(Sosial) Jurnal Ilmiah Pendidikan IPS
Volume 3, Nomor 4, Desember 2025
e-ISSN: 3032-2200; p-ISSN: 3032-2553, Hal. 01-15
DOI: https://doi.org/10.62383/sosial.v3i4.1228
Tersedia: https://journal.appisi.or.id/index.php/sosial

Implementasi Algoritma *K-Means Clustering* dengan Python untuk Analisis Produksi Bawang Merah di Indonesia

Jafar Pahrudin[1*], Sri Mulyeni[2]
[1]Fakultas Ilmu Komputer, Universitas Nasional Pasim, Indonesia
[2]Fakultas Ekonomi, Universitas Nasional Pasim, Indonesia
**Penulis Korespondensi: jafar.pahrudin@gmail.com*

Abstract. *Shallots are one of the most strategic horticultural commodities in Indonesia, with high demand and varying production levels across regions. Differences in productivity between areas often create challenges in managing distribution and formulating national food policies. This study aims to analyze shallot production data in Indonesia by applying the K-Means Clustering algorithm using Python. The production data were collected from official agricultural statistics publications, followed by preprocessing, normalization, and determination of the optimal number of clusters using the Elbow method and Silhouette Score. The clustering results show the formation of several groups representing regions with high, medium, and low production levels. Visualization of the clustering results reveals the distribution patterns of shallot production, which can serve as a basis for supporting policy formulation in the development of shallot production centers in Indonesia. Thus, the application of K-Means Clustering with Python proves to be an effective approach to provide clearer insights into regional production variations and can be utilized as an analytical tool to support decision-making in the agricultural sector.*

Keywords: *Shallot Production; K-Means Clustering; Data Mining; Python; Agricultural Analysis*

Abstrak. Bawang merah merupakan salah satu komoditas hortikultura strategis di Indonesia dengan tingkat permintaan yang tinggi dan produksi yang beragam di setiap daerah. Perbedaan produktivitas antarwilayah sering menimbulkan tantangan dalam pengelolaan distribusi dan perencanaan kebijakan pangan nasional. Penelitian ini bertujuan untuk menganalisis data produksi bawang merah di Indonesia melalui penerapan algoritma K-Means Clustering dengan menggunakan Python. Data produksi dikumpulkan dari publikasi resmi statistik pertanian, kemudian dilakukan tahap praproses, normalisasi, serta penentuan jumlah klaster optimal dengan metode *Elbow* dan *Silhouette Score*. Hasil pengelompokan menunjukkan terbentuknya beberapa klaster yang merepresentasikan wilayah dengan tingkat produksi tinggi, sedang, dan rendah. Visualisasi hasil clustering memperlihatkan pola distribusi produksi yang dapat menjadi dasar dalam mendukung perumusan kebijakan pengembangan sentra bawang merah di Indonesia. Dengan demikian, penerapan K-Means Clustering berbasis Python terbukti mampu memberikan gambaran yang lebih jelas terkait variasi produksi bawang merah antarwilayah, serta dapat digunakan sebagai pendekatan analitik dalam mendukung pengambilan keputusan di sektor pertanian.

Kata Kunci: Produksi Bawang Merah; *K-Means Clustering*; Data Mining; Python; Analisis Pertanian

1. PENDAHULUAN

Indonesia memiliki karakteristik sebagai negara agraris karena sebagian besar wilayahnya berupa lahan pertanian. Sektor pertanian menjadi salah satu penyumbang utama pendapatan sumber daya nasional. Di dalamnya, subsektor perkebunan memegang peran penting dengan kontribusi yang cukup besar terhadap pembangunan ekonomi Indonesia (Bode, 2017). Subsektor hortikultura menunjukkan tren peningkatan kontribusi terhadap Produk Domestik Bruto (PDB), khususnya melalui produksi komoditas sayuran. Sayuran termasuk komoditas bernilai ekonomi tinggi yang berperan penting dalam mendukung kebutuhan pangan rumah tangga petani. Kondisi ini terlihat dari beberapa karakteristik, seperti masa tanam yang relatif singkat sehingga cepat memberikan hasil, kemudahan dalam pengelolaan dengan teknologi sederhana, serta tingginya permintaan pasar karena sayuran merupakan bagian pokok

Naskah Masuk: 29 September, 2025; Revisi: 30 September, 2025; Diterima: 02 Oktober, 2025; Tersedia: 07 Oktober, 2025

dalam konsumsi harian keluarga (Rahmadona et al., 2015). Komoditas hortikultura di Indonesia memiliki prospek yang cerah apabila dikelola secara optimal dengan dukungan kebijakan yang mampu menciptakan iklim usaha yang kondusif, baik pada level makro maupun mikro. Selain bernilai ekonomi tinggi, subsektor hortikultura juga berpotensi besar dalam meningkatkan kesejahteraan petani sekaligus menjadi salah satu sumber devisa bagi negara (Elfianto et al., 2020).

Bawang merah yang memiliki nama ilmiah *Allium cepa L. var. aggregatum* merupakan salah satu komoditas hortikultura utama di Indonesia dengan nilai ekonomi yang tinggi serta peran penting dalam memenuhi kebutuhan pangan dan menjadi sumber penghasilan bagi petani. Tingkat produktivitasnya dipengaruhi oleh beragam faktor, baik faktor internal seperti varietas dan mutu benih, maupun faktor eksternal seperti kondisi iklim, karakteristik tanah, serta teknik budidaya yang diterapkan (Sofyan & Sitorus, 2025). Budidaya bawang merah di Indonesia pada dasarnya dapat dilakukan hampir di seluruh wilayah. Namun, rata-rata produktivitas nasional masih berada pada kisaran 9,69 ton per hektar, yang dinilai relatif rendah apabila dibandingkan dengan potensi hasil optimal. Berdasarkan penelitian dari Balai Penelitian Tanaman Sayuran, bawang merah berpotensi menghasilkan antara 12 hingga 15 ton per hektar dengan penerapan teknik budidaya yang tepat (Pada et al., 2017).

Perkembangan teknologi mendorong pemanfaatan analisis berbasis data secara lebih luas dalam sektor pertanian. Dalam penelitian ini, proses *data mining* diterapkan untuk menganalisis produksi bawang merah di Indonesia. Salah satu teknik yang digunakan adalah *clustering*, yang bertujuan mengelompokkan *data* sehingga pola umum dari distribusi produksi dapat diidentifikasi dengan lebih jelas (Bode, 2017). *Clustering* telah terbukti sebagai metode yang efektif dalam menyelesaikan permasalahan kompleks di bidang ilmu komputer dan statistika. Terdapat berbagai algoritma dalam teknik ini, dan salah satu yang paling banyak digunakan adalah *K-Means*. Pada penelitian ini, algoritma tersebut dipilih karena dinilai relevan untuk mengidentifikasi serta memahami pola produksi bawang merah di Indonesia (Abdullah & Fatah, 2024). *K-Means* merupakan algoritma iteratif yang relatif sederhana dan digunakan untuk membagi himpunan data ke dalam sejumlah klaster sesuai dengan jumlah yang telah ditentukan sebelumnya oleh pengguna. Metode ini telah dikembangkan dan digunakan oleh berbagai peneliti dari beragam disiplin ilmu (Kurniawan, 2019). Kelebihan algoritma *K-Means* terletak pada sifatnya yang mudah dipahami, sederhana untuk diimplementasikan, serta mampu menghasilkan pengelompokan yang efektif. Hal tersebut menjadikan *K-Means* banyak dimanfaatkan dalam berbagai penelitian. Meski demikian, algoritma ini juga memiliki keterbatasan, yaitu sangat bergantung pada penentuan titik pusat

klaster awal yang dapat memengaruhi hasil akhir proses pengelompokan (Sekar Setyaningtyas et al., 2022).

Penentuan jumlah klaster yang optimal dalam proses *clustering* dapat dilakukan dengan beberapa pendekatan agar hasil pengelompokan menjadi lebih representatif. Beberapa metode yang umum digunakan antara lain *Elbow Method* dan *Silhouette Score*. Kedua metode ini merupakan teknik evaluasi yang paling sering diaplikasikan dalam *data mining* untuk menentukan jumlah klaster terbaik. Baik *Elbow* maupun *Silhouette* memiliki karakteristik masing-masing, sehingga diperlukan ketelitian dalam mengombinasikannya dengan algoritma *clustering* yang digunakan (Vania & Nurina Sari, 2023).

Penelitian ini bertujuan untuk mengidentifikasi pusat-pusat produksi bawang merah di Indonesia dengan cara menganalisa data produksi bawang merah dari tahun 2020 – 2024 di Indonesia dengan menggunakan metode data mining. Untuk data yang akan di analisa, peneliti akan mengambil sumber data yang di sediakan oleh Badan Pusat Statistik (*https://www.bps.go.id/*). Selain metode dan sumber data peneliti akan menggunakan Python untuk melakukan proses analisa, dimulai dari pengambilan data, pemrosesan data, hingga ke penyajian laporan atau hasil pemrosesan data. Evolusi *Python* sebagai bahasa pemrograman untuk analisis data dapat ditelusuri kembali ke awal mulanya awal 1990an oleh Guido van Rossum. Awalnya dikembangkan untuk menjawab kebutuhan Bahasa tingkat tinggi yang mudah dipahami bahasa *scripting*, *Python* dengan cepatnya menjadi salah satu Bahasa *pemrograman* populer karena simpel dan mudah dalam pembacaan *code*-nya (Kabir et al., 2024).

2. KAJIAN TEORI

Data Mining

Data mining memiliki tujuan utama untuk menemukan korelasi serta pola yang baru, valid, bermanfaat, dan mudah dipahami dari kumpulan data yang tersedia. Proses penemuan pola dalam data ini dikenal dengan berbagai istilah di lingkungan akademik, seperti *knowledge extraction*, *information harvesting*, *information discovery*, data *archeology*, maupun *data pattern processing*. Istilah "*data mining*" sendiri lebih umum digunakan oleh kalangan statistikawan, peneliti basis data, serta komunitas Sistem Informasi Manajemen (SIM) dan bisnis (Jackson, 2002).

Data mining merupakan salah satu bagian dari aktivitas analitik data. Dalam beberapa literatur, konsep ini juga dikenal dengan istilah *Knowledge Discovery in Databases* (KDD) (Kurniawan, 2019). Secara umum, istilah *Knowledge Discovery in Databases* (KDD) merujuk

pada keseluruhan rangkaian proses untuk memperoleh pengetahuan dari data, di mana *data mining* merupakan salah satu tahap inti di dalamnya. Proses KDD mencakup beberapa langkah tambahan, seperti persiapan, seleksi, pembersihan data, hingga interpretasi hasil, sehingga informasi yang dihasilkan dapat memberikan manfaat yang optimal (Jackson, 2002).

Data mining, yang juga dikenal dengan istilah *Knowledge Discovery in Databases* (KDD), merupakan proses yang mencakup pengumpulan serta pemanfaatan data historis untuk mengidentifikasi keteraturan dan pola hubungan pada himpunan *data* berukuran besar. Hasil dari proses ini dapat dimanfaatkan sebagai dasar dalam pengambilan keputusan di masa mendatang. Salah satu teknik yang banyak digunakan dalam *data mining* adalah *clustering,* yang akan kita gunakan untuk menganalisis data produksi bawang merah di Indonesia (Handoko et al., 2020).

Data mining dapat dipahami sebagai pengembangan dari metode analisis data tradisional dan pendekatan statistika, dengan mengintegrasikan berbagai teknik dari disiplin ilmu lain, seperti analisis numerik, *pattern matching*, serta kecerdasan buatan yang mencakup *machine learning*, *neural networks*, hingga *genetic algorithms*. Tidak seperti analisis tradisional yang berorientasi pada hipotesis, *data mining* lebih menekankan pendekatan berbasis *data* (*data-driven*), di mana algoritma secara adaptif mencari pola, tren, dan hubungan yang tersembunyi dalam kumpulan data (Jackson, 2002).

Clustering

Dalam *data mining*, *clustering* merupakan salah satu teknik yang digunakan untuk menganalisis dan membagi data ke dalam sejumlah kelompok. Tujuan utamanya adalah mempartisi data sehingga objek dengan karakteristik yang serupa ditempatkan dalam klaster yang sama, sementara objek yang berbeda dikelompokkan ke klaster lain. Metode ini telah banyak diaplikasikan di berbagai bidang, seperti penelitian ilmiah, identifikasi objek, pengolahan citra, hingga pemetaan wilayah (Sekar Setyaningtyas et al., 2022).

Menurut (Vania & Nurina Sari, 2023), *clustering* adalah proses mengelompokkan data berdasarkan keterkaitannya dengan memperhatikan kesamaan antar *data* dalam satu kelompok serta perbedaannya dengan kelompok lain. Sementara itu, (Handoko et al., 2020) mendefinisikan *clustering* sebagai metode pengelompokan data atau objek ke dalam beberapa klaster, di mana setiap klaster berisi *data* yang memiliki kemiripan tinggi, dan berbeda dengan *data* pada klaster lainnya.

Algoritma *K-Means Clustering*

K-Means merupakan salah satu algoritma *clustering* dengan pendekatan partisi yang membagi *data* ke dalam sejumlah klaster berdasarkan kedekatan terhadap pusat titik (*centroid*).

Metode ini berupaya mengelompokkan data sehingga objek dengan karakteristik yang serupa ditempatkan pada klaster yang sama, sementara objek dengan karakteristik yang berbeda dimasukkan ke klaster yang lain (Meliza & Susanti, n.d.). Proses *K-Means* dilakukan secara bertahap dan berulang (*iteratif*) melalui beberapa langkah utama, yaitu: (1) menentukan jumlah klaster (K), (2) menginisialisasi titik pusat klaster (*centroid*), (3) menghitung jarak setiap data terhadap *centroid* dengan ukuran jarak tertentu, umumnya menggunakan *Euclidean Distance*, (4) mengalokasikan data ke klaster terdekat, (5) memperbarui posisi *centroid* berdasarkan nilai rata-rata data dalam klaster, serta (6) mengulangi langkah 3–5 hingga posisi *centroid* tidak lagi berubah secara signifikan atau telah mencapai kondisi stabil.

Kelebihan algoritma *K-Means* terletak pada kemampuannya mengungkap segmentasi dalam suatu *data* tanpa memerlukan kriteria khusus untuk setiap kelompok. Namun, algoritma ini juga memiliki keterbatasan, yaitu memerlukan proses iterasi berulang hingga hasilnya stabil, serta membutuhkan pengaturan parameter tambahan agar dapat menghasilkan pengelompokan yang lebih optimal (Kamila et al., 2021).

Evaluasi Jumlah Klaster

Pemilihan jumlah cluster yang tepat merupakan tahap penting dalam proses *clustering*. Berikut ini adalah beberapa metode yang akan kita gunakan untuk menentukan jumlah *cluster* optimal adalah *Metode elbow* dan *Metode Silhouette.*

Metode *Elbow* merupakan salah satu teknik yang umum digunakan untuk menentukan jumlah klaster dalam proses *clustering*. Tujuannya adalah menemukan jumlah klaster yang paling optimal agar hasil pengelompokan lebih representatif. Prinsip kerjanya dilakukan dengan menghitung tingkat kohesi dan pemisahan antar klaster. Kohesi menunjukkan seberapa kuat keterkaitan data dalam satu klaster, sementara pemisahan menggambarkan sejauh mana perbedaan antar klaster. Kedua ukuran ini biasanya dievaluasi menggunakan nilai Sum of Squares Error (SSE) (Vania & Nurina Sari, 2023)

$$SSE = \sum_{k=1}^{k} \sum_{x_i} |x_i - C_k|^2$$

Keterangan:

K = *Custer* ke-C

x_i = Jarak data pada obyek ke-i

C_k = Pusat *Cluster* ke-i

Metode Silhouette mengevaluasi kualitas klaster dengan memanfaatkan koefisien siluet yang mengombinasikan ukuran pemisahan antar klaster dan kohesi dalam klaster. Nilai *koefisien* yang semakin tinggi menunjukkan bahwa klaster yang terbentuk semakin baik kualitasnya.

$SC = maks_k\ SI\ (k)$

Keterangan:

SC = *Silhouette Coefficient*

SI = *Silhouette Index Global*

k = Jumlah *cluster*

Python Dalam Analisis Data Mining

Python merupakan bahasa pemrograman yang paling luas digunakan saat ini. Dalam konteks penyelesaian tugas dan tantangan pada bidang *data science*, *Python* senantiasa memberikan kemudahan dan fleksibilitas bagi penggunanya. Sebagian besar ilmuwan data (*data scientists*) telah memanfaatkan kemampuan *Python* dalam aktivitas sehari-hari. *Python* dikenal sebagai bahasa pemrograman yang mudah dipelajari, mudah ditelusuri kesalahannya (*debugging*), bersifat *object-oriented*, bersumber terbuka (*open-source*), memiliki kinerja tinggi, serta menawarkan berbagai keuntungan lainnya. Selain itu, Python dilengkapi dengan pustaka-pustaka unggulan yang digunakan secara luas dalam pemecahan permasalahan komputasi (Saabith et al., 2021). Berikut ini daftar pustaka yang akan peneliti gunakan dalam penelitian ini:

NumPy (*Numerical Python*) merupakan pustaka fundamental dalam komputasi numerik menggunakan *Python*, dengan dukungan sekitar 18.000 komentar dan 700 kontributor aktif. *NumPy* menyediakan objek array N-dimensi berkinerja tinggi dan seperangkat alat untuk bekerja dengan array tersebut. Pustaka ini menjadi solusi atas keterbatasan kecepatan komputasi Python dengan menyediakan fungsi serta operator yang dioptimalkan untuk operasi pada array multidimensi. Fitur utama *NumPy* mencakup: fungsi numerik yang telah dikompilasi sebelumnya dengan kecepatan tinggi, komputasi berbasis array yang efisien, dukungan terhadap paradigma *object-oriented*, serta perhitungan yang ringkas dan cepat dengan teknik *vectorization*. *NumPy* sangat berguna dalam analisis data, pembuatan array N-dimensi, menjadi dasar dari pustaka lain seperti *SciPy* dan *Scikit-learn*, serta berfungsi sebagai pengganti *MATLAB* ketika dipadukan dengan *SciPy* dan *Matplotlib* (Saabith et al., 2021).

Pandas (*Python Data Analysis*) merupakan pustaka penting dalam *life cycle data science*. Bersama dengan *NumPy* dan *Matplotlib*, *Pandas* menjadi salah satu pustaka yang paling populer untuk analisis dan pembersihan data, dengan sekitar 17.000 komentar dan 1.200 kontributor aktif. *Pandas* menyediakan struktur data yang cepat dan fleksibel, seperti *data*

frame, yang dirancang untuk bekerja dengan data terstruktur secara intuitif. Fitur utama *Pandas* mencakup sintaksis yang ekspresif dengan fungsi yang kaya, dukungan terhadap penanganan data yang hilang, pembuatan serta penerapan fungsi pada rangkaian data, abstraksi tingkat tinggi, serta alat manipulasi data yang komprehensif. *Pandas* banyak digunakan dalam pembersihan dan pengolahan data, pekerjaan *ETL* (*extract*, *transform*, *load*), serta mendukung pemuatan data dalam *format CSV* ke dalam *data frame*. *Pandas* juga dimanfaatkan dalam berbagai bidang akademik dan komersial, termasuk statistik, keuangan, dan ilmu saraf, dengan dukungan fungsi khusus untuk analisis deret waktu, regresi linear, hingga pergeseran tanggal (Saabith et al., 2021).

Matplotlib merupakan pustaka visualisasi data dengan kemampuan grafis yang kuat dan representasi visual yang menarik. Pustaka ini memiliki sekitar 26.000 komentar dan 700 kontributor aktif di *GitHub*. *Matplotlib* digunakan secara luas untuk visualisasi data karena kemampuannya menghasilkan grafik dan plot yang informatif. Selain itu, *Matplotlib* menyediakan *API object-oriented* yang memungkinkan integrasi plot ke dalam aplikasi. Fitur utama *Matplotlib* mencakup: kompatibilitas sebagai pengganti *MATLAB* yang gratis dan *open-source*, dukungan berbagai *backend* dan format keluaran, konsumsi memori rendah, serta kinerja waktu eksekusi yang baik. *Matplotlib* sangat berguna untuk analisis korelasi variabel, visualisasi interval kepercayaan 95% pada model, deteksi pencilan menggunakan scatter plot, serta eksplorasi distribusi data untuk memperoleh wawasan instan (Saabith et al., 2021).

Scikit-learn merupakan salah satu pustaka *open-source* berbasis *Python* yang banyak digunakan dalam penelitian dan praktik *machine learning* karena menyediakan berbagai algoritma dan metode analisis data dengan antarmuka pemrograman yang konsisten serta mudah diimplementasikan. Pustaka ini mendukung beragam teknik pembelajaran terawasi (*supervised learning*) seperti regresi dan klasifikasi, pembelajaran tanpa pengawasan (*unsupervised learning*) seperti *clustering* dan reduksi dimensi, serta fasilitas praproses data melalui normalisasi, standarisasi, dan *encoding* variabel kategorikal. Selain itu, *Scikit-learn* juga dilengkapi dengan modul evaluasi model, model selection, serta pipeline yang memungkinkan integrasi alur kerja secara sistematis, sehingga mempermudah peneliti maupun praktisi dalam melakukan eksperimen hingga penerapan model dalam skala kecil hingga menengah.

Streamlit merupakan pustaka *Python* sumber terbuka (*open-source*) yang memungkinkan pengembang untuk membuat aplikasi web yang interaktif dan menarik secara cepat dengan kebutuhan penulisan kode yang minimal. Pustaka ini dirancang khusus untuk membantu ilmuwan data (*data scientists*) dan insinyur pembelajaran mesin (*machine learning*

engineers) dalam mengubah skrip data menjadi aplikasi web yang dapat dibagikan dengan mudah. Daya tarik utama *Streamlit* terletak pada kesederhanaan dan efisiensinya dalam membangun aplikasi berbasis data. Pengembangan web tradisional, khususnya untuk kebutuhan visualisasi data dan penerapan pembelajaran mesin, biasanya memerlukan pengetahuan tentang kerangka kerja *web*, *HTML*, *CSS*, serta *JavaScript*, yang sering kali dirasa rumit bagi mereka yang lebih berfokus pada analisis data. *Streamlit* menyederhanakan kompleksitas tersebut dengan menyediakan antarmuka yang memungkinkan pengguna membangun aplikasi hanya melalui skrip *Python* sederhana (Akkem et al., 2023).

3. METODE PENELITIAN

Dalam penelitian ini, Teknik yang digunakan adalah metode data mining dengan tahapan sebagai berikut: (a) tahap pengumpulan data, (b) tahap pengolahan data, (c) tahap *clustering*, dan (d) tahap analisis. Rangkaian tahapan metode tersebut dijelaskan lebih lanjut sebagai berikut.

Tahap Pengumpulan Data

Pada penerapan data mining terhadap data produksi bawang merah di Indonesia, diperlukan data yang relevan. Sumber data penelitian diperoleh dari publikasi resmi Badan Pusat Statistik (https://www.bps.go.id). Data yang digunakan adalah data produksi bawang merah nasional pada periode 2022–2024 yang mencakup berbagai daerah sentra produksi. Variabel yang digunakan meliputi: (1) jumlah produksi bawang merah (ton), dan (2) luas panen (hektar).

Tahap Pengolahan Data

Data tersebut kemudian diolah dengan metode clustering untuk mengelompokkan daerah produksi bawang merah ke dalam tiga kelompok, yaitu klaster produksi tinggi, klaster produksi sedang, dan klaster produksi rendah. Sebelum dilakukan clustering, data setiap daerah diproses melalui tahap prapengolahan, meliputi pembersihan data, normalisasi, serta penggabungan variabel agar diperoleh nilai yang siap digunakan pada tahap pengelompokan.

Tahap *Clustering*

Clustering merupakan bentuk klasifikasi tanpa pengawasan (*unsupervised classification*) yang mempartisi sekumpulan objek data ke dalam beberapa kelas atau kelompok. Proses ini dilakukan dengan menerapkan ukuran jarak tertentu, dalam hal ini digunakan *Euclidean Distance*. Analisis klaster bertujuan untuk membagi data produksi bawang merah ke dalam kelompok-kelompok berdasarkan kesamaan karakteristik produksi.

Tahap Analisis

Dalam penentuan jumlah klaster optimal, digunakan metode *Elbow* dan *Silhouette Coefficient* untuk memastikan hasil pengelompokan yang representatif. Proses analisis dilakukan menggunakan bahasa pemrograman *Python*, dengan memanfaatkan pustaka seperti *NumPy*, *Pandas*, *Matplotlib*, dan *Scikit-learn*. Data yang telah dikelompokkan kemudian divisualisasikan untuk memperlihatkan pola distribusi produksi bawang merah di berbagai daerah. Pada tahap akhir, hasil klaster dianalisis untuk mengidentifikasi wilayah dengan tingkat produksi tinggi, sedang, dan rendah, sehingga dapat menjadi dasar dalam perumusan kebijakan pengembangan sentra bawang merah di Indonesia.

4. HASIL DAN PEMBAHASAN

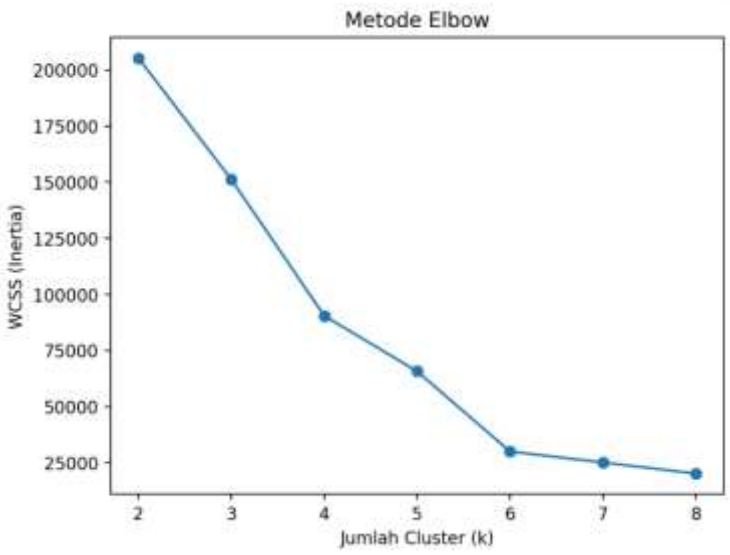
Tahapan pertama adalah pengambilan data dari sumber data, di sini peneliti memilih untuk mengambil data ke sumbernya langsung dengan menggunakan layanan *WEB API* yang disediakan oleh BPS. Data yang diambil adalah informasi produksi bawang merah dan luas lahan produksi selama tiga tahun yaitu dari tahun 2022 – 2024.

Data Produksi Bawang Merah (BPS)

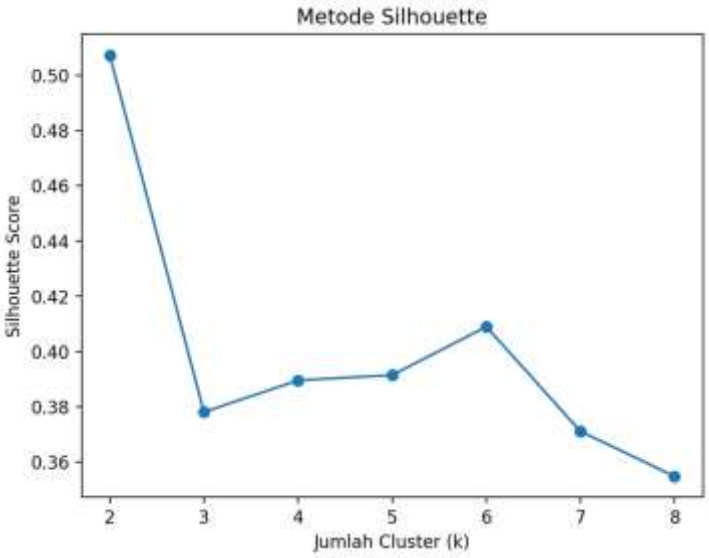
	ID	Provinsi	produksi_2022_Kw	wilayah_2022_Ha	prod_lahan_2022	produksi_2023_Kw	wilayah_2023_Ha	prod_lahan_2023	produksi_2024_Kw	wilayah_2024_Ha	prod_lahan_2024
0	1100000	Aceh	100703	1236	81.4749	136728	1474	92.7585	131843.96	1418.03	92.9664
1	1200000	Sumatera Utara	646049	4249	152.0486	655852	4296	152.6657	580699.86	3861.34	150.3882
2	1300000	Sumatera Barat	2073758	14033	147.7772	2339173	15427	151.6285	2307184.11	14843.15	155.4376
3	1400000	Riau	1953	34	57.4412	3227	74	43.6081	2456.15	64.4	55.2962
4	1500000	Jambi	160502	2125	75.5304	184013	2128	86.4723	122878.65	1554.9	79.0254
5	1600000	Sumatera Selatan	11299	180	62.7722	11978	165	72.5455	6965.09	118.32	75.7694
6	1700000	Bengkulu	10233	94	108.8617	6708	97	69.134	7222.15	152.8	47.2654
7	1800000	Lampung	17267	235	73.4766	21930	300	73.1	16487.2	223.39	73.8046
8	1900000	Kepulauan Bangka Belitung	787	16	49.1875	1757	32	54.9063	1186.48	26.7	44.4082
9	2100000	Kepulauan Riau	414	7	59.1429	170	6	28.3333	1022.9	8.33	122.7971
10	3100000	DKI Jakarta	14	1	14	1	0	0	128.27	3.27	39.2321
11	3200000	Jawa Barat	1955183	17990	111.1864	1793564	18010	112.0271	1998916.88	17917.88	111.5399
12	3300000	Jawa Tengah	5565068	53593	103.84	4760908	46795	102.3808	6078567.21	52240.97	116.3646

Gambar 1. Hasil pengambilan data produksi dan luas lahan produksi bawang merah yang di abmil dari Badan. Pusat Statistik (*https://www.bps.go.id/*).

Sebelum memulai proses *clustering*, peneliti menentukan terlebih dahulu jumlah *cluster* dengan menggunakan metode *Elbow* dan *Silhouette Score.* Dari hasil methode ini didapatkan rekomendasi jumlah *cluster* adalah 2.

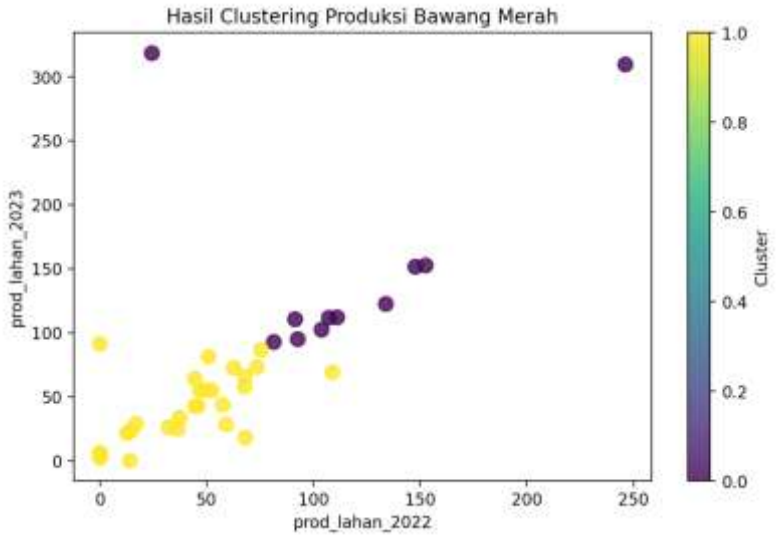


Gambar 2. Grafik penentuan cluster dengan *Metode Elbow*.



Gambar 3. Grafik penentuan cluster dengan *Metode Silhouette.*

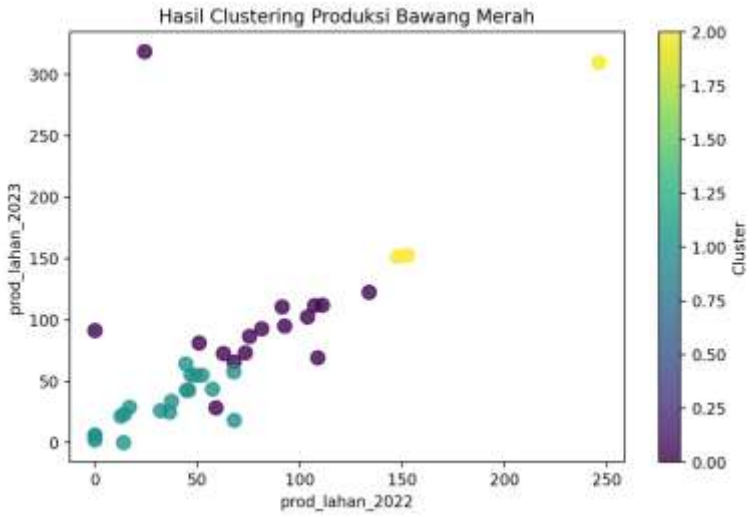
Untuk *clustering* peneliti akan membandingkan hasil dari 2 *cluster* dan 3 *cluster*, berikut ini hasil dari 2 *cluster*.



Gambar 4. Visualisasi *clustering* dengan 2 *cluster*.

Tabel 1. hasil *clustering* dengan 2 *cluster*.

Cluster	*ID*	*Provinsi*	*Prod 2022 (Kw/Ha)*	*Prod 2023 (Kw/Ha)*	*Prod 2024 (Kw/Ha)*
0	1100000	Aceh	81.47	92.76	92.97
0	1200000	Sumatera Utara	152.59	152.67	150.39
0	1300000	Sumatera Barat	147.78	151.63	155.44
0	3200000	Jawa Barat	111.17	112.03	111.56
0	3300000	Jawa Tengah	103.84	102.38	116.36
0	3400000	DI Yogyakarta	91.46	110.58	114.1
0	3500000	Jawa Timur	92.7	95	88.8
0	5100000	Bali	246.23	309.75	285.82
0	5200000	Nusa Tenggara Barat	107.33	111.72	105.46
0	6500000	Kalimantan Utara	24.33	318.64	None
0	7300000	Sulawesi Selatan	133.97	122.51	123.95
1	1400000	Riau	57.44	43.61	55.3
1	1500000	Jambi	75.53	86.47	79.03
1	1600000	Sumatera Selatan	62.77	72.55	75.77
1	1700000	Bengkulu	108.86	69.13	47.27
1	1800000	Lampung	73.48	73.1	73.8
1	1900000	Kepulauan Bangka Belitung	49.19	54.91	44.89
1	2100000	Kepulauan Riau	59.14	28.33	122.8
1	3100000	DKI Jakarta	14	0	39.53
1	3600000	Banten	67.94	65.65	52.35
1	5300000	Nusa Tenggara Timur	50.9	81.15	58.39
1	6100000	Kalimantan Barat	16.92	28.82	17.85
1	6200000	Kalimantan Tengah	44.72	42.53	24.31
1	6300000	Kalimantan Selatan	52.34	54.94	56.85
1	6400000	Kalimantan Timur	45.8	42.75	61.72
1	7100000	Sulawesi Utara	67.83	57.64	41.49
1	7200000	Sulawesi Tengah	46.87	55.17	61.36
1	7400000	Sulawesi Tenggara	36.53	24.78	33.72
1	7500000	Gorontalo	44.43	64.16	66.08
1	7600000	Sulawesi Barat	12.68	21.51	17.9
1	8100000	Maluku	37.4	33.71	21.99
1	8200000	Maluku Utara	32.06	25.98	26.39
1	9100000	Papua Barat	14.37	23.27	12.47
1	9200000	Papua Barat Daya	None	5.71	3.84
1	9400000	Papua	68.1	18.14	43.01
1	9500000	Papua Selatan	None	91.31	94.46
1	9600000	Papua Tengah	None	2	5.25
1	9700000	Papua Pegunungan	None	6.17	None



Gambar 5. Visualisasi *clustering* dengan 3 *cluster*.

Tabel 2. hasil *clustering* dengan 3 *cluster*.

Cluster	*ID*	*Provinsi*	*Prod 2022 (Kw/Ha)*	*Prod 2023 (Kw/Ha)*	*Prod 2024 (Kw/Ha)*
0	1100000	Aceh	81.47	92.76	92.97
0	1500000	Jambi	75.53	86.47	79.03
0	1600000	Sumatera Selatan	62.77	72.55	75.77
0	1700000	Bengkulu	108.86	69.13	47.27
0	1800000	Lampung	73.48	73.1	73.8
0	2100000	Kepulauan Riau	59.14	28.33	122.8
0	3200000	Jawa Barat	111.17	112.03	111.56
0	3300000	Jawa Tengah	103.84	102.38	116.36
0	3400000	DI Yogyakarta	91.46	110.58	114.1
0	3500000	Jawa Timur	92.7	95	88.8
0	3600000	Banten	67.94	65.65	52.35
0	5200000	Nusa Tenggara Barat	107.33	111.72	105.46
0	5300000	Nusa Tenggara Timur	50.9	81.15	58.39
0	6500000	Kalimantan Utara	24.33	318.64	None
0	7300000	Sulawesi Selatan	133.97	122.51	123.95
0	9500000	Papua Selatan	None	91.31	94.46
1	1400000	Riau	57.44	43.61	55.3
1	1900000	Kepulauan Bangka Belitung	49.19	54.91	44.89
1	3100000	DKI Jakarta	14	0	39.53
1	6100000	Kalimantan Barat	16.92	28.82	17.85
1	6200000	Kalimantan Tengah	44.72	42.53	24.31
1	6300000	Kalimantan Selatan	52.34	54.94	56.85
1	6400000	Kalimantan Timur	45.8	42.75	61.72
1	7100000	Sulawesi Utara	67.83	57.64	41.49
1	7200000	Sulawesi Tengah	46.87	55.17	61.36
1	7400000	Sulawesi Tenggara	36.53	24.78	33.72
1	7500000	Gorontalo	44.43	64.16	66.08
1	7600000	Sulawesi Barat	12.68	21.51	17.9
1	8100000	Maluku	37.4	33.71	21.99
1	8200000	Maluku Utara	32.06	25.98	26.39
1	9100000	Papua Barat	14.37	23.27	12.47
1	9200000	Papua Barat Daya	None	5.71	3.84
1	9400000	Papua	68.1	18.14	43.01
1	9600000	Papua Tengah	None	2	5.25
1	9700000	Papua Pegunungan	None	6.17	None
2	1200000	Sumatera Utara	152.59	152.67	150.39
2	1300000	Sumatera Barat	147.78	151.63	155.44
2	5100000	Bali	246.23	309.75	285.82

Hasil analisis *K-Means Clustering* menunjukkan bahwa penggunaan dua klaster menghasilkan pemisahan data yang relatif sederhana, yakni membagi provinsi ke dalam kelompok produksi tinggi dan kelompok produksi rendah. Model dengan dua klaster ini lebih mudah dipahami, karena secara garis besar hanya menekankan pada perbedaan signifikan antar wilayah dalam hal produktivitas bawang merah. Dengan demikian, pola distribusi produksi dapat terlihat secara umum, di mana beberapa provinsi konsisten masuk ke kelompok produksi tinggi sementara sebagian besar lainnya berada pada kelompok produksi rendah.

Namun, ketika jumlah klaster ditingkatkan menjadi tiga, hasil analisis memberikan variasi yang lebih detail. Pada model tiga klaster, terdapat kelompok tambahan yang dapat merepresentasikan provinsi dengan tingkat produksi sedang. Hal ini bermanfaat untuk memberikan pemahaman yang lebih komprehensif, terutama bagi pengambil kebijakan yang membutuhkan informasi tidak hanya tentang wilayah yang sangat produktif atau kurang produktif, tetapi juga wilayah dengan potensi berkembang. Dengan adanya klaster sedang, pemerintah atau pemangku kepentingan dapat merumuskan strategi intervensi yang lebih spesifik, seperti peningkatan teknologi pertanian atau pemberian dukungan infrastruktur untuk provinsi yang berada di kelompok tersebut.

Perbedaan utama antara penggunaan dua klaster dan tiga klaster terletak pada tingkat detailnya informasi yang disajikan. Dua klaster lebih efektif untuk analisis makro karena sederhana dan mudah diinterpretasikan, sedangkan tiga klaster lebih bermanfaat untuk analisis mikro karena memungkinkan identifikasi wilayah transisi yang berpotensi ditingkatkan produksinya. Dengan kata lain, pemilihan jumlah klaster perlu disesuaikan dengan tujuan analisis: apabila tujuan utamanya adalah segmentasi umum, maka dua klaster sudah memadai; tetapi apabila tujuannya adalah mendukung perencanaan pengembangan wilayah secara bertahap, maka tiga klaster lebih memberikan nilai tambah.

5. KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan algoritma K-Means Clustering dengan Python mampu mengelompokkan daerah produksi bawang merah di Indonesia ke dalam klaster produksi tinggi, sedang, dan rendah. Berdasarkan evaluasi menggunakan metode Elbow dan Silhouette Score, jumlah klaster optimal diperoleh sebanyak dua klaster, meskipun pengujian dengan tiga klaster juga memberikan gambaran variasi yang lebih detail. Hasil clustering memperlihatkan adanya perbedaan signifikan antarwilayah dalam hal produktivitas bawang merah, di mana beberapa provinsi konsisten menempati kelompok produksi tinggi, sementara provinsi lainnya cenderung berada pada kelompok produksi rendah. Temuan ini dapat

dijadikan dasar dalam mendukung pengambilan keputusan strategis, khususnya dalam perumusan kebijakan pengembangan sentra produksi dan distribusi bawang merah di Indonesia. Dengan demikian, pendekatan data mining berbasis K-Means terbukti efektif sebagai alat analisis untuk memahami pola distribusi produksi komoditas hortikultura.

SARAN

Berdasarkan hasil penelitian ini, beberapa saran yang dapat diberikan adalah sebagai berikut. Pertama, penelitian selanjutnya disarankan untuk memperluas cakupan data hingga tingkat kabupaten/kota agar analisis klaster dapat memberikan gambaran yang lebih detail dan aplikatif bagi pengambil kebijakan daerah. Kedua, integrasi variabel tambahan seperti faktor iklim, curah hujan, serta ketersediaan sarana produksi dapat meningkatkan akurasi model dalam menjelaskan variasi produktivitas bawang merah. Ketiga, penggunaan metode clustering lain seperti DBSCAN atau Hierarchical Clustering dapat dibandingkan dengan K-Means untuk memperoleh sudut pandang yang lebih komprehensif. Terakhir, penerapan sistem visualisasi interaktif berbasis web dapat mempermudah pemangku kepentingan dalam memahami hasil analisis dan mendukung pengambilan keputusan strategis terkait pengelolaan produksi bawang merah di Indonesia.

DAFTAR PUSTAKA

Abdullah, S. H., & Fatah, Z. (2024). Analisis produksi cabai rawit Indonesia menggunakan algoritma K-means clustering. Jurnal Ilmiah Sains Teknologi dan Informasi, 3(1), 66–74. https://doi.org/10.59024/jiti.v3i1.1024

Akkem, Y., Kumar, B. S., & Varanasi, A. (2023). Streamlit application for advanced ensemble learning methods in crop recommendation systems: A review and implementation. Indian Journal of Science and Technology, 16(48), 4688–4702. https://doi.org/10.17485/IJST/v16i48.2850

Bode, A. (2017). K-nearest neighbor dengan feature selection menggunakan backward elimination untuk prediksi harga komoditi kopi Arabika. ILKOM Jurnal Ilmiah, 9(2), 188–195. https://doi.org/10.33096/ilkom.v9i2.139.188-195

Elfianto, Syahni, R., Asmawi, & Ifdal. (2020). Perilaku petani bawang merah dalam penggunaan pestisida: Sebuah literature review. Jurnal Ilmiah Pertanian, 11(2). https://doi.org/10.31317

Handoko, S., Fauziah, F., & Handayani, E. T. E. (2020). Implementasi data mining untuk menentukan tingkat penjualan paket data Telkomsel menggunakan metode K-means clustering. Jurnal Ilmiah Teknologi dan Rekayasa, 25(1), 76–88. https://doi.org/10.35760/tr.2020.v25i1.2677

Jackson, J. (2002). Data mining: A conceptual overview. Communications of the Association for Information Systems, 8, 1–20. https://doi.org/10.17705/1CAIS.00819

Kabir, M. A., Ahmed, F., Islam, M. M., & Ahmed, M. R. (2024). Python for data analytics: A systematic literature review of tools, techniques, and applications. Academic Journal on Science, Technology, Engineering & Mathematics Education, 4(4), 134–154. https://doi.org/10.69593/ajsteme.v4i04.146

Kamila, C., Adiyatma, M., Namang, G. R., Ramadhan, R., Syah, F., & Redaksi, D. (2021). Pendidikan teknik informatika dan komputer, fakultas teknik. Jurnal Intech, 2(1), 1.

Kurniawan, C. (2019). A survey on big data analytics model. ITEJ (Information Technology Engineering Journal), 4(1), 1–13. https://doi.org/10.24235/itej.v4i1.46

Meliza, O., & Susanti, T. (n.d.). Implementasi K-means: Sebuah studi literatur. Jurnal Informatika (JURI). https://doi.org/10.12345/juri

Pada, P., Pengkajian, B., Pertanian, T., Kompleks, S. B., Gubernur, P., Barat, S., Abdul, J., & Pattanaendeng, M. (2017). Perbaikan usaha tani bawang merah dataran rendah dengan perbandingan paket teknologi petani dengan paket teknologi introduksi di Kabupaten Majene. Jurnal Agrotan, 3(1), 56–66.

Rahmadona, L., Fariyanti, A., & Burhanuddin. (2015). Analisis pendapatan usahatani bawang merah di Kabupaten Majalengka (Income analysis of shallot farming in Majalengka Regency). Jurnal Agribisnis, 2(1), 1–10.

Saabith, S., Thangarajah, V., Saabith, A. S., Vinothraj, T., & Fareez, M. (2021). A review on Python libraries and IDEs for data science. International Journal of Research in Engineering and Science (IJRES). www.ijres.org

Setyaningtyas, S., Nugroho, I., & Arif, Z. (2022). Tinjauan pustaka sistematis: Penerapan data mining teknik clustering algoritma K-means. Jurnal Teknoif Teknik Informatika Institut Teknologi Padang, 10(2), 52–61. https://doi.org/10.21063/jtif.2022.V10.2.52-61

Sofyan, S. N., & Sitorus, Z. (2025). Implementasi data mining untuk clustering produktivitas bawang merah menggunakan metode K-means. Jurnal Jatilima, 7(2), 1–9. https://doi.org/10.54209/jatilima.v7i02.1442

Vania, P., & Nurina Sari, B. (2023). Perbandingan metode elbow dan silhouette untuk penentuan jumlah klaster yang optimal pada clustering produksi padi menggunakan algoritma K-means. Jurnal Ilmiah Wahana Pendidikan, 9(21), 547–558. https://doi.org/10.5281/zenodo.10081332