

BAB II

LANDASAN TEORI

Machine learning merupakan cabang kecerdasan buatan yang berfokus pada pengembangan algoritma agar komputer dapat mengenali pola dari data dan mendukung pengambilan keputusan. Dalam penelitian ini, machine learning digunakan dalam bentuk unsupervised learning karena data tidak memiliki label kelas sebelumnya (Moreno et al., 2025; Aldino et al., 2021).

2.1 Teori–Teori dan Konsep K-Means pada Segmentasi Wilayah Pertanian

(Moreno et al., 2025)(Aldino et al., 2021)Machine learning merupakan cabang dari kecerdasan buatan (artificial intelligence) yang berfokus pada kemampuan sistem komputer untuk belajar dari data dan meningkatkan kinerjanya tanpa pemrograman eksplisit. Moreno mendefinisikan machine learning sebagai studi tentang algoritma yang memungkinkan komputer mempelajari pola dari data untuk membuat prediksi atau keputusan. Algoritma ini bekerja dengan menganalisis data dalam jumlah besar untuk menemukan hubungan antar variabel, yang kemudian digunakan dalam membangun model prediktif. Salah satu keunggulan utama machine learning ialah kemampuannya menyesuaikan model secara otomatis ketika data baru tersedia, menjadikannya alat penting dalam pengambilan keputusan berbasis data.

(Mukhtar et al., 2024)Segmentasi data adalah proses mengelompokkan sekumpulan objek ke dalam kelompok yang relatif homogen berdasarkan kemiripan atribut. Dalam konteks penelitian ini, objek berupa kabupaten/kota

dikelompokkan berdasarkan luas panen, produktivitas, dan produksi padi (Rahmanita et al., 2025).

2.1.1 Teori Segmentasi Data dan Wilayah Pertanian

(Rahmanita et al., 2025)(Yashchyshyn et al., 2021)Segmentasi data adalah proses mengelompokkan sekumpulan data menjadi beberapa bagian homogen berdasarkan kemiripan atribut tertentu. Rahmanita et al. menyebutkan bahwa tujuan utama segmentasi adalah menemukan struktur alami yang tersembunyi dalam data. Dalam konteks pertanian, segmentasi wilayah menjadi kunci untuk memahami variasi spasial seperti perbedaan kesuburan tanah, tingkat produktivitas, curah hujan, serta jenis tanaman yang cocok. Melalui analisis segmentasi, para pembuat kebijakan dapat menentukan wilayah prioritas untuk pengembangan, distribusi pupuk, dan alokasi sumber daya lainnya.

Penerapan segmentasi data dalam bidang pertanian telah terbukti meningkatkan efisiensi pengelolaan lahan. (Septianto et al., 2025) menunjukkan bahwa pendekatan berbasis data membantu menentukan zona produksi unggulan dengan memanfaatkan data spasial dan historis hasil panen. Segmentasi wilayah tidak hanya meningkatkan produktivitas, tetapi juga mengurangi ketimpangan pengelolaan sumber daya antar daerah.

2.1.2 Peran *Unsupervised Learning* dalam Analisis Data Pertanian

(Barbosa et al., 2023)(T. Santoso et al., 2023)Pendekatan unsupervised learning berperan penting dalam menganalisis data pertanian karena mampu mengelompokkan data tanpa memerlukan label yang telah ditentukan sebelumnya. Teknik ini digunakan untuk mengeksplorasi pola-pola alami dan

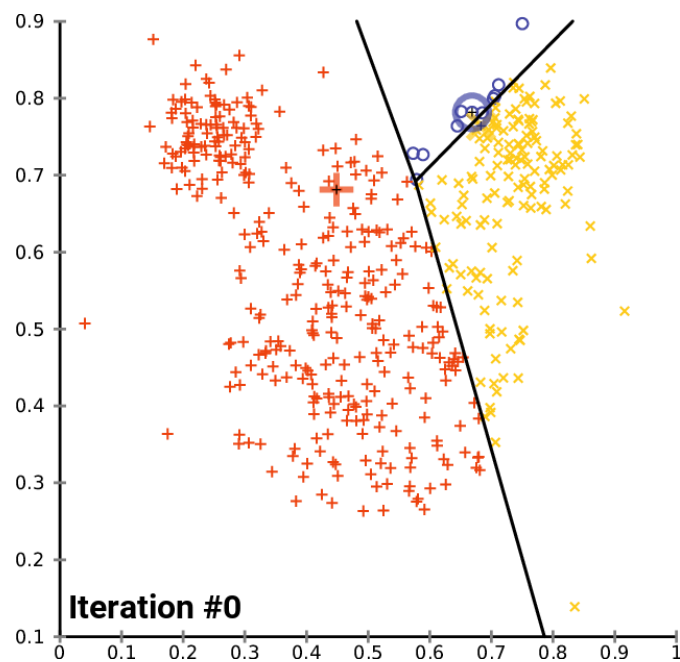
struktur tersembunyi dalam data pertanian. Salah satu metode utama dalam pendekatan ini adalah clustering, yaitu pengelompokan data berdasarkan kesamaan karakteristik antar objek.

Dalam konteks pertanian, *unsupervised learning* membantu menemukan keterkaitan antara kondisi tanah, curah hujan, dan produktivitas. Misalnya, data mengenai kelembapan tanah dan hasil panen dapat dikelompokkan menjadi beberapa klaster untuk mengidentifikasi wilayah dengan potensi pertanian yang mirip. (Barbosa et al., 2023) menunjukkan bahwa penggunaan *unsupervised learning* dalam pertanian mampu mengungkap hubungan kompleks antara faktor iklim dan hasil pertanian, sehingga memudahkan perencanaan pengelolaan lahan secara berkelanjutan.

2.1.3 Penerapan Algoritma *K-Means Clustering* dalam Penelitian Pertanian

Algoritma K-Means clustering merupakan metode unsupervised learning yang membagi data ke dalam sejumlah klaster berdasarkan jarak terdekat terhadap pusat klaster (centroid). Proses dilakukan secara iteratif sampai keanggotaan klaster dan nilai centroid tidak lagi berubah secara berarti.

(Barbosa et al., 2023)(Septianto et al., 2025) Dalam bidang pertanian, K-Means dapat digunakan untuk mengelompokkan wilayah berdasarkan kemiripan indikator produksi atau karakteristik agronominya. Pada penelitian ini, penggunaan K-Means dibatasi pada tiga variabel yang tersedia, yaitu luas panen, produktivitas, dan produksi padi; karena itu, hasilnya diinterpretasikan sebagai kapasitas produksi padi relatif, bukan keseluruhan potensi pertanian.



Gambar 2.1
Visualisasi proses iterasi algoritma *K-Means*

2.1.4 Hubungan Segmentasi Wilayah dengan Pengelolaan Sumber Daya

Pertanian

Segmentasi wilayah pertanian berbasis algoritma *K-Means* menghasilkan peta pengelompokan wilayah berdasarkan kemiripan data produktivitas dan hasil panen. Pemetaan ini menjadi dasar dalam pengambilan keputusan berbasis data, memungkinkan pemerintah dan petani menentukan strategi pengelolaan sumber daya secara efektif. Wilayah dengan kesuburan rendah dapat difokuskan pada rehabilitasi lahan dan peningkatan infrastruktur pertanian, sedangkan wilayah dengan produktivitas tinggi dapat diarahkan untuk pengembangan komoditas bernilai ekonomi tinggi.

(Mukhtar et al., 2024) Pendekatan ini juga berkontribusi terhadap pencapaian tujuan pembangunan berkelanjutan (Sustainable Development Goals),

terutama dalam bidang ketahanan pangan dan pengelolaan sumber daya alam. Dengan demikian, penerapan machine learning, khususnya metode K-Means clustering, memberikan dasar ilmiah bagi kebijakan pertanian yang lebih adaptif, efisien, dan berbasis bukti.

2.2 Algoritma yang Digunakan

Algoritma *K-Means clustering* merupakan salah satu metode *unsupervised learning* yang digunakan untuk mengelompokkan data berdasarkan kesamaan karakteristik. Prinsip utama algoritma ini adalah meminimalkan jarak antar data dengan pusat kelompok (*centroid*) sehingga setiap klaster memiliki tingkat kesamaan internal yang tinggi. Proses pengelompokan dilakukan secara iteratif untuk mencari posisi *centroid* yang optimal dengan menggunakan rumus jarak Euclidean, seperti pada Persamaan (2.1):

$$d(x_i, \mu_j) = \sqrt{\sum_{k=1}^n (x_{ik} - \mu_{jk})^2}$$

Keterangan: (x_i) merupakan data ke- i , (μ_j) adalah pusat klaster ke- j , dan (n) adalah jumlah variabel. Proses ini berulang hingga posisi *centroid* tidak lagi berubah secara signifikan atau jumlah iterasi maksimum tercapai. Prinsip ini pertama kali diperkenalkan oleh dan kemudian dikembangkan lebih lanjut untuk berbagai aplikasi, termasuk bidang pertanian dan analisis spasial.

Algoritma K-Means bekerja melalui lima tahapan utama: menentukan jumlah klaster (k), memilih centroid awal, menghitung jarak setiap objek terhadap seluruh centroid, menempatkan objek ke centroid terdekat, dan menghitung

centroid baru sebagai rata-rata anggota kluster. Proses diulang sampai tidak terjadi perpindahan anggota atau perubahan centroid yang berarti (Rathore et al., 2021).

Rathore, R., et al. (2021) Rathore, R., et al. (2021) Algoritma K-Means bekerja melalui beberapa tahapan utama. Pertama, menentukan jumlah kluster (k) yang akan dibentuk berdasarkan analisis awal atau menggunakan metode Elbow untuk menentukan nilai optimal. Kedua, memilih centroid awal secara acak dari dataset. Ketiga, menghitung jarak setiap data terhadap masing-masing centroid untuk menentukan keanggotaan kluster. Keempat, memperbarui posisi centroid berdasarkan rata-rata posisi anggota dalam kluster. Kelima, mengulangi proses hingga tidak ada perubahan signifikan pada nilai centroid.

Wang, M., et al. (2020). Riaz, U., et al. (2020). Pemilihan nilai k perlu mempertimbangkan tujuan analisis dan kualitas pemisahan kluster. Nilai k yang terlalu kecil dapat menyederhanakan struktur data secara berlebihan, sedangkan nilai k yang terlalu besar dapat menghasilkan kelompok yang terlalu rinci dan sulit diinterpretasikan (Wang et al., 2020; Riaz et al., 2020).

2.2.2 Keunggulan dan Keterbatasan Algoritma *K-Means*

Saxena, A., et al. (2020). Algoritma K-Means memiliki sejumlah keunggulan, antara lain kesederhanaan implementasi, efisiensi komputasi, dan kemampuan untuk menangani dataset berukuran besar. Proses perhitungan berbasis jarak Euclidean memungkinkan algoritma ini bekerja cepat pada data numerik dengan dimensi rendah hingga menengah. Keunggulan lainnya adalah kemampuannya dalam memberikan hasil pengelompokan yang intuitif dan mudah divisualisasikan.

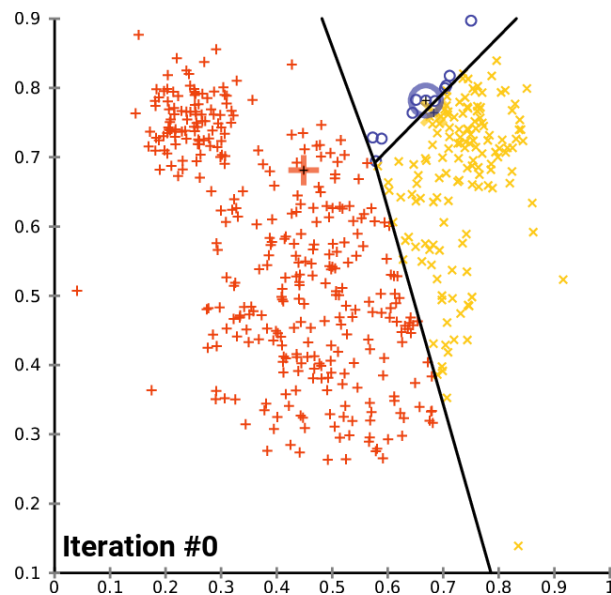
Zhang, Z., et al. (2020)Bhat, K. A., et al. (2021). Namun, algoritma ini juga memiliki beberapa keterbatasan. Hasil akhir sangat bergantung pada pemilihan centroid awal yang dapat menyebabkan variasi hasil antar percobaan. Selain itu, K-Means kurang efektif pada data yang memiliki distribusi non-linier atau bentuk klaster yang tidak bulat. menambahkan bahwa sensitivitas terhadap keberadaan nilai ekstrem (outlier) juga menjadi salah satu kelemahan yang perlu diatasi. Outlier dapat menggeser posisi centroid dan menurunkan kualitas hasil klaster. Oleh karena itu, praproses data seperti normalisasi dan penghapusan outlier menjadi tahap penting sebelum penerapan algoritma.

2.2.3 Penerapan *K-Means* dalam Analisis Data Pertanian

He, S., et al. (2020).)Penggunaan algoritma K-Means dalam sektor pertanian telah diterapkan untuk mengidentifikasi kelompok wilayah berdasarkan indikator numerik, seperti luas panen, produktivitas, kondisi tanah, dan curah hujan. Namun, interpretasi hasil harus disesuaikan dengan variabel yang benar-benar dianalisis dan tidak boleh diperluas menjadi kesimpulan sebab-akibat.

Implementasi algoritma *K-Means* juga dimanfaatkan untuk mendukung pertanian presisi (*precision agriculture*). Hasil penelitian Wang, M., et al. (2020). membuktikan bahwa *K-Means* mampu memberikan rekomendasi pola tanam yang sesuai berdasarkan hasil klaster wilayah dengan kondisi tanah yang serupa. Selain itu, penelitian oleh Riaz, U., et al. (2020). menunjukkan bahwa metode ini mampu memetakan zona pertanian potensial berbasis analisis spasial dan citra satelit, memberikan kontribusi signifikan terhadap sistem pertanian berkelanjutan.

Pendekatan ini memberikan nilai tambah bagi pengembangan sistem pendukung keputusan di bidang pertanian modern.



Gambar 2.2

K-Means dalam menemukan posisi *centroid* optimal.

2.2.4 Faktor yang Mempengaruhi Kinerja Algoritma *K-Means*

Kinerja algoritma *K-Means* sangat dipengaruhi oleh kualitas data yang digunakan. Data yang tidak ternormalisasi dapat menyebabkan variabel dengan skala besar mendominasi hasil pengelompokan. Cheng, J., & Wang, L. (2020). menegaskan bahwa normalisasi dan pembersihan data merupakan tahap prasyarat yang penting sebelum algoritma diterapkan. Selain itu, ukuran dataset juga memengaruhi performa karena jumlah data yang besar meningkatkan beban komputasi. Pemilihan nilai k optimal melalui metode *Elbow* atau *Silhouette* juga memengaruhi kualitas hasil kluster.

Performa algoritma dapat ditingkatkan melalui inisialisasi *centroid* yang baik. Metode *K-Means++* yang dikembangkan oleh Arthur dan Vassilvitskii menawarkan peningkatan stabilitas hasil dengan memilih *centroid* awal berdasarkan probabilitas jarak antar titik.

Evaluasi hasil K-Means diperlukan untuk menilai kekompakan anggota dalam kluster dan keterpisahan antar-kluster. Penelitian ini menggunakan WCSS, Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Index sebagai evaluasi internal; tidak dilakukan validasi eksternal menggunakan data lapangan.

Evaluasi hasil pengelompokan pada algoritma *K-Means* menjadi langkah penting untuk memastikan bahwa kluster yang terbentuk benar-benar mencerminkan pola alami dalam data. Dua metode yang paling umum digunakan dalam penentuan kualitas kluster adalah **metode *Elbow*** dan **metode *Silhouette Score*** (Suarsana et al., 2019b). Kedua pendekatan ini memberikan tolok ukur kuantitatif untuk menilai kesesuaian jumlah kluster (k) serta tingkat homogenitas di dalam kluster.

2.2.6 Evaluasi Kualitas Kluster dengan *Elbow Method* dan *Silhouette Score*

a) Metode *Elbow*

Algoritma K-Means clustering membutuhkan jumlah kluster (k) yang telah ditentukan sebelumnya agar dapat bekerja optimal. Namun, penentuan nilai k yang tepat menjadi tantangan utama dalam analisis pengelompokan data. Salah satu metode yang banyak digunakan untuk menentukan jumlah kluster ideal adalah Elbow Method. Prinsip utama metode ini ialah menghitung Within-Cluster

Sum of Squares (WCSS), yaitu total variasi antar data di dalam satu klaster. Semakin kecil nilai WCSS, semakin homogen data di dalam klaster tersebut.

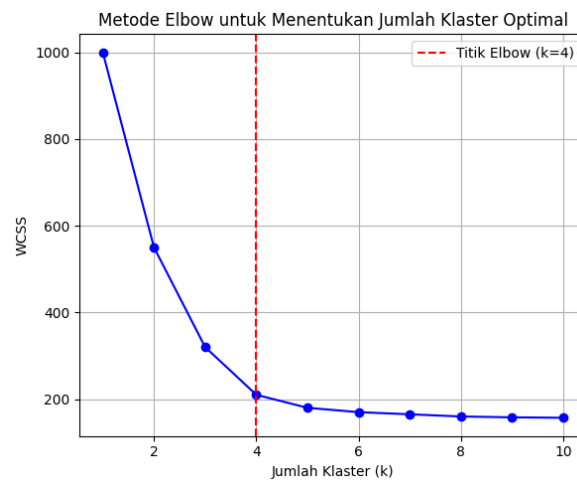
Proses *Elbow Method* dilakukan dengan menghitung WCSS untuk beberapa nilai k (biasanya 1–10) dan memvisualisasikan hasilnya dalam bentuk grafik. Kurva yang terbentuk menunjukkan penurunan nilai WCSS secara tajam pada awalnya, kemudian menurun lebih lambat hingga mencapai titik belok (*elbow point*). Titik inilah yang merepresentasikan jumlah klaster optimal karena setelah titik tersebut, penambahan klaster tidak memberikan peningkatan signifikan terhadap kualitas pengelompokan.

Metode ini membantu mengidentifikasi struktur kelompok berdasarkan karakteristik wilayah seperti luas panen, produktivitas, dan produksi padi. Pada contoh ilustratif, apabila penurunan WCSS terlihat tajam hingga $k = 4$ lalu melandai, $k = 4$ dapat dipertimbangkan sebagai titik elbow. Contoh tersebut hanya menjelaskan konsep dan bukan merupakan hasil pengujian data penelitian ini.

Selain *Elbow Method*, kualitas klaster dapat dinilai menggunakan *Silhouette Score*, yaitu ukuran yang membandingkan kedekatan suatu objek dengan klasternya sendiri dan jaraknya terhadap klaster terdekat lainnya (Kumar et al., 2019). Nilainya berkisar dari -1 sampai 1; nilai yang semakin mendekati 1 menunjukkan pemisahan klaster yang semakin baik

Gambar ini merupakan ilustrasi penggunaan *Elbow Method*. WCSS diplot pada sumbu y dan jumlah klaster (k) pada sumbu x. Garis merah putus-putus menunjukkan contoh titik elbow pada $k = 4$. Gambar ini hanya digunakan untuk

menjelaskan konsep; penentuan jumlah kluster pada data penelitian dibahas pada Bab IV.



Gambar 2.3

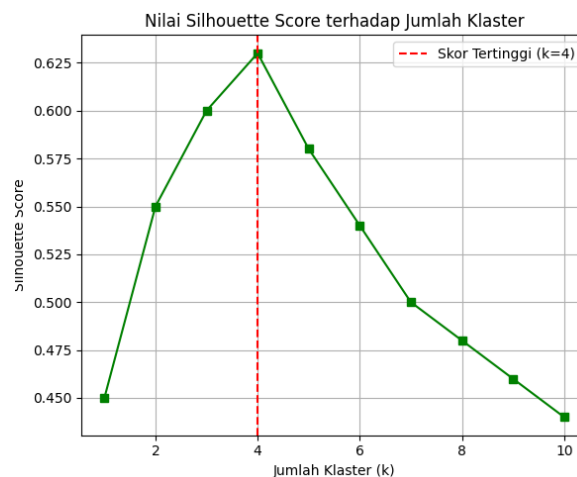
Diagram visualisasi metode Elbow.

b) Metode *Silhouette Score*

Selain menggunakan *Elbow Method*, kualitas hasil pengelompokan juga dapat dievaluasi menggunakan *Silhouette Score*. Metode ini menilai seberapa baik suatu data cocok dengan klasternya sendiri dibandingkan dengan kluster lain Kumar, A., et al. (2019). . Nilai *Silhouette Score* berkisar antara -1 hingga 1. Nilai mendekati 1 menunjukkan bahwa data berada dalam kluster yang tepat, nilai sekitar 0 menandakan data berada di batas antar kluster, sedangkan nilai negatif menunjukkan kemungkinan salah pengelompokan.

Dalam konteks analisis wilayah pertanian, *Silhouette Score* menjadi indikator penting untuk menilai apakah hasil klusterisasi benar-benar mencerminkan perbedaan nyata antar wilayah. Misalnya, wilayah dengan kesuburan tanah tinggi, curah hujan ideal, dan produktivitas tinggi harus

tergabung dalam satu klaster yang jelas terpisah dari wilayah dengan produktivitas rendah. Jika nilai *Silhouette Score* mencapai 0,63, maka kualitas klasterisasi tergolong baik dan menandakan bahwa hasil pengelompokan sudah cukup efektif dalam memisahkan karakteristik wilayah.



Gambar 2.4
Diagram visualisasi Silhouette Score

Gambar ini merupakan ilustrasi penggunaan Silhouette Score. Nilai skor diplot pada sumbu y dan jumlah klaster pada sumbu x. Pada contoh tersebut, nilai tertinggi berada pada $k = 4$. Gambar ini hanya menjelaskan konsep; hasil evaluasi data penelitian disajikan pada Bab IV.

c) Integrasi Evaluasi pada Analisis Data Pertanian

Kombinasi *Elbow Method* dan *Silhouette Score* menjadi pendekatan evaluatif yang saling melengkapi. *Elbow Method* memberikan dasar kuantitatif untuk menentukan jumlah klaster yang efisien, sedangkan *Silhouette Score* menilai kejelasan dan konsistensi hasil klasterisasi. Penerapan kedua metode ini pada data spasial pertanian menghasilkan pengelompokan wilayah yang dapat digunakan untuk mendukung kebijakan berbasis data.

Zhang, Y., & Sun, J. (2019). Elbow Method dan Silhouette Score dapat digunakan secara bersama untuk membantu membandingkan beberapa kandidat jumlah klaster. Elbow Method menilai penurunan WCSS, sedangkan Silhouette Score menilai kekompakan dan keterpisahan klaster. Hasil kedua metode perlu ditafsirkan bersama tujuan penelitian, bukan digunakan secara mekanis sebagai satu-satunya dasar pemilihan k.

Pendekatan evaluasi ini juga menjadi bagian penting dalam pengembangan *precision agriculture*, di mana keputusan berbasis data digunakan untuk meningkatkan efisiensi lahan dan produktivitas. Dengan demikian, evaluasi hasil klasterisasi melalui kedua metode ini tidak hanya menambah validitas ilmiah penelitian, tetapi juga berkontribusi langsung terhadap pengelolaan sumber daya pertanian yang berkelanjutan dan efektif.

2.3 Teknik Evaluasi Algoritma di *Machine Learning*

Evaluasi algoritma berfungsi untuk mengukur seberapa efektif model *machine learning* dalam menggambarkan pola atau hubungan antar data. Pada konteks klasterisasi, penilaian algoritma menjadi kompleks karena tidak adanya label data yang benar (*ground truth*). Oleh sebab itu, berbagai metrik evaluasi internal dan eksternal digunakan untuk menilai kualitas hasil pengelompokan tanpa harus membandingkannya dengan data referensi. Proses evaluasi ini memastikan bahwa klaster yang terbentuk memiliki tingkat kesamaan internal yang tinggi dan perbedaan antar klaster yang signifikan.

Pendekatan evaluasi algoritma klasterisasi pada penelitian ini mengacu pada konsep unsupervised learning, di mana data pertanian dikelompokkan

berdasarkan karakteristik yang serupa. Evaluasi tidak hanya berfokus pada hasil numerik, tetapi juga mencakup interpretasi terhadap data pertanian seperti produktivitas, luas panen, serta hasil produksi. Teknik ini membantu peneliti memahami sejauh mana model dapat merepresentasikan struktur alami dalam data yang digunakan.

Kinerja algoritma *K-Means clustering* umumnya dinilai menggunakan beberapa metrik utama, yaitu *Within-Cluster Sum of Squares (WCSS)*, *Silhouette Score*, dan *Davies-Bouldin Index*. Nilai WCSS menggambarkan total jarak kuadrat antara setiap titik data dan pusat klasternya. Semakin kecil nilai WCSS, semakin baik kualitas klaster yang dihasilkan karena menunjukkan bahwa anggota klaster memiliki tingkat kesamaan tinggi.

Patel, H., et al. (2019). Metode Silhouette Score digunakan untuk menilai seberapa baik pemisahan antar klaster. Nilai skor yang mendekati 1 menandakan bahwa objek berada di dalam klaster yang tepat, sedangkan nilai mendekati 0 menunjukkan ketidaktegasan batas antar klaster. Di sisi lain, Davies-Bouldin Index (DBI) mengukur rasio antara jarak dalam klaster dan jarak antar klaster. Nilai DBI yang lebih rendah menunjukkan hasil pengelompokan yang lebih baik. Kombinasi ketiga metrik ini membantu memastikan model K-Means menghasilkan segmentasi wilayah pertanian yang valid dan representatif.

Zhang, T., et al. (2021). Evaluasi klasterisasi pada data pertanian menghadapi tantangan khusus karena karakteristik data yang cenderung heterogen dan bergantung pada kondisi geografis, iklim, serta musim. Fluktuasi produktivitas yang tinggi antar wilayah menyebabkan perbedaan distribusi data

yang signifikan. Kondisi ini mengharuskan evaluasi dilakukan dengan mempertimbangkan konteks spasial dan temporal agar hasil pengelompokan tidak bias.

Selain itu, keterbatasan data lapangan yang lengkap dan terkini menjadi kendala lain dalam menilai efektivitas algoritma. Penggunaan data sekunder dari instansi pemerintah atau lembaga statistik sering kali memiliki format dan skala yang berbeda, sehingga proses normalisasi data menjadi penting. Evaluasi model yang mempertimbangkan ketiga aspek tersebut—kualitas data, variasi spasial, dan dinamika waktu—akan menghasilkan segmentasi wilayah pertanian yang lebih akurat dan adaptif.

Validasi hasil klasterisasi merupakan tahap penting untuk memastikan bahwa model yang dihasilkan memiliki relevansi praktis terhadap kondisi dunia nyata. Proses validasi dilakukan dengan membandingkan hasil pengelompokan terhadap indikator pertanian faktual seperti tingkat produktivitas, jenis tanah, dan curah hujan tahunan. Jika hasil klasterisasi konsisten dengan perbedaan nyata di lapangan, maka model dianggap valid secara empiris.

Pendekatan validasi ini mendukung terciptanya model pertanian berbasis data yang dapat diterapkan langsung dalam kebijakan publik. Misalnya, jika satu klaster menunjukkan wilayah dengan tingkat produktivitas tinggi dan kebutuhan pupuk rendah, maka pemerintah dapat mengarahkan strategi distribusi sumber daya berdasarkan hasil tersebut. Teknik evaluasi ini tidak hanya menilai ketepatan model dari sisi matematis, tetapi juga dari sisi kebermanfaatan dan aplikasinya terhadap peningkatan efisiensi sektor pertanian.

Evaluasi algoritma yang komprehensif, seperti yang diterapkan pada *K-Means clustering*, memperkuat dasar ilmiah penelitian ini. Pendekatan evaluatif yang sistematis dan berbasis metrik kuantitatif memberikan gambaran objektif mengenai kualitas model serta menjamin bahwa hasil analisis dapat mendukung pengambilan keputusan berbasis bukti (*evidence-based decision making*) di bidang pertanian.

2.4 Langkah-Langkah *Machine Learning* dalam Segmentasi Wilayah Pertanian

2.4.1 Tahap Persiapan dan Praproses Data

Keberhasilan penerapan *machine learning* dalam segmentasi wilayah pertanian sangat bergantung pada kualitas data yang digunakan. Proses praproses data (*data preprocessing*) menjadi langkah fundamental dalam memastikan bahwa data mentah yang dikumpulkan dari berbagai sumber seperti Badan Pusat Statistik (BPS), citra satelit, dan hasil survei lapangan berada dalam kondisi bersih dan konsisten. Data agronomi yang umum digunakan meliputi variabel luas panen, curah hujan, jenis tanah, produktivitas, dan indeks kesuburan. Tahapan ini bertujuan untuk menghapus duplikasi, memperbaiki kesalahan input, serta mengatasi nilai hilang (*missing values*) yang sering kali menjadi kendala utama dalam analisis spasial pertanian.

(Septianto et al., 2025) Normalisasi merupakan teknik yang secara umum dapat digunakan untuk menyetarakan skala variabel sebelum penerapan K-Means. Namun, penelitian ini menggunakan data dalam skala asli agar perhitungan manual tetap sesuai dengan satuan BPS. Konsekuensinya, variabel luas panen dan

produksi yang memiliki rentang lebih besar memberikan pengaruh lebih dominan terhadap Euclidean Distance dibandingkan produktivitas.

2.4.2 Tahap Pemilihan dan Pelatihan Model

Pemilihan model merupakan langkah strategis dalam siklus *machine learning*. Untuk segmentasi wilayah pertanian, algoritma *K-Means clustering* banyak dipilih karena kesederhanaan, efisiensi komputasi, serta kemampuannya mengidentifikasi pola alami dalam data multivariat. Sebelum pelatihan model dimulai, peneliti menentukan jumlah kluster optimal (k) menggunakan metode *Elbow* dan *Silhouette Score*.

Proses pelatihan model melibatkan inisialisasi centroid secara acak, perhitungan jarak Euclidean antar titik, serta pembaruan posisi centroid hingga mencapai konvergensi. Iterasi ini dilakukan berulang untuk memastikan bahwa posisi centroid berada pada titik yang paling representatif terhadap data. Evaluasi awal model dilakukan melalui nilai *inertia* (total jarak kuadrat antar titik dan centroid) dan *silhouette score* untuk menilai seberapa baik data dalam kluster terpisah satu sama lain. (Mukhtar et al., 2024) menunjukkan bahwa nilai *silhouette score* di atas 0,6 menandakan pengelompokan yang kuat dan struktur data yang jelas.

2.4.3 Tahap Evaluasi dan Validasi Model

Zhang, X., et al. (2020). Evaluasi dan validasi model merupakan langkah penting untuk memastikan reliabilitas hasil klasterisasi sebelum diterapkan dalam konteks nyata. Metode evaluasi yang umum digunakan antara lain *Silhouette Score*, *Davies-Bouldin Index (DBI)*, dan *Calinski-Harabasz Index*. Metrik-metrik

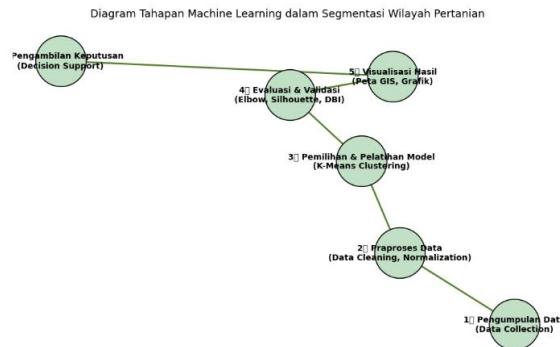
tersebut berfungsi untuk mengukur kualitas pembentukan klaster berdasarkan jarak antar klaster dan kedekatan internal antar data. Nilai DBI yang lebih rendah menandakan bahwa klaster memiliki jarak antar pusat yang besar dan distribusi internal yang homogen, yang berarti hasil segmentasi berkualitas tinggi.

Penelitian ini menggunakan evaluasi internal karena tidak tersedia label kelas atau data lapangan pembanding. Validitas hasil dinilai melalui konsistensi proses iterasi, WCSS, Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, serta interpretasi centroid berdasarkan tiga variabel penelitian.

2.4.4 Tahap Interpretasi dan Visualisasi Hasil

Interpretasi hasil klasterisasi dilakukan melalui tabel keanggotaan klaster, centroid, scatter plot, dan box plot. Penelitian ini belum menggunakan peta tematik atau analisis spasial berbasis GIS, sehingga hasil yang diperoleh merupakan pengelompokan atribut kabupaten/kota dan bukan spatially constrained clustering.

(Septianto et al., 2025) Peta digital berbasis GIS dapat digunakan pada penelitian lanjutan untuk menampilkan lokasi setiap kabupaten/kota menurut klasternya. Namun, GIS tidak digunakan sebagai bagian dari proses pengelompokan pada penelitian ini.



Gambar 2.5

Diagram Tahapan *Machine Learning* dalam Segmentasi Wilayah Pertanian

Gambar ini menampilkan alur kerja sistematis penerapan *machine learning* untuk analisis segmentasi wilayah pertanian. Tahapan dimulai dari praproses data, pemilihan model, evaluasi performa, hingga visualisasi hasil. Setiap tahap terhubung secara iteratif membentuk siklus analisis yang dinamis untuk mendukung pengambilan keputusan berbasis data.

2.5 Alat Bantu Pemrograman dan *Tools* Pendukung

Bahasa pemrograman berfungsi sebagai tulang punggung dalam pembangunan sistem analitik berbasis *machine learning* untuk sektor pertanian. Python menjadi bahasa yang paling dominan digunakan karena bersifat terbuka (*open source*), mudah dipelajari, dan memiliki komunitas besar yang aktif dalam mengembangkan pustaka dan dokumentasi. Keunggulan Python terletak pada fleksibilitasnya dalam menangani data numerik dan spasial, serta kemampuannya berintegrasi dengan sistem informasi geografis (*Geographic Information System* atau GIS) dan platform *cloud computing* seperti Google Colab.

(T. Santoso et al., 2023) menyatakan bahwa Python mendukung berbagai paradigma pemrograman seperti object-oriented, procedural, dan functional programming, sehingga memudahkan peneliti untuk menyesuaikan pendekatan analisis sesuai kebutuhan dataset. Bahasa ini juga mendukung visualisasi interaktif yang memungkinkan pengguna memantau hasil model secara langsung. Dibandingkan dengan bahasa lain seperti R atau MATLAB, Python menawarkan keseimbangan antara kekuatan komputasi, kemudahan sintaksis, dan ketersediaan pustaka analitik yang luas. Keunggulan ini menjadikan Python pilihan utama untuk penelitian agrikultur modern yang berfokus pada efisiensi, otomatisasi, dan skalabilitas.

Santoso et al. (2023) menyatakan bahwa Python mendukung berbagai paradigma pemrograman dan memiliki ekosistem pustaka yang luas sehingga memudahkan pengolahan data, penerapan algoritma, dan visualisasi hasil.

(Garuda2945045, n.d.-a) Ekosistem Python menyediakan berbagai pustaka yang saling melengkapi untuk mendukung analisis data pertanian. Pustaka pandas berperan dalam manipulasi dan pengelolaan data dalam format Data Frame, memungkinkan pengguna melakukan pembersihan, transformasi, serta penggabungan data dengan mudah. Num Py menyediakan operasi numerik yang cepat dan efisien untuk perhitungan vektor dan matriks, sedangkan Matplotlib dan Seaborn digunakan untuk visualisasi data dalam bentuk grafik dan diagram yang informatif.

(T. Santoso et al., 2023) Pustaka scikit-learn merupakan inti dalam penerapan algoritma machine learning. Library ini menyediakan algoritma

pengelompokan K-Means serta fungsi evaluasi seperti Silhouette Score dan Davies-Bouldin Index. Pandas digunakan untuk pengelolaan data, Num Py untuk operasi numerik, sedangkan Matplotlib digunakan untuk visualisasi. Kombinasi pustaka tersebut mendukung proses analisis yang transparan dan dapat direplikasi.

2.5.2 Keunggulan Penggunaan *Open-Source Tools* dalam Penelitian Pertanian

Patel, H., et al. (2019). Pustaka scikit-learn menyediakan implementasi K-Means serta metrik evaluasi seperti Silhouette Score dan Davies-Bouldin Index. Pandas digunakan untuk pengelolaan data, Num Py untuk operasi numerik, dan Matplotlib untuk visualisasi. Kombinasi pustaka tersebut mendukung analisis yang transparan dan dapat direplikasi (Santoso et al., 2023).

(Aldino et al., 2021) Kelebihan lain perangkat lunak open-source terletak pada fleksibilitas integrasi. Skrip dapat disesuaikan dengan konteks dan kebutuhan analisis, sementara keterbukaan kode mendukung replikasi hasil dan pemeriksaan ulang prosedur penelitian (Aldino et al., 2021).

2.5.3 Otomatisasi dan Skalabilitas Pemrosesan Data

Liu, H., et al. (2020) Kelebihan perangkat lunak open-source terletak pada fleksibilitas, keterbukaan kode, dan kemudahan replikasi. Skrip dapat diperiksa dan disesuaikan dengan kebutuhan analisis sehingga prosedur penelitian lebih transparan (Aldino et al., 2021).

Skalabilitas sistem analisis menjadi penting untuk menyesuaikan kapasitas komputasi dengan pertumbuhan data yang terus meningkat. Python mendukung koneksi dengan layanan komputasi awan seperti Google Colab dan Amazon Web Services (AWS) untuk mempercepat proses analisis spasial. Penelitian (Aldino et

al., 2021) menunjukkan bahwa pemanfaatan komputasi awan dapat meningkatkan efisiensi pemrosesan hingga 60%, terutama pada data spasial multidimensi. Pendekatan ini memungkinkan penerapan analitik waktu nyata (*real-time analytics*) dalam memantau produktivitas lahan dan mendeteksi potensi anomali pertanian.

2.6 Penelitian Terdahulu dan Kelebihan Penelitian Terdahulu

Penelitian yang memanfaatkan *K-Means* untuk analisis wilayah pertanian menegaskan efektivitas algoritme tersebut dalam mengidentifikasi kelompok wilayah dengan karakteristik agroekologi yang serupa, seperti pola curah hujan, indeks kelembapan, dan variasi produktivitas tanaman. Implementasi *K-Means* dengan kombinasi data citra satelit dan data administrasi menghasilkan kluster yang dapat digunakan untuk zonasi pengelolaan wilayah pertanian Shinde, M. R., & Mane, D. G. (2021). . Klasterisasi berbasis data semacam ini terbukti membantu pemerintah daerah dan lembaga pertanian dalam menurunkan ketidakefisienan alokasi pupuk dan air irigasi ketika rekomendasi hasil kluster diterapkan di lapangan.

Studi-studi terdahulu memanfaatkan *data fusion* antara data spasial dan non-spasial untuk menangkap dinamika musiman dan heterogenitas lokal. Penggabungan data satelit (Sentinel/Landsat) dengan data statistik produksi kabupaten meningkatkan ketepatan model. Proses praproses menjadi tahap penting karena *K-Means* sensitif terhadap skala dan *outlier*. Peneliti menerapkan normalisasi Min–Max atau *Z-score*, deteksi nilai ekstrem menggunakan *IQR*, dan imputasi data hilang berbasis median atau *k-NN* untuk menjaga integritas dataset.

Pendekatan ini meningkatkan kestabilan hasil klaster dan mendukung interpretasi wilayah prioritas untuk pengembangan.

Keterbatasan utama penelitian terdahulu terletak pada kualitas data (resolusi spasial rendah dan cakupan waktu terbatas) serta asumsi bentuk klaster yang dihasilkan *K-Means*. Pola klaster yang dianggap berbentuk bulat sering kali tidak sesuai dengan distribusi spasial yang kompleks. Selain itu, sulitnya mengintegrasikan variabel kualitatif seperti jenis pengairan atau praktik budidaya menyebabkan hasil klasterisasi kurang mencerminkan kondisi sosial-teknis wilayah. Beberapa penelitian mengusulkan metode hibrid atau *constrained clustering* untuk mengatasi kelemahan tersebut.

2.6.1 Kelebihan dan Inovasi Penelitian Ini

Bhat, K. A., et al. (2021). Kontribusi penelitian ini terletak pada penerapan *K-Means* terhadap data luas panen, produktivitas, dan produksi padi kabupaten/kota di Provinsi Sumatera Utara tahun 2023, disertai perhitungan manual, verifikasi menggunakan Python, serta perbandingan beberapa kandidat jumlah klaster dengan evaluasi internal. Penelitian ini tidak melakukan validasi eksternal menggunakan data lapangan.

2.6.2 Ringkasan Penelitian Terdahulu

Tabel berikut merangkum lima penelitian terdahulu yang relevan dengan penerapan algoritma *K-Means* di sektor pertanian, serta satu penelitian ini sebagai pembeda dan penguat kontribusi.

Tabel 2.1 Ringkasan Penelitian Terdahulu

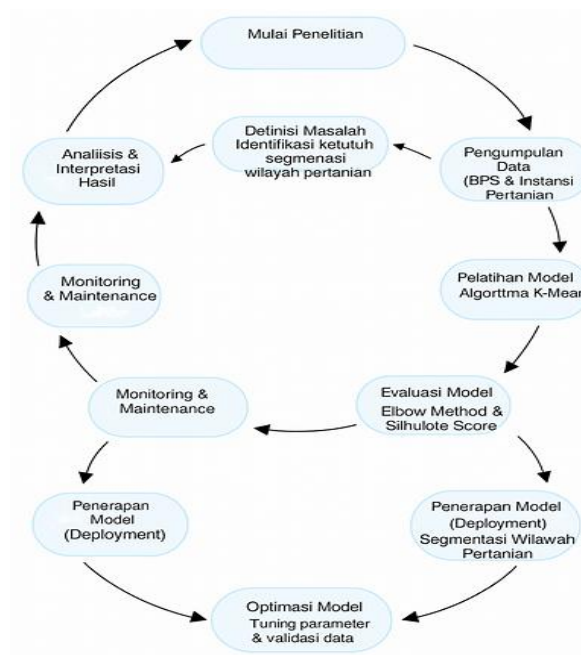
NO	Peneliti	Judul	Tahun	Data dan Algoritma	Hasil Utama
1	Swain, K. P.	<i>Empowering Crop Selection with Ensemble Learning and K-Means Clustering</i>	2024	Data hasil panen dan kondisi tanah, menggunakan <i>K-Means</i> dan <i>Ensemble Learning</i>	Menghasilkan klaster wilayah optimal untuk rekomendasi tanaman unggulan berdasarkan kondisi tanah dan produktivitas.
2	Zhu, F.	<i>Using Constrained K-Means Clustering for Soil Texture Mapping</i>	2025	Data tekstur tanah dan citra satelit, algoritma <i>Constrained K-Means</i>	Pemetaan tekstur tanah secara otomatis untuk mendukung pengelolaan lahan pertanian presisi.
3	Rathore, R. et al.	<i>Machine Learning in Agriculture and Crop Management: A Comprehensive Review</i>	2025	Data berbagai penelitian pertanian dengan algoritma ML (termasuk <i>K-Means</i>)	Menunjukkan potensi besar <i>K-Means</i> untuk pengelompokan wilayah pertanian dan manajemen hasil pertanian.
4	Razhmanita, E.	<i>Optimization of Village Grouping Using Comparison of K-Means and K-Medoids</i>	2025	Data demografi desa, membandingkan <i>K-Means</i> dan <i>K-Medoids</i>	<i>K-Means</i> menghasilkan pembagian wilayah yang lebih efisien dalam pengelompokan wilayah administratif.
5	Botero-Valencia, J. et al.	<i>Machine Learning in Sustainable Agriculture: Systematic Bibliometric</i>	2025	Data bibliometrik publikasi pertanian, berbagai algoritma ML	Mengidentifikasi tren penggunaan <i>K-Means</i> dan algoritma lain untuk keberlanjutan

		<i>Review</i>			pertanian global.
6	Penelitian Ini	<i>Segmentasi Wilayah Pertanian di Provinsi Sumatera Utara Menggunakan Metode K-Means</i>	2025	Data luas panen, produktivitas, produksi padi (tahun 2023), algoritma K-Means	Menghasilkan segmentasi wilayah pertanian yang efisien untuk mendukung kebijakan pertanian berbasis data dan validasi lapangan.

2.7 Flowchart Kerangka Penelitian

Flowchart penelitian berperan penting dalam menggambarkan alur kerja penelitian secara visual dan sistematis. Representasi ini memungkinkan pembaca memahami urutan logis dari tahapan penelitian mulai dari pengumpulan data hingga analisis hasil. Setiap langkah dalam flowchart merepresentasikan keputusan atau proses yang saling berkaitan, yang memudahkan identifikasi hubungan sebab-akibat di antara tahap penelitian. Penggunaan flowchart dalam penelitian pertanian berbasis *machine learning* telah menjadi praktik umum karena mampu memperjelas integrasi antara proses teknis dan analisis ilmiah.

Gandhi, S., et al. (2020). Flowchart juga memberikan pemetaan yang ringkas mengenai metodologi penelitian yang kompleks. Diagram ini menampilkan hubungan antarproses seperti tahap praproses data, pemilihan model, evaluasi, dan validasi, yang secara keseluruhan membentuk satu sistem penelitian terpadu. Penerapan flowchart dalam penelitian segmentasi wilayah pertanian menggunakan algoritma K-Means berfungsi untuk memperjelas bagaimana data mentah diolah hingga menjadi informasi yang dapat digunakan untuk pengambilan keputusan.



Gambar 2.6

Flowchart Penelitian Segmentasi Wilayah Pertanian Menggunakan K-Means

Deskripsi: Flowchart menggambarkan tahapan mulai dari *Problem Definition*, *Data Collection*, *Preprocessing*, *Modeling*, hingga *Evaluation* dan *Deployment*. Setiap blok proses dihubungkan oleh panah yang menunjukkan aliran data dan informasi secara sistematis.

Flowchart ini menggambarkan tahapan penelitian secara menyeluruh mulai dari pendefinisian masalah hingga interpretasi hasil akhir. Setiap blok proses menunjukkan alur kegiatan penelitian yang meliputi *Problem Definition*, *Data Collection*, *Preprocessing*, *Modeling*, *Evaluation*, serta *Deployment*. Diagram ini menampilkan keterkaitan logis antara tahap konseptual dan tahap teknis penelitian, memperlihatkan bagaimana data pertanian dikumpulkan, diproses, dan diklasterisasi menggunakan algoritma *K-Means*. Visualisasi ini memperkuat pemahaman pembaca terhadap sistematika metodologi penelitian

yang digunakan, sekaligus menegaskan hubungan antara pengolahan data dan pengambilan keputusan berbasis informasi yang dihasilkan dari model klasterisasi tersebut.

2.7.1 Struktur dan Komponen Flowchart Penelitian

Flowchart penelitian ini terdiri dari beberapa komponen utama, yaitu identifikasi masalah, pengumpulan data, pra-pemrosesan, penerapan algoritma, evaluasi, serta analisis hasil. Tahapan-tahapan tersebut membentuk urutan logis yang saling berkaitan dan mendukung pencapaian tujuan penelitian. Komponen-komponen tersebut juga mencerminkan penerapan pendekatan ilmiah yang sistematis dalam konteks penelitian pertanian berbasis *machine learning*.

1. Identifikasi Masalah dan Tujuan Penelitian

Tahap ini berfokus pada pengenalan masalah utama dalam pengelolaan data pertanian serta perumusan tujuan penelitian yang akan dicapai. Identifikasi yang tepat menjadi dasar bagi efektivitas tahapan selanjutnya.

2. Pengumpulan Data Spasial Pertanian

Pengumpulan data mencakup variabel seperti luas lahan, hasil panen, jenis tanaman, serta koordinat wilayah pertanian. Data diperoleh melalui sumber terbuka maupun instansi terkait, dengan memperhatikan validitas dan kelengkapan informasi.

3. Pra-pemrosesan Data

Proses ini mencakup pemeriksaan kelengkapan data, konsistensi satuan, konversi nilai ke format numerik, dan pemeriksaan nilai ekstrem. Pada analisis utama, data digunakan dalam skala asli agar konsisten dengan perhitungan

manual; konsekuensinya, variabel dengan rentang terbesar dapat memberikan pengaruh lebih kuat terhadap jarak Euclidean.

4. Penerapan Algoritma *K-Means Clustering*

Algoritma ini digunakan untuk mengelompokkan wilayah berdasarkan kesamaan karakteristik data. Proses iteratif dilakukan hingga mencapai konvergensi, di mana pusat kluster stabil.

5. Evaluasi Model dengan *Elbow Method* dan *Silhouette Score*

Evaluasi ini bertujuan menentukan jumlah kluster optimal dan mengukur tingkat pemisahan antar kluster. Kombinasi kedua metode tersebut menghasilkan hasil segmentasi yang lebih objektif dan akurat.

6. Analisis dan Interpretasi Hasil Segmentasi

Tahap analisis berfungsi untuk memahami makna dan pola dari hasil pengelompokan. Interpretasi hasil ini mendukung pengambilan keputusan dalam perencanaan kebijakan pertanian berbasis data.

7. Kesimpulan dan Rekomendasi

Bagian akhir penelitian menyajikan temuan utama serta memberikan saran untuk penelitian lanjutan atau penerapan hasil dalam konteks nyata.