, doi: 10.3390/s24165084.

[22] A. F. Dewi and K. Ahadiyah, "Agglomerative Hierarchy Clustering Pada Penentuan Kelompok Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Pendidikan," *Zeta-Math Journal*, vol. 7, no. 2, pp. 57–63, 2022.

[23] L. P. W. Adnyani and P. R. Sihombing, "Analisis cluster time series dalam pengelompokan provinsi di Indonesia berdasarkan nilai PDRB," *Jurnal Bayesian: Jurnal Ilmiah Statistika dan Ekonometrika*, vol. 1, no. 1, pp. 47–54, 2021.

[24] A. M. Jarman, "Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method," *Georgia Southern University*, vol. 29, p. 32, 2020.

[25] B. K. Mishra *et al.*, "An Efficient Framework for Obtaining The Initial Cluster Centers," *Sci Rep*, vol. 13, no. 1, p. 20821, Nov. 2023, doi: 10.1038/s41598-023-48220-3.

[26] R. Ishak, N. Nurmawanti, and A. Bengnga, "Optimization of K-Means in Disease Clustering of Pregnant Women Using Random Forest," *Jambura Journal of Electrical and Electronics Engineering*, vol. 7, no. 1, pp. 41–47, 2025.

[27] K. Amrulloh, T. H. Pudjiantoro, P. N. Sabrina, and A. I. Hadiana, "Comparison Between Davies-Bouldin Index and Silhouette Coefficient Evaluation Methods in Retail Store Sales Transaction Data Clusterization Using K-Medoids Algorithm," in *Proceedings of the 3rd South American International Industrial Engineering and Operations Management Conference, Asuncion, Paraguay*, 2022, pp. 19–21.

[28] I. T. Utami, F. Suryaningrum, and D. Ispriyanti, "K-Means Cluster Count Optimization With Silhouette Index Validation And Davies Bouldin Index (Case Study: Coverage Of Pregnant Women, Childbirth, And Postpartum Health Services In Indonesia In 2020)," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 2, pp. 707–716, 2023.

[29] S. Lin, X. Wu, G. Martinez, and N. V. Chawla, "Filling Missing Values on Wearable-Sensory Time Series Data," in *Proceedings of the 2020 SIAM International Conference on Data Mining*, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2020, pp. 46–54. doi: 10.1137/1.9781611976236.6.

[30] I. K. Omurlu, B. Varol, and M. Ture, "Comparing The Performance of Eight Imputation Methods for Propensity Score Matching in Missing Data Problem," *Journal of Statistics and Management Systems*, vol. 26, no. 4, pp. 915–927, 2023, doi: 10.47974/JSMS-971