TUGAS LAB
DATA CLUSTERING & CLASSIFICATION

PENAMBANGAN DATA DAN INTELIGENSIA BISNIS (PDIB) FAKULTAS ILMU KOMPUTER, UNIVERSITAS INDONESIA

Petunjuk Pengerjaan

A. Tugas ini dikerjakan secara **kelompok 2-4 orang.**
B. Jangan lupa menuliskan nama dan NPM di awal pekerjaan Anda.
C. Kumpulkan jawaban sebelum **Selasa 29 November 2022**, pukul **22:00 WIB** pada *dropbox* Scele, serta **Video** maksimal satu minggu setelah pengumpulan jawaban (Selasa, 6 Desember 2022 pukul 22.00), sebagai pengganti presentasi.
D. Fase 1 kumpulkan dalam format .zip yang terdiri dari:
    - File csv yang berisi data yang sudah dikerjakan.
    - **Dokumen pdf** yang berisi laporan beserta npm dan nama yang mengerjakan.
    - file .ipynb
    - file .twbx , yang berisi grafik visual tableau
E. Gunakan format di bawah ini untuk melakukan penamaan dokumen:
    - Lab4_NPM_NamaPengumpul.zip
    - Contoh: Lab4_2106557535_FaisalRahmanto.zip
F. Untuk video, upload video di Youtube, dan tuliskan pada forum link yang Anda gunakan. Pada video presentasi, setiap orang harus menampilkan wajah dan melakukan presentasi. Jangan lupa perkenalkan diri Anda di awal. Untuk presentasi kelompok Anda boleh menggunakan pdf yang sebelumnya telah dikumpulkan, ataupun membuat ppt yang baru.

Persiapan Tugas
Tugas yang dikerjakan pada lab kali ini adalah *prediction* dan *clustering* dengan menggunakan python atau tableau. Untuk dapat mengerjakan tugas ini dengan baik, pastikan Anda telah melakukan hal berikut:

a. Instalasi Tableau

Pastikan Tableau sudah terinstall pada komputer Anda. Seperti pada tugas Lab sebelumnya.

b. Instalasi Python

Pastikan python sudah terinstall pada komputer Anda. Seperti pada tugas Lab sebelumnya mengenai Data Preparation.

c. Install Environment

- **Instalasi Anaconda dan Jupyter Notebook**

  Tugas ini akan menggunakan file .ipynb atau python jupyter notebook sebagai file utamanya. Sudah dijelaskan pada tugas Lab sebelumnya mengenai Data Preparation.

- **Alternatif, menggunakan Google Collab**

  Google Collaboratory adalah sebuah service yang disediakan Google untuk membantu membuka file Jupyter Notebooks. Google Collab membaca file dari Google Drive Anda, sehingga Anda bisa langsung berbagi file dengan teman sekelompok untuk mengerjakan tugas ini. Tautan-tautan berikut dapat membantu Anda menggunakan Google Collab :
  - https://colab.research.google.com/
  - https://towardsdatascience.com/getting-started-with-google-colab-f2fff97f594c

Tutorial
Langkah-langkah pada tutorial dapat dilihat pada file .ipynb yang tersedia pada link ini.

1. PREDIKSI

Lakukan *prediction* terhadap dataset kolom harga, di table **dataset_harga_rumah.csv.zip** menggunakan python. Gunakan data **y**ang sudah dibersihkan dari tugas lab data preparation sebelumnya. Karena tugas ini dilakukan secara kelompok, Anda cukup memilih satu data, yang cleansingnya paling baik.
Ketika sudah melakukan prediksi, Anda dapat menggunakan data ini untuk mengetes akurasi *prediction model* yang sudah Anda buat.

Deskripsi kolom untuk: dataset_harga_rumah.csv.zip

- **diposting_oleh** Kolom yang berisi mengenai value tentang siapakah pemilik property rumahnya: misalnya ber-value: *dealer*
- **sedang_pembangunan:** Sedang dalam masa pembangunan (bernilai 1) atau tidak dalam masa pembangunan (bernilai 0)
- **disetujui_pemerintah** - Menjelaskan mengenai apakah properti atau rumahnya sudah disetujui oleh pemerintah setempat atau belum. Jika sudah disetujui maka bernilai 1, jika belum disetujui maka bernilai 0.
- **total_ruangan -** Total banyaknya ruangan pada suatu property atau rumahnya
- **tipe_a_atau_b -** Tipe dari suatu properti rumah
    - 1 Tipe a = Terdiri atas 1 kamar tidur, 1 ruang tengah / aula, dan 1 dapur
    - 1 Tipe b = Terdiri atas 1 ruangan dan 1 dapur
        - Perbedaan mendasar antara keduanya adalah tidak adanya aula di Tipe b. 1 Tipe a sangat cocok untuk keluarga kecil dan menengah ke bawah karena ada kamar, ruang tengah / aula, dan dapur terpisah sehingga keluarga pada akhirnya akan mengubah aula menjadi ruangan lain saat anak-anak keluarga tumbuh dewasa. Tapi di Tipe b hanya ada kamar dan dapur terpisah tanpa aula yang bisa diubah menjadi kamar lain untuk anak-anak Anda yang sudah dewasa, jadi mungkin hanya bagus untuk para bujangan. Tipe a menawarkan lebih banyak ruang sementara Tipe b terbatas dan biaya Tipe a biasanya lebih tinggi dari Tipe b. Sehingga biasanya disarankan untuk memiliki rumah Tipe a daripada Tipe b.
- **kaki_persegi** - Luas total rumah dalam satuan square feet atau kaki persegi

    - Contohnya: 500 square feet = 46.45152 square meters (https://www.calculateme.com/area/square-feet/to-square-meters/500)
- **siap_pindah** - Kolom yang berisi value tentang apakah rumahnya siap untuk dihuni untuk pindahan (bernilai 1) atau tidak (bernilai 0)
- **dijual_kembali** - Kolom yang berisi value tentang apakah rumah atau property nya merupakan rumah yang dijual kembali (bernilai 1) atau rumah baru (sebelumnya belum pernah terjual) (bernilai 0)
- **alamat** - Alamat dari suatu property atau rumahnya.
- **longitude** - Longitude dari suatu property rumahnya
- **latitude** - Latitude dari suatu property rumahnya
- **harga -** Harga rumah dalam satuan rupee lakh, misalnya jika value = 55, artinya bernilai 5.500.000 rupee India.

Tugas Nomor 1.

Pada bagian prediksi ini, Anda diminta mengerjakan menggunakan **Python.** Pada laporan, jelaskan setidaknya:

- Deskripsi data
- Penjelasan singkat data preparation sebelumnya
- Variabel yang digunakan/tidak digunakan untuk prediksi dan alasannya
- Langkah per langkah pengerjaan dan screen shoot pendukung
- Algoritma yang dipilih (minimal mencoba 3 algoritma), serta akurasi dari algoritma yang Anda dipilih, dibandingkan dengan data yang sudah tersedia.
- Kesimpulan

2. CLUSTERING

Pada clustering ini, kelompok Anda diberikan tiga tabel data, yang ada pada folder **dataset_pesanan_produk.zip**

Skenario singkat untuk memahami dataset_pesanan_produk.zip:

Tentang Fia dan Vilip, pasangan suami istri. Mereka memulai bisnis e-commerce baru mereka selama pandemi pada tahun 2020 dengan menawarkan barang-barang kebutuhan sehari-hari secara online. Mereka mulai dengan menjual bermacam-macam masker dan disinfektan, ternyata secara cepat bisnis mereka berkembang ke berbagai komoditas sehari-hari yang lebih luas.

Deskripsi kolom untuk: dataset_pesanan_produk.zip

Table: produk.csv

Nama Kolom	Deskripsi	Value Range / Format
id_produk	Identifier unik untuk setiap produk	Integer positif [0, ..., 32775]
merek	Merek dari suatu produk	Integer* [-1, ..., 1513]
fitur1	fitur deskriptif (kategorikal)	Integer* [-1, ..., 10]
fitur2	fitur deskriptif (kategorikal)	Integer positif {0, 1, 2, 3}
fitur3	fitur deskriptif (kategorikal)	Integer* [-1, ..., 538]
fitur4	fitur deskriptif (kategorikal)	Integer* {-1, 0, 1, 2, 3, 4}
fitur5	fitur deskriptif (kategorikal)	Integer* [-1, ..., 190]
daftar_kategori	Daftar kategori terkait, berbentuk array	Daftar bilangan integer yang dipisahkan dengan koma

* Nilai -1 = nilai fitur tidak ada

Table: kategori.csv

Nama Kolom	Deskripsi	Value Range / Format
kategori	Identifier kategori unik untuk setiap kategori	Integer positif [0, ..., 4299]
kategori_induk	Kategori yang lebih generik/umum (induk/parent) terkait dengan identifier kategori	Subset Integer Positif dari rentang integer dari kolom 'kategori'

Table: pesanan.csv

Nama Kolom	Deskripsi	Value Range / Format
tanggal	Tanggal transaksi	YYYY-MM-dd dari 2020-06-01 hingga 2021-01-31
id_pengguna	Identifier unik untuk pengguna	Integer positif [0, ..., 46137]
id_produk	Identifier produk dari suatu produk yang sudah terjual	Integer positif [0, ..., 32775]
pesanan	Jumlah atau total pesanan produk	Integer positif

	pada tanggal tertentu oleh pengguna tertentu	

Tugas Nomor 2

Pada bagian clustering ini, Anda diminta mengerjakan menggunakan **Tableau.** Salah satu contoh cara melakukan clustering pada Tableau dapat dilihat di sini. Pada laporan, jelaskan setidaknya:

- Background perlunya melakukan clustering dari sisi bisnis dikaitkan dengan tabel yang tersedia.
- Deskripsi data.
- Penjelasan singkat data preparation yang dilakukan
- Variabel yang digunakan/tidak digunakan untuk clustering dan alasannya
- Langkah pengerjaan dan screen shoot pendukung
- Manfaat yang bisa didapat, dan *what-to-do-next* dari dilakukannya clustering ini.
- Kesimpulan

Semangat!