TRƯỜNG CAO ĐẲNG CÔNG NGHỆ   
THỦ ĐỨC

----- □ & □ -----



Khoa Công Nghệ Thông Tin

**ĐỒ ÁN LẬP TRÌNH PYTHON NÂNG CAO**

**Giảng viên:** Phan Thị Thể

|  |  |  |
| --- | --- | --- |
| **STT** | **Họ và tên** | **MSSV** |
| 1 | Lê Thanh Sơn | 21211TT4174 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**2023**

**CHƯƠNG 1.GIỚI THIỆU VỀ THUẬT TOÁN LINEAR REGRESSION**

**Linear Regression (Hồi quy tuyến tính)** là một trong những thuật toán cơ bản và phổ biến nhất của **Supervised Learning (Học có giám sát)**, trong đó **đầu ra dự đoán là liên tục**. Thuật toán này thích hợp để dự đoán các giá trị đầu ra là các đại lượng liên tục như **doanh số hay giá cả** thay vì cố gắng phân loại chúng thành các đại lượng rời rạc như màu sắc và chất liệu của quần áo, hay xác định đối tượng trong một bức ảnh là mèo hay là cún.

Linear Regression được ứng dụng rộng rãi trong nhiều lĩnh vực, như: Dự báo giá cả: dự đoán giá nhà, giá cổ phiếu, giá nhiên liệu dựa trên các yếu tố như vị trí, kích thước, chất lượng, lượng cung cầu, … Dự báo điểm số: dự đoán điểm số của học sinh dựa trên thời gian học, nỗ lực, kỹ năng, trình độ giáo viên

**Chương 2. Mô tả đồ án**

**2.1 Giới thiệu**

**Housing Price Prediction Data** là dữ liệu dự đoán giá nhà có thể giúp chủ đầu tư xác định giá bán căn nhà và giúp khách hàng sắp xếp thời điểm thích hợp để mua nhà. Có ba yếu tố ảnh hưởng đến giá của một ngôi nhà bao gồm điều kiện vật chất, ý tưởng và vị trí.ở tập dữ liệu này cung cấp một bộ tính năng đa dạng, bao gồm diện tích, phòng ngủ, phòng tắm, loại khu vực ,năm xây dựng và giá của căn nhà. Dự án này giúp ta thử thách dự đoán giá nhà và nâng cao kỹ năng phân tích hồi quy của chúng ta.

**2.1** **Dữ liệu đầu vào**

Dữ liệu đầu vào được lưu trong tệp "housing\_price\_dataset.csv". Dữ liệu này chứa thông tin bao gồm diện tích, phòng ngủ, phòng tắm, loại khu vực ,năm xây dựng và giá . Dưới đây là một số điểm chính về dữ liệu:

Trong file bao gồm housing\_price\_dataset.csv.  
- SquareFeet: Diện tích của nhà được xây dựng.

- Bedroom: Phòng tắm của căn nhà .

- Bathroom: Phòng ngủ của căn nhà.

- Neighborhood: Ở đây chỉ là khu vực vùng(nông thôn , đô thị , ngoại ô, thành phố).

- Year Built: Năm xây dựng nhà x.

- Price: Giá của căn nhà.

## **2.3 Vấn đề cần giải quyết**

Vấn đề cần giải quyết trong dự án Housing Price Prediction Data là xây dựng một mô hình dự đoán giá nhà dựa trên các yếu tố như diện tích, số phòng ngủ, số phòng tắm, loại khu vực, năm xây dựng để cung cấp khả năng ước tính giá nhà một cách chính xác.

Để giải quyết vấn đề này, bạn cần:

1. **Phân tích và tiền xử lý dữ liệu**: Đánh giá dữ liệu, kiểm tra các giá trị thiếu, xử lý outliers nếu có, chuẩn hóa dữ liệu để sử dụng trong việc xây dựng mô hình.

2. **Chọn và huấn luyện mô hình**: Sử dụng các thuật toán học máy phù hợp như hồi quy tuyến tính, hồi quy đa biến, hoặc các mô hình phức tạp hơn như Random Forests, Gradient Boosting để huấn luyện mô hình dự đoán giá nhà.

3. **Đánh giá hiệu suất của mô hình**: Sử dụng các phép đo như RMSE, MAE, R-squared để đánh giá mức độ chính xác của mô hình.

4. **Tinh chỉnh và cải thiện mô hình**: Điều chỉnh tham số của mô hình, xem xét các cải tiến để tăng độ chính xác của dự đoán.

5. **Áp dụng mô hình vào dự đoán thực tế**: Khi mô hình đã được đào tạo và kiểm tra, sử dụng nó để dự đoán giá nhà mới dựa trên thông tin đầu vào như diện tích, số phòng ngủ, vị trí, vv.

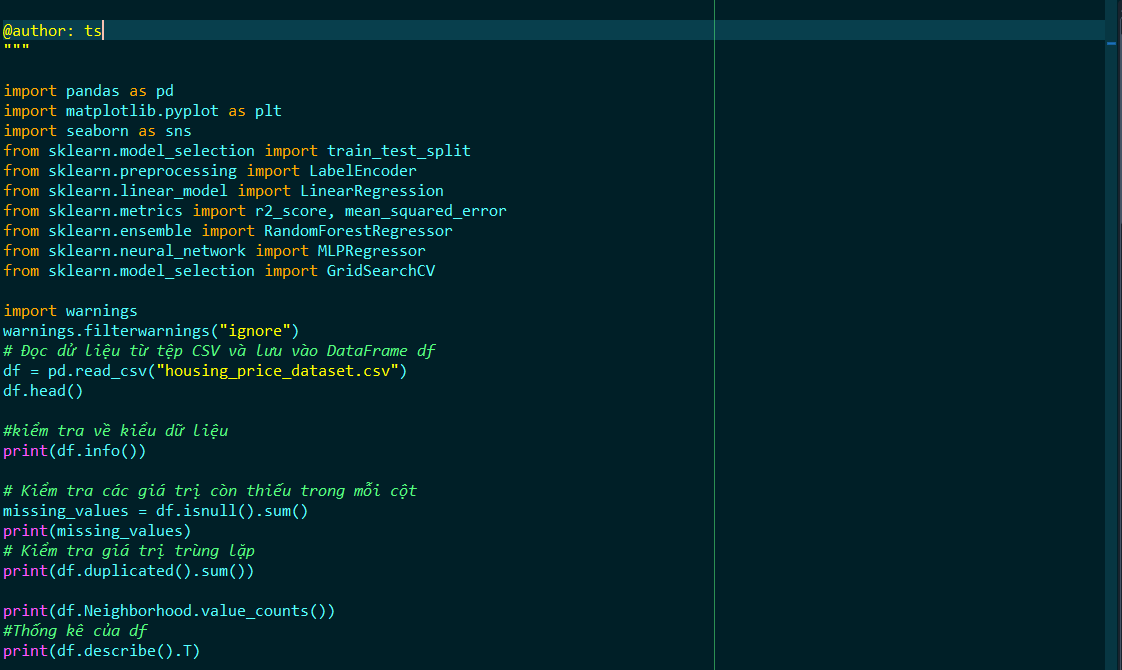
Với việc giải quyết vấn đề này, bạn sẽ có cơ hội nâng cao kỹ năng phân tích hồi quy và cung cấp một công cụ hữu ích cho chủ đầu tư và người mua bất động sản để đưa ra quyết định thông minh.

Tôi sử dụng Hồi Quy Tuyến Tính để mô tả mối quan hệ tuyến tính giữa các biến đọc lập và biến phụ thuộc, qua đó mô hình được huấn luyện trên tập dữ liệu.

Thuật toán hồi quy tuyến tính là một phương pháp trong thống kê để mô hình hóa mối quan hệ tuyến tính giữa một biến độc lập và một biến phụ thuộc. Mục tiêu của thuật toán là tìm ra đường hồi quy tuyến tính sao cho tổng bình phương của sự chênh lệch giữa các giá trị dự đoán và giá trị thực tế của biến phụ thuộc là nhỏ nhất.

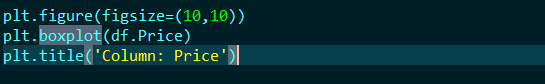
## **Mô tả các hàm chức năng**

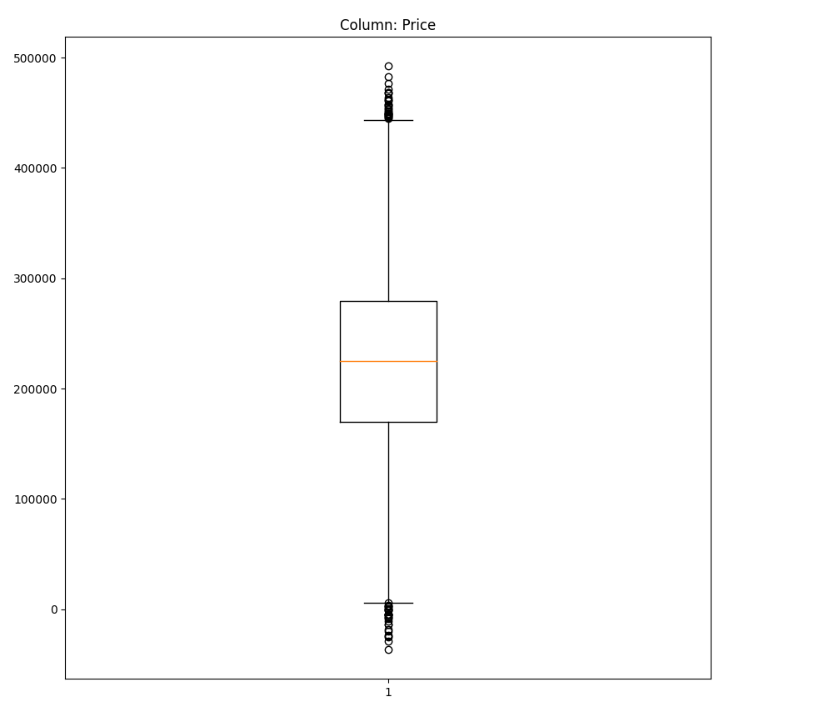
Dữ liệu đầu vào được đọc từ tệp ‘housing\_price\_dataset.csv ’và kiểm tra về kiểu dữ liệu, giá trị còn thiếu, giá trị trùng lặp.



## **Biểu Đồ Hộp (Boxplot)**

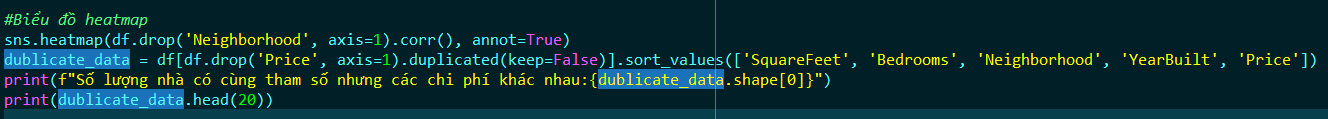
Biểu đồ hộp dùng để hiển thị phân phối giá trị của căn nhà.





## **Biểu Đồ Heatmap**

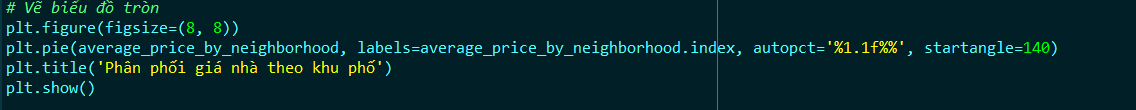
Dùng để hiển thị mức độ tương quan giữa diện tích, số phòng ngủ, số phòng tắm, năm xây dựng và giá nhà. Điều này giúp bạn nhận ra những đặc trưng nào có mối tương quan cao với giá nhà, từ đó có thể chọn ra các đặc trưng quan trọng để sử dụng trong mô hình dự đoán.

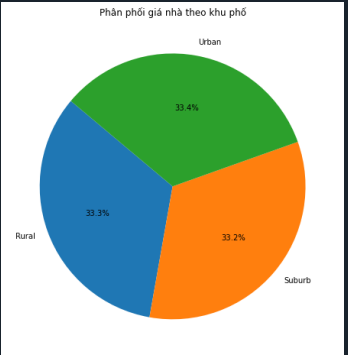




## **Biểu Đồ Tròn**

Biểu đồ tròn này dùng để phân phối giá của nhà theo từng khu phố.

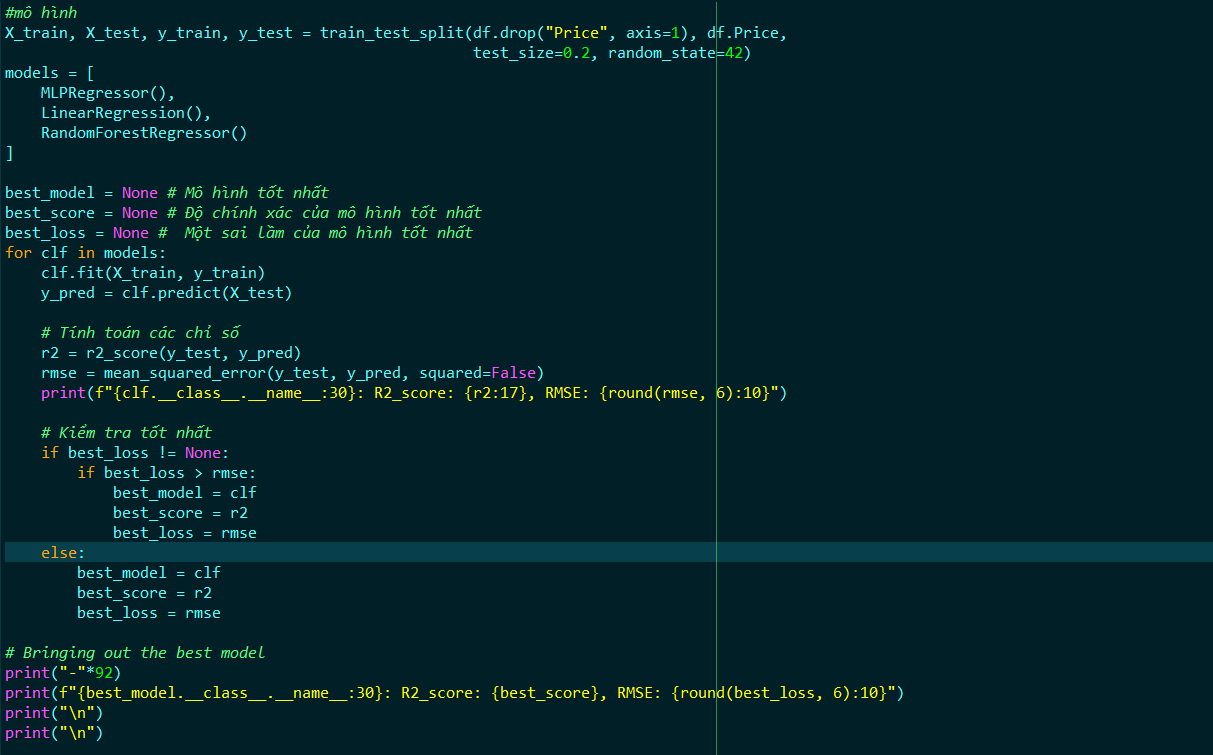




* 1. **Xây Dựng Mô Hình**

Chúng ta sử dụng Hồi Quy Tuyến Tính để mô tả mối quan hệ tuyến tính giữa các biến đọc lập và biến phụ thuộc, qua đó mô hình được huấn luyện trên tập dữ liệu.

Thuật toán hồi quy tuyến tính là một phương pháp trong thống kê để mô hình hóa mối quan hệ tuyến tính giữa một biến độc lập và một biến phụ thuộc.



Kết quả in ra ban đầu:

MLPRegressor : R2\_score: 0.5222290164310276, RMSE: 52221.683671

LinearRegression : R2\_score: 0.5424884689862141, RMSE: 51102.483516

RandomForestRegressor : R2\_score: 0.4555318992061206, RMSE: 55747.732165

--------------------------------------------------------------------------------------------

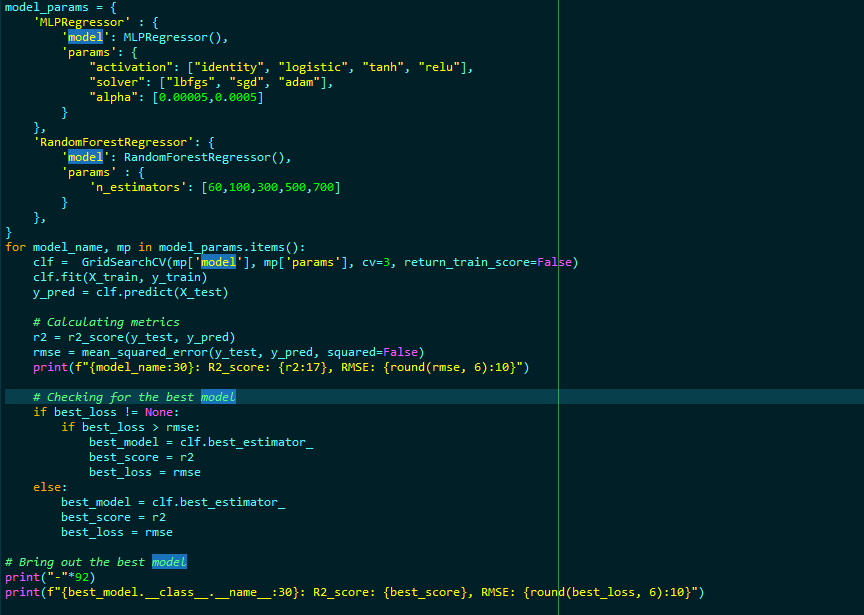
LinearRegression : R2\_score: 0.5424884689862141, RMSE: 51102.483516

-------------------------------------------------- ------------------------------------------

Hồi quy tuyến tính: R2\_score: 0.5424884689862141, RMSE: 51102.483516

Mô hình LinearRegression cho thấy kết quả tốt nhất khi đào tạo các mô hình có sẵn.

Chưa tham số

Tiếp theo ta se lựa chọn từ các mô hình

MLPRegressor : R2\_score: 0.542562862048989, RMSE: 51098.328621

RandomForestRegressor : R2\_score: 0.4603180189503191, RMSE: 55502.167485

--------------------------------------------------------------------------------------------

MLPRegressor : R2\_score: 0.542562862048989, RMSE: 51098.328621

Sau khi khớp các siêu tham số, mô hình MLPRegressor cho thấy độ chính xác tốt nhất trong số tất cả các mô hình.

Đả tham số

1. **Giải Pháp**

Dự án Housing Price Prediction Data có thể được tiếp cận thông qua các bước sau:

1.Khám phá Dữ liệu

Tải và Kiểm tra Dữ liệu,Phân tích mô tả vàBiểu đồ và Tương quan

2. Tiền Xử lý Dữ liệu

Xử lý Dữ liệu Thiếu,Chuẩn hóa Dữ liệu và Mã hóa Categorical Variables

3. Tạo và Đánh giá Mô hình

Chia Dữ liệu ,Huấn luyện Mô hình,Đánh giá Mô hình và Tinh chỉnh Mô hình

4. Dự đoán và Triển khai

Dự đoán Giá nhà và Triển khai và Sử dụng5. Tinh chỉnh và Cải thiện

5 Tinh chỉnh Mô hình và Đánh giá và Cập nhật

Quy trình trên sẽ giúp bạn tiếp cận vấn đề dự đoán giá nhà một cách toàn diện từ việc tiền xử lý dữ liệu đến xây dựng và đánh giá mô hình, và cuối cùng là triển khai mô hình để sử dụng trong thực tế.

**4. Kết Luận**

Dự án Housing Price Prediction Data là một cơ hội để dự đoán giá nhà dựa trên các yếu tố như diện tích, số lượng phòng ngủ, vị trí, và nhiều đặc trưng khác. Quá trình xây dựng mô hình không chỉ giúp chúng ta hiểu rõ hơn về dữ liệu mà còn tạo ra một công cụ hữu ích cho các chủ đầu tư và người mua bất động sản.

Cuối cùng, dự án này cung cấp cơ hội phát triển kỹ năng phân tích dữ liệu và xây dựng mô hình học máy, đồng thời đem lại giá trị thực tiễn trong việc dự đoán giá nhà, hỗ trợ quyết định đầu tư và mua bán bất động sản.

**5 Demo Chương Trình**