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 và phân loại đất nền ở Đà Nẵng

Nhóm	12
Họ Và Tên Sinh Viên	Lớp Học Phần
Nguyễn Quốc Cường	21N12
Võ Đức Việt	

ĐÀ NẴNG, 05/2024

TÓM TẮT

Đề tài "Dự đoán và phân loại giá đất nền ở Đà Nẵng" giải quyết vấn đề biến động giá đất và sự thiếu hụt thông tin chính xác, gây rủi ro khi quyết định đầu tư. Nghiên cứu này bắt đầu bằng việc thu thập và xử lý các dữ liệu từ nhiều nguồn khác nhau. Dữ liệu sau đó được phân tích bằng các công cụ thống kê và thuật toán như hồi quy tuyến tính, phân cụm K-means, phân lớp để xây dựng mô hình dự đoán và phân loại giá đất. Kết quả đạt được cho thấy mô hình dự đoán đạt độ chính xác cao, giúp dự đoán giá đất một cách hiệu quả. Các giá được phân loại rõ ràng, tạo điều kiện cho việc so sánh và đánh giá giá trị đất nền.

BẢNG PHÂN CÔNG NHIỆM VỤ

Sinh viên thực hiện	Các nhiệm vụ	Tự đánh giá theo 3 mức (Đã hoàn thành/Chưa hoàn thành/Không triển khai)
Nguyễn Quốc Cường	- Crawl Data - Data Visualization - Data Classification - Viết báo cáo	- Đã hoàn thành - Đã hoàn thành - Đã hoàn thành - Đã hoàn thành
Võ Đức Việt	- Clean Data -Data Visualization -Data Clustering - Viết báo cáo	- Đã hoàn thành - Đã hoàn thành - Đã hoàn thành - Đã hoàn thành

MỤC LỤC

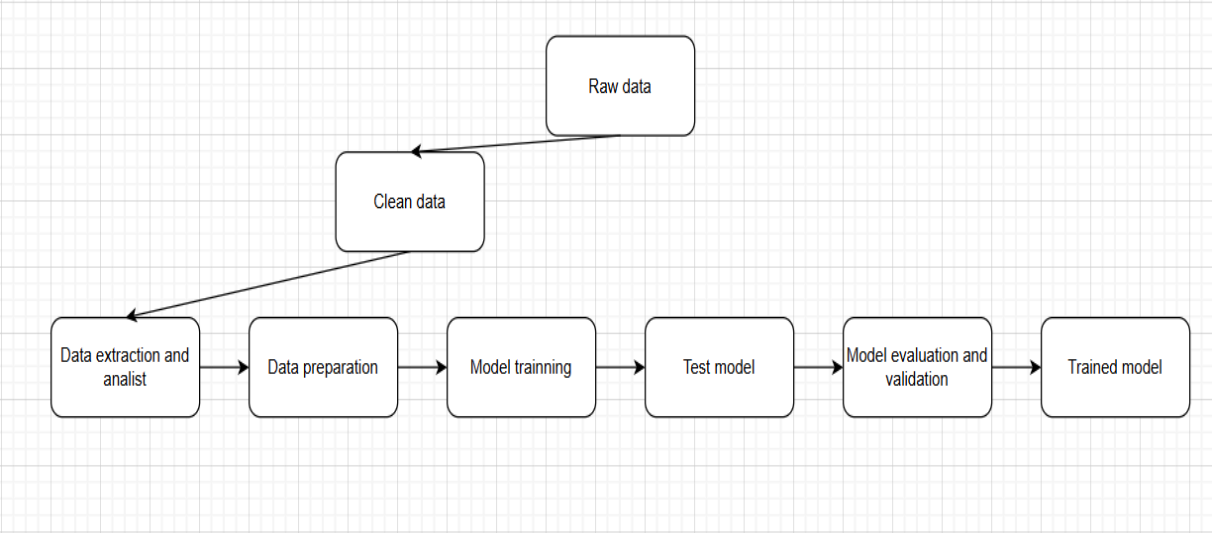
DANH MỤC HÌNH VẼ

1. Giới thiệu

1.1 Giới thiệu các bài toán.

- **Mục tiêu:** xây dựng mô hình dự đoán giá và phân loại đất nền ở Đà Nẵng dựa trên các yếu tố như giá, diện tích, loại hình đất, hướng đất, chiều ngang, chiều rộng, quận.
- **Dự đoán giá đất nền:**
    - **Giải pháp**
        - Thu thập dữ liệu
        - Xử lý và làm sạch dữ liệu.
        - Xây dựng các mô hình và tối ưu mô hình.
- **Phân loại đất nền**
    - **Giải pháp**
        - Thu thập dữ liệu
        - Xử lý và làm sạch dữ liệu.
        - Phân tích dữ liệu.
        - Phân cụm dữ liệu.
        - Phân loại dữ liệu dựa trên kết quả phân cụm.

1.2 Sơ đồ khối:



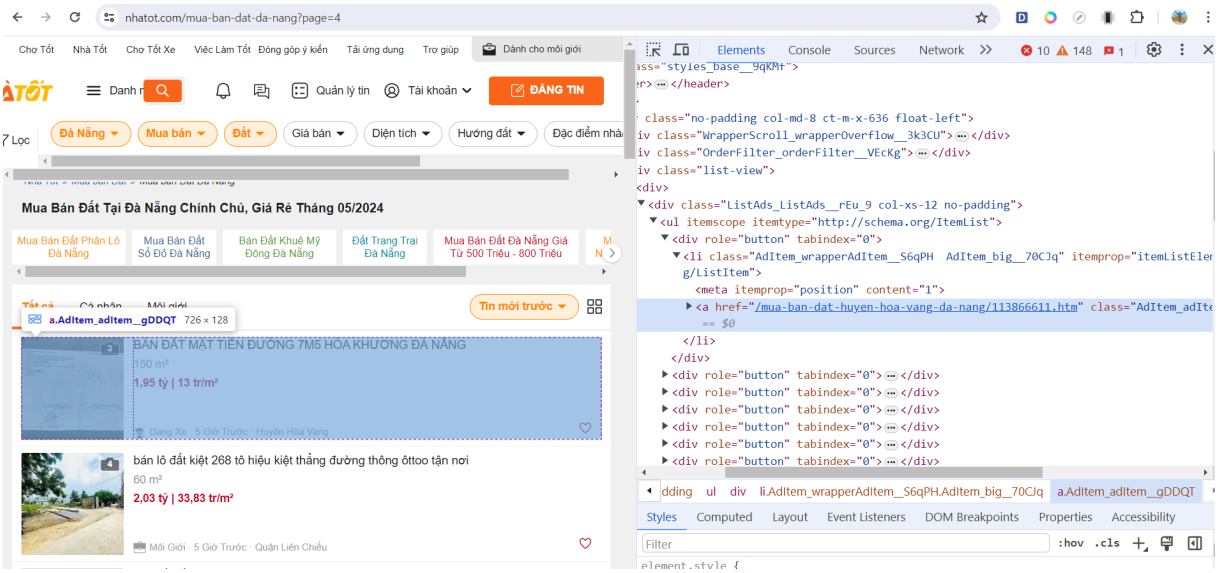
Hình 1: Giải pháp tổng quan về dữ liệu dưới dạng sơ đồ khối

2. Thu thập và mô tả dữ liệu

2.1. Thu thập dữ liệu

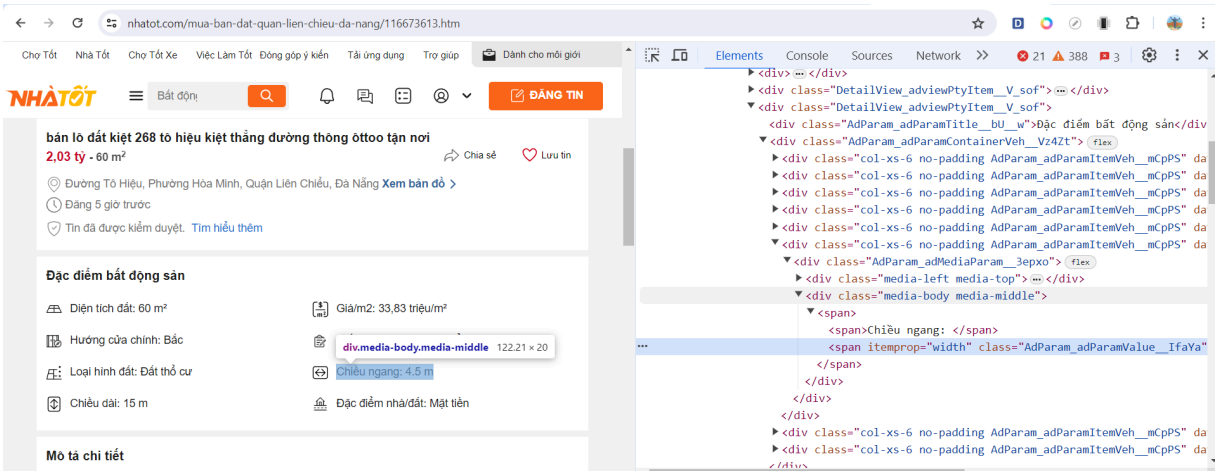
- **Giải pháp thu thập dữ liệu**
    - Nguồn dữ liệu:
        - train:
            - trang https://www.nhatot.com
            - trang https://nhadat24h.net
        - test:
            - trang https://www.nhatot.com
    - Công cụ thu thập:
        - BeautifulSoup và Selenium
    - Cách thức sử dụng công cụ
        - Sử dụng Selenium để thu thập liên kết các bài đăng từ danh sách trang.
        - Sử dụng BeautifulSoup và Selenium để thu thập thông tin chi tiết từ từng liên kết bài đăng.
    - Đầu vào
        - URL của trang web cần thu thập dữ liệu
        - Số lượng trang cần duyệt
        - Các trường thông tin cần thu thập: Bao gồm giá, địa chỉ, diện tích, giá/m², hướng đất, loại hình đất, chiều ngang, chiều dài.
    - Đầu ra
        - Danh sách các liên kết bài đăng: Được lưu trữ trong file CSV (link_nhatot_test.csv) chứa các liên kết đến từng bài đăng chi tiết.
        - Thông tin chi tiết của từng bài đăng: Được lưu trữ trong file CSV (data_crawl_test.csv) chứa các thông tin như giá, địa chỉ, diện tích, giá/m², hướng đất, loại hình đất, chiều ngang, chiều dài.

- Ví dụ minh họa



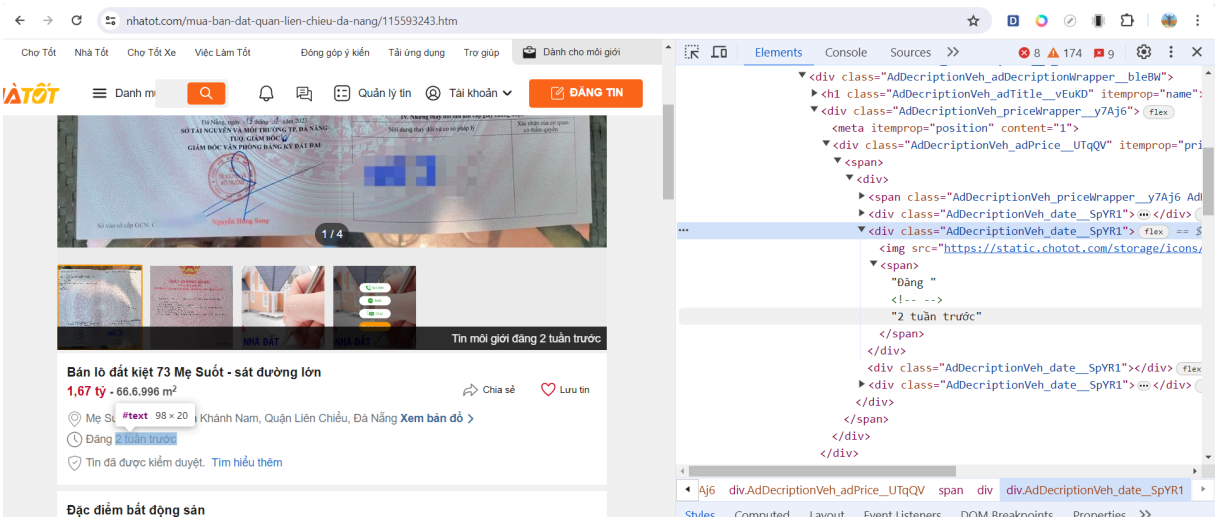
Hình 2: Hình ảnh minh họa cách lấy đường dẫn

- Với mỗi trang page ta sẽ lấy được đường dẫn các đường link vào các bài viết cụ thể, đường link được lưu vào file link_nhatot.csv hoặc link_nhatot_test.csv



Hình 3: Hình ảnh minh họa cách lấy thông tin

- Với mỗi bài đăng cụ thể ta thu được các thông tin cần thu thập được lưu vào file data_crawl_nhatot.csv hoặc data_crawl_test.csv



Hình 4: Hình ảnh minh họa cách dừng thu thập dữ liệu

- Điều kiện để dừng thu nhập dữ liệu tập test là thời gian đăng bài là 2 tuần trước, ví dụ này là link thứ 1588/1704 trong file link_nhatot_test.csv

```
1     ,links
1588  1586,https://nhatot.com//mua-ban-dat-quan-ngu-hanh-son-da-nang/115150522.htm
1589  1587,https://nhatot.com//mua-ban-dat-quan-ngu-hanh-son-da-nang/116397061.htm
1590  1588,https://nhatot.com//mua-ban-dat-quan-lien-chieu-da-nang/115593243.htm
1591  1589,https://nhatot.com//mua-ban-dat-quan-hai-chau-da-nang/116396454.htm
1592  1590,https://nhatot.com//mua-ban-dat-quan-ngu-hanh-son-da-nang/116395994.htm
```

```
1587  5 tỷ - 126 m2,"Đường Hồ Sĩ Tân, Phường Nại Hiên Đông, Quận Sơn Trà
1588  "4,8 tỷ - 100 m2","Đường Mỹ Đa Tây 5, Phường Khuê Mỹ, Quận Ngũ Hàn
1589  "4,4 tỷ - 62 m2","Đường Tôn Thất Thiệp, Phường Mỹ An, Quận Ngũ Hàn
1590
```

Hình 5: Hình ảnh dẫn chứng cụ thể trong data

2.2. Mô tả và trực quan hoá dữ liệu

Tổng quan về tập dữ liệu:

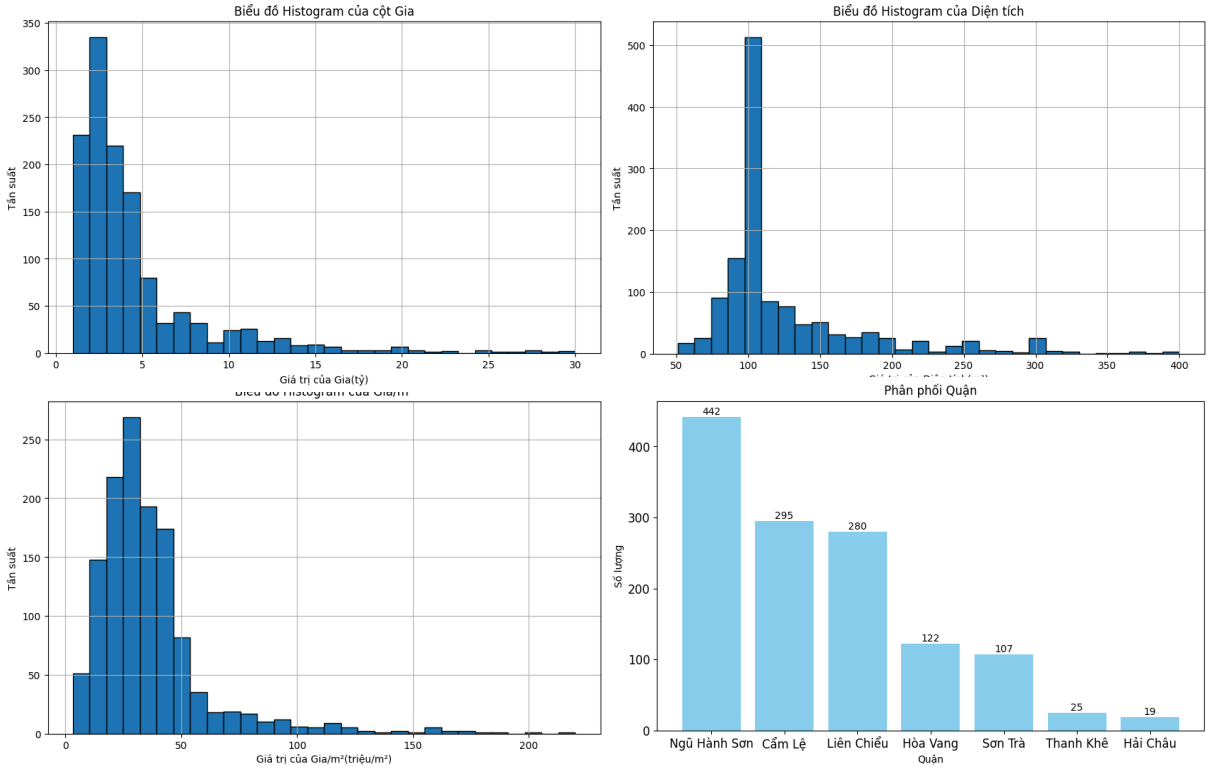
- **Tập huấn luyện:**

Số mẫu dữ liệu	Số đặc trưng
1290	9

Bảng 1: Số mẫu dữ liệu, đặc trưng của tập huấn luyện

STT	Tên cột	Số mẫu trống	Kiểu dữ liệu
1	Gia	0	float64
2	Dia chi	0	object
3	Dien tich	0	float64
4	Gia/m²	0	float64
5	Huong dat	334	object
6	Loai hinh dat	0	object
7	Chieu ngang	0	float64
8	Chieu dai	0	float64
9	Quan	0	object

Bảng 2: tên cột, số mẫu trống, kiểu dữ liệu của tập huấn luyện.



Hình 6: Biểu đồ các biến mục tiêu và biến quan trọng của tập huấn luyện.

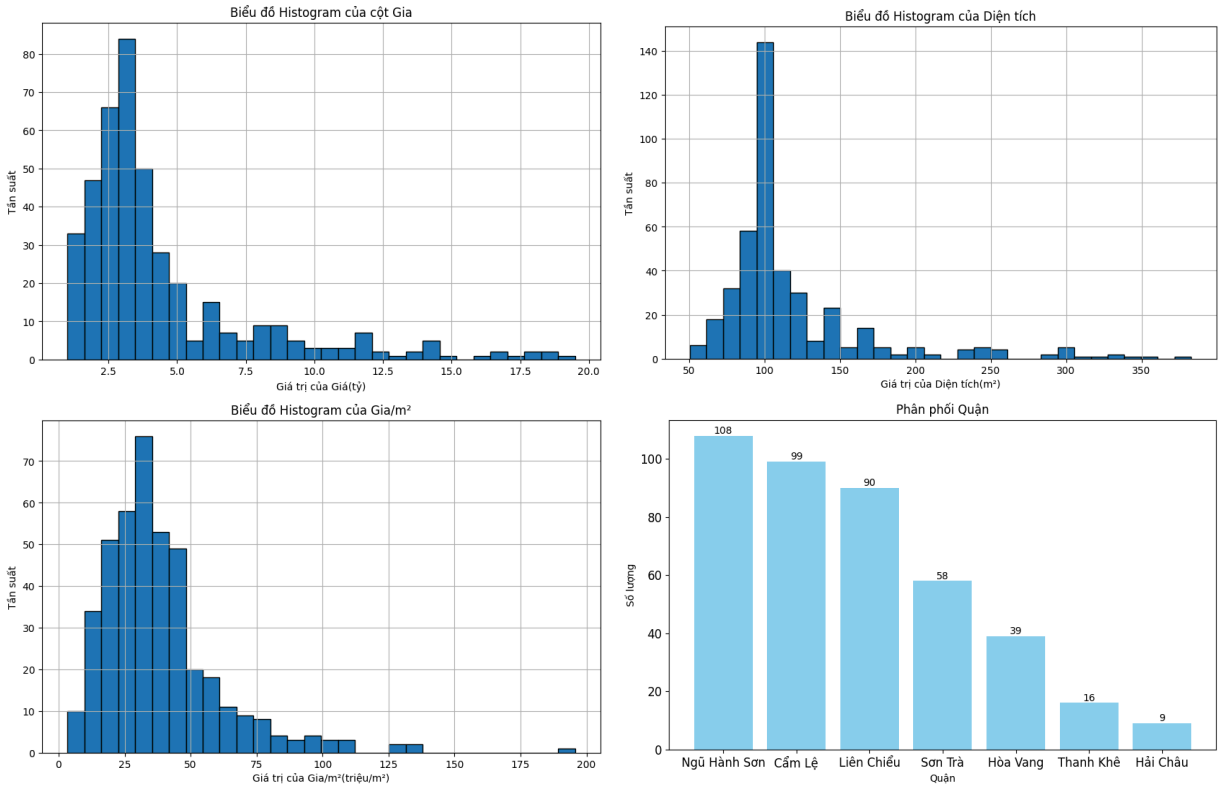
- **Tập kiểm thử:**

Số mẫu dữ liệu	**Số đặc trưng**
419	9

Bảng 3: Số mẫu dữ liệu, đặc trưng của tập kiểm thử

STT	**Tên cột**	**Số mẫu trống**	**Kiểu dữ liệu**
1	Gia	0	float64
2	Dia chi	0	object
3	Dien tich	0	float64
4	Gia/m²	0	float64
5	Huong dat	77	object
6	Loai hinh dat	0	object
7	Chieu ngang	0	float64
8	Chieu dai	0	float64
9	Quan	0	object

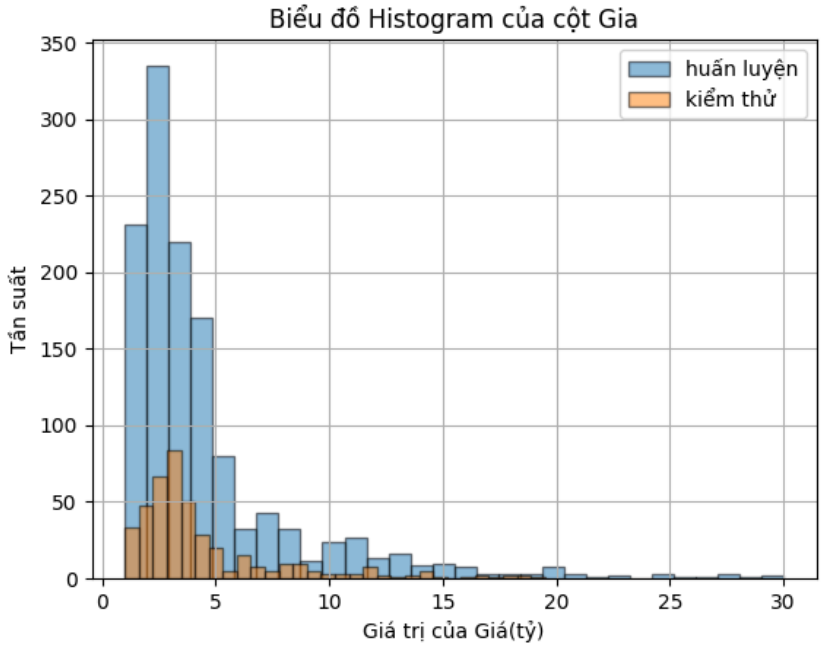
Bảng 4: tên cột, số mẫu trống, kiểu dữ liệu của tập kiểm thử.



Hình 7: Biểu đồ các biến mục tiêu và biến quan trọng của tập kiểm thử.

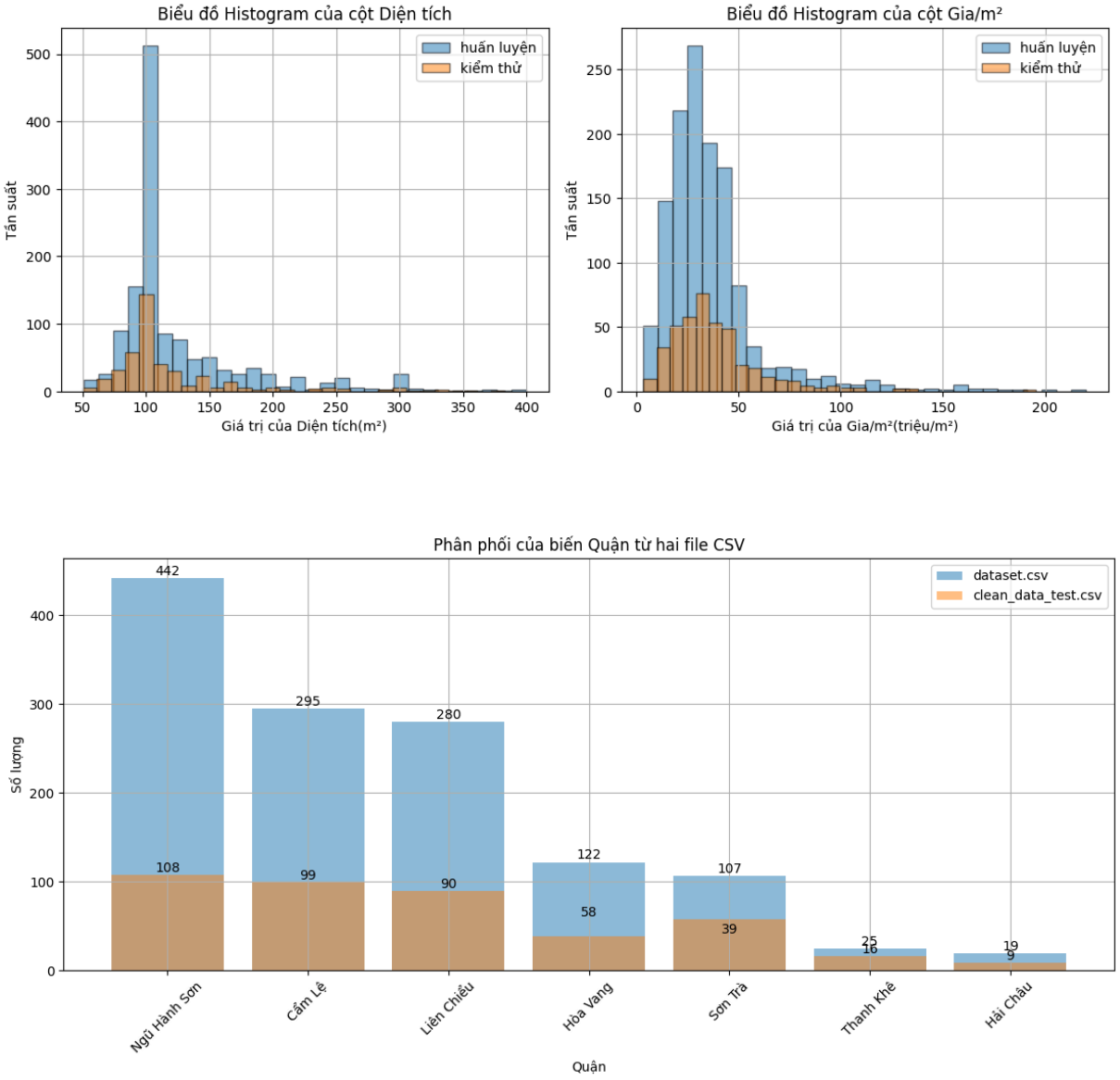
So sánh giữa tập huấn luyện và tập kiểm thử:

- Biến mục tiêu:



Hình 8: Biểu đồ so sánh biến mục tiêu của hai tập dữ liệu.

- Biến quan trọng:



Hình 9: Biểu đồ so sánh biến quan trọng của hai tập dữ liệu.

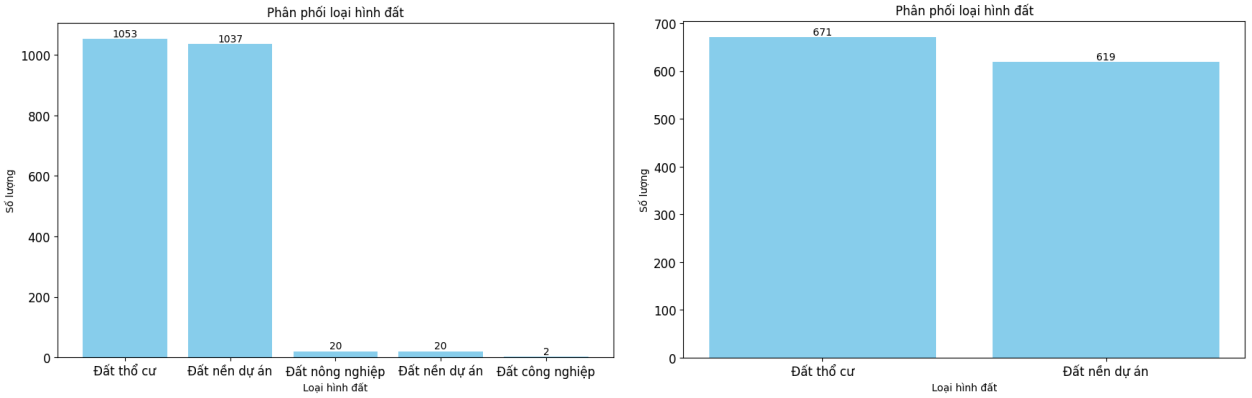
- **nhận xét:** Phân bố dữ liệu của biến mục tiêu và biến quan trọng trên cả tập huấn luyện và tập kiểm thử đều khá đồng đều.

3. Trích xuất đặc trưng

3.1 Làm sạch dữ liệu:

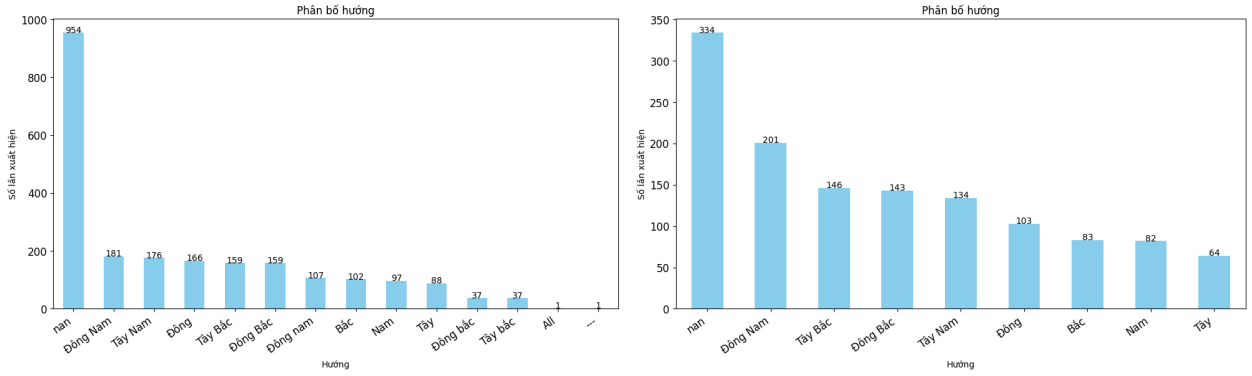
- **Giá:**
  - Loại bỏ chuỗi tỷ và triệu, giá trị có thỏa thuận.
  - Thay đổi dấu phẩy thành dấu chấm.
  - Chuyển đổi giá trị của triệu sang giá trị tỷ.
  - Lấy giá trị trước dấu gạch ngang và lấy từ 1 đến 30 tỷ.
- **Địa chỉ:**
  - Biến đổi chuỗi thành chữ cái đầu tiên của mỗi từ là in hoa, các chữ cái còn lại in thường.
  - Loại bỏ các dữ liệu có kí tự đặc biệt ở đầu chuỗi, xóa khoảng trắng và xóa chuỗi" xem biểu đồ".
- **Diện tích:**
  - Loại bỏ các chuỗi m^2 và m2.
  - Lấy giá trị bé hơn bằng 400.
- **Giá trên m^2:**
  - Loại bỏ các chuỗi "triệu/m^2", "triệu/m2", "đ/m^2".
  - Tính toán lại giá trị và làm tròn.
- **Hướng đất:**
  - Loại bỏ các chuỗi "Hướng:".
  - Biến đổi chuỗi thành chữ cái đầu tiên của mỗi từ là in hoa, các chữ cái còn lại in thường.
- **Loại hình đất:**
  - Loại bỏ dữ liệu của Đất công nghiệp.
- **Chiều ngang:**
  - Loại bỏ các chuỗi "m", "Mặt tiền:".
  - Lấy giá trị từ 4 đến 20.
  - Làm tròn giá trị.
- **Chiều dài:**
  - Loại bỏ các chuỗi "m".
  - Thay các giá trị trống bằng cách lấy diện tích chia cho chiều ngang.
  - Lấy giá trị từ 10 đến 40.
  - Làm tròn giá trị.

- **Quận:**
    - Thêm dữ liệu mới bằng cách lấy dữ liệu từ cột Địa chỉ, lấy giá trị của quận hoặc huyện.
    - Nếu có nhiều hơn 1 dấu phẩy thì loại bỏ các giá trị sau dấu phẩy thứ nhất.
- **Trực quan dữ liệu trước và sau khi khi làm sạch:**



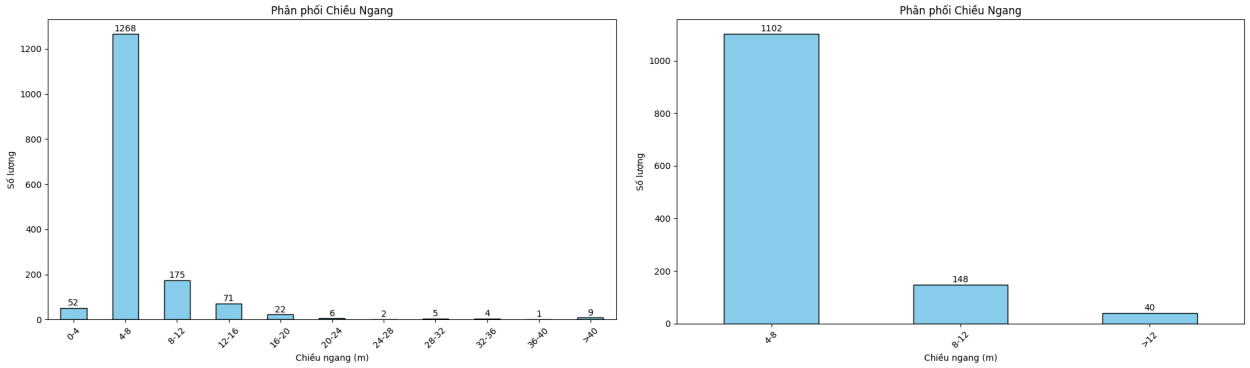
Hình 10: Biểu đồ phân loại hình đất của trước và sau khi làm sạch.

- **Nhận xét:** các giá trị sau khi làm làm sạch sẽ loại bỏ bớt các giá trị xuất hiện ít lần, chỉ tập trung vào hai giá trị nhiều.



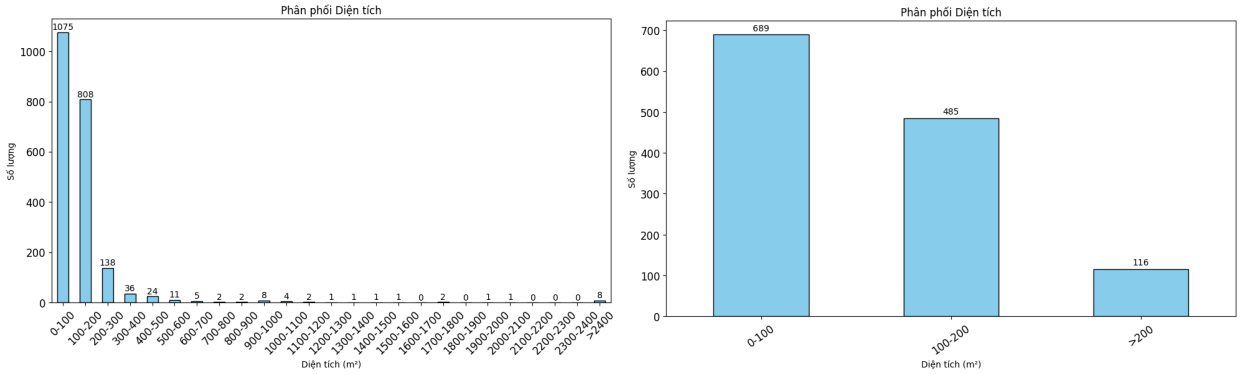
Hình 11: Biểu đồ phân loại hướng của trước và sau khi làm sạch.

- **Nhận xét:** các chuỗi được làm đồng nhất một phông chữ và bỏ những ngoại lệ.



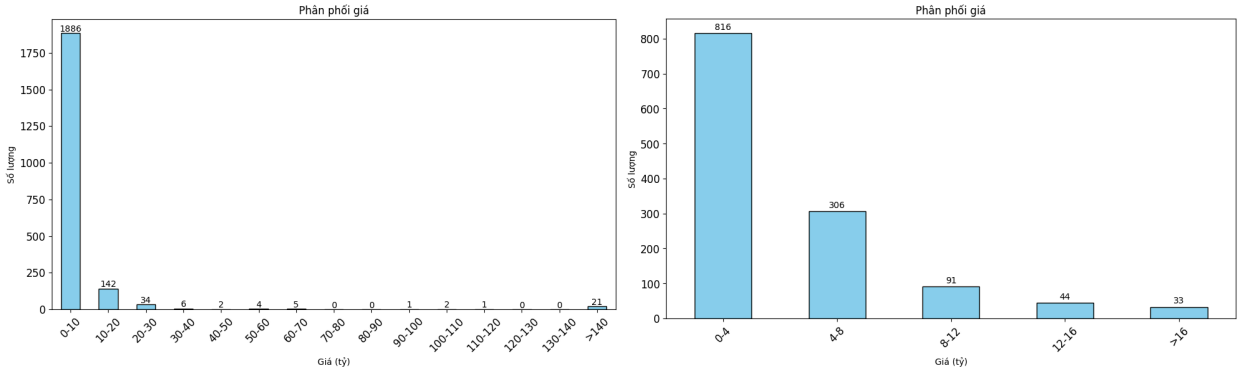
Hình 12: Biểu đồ phân loại chiều ngang của trước và sau khi làm sạch.

- **Nhận xét:** dữ liệu được phân bố hẹp lại, loại bỏ những ngoại lệ, loại bỏ những giá trị lớn hơn 20.



Hình 13: Biểu đồ phân loại diện tích của trước và sau khi làm sạch.

- **Nhận xét:** dữ liệu được phân bố hẹp lại, loại bỏ những ngoại lệ.



Hình 14: Biểu đồ phân loại giá của trước và sau khi làm sạch.

- **Nhận xét:** dữ liệu được phân bố hẹp lại, loại bỏ những ngoại lệ.

3.2 Mã hóa dữ liệu:

Target Encoding:

- Là một phương pháp mã hóa biến categorical dựa trên giá trị trung bình của biến mục tiêu (target variable) tương ứng với từng giá trị trong biến categorical. Phương pháp này có thể cải thiện hiệu suất mô hình trong trường hợp biến categorical có mối quan hệ mạnh với biến mục tiêu

StandardScaler:

- Là một kỹ thuật được sử dụng trong việc chuẩn hóa dữ liệu trước khi đưa vào các thuật toán học máy.
- Các bước thực hiện:
    - Bước 1: Tính toán giá trị trung bình của từng thuộc tính (cột) trong dữ liệu đầu vào.
    - Bước 2: Tính toán độ lệch chuẩn của mỗi thuộc tính.
    - Bước 3: Chuẩn hóa mỗi thuộc tính của dữ liệu đầu vào bằng công thức

$$Z = (X - mean) / std$$

Trong đó:

- Z là giá trị chuẩn hóa.
- X là giá trị của một thuộc tính bất kỳ.
- mean là giá trị trung bình của thuộc tính đó.
- std là độ lệch chuẩn của thuộc tính đó.

- Kết quả:
    - Trước khi chưa mã hóa:

Gia	Dien tich	Gia/m²	Chieu ngang	Chieu dai	Huong dat	Loai hinh dat	Quan
1.59	70	22.71	5.1	14	Tây Bắc	Đất thổ cư	Liên Chiểu
2.96	105	28.19	5	21	Tây Bắc	Đất nền dự án	Ngũ Hành Sơn
2.85	111	25.68	6	19	Đông Nam	Đất thổ cư	Ngũ Hành Sơn
2.15	168	12.8	7	24	Nam	Đất thổ cư	Liên Chiểu
10	100	100	5	20	Đông Nam	Đất thổ cư	Ngũ Hành Sơn

Bảng 5: Bảng dữ liệu trước khi dùng StandardScaler.

- Sau khi mã hóa:

Gia	Dien tich	Gia/m²	Chieu ngang	Chieu dai	Huong dat	Loai hinh dat	Quan
1.59	-0.9927	-0.5995	-0.4455	-1.5125	-0.3242	0.5536	-0.6870
2.96	-0.3036	-0.3617	-0.4949	0.2410	-0.3242	-1.8064	0.0107
2.85	-0.1855	-0.4706	-0.0014	-0.2600	0.4766	0.5536	0.0170
2.15	0.9365	-1.0295	0.4921	0.9935	-0.9257	0.5536	-0.6870
10	-0.4021	2.7547	-0.4949	-0.0095	0.4766	0.5536	0.0170

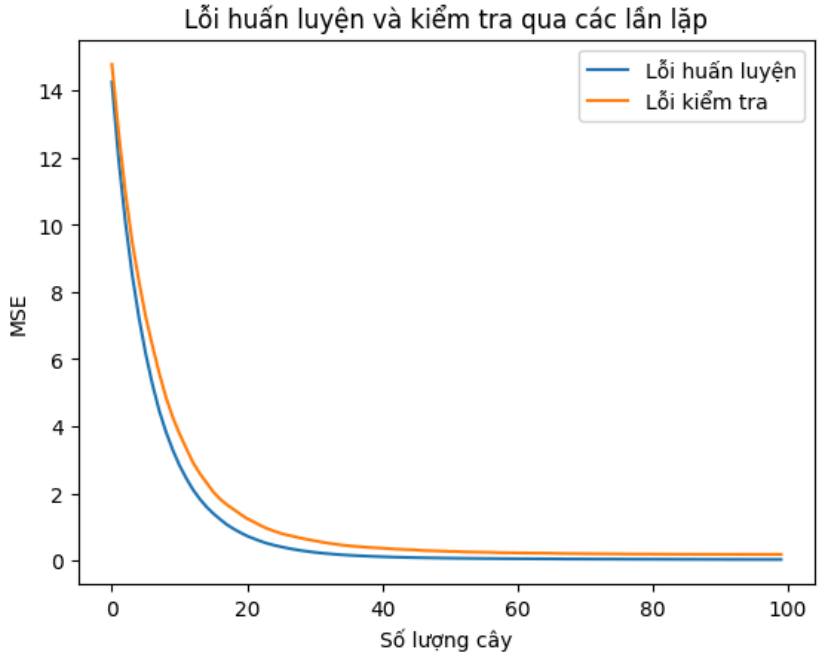
Bảng 6: Bảng dữ liệu sau khi dùng StandardScaler.

4. Mô hình hóa dữ liệu

4.1 Mô hình Gradient Boosting Regression

- Gradient boosting là một thuật toán học máy mạnh mẽ, kết hợp nhiều mô hình yếu thành một mô hình mạnh hơn. Trong mỗi lần lặp, thuật toán tính toán độ dốc của hàm mất mát theo dự đoán của mô hình hiện tại, sau đó huấn luyện một mô hình yếu mới để giảm thiểu độ dốc này. Các dự đoán của mô hình mới sau đó được thêm vào tập hợp và quá trình này được lặp lại cho đến khi đáp ứng tiêu chí dừng, như đạt được số lượng lặp tối đa hoặc không có sự cải thiện đáng kể trên tập kiểm tra.
- Bộ tham số:
    - random_state: được sử dụng để kiểm soát hạt giống ngẫu nhiên được cung cấp cho mỗi công cụ ước tính cây Gradient Boosting tại mỗi tăng cường lặp lại trong quá trình huấn luyện.
    - n_estimators: Số lượng cây trong mô hình Gradient Boosting. Càng nhiều cây thì mô hình càng phức tạp và có khả năng tổng hợp tốt hơn, nhưng cũng có thể dẫn đến overfitting.
    - learning_rate: Tốc độ học của mô hình, tức là mức độ cập nhật trọng số của mỗi cây trong quá trình huấn luyện
    - max_depth: Độ sâu tối đa của mỗi cây. Càng sâu, mô hình càng phức tạp và có khả năng học được nhiều mẫu dữ liệu hơn, nhưng cũng có thể dẫn đến overfitting.
    - validation_fraction: Phần trăm dữ liệu được sử dụng để tạo tập validation. Tập validation được sử dụng để đánh giá hiệu suất của mô hình trong quá trình huấn luyện và sử dụng để chọn các siêu tham số tốt nhất.
    - n_iter_no_change: Số lượng vòng lặp mà không có cải thiện đáng kể trên tập validation trước khi dừng lại.
    - Tol: Ngưỡng dừng sớm, tức là giá trị mức độ dừng của lỗi. Nếu không có cải thiện đáng kể trên tập validation sau mỗi vòng lặp, huấn luyện sẽ dừng sớm nếu lỗi giảm dưới ngưỡng này.
- Chia tập huấn luyện và tập kiểm thử theo tỉ lệ 80: 20

- **Quá trình huấn luyện:**



Hình 15: Hình ảnh so sánh lỗi kiểm tra và lỗi huấn luyện trên tập dữ liệu huấn luyện

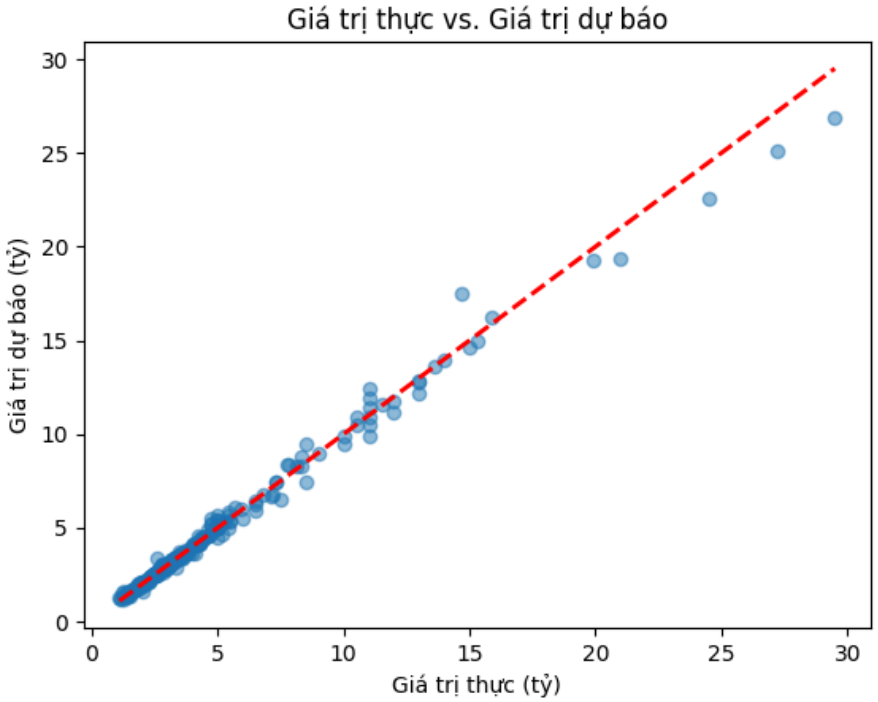
- **Nhận xét:**Lỗi của tập kiểm tra cao hơn tập huấn luyện trên mọi giá trị số lượng

	Train	Test
MAE	0.1072	0.1952
MSE	0.0334	0.1717
RMSE	0.1826	0.4143
R^2	0.9980	0.9899

Bảng: Bảng thể hiện các chỉ số khi huấn luyện mô hình

- Nhận xét:

- MAE của tập huấn luyện là 0.1072, thấp hơn đáng kể so với MAE của tập kiểm tra là 0.1952
- MSE của tập huấn luyện là 0.0334, rất thấp so với của tập kiểm tra là 0.1717.
- Giá trị R^2 của tập huấn luyện là 0.998. R^2 của tập kiểm tra là 0.9899, gần với giá trị hoàn hảo 1



	Gia du doan	Gia thuc	% sai lech
878	1.479976	1.48	0.0016
768	1.349840	1.35	0.0119
344	2.250515	2.25	0.0229
655	2.451150	2.45	0.0470
641	3.497870	3.50	0.0609
...	...	...	...
225	17.520328	14.70	19.1859
368	1.598508	2.00	20.0746
846	1.504245	1.20	25.3538
892	1.588743	1.25	27.0994
1215	3.377514	2.60	29.9044

```
count    258.000000
mean       3.874496
std        4.501251
min        0.001600
25%        0.860475
50%        2.145600
75%        5.614975
max       29.904400
```

Hình 16: Hình ảnh dự đoán giá của mô hình Gradient Boosting Regression trên tập dữ liệu huấn luyện

- **Nhận xét:** Giá trị dự đoán tập trung quanh đường chéo 45 độ. Giá càng tăng khả năng rời xa đường chéo càng lớn.
- Giá trị sai lệch trung bình 0.414 triệu

- **Kiểm tra mô hình bằng kỹ thuật multiple test sets**

	MAE	MSE	RMSE	R^2
Train	0.1183	0.0442	0.2102	0.9975
Split 1	0.1879	0.1908	0.4368	0.9919
Split 2	03002	0.3295	0.574	0.9853
Split 3	0.1915	0.1195	0.3457	0.9859
Split 4	0.1823	0.0998	0.3159	0.9878
Split 5	0.2399	0.3232	0.5685	0.9707

Bảng 8: Bảng kết quả các chỉ số kiểm tra mô hình bằng kỹ thuật multiple test sets.

- Nhận xét:
  - R^2 vẫn duy trì ở mức cao, tuy nhiên các chỉ số còn lại khá cao, cho thấy mô hình có thể dự đoán sai lệch trên các giá trị mới.

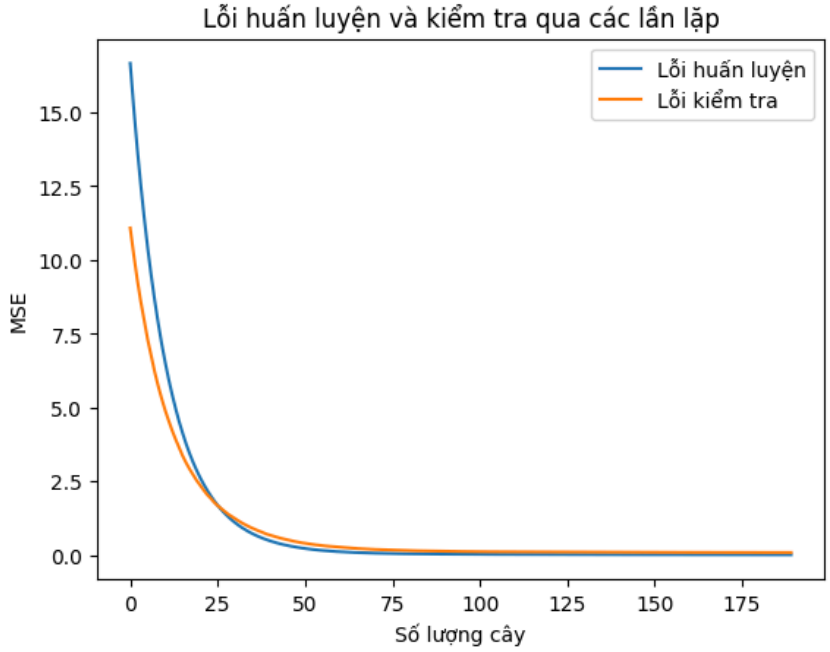
- **Hiệu chỉnh mô hinh**
  - Sử dụng phương pháp GridSearchCV để tìm bộ tham số tối ưu nhất. Kết quả thu được "random_state=2345,learning_rate=0.05, ,n_estimators=1000 max_depth=4, n_iter_no_change=20, tol=0.0001, validation_fraction=0.2"

- **Kết quả sau hiệu chỉnh**

	Train	Test
MAE	0.0954	0.1300
MSE	0.1097	0.0874
RMSE	0.3312	0.2956
R^2	0.9940	0.9921

Bảng: Bảng thể hiện các chỉ số khi huấn luyện mô hình sau hiệu chỉnh

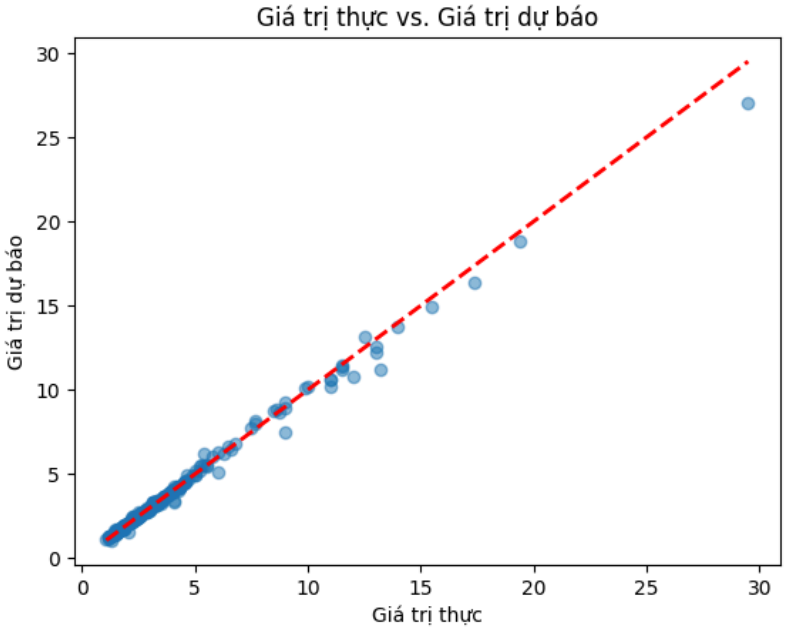
- Số vòng lặp trước khi dừng sớm: 190



-

Hình 17: Hình ảnh so sánh lỗi kiểm tra và lỗi huấn luyện trên tập dữ liệu huấn luyện sau hiệu chỉnh

- Nhận xét: Giá trị lỗi của kiểm tra gần sát với huấn luyện hơn

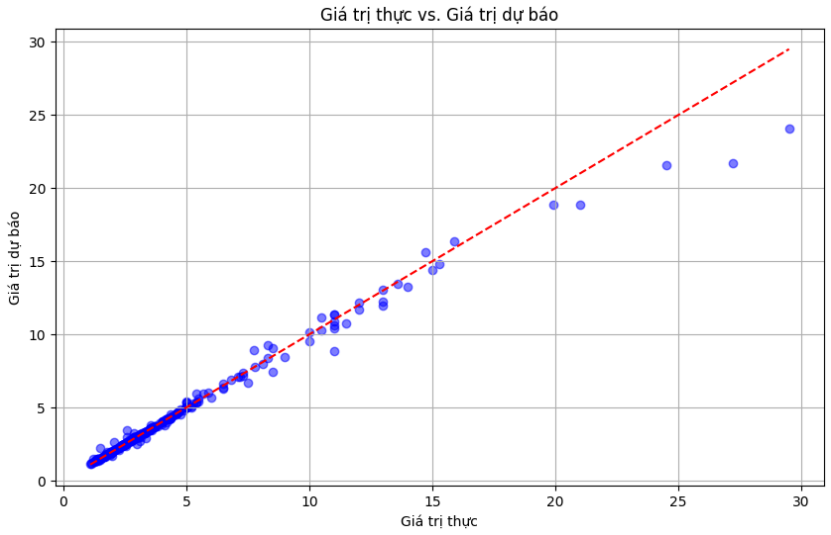


Hình 18: Hình ảnh dự đoán giá của mô hình Gradient Boosting Regression trên tập dữ liệu huấn luyện sau hiệu chỉnh

- Nhận xét: Giá trị dự đoán tập trung quanh đường chéo 45 độ hơn so với ban đầu

4.2 Mô hình hồi quy Random Forest:

- một kỹ thuật tổng hợp có khả năng thực hiện cả nhiệm vụ nhiệm vụ hồi quy và phân loại bằng cách sử dụng nhiều cây quyết định và một kỹ thuật gọi là Bootstrap và Aggregation, thường được gọi là đóng bao. Ý tưởng cơ bản đằng sau điều này là kết hợp nhiều cây quyết định để xác định đầu ra cuối cùng thay vì dựa vào các cây quyết định riêng lẻ.
- Các bước thực hiện:
    - Bước-1: Nhập thư viện.
    - Bước 2: Nhập dữ liệu.
    - Bước 3: Chuẩn bị dữ liệu.
    - Bước 4: Mô hình hồi quy rừng ngẫu nhiên.
    - Bước-5: Đưa ra dự đoán và đánh giá.
- Bộ tham số:
    - random_state: Kiểm soát hạt giống ngẫu nhiên được cung cấp cho mỗi công cụ ước tính Cây tại mỗi tăng cường lặp lại.
- Các đồ thị:



Hình 19: Hình ảnh dự đoán giá của mô hình Random Forest trên tập dữ liệu huấn luyện

- Nhận xét: Biểu đồ này thể hiện mối quan hệ giữa giá trị thực và giá trị dự báo. Có thể thấy rằng dữ liệu thực tế và dự báo có sự biến động và tương quan. Đường xu hướng đỏ cho thấy một mô hình dự đoán tương đối tốt, nhưng còn một số điểm ngoại lai cách xa đường đỏ.

- Các chỉ số MAE, MSE, RMSE, R^2:

	Huấn luyện	**Kiểm thử**
MAE	0.0777	0.1952
MSE	0.0612	0.3647
RMSE	0.2475	0.6039
R^2	0.9964	0.9786

Bảng 9: So sánh các chỉ số giữa hai tập huấn luyện và kiểm tra

- Nhận xét:
    - MAE của tập huấn luyện là 0.0777, thấp hơn đáng kể so với MAE của tập kiểm tra là 0.1952 Điều này cho thấy mô hình hoạt động rất tốt trên tập huấn luyện nhưng không có khả năng tổng quát hóa tốt trên tập kiểm tra và kiểm định chéo.
    - MSE của tập huấn luyện là 0.0612, rất thấp so với MSE của tập kiểm tra là 0.3647.
    - Giá trị R^2 của tập huấn luyện là 0.9964, rất gần với giá trị hoàn hảo là 1. Trong khi đó, R^2 của tập kiểm tra là 0.9786, vẫn là một giá trị cao nhưng thấp hơn so với tập huấn luyện.

4.3 So sánh hai mô hình hồi quy:

	Gradient Boosting	**Random Forest**
MAE	0.1098	0.0777
MSE	0.0328	0.0612
RMSE	0.1812	0.2475
R^2	0.9981	0.9964

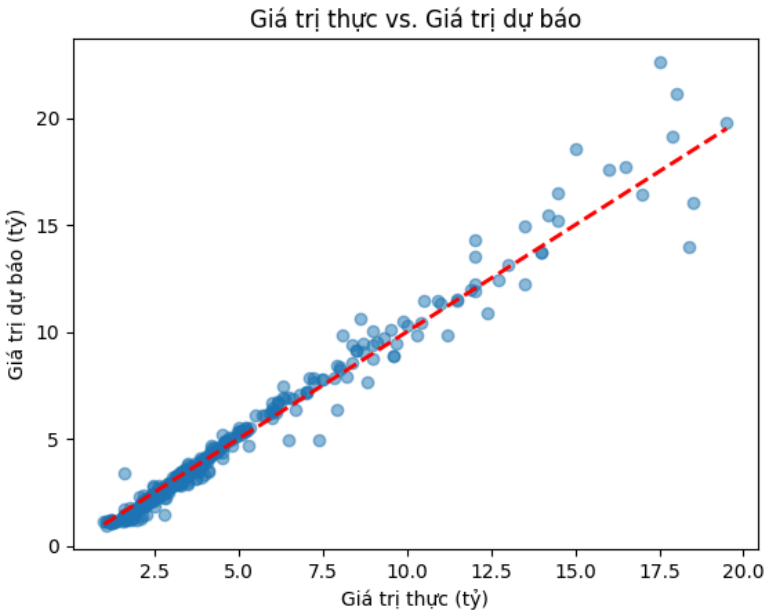
Bảng 10: So sánh chỉ số của hai mô hình hồi quy

- Nhận xét:
    - Giá trị MAE của Gradient Boosting cao hơn Random Forest.
    - Giá trị MSE của Gradient Boosting thấp gần gấp đôi Random Forest.
    - Giá trị RMSE của Gradient Boosting thấp hơn Random Forest.

- Giá trị R^2 của Gradient Boosting cao hơn Random Forest. Nên sẽ chọn mô hình Gradient Boosting

4.4 Thực hiện kiểm tra mô hình Gradient Boosting Regression trên tập dữ liệu mới:

- Các chỉ số MAE:0.3478, MSE: 0,2917, RMSE:0,6259, R^2=0.966
- R^2 giảm còn các chỉ số còn lại đều tăng lên, cho thấy mô hình đã dự đoán sai lệch một vài giá trị trên tập dữ liệu mới hoàn toàn



	Gia du doan	Gia thuc	% sai lech
18	11.497606	11.50	0.0208
181	7.847283	7.85	0.0346
35	10.404360	10.40	0.0419
43	3.402039	3.40	0.0600
9	2.948058	2.95	0.0658
...	...	...	...
167	1.436509	2.25	36.1552
106	1.188117	1.95	39.0709
16	1.240495	2.09	40.6462
245	1.434340	2.80	48.7736
267	3.407458	1.59	114.3056

```
count    419.000000
mean       8.454273
std        9.045015
min        0.020800
25%        3.028600
50%        6.233300
75%       11.340650
max      114.305600
Name: % sai lech, dtype: float64
```

Hình 20: Hình ảnh dự đoán giá của mô hình Gradient Boosting Regression trên tập dữ liệu kiểm tra

- Nhận xét: Giá trị dự đoán khá bám sát đường chéo 45 độ, tuy nhiên có một vài giá trị rời xa đường chéo tập trung ở các mức giá cao

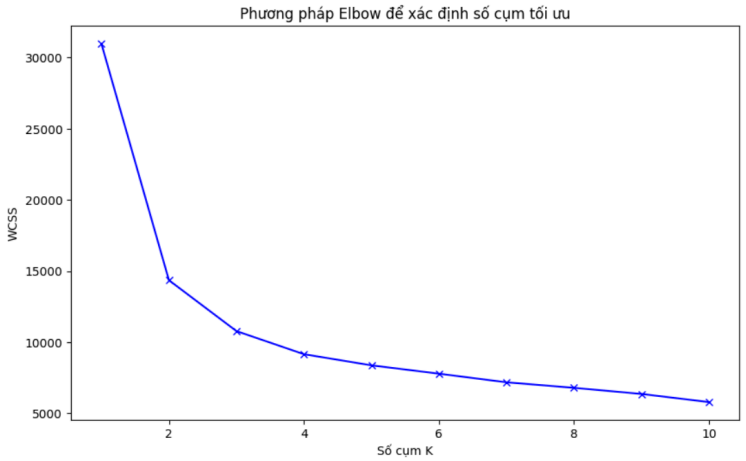
- Giá trị phần trăm sai lệch giao động 0,03% đến 114%
- Giá trị lệch trung bình 8,45%, 75% giá trị có sự sai lệch < 11,34%
- Giá trị sai lệch trung bình: 0.626 tỉ

4.5 Thuật toán K-means:

- Thuật toán K-means là một thuật toán lặp lại cố gắng phân vùng tập dữ liệu thành K nhóm con (cụm) riêng biệt không chồng chéo được xác định trước, trong đó mỗi điểm dữ liệu chỉ thuộc về một nhóm . Nó cố gắng làm cho các điểm dữ liệu trong cụm càng giống nhau càng tốt đồng thời giữ cho các cụm càng khác biệt (càng xa) càng tốt. Nó chỉ định các điểm dữ liệu cho một cụm sao cho tổng bình phương khoảng cách giữa các điểm dữ liệu và trung tâm của cụm (trung bình cộng của tất cả các điểm dữ liệu thuộc cụm đó) là nhỏ nhất. Chúng ta càng có ít biến thể trong các cụm, các điểm dữ liệu càng đồng nhất (tương tự) trong cùng một cụm.

- Cách thức hoạt động của thuật toán kmeans như sau:
    - Xác định số lượng các cụm K .
    - Khởi tạo các trung tâm bằng cách xáo trộn tập dữ liệu trước và sau đó chọn ngẫu nhiên K điểm dữ liệu cho các trung tâm mà không cần thay thế.
    - Tiếp tục lặp lại cho đến khi không có thay đổi đối với centroid. tức là việc gán các điểm dữ liệu cho các cụm không thay đổi.
    - Tính tổng bình phương khoảng cách giữa các điểm dữ liệu và tất cả các centroid.
    - Gán mỗi điểm dữ liệu cho cụm gần nhất (centroid).
    - Tính toán trọng tâm cho các cụm bằng cách lấy giá trị trung bình của tất cả các điểm dữ liệu thuộc mỗi cụm.
- Các tham số:
    - n_clusters=K: là số cụm mà thuật toán sẽ cố gắng tìm.
    - init='k-means++': là phương pháp khởi tạo các tâm cụm.
    - max_iter=1000: là số lần lặp tối đa mà thuật toán sẽ thực hiện.

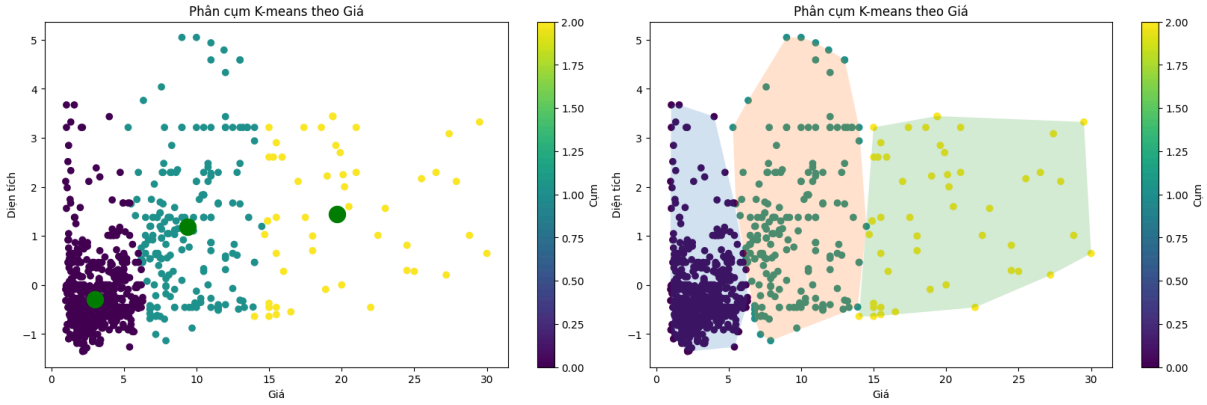
- Cách chọn tham số:



Hình 21: Đồ thị Elbow.

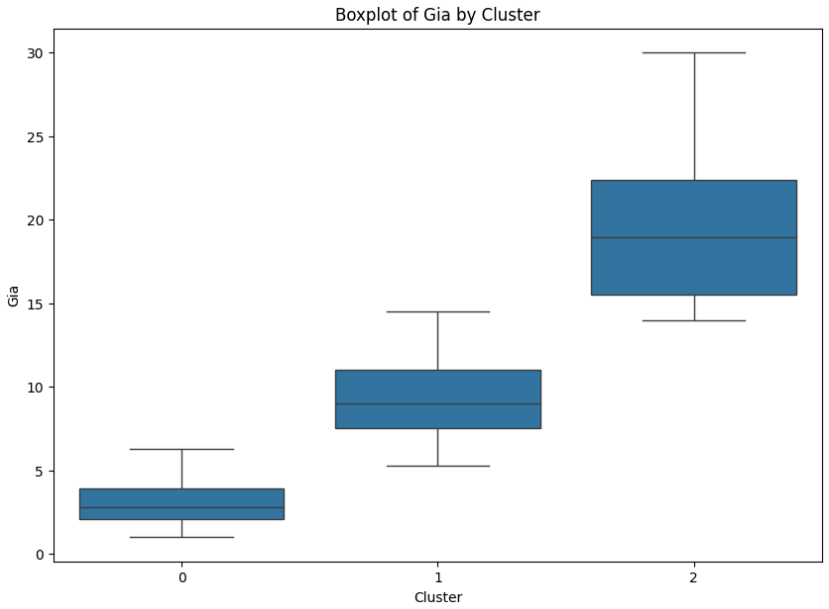
- cách chọn tham số k: Điểm khuỷu tay là điểm mà ở đó tốc độ suy giảm của hàm biến dạng sẽ thay đổi nhiều nhất. Tức là kể từ sau vị trí này thì gia tăng thêm số lượng cụm cũng không giúp hàm biến dạng giảm đáng kể.

- Đồ thị:



Hình 22: đồ thị của kết quả phân cụm.

- Nhận xét: dữ liệu được chia thành ba cụm. Các cụm được chia theo giá.



Hình 23: Biểu đồ boxplot của ba cụm.

- Nhận xét: cụm 0 khoảng cách giá trị rất ngắn so với hai cụm còn lại, cụm 2 thì khoảng cách của box rất rộng.

4.6 Mô hình GaussianNB

- Là một mô hình học máy được sử dụng rộng rãi trong phân loại. Trong mô hình này, chúng ta giả định rằng các đặc trưng (features) của dữ liệu được phân phối theo phân phối Gaussian
- Cách hoạt động:
  - Giả định đặc trưng độc lập: Mô hình giả định rằng các đặc trưng (features) trong dữ liệu là độc lập có điều kiện
  - Huấn luyện: Trong quá trình huấn luyện, mô hình ước lượng các tham số của phân phối Gaussian cho mỗi lớp dựa trên dữ liệu huấn luyện.
  - Dự đoán: Khi có một mẫu mới cần dự đoán, mô hình tính toán xác suất cho mẫu thuộc vào mỗi lớp bằng cách sử dụng phân phối Gaussian đã học được và giả định đặc trưng độc lập.
  - Chọn lớp: Mô hình chọn lớp có xác suất cao nhất cho mẫu mới. Cụ thể, nó chọn lớp mà mẫu có xác suất cao nhất thuộc vào lớp đó.
- Bộ tham số:
  - priors: Một mảng các xác suất tiền nghiệm cho từng lớp, không được chỉ định, tất cả các lớp được giả định có cùng xác suất.

- var_smoothing: Một hằng số được thêm vào phương sai (variance) của mỗi đặc trưng để đảm bảo tính chính xác khi tính toán xác suất.

- Giá trị Accuracy: 0.9623

	precision	**recall**	**f1-score**
0	1.00	**0.97**	**0.98**
1	0.84	0.93	0.88
2	0.74	0.93	0.82
accuracy			0.96
macro avg	0.86	0.94	0.90
weighted avg	0.97	0.96	0.96

Bảng 11: Bảng đánh giá hiệu suất của mô hình GaussianNB

4.7 Mô hình Random Forest Classifier

- Là tạo một tập hợp các cây quyết định từ một tập hợp con được chọn ngẫu nhiên của tập huấn luyện.
- Bộ tham số:
    - n_estimators: Số lượng cây trong rừng.
    - random_state: Kiểm soát hạt giống ngẫu nhiên được cung cấp cho mỗi công cụ ước tính cây tại mỗi tăng cường lặp lại.
- Chia tập huấn luyện: kiểm thử theo tỉ lệ 63:37
- Giá trị Accuracy: 0.997

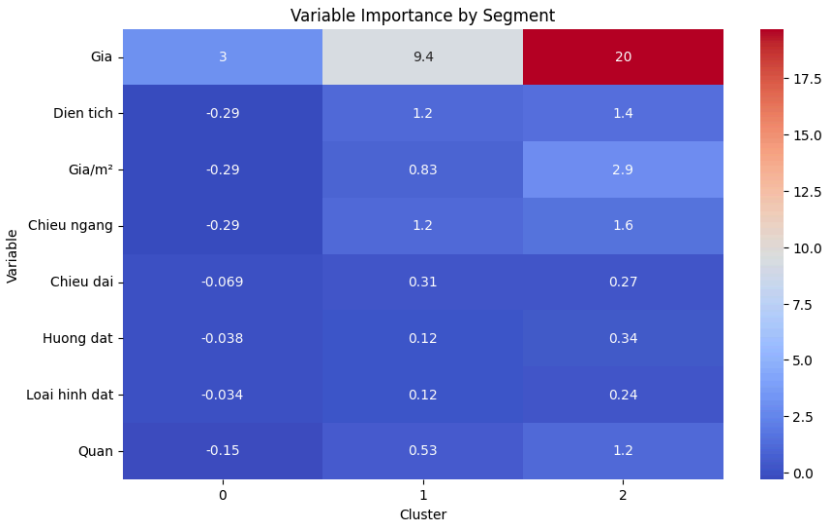
	precision	**recall**	**f1-score**
0	1.00	1.00	1.00
1	0.99	1.00	0.99
2	1.00	1.00	1.00
accuracy			1.00
macro avg	1.00	1.00	1.00
weighted avg	1.00	1.00	1.00

Bảng 12: Bảng đánh giá hiệu suất của mô hình Random Forest Classifier

- **Kiểm tra mô hình bằng kỹ thuật cross-validation**

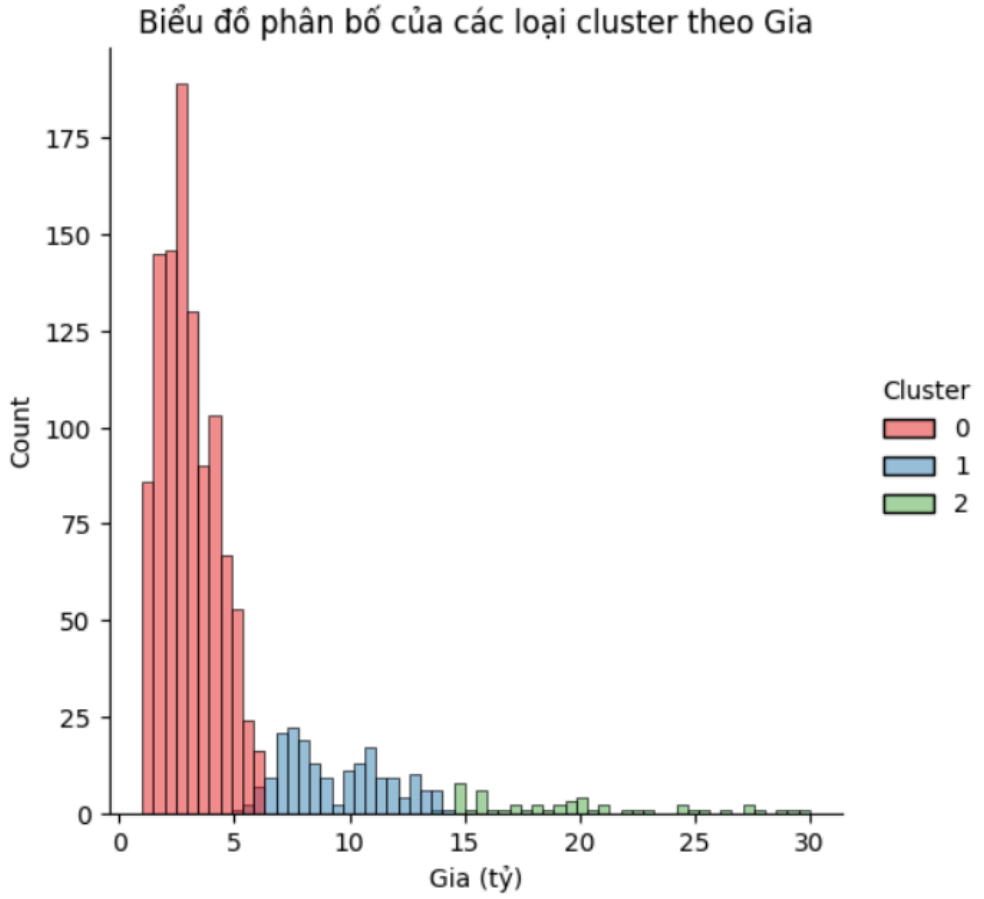
	Accuracy
Fold 1	1.0
Fold 2	0.9961
Fold 3	0.9883
Fold 4	0.9961
Fold 5	0.9844

Bảng 13: Bảng kết quả hiệu suất mô hình bằng kỹ thuật cross-validation.

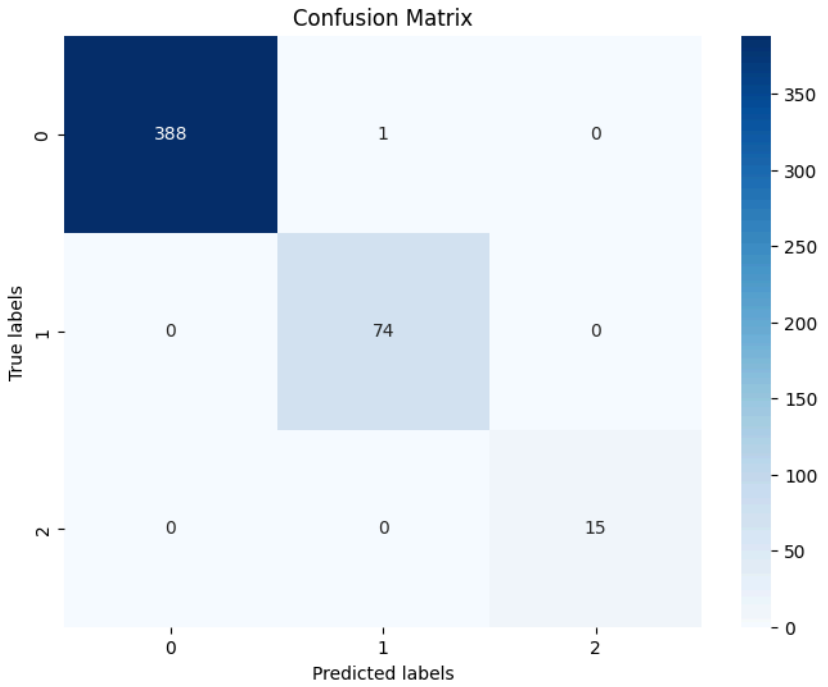


Hình 24: Hình ảnh tầm quan trọng của biến thay đổi theo phân khúc trên tập dữ liệu huấn luyện

- Nhận xét:
  - biến Gia có mức độ quan trọng cao nhất đối với mỗi loại
  - biến loai hinh dat có mức độ quan trọng thấp nhất đối với mỗi loại



Hình 25: Hình ảnh phân bố của các loại cluster theo giá



Hình 26: Hình ảnh ma trận nhầm lẫn của mô hình trên tập dữ liệu huấn luyện

- Nhận xét:
    - Mô hình dự đoán tốt, chỉ nhầm lẫn rất ít mẫu
- **Kiểm tra mô hình** bằng kỹ thuật cross-validation, cho ta kết quả giao động từ 97.67% đến 100%. Giá trị trung bình đạt được là 99.3%

4.8 So sánh hai mô hình phân loại:

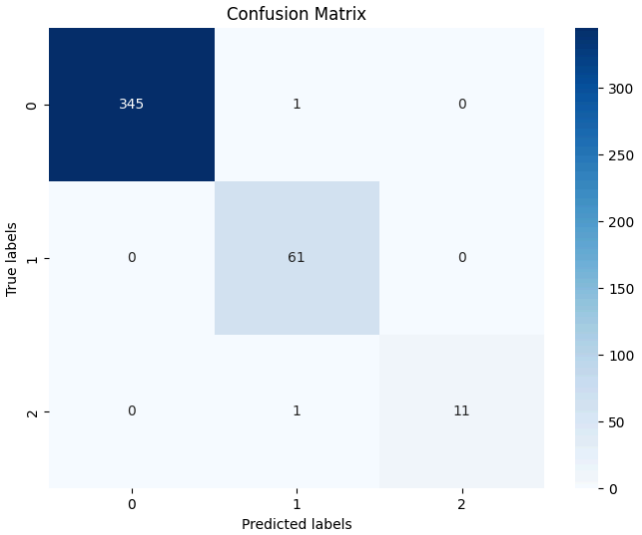
	GaussianNB	**Random Forest Classifier**
Accuracy	0.9623	0.9979
Weighted avg precision	0.97	1.0
Weighted avg recall	0.96	1.0
Weighted avg f1-score	0.96	1.0

Bảng 11: So sánh hai mô hình phân loại

- Nhận xét: Tất cả giá trị của Random Forest Classifier đều tốt hơn GaussianNB. Nên ta sẽ chọn mô hình Random Forest Classifier

4.9 Thực hiện kiểm tra mô hình Random Forest Classifier trên tập dữ liệu mới:

- Giá trị Accuracy: 0.9928



Hình 27: Hình ảnh ma trận nhầm lẫn của mô hình trên tập dữ liệu kiểm tra

- Nhận xét: Mô hình dự đoán chính xác, nhầm lẫn rất ít

5. Kết luận.

- **Kết quả đạt được:**
    - Đạt được một mô hình dự đoán giá đất nền với độ chính xác khá cao
    - Mô hình có thể hiển thị được mối quan hệ giữa các biến độc lập và giá đất nền, giúp người dùng hiểu rõ hơn về yếu tố ảnh hưởng đến giá đất.
    - Xây dựng một mô hình phân loại giá đất nền với độ chính xác cao
    - Mô hình có khả năng phân loại các khu vực thành các nhóm giá đất khác nhau, giúp người dùng hiểu rõ hơn về phân bố giá đất trên thị trường.
- **Hướng phát triển:**
    - Tăng cường thu thập dữ liệu: Thu thập thêm dữ liệu từ nhiều nguồn khác nhau để tăng tính đa dạng và độ chính xác của mô hình.
    - Cải thiện tiền xử lý dữ liệu: Xử lý dữ liệu thiếu và dữ liệu nhiễu một cách kỹ lưỡng để tăng chất lượng dữ liệu.
    - Tích hợp tính năng dự báo thời gian thực: Phát triển các ứng dụng hoặc hệ thống có khả năng dự báo giá đất nền theo thời gian thực dựa trên dữ liệu đầu vào cập nhật liên tục.
    - Phân tích sâu hơn về yếu tố ảnh hưởng: Thực hiện phân tích sâu hơn về từng yếu tố ảnh hưởng đến giá đất nền để hiểu rõ hơn về mối quan hệ giữa các biến số và giá đất

6. Tài liệu tham khảo

[1] Feature Engineering and Selection: A Practical Approach for Predictive Models (http://www.feat.engineering/)

[2] A Deep-Learned Embedding Technique for Categorical Features Encoding (https://drive.google.com/file/d/1RBfFRSccYSInrl-4ed0LwhCrvLLdf819/view)

[3] Sự đánh đổi giữa độ lệch và phương sai (https://phamdinhkhanh.github.io/deepai-book/ch_ml/OvfAndUdf.html)

[4] An Overview of Categorical Encoding Methods (An Overview of Categorical Encoding Methods (kaggle.com))

[5] Feature Engineering (11.1. Feature Engineering — Deep AI KhanhBlog (phamdinhkhanh.github.io)

[6] Học máy và khai phá dữ liệu (users.soict.hust.edu.vn/khoattq/ml-dm-course/)

[7] Random Forest Classifier Tutorial (Random Forest Classifier Tutorial (kaggle.com))