

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

ĐH * ĐHTT



**BÁO CÁO ĐỒ ÁN MÔN HỌC KỸ THUẬT LẬP
TRÌNH PYTHON
MÔ HÌNH WEB DỰ ĐOÁN GIÁ NHÀ**

Sinh viên thực hiện:

STT	Họ tên	MSSV
1	Nguyễn Việt Thư	19522309
2	Trương Thị Kim Thoa	19522295

TP. HỒ CHÍ MINH – 05/2022

MỤC LỤC

1. GIỚI THIỆU	3
2. NỘI DUNG	4
2.1. QUY TRÌNH ÁP DỤNG XÂY DỰNG MÔ HÌNH.....	4
SƠ ĐỒ HOẠT ĐỘNG CỦA MÔ HÌNH.....	4
2.2. QUY TRÌNH ÁP DỤNG XÂY DỰNG MÔ HÌNH.....	4
2.2.1. Crawl dữ liệu	4
2.2.2. Tổng quan bộ dữ liệu.....	5
2.2.3. Làm sạch dữ liệu	6
2.2.4. Xây dựng mô hình	10
2.2.5. Xây dựng web.....	10
2.3. ĐÁNH GIÁ MÔ HÌNH	11
2.3.1. Các mô hình mặc định.....	11
2.3.2. Thực hiện tuning Gridsearch	11
2.3.3. Thực hiện đánh giá web	12
3. KẾT LUẬN	12
TÀI LIỆU THAM KHẢO	13
PHỤ LỤC	13
PHỤ LỤC PHÂN CÔNG NHIỆM VỤ	13

MỤC LỤC HÌNH

Hình 1. Giao diện website dự đoán giá mặc định.....	3
Hình 2. Sơ đồ hoạt động của mô hình dự đoán	4
Hình 3. DataFrame gốc chưa qua xử lý.....	5
Hình 4. Mô tả tổng quát về DataFrame gốc	5
Hình 5. Thống kê trên DataFrame	6
Hình 6. DataFrame sau khi được fill, drop NULL và đổi tên	6
Hình 7. Bộ dữ liệu sau khi được làm sạch.....	7
Hình 8. Thông tin của dữ liệu khi được làm sạch	7
Hình 9. Biểu đồ phân bố của các thuộc tính số liên tục	8
Hình 10. Biểu đồ phân bố của trường thuộc tính district	8
Hình 11. Biểu đồ phân bố của trường thuộc tính type_of_housing	9
Hình 12. Biểu đồ phân bố của trường thuộc tính legal_paper.....	9
Hình 13. Giao diện website dự đoán	12

1. GIỚI THIỆU

Phát triển xây dựng mô hình Dự đoán giá nhà ở Thành phố HCM là một chủ đề khá hot trong những năm gần đây, nhất là trong thời đại mà giá của BĐS ngày càng tăng vọt. Trong báo cáo này, chúng tôi tập trung trình bày ba nội dung chính: (1) Quy trình áp dụng xây dựng mô hình, (2) Thiết lập từng phần cho mô hình, (3) Đánh giá mô hình.



Nhóm bắt tay vào việc tiến hành crawl dữ liệu trên trang web về Bất Động Sản, dữ liệu thô sau khi qua các bước làm sạch có thể cho vào model để training. Nhóm tiến hành training trên nhiều mô hình để chọn được mô hình nào tốt nhất, phù hợp nhất với bộ dữ liệu. Cuối cùng là kết hợp xây dựng mô hình có được lên trang web để tạo ra ứng dụng trực quan cho người dùng sử dụng. Ứng dụng này giúp người dùng dự đoán được giá của Căn hộ mà mình muốn mua khi được người dùng cung cấp 1 vài thông tin ngôi nhà như: Vị trí căn hộ? (vd: Quận 1), Số phòng ngủ, số tầng, diện tích, ...

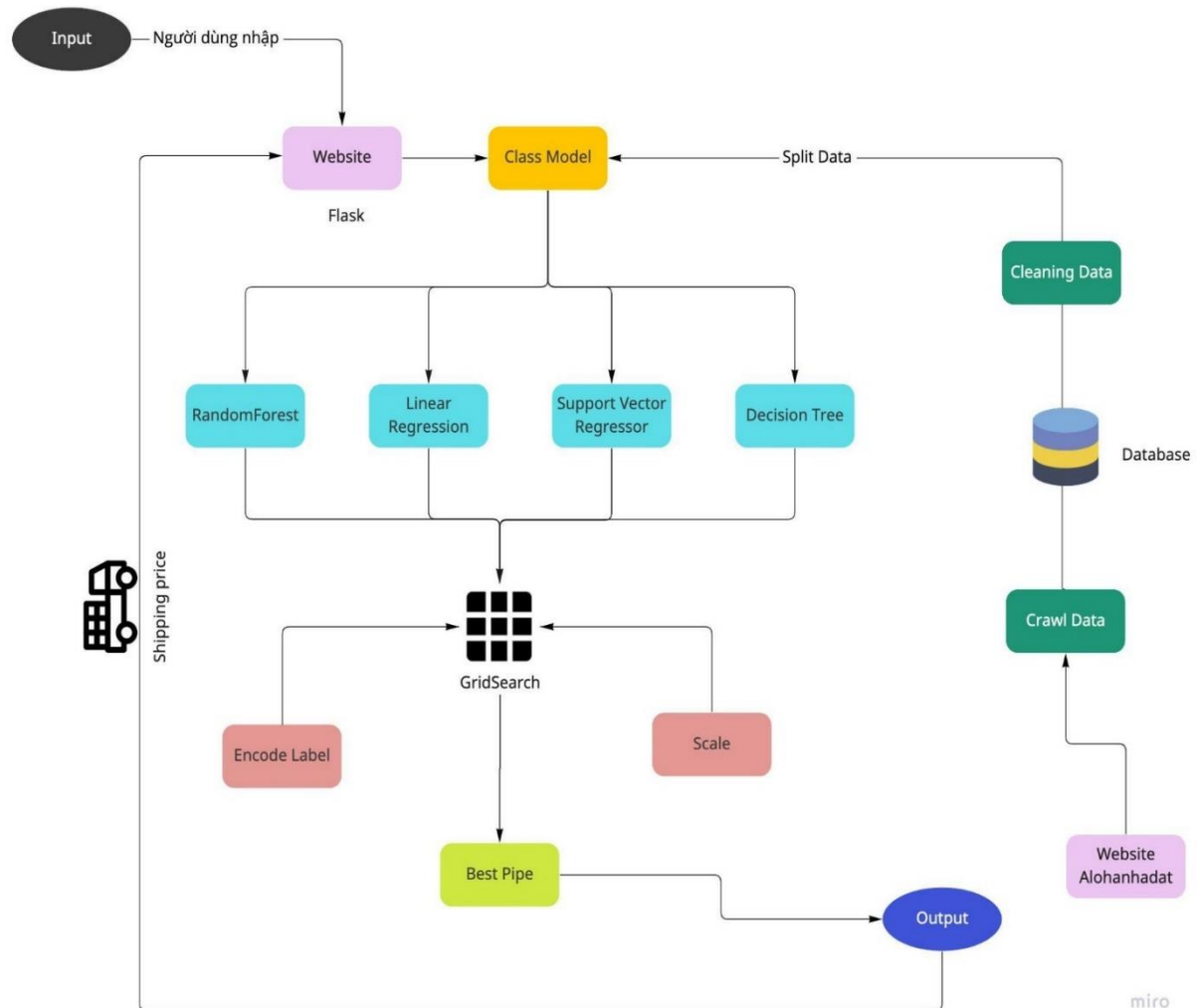
Welcome to Ho Chi Minh city house predict	
Select District:	Select Type of Housing:
Quận 1	Nhà mặt tiền
Select legal paper:	House Size
Số hồng/ Số đỏ	Enter Size
Number of floor	Number of room
Enter Number of Floor	Enter Number of Room
Width Meter	Length Meter
Enter Width Meter	Enter Length Meter
Predict Price	

Hình 1. Giao diện website dự đoán giá mặc định

2. NỘI DUNG

2.1. QUY TRÌNH ÁP DỤNG XÂY DỰNG MÔ HÌNH

SƠ ĐỒ HOẠT ĐỘNG CỦA MÔ HÌNH



Hình 2. Sơ đồ hoạt động của mô hình dự đoán

2.2. QUY TRÌNH ÁP DỤNG XÂY DỰNG MÔ HÌNH

2.2.1. Crawl dữ liệu

Việc lựa chọn trang web tin cậy về bất động sản là vô cùng cần thiết. Sau quá trình nghiên cứu chúng em quyết định chọn trang web: <https://alanhadat.com.vn/>. Và dữ liệu được crawl về là những bài đăng mới nhất về buôn bán nhà đất. Đảm bảo tính thực tế.

- Cách tiến hành:

- Sử dụng thư viện Selenium thao tác trực tiếp trên trang web, truy cập vào các open source của trang web.
- Sử dụng BeautifulSoup đọc và lấy dữ liệu.
- Kết nối với MySQL để lưu trữ dữ liệu thành file csv.

2.2.2. Tổng quan bộ dữ liệu

Số lượng dữ liệu thu thập: 20 000 dòng với 9 trường thuộc tính: [“Quận”, “Loại_BDS”, “Pháp_lý”, “Số_tầng”, “Số_phòng”, “Diện_tích”, “Dài”, “Rộng”, “Giá”].

	Quận	Loại_BDS	Pháp_lý	Số_tầng	Số_phòng	Diện_tích	Dài	Rộng	Giá
0	Huyện Bình Chánh	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	1	4	100 m2	20m	5m	1,2 tỷ
1	Huyện Củ Chi	Đất thổ cư, đất ở	Sổ hồng/ Sổ đỏ	---	---	3.100 m2	50m	30m	2,5 triệu / m2
2	Huyện Củ Chi	Đất thổ cư, đất ở	Sổ hồng/ Sổ đỏ	---	---	5.002 m2	205m	45m	12,75 tỷ
3	Huyện Củ Chi	Đất thổ cư, đất ở	---	---	---	4.147 m2	---	---	6,39 tỷ
4	Huyện Củ Chi	Đất thổ cư, đất ở	Sổ hồng/ Sổ đỏ	---	---	1.635 m2	82m	20m	9 tỷ

Hình 3. DataFrame gốc chưa qua xử lý

- ➔ Nhận xét: với 5 dòng dữ liệu đầu tiên ta có thể thấy được dữ liệu còn chứa nhiều giá trị NULL (Pháp_lý, Số_tầng, Số_phòng), và giá trị trong cột không thống nhất về đơn vị hay format (Diện_tích, Dài, Rộng, Giá).

- Thông tin về dataframe:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 20000 entries, 0 to 9999
Data columns (total 9 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   Quận      20000 non-null  object
1   Loại_BDS     20000 non-null  object
2   Pháp_lý    14109 non-null  object
3   Số_tầng    17092 non-null  object
4   Số_phòng   17079 non-null  object
5   Diện_tích  20000 non-null  object
6   Dài        16847 non-null  object
7   Rộng       17485 non-null  object
8   Giá        20000 non-null  object
dtypes: object(9)
memory usage: 1.5+ MB
```

Hình 4. Mô tả tổng quát về DataFrame gốc

- ➔ Nhận xét:

- Pháp_lý chứa nhiều dữ liệu Null nhất, nhiều căn hộ được đăng bán chưa cung cấp giấy.
- Kiểu dữ liệu của các trường thuộc tính chưa phù hợp.

➤ Thực hiện thống kê trên dữ liệu:

	Quận	Loại_BDS	Pháp_lý	Số_tầng	Số_phòng	Diện_tích	Dài	Rộng	Giá
count	20000	20000	14109	17092	17079	20000	16847	17485	20000
unique	45	14	4	37	73	740	330	281	1159
top	Quận Gò Vấp	Nhà trong hẻm	Sổ hồng/ Sổ đỏ	2	4	100 m2	20m	4m	25 tỷ
freq	2891	8409	13780	3849	3923	948	2097	4039	211

Hình 5. Thống kê trên DataFrame

→ Nhận xét:

- Ở HCM không có tới 45 quận huyện.
- Các giá trị NULL và sai format còn nhiều, cần tiến hành bước làm sạch tiếp theo.

2.2.3. Làm sạch dữ liệu

Dữ liệu thô được crawl từ trang web chưa thể sử dụng, ở bước phân tích tổng quan ta có thể thấy dữ liệu chứa nhiều giá trị Null, viết với nhiều định dạng khác nhau. Ta cần tiến hành xử lý trước khi đưa vào model huấn luyện.

- Cách tiến hành:

➤ Thay đổi tên trường dữ liệu theo ý muốn, drop bỏ những dòng dữ liệu chứa NULL. (Không thể fill 1 giá trị bất kì vì nó ảnh hưởng rất nhiều đến độ tin cậy của bài toán).

	index	district	type_of_housing	legal_paper	num_floors	num_bed_rooms	squared_meter_area	length_meter	width_meter	price
0	0	Huyện Bình Chánh	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	1	4	100 m2	20m	5m	1,2 tỷ
1	6	Thành phố Thủ Đức	Nhà mặt tiền	Không có giấy	3	4	55 m2	11m	5m	7 tỷ
2	8	Thành phố Thủ Đức	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	4	4	72 m2	14m	5m	9,2 tỷ
3	11	Thành phố Thủ Đức	Nhà trong hẻm	Sổ hồng/ Sổ đỏ	3	3	73 m2	12m	6,6m	4,99 tỷ
4	14	Huyện Bình Chánh	Nhà mặt tiền	Không có giấy	1	4	100 m2	20m	5m	1 tỷ

Hình 6. DataFrame sau khi được fill, drop NULL và đổi tên

➤ Tiến hành làm sạch trên từng trường dữ liệu:

- “district”: Chỉ giữ lại tên quận huyện, vd: “Đường số 5, Quận 9” -> “Quận 9”.

- “type_of_housing”: có khá nhiều loại hình căn hộ ở đây, sử dụng label encoder để lưu các giá trị về những con số nhất định.
- “legal_paper”: fill những dòng chứa Null = “Không có giấy” (thực hiện đầu tiên, trước khi drop tất cả giá trị null của bảng).
- “num_floors”, “num_bed_room”: đổi thành kiểu int.
- “squared_meter_area”: loại bỏ kí tự đơn vị mét vuông, đổi thành kiểu float.
- “length_meter”, “width_meter”: loại bỏ đơn vị mét, xóa bỏ những dòng chứa giá trị sai format (vd: 09.,5m), chuyển thành kiểu float.
- “price”: xóa bỏ những dòng chứa giá trị sai format, chuyển lại kiểu float và chuyển đơn vị tính thành kiểu triệu đồng.

KẾT QUẢ VỀ BỘ DỮ LIỆU SAU KHI ĐƯỢC LÀM SẠCH

	index	district	type_of_housing	legal_paper	num_floors	num_bed_rooms	squared_meter_area	length_meter	width_meter	price
0	0	Huyện Bình Chánh	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	1	4	100.0	20.0	5.0	1200.0
1	6	Thành phố Thủ Đức	Nhà mặt tiền	Không có giấy	3	4	55.0	11.0	5.0	7000.0
2	8	Thành phố Thủ Đức	Nhà mặt tiền	Sổ hồng/ Sổ đỏ	4	4	72.0	14.0	5.0	9200.0
3	11	Thành phố Thủ Đức	Nhà trong hẻm	Sổ hồng/ Sổ đỏ	3	3	73.0	12.0	6.6	4990.0
4	14	Huyện Bình Chánh	Nhà mặt tiền	Không có giấy	1	4	100.0	20.0	5.0	1000.0

Hình 7. Bộ dữ liệu sau khi được làm sạch

→ Bảng dữ liệu sau khi clean đã về đúng định dạng và có thể sử dụng.

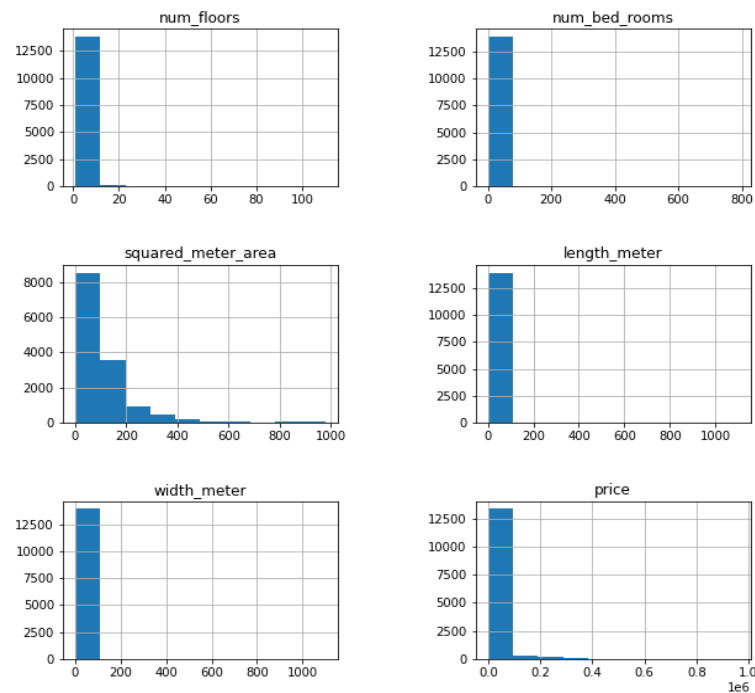
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13951 entries, 0 to 14098
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  -
 0   district            13951 non-null  object
 1   type_of_housing     13951 non-null  object
 2   legal_paper         13951 non-null  object
 3   num_floors          13951 non-null  int64
 4   num_bed_rooms       13951 non-null  int64
 5   squared_meter_area  13951 non-null  float64
 6   length_meter        13951 non-null  float64
 7   width_meter         13951 non-null  float64
 8   price               13951 non-null  float64
dtypes: float64(4), int64(2), object(3)
memory usage: 1.1+ MB
```

Hình 8. Thông tin của dữ liệu khi được làm sạch

→ Giảm từ 20 000 dòng dữ liệu về còn 13 951 dòng dữ liệu sạch.

VISUALIZATION VỀ BỘ DỮ LIỆU NÀY

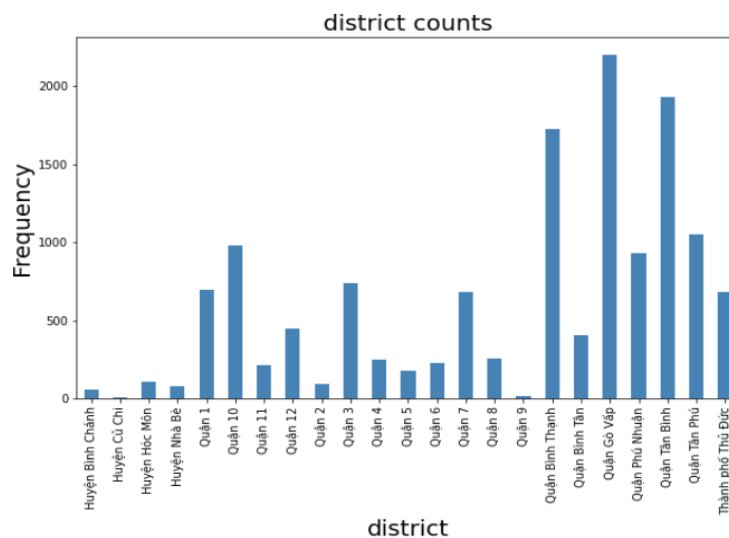
❖ Sự phân bố của các cột dữ liệu số liên tục:



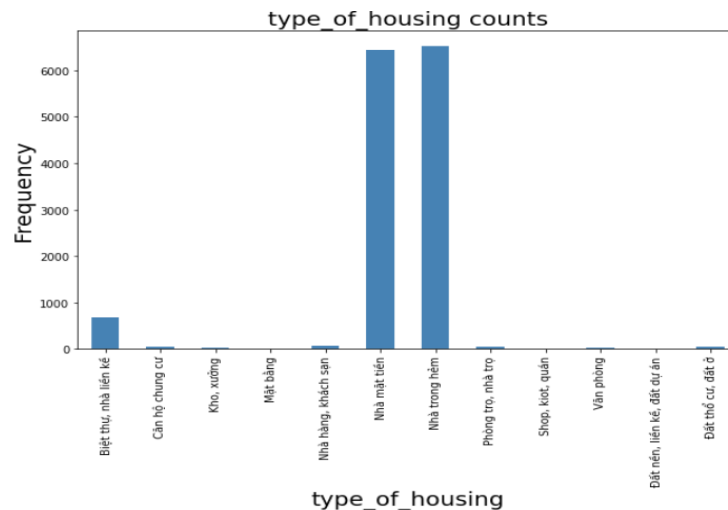
Hình 9. Biểu đồ phân bố của các thuộc tính số liên tục

→ Nhận xét: Các biểu đồ cột cho thấy các trường dữ liệu phân bố đơn giản, có tính tập trung ở phần đầu bảng (nhỏ nhất).

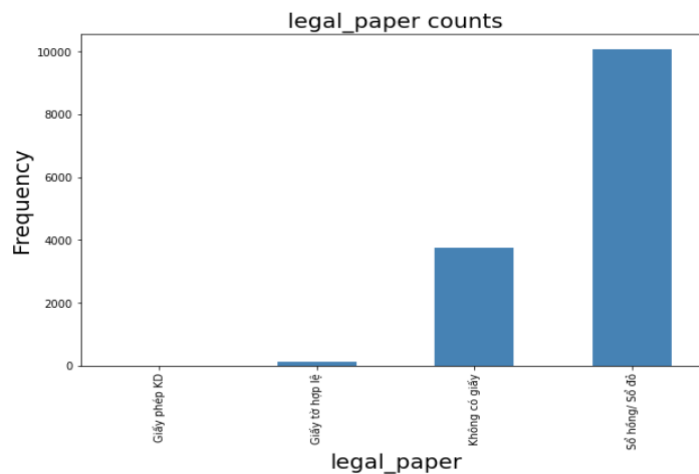
❖ Phân bố của các cột dữ liệu phân loại:



Hình 10. Biểu đồ phân bố của trường thuộc tính district



Hình 11. Biểu đồ phân bố của trường thuộc tính *type_of_housing*



Hình 12. Biểu đồ phân bố của trường thuộc tính *legal_paper*

→ Nhận xét:

- District: phân bố khá đều, với phần đông căn hộ ở “Gò Vấp”, ít nhất là ở “huyện Củ Chi”.
- Type_of_housing: các căn hộ trong bộ dữ liệu này chủ yếu là “Nhà mặt tiền”, “Nhà trong hẻm”. Các loại căn hộ còn lại là khá thấp, nên có thể ảnh hưởng đến kết quả mô hình.
- Legal_paper: hầu hết căn nhà đều có sổ hồng, bên cạnh đó còn tồn tại nhiều căn hộ chưa có giấy, người dùng sẽ đắn đo khi lựa chọn những căn nhà thế này.

2.2.4. Xây dựng mô hình

Nhóm tiến hành huấn luyện nhiều mô hình trên tập dữ liệu training, sau đó tiến hành đánh giá độ chính xác mô hình trên tập test. Cuối cùng chọn được mô hình hiệu quả nhất cho tập dữ liệu dựa trên accuracy. Mô hình tốt nhất được chọn sẽ được thực hiện GridSearch trên tập parameter mẫu để góp phần nâng cao độ chính xác.

- Cách tiến hành:
 - Bước 1: Sử dụng khoảng quartile Q1-Q3 loại bỏ các outlier, đưa bộ dữ liệu về dạng chuẩn ít các giá trị nhiễu.
 - Bước 2: Số hóa các cột dữ liệu category thông qua Label Encoder, sử dụng Standard Scaler chuẩn hóa các cột dữ liệu số liên tục.
 - Bước 3: Phân chia dữ liệu train, test (8:2).
 - Bước 4: Tiến hành huấn luyện trên 4 loại mô hình: Random Forest, Linear Regression, Support Vector Regressor, Decision Tree.
 - Bước 5: Đánh giá độ chính xác của 4 mô hình này thông qua R2_Score trên tập test, lựa chọn mô hình có độ chính xác cao nhất.
 - Bước 6: Thực hiện gridsearch cho mô hình tốt nhất với tập parameter mẫu mà nhóm tìm được, huấn luyện lại mô hình này với best_parameter và tiến hành xem xét kết quả.
 - Bước 7: Save mô hình tốt nhất có được thông qua pickle để tái sử dụng và dễ dàng gọi để kết hợp xây dựng web.

2.2.5. Xây dựng web

Nhóm sử dụng Web Framework Flask để xây dựng trang web, lý do nhóm chọn Framework này vì Flask rất nhẹ, dễ tiếp cận cho người mới bắt đầu tạo website nhỏ và cũng dễ mở rộng website phức tạp hơn.

Bước đầu tiên để xây dựng trang web nhóm đã build một template giao diện cho trang web của mình, giao diện web cho phép người dùng input vào những dữ liệu cần thiết cho model dự đoán bao gồm: quận, loại nhà, loại giấy tờ, diện tích, số tầng, số phòng, chiều dài, chiều rộng.

Cuối cùng load model đã tạo từ pickle và đánh giá mô hình từ dữ liệu người dùng nhập và xuất ra số tiền (đơn vị triệu đồng).

2.3. ĐÁNH GIÁ MÔ HÌNH

2.3.1. Các mô hình mặc định

BẢNG SO SÁNH SCORE GIỮA CÁC DEFAULT MODEL

MÔ HÌNH	LR	RF	DT	SVR
SCORE	0.7109	0.8944	0.8132	0.7882

→ Nhận xét:

- RF là mô hình tốt nhất với các tham số mặc định, ta chọn nó để tiến hành thực hiện tuning trên nhiều tham số.
- LR có accuracy thấp nhất do mô hình này khá đơn giản.
- Mặc dù các mô hình có accuracy tương đối cao, nhưng để đưa vào sử dụng trong thực tế thì cần cải thiện thêm.

2.3.2. Thực hiện tuning Gridsearch

❖ Parameter được sử dụng cho gridsearch:

```
{ 'n_estimators': [100,200,300,400,500],  
  'max_features': [ 'sqrt', 'log2'],  
  'max_depth': [10,20,30,40,50,60,70,80],  
  'min_samples_split': [1,2,3,4,5],  
  'min_samples_leaf': [1,2,3,4,5], }
```

❖ Best score: 0.8997

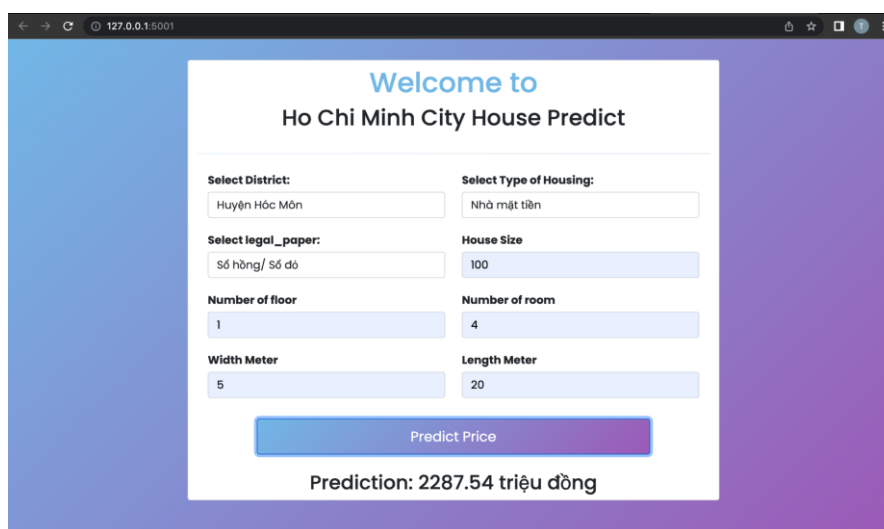
❖ Best parameter:

```
{ 'max_depth': 40,  
  'max_features': 'log2',  
  'min_samples_leaf': 1,  
  'min_samples_split': 2,  
  'n_estimators': 400 }
```

→ Nhận xét: Mặc dù accuracy cải thiện là rất ít (89.4 lên 89.99) nhưng đã góp phần vào ý tưởng cải thiện mô hình, Random Forest là mô hình phức tạp và tốn nhiều thời gian huấn luyện, gây nhiều khó khăn trong việc thực hiện tuning tham số tối ưu.

2.3.3. Thực hiện đánh giá web

Trang web được thiết kế với giao diện cơ bản, màu sắc dịu mắt, vị trí đặt các thanh tùy chọn cũng như kích cỡ tối ưu, có thể đem vào sử dụng trong thực tế.



Hình 13. Giao diện website dự đoán

3. KẾT LUẬN

Thông qua đề tài, chúng em có cơ hội tìm hiểu về việc xây dựng 1 mô hình dự đoán giá nhà ở Thành phố Hồ Chí Minh. Bắt đầu từ việc crawl dữ liệu, làm sạch, huấn luyện mô hình và build web trực quan. Đề tài này khá sát thực tế, đem lại nhiều hứng thú cho nhóm em trong quá trình thực hiện.

Nhìn chung nhóm đã hoàn thiện cơ bản những yêu cầu của bài toán, thực hiện đầy đủ các phần được ghi trong phần đăng ký đồ án. Bên cạnh đó còn tồn tại nhiều khó khăn trong việc cleaning dữ liệu sao cho tối ưu, hợp lý, vấn đề độ chính xác là những gì nhóm cần cải thiện.

Mô hình web dự đoán giá nhà với độ chính xác 90% là cơ sở, cũng như tiền đề để nhóm nỗ lực phát triển thêm và có thể đưa vào ứng dụng sử dụng trong tương lai.

TÀI LIỆU THAM KHẢO

[1] HaNoi Housing Market: Predict House Price.

Link: <https://rpubs.com/chidungkt/538677> (3/5/2022)

[2] Selenium with Python.

Link: <https://selenium-python.readthedocs.io/>(7/5/2022)

[3] Scikit-learn(Machine Learning in Python).

Link: <https://scikit-learn.org/stable/> (9/5/2022)

PHỤ LỤC

//

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

ST T	Thành viên	Nhiệm vụ
1	Nguyễn Việt Thư	Build source code model, viết báo cáo, slide
2	Trương Thị Kim Thoa	Scraping Data, Build Website, viết báo cáo, slide