**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA CÔNG NGHỆ PHẦN MỀM**

****

**BÁO CÁO ĐỒ ÁN 1**

**ĐỀ TÀI:**

**NGHIÊN CỨU VÀ THỬ NGHIỆM PHƯƠNG PHÁP**

**ĐỊNH GIÁ BẤT ĐỘNG SẢN**

**Giảng viên: Trần Anh Dũng**

**Sinh viên thực hiện: Nguyễn Ngọc Thạch 19520265**

**MỤC LỤC**

**DANH MỤC HÌNH ẢNH**

**DANH TỪ VIẾT TẮT**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Từ viết tắt** | **Từ đầy đủ** | **Ý nghĩa** |
| **1** | **BĐS** | **Bất động sản** |  |
| **2** |  |  |  |
| **3** |  |  |  |
| **4** |  |  |  |
| **5** |  |  |  |
| **6** |  |  |  |
| **7** |  |  |  |

**LỜI CẢM ƠN**

**ĐỀ CƯƠNG CHI TIẾT**

|  |
| --- |
| **TÊN ĐỀ TÀI:** Nghiên cứu và thử nghiệm một số phương pháp định giá bất động sản |
| **Giảng Viên Hướng Dẫn:** ThS. Trần Anh Dũng |
| **Thời gian thực hiện:** Từ ngày 2/3/2022 đến ngày x/x/2022 |
| **Sinh viên thực hiện:**  Nguyễn Ngọc Thạch – 19520265 |
| **Nội dung đề tài**  **1. Lý do chọn đề tài:**  Bất động sản là lĩnh vực có lực lượng người quan tâm hùng hậu. Bất động sản đóng vai trò quan trọng trong nền kinh tế, góp phần thu hút vốn đầu tư trong và ngoài nước, là cầu nối giữa người mua và bán, thúc đẩy phát triển khoa học kỹ thuật, … Hay cụ thể hơn, bất động sản có mặt và ảnh hưởng đến hầu hết mọi hoạt động của con người như: cư trú, làm việc, khai thác, canh tác, … Vì vậy, định giá bất động sản là bài toán được chú ý rất nhiều bởi những người có nhu cầu giao dịch bất động sản. Nhưng giá trị của bất động sản bị tác động bởi rất nhiều yếu tố. Đặc biệt các “cá thể” bất động sản là không giống nhau gây ra việc định giá nhiều lúc không phù hợp. Nhận thấy điều đó, em quyết định chọn đề tài này, nhằm tạo ra ứng dụng và đóng góp vào phương pháp, đưa ra số liệu tham khảo về giá bất động sản.  Ứng dụng sẽ lấy dữ liệu từ các trang nổi tiếng đăng tin mua bán bất động sản, người dùng có nhu cầu tham khảo về giá hay định giá bất động sản của bản thân có thể cung cấp thông tin và nhận kết quả.  **2. Mục tiêu thực hiện:**  - Thu thập và xây dựng bộ dữ liệu bất động sản.  - Nghiên cứu các phương pháp để định giá bất động sản.  - Thử nghiệm các phương pháp để định giá bất động sản.  - Trực quan hóa dữ liệu bất động sản.  - Xây dựng ứng dụng định giá bất động sản.  **3. Phạm vi:**  Phạm vi môi trường: Ứng dụng desktop.  Phạm vi chức năng:  - Đối với thu thập và xử lý dữ liệu:  + Ứng dụng có khả năng tự thu thập dữ liệu và lọc theo ý muốn của người dùng.  + Ứng dụng tự động xử lý dữ liệu, chuyển đổi sang định dạng phù hợp.  - Đối với trực quan hóa dữ liệu:  + Thể hiện biểu đồ từng thuộc tính BĐS, giúp người dùng quan sát và nhận biết được BĐS của mình có “thứ hạng” bao nhiêu, tổng kết.  + Thể hiện được kết quả so sánh giữa BĐS của người dùng và BĐS khác.  + Tích hợp google map, hiển thị danh sách các BĐS trong khu vực.  - Đối với ứng dụng định giá bất động sản:  + Cho phép người dùng tùy chọn phạm vi bộ dữ liệu.  + Đưa ra kết quả định giá tham khảo cho người dùng.  + Để người dùng nắm bắt được ưu, nhược, so sánh BĐS của bản thân với các BĐS khác cùng loại trong khu vực. Rồi nắm được vì sao ứng dụng đưa ra kết quả định giá.  + Đánh giá giá trị BĐS của người dùng mà họ cung cấp có hợp lý không.  **4. Đối tượng sử dụng:**  Tất cả mọi người có nhu cầu đánh giá giá trị, định giá bất động sản tại Thành Phố Hồ Chí Minh.  **5. Phương pháp thực hiện:**  Phương pháp làm việc:  - Làm việc 1 thành viên, dưới sự chỉ dẫn của giảng viên hướng dẫn.  Phương pháp nghiên cứu:  - Phân tích nhu cầu sử dụng của đối tượng dùng.  Phương pháp công nghệ:  - Tìm hiểu xây dựng ứng dụng desktop bằng Tkinter.  - Sử dụng Azure Devops để quản lý tiến trình và tài liệu.  - Sử dụng Github, kết nối với Azure Devops để quản lý source code.  **6. Nền tảng công nghệ:**  - Python, tkinter.  - Sklearn: hỗ trợ xây dự dữ liệu và dựng model.  **7. Kết quả mong đợi:**  - Nghiên cứu và thử nghiệm được nhiều phương pháp định giá bất động sản.  - Nhiều phương pháp và kết quả gần nhau nhất có thể.  - Ứng dụng giúp người dụng dễ dàng quan sát dữ liệu, nắm bắt được các ưu, nhược điểm của BĐS của mình và nắm được cách ứng dụng định giá.  **8. Hướng phát triển của đề tài:**  - Nghiên cứu và thử nghiệm, cập nhật nhiều phương pháp hơn.  - Mở rộng số trang có thể thu thập dữ liệu.  **9. Kế hoạch làm việc:**  Thời gian thực hiện từ ngày 2/3/2021 đến ngày kết thúc môn Đồ Án 1.  Chia thành 5 giai đoạn:  Giai đoạn 1: Thu thập dữ liệu và xây dựng bộ dữ liệu.  Giai đoạn 2: Tìm kiếm và nghiên cứu các phương pháp định giá bất động sản hiện có.  Giai đoạn 3: Từ các phương pháp đã tìm hiểu, cải thiện hoặc ít nhất áp dụng vào bộ dữ liệu đã thu thập.  Giai đoạn 4: Phân tích yêu cầu và chức năng, xây dựng ứng dụng.  Giai đoạn 5: Hoàn thiện tính năng, viết báo cáo.   |  |  |  | | --- | --- | --- | | Giai đoạn | Thời gian | Công việc dự kiến | | Giai đoạn 1 | 1 tuần | - Tìm hiểu các yếu tố ảnh hưởng đến giá trị BĐS.  - Xây dựng data model và khai thác dữ liệu từ trang đăng tin BĐS nổi tiếng nhất.  - Chọn cách lưu trữ, xuất dữ liệu. | | Giai đoạn 2 | 3 tuần | - Tìm kiếm, tổng hợp các phương pháp định giá bất động sản hiện có.  - Hiểu và nhận biết được các ưu, nhược điểm của từng phương pháp. | | Giai đoạn 3 | 3 tuần | - Áp dụng phương pháp đã tìm được vào dữ liệu.  - Cải thiện phương pháp.  - Tạo thư viện với từng phương pháp. | | Giai đoạn 4 | 1 tuần | - Viết báo cáo.  - Phân tích yêu cầu và chức năng của ứng dụng.  - Viết tài liệu mô tả.  - Tìm hiểu các cách trực quan hóa dữ liệu cho phù hợp. | | Giai đoạn 5 | 4 tuần | - Xây dựng giao diện ứng dụng.  - Xây dựng chức năng trực quan hóa dữ liệu.  - Xây dựng chức năng lưu trữ không gian làm việc  - Kiểm thử ứng dụng.  - Viết báo cáo.  - Bổ sung các chức năng như: tích hợp google map, … | |  |  |  | |  |  |  | |
|  |
|  |
|  |

**NỘI DUNG**

**CHƯƠNG 1: GIỚI THIỆU CHUNG**

* 1. **Tổng quan đề tài:**
* Tên ứng dụng: Định giá bất động sản
* Chủ đề chính: Thu thâp dữ liệu và định giá bất động sản
* Nền tảng phát triển: Desktop.
  1. **Lý do chọn đề tài:**

- Bất động sản là lĩnh vực có lực lượng người quan tâm hùng hậu. Bất động sản đóng vai trò quan trọng trong nền kinh tế, góp phần thu hút vốn đầu tư trong và ngoài nước, là cầu nối giữa người mua và bán, thúc đẩy phát triển khoa học kỹ thuật, … Hay cụ thể hơn, bất động sản có mặt và ảnh hưởng đến hầu hết mọi hoạt động của con người như: cư trú, làm việc, khai thác, canh tác, … Vì vậy, định giá bất động sản là bài toán được chú ý rất nhiều bởi những người có nhu cầu giao dịch bất động sản. Nhưng giá trị của bất động sản bị tác động bởi rất nhiều yếu tố. Đặc biệt các “cá thể” bất động sản là không giống nhau gây ra việc định giá nhiều lúc không phù hợp. Nhận thấy điều đó, em quyết định chọn đề tài này, nhằm tạo ra ứng dụng và đóng góp vào phương pháp, đưa ra số liệu tham khảo về giá bất động sản.

- Ứng dụng sẽ lấy dữ liệu từ các trang nổi tiếng về đăng tin mua bán bất động sản, người dùng có nhu cầu tham khảo về giá hay định giá bất động sản của bản thân có thể cung cấp thông tin và nhận kết quả.

* 1. **Đối tượng sử dụng:**
* Tất cả mọi người có nhu cầu đánh giá giá trị, định giá bất động sản tại Việt Nam.
  1. **Phạm vi nghiên cứu:**
     1. **Phạm vi môi trường:** Ứng dụng desktop
     2. **Phạm vi chức năng:**

- Đối với thu thập và xử lý dữ liệu:

+ Ứng dụng có khả năng tự thu thập dữ liệu và lọc theo ý muốn của người dùng.

+ Ứng dụng tự động xử lý dữ liệu, chuyển đổi sang định dạng phù hợp.

- Đối với trực quan hóa dữ liệu:

+ Thể hiện biểu đồ từng thuộc tính BĐS, giúp người dùng quan sát và nhận biết được BĐS của mình có “thứ hạng” bao nhiêu, tổng kết.

+ Thể hiện được kết quả so sánh giữa BĐS của người dùng và BĐS khác.

+ Tích hợp google map, hiển thị danh sách các BĐS trong khu vực.

- Đối với ứng dụng định giá bất động sản:

+ Cho phép người dùng tùy chọn phạm vi bộ dữ liệu.

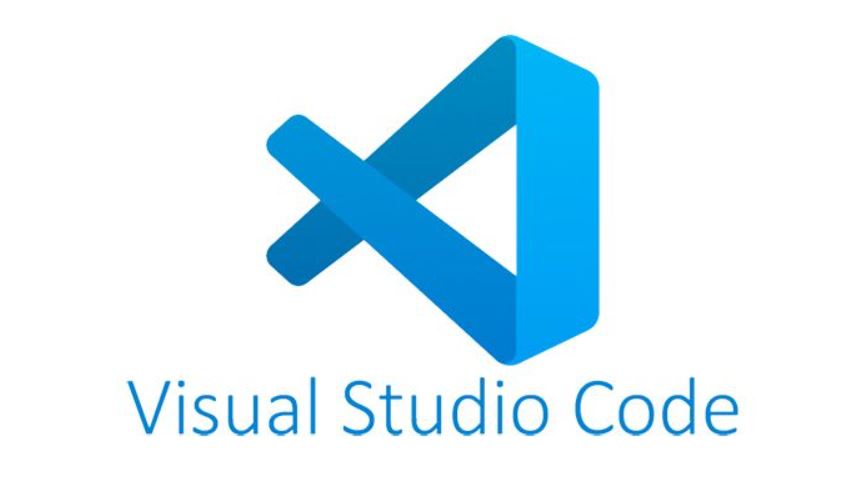
+ Đưa ra kết quả định giá tham khảo cho người dùng.

+ Để người dùng nắm bắt được ưu, nhược, so sánh BĐS của bản thân với các BĐS khác cùng loại trong khu vực. Rồi nắm được vì sao ứng dụng đưa ra kết quả định giá.

+ Đánh giá giá trị BĐS của người dùng mà họ cung cấp có hợp lý không.

**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

**2.1 Visual Studio Code:**



Hình 2.1-1: Visual studio code

Visual studio code là trình soạn thảo phổ biến nhất hiện nay. Hỗ trợ đầy đủ các chức năng cơ bản như debug, kết nối với source control, có syntax highlighting, gợi ý mã, snippets, ... Thêm vào đó, Visual studio code có cộng đồng người dùng lớn và đóng góp phát triển, tạo ra nhiều extention để hỗ trợ quá trình phát triển ứng dụng.

Hỗ trợ đa nền tảng như Linux, Window, MacOS.

**2.2 Python:**



Hình 2.2-1: Python

Python là ngôn ngữ lập trình hỗ trợ lập trình hướng đối tượng và lập trình hàm. Python thường được sử dụng cho máy học, phân tích dữ liệu, trực quan hóa dữ liệu, phát triển ứng dụng, phát triển ứng dụng web, phân tích định lượng và định tính trong tài chính, ...

Python là ngôn ngữ dễ học, quen thuộc với đa số người lập trình. Một chương trình được viết bằng python ngắn gọn, dễ hiểu hơn rất nhiều so với ngôn ngữ lập trình khác. Đặc biệt, python là 1 trong những lựa chọn hàng đầu cho các project về máy học vì số lượng framework, thư viện hỗ trợ liên quan có thể được tiếp cận rất dễ dàng.

**2.3 Tkinter:**



Hình 2.3-1: Tkinter

Tkinter là 1 GUI package của python. Tkinter là lựa chọn phù hợp cho các dự án đơn giản, có giao diện không quá phức tạp. Nó rất dễ tiếp cận, quá trình làm quen sử dụng nhanh hơn rất nhiều so với các GUI package khác như PyQt, wxPython, Pyforms, ...

**2.4 Scikit-learn (Sklearn):**

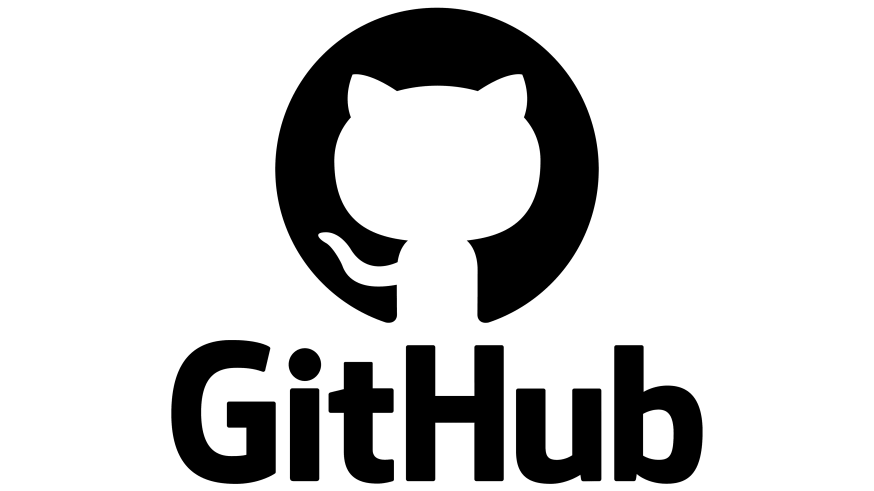


Hình 2.4-1: Sciki-learn (Sklearn)

Sklearn là thư viện hỗ trợ các thuật toán máy học của python. Thư viện này tích hợp rất nhiều thuật toán để người dùng có thể giải quyết bài toán của mình bằng cách đơn giản nhất.

Đây là 1 trong những thư viện mạnh nhất về máy học, được tin tưởng sử dụng bởi nhiều công ty nổi tiếng như: Spotify, Booking.com, J.P Morgan, ... Thư viện này được nghiên cứu quản lý rất chặt chẽ, do đó nó rất phù hợp kể cả là sản phẩm prototype hay production.

**2.5 Github:**



Hình 2.5-1: GitHub

GitHub là một hệ thống quản lý dự án và phiên bản code, hoạt động giống như một mạng xã hội cho lập trình viên. Các lập trình viên có thể clone lại mã nguồn từ một repository và Github chính là một dịch vụ máy chủ repository công cộng, mỗi người có thể tạo tài khoản trên đó để tạo ra các kho chứa của riêng mình để có thể làm việc. Github có đầy đủ những tính năng của Git, ngoài ra nó còn bổ sung những tính năng về social để các developer tương tác với nhau.

Github cung cấp các tính năng social networking như feeds, followers, và network graph để các developer học hỏi kinh nghiệm của nhau thông qua lịch sử commit

**2.6 Selenium:**



Hình 2.6-1: Selenium

Selenium là một trong những công cụ kiểm thử tự động tốt nhất hiện nay cho ứng dụng web.

Không chỉ kiểm thử, Selenium còn được sử dụng để crawl dữ liệu. Điểm mạnh của selenium là có thể tương tác với trang web như người dùng thật, giúp ta truy cập các thông tin bị ẩn đi bởi javascript.

**CHƯƠNG 3: THU THẬP VÀ XỬ LÝ DỮ LIỆU**

**3.1 Thu thập dữ liệu:**

**3.1.1: Thông tin nơi thu thập dữ liệu:**

- Thu thập dữ liệu từ trang batdongsan.com.vn.

- Trang batdongsan.com.vn không có API (hoặc đã ẩn), không thể lấy dữ liệu bằng API.

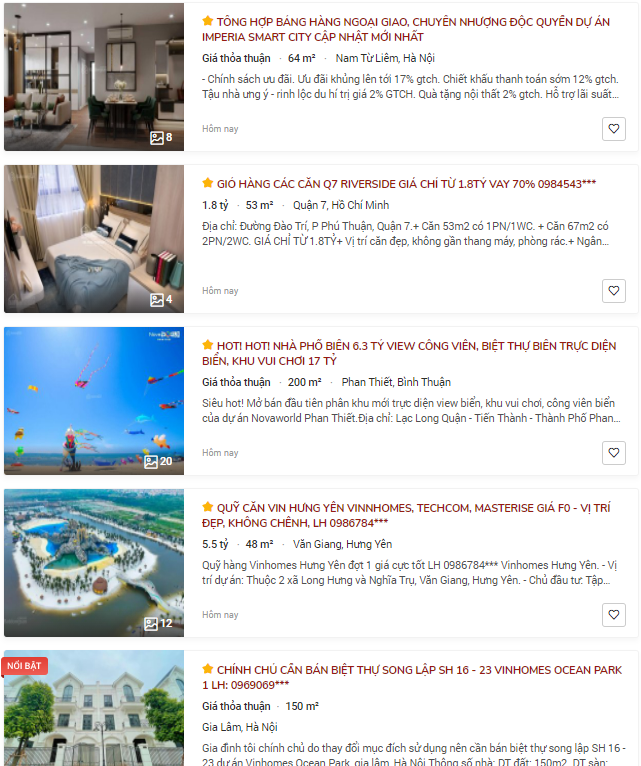
- Thay vào đó phân tích cấu trúc html của website, tách từng thông tin để có được dữ liệu.

- Trang batdongsan.com.vn có bảo mật khá tốt, họ có thể chặn các request từ nhiều thư viện. Thư viện có thể dùng được là selenium.

**3.1.2: Thu thập dữ liệu với Selenium:**

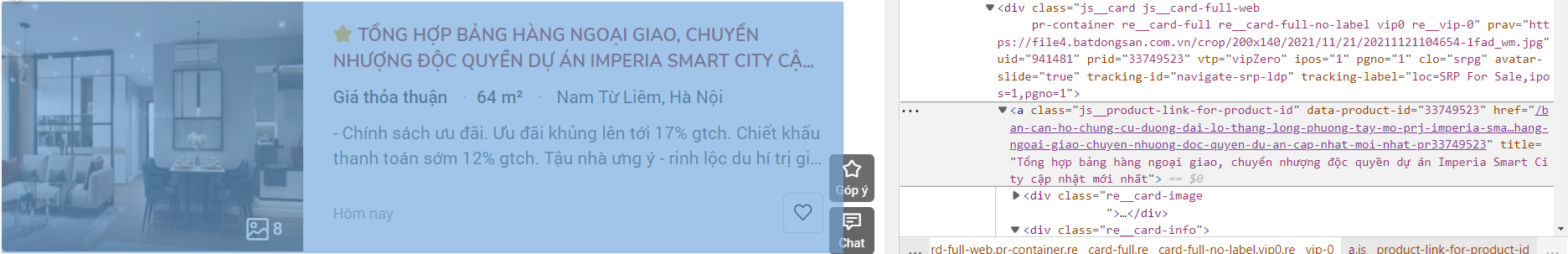
- Bước 1: Tìm hiểu cấu trúc trang batdongsan.com.vn

+ Các bài đăng được hiển thị theo danh sách:



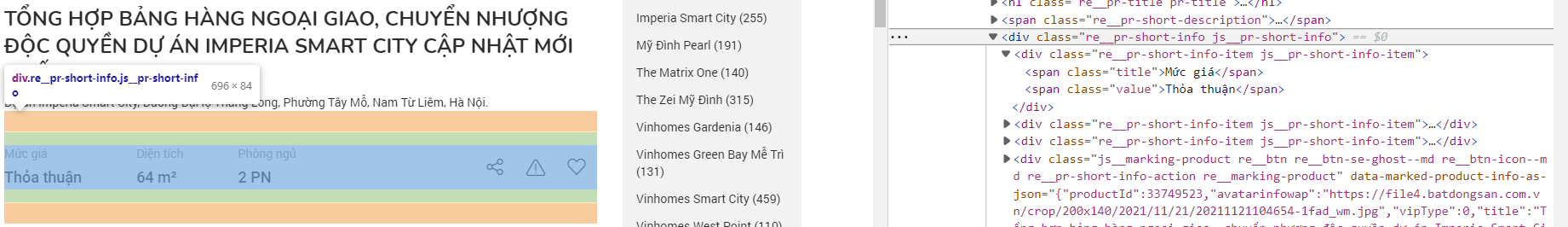
Hình 3.1.2-1: Danh sách tin đăng bất động sản

+ Ta tìm được đường dẫn đến chi tiết bài đăng:



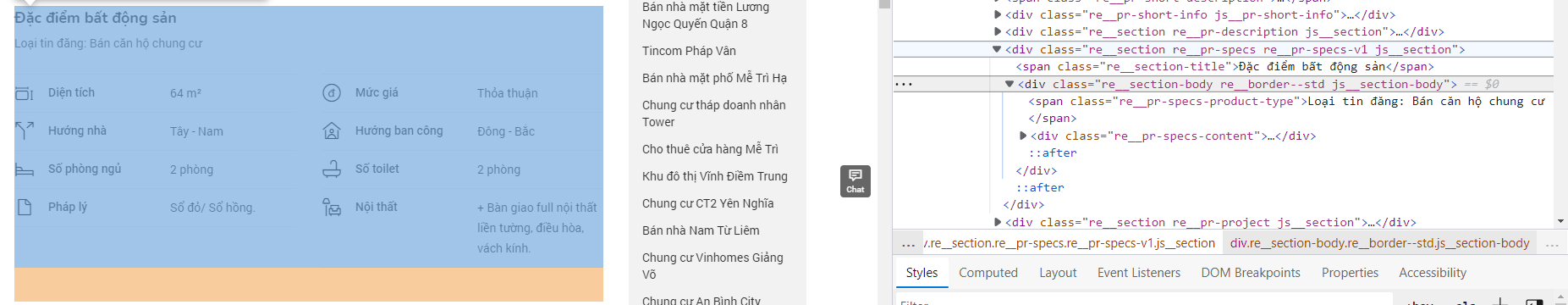
Hình 3.1.2-2: Đường dẫn đến chi tiết bài đăng

+ Trong trang chi tiết bài đăng, ta tìm được thông tin nhanh của bất động sản:



Hình 3.1.2-3: Thông tin nhanh của chi tiết bài đăng

+ Trong trang chi tiết bài đăng, ta tìm được đặc điểm bất động sản (đây là nơi ta sẽ thu thập dữ liệu). Class re\_\_pr-specs-content-item-title chứa tên “đặc trưng”, re\_\_pr-specs-content-item-value chứa dữ liệu.

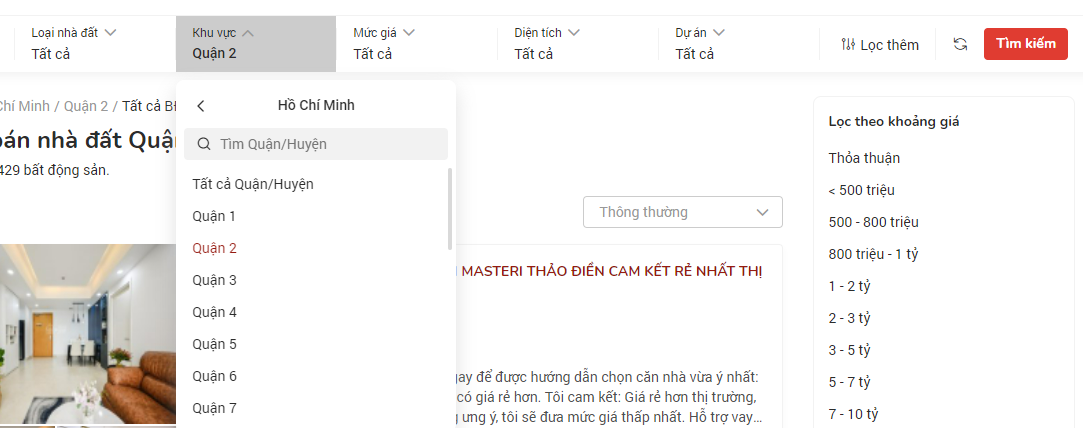


Hình 3.1.2-4: Đặc điểm bất động sản

**Ta đã tìm được nơi để thu thập dữ liệu, nhưng làm sao để lọc/tìm kiếm bài đăng phù hợp?**

- Bước 2: Thu thập các đường dẫn để “tìm kiếm/filter”

+ Cho selenium click từng mục trong combobox ở phần tìm kiếm, và lưu trữ lại đường dẫn.



Hình 3.1.2-5: Combobox thanh tìm kiếm





Hình 3.1.2-6: Đường dẫn cần lưu lại

**Vậy ta đã có được các đường dẫn của trang web phục vụ cho chức năng “tìm kiếm/filter” của ứng dụng**

- Bước 3: Sau khi đã hiểu trang web và có được danh sách để “tìm kiếm/filter” ta có thể xây dựng ứng dụng hoàn toàn tự động để lấy dữ liệu.

**3.2 Quan sát và tiền xử lý dữ liệu:**

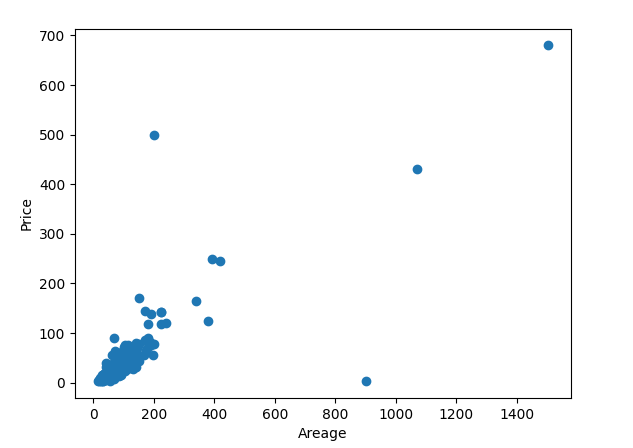
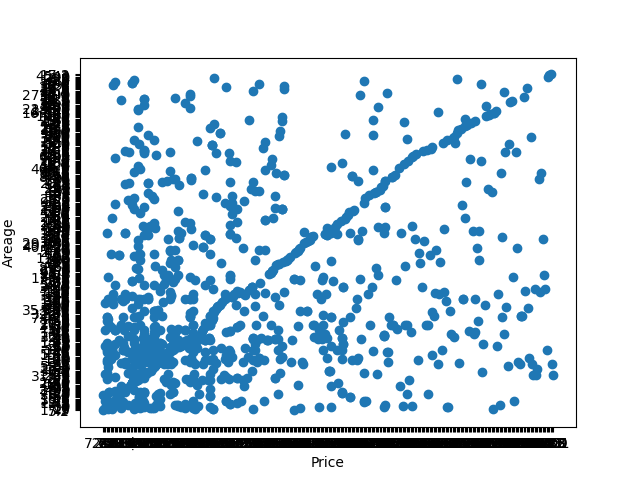
- Tiền xử lý dữ liệu là giai đoạn rất quan trọng để giải quyết các bài toán machine learning. Khi thu thập dữ liệu, khả năng cao các dữ liệu sẽ không được đầy đủ, nhiều thuộc tính sẽ bị thiếu đi. Đặc biệt là khi lấy dữ liệu từ trang đăng tin mua bán, mỗi cá nhân sẽ có cách viết và cung cấp thông tin bất động sản theo cách khác nhau, việc này làm cho dữ liệu thu thập được không đồng nhất.

- Các vấn đề đã gặp và kỹ thuật tiền xử lý dữ liệu:

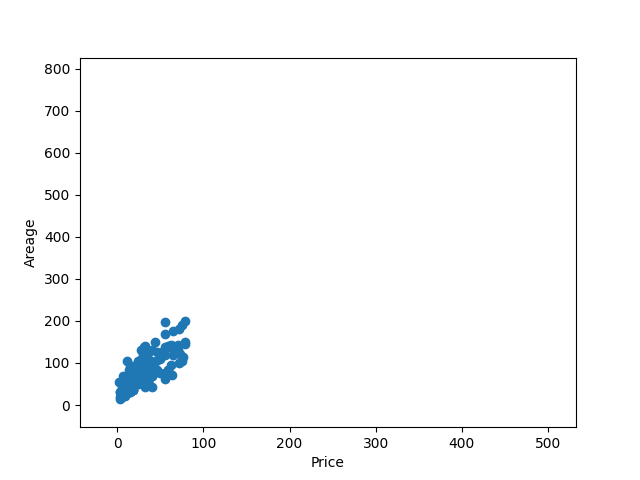
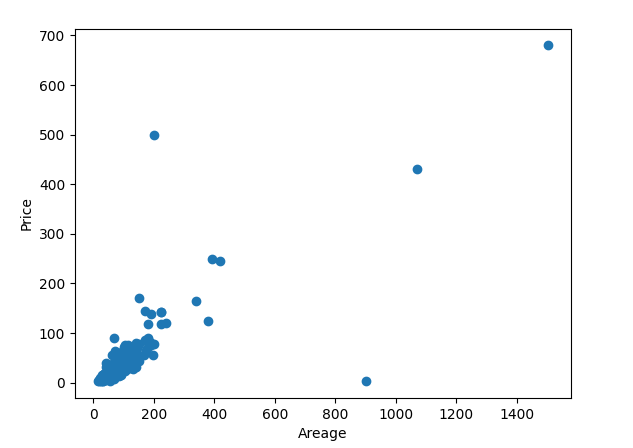
+ Dữ liệu bị khuyết: Các ô dữ liệu có giá trị NULL là các ô bị khuyết, có các phương pháp để xử lý vấn đề này:

+ Giữ nguyên dữ liệu bị khuyết và sử dụng model có khả năng hiểu và xử lý phần khuyết.

+ Xóa bớt dữ liệu bị khuyết, hoặc xóa bớt đặc trưng của dữ liệu.



+ Xóa đi các dòng dữ liệu nhiễu.



+ Điền dữ liệu bị khuyết bằng giá trị trung bình hoặc bằng số giá trị có tần suất xuất hiện lớn nhất.

**Ví dụ: ở dòng dữ liệu 1, số nhà vệ sinh là NULL**, 3 dòng còn lại lần lượt là 3, 2, 2 nhà vệ sinh. Ta chọn 1 trong 2 cách sau:

Điền dữ liệu khuyết bằng giá trị trung bình: [(3 + 3 + 2) / 3] = 2

Điền dữ liệu khuyết bằng tầng suất xuất hiện:

Giá trị 2 xuất hiện 2 lần. **(Chọn)**

Giá trị 3 xuất hiện 1 lần.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | Giá | Diện Tích | Phòng Ngủ | Nhà Vệ Sinh | Tầng |
| 1 | 10 | 126 | NULL | NULL | 1 |
| 2 | 15 | 152 | 2 | 3 | 1 |
| 3 | 17 | 116 | 1 | 2 | 3 |
| 4 | 11 | 133 | 3 | 2 | 1 |

Sau khi sử lí, bảng sẽ thành:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | Giá | Diện Tích | Phòng Ngủ | Nhà Vệ Sinh | Tầng |
| 1 | 10 | 126 | 2 | 2 | 1 |
| 2 | 15 | 152 | 2 | 3 | 1 |
| 3 | 17 | 116 | 1 | 2 | 3 |
| 4 | 11 | 133 | 3 | 2 | 1 |

+ Dữ liệu phân bố rời rạc:

Lấy lại mẫu.

Áp dụng thuật toán phân cụm vào dữ liệu

**3.3 Lưu trữ dữ liệu:**

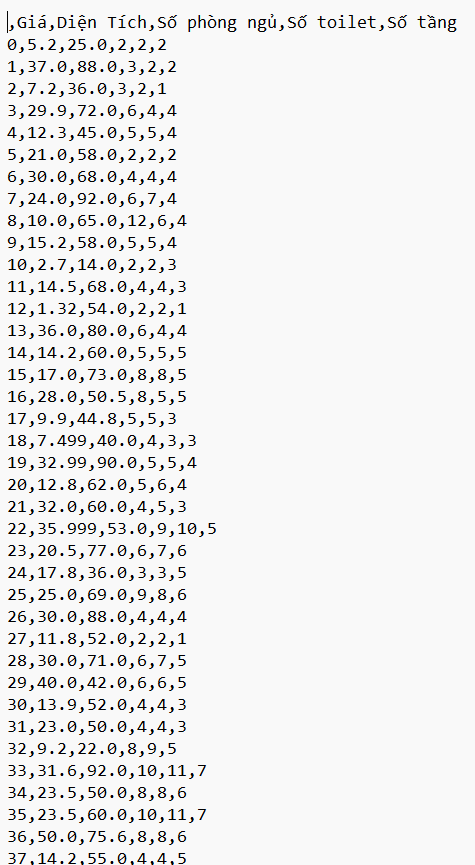
- Dữ liệu thu thập được từ selenium chưa thực sự ở dạng có thể lưu trữ và tái sử dụng. Ta sẽ xử lý và lưu trữ dữ liệu ở dạng file (\*.csv).

- File (\*.csv) có tên đầy đủ là Comma Separated Values, tức các giá trị được phân cách bởi dấu phẩy. Tuy nhiên, ta vẫn có thể linh động điều chỉnh thành các ký hiệu khác.

- Cấu trúc file (\*.csv):

+ **Dòng đầu:** Là cột đầu trong bảng tính, biểu thị tên cột.

**+ Các dòng sau:** Có cấu trúc tương đương và thứ tự tương ứng với nội dung của các cột giá trị trong bảng.



Hình 3.3-1: Cấu trúc file (\*.csv)

**4. Thuật toán định giá bất động sản:**

**4.1 Cây hồi quy (Regression Tree):** Cây hồi quy được dựa trên cây quyết định (decision tree).

**a/ Cây hồi quy là cây mà:**

- Là cây nhị phân.

- Root (gốc): là nút đại diện cho mẫu, là nút không có nút cha.

- Decision node (nút quyết định): là nút có nút con, hay nút được tách ra thành nút con từ hành động “splitting”.

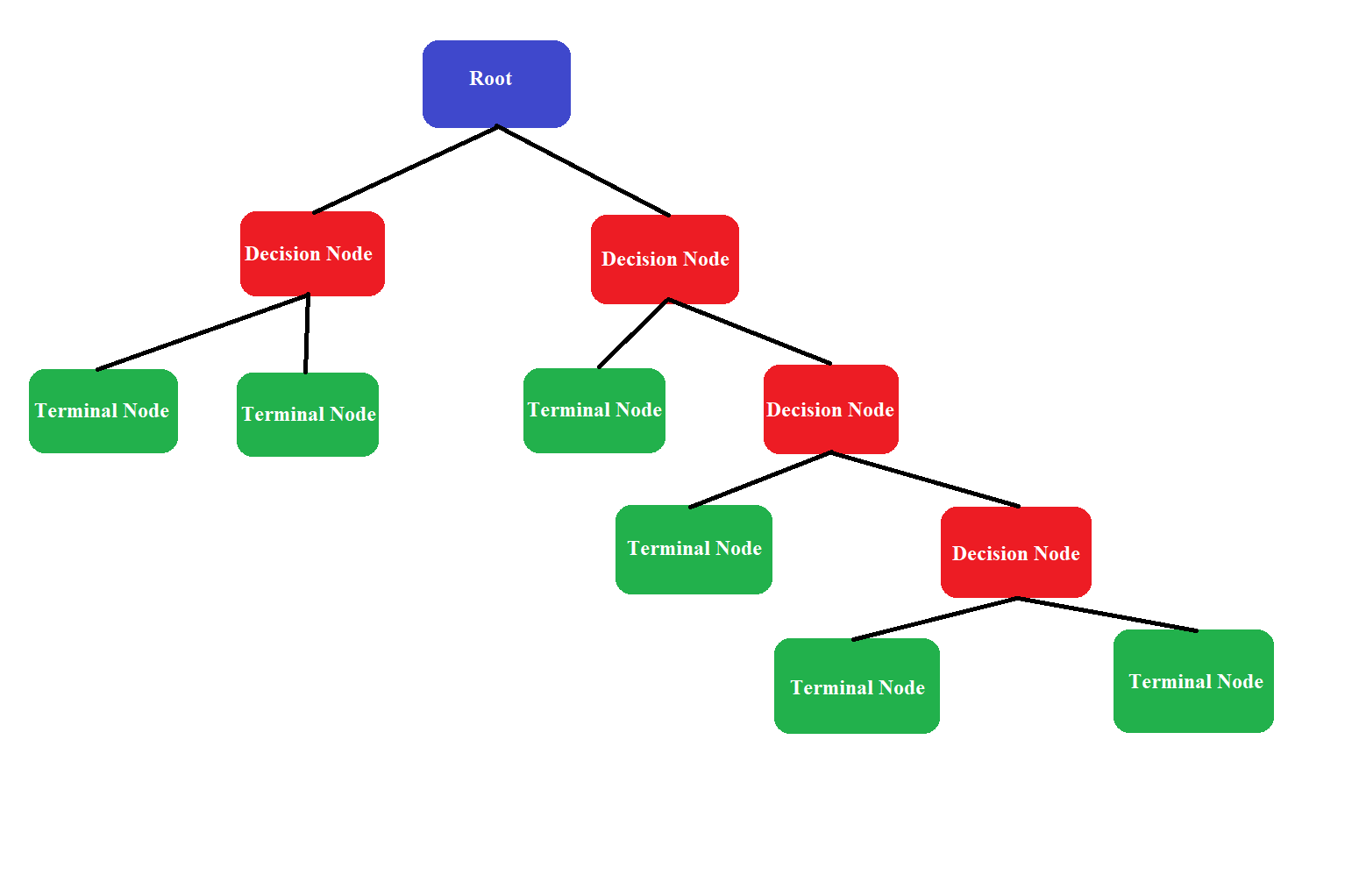
- Leaf / Terminal node (lá, nút cuối): là nút không có nút con, nút lá, hay nút này ta không “splitting”.

- Lá con bên trái là “Đúng”, lá con bên phải là “Sai”.

**b/ Các thao tác với cây hồi quy:**

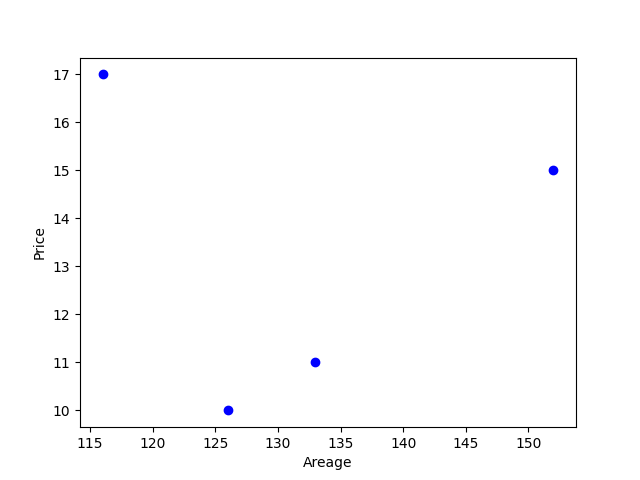
- Splitting (tách, chẻ): là hành động khi chia 1 nút thành nhiều nút con.

- Pruning (cắt tỉa): là hành động khi ta xóa đi nút con của nút quyết định, ngược với hành động “splitting”.



Hình 4-1: Cây hồi quy

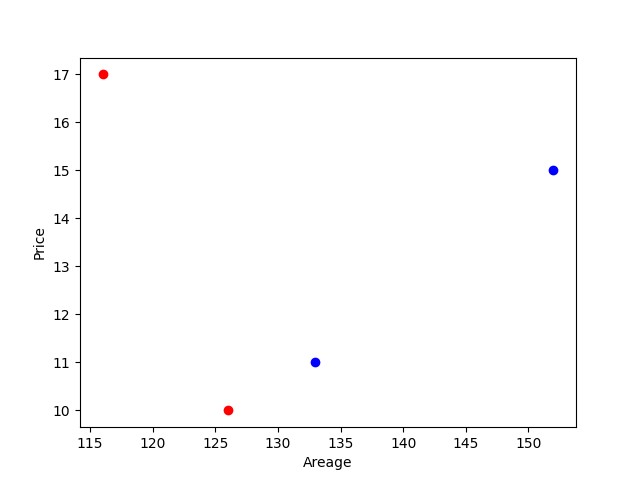
**c/ Dựng cây hồi quy:**

 Ví dụ ta có dữ liệu như sau:

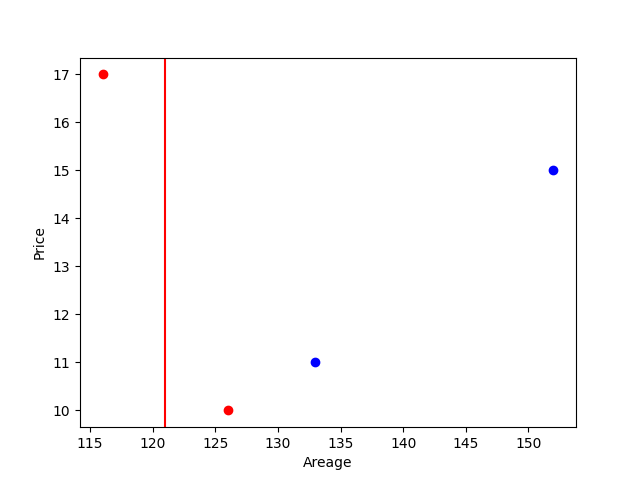
|  |  |
| --- | --- |
| Giá | Diện Tích |
| 10 | 126 |
| 15 | 152 |
| 17 | 116 |
| 11 | 133 |

**Ta sẽ phải tạo root cho cây (hay tìm ứng cử viên cho splitting), nhưng bằng cách nào?**

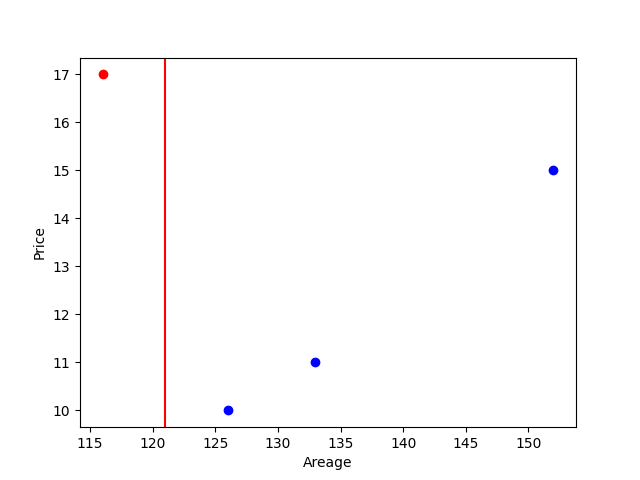
Ban đầu ta có 1 tập gồm 4 điểm (giá, diện tích) là (10;126) (11;133) (15;152) (17;116) đã được sắp xếp theo diện tích.



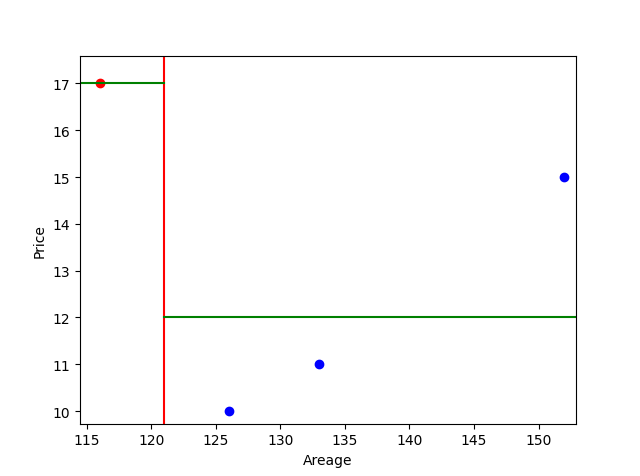
Ta chọn 2 điểm đầu tiên là (17;116) (10;126)



Trung bình diện tích của 2 điểm này là (116 + 126) / 2 = 121 m2



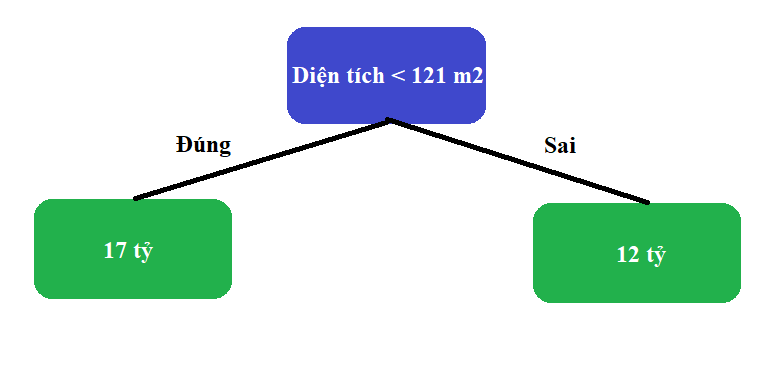
Chia 4 điểm thành 2 tập là đỏ S1 và xanh S2, tách nhau bởi đường thẳng x = 121



Trung bình “Giá” của tập S1 = 17/1 = 17 tỷ.

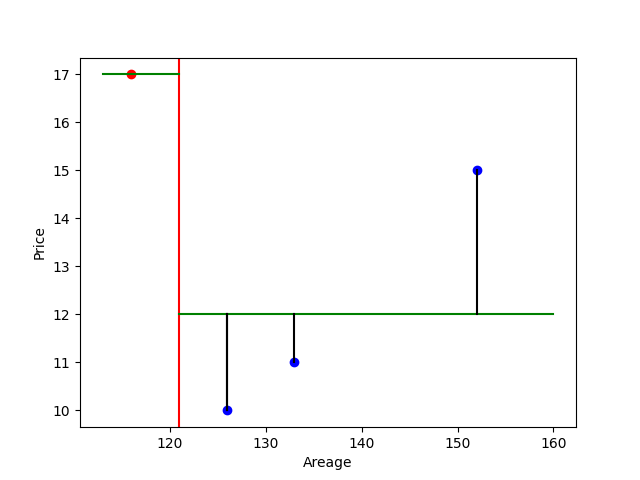
Trung bình “Giá” của tập S2 = (10+11+15)/3 = 12 tỷ.

Lúc này ta có cây là:



Nhưng đây liệu có là “root” tốt? Để tìm được root tốt, ta tính giá trị “residual”. Residual là khoảng cách thẳng đứng từ điểm đến đường hồi quy (đường màu xanh lá trên đồ thị).

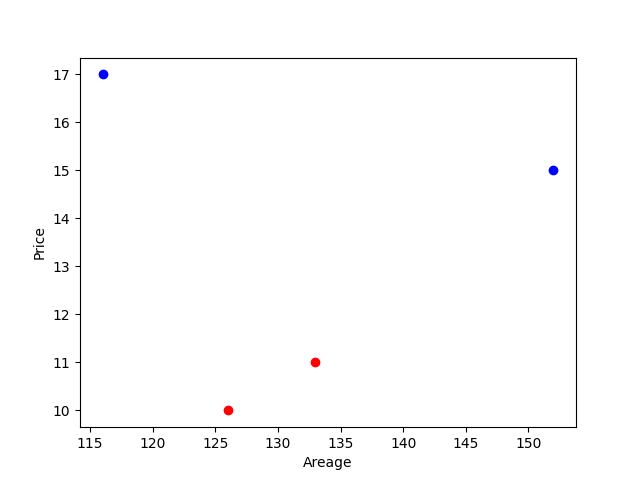
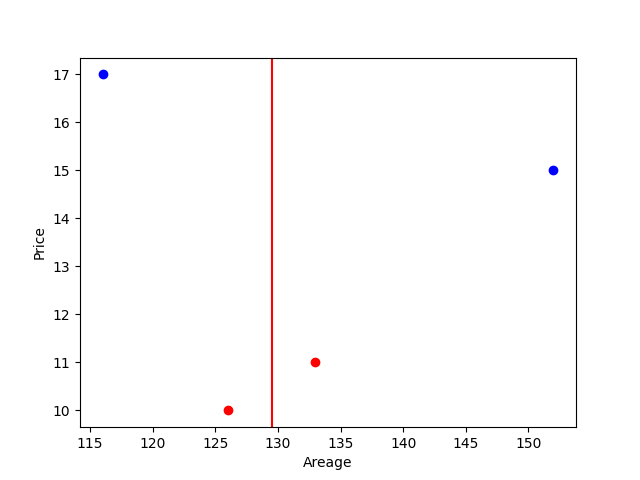
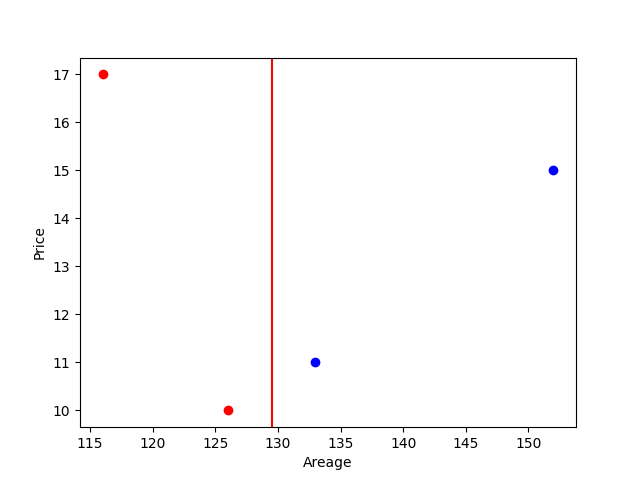
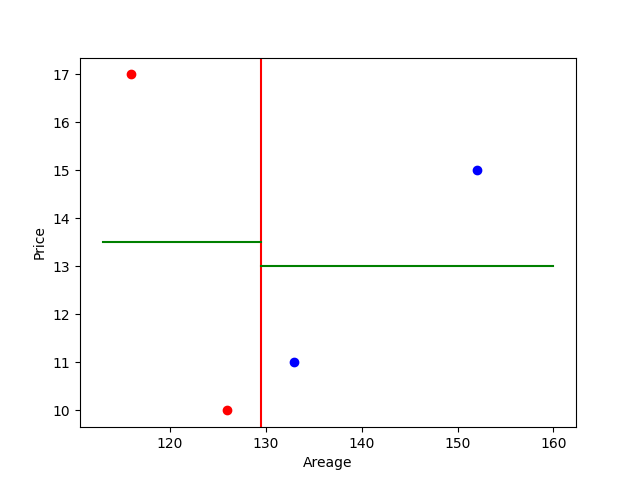
Ta tính tổng các residual khi chọn root là trung bình diện tích của 2 điểm đầu tiên.

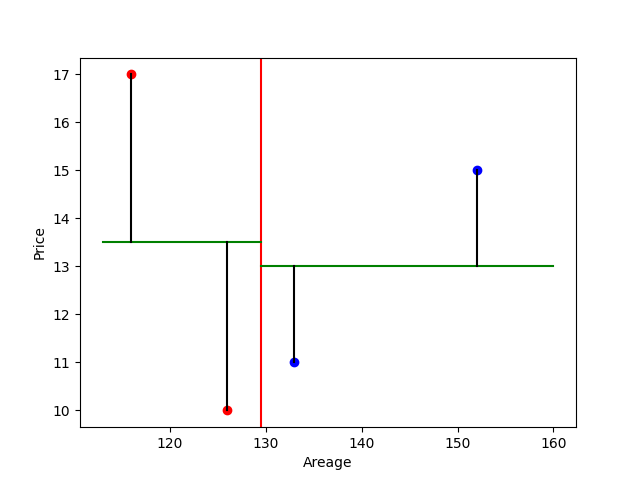


Tổng residual = (17 – 17)2 + (10 – 12)2 + (11 – 12)2 + (15 – 12)2 = 4 + 1 + 9 = 14

Tương tự, ta lần lượt chọn các cặp điểm tiếp theo, lần lượt chia thành 2 tập đỏ và xanh, vẽ đường hồi quy, tính tổng residual. Cặp nào có tổng residual bé nhất thì trung bình diện tích của cặp đó sẽ là root.

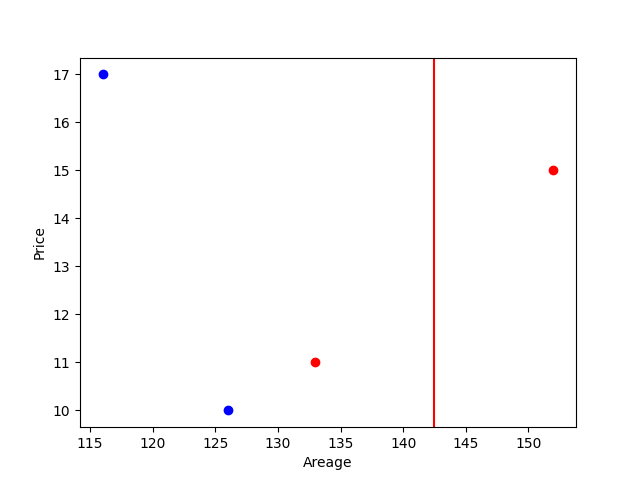
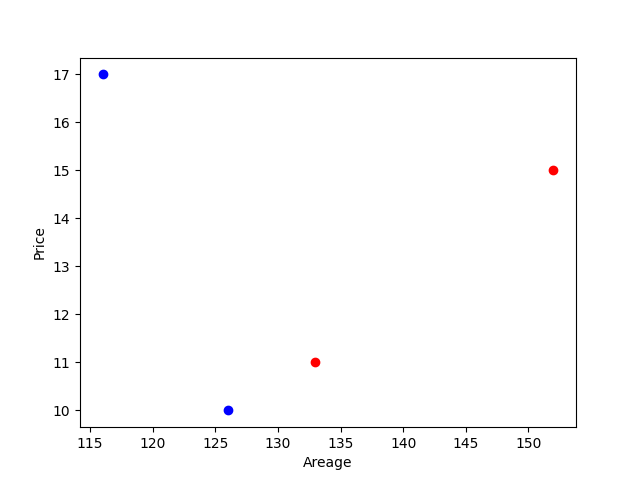
Cặp điểm tiếp theo (10,126) (11,133)

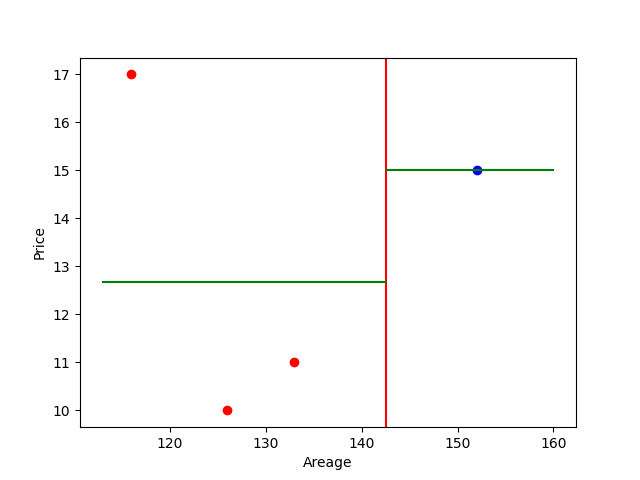
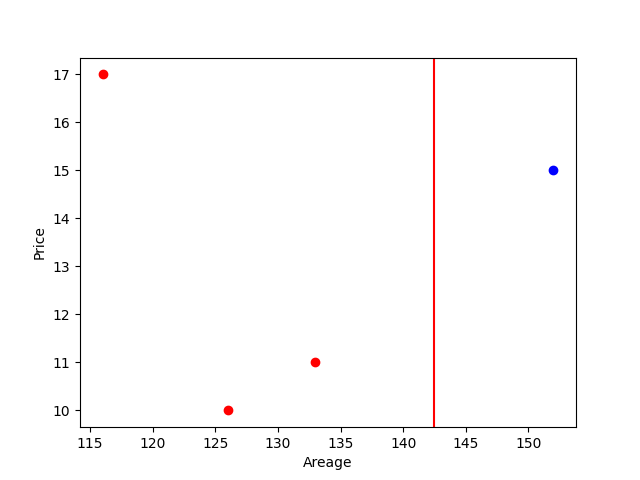
   

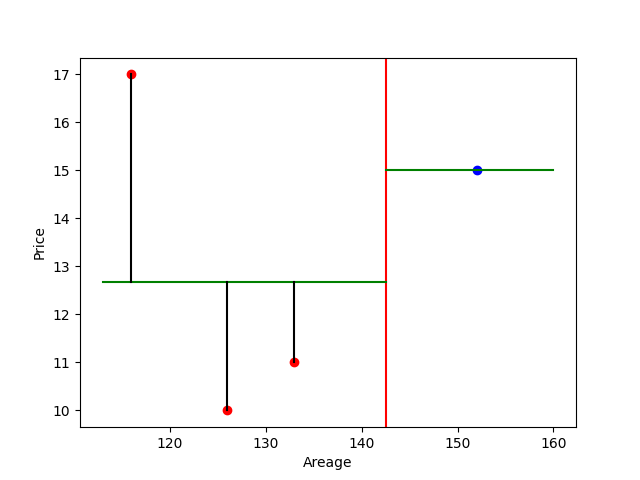


Tổng residual = (17 – 13.5)2 + (10 – 13.5)2 + (11 – 13)2 + (15 – 13)2 = 12.25 + 12.25 + 4 + 4 = **32.5**

Cặp tiếp theo (11, 133) (15, 152)

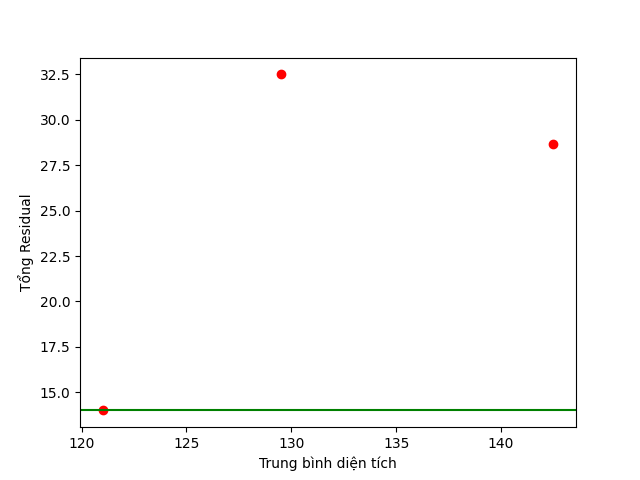




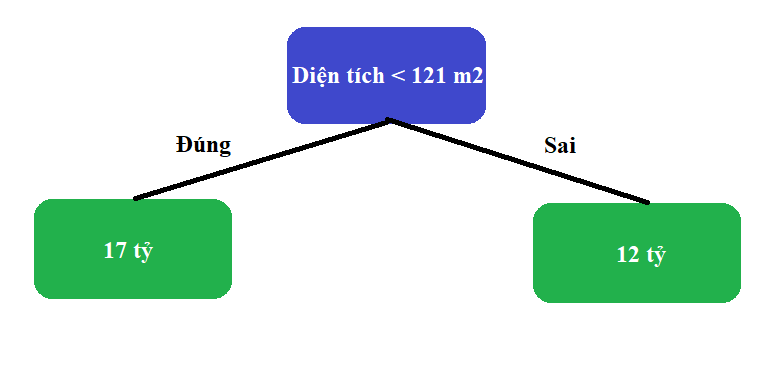


Tổng residual = (17 – 12.67)2 + (10 – 12.67)2 + (11 – 12.67)2 + (15 – 15)2 = 28.67

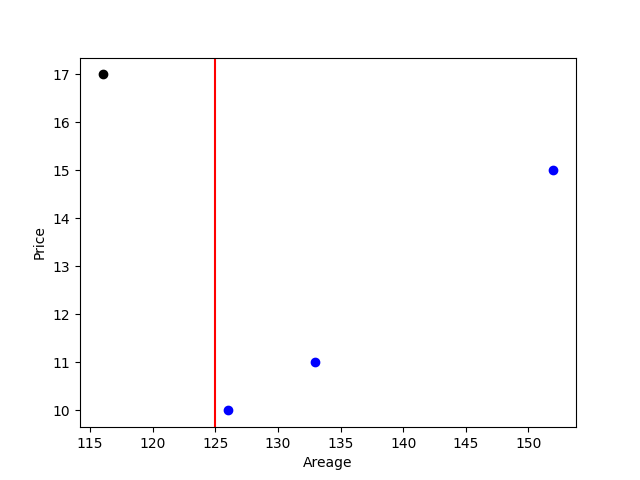
Từ 3 cặp điểm trên ta có:



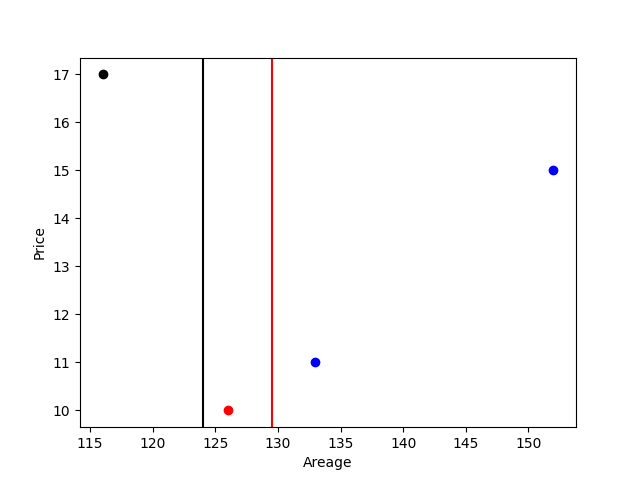
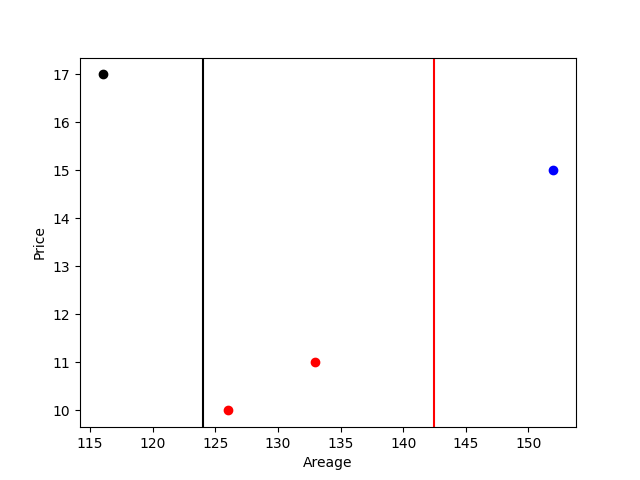
Trung bình diện tích của cặp điểm có tổng residual là 14 (bé nhất) sẽ là root (giống như kết quả ban đầu)



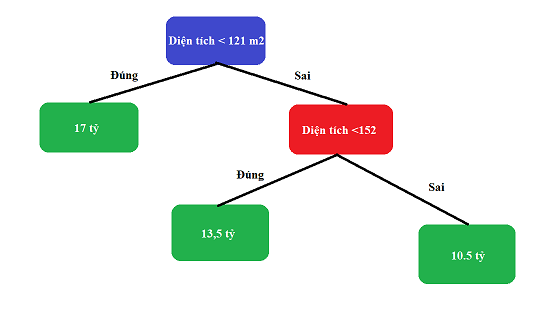
Kết quả ở đây ta chỉ mới có 3 node và với diện tích >= 121 m2, cây hiện tại cho ra kết quả 12 tỷ. Cây vẫn chưa “ổn”. Ta tiến hành “splitting” nút lá 12 tỷ.



Tập điểm với diện tích >= 121m2 gồm 3 điểm màu xanh. Ta sẽ tiến hành chọn lần lượt từng cặp điểm màu xanh rồi tính trung bình diện tích, chia 2 tập đỏ và xanh, tính tổng residual như trên.

Ta được cây là:

****

- Ta có thể split các nút, mỗi lần split sẽ tăng 1 đơn vị chiều cao của cây. Split quá ít làm cây không hoàn thiện, nhưng split quá nhiều cũng không tốt, có thể gây ra overfitting. Nên việc lựa chọn độ cao phù hợp cho cây rất quan trọng.

- Một số phương pháp hạn chế split:

+ Khi tập dữ liệu được chia đủ nhỏ, ta ngưng không cho cây split nữa.

- Với bộ dữ liệu có nhiều đặc tính hơn, không chỉ mỗi diện tích:

+ Với mỗi đặc tính, tìm root/ứng cử viên tốt nhất.

+ Sắp xếp các ứng cử viên và split cây theo tổng residual của root/ứng cử viên tốt nhất giữa các đặc tính.

**4.2 Hồi quy tuyến tính (Linear Regression)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | Giá | Diện Tích | Phòng Ngủ | Nhà Vệ Sinh | Tầng |
| 1 | 10 | 126 | 2 | 2 | 1 |
| 2 | 15 | 152 | 2 | 3 | 1 |
| 3 | 17 | 116 | 1 | 2 | 3 |
| 4 | 11 | 133 | 3 | 2 | 1 |

**Giới thiệu:**

***Ký hiệu:***

* Biến phụ thuộc, giá trị thực: Giá (y).
* Biến độc lập, đặc trưng: Diện tích (x1), Phòng Ngủ (x2), Nhà Vệ Sinh (x3), Tầng (x4).
* Kết quả dự đoán của mô hình: y’

***Mục tiêu:***

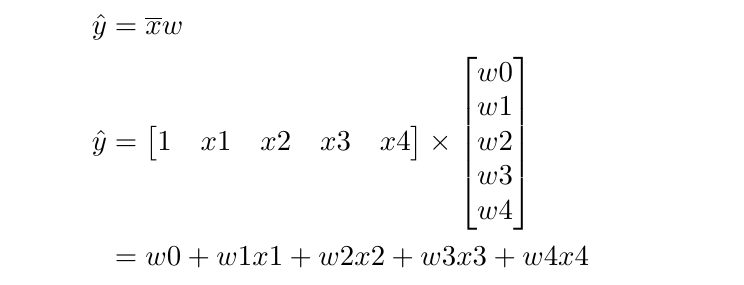
* Kết quả dự đoán gần nhất với giá trị thực: y ≈ y’

***Phương trình hồi quy tuyến tính:***

y' = f(x) = w1x1 + w2x2 + w3x3 + w4x4 + w0

- Vì hàm số f(x) có dạng đường thẳng, nên được gọi là tuyến tính.

- **w** = {w0, w1, w2, w3, w4} là các hằng số, w0 gọi là bias. Là các số cần phải tìm.



**Bài toán:** Tìm **w** = {w0, w1, w2, w3, w4} sao cho sai số giữa y’ và y gần 0 nhất:

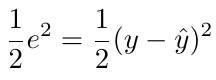


Để giá trị không âm, ta có 2 cách:

* Lấy giá trị tuyệt đối.
* Bình phương. (Ta chọn cách này, vì phương trình bậc 2 luôn có đạo hàm).



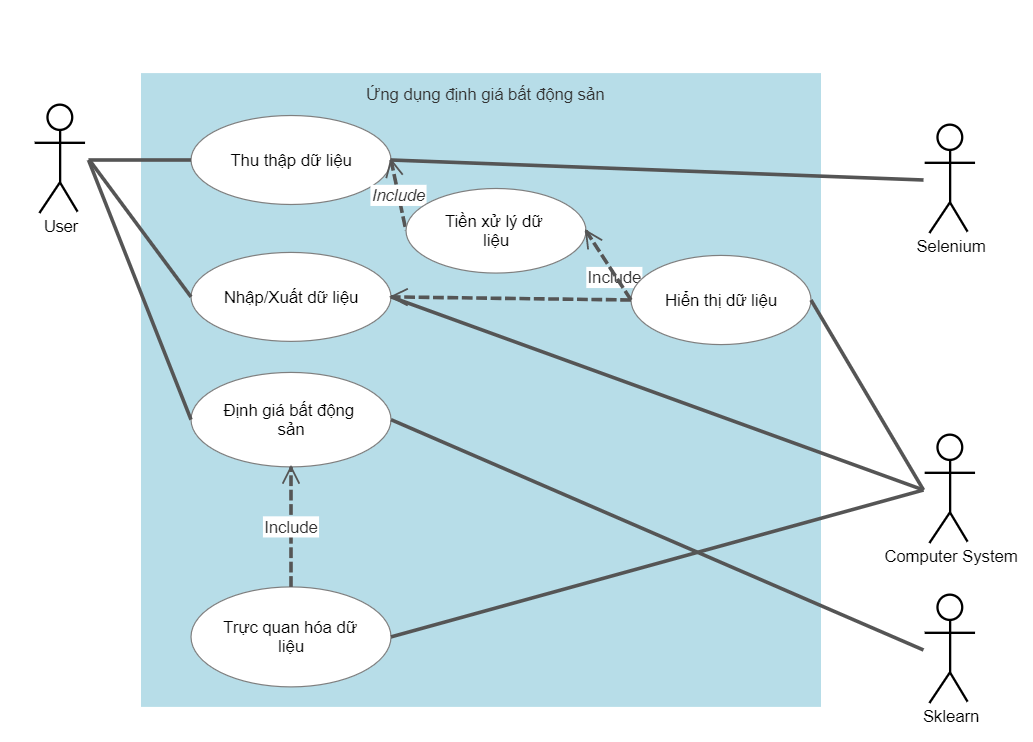
Vì bước sau sẽ tiến hành đạo hàm nên ta nhân thêm ½ để kết quả đẹp hơn.



Chưa xong

CHƯƠNG 5: ĐẶC TẢ USE CASE

5.1 Sơ đồ Use Case:



5.2 Danh sách tác nhân:

|  |  |  |
| --- | --- | --- |
| **STT** | **Tác nhân** | **Mô tả/ Ghi chú** |
| 1 | User | Người dùng ứng dụng |
| 2 | Selenium | Phần hệ thống thu thập và xử lý dữ liệu |
| 3 | Sklearn | Phần hệ thống train model và trả kết quả dự đoán |
| 4 | Computer System | Phần hệ thống xử lý giao tiếp với người dùng |

5.3 Danh sách Use Case:

|  |  |  |
| --- | --- | --- |
| STT | Tính năng | Mô tả/ Ghi chú |
| 1 | Thu thập dữ liệu | Cho phép User thu thập dữ liệu và lọc. Do Selenium thu thập theo yêu cầu User. |
| 2 | Tiền xử lý dữ liệu | Computer System tự động xử lý dữ liệu thu được từ Selenium. |
| 3 | Hiển thị dữ liệu | Computer System hiển thị dữ đã được xử lý. |
| 4 | Nhập/Xuất dữ liệu | Cho phép User thêm, xuất dữ liệu từ file. |
| 5 | Định giá bất động sản | Cho phép User nhập thông tin bất động sản cần dự đoán. Do Sklearn chạy thuật toán và trả kết quả định giá cho User. |
| 6 | Trực quan hóa dữ liệu | Hiển thị dữ liệu cho User sau khi Sklearn trả kết quả định giá. |

**5.4 Mô tả chi tiết Use Case:**

|  |  |
| --- | --- |
| **Mô tả chi tiết tính năng “Thu thập dữ liệu”** | |
| Mô tả chi tiết | - Người dùng cung cấp thông tin các đặc điểm bất động sản tương đồng với bất động sản của mình. Người dùng đặt ra các giới hạn để lọc dữ liệu sẽ lấy.  - Computer System sẽ nhận các thông tin của Người dùng đã cung cấp và chuyển yêu cầu cho Selenium.  - Selenium thu thập dữ liệu và trả về cho Computer System. |
| **Luồng sự kiện** | |
| Luồng chính | 1. Người dùng chọn thành phố từ combobox.  2. Người dùng nhập dữ liệu vào textbox để đưa ra giới hạn cho các đặc trưng là giá trị số.  3. Người dùng bấm nút “Crawl Data”.  4. Computer System chuyển dữ liệu người dùng đã nhập cho Selenium.  5. Selenium thu thập dữ liệu theo yêu cầu của người dùng và trả dữ liệu về cho Computer System.  6. Chuyển sang Use Case: “Tiền xử lý dữ liệu” |
| Luồng phụ | 3. Người dùng chưa nhập thông tin hay nhập sai định dạng.  5. Selenium gặp vấn đề khi thu thập dữ liệu do người dùng can thiệp. |
| Yêu cầu đặc biệt | Không |
| Điều kiện trước | Không |
| Điều kiện sau | Selenium thành công thu thập dữ liệu |
| Mở rộng | Không |

|  |  |
| --- | --- |
| **Mô tả chi tiết tính năng “Tiền xử lý dữ liệu”** | |
| Mô tả chi tiết | - Computer System nhận dữ liệu từ Selenium và tiến hành tiền xử lý dữ liệu. |
| **Luồng sự kiện** | |
| Luồng chính | 1. Computer System nhận dữ liệu.  2. Computer System chuyển dữ liệu thô nhận được thành bảng.  3. Computer System áp dụng các phương pháp tiền xử lý dữ liệu đã được cài đặt sẵn vào bảng.  4. Chuyển sang Use Case: “Hiển thị dữ liệu” |
| Luồng phụ | 1. Computer System không nhận được dữ liệu.  2. Lỗi sai định dạng dữ liệu.  3. Lỗi sai định dạng dữ liệu. |
| Yêu cầu đặc biệt | Không |
| Điều kiện trước | Selenium thành công thu thập dữ liệu |
| Điều kiện sau | Computer system chạy xong các thuật toán |
| Mở rộng | Không |

|  |  |
| --- | --- |
| **Mô tả chi tiết tính năng “Hiển thị dữ liệu”** | |
| Mô tả chi tiết | - Computer System xử lý xong dữ liệu nhận được từ Selenium.  - Hoặc Computer System đọc dữ liệu từ file người dùng cung cấp. |
| **Luồng sự kiện** | |
| Luồng chính | 1.  a/ Computer System xử lý xong dữ liệu nhận được từ Selenium.  b/ Computer System đọc dữ liệu từ file với đường dẫn người dùng cung cấp.  2. Computer System đổ dữ liệu vào bảng và hiển thị. |
| Luồng phụ | 1.  b/ Dữ liệu của người dùng không hợp lệ hoặc file không tồn tại. |
| Yêu cầu đặc biệt | Không |
| Điều kiện trước | Computer System xử lý dữ liệu từ Selenium  Hoặc  Người dùng chọn đọc dữ liệu từ file |
| Điều kiện sau | Hiển thị thành công dữ liệu lên bảng |
| Mở rộng | Không |

|  |  |
| --- | --- |
| **Mô tả chi tiết tính năng “Nhập/Xuất”** | |
| Mô tả chi tiết | - Người dùng có thể thêm dữ liệu vào bảng từ file.  - Người dùng có thể xuất dữ liệu ở bảng ra file. |
| **Luồng sự kiện** | |
| Luồng chính | 1.  a/ Người dùng chọn nhập dữ liệu.  b/ Người dùng chọn xuất dữ liệu.  2. Computer System hiển thị explorer để người dùng chọn đường dẫn.  3.  a/ Computer System đọc file từ đường dẫn và thực hiện Use Case “Hiển thị dữ liệu”  b/ Computer System xử lý dữ liệu từ bảng và xuất ra file (\*.csv) |
| Luồng phụ | 2. Người dùng tắt explorer.  3.  a/ File không tồn tại, File sai định dạng.  b/ Đường dẫn không tồn tại. |
| Yêu cầu đặc biệt | Không |
| Điều kiện trước |  |
| Điều kiện sau | a/ Hiển thị thành công dữ liệu lên bảng  b/ Xuất thành công dữ liệu ra file (\*.csv) |
| Mở rộng | Không |

|  |  |
| --- | --- |
| **Mô tả chi tiết tính năng “Định giá bất động sản”** | |
| Mô tả chi tiết | - Người dùng có thể điền thông tin bất động sản để định giá.  - Sklearn thực hiện định giá bất động sản. |
| **Luồng sự kiện** | |
| Luồng chính | 1. Người dùng nhập dữ liệu bất động sản.  2. Người dùng bấm nút “Predict”.  3. Computer System chuyển dữ liệu cho Sklearn để dự đoán.  4. Sklearn hoàn thành dự đoán và trả kết quả cho Computer System.  5. Thực hiện Use Case: “Trực quan hóa dữ liệu”. |
| Luồng phụ | 2. Dữ liệu nhập từ người dùng không hợp lệ. |
| Yêu cầu đặc biệt | Không |
| Điều kiện trước | Bảng dữ liệu phải có dữ liệu. |
| Điều kiện sau | Hiển thị trang trực quan hóa dữ liệu |
| Mở rộng | Không |

|  |  |
| --- | --- |
| **Mô tả chi tiết tính năng “Trực quan hóa dữ liệu”** | |
| Mô tả chi tiết | - Computer System hiển thị kết quả định giá.  - Computer System hiển thị các sơ đồ thể hiện sự liên quan giữa dữ liệu và bất động sản của người dùng cũng như kết quả dữ đoán. |
| **Luồng sự kiện** | |
| Luồng chính | 1. Computer System hiển thị kết quả dự đoán và các sơ đồ.  2. Người dùng chọn nút xem các loại sơ đồ khác nhau. |
| Luồng phụ | Không |
| Yêu cầu đặc biệt | Không |
| Điều kiện trước | Sklearn thành công định giá bất động sản. |
| Điều kiện sau | Thành công hiển thị kết quả định giá và biểu đồ |
| Mở rộng | Không |

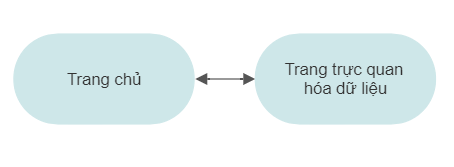
**CHƯƠNG 6: THIẾT KẾ DỮ LIỆU**

**6.1 Danh sách các dữ liệu:**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thực thể** | **Kiểu dữ liệu** | **Ghi chú đơn vị** |
| 1 | Gia | float | tỷ |
| 2 | DienTich | float | m2 |
| 3 | SoPhongNgủ | int | phòng |
| 4 | SoToilet | int | phòng |
| 5 | SoTang | int | tầng |

**CHƯƠNG 7: THIẾT KẾ GIAO DIỆN NGƯỜI DÙNG**

**7.1 Luồng giao diện:**



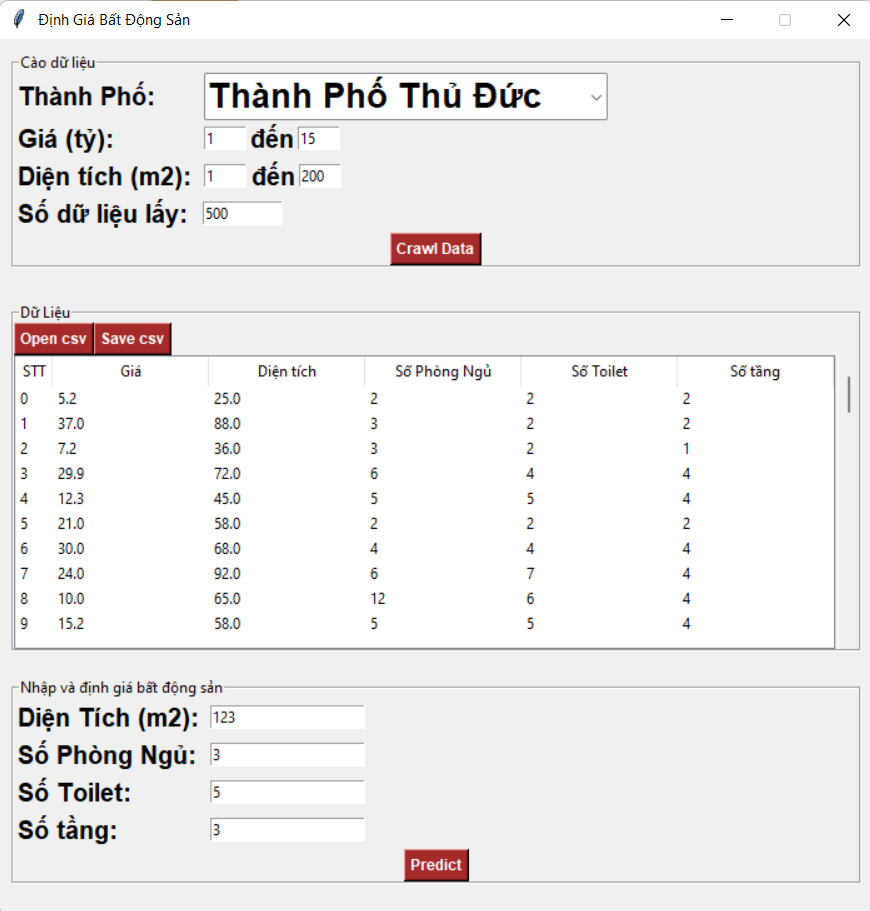
Hình 7.1-1: Luồng giao diện

**7.2 Danh sách màn hình:**

|  |  |  |
| --- | --- | --- |
| STT | Tên màn hình | Loại |
| 1 | Trang chủ | Hiện thị và Nhập liệu |
| 2 | Trang trực quan hóa dữ liệu | Hiển thị |

**7.3 Mô tả chi tiết từng màn hình:**

* + 1. **Trang chủ:**



Hình 7.3.1-1: Trang chủ

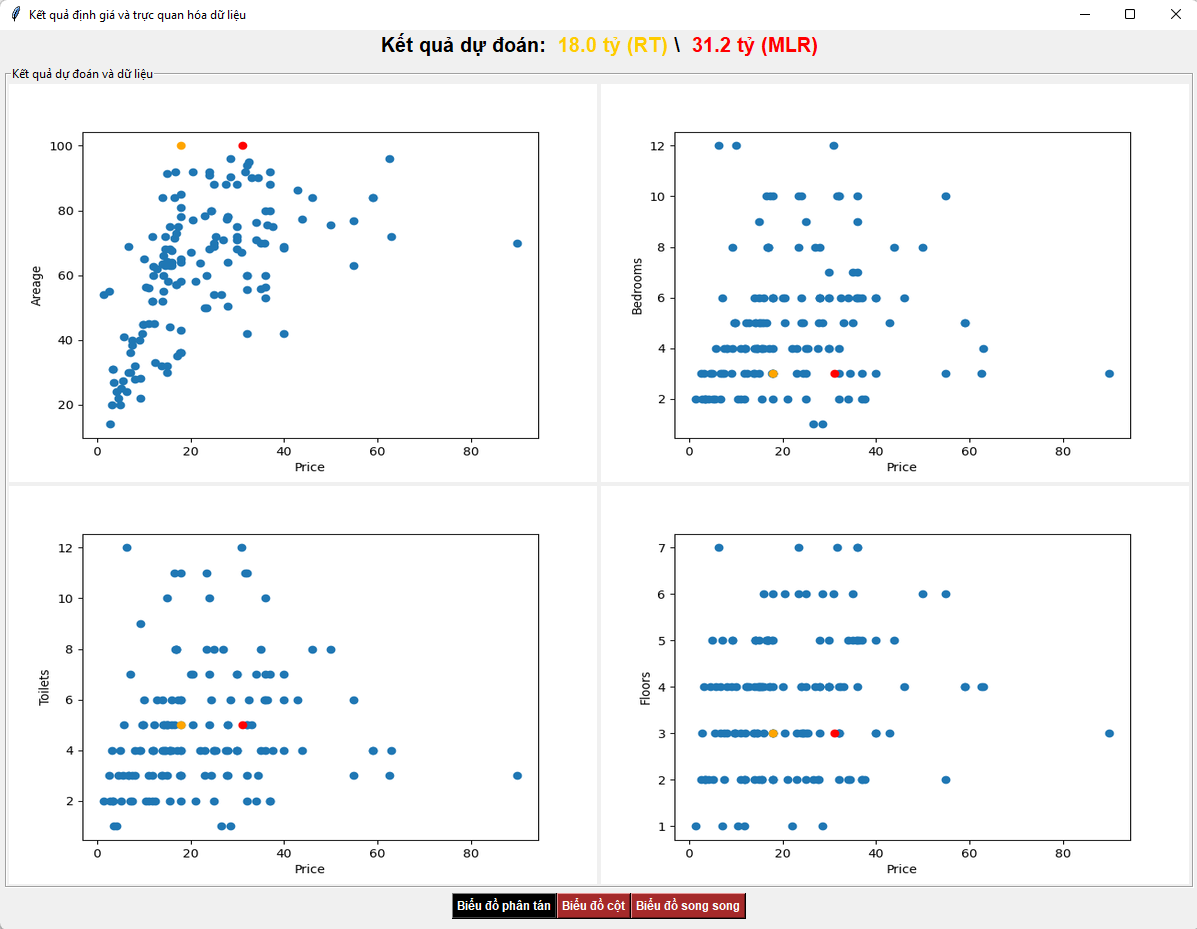
**a. Mô tả tổng quát:**

Đây là nơi cho phép người thu thập và “filter” dữ liệu, nhập bất động sản cần dự đoán.

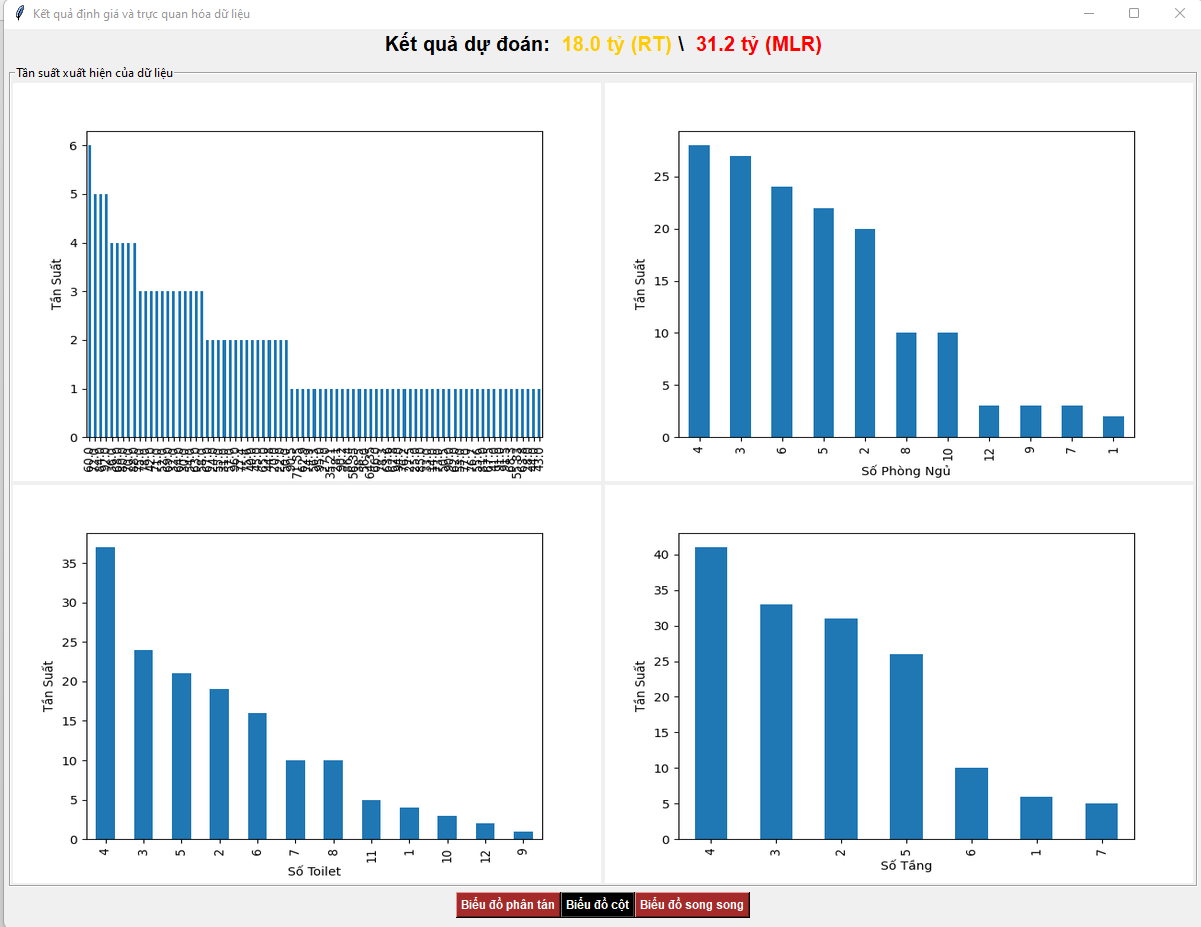
**b. Các đối tượng trên màn hình:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dành cho người dùng** | | | |
| **STT** | **Tên đối tượng** | **Loại** | **Chức năng** |
| 1 | Cào dữ liệu - Thành phố | Combobox | Filter – Thành phố |
| 2 | Cào dữ liệu – Giá | TextBox | Filter – Giá |
| 3 | Cào dữ liệu – Diện tích | TextBox | Filter – Diện Tích |
| 4 | Cào dữ liệu – Số dữ liệu lấy | TextBox | Giới hạn số lượng dữ liệu lấy |
| 5 | Cào dữ liệu – Crawl Data | Button | Tự động lấy dữ liệu và filter dữ liệu |
| 6 | Hiển thị dữ liệu – Bảng dữ liệu | TreeView | Hiển thị dữ liệu đã thu thập hoặc từ file (\*.csv) |
| 7 | Hiển thị dữ liệu – Open csv | Button | Đọc dữ liệu từ file (\*.csv) |
| 8 | Hiển thị dữ liệu – Save csv | Button | Lưu dữ liệu từ bảng thành file (\*.csv) |
| 9 | Nhập và định giá bđs – Diện tích | TextBox | Đặc điểm bất động sản cần dự đoán |
| 10 | Nhập và định giá bđs – Số Phòng Ngủ | TextBox | Đặc điểm bất động sản cần dự đoán |
| 11 | Nhập và định giá bđs – Số Toilet | TextBox | Đặc điểm bất động sản cần dự đoán |
| 11 | Nhập và định giá bđs – Số Tầng | TextBox | Đặc điểm bất động sản cần dự đoán |
| 12 | Nhập và định giá bđs – Predict | Button | Dự đoán |

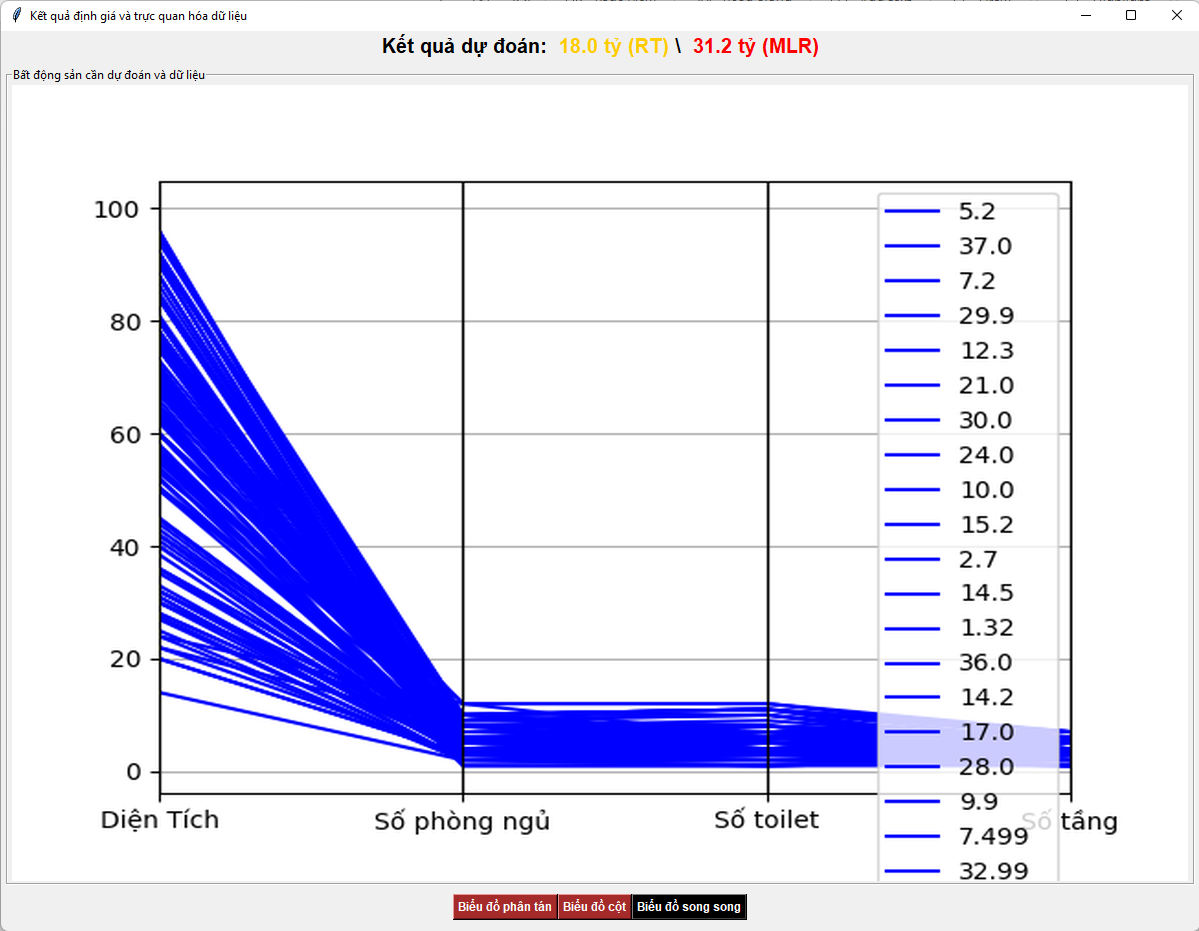
**7.3.2: Trang trực quan hóa dữ liệu – Biểu đồ phân tán:**



Hình 7.3.2-1: Trang trực quan hóa dữ liệu – Biểu đồ phân tán



Hình 7.3.2-2: Trang trực quan hóa dữ liệu – Biểu đồ cột

 Hình 7.3.2-3: Trang trực quan hóa dữ liệu – Biểu đồ song song

a. Mô tả tổng quan:

Đây là nơi người dùng nhận được kết quả dự đoán và các biểu đồ.

b. Các đối tượng trên màn hình

|  |  |  |  |
| --- | --- | --- | --- |
| **Dành cho người dùng** | | | |
| **STT** | **Tên đối tượng** | **Loại** | **Chức năng** |
| 1 | Kết quả dự doán - RT | Text | Kết quả định giá bởi Regression Tree |
| 2 | Kết quả dự doán - MLR | Text | Kết quả định giá bởi Multiple Linear Regression |
| 3 | Biểu đồ cột | TextBox | Trực quan hóa dữ liệu |
| 4 | Biểu đồ phân tán | TextBox | Trực quan hóa dữ liệu |
| 5 | Biểu đồ song song | TextBox | Trực quan hóa dữ liệu |
| 6 | Button - Biểu đồ cột | TextBox | Chuyển đổi biểu đồ |
| 7 | Button - Biểu đồ phân tán | Button | Chuyển đổi biểu đồ |
| 8 | Button - Biểu đồ song song | Button | Chuyển đổi biểu đồ |

**CHƯƠNG 8: CÀI ĐẶT VÀ KIỂM THỬ**

8.1 Môi trường cài đặt và kiểm thử:

- Thiết bị cài đặt: Window, MacOs, Linux.

8.2 Kết quả kiểm thử:

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tính năng** | **Mức độ hoàn thiện** | **Ghi chú** |
| 1 | Crawl dữ liệu |  |  |
| 2 | Save dữ liệu |  |  |
| 3 | Đọc dữ liệu từ file |  |  |
| 4 | Định giá bất động sản |  |  |
| 5 | Trực quan hóa dữ liệu |  |  |

**CHƯƠNG 9: KẾT LUẬN**

**9.1 Tổng kết:**

- Dữ liệu rất quan trọng cho việc định giá bất động sản, việc có 1 bộ dữ liệu đầy đủ và chính xác là rất khó, thậm chí phải bỏ tiền ra mua mới có được dữ liệu phù hợp.

- Cần đầu tư hơn cho việc thu thập và xử lý dữ liệu, ban đầu chỉ xác định 1 tuần cho việc thu thập nhưng kết quả thực là 1 tháng.

- Phải xác định thuật toán áp dụng để định giá, thuật toán phải phù hợp với bài toán, phải phù hợp với bộ dữ liệu thì kết quả mô hình mới tốt.

- Trực quan hóa dữ liệu cũng rất quan trọng, đóng vai trò lớn trong việc hiểu dữ liệu từ người lập trình và hiểu dữ liệu, kết quả từ người sử dụng.

- Cân nhắc thay đổi lưu trữ dữ liệu bằng file csv thành cơ sở dữ liệu, vì file csv không thể lưu kiểu dữ liệu, truy xuất chậm, ...

- Có khả năng đọc được tài liệu tiếng anh là một lợi thế vì tất cả nguồn tài liệu tìm hiểu trong quá trình nghiên cứu, thử nghiệm và tạo nên ứng dụng đều được viết bằng tiếng anh.

**9.2 Khó khăn:**

- Không có kinh nghiệm trong hầu hết giai đoạn.

- Không đủ nguồn để thu thập thêm nhiều đặc trưng.

- Khó khăn trong việc trích xuất địa chỉ bất động sản để đưa vào google map.

**9.3 Kết quả đạt được:**

**-** Chưa hoàn thành hết các chức năng đưa ra từ đầu.

- Học được ngôn ngữ python và nhiều thư viện về machine learning.

- Cải thiện khả năng đọc tiếng anh.

9.4 Hạn chế của đồ án:

- Ứng dụng còn sơ sài, không có nhiều chức năng.

- Chưa có chức năng cải thiện mô hình đã train.

- Chưa cho phép người dùng sở hữu bộ dữ liệu chuẩn.

9.5 Hướng phát triển:

- Thu thập nhiều dữ liệu và đặc tính hơn để tăng độ chính xác.

- Cung cấp cho người dùng khả năng tải bộ dữ liệu chuẩn có sẵn từ nhà phát triển.

- Đầu tư giao diện.

- Tối ưu hóa code.

**TÀI LIỆU THAM KHẢO**

Tkinter

[1] [tkinter — Python interface to Tcl/Tk — Python 3.10.4 documentation](https://docs.python.org/3/library/tkinter.html)

[2] [Python - GUI Programming (Tkinter) (tutorialspoint.com)](https://www.tutorialspoint.com/python/python_gui_programming.htm)

Sklearn

[1] [API Reference — scikit-learn 1.1.1 documentation](https://scikit-learn.org/stable/modules/classes.html)

[2] [StatQuest!!! - An epic journey through statistics and machine learning](https://statquest.org/)

Pandas

[1] [pandas documentation — pandas 1.4.2 documentation (pydata.org)](https://pandas.pydata.org/docs/)

Numpy

[1] [NumPy documentation — NumPy v1.22 Manual](https://numpy.org/doc/stable/)

Selenium

[1] [The Selenium Browser Automation Project | Selenium](https://www.selenium.dev/documentation/)

Python

[1] [Python Tutorial (tutorialspoint.com)](https://www.tutorialspoint.com/python/index.htm)

[2] [Python Tutorial (w3schools.com)](https://www.w3schools.com/python/)

PyInstaller

[1] [Standalone Python EXE Executable - Python Tkinter GUI Tutorial #40 - YouTube](https://www.youtube.com/watch?v=QWqxRchawZY)

Machine learning Algorithm

[1] [Machine Learning cơ bản (machinelearningcoban.com)](https://machinelearningcoban.com/)

[2] Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems