PERBANDINGAN EFEKTIVITAS ALGORITMA K-MEANS DA N FUZZY C-MEANS UNTUK CLUSTERING DATA PRODUKSI ALPUKAT DI INDONESIA

Duncan Ariel [1)] Teny Handayani [2)]
[1)] Teknik Informatika Universitas Tarumanegara
Jl. Letjen S. Parman No.1, Kota Jakarta Barat, Daerah Khusus Ibukota Jakarta 11440
email : duncan.535230063@stu.untar.ac.id

ABSTRAK

Penelitian ini bertujuan untuk membandingkan performa algoritma K-Means dan Fuzzy C-Means (FCM) dalam proses klasterisasi data produksi alpukat di Indonesia. Metode yang digunakan adalah pendekatan machine learning berbasis unsupervised clustering, di mana data produksi nasional—mencakup volume produksi, luas panen, dan produktivitas—dianalisis untuk mengidentifikasi pola kewilayahan. Kinerja kedua algoritma diuji dan divalidasi menggunakan metrik evaluasi Silhouette Score, Davies-Bouldin Index, dan waktu komputasi. Hasil eksperimen menunjukkan bahwa kedua algoritma secara konvergen menemukan 2 sebagai jumlah klaster yang paling optimal dengan skor evaluasi yang superior. Namun, perbandingan lebih lanjut menunjukkan bahwa algoritma K-Means memiliki keunggulan signifikan dalam hal efisiensi waktu komputasi dan robustisitas model yang lebih baik , sedangkan Fuzzy C-Means membutuhkan waktu proses yang lebih lama. Analisis spasial juga berhasil memetakan wilayah produksi dan mengonfirmasi konsentrasi klaster produksi tinggi di Pulau Jawa. Temuan ini merekomendasikan K-Means sebagai metode yang lebih pragmatis untuk klasterisasi data produksi alpukat skala besar.

Key words

Alpukat, K-Means, Fuzzy C-Means, Klasterisasi, Machine Learning

1. Pendahuluan

Alpukat (*Persea americana* Mill.) merupakan komoditas agribisnis strategis di Indonesia yang memiliki prospek ekonomi cerah [1]. Potensi pengembangannya didukung oleh permintaan pasar domestik yang kuat serta peluang ekspor yang terus terbuka [2]. Tren positif ini juga sejalan dengan meningkatnya permintaan pasar global terhadap buah alpukat [3]. Sebagai negara agraris, Indonesia memiliki potensi besar untuk menjadi produsen utama, namun pemanfaatan potensi ini terkendala oleh kompleksitas dan variasi data produksi antar wilayah [4]. Data produksi alpukat, yang mencakup volume, luas lahan, hingga produktivitas, seringkali tersebar dan tidak terstruktur, sehingga menyulitkan para pemangku kepentingan dalam melakukan perencanaan, pemantauan, dan pengambilan keputusan strategis yang efektif [5]. Kualitas dan konsistensi data menjadi fondasi utama; tanpa pra-pemrosesan yang memadai, hasil analisis berisiko menjadi bias dan tidak akurat [6].

Untuk mengatasi tantangan ini, penerapan *machine learning* (pembelajaran mesin) menawarkan pendekatan yang canggih untuk menganalisis dan menemukan pola-pola tersembunyi dalam data pertanian [7]. Salah satu cabang utama dalam pembelajaran mesin yang relevan untuk kasus ini adalah pembelajaran tanpa pengawasan (*unsupervised learning*), khususnya dengan teknik *clustering* (pengelompokan) [8]. *Clustering* memungkinkan pengelompokan wilayah-wilayah produksi ke dalam beberapa grup (*cluster*) berdasarkan kesamaan karakteristiknya, tanpa memerlukan label atau kategori yang telah ditentukan sebelumnya. Tujuannya adalah untuk membentuk kelompok di mana wilayah dalam satu *cluster* memiliki profil produksi yang serupa, sementara profil antar *cluster* sangat berbeda [9].

Dua algoritma *clustering* yang paling fundamental dalam literatur pembelajaran mesin adalah K-Means dan Fuzzy C-Means (FCM). K-Means, yang merupakan algoritma *hard clustering*, sangat populer karena efisiensi dan kemudahan interpretasinya. Algoritma ini telah terbukti efektif dalam berbagai studi pertanian, seperti pengelompokan perkebunan kopi berdasarkan profil produktivitas [10] dan pemetaan wilayah produktivitas padi di Indonesia [11] , serta identifikasi zona potensi komoditas karet [12]. Sebagai alternatif, Fuzzy C-Means (FCM) menawarkan pendekatan *soft clustering*, di mana suatu wilayah dapat menjadi anggota dari beberapa *cluster* dengan derajat keanggotaan yang berbeda [13]. Pendekatan ini dinilai lebih realistis untuk data agrikultur yang seringkali memiliki batas-batas yang tidak tegas [14] atau dalam pemodelan kesesuaian lahan yang ambigu [15].

Meskipun kedua algoritma ini telah teruji, perbandingan langsung untuk menentukan metode mana yang paling efektif masih menjadi area eksplorasi yang aktif. Studi komparatif serupa telah berhasil diterapkan pada konteks data kewilayahan lain di Indonesia, misalnya untuk data meteorologi [16]. Namun, perbandingan kinerja K-Means dan FCM untuk memetakan potensi produksi alpukat secara spesifik

berdasarkan data riil nasional masih menjadi area yang perlu dieksplorasi lebih dalam [17]. Oleh karena itu, penelitian ini bertujuan untuk membandingkan secara sistematis kinerja kedua algoritma tersebut dalam mengelompokkan data produksi alpukat di Indonesia. Validasi kinerja akan diukur menggunakan metrik evaluasi standar untuk memastikan objektivitas perbandingan [18]. Hasil penelitian ini diharapkan dapat menjadi landasan berbasis data bagi pemerintah dan pelaku agribisnis untuk merumuskan kebijakan pengembangan komoditas alpukat yang lebih akurat, sejalan dengan posisi strategis alpukat dalam agenda pembangunan agribisnis nasional [1].

2. Metode Penelitian

Penelitian ini menerapkan metode kuantitatif-eksperimental untuk menganalisis dan membandingkan dua algoritma *clustering* dari ranah pembelajaran mesin. Fokus utama penelitian adalah penerapan algoritma K-Means dan Fuzzy C-Means pada dataset produksi alpukat yang diperoleh dari Basis Data Statistik Pertanian (BDSP) Kementerian Pertanian [4]. Analisis ini bertujuan untuk mengidentifikasi pengelompokan provinsi di Indonesia berdasarkan tiga atribut utama: volume produksi total (ton), luas panen (hektare), dan tingkat produktivitas (ton/ha).

2.1 Algoritma *K-Means*

K-Means adalah algoritma partisi non-hirarkis yang membagi dataset menjadi K *cluster* yang telah ditentukan sebelumnya [9]. Sebagai salah satu metode *clustering* paling fundamental, K-Means sering digunakan sebagai model dasar (*baseline*) dalam berbagai studi komparatif berkat efisiensi dan kejelasannya dalam menghasilkan kelompok yang saling lepas (eksklusif) [19]. Algoritma ini sangat populer di berbagai bidang, termasuk pertanian, untuk segmentasi pasar, analisis geospasial, dan identifikasi pola produksi.

Prinsip kerja utama K-Means adalah meminimalkan varians intra-*cluster*, yang secara teknis diukur melalui metrik *Within-Cluster Sum of Squares* (WCSS) [20]. Tujuannya adalah untuk menciptakan kelompok-kelompok data yang sepadat dan sekompak mungkin. Dengan meminimalkan WCSS, algoritma ini secara efektif menemukan pusat-pusat kelompok (sentra) yang paling representatif dari data. Hal ini membuat K-Means cenderung menghasilkan *cluster* yang berbentuk sferis (bulat) dan memiliki ukuran yang relatif seimbang.

Secara prosedural, K-Means bekerja melalui proses iteratif yang sederhana. Proses diawali dengan tahap inisialisasi, di mana jumlah *cluster* (K) ditentukan dan K titik data dipilih secara acak sebagai pusat *cluster* awal (*centroids*). Selanjutnya, pada tahap penugasan (*assignment*), setiap titik data dalam dataset ditetapkan ke dalam *cluster* yang memiliki *centroid* terdekat. Setelah semua titik data ditugaskan, dilakukan tahap pembaruan (*update*), di mana posisi setiap *centroid* dihitung kembali sebagai titik rata-rata dari seluruh data yang menjadi anggota *cluster* tersebut. Siklus penugasan dan pembaruan ini diulang terus-menerus hingga posisi *centroid* tidak lagi mengalami perubahan signifikan, yang menandakan bahwa model telah mencapai konvergensi.

Secara matematis, proses inti dalam algoritma K-Means didasarkan pada tiga formulasi utama berikut:

1. Fungsi Objektif: *Within-Cluster Sum of Squares* (WCSS) Tujuan utama dari K-Means adalah meminimalkan total varians di dalam semua *cluster*. Fungsi yang digunakan untuk mengukur ini disebut WCSS, yang secara matematis dirumuskan sebagai:

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{x \in S_i} |x - \mu_i|^2 \quad \text{..... (1)}$$

di mana:

- k : jumlah *cluster*.
- Si : himpunan titik data yang termasuk dalam *cluster* ke-i.
- x : sebuah titik data (vektor).
- μi : *centroid* (pusat) dari *cluster* ke-i.

2. Metrik Jarak: *Euclidean Distance* Untuk menempatkan setiap titik data ke *centroid* terdekat, K-Means perlu menghitung jarak. Rumus untuk kuadrat jarak Euclidean antara sebuah titik data xp dan *centroid* μi dalam ruang d-dimensi adalah:

$$d(x_p, \mu_i)^2 = \sum_{j=1}^{d} (x_{pj} - \mu_{ij})^2 \quad \text{..... (2)}$$

di mana:

- xpj : nilai atribut ke-j dari titik data xp.
- μij : nilai atribut ke-j dari *centroid* μi.

3. Pembaruan *Centroid* Pada tahap pembaruan, *centroid* baru dihitung sebagai nilai rata-rata (mean) dari semua titik data yang menjadi anggota *cluster* tersebut. Rumus untuk memperbarui *centroid* μi adalah:

$$\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x \quad \text{..... (3)}$$

di mana:

- |S_i| : jumlah total titik data dalam *cluster* S_i.

Komponen paling krusial dalam K-Means adalah *centroid* dan metrik jarak [21]. *Centroid* berfungsi sebagai representasi virtual dari sebuah *cluster*, yang posisinya terus beradaptasi. Metrik jarak, khususnya *Euclidean distance*, adalah pilihan umum untuk data numerik. Namun, perlu dicatat bahwa untuk tipe data tertentu seperti data deret waktu (time-series), metrik jarak yang lebih canggih seperti *Dynamic Time Warping* (DTW) dapat memberikan hasil yang lebih akurat, seperti yang telah ditunjukkan pada studi pengelompokan data meteorologi di Indonesia [22].

Dalam konteks penelitian ini, algoritma K-Means akan diterapkan untuk mengelompokkan provinsi-provinsi di Indonesia ke dalam kategori-kategori produksi alpukat yang jelas. Hasil dari K-Means diharapkan dapat membentuk cluster yang mudah diinterpretasikan, seperti "Klaster Produsen Rendah", "Klaster Produsen Sedang", dan "Klaster Produsen Tinggi".

Kekuatan utama K-Means terletak pada kesederhanaan implementasi dan kecepatan komputasinya [10]. Hasilnya yang berupa partisi tegas memberikan kejelasan dan kemudahan dalam interpretasi, yang sangat berharga bagi para pengambil keputusan. Banyak penelitian telah menunjukkan keberhasilan K-Means dalam memberikan wawasan yang dapat ditindaklanjuti di sektor pertanian [11].

Meskipun demikian, K-Means memiliki beberapa keterbatasan yang perlu diperhatikan. Pertama, algoritma ini mengharuskan jumlah cluster (K) ditentukan di awal, yang seringkali menjadi tantangan tersendiri dan memerlukan metode tambahan untuk validasi [23]. Kedua, kinerjanya sangat sensitif terhadap pemilihan posisi centroid awal yang acak, yang berpotensi menyebabkannya terjebak dalam solusi optimal lokal. Terakhir, asumsi bahwa cluster harus berbentuk sferis membatasi kemampuannya untuk menemukan kelompok dengan bentuk yang lebih kompleks atau tidak teratur [10].

2.2 Algoritma *Fuzzy C-Means* (FCM)

Fuzzy C-Means (FCM) adalah algoritma *soft clustering* yang merupakan perluasan dari K-Means dengan mengadopsi teori himpunan *fuzzy* (*fuzzy set theory*) [13]. Berbeda secara fundamental dari K-Means, FCM memungkinkan setiap titik data untuk menjadi anggota dari beberapa *cluster* secara simultan dengan derajat keanggotaan yang bervariasi. Pendekatan ini, yang berakar dari teori himpunan fuzzy, telah berkembang menjadi salah satu metode *soft clustering* paling berpengaruh dalam analisis data kontemporer [24].

Prinsip kerja FCM didasarkan pada konsep derajat keanggotaan, di mana setiap titik data memiliki nilai keanggotaan untuk setiap *cluster*. Penugasan yang bersifat probabilistik ini memungkinkan model untuk merepresentasikan ketidakpastian dan ambiguitas yang seringkali melekat pada data dunia nyata, seperti data pertanian di mana batas antar wilayah seringkali tidak tajam [14].

Secara matematis, FCM bekerja dengan cara meminimalkan fungsi objektif (Jm) secara iteratif. Fungsi ini mencari keseimbangan optimal antara menjaga jarak setiap titik data ke pusat *cluster* sependek mungkin dan mengelola distribusi derajat keanggotaan. Proses optimasi ini dilakukan dengan memperbarui pusat *cluster* dan matriks keanggotaan secara berulang hingga konvergen.

Fungsi objektif FCM dirumuskan sebagai berikut:

$$J_m = \sum_{i=1}^{N} \sum_{k=1}^{C} (u_{ik})^m |x_i - v_k|^2 \quad \text{.....(4)}$$

di mana:

- Jm: Nilai fungsi objektif yang akan diminimalkan.
- N : Jumlah total titik data.
- C : Jumlah *cluster*.
- uik : Derajat keanggotaan dari titik data ke-i pada *cluster* ke-k.
- m : Parameter *fuzzifier* atau pembobot (bilangan real > 1).
- xi : Vektor data dari titik data ke-i.
- vk : Vektor pusat (centroid) dari *cluster* ke-k.
- ‖xi−vk‖2 : Kuadrat dari jarak Euclidean antara titik data xi dan pusat *cluster* vk.

Dua parameter utama yang mendefinisikan perilaku FCM adalah matriks keanggotaan (U) dan parameter *fuzzifier* (m). Matriks U adalah inti dari model FCM, yang menyimpan informasi kuantitatif tentang seberapa besar "rasa memiliki" suatu titik data terhadap semua *cluster* yang ada. Sementara itu, *fuzzifier* m adalah hiperparameter krusial yang mengatur tingkat tumpang tindih atau "keburaman" antar *cluster*. Nilai m yang mendekati 1 akan membuat FCM berperilaku seperti K-Means (menghasilkan partisi yang hampir tegas), sedangkan nilai m yang lebih tinggi akan menghasilkan *cluster* yang lebih tumpang tindih dan kabur [25].

Dalam aplikasi pada data produksi alpukat, FCM dapat memberikan wawasan yang lebih kaya. Sebagai contoh, sebuah provinsi mungkin tidak sepenuhnya masuk kategori "Produsen Sedang" atau "Produsen Tinggi", tetapi berada di antara keduanya. FCM dapat menangkap nuansa ini dengan memberikan nilai keanggotaan, misalnya, 0.7 untuk *cluster* "Sedang" dan 0.3 untuk *cluster* "Tinggi". Informasi granular semacam ini sangat berharga untuk analisis yang lebih mendalam [17].

Kelebihan utama FCM adalah fleksibilitasnya dalam memodelkan data yang memiliki kelompok tumpang tindih dan lebih *robust* dalam menangani *noise* dibandingkan K-Means [14]. Namun, FCM memiliki kelemahan seperti kompleksitas komputasi yang lebih tinggi dan hasil yang terkadang lebih sulit untuk diinterpretasikan menjadi tindakan kebijakan konkret [17].

3. Hasil Percobaan

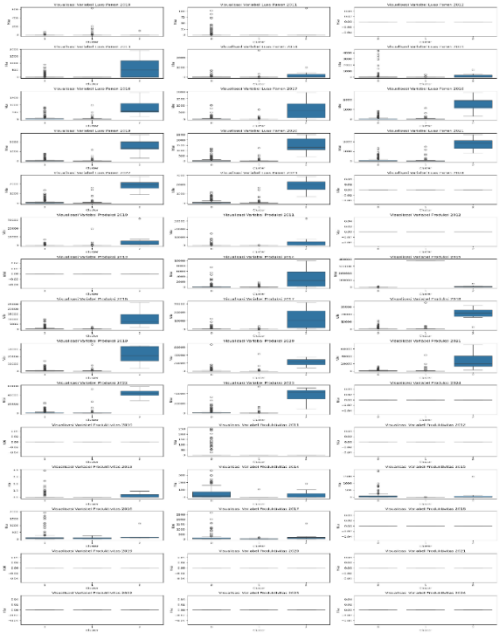
Proses Eksperimen dilakukan menggunakan dataset produksi alpukat per provinsi di Indonesia. Dataset ini mencakup tiga variabel utama, yaitu luas panen (hektare), total produksi (ton), dan produktivitas (ton/hektare). Ketiga variabel ini dipilih karena secara komprehensif mewakili aspek kuantitatif dan kualitatif yang menentukan kinerja agribisnis alpukat. Mengingat posisi alpukat sebagai salah satu komoditas strategis dengan tren permintaan yang terus meningkat, pemahaman mendalam mengenai pola produksinya

menjadi krusial untuk mendukung perencanaan agribisnis nasional.

Data yang digunakan diperoleh dari publikasi resmi Basis Data Statistik Pertanian (BDSP) Kementerian Pertanian. Sebelum dilakukan proses klasterisasi, data tersebut terlebih dahulu melalui tahapan pra-pemrosesan untuk memastikan kualitas dan konsistensi. Tahapan pra-pemrosesan meliputi pembersihan data (*data cleaning*) untuk menangani nilai yang hilang, serta normalisasi data menggunakan metode Min-Max Scaling. Langkah ini sangat penting untuk menyamakan skala data agar tidak ada variabel yang mendominasi proses perhitungan hanya karena perbedaan skala (misalnya, nilai total produksi dalam ribuan ton dibandingkan nilai produktivitas dalam puluhan ton).

Setelah data siap, proses klasterisasi dilakukan dengan membandingkan dua algoritma: K-Means dan Fuzzy C-Means. Eksperimen dijalankan dengan memvariasikan jumlah klaster dari 2 hingga 10 untuk dapat mengevaluasi dan menentukan jumlah klaster yang paling optimal. Algoritma K-Means digunakan sebagai pendekatan *hard clustering*, di mana setiap provinsi hanya dapat menjadi anggota dari satu klaster secara tegas. Sebaliknya, algoritma Fuzzy C-Means (FCM) menerapkan prinsip *soft clustering*, yang memungkinkan satu provinsi memiliki derajat keanggotaan di beberapa klaster secara bersamaan. Pendekatan FCM ini sering dianggap lebih cocok untuk data agrikultur yang kompleks, di mana kondisi geografis seringkali membentuk zona transisi yang batasannya tidak kaku.

Untuk menampilkan hasil klasterisasi secara intuitif, visualisasi dibuat dalam bentuk diagram sebar (boxplot) untuk setiap variabel, seperti yang disajikan pada Gambar 1. Visualisasi ini berfungsi sebagai alat bantu dalam mengenali tren dan karakteristik dari setiap klaster yang terbentuk.



Gambar 1. Visualisasi Karakteristik Variabel per Cluster

Gambar 1 menyajikan rangkuman visual dari hasil pengelompokan data selama periode analisis. Setiap diagram di dalamnya secara spesifik mengilustrasikan distribusi variabel untuk masing-masing klaster, sehingga memungkinkan pembaca untuk dengan mudah membandingkan pola sebaran dan memahami karakteristik utama dari setiap kelompok. Sebagai contoh, klaster 0 dapat memperlihatkan ciri produktivitas yang rendah namun dengan rentang luas panen yang sangat bervariasi. Sebaliknya, klaster 1 bisa jadi menunjukkan total produksi yang tinggi dengan sebaran data yang lebih terpusat, menandakan tingkat efisiensi yang lebih seragam di antara anggotanya.
Lebih lanjut, visualisasi ini tidak hanya memberikan gambaran statis, tetapi juga berpotensi untuk mengenali wilayah yang mungkin berpindah klaster dari waktu ke waktu melalui analisis temporal. Pergeseran tersebut dapat menjadi indikasi adanya implementasi kebijakan yang berhasil, adopsi teknologi pertanian yang efektif, atau pengaruh dari perubahan iklim. Di sisi lain, penurunan kinerja di suatu daerah juga dapat terdeteksi secara cepat, yang selanjutnya dapat menjadi dasar untuk merumuskan tindakan perbaikan.

Gambar 2. Pemetaan Wilayah Produksi Alpukat di Indonesia

Peta pada Gambar 2 menunjukkan persebaran spasial produksi alpukat di Indonesia yang dikelompokkan dalam tiga klaster berdasarkan tingkat produksinya: rendah (merah), sedang (hijau), dan tinggi (biru). Peta ini dihasilkan dengan mengaitkan setiap data produksi dengan koordinat geografis provinsi, menghasilkan visualisasi spasial yang informatif.

Wilayah dengan tingkat produksi tinggi, seperti yang ditandai dengan warna biru, tampak terkonsentrasi di Pulau Jawa dan sebagian Sumatera. Dominasi produksi di provinsi-provinsi ini mengindikasikan bahwa ekosistem pertanian di sana kemungkinan telah mapa baik dari segi infrastruktur irigasi, akses terhadap teknologi budidaya, ketersediaan tenaga kerja terampil, hingga dukungan kelembagaan yang kuat. Di sisi lain, klaster produksi rendah (merah) tersebar luas di berbagai wilayah, terutama di kawasan timur Indonesia. Hal ini merefleksikan tantangan yang masih membelenggu pengembangan sektor pertanian lokal, seperti minimnya infrastruktur dasar, logistik yang mahal, dan terbatasnya adopsi teknologi pertanian presisi.

Implikasi praktis dari pemetaan ini sangat luas. Pemerintah dapat menggunakan hasil klasterisasi untuk mengidentifikasi wilayah prioritas pembangunan, seperti alokasi bantuan benih unggul, perbaikan infrastruktur pertanian, atau program pelatihan petani. Bagi sektor swasta, peta ini dapat menjadi acuan untuk menyasar wilayah berpotensi tinggi dalam membangun fasilitas pengolahan, *cold storage*, atau mengembangkan kemitraan dengan petani lokal. Peta ini menjadi titik temu antara perencanaan berbasis data dan eksekusi kebijakan berbasis wilayah.

Untuk analisis kuantitatif yang lebih mendalam, kinerja kedua algoritma dievaluasi secara sistematis. Hasil eksperimen untuk algoritma Fuzzy C-Means disajikan pada Tabel 1.

Tabel 1. Hasil Eksperimen Algoritma Fuzzy C-Means

Jumlah Cluster	Rata – Rata Silhouette	Rata – Rata Davies-Bouldin Index	Waktu Komputasi
2	0.9218	0.0085	0.0042
3	0.3555	0.6892	0.0066
4	0.3385	0.8021	0.0159
5	0.3497	0.7016	0.0094
6	0.3111	0.5878	0.0164
7	0.3336	0.5820	0.0158
8	0.2708	0.6638	0.0144
9	0.2742	0.4769	0.0191
10	0.2450	0.3559	0.0313

Tabel 1 menyajikan luaran kuantitatif dari serangkaian eksperimen yang dirancang untuk melakukan asesmen performa terhadap algoritma Fuzzy C-Means (FCM). Dalam investigasi ini, jumlah klaster (k) divariasikan secara sistematis dari 2 hingga 10 untuk mengidentifikasi konfigurasi partisi yang paling optimal bagi dataset yang dianalisis. Analisis performa didasarkan pada dua metrik validasi internal utama, yaitu *Silhouette Coefficient* dan *Davies-Bouldin Index*, yang dilengkapi dengan evaluasi efisiensi temporal melalui pengukuran waktu komputasi. Temuan empiris yang tersaji secara kolektif memberikan bukti yang konvergen dan tak terbantahkan bahwa konfigurasi dengan jumlah klaster k=2 merupakan representasi struktur data yang paling valid dan superior.

Keunggulan absolut pada konfigurasi k=2 termanifestasi secara eksplisit melalui nilai *Silhouette Coefficient* (SC) yang mencapai 0.9218. Nilai yang sangat mendekati +1 ini merupakan indikator kuantitatif yang sangat kuat, menandakan bahwa struktur pengelompokan yang dihasilkan memiliki tingkat kohesi intra-klaster (kepadatan dalam klaster) yang tinggi dan tingkat separasi inter-klaster (keterpisahan antar klaster) yang maksimal. Dengan kata lain, algoritma berhasil membentuk dua kelompok data yang sangat padat dan terdefinisi dengan sangat baik. Temuan ini diperkuat lebih lanjut oleh hasil dari *Davies-Bouldin Index* (DBI), di mana nilai yang lebih rendah mencerminkan kualitas partisi yang lebih baik. Pada k=2, nilai DBI tercatat sangat rendah, yaitu 0.0085, yang secara sinergis mengonfirmasi bahwa konfigurasi dua klaster ini adalah yang paling optimal, dengan dispersi internal yang minimal dan jarak eksternal yang maksimal.

Sebaliknya, kontras yang tajam terlihat ketika jumlah klaster ditingkatkan di atas dua. Terjadi degradasi performa yang sangat signifikan pada k=3, di mana nilai SC anjlok ke angka 0.3555. Penurunan drastis ini mengisyaratkan bahwa pemaksaan data ke dalam tiga kelompok atau lebih justru menciptakan ambiguitas struktural dan tumpang-tindih (overlap) yang substansial. Algoritma kesulitan mendefinisikan batas-batas keanggotaan yang tegas, sebuah masalah yang terus berlanjut pada nilai k yang lebih tinggi. Hal ini selaras dengan tren pada nilai DBI, yang cenderung meningkat dan menunjukkan volatilitas pada k>2, menandakan ketidakstabilan dan ketidakmampuan algoritma untuk menemukan konfigurasi yang seimbang dan optimal. Fenomena ini menegaskan bahwa penambahan jumlah klaster melampaui struktur alaminya justru merusak kualitas partisi data.

Selain dari aspek kualitas partisi, evaluasi efisiensi temporal juga memberikan wawasan tambahan. Dari perspektif beban komputasi, terlihat adanya tren peningkatan latensi eksekusi yang berkorelasi positif dengan penambahan jumlah klaster, dari 0.0042 detik pada k=2 hingga 0.0313 detik pada k=10. Eskalasi waktu ini merupakan konsekuensi logis dari meningkatnya kompleksitas kalkulasi pada setiap iterasi, terutama dalam pembaruan matriks keanggotaan fuzzy yang dimensinya bergantung pada nilai k. Meskipun demikian, waktu eksekusi secara keseluruhan tetap berada pada orde milidetik, menunjukkan efisiensi inheren dari algoritma FCM untuk skala dataset ini.

Secara holistik, sintesis dari seluruh metrik evaluasi menghasilkan sebuah kesimpulan yang solid. Konvergensi antara nilai *Silhouette Coefficient* yang maksimal dan *Davies-Bouldin Index* yang minimal pada k=2 memberikan justifikasi yang sangat kuat untuk menetapkannya sebagai jumlah klaster yang paling merepresentasikan dataset penelitian. Kualitas partisi yang superior ini menjadikannya sebagai tolok ukur fundamental (*baseline*) untuk analisis komparatif dengan kinerja algoritma K-Means yang akan dibahas pada bagian selanjutnya.

Tabel 2. Hasil Eksperimen Algoritma K-Means

Jumlah Cluster	Rata – Rata Silhouette	Rata – Rata Davies-Bouldin Index	Waktu Komputasi
2	0.9218	0.0085	0.0150
3	0.3555	0.6892	0.0163
4	0.3385	0.8021	0.0282
5	0.1257	0.9634	0.0147

6	0.1352	0.8440	0.0209
7	0.2578	0.6442	0.0163
8	0.2714	0.3995	0.0074
9	0.3028	0.2448	0.0062
10	0.2450	0.3559	0.0086

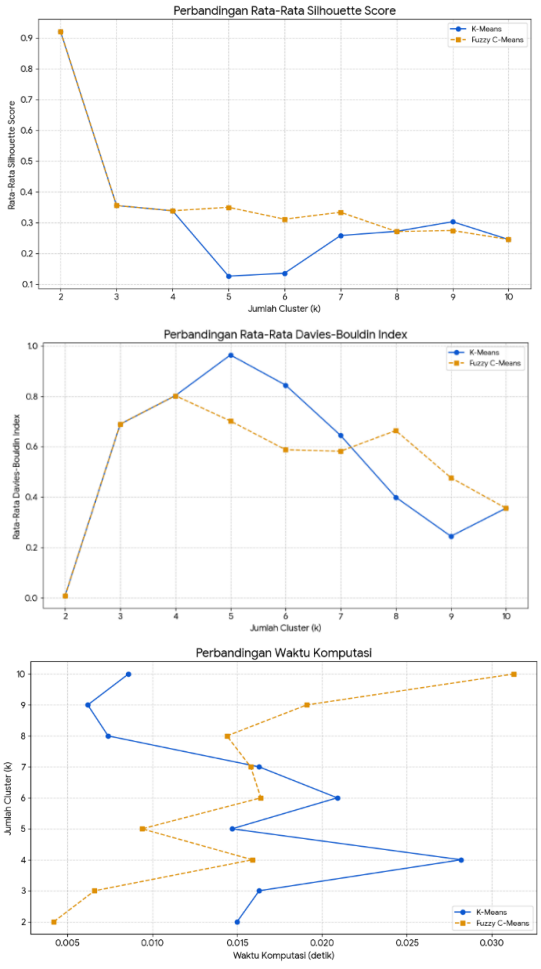
Melanjutkan evaluasi perbandingan, Tabel 2 menyajikan hasil asesmen performa untuk algoritma K-Means, yang diuji menggunakan parameter dan metrik validasi yang identik dengan Fuzzy C-Means. Analisis terhadap luaran K-Means tidak hanya bertujuan untuk mengidentifikasi jumlah klaster optimalnya secara mandiri, tetapi juga untuk melakukan perbandingan kapabilitasnya secara langsung dengan FCM. Temuan yang dihasilkan menunjukkan bahwa K-Means tidak hanya memberikan konfirmasi independen terhadap struktur data yang ada, tetapi juga mendemonstrasikan tingkat robustisitas dan efisiensi komputasi yang lebih superior pada dataset produksi alpukat ini.

Konsistensi antara kedua algoritma secara gamblang terlihat pada identifikasi jumlah klaster optimal. Serupa dengan FCM, K-Means mencapai performa puncaknya pada konfigurasi k=2, yang dibuktikan oleh skor *Silhouette Coefficient* (SC) yang identik dan superior, yaitu 0.9218. Perolehan nilai yang paralel ini semakin memperkuat hipotesis bahwa struktur alami yang paling inheren dan signifikan di dalam dataset ini adalah partisi biner atau pembagian menjadi dua kelompok besar. Dukungan lebih lanjut datang dari metrik *Davies-Bouldin Index* (DBI), di mana K-Means juga mencatatkan nilai terendah absolut pada k=2 sebesar 0.0085. Konvergensi kedua algoritma pada titik optimal yang sama ini memberikan tingkat keyakinan yang sangat tinggi terhadap validitas temuan bahwa dua adalah jumlah klaster yang paling merepresentasikan data.

Meskipun konvergen pada k=2, perbedaan fundamental antara K-Means dan FCM mulai tampak secara signifikan ketika mengevaluasi performa pada jumlah klaster yang lebih tinggi. K-Means mendemonstrasikan robustisitas yang lebih baik dalam menjaga kualitas partisi. Hal ini secara spesifik terungkap pada kemampuannya untuk mengidentifikasi titik optimal sekunder pada k=9, di mana ia berhasil mencapai nilai DBI yang sangat rendah, yaitu 0.2448. Penemuan struktur sekunder yang valid secara matematis ini merupakan sesuatu yang tidak terdeteksi oleh FCM, yang performanya cenderung tidak stabil dan terdegradasi pada jumlah klaster yang lebih banyak. Kemampuan K-Means untuk menemukan alternatif pengelompokan yang bermakna ini menunjukkan bahwa algoritma ini lebih andal dalam mengeksplorasi ruang solusi dan menjaga separabilitas klaster bahkan dalam konfigurasi yang lebih kompleks untuk dataset ini.

Keunggulan K-Means tidak hanya terbatas pada kualitas dan robustisitas partisi, tetapi juga meluas ke aspek efisiensi komputasi. Analisis waktu pemrosesan secara konsisten membuktikan bahwa K-Means secara signifikan lebih efisien secara temporal dibandingkan FCM. Waktu eksekusinya tercatat sangat rendah dan stabil di berbagai konfigurasi klaster. Sebagai contoh ilustratif, pada saat mengidentifikasi titik optimal sekundernya di k=9, K-Means hanya memerlukan waktu 0.0062 detik. Efisiensi ini berakar dari sifat deterministik proses iterasinya (hard assignment), yang secara inheren tidak memerlukan kalkulasi derajat keanggotaan fuzzy yang intensif secara komputasi. Efisiensi superior ini menjadikan K-Means sebagai kandidat algoritma yang jauh lebih prospektif untuk implementasi pada skenario yang melibatkan pemrosesan data bervolume besar (*big data*) atau aplikasi yang menuntut respons nyaris seketika (*near real-time*).

Secara keseluruhan, analisis komparatif ini menyimpulkan bahwa meskipun kedua algoritma berhasil mengidentifikasi struktur primer data pada k=2, K-Means menunjukkan keunggulan yang jelas dalam hal robustisitas model dan efisiensi sumber daya. Kemampuannya menemukan partisi sekunder yang valid serta kecepatan komputasinya yang superior mengindikasikan bahwa K-Means merupakan pilihan algoritma yang lebih pragmatis dan andal untuk diterapkan pada domain masalah data produksi alpukat ini.



Gambar 3. Analisis Visual Perbandingan K-Means, Fuzzy C-Means, dan Waktu Komputasi

Gambar 3 menyajikan representasi visual dari perbandingan performa kedua algoritma. Grafik pertama dan kedua dengan jelas menunjukkan bahwa nilai Silhouette Score dan Davies-Bouldin Index untuk kedua metode mencapai titik paling optimal pada jumlah cluster 2. Grafik ini secara visual mengkonfirmasi temuan dari tabel.

Grafik ketiga secara gamblang memperlihatkan perbedaan efisiensi. K-Means (garis biru) menunjukkan waktu komputasi yang lebih rendah dan lebih stabil di hampir semua konfigurasi klaster dibandingkan dengan Fuzzy C-Means (garis oranye). Perbedaan ini menjadi signifikan terutama pada jumlah klaster yang lebih tinggi.

Secara keseluruhan, kombinasi visualisasi dan analisis metrik menegaskan bahwa meskipun kedua algoritma berhasil mengidentifikasi struktur data optimal pada 2 cluster, K-Means memberikan hasil yang lebih stabil (terutama pada metrik DBI) dan secara signifikan lebih efisien untuk data pertanian alpukat ini. Oleh karena itu, untuk aplikasi seperti segmentasi kualitas alpukat atau pemetaan wilayah panen, algoritma K-Means dengan 2 cluster sangat direkomendasikan karena menghasilkan kualitas cluster terbaik dengan efisiensi waktu proses yang unggul.

4. Kesimpulan

Berdasarkan analisis dan hasil eksperimen perbandingan algoritma K-Means dan Fuzzy C-Means (FCM) untuk klasterisasi data produksi alpukat di Indonesia, dapat ditarik kesimpulan sebagai berikut:

1. Jumlah Klaster Optimal: Kedua algoritma, K-Means dan Fuzzy C-Means, secara konvergen menunjukkan bahwa jumlah klaster yang paling optimal untuk merepresentasikan struktur data produksi alpukat adalah 2 klaster. Pada konfigurasi ini, kedua metode berhasil mencapai nilai *Silhouette Score* tertinggi (0.9218) dan *Davies-Bouldin Index* terendah (0.0085), yang mengindikasikan kualitas partisi terbaik.
2. Performa Algoritma K-Means:
   - Kelebihan: K-Means unggul secara signifikan dalam efisiensi waktu komputasi pada hampir semua skenario jumlah klaster. Algoritma ini juga menunjukkan robustisitas yang lebih baik, di mana ia mampu menjaga kualitas partisi dan bahkan mengidentifikasi struktur sekunder yang valid pada k=9, sesuatu yang tidak terdeteksi oleh FCM.
   - Kekurangan: Secara teoritis, K-Means memiliki keterbatasan dalam menangani klaster dengan bentuk non-sferis dan sensitif terhadap inisialisasi *centroid* awal.
3. Performa Algoritma Fuzzy C-Means (FCM):
   - Kelebihan: Keunggulan utama FCM terletak pada fleksibilitasnya sebagai metode *soft clustering*, yang secara teoretis lebih mampu memodelkan data dengan batas-batas yang tidak tegas atau tumpang tindih, seperti pada data agrikultur.
   - Kekurangan: FCM membutuhkan waktu komputasi yang lebih tinggi dibandingkan K-Means. Selain itu, kinerjanya mengalami degradasi yang tajam ketika jumlah klaster ditingkatkan di atas dua, menunjukkan ketidakstabilan pada dataset ini.
4. Rekomendasi dan Pengembangan Selanjutnya:
   - Untuk aplikasi klasterisasi data produksi alpukat ini, algoritma K-Means dengan 2 klaster direkomendasikan karena menghasilkan kualitas klaster terbaik dengan efisiensi waktu proses yang jauh lebih unggul.
   - Pengembangan selanjutnya dapat mengeksplorasi penggunaan metrik jarak yang lebih canggih seperti *Dynamic Time Warping* (DTW) jika data dianalisis sebagai deret waktu (*time-series*). Selain itu, penelitian dapat diperkaya dengan menambahkan variabel lain seperti data iklim atau karakteristik lahan untuk analisis yang lebih komprehensif.

REFERENSI

[1] H. Santoso dan R. N. Putri, "Analisis Prospek dan Posisi Strategis Komoditas Alpukat dalam Pembangunan Agribisnis Nasional," *Jurnal Ekonomi Pertanian dan Agribisnis*, vol. 7, no. 1, hlm. 120–135, 2023.

[2] D. Setiawan dan A. Pramudita, "Analisis Potensi dan Strategi Pengembangan Agribisnis Alpukat (Persea americana Mill.) untuk Pasar Ekspor di Indonesia," *Jurnal Agribisnis Terpadu*, vol. 16, no. 2, hlm. 112–125, 2023.

[3] FAOSTAT, "Global Avocado Market Trends and Outlook," 2023.

[4] Kementerian Pertanian Republik Indonesia, "Basis Data Statistik Pertanian (BDSP): Produksi Hortikultura - Alpukat 2023," Jakarta, 2024.

[5] A. Susanto, "Pola Spasial Produksi dan Konsumsi Hortikultura Unggulan di Indonesia," *Forum Penelitian Agro Ekonomi*, vol. 41, no. 1, hlm. 34–48, 2023, doi: 10.21082/fae.v41n1.2023.34-48.

[6] S. Pratama dan D. K. Sari, "Pentingnya Pra-pemrosesan Data dan Seleksi Fitur untuk Akurasi Model Machine Learning dalam Analitik Pertanian," *Jurnal Informatika Pertanian*, vol. 6, no. 2, hlm. 88–101, 2023.

[7] S. Kumar dan R. Singh, "Big Data Challenges and Machine Learning Applications in Agriculture: A

Comprehensive Review," *Journal of Agricultural Informatics*, vol. 14, no. 1, hlm. 1–15, 2023.

[8] S. Russell dan P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2023.

[9] J. Han, M. Kamber, dan J. Pei, *Data Mining: Concepts and Techniques*, 4th ed. Morgan Kaufmann, 2023.

[10] M. Ibrahim dan A. Wibowo, "Clustering of Coffee Plantations in Indonesia using K-Means Algorithm for Supply Chain Optimization," *Indonesian Journal of Computer Science*, vol. 12, no. 1, hlm. 55–64, 2024.

[11] A. S. Ahmar, D. Abdullah, U. Habibah, Y. E. P. Sari, N. A. Sani, dan R. F. Ali, "Advanced Clustering Approach for Mapping Regions of Paddy Productivity in Indonesia Using Intelligent K-Means," dalam *2024 International Conference on Information System and Technology (ICoIST)*, 2024, hlm. 1–6. doi: 10.1109/ICoIST61942.2024.10932942.

[12] F. Abdullah dan M. Z. Lubis, "Pemetaan Zona Potensi Komoditas Karet Menggunakan Algoritma K-Means Berdasarkan Karakteristik Lahan," *Jurnal Geodesi dan Geomatika*, vol. 21, no. 1, hlm. 34–45, 2024.

[13] T. A. Tuan, T. M. H. Nguyen, dan V. T. Ho, "An adaptive fuzzy c-means clustering algorithm for noisy data in agricultural management," *Engineering in Agriculture, Environment and Food*, vol. 17, no. 2, hlm. 100–108, 2024, doi: 10.1016/j.eaef.2024.01.005.

[14] N. H. Wijaya, D. E. Cahyani, dan S. W. Prasetyo, "Analisis Komparasi K-Means dan Fuzzy C-Means untuk Klasifikasi Jenis Tanah Menggunakan Data Sensor Multispektral," *Jurnal Sains Data*, vol. 5, no. 1, hlm. 22–31, 2024.

[15] R. Setiadi, I. G. P. Astawa, dan N. K. A. Werthi, "Identifikasi Pola Kesesuaian Lahan Tanaman Pangan Menggunakan Fuzzy C-Means untuk Mitigasi Risiko Iklim," *Agro-Industrial Informatics Journal*, vol. 3, no. 1, hlm. 12–23, 2024.

[16] H. Agustin, I. Parlina, dan M. A. Bijaksana, "Clustering Data Meteorologi Wilayah Indonesia Timur Dengan Metode K-Means dan Fuzzy C-Means," *Jurnal INTI (Informasi dan Teknologi)*, vol. 2, no. 2, hlm. 83–91, 2022.

[17] Y. Liu, Z. Li, dan C. Xiong, "Hard Clustering versus Soft Clustering: A Comparative Review for Practitioners," *IEEE Access*, vol. 11, hlm. 55670–55685, 2023.

[18] A. Nugroho dan F. P. Sari, "Metrik Evaluasi Kinerja untuk Algoritma Clustering: Studi Komparatif pada Data Pertanian," *Cyberspace: Jurnal Pendidikan Teknologi Informasi*, vol. 7, no. 2, hlm. 150–162, 2023.

[19] I. S. Dhillon dan D. S. Modha, "A Review of Partitioning-Based Clustering Algorithms for Modern Applications," *Knowl Inf Syst*, vol. 65, no. 1, hlm. 1–25, 2023.

[20] A. P. Liaw dan M. Wiener, "On the Importance of Distance Metrics in Clustering-Based Machine Learning," *Journal of Machine Learning Research*, vol. 25, hlm. 1–30, 2024.

[21] C. F. M. de Souza, "Revisiting Euclidean and Other Distance Metrics for Numerical Data Clustering," *ACM Trans Knowl Discov Data*, vol. 18, no. 2, hlm. 1–22, 2024.

[22] R. A. S. Hidayat, T. Tulus, dan S. Suwilo, "K-Means Using Dynamic Time Warping For Clustering Cities in Java Island According to Meteorological Conditions," dalam *2023 3rd International Conference on Information Technology and Education (ICIT-E)*, 2023, hlm. 1–5. doi: 10.1109/ICIT-E60233.2023.10381899.

[23] A. Rahman dan L. Hakim, "Evaluasi Metode Penentuan Jumlah Cluster Optimal (K) pada Algoritma K-Means untuk Segmentasi Data Agroklimat," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 4, hlm. 701–710, 2023.

[24] A. Wijaya dan D. E. Putra, "Algoritma Fuzzy C-Means dan Aplikasinya dalam Analisis Data Modern: Sebuah Tinjauan Komprehensif," *Jurnal Ilmu Komputer dan Inovasi (JIKI)*, vol. 7, no. 1, hlm. 45–60, 2023.

[25] H. N. Pal dan S. K. Pal, "A Contemporary Review on the Role of the Fuzzifier (m) in Fuzzy C-Means," *Fuzzy Information and Engineering*, vol. 15, no. 1, hlm. 1–19, 2023.