

**TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP THÁI NGUYÊN**

**KHOA: ĐIỆN TỬ**

**BỘ MÔN: CÔNG NGHỆ THÔNG TIN**



## **ĐỒ ÁN MÔN HỌC**

**MÔN HỌC**

# **THỰC TẬP CHUYÊN NGÀNH TIN HỌC CÔNG NGHIỆP**

Sinh viên: Nguyễn Ngọc Hiếu

Lớp: K55KMT

Giáo viên hướng dẫn: TS. Nghiêm Văn Tính

**Thái Nguyên 2023**

## ĐỒ ÁN MÔN HỌC

**MÔN HỌC: THỰC TẬP CHUYÊN NGÀNH**  
**TIN HỌC CÔNG NGHIỆP**  
**BỘ MÔN: CÔNG NGHỆ THÔNG TIN**

*Sinh viên: Nguyễn Ngọc Hiếu.*

*Lớp: K55KMT.*

*Nghành: Kỹ Thuật Máy Tính.*

*Giáo viên hướng dẫn: TS. Nghiêm Văn Tính.*

*Ngày giao đề: 09/05/2023*

*Ngày hoàn thành: 16/06/2023*

1. Tên đề tài: Dự đoán giá nhà, chung cư khu vực Hà Nội.

2. Yêu cầu:

.....

.....

.....

.....

.....

.....

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

**TRƯỞNG BỘ MÔN**  
( Ký và ghi rõ họ, tên)

**GIÁO VIÊN HƯỚNG DẪN**  
( Ký và ghi rõ họ, tên)

## NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

*Thái Nguyên, ngày... tháng ... năm 2023*

**GIÁO VIÊN HƯỚNG DẪN**

*( Ký và ghi rõ họ, tên)*

## NHẬN XÉT CỦA GIÁO VIÊN CHĂM

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

*Thái Nguyên, ngày... tháng ... năm 2023*

**GIÁO VIÊN CHĂM**

*( Ký và ghi rõ họ, tên)*

## Mục Lục

LỜI NÓI ĐẦU .....	6
CHƯƠNG I: GIỚI THIỆU CHUNG.....	7
1.    Đoán giá là gì? .....	7
2.    Ứng dụng.....	7
3.    Công nghệ .....	8
4.    Áp dụng.....	9
CHƯƠNG II: CƠ SỞ LÝ THUYẾT .....	10
1.    Phân tích bài toán .....	10
2.    Phân tích yêu cầu .....	10
2.1.    Xử lý dữ liệu. ....	10
2.2.    Các phương pháp dự báo .....	17
2.3.    Các bước dự đoán .....	23
2.4.    Giao diện .....	23
3.    Phân tích hướng đối tượng.....	24
3.1.    Biểu đồ use case .....	24
3.2.    Biểu đồ tuần tự .....	25
CHƯƠNG III: XÂY DỰNG CHƯƠNG TRÌNH .....	26
1.    Streamlit.....	26
2.    Xây dựng chương trình.....	26
CHƯƠNG IV: KẾT LUẬN.....	29
1.    Tổng quan .....	29
2.    Ưu, nhược điểm .....	29
3.    Hướng phát triển.....	29
TÀI LIỆU THAM KHẢO .....	30

## **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 nhà, 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 nhà, 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.

## 2. 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.

### 3. 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.

#### **4. Áp dụng**

Đề tài “Dự đoán giá nhà, 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 dựng 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. Phân tích bài toán

Bài toán "Dự đoán giá nhà, chung cư khu vực Hà Nội" đòi hỏi sự ứng dụng của các phương pháp dự đoán và phân tích dữ liệu để xác định giá trị thực của nhà, chung cư. Điều này mang lại lợi ích cho người mua, người bán và các chuyên gia trong lĩnh vực này.

Để giải quyết bài toán này, quá trình thu thập dữ liệu về các yếu tố quan trọng như đường phố, xã (phường), quận( huyện), số phòng, diện tích, số tầng và số toilet. Dữ liệu này được sử dụng để xây dựng các mô hình dự đoán và áp dụng các phương pháp thống kê để đánh giá giá trị ngôi nhà, chung cư. Các yếu tố thị trường khác như cung và cầu, xu hướng mua bán cũng được xem xét để có cái nhìn tổng quan về thị trường.

Việc dự đoán giá nhà, chung cư mang lại nhiều lợi ích cho người mua. Bằng cách có được thông tin về giá trị thực của ngôi nhà, chung cư người mua có thể đánh giá xem giá đề xuất có phản ánh đúng giá trị hay không. Điều này giúp tránh trường hợp trả giá quá cao hoặc không đúng giá trị thực. Ngoài ra, dự đoán giá cũng cung cấp cho người mua sự tự tin trong quyết định mua và giúp tiết kiệm tài chính.

Đối với người bán, dự đoán giá nhà, chung cư giúp họ định giá xe một cách công bằng và hợp lý. Thay vì dựa vào cảm tính, người bán có thể sử dụng thông tin dự đoán để đưa ra giá bán hợp lý, thu hút khách hàng và nhanh chóng bán nhà, chung cư với giá trị thích hợp.

### 2. Phân tích yêu cầu

Đầu vào của hệ thống này bao gồm các thông tin về đường phố, xã (phường), quận( huyện), số phòng, diện tích, số tầng và số toilet. Để đưa thông tin vào hệ thống, người dùng phải nhập thông tin vào các ô tương ứng trên trang web.

Đầu ra của hệ thống là giá trị dự đoán về giá bán của ngôi nhà, chung cư. Kết quả được hiển thị trên trang web và người dùng có thể xem kết quả bằng cách đọc giá trị hiển thị trên trang web.

#### 2.1. Xử lý dữ liệu.

##### a. 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/>.

#	Field1	Field2	Field3	
1	Đường Dương Khuê, Phường Mai Dịch, Cầu Giấy, Hà Nội	Mức giá 110 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
2	Đường Đội Cấn, Phường Đội Cấn, Ba Đình, Hà Nội	Mức giá 95 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
3	Đường Xuân Thủy, Phường Dịch Vọng Hậu, Cầu Giấy, Hà Nội	Mức giá Thỏa thuận ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
4	Đường Lạc Long Quân, Phường Xuân La, Tây Hồ, Hà Nội	Mức giá 100 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
5	Đường Nguyễn Quốc Trị, Phường Trung Hòa, Cầu Giấy, Hà Nội	Mức giá 165 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
6	Phố Trung Kính, Phường Yên Hòa, Cầu Giấy, Hà Nội	Mức giá 75 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
7	Đường Nguyễn Quốc Trị, Phường Trung Hòa, Cầu Giấy, Hà Nội	Mức giá 50 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
8	Đường Nguyễn Văn Huyền, Phường Quan Hoa, Cầu Giấy, Hà Nội	Mức giá 45 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
9	Đường An Dương Vương, Phường Phú Thượng, Tây Hồ, Hà Nội	Mức giá 10.5 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Nhà tr...	More
10	Đường Nguyễn Du, Phường Bưởi Thị Xuân, Hai Bà Trưng, Hà Nội	Mức giá 3.5 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
11	Đường Phan Kế Bính, Phường Liễu Giai, Ba Đình, Hà Nội	Mức giá 20 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Sở toil...	More
12	Phường Phúc Lợi, Long Biên, Hà Nội	Mức giá 30 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Sở tăn...	More
13	Phường Đại Kim, Hoàng Mai, Hà Nội	Mức giá 17 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Sở ph...	More
14	Phường Phú Thượng, Quận Tây Hồ, Hà Nội	Mức giá 20 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
15	đường Mễ Trì, Phường Mễ Trì, Nam Từ Liêm, Hà Nội	Mức giá 27 triệu/tháng ... Sao chép liên kết	Đặc điểm bất động sản ... Mức gi...	More

Total 2004 lines, showing 1-50 < 1 2 3 ... 41 > Go to Page

Cloud Backup Export Data

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

## b. Xử lý data trên Excel.

	A	B	C	D	E
1	DC	TT1		TT2	
2	Ngõ 70, Đường Nguyễn Hoàng, Phường Mỹ Đình 1, Múc giá 32 triệu/tháng Diện tích 55 m² 7 tầng Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 55 m² Múc giá 32 triệu/tháng	
3	Đường Tô Ngọc Vân, Phường Quảng An, Tây Hồ, Hà N Múc giá 160 triệu/tháng Diện tích 200 m² Mặt tiền 7 m Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 200 m² Múc giá 160 triệu/tháng	
4	Đường Xóm Chùa, Phường Quảng An, Tây Hồ, Hà N Múc giá 28 triệu/tháng Diện tích 400 m² Mặt tiền 10 m Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 400 m² Múc giá 28 triệu/tháng	
5	Đường Võ Chí Công, Phường Nghĩa Đô, Cầu Giấy, H Múc giá 75 triệu/tháng Diện tích 125 m² Mặt tiền 7 m Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 125 m² Múc giá 75 triệu/tháng	
6	Đường Mạc Thái Tông, Phường Trung Hòa, Cầu Giấy, Hà N Múc giá 45 triệu/tháng Diện tích 95 m² Mặt tiền 7 m Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 95 m² Múc giá 45 triệu/tháng	
7	Phố Trung Kính, Phường Trung Hòa, Cầu Giấy, Hà N Múc giá 26 triệu/tháng Diện tích 50 m² Mặt tiền 5 m Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 50 m² Múc giá 26 triệu/tháng	
8	Ngõ 2, Đường Nguyễn Chánh, Phường Trung Hòa, Cầu Giấy, Hà N Múc giá 30 triệu/tháng Diện tích 60 m² Mặt tiền 4 m Phòng ngủ 6 PN Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 60 m² Múc giá 30 triệu/tháng	
9	Đường Nguyễn Hoàng, Phường Mai Dịch, Cầu Giấy, Hà N Múc giá 42 triệu/tháng Diện tích 70 m² Mặt tiền 7 m Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 70 m² Múc giá 42 triệu/tháng	
10	Dự án The Manor Central Park, Đường Nguyễn Xiển Múc giá 38 triệu/tháng Diện tích 75 m² 5 tầng Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 75 m² Múc giá 38 triệu/tháng	
11	Phố Bà Triệu, Phường Phúc La, Hà Đông, Hà Nội Múc giá 10 triệu/tháng Diện tích 55 m² Phòng ngủ 4 PN Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 55 m² Múc giá 10 triệu/tháng	
12	Đường Nguyễn Cảnh Di, Phường Đại Kim, Hoàng M Múc giá 14 triệu/tháng Diện tích 55 m² Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 55 m² Múc giá 14 triệu/tháng	
13	Đường Tân Mai, Phường Tân Mai, Hoàng Mai, Hà N Múc giá 10 triệu/tháng Diện tích 40 m² 4 tầng Phòng ngủ 4 PN Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 40 m² Múc giá 10 triệu/tháng	
14	Đường Ngô Thì Nhậm, Phường Trần Hưng Đạo, Hoàn M Múc giá 6.5 triệu/tháng Diện tích 60 m² 2 tầng Phòng ngủ 2 PN Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 60 m² Múc giá 6.5 triệu/tháng	
15	Dự án Vinhomes Gardenia, Đường Hàm Nghi, Thủ T Múc giá 50 triệu/tháng Diện tích 95 m² 5 tầng Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 95 m² Múc giá 50 triệu/tháng	
16	Đường Nguyễn Khoái, Phường Thanh Trì, Hoàng M Múc giá 12 triệu/tháng Diện tích 45 m² 4 tầng Phòng ngủ 3 PN Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 45 m² Múc giá 12 triệu/tháng	
17	Phố Kim Mã, Phường Kim Mã, Ba Đình, Hà Nội Múc giá 18 triệu/tháng Diện tích 85 m² Mặt tiền 4 m Phòng ngủ 4 PN Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 85 m² Múc giá 18 triệu/tháng	
18	Dự án Mỗ Lao, Đường Nguyễn Văn Lộc, Phường M Múc giá 18 triệu/tháng Diện tích 50 m² Mặt tiền 5 m Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 50 m² Múc giá 18 triệu/tháng	
19	Đường Mễ Trì, Phường Mễ Trì, Nam Từ Li M Múc giá 30 triệu/tháng Diện tích 50 m² 6 tầng Phòng ngủ 5 PN Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 50 m² Múc giá 30 triệu/tháng	
20	Đường Mỹ Đình, Phường Mỹ Đình 1, Nam Từ Li M Múc giá 19 triệu/tháng Diện tích 54 m² 5 tầng Phòng ngủ 6 PN Facebook Zalo Sao chép liên kết			Đặc điểm bất động sản Diện tích 54 m² Múc giá 19 triệu/tháng	

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.

	A	B	C	D	E	F	G	H	I	J
1	Đường/Ph	Xã/Phườn	Quận/Huy	Diện tích	Giá	Số tầng	Số phòng r	Số toilet		
2	Nguyễn Hc	Mỹ Đình 1	Nam Từ Li	55	32	7				
3	Tô Ngọc V	Quảng An	Tây Hồ	200	160					
4	Xóm Chùa	Quảng An	Tây Hồ	400	28					
5	Võ Chí Cô	Nghĩa Đô	Cầu Giấy	125	75					
6	Mạc Thái T	Trung Hòa	Cầu Giấy	95	45	4				
7	Trung Kính	Trung Hòa	Cầu Giấy	50	26	5				
8	Nguyễn Ch	Trung Hòa	Cầu Giấy	60	30	5	6			
9	Nguyễn Hc	Mai Dịch	Cầu Giấy	70	42	7				
10	Nguyễn Xi	Đại Kim	Hoàng Ma	75	38	5				
11	Bà Triệu	Phúc La	Hà Đông	55	10		4			
12	Nguyễn C	Đại Kim	Hoàng Ma	55	14					
13	Tân Mai	Tân Mai	Hoàng Ma	40	10	4	4	3		
14	Ngô Thì N	Trần Hưng	Hoàn Kiếm	60	6.5	2	2	2		
15	Hàm Nghi	Cầu Diễn	Nam Từ Li	95	50	5				
16	Nguyễn Kh	Thanh Trì	Hoàng Ma	45	12	4	3	3		
17	Kim Mã	Kim Mã	Ba Đình	85	18	4	4	4		
18	Nguyễn V	Mỗ Lao	Hà Đông	50	18	6				
19	Mễ Trì	Mễ Trì	Nam Từ Li	50	30	6	5	4		
20	Mỹ Đình	Mỹ Đình 1	Nam Từ Li	54	19	5	6	4		
21	Trịnh Văn	Xuân Phư	Nam Từ Li	100	17	4	6	4		
22	Trịnh Văn	Xuân Phư	Nam Từ Li	150	30	7	6	6		
23	Nguyễn Xi	Đại Kim	Hoàng Ma	80	33	4				
24	Đức Giang	Đức Giang	Long Biên	110	12	3	4	3		
25	Nguyễn Th	Trung Hòa	Cầu Giấy	45	16	5	7	4		
26	Đội Cấn	Đội Cấn	Ba Đình	90	25	4	4	4		
27	Huỳnh Th	Thành Côn	Ba Đình	70	25	5	5	5		
28	Hoàng Ng	Trung Hòa	Cầu Giấy	55	22	5	6	4		
29	Nguyễn Th	Trung Hòa	Cầu Giấy	50	23	5				
30	Xã Đan	Trung Phư	Đống Đa	80	32	6				
31	Vạn Bảo	Liễu Giai	Ba Đình	130	36					
32	Đội Cấn	Vĩnh Phúc	Ba Đình	85	40	4				
33	Trần Thái	Dịch Vọng	Cầu Giấy	160	110	7				
34	Trần Bình	Mai Dịch	Cầu Giấy	80	42	7				

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

### c. Xử lý data với python.

```

1 import pandas as pd
2 import numpy as np
3 import seaborn as sns
4 import matplotlib.pyplot as plt
5
6 data = pd.read_csv('price_rent_department_in_hanoi.csv')
7 data.head()

```

Executed at 2023.05.30 20:24:03 in 172ms

	Đường/Phố	Xã/Phường/Thị Trấn	Quận/Huyện	Diện tích	Giá	Số tầng	Số phòng ngủ	Số toilet
0	Nguyễn Hoàng	Mỹ Đình 1	Nam Từ Liêm	55.0	32.0	7.0	NaN	NaN
1	Tô Ngọc Vân	Quảng An	Tây Hồ	200.0	160.0	NaN	NaN	NaN
2	Xóm Chùa	Quảng An	Tây Hồ	400.0	28.0	NaN	NaN	NaN
3	Võ Chí Công	Nghĩa Đô	Cầu Giấy	125.0	75.0	NaN	NaN	NaN
4	Mạc Thái Tông	Trung Hòa	Cầu Giấy	95.0	45.0	4.0	NaN	NaN

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.

```

1 data.info()

```

Executed at 2023.05.30 20:24:03 in 122ms

		non-null count	dtype
0	Đường/Phố	1463 non-null	object
1	Xã/Phường/Thị Trấn	1463 non-null	object
2	Quận/Huyện	1463 non-null	object
3	Diện tích	1463 non-null	float64
4	Giá	1429 non-null	float64
5	Số tầng	1218 non-null	float64
6	Số phòng ngủ	898 non-null	float64
7	Số toilet	812 non-null	float64

dtypes: float64(5), object(3)  
memory usage: 91.6+ KB

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.

```
In 15 1 data.nunique()
      Executed at 2023.05.30 20:24:03 in 127ms
```

Out 15

	<unnamed>
Đường/Phố	348
Xã/Phường/Thị trấn	160
Quận/Huyện	18
Diện tích	127
Giá	112
Số tầng	11
Số phòng ngủ	16
Số toilet	15

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

```
In 16 1 for i in list(data.columns):
      2     if data[i].isnull().sum() >= 0:
      3         print(i, ': ', data[i].isnull().sum())
      Executed at 2023.05.30 20:24:03 in 238ms
```

Đường/Phố : 0  
Xã/Phường/Thị trấn : 0  
Quận/Huyện : 0  
Diện tích : 0  
Giá : 34  
Số tầng : 245  
Số phòng ngủ : 565  
Số toilet : 651

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

Đối với trường hợp missing 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.

```
In 19 1 data['Giá'] = data['Giá'].replace(np.nan, data['Giá'].mean())
2 data['Số tầng'] = data['Số tầng'].replace(np.nan, np.trunc(data['Số tầng'].mean()))
3 data['Số phòng ngủ'] = data['Số phòng ngủ'].replace(np.nan, np.trunc(data['Số phòng ngủ'].mean()))
4 data['Số toilet'] = data['Số toilet'].replace(np.nan, np.trunc(data['Số toilet'].mean()))
5 data
```

Executed at 2023.05.30 20:24:04 in 2ms

Out 19 1463 rows x 8 columns pd.DataFrame

	Đường/Phố	Xã/Phường/Thị Trấn	Quận/Huyện	Diện tích	Giá	Số tầng	Số phòng ngủ	Số toilet
0	Nguyễn Hoàng	Mỹ Đình 1	Nam Từ Liêm	55.0	32.0	7.0	4.0	4.0
2	Xóm Chùa	Quảng An	Tây Hồ	400.0	28.0	4.0	4.0	4.0
3	Võ Chí Công	Nghĩa Đô	Cầu Giấy	125.0	75.0	4.0	4.0	4.0
4	Mạc Thái Tông	Trung Hòa	Cầu Giấy	95.0	45.0	4.0	4.0	4.0
5	Trung Kính	Trung Hòa	Cầu Giấy	50.0	26.0	5.0	4.0	4.0
6	Nguyễn Chánh	Trung Hòa	Cầu Giấy	60.0	30.0	5.0	6.0	4.0
7	Nguyễn Hoàng	Mai Dịch	Cầu Giấy	70.0	42.0	7.0	4.0	4.0
8	Nguyễn Xiển	Đại Kim	Hoàng Mai	75.0	38.0	5.0	4.0	4.0

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.

```
In 21 1 def outliers(data_temp, k=3):
2     q1, q3 = np.percentile(data_temp, [25, 75])
3
4     iqr = q3 - q1
5     lower_bound = q1 - (k * iqr)
6     upper_bound = q3 + (k * iqr)
7
8     outliers = [index for index, x in enumerate(data_temp) if x < lower_bound or x > upper_bound]
9
10    return outliers
```

Executed at 2023.05.30 20:25:43 in 0.001ms

Hàm này 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 dữ liệu dưới và lớn hơn dữ liệu trên. Sau khi lấy được mảng sẽ loại bỏ các dữ liệu đó.



```

22 1 cont_features = np.array([i for i in data.columns.tolist() if data[i].dtype != 'object'])
2   rows = []
3   # for i in cont_features:
4   #     rows += outliers(data[i])
5   rows += outliers(data['Giá'])
6   len(set(rows))
    Executed at 2023.05.30 20:24:18 in 296ms

```

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.

```

from sklearn.preprocessing import LabelEncoder
for i in cat_features:
    enc = LabelEncoder()
    X[i] = enc.fit_transform(X[i])

```

## 2.2. Các phương pháp dự báo

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

### a) 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.

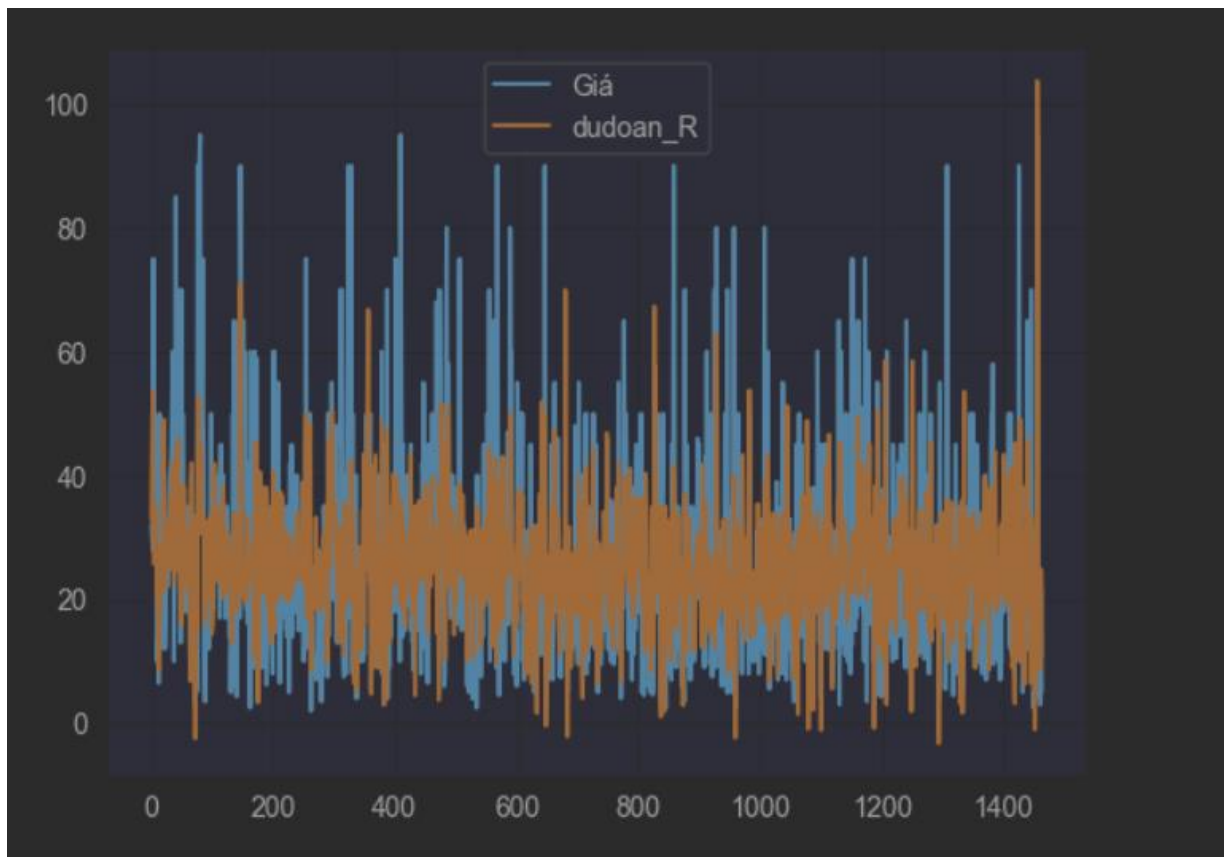
```

n 31 1 from sklearn.linear_model import LinearRegression
      2 from sklearn.model_selection import train_test_split
      3 import pickle
      4
      5 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
      6
      7 model = LinearRegression()
      8 model.fit(X_train, y_train)
      9
     10 filename = 'model.sav'
     11 pickle.dump(model, open(filename, 'wb'))
     12
     13 data['dudoan_R'] = model.predict(X)
     14 data[['Giá', 'dudoan_R']].plot()

```

Executed at 2023.05.30 20:24:26 in 1s 19ms

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 quả không được khả quan lắm. Dưới đây là là kết quả đánh giá độ sai lệch sử dụng MSE:

```
h 32 1 a=np.sqrt(pow(data['dudoan_R']-data['Giá'],2).mean())
      2 print(a)
      Executed at 2023.05.30 20:24:26 in 14ms
      13.259222281843469
```

Sai số lên tới tận 13.25 tức là độ sai lệch giá.

Công thức:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

Trong đó:

y là biến mục tiêu (giá nhà)

$x_1, x_2, \dots, x_n$  là các biến đầu vào (các thuộc tính như số tầng, số phòng, số toilet, v.v.)

$\beta_0, \beta_1, \beta_2, \dots, \beta_n$  là các hệ số tương ứng với từng biến đầu vào

Mô hình Linear Regression tìm cách xác định các hệ số  $\beta_0, \beta_1, \beta_2, \dots, \beta_n$  sao cho tổng bình phương sai số giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất (phương pháp bình phương tối thiểu).

Các bước thực hiện:

- Đọc dữ liệu từ file CSV vào DataFrame data sử dụng `pd.read_csv('hanoi.csv')`.
- Hiển thị và kiểm tra thông tin về dữ liệu bằng `data.info()`.
- Tìm số lượng giá trị duy nhất trong mỗi cột bằng `data.nunique()`.
- Kiểm tra và hiển thị số lượng giá trị thiếu trong từng cột sử dụng vòng lặp `for`.
- Hiển thị các giá trị duy nhất trong cột 'Giá' bằng `data['Giá'].unique()`.
- Vẽ biểu đồ đường của cột 'Giá' bằng `plt.plot(data['Giá'])`.

- Xử lý các giá trị thiếu trong dữ liệu bằng cách thay thế chúng bằng giá trị trung bình của cột tương ứng. Các cột 'Số tầng', 'Số phòng ngủ' và 'Số toilet' được thay thế bằng giá trị trung bình nguyên (sử dụng `np.trunc`) của cột đó.
- Vẽ biểu đồ đếm các giá trị duy nhất trong các cột dữ liệu phân loại sử dụng `sns.countplot`.
- Xác định các ngoại lệ (outliers) trong dữ liệu sử dụng hàm `outliers` và lưu các chỉ mục của chúng vào danh sách `raws`.
- Loại bỏ các dòng dữ liệu chứa ngoại lệ khỏi data sử dụng `data.drop`.
- Vẽ các biểu đồ scatterplot giữa các biến đầu vào và biến mục tiêu 'Giá' sử dụng `scatter`.
- Tiến hành chuẩn bị dữ liệu để huấn luyện mô hình. Xác định các biến đầu vào (X) bằng cách loại bỏ cột 'Giá' từ data, và biến mục tiêu (y) là cột 'Giá'.
- Chuyển đổi các biến đầu vào phân loại thành dạng số sử dụng `LabelEncoder` từ `sklearn.preprocessing`.
- Vẽ ma trận tương quan giữa các biến đầu vào sử dụng `sns.heatmap`.
- Huấn luyện mô hình `Linear Regression` trên tập huấn luyện (`X_train`, `y_train`) sử dụng `fit`.
- Lưu mô hình đã huấn luyện vào file '`model.sav`' sử dụng.

## **b) 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.

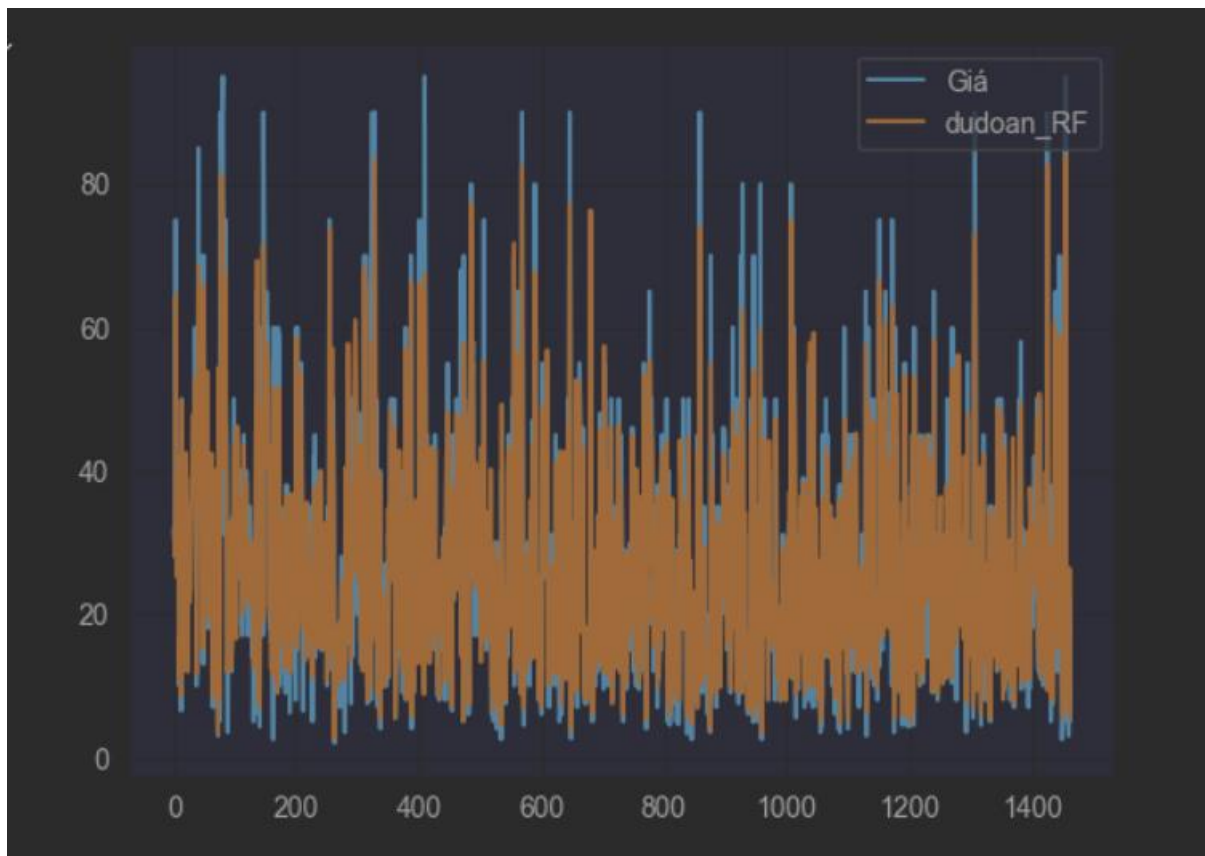
Mô hình `RandomForestRegressor` là một tập hợp các cây quyết định độc lập với nhau. Mỗi cây quyết định được huấn luyện trên một tập con của dữ liệu và sử dụng một số thuộc tính ngẫu nhiên để tạo ra các quyết định. Kết quả cuối cùng của `RandomForestRegressor` là sự kết hợp của dự đoán từ tất cả các cây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

```
13.259222281843469

In [33]: 1 from sklearn.ensemble import RandomForestRegressor
2 rf = RandomForestRegressor()
3 rf.fit(X_train,y_train)
4 filename1 = 'model_Random_Forest.sav'
5 pickle.dump(rf, open(filename1, 'wb'))
6 data['dudoan_RF']=rf.predict(X)
7 data[['Giá', 'dudoan_RF']].plot()

Executed at 2023.05.30 20:24:28 in 2s 227ms
```

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



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

```

1 b=np.sqrt(pow(data['dudoan_RF']-data['Giá'],2).mean())
2 print(b)

```

Executed at 2023.05.30 21:26:03 in 88ms

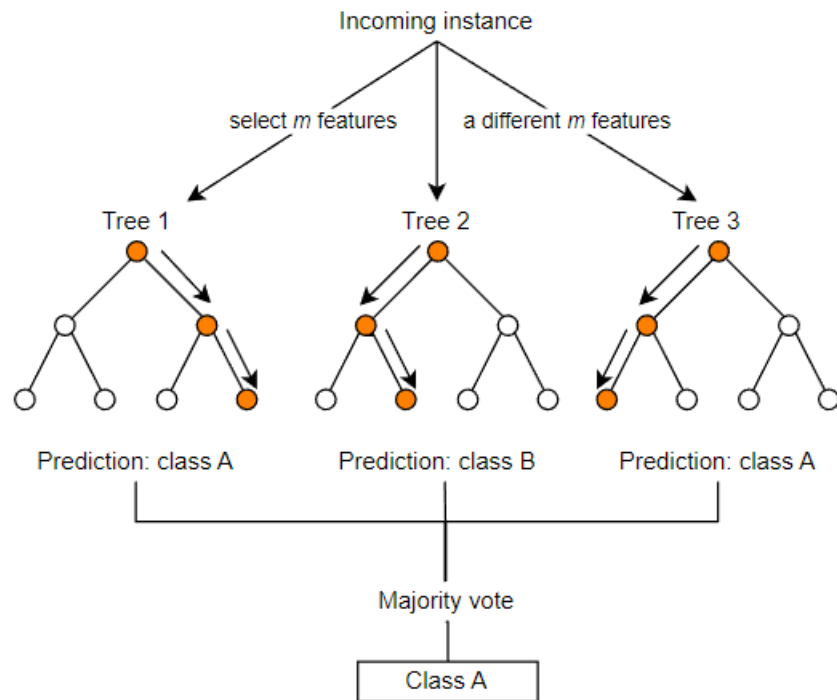
6.208237702570439

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

Công thức:

Thuật toán Random Forest sử dụng nhiều cây quyết định để đưa ra dự đoán. Quá trình xây dựng một Random Forest bao gồm các bước sau đây:

1. Tạo ra k cái bootstrap (k tập con chứa n phần tử được lấy ngẫu nhiên từ tập huấn luyện).
2. Với mỗi tập bootstrap, huấn luyện một cây quyết định.
3. Khi phân loại một điểm dữ liệu mới, dùng kết quả phân loại của các cây để đưa ra dự đoán cuối cùng.



Các bước thực hiện:

- Khởi tạo mô hình RandomForestRegressor và fit dữ liệu X\_train và y\_train.
- Lưu mô hình vào tệp model\_Random\_Forest.sav.
- Dự đoán giá trị 'dudoan\_RF' dựa trên X và vẽ đồ thị so sánh 'Giá' và 'dudoan\_RF'.
- Tính độ lỗi bằng cách tính căn bậc hai của trung bình bình phương sai số giữa 'dudoan\_RF' và 'Giá' và in ra giá trị đó.

### **2.3. Các bước dự đoán**

Bước 1: Chọn các thông tin đầu vào như : đường phố, xã (phường), quận(huyện), số phòng, diện tích, số tầng và số toilet.

Bước 2 : Sau khi chọn đầy đủ thông tin đầu vào thì sẽ ấn nút ‘Gửi’ thì sẽ hiện ra kết quả dự đoán.

### **2.4. Giao diện**

Giao diện hiển thị của hệ thống dự đoán giá nhà, chung cư được thiết kế để cung cấp một trải nghiệm người dùng thuận tiện và trực quan. Giao diện giúp người dùng tương tác với hệ thống thông qua các thành phần và chức năng có sẵn.

Giao diện hiển thị chính của hệ thống gồm một trang web đơn giản, dễ sử dụng và thân thiện với người dùng. Trang web chứa một biểu mẫu cho phép người dùng nhập thông tin về đường phố, xã (phường), quận( huyện), số phòng, diện tích, số tầng và số toilet. Người dùng có thể lựa chọn từ các tùy chọn được cung cấp trong các danh sách thả xuống.

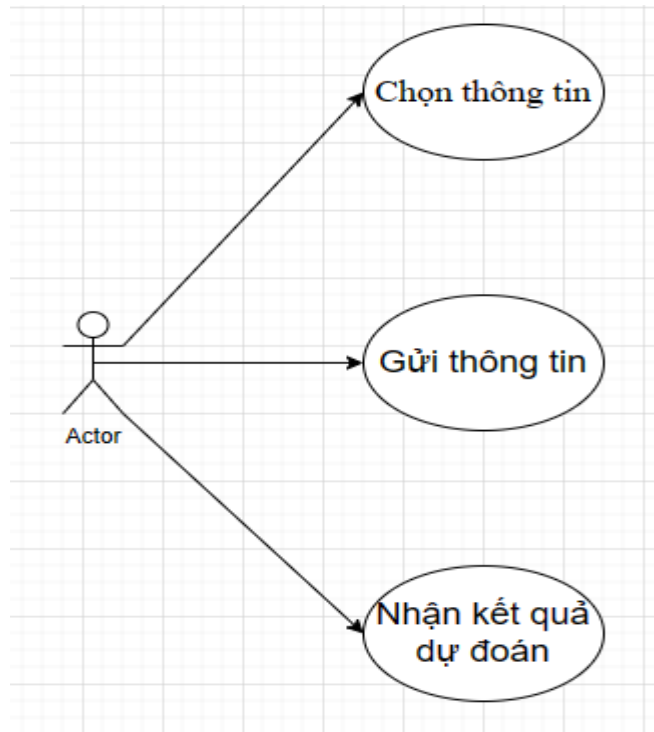
Sau khi người dùng đã nhập đầy đủ thông tin, họ có thể nhấn vào nút "Gửi" để gửi yêu cầu dự đoán giá đến hệ thống. Khi đó, hệ thống sẽ xử lý yêu cầu và hiển thị kết quả dự đoán trên cùng trang web. Kết quả sẽ được hiển thị dưới dạng một con số, đại diện cho giá trị dự đoán của nhà, chung cư.

Giao diện hiển thị của hệ thống dự đoán giá nhà, chung cư được thiết kế đơn giản, tối giản hóa các yếu tố phức tạp để tạo ra trải nghiệm người dùng trực

quan và dễ sử dụng. Nó giúp người dùng tương tác và nhận được kết quả dự đoán một cách nhanh chóng và dễ dàng.

### 3. Phân tích hướng đối tượng

#### 3.1. Biểu đồ use case



Người dùng: Đại diện cho người dùng cuối, người sử dụng hệ thống. Có các hành động sau:

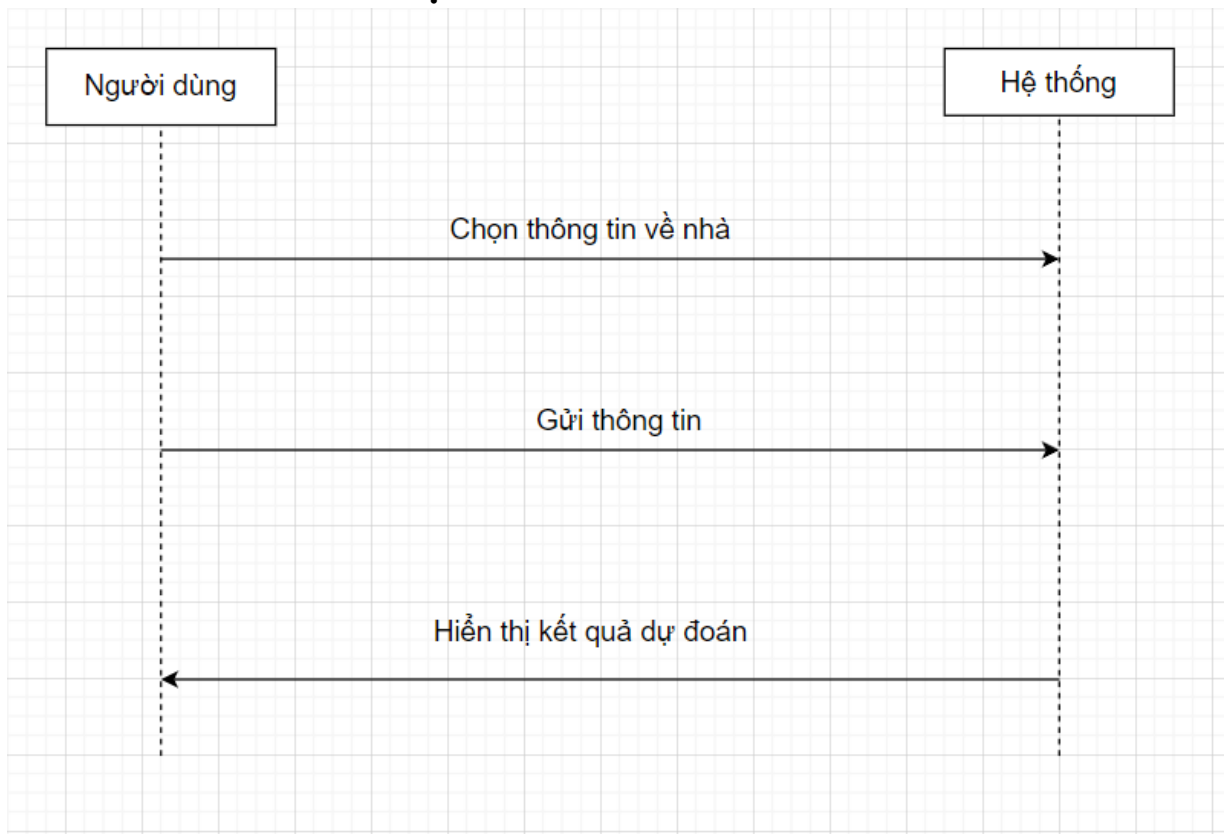
Chọn thông tin xe: Người dùng chọn các thông tin về đường phố, xã (phường), quận( huyện), số phòng, diện tích, số tầng và số toilet.

Gửi thông tin xe: Người dùng gửi thông tin xe đã chọn đến hệ thống để thực hiện dự đoán giá.

Nhận kết quả dự đoán giá: Người dùng nhận kết quả dự đoán giá của xe từ hệ thống và hiển thị cho người dùng.



### 3.2. Biểu đồ tuần tự



Trong biểu đồ trên, có hai đối tượng chính là "Người dùng" và "Hệ thống". Các thông điệp (message) được gửi từ người dùng đến hệ thống và ngược lại

## 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.

### 2. Xây dựng chương trình

```
import streamlit as st
import pickle
import numpy as np

load = pickle.load(open('model_Random_Forest.sav', 'rb'))

1 usage
def main():
    # Mảng lựa chọn
    quanhuyen = [...]
    xaphuong = [...]
    duongpho = [...]

    # Mảng mã
    code_quanhuyen = [...]
    code_xaphuong = [...]
    code_duongpho = [...]
    # Tạo từ điển từ hai mảng
    quanhuyen_dict = dict(zip(quanhuyen, code_quanhuyen))
    xaphuong_dict = dict(zip(xaphuong, code_xaphuong))
    duongpho_dict = dict(zip(duongpho, code_duongpho))

    # Tạo select box
    selected_quanhuyen = st.selectbox("Chọn Quận(Huyện):", quanhuyen)
    selected_xaphuong = st.selectbox("Chọn Xã(Phường, Thị trấn):", xaphuong)
    selected_duongpho = st.selectbox("Chọn Đường(Phố):", duongpho)
```

```

# Chuyển đổi lựa chọn thành mã
change_code_quanhuyen = quanhuyen_dict[selected_quanhuyen]
change_code_xaphuong = xaphuong_dict[selected_xaphuong]
change_code_duongpho = duongpho_dict[selected_duongpho]

# Lựa chọn số tầng
sotang = ["{}"].format(i) for i in range(1,10)]
luachontang = st.selectbox("Số tầng:", sotang)

# Nhập diện tích
text_dientich = st.number_input("Diện tích:", value=0, step=1)
text_sophong = st.number_input("Số phòng:", value=0, step=1)
text_sotoilet = st.number_input("Số Toilet:", value=0, step=1)

button = st.button('GỬI')
if button:
    data = np.array([change_code_duongpho, change_code_xaphuong, change_code_quanhuyen, text_dientich, luachontang, text_sophong, text_sotoilet])
    predict = load.predict(data)
    st.write('Giá dự đoán: ' + str(predict))
pass

if __name__ == '__main__':
    main()

```

Do code còn hơi dài nên đã thu gọn một số chỗ để 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.

giaodien - Streamlit

localhost:8501

Chọn Quận(Huyện):

Nam Từ Liêm

Chọn Xã(Phường, Thị Trấn):

Mỹ Đình 1

Chọn Đường(Phố):

Nguyễn Hoàng

Số tầng:

4

Diện tích:

20 - +

Số phòng:

4 - +

Số Toilet:

3 - +

Gửi

Giá dự đoán: [9,761]

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

### **2. Ưu, nhược điểm**

#### **a) Ưu điểm**

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

#### **b) Nhược điểm.**

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

### **3. 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/>

[2] <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/>

[5] <https://solieu.vip/mse-va-rmse-la-gi-va-cach-tinh-tren-stata/>