CAPSTONE PROJECT PROPOSAL
HOUSE PRICE PREDICTION

0. TEAM MEMBERS:

- Nguyễn Văn Thanh Tùng – 20190090
- Nguyễn Huy Hoàng – 20194433
- Nguyễn Mậu Trà – 20200624
- Nguyễn Thị Linh – 20200349

1. PROBLEM DESCRIPTION

Hanoi is the capital city of Vietnam, as well as one of the largest economic centers in the country. Therefore, many people, especially fresh graduates, are looking for a job opportunity in the city, which leads to problems involving accommodation. However, predicting house prices can help to determine the selling price of a house in a particular region and can help people to find the right time to buy a home. In this project on House Price Prediction, our task is to **predict house prices in Hanoi using different approaches**.

2. DATASET DESCRIPTION: VIETNAM HOUSING DATASET (HANOI)

Source: https://www.kaggle.com/code/kwonhoang/predicting-hanoi-housing-price-ann-rf/data

This is a raw dataset which is a set of house prices in Hanoi, Vietnam taken from 23/05/2020 to 05/08/2020.

- Dataset characteristic: Multivariate
- Attribute characteristics: Integer, Real, String, Date
- Number of columns: 13
- Number of rows: 82497

There are 12 attributes in each record of dataset:

- Ngày (date): the time the house was offered for sale
- Địa chỉ (address): detail address of the house (Street, Ward, District)
- Quận (district): the district the house is located in
- Phường (ward): the ward the house is located in

- Loại hình nhà ở (type of accomodation): base on the location and the design the houses, they are split into 4 categories: villa, front house, alley house, townhouse
- Giấy tờ pháp lý (legal papers)
- Số tầng (number of floors)
- Số phòng ngủ (number of bedrooms)
- Diện tích (area)
- Dài (length)
- Rộng (width)
- Giá (price): the price of the house in Vietnam Dong per meter square. ***This is the response value that we have to predict.***

3. INPUT, OUTPUT, METRIC DESCRIPTION:

- **Input:** Representation of the features of a house (a vector of considerable features).
- **Output:** A predicted price for the house (VND/m^2).
- **Metric:**
    - Mean Squared Error (MSE).
    - Mean Absolute Error (MAE).

4. ALGORIHM & APPROACH PROPOSAL

- Approach 1 (simple approach): build multiple regression models (linear regression, decision tree, random forest, neural network) to choose the best one.
- Approach 2 (extended approach):
    - 1st step: apply a K-mean clustering algorithm to figure out some kind of similarity or relation between all the data points in the dataset.
    - 2nd step: apply regression models for each of the clusters to find the corresponding optimal regression function.
    - 3rd step: in the testing phase, classify the test data point into one of the clusters defined, and then apply the corresponding regression function learned to inference.