Guideline Final Project

# Pembagian Dataset dan Kelompok

Setiap kelompok terdiri dari **2 - 4 peserta didik** yang **bebas menentukan** rekan satu kelompok.

Setiap kelompok akan diminta untuk memilih satu dataset yang ingin dikerjakan. Pilihan dataset yang *available* adalah:

* Classification:
  + Data 1: Diabetes Dataset 2019
  + Data 2: Loan Default Dataset
* Regression:
  + Data 3: The Boston Housing Dataset
  + Data 4: Singapore HDB Resale Price

Detil dataset dapat dilihat di bagian dokumen ini setelah bab ‘Guideline Final Project’.

Catatan: setiap kelompok hanya memilih salah satu dari empat dataset di atas, jadi tidak usah memilih satu data klasifikasi dan satu data regresi.

Masing-masing **dataset hanya boleh digunakan 3 (tiga) kelompok,** kecuali untuk **satu dataset** **dapat digunakan oleh** **4 kelompok**.

# Langkah Pengerjaan

Langkah-langkah pengerjaan adalah sebagai berikut:

1. **Daftarkan anggota kelompok Anda** dan pilihan dataset yang ingin diambil melalui satu perwakilan anggota. Pendaftaran dilakukan melalui g-form yang telah disediakan.
2. Bila kelompok telah disetujui, maka kelompok Anda dipersilakan untuk **mengunduh dataset** yang sesuai.
3. Jika ada dataset yang dipilih oleh lebih dari 3 atau 4 kelompok, maka kelompok yang lebih dahulu melakukan pendaftaran dengan dataset tersebutlah yang berhak menggunakan datasetnya.
4. Setiap dataset memiliki ‘*target column*’ yang ingin diprediksi menggunakan informasi-informasi dari kolom-kolom lain pada data tersebut. Tugas kelompok Anda adalah membuat sebuah ***machine learning model* yang mampu memprediksi *target column***.
5. Lakukanlah ***data cleansing*** **apabila diperlukan**, seperti:
   1. Memperbaiki format kolom yang salah!
   2. Memperbaiki *missing values* dan *outlier* dengan cara-cara yang tepat!
6. Lakukanlah **EDA** terlebih dahulu untuk memperoleh pemahaman yang mendalam dari data yang dikerjakan.
7. Buatlah **visualisasi** yang menarik agar:
   1. Anda memperoleh pemahaman yang baik pada data
   2. Dapat digunakan dalam melakukan presentasi tugas akhir
8. Lakukanlah **pemodelan** agar *target column* dapat diprediksi menggunakan kolom-kolom lain yang ada. Boleh tidak menggunakan semua kolom, jika dirasa ada beberapa kolom yang tidak relevan / tidak berkontribusi dengan baik terhadap pemodelan (harus memiliki penjelasan).
9. Pada langkah pemodelan ini, perhatikan beberapa hal:
   1. Apakah model Anda termasuk ***classification* atau *regression***
   2. Bagaimana Anda melakukan ***train test* split**
   3. Bagaimana Anda **mengevaluasi performa** dari model tersebut
   4. **Metrik** apa saja yang Anda gunakan untuk mengevaluasi model tersebut
   5. Apa saja upaya-upaya untuk melakukan ***improvement*** terhadap model awal? Bisa berupa *feature selection*, *feature engineering*, *hyperparameter tuning*, *training* modeldengan teknik *cross validation* atau mencoba model-model lain (misal, mencoba Extreme Gradient Boosting atau Support Vector Machine, dll.)
   6. Wajib setidaknya mencoba 4 model (2 model boleh yang telah dipelajari, dan 2 model lain selain yang pernah diajarkan pada pertemuan sebelumnya). Semakin beragam model yang dieksplorasi, maka semakin bagus!
   7. Model dan *setting*-an *hyperparameter* apa, *feature* apa, serta *feature engineering* apa yang menjadi **pilihan terakhir** Anda, dan mengapa itu dijadikan pilihan akhir Anda?
10. Laporkanlah model akhir Anda, beserta kolom-kolom apa saja yang digunakan dalam model tersebut, serta akurasi atau metrik lain hasil evaluasi dari model Anda (dan bagaimana cara Anda mengukur akurasi model Anda)!
11. Untuk bagian ‘*conclusion’*, siapkanlah poin-poin berikut:
    1. **Interesting Insights** (apa saja pola-pola / trend menarik yang Anda perhatikan pada data?)
    2. **Rekomendasi** kepada *stakeholder / audience* (contoh: jika dataset yang digunakan adalah dataset *churn*, kira-kira langkah apa saja yang dapat dilakukan sebuah perusahaan agar *customer* mereka tidak *churn*? Apa saja faktor-faktor utama yang membedakan *customer* yang *churn* dan tidak *churn*?)
12. Buatlah sebuah **presentasi** (sesuai *template* isi presentasi yang dapat dilihat di link berikut: <https://docs.google.com/presentation/d/18Zz2G_jyy16P7BmMt32pHlGREYCkm_Sh/edit?usp=share_link&ouid=118053588992983113638&rtpof=true&sd=true>) yang nanti akan dipresentasikan pada saat hari presentasi.

Langkah-langkah pengumpulan adalah sebagai berikut:

1. Kumpulkan **semua** file *coding* yang Anda kerjakan (baik itu berupa Python Script, atau Jupyter Notebook), file dataset, dan file *slides* presentasi (**.pdf**). Pastikan semua *file* kodingan dapat di-*run* tanpa *error*!
2. File-file tersebut dikumpulkan dalam file arsip (.zip atau .rar) dengan format nama: **Final\_Project\_Kelompok\_[nama\_kelompok].zip/rar**
3. Unggah file arsip tersebut di gform yang akan disediakan.
4. Deadline pengumpulan file arsip adalah **20 Desember 2024 pukul 23.59 WIB**.

Pilihan Dataset

# Data 3: The Boston Housing Dataset

Link: <https://drive.google.com/file/d/1yfx0w0MnvNQ7ji-liF7p4_W12HeMA-1b/view?usp=sharing>

Problem: **regression**

Context:

The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA. The following describes the dataset columns:

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV - Median value of owner-occupied homes in $1000's

Instruksi spesifik:

1. Kolom yang diprediksi adalah *‘*MEDV*’*
2. Insights yang dapat digali meliputi faktor-faktor apa saja yang memengaruhi nilai median harga rumah. Apakah tingkat kriminalitas, pajak, akses terhadap jalan raya, dll. dapat mempengaruhi harga rumah? Faktor-faktor apa saja yang perlu diperhatikan oleh stackholder / audience untuk mempertimbangkan harga rumah jika ingin membeli atau menjual rumah?

Berikut penjelasan untuk setiap kolom dalam dataset Boston Housing:

1. **CRIM** - **Tingkat Kejahatan per Kapita per Kota**: Ini menunjukkan tingkat kejahatan yang terjadi per kapita di masing-masing wilayah kota. Angka yang lebih tinggi menunjukkan tingkat kejahatan yang lebih tinggi.
2. **ZN** - **Proporsi Tanah Permukiman yang Dizonasi untuk Lot di Atas 25.000 sq.ft.**: Ini mengindikasikan proporsi tanah yang diizinkan untuk pembangunan rumah dengan ukuran lebih besar dari 25.000 kaki persegi. Tanah yang lebih banyak zonasi untuk lot besar cenderung berkaitan dengan wilayah yang lebih sejahtera.
3. **INDUS** - **Proporsi Tanah Bisnis Non-Ritel per Kota**: Ini mengukur proporsi tanah yang digunakan untuk keperluan bisnis non-ritel, seperti industri dan pabrik, dalam suatu kota.
4. **CHAS** - **Variabel Dummy Charles River**: Ini adalah variabel biner (0 atau 1) yang menunjukkan apakah wilayah kota terletak di dekat sungai Charles. Nilai 1 berarti kota terletak di dekat sungai, sedangkan 0 berarti tidak.
5. **NOX** - **Konsentrasi Nitrogen Oksida (dalam bagian per 10 juta)**: Mengukur kadar polusi udara yang disebabkan oleh nitrogen oksida, yang dapat memengaruhi kualitas udara dan kesehatan.
6. **RM** - **Rata-rata Jumlah Kamar per Tempat Tinggal**: Ini menunjukkan rata-rata jumlah kamar tidur di rumah-rumah dalam wilayah tersebut. Rumah dengan lebih banyak kamar cenderung lebih besar dan lebih mahal.
7. **AGE** - **Proporsi Rumah yang Dibangun Sebelum 1940**: Ini mengukur seberapa banyak rumah di wilayah tersebut yang dibangun sebelum tahun 1940. Rumah yang lebih tua mungkin membutuhkan perawatan lebih banyak.
8. **DIS** - **Jarak Terhitung ke Lima Pusat Pekerjaan di Boston**: Ini menunjukkan jarak tertimbang dari wilayah ke pusat-pusat pekerjaan utama di Boston. Semakin dekat ke pusat kota, semakin mudah aksesibilitasnya.
9. **RAD** - **Indeks Aksesibilitas ke Jalan Raya Radial**: Ini mengukur sejauh mana suatu wilayah terhubung dengan jalan raya utama yang mengarah ke pusat kota, yang menunjukkan kemudahan akses transportasi.
10. **TAX** - **Tarif Pajak Properti Berdasarkan Nilai Penuh per $10.000**: Menunjukkan tingkat pajak properti di wilayah tersebut. Angka yang lebih tinggi menunjukkan tarif pajak properti yang lebih tinggi.
11. **PTRATIO** - **Rasio Siswa-Guru per Kota**: Mengukur rasio jumlah siswa terhadap jumlah guru di sekolah-sekolah kota tersebut. Rasio yang lebih rendah mengindikasikan lebih banyak perhatian dan sumber daya yang diberikan kepada siswa.
12. **B** - **Proporsi Penduduk Kulit Hitam per Kota**: Ini mengukur proporsi penduduk kulit hitam di suatu wilayah dengan rumus khusus: B=1000×(Bk−0.63)2B = 1000 \times (Bk - 0.63)^2B=1000×(Bk−0.63)2, di mana BkBkBk adalah proporsi warga kulit hitam di kota tersebut.
13. **LSTAT** - **Persentase Penduduk dengan Status Sosial Ekonomi Rendah**: Ini mengukur persentase penduduk dengan status sosial ekonomi rendah. Wilayah dengan nilai lebih tinggi menunjukkan ketimpangan sosial yang lebih besar.
14. **MEDV** - **Nilai Median Rumah yang Dimiliki oleh Pemilik dalam $1000**: Ini adalah target variabel yang menunjukkan nilai median rumah yang dimiliki oleh pemilik di wilayah tersebut, diukur dalam ribuan dolar. Variabel ini sering digunakan untuk prediksi harga properti.