Phân tích, đánh giá, dự đoán giá nhà

### 1. Giới thiệu

* Mục tiêu: Đưa ra dự đoán giá nhà dựa trên các yếu tố ảnh hưởng như diện tích, vị trí, số phòng ngủ, năm xây dựng, v.v.
* Ứng dụng: Hỗ trợ các bên mua bán, các nhà đầu tư và các công ty bất động sản trong việc ra quyết định.
* Mục tiêu của bài toán là xây dựng một mô hình dự đoán giá nhà dựa trên các đặc trưng của dữ liệu nhà ở. Mô hình sử dụng hồi quy tuyến tính sau khi đã giảm chiều dữ liệu bằng phương pháp Truncated Singular Value Decomposition (SVD).

#### **2. Phương pháp và Quy trình**

##### **2.1. Chuẩn bị dữ liệu**

**Dữ liệu đầu vào**: Bộ dữ liệu đầu vào df\_data\_house.csv chứa các thông tin về nhà ở với nhiều đặc trưng số(diện tích,phòng ngủ,....).

**Biến mục tiêu**: Biến mục tiêu là price, đại diện cho giá nhà mà chúng ta cần dự đoán.

**Xử lý dữ liệu**:

+Các đặc trưng số học đã được chuẩn hóa bằng **StandardScaler** để đảm bảo tính đồng nhất trong phạm vi của các giá trị.

+Dữ liệu đã được chia thành hai tập: **tập huấn luyện** (training set) và **tập kiểm tra** (test set), với tỷ lệ 80-20.

##### **2.2. Áp dụng SVD**

-Để giảm số lượng đặc trưng và tối ưu hóa quá trình huấn luyện, chúng ta sử dụng phương pháp **Giảm chiều SVD** (Singular Value Decomposition). SVD giúp giảm số lượng thành phần chính cần thiết để duy trì khả năng giải thích dữ liệu, từ đó giảm độ phức tạp của mô hình.

-SVD đã giảm chiều dữ liệu xuống chỉ còn **15 thành phần chính**.

### 3.Xây Dựng Mô Hình Hồi Quy Tuyến Tính OLS

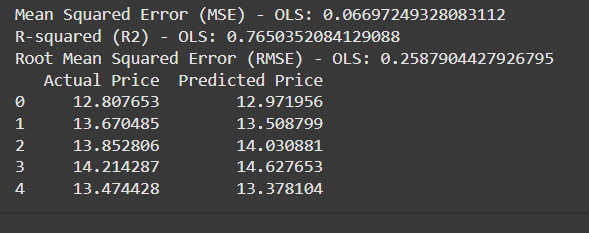
Dữ liệu được chia thành tập huấn luyện và tập kiểm tra với tỷ lệ 80:20.

**Mô hình OLS (Hồi Quy Tuyến Tính)** được huấn luyện trên các thành phần chính đã được giảm chiều qua SVD.

Mô hình **LinearRegression** của scikit-learn được sử dụng để thực hiện hồi quy tuyến tính mà không có regularization.

Hằng số được thêm tự động trong quá trình huấn luyện mà không cần thêm cột chứa giá trị 1 vào dữ liệu.

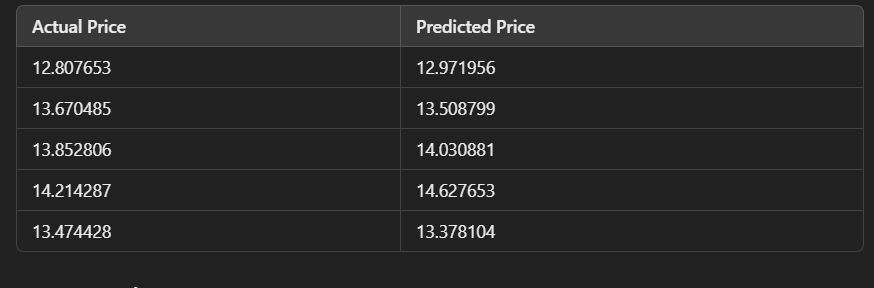
### 4.Kết quả mô hình

****

Sau khi huấn luyện mô hình, chúng ta tiến hành đánh giá mô hình trên tập kiểm tra bằng các chỉ số sau:

* **Mean Squared Error (MSE)**: 0.06697  
  Đây là một giá trị khá thấp, cho thấy mô hình đã học tốt và sai số giữa giá trị thực tế và giá trị dự đoán là rất nhỏ.
* **Root Mean Squared Error (RMSE)**: 0.2588  
  RMSE cho thấy độ lệch giữa giá trị dự đoán và giá trị thực tế là rất nhỏ, giúp ta đánh giá được mức độ chính xác của mô hình.
* **R-squared (R²)**: 0.7650  
  Mô hình giải thích được khoảng **76.5%** sự biến thiên trong dữ liệu, cho thấy mô hình có khả năng dự đoán tương đối chính xác.

### 5.Dữ Liệu Thực Tế và Dự Đoán

****

### 6.Dữ Liệu Thực Tế và Dự Đoán

-Mô hình OLS với **SVD** đã cung cấp một kết quả tốt cho bài toán dự đoán giá nhà. Chỉ số **R²** đạt **76.5%**, cho thấy mô hình có khả năng giải thích phần lớn sự biến thiên của giá nhà dựa trên các đặc trưng đầu vào.

**-MSE =**0.06697249328083112 và **RMSE=**0.2587904427926795 thấp chứng tỏ mô hình đang hoạt động hiệu quả với sai số dự đoán nhỏ.

### 7.Link data và link code

-Link code:[Untitled0.ipynb](https://colab.research.google.com/drive/1VxcMzvYJZ79UiAYZCD89jJ-W1Y5x8Lsm?usp=sharing)

-Link data:[link\_data](https://drive.google.com/drive/folders/1V0b508Le6iexJcscbGmrj8Ws-LnzIMsn?usp=sharing)