Rakamin Academy

UNSUPERVISED LEARNING

Homework

Estimasi Waktu Pengerjaan

 2 - 3 jam

Jumlah Soal

 4 Bagian

Total Point

 100 poin

Teknis Pengerjaan

1. Pekerjaan dilakukan secara **berkelompok**, dengan output berupa:
   a. File .ipynb yang berisi hasil analisis/modeling
   b. Sebuah presentasi .pdf yang berisi ringkasan dari poin-poin penting yang dijabarkan pada slide-slide selanjutnya

2. Homework ini berupa soal clustering **end-to-end pertanyaan bisnis** dimana teman-teman akan diberikan sebuah dataset mentah berisi data customer. Teman-teman diharapkan melakukan 4 hal berikut dengan menggunakan dataset tersebut
   - ***EDA***
   - ***Feature Engineering***
   - ***Modeling + Evaluasi***
   - ***Interpretasi model + Rekomendasi***

3. Upload hasil pengerjaanmu melalui LMS dengan format nama file sebagai berikut **Nama_Lengkap_Batch_XX** dalam format .html (cara save dalam format .html disini)

Dataset yang digunakan

Airline Customer Value Analysis Case

- **Deskripsi**:

  Dataset ini berisi data customer sebuah perusahaan penerbangan dan beberapa fitur yang dapat menggambarkan value dari customer tersebut.

- **Data**:

  Setiap baris mewakili customer, setiap kolom berisi atribut customer.

- **Link download disini**

Deskripsi Dataset

Code	Description
MEMBER_NO-b	: ID Member
FFP_DATE	: Frequent Flyer Program Join Date
FIRST_FLIGHT_DATE	: Tanggal Penerbangan pertama
GENDER	: Jenis Kelamin
FFP_TIER	: Tier dari Frequent Flyer Program
WORK_CITY	: Kota Asal
WORK_PROVINCE	: Provinsi Asal
WORK_COUNTRY	: Negara Asal
AGE	: Umur Customer
LOAD_TIME	: Tanggal data diambil
FLIGHT_COUNT	: Jumlah penerbangan Customer
BP_SUM	: Rencana Perjalanan
SUM_YR_1	: Fare Revenue
SUM_YR_2	: Votes Prices
SEG_KM_SUM	: Total jarak(km) penerbangan yg sudah dilakukan
LAST_FLIGHT_DATE	: Tanggal penerbangan terakhir
LAST_TO_END	: Jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir
AVG_INTERVAL	: Rata-rata jarak waktu
MAX_INTERVAL	: Maksimal jarak waktu
EXCHANGE_COUNT	: Jumlah penukaran
avg_discount	: Rata rata discount yang didapat customer
Points_Sum	: Jumlah poin yang didapat customer
Point_NotFlight	: point yang tidak digunakan oleh members

Instruksi Detil

1. **Lakukan EDA pada dataset untuk mendapatkan pemahaman umum mengenai data dan memandu proses feature engineering (20 poin)**

Langkah-langkah:

a. Pastikan setiap kolom dataset memiliki tipe data yang tepat, tidak ada data kosong, bebas dari duplikat, dan berada di *range* value yang tepat
b. Keluarkan statistik kolom baik numerik maupun kategorikal, cari bentuk distribusi setiap kolom (numerik), dan jumlah baris untuk setiap *unique* value (kategorikal)
c. Cari tahu apakah ada kolom-kolom yang berkorelasi kuat satu sama lain

Untuk mempermudah kamu, yuk lihat resource di bawah ini:

- Topic Machine Learning Preparation - EDA

30-45 menit

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

Langkah-langkah:

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

Untuk mempermudah kamu, yuk lihat resource di bawah ini:

- Topic Machine Learning Preparation - Feature Engineering

30-45 menit

Rakamin Academy

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)

Langkah-langkah:

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

Untuk mempermudah kamu, yuk lihat resource di bawah ini:

- Topic Unsupervised Learning - Clustering bagian KMeans dan Evaluasi

20-30 menit

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

Langkah-langkah:

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

Untuk mempermudah kamu, yuk lihat resource di bawah ini:

- Topic Unsupervised Learning - Clustering

20-30 menit

Rakamin Academy

Selamat Mengerjakan!