

# Dự Đoán Giá Nhà Bằng Machine Learning

- Nguyễn Tuấn Đạt, Châu Hải Đăng, Trần Đại Thắng, Nguyễn Xuân Đăng -

## Problem Review

Thuộc dạng bài toán hồi quy (Regression)

Input:

- Dữ liệu nhà cửa và khu vực xung quanh gồm thông tin cá nhân từ **Kaggle House Prices - Advanced Regression Techniques Dataset**.

Output:

- Dự đoán giá nhà.

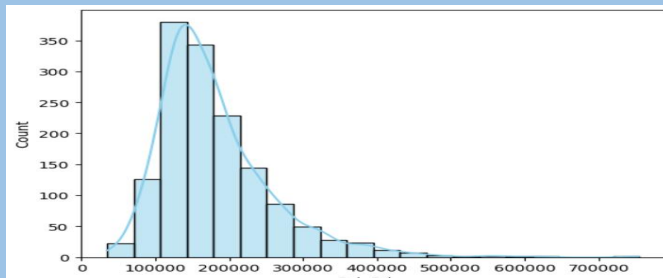
Notice:

- Tập dữ liệu bị thiếu, mất mát và nhiễu. Cần tiền xử lý và phân tích đặc trưng.

## Dataset

Source:

- Kaggle - House Prices - Advanced Regression Techniques.  
<https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data>

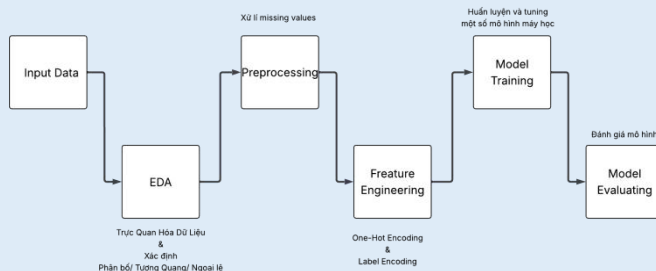


Mô tả:

- Gồm 1460 bộ train và 1459 bộ test.
- Dữ liệu output tập train phân bố lệch về bên trái.
- Có bao gồm dữ liệu về thời gian

Loại Dữ Liệu	Số lượng
Số	37
Danh mục	43

## Propose Method



## Exploring Data Analysis

Inspecting Data:

Loại Đặc Trưng	Cân Chú Ý
Danh Mục	Có dữ liệu thiếu, gồm các giá trị Nan hợp lệ nhưng ảnh hưởng đến việc xác định các dữ liệu thiếu.
Thời gian	Do sự phân bố giữa tháng 1 và 12 theo số học xa nhau nhưng trên thực tế là gần nhau nên sẽ ảnh hưởng đến học máy.
Các giá trị số còn lại	Có dữ liệu thiếu và Nan. Có độ phân bố rộng và xiên.

## Preprocessing

Misscorrected & Missing Datas:

Loại Đặc Trưng	Phương Thức xử lý
Danh Mục	Mode
Thời gian	Đưa về 0 đối với các dữ liệu cho là không có.
Các giá trị số còn lại	Theo Median hoặc fill 0.

Features Engineering:

Phương thức	Thay đổi
Tổng hợp các đặc trưng số mật thiết	Tạo nên các đặc trưng số mới có ý nghĩa như TotalArea, TotalBaths, TotalPorch
Chiết xuất đặt trưng binary có ý nghĩa	Thêm Pool, 2ndFloor, Garage, Bsmt, Fireplace, Porch
Chuyển đổi đặc trưng số về danh mục	Loại số về danh mục MSSubClass, YrSold
Chuẩn hóa dữ liệu thẳng theo sin, cos	Chuẩn hóa MoSold thành MoSoldsin, MoSoldcos
One-hot tất cả các danh mục đã tạo	Tổng thể có 330 thuộc tính sau one-hot.

Skewness Reduction:



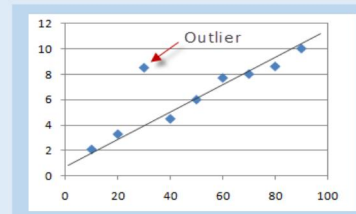
Standardization

$$z = \frac{x - \mu}{\sigma}$$

$\mu$  = Mean  
 $\sigma$  = Standard Deviation

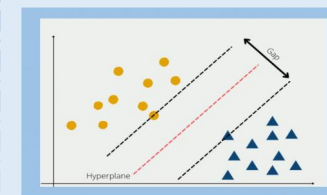
Áp dụng phân phối chuẩn cho việc chuẩn hóa dữ liệu.

Outlier Detection

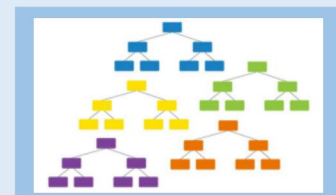


## Model Training

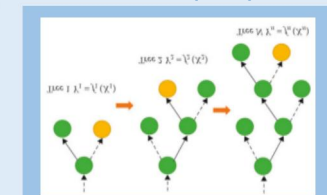
Proposed Regression models:



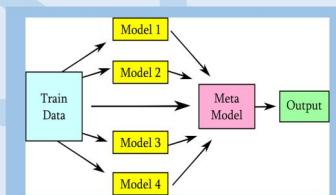
- Support Vector Machine (SVM).



- Random Forest



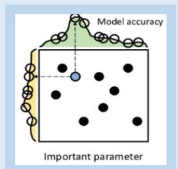
- Gradient boosting



Stacked Model

Tuning Method:

- Kết hợp giữa K-Fold và RandomSearch Cho các hyperparameters của các model.



## Model Evaluating

- K-Fold Cross validation by Mean Square Error Evaluating.

Models	Baseline	Optimized	Kaggle
Support Vector Machine	0.1153	0.1065	0.12711
Random Forest	0.1278	0.1216	0.14036
Gradient boosting	0.1150	0.1068	0.12303
Stacked Model (Optimized)			0.12226

## Result

exp1.csv Complete - now	0.12226
----------------------------	---------