LAPORAN TUGAS BESAR PEMBELAJARAN MESIN
CLUSTERING DATASET “SHIP PERFORMANCE DATASET”
MEMBANDINGKAN METODE K-MEANS DAN DBSCAN

DISUSUN OLEH:

ADINDA EKA RAHAYU (1206230025)

SUN KAYLA ZELIKHA AZ ZAHRA (1206230031)

ABIDA RAMADHANI MURYANSILA (1206230045)

PROGRAM STUDI SAINS DATA

FAKULTAS INFORMATIKA

TELKOM UNIVERSITY SURABAYA

2025

A. Formulasi Masalah

Dalam industri pelayaran modern, efisiensi energi menjadi salah satu indikator kunci keberhasilan operasional. Kapal-kapal niaga atau logistik menempuh jarak yang sangat jauh dengan beban dan kondisi laut yang bervariasi, sehingga penggunaan energi menjadi salah satu komponen biaya terbesar. Jika konsumsi energi tidak dioptimalkan, maka akan timbul dampak langsung berupa peningkatan biaya bahan bakar, penurunan umur mesin karena beban berlebih, serta kontribusi signifikan terhadap emisi karbon di atmosfer. Oleh karena itu, sangat penting bagi perusahaan pelayaran untuk memahami karakteristik konsumsi energi dari tiap kapal, agar bisa mengelompokkan kapal berdasarkan performa aktual mereka di lapangan.

Permasalahan utama yang muncul adalah bagaimana cara mengidentifikasi kapal-kapal mana yang termasuk efisien, mana yang boros karena operasional (misalnya karena rute jauh), dan mana yang boros karena tidak efisien secara teknis. Dengan kata lain, perusahaan tidak hanya butuh melihat konsumsi energi total semata, tetapi juga perlu mempertimbangkan efisiensi energi, yaitu seberapa jauh kapal mampu berlayar per satuan energi yang dikonsumsi. Dua kapal yang mengonsumsi energi besar bisa saja sangat berbeda: satu menempuh ribuan mil laut dengan performa tinggi, satu lagi mungkin tidak efisien meskipun menempuh jarak lebih pendek. Pola-pola semacam ini tidak mudah dikenali tanpa pendekatan analitik seperti *clustering*.

Oleh karena itu, pendekatan *unsupervised learning* seperti KMeans dan DBSCAN digunakan untuk melakukan segmentasi terhadap kapal-kapal berdasarkan kombinasi dua fitur utama: total energi yang digunakan (estimated_energy_used) dan efisiensi energi (Efficiency_nm_per_kWh). Clustering akan membantu memetakan kapal ke dalam kelompok-kelompok dengan pola penggunaan energi yang mirip. Dengan cara ini, kita tidak hanya bisa mengidentifikasi kelompok kapal boros dan hemat, tetapi juga memahami apakah ada kelompok yang menunjukkan pola anomali atau outlier yang bisa jadi merupakan kapal yang memerlukan perhatian khusus dalam hal maintenance atau operasional.

Tujuan akhir dari formulasi masalah ini adalah menghasilkan sistem segmentasi performa energi kapal secara otomatis yang dapat menjadi dasar pengambilan keputusan teknis dan strategis. Misalnya, kapal dalam cluster "tidak efisien" bisa dijadikan target evaluasi teknis atau rencana retrofit mesin. Kapal dalam cluster "efisien" bisa dijadikan standar benchmarking. Sementara kapal dalam cluster

"anomali" atau "outlier" bisa diperiksa lebih lanjut apakah penyebabnya bersifat operasional, lingkungan, atau kerusakan sistem. Dengan demikian, metode clustering memberikan pendekatan berbasis data untuk pengambilan keputusan yang lebih cerdas dan akurat dalam pengelolaan armada kapal.

B. Eksplorasi dan Persiapan Data

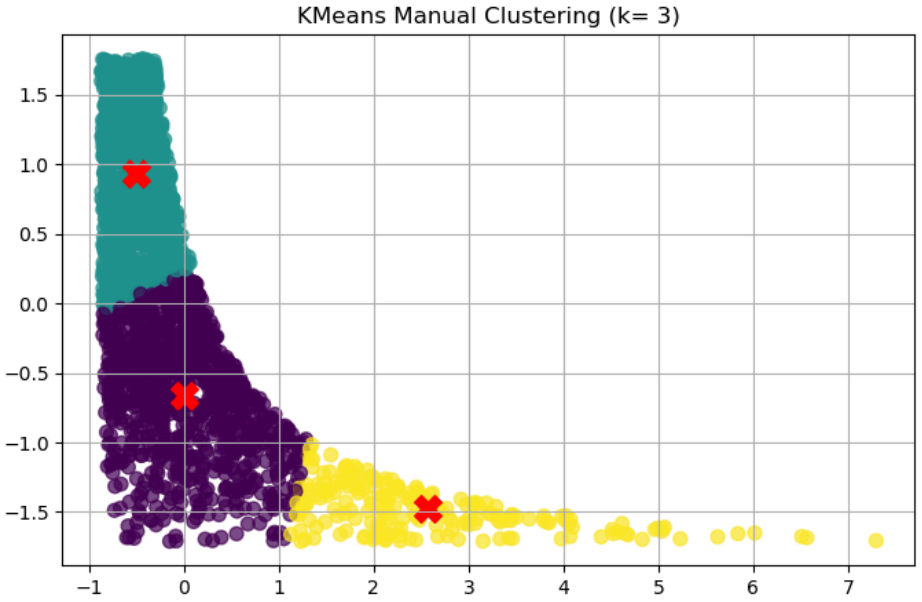
Langkah preprocessing dimulai dengan menghapus data yang memiliki nilai kosong untuk memastikan kelengkapan data. Selanjutnya, dilakukan encoding terhadap kolom kategorikal seperti jenis kapal dan kondisi cuaca menggunakan LabelEncoder agar dapat diproses oleh algoritma machine learning. Untuk meningkatkan relevansi analisis, ditambahkan fitur baru bernama estimated_energy_used yang dihitung dari pembagian jarak tempuh kapal dengan efisiensinya. Fitur-fitur utama yang digunakan untuk clustering adalah estimated_energy_used dan Efficiency_nm_per_kWh. Kedua fitur ini kemudian distandarisasi menggunakan Z-score normalization agar memiliki skala yang sama, sehingga proses clustering menjadi lebih adil dan akurat.

C. Proses Clustering

Pada tahap ini, dilakukan proses clustering menggunakan algoritma KMeans yang diimplementasikan secara manual. Algoritma ini bekerja dengan cara menginisialisasi k buah centroid secara acak dari data, kemudian mengelompokkan setiap titik data ke centroid terdekat berdasarkan jarak Euclidean. Proses dimulai dengan memilih dua fitur utama, yaitu estimated_energy_used dan Efficiency_nm_per_kWh, yang telah distandarisasi menggunakan z-score normalization agar kedua fitur memiliki skala yang sebanding dan tidak mendominasi dalam perhitungan jarak. Setelah data distandarisasi, proses clustering dimulai dengan menentukan jumlah cluster (k) sebanyak 3, yang berarti model akan mencoba mengelompokkan data ke dalam tiga grup yang berbeda.

Langkah pertama dari KMeans adalah inisialisasi centroid secara acak. Tiga titik dipilih secara acak dari dataset sebagai centroid awal. Kemudian, untuk setiap titik data, algoritma menghitung jarak Euclidean ke masing-masing centroid dan menetapkan setiap titik ke cluster terdekat. Setelah seluruh data diberi label cluster, centroid dihitung ulang dengan mengambil rata-rata dari semua titik yang berada dalam masing-masing cluster. Proses ini diulang terus menerus, yakni perhitungan jarak, penentuan cluster, dan pembaruan centroid, hingga posisi centroid tidak lagi berubah secara signifikan (konvergen) atau jumlah iterasi maksimum tercapai.

Setelah model selesai dilatih, diperoleh hasil pengelompokan di mana setiap titik data memiliki label cluster masing-masing. Hasil ini divisualisasikan dalam bentuk scatter plot dua dimensi, di mana data ditampilkan berdasarkan dua fitur yang digunakan (sumbu x dan y), dan tiap cluster diberi warna berbeda untuk membedakan. Titik centroid ditandai secara eksplisit menggunakan simbol 'X' berwarna merah untuk menunjukkan pusat dari masing-masing cluster.



Dari hasil visualisasi tersebut, terlihat bahwa data terbagi ke dalam tiga kelompok yang relatif terpisah dan terorganisir, menunjukkan bahwa proses clustering berjalan dengan baik dan centroid berhasil menemukan posisi optimalnya berdasarkan struktur distribusi data.

Berdasarkan hasil implementasi KMeans manual dengan jumlah klaster (k) = 3, didapatkan pengelompokan data kapal berdasarkan dua fitur utama, yaitu estimated_energy_used (jumlah energi yang diperkirakan digunakan) dan Efficiency_nm_per_kWh (efisiensi jarak tempuh per satuan energi). Didapatkan Tiga klaster yang memiliki distribusi yang dapat dibedakan berdasarkan posisi relatif antar titik data dan centroid-nya. Misalnya, satu klaster terdiri dari kapal-kapal dengan efisiensi tinggi namun penggunaan energi rendah, yang mengindikasikan kapal tersebut hemat energi. Klaster lainnya mungkin berisi kapal dengan konsumsi energi tinggi namun efisiensinya rendah, yang bisa menunjukkan adanya ketidakefisienan operasional. Sedangkan klaster ketiga kemungkinan terdiri dari kapal yang berada di antara dua kategori tersebut atau memiliki karakteristik unik lain yang membedakannya.

D. Evaluasi

Dalam proses evaluasi clustering, metode KMeans manual menunjukkan performa yang sangat baik dengan Silhouette Score tertinggi sebesar 0.7669 pada jumlah cluster K = 2. Nilai ini mengindikasikan bahwa objek-objek dalam satu cluster memiliki kemiripan yang tinggi dan cukup terpisah secara signifikan dari cluster lain, mencerminkan struktur klaster yang kompak dan jelas. Selain itu, Davies-Bouldin Index yang diperoleh adalah 0.6426, yang juga merupakan indikasi positif karena semakin kecil nilai indeks ini, semakin baik pemisahan antar cluster dan semakin kecil variasi internal di dalamnya. Hal ini menunjukkan bahwa KMeans cukup efektif dalam membentuk cluster yang representatif terhadap pola performa kapal berdasarkan efisiensi dan konsumsi energinya.

Sementara itu, metode DBSCAN menghasilkan Silhouette Score sebesar 0.6079 dan Davies-Bouldin Index sebesar 1.0505. Meskipun nilainya tidak sebaik KMeans, DBSCAN tetap memberikan hasil yang layak, terutama karena pendekatannya yang fleksibel dalam mendeteksi bentuk cluster yang tidak reguler dan kemampuannya dalam menangani noise (outlier). Namun, skor evaluasi tersebut menunjukkan bahwa struktur cluster yang terbentuk oleh DBSCAN kurang terpisah secara tajam dan memiliki variasi internal yang lebih besar dibandingkan KMeans. Hal ini bisa disebabkan oleh sensitivitas parameter DBSCAN terhadap skala dan distribusi data, serta kemungkinan bahwa data yang digunakan memang lebih cocok dipetakan secara linier seperti pada KMeans.

Dalam hal pemilihan metrik evaluasi, Silhouette Score digunakan karena tidak memerlukan label asli (ground truth) dan mampu mengevaluasi seberapa baik sebuah titik cocok dengan cluster-nya dibandingkan dengan cluster lainnya, dengan rentang nilai dari -1 hingga 1 yang memudahkan interpretasi. Di sisi lain, Davies-Bouldin Index dipilih sebagai metrik pelengkap karena mempertimbangkan rasio antara jarak antar centroid dengan sebaran internal masing-masing cluster, sehingga memberikan perspektif tambahan mengenai kepadatan dan pemisahan cluster. Kombinasi kedua metrik ini memberikan gambaran evaluatif yang lebih seimbang dalam membandingkan kualitas model KMeans dan DBSCAN.

E. Eksperimen dan Perbandingan

Metode	Parameter Utama	Silhouette Score	Davies-Bouldin Index	Jumlah Cluster Terbentuk	Keterangan
KMeans Manual	K = 2	0.7669	0.6426	2	Cluster sangat terpisah, visualisasi rapi. Hasil terbaik menurut metrik evaluasi.
KMeans Manual	K = 3–10	0.65–0.55	> 0.7 (rata-rata)	3-10	Skor cenderung menurun, terjadi overclustering, separasi antar cluster menurun.
DBSCAN	ε = 0.8, min_samples = 5	0.6079	1.0505	3 (1 utama + 2 noise)	Sebagian besar data terkumpul dalam 1 cluster, sisanya noise. Kurang optimal untuk pola distribusi data ini.

Jika dibandingkan secara langsung, metode KMeans menunjukkan performa yang lebih unggul pada dataset ini dibandingkan DBSCAN, terutama dalam hal skor evaluasi dan kejelasan visualisasi cluster. Kelebihan KMeans adalah kemampuannya membentuk cluster dengan batas yang jelas dan efisien untuk data yang tersebar secara merata. Sebaliknya, DBSCAN lebih cocok untuk mendeteksi cluster dengan bentuk tidak teratur atau ketika data mengandung noise yang signifikan. Namun, pada kasus ini, data yang digunakan lebih sesuai dengan asumsi KMeans, yaitu distribusi berbentuk bundar (globular), sehingga DBSCAN tidak memberikan hasil optimal.

Dengan eksperimen ini, dapat disimpulkan bahwa KMeans lebih sesuai digunakan untuk dataset Ship Performance yang telah dinormalisasi dan memiliki

fitur efisiensi dan konsumsi energi, karena mampu mengelompokkan kapal berdasarkan pola efisiensinya secara lebih terstruktur dan mudah diinterpretasikan.

F. Kesimpulan

Berdasarkan hasil analisis dan eksperimen yang telah dilakukan, dapat disimpulkan bahwa algoritma KMeans manual memberikan hasil clustering yang lebih optimal dibandingkan DBSCAN untuk dataset performa kapal ini. Dengan menggunakan dua fitur utama, yaitu estimated_energy_used dan Efficiency_nm_per_kWh, proses clustering berhasil mengelompokkan kapal-kapal berdasarkan pola efisiensi dan konsumsi energinya. KMeans menghasilkan Silhouette Score tertinggi sebesar 0.7669 pada K = 2, yang menunjukkan bahwa hasil klasterisasi cukup terpisah dan kompak. Sementara itu, DBSCAN hanya menghasilkan Silhouette Score sebesar 0.6079 dan menunjukkan kecenderungan mengelompokkan sebagian besar data ke dalam satu cluster, sementara sisanya dianggap noise.

Dalam hal visualisasi dan interpretasi, hasil clustering KMeans juga lebih mudah dipahami dan menunjukkan distribusi cluster yang lebih terstruktur. Meskipun DBSCAN memiliki keunggulan dalam mendeteksi outlier dan bentuk cluster yang tidak beraturan, pada kasus ini struktur data lebih cocok untuk pendekatan berbentuk bundar seperti asumsi KMeans. Oleh karena itu, dapat disimpulkan bahwa KMeans adalah metode clustering yang lebih tepat untuk digunakan pada dataset Ship Performance yang telah dinormalisasi dan difokuskan pada efisiensi energi kapal.