ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH VÀ DỰ ĐOÁN GIÁ CĂN HỘ TRÊN ĐỊA BÀN THÀNH PHỐ HỒ CHÍ MINH

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Nguyễn Ngọc Lương	21522311
2	Nguyễn Thiện Trí	21522707
3	Đoàn Bảo Long	21520332

TP. HỒ CHÍ MINH – 12/2023

1. GIỚI THIỆU

Trong thời đại công nghệ ngày nay, sự phát triển của thị trường bất động sản đang diễn ra với tốc độ nhanh chóng. Điều này tạo ra nhu cầu lớn cho các công cụ dự đoán giá nhà hiệu quả, giúp người mua, người bán và các nhà đầu tư có cái nhìn chính xác và toàn diện về giá trị của bất động sản. Vì vậy, trong khuôn khổ đồ án này, chúng tôi thực hiện đề tài này để thực hiện phân tích trực quan các yếu tố ảnh hưởng đến giá cả bất động sản cũng như xây dựng các mô hình dự đoán cho bộ dữ liệu mà nhóm đã thu thập.

Bộ dữ liệu được nhóm tự thu thập tại website Nhà Tốt [1], đây là một website được liên kết với website Chợ Tốt [2] - một trang thương mại điện tử phổ biến tại Việt Nam, nơi người dùng có thể mua bán các sản phẩm và dịch vụ khác nhau. Trang web hoạt động như một nền tảng quảng cáo phân loại, cho phép cá nhân và doanh nghiệp đăng thông tin về sản phẩm hoặc dịch vụ mà họ muốn bán.

2. MÔ TẢ BỘ DỮ LIỆU

Bộ dữ liệu chứa thông tin của các bài đăng quảng cáo (advertisement) rao bán căn hộ chung cư trên địa bàn thành phố Hồ Chí Minh. Bộ dữ liệu được nhóm tự thu thập tại trang web Nhà Tốt bằng cách sử dụng thư viện selenium, request trong Python và hidden API từ Chợ Tốt.

1.1. Quy trình thu thập bộ dữ liệu

- Dùng thư viện selenium để tạo 1 trình duyệt web giúp tự động thu thập các đường link tới các bài đăng trên website Nhà Tốt.
- Trích xuất ra ID của bài đăng từ văn bản liên kết, ví dụ: trang web có đường link https://www.nhatot.com/mua-ban-can-ho-chung-cu-quan-binh-tan-tp-ho-chi-minh/109701596.htm sẽ có ID là 109701596.
- Sử dụng công cụ Developer Tools [3] của trình duyệt web và vào phần network inspector để bắt các gói tin trả về từ đường link trên, sau đó trích xuất ra gói tin chứa hidden API của Chợ Tốt.
- Từ danh sách các ID của các bài đăng, nhóm sử dụng thư viện request để gửi yêu cầu đến hidden API của Chợ Tốt như đã nói ở trên và lấy dữ liệu của căn hộ từ bài đăng dưới dạng một đối tượng JSON. Ví dụ: Bài đăng có ID là 109701596 sẽ chứa dữ liệu ở https://gateway.chotot.com/v2/public/ad-listing/109701596?adview_position=true&tm=treatment2 .
- Chọn ra các thuộc tính cần thiết cho việc phân tích và chạy mô hình, sau đó lưu lại bộ dữ liệu ra file Excel.

1.2. Thông tin các thuộc tính

Bộ dữ liệu được thu thập vào ngày 14-11-2023, gồm tổng cộng 9550 dòng và 35 thuộc tính. Trong đó các thuộc tính quan trọng được dùng để phân tích và huấn luyện mô hình gồm:

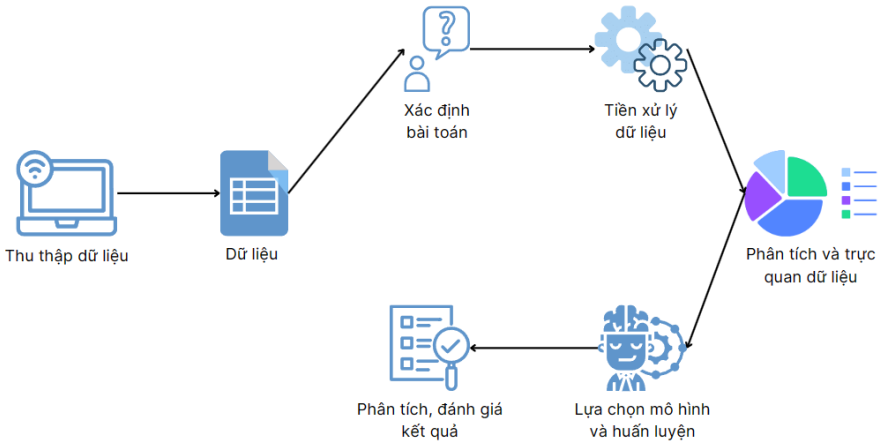
Tên cột	Desription	Kiểu dữ liệu
ad_price	Giá căn hộ	Số
ad_projectid	Mã dự án của căn hộ	Phân loại

ad_company_ad	Tài khoản đăng tin là môi giới (True) hay cá nhân (False)?	Phân loại
ad_rooms	Số phòng ngủ	Số
ad_size	Diện tích căn hộ	Số
ad_ward_name	Tên phường	Phân loại
ad_area_name	Tên quận	Phân loại
ad_floornumber	Căn hộ nằm ở tầng số mấy?	Số
ad_longitude	Kinh độ	Số
ad_latitude	Vĩ độ	Số
ad_protection_entitlement	Bài đăng có phải từ đối tác Nhà Tốt hay không?	Phân loại
ad_street_name	Tên đường	Phân loại
ad_params_apartment_feature_value	Đặc điểm căn hộ	Phân loại
ad_params_apartment_type_value	Loại căn hộ	Phân loại
ad_params_property_status_value	Tình trạng bất động sản	Phân loại
ad_toilets	Số phòng vệ sinh	Số
ad_params_property_legal_document_value	Tình trạng giấy tờ pháp lý	Phân loại
ad_params_furnishing_sell_value	Tình trạng nội thất	Phân loại

3. PHƯƠNG PHÁP PHÂN TÍCH

3.1. Tổng quan quy trình

Phương pháp phân tích dữ liệu tập trung vào phân tích thăm dò:



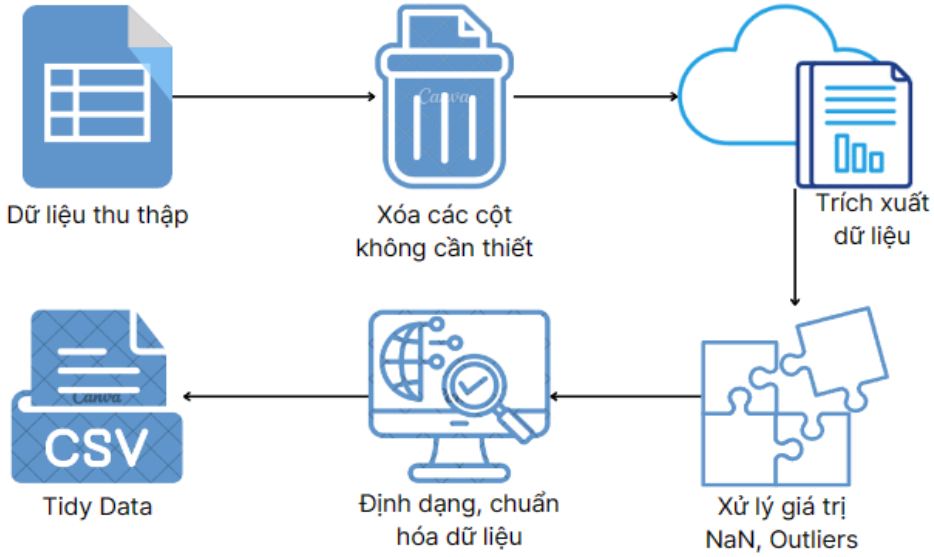
Hình 1: Quy trình phân tích dữ liệu

Mô tả: nhóm bắt đầu với việc thu thập dữ liệu, khi đã có bộ dữ liệu thô nhóm tiến hành xác định bài toán. Sau đó, thực hiện phân tích, thăm dò dữ liệu, làm sạch và tiền xử lý dữ liệu. Sau khi có được bộ dữ liệu sạch, nhóm tiến hành phân tích trực quan để có cái nhìn rõ hơn về bộ dữ liệu, tiếp theo chia bộ dữ liệu và chọn các mô hình phù hợp để huấn luyện dữ liệu. Cuối cùng, nhóm đánh giá và phân tích kết quả của các mô hình, từ đó rút ra được kết luận và hướng phát triển trong tương lai.

3.2. Tiền xử lý bộ dữ liệu

Phần tiền xử lí dữ liệu gồm các bước:

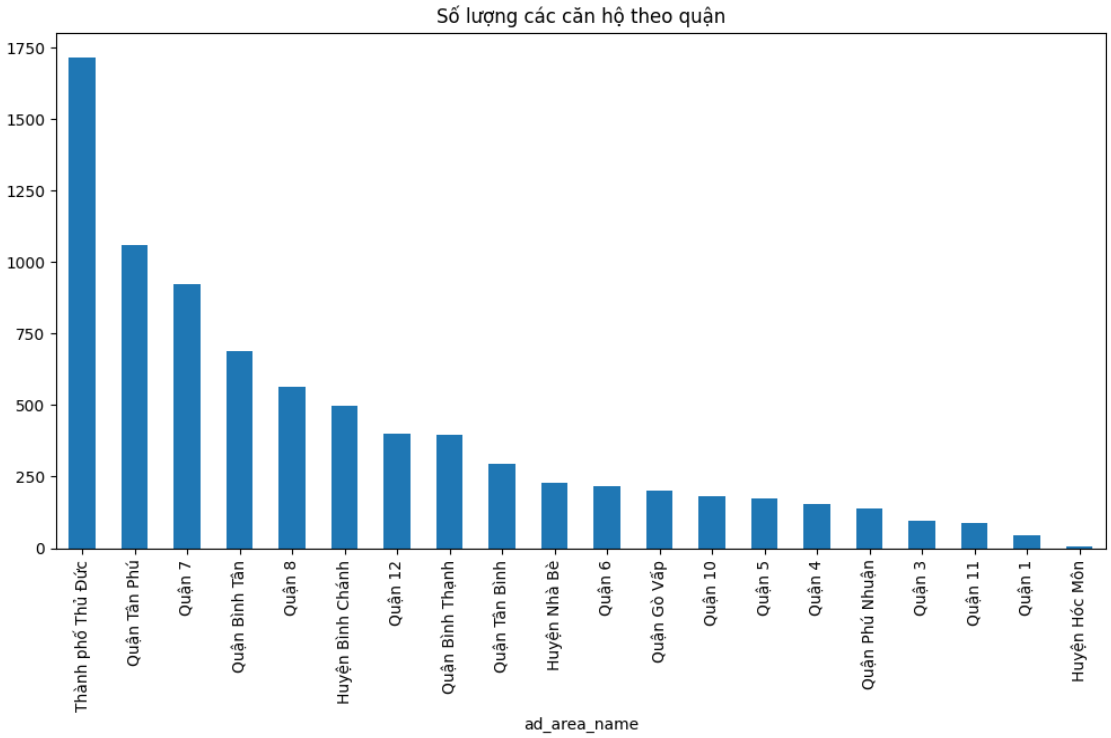
- **Xóa các cột dư thừa và các cột có số lượng Na value lớn**: Nhóm tiến hành xóa các cột chứa thông tin về ID, tên dự án, tên người đăng,... và những cột không có ý nghĩa như thời điểm đăng tin, link ảnh ... Ngoài ra, những cột chứa các thông tin bị trùng lặp với cột khác (VD: cột ad_street_name chứa tên đường là không cần thiết bởi những thông tin này đã có trong ad_detail_address) và những cột chứa quá nhiều missing data đều bị loại bỏ.
- **Trích xuất thông tin từ 1 số cột, tạo cột mới, feature engineering**: Cột ad_detail_address chứa thông tin về địa chỉ chi tiết của chung cư. Ban đầu nhóm dự định tách tên đường, tên phường và quận ra làm 3 cột mới, tuy nhiên số lượng đường và phường là rất lớn, dẫn đến việc khó khăn trong phân tích và thông kê nên nhóm quyết địch chỉ tách tên quận ra làm một cột mới là ad_area_name và xóa cột ad_detail_address.
- **Điền các giá trị bị khuyết**: Nhóm tiến hành điền giá trị bị khuyết cho các cột bằng các phương pháp khác nhau tùy vào đặc tính của từng cột. Nhóm chủ yếu sử dụng 2 phương pháp là "ffill" và điền bằng giá trị mode. Đối với biến "ad_params_apartment_feature_value" chứa thông tin về đặc điểm căn hộ(chỉ chứa các giá trị của căn góc, những căn thường mang giá trị mặc định là Null) nhóm thực hiện thay thế các giá trị Null bằng "Căn thường".
- **Định dạng dữ liệu**: xử lý một số cột chứa các giá trị trùng lặp như tên quận("Quận Tân Phú" và "Quận Tân Phú.").
- **Xử lí các giá trị nhiễu trong bộ dữ liệu**: Tiến hành xử lý giá trị nhiễu của các cột có biến kiểu số trong bộ dữ liệu. Xây dựng hàm trích xuất và xác định được các giá trị nhiễu, sau đó xóa ra khỏi bộ dữ liệu. Cột "ad_price" chứa thông tin về giá nhà có rất nhiều mẫu chứa thông tin sai lệch, vì vậy nhóm đã tiến hành loại bỏ thủ công các mẫu sai(những mẫu đăng cho thuê nhà trong mục mua nhà hoặc những mẫu đưa ra giá nhà ảo, không đúng với sự thật) rồi mới tiến hành loại bỏ các Outliers.
- **Chuẩn hóa dữ liệu, chuyển đổi các biến kiểu phân loại thành các biến kiểu số:** Nhóm quyết định sử dụng phương pháp chuẩn hóa MinMaxScaler cho các biến kiểu số. Đối với các biến kiểu phân loại, nhóm sử dụng phương pháp LabelEncoder để chuyển sang biến kiểu số.



Hình 2: Quy trình tiền xử lý dữ liệu

3.3. Phân tích thăm dò

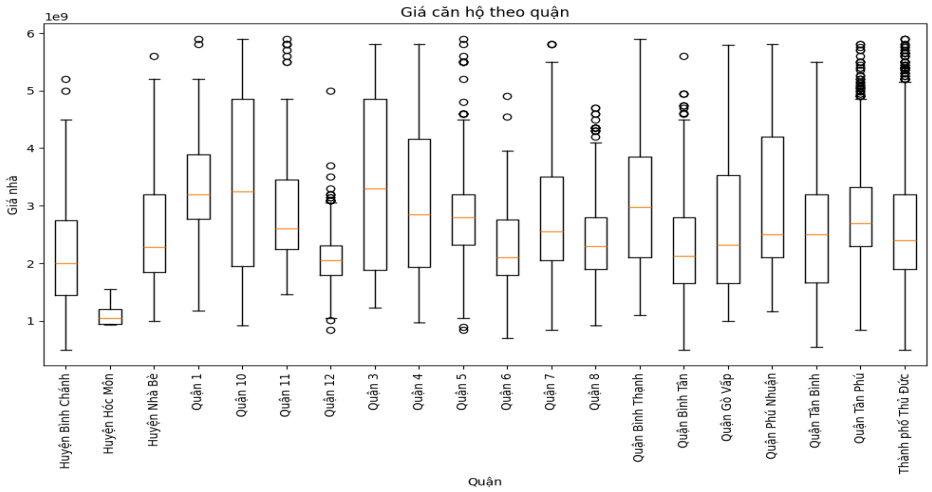
Số lượng căn hộ được bán ở thành phố Thủ đức là nhiều nhất:



Hình 3. Ảnh hưởng của dữ liệu nhiễu và outliers đến mô hình hồi quy.

Điều này cũng dễ hiểu bởi diện tích thành phố Thủ Đức lớn hơn rất nhiều so với các quận khác, cũng như đông dân hơn.

Tuy nhiên, giá căn hộ trung bình ở các quận không có sự chênh lệch quá lớn:



Hình 4.Tình hình giá căn hộ ở mỗi quận.

Giá căn hộ ở tất cả các quận phân bố rất không đều (ta sẽ không quá để ý tới Huyện Hóc Môn vì có số nhà rất thấp). Đặc biệt là ở Quận Tân Phú, Thành phố Thủ Đức có nhiều căn hộ có giá cực kì hơn so với các căn trong khu vực (số lượng positive outlier nhiều). Bởi vì đây đều là những khu vực đông dân cư và có nhiều trung tâm thương mại cũng như các vị trí đắt đỏ.

Sau khi thực hiện tính toán Hệ số tương quan, ta nhận thấy có 3 biến liên tục có khả năng ảnh hưởng tới giá là: Số phòng, Kích thước và Số toilets. Các yếu tố này khá dễ hiểu khi chúng có ảnh hưởng khá lớn đến giá của căn hộ

	Correlation	P-value
ad_rooms | ad_price	0.35933216133756507	1.5322736297949094e-244
ad_size | ad_price	0.5520392764274383	0.0
ad_toilets | ad_price	0.24441121509518968	4.443633545673152e-110

Hình 5.Các biến liên tục có ảnh hưởng đến giá căn hộ.

Nhóm còn nhận thấy được sự tương quan của các cặp giá trị trong các giá trị ad_rooms, ad_size, ad_toilets.

	Correlation	P-value
ad_rooms | ad_size	0.6857178656246309	0.0
ad_size | ad_toilets	0.5105750693816791	0.0
ad_toilets | ad_rooms	0.4462727649678429	0.0

Hình 6.Tương quan giữa ad_toilets, ad_price, ad_rooms.

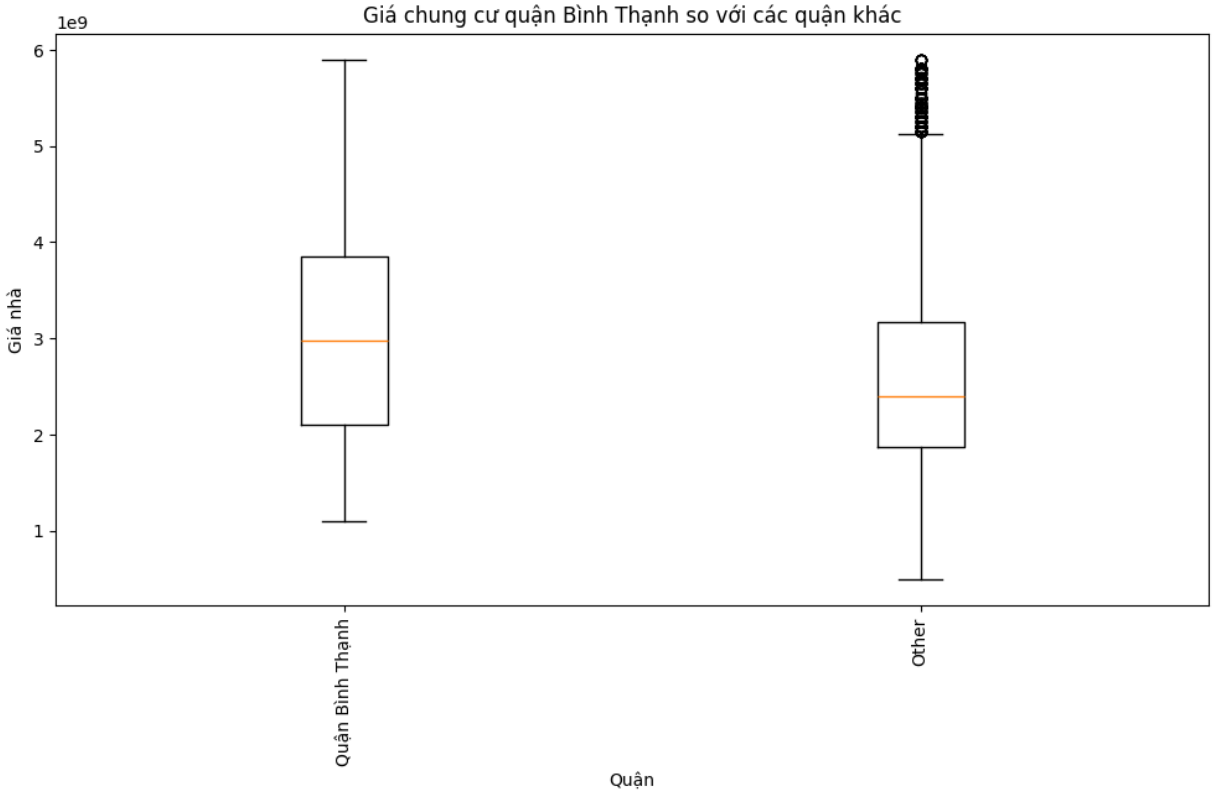
Điều này cũng dễ giải thích khi các giá trị về Số phòng, Số phòng vệ sinh hay Kích thước đều có liên quan với nhau khi căn hộ chỉ có diện tích khá giới hạn.

Sau khi thực nghiệm bằng phương pháp ANOVA cho các biến phân loại, ta tìm được các biến phân loại có tầm ảnh hưởng tới giá căn hộ là: Đặc điểm căn hộ, Loại căn hộ, Tình trạng bất động sản, Tình trạng giấy tờ pháp lý, Tình trạng nột thất, Khu vực (Quận).

	F-statistic	P-value
ad_params_apartment_feature_value	143.43006282236274	8.966496084814862e-33
ad_params_apartment_type_value	29.675245031997104	5.6530209647746384e-30
ad_params_property_status_value	18.145623925854643	2.069624995620055e-05
ad_params_property_legal_document_value	4.774207238248422	0.00846865145500523
ad_params_furnishing_sell_value	20.495021663628815	3.169062123515327e-13
ad_area_name	44.41892094178392	4.017343551667855e-158

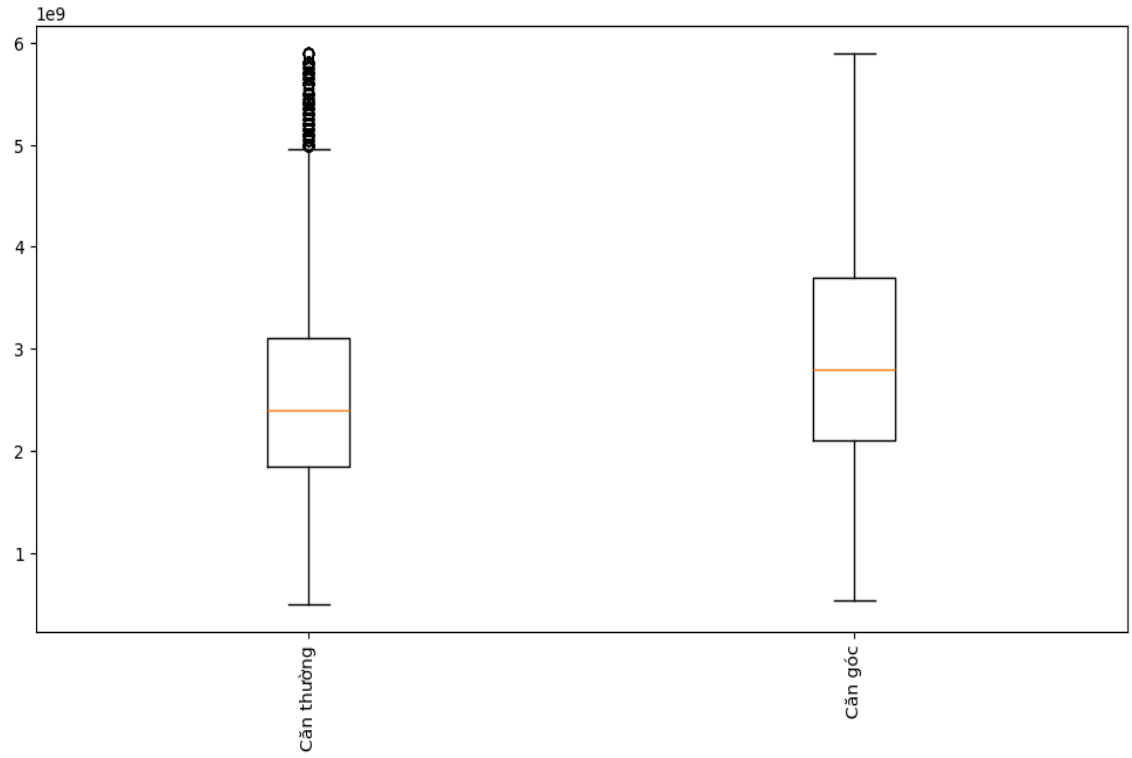
Hình 7.Kết quả thực nghiệm ANOVA cho các biến phân loại.

Ngoài ra thì nhóm còn thực hiện việc thực nghiệm ANOVA với giá căn hộ của từng quận thì nhận thấy giá chung cư tốt nhất là ở Quận Bình Thạnh (vì Quận Bình Thạnh ít ảnh hưởng tới căn hộ nhất so với các quận còn lại).



Hình 8. Giá của các căn hộ ở Quận Bình Thạnh so với các quận còn lại.

Có sự khác nhau về giá khá lớn giữa những căn hộ có là Căn thường và Căn góc:



Hình 9.Khác biệt về giá của Căn thường và Căn góc.

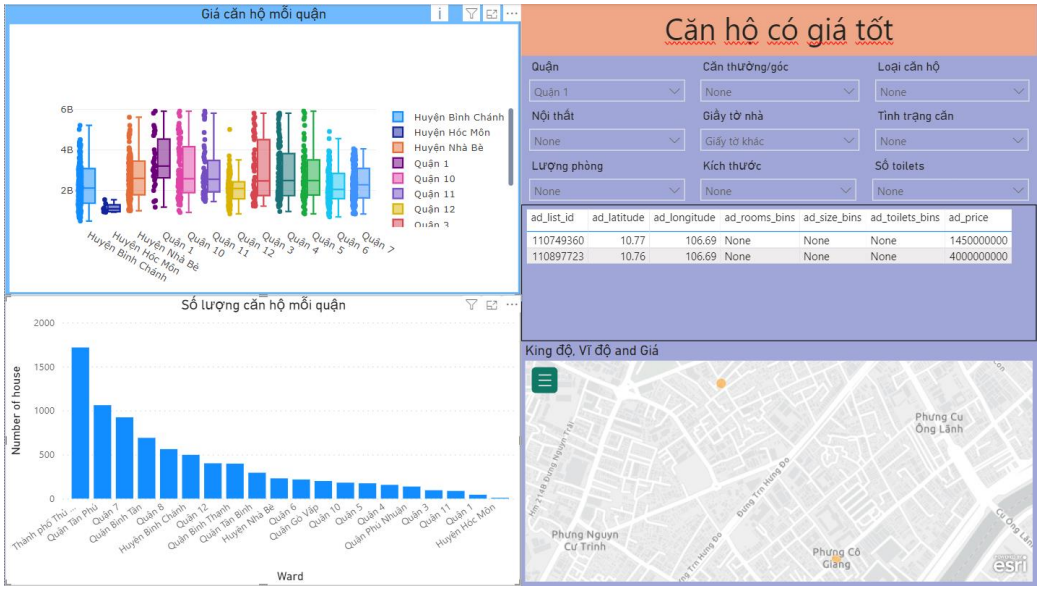
Bởi vì Căn góc là những căn ở góc của chung cư nên có Thiết kế khác biệt so với các Căn thường, và có quang cảnh đẹp hơn, vì là ở góc của chung cư nên số lượng cũng ít hơn nhiều, từ đó dẫn đến sự chênh lệch giá.

Nhóm còn thực nghiệm xem căn chung cư có các đặc điểm như thế nào sẽ có giá tốt nhất. Và kết quả là các căn hộ có các đặc điểm sau:

Đặc điểm	**Giá trị**
Đặc điểm chính	Căn thường
Loại căn hộ	Chung cư
Tình trạng căn hộ	Đã bàn giao
Giấy tờ pháp lý	Đã có sổ
Nội thất	Nội thất đầy đủ
Khu vực	Quận 10
Số phòng	0-2
Kích cỡ	0-40(m2)
Số toilet	0-2

Hình 10. Tập đặc điểm mà căn hộ giá tốt nhất có.

3.4. Dashboard trực quan hóa dữ liệu



Hình 11.Tập đặc điểm mà căn hộ giá tốt nhất có.

Dashboard sẽ gồm có 2 phần:

- Bên trái gồm 2 biểu đồ: Một biểu đồ hộp biểu diễn giá căn hộ theo quận và một biểu đồ thanh biểu diễn số lượng căn hộ theo quận

- Bên phải sẽ biểu diễn thông tin về các căn hộ có giá tốt được nhóm phân tích bằng phương pháp ANOVA, dữ liệu sẽ được biểu diễn thông qua các bộ lọc dữ liệu.

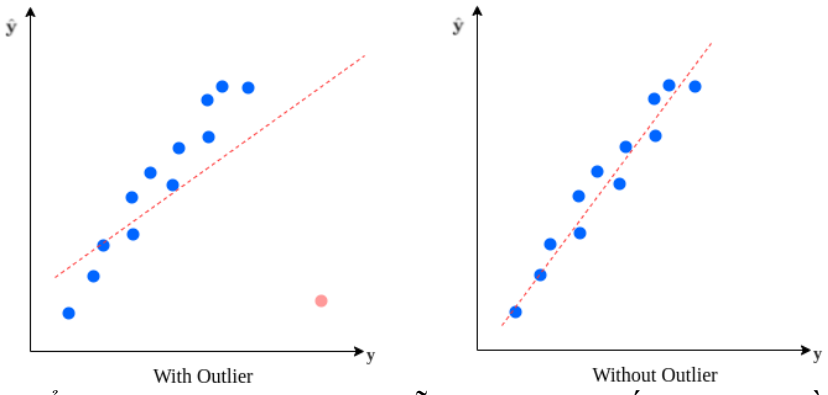
4. MÔ HÌNH HỒI QUY

Nhóm đã sử dụng một số mô hình hồi quy từ thư viện sklearn [4] và các thư viện mô hình hồi quy khác trong Python để dự đoán giá của các căn hộ dựa vào các biến đầu vào trong bộ dữ liệu, output của bài toán là biến mục tiêu được dùng để dự đoán ‘ad_price’. Nhóm cũng thực hiện việc nghiên cứu bộ dữ liệu để tìm ra mô hình hồi quy tốt nhất cho bài toán, từ đó có được một số nhận xét và đánh giá cho mức độ hiệu quả của các mô hình.

4.1. Nghiên cứu xây dựng mô hình

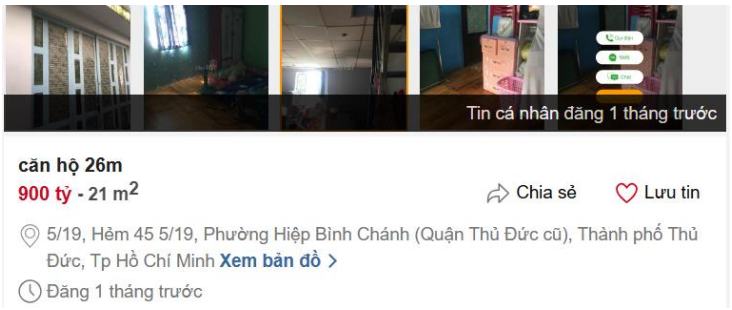
Để dự đoán giá căn hộ, trước tiên nhóm sử dụng mô hình hồi quy tuyến tính cơ bản là Linear Regression với input là các biến số có trong bộ dữ liệu, sau đó nhóm thực hiện kiểm tra lại hệ số tương quan pearsonr giữa các thuộc tính có trong bộ dữ liệu và nhận thấy có sự tương quan khá cao ở một số cặp thuộc tính, ví dụ: ‘ad_rooms’ và ‘ad_size’. Vì vậy nhóm quyết định sử dụng thêm 2 mô hình hồi quy là hồi quy Ridge [5] và hồi quy Lasso [6], Ridge và Lasso áp dụng hai phương pháp regularization L1 và L2 để giảm overfitting, giúp cải thiện hiệu suất của mô hình khi các biến đầu vào có tương quan cao. Nhóm cũng xây dựng thêm một số mô hình hồi quy như DecisionTreeRegressor [7], RandomForestRegressor [8], GradientBoostingRegressor [9], XGBRegressor [10] và CatBoostRegressor [11] nhằm xử lý các thuộc phân loại từ dữ liệu ban đầu, vì không chỉ có các thuộc tính số mà bên cạnh đó các thuộc tính phân loại cũng có thể chứa nhiều thông tin hữu ích giúp cho việc dự đoán biến mục tiêu được chính xác hơn.

Ngoài ra, việc loại bỏ dữ liệu nhiễu (noise) và outliers cũng đóng góp một phần quan trọng trong quá trình xây dựng mô hình hồi quy. Dữ liệu nhiễu và outliers có thể tạo ra sự biased (chệch) và không chính xác trong mô hình hồi quy. Ta có thể xem qua hình minh họa dưới đây để so sánh ảnh hưởng của các điểm dữ liệu nhiễu và outliers tới mô hình.



Hình 12. Ảnh hưởng của dữ liệu nhiễu và outliers đến mô hình hồi quy.

Qua khảo sát của nhóm, bộ dữ liệu gốc chứa khá nhiều thông tin nhiễu và outliers, đây là những điểm dữ liệu bị sai thông tin khá nghiêm trọng về giá và một số thuộc tính kiểu số khác, điều này có thể là do lỗi nhập liệu khách quan từ người dùng hoặc do người dùng cố tình nhập sai (nhằm bán được căn hộ với giá cao). Ngoài ra, nền tảng Chợ Tốt cũng không kiểm duyệt thông tin bài đăng từ người bán, vì vậy họ có thể tự chỉnh sửa thông tin trong bài đăng theo ý của mình.



Hình 13. Hình minh họa một điểm dữ liệu nhiễu trong bộ dữ liệu

4.2. Kết quả dự đoán của mô hình

Kết quả dự đoán của các mô hình trên tập thử nghiệm (test) được trình bày trong bảng dưới đây, nhóm sử dụng độ đo Mean Squared Error và R2 Score để đánh giá độ hiệu quả của mô hình, mỗi mô hình sẽ có 2 cột kết quả tương ứng với kết quả khi chưa loại bỏ dữ liệu nhiễu và đã loại bỏ dữ liệu nhiễu:

Model	Chưa loại bỏ nhiễu		Đã loại bỏ nhiễu	
	MSE	R2 Score	MSE	R2 Score
Linear Regression	2.837706e+07	0.145286	7.547920e+05	0.316503
Ridge Regression	2.837658e+07	0.145301	7.549477e+05	0.316362
Lasso Regression	2.837599e+07	0.145318	7.562183e+05	0.315212
DecisionTree Regressor	2.850025e+07	0.141575	7.249634e+05	0.343514

RandomForest Regressor	2.773615e+07	0.164590	5.123438e+05	0.536051
GradientBoosting Regressor	2.429517e+07	0.268232	5.036971e+05	0.543881
XGB Regressor	2.323104e+07	0.300283	4.015732e+05	0.636358
CatBoost Regressor	2.384278e+07	0.281858	3.736828e+05	0.661614

Hình 13. Kết quả các mô hình

4.3. Nhận xét và hướng phát triển

Bảng kết quả một lần nữa cho thấy sự khác biệt rõ rệt trong việc loại bỏ các điểm dữ liệu đối với sự chính xác của các mô hình hồi quy. Bên cạnh đó, ta cũng có thể nhận thấy sự chênh lệch khá lớn khi so sánh các độ chính xác của các mô hình hồi quy tuyến tính thông thường, chỉ xử lý các thuộc tính kiểu số như (Linear Regression, Ridge Regression, Lasso Regression) với các mô hình hồi quy còn lại như (DecisionTree Regressor, RandomForest Regressor, GradientBoosting Regressor, XGB Regressor, CatBoost Regressor), vì các mô hình này có khả năng xử lý các biến phân loại một cách hiệu quả. Nhờ tận dụng được giải thuật của cây quyết định, chúng có thể tạo các quy tắc phi tuyến tính và phức tạp hơn, giúp các mô hình này phù hợp hơn với bộ dữ liệu gồm cả thuộc tính phân loại và thuộc tính số.

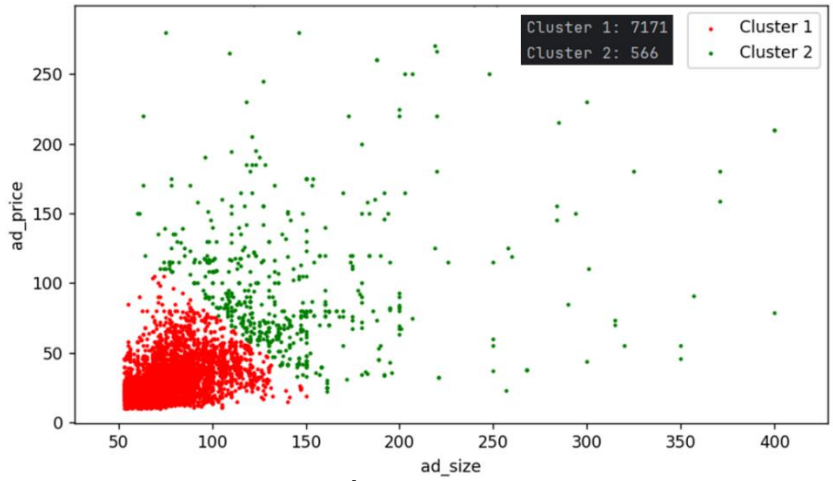
Trong quá trình chạy mô hình, nhóm đã lần lượt chọn ra ngẫu nhiên 2000, 5000 và 8000 mẫu từ bộ dữ liệu, sau đó chạy thử nghiệm để thu về kết quả và có nhận xét sau: Khi số mẫu trong bộ dữ liệu tăng, kết quả các mô hình cũng lần lượt tăng theo, khoảng 5% R2 Score được tăng thêm tương ứng với mỗi 3000 mẫu cộng thêm. Vì vậy hướng phát triển tiếp theo của nhóm là tiếp tục thu thập thêm nhiều mẫu dữ liệu hơn tại thành phố Hồ Chí Minh và cũng có thể là các tỉnh thành lân cận khác để cải thiện thêm độ chính xác cho các mô hình hồi quy.

5. PHÂN CỤM DỮ LIỆU SỬ DỤNG K-MEANS CLUSTERING

Phân cụm dữ liệu trong bộ dữ liệu có thể giúp chúng ta hiểu rõ hơn về các đặc trưng và mối quan hệ giữa các điểm dữ liệu, đồng thời giúp ta xác định được các nhóm dữ liệu có đặc điểm tương tự và nhận biết các yếu tố quan trọng ảnh hưởng đến giá căn hộ trong từng phân khúc thị trường cụ thể. Trong bài nghiên cứu này nhóm sử dụng thuật toán phân cụm K-means Clustering [12] để nhóm các căn hộ có đặc điểm tương tự vào cùng một nhóm, 2 thuộc tính được sử dụng để gom nhóm là diện tích căn hộ (ad_size) và giá của căn hộ (ad_price).

5.1. Kết quả phân cụm và nhận xét

Sử dụng phương pháp Elbow [13] để phân tích và đánh giá trong quá trình thực nghiệm, nhóm nhận thấy tại K=2, giá trị của tổng bình phương khoảng cách giảm đột ngột và việc thêm cụm mới không còn cải thiện khả năng phân cụm của thuật toán nhiều hơn nữa, vì vậy nhóm quyết định chọn 2 làm giá trị khởi tạo cho K (số lượng cụm). Kết quả phân cụm của thuật toán K-means Clustering được thể hiện trong hình dưới đây:



Hình 14. Kết quả phân cụm.

Nhóm đã thực hiện scale lại miền giá trị của cột ad_price theo tỷ lệ $10^8 \div 1$ vì 2 thuộc tính ở trên có độ đo khác nhau (VND và m^2) và phân phối cũng khác nhau, việc scale giúp cho hàm tính tổng bình phương không bị chi phối quá nhiều từ sự biến thiên của các giá trị trong cột ad_price.

5.1. Nhận xét

Từ kết quả phân cụm trên nhóm rút ra được một số nhận xét sau: các căn hộ được phân vào 2 cụm chính, gồm có 7171 căn hộ thuộc cụm thứ nhất (cụm màu đỏ) và 566 căn hộ thuộc cụm thứ hai (cụm màu xanh). Phần lớn các căn hộ có giá nằm trong khoảng từ 3-5 tỷ VND và có diện tích từ 50-100 m^2, khi diện tích dần tiến ra ngoài 100 m^2, các căn hộ có giá từ 3-5 tỷ VND dần ít đi và xuất hiện khá nhiền các căn hộ có giá từ 5-10 tỷ VND. Thông qua phân tích này, chúng ta có thể nhận thức được sự phân chia khá rõ ràng giữa hai nhóm căn hộ dựa trên giá và diện tích, cũng như chỉ ra sự tăng giá sẽ tương ứng với diện tích lớn hơn.

6. KẾT LUẬN

Trong quá trình nghiên cứu, nhóm đã đạt được những kết quả rất khả quan, đáp ứng những tiêu đã đề ra ban đầu. Nhóm đã tự thu thập bộ dữ liệu, tìm hiểu và thực hiện được các phương pháp tiền xử lý, phân tích trực quan, thực hiện các phép thực nghiệm cũng như đã xây dựng được các mô hình máy học áp dụng bộ dữ liệu đã xây dựng. Ngoài việc giúp mọi người tiếp được căn hộ có giá tốt nhất, cũng như dự đoán giá căn hộ thì tính ứng dụng của mô hình không chỉ giới hạn trong việc hỗ trợ quyết định mua bán bất động sản mà còn có thể tạo ra những ứng dụng và dịch vụ mới, làm giàu hơn nguồn thông tin cho người dùng và thúc đẩy sự phát triển bền vững của thị trường bất động sản.

Chúng tôi nhận thức về sự quan trọng của sự tiếp tục học hỏi và cải thiện, và mong muốn được đóng góp vào cộng đồng nghiên cứu và phát triển. Chúng tôi hy vọng rằng bài báo cáo này không chỉ là một sản phẩm của quá trình học tập, mà còn là sự bắt đầu cho những hướng đi mới và những thách thức mới trong tương lai. Bằng cách này, chúng tôi mong muốn rằng công trình của chúng tôi có thể làm phong phú thêm kiến thức trong cộng đồng sinh viên và góp phần vào sự phát triển của lĩnh vực khoa học dữ liệu và máy học.

TÀI LIỆU THAM KHẢO

[1] Website Nhà Tốt. Link: https://www.nhatot.com/mua-ban-can-ho-chung-cu-tp-ho-chi-minh (Ngày truy cập: 14/11/2023).

[2] Website Chợ Tốt. Link: Chợ Tốt - Website Mua Bán, Rao Vặt Trực Tuyến Hàng Đầu Của Người Việt (chotot.com)

[3] Developer Tool. Link: https://learn.microsoft.com/en-us/microsoft-edge/devtools-guide-chromium/landing/.

[4] Thư viện sklearn. Link: https://scikit-learn.org/ .

[5] Ridge Regression. Link: sklearn.linear_model.Ridge — scikit-learn 1.3.2 documentation

[6] Lasso Regression. Link: sklearn.linear_model.Lasso — scikit-learn 1.3.2 documentation

[7] DecisionTreeRegressor. Link: sklearn.tree.DecisionTreeRegressor — scikit-learn 1.3.2 documentation

[8] RandomForestRegressor. Link: sklearn.ensemble.RandomForestRegressor — scikit-learn 1.3.2 documentation

[9] GradientBoostingRegressor.Link: sklearn.ensemble.GradientBoostingRegressor — scikit-learn 1.3.2 documentation

[10] XGBRegressor. Link: https://xgboost.readthedocs.io/en/stable/python/ .

[11] CatBoostRegressor. Link: Overview - CatBoostRegressor | CatBoost .

[12] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, *8*, 80716-80727.

[13] Cui, M. (2020). Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, *1*(1), 5-8.

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Ngọc Lương	Phân tích tổng quan và tiền xử lý trên bộ dữ liệu, tổng hợp code
2	Nguyễn Thiện Trí	Thu thập dữ liệu, nghiên cứu xây dựng mô hình hồi quy, đánh giá mô hình, phân tích cụm dữ liệu.
3	Đoàn Bảo Long	Phân tích bộ dữ liệu sau khi được tiền xử lí, thực hiện các thực nghiệm giúp để tìm ra được những căn hộ có giá tốt và đưa những dữ liệu này lên Dashboard.