Homework Solution - Unsupervised Learning

I Gusti Ngurah Agung Hari Vijaya Kusuma Batch 57

August 17, 2025

Submission Links

- Repository: github.com/AgungHari

1 Pendahuluan

1.1 Latar Belakang

Diberikan tugas untuk membuat model unsupervised learning dengan menggunakan dataset yang berisi data customer sebuah perusahaan penerbangan. Dataset ini mencakup berbagai fitur yang dapat menggambarkan nilai dari setiap customer, seperti ID Member, tanggal bergabung dalam program Frequent Flyer, jenis kelamin, tier program, kota asal, provinsi asal, negara asal, umur, jumlah penerbangan yang telah dilakukan, dan informasi terkait jarak penerbangan serta poin yang diperoleh.

Tujuan dari tugas ini adalah untuk menjawab Soal soal yang diberikan oleh Rakamin Academy. Dimana diharapkan outputnya dapat memberikan wawasan yang lebih dalam mengenai pola perilaku customer, segmentasi pasar, dan rekomendasi bisnis yang relevan berdasarkan hasil clustering. Dengan demikian, perusahaan penerbangan dapat mengoptimalkan strategi pemasaran dan meningkatkan pengalaman pelanggan.

1.2 Homework Unsupervised Learning

Adapun beberapa soal yang diberikan dalam tugas ini adalah sebagai berikut:

1. Lakukan EDA pada dataset untuk mendapatkan pemahaman umum mengenai data dan memandu proses feature engineering (20 poin)
    - Pastikan setiap kolom dataset memiliki tipe data yang tepat, tidak ada data kosong, bebas dari duplikat, dan berada di range value yang tepat.
    - Keluarkan statistik kolom baik numerik maupun kategorikal, cari bentuk distribusi setiap kolom (numerik), dan jumlah baris untuk setiap unique value (kategorikal).
    - Cari tahu apakah ada kolom-kolom yang berkorelasi kuat satu sama lain.

2. Pilih fitur-fitur yang menurut teman-teman masuk akal secara bisnis untuk digunakan sebagai fitur clustering. Lakukan feature engineering! (20 poin)

    - Dari sekian banyak kolom yang ada, tentukan 3-6 fitur untuk digunakan sebagai fitur clustering. Tulis alasan teman-teman memilih fitur tersebut.
    - Lakukan preprocessing dan feature engineering (apabila fitur yang teman-teman pilih merupakan fitur baru yang dihasilkan dari fitur-fitur yang sudah ada).

3. Lakukan clustering K-means! Temukan jumlah cluster yang menurut teman-teman optimal dan evaluasi cluster yang dihasilkan dengan visualisasi dan silhouette score (30 poin)

    - Temukan jumlah cluster yang optimal dengan menggunakan elbow method.
    - Lakukan clustering menggunakan K-means.
    - Evaluasi cluster yang dihasilkan dengan menggunakan visualisasi, gunakan PCA apabila diperlukan.

4. Interpretasi cluster yang dihasilkan secara bisnis dan berikan rekomendasi yang sesuai dengan cluster yang dihasilkan (30 poin)

    - Tempelkan kembali label yang dihasilkan ke dataframe asal, dan keluarkan statistik fitur dari setiap cluster.
    - Deskripsikan secara kontekstual customer seperti apa yang ada di masing-masing cluster.
    - Berdasarkan cluster tersebut, berikan 1-2 rekomendasi bisnis.

2 Tinjauan Pustaka

2.1 Frequent Flyer Program

Frequent Flyer Program (FFP) adalah program loyalitas yang ditawarkan oleh maskapai penerbangan kepada pelanggan setia mereka. Program ini memberikan berbagai keuntungan, seperti akumulasi poin atau miles yang dapat ditukarkan dengan tiket penerbangan gratis, peningkatan kelas penerbangan, akses ke lounge bandara, dan layanan prioritas lainnya. FFP dirancang untuk mendorong pelanggan agar terus menggunakan layanan maskapai tertentu, sehingga meningkatkan retensi pelanggan dan loyalitas merek.

Program ini biasanya memiliki beberapa tingkatan atau tier, yang memberikan manfaat tambahan kepada anggota yang mencapai tingkat tertentu berdasarkan frekuensi penerbangan atau jumlah poin yang dikumpulkan. Dengan demikian, FFP tidak hanya memberikan insentif bagi pelanggan untuk terbang lebih sering, tetapi juga menciptakan hubungan jangka panjang antara maskapai dan pelanggannya.

Gambar 1: Contoh Frequent Flyer Program

2.2 EDA (Exploratory Data Analysis)

Exploratory Data Analysis (EDA) adalah proses analisis data yang bertujuan untuk memahami struktur, pola, dan hubungan dalam dataset sebelum melakukan analisis lebih lanjut atau membangun model. EDA melibatkan penggunaan berbagai teknik statistik dan visualisasi untuk mengeksplorasi data, mengidentifikasi anomali, dan mendapatkan wawasan awal tentang karakteristik data.

Proses EDA biasanya mencakup langkah-langkah seperti:

- **Pemeriksaan Data:** Memeriksa tipe data, nilai yang hilang, dan distribusi variabel.
- **Statistik Deskriptif:** Menghitung ukuran pusat (mean, median) dan ukuran dispersi (standar deviasi, rentang).
- **Visualisasi Data:** Menggunakan grafik seperti histogram, boxplot, dan scatter plot untuk memahami distribusi dan hubungan antar variabel.
- **Identifikasi Outlier:** Mendeteksi nilai-nilai yang tidak biasa yang dapat mempengaruhi analisis.

- **Korelasi:** Menganalisis hubungan antar variabel untuk mengidentifikasi pola yang mungkin ada.

EDA sangat penting dalam tahap awal analisis data karena membantu peneliti atau analis untuk memahami data secara mendalam, mengarahkan fokus pada area yang relevan, dan menginformasikan keputusan tentang teknik analisis yang akan digunakan selanjutnya. Dengan demikian, EDA merupakan langkah krusial dalam proses analisis data yang efektif.

2.3 Data Preprocessing

Data preprocessing adalah langkah penting dalam analisis data yang melibatkan pembersihan, transformasi, dan persiapan data sebelum digunakan dalam analisis atau pemodelan. Tujuan dari preprocessing adalah untuk memastikan bahwa data dalam kondisi yang baik, konsisten, dan siap untuk dianalisis. Langkah-langkah umum dalam data preprocessing meliputi:

- **Pembersihan Data:** Menghapus atau memperbaiki data yang tidak lengkap, duplikat, atau tidak konsisten. Ini termasuk menangani nilai yang hilang, mengoreksi kesalahan pengetikan, dan menghapus outlier yang tidak relevan.
- **Transformasi Data:** Mengubah format data agar sesuai dengan kebutuhan analisis. Ini bisa meliputi normalisasi atau standarisasi nilai numerik, pengkodean variabel kategorikal, dan konversi tipe data.
- **Penggabungan Data:** Menggabungkan beberapa sumber data menjadi satu dataset yang kohesif, jika diperlukan.
- **Pemisahan Data:** Membagi dataset menjadi subset untuk pelatihan dan pengujian model, terutama dalam konteks machine learning.
- **Feature Engineering:** Membuat fitur baru dari data yang ada untuk meningkatkan kinerja model. Ini bisa meliputi penggabungan variabel, ekstraksi informasi dari teks, atau pembuatan variabel waktu.
- **Skalasi Data:** Mengubah skala fitur numerik agar berada dalam rentang yang sama, yang penting untuk algoritma yang sensitif terhadap skala, seperti K-Means atau SVM.

2.4 K-Means Clustering

K-Means Clustering adalah algoritma pembelajaran tidak terawasi yang digunakan untuk mengelompokkan data ke dalam sejumlah kluster berdasarkan kesamaan fitur. Algoritma ini bekerja dengan cara membagi dataset menjadi K kluster, di mana setiap kluster diwakili oleh centroid (titik pusat kluster). Proses K-Means Clustering melibatkan langkah-langkah berikut:

1. **Inisialisasi Centroid:** Memilih K titik acak dari dataset sebagai centroid awal.

2. **Penugasan Kluster:** Menghitung jarak antara setiap titik data dan centroid, lalu mengelompokkan setiap titik ke kluster terdekat.

3. **Pembaruan Centroid:** Menghitung ulang centroid untuk setiap kluster berdasarkan rata-rata posisi titik-titik dalam kluster tersebut.

4. **Iterasi:** Mengulangi langkah 2 dan 3 hingga centroid tidak berubah secara signifikan atau jumlah iterasi maksimum tercapai.

5. **Output:** Menghasilkan kluster yang berisi titik-titik data yang dikelompokkan berdasarkan kesamaan fitur.

Adapun rumus untuk menghitung jarak antara titik data dan centroid biasanya menggunakan Euclidean distance, yang didefinisikan sebagai:

$$d(x, c) = \sqrt{\sum_{i=1}^{n} (x_i - c_i)^2} \tag{1}$$

Di mana :

- $d(x, c)$ adalah jarak antara titik data x dan centroid c,
- x_i adalah nilai fitur ke-i dari titik data x,
- c_i adalah nilai fitur ke-i dari centroid c,
- n adalah jumlah fitur.
- x adalah titik data yang akan dikelompokkan.
- c adalah centroid dari kluster yang sedang dianalisis.

K-Means Clustering banyak digunakan dalam berbagai aplikasi, seperti segmentasi pasar, pengelompokan dokumen, dan analisis citra. Kelebihan dari algoritma ini adalah kesederhanaannya dan efisiensi dalam menangani dataset besar. Namun, K-Means juga memiliki beberapa kelemahan, seperti ketergantungan pada pemilihan jumlah kluster K yang tepat dan sensitivitas terhadap outlier.

2.5 Elbow Method

Elbow Method adalah teknik yang digunakan untuk menentukan jumlah optimal kluster K dalam algoritma K-Means Clustering. Metode ini melibatkan pengukuran varians dalam kluster (inertia) untuk berbagai nilai K dan kemudian memplot hasilnya. Tujuan dari Elbow Method adalah untuk menemukan titik di mana penambahan kluster baru tidak memberikan peningkatan signifikan dalam pengurangan varians.

Proses Elbow Method meliputi langkah-langkah berikut:

1. **Inisialisasi K-Means:** Jalankan algoritma K-Means untuk berbagai nilai K (misalnya, dari 1 hingga 10).

2. **Hitung Inertia:** Untuk setiap nilai K, hitung inertia, yaitu jumlah jarak kuadrat antara titik data dan centroid kluster mereka.

3. **Plot Inertia:** Buat plot dengan nilai K pada sumbu x dan inertia pada sumbu y.

4. **Identifikasi Elbow:** Cari titik di mana penurunan inertia mulai melambat, yang biasanya terlihat seperti "siku" pada plot. Titik ini menunjukkan jumlah kluster optimal.

5. **Pilih K Optimal:** Nilai K pada titik siku ini dianggap sebagai jumlah kluster yang paling sesuai untuk dataset.

Elbow Method membantu dalam menghindari overfitting dengan memilih jumlah kluster yang tepat, sehingga model K-Means dapat menangkap struktur data dengan baik tanpa menjadi terlalu kompleks. Meskipun metode ini sederhana dan intuitif, hasilnya dapat bervariasi tergantung pada dataset dan distribusi data, sehingga penting untuk mempertimbangkan konteks analisis saat menentukan jumlah kluster optimal.

2.6 Silhouette Score

Silhouette Score adalah metrik yang digunakan untuk mengevaluasi kualitas klustering dalam algoritma K-Means atau metode klustering lainnya. Metrik ini mengukur seberapa baik setiap titik data dikelompokkan dalam kluster yang benar, dengan mempertimbangkan jarak antar titik dalam kluster dan jarak ke titik di kluster lain. Silhouette Score untuk setiap titik data dihitung dengan rumus berikut:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{2}$$

Di mana:

- $s(i)$ adalah Silhouette Score untuk titik data i,
- $a(i)$ adalah rata-rata jarak antara titik data i dan semua titik lain dalam kluster yang sama (intra-kluster),
- $b(i)$ adalah rata-rata jarak antara titik data i dan titik-titik di kluster terdekat lainnya (inter-kluster).
- i adalah indeks dari titik data yang sedang dianalisis.

Nilai Silhouette Score berkisar antara -1 hingga 1:

- Nilai mendekati 1 menunjukkan bahwa titik data berada jauh dari kluster lain dan dekat dengan kluster yang benar.
- Nilai mendekati 0 menunjukkan bahwa titik data berada di batas antara dua kluster.
- Nilai negatif menunjukkan bahwa titik data mungkin telah dikelompokkan ke kluster yang salah.

Silhouette Score memberikan gambaran tentang seberapa baik klustering dilakukan, dengan nilai yang lebih tinggi menunjukkan klustering yang lebih baik. Metrik ini berguna untuk membandingkan hasil klustering dengan jumlah kluster yang berbeda dan membantu dalam memilih jumlah kluster yang optimal.

2.7 PCA (Principal Component Analysis)

Principal Component Analysis (PCA) adalah teknik reduksi dimensi yang digunakan untuk mengurangi jumlah variabel dalam dataset sambil mempertahankan sebanyak mungkin informasi yang ada. PCA bekerja dengan mengidentifikasi arah (komponen utama) di mana data memiliki varians terbesar, sehingga memungkinkan representasi data dalam ruang dimensi yang lebih rendah.

Proses PCA melibatkan langkah-langkah berikut:

1. **Standardisasi Data:** Mengubah data sehingga memiliki rata-rata 0 dan deviasi standar 1 untuk setiap fitur, agar setiap fitur berkontribusi secara setara.

2. **Kovarians Matriks:** Menghitung matriks kovarians untuk memahami hubungan antar fitur dalam dataset.

3. **Eigen Decomposition:** Menghitung nilai eigen (eigenvalues) dan vektor eigen (eigenvectors) dari matriks kovarians. Vektor eigen menunjukkan arah komponen utama, sedangkan nilai eigen menunjukkan seberapa banyak varians yang dijelaskan oleh masing-masing komponen.

4. **Pemilihan Komponen Utama:** Memilih sejumlah komponen utama berdasarkan nilai eigen terbesar, yang akan digunakan untuk merepresentasikan data dalam dimensi yang lebih rendah.

5. **Transformasi Data:** Mengalikan data asli dengan vektor eigen terpilih untuk mendapatkan representasi baru dalam ruang dimensi yang lebih rendah.

PCA sangat berguna dalam mengurangi kompleksitas data, menghilangkan redundansi, dan meningkatkan efisiensi komputasi dalam analisis data. Selain itu, PCA juga membantu dalam visualisasi data dengan mengurangi dimensi menjadi 2 atau 3 komponen utama, sehingga memudahkan pemahaman pola dan struktur dalam dataset. Namun, penting untuk diingat bahwa PCA adalah teknik linier, sehingga mungkin tidak cocok untuk semua jenis data, terutama yang memiliki hubungan non-linier yang kompleks.

2.8 RFM (Recency, Frequency, Monetary)

RFM (Recency, Frequency, Monetary) adalah metode analisis yang digunakan untuk mengukur nilai pelanggan berdasarkan tiga dimensi utama:

- **Recency (R):** Mengukur seberapa baru pelanggan melakukan pembelian. Semakin baru pembelian, semakin tinggi nilai recency.

- **Frequency (F):** Mengukur seberapa sering pelanggan melakukan pembelian dalam periode tertentu. Semakin sering pembelian, semakin tinggi nilai frequency.
- **Monetary (M):** Mengukur total pengeluaran pelanggan dalam periode tertentu. Semakin besar pengeluaran, semakin tinggi nilai monetary.

3 Desain dan Implementasi

3.1 Deskripsi sistem

4 Pengujian dan analisis

4.1 Pengujian Sistem

5 Kesimpulan dan Saran