

TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP THÁI NGUYÊN

KHOA: ĐIỆN TỬ

BỘ MÔN: CÔNG NGHỆ THÔNG TIN



ĐỒ ÁN

MÔN HỌC
THỰC TẬP CHUYÊN NGÀNH

Sinh viên: Nguyễn Đình Đức

Lớp: K55KMT

Giáo viên hướng dẫn: TS.Nguyễn Văn Huy

Thái Nguyên - 2023

ĐỒ ÁN

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

Sinh viên: Nguyễn Đình Đức.

Lớp: K55KMT.

Nghành: Kỹ Thuật Máy Tính.

Giáo viên hướng dẫn: TS.Nguyễn Văn Huy.

Ngày giao đề: .../.../.....

Ngày hoàn thành: .../.../.....

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

2. Yêu cầu:

- Thu thập dữ liệu từ website.
- Phân tích dữ liệu.
- Chọn mô hình để huấn luyện.
- Xây dựng website dự đoán giá chung cư.

3. 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)

NHẬN XÉT CỦA GIÁO VIÊN CHĂM

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Thái Nguyên, ngày... tháng ... năm 20...

GIÁO VIÊN CHĂM

(Ký và ghi rõ họ, tên)

MỤC LỤC

MỤC LỤC	5
MỤC LỤC HÌNH ẢNH.....	6
LỜI NÓI ĐẦU	7
CHƯƠNG I: GIỚI THIỆU CHUNG	8
1. Đoán giá là gì?	8
2. Ứng dụng	8
3. Công nghệ	9
4. Áp dụng	11
CHƯƠNG II: CƠ SỞ LÝ THUYẾT	12
1. Xử lý dữ liệu.	12
a) Crawl data.	12
b) Xử lý data trên Excel.	13
c) Xử lý data với python.	14
2. Model Random Forest	18
CHƯƠNG III: XÂY DỰNG CHƯƠNG TRÌNH.....	20
1. Framework Django.	20
2. Xây dựng web.....	20
a) Front-end.....	20
b) Back-end.....	21
c) Kiểm thử.....	22
CHƯƠNG IV: KẾT LUẬN	23
1. Tổng quan.....	23
2. Ưu, nhược điểm.....	23
a) Ưu điểm	23
b) Nhược điểm.	23
3. Hướng phát triển.....	23
TÀI LIỆU THAM KHẢO.....	24

MỤC LỤC HÌNH ẢNH

Hình 1: Đoán giá là gì?.....	8
Hình 2: https://batdongsan.com.vn/	11
Hình 3: Dữ liệu crawl trên Octoparse 8.	12
Hình 4: Dữ liệu thô được đưa về Excel xử lý.	13
Hình 5: Dữ liệu đã qua xử lý trên Excel.	13
Hình 6: Đọc file CSV bằng thư viện Pandas.....	14
Hình 7: Thông tin kiểu dữ liệu.....	14
Hình 8: Dữ liệu duy nhất.....	15
Hình 9: Dữ liệu bị trống (Nan).	15
Hình 10: Điền dữ liệu vào dữ liệu bị trống bằng hàm mean().	16
Hình 11: Tạo hàm Outlier.....	16
Hình 12: Tìm các dữ liệu khác thường.	17
Hình 13: Mã hoá các dữ liệu object.	17
Hình 14: Dữ liệu đã xử lý xong, chuẩn bị training.....	17
Hình 15: Xây dựng mô hình bằng phương pháp Random Forest Regressor.....	18
Hình 16: Sơ đồ dự đoán giá so với dữ liệu đầu vào của mô hình được xây dựng bằng phương pháp Random Forest Regressor.....	19
Hình 17: Chỉ số RMSE và MSE của mô hình xây dựng bằng phương pháp Random Forest Regressor.....	19
Hình 18: HTML phần giao diện web.....	21
Hình 19: Hàm Home() để đưa dữ liệu nhập từ web vào mô hình dự đoán và đưa kết quả ra web.	21
Hình 20: Giao diện dự đoán và thử dự đoán.	22
Hình 21: Giá của chung cư trên web chính.....	22

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 chung cư 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 chung cư khu vực Hà Nội”. Để làm đề tài này tôi dùng kiến thức của khoa học dữ liệu, học máy và lập trình web.

Đề này nhằm mục tiêu sử dụng dữ liệu được lấy trên website để phân tích và xây dựng mô hình dự đoán giá 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 chung cư và xây dựng mô hình dự đoán chính xác.

CHƯƠNG I: GIỚI THIỆU CHUNG

1. Đoán giá là gì?



Hình 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

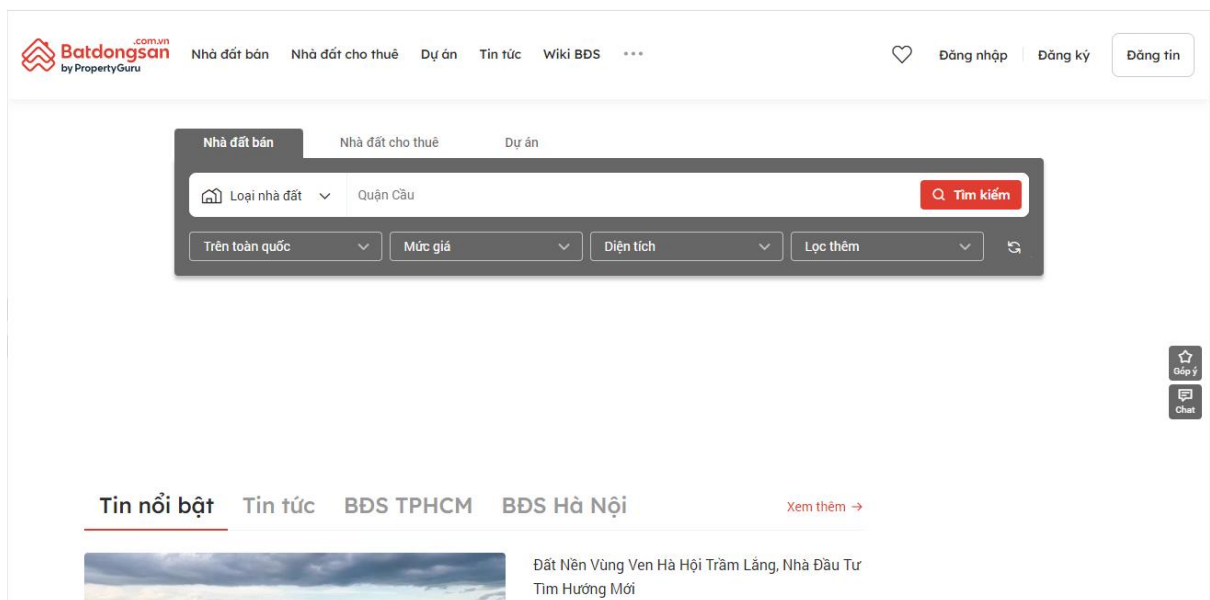
Công nghệ đóng vai trò quan trọng trong việc đoán giá và các ứng dụng liên quan. Dưới đây là một số công nghệ thường được sử dụng trong việc đoán giá:

- 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



Hình 2: <https://batdongsan.com.vn/>.

Đề tài “Dự đoán giá 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 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. 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	Dự án Le Grand Jardin Sài Đồng...	Mức giá 2.5 tỷ ... Sao chép liên kết	Đặc điểm bất động sản ... Số ph...	More
2	Dự án Golden Palace, đường Mê...	Mức giá 28 triệu/m ² ... Sao chép ...	Đặc điểm bất động sản ... lý Số đ...	More
3	Phường Nguyễn Du, Hai Bà Trun...	Mức giá 1.2 tỷ ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
4	Dự án C14 - Bộ Công An, Đường...	Mức giá 3.2 tỷ ... Sao chép liên kết	Thông tin mô tả Gia đình tôi cần...	More
5	Dự án Tecco Garden, đường Tứ ...	Mức giá 25 triệu/m ² ... Sao chép ...	Đặc điểm bất động sản ... Số toil...	More
6	Dự án Moonlight 1 - An Lạc Gre...	Mức giá 3.3 tỷ ... Sao chép liên kết	Đặc điểm bất động sản ... Số ph...	More
7	Dự án Tecco Garden, đường Tứ ...	Mức giá 3.5 tỷ ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
8	Dự án Hoàng Thành Pearl, Dườn...	Mức giá 3.9 tỷ ... Sao chép liên kết	Đặc điểm bất động sản ... tủ bếp...	More
9	Dự án Masteri West Heights, Ph...	Mức giá 2.9 tỷ ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
10	Dự án Moonlight I - An Lạc Gree...	Mức giá 2.45 tỷ ... Sao chép liên ...	Đặc điểm bất động sản ... Nội th...	More
11	Dự án Masteri West Heights, Ph...	Mức giá 1.8 tỷ ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
12	Dự án Sunshine City, Phường Đò...	Mức giá 45 triệu/m ² ... Sao chép ...	Đặc điểm bất động sản ... Nội th...	More
13	Dự án Vinhomes Ocean Park Gia...	Mức giá 1.7 tỷ ... Sao chép liên kết	Đặc điểm bất động sản ... Nội th...	More
14	Dự án Sunshine Garden, Đường ...	Mức giá 5.57 tỷ ... Sao chép liên ...	Đặc điểm bất động sản ... thươn...	More
15	Dự án Park Kiara, Đường Lê Tron...	Mức giá 3.2 tỷ ... Sao chép liên kết	Đặc điểm bất động sản ... lý Số đ...	More

Hình 3: Dữ liệu crawl trên Octoparse 8.

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.

hanoi_department_price - Copy.csv

DC	Dự án/ khu	Số	Ngõ	Đường/Phố	Xã/ Phường/ Thị trấn Quận, Huyện
	Dự án Le Grand Jardin	Dự án Le C	Đường Huỳnh Văn N	Đường Huỳnh Văn N	Đường Huỳnh Văn N
	Dự án Golden Palace	Dự án Gol	đường Mê Trì	Phườn đường Mê Trì	Phườn đường Mê Trì
	Phường Nguyễn Du	Hai Bà Trư	Phường Nguyễn Du	Phường Nguyễn Du	Phường Nguyễn Du
	Dự án Tecco Garden	Dự án Tec	đường Từ Hiệp	Xã T đường Từ Hiệp	Thanh Tr Xã Từ Hiệp
	Dự án Moonlight 1	Dự án Moi	Xã An Khánh	Hoài Đ Xã An Khánh	Hoài Đ Xã An Khánh
	Dự án Tecco Garden	Dự án Tec	đường Từ Hiệp	Xã T đường Từ Hiệp	Thanh Tr Xã Từ Hiệp
	Dự án Hoàng Thành I	Dự án Hoà	Đường Nguyễn Văn	Đường Nguyễn Văn	Đường Nguyễn Văn
	Dự án Masteri West	Dự án Mai	Phường Tây Mỗ	Nar Phường Tây Mỗ	Nar Phường Tây Mỗ
	Dự án Moonlight 1	Dự án Moi	Xã An Khánh	Hoài Đ Xã An Khánh	Hoài Đ Xã An Khánh
	Dự án Masteri West	Dự án Mai	Phường Tây Mỗ	Nar Phường Tây Mỗ	Nar Phường Tây Mỗ
	Dự án Sunshine City	Dự án Sun	Phường Đồng Ngạc	Phường Đồng Ngạc	Phường Đồng Ngạc
	Dự án Vinhomes Ocean Park	Dự án Vin	Xã Dương Xá	Gia Lãi Xã Dương Xá	Gia Lãi Xã Dương Xá
	Dự án Sunshine Garden	Dự án Sun	Đường Dương Văn B	Đường Dương Văn B	Đường Dương Văn B
	Dự án Park Kiara	Dự án Parl	Đường Lê Trọng Tấn	Đường Lê Trọng Tấn	Đường Lê Trọng Tấn
	Dự án Moonlight 1	Dự án Moi	Xã An Khánh	Hoài Đ Xã An Khánh	Hoài Đ Xã An Khánh
	Dự án The Manor	Dự án The	Đường Mê Trì	Phườn đường Mê Trì	Phườn đường Mê Trì
	Discovery Complex	Discovery	302	Đường Cầu Giấy	Phi Đường Cầ
	Dự án N01-T7 Ngoại	Dự án N01	Phường Xuân Tảo	Bí Phường Xuân Tảo	Bí Phường Xuân Tảo
	Dự án Feliz Homes	Dự án Feli	Đường Hoàng Mai	P Đường Ho	Đường Ho

Hình 4: Dữ liệu thô được đưa về Excel xử lý.

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.

price_department_in_hanoi.csv

Dự án/ khu	Số	Ngõ	Đường/Phố	Xã/ Phường/ Thị	Quận/Huyện	Diện tích(m2)	Giá (tỷ đồng)	Hướng nhà	Hướng ban công	Số phòng ngủ	Số toilet
Le Grand Jardin Sài Đồng			Huỳnh Văn Nghệ	Sài Đồng	Long Biên	65	2.5			2	
Golden Palace			Mê Trì	Mê Trì	Nam Từ Liêm	118	3.3	Bắc	Nam	3	2
Tecco Garden			Nguyễn Du	Nguyễn Du	Hai Bà Trưng	30	1.2				
Moonlight 1 - An Lạc Green Symphony			Tứ Hiệp	Tứ Hiệp	Thanh Trì	127	3.18			3	2
Tecco Garden			An Khánh	Tứ Hiệp	Hoài Đức	81	3.3		Đông - Nam	3	
Hoàng Thành Pearl			Nguyễn Văn Giáp	Tứ Hiệp	Thanh Trì	141	3.5	Đông - Bắc		4	
Masteri West Heights				Nguyễn Văn Giáp	Nam Từ Liêm	80	3.9			2	2
Moonlight 1 - An Lạc Green Symphony				Tây Mỗ	Nam Từ Liêm	54	2.9	Tây - Bắc	Đông - Nam	2	2
Masteri West Heights				An Khánh	Hoài Đức	66.68	2.45	Đông - Nam	Tây - Bắc	2	2
Sunshine City				Tây Mỗ	Nam Từ Liêm	30	1.8	Tây - Bắc	Đông - Nam	1	1
Vinhomes Ocean Park Gia Lâm				Đông Ngạc	Bắc Từ Liêm	97	4.36			3	2
Sunshine Garden				Dương Xá	Gia Lâm	60	1.7			2	2
Park Kiara				Dương Văn Bé	Vĩnh Tuy	136	5.57			3	2
Moonlight 1 - An Lạc Green Symphony				Lê Trọng Tấn	La Khê	60	3.2			2	1
The Manor				An Khánh	Hoài Đức	66.68	2.433			2	2
Discovery Comple	302			Mê Trì	Mỹ Đình 1	189	9			3	
N01-T7 Ngoại Giao Đoàn				Cầu Giấy	Dịch Vọng	93	4.6			2	
Feliz Homes				Cầu Giấy	Xuân Tảo	104	7.49	Đông		3	
				Hoàng Mai	Hoàng Văn Thu	45	2.1	Nam	Nam	1	1

Hình 5: Dữ liệu đã qua xử lý trên Excel.

Qua xử lý dữ liệu trên Excel thì dữ liệu còn khoảng hơn 5000 dòng.

c) Xử lý data với python.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

data = pd.read_csv('price_department_in_hanoi.csv')
data.head()
```

[70] Python

...

	Dự án/ khu	Số	Ngõ	Đường/Phố	Xã/ Phường/ Thị trấn	Quận/Huyện	Diện tích(m2)	Giá (tỷ đồng)	Hướng nhà	Hướng ban công	Số phòng ngủ	Số toilet
0	Le Grand Jardin Sài Đông	NaN	NaN	Huỳnh Văn Nghệ	Sài Đông	Long Biên	65.0	2.50	NaN	NaN	2.0	NaN
1	Golden Palace	NaN	NaN	Mễ Trì	Mễ Trì	Nam Từ Liêm	118.0	3.30	Bắc	Nam	3.0	2.0
2	NaN	NaN	NaN	NaN	Nguyễn Du	Hai Bà Trưng	30.0	1.20	NaN	NaN	NaN	NaN
3	Tecco Garden	NaN	NaN	Tứ Hiệp	Tứ Hiệp	Thanh Trì	127.0	3.18	NaN	NaN	3.0	2.0
4	Moonlight 1 - An Lạc Green Symphony	NaN	NaN	NaN	An Khánh	Hoài Đức	81.0	3.30	NaN	Đông - Nam	3.0	NaN

Hình 6: Đọc file CSV bằng thư viện Pandas.

Đầu tiên sử 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.

```
data.info()
```

[71]

...

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4884 entries, 0 to 4883
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   Dự án/ khu            4547 non-null  object
1   Số                    723 non-null   object
2   Ngõ                   12 non-null    object
3   Đường/Phố            3898 non-null  object
4   Xã/ Phường/ Thị trấn 4752 non-null  object
5   Quận/Huyện          4884 non-null  object
6   Diện tích(m2)         4884 non-null  float64
7   Giá (tỷ đồng)         4884 non-null  float64
8   Hướng nhà            2368 non-null  object
9   Hướng ban công        2550 non-null  object
10  Số phòng ngủ          4508 non-null  float64
11  Số toilet             3736 non-null  float64
dtypes: float64(4), object(8)
memory usage: 458.0+ KB
```

Hình 7: Thông tin kiểu dữ liệu.

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.

```
[72] data.nunique()
...
Dự án/ khu      511
Số              134
Ngõ             11
Đường/Phố      294
Xã/ Phường/ Thị trấn  167
Quận/Huyện      20
Diện tích(m2)   584
Giá (tỷ đồng)   698
Hướng nhà       8
Hướng ban công  8
Số phòng ngủ    7
Số toilet       6
dtype: int64
```

Hình 8: Dữ liệu duy nhất.

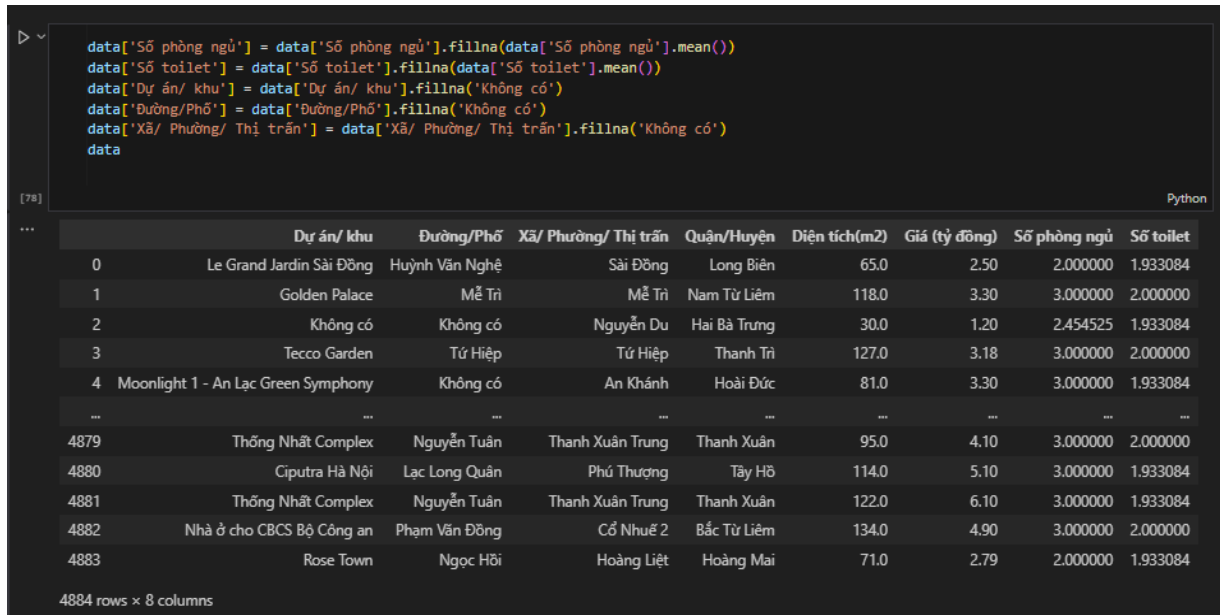
Sau khi xem kiểu dữ liệu của dữ liệu đầu vào, dùng nunique() để xem có bao nhiêu dữ liệu duy nhất trong dữ liệu đầu vào.

```
[74] data.isnull().sum()
...
Dự án/ khu      337
Số              4161
Ngõ             4872
Đường/Phố      986
Xã/ Phường/ Thị trấn  132
Quận/Huyện      0
Diện tích(m2)   0
Giá (tỷ đồng)   0
Hướng nhà       2516
Hướng ban công  2334
Số phòng ngủ     376
Số toilet       1148
dtype: int64
```

Hình 9: Dữ liệu bị trống (Nan).

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 dòng đó sẽ bị xóa 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. Đối với dữ liệu kiểu object thì tôi điền là không có.



```
data['Số phòng ngủ'] = data['Số phòng ngủ'].fillna(data['Số phòng ngủ'].mean())
data['Số toilet'] = data['Số toilet'].fillna(data['Số toilet'].mean())
data['Dự án/ khu'] = data['Dự án/ khu'].fillna('Không có')
data['Đường/Phố'] = data['Đường/Phố'].fillna('Không có')
data['Xã/ Phường/ Thị trấn'] = data['Xã/ Phường/ Thị trấn'].fillna('Không có')
```

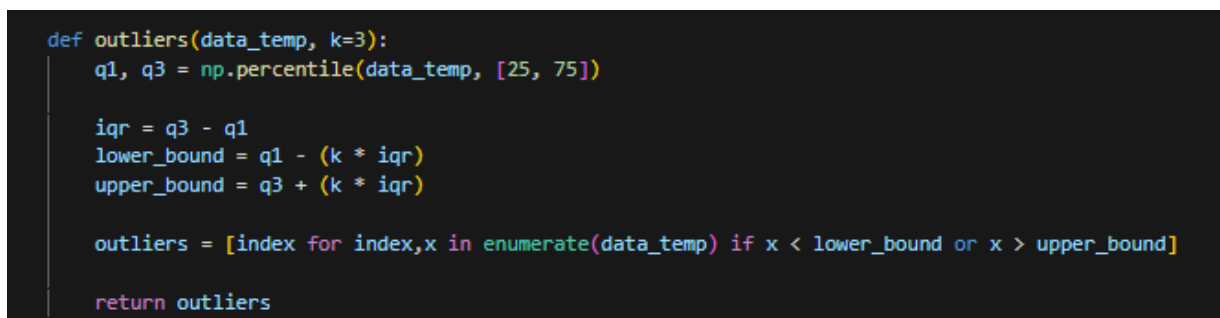
	Dự án/ khu	Đường/Phố	Xã/ Phường/ Thị trấn	Quận/Huyện	Diện tích(m2)	Giá (tỷ đồng)	Số phòng ngủ	Số toilet
0	Le Grand Jardin Sài Đồng	Huỳnh Văn Nghệ	Sài Đồng	Long Biên	65.0	2.50	2.000000	1.933084
1	Golden Palace	Mễ Trì	Mễ Trì	Nam Từ Liêm	118.0	3.30	3.000000	2.000000
2	Không có	Không có	Nguyễn Du	Hai Bà Trưng	30.0	1.20	2.454525	1.933084
3	Tecco Garden	Tứ Hiệp	Tứ Hiệp	Thanh Trì	127.0	3.18	3.000000	2.000000
4	Moonlight 1 - An Lạc Green Symphony	Không có	An Khánh	Hoài Đức	81.0	3.30	3.000000	1.933084
...
4879	Thống Nhất Complex	Nguyễn Tuấn	Thanh Xuân Trung	Thanh Xuân	95.0	4.10	3.000000	2.000000
4880	Ciputra Hà Nội	Lạc Long Quân	Phú Thượng	Tây Hồ	114.0	5.10	3.000000	1.933084
4881	Thống Nhất Complex	Nguyễn Tuấn	Thanh Xuân Trung	Thanh Xuân	122.0	6.10	3.000000	1.933084
4882	Nhà ở cho CBCS Bộ Công an	Phạm Văn Đồng	Cổ Nhuế 2	Bắc Từ Liêm	134.0	4.90	3.000000	2.000000
4883	Rose Town	Ngọc Hồi	Hoàng Liệt	Hoàng Mai	71.0	2.79	2.000000	1.933084

4884 rows x 8 columns

Hình 10: Điền dữ liệu vào dữ liệu bị trống bằng hàm `mean()`.

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.



```
def outliers(data_temp, k=3):
    q1, q3 = np.percentile(data_temp, [25, 75])

    iqr = q3 - q1
    lower_bound = q1 - (k * iqr)
    upper_bound = q3 + (k * iqr)

    outliers = [index for index, x in enumerate(data_temp) if x < lower_bound or x > upper_bound]

    return outliers
```

Hình 11: Tạo hàm `Outlier`.

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.

```
cont_features = np.array([i for i in data.columns.tolist() if data[i].dtype != 'object'])
rows = []
rows += outliers(data['Giá (tỷ đồng)'])
len(set(rows))
```

[81] Python

Hình 12: Tìm các dữ liệu khác thường.

Trong bài thì tôi dùng tìm các outlier cho cột giá và tìm được 101 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])
```

Hình 13: Mã hoá các dữ liệu object.

Dưới đây là dữ liệu đã được xử lý qua tất các công đoạn ở trên:

X.head([10])

[31] ✓ 0.0s Python

	Dự án/ khu	Đường/Phố	Xã/ Phường/ Thị trấn	Quận/Huyện	Diện tích(m2)	Số phòng ngủ	Số toilet
0	245	52	95	9	65.00	2.000000	1.933084
1	140	115	56	11	118.00	3.000000	2.000000
2	221	72	66	4	30.00	2.454525	1.933084
3	391	236	129	12	127.00	3.000000	2.000000
4	269	72	0	5	81.00	3.000000	1.933084
5	391	236	129	12	141.00	4.000000	1.933084
6	179	148	12	11	80.00	2.000000	2.000000
7	263	72	126	11	54.00	2.000000	2.000000
8	270	72	0	5	66.68	2.000000	2.000000
9	263	72	126	11	30.00	1.000000	1.000000

Hình 14: Dữ liệu đã xử lý xong, chuẩn bị training.

2. Model Random Forest

Random Forest Regressor 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 Random Forest (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 Random Forest Regressor 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 Random Forest Regressor là sự kết hợp của dự đoán từ tất cả các cây quyết định.

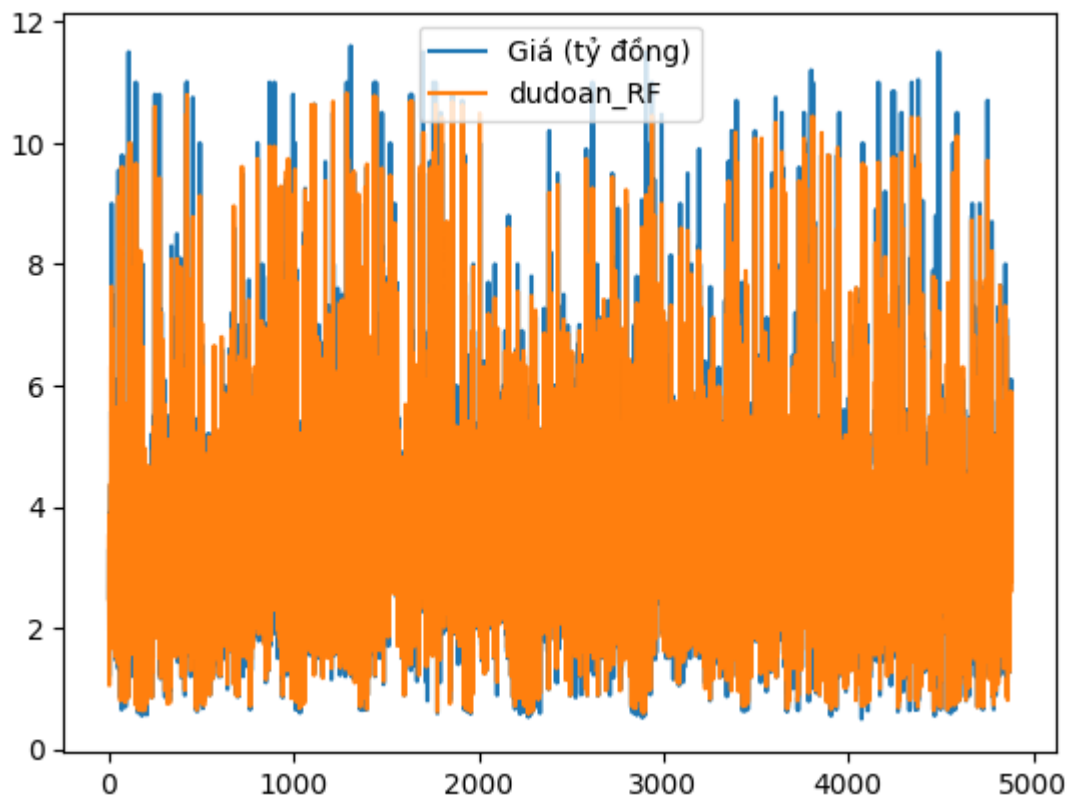
```
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
import pickle

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

rf = RandomForestRegressor()
rf.fit(X_train, y_train)
filename1 = 'model_Random_Forest.sav'
pickle.dump(rf, open(filename1, 'wb'))
data['dudoan_RF'] = rf.predict(X)
data[['Giá (tỷ đồng)', 'dudoan_RF']].plot()
```

✓ 4.0s Python

Hình 15: Xây dựng mô hình bằng phương pháp Random Forest Regressor.



Hình 16: Sơ đồ dự đoán giá so với dữ liệu đầu vào của mô hình được xây dựng bằng phương pháp Random Forest Regressor.

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

```

c=np.sqrt(pow(data['dudoan_RF']-data['Giá (tỷ đồng)'],2).mean())
d = pow(data['dudoan_RF']-data['Giá (tỷ đồng)'],2).mean()
print(c)
print(d)

```

[27] ✓ 0.0s

... 0.37258586280150463
0.13882022515954162

Hình 17: Chỉ số RMSE và MSE của mô hình xây dựng bằng phương pháp Random Forest Regressor.

Kết quả với chỉ số MSE là 0.138 và RMSE là 0.3725. Hai chỉ số này đều khá gần 0 nên có thể nói rằng mô hình dự đoán khá chính xác.

CHƯƠNG III: XÂY DỰNG CHƯƠNG TRÌNH

1. Framework Django.

Django là một khung Web Python cấp cao, khuyến khích phát triển nhanh chóng và thiết kế thực dụng, gọn gàng. Được xây dựng bởi các nhà phát triển có kinh nghiệm, nó xử lý nhiều rắc rối của việc phát triển Web, vì vậy bạn có thể tập trung vào viết ứng dụng của mình mà không cần phải phát minh lại bánh xe. Nó có nguồn mở và miễn phí.

Với django bạn có thể lấy các ứng dụng web từ ý tưởng để khởi chạy trong vài phút. Và để làm được điều này thì django có một vài tính năng nhẹ như sau:

- Nhanh: Django được thiết kế để giúp các nhà phát triển đưa các ứng dụng từ ý tưởng đến hoàn thành càng nhanh càng tốt..
- Có đầy đủ các thư viện/module cần thiết: Django bao gồm hàng tá các tính năng bổ sung mà bạn có thể sử dụng để xử lý các tác vụ phát triển Web phổ biến. Django chăm sóc xác thực người dùng, quản trị nội dung, bản đồ trang web, nguồn cấp dữ liệu RSS và nhiều tác vụ khác - ngay lập tức.
- Đảm bảo về tính bảo mật: Django rất coi trọng vấn đề bảo mật và giúp các nhà phát triển tránh được nhiều lỗi bảo mật phổ biến, chẳng hạn như SQL SQL, kịch bản chéo trang, giả mạo yêu cầu chéo trang và nhấp chuột. Hệ thống xác thực người dùng của nó cung cấp một cách an toàn để quản lý tài khoản và mật khẩu người dùng.
- Khả năng mở rộng tốt: Một số địa điểm bận rộn nhất trên hành tinh sử dụng khả năng có thể mở rộng nhanh chóng và linh hoạt của django để đáp ứng nhu cầu giao thông nặng nhất.
- Tính linh hoạt: Các công ty, tổ chức và chính phủ đã sử dụng Django để xây dựng tất cả mọi thứ - từ hệ thống quản lý nội dung đến mạng xã hội đến nền tảng điện toán khoa học.

2. Xây dựng web.

a) Front-end.

```

<body class="center">

    <form method="post" action="">
        {% csrf_token %}
        <div class="khung" style="display: flex;flex-direction: column;">
            <h4>Giá dự đoán: {{predict}} tỷ đồng</h4>
            <div><label for="id_district_quanhuyen">Quận/Huyện:</label>
                {{form.district_quanhuyen}}</div>
            <div><label for="id_district_xaphuong">Xã/Phường/Thị Trấn:</label>
                {{form.district_xaphuong}}</div>
            <div><label for="id_district_duongpho">Đường/Phố:</label>
                {{form.district_duongpho}}</div>
            <div><label for="id_district_duan">Dự án/Khu:</label>
                {{form.district_duan}}</div>
            <div><label for="id_district_dientich">Diện tích:</label>
                {{form.district_dientich}}</div>
            <div><label for="id_district_phongngu">Số phòng ngủ:</label>
                {{form.district_phongngu}}</div>
            <div><label for="id_district_toilet">Số Toilet:</label>
                {{form.district_toilet}}</div>
            <button class="button-27" role="button" type="submit">Dự đoán</button>
        </div>
    </form>

</body>

```

Hình 18: HTML phân giao diện web.

b) Back-end.

```

def Home(request):
    predict = 0
    if request.method == 'POST':
        form = DistrictForm(request.POST)

        if form.is_valid():
            selected_district_code_quanhuyen = form.cleaned_data['district_quanhuyen']
            selected_district_quanhuyen = dict(form.fields['district_quanhuyen'].choices)[int(selected_district_code_quanhuyen)]
            selected_district_code_xaphuong = form.cleaned_data['district_xaphuong']
            selected_district_xaphuong = dict(form.fields['district_xaphuong'].choices)[int(selected_district_code_xaphuong)]
            selected_district_code_duongpho = form.cleaned_data['district_duongpho']
            selected_district_duongpho = dict(form.fields['district_duongpho'].choices)[int(selected_district_code_duongpho)]
            selected_district_code_duan = form.cleaned_data['district_duan']
            selected_district_duan = dict(form.fields['district_duan'].choices)[int(selected_district_code_duan)]
            selected_district_code_dientich = request.POST.get('district_dientich')
            selected_district_code_phongngu = request.POST.get('district_phongngu')
            selected_district_code_toilet = request.POST.get('district_toilet')
            data = np.array([selected_district_code_duan,selected_district_code_duongpho,
                            selected_district_code_xaphuong,selected_district_code_quanhuyen,
                            selected_district_code_dientich,selected_district_code_phongngu,
                            selected_district_code_toilet]).reshape(1,-1)
            predict = load.predict(data)

        else:
            form = DistrictForm()
            context = {'form': form, 'predict':predict}
            return render(request,'index.html',context)

```

Hình 19: Hàm Home() để đưa dữ liệu nhập từ web vào mô hình dự đoán và đưa kết quả ra web.

c) Kiểm thử.

Giá dự đoán: [3.90015429] tỷ đồng

Quận/Huyện:

Xã/Phường/Thị trấn:

Đường/Phố:

Dự án/Khu:

Diện tích:

Số phòng ngủ:

Số Toilet:

Dự đoán

Hình 20: Giao diện dự đoán và thử dự đoán.

Bán / Hà Nội / Hai Bà Trưng / Căn hộ chung cư tại Times City

Danh sách các căn hộ đang bán rẻ nhất tại Times City - Park Hill tháng 6/2023 LH 0906 289 ***

Dự án Times City, Đường Minh Khai, Phường Vĩnh Tuy, Hai Bà Trưng, Hà Nội

Mức giá	Diện tích	Phòng ngủ
3,7 tỷ	83 m ²	2 PN
~44,58 triệu/m ²		

Thông tin mô tả

C&C Homes đơn vị chuyển nhượng và cho thuê số 1 Times City Park Hill.

Với đội ngũ chuyên viên tư vấn chuyên sâu và am hiểu nhất thị trường, các chuyên gia của C&C Homes sẽ giúp bạn tìm kiếm được căn hộ phù hợp nhất cho nhu cầu của mình.

Sidebar:

- Hòa Bình Green City (18)
- Imperia Sky Garden (52)
- Park Hill Premium - Times City (55)
- Sun Grand City Ancora Residence (29)
- Sunshine Garden (41)
- Times City (168)
- Toà nhà 93 Lò Đúc - Kinh Đô Tower (13)
- Vinhomes Times City - Park

Hình 21: Giá của chung cư trên web chính.

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 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.

- Vẫn có một số giá khi dự đoán còn cách xa giá cũ.
- Dữ liệu còn ít.
- Dữ liệu còn 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
- [4] <https://machinelearningcoban.com/2016/12/28/linearregression/>
- [5] <https://solieu.vip/mse-va-rmse-la-gi-va-cach-tinh-tren-stata/>