RESEARCH ARTICLE

ỨNG DỤNG PHÂN TÍCH VÀ TIỀN XỬ LÝ DỮ LIỆU TRONG PYTHON VÀO BÀI TOÁN DỰ ĐOÁN GIÁ NHÀ

Nguyễn Mai Khánh Vy, Trần Ngọc Thảo Ngân, Trần Minh Thái

Khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ -Tin học TP.HCM

21dh114581@st.huflit.edu.vn, 21dh114460@st.huflit.edu.vn, thaitm@huflit.edu.vn

TÓM TẮT— Bài toán dự đoán là một trong những bài toán quan trọng và có ứng dụng rộng rãi nhất trong lĩnh vực học máy. Nó đóng vai trò nền tảng cho nhiều ứng dụng quan trọng trong đời sống con người, từ những lĩnh vực quen thuộc như dự báo thời tiết, dự đoán giá cả đến những lĩnh vực phức tạp hơn như chẩn đoán bệnh, phát hiện gian lận, và lái xe tự động. Bài toán dự đoán tập trung vào việc dự đoán kết quả của một sự kiện hoặc một biến số trong tương lai dựa trên dữ liệu lịch sử bằng cách tự động học dữ liệu và xây dựng mô hình dự đoán. Bài báo này tập trung vào việc xây dựng mô hình dự đoán giá nhà ở khu vực thành phố Hồ Chí Minh. Thông qua áp dụng các kỹ thuật phân tích và tiền xử lý dữ liệu trên các thư viện của ngôn ngữ lập trình Python bao gồm làm sạch dữ liệu, xử lý giá trị thiếu, giá trị trùng, giá trị ngoại lai, mã hóa biến phân loại, chuẩn hóa dữ liệu, trích chọn đặc trưng, giảm chiều dữ liệu. Sau đó, huấn luyện các mô hình học máy để dự đoán giá nhà thông qua phương pháp Support Vector Regressor (SVR) và Random Forest Regressor (RFR). Kết quả thực nghiệm cho thấy RFR có khả năng nắm bắt các mối quan hệ phức tạp và phi tuyến tính, ít bị ảnh hưởng bởi các giá trị ngoại lai và nhiễu, và có hiệu suất cao hơn so với SVR.

Từ khóa — Dự đoán giá nhà, phân tích dữ liệu, tiền xử lý dữ liệu, Support Vector Regressor, Random Forest Regressor, Python.

I. GIỚI THIỆU

Trong thời đại công nghệ số hiện nay, việc ứng dụng các kỹ thuật phân tích và tiền xử lý dữ liệu đóng vai trò ngày càng quan trọng trong nhiều lĩnh vực, đặc biệt là trong các bài toán dự đoán và ra quyết định. Một trong những ứng dụng quan trọng là dự đoán giá nhà. Dự đoán giá nhà là một bài toán phổ biến và có ý nghĩa thực tiễn cao trong lĩnh vực bất động sản. Việc có thể dự đoán chính xác giá nhà không chỉ giúp người mua và người bán đưa ra quyết định sáng suốt, mà còn hỗ trợ các nhà đầu tư và các tổ chức tài chính trong việc đánh giá rủi ro và cơ hội đầu tư.

Trong bài báo này, chúng tôi sẽ trình bày quy trình phân tích và tiền xử lý dữ liệu bằng Python, bao gồm các bước như làm sạch dữ liệu, xử lý giá trị thiếu, xử lý giá trị trùng, xử lý dữ liệu ngoại lai, mã hóa biến phân loại, chuẩn hóa dữ liệu, trích chọn đặc trưng và giảm số chiều dữ liệu. Chúng tôi cũng sử dụng các kỹ thuật trực quan hóa dữ liệu để khám phá mối quan hệ giữa các biến và phát hiện các xu hướng tiềm ẩn trong dữ liệu. Cuối cùng, chúng tôi sẽ đánh giá hiệu quả của quá trình tiền xử lý dữ liệu bằng cách so sánh kết quả dự đoán giá nhà trước và sau khi áp dụng các kỹ thuật này. Kết quả của nghiên cứu này sẽ cung cấp cái nhìn sâu sắc về tầm quan trọng của việc phân tích và tiền xử lý dữ liệu trong bài toán dự đoán giá nhà, đồng thời đề xuất các hướng phát triển tiếp theo cho việc cải thiện độ chính xác của mô hình dự đoán.

Phần còn lại của bài báo được tổ chức với bố cục bao gồm: Mục II trình bày tóm tắt các nghiên cứu liên quan. Mục III mô tả bài toán và các khái niệm liên quan. Mục IV thể hiện mô hình đề xuất. Tiếp theo, mục V của bài báo trình bày kết quả đánh giá thực nghiệm. Cuối cùng, kết luận và các hướng nghiên cứu tiếp theo được thể hiện trong mục VI.

II. CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Có nhiều nghiên cứu đã tập trung vào việc ứng dụng phân tích và tiền xử lý dữ liệu trong các bài toán học máy, đặc biệt là trong dự đoán giá nhà. Những nghiên cứu này thường đề cập đến các phương pháp khác nhau để làm sạch dữ liệu, xử lý các giá trị thiếu, chuẩn hóa dữ liệu và biến đổi dữ liệu để phù hợp với các mô hình học máy.

Một nghiên cứu tiêu biểu của Kok, Monkkonen và Quigley [1] đã áp dụng phương pháp hồi quy tuyến tính để dự đoán giá nhà ở California. Nghiên cứu này nhấn mạnh tầm quan trọng của việc chọn lựa các biến đầu vào thích hợp như diện tích nhà, số phòng, vị trí địa lý và các tiện ích xung quanh. Việc làm sạch và chuẩn hóa dữ liệu cũng được coi là bước quan trọng giúp cải thiện độ chính xác của mô hình.

Zhang, Li và Sun [2] đã khám phá việc sử dụng mạng nơ-ron tích chập (Convolutional Neural Network - CNN) để dự đoán giá nhà ở Bắc Kinh. Nghiên cứu này đặc biệt chú trọng vào việc xử lý dữ liệu phi cấu trúc như hình ảnh của các căn nhà và khu vực xung quanh. Họ đã chỉ ra rằng việc kết hợp dữ liệu phi cấu trúc và cấu trúc sau tiền xử lý một cách cẩn thận có thể cải thiện đáng kể độ chính xác của dự đoán.

Một nghiên cứu khác là "An Optimal House Price Prediction Algorithm: XGBoost" [3]. Các tác giả đã thực nghiệm so sánh các thuật toán như SVR, Random Forest, XGBoost, multilayer perceptron, và multiple linear regression. Kết quả cho thấy XGBoost là mô hình có hiệu suất cao nhất trong dự đoán giá nhà trên tập dữ liệu thực nghiệm.

Phương pháp học máy cũng được các tác giả áp dụng để dự đoán giá nhà ở Kuala Lumpur [4]. Trong đó, hai thuật toán LightGBM và XGBoost được so sánh với multiple regression analysis và ridge regression. Kết quả cho thấy mô hình dự đoán giá nhà dựa trên XGBoost đạt hiệu suất cao nhất. Một cách tiếp cận khác của nhóm tác giả trong công trình [5] đề xuất dự đoán giá nhà thông qua các mô tả dạng văn bản với ba kỹ thuật TF-IDF, Word2Vec và BERT kết hợp với bốn mô hình hồi quy được huấn luyện dựa trên các dạng dữ liệu khác nhau (dữ liệu văn bản, dữ liệu không phải văn bản, hoặc cả hai). Kết quả thực nghiệm cho thấy việc sử dụng dữ liệu mô tả với Word2Vec và mô hình học sâu đạt hiệu suất tốt nhất. Hay trong nghiên cứu [6], nhóm tác giả cũng áp dụng các phương pháp cho việc dự đoán giá nhà ở Quận King, Washington bao gồm linear regression, random forest, neural network và XGBoost. Kết quả cho thấy XGBoost đạt được độ chính xác cao nhất. Ngoài ra, trong nghiên cứu [7], nhóm tác giả sử dụng các phương pháp học máy như Random Forest, Stacked Generalization Regression, Linear Regression và XGBoost để dự đoán giá nhà. Kết quả cho thấy hiệu suất vượt trội của các mô hình trong các giả thuyết khác nhau và cung cấp cái nhìn sâu sắc về hiệu quả của chúng để dự đoán giá nhà một cách chính xác nhất.

Nhìn chung, các nghiên cứu trên đã chứng minh rằng việc sử dụng các phương pháp phân tích và tiền xử lý dữ liệu hiện đại, kết hợp với các thuật toán học máy tiên tiến, có thể mang lại những kết quả đáng kể trong việc dự đoán giá nhà. Những phương pháp này không chỉ giúp cải thiện độ chính xác của mô hình mà còn mở ra nhiều hướng nghiên cứu mới trong việc tối ưu hóa các quy trình tiền xử lý và mô hình hóa dữ liệu.

III. MÔ TẢ BÀI TOÁN

Để giải quyết bài toán dự đoán giá nhà, việc sử dụng các công cụ phân tích và tiền xử lý dữ liệu mạnh mẽ là rất cần thiết. Python là ngôn ngữ lập trình lý tưởng cho mục tiêu này nhờ vào thư viện phong phú và khả năng hỗ trợ các thuật toán học máy hiệu quả. Dữ liệu về giá nhà thường bao gồm nhiều biến số như diện tích, số phòng ngủ, vị trí và nhiều yếu tố khác. Việc tiền xử lý dữ liệu bao gồm các bước như xử lý dữ liệu thiếu, dữ liệu trùng, mã hóa biến số phân loại, chuẩn hóa dữ liệu, và tách tập dữ liệu thành các tập huấn luyện và kiểm tra. Để thực hiện các bước này, chúng ta sẽ sử dụng các thư viện mạnh mẽ của Python như Pandas, NumPy và Scikit-learn. Pandas giúp xử lý và khám phá dữ liệu, xử lý các dữ liệu thiếu, lọc và thao tác với các DataFrame. NumPy hỗ trợ các thao tác số học cơ bản và nâng cao, làm việc với các mảng nhiều chiều. Scikit-learn là thư viện chính để xây dựng và đánh giá các mô hình học máy.

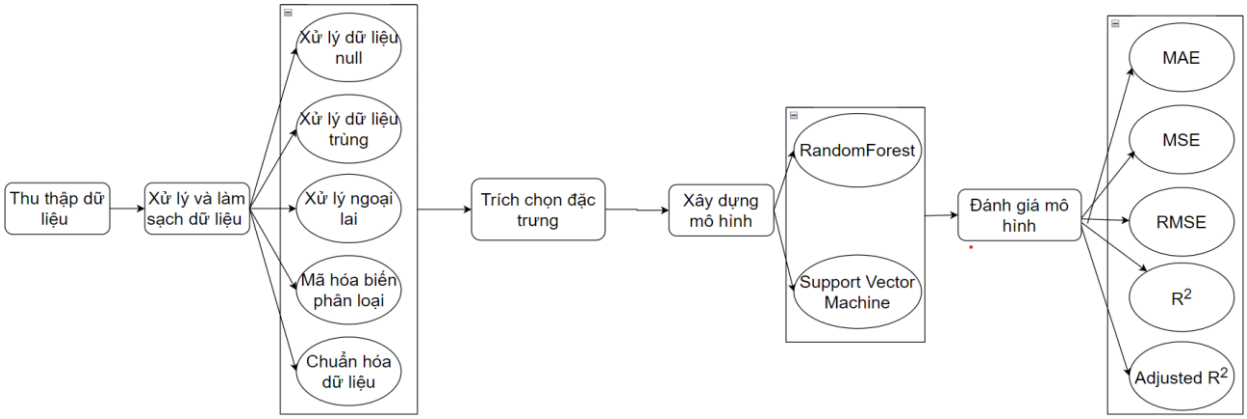
Với mục tiêu là có thể áp dụng phương pháp xây dựng mô hình dự đoán một cách đơn giản và hiệu quả. Chúng tôi xem xét sử dụng hai thuật toán phổ biến SVR [8] và RFR [9]. Thuật toán SVR là một thuật toán cơ bản có ưu điểm trong việc tìm ra ranh giới phân lớp tối ưu, phù hợp với các bài toán có không gian mẫu phức tạp. Đối với thuật toán RFR có khả năng nắm bắt các mối quan hệ phức tạp và phi tuyến tính giữa các đặc trưng, điều này rất phù hợp với dữ liệu bất động sản thường có nhiều yếu tố ảnh hưởng phức tạp đến giá. Bên cạnh đó, RFR ít bị ảnh hưởng bởi các giá trị ngoại lai và nhiễu trong dữ liệu, giúp mô hình ổn định hơn trên dữ liệu thực tế. Ngoài ra, RFR có khả năng đánh giá tầm quan trọng của các đặc trưng, giúp chúng ta hiểu rõ hơn về các yếu tố ảnh hưởng đến giá nhà. Điều đặc biệt là Random Forest có thể dễ dàng được song song hóa, giúp tăng tốc độ huấn luyện trên các tập dữ liệu lớn.

Thuật toán SVR là một thuật toán học máy có giám sát được sử dụng cho hồi quy. Mục tiêu chính của thuật toán SVR là tìm ra siêu phẳng tối ưu (hyperplane) trong không gian N-chiều để có thể tách biệt các điểm dữ liệu thuộc các lớp khác nhau trong không gian đặc trưng. Siêu phẳng này cố gắng tối đa hóa khoảng cách giữa các điểm gần nhất của các lớp khác nhau. Kích thước của siêu phẳng phụ thuộc vào số lượng đặc trưng đầu vào. Nếu có hai đặc trưng đầu vào, siêu phẳng chỉ là một đường thẳng. Nếu có ba đặc trưng đầu vào, siêu phẳng trở thành một mặt phẳng 2-chiều. Khi số đặc trưng đầu vào lớn hơn ba, việc tưởng tượng siêu phẳng này trở nên khó khăn.

Thuật toán RFR là một kỹ thuật mạnh mẽ trong học máy, sử dụng nhiều cây quyết định (decision trees) để đưa ra dự đoán. Trong giai đoạn huấn luyện, thuật toán tạo ra nhiều cây quyết định. Mỗi cây được huấn luyện trên một tập con ngẫu nhiên của tập dữ liệu. Khi xây dựng mỗi cây, thuật toán chọn ngẫu nhiên một tập con của các đặc trưng (features) trong dữ liệu. Điều này tạo ra sự khác biệt giữa các cây và giúp giảm nguy cơ quá khớp (overfitting). Khi dự đoán, thuật toán sẽ kết hợp kết quả từ tất cả các cây bằng cách tính trung bình kết quả từ các cây. Quá trình này giúp tạo ra các dự đoán ổn định và chính xác hơn vì nó tận dụng thông tin từ nhiều cây quyết định khác nhau. Rừng ngẫu nhiên rất hiệu quả trong việc xử lý dữ liệu phức tạp, giảm thiểu nguy cơ quá khớp, và cung cấp dự đoán đáng tin cậy cho nhiều ứng dụng khác nhau như phân loại và hồi quy.

IV. MÔ HÌNH ĐỀ XUẤT

Hình 1 thể hiện sơ đồ xử lý của quá trình dự đoán giá nhà. Quá trình xử lý gồm 5 giai đoạn: (i) Thu thập dữ liệu, (ii) phân tích và tiền xử lý dữ liệu, (iii) trích chọn đặc trưng chính, (iv) xây dựng mô hình và cuối cùng (v) đánh giá mô hình dự đoán.



Hình 1. *Sơ đồ quá trình dự đoán giá nhà.*

Bước 1: Thu thập dữ liệu

Tập dữ liệu House_price_predict được thu thập từ GitHub [10]. Tập dữ liệu mô tả giá của các căn chung cư ở khu vực TP. Hồ Chí Minh gồm 24,949 dòng với 16 cột được mô tả trong Bảng 1. Dữ liệu mẫu của 5 dòng đầu tiên được thể hiện trong Hình 2.

Bảng 1. *Mô tả tập dữ liệu*

Thứ tự	Thuộc tính	Ý nghĩa
0	DiaChi	Mô tả địa chỉ chung cư
1	TinhTrangBDS	Mô tả tình trạng của chung cư
2	DienTich	Mô tả diện tích của chung cư
3	Gia/m2	Mô tả giá của 1 mét vuông
4	Phongngu	Mô tả số lượng phòng ngủ
5	TenPhanKhu	Mô tả block của căn chung cư
6	SoTang	Mô tả chung cư nằm ở tầng bao nhiêu
7	PhongTam	Số lượng nhà vệ sinh
8	Loai	Mô tả loại hình chung cư
9	GiayTo	Tình trạng pháp lý của căn nhà
10	MaCanHo	Mã căn hộ
11	TinhTrangNoiThat	Mô tả tình trạng nội thất của căn chung cư
12	HuongCuaChinh	Hướng cửa chính của chung cư
13	HuongBanCong	Hướng ban công của chung cư
14	DacDiem	Căn trong góc hoặc căn chính giữa
15	Gia	Giá bán của căn chung cư

	DiaChi	TinhTrangBDS	DienTich	Gia/m2	Phongngu	TenPhanKhu	SoTang	PhongTam	Loai	GiayTo	MaCanHo	TinhTrangNoiThat	HuongCuaChinh	HuongBanCong	DacDiem	Gia
0	Đường Nguyễn Văn Quỳ, Phường Phú Thuận, Quận 7...	Đã bàn giao	62 m^2	32,26 triệu/m^2	2 phòng	NaN	NaN	2 phòng	Chung cư	Đã có sổ	NaN	NaN	NaN	NaN	NaN	2 tỷ- 62 m2đ
1	Đường Nguyễn Văn Linh, Phường Tân Thuận Tây, Q...	Đã bàn giao	95 m^2	55,79 triệu/m^2	3 phòng	NaN	NaN	2 phòng	Chung cư	Đang chờ sổ	NaN	Nội thất cao cấp	NaN	NaN	Căn góc	5,3 tỷ- 95 m2đ
2	Đường Võ Văn Kiệt, Phường An Lạc, Quận Bình Tâ...	Chưa bàn giao	75 m^2	34,4 triệu/m^2	2 phòng	2	5.0	2 phòng	Chung cư	Giấy tờ khác	17	NaN	Đông Nam	Đông Nam	NaN	2,58 tỷ- 75 m2đ
3	108, Đường Hồng Hà, Phường 2, Quận Tân Bình, T...	Đã bàn giao	70 m^2	57,14 triệu/m^2	1 phòng	A	7.0	1 phòng	Chung cư	Đang chờ sổ	BPA - 0712	Nội thất cao cấp	Đông Nam	Tây Bắc	NaN	4 tỷ- 70 m2đ
4	Đường Hậu Giang, Phường 11, Quận 6, Tp Hồ Chí ...	Đã bàn giao	83 m^2	35,54 triệu/m^2	2 phòng	NaN	NaN	2 phòng	Chung cư	Đã có sổ	NaN	Nội thất cao cấp	Tây Bắc	NaN	NaN	2,95 tỷ- 83 m2đ

Hình 2. *Dữ liệu mẫu của 5 dòng đầu tiên..*

Bước 2: Phân tích và làm sạch dữ liệu

Với dữ liệu ở Bước 1, chúng tôi tiến hành xử lý và làm sạch dữ liệu. Thông tin (*data.info()*) về số lượng hàng, cột, tên cột, số lượng giá trị không null, kiểu dữ liệu được thể hiện trong Bảng 2. Thống kê tóm tắt của các cột dữ liệu số (*data.describe()*) được thể hiện trong Bảng 3.

Bảng 2. *Thông tin về tập dữ liệu*

RangeIndex: 24949 entries, 0 to 24948			
Data columns (total 16 columns):			
	Column	Non-Null Count	Dtype
0	DiaChi	24624 non-null	object
1	TinhTrangBDS	24924 non-null	object
2	DienTich	24917 non-null	float64
3	Gia/m2	24916 non-null	float64
4	Phongngu	24926 non-null	float64
5	TenPhanKhu	7036 non-null	object
6	SoTang	6726 non-null	float64
7	PhongTam	24388 non-null	float64
8	Loai	24926 non-null	object
9	GiayTo	18852 non-null	object
10	MaCanHo	3358 non-null	object
11	TinhTrangNoiThat	12790 non-null	object
12	HuongCuaChinh	9370 non-null	object
13	HuongBanCong	8670 non-null	object
14	DacDiem	5601 non-null	object
15	Gia	24949 non-null	float64

Bảng 3. *Tóm tắt thống kê cho tập dữ liệu*

	DienTich	Gia/m2	Phongngu	SoTang	PhongTam	Gia
count	24917.00000	2.491600e+0	24926.00000	6726.000000	2488.000000	24949.00000
mean	193.907179	6.803146e+0	2.045134	11.482307	1.750656	2.865201
std	7465.948823	5.424467e+0	0.730980	18.142728	0.596897	7.827805
min	1.000000	0.000000e+0	1.000000	1.000000	1.000000	0.000000
25%	56.000000	1.412000e+0	2.000000	5.000000	1.000000	1.600000
50%	68.000000	2.973500e+0	2.000000	9.000000	2.000000	2.250000
75%	80.000000	4.151000e+0	2.000000	16.000000	2.000000	3.300000
max	780000.000000	5.428571e+1	10.000000	789.000000	6.000000	980.000000

Do các cột 'DienTich', 'Gia/m2', 'Phongngu' và 'PhongTam' thuộc dạng dòng ký tự (string) bao gồm cả số và chữ nên cần chuyển chúng sang dạng số (float). Tiếp theo, cần xử lý trích lọc ra tên Quận và Huyện ở cột 'DiaChi' vì không cần địa chỉ chi tiết của các căn hộ. Tương tự, cột 'Gia' vừa chứa số và chữ nên cũng cần Convert lại cột 'Gia' theo đơn vị tỷ đồng. Kết quả biến đổi dữ liệu được thể hiện trong Hình 3.

	DiaChi	TinhTrangBDS	DienTich	Gia/m2	Phongngu	TenPhanKhu	SoTang	PhongTam	Loai	GiayTo	MaCanHo	TinhTrangNoiThat	HuongCuaChinh	HuongBanCong	DacDiem	Gia
0	Quận 7	Đã bàn giao	62.0	3226.0	2.0	NaN	NaN	2.0	Chung cư	Đã có sổ	NaN	NaN	NaN	NaN	NaN	2.00
1	Quận 7	Đã bàn giao	95.0	5579.0	3.0	NaN	NaN	2.0	Chung cư	Đang chờ sổ	NaN	Nội thất cao cấp	NaN	NaN	Căn góc	5.30
2	Quận Bình Tân	Chưa bàn giao	75.0	344.0	2.0	2	5.0	2.0	Chung cư	Giấy tờ khác	17	NaN	Đông Nam	Đông Nam	NaN	2.58
3	Quận Tân Bình	Đã bàn giao	70.0	5714.0	1.0	A	7.0	1.0	Chung cư	Đang chờ sổ	BPA - 0712	Nội thất cao cấp	Đông Nam	Tây Bắc	NaN	4.00
4	Quận 6	Đã bàn giao	83.0	3554.0	2.0	NaN	NaN	2.0	Chung cư	Đã có sổ	NaN	Nội thất cao cấp	Tây Bắc	NaN	NaN	2.95

Hình 3. *Dữ liệu liệu mẫu sau khi biến đổi 5 dòng đầu tiên..*

Thông qua phân tích, tập dữ liệu có tồn tại nhiều giá trị null (Bảng 4). Do vậy, cần xử lý những giá trị null này.

Đầu tiên, các cột TenPhanKhu, SoTang, MaCanHo, HuongCuaChinh, HuongBanCong, DacDiem cần phải xóa vì các cột này có các giá trị missing rất lớn (hơn 50%). Tiếp theo, sử dụng 'dropna' để loại bỏ các giá trị null cho các cột DiaChi, TinhTrangBDS, DienTich, Gia/m2, Phongngu, PhongTam, Loai vì các cột này có số lượng các giá trị bị thiếu nhỏ hơn 5%. Đối với cột 'GiayTo' và 'TinhTrangNoiThat' có phần trăm giá trị null lần lượt là 24.43% và 48.73% thì ta tiến hành xử lý bằng cách thay thế giá trị null bằng giá trị mode vì kiểu dữ liệu của 2 cột này là object. Bảng 5 thể hiện kết quả sau khi xử lý những giá trị null.

Bảng 4. *Tỷ lệ phần trăm các giá trị null.*

DiaChi	1.302657
TinhTrangBDS	0.100204
DienTich	0.128262
Gia/m2	0.132270
Phongngu	0.092188
TenPhanKhu	71.798469
SoTang	73.041004
PhongTam	2.248587
Loai	0.092188
GiayTo	24.437853
MaCanHo	86.540543
TinhTrangNoiThat	48.735420
HuongCuaChinh	62.443385
HuongBanCong	65.249108
DacDiem	77.550202
Gia	0.000000
dtype:	float64

Bảng 5. *Kết quả sau khi xử lý các giá trị null.*

DiaChi	0
TinhTrangBDS	0
DienTich	0
Gia/m2	0
Phongngu	0
PhongTam	0
Loai	0
GiayTo	0
TinhTrangNoiThat	0
Gia	0
dtype:	int64

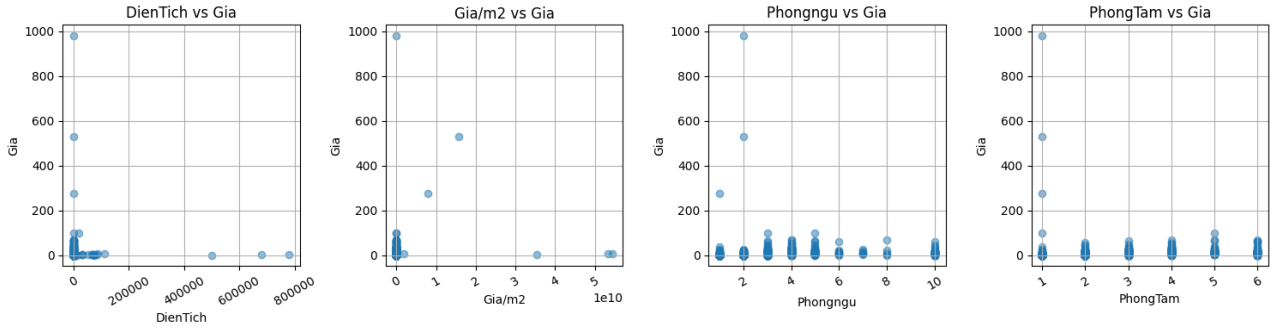
Ngoài ra, dữ liệu trùng cũng cần được loại bỏ. Hình 5 trình bày thông tin nhưng dữ liệu bị trùng nhau (7,358 dòng bị trùng)

```
Number of duplicate rows: 7358
```

	DiaChi	TinhTrangBDS	DienTich	Gia/m2	Phongngu	PhongTam	Loai	GiayTo	TinhTrangNoiThat	Gia
25	Quận 6	Đã bàn giao	83.0	3554.0	2.0	2.0	Chung cư	Đã có sổ	Nội thất cao cấp	2.95
29	Quận 7	Đã bàn giao	95.0	5579.0	3.0	2.0	Chung cư	Đang chờ sổ	Nội thất cao cấp	5.30
51	Quận 6	Đã bàn giao	83.0	3554.0	2.0	2.0	Chung cư	Đã có sổ	Nội thất cao cấp	2.95
53	Quận 7	Đã bàn giao	62.0	3226.0	2.0	2.0	Chung cư	Đã có sổ	Hoàn thiện cơ bản	2.00
75	Quận 8	Đã bàn giao	65.0	2769.0	2.0	1.0	Chung cư	Đã có sổ	Hoàn thiện cơ bản	1.80

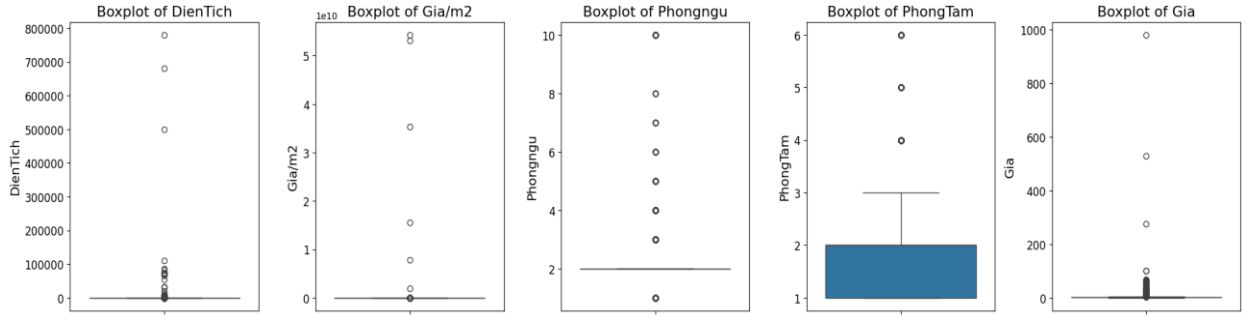
Hình 4. *Một phần các giá trị trùng trong tập dữ liệu.*

Sau đó, chúng tôi phân tích mối tương quan giữa các biến dạng số với giá trị "Gia" thông qua biểu đồ Scatter (Hình 5).

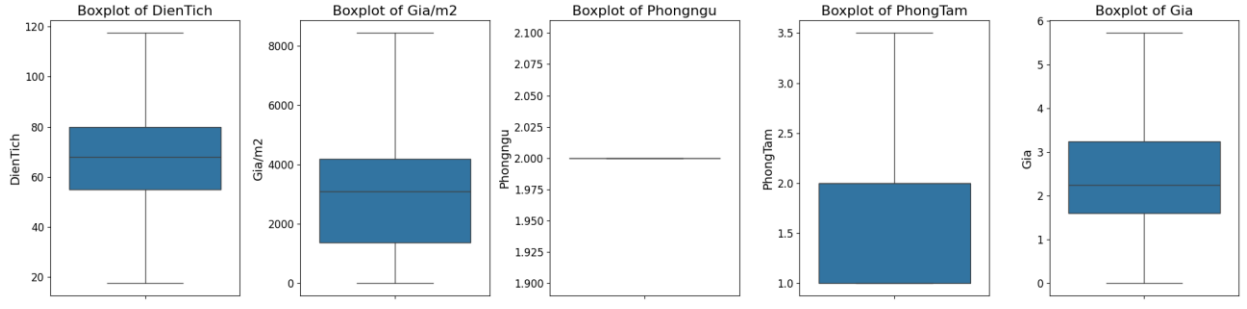


Hình 5. *Biểu đồ Scatter trực quan mối quan hệ giữa các biến dạng số với giá trị "Gia".*

Theo Hình 5, các điểm dữ liệu tập trung nhiều ở các giá trị thấp của DienTich và Gia, cho thấy hầu hết các bất động sản có diện tích và giá trị thấp. Tuy nhiên, có một vài điểm ngoại lệ với giá trị rất cao. Hầu hết các điểm dữ liệu tập trung ở giá trị thấp của Gia/m2 và Gia, cho thấy các bất động sản có giá cao trên mỗi mét vuông ít phổ biến hơn. Các điểm dữ liệu tập trung đa số ở các bất động sản có 2-4 phòng ngủ, và giá cả tương đối thấp. Trên đây cũng xuất hiện một số điểm ngoại lệ với giá cao hơn. Hầu hết các điểm dữ liệu tập trung ở các bất động sản có 1-3 phòng tắm, và giá cả tương đối thấp. Trong mối biểu đồ này cũng có một vài điểm ngoại lệ với nhiều phòng tắm hơn và giá cao hơn.

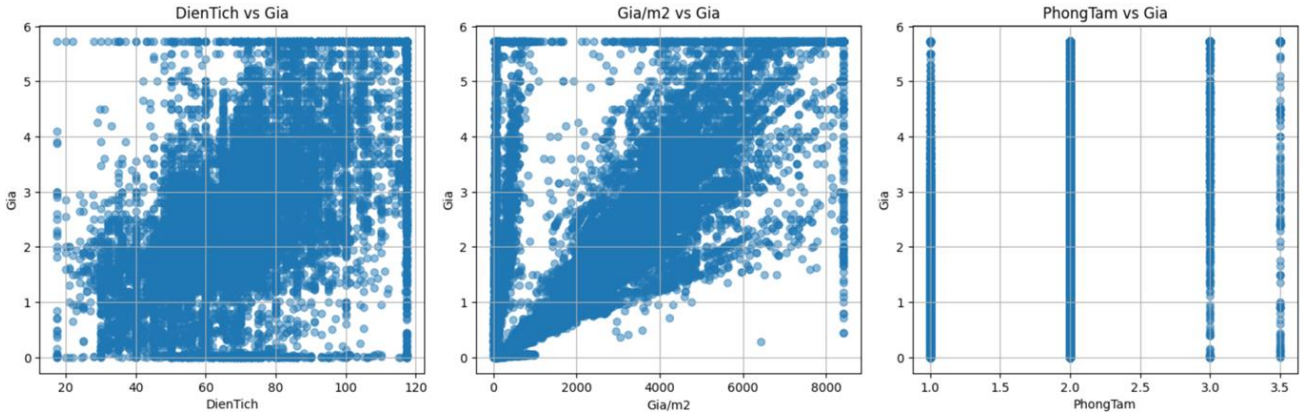


Hình 6. *Biểu đồ hộp trước khi xử lý ngoại lệ.*



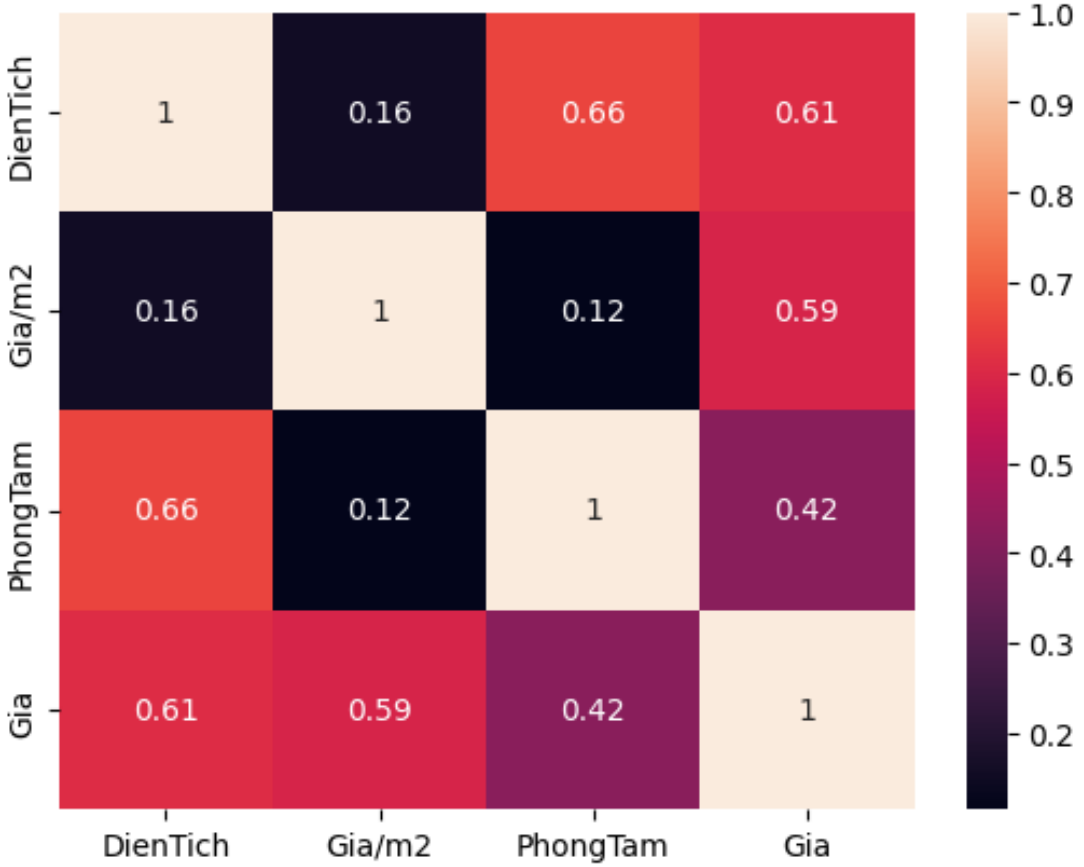
Hình 7. *Biểu đồ hộp sau khi xử lý ngoại lệ.*

Do đó, chúng tôi sử dụng biểu đồ boxplot (Hình 6) cho các cột 'DienTich', 'Gia/m2', 'Phongngu', 'PhongTam', 'Gia' để xác định outliers và tiến hành xử lý trên những ngoại lệ này. Biểu đồ trong Hình 6 cho thấy sự hiện diện của nhiều giá trị ngoại lệ, đặc biệt là ở các biến DienTich, Gia/m2 và Gia. Phần hộp chính của tất cả các biến số đều rất nhỏ, chứng tỏ sự tập trung cao độ của dữ liệu ở phạm vi thấp của các biến số. Xử lý các giá trị ngoại lệ bằng phương pháp IQR. Thay thế các giá trị lớn hơn upper_limit bằng upper_limit và các giá trị nhỏ hơn lower_limit bằng lower_limit. Kết quả xử lý ngoại lệ được thể hiện trong Hình 7. Sau khi xử lý outlier, biến "Phongngu" phân phối đồng nhất cho thấy sự thiếu đa dạng về số phòng ngủ trong tập dữ liệu, điều này có thể ảnh hưởng đến khả năng phân tích và dự báo vì vậy biến này cần loại bỏ. Hình 8 và Hình 9 thể hiện biểu đồ Scatter và biểu đồ Heatmap trực quan mối quan hệ giữa các biến dạng số với giá trị "Gia".



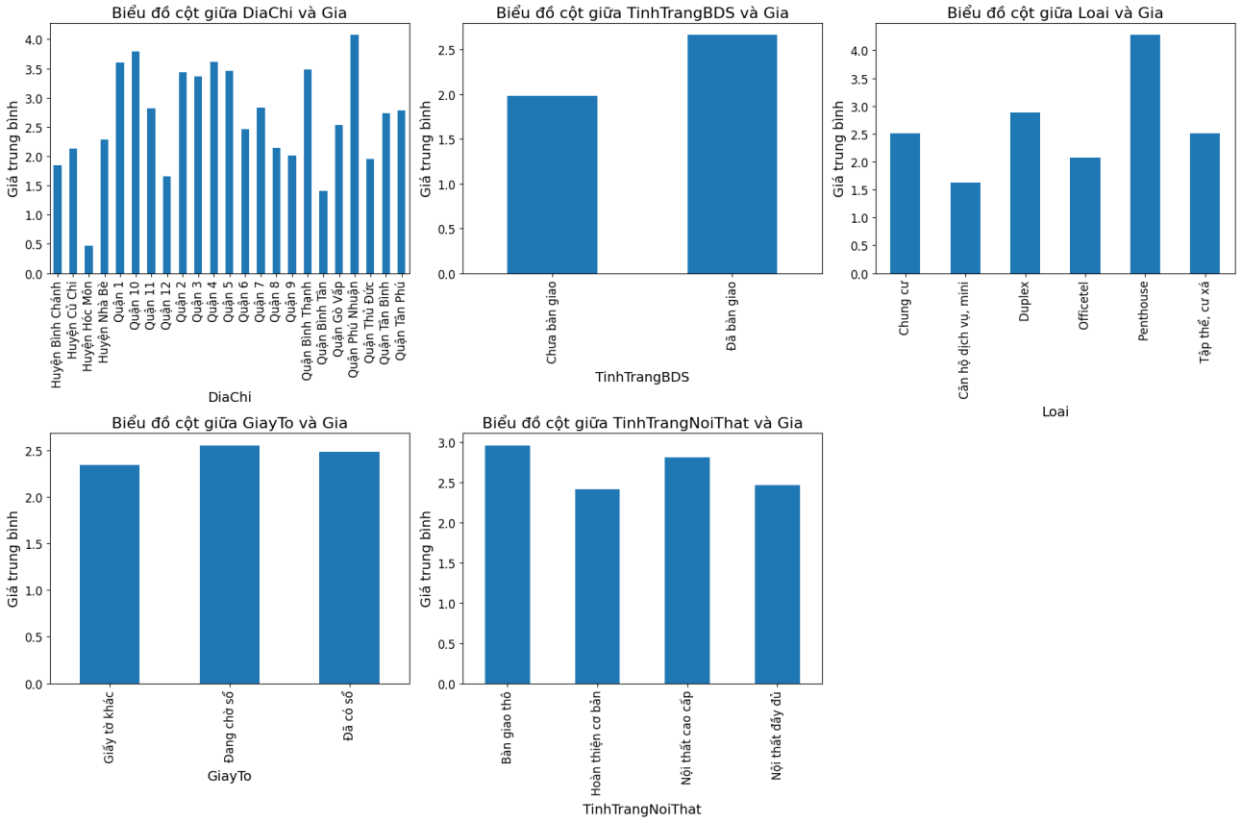
Hình 8. *Biểu đồ Scatter trực quan mối quan hệ giữa các biến số với cột giá "Gia".*

Hình 8 cho thấy sự phân tán dữ liệu rộng, đặc biệt ở các biến DienTich và Gia/m2, điều này cho thấy rằng các yếu tố này có thể có ảnh hưởng đáng kể đến giá. Hình 9 cho thấy các biến DienTich, Gia/m2 và PhongTam đều có ảnh hưởng nhất định đến Gia.



Hình 9. *Biểu đồ heatmap thể hiện sự tương quan giữa các biến.*

Hình 10 trực quan hóa mối quan hệ giữa các cột dữ liệu phân loại và giá trị trung bình của giá nhà để xác định mức độ ảnh hưởng của các biến phân loại này. Biểu đồ cột giữa DiaChi và Gia thể hiện các quận trung tâm thường có giá trung bình cao hơn các huyện ngoại thành. Đối với Biểu đồ cột giữa TinhTrangBDS và Gia thể hiện tình trạng của bất động sản cũng ảnh hưởng đến giá. Loại hình bất động sản cũng ảnh hưởng lớn đến giá (iểu đồ cột giữa Loai và Gia), với penthouse và duplex có giá cao nhất. Riêng Biểu đồ cột giữa GiayTo và Gia, loại giấy tờ không ảnh hưởng nhiều đến giá trung bình. Cuối cùng, là Biểu đồ cột giữa TinhTrangNoiThat và Gia thể hiện tình trạng nội thất có ảnh hưởng đáng kể đến giá trị của bất động sản. Các căn hộ "Bàn giao thô" và "Nội thất cao cấp" có giá trung bình cao hơn so với các căn hộ "Hoàn thiện cơ bản" và "Nội thất đầy đủ".



Hình 10. *Biểu đồ cột giữa các dữ liệu phân loại và Gia.*

Chính vì vậy, chúng tôi cần mã hoá các biến phân loại này để có thể áp dụng vào mô hình dự đoán. Trong đó, biến phân loại 'DiaChi', 'TinhTrangBDS', 'Loai', 'GiayTo', 'TinhTrangNoiThat' được mã hóa thông qua phương pháp One-hot Encoder (Bảng 6).

Bảng 6. *Kết quả sau khi mã hóa*

RangeIndex: 16728 entries, 0 to 16727			
Data columns (total 41 columns):			
	Column	Non-Null Count	Dtype
0	DienTich	16728 non-null	float64
1	Gia/m2	16728 non-null	float64
2	PhongTam	16728 non-null	float64
3	DiaChi_Huyện Bình Chánh	16728 non-null	float64
4	DiaChi_Huyện Củ Chi	16728 non-null	float64
5	DiaChi_Huyện Hóc Môn	16728 non-null	float64
6	DiaChi_Huyện Nhà Bè	16728 non-null	float64
7	DiaChi_Quận 1	16728 non-null	float64
8	DiaChi_Quận 10	16728 non-null	float64
9	DiaChi_Quận 11	16728 non-null	float64
10	DiaChi_Quận 12	16728 non-null	float64
11	DiaChi_Quận 2	16728 non-null	float64
12	DiaChi_Quận 3	16728 non-null	float64
13	DiaChi_Quận 4	16728 non-null	float64
14	DiaChi_Quận 5	16728 non-null	float64
15	DiaChi_Quận 6	16728 non-null	float64
16	DiaChi_Quận 7	16728 non-null	float64
17	DiaChi_Quận 8	16728 non-null	float64
18	DiaChi_Quận 9	16728 non-null	float64
19	DiaChi_Quận Bình Thạnh	16728 non-null	float64
20	DiaChi_Quận Bình Tân	16728 non-null	float64

21	DiaChi_Quận Gò Vấp	16728 non-null	float64
22	DiaChi_Quận Phú Nhuận	16728 non-null	float64
23	DiaChi_Quận Thủ Đức	16728 non-null	float64
24	DiaChi_Quận Tân Bình	16728 non-null	float64
25	DiaChi_Quận Tân Phú	16728 non-null	float64
26	TinhTrangBDS_ Chưa bàn giao	16728 non-null	float64
27	TinhTrangBDS_ Đã bàn giao	16728 non-null	float64
28	Loai_ Chung cư	16728 non-null	float64
29	Loai_ Căn hộ dịch vụ, mini	16728 non-null	float64
30	Loai_ Duplex	16728 non-null	float64
31	Loai_ Officetel	16728 non-null	float64
32	Loai_ Penthouse	16728 non-null	float64
33	Loai_ Tập thể, cư xá	16728 non-null	float64
34	GiayTo_ Giấy tờ khác	16728 non-null	float64
35	GiayTo_ Đang chờ sổ	16728 non-null	float64
36	GiayTo_ Đã có sổ	16728 non-null	float64
37	TinhTrangNoiThat_ Bàn giao thô	16728 non-null	float64
38	TinhTrangNoiThat_ Hoàn thiện cơ bản	16728 non-null	float64
39	TinhTrangNoiThat_ Nội thất cao cấp	16728 non-null	float64
40	TinhTrangNoiThat_ Nội thất đầy đủ	16728 non-null	float64
dtypes:	float64(41)		

Cuối cùng, công cụ StandardScaler được dùng để chuẩn hóa dữ liệu nhằm đưa giá trị các cột dữ liệu số về một thang đo với giá trị trung bình là 0 và độ lệch chuẩn là 1. Bảng 7 thể hiện kết quả chuẩn hóa cho 5 mẫu dữ liệu đầu tiên.

Bảng 7. *Sau khi Feature Scaling*

	DienTich	Gia/m2	PhongTam
0	-0.337101	0.107491	0.497391
1	1.207092	1.255721	0.497391
2	0.271218	-1.298883	0.497391
3	0.037249	1.321599	-1.278945
4	0.645568	0.267550	0.497391
5 rows × 41 columns			

Bước 3: Trích chọn đặc trưng

Trong nghiên cứu này, phương pháp Recursive Feature Elimination (RFE) được chọn làm kỹ thuật trích chọn đặc trưng chính, mặc dù có nhiều phương pháp khác như phương pháp Filter (Chi-Square Test, Mutual Information, ANOVA), phương pháp Wrapper (Forward Selection, Backward Elimination), và phương pháp Embedded (Lasso, Ridge Regression). RFE đặc biệt phù hợp với bài toán dự đoán giá nhà do khả năng xử lý hiệu quả các tập dữ liệu có nhiều đặc trưng và phức tạp. So với các phương pháp Filter như Chi-Square Test hay Mutual Information có thể bỏ qua mối tương tác phức tạp giữa các đặc trưng. Trong khi đó, RFE thực hiện quá trình loại bỏ đặc trưng một cách có hệ thống và lặp đi lặp lại, giúp xác định chính xác hơn tầm quan trọng của từng đặc trưng đối với giá nhà. RFE cũng vượt trội hơn các phương pháp Wrapper khác như Forward Selection hay Backward Elimination trong việc cân bằng giữa hiệu suất tính toán và độ chính xác. Đồng thời, RFE có thể kết hợp với nhiều loại mô hình khác nhau, không bị giới hạn bởi một mô hình cụ thể như với một số phương pháp Embedded như Lasso hay Ridge Regression [11] [12]. Trong nghiên cứu này, mô hình Linear Regression được sử dụng như một mô hình hồi quy cơ bản để đánh giá các tập hợp đặc trưng thông qua RFE. RFECV (Recursive Feature Elimination with Cross-Validation) được áp dụng với các thiết lập cụ thể: mô hình Linear Regression được chọn làm estimator, step=1: mỗi lần loại bỏ một đặc trưng kém nhất, cv=KFold(5): sử dụng K-Fold Cross-Validation với 5 fold để đánh giá mô hình, scoring='r2': sử dụng R-squared làm thước đo hiệu suất mô hình. Sau khi thực thực hiện Feature Selection ta giảm được số chiều dữ liệu từ 41 xuống còn 35 (Hình 11).

```
Optimal number of features: 35
Selected Features: Index(['DienTich', 'Gia/m2', 'PhongTam', 'DiaChi_Huyện Bình Chánh',
       'DiaChi_Huyện Củ Chi', 'DiaChi_Huyện Hóc Môn', 'DiaChi_Huyện Nhà Bè',
       'DiaChi_Quận 1', 'DiaChi_Quận 10', 'DiaChi_Quận 12', 'DiaChi_Quận 2',
       'DiaChi_Quận 3', 'DiaChi_Quận 4', 'DiaChi_Quận 5', 'DiaChi_Quận 6',
       'DiaChi_Quận 8', 'DiaChi_Quận 9', 'DiaChi_Quận Bình Thạnh',
       'DiaChi_Quận Bình Tân', 'DiaChi_Quận Gò Vấp', 'DiaChi_Quận Phú Nhuận',
       'DiaChi_Quận Thủ Đức', 'DiaChi_Quận Tân Bình',
       'TinhTrangBDS_ Chưa bàn giao', 'TinhTrangBDS_ Đã bàn giao',
       'Loai_ Chung cư', 'Loai_ Căn hộ dịch vụ, mini', 'Loai_ Officetel',
       'Loai_ Penthouse', 'Loai_ Tập thể, cư xá', 'GiayTo_ Đã có sổ',
       'TinhTrangNoiThat_ Bàn giao thô', 'TinhTrangNoiThat_ Hoàn thiện cơ bản',
       'TinhTrangNoiThat_ Nội thất cao cấp',
       'TinhTrangNoiThat_ Nội thất đầy đủ'],
      dtype='object')
Feature Ranking: [1 1 1 1 1 1 1 1 6 1 1 1 1 1 1 4 1 1 1 1 1 1 1 1 5 1 1 1 1 2 1 1 1 3 7 1
 1 1 1 1]
```

Hình 11. *Kết quả sau khi trích chọn đặc trưng.*

V. KẾT QUẢ THỰC NGHIỆM

Sau khi thực hiện phân tích và tiền xử lý cho tập dữ liệu, chúng tôi tiến hành thực nghiệm hai mô hình học máy để dự đoán giá nhà: SVR và RFR. Chương trình được viết bằng ngôn ngữ Python phiên bản 3.10.12, chạy trên nền tảng Google Colab, máy tính Intel Core i7-12700H, bộ nhớ 16GB và hệ điều hành Windows 11. Chúng tôi sử dụng kỹ thuật Cross-Validation để đảm bảo tính khách quan và chính xác của kết quả đánh giá. Các độ đo bao gồm Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2), độ lệch chuẩn của R^2, và R-squared hiệu chỉnh (Adjusted R^2).

(i) MAE [13] là 1 metric đánh giá mô hình bằng cách tính trung bình giá trị tuyệt đối sai số giữa giá trị thực tế và giá trị dự đoán được tính theo Công thức 1. Trong đó: x_i là các giá trị thực tế và y_i là các giá trị dự đoán.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \quad (1)$$

(ii) MSE [13] là một chỉ số phổ biến trong thống kê và học máy. Nó tính trung bình của bình phương sai số giữa các giá trị thực tế và các giá trị dự đoán của một tập dữ liệu. MSE thường được sử dụng trong các vấn đề hồi quy để đánh giá hiệu suất của các mô hình dự đoán được tính theo Công thức 2.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \quad (2)$$

(iii) RMSE [13] là căn bậc hai của MSE. RMSE đo lường mức độ phù hợp của các giá trị dự đoán từ mô hình so với các giá trị thực tế được quan sát trong tập dữ liệu được tính theo Công thức 3.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

(iv) R^2 [14] là một chỉ số thống kê được sử dụng để đánh giá mức độ phù hợp của một mô hình hồi quy tuyến tính. Nó đo lường tỷ lệ biến thiên trong biến phụ thuộc (biến kết quả) được giải thích bởi biến độc lập (biến dự đoán) trong mô hình. R^2 có giá trị từ 0 đến 1, giá trị R^2 càng cao, mô hình càng phù hợp với dữ liệu và giải thích được nhiều biến thiên hơn (Công thức 4).

$$R^2 = 1 - \frac{RSS}{TSS} \quad (4)$$

Trong đó: RSS (Residual Sum of Squares): Tổng bình phương của các sai số (phần dư) giữa giá trị dự đoán và giá trị thực tế được tính theo Công thức 5 ($y_i$là giá trị thực tế của biến phụ thuộc và $\hat{y}_i$ là giá trị dự đoán từ mô hình). TSS (Total Sum of Squares): Tổng bình phương của sự khác biệt giữa giá trị thực tế và giá trị trung bình của chúng được tính theo Công thức 6 ($\overline{y}_i$ là giá trị trung bình của biến phụ thuộc và n là số lượng quan sát).

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (5)$$

$$TSS = \sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2 \quad (6)$$

(v) Adjusted R^2 [14] là một phiên bản được điều chỉnh của R-squared, được sử dụng để đánh giá mức độ phù hợp của mô hình hồi quy tuyến tính. R^2 thông thường chỉ đo lường tỷ lệ biến thiên của biến phụ thuộc được giải thích bởi mô hình hồi quy. Tuy nhiên, một hạn chế lớn của R^2 là nó luôn tăng hoặc không thay đổi khi thêm biến giải thích mới vào mô hình, ngay cả khi biến đó không thực sự có ý nghĩa. Để khắc phục vấn đề này, Adjusted R-squared trong Công thức 7 được sử dụng. Với n là số lượng điểm dữ liệu, k là số lượng biến độc lập và R^2 là giá trị R-squared

$$Adjusted\ R^2 = \left(\frac{(1 - R^2)(n - 1)}{n - k - 1}\right) \quad (7)$$

Trong nghiên cứu này, chúng tôi sử dụng phương pháp RandomizedSearchCV để xác định các siêu tham số phù hợp cho hai mô hình SVR và RFR [15]. RandomizedSearchCV thực hiện tìm kiếm ngẫu nhiên các tổ hợp siêu tham số từ các lưới tham số đã định nghĩa và đánh giá mô hình trên tập dữ liệu sử dụng phương pháp cross-validation. RandomizedSearchCV được chọn thay vì GridSearchCV, do nó hiệu quả hơn về mặt thời gian khi làm việc với không gian tham số lớn, đồng thời vẫn cung cấp kết quả tối ưu gần như tương đương.

Cả hai quá trình tìm kiếm siêu tham số tối ưu và đánh giá mô hình đều sử dụng K-Fold Cross-Validation với k = 5. Kỹ thuật này chia dữ liệu thành 5 phần, mỗi phần lần lượt được sử dụng làm tập kiểm tra trong khi các phần còn lại được sử dụng làm tập huấn luyện. Việc sử dụng kỹ thuật này giúp giảm thiểu hiện tượng quá khớp và đảm bảo mô hình có tính tổng quát tốt hơn. Tiêu chí đánh giá (scoring) là 'neg_root_mean_squared_error', nhằm tìm kiếm mô hình có khả năng dự đoán chính xác nhất bằng cách tối thiểu hóa sai số RMSE.

Đối với mô hình SVR, chúng tôi thiết lập không gian tìm kiếm với các giá trị C (0.1, 1, 10, 100) và các loại kernel ('rbf', 'linear', 'poly', 'sigmoid'), thực hiện 16 lần lặp ngẫu nhiên. Kết quả cho thấy siêu tham số tốt nhất cho SVR là kernel 'rbf' và C = 1. Hàm kernel 'rbf' giúp biến đổi không gian đặc trưng để mô hình có thể tìm ra các quan hệ phi tuyến giữa các đặc trưng, trong khi tham số C điều chỉnh mức độ phạt áp dụng cho các lỗi trong dữ liệu huấn luyện, giúp đạt được sự cân bằng giữa độ phức tạp của mô hình và khả năng khái quát hóa.

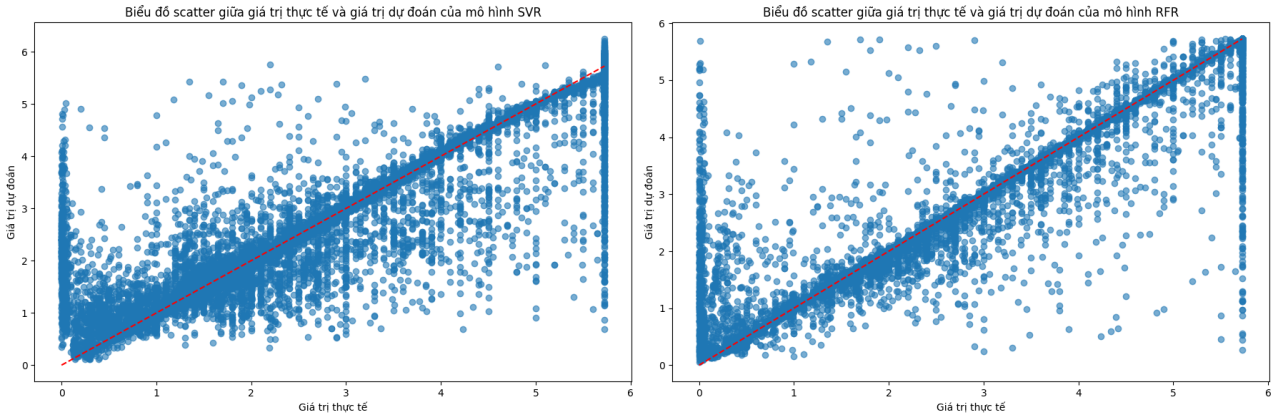
Đối với mô hình RFR, chúng tôi thiết lập không gian tìm kiếm với các giá trị n_estimators (số lượng cây) từ 50 đến 300 với bước nhảy 50, max_depth (độ sâu tối đa của cây) từ 10 đến 30 với bước nhảy 1 và thêm giá trị None, min_samples_split (số mẫu tối thiểu để phân tách) từ 2 đến 10 với bước nhảy 2, và min_samples_leaf (số mẫu tối thiểu ở nút lá) từ 1 đến 4 với bước nhảy 1. Quá trình tìm kiếm này thực hiện 50 lần lặp ngẫu nhiên. Kết quả các siêu tham số tối ưu cho RFR là n_estimators = 300, min_samples_split = 4, min_samples_leaf = 1, và max_depth = 22. Chúng tôi sử dụng random_state = 42 cho cả RandomizedSearchCV và mô hình RFR để đảm bảo tính tái lập của kết quả.

Bảng 8. *So sánh kết quả của SVR và RFR*

Độ đo	SVR	RFR
MAE	0,2978	0,1617
MSE	0,4278	0,2817
RMSE	0,6541	0,5307
R^2	0,7892	0,8611
Adjusted R^2	0,7887	0,8608
Thời gian thực thi	97,07 s	212,14 s

RFR vượt trội hơn SVR ở hầu hết các chỉ số đánh giá. Cụ thể, RFR có các giá trị MAE, MSE và RMSE thấp hơn, đồng nghĩa với việc có độ chính xác cao hơn và ít lỗi hơn trong dự đoán. Bên cạnh đó, giá trị R^2 và Adjusted R^2 của RFR cũng cao hơn, cho thấy khả năng giải thích biến thiên của dữ liệu tốt hơn. Mặc dù RFR có thời gian thực thi dài hơn so với SVR, tuy nhiên, sự chênh lệch này là không đáng kể so với lợi ích mà RFR mang lại về mặt độ chính xác và khả năng giải thích. Nhìn chung, mô hình RFR là lựa chọn tốt hơn cho bài toán dự đoán giá nhà trong trường hợp này.

Hình 12 thể hiện mối quan hệ giữa giá trị thực tế và giá trị dự đoán của hai mô hình dự đoán giá nhà: SVR và RFR. Biểu đồ scatter giữa giá trị thực tế và giá trị dự đoán của mô hình SVR cho thấy các điểm dữ liệu nằm rải rác và có độ phân tán lớn xung quanh đường lý tưởng (đường màu đỏ), đặc biệt là ở các giá trị thấp và cao. Điều này cho thấy mô hình SVR có sự bất ổn định và độ chính xác không cao trong việc dự đoán giá trị thực tế. Trong khi đó, Biểu đồ scatter giữa giá trị thực tế và giá trị dự đoán của mô hình RFR cho thấy các điểm dữ liệu tập trung gần đường lý tưởng hơn, cho thấy mô hình này có khả năng dự đoán chính xác và ổn định hơn. Các điểm dữ liệu ít bị phân tán hơn, đặc biệt là ở các giá trị thấp và cao, chứng tỏ mô hình Random Forest Regressor có độ chính xác cao hơn trong việc dự đoán giá trị thực tế. Do đó, RFR cho thấy sự ưu việt hơn so với SVR trong việc dự đoán giá trị thực tế, với độ chính xác cao hơn và sự ổn định tốt hơn. Điều này cho thấy RFR là một mô hình dự đoán hiệu quả và đáng tin cậy hơn so với SVR trong bài toán dự đoán giá nhà.



Hình 12. *Biểu đồ scatter thể hiện mối quan hệ giữa giá trị thực tế và giá trị dự đoán của mô hình SVR và RFR.*

VI. KẾT LUẬN

Bài báo đã trình bày việc áp dụng các kỹ thuật phân tích và tiền xử lý dữ liệu trong Python để dự đoán giá nhà tại thành phố Hồ Chí Minh. Qua việc phân tích và tiền xử lý dữ liệu, dữ liệu được làm sạch để huấn luyện mô hình. Trong nghiên cứu này, hai mô hình học máy được thử nghiệm là SVR và RFR. Kết quả cho thấy mô hình RFR hiệu quả hơn khi nắm bắt các mối quan hệ phức tạp và phi tuyến tính, ít bị ảnh hưởng bởi ngoại lệ và nhiễu, và đạt hiệu suất tốt hơn so với mô hình SVR. Quá trình tiền xử lý và chuẩn hóa dữ liệu đã chứng minh được sự quan trọng trong việc cải thiện độ chính xác của các mô hình học máy. Các bước tiền xử lý dữ liệu như mã hóa biến số, chuẩn hóa dữ liệu và giảm chiều dữ liệu đã đóng vai trò quan trọng trong việc cải thiện hiệu suất dự đoán.

Nhìn chung, nghiên cứu cho thấy rằng việc sử dụng các phương pháp phân tích và tiền xử lý dữ liệu hiện đại kết hợp với các thuật toán học máy tiên tiến có thể mang lại những kết quả đáng kể trong việc dự đoán giá nhà. Điều này không chỉ giúp cải thiện độ chính xác của mô hình mà còn mở ra nhiều hướng nghiên cứu mới trong việc tối ưu hóa các quy trình tiền xử lý và mô hình hóa dữ liệu. Nghiên cứu trong tương lai có thể tập trung vào việc thử nghiệm các phương pháp học máy khác, cũng như áp dụng các kỹ thuật tiền xử lý và phân tích dữ liệu mới để tiếp tục cải thiện độ chính xác và hiệu quả của mô hình dự đoán giá nhà.

VII. TÀI LIỆU THAM KHẢO

[1] N. Kok, P. Monkkonen and John M. Quigley , "Economic Geography, Jobs, and Regulations: The Value of Land and Housing," 3 2011. [Online]. Available: https://www.researchgate.net/publication/228558115_Economic_Geography_Jobs_and_Regulations_The_Value_of_Land_and_Housing. [Accessed 06 07 2024].

[2] C. Zhan, Zeqiong Wu, Yonglin Liu, Zefeng Xie and Wangling Chen, "Housing prices prediction with deep learning: an application for the real estate market in Taiwan," 20 07 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9442244. [Accessed 06 07 2024].

[3] H. Sharma, H. Harsora and B. Ogunleye, "An Optimal House Price Prediction Algorithm: XGBoost," *Analytics*, vol. 3, no. 1, pp. 30-45, 06 02 2024.

[4] S. Abdul-Rahman, S. Mutalib, N. H. Zulkifley and I. Ibrahim, "Advanced Machine Learning Algorithms for House," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 12, no. 12, pp. 736-745, 2021.

[5] H. Zhang, Yansong Li and Paula Branco, "Describe the house and I will tell you the price: House price prediction with textual description data," *Natural Language Engineering*, p. 1–35, 2023.

[6] C. Li, "House price prediction using machine learning," *Proceedings of the 4th International Conference on Signal Processing and Machine Learning*, vol. 53, no. 1, pp. 225-237, 2024.

[7] R. Ritu, "Machine learning techniques for house price prediction: A literature review," 12 2023. [Online]. Available: https://www.researchgate.net/publication/376519663_Machine_learning_techniques_for_house_price_prediction_A_literature_review.

[8] subhambhansali2000, "When to use Random Forest over SVM and vice versa?," [Online]. Available: https://www.geeksforgeeks.org/when-to-use-random-forest-over-svm-and-vice-versa/.

[9] susmit_sekhar_bhakta, "Random Forest Algorithm in Machine Learning," [Online]. Available: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/.

[10] HungTrinhIT, "FinalProject-Datascience," 2020. [Online]. Available: https://github.com/HungTrinhIT/FinalProject-Datascience/blob/main/Chotot/rawdata.csv. [Accessed 2024].

[11] GeeksforGeeks, "Feature selection techniques in machine learning," 19 03 2024. [Online]. Available: https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/. [Accessed 08 2024].

[12] Javatpoint, "Feature selection techniques in Machine Learning - JavatPoint.," [Online]. Available: https://www.javatpoint.com/feature-selection-techniques-in-machine-learning. [Accessed 6 2024].

[13] nandakishorereddy, "Regression Metrics," [Online]. Available: https://www.geeksforgeeks.org/regression-metrics/.

[14] V. Singh, "How to Calculate Adjusted R-Squared," [Online]. Available: https://www.shiksha.com/online-courses/articles/adjusted-r-squared/.

[15] swapnilvishwakarma7, "Comparing Randomized Search and Grid Search for Hyperparameter Estimation in Scikit Learn," 30 12 2022. [Online]. Available: https://www.geeksforgeeks.org/comparing-randomized-search-and-grid-search-for-hyperparameter-estimation-in-scikit-learn/.

APPLICATION OF DATA ANALYSIS AND PREPROCESSING IN PYTHON TO THE HOUSING PRICE PREDICTION PROBLEM

Nguyen Mai Khanh Vy, Tran Ngoc Thao Ngan, Tran Minh Thai

ABSTRACT— Predictive modelling is one of the most important and widely applicable problems in machine learning. It serves as the foundation for many important applications in human life, ranging from familiar areas such as weather forecasting and price prediction to more complex areas such as disease diagnosis, fraud detection, and autonomous driving. The focus of predictive modelling is to predict the outcome of an event or a variable in the future based on historical data by automatically learning from the data and building a prediction model. This paper proposes a model to predict housing prices in Ho Chi Minh City. Through the application of analysis and data preprocessing techniques using Python programming language libraries, the data is cleaned, missing values are handled, duplicates and outliers are addressed, categorical variables are encoded, data is normalized, and feature selection and dimensionality reduction are performed. Next, machine learning models are trained to predict housing prices using the Support Vector Regressor (SVR) and Random Forest Regressor (RFR) methods. Experimental results show that RFR is capable of capturing complex and nonlinear relationships, is less affected by outliers and noise, and outperforms SVR in terms of performance.

Keywords— House price prediction, data analysis, data preprocessing, Support Vector Regressor, Random Forest Regressor, Python.

TS. Trần Minh Thái tốt nghiệp cử nhân Công nghệ thông tin (CNTT) năm 2001 và thạc sỹ Tin học năm 2006 Trường ĐH Khoa học tự nhiên – ĐH Quốc gia TP.HCM, nhận bằng tiến sỹ CNTT năm 2017 do ĐH Quốc gia TP.HCM cấp. TS. Thái từng là giảng viên và quản lý khoa CNTT trường Cao đẳng CNTT TP.HCM từ 2002 đến 2014. Từ 2015 đến nay, anh là giảng viên và là trưởng bộ môn Hệ thống Thông tin thuộc khoa CNTT Trường ĐH Ngoại ngữ - Tin học TP.HCM. Lĩnh vực nghiên cứu chính hiện tại của TS. Thái liên quan đến vấn đề khai thác dữ liệu, ẩn dữ liệu, xử lý dữ liệu lớn và nhận dạng.

Nguyễn Mai Khánh Vy hiện là sinh viên chuyên ngành Khoa học dữ liệu, ngành Công nghệ thông tin tại Trường ĐH Ngoại ngữ - Tin học TP.HCM (HUFLIT).

Hướng nghiên cứu chính: Phân tích và tiền xử lý dữ liệu, trí tuệ nhân tạo.

Trần Ngọc Thảo Ngân hiện là sinh viên chuyên ngành Khoa học dữ liệu, ngành Công nghệ thông tin tại Trường ĐH Ngoại ngữ - Tin học TP.HCM (HUFLIT).

Hướng nghiên cứu chính: Phân tích và tiền xử lý dữ liệu, trí tuệ nhân tạo.