TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN CUỐI KỲ

HỌC PHẦN: KHOA HỌC DỮ LIỆU

TÊN ĐỀ TÀI: DỰ ĐOÁN GIÁ CẢ BẤT ĐỘNG SẢN ĐÀ NẴNG

Nhóm	7
Họ Và Tên Sinh Viên	Lớp Học Phần
Trương Quang Khang	2012A
Trần Thị Hương Trinh	
Dương Trí Đức	

ĐÀ NẴNG, 06/2023

TÓM TẮT

Bài báo cáo lần này chúng em giải quyết vấn đề về dự đoán giá cả nhà đất trong khu vực Đà Nẵng, một trong những thành phố đang phát triển nhanh chóng và thu hút nhiều sự quan tâm của nhiều nhà đầu tư.

Để giải quyết vấn đề này, chúng em sử dụng 3 mô hình là Linear regression, Random Forest và Decision Tree Regression. Linear regression là mô hình hồi quy tuyến tính đơn giản đưa ra dự đoán dựa trên các biến độc lập, Random Forest và Decision Tree regression là các mô hình học máy phức tạp hơn, sử dụng nhiều cây quyết định để đưa ra giá cả dự đoán. Cả ba mô hình đều dựa trên các đặc trưng độc lập.

Sau khi áp dụng các mô hình và sử dụng các thuật toán đánh giá mô hình như R2-score, RMSE, MAE thì độ chính xác của kết quả dự đoán trong mô hình Random Forest là cao nhất với kết quả chính xác 76%

BẢNG PHÂN CÔNG NHIỆM VỤ

Sinh viên thực hiện	Các nhiệm vụ	Tự đánh giá
Trương Quang Khang	Craw data	Hoàn thành
	Clean data	Hoàn thành
	Tạo đặc trưng mới	Hoàn thành
	Tiền xử lý dữ liệu	Hoàn thành
	Chuẩn hóa dữ liệu	Hoàn thành
	Lựa chọn đặc trưng	Hoàn thành
	Xây dựng mô hình random forest regressor	Hoàn thành
Trần Thị Hương Trinh	Clean data	Hoàn thành
	Xây dựng mô hình linear regression	Hoàn thành
	So sánh 3 mô hình và đưa ra nhận xét	Hoàn thành
	Viết báo cáo	Hoàn thành
Dương Trí Đức	Xây dựng mô hình decision tree regression	Hoàn thành
	Đánh giá thuật toán theo các metric	Hoàn thành
	Làm slide	

MỤC LỤC

1. Giới thiệu

Như chúng ta đã biết dự đoán giá cả nhà đất là một công cụ quan trọng trong việc định giá bất động sản, giúp cho người mua, người bán và nhà đầu tư có thể đưa ra quyết định đúng đắn về giá cả cũng như đánh giá tiềm năng đúng đắn về các mảnh đất/nhà mà mình định đầu tư, tránh các rủi ro và tối ưu hóa lợi nhuận.

Việc không dự đoán chính xác được giá cả của nhà đất là một bất lợi và gây nhiều rủi ro đối với các nhà đầu tư, người có ý định mua nhà, bán nhà.

Bài toán dự đoán giá trị của đầu ra dựa trên vector đặc trưng đầu vào, ngoài ra giá trị của đầu ra có thể nhận rất nhiều giá trị dương thực khác nhau. Vì vậy đây là một bài toán regression. Vì vậy đề tài lần này chúng em sử dụng ba mô hình học máy là Linear Regression, Random Forest, Decision Tree Regression để dự đoán giá cả nhà .

Việc dự đoán giá cả nhà đất phụ thuộc vào nhiều yếu tố, sau khi trực quan hoá và so sánh mức độ tương quan của từng biến so với biến mục đích là biến giá cả, mô hình của chúng em sử dụng các yếu tố sau để dự đoán: thuộc quận nào, diện tích, loại hình nhà, có giấy chứng nhận không, chiều dài nhà, chiều rộng nhà, số tầng, khoảng cách đến trung tâm, khoảng cách đến biển.

Để biết được mô hình mà chúng ta sử dụng để dư đoán giá cả là có chính xác hay không và độ chính xác là bao nhiêu thì chúng em sử dụng các thước đo để đo độ chính xác của mô hình hồi quy là R2-score kết hợp phương pháp cross -validation để đám bảo tính đúng đắn và tổng quát mô hình. Thêm vào đó, chúng em sử dụng thước đo đánh giá độ lỗi của mô hình dự đoán là RMSE và MAE .

2. Thu thập và mô tả dữ liệu

2.1 Thu thập dữ liệu

Dữ liệu dự án sử dụng là dữ liệu nhà ở được bán ở Đà Nẵng, Ta sẽ crawl dữ liệu từ trang: Bán nhà Đà Nẵng tháng 5/2023 (alonhadat.com.vn)
Sử dụng ba thư viện là Beautifulsoup4, CSV và Request để Crawl Dữ liệu:

- Thư viện request: ta sẽ sử dụng để gửi một request đến trang muốn crawl và sẽ nhận về một response là một trang html và các thẻ trong trang đó

- Thư viện beautifulsoup4: sử dụng để thao tác trên các trang html mới lấy về như tìm dữ liệu theo các thẻ, tìm dữ liệu theo tên của các class …

- Thư viện CSV: sau khi toàn tất crawl ta sẽ sử dụng thư viện này để ghi tất cả các sample vào file raw_data.csv

Sau khi crawl dữ liệu ta thu được một file raw_data.csv chứa (14942 sample) và 9 đặc trưng bao gồm:

- Address: Địa chỉ nhà
- Prices: Giá nhà
- Area: Diện tích
- toFace: Hướng của ngôi nhà
- type: Loại nhà (Mặt tiền/trong hẻm)
- certificate: Giấy tờ pháp lý của ngôi nhà (Sổ đỏ/ Không có)
- width: Chiều rộng của ngôi nhà
- length: Chiều dài của ngôi nhà
- floors: Số tầng của ngôi nhà

	Address	Prices	Area	toFace	type	certificate	width	length	floors
0	Phường Khuê Trung, Quận Cẩm Lệ, Đà Nẵng	35 tỷ	600 m2	_	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	33m	---	4
1	Đường Võ Nguyên Giáp, Phường Hòa Hải, Quận Ng...	28 tỷ	802 m2	_	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	---	---	3
2	Đường Nguyễn Giản Thanh, Phường An Khê, Quận ...	3,55 tỷ	70 m2	_	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	4m	17m	2
3	Đường Mai Đăng Chơn, Phường Hòa Quý, Quận Ngũ...	4 tỷ	100 m2	_	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	5m	20m	1
4	Đường Phan Văn Trị, Phường Khuê Trung, Quận C...	35 tỷ	600 m2	_	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	33m	18m	4
5	Đường Ỷ Lan Nguyên Phi, Phường Hòa Cường Bắc,...	7,75 tỷ	90 m2	_	Nhà mặt tiền	---	4,5m	20m	3
6	Đường Tùng Thiện Vương, Phường Khuê Mỹ, Quận ...	5,1 tỷ	95 m2	_	Nhà mặt tiền	---	---	---	---
7	Đường Trương Định, Phường Mân Thái, Quận Sơn ...	138 tỷ	3.000 m2	_	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	70m	---	1
8	Đường Đặng Vũ Hỷ, Phường Phước Mỹ, Quận Sơn T...	42 tỷ	580 m2	_	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	12m	---	1
9	Đường Yết Kiêu, Phường Thọ Quang, Quận Sơn Tr...	8,22 tỷ	137 m2	_	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	13m	---	3
10	Đường Nguyễn Phước Lan, Phường Hòa Xuân, Quận...	8,5 tỷ	130 m2	_	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	25m	---	3
11	Đường Khúc Hạo, Phường Nại Hiên Đông, Quận Sơ...	7,5 tỷ	90 m2	_	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	5m	18m	3
12	Đường Hà Mục, Phường Hòa Thọ Đông, Quận Cẩm L...	5,15 tỷ	100 m2	Tây Nam	Nhà mặt tiền	---	5m	20m	4
13	Đường An Thượng 29, Phường Mỹ An, Quận Ngũ Hà...	8,3 tỷ	71 m2	Bắc	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	5m	14m	3
14	Đường Nguyễn Quý Đức, Phường Khuê Trung, Quận...	35 tỷ	600 m2	_	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	33m	18m	4
15	Đường Nại Hiên Đông, Phường Nại Hiên Đông, Qu...	3,5 tỷ	65 m2	Tây Bắc	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	4,5m	15,3m	3
16	Đường Mai Lão Bạng, Phường Thuận Phước, Quậ...	7,3 tỷ	93 m2	_	Nhà mặt tiền	---	7,8m	12m	4
17	Đường Ngô Chi Lan, Phường Thuận Phước, Quận H...	3 tỷ	61 m2	_	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	3,5m	20m	---
18	Đường Yên Khê 2, Phường Thanh Khê Tây, Quận T...	3,65 tỷ	50 m2	_	Nhà mặt tiền	---	5m	---	3
19	Đường Lương Thế Vinh, Phường An Hải Đông, Quậ...	2,6 tỷ	50 m2	_	Nhà trong hẻm	Sổ hồng/ Sổ đỏ	4m	13m	3

Hình 1: 20 dữ liệu đầu tiên được crawl

2.2 Mô tả dữ liệu

Sau khi có dữ liệu thô, tiến hành làm sạch dữ liệu bằng cách tạo các cột dữ liệu mới, trích xuất giá trị hữu ích từ các dữ liệu có sẵn và ép kiểu dữ liệu sang kiểu dữ liệu thích hợp. Kết quả thu được tập dữ liệu sau khi làm sạch có:

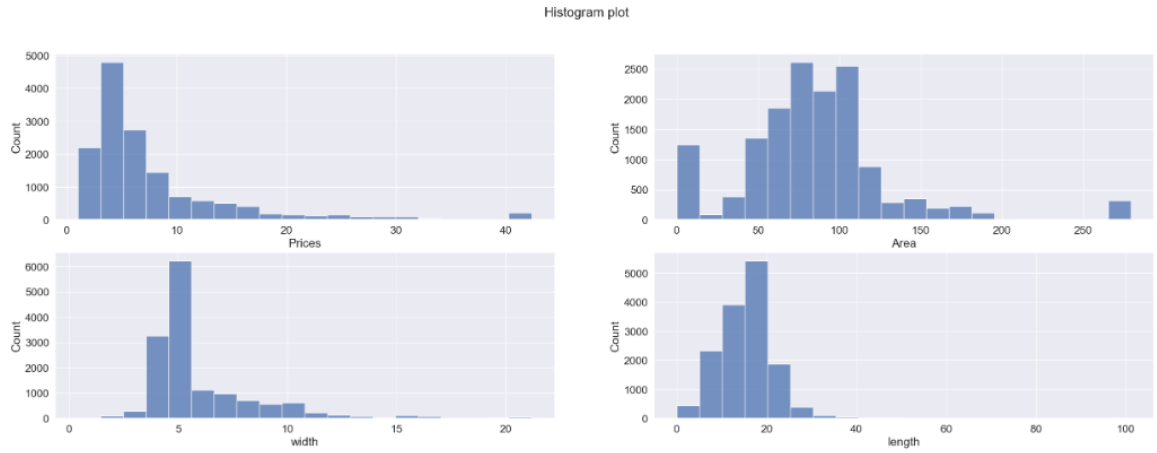
- Số lượng mẫu: 14611
- Số lựng đặc trưng: 17

Sau đây là cách làm sạch dữ liệu:

STT	Đặc trưng	Mô tả	Kiểu dữ liệu	Số mẫu dữ liệu có định dạng "---"
1	Street	Tên đường	String	0
2	Ward	Tên phường	String	0
3	District	Tên quận	String	0
4	Prices	Giá cả nhà (tỷ)	Float	65
5	Area	Diện tích (mét vuông)	Float	148
6	toFace	Hướng	String	0
7	type	Loại nhà	String	9
8	certificate	Có sổ đỏ hay không có	String	0
9	width	Chiều rộng nhà (mét)	Float	2003
10	length	Chiều dài nhà (mét)	Float	4033

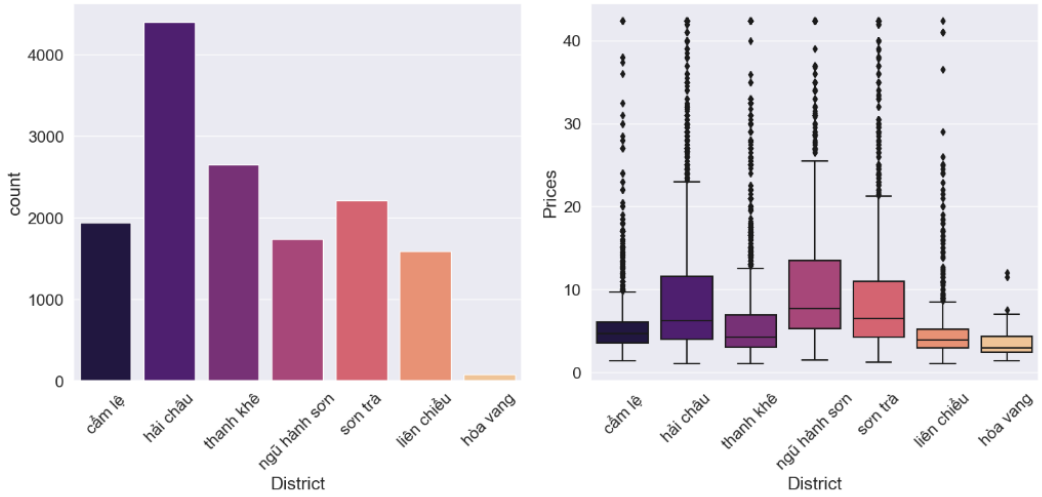
11	floors	Số tầng	int	343
12	Location	Vị trí trên bản đồ	String	0
13	Latitude	Vĩ độ	Float	0
14	Longitude	Kinh độ	Float	0
15	DistanceToCenter	Khoảng cách đến trung tâm	Float	0
16	DistanceToBeach	Khoảng cách đến biển	Float	0

Table 1: Bảng mô tả các làm sạch dữ liệu



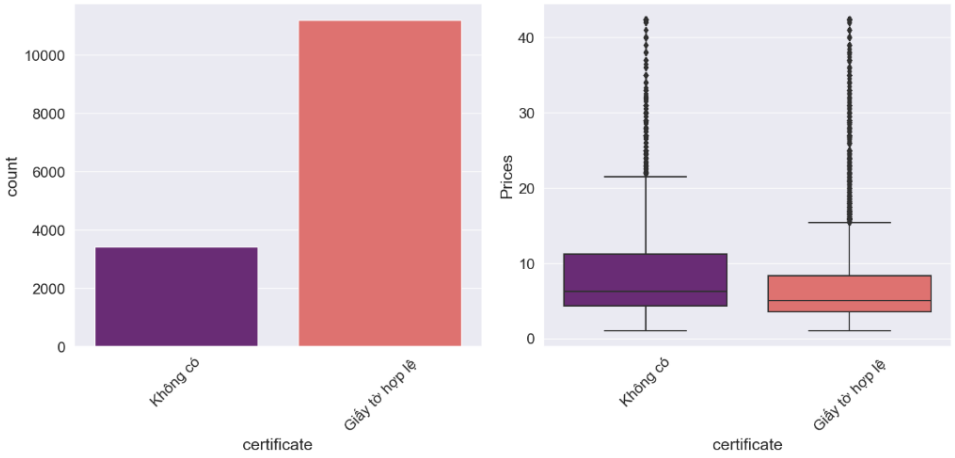
Hình 2: Biểu đồ phân bố của giá, diện tích, chiều dài, chiều rộng nhà

Nhận xét: Giá cả nhà ở Đà Nẵng với có giá < 50 tỷ, số lượng lớn giá nhà nằm ở mức từ 2 – 10 tỷ.



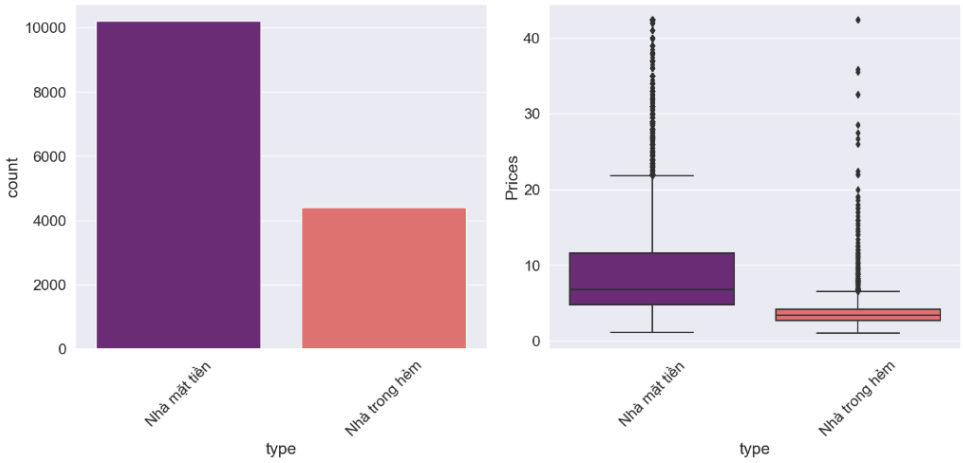
Hình 3: Biểu đồ phân bố của quận tại Đà nẵng và biểu đồ hộp tương quan giữa các quận với giá bán

Nhận xét: Số lượng nhà ở quận Hải Châu là nhiều nhất nhưng giá nhà ở quận Ngũ hành Sơn là cao nhất



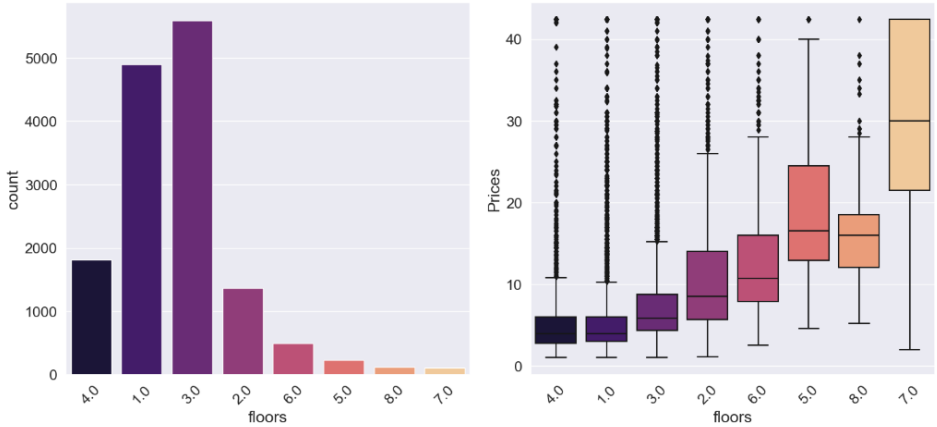
Hình 4: Biểu đồ phân bố của việc có/không có giấy chứng nhận và biểu đồ hộp tương quan với giá bán

Nhận xét: Nhà có giấy tờ hợp lệ chiếm số lượng nhiều hơn nhà không có giấy tờ và nhà có giấy tờ bán với giá cao hơn



Hình 5: Biểu đồ phân bố của loại nhà và biểu đồ hộp tương quan của loại hình nhà với giá bán

Nhận xét: Đà Nẵng có số lượng nhà mặt tiền nhiều hơn nhà trong hẻm,



Hình 6: Biểu đồ phân bố của số tầng nhà và biểu đồ hộp tương quan giữa số tầng nhà và giá

Nhận xét: nhà 1 tầng và nhà 2 tầng được bán nhiều nhất và nhà càng nhiều tầng thì giá càng cao

3. Trích xuất đặc trưng

3.1 Làm sạch dữ liệu

Dữ liệu sau khi thu thập về có thể lẫn vào những mẫu dữ liệu không đảm bảo chất lượng, hoặc trống dữ liệu, vì vậy việc làm sạch dữ liệu là điều cần thiết.

Tên đặc trưng	Cách xử lý	Định dang lại kiểu dữ liệu
Address	Không xử lý	String
Prices	- Xoá đơn vị "tỷ" - Thay dấu phẩy thành dấu chấm - Những cột nào định dạng sai không phải là số thì chuyển thành NaN - Chuyển lại kiểu dữ liệu từ thành float	Float
Area	- Xoá đơn vị "m2" - Những biến Area nào không phải định dạng là số thì chuyển vè NaN - Những giá trị area nào bằng 0 thì chuyển về NaN - Chuyển kiểu dữ liệu thành float	Float
toFace	- Chuyển '_' về là 'Không' - Đổi kiểu dữ liệu thành category	String
Type	- Đổi kiểu dữ liệu thành category	String
Certificate	- Thay đổi giá trị '---' thành 'Không có' - Thay đổi giá trị 'Giấy tờ hợp lệ' thành 'Sổ hồng/ Sổ đỏ' - Đổi kiểu dữ liệu thành category	String
Width	- Xoá đơn vị "m" - Loại bỏ các kí tự thừa ví dụ .x.y đổi thành x.y - Những giá trị không phải là số thì Chuyển thành NaN - Những giá trị nào bằng 0 thì chuyển thành NaN - Đổi kiểu dư liệu thành float	Float
length	- Xoá đơn vị "m" - Loại bỏ các kí tự thừa ví dụ x.y. đổi thành x.y - Những giá trị không phải là số thì Chuyển thành NaN - Những giá trị nào bằng 0 thì chuyển thành NaN - Đổi kiểu dư liệu thành float	Float
Floors	- Chuyển những giá trị '---' thành NaN - Đổi kiểu dữ liệu thành int	Int

Table 2: Bảng mô tả làm sạch dữ liệu

3.2 Tạo mới đặc trưng

Trong bài toán về dự đoán giá cả việc tạo mới đặc trưng là quan trọng bởi vì các đặc trưng thu thập có thể không đủ để mô tả đầy đủ các yếu tố ảnh hưởng đến giá cả. Thêm vào đó việc tạo mới đặc trưng có thể tăng độ chính xác của mô hình dự đoán

Tiến hành tách trường Address thành:

- Street (Đường)
- Ward (Phường-Xã)
- District (Quận - Huyện)
- City (Thành Phố)

Sau đó loại bỏ các trường "City", "Address", đồng thời loại bỏ những giá trị không phải là quận, huyện ở Đà Nẵng.
Thêm một trường Location theo Address của ngôi nhà bằng cách sử dụng API của BingMap để lấy latitude, Longitude theo địa chỉ của ngôi nhà. Tiếp theo ta tạo thêm 2 trường

- DistanceToCenter (Khoảng cách đến trung tâm thành phố)
- DistanceToBeach (Khoảng cách đến biển)

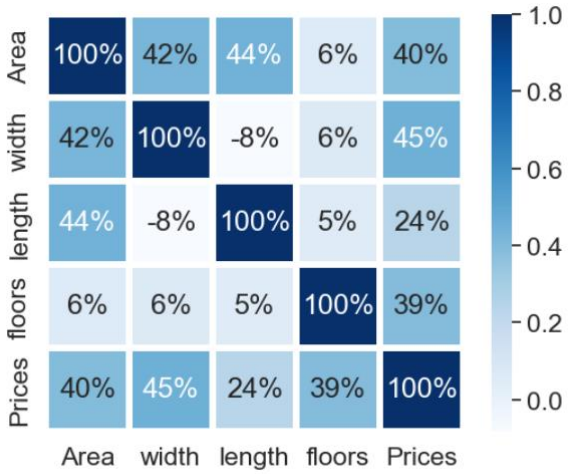
3.3 Xử lý dữ liệu trống

Đối với các dữ liệu trống của trường Prices, Area, floor ta thay thế các giá trị thiếu bằng mean. Tuy nhiên ta nhận thấy mối quan hệ giữa ba trường Area, width, length là Area = width * length. Vậy thì:

- Nếu giá trị width khác rỗng và length là rỗng thì: length = area/width
- Ngược lại thì: width = area/length
- Cả hai là rỗng thì length = width = sqrt(area)

3.4 Lựa chọn đặc trưng

Lựa chọn đặc trưng là một bước quan trọng trong quá trình xây dựng mô hình máy học, vì nó ảnh hưởng trực tiếp đến độ chính xác của mô hình. Nếu chọn những đặc trưng không phù hợp hoặc không đủ, mô hình sẽ không thể học được mối quan hệ giữa các đặc trưng và đầu ra, dẫn đến kết quả dự đoán không chính xác.



Hình 7: Mối tương quan giữa các đặc trưng

Từ những dữ liệu thống kê trực quan hóa từ sự tương quan giữa các đặc trưng với biến mục tiêu là giá cả, các đặc trưng ảnh hưởng đến giá cả là : Area, length, width, District, certificate, floors

3.5 Mã hóa nhãn

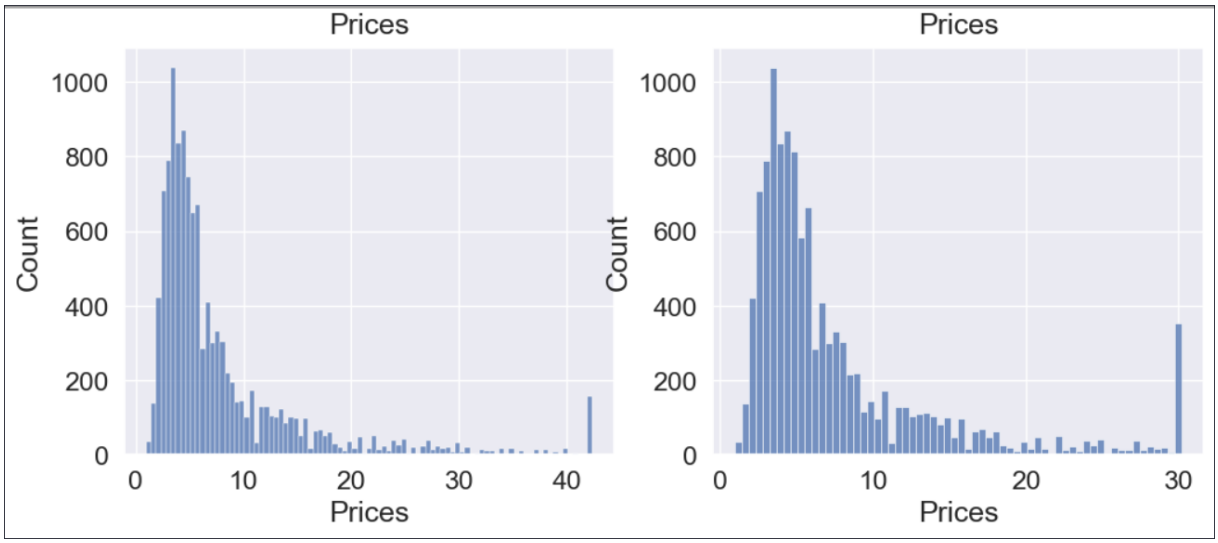
Mã hóa nhãn hay còn gọi là label encoding là một kỹ thuật chuyển đổi các giá trị của một biến thành các số nguyên, để dễ dàng cho việc xử lý trong các mô hình học máy. Label

encoding thường được sử dụng cho các biến định danh hoặc biến phân loại có giá trị không phải số.

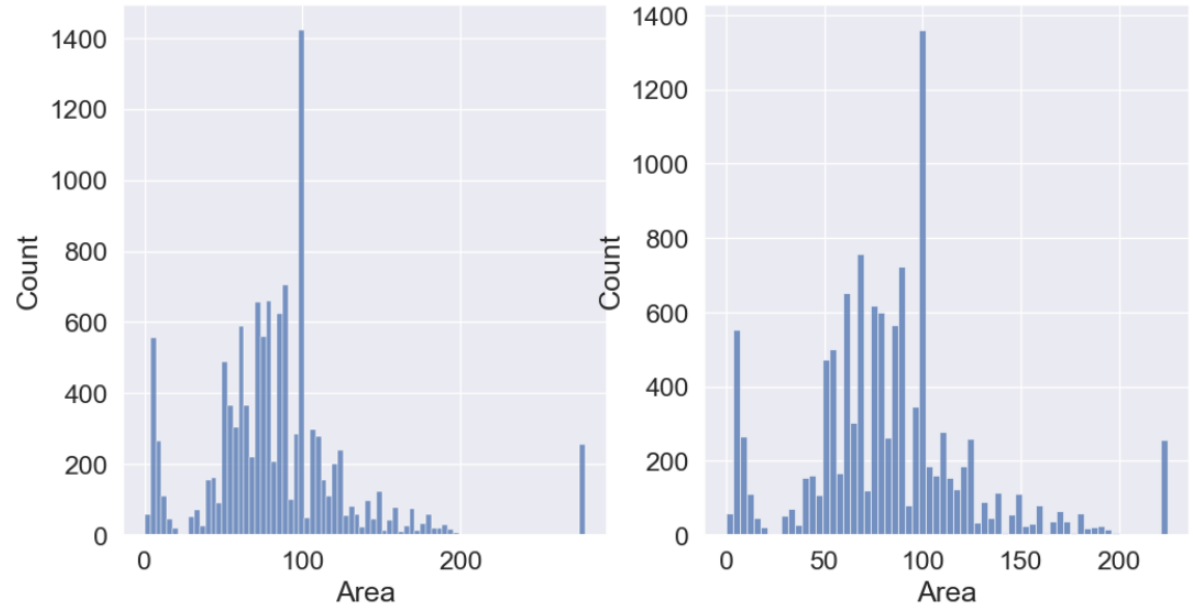
Trong các đặc trưng, các đặc trưng cần mã hóa là District, type, certificate
Chuyển dữ liệu đặc trưng "District" về dạnh số
Chuyển dữ liệu đặc trưng "type" với "Nhà mặt tiền" : 1, "Nhà trong hẻm" : 0
Chuyển dữ liệu đặc trưng certificae với "Giấy tờ hợp lệ": 1, "Không có": 0

3.6 Xử lý ngoại lệ

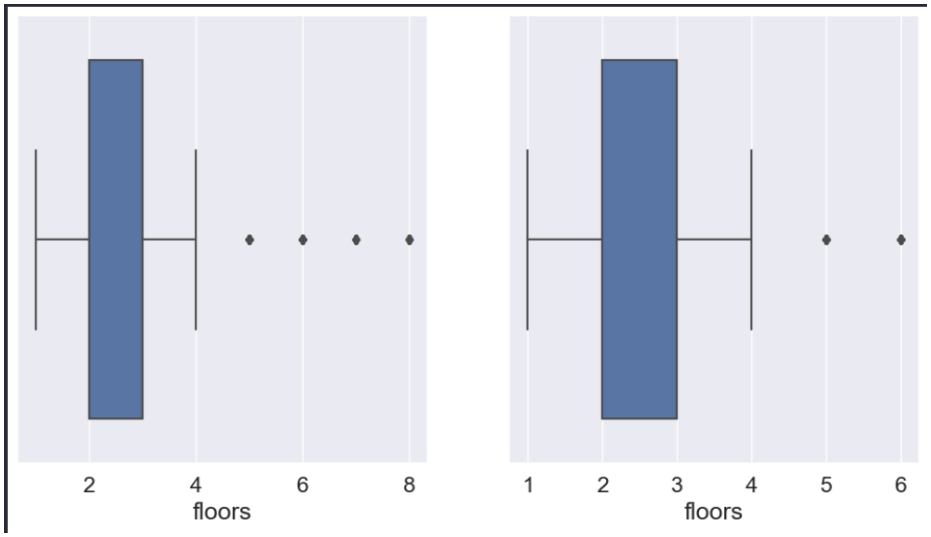
Áp dụng ngoại lệ lên các biến: Prices, Area, floors, width, length



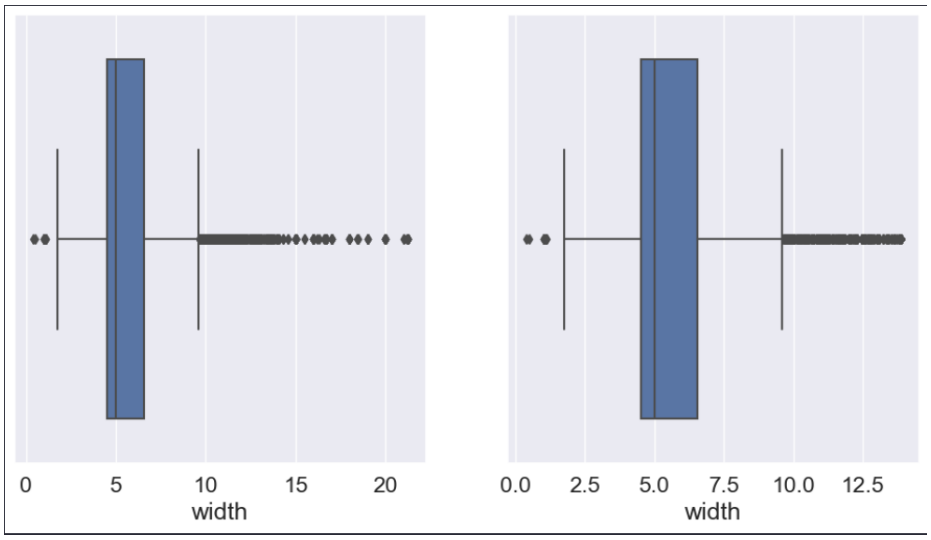
Hình 8: Áp dụng xử lý ngoại lệ lên đặc trưng giá



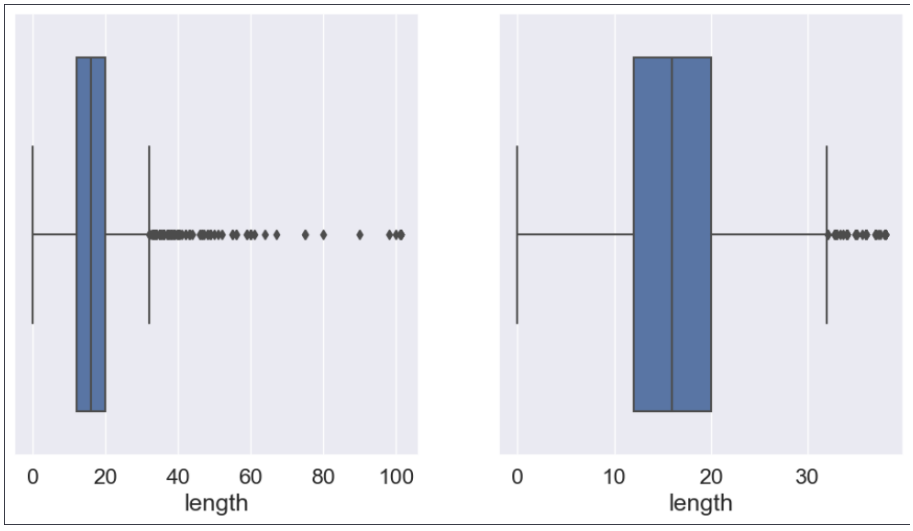
Hình 9: Áp dụng xử lý ngoại lệ lên đặc trưng diện tích



Hình 10: Áp dụng xử lý ngoại lệ lên đặc trưng số tầng



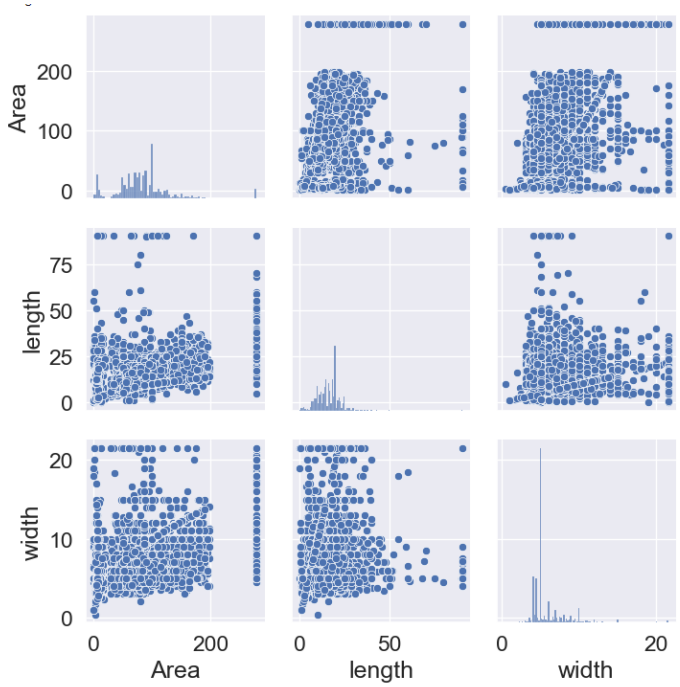
Hình 11: Áp dụng xử lý ngoại lệ lên đặc trưng chiều rộng



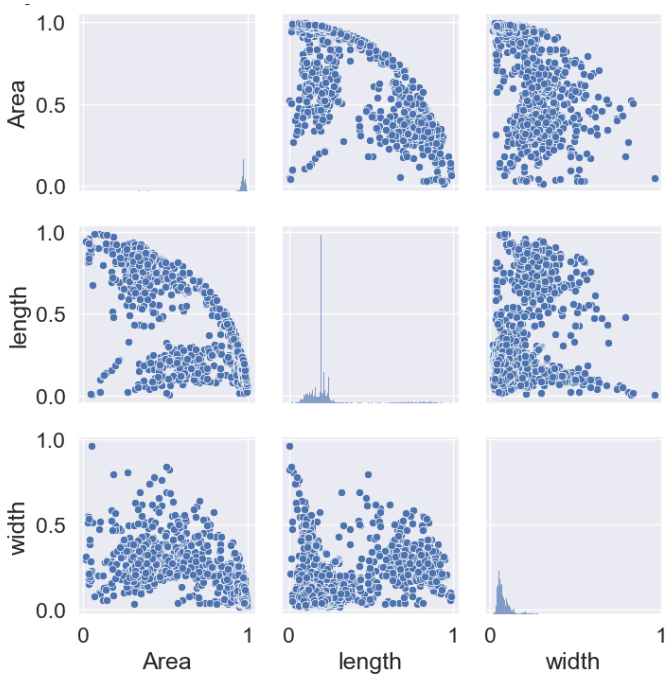
Hình 12: Áp dụng xử lý ngoại lệ lên đặc trưng chiều dài

3.7 Chuẩn hóa dữ liệu

Dữ liệu sử dụng chuẩn hóa Nomarlizer của sklearn preprocessing



Hình 13: Trực quan hóa sự tương quan của Area, length, width trước khi chuẩn hóa



Hình 14: Trực quan hóa sự tương quan của Area, length, width sau khi chuẩn hóa

4. Mô hình hóa dữ liệu

4.1 Mô hình sử dụng

4.1.1 Hồi quy tuyến tính (Linear Regression)

Hồi quy tuyến tính là một thuật toán học máy cơ bản thuộc loại học có giám sát. Đây là phương pháp thống kê để hồi quy dữ liệu với những biến phụ thuộc có giá trị liên tục dựa vào những biến độc lập (có thể liên tục hoặc không liên tục).

4.1.2 Random Forest

Random Forest là một thuật toán học máy dựa trên phương pháp Ensemble Learning, kết hợp nhiều cây quyết định để cải thiện độ chính xác và giảm thiểu overfitting.

4.1.3 Decision Tree Regression

Decision Tree Regression là một thuật toán học máy dựa trên cây quyết định, phân loại hoặc dự đoán dựa trên các quyết định đưa ra bằng cách phân tách dữ liệu thành các nhóm dựa trên các đặc tính của chúng.

4.2 Chia dữ liệu

Dữ liệu được chia với training set 80% và test set 20%

Phần dữ liệu training set tiến hành dùng các kĩ thuật mã hóa nhãn, chuẩn hóa dữ liệu rồi mới đưa vào mô hình huấn luyện.

4.3 Tham số huấn luyện

4.3.1 Linear Regression

Các bộ tham số chính:

- fit_intercept: Xác định xem mô hình có sử dụng giá trị y tại điểm cắt trục y hay không. Mặc định là True.
- Tham số positive quyết định liệu các hệ số trong mô hình Linear Regression có bị giới hạn là các giá trị dương hay không.. Mặc định, positive=False.
- copy_X: Xác định xem dữ liệu đầu vào có được sao chép hay không. Mặc định là True.
- n_jobs: Số lượng CPU được sử dụng khi tính toán. Mặc định là 1.

➢ *Sử dụng GridSearchCV thu được bộ tham số tốt nhất:*

- Fit_intercept : True
- Positive: False
- Copy_X ; True
- n_jobs: 1

4.3.2 Random Forest

Bộ tham số:

- n_estimators: Số lượng cây quyết định
- max_depth: Độ sâu tối đa của các cây quyết định
- min_samples_split: Số lượng mẫu tối thiểu yêu cầu để phân chia một nút.
- min_samples_leaf: Số lượng mẫu tối thiểu yêu cầu để tạo ra một lá
- max_features: Số lượng đặc trưng tối đa được xem xét để tìm kiếm phân chia tốt nhất (sqrt, log2)

➢ *Sử dụng GridSearchCV thu được bộ tham số tốt nhất:*

- max_depth **: None**
- min_samples_split **:** 2
- max_features **:** 'sqrt'
- min_samples_leaf **:** 1

4.3.3 Decision Tree Regression

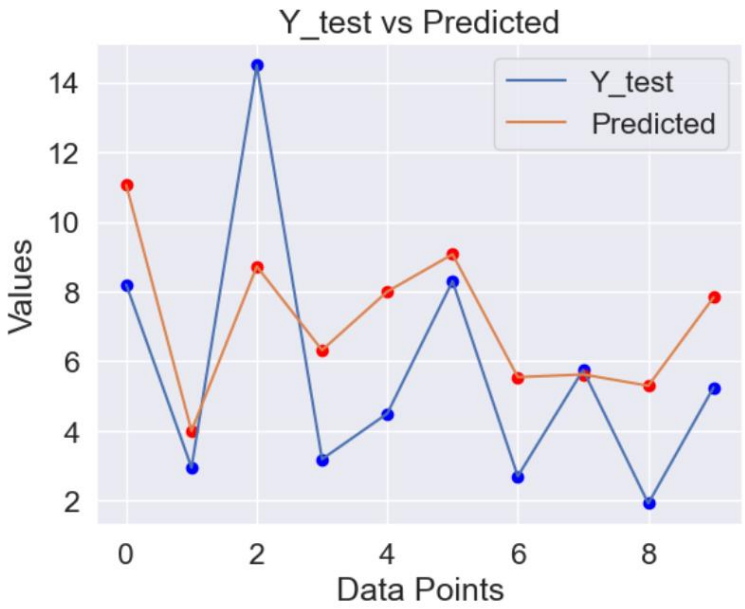
Bộ tham số:

- max_depth: Độ sâu tối đa của cây quyết định. Mặc định là None, điều này có nghĩa là các nút lá không bị giới hạn bởi độ sâu.
- min_samples_split: Số lượng mẫu tối thiểu để một nút có thể được phân chia thành hai nút con. Mặc định là 2.
- min_samples_leaf: Số lượng mẫu tối thiểu trong mỗi nút lá. Mặc định là 1

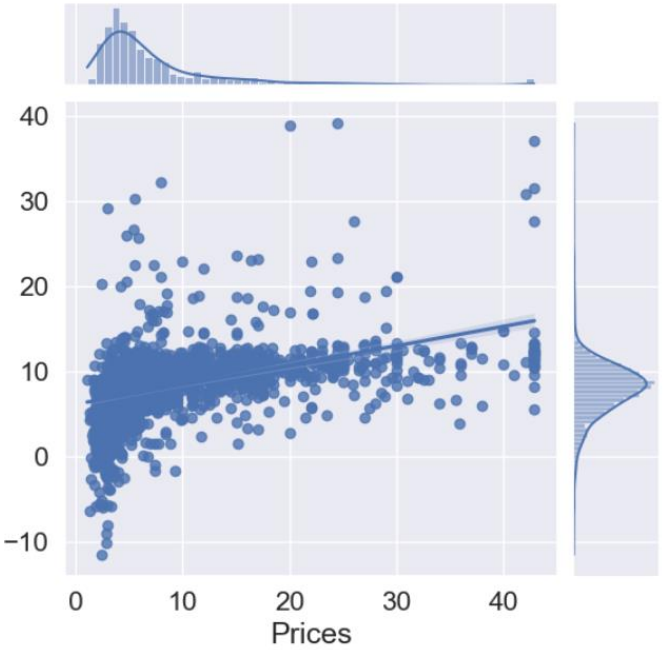
- *Sử dụng GridSearchCV thu được bộ tham số tốt nhất:*
  - max_depth: None
  - min_samples_leaf: 4
  - min_samples_split: 10

4.4 Kết quả

4.4.1 Hồi quy tuyến tính (Linear Regression)

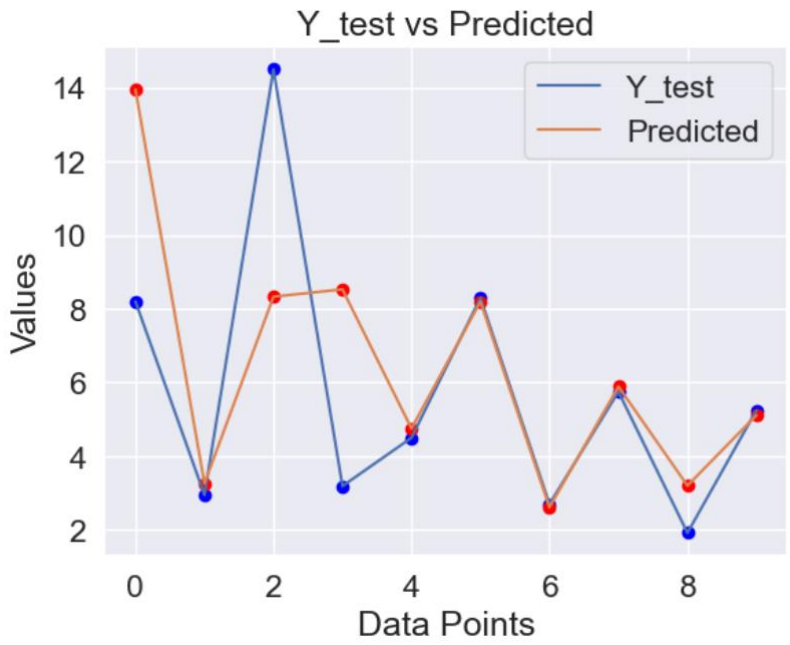


Hình 15: Đồ thị biểu mối liên hệ dữ liệu dự đoán và dữ liệu thực tế ở tập test với 10 điểm đầu tiên mô hình Linear Regression

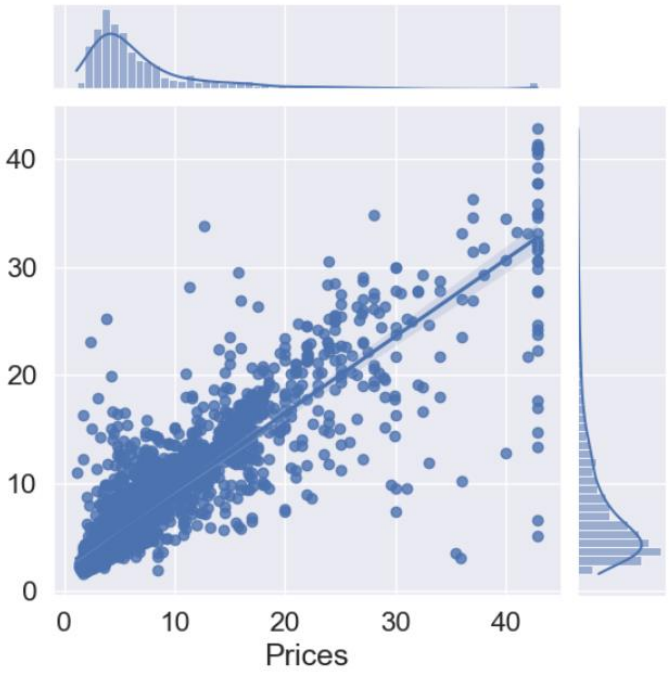


Hình 16: Tương quan giữa prices dự đoán theo mô hình linear regression và thực tế

4.4.2 Random Forest



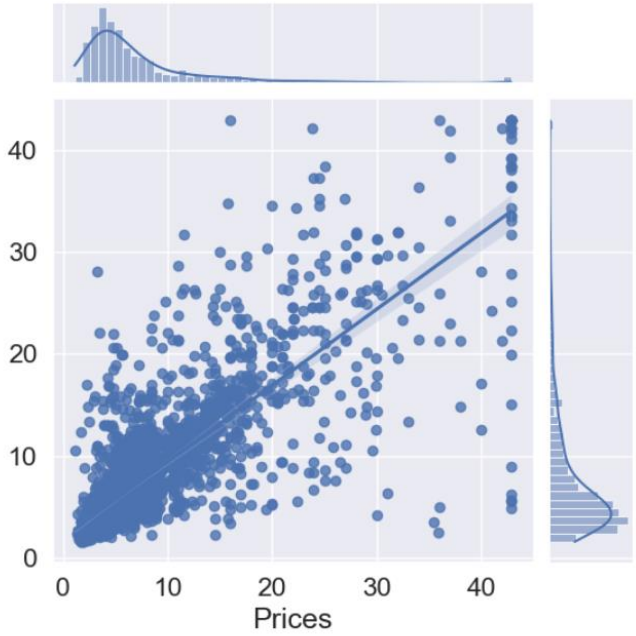
Hình 17: Đồ thị biểu mối liên hệ dữ liệu dự đoán và dữ liệu thực tế ở tập test với 10 điểm đầu tiên mô hình Random Forest



Hình 18: Tương quan giữa giá dự đoán theo mô hình Random Forest và giá thực tế

4.4.3 Decision Tree Regression

Hình 19: Đồ thị biểu mối liên hệ dữ liệu dự đoán và dữ liệu thực tế ở tập test với 10 điểm đầu tiên mô hình Decision Tree Regression



Hình 20: Tương quan giữa giá dự đoán theo mô hình Decision Tree Regression và giá thực tế

4.5 Metric đánh giá

4.5.1 Khái niệm và mô tả

- **R2 Score** (hay còn gọi là Coefficient of determination) là một số đo độ chính xác của một mô hình hồi quy tuyến tính. R2 Score thường được sử dụng để đánh giá khả năng giải thích của một mô hình và đo lường sự giống nhau giữa giá trị dự đoán và giá trị thực tế. Công thức như sau:

R2_Score = 1 - (SS_res / SS_tot)

Trong đó :

- SS_res là tổng bình phương sai số giữa giá trị dự đoán và giá trị thực tế
- SS_tot là tổng bình phương của hiệu giá trị thực tế và giá trị trung bình

- **RMSE** (Root Mean Squared Error) là một số đo lường độ lỗi của một mô hình dự đoán. Nó thường được sử dụng để đo lường khoảng cách giữa giá trị thực tế và giá trị dự đoán, và được tính bằng căn bậc hai của trung bình tổng bình phương sai số giữa giá trị thực tế và giá trị dự đoán.Công thức như sau:

```
RMSE = sqrt(mean((y_true - y_pred)^2))
```

Trong đó:

- y_true là danh sách các giá trị thực tế.
- y_pred là danh sách các giá trị dự đoán tương ứng.
- mean() là hàm tính trung bình.

- MAE (Mean Absolute Error) là một số đo lường độ lỗi của một mô hình dự đoán. Nó đo khoảng cách trung bình giữa giá trị thực tế và giá trị dự đoán, và được tính bằng trung bình của giá trị tuyệt đối của sai số giữa giá trị thực tế và giá trị dự đoán.

Công thức tính MAE như sau:

```
MAE = mean(abs(y_true - y_pred))
```

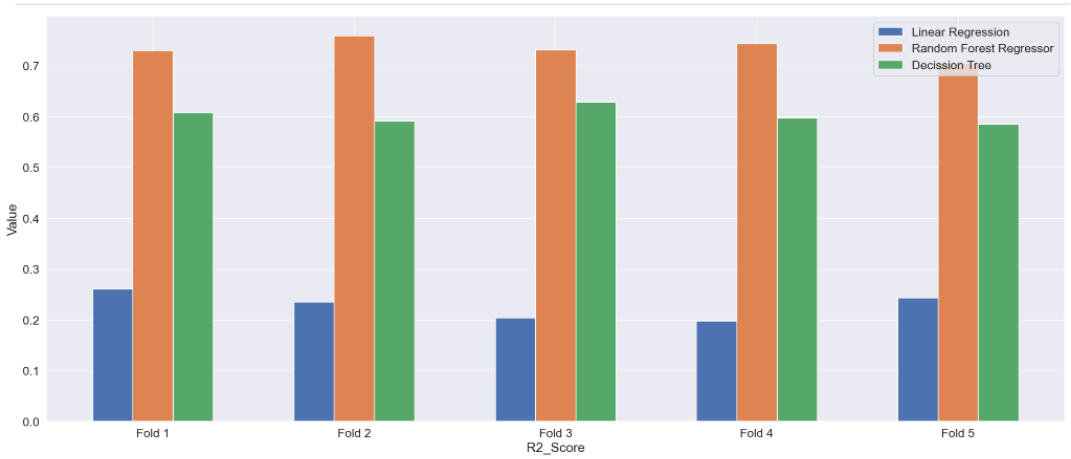
Trong đó:

- y_true là danh sách các giá trị thực tế.
- y_pred là danh sách các giá trị dự đoán tương ứng.
- mean() là hàm tính trung bình.
- abs() là hàm trị tuyệt đối.

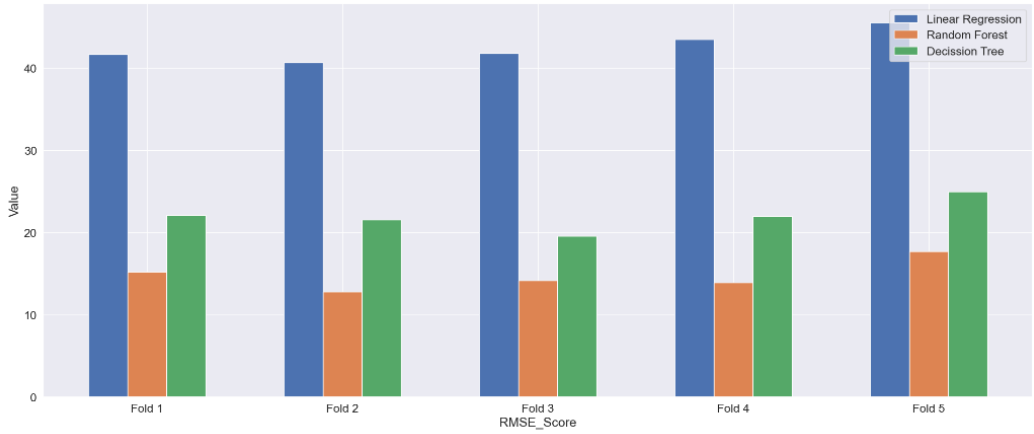
4.5.2 Metric các mô hình

	Linear Rrgression	Random Forest	Decision Tree Regression
R2 Score	0.212031	0.766034	0.655481
RMSE Score	6.663245	3.630845	4.405933
MAE	4.320418	1.732743	2.149602

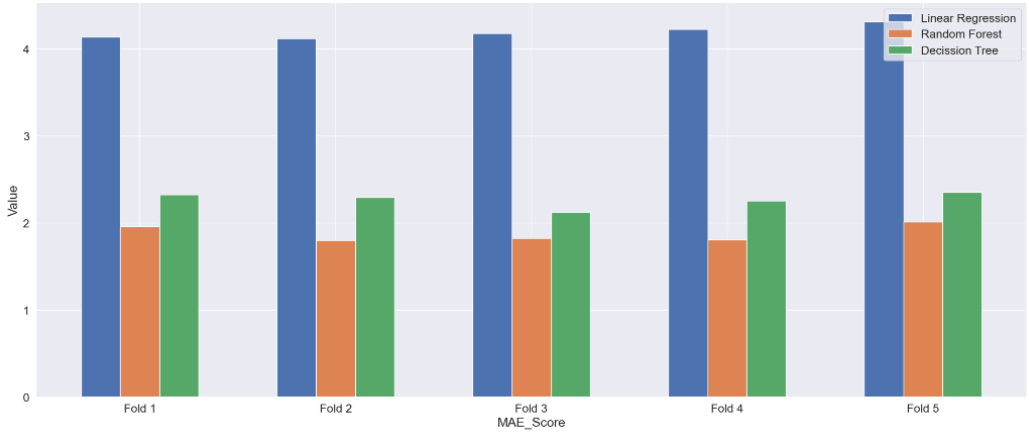
Table 3: Bảng so sánh metric đánh giá trên tập test của các mô hình



Hình 21: Đồ thị thể hiện R2 Score của 3 mô hình theo kỹ thuật cross validation

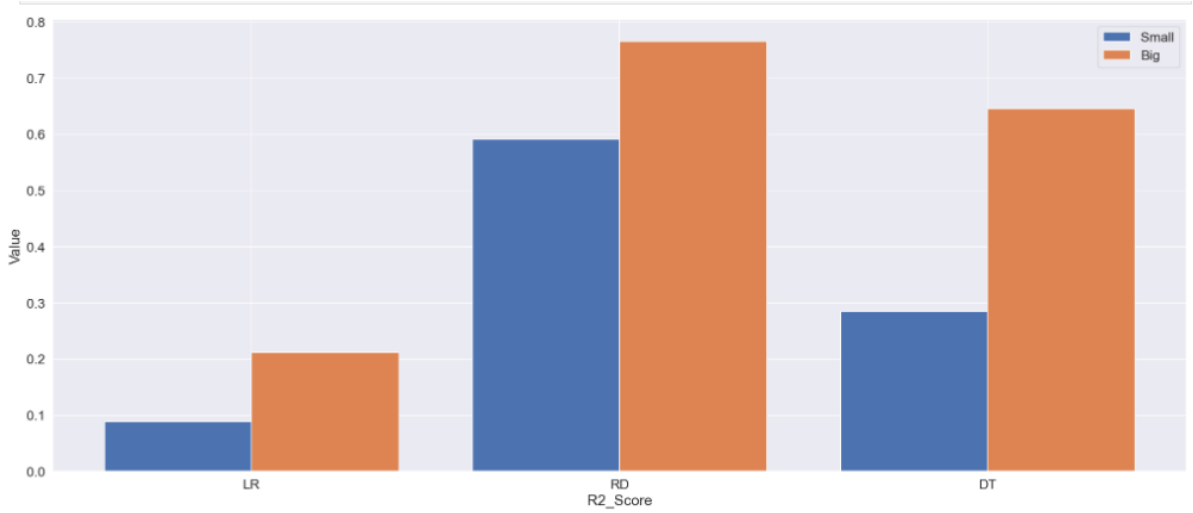


Hình 22: Đồ thị thể hiện RMSE Score của 3 mô hình theo kĩ thuật cross validation

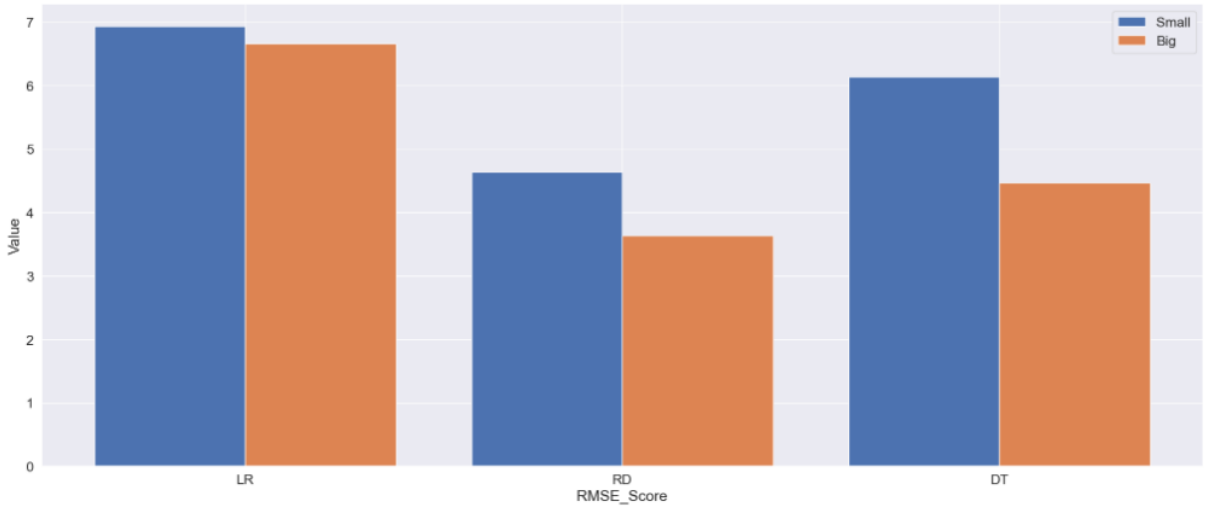


Hình 23: Đồ thị thể hiện MAE Score của 3 mô hình theo kĩ thuật cross validation

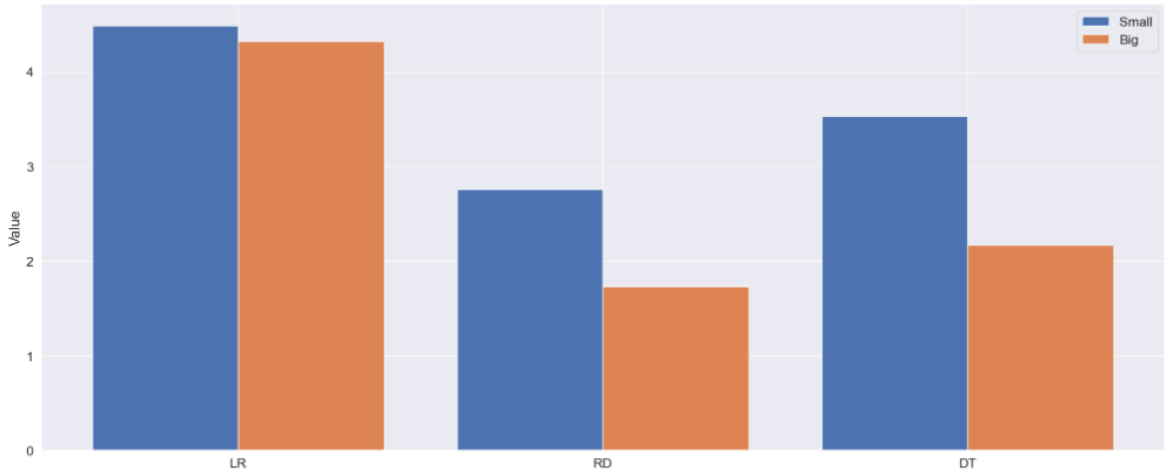
4.6 So sánh hiệu suất theo độ lớn dữ liệu



Hình 24: So sánh hiệu suất R2 Score các mô hình dựa trên độ phức tạp dữ liệu



Hình 25: So sánh hiệu suất RMSE Score các mô hình dựa trên độ phức tạp dữ liệu



Hình 26: So sánh hiệu suất MAE Score các mô hình dựa trên độ phức tạp dữ liệu

5. Kết luận

5.1 Hiệu suất mô hình

Mô hình Random Forest, Decision Tree Regression, Linear Regression có hiệu suất giảm dần, trong đó Random Forest có hiệu suất cao nhất là 76%.

Mô hình Random Forest Regression và Decission Tree Regression thích hợp để dự đoán giá nhà tại Đà Nẵng trong đó Random Forest là tối ưu nhất.

Mô hình Linear Regression là mô hình không thích hợp

Càng nhiều dự liệu huấn luyện thì độ chính xác sẽ càng cao Khi thay đổi tập dữ liệu thì mô hình Random Forest Regression vẫn cho kết quả tối ưu nhất

5.2 Giải thích, dự đoán, nguyên nhân

- Vì dữ liệu trong mô hình phi tuyến không thích hợp với mô hình linear nên dẫn đến hiệu suất của mô hình linear regression thấp

5.3 Hướng Phát triển

- Cần thu thập trên nhiều nguồn để dữ liệu đa dạng hơn

- Sử dụng thêm nhiều mô hình khác như Support Vector Regression (SVR), XGBoost

6. Tài liệu tham khảo

[1] Vũ Khắc Tiệp, Machine Learning cơ bản 2018

[2] https://dothanhlong.org/mo-hinh-hoi-quy-ung-dung-trong-bai-toan-du-doan-gia-bat-dong-san-machine-learning-phan-2/

[3] https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html

[4] https://www.analyticsvidhya.com/blog/2020/03/beginners-guide-random-forest-hyperparameter-tuning/