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: Mobile Price Classification
  + Data 2: Loan Default
* Regression:
  + Data 3: Walmart 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**.

# 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 *submission* “Registrasi Kelompok” di e-learning.
2. Bila kelompok telah disetujui, maka kelompok Anda dipersilakan untuk **mengunduh dataset** yang sesuai. Persetujuan dapat dilihat pada *feedback submission* registrasi (jika disetujui maka akan mendapatkan grade >= 80).
3. Jika ada dataset yang dipilih oleh lebih dari 3 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 **EDA** terlebih dahulu untuk memperoleh pemahaman yang mendalam dari data yang dikerjakan.
6. Lakukanlah ***data cleansing*** **apabila diperlukan**, seperti:
   1. Memperbaiki format kolom yang salah
   2. Memperbaiki *missing values* dan *outlier* dengan cara-cara yang tepat
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. Tidak usah menggunakan semua kolom, jika dirasa ada beberapa kolom yang tidak relevan / tidak berkontribusi dengan baik terhadap pemodelan.
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*, *hyperparameter tuning*, *feature engineering*, atau mencoba model-model lain (misal, mencoba Linear Regression, dan Random Forest Regression, dll.)
   6. Wajib setidaknya mencoba satu model lain selain yang pernah diajarkan pada pertemuan-pertemuan sebelumnya. Semakin beragam model yang dieksplorasi, maka semakin bagus
   7. Model apa dan *setting*-an *hyperparameter* apa yang menjadi **pilihan terakhir** Anda, dan mengapa model 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 dalam format **.pdf**.
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 e-learning pada *submission* “Submission Final Project”.
4. Deadline pengumpulan file arsip adalah **28 Mei 2023 pukul 23.59 WIB**.

Pilihan Dataset

# Data 1: Mobile Price Classification

Link: <https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification>

Problem: **classification**

Instruksi spesifik:

1. Download **hanya** ‘***train.csv***’ saja, karena ‘test.csv’ dia tidak memiliki *target column*, jadi kita tidak bisa melakukan evaluasi pada *test.csv* yang diberikan (tidak bisa mengecek apakah hasil prediksi kita benar atau salah). Jadi, anggaplah *train.csv* sebagai **keseluruhan data**, yang nanti dibagi menjadi *train-test-split*.
2. Target column adalah kolom paling kanan (terakhir), yaitu *‘price\_range*’
3. Insights yang dapat digali meliputi faktor apa saja yang paling memengaruhi *price\_range* dari sebuah *smartphone*? (Fitur apa saja yang paling membedakan antara *smartphone* murah dan *smartphone* mahal)

# Data 2: Loan Default Dataset

Link: <https://www.kaggle.com/datasets/yasserh/loan-default-dataset>

Problem: **classification**

Instruksi spesifik:

1. Target column adalah kolom dengan nama ‘*Status*’
2. Insights yang dapat digali meliputi faktor apa saja yang paling membedakan antara nasabah yang akan gagal bayar, atau nasabah yang akan berhasil bayar. Selain itu, dari sekian banyak kolom prediktor yang ada, faktor apa saja yang perlu diperhatikan perbankan agar mereka tidak salah dalam memberikan *loan* kepada nasabah?
3. Perhatikan baik-baik definisi *False Positive, False Negative*, serta penggunaan metrik-metrik seperti *precision, recall, dst*.

# Data 3: Walmart Dataset

Link: <https://www.kaggle.com/datasets/yasserh/walmart-dataset>

Problem: **regression**

Instruksi spesifik:

1. Kolom yang diprediksi adalah *‘Weekly\_Sales’*
2. Insights yang dapat digali meliputi faktor-faktor apa saja yang memengaruhi penjualan mingguan. Perlu dicari tahu lebih lanjut pula apakah faktor tanggal juga memengaruhi penjualan mingguan. Feature engineering yang dapat dilakukan meliputi memecah kolom ‘Date’ ke dalam 4 kolom baru:
   1. Date (tanggal numerik)
   2. Day (hari, i.e. Senin, Selasa, Rabu, dst)
   3. Month (bulan)
   4. Year (tahun)

# Data 4: Singapore HDB Resale Price

Link: <https://www.kaggle.com/datasets/teyang/singapore-hdb-flat-resale-prices-19902020>

Problem: **regression**

Instruksi spesifik:

1. **Filter lah data sehingga Anda hanya menggunakan data dari tahun 2016 - 2020. Buang data dari tahun 1990 - 2015.**
2. Kolom yang diprediksi adalah ‘*resale\_price*’
3. Insights yang dapat diambil meliputi properti di daerah mana saja yang *paling worth it*, faktor apa saja yang paling memengaruhi tinggi rendahnya harga properti, dan pergolakan trend serta volume transaksi properti dari tahun 2016 sampai tahun 2020.