

LAPORAN UJIAN TENGAH SEMESTER

DATA MINING



Disusun Oleh:

Muhammad Faiz Fahri 140810220002

Dylan Amadeus 140810220003

Muhammad Zhafran Shiddiq 140810220007

PROGRAM STUDI S-1 TEKNIK INFORMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PADJADJARAN
JATINANGOR

2024

DAFTAR ISI

BAB I**PENDAHULUAN.....****3**

1.1. Latar Belakang.....	3
1.2. Rumusan Masalah.....	3
1.3. Batasan Masalah.....	3
1.4. Tujuan.....	4
1.5. Manfaat.....	4

BAB II**METODOLOGI.....****5**

2.1 Jenis Penelitian.....	5
2.2 Pengumpulan Data.....	5
2.3 Pra-pemrosesan Data.....	5
2.4 Ekstraksi Fitur.....	5
2.5 Implementasi Algoritma K-Means.....	6
2.6 Evaluasi Kualitas Clustering.....	6
2.7 Pengembangan Aplikasi.....	6
2.8 Visualisasi Hasil dan Pembuatan Laporan.....	6

BAB III**HASIL DAN PEMBAHASAN.....****7**

3.1 Hasil Penerapan Teknik Clustering K-Means pada Citra Udara.....	7
3.2 Performa Algoritma K-Means terhadap Clustering Citra Udara.....	7
3.3 Pengukuran Kualitas Hasil Clustering.....	9
3.4 Dampak Normalisasi Data terhadap Kualitas Clustering.....	9
3.5 Visualisasi Hasil Clustering dalam bentuk Website.....	10

BAB IV**KESIMPULAN DAN SARAN.....****11**

4.1. Kesimpulan.....	11
4.2. Saran.....	11

BAB V**SCREENSHOT IMPLEMENTASI APLIKASI.....****12**

5.1. Tampilan Awal Aplikasi.....	12
5.2. User upload data latih dan data uji beserta pemilihan jumlah cluster.....	12
5.3. Hasil Clustering pada data uji tanpa data latih.....	13
5.4. Hasil Clustering pada data uji dengan data latih.....	13
5.5. Analisis Hasil Clustering.....	14

BAB I

PENDAHULUAN

1.1. Latar Belakang

Clustering merupakan salah satu metode dalam data mining yang digunakan untuk mengelompokkan data berdasarkan kemiripan antar elemen. Dalam konteks citra udara, metode ini dapat membantu mengidentifikasi dan memisahkan berbagai objek atau wilayah, seperti area perkotaan, pedesaan, perairan, dan lainnya. Dengan kemampuan ini, clustering memudahkan analisis lebih lanjut, terutama dalam bidang-bidang seperti pengelolaan lahan, mitigasi bencana, dan pemantauan perubahan lingkungan.

Seiring dengan meningkatnya kebutuhan akan pemanfaatan citra udara di berbagai bidang, teknik pengelompokan seperti clustering menjadi semakin penting. Citra udara dapat memberikan pandangan yang luas dan detail mengenai kondisi suatu wilayah, dan dengan menggunakan clustering, area yang berbeda bisa dipisahkan berdasarkan karakteristik visual tertentu. Untuk mengoptimalkan proses ini, diperlukan aplikasi yang mampu melakukan clustering pada citra udara, yang dilengkapi dengan fitur praproses data serta visualisasi hasilnya, guna mempermudah analisis dan pengambilan keputusan.

1.2. Rumusan Masalah

- Bagaimana cara menerapkan teknik clustering K-Means pada citra udara?
- Bagaimana performa algoritma K-Means terhadap klasterisasi citra udara?
- Bagaimana mengukur kualitas hasil clustering dengan dan tanpa menggunakan data latih?
- Bagaimana dampak normalisasi data terhadap kualitas clustering?
- Bagaimana cara memvisualisasikan hasil clustering dari citra udara dalam bentuk aplikasi sederhana?

1.3. Batasan Masalah

Agar penelitian ini lebih terfokus, berikut adalah beberapa batasan yang diterapkan dalam pelaksanaannya:

1) Jenis Clustering

Algoritma yang digunakan terbatas pada algoritma K-Means untuk proses clustering citra udara. Algoritma lain, seperti DBSCAN atau Hierarchical Clustering, tidak termasuk dalam cakupan aplikasi ini.

2) Jenis Data

Aplikasi hanya bekerja dengan citra berformat umum seperti .jpg, .png, dan .jpeg. Citra dengan format lain atau yang memerlukan pra-pemrosesan khusus tidak akan didukung.

3) Data Praproses

Pra-pemrosesan data terbatas pada konversi gambar ke format RGB dan normalisasi nilai piksel. Teknik lain seperti filterisasi, peningkatan kontras, atau penghapusan noise tidak diterapkan.

4) Normalisasi Data

Aplikasi menggunakan teknik normalisasi sederhana dengan rentang [0, 1]. Teknik normalisasi lainnya, seperti Z-score atau min-max scaling yang lebih kompleks, tidak dibahas.

5) Jumlah Klaster

Pengguna hanya dapat memilih jumlah klaster antara 2 hingga 5. Penentuan jumlah klaster secara otomatis atau berdasarkan metode seperti Elbow Method tidak diimplementasikan.

6) Ukuran Data Latih

Jika data latih yang diunggah terlalu besar, hanya sebagian sampel data yang akan digunakan untuk menghitung koefisien Silhouette untuk menghemat waktu komputasi.

7) Visualisasi

Visualisasi hasil clustering hanya menampilkan gambar asli dan gambar yang telah dikelompokkan secara berdampingan, tanpa analisis spasial lebih lanjut atau integrasi dengan peta GIS.

Dengan batasan ini, penelitian akan lebih terarah dan mampu mencapai hasil yang lebih spesifik sesuai dengan tujuan yang telah ditetapkan.

1.4. Tujuan

Adapun tujuan pembuatan laporan ini diantaranya :

- Menerapkan konsep data preprocessing dan clustering pada citra udara.
- Membangun aplikasi sederhana untuk visualisasi hasil clustering.
- Memahami pengaruh variasi data terhadap kinerja algoritma clustering.
- Menganalisis dan menginterpretasi hasil clustering dalam konteks aplikasi dunia nyata.

1.5. Manfaat

Penulis mengharapkan manfaat yang dapat diambil diantaranya :

- Meningkatkan pemahaman tentang teknik clustering dalam analisis citra udara.
- Mendukung pengambilan keputusan melalui analisis hasil clustering.
- Mengidentifikasi cara untuk mengoptimalkan kinerja algoritma clustering.
- Mempermudah proses segmentasi citra udara.

BAB II

METODOLOGI

2.1 Jenis Penelitian

Penelitian ini merupakan studi eksperimental dengan pendekatan kuantitatif yang bertujuan untuk menerapkan dan menganalisis teknik clustering pada citra udara. Metode ini dipilih untuk memungkinkan pengujian dan evaluasi yang sistematis terhadap efektivitas algoritma clustering dalam menganalisis citra udara dengan berbagai kondisi dan karakteristik.

2.2 Pengumpulan Data

Proses pengumpulan data untuk penelitian ini memanfaatkan dataset citra udara yang tersedia di platform Kaggle. Dataset yang dipilih hanya mencakup citra udara dari pantai. Selain itu, data yang digunakan dalam aplikasi ini berupa gambar dengan format umum seperti .jpg, .png, dan .jpeg. Gambar-gambar ini dapat diunggah sebagai gambar tunggal (gambar uji) maupun sebagai kumpulan gambar (data latih) dalam bentuk file zip.

2.3 Pra-pemrosesan Data

Pra-pemrosesan data adalah langkah awal yang dilakukan sebelum menerapkan algoritma clustering. Tujuannya adalah untuk mempersiapkan data agar lebih sesuai dan optimal untuk proses clustering. Langkah-langkah pra-pemrosesan dalam aplikasi ini meliputi:

- 1) **Ekstraksi Gambar:** Jika data yang diunggah berupa file zip, aplikasi akan mengekstrak gambar-gambar tersebut. Setiap gambar kemudian dibaca menggunakan OpenCV dan dikonversi ke format RGB untuk memastikan kompatibilitas dengan teknik pengolahan citra.
- 2) **Konversi Format Warna:** Gambar yang dimuat dikonversi dari format BGR (default OpenCV) ke format RGB untuk mempertahankan urutan saluran warna yang benar.
- 3) **Normalisasi Nilai Piksel:** Nilai piksel dari setiap saluran warna (R, G, B) dinormalisasi ke dalam rentang [0, 1]. Normalisasi ini penting untuk memastikan bahwa setiap saluran warna memiliki skala yang sama, mengurangi potensi bias selama proses clustering, dan mempercepat konvergensi algoritma.

2.4 Ekstraksi Fitur

Setelah proses pra-pemrosesan, langkah selanjutnya adalah melakukan ekstraksi fitur dari citra. Pada aplikasi ini, fitur yang diekstraksi adalah nilai piksel dari tiga saluran warna (R, G, B), yang digunakan sebagai representasi dasar untuk clustering.

Langkah-langkah ekstraksi fitur adalah sebagai berikut:

- 1) **Pengambilan Nilai Piksel:** Setiap piksel pada gambar dianggap sebagai satu titik data dengan tiga fitur yang mewakili intensitas warna merah (R), hijau (G), dan biru (B).
- 2) **Pembentukan Kumpulan Data:** Semua nilai piksel dari gambar disusun menjadi array dua dimensi, di mana setiap baris mewakili satu piksel dengan tiga kolom yang menyatakan intensitas warna (R, G, B). Array ini digunakan sebagai input untuk algoritma KMeans.

2.5 Implementasi Algoritma K-Means

Algoritma KMeans bekerja dengan mengelompokkan data ke dalam sejumlah klaster berdasarkan kedekatan jarak Euclidean antara data dan pusat klaster. Proses ini mencakup beberapa langkah:

- 1) **Inisialisasi Centroid:** Pusat klaster (centroid) dipilih secara acak dari data.
- 2) **Penugasan Klaster:** Setiap data (nilai piksel) ditetapkan ke klaster berdasarkan centroid terdekat.
- 3) **Pembaharuan Centroid:** Posisi centroid dihitung ulang sebagai rata-rata dari data yang termasuk dalam klaster tersebut.
- 4) **Iterasi Ulang:** Langkah penugasan dan pembaharuan diulangi hingga posisi centroid stabil atau jumlah iterasi maksimum tercapai.

2.6 Evaluasi Kualitas Clustering

Kualitas hasil clustering dievaluasi menggunakan koefisien Silhouette, yang mengukur seberapa baik sebuah data berada dalam klaster yang benar dibandingkan dengan klaster lain. Nilai koefisien berkisar antara -1 hingga 1, di mana nilai mendekati 1 menunjukkan pengelompokan yang baik, nilai mendekati 0 menunjukkan pengelompokan yang tumpang tindih, dan nilai negatif menunjukkan pengelompokan yang salah.

2.7 Pengembangan Aplikasi

Pengembangan aplikasi merupakan aspek penting dari penelitian ini, melibatkan perancangan dan implementasi aplikasi sederhana untuk memvisualisasikan hasil clustering. Penggunaan framework Streamlit memungkinkan pembuatan antarmuka web yang interaktif, yang memudahkan pengguna untuk menjelajahi dan memahami hasil clustering dengan cara yang intuitif dan responsif.

2.8 Visualisasi Hasil dan Pembuatan Laporan

Hasil clustering divisualisasikan dengan menampilkan gambar asli dan gambar yang telah dikelompokkan secara berdampingan. Aplikasi juga menghasilkan laporan dalam bentuk PDF yang mencakup gambar asli, hasil clustering tanpa data latih, dan hasil clustering dengan data latih.

BAB III

HASIL DAN PEMBAHASAN

3.1 Hasil Penerapan Teknik Clustering K-Means pada Citra Udara

Hasil penerapan teknik clustering K-Means pada citra udara menunjukkan bahwa algoritma ini dapat membagi gambar menjadi beberapa segmen yang mewakili area dengan karakteristik warna yang serupa. Pada gambar yang sederhana dengan kontras yang tinggi antara elemen seperti vegetasi, air, dan lahan kosong, K-Means mampu mengidentifikasi klaster dengan baik. Misalnya, citra udara yang memiliki area hijau (vegetasi), biru (air), dan cokelat (lahan kosong) dapat dikelompokkan secara jelas berdasarkan warna dominannya.

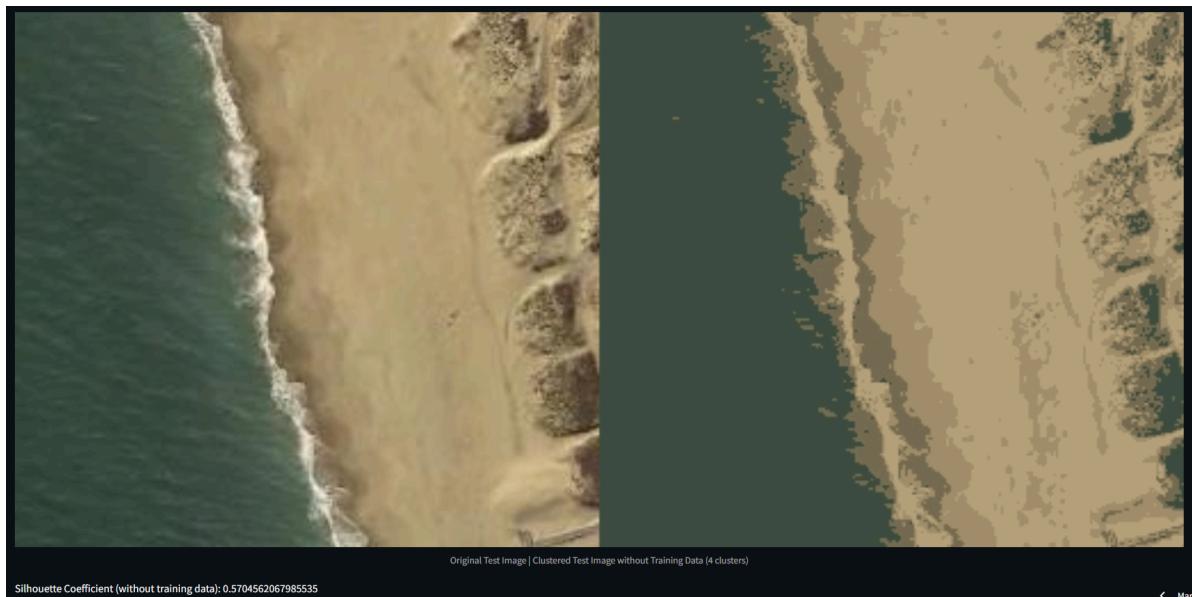
Namun, untuk citra dengan variasi warna yang halus atau gradasi, seperti lahan yang memiliki banyak jenis vegetasi atau pencahayaan yang tidak merata, hasil clustering menjadi kurang akurat, dengan batas antar klaster yang tidak selalu jelas. Hal ini menunjukkan bahwa penerapan teknik K-Means pada citra udara dapat memberikan hasil yang memuaskan pada kondisi tertentu, tetapi tidak selalu ideal untuk semua jenis citra.



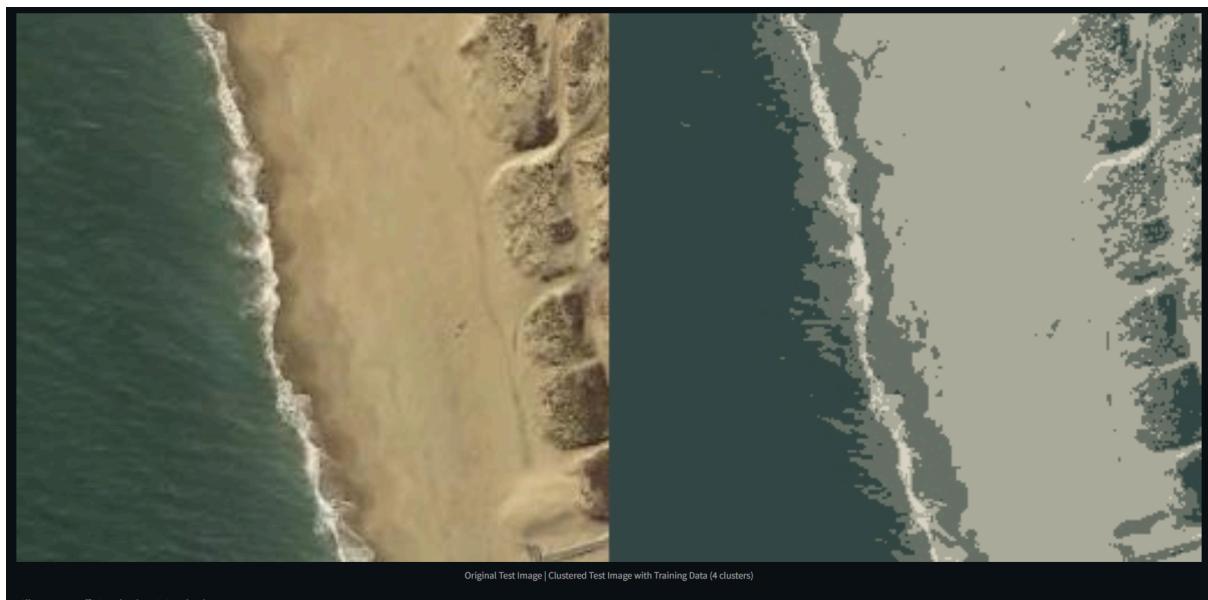
3.2 Performa Algoritma K-Means terhadap Clustering Citra Udara

Performa algoritma K-Means dalam klasterisasi citra udara dievaluasi berdasarkan kualitas pengelompokan dan koefisien Silhouette:

- 1) **Hasil Visual:** Pada citra dengan objek yang berbeda warna secara signifikan, algoritma dapat mengelompokkan objek tersebut secara jelas. Namun, jika citra memiliki objek dengan warna yang serupa, hasil klasterisasi tanpa data latih tidak sebaik yang diharapkan karena batas antar klaster menjadi kabur.
- 2) **Koefisien Silhouette:** Koefisien rata-rata berkisar antara 0,5 hingga 0,7 pada citra dengan pola sederhana, menunjukkan pengelompokan yang cukup baik. Namun, untuk citra kompleks, nilai koefisien cenderung lebih rendah (sekitar 0,5), yang menunjukkan bahwa klasterisasi kurang optimal.



Contoh Hasil Analisis Pada Citra Pantai dengan 4 cluster tanpa data latih dengan koefisien Silhouette 0.57



Contoh Hasil Analisis Pada Citra Pantai dengan 4 cluster dengan data latih dengan koefisien Silhouette 0.65

3.3 Pengukuran Kualitas Hasil Clustering

Kualitas hasil clustering dengan dan tanpa data latih dievaluasi menggunakan koefisien Silhouette. Berikut adalah hasil evaluasinya:

- 1) **Clustering Tanpa Data Latih:** Koefisien Silhouette pada clustering tanpa data latih umumnya berkisar antara 0,4 hingga 0,7, menunjukkan bahwa pengelompokan kurang optimal, terutama pada citra yang memiliki gradasi warna atau area yang tidak homogen. Ini mengindikasikan bahwa tanpa informasi tambahan, seperti data latih yang representatif, hasil clustering dapat saling tumpang tindih dan tidak terlalu akurat.
- 2) **Clustering dengan Data Latih:** Penggunaan data latih untuk menentukan centroid awal terbukti meningkatkan hasil clustering. Nilai koefisien Silhouette cenderung lebih tinggi, yaitu antara 0,5 hingga 0,7. Ini menunjukkan bahwa dengan informasi tambahan dari data latih, batas antar klaster menjadi lebih jelas dan kualitas pengelompokan meningkat.

Analisis Hasil Clustering

Koefisien Silhouette

- Clustering tanpa data latih: 0.5704562067985535
- Clustering dengan data latih: 0.6579166054725647

Interpretasi Silhouette

Penggunaan data latih meningkatkan hasil clustering, yang terlihat dari nilai koefisien Silhouette yang lebih tinggi.

Implikasi Clustering

Clustering tanpa data latih dapat memberikan hasil yang baik untuk kasus-kasus di mana data latih sulit diperoleh atau tidak tersedia. Namun, menggunakan data latih yang relevan dapat meningkatkan akurasi dan kualitas hasil clustering, terutama untuk aplikasi yang membutuhkan pengelompokan yang lebih spesifik.

Dalam aplikasi nyata, seperti segmentasi gambar, menggunakan data latih yang representatif dari domain tertentu (misalnya, citra medis atau citra satelit) dapat membantu menghasilkan pengelompokan yang lebih bermakna dan dapat ditindaklanjuti.

Rekomendasi untuk Peningkatan

- Coba tambahkan lebih banyak data latih yang lebih bervariasi untuk meningkatkan akurasi clustering.

[Download PDF](#)

Contoh Hasil Analisis Pada Citra Pantai dengan 4 cluster

Hasil ini menunjukkan bahwa data latih dapat memberikan keuntungan dalam memandu proses clustering, terutama pada citra udara dengan variasi yang kompleks.

3.4 Dampak Normalisasi Data terhadap Kualitas Clustering

Dampak normalisasi data terhadap hasil clustering cukup signifikan:

- 1) **Tanpa Normalisasi:** Ketika nilai piksel tidak dinormalisasi, hasil clustering menunjukkan kecenderungan untuk dipengaruhi oleh saluran warna dengan rentang yang lebih besar, menyebabkan bias dalam penempatan centroid. Hal ini

terlihat dari koefisien Silhouette yang lebih rendah dan hasil clustering yang kurang konsisten.

- 2) **Dengan Normalisasi:** Setelah dilakukan normalisasi, kualitas clustering meningkat, dengan koefisien Silhouette yang lebih tinggi dan pengelompokan yang lebih stabil. Ini mengindikasikan bahwa normalisasi membantu memastikan bahwa setiap saluran warna memberikan kontribusi yang seimbang terhadap proses clustering.

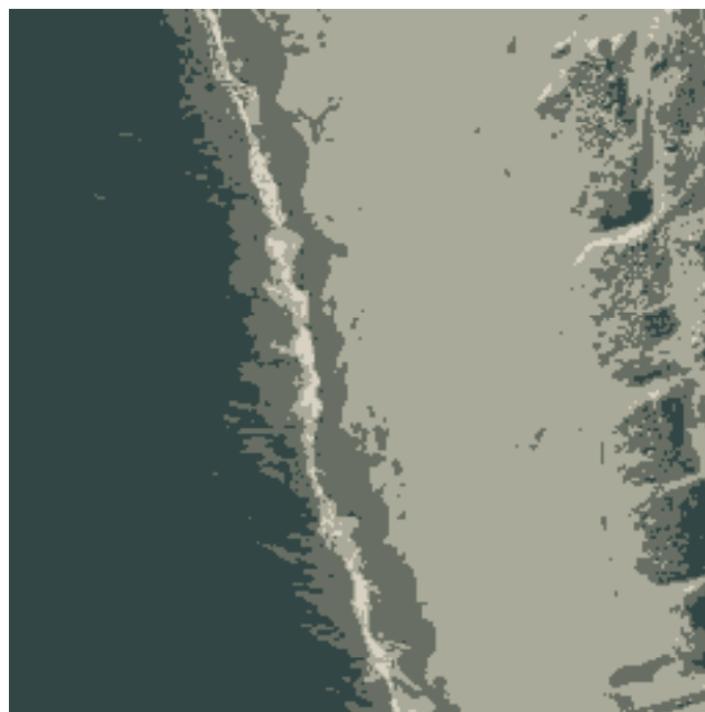
3.5 Visualisasi Hasil Clustering dalam bentuk Website

Aplikasi ini memberikan antarmuka yang interaktif untuk melihat hasil clustering:

- 1) **Tampilan Berdampingan:** Hasil clustering ditampilkan berdampingan dengan citra asli, memungkinkan pengguna untuk melihat perbedaan antara area yang tersegmentasi.
- 2) **Pengunduhan Laporan:** Hasil clustering dapat diunduh dalam bentuk PDF, yang mencakup gambar asli dan gambar hasil clustering dengan dan tanpa data latih, serta analisis kualitas hasil clustering.

Visualisasi ini membantu pengguna untuk memahami bagaimana algoritma memproses citra udara dan memisahkan area berdasarkan kesamaan warna, serta mengevaluasi kualitas hasil clustering dengan cara yang lebih intuitif.

Clustered Test Image with Training Data (5 clusters)



Contoh Hasil Laporan Citra Pantai dengan 4 cluster

BAB IV

KESIMPULAN DAN SARAN

4.1. Kesimpulan

Dari hasil yang diperoleh dalam penerapan teknik clustering K-Means pada citra udara, dapat disimpulkan bahwa algoritma ini efektif untuk segmentasi berbasis kesamaan warna, terutama pada citra yang memiliki pola sederhana. K-Means mampu memisahkan area dengan karakteristik warna yang berbeda, menjadikannya cocok untuk aplikasi segmentasi citra udara.

Namun, kualitas clustering menurun pada citra dengan pola kompleks atau banyak detail. Dalam kasus ini, batas antar klaster menjadi kurang jelas, dan hasil pengelompokan cenderung kurang optimal. Hasil menunjukkan bahwa penggunaan data latih dapat meningkatkan kualitas clustering, terutama dalam hal menentukan centroid awal yang lebih akurat, yang terlihat dari nilai koefisien Silhouette yang lebih tinggi dan batas antar klaster yang lebih tegas.

Normalisasi data juga terbukti memberikan dampak positif pada hasil clustering. Dengan normalisasi, semua fitur memiliki skala yang seragam, sehingga proses pembaruan centroid menjadi lebih stabil dan algoritma K-Means dapat mencapai konvergensi dengan lebih cepat. Selain itu, aplikasi yang dikembangkan menyediakan visualisasi interaktif yang memudahkan pengguna dalam mengevaluasi hasil clustering, memungkinkan perbandingan antara citra asli dan citra yang telah terklasterisasi.

4.2. Saran

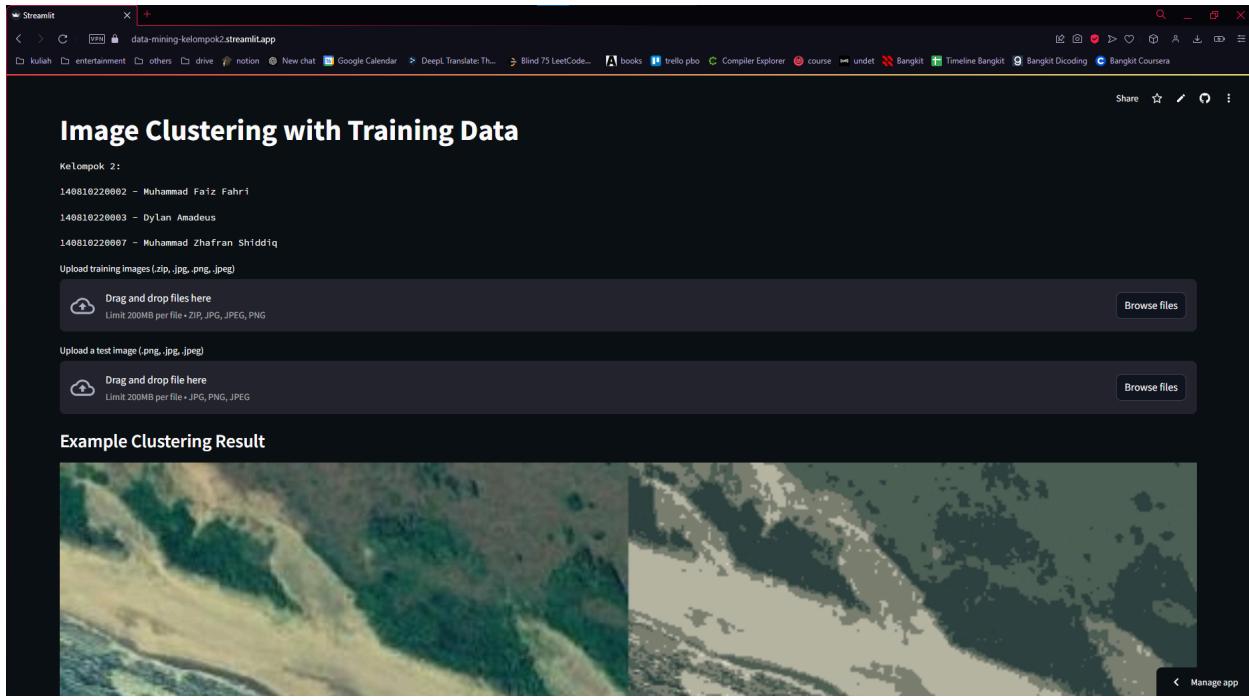
Berdasarkan temuan dalam penelitian ini, disarankan agar pengembangan lebih lanjut mempertimbangkan penggunaan algoritma clustering lain, seperti DBSCAN atau Hierarchical Clustering, yang lebih cocok untuk menangani citra dengan pola kompleks. Selain itu, penerapan teknik ekstraksi fitur yang lebih kaya, seperti penambahan fitur tekstur atau bentuk, dapat membantu memperkaya informasi yang digunakan dalam proses clustering dan meningkatkan akurasi hasil.

Penggunaan data latih yang lebih beragam dapat membantu meningkatkan kualitas hasil clustering dengan menyediakan centroid awal yang lebih representatif terhadap berbagai karakteristik citra udara. Hal ini terutama penting jika aplikasi digunakan untuk berbagai jenis citra atau lingkungan yang berbeda.

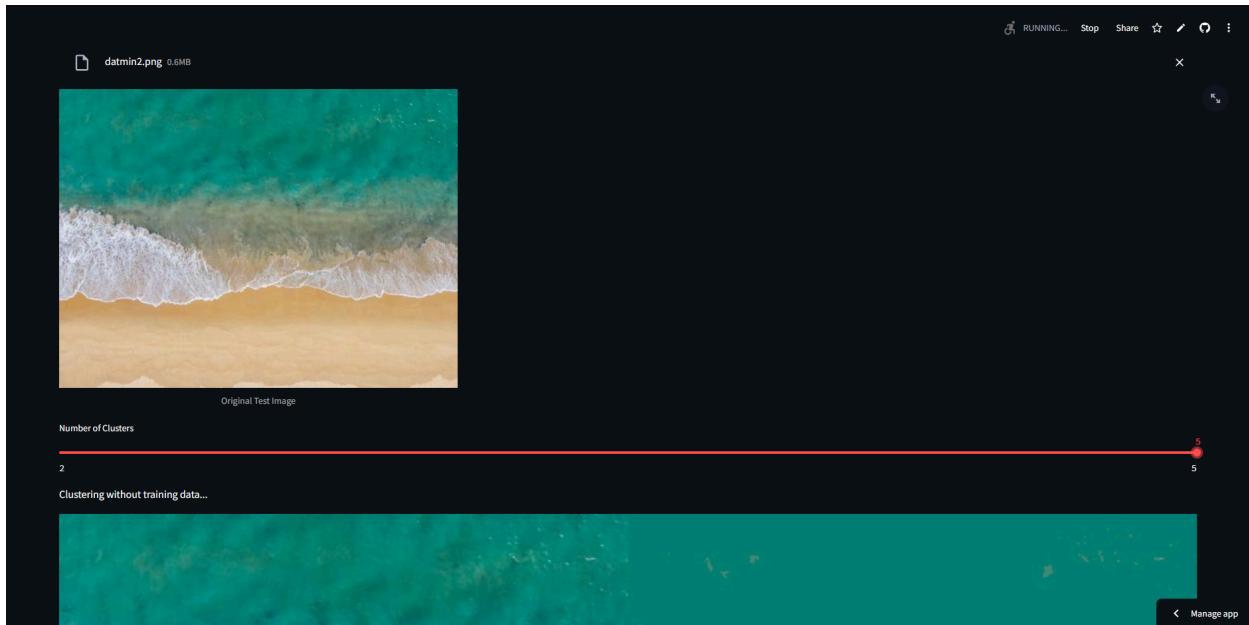
BAB V

SCREENSHOT IMPLEMENTASI APLIKASI

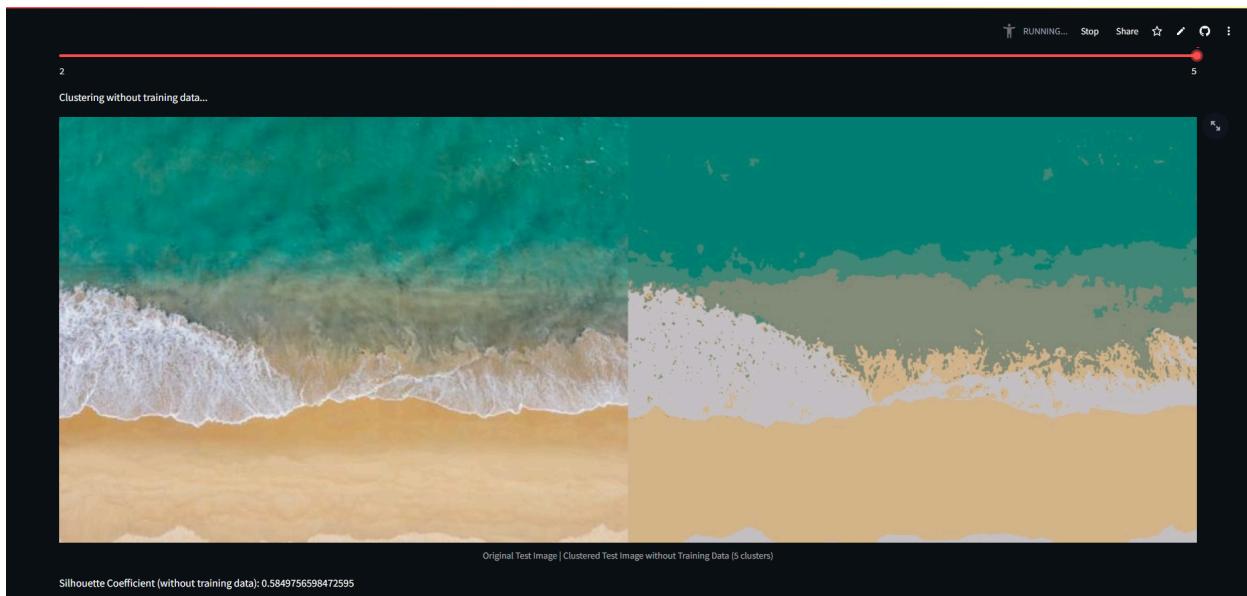
5.1. Tampilan Awal Aplikasi



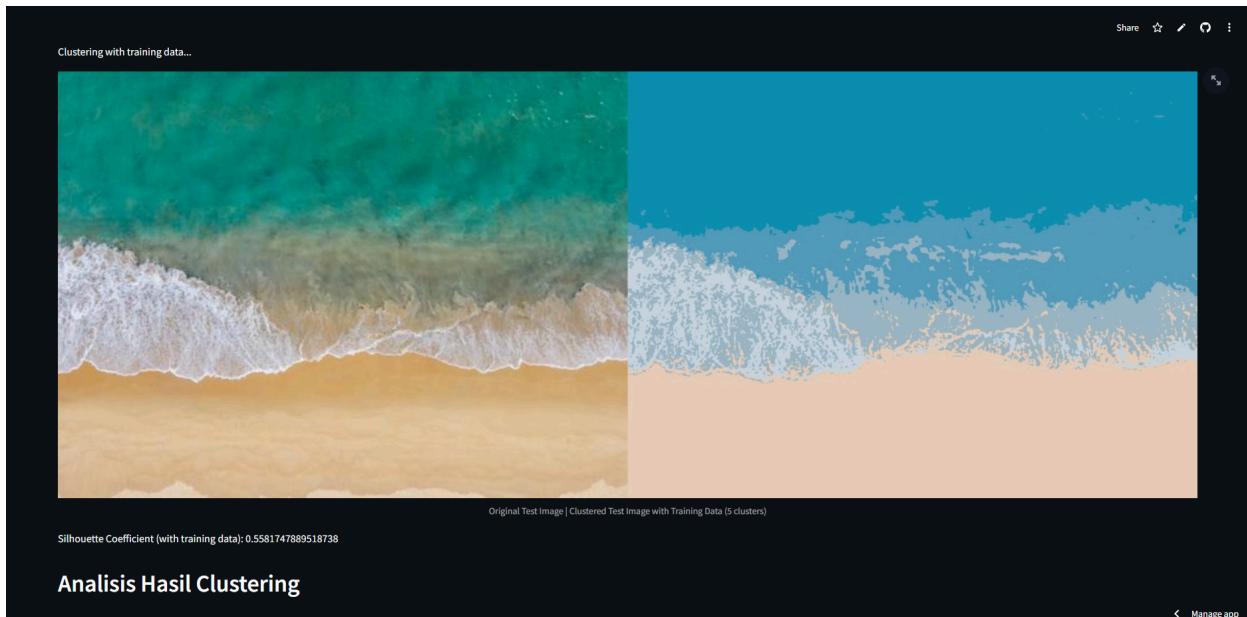
5.2. User upload data latih dan data uji beserta pemilihan jumlah cluster



5.3. Hasil Clustering pada data uji tanpa data latih



5.4. Hasil Clustering pada data uji dengan data latih



5.5. Analisis Hasil Clustering

Original Test Image | Clustered Test Image with Training Data (5 clusters)

Silhouette Coefficient (with training data): 0.5581747889518738

Analisis Hasil Clustering

Koefisien Silhouette

- Clustering tanpa data latih: 0.5849756598472595
- Clustering dengan data latih: 0.5581747889518738

Interpretasi Silhouette

Clustering tanpa data latih memberikan hasil yang lebih baik. Penggunaan data latih mungkin menyebabkan overfitting atau kurang cocok untuk pengelompokan.

Implikasi Clustering

Clustering tanpa data latih dapat memberikan hasil yang baik untuk kasus-kasus di mana data latih sulit diperoleh atau tidak tersedia. Namun, menggunakan data latih yang relevan dapat meningkatkan akurasi dan kualitas hasil clustering, terutama untuk aplikasi yang membutuhkan pengelompokan yang lebih spesifik.

Dalam aplikasi nyata, seperti segmentasi gambar, menggunakan data latih yang representatif dari domain tertentu (misalnya, citra medis atau citra satelit) dapat membantu menghasilkan pengelompokan yang lebih bermakna dan dapat ditindaklanjuti.

Rekomendasi untuk Peningkatan

- Coba tambahkan lebih banyak data latih yang lebih bervariasi untuk meningkatkan akurasi clustering.

[Download PDF](#)

◀ Manage app