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 e-learning pada *submission* “Submission Final Project”.
4. Deadline pengumpulan file arsip adalah **6 Desember 2023 pukul 23.59 WIB**.

Pilihan Dataset

# Data 1: Diabetes Dataset 2019

Link: <https://drive.google.com/file/d/14ySas0DRD6eJ7ysw22JSN2G_OZ5tLHlR/view?usp=sharing>

Problem: **classification**

Context:

This dataset was collected by Neha Prerna Tigga and Dr. Shruti Garg of the Department of Computer Science and Engineering, BIT Mesra, Ranchi-835215 for research, non-commercial purposes only. An article is also published implementing this dataset. For more information for this dataset please refer:

Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. Procedia Computer Science, 167, 706-716. DOI: https://doi.org/10.1016/j.procs.2020.03.336

Content:

There is a total of 952 instances with 17 independent predictor variables and one binary target or dependent variable, Diabetes.

Instruksi spesifik:

1. Target column adalah kolom *‘Diabetic*’.
2. Insights yang dapat digali meliputi faktor apa saja yang memengaruhi penyandang diabetes atau tidak. Faktor seperti apa yang dapat dipertimbangkan oleh pemerintah / *stackholder* untuk menghindari penyakit diabetes?
3. Perhatikan bagaimana memperlakukan kasus data dengan jumlah kelas tidak seimbang (*imbalance data problem*). Apakah menggunakan metrik evaluasi khusus atau dengan mengatasi permasalahan *imbalance data* tersebut!

# Data 2: Loan Default Dataset

Link: <https://drive.google.com/file/d/1msCuf24AxJqAwYi8P_sarWUAfelAIXF8/view?usp=sharing>

Problem: **classification**

**Description**:

Banks earn a major revenue from lending loans. But it is often associated with risk. The borrower's may default on the loan. To mitigate this issue, the banks have decided to use Machine Learning to overcome this issue. They have collected past data on the loan borrowers & would like you to develop a strong ML Model to classify if any new borrower is likely to default or not.

The dataset is enormous & consists of multiple deteministic factors like borrowe's income, gender, loan pupose etc. The dataset is subject to strong multicollinearity & empty values. Can you overcome these factors & build a strong classifier to predict defaulters?

**Objective**:

* Understand the Dataset & cleanup (if required).
* Build classification model to predict weather the loan borrower will default or not.
* Also fine-tune the hyperparameters & compare the evaluation metrics of vaious classification algorithms.

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: 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?

# Data 4: Singapore HDB Resale Price

Link: <https://drive.google.com/file/d/15Xwkyf5IkQiCoNP1xbLgNbOKwfNi2P73/view?usp=sharing>

Problem: **regression**

**Context**

Every day in the small country of Singapore, people are struggling to purchase houses due to a number of factors, but primarily because of Singapore's limited land. Prices are competitive and new houses and flats are very expensive. Thus, many people have turned to purchasing resale Housing Development Board (HDB) flats. Researching into the pricing and drivers of these HDB resale prices is therefore important, not just for individuals or families wanting to purchase a new house, but also for analysts and economists trying to understand the stability and performance of this sector.

**Content**

The dataset contains 2 CSV files, which covers transactions from 2015 - 2020. It includes features of the flat and sale, such as the year of sale, location of the flat, flat type, street name, block number, area of the flat, lease and resale price.

**Notes**

* The approximate floor area includes any recess area purchased, space adding item under HDB’s upgrading programmes, roof terrace, etc.
* The transactions exclude resale transactions that may not reflect the full market price such as resale between relatives and resale of part shares.
* Resale prices should be taken as indicative only as the resale prices agreed between buyers and sellers are dependent on many factors.
* Remaining lease is the number of years, months and days left before the lease expires. This information is computed as at the resale flat application and has been rounded up to the nearest month for the purpose of CPF monies usage and HDB loan application.

**Acknowledgements**

The HDB resale price data was downloaded from Data.gov.sg, accessed on 4 November 2020 from Data.gov.sg which is made available under the terms of the Singapore Open Data Licence version 1.0 https://data.gov.sg/open-data-licence.

**Inspiration**

Explore the data and help improve the understanding of Singapore's housing industry!

Instruksi spesifik:

1. **Filter lah data sehingga Anda hanya menggunakan data dari tahun 2016 - 2020.**
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.