TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP THÁI NGUYÊN

**KHOA: ĐIỆN TỬ**

**BỘ MÔN: CÔNG NGHỆ THÔNG TIN**

****

**BÀI TẬP LỚN**

**MÔN HỌC**

**KHOA HỌC DỮ LIỆU**

Sinh viên: Nguyễn Đình Đức

Lớp: K55KMT

Giáo viên hướng dẫn: Nguyễn Văn Huy

**Thái Nguyên 2023**

**TRƯỜNG ĐHKTCN CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**

**KHOA: ĐIỆN TỬ *Độc lập - Tự do - Hạnh phúc***

**BÀI TẬP LỚN**

**MÔN HỌC: KHOA HỌC DỮ LIỆU**

**BỘ MÔN: CÔNG NGHỆ THÔNG TIN**

*Sinh viên: Nguyễn Đình Đức. Lớp: K55KMT.*

*Nghành: Kỹ Thuật Máy Tính.*

*Giáo viên hướng dẫn: Nguyễn Văn Huy.*

*Ngày giao đề:…/…/…… Ngày hoàn thành: …/…/……*

1. Tên đề tài: Dự đoán giá thuê chung cư khu vực Hà Nội.
2. Yêu cầu:

………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

1. Các bản vẽ, chương trình và đồ thị: Kiểm thử chương trình.

**GIÁO VIÊN HƯỚNG DẪN**

*( Ký và ghi rõ họ, tên)*

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

..............................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................

*Thái Nguyên, ngày… tháng … năm 20…*

**GIÁO VIÊN HƯỚNG DẪN**

*( Ký và ghi rõ họ, tên)*

**LỜI NÓI ĐẦU**

Trong bối cảnh tăng trưởng đô thị và phát triển kinh tế, giá nhà trở thành một trong những yếu tố quan trọng ảnh hưởng đến quyết định mua bán và đầu tư trong lĩnh vực bất động sản. Việc dự đoán giá mua và thuê nhà trở thành một thách thức lớn đối với các nhà phát triển, chính phủ, và cả những người đang tìm kiếm mua nhà.

Để tìm hiểu sâu hơn về lĩnh vực giá trên, tôi chọn đề tài “Dự đoán giá thuê chung cư khu vực Hà Nội” để hiểu thêm về lĩnh vực khoa học dữ liệu.

Đề này nhằm mục tiêu sử dụng dữ liệu được lấy trên website để phân tích và xây dựng mô hình dự đoán giá thuê chung cư khu vực Hà Nội. Chúng ta sẽ áp dụng các kỹ thuật và phương pháp trong lĩnh vực khoa học dữ liệu để tìm hiểu các yếu tố ảnh hưởng đến giá thuê chung cư và xây dựng mô hình dự đoán chính xác.

**CHƯƠNG I: GIỚI THIỆU CHUNG**

1. **Đoán giá là gì?**



"Đoán giá" có thể ám chỉ việc ước lượng hoặc dự đoán giá trị của một sản phẩm, tài sản hoặc dịch vụ. Việc đoán giá thường được thực hiện dựa trên các yếu tố như dữ liệu thị trường, thông tin liên quan, xu hướng và mô hình phân tích. Các phương pháp và kỹ thuật trong việc đoán giá có thể được áp dụng từ nhiều lĩnh vực, bao gồm kinh tế, tài chính, bất động sản, hàng hóa và các thị trường tài sản khác.

Tuy nhiên, để có dự đoán giá chính xác, thông tin cụ thể về sản phẩm hoặc dịch vụ cần được xem xét, cũng như các yếu tố kinh tế và thị trường liên quan. Trong một số trường hợp, các mô hình dự đoán hoặc thuật toán máy học có thể được sử dụng để phân tích dữ liệu và đưa ra dự đoán về giá trị trong tương lai.

1. **Ứng dụng**

Có nhiều ứng dụng của việc đoán giá trong các lĩnh vực khác nhau. Dưới đây là một số ví dụ phổ biến về các ứng dụng của đoán giá:

* Tài chính và chứng khoán: Trong lĩnh vực tài chính, việc đoán giá có thể được sử dụng để dự đoán giá cổ phiếu, hàng hóa, ngoại tệ và các tài sản tài chính khác. Các nhà đầu tư và các công ty tài chính thường sử dụng các mô hình và thuật toán dự đoán để đưa ra quyết định giao dịch và đầu tư.
* Bất động sản: Đoán giá cũng được áp dụng trong lĩnh vực bất động sản để ước lượng giá trị các tài sản như căn hộ, nhà đất hoặc tòa nhà. Các nhà phát triển, nhà môi giới và các chuyên gia bất động sản sử dụng các mô hình đoán giá để đưa ra quyết định về mua, bán hoặc đầu tư vào bất động sản.
* Thương mại điện tử: Trong lĩnh vực thương mại điện tử, việc đoán giá có thể được sử dụng để ước lượng giá sản phẩm và dịch vụ trên các nền tảng mua sắm trực tuyến. Các công ty thương mại điện tử thường sử dụng các thuật toán dự đoán giá để đề xuất giá cả cạnh tranh và tùy chỉnh cho khách hàng.
* Ngành y tế: Đoán giá cũng có thể được áp dụng trong ngành y tế để ước lượng giá trị của các dịch vụ y tế và sản phẩm dược phẩm. Các công ty bảo hiểm y tế, bệnh viện và nhà nghiên cứu y tế sử dụng các mô hình đoán giá để đưa ra quyết định về giá cả, chi phí và quản lý tài chính.
* Du lịch và khách sạn: Trong ngành du lịch và khách sạn, việc đoán giá có thể được sử dụng để ước lượng giá phòng khách sạn, vé máy bay, tour du lịch và các dịch vụ du lịch khác. Các công ty du lịch, đại lý du lịch và các trang web đặt phòng sử dụng các mô hình đoán giá để quản lý giá cả, tối ưu hóa doanh thu và cung cấp ưu đãi cho khách hàng.

1. **Công nghệ**

Công nghệ đóng vai trò quan trọng trong việc đoán giá và các ứng dụng liên quan. Dưới đây là một số công nghệ thường được sử dụng trong việc đoán giá:

* Học máy (Machine Learning): Học máy là một phương pháp sử dụng các thuật toán và mô hình để học từ dữ liệu và dự đoán kết quả trong tương lai. Các thuật toán phổ biến trong lĩnh vực đoán giá bao gồm hồi quy tuyến tính, hồi quy logistic, máy vector hỗ trợ (SVM), cây quyết định và các thuật toán học sâu như mạng neural.
* Học sâu (Deep Learning): Học sâu là một lĩnh vực của trí tuệ nhân tạo (AI) tập trung vào việc xây dựng và huấn luyện các mạng neural sâu để hiểu và giải quyết các bài toán phức tạp. Trong lĩnh vực đoán giá, mạng neural sâu, chẳng hạn như mạng neural tích chập (CNN) hoặc mạng neural hồi quy (RNN), có thể được sử dụng để học từ dữ liệu và dự đoán giá trị.
* Kỹ thuật tăng cường (Ensemble Techniques): Kỹ thuật tăng cường là một phương pháp kết hợp nhiều mô hình dự đoán để tạo ra một dự đoán tốt hơn. Ví dụ: Kỹ thuật Bagging (Bootstrap Aggregating) sử dụng nhiều mô hình học máy độc lập và kết hợp kết quả để đưa ra dự đoán cuối cùng.
* Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP): Trong một số trường hợp, đoán giá có thể dựa trên thông tin và mô tả liên quan đến sản phẩm hoặc dịch vụ. Xử lý ngôn ngữ tự nhiên là một lĩnh vực của AI tập trung vào việc xử lý và hiểu ngôn ngữ tự nhiên của con người. Trong đoán giá, NLP có thể được sử dụng để phân tích và rút trích thông tin từ các mô tả sản phẩm hoặc nhận xét của người dùng để đưa ra dự đoán về giá trị.
* Kỹ thuật trích xuất đặc trưng (Feature Extraction): Trích xuất đặc trưng là quá trình chuyển đổi dữ liệu đầu vào thành một tập hợp các đặc trưng có ý nghĩa và dễ dùng cho việc dự đoán. Trong đoán giá, kỹ thuật trích xuất đặc trưng có thể bao gồm việc chuyển đổi dữ liệu định tính thành dữ liệu số, trích xuất thông tin từ hình ảnh hoặc văn bản, và xử lý các biến đầu vào để tạo ra các đặc trưng mới..

Những công nghệ này có thể được kết hợp và tùy chỉnh phù hợp với bài toán đoán giá cụ thể để đạt được kết quả tốt nhất. Việc lựa chọn công nghệ phụ thuộc vào tính chất của dữ liệu, độ phức tạp của bài toán, và tài nguyên có sẵn.

1. **Áp dụng**

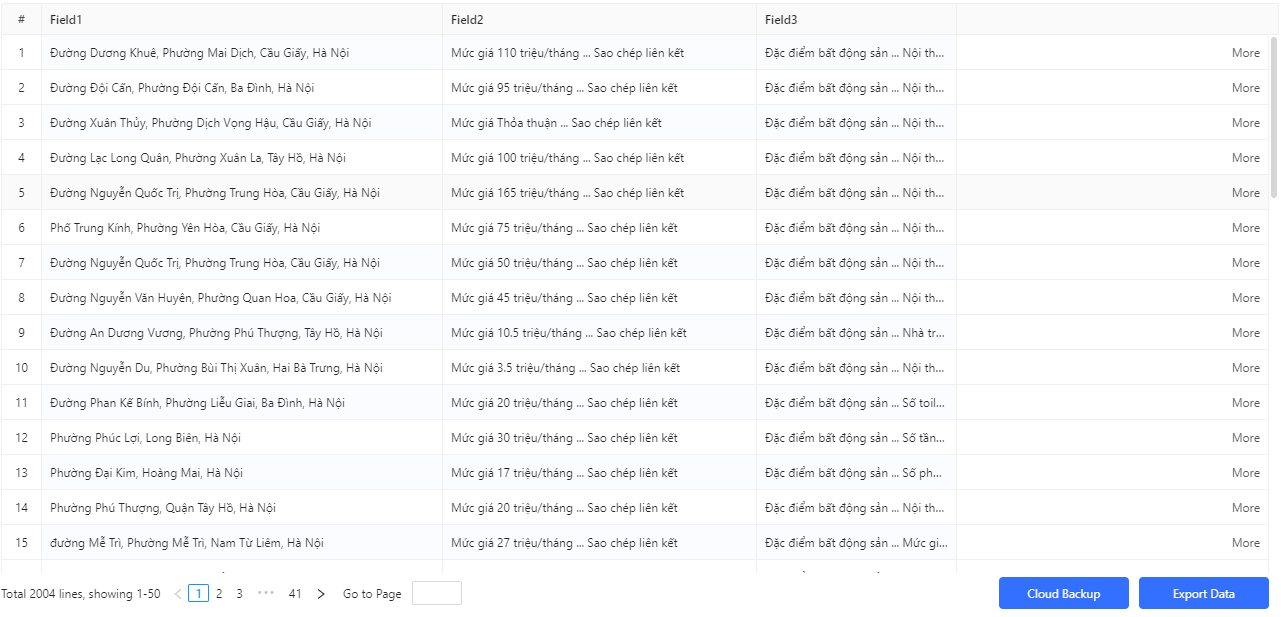
Đề tài “Dự đoán giá thuê chung cư khu vực Hà Nội” là một đề tài trong lĩnh vực bất động sản. Để có thể hoàn thành đề tài này, thì đầu tiên tôi đã lấy dữ liệu về giá thuê chung cư trên trang <https://batdongsan.com.vn/>. Sau khi lấy được dữ liệu thô, thông qua excel để xử lý dữ liệu thô xong rồi đưa dữ liệu đã được xử lý vào bài làm để xây dưng mô hình. Phương pháp được sử dụng để xây dựng mô hình đề sử dụng ở đây là Random Forest. Đây là một phương pháp trong lĩnh vực Học máy và thuộc vào kỹ thuật tăng cường.

**CHƯƠNG II: CƠ SỞ LÝ THUYẾT**

1. **Xử lý dữ liệu.**
2. **Crawl data.**

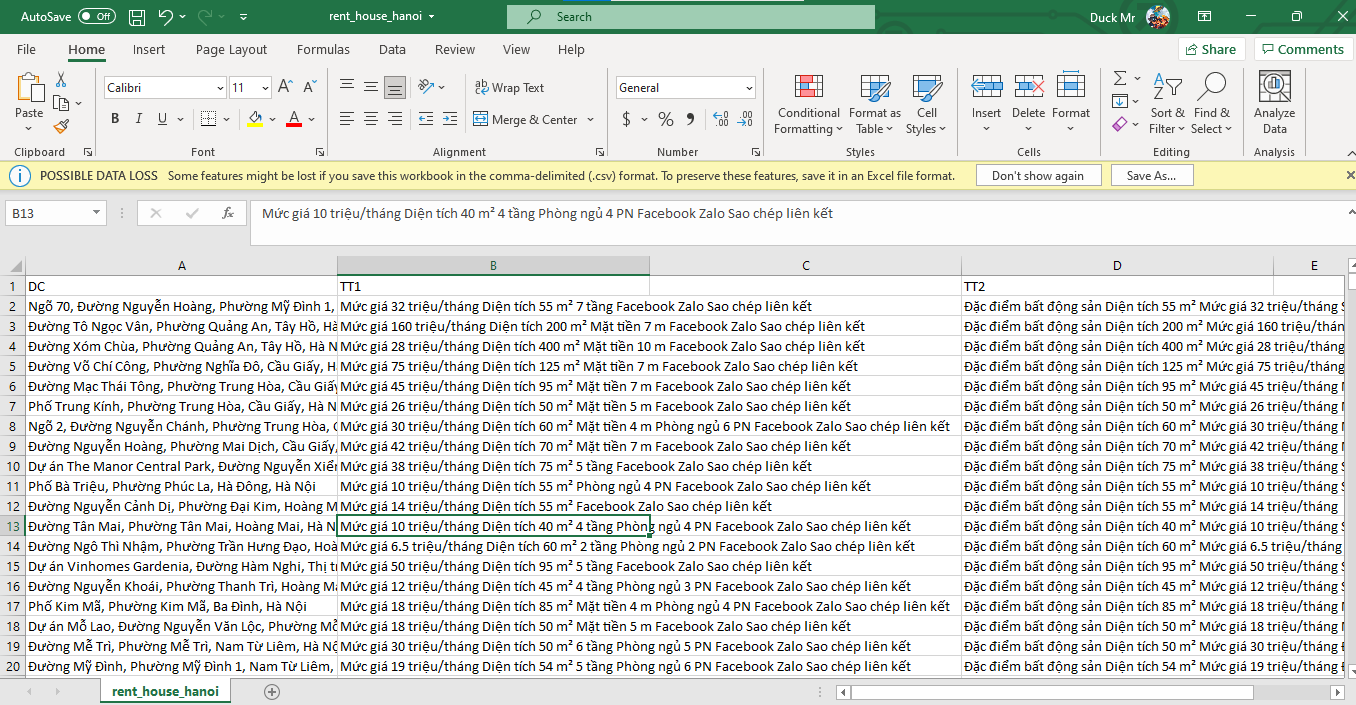
Crawl data là quá trình thu thập dữ liệu và thông tin website nhằm phục vụ nhiều mục tiêu khác nhau. Theo đó, các bot của công cụ tìm kiếm (Search Engine) như Google, Bing,… sẽ lần lượt truy cập vào tất cả trang trên website cũng như liên kết liên quan để thống kê dữ liệu.

Ở đây tôi sử dụng Octoparse 8 để có thể thu thập thông tin từ <https://batdongsan.com.vn/>.

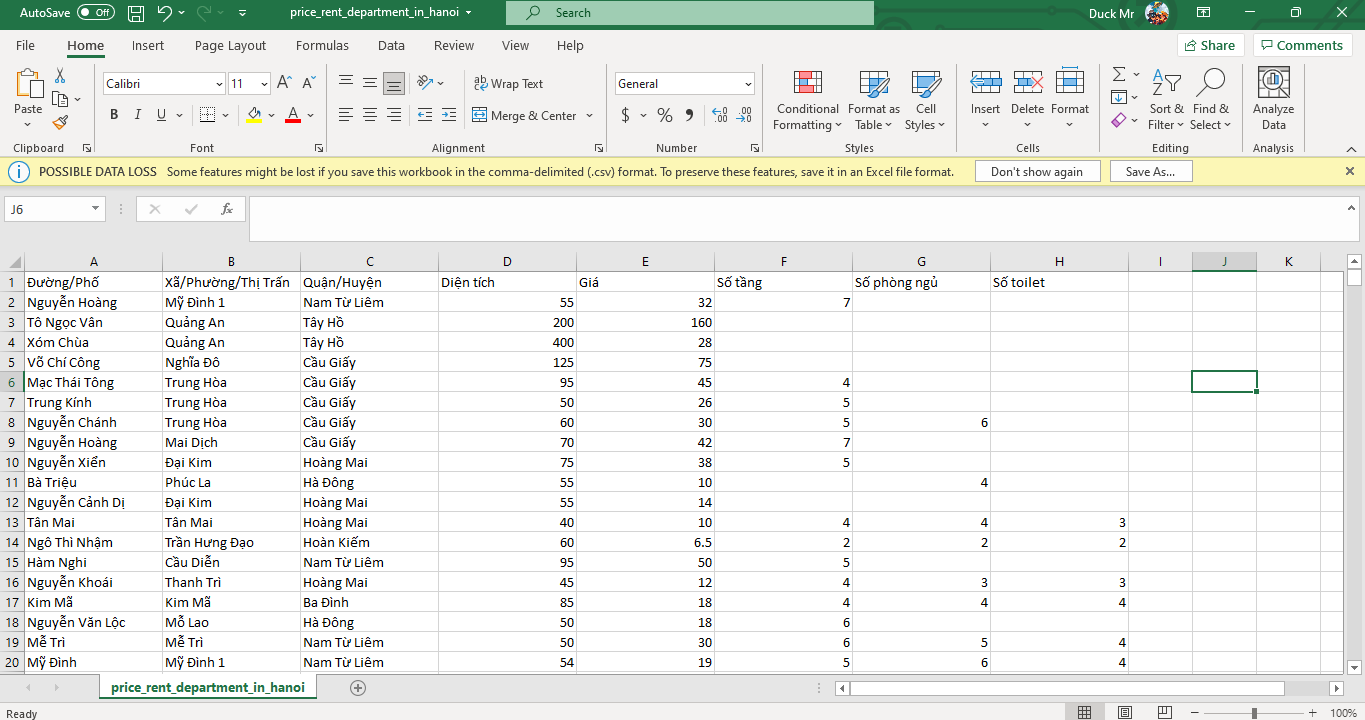


Sau khi Crawl data xong thì sẽ export data về file excel, csv … để bắt đầu quá trình xử lý data thô.

1. **Xử lý data trên Excel.**

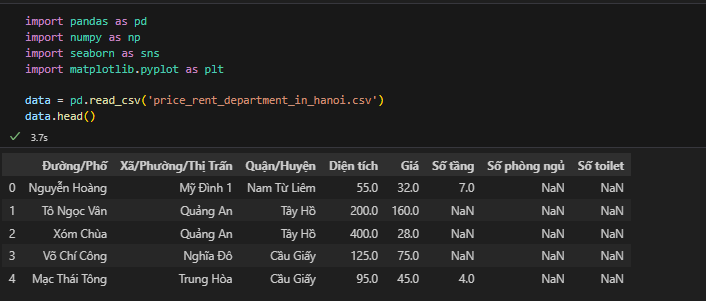


Với dữ liệu thô như hình trên, thông qua quá dùng các hàm có trên Excel như TRIM(), SUBSTITUTE(), LEFT(), RIGHT(), MID(),… thì dữ liệu đã được tách ra làm các trường phù hợp để có thể vào công đoạn xử lý data cho mô hình.

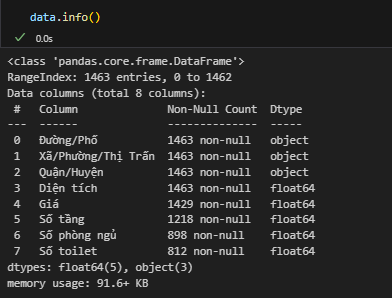


Qua xử lý dữu liệu trên Excel thì dữ liệu còn khoảng hơn 1400 dòng.

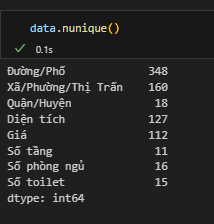
1. **Xử lý data với python.**

****

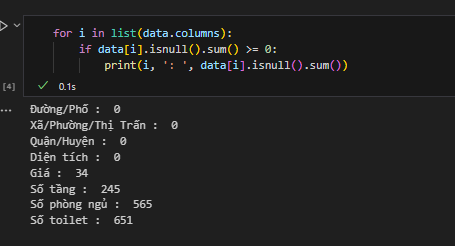
Đầu tiên xử dụng thư viện pandas để đọc file csv. Sau khi xem qua dữ liệu đầu vào, sẽ xem các thông số thuộc tính,… để từ đó đánh giá dữ liệu đầu vào nên làm những gì để phù hợp với mô hình.



Ở đây dữ liệu đầu vào có hai kiểu là object và float64. Float là kiểu dữ liệu số dạng thập phân, còn object là kiểu dữ liệu gốc và là cơ sở cho tất cả các kiểu dữ liệu khác. Trong dữ liệu đầu vào có 3 cột Đường/Phố, Xã/Phường/Thị Trấn, Quận/Huyện là thuộc kiểu object, cụ thể ở đây là kiểu text.

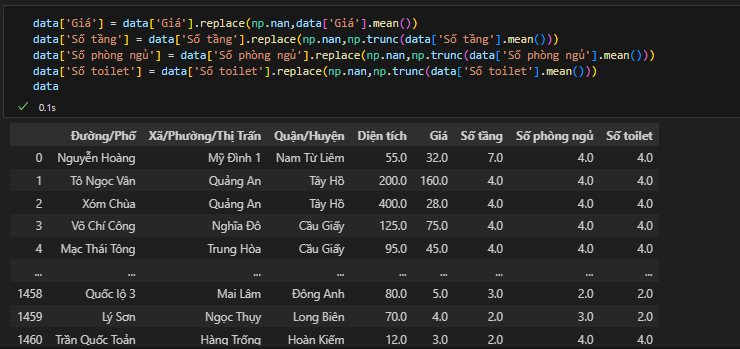


Sau khi xem kiểu dữ liệu của dữ liệu đầu vào, dùng nunique() để xem cáo bao nhiêu dữ liệu duy nhất trong dư liệu đầu vào.



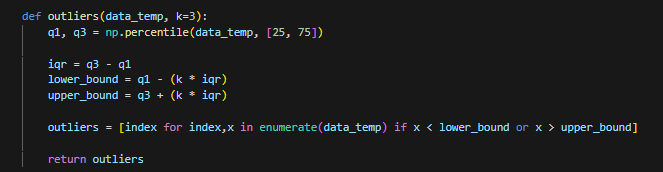
Dùng vòng lặp để xem số các dòng dữ liệu của tùng cột thuộc tính không có giá trị, hay còn là giá trị Nan (Các dữ liệu trống đó sẽ được gọi là bị mising value). Do số lượng dữ liệu trống khá lớn nên sẽ phải xử lý các dữ liệu đó.

Đối với trường hợp mising value ít so với tập dữ liệu thì các dfng đó sẽ bị xoá bằng hàm dropna(). Nhưng ở đây dữ liệu trống so với tập dữ liệu khá lớn nên sẽ phải dùng cách khác, đó là dùng giá trị trung bình của cột để điền vào các ô bị thiếu dữ liệu.

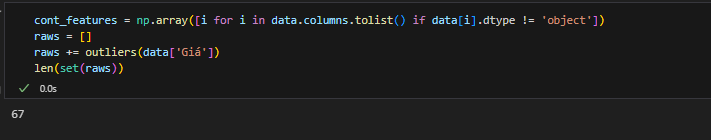


Sau khi xử lý các dữ liệu bị thiếu xong thì chuyển qua xử lý dữ liệu bị outliers, tức là các dữ liệu khác thường trong dữ liệu đầu vào.

Để tìm các dòng dữ liệu bị outliers, tôi sử dụng hàm outliers dựa trên phương pháp IQR để xác định ngưỡng cho outliers.

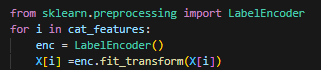


Hàm nãy sẽ tính các phân vị q1, q3 sau đó sẽ tìm giới hạn trên và dưới từ các phân vị và tham số k. Cuối cùng đưa ra mảng các dữ liệu nhỏ hơn dữu liệu dưới và lớn hơn dữ liệu trên. Sau khi lấy được mảng sẽ loại bỏ các dữ liệu đó.

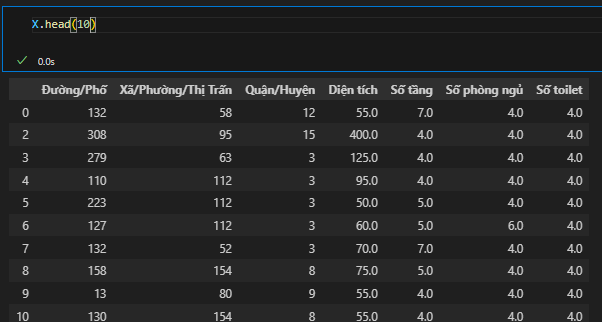


Trong bài thì tôi dùng tìm các outlier cho cột giá và tìm đc 67 dữ liệu outlier.

Sau khi xử lý xong các dữ liệu kiểu float xong thì tôi đi vào xử lý dữ liệu kiểu object. Đối với các dữ liệu này tôi dùng hàm LabelEncoder() để mã hoá các biến thành các số nguyên vì mô hình cần dữ liệu là số để xây dựng.



Dưới đấy là dữ liệu đã được xử lý qua tất các công đoan ở trên:

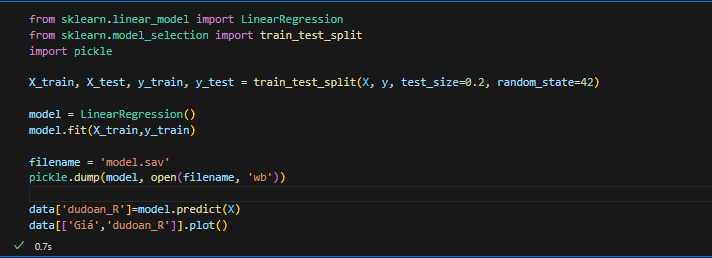


1. **Model**

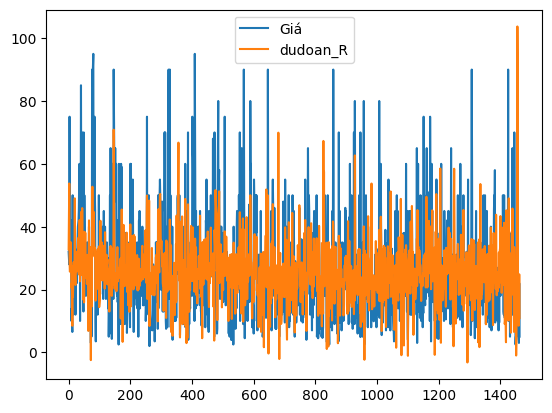
Trong bài thì tôi có thử xử dụng LinearRegression và RandomForestRegressor. Tuy LinearRegression kết quả không tốt lắm nhưngtooi sẽ vẫn giới thiệu nó trong phần này.

1. **LinearRegression (hồi quy tuyến tính).**

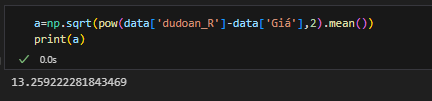
Linear regression (Hồi quy tuyến tính) là một thuật toán trong lĩnh vực học máy được sử dụng để xây dựng mô hình dự đoán dựa trên mối quan hệ tuyến tính giữa biến đầu vào (biến độc lập) và biến đầu ra (biến phụ thuộc). Mô hình hồi quy tuyến tính giả định rằng có một mối quan hệ tuyến tính giữa các biến đầu vào và đầu ra.



Sau khi xây dựng ô hình xong thì tôi vẽ biểu đồ dự đoán theo các thuộc tính ở dữ liệu train X và so nó với giá ban đầu trong tệp data.



Nhìn qua biểu đồ có thể thấy được dự đoán bằng có kết quá không được khả quan lắm. Dưới đây là là kết quả đánh giá độ sai lệch sử dụng MSE:

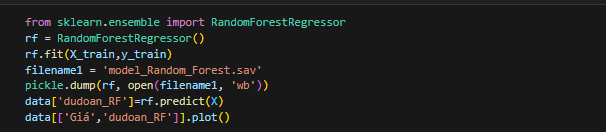


Sai số lên tới tận 13.25, tức là độ sai lệch giá thuê khoảng trên 13,25 triệu.

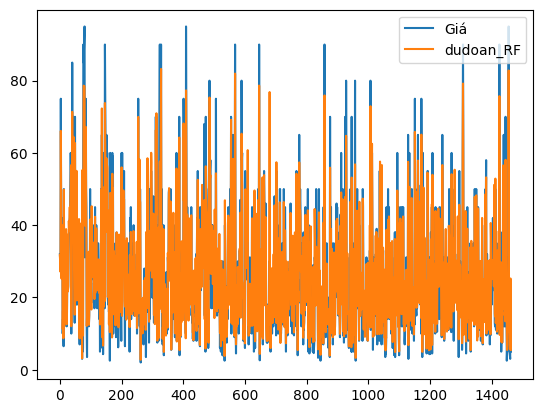
1. **RandomForestRegressor**

RandomForestRegressor là một mô hình học máy trong scikit-learn được dùng để huấn luyện và dự đoán trên dữ liệu dạng hồi quy (regression). Nó sử dụng phương pháp RandomForest (rừng ngẫu nhiên) để xây dựng nhiều cây quyết định (decision trees) và kết hợp kết quả của chúng để đưa ra dự đoán cuối cùng.

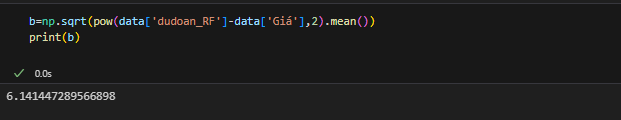
Mô hình RandomForestRegressor là một tập hợp các cây quyết định độc lập với nhau. Mỗi cây quyết định được huấn luyện trên một tập con của dữ liệu và sử dụng một số thuộc tính ngẫu nhiên để tạo ra các quyết định. Kết quả cuối cùng của RandomForestRegressor là sự kết hợp của dự đoán từ tất cả các cây quyết định.



Cũng giống như hồi quy tuyến tính, tooi sẽ so sánh kết quả dự đoán so với dữ liệu gốc.



Các dường dự đoán đã khả quan hơn so với hồi quy tuyến tính.



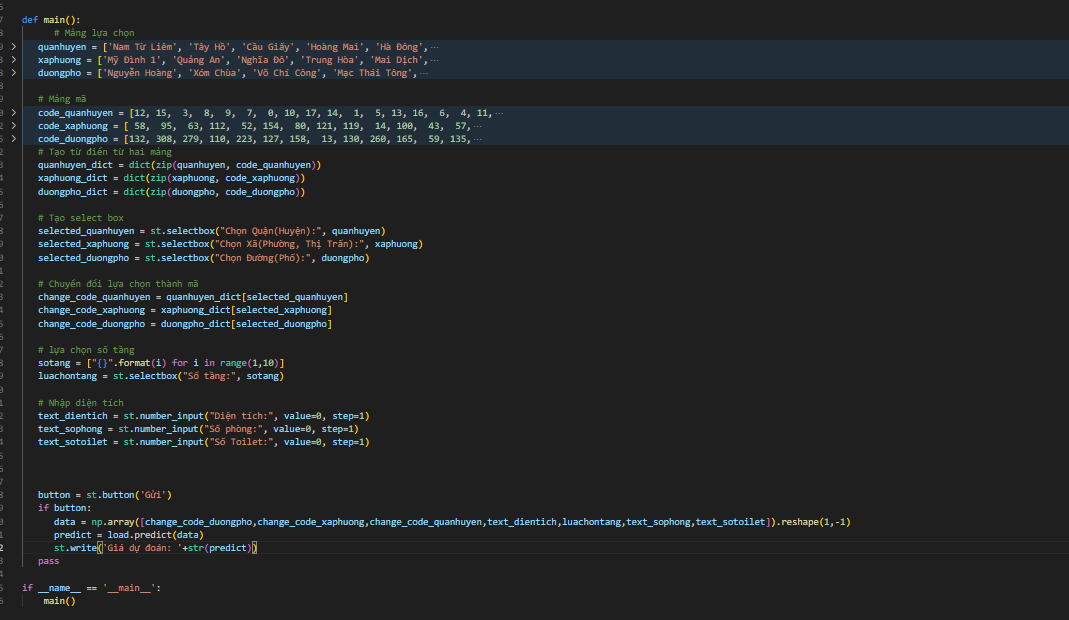
Tuy kết quả đánh giá độ sai lệch là 6.14 không quá thấp nhưng kết quả đó cũng không hẳn là quá cao.

**CHƯƠNG III: XÂY DỰNG CHƯƠNG TRÌNH**

1. **Streamlit**

Streamlit là một framework Python mã nguồn mở được sử dụng để xây dựng ứng dụng web tương tác dễ dàng và nhanh chóng. Với Streamlit, bạn có thể biến mã Python thành các ứng dụng web thân thiện với người dùng một cách nhanh chóng và dễ dàng.

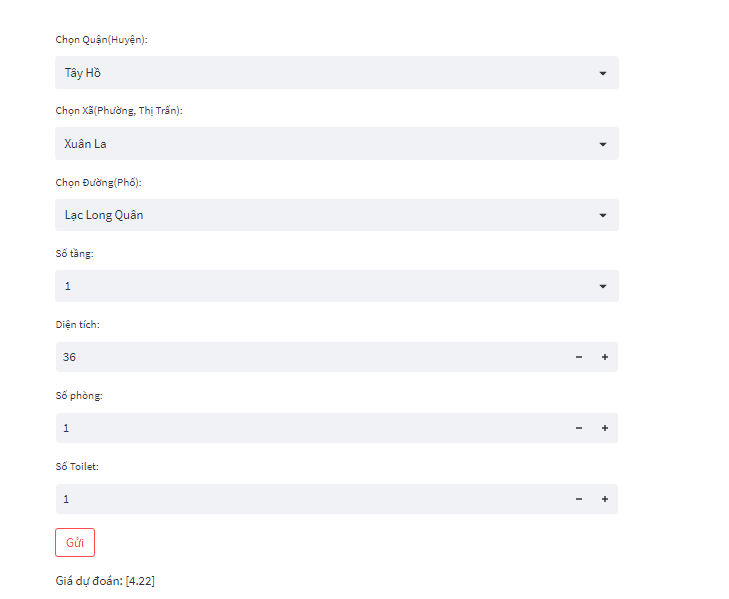
1. **Xây dựng chương trình**

****

Do code còn hơi dài nên đã thu gọn một số chỗ đễ có thể chụp toàn bộ code để mang vào bài.

Code gồm các mảng mã và mảng lựa chọn. mảng mã chứa các mã đã được mã hoá khi dùng LabelEncode. Sau khi chọn các địa điểm trong phần lựa chọn thì sẽ chuyển đổi thành mã để đưa vào model dữ đoán.

Sau khi điền các thông tin cần thì bấm gửi sẽ đưa các thông số đó vào model để dự đoán.



Sau khi dự đoán kết quả sẽ được trả về in ra mà hình.

**CHƯƠNG IV: KẾT LUẬN**

1. **Tổng quan**

Chương trình đã có thể thu thập các thông tin được nhập qua giao diện web của streamlit rồi đưa vào model đã được xây dựng để dự đoán giá thuê chung cư trong khu vực Hà Nội.

1. **Ưu, nhược điểm**
2. **Ưu điểm**

* Giao diện thân thiện dễ sử dụng.
* Tốc độ dự đoán nhanh chóng, không bị gián đoạn.

1. **Nhược điểm.**

* Độ sai lệch vẫn còn hơi lớn.
* Dữ liệu còn ít.
* Dữ liệu còn hơi ít thuộc tính.

1. **Hướng phát triển.**

Chương trình sẽ thu thập nhiều dữ liệu, bổ sung thêm các thuộc tính có tính tương quan với giá nhà, giảm độ lệch sai số thấp hơn so với kết quả hiện tại.

**TÀI LIỆU THAM KHẢO**

**[1]** [**https://www.kaggle.com/**](https://www.kaggle.com/)

**[2]** [**https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html**](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)

**[3]** [**https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html**](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html)

**[4]** [**https://machinelearningcoban.com/2016/12/28/linearregression/**](https://machinelearningcoban.com/2016/12/28/linearregression/)

**[5]** [**https://solieu.vip/mse-va-rmse-la-gi-va-cach-tinh-tren-stata/**](https://solieu.vip/mse-va-rmse-la-gi-va-cach-tinh-tren-stata/)