CAPSTONE PROJECT PROPOSAL

HOUSING PRICE PREDICTION

Advisor: Prof. Than Quang Khoat

Hanoi University of Science and Technology, Vietnam

0 TEAM MEMBERS:

- Nguyen Van Thanh Tung – 20190090
- Nguyen Huy Hoang – 20194433
- Nguyen Mau Tra – 20200624
- Nguyen Thi Linh – 20200349

1 PROBLEM DESCRIPTION

House price prediction is an important concept in the real estate industry and has been a popular problem in research for years since the traditional house price prediction depend on cost and sale price comparison does not satisfy the accepted standards and certification process. In addition to getting accurate prediction, it is important to know the factors that have a significant impact on the house price. In this project on House Price Prediction, our task is to **predict house prices in Boston using different approaches**.

2 DATASET DESCRIPTION: BOSTON HOUSING DATASET

- Number of Instances: 506
- Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.
- Attribute Information (in order):
  - CRIM per capita crime rate by town
  - ZN proportion of residential land zoned for lots over 25,000 sq.ft.
  - INDUS proportion of non-retail business acres per town
  - CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
  - NOX nitric oxides concentration (parts per 10 million)
  - RM average number of rooms per dwelling
  - AGE proportion of owner-occupied units built prior to 1940
  - DIS weighted distances to five Boston employment centres
  - RAD index of accessibility to radial highways
  - TAX full-value property-tax rate per $10,000
  - PTRATIO : pupil-teacher ratio by town
  - B 1000(Bk - 0.63)ˆ2 where Bk is the proportion of black people by town
  - LSTAT % lower status of the population
  - MEDV Median value of owner-occupied homes in $1000's
- Missing Attribute Values: None

3 INPUT, OUTPUT, METRIC DESCRIPTION:

- **Input:** Representation of the features of a house (a vector of considerable features).
- **Output:** A number (MEDV) represents the mean value of owner-occupied home in that neighborhood.
- **Metric:**
  - R2 Score.
  - Mean Squared Error (MSE).
  - Mean Absolute Error (MAE).

4 ALGORITHM & 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.