TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO DỰ ÁN CNTT 2**

**DỰ ĐOÁN SỰ THAY ĐỔI VÀ GIÁ NHÀ ĐẤT TRONG THỊ TRƯỜNG BẤT ĐỘNG SẢN**

*Giảng viên hướng dẫn*: **Tiến sĩ** **TRỊNH HÙNG CƯỜNG**

*Người thực hiện*: **NGUYỄN THÀNH ĐẠT - 51600012**

**LÊ TRẦN NHẬT VIỆT - 51603376**

Lớp **: 16050311&16050304**

Khoá  **: 20**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO DỰ ÁN CNTT 2**

**DỰ ĐOÁN SỰ THAY ĐỔI VÀ GIÁ NHÀ ĐẤT TRONG THỊ TRƯỜNG BẤT ĐỘNG SẢN**

*Giảng viên hướng dẫn*: **Tiến sĩ** **TRỊNH HÙNG CƯỜNG**

*Người thực hiện*: **NGUYỄN THÀNH ĐẠT – 51600012**

**LÊ TRẦN NHẬT VIỆT - 51603376**

Lớp **: 16050311&16050304**

Khoá  **: 20**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

LỜI CẢM ƠN

Để hoàn thành đề tài luận văn và kết thúc khóa học, em xin bày tỏ lòng biết ơn sâu sắc tới trường Đại học Tôn Đức Thắng đã tạo điều kiện cho em có môi trường học tập tốt trong suốt thời gian em học tập, nghiên cứu tại trường.

Lời nói đầu tiên em xin chân thành gởi lời cảm ơn đến thầy Trịnh Hùng Cường đã tận tình hướng dẫn và giúp đỡ bọn em trong thời gian hoàn thành đồ án. Trong quá trình làm đồ án em có nhiều sai xót mong thầy thông cảm, bỏ qua và góp ý cho em để em có thể hoàn thiện tốt hơn nữa. Đồng thời, em xin bày tỏ lòng cảm ơn tới thầy cô trong Khoa Công Nghệ Thông Tin, bạn bè đã giúp đỡ, tạo điều kiện cho em trong suốt quá trình học tập và hoàn thành Khóa luận tốt nghiệp lần này. Em xin chân thành cảm ơn.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi / chúng tôi và được sự hướng dẫn của thầy Trịnh Hùng Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Nguyễn Thành Đạt*

*Lê Trần Nhật Việt*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Thị trường bất động sản (BĐS) là một trong những thị trường quan trọng của nền kinh tế thị trường vì thị trường này liên quan trực tiếp tới một lượng tài sản cực lớn cả về quy mô, tính chất cũng như giá trị của các mặt trong nền kinh tế quốc dân. Tỷ trọng BĐS trong tổng số của cải xã hội ở các nước có khác nhau nhưng thường chiếm trên dưới 40% lượng của cải vật chất của mỗi nước. Các hoạt động liên quan đến BĐS chiếm tới 30% tổng hoạt động của nền kinh tế. Theo thống kê, ở các nước phát triển lượng tiền ngân hàng cho vay qua thế chấp bằng BĐS chiếm trên 80% trong tổng lượng vốn cho vay. Vì vậy, phát triển đầu tư, kinh doanh BĐS đóng vai trò quan trọng trong việc chuyển các tài sản thành nguồn tài chính dồi dào phục vụ cho yêu cầu phát triển kinh tế - xã hội đặc biệt là đầu tư phát triển cơ sở hạ tầng của nền kinh tế.

Đồ án Dự đoán sự thay đổi và giá bất động sản sẽ đưa ra sự dự đoán giá cả bất động sản dựa trên những dữ liệu đã thu thập được. Qua đó có thể một phần nào đưa ra sự tham khảo cần thiết cho những nghiên cứu sau này.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc387692905)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc387692906)

[TÓM TẮT i](#_Toc387692907)v

[MỤC LỤC 1](#_Toc387692908)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 4](#_Toc387692909)

[CHƯƠNG 1 – TỔNG QUAN VỀ ĐỀ TÀI 6](#_Toc387692910)

[1.1 Giới thiệu đề tài 6](#_Toc387692911)

[1.2 Ý tưởng 7](#_Toc387692916)

[CHƯƠNG 2 – TỔNG QUAN VỀ THỊ TRƯỜNG BẤT ĐỘNG SẢN 8](#_Toc387692917)

[2.1 Giới thiệu về thị trường bất động sản 8](#_Toc387692918)

[2.2 Những yếu tố ảnh hưởng đến thị trường bất động sản 12](#_Toc387692919)

2.3 Báo cáo những biến động của thị trường trong năm 2020………………....…..…13

[CHƯƠNG 3 – CƠ SỞ LÝ THUYẾT .20](#_Toc387692920)

[3.1 Kỹ thuật Crawl Data .20](#_Toc387692921)

3.1.1 Web Crawling là gì ……………………………………………….….…20

3.1.2 Thư viện BeautifulSoup ………………………………………...………20

3.1.1 Demo…………….. ………………………………………………..……21

[3.2 Các thư viện cần thiết 25](#_Toc387692922)

[3.3 Giới thiệu về Machine Learning 28](#_Toc387692922)

[3.4 Mô hình hồi quy trong Machine Learning 30](#_Toc387692923)

[CHƯƠNG 4 – QÚA TRÌNH THỰC HIỆN ĐỀ TÀI 31](#_Toc387692920)

[4.1 Thu thập dữ liệu 31](#_Toc387692921)

[4.2 Làm sạch và chuẩn bị dữ liệu 37](#_Toc387692922)

[4.3 Trực quan hóa dữ liệu 40](#_Toc387692923)

4.4 Áp dụng vào mô hình dự đoán bằng phương pháp hồi quy ………….........46

[CHƯƠNG 5 – KẾT QUẢ ĐẠT ĐƯỢC 52](#_Toc387692920)

[TÀI LIỆU THAM KHẢO 53](#_Toc387692920)

[PHỤ LỤC 54](#_Toc387692920)

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC KÝ HIỆU**

**CÁC CHỮ VIẾT TẮT**

BĐS Bất động sản

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

Hình 1.1: Đề tài dự đoán giá nhà đất 6

Hình 1.2: Thị trường bất động sản……………………………………………………...8

Hình 1.3: Biến động của thị trường bất động sản 14

Hình 1.4: Biến động của thị trường bất động sản tại một số tỉnh thành 15

Hình 1.5: Biến động của thị trường tại HCM 16

Hình 1.6: Biến động của thị trường tại Hà Nội 17

Hình 1.7: Dự đoán trong tương lai 18

Hình 1.8: Thư viện BeautifulSoup 20

Hình 1.9: Source code web cần lấy dữ liệu. 21

Hình 1.10: Thư viện Pandas. 26

Hình 1.11: Thư viện Scikit-learn 27

Hình 1.12:Machine Learning 29

Hình 1.13: Các mô hình hồi quy 30

Hình 1.14: Trang web được chọn để thu thập dữ liệu 32

Hình 1.15: Xác định các phần tử html chứa dữ liệu 33

Hình 1.16: Code crawl dữ liệu 34

Hình 1.17: Ghi dữ liệu vào file 35

Hình 1.18: Chạy chương trình 36

Hình 1.19: Dữ liệu đã thu được 36

Hình 1.20: Trường dữ liệu có thể đem vào mô hình dự đoán 37

Hình 1.21: Dữ liệu đã thu được 37

Hình 1.22: Code giữ lại phần số 38

Hình 1.23: Kết quả đã giữ lại phần số 38

Hình 1.24: Phần code chuyển đổi mức giá cùng đơn vị 39

Hình 1.25: Phần code chuyển đổi kiểu dữ liệu 39

Hình 1.26: Đồ thị thể hiện tuần suất xuất hiện của từng mệnh giá. 40

Hình 1.27: Kết quả tính trung vị, độ lệch chuẩn và phương sai 41

Hình 1.28: Đồ thị phân bố dữ liệu giữa giá nhà và số phòng ngủ 42

Hình 1.29: Đồ thị phân bố dữ liệu giữa giá nhà và số lầu 43

Hình 1.30: Đồ thị phân bố dữ liệu giữa giá nhà và chiều dài đường trước nhà 44

Hình 1.31: Đồ thị phân bố dữ liệu giữa giá nhà và diện tích. 45

Hình 1.32: Dữ liệu nhiễu 46

Hình 1.33: Code xử lý dữ liệu giá nhà đất nhiễu 47

Hình 1.34: Triển khai mô hình dự đoán Linear Regression và Lasso Regression 48

Hình 1.35: Kết quả mô hình dự đoán Linear Regression và Lasso Regression 49

Hình 1.36: Code triển khai mô hình gradient boosting regression 49

Hình 1.37: Code áp dụng kỹ thuật GridSearch vàomô hình Gradient boosting regression 50

Hình 1.38: Các tham số tốt nhất để tối ưu hiệu quả dự đoán của mô hình Gradient boosting regression 50

Hình 1.39: Kết quả dự đoán giá nhà với mô hình Gradient boosting regression 51

**DANH MỤC BẢNG**

CHƯƠNG 1 – TỔNG QUAN VỀ ĐỀ TÀI

1.1 Giới thiệu đề tài

Thị trường giá nhà đất ở Việt Nam đặc biệt là ở Thành phố Hồ Chí Minh luôn là vấn đề được quan tâm rất lớn. Chính vì đều này nên đã có rất nhiều chuyên gia phân tích hay dự đoán giá nhà đất nghiên cứu các tác động từ những dữ liệu đã thu thập được nhằm mục đích phục vụ cho các vấn đề liên quan đến kinh tế.

Đề tài dự đoán sự thay đổi và giá nhà đất trong thị trường bất động sản này sẽ mở ra một hướng đi hay hướng nghiên cứu mới cho tương lai. Múc đích của đề tài là phân tích và ảo hóa dữ liệu thông qua biểu đồ để có một cái nhìn tổng quan về giá nhà đất trong thị trường bất động sản.



Hình 1.1: Đề tài dự đoán giá nhà đất

Nguồn: [<https://biggee.vn/dinh-gia-nha-dat>]

* 1. Ý tưởng

Không như thị trường bất động sản ở các nước tiên tiến khác, ở Việt Nam bất động sản vẫn chưa thể có một sản giao dịch chính thức nào. Chính vì thế nên dữ liệu để thực hiện đề tài dự đoán sự thay đổi và giá nhà đất trong thị trường bất động sản phải được “cào” từ một website rao bán bất động sản. Các bước để thực hiện đề tài sẽ bao gồm:

* Thu thập dữ liệu giá nhà ở thành phố HCM
* Tiến hành làm sạch và chuẩn bị dữ liệu
* Trực quan hóa dữ liệu
* Áp dụng mô hình dự đoán giá

CHƯƠNG 2 – TỔNG QUAN VỀ THỊ TRƯỜNG BẤT ĐỘNG SẢN

2.1 Giới thiệu về thị trường bất động sản

Thị trường bất động sản (BĐS) là một trong những thị trường quan trọng của nền kinh tế thị trường vì thị trường này liên quan trực tiếp tới một lượng tài sản cực lớn cả về quy mô, tính chất cũng như giá trị của các mặt trong nền kinh tế quốc dân. Tỷ trọng BĐS trong tổng số của cải xã hội ở các nước có khác nhau nhưng thường chiếm trên dưới 40% lượng của cải vật chất của mỗi nước. Các hoạt động liên quan đến BĐS chiếm tới 30% tổng hoạt động của nền kinh tế. Theo thống kê, ở các nước phát triển lượng tiền ngân hàng cho vay qua thế chấp bằng BĐS chiếm trên 80% trong tổng lượng vốn cho vay. Vì vậy, phát triển đầu tư, kinh doanh BĐS đóng vai trò quan trọng trong việc chuyển các tài sản thành nguồn tài chính dồi dào phục vụ cho yêu cầu phát triển kinh tế - xã hội đặc biệt là đầu tư phát triển cơ sở hạ tầng của nền kinh tế.



Hình 1.2: Thị trường bất động sản

Nguồn: [[https://vietnamfinance.vn](https://vietnamfinance.vn/bao-cao-nghien-cuu-thi-truong-bat-dong-san-viet-nam-quy-iv2019-cua-colliers-international-20180504224234515.htm)]

**Lịch sử hình thành**

Lịch sử hình thành ngành BĐS tại Việt Nam: chia làm 6 giai đoạn chính:

**Trước năm 1990**

Thời kỳ này kinh tế chưa phát triển, quỹ đất còn nhiều, quá trình đô thị hóa diễn ra còn chậm, do đó nhu cầu về sử dụng đất của các tổ chức, cá nhân trong nước chưa cao. Trong giai đoạn này hầu như không tồn tại thị trường bất động sản. Các giao dịch bất động sản trong thời gian này chỉ là các giao dịch ngầm, phi thị trường.

**Cơn sốt lần thứ nhất (1993-1994)**

Thị trường Bất động sản Việt Nam đã có “Cơn sốt lần thứ nhất” - diễn ra trong hai năm (1993-1994), trong đó Thị trường Bất động sản đặc biệt “sốt” mạnh ở phân khúc Đất và Quyền sử dụng đất. Nguyên nhân là do nhu cầu nhà ở, đất ở của dân cư và đất để sản xuất kinh doanh ngày càng tăng, nhất là ở các đô thị, mà quan hệ chuyển nhượng, mua bán đất đai trong dân cư đã xuất hiện từ những năm trước khi Luật đất đai (1993) ra đời.

**Đóng băng lần thứ nhất (1995 – 1999)**

Trước tình hình thị trường sốt nóng, nhà nước đã can thiệp vào thị trường bằng 2 nghị định là Nghị định số 18 và 87 về thuê đất, nhằm ngăn chặn hiện tượng đầu cơ. Theo đó, người sử dụng đất phải trả tiền hai lần ( tiền chuyển quyền sử dụng đất và tiền thuê đất ). Với sự tác động từ 2 nghị định trên đã buộc các nhà đầu cơ sử dụng đòn bẩy tài chính để tham gia thị trường phải bán tháo đất đai đang nắm giữ để trả tiền cho Ngân hàng. Làn sóng xả hàng của các nhà đầu cơ đã làm cho thị trường ở trạng thái cung vượt cầu, thị trường lao dốc. Bên cạnh đó, khủng hoảng kinh tế châu Á khởi nguồn từ Thái Lan đã làm cho một số dự án bất động sản của nước ngoài đầu tư vào Việt Nam đã thất bại, góp phần làm cho thị trường suy thoái.

**Cơn sốt lần thứ 2 ( 2001 – 2002)**

Sau một thời gian dài bình lặng, từ năm 2000, giá nhà đất bắt đầu biến động, tiếp đó giá cả tăng nhanh liên tục và đạt đỉnh cao vào khoảng Quý II năm 2001. Cơn sốt đất lần thứ hai diễn ra do người đầu tư dự đoán và đánh giá chủ trương cho Việt kiều mua nhà và ban hành giá đất mới sẽ có triển vọng cho thị trường nhà đất nên nhiều người đầu tư mua đất khắp nơi ở vùng ven.

Đóng băng lần thứ hai (2002 – 2006) Từ cuối năm 2002 đến 2006 Thị trường Bất động sản diễn ra trầm lắng, thậm chí “đóng băng” như nhiều người nhận xét. Thị trường Bất động sản năm sau lạnh hơn năm trước. Theo số liệu thống kê: Năm 2003 giao dịch địa ốc thành công giảm 28%, năm 2004 giảm 56% và năm 2005 giảm 78%! Thị trường tiếp tục lên cơn sốt vào năm 2001 – 2002, trước thực trạng đó, nhà nước một lần nữa ra tay điều tiết thị trường bằng biện pháp hành chính với “ luật đất đai 2003” và “nghị định 181” hướng dẫn thi hành luật đất đai. Nghị định 181 đã chấm dứt tình trạng “phân lô bán nền”. đã làm thị trường nhà đất hạ nhiệt vì "đánh" trúng vào lợi ích tạo ra giá đầu cơ. Nhưng vì phần lớn vốn đầu tư này là vốn nhàn rỗi của tư nhân nên nhu cầu phải bán nhanh thu hồi vốn thanh toán nợ không xảy ra như lần thứ nhất. Do đó dù cơn sốt đã dừng lại nhưng giá đất vẫn BÁO CÁO NGÀNH 2 không hạ xuống bằng giá trước khi xảy ra cơn sốt.Tác động bằng chính sách của Nhà nước từ năm 2002 đã gây ra lần đóng băng thứ hai của Thị trường Bất động sản Việt Nam 2002-2006 với việc dòng vốn chuyển sang Thị trường Chứng khoán.

**Cơn sốt đất lần thứ ba (2007 -2008)**

Không giống như hai lần sốt đất trước đó, lần sốt đất này tập trung mạnh mẽ vào phân khúc Căn hộ cao cấp và Biệt thự. Một số nguyên nhân chủ chốt :

• Nguồn vốn FDI tăng trưởng mạnh đầu tư vào Việt Nam tạo nên tăng trưởng kinh tế ấn tượng trong 2003-2007.

• Trong năm 2006-2007, có thể xem là năm khá huy hoàng đối với Thị trường Chứng khoán Việt Nam và mọi người đều tham gia cũng như kiếm tiền dễ dàng từ Thị trường Chứng khoán. Điều này tạo ra một nguồn vốn thặng dư lớn của người thắng chứng khoán chuyển dịch sang Thị trường Bất động sản tạo điều kiện cho phân khúc cao cấp “sốt” mạnh.

**Đóng băng lần 3 ( 2008 – nay )**.

Đứng trước thực trạng bong bóng thị trường bất động sản ngày càng lớn cùng với tốc độ lạm phát tăng một cách chóng mặt, chính phủ đã tiến hành điều tiết thị trường bằng các chính sách tiền tệ với việc kiểm soát tín dụng chặt chẽ, đặc biệt là tín dụng phi sản xuất với hàng loạt các biện pháp được chính phủ thực hiện nhằm ổn định thị trường và kiềm chế lạm phát. Ngay sau khi có các chính sách thắt chặt tài chính thực hiện thị trường Bất động sản miền Nam lập tức bị ảnh hưởng nặng nề và sự suy giảm cả về giá lẫn giao dịch là điều không tránh khỏi. Với thị trường miền Bắc, khác biệt cơ bản cũng lại là dòng tài chính. Có thể nói, lượng tài chính dư dả tại miền Bắc đủ để Thị trường Bất động sản chỉ bị ảnh hưởng tức thời, giao dịch chững lại nhưng giá cả chỉ sụt giảm khoảng 10% vào năm 2008. Nghị định 71 và 69 ra đời năm 2010 hướng dẫn thi hành luật đất đai có sửa đổi 2009 và việc thu thuế sử dụng đất càng làm cho thị trường đã trầm lắng nay lại thêm phần ảm đạm. Bước sang 2011, trước nguy cơ lạm phát tăng cao cùng sự leo dốc không mãnh liệt của giá vàng và USD đã buộc chính phủ phải có nghị quyết số 11 về các biện pháp kiềm chế lạm pháp, ổn định tình hình kinh tế vĩ mô. Theo đó tăng trưởng tín dụng năm 2011 bị giới hạn ở con số 20%, dư nợ tín dụng cho vay đối với lĩnh vực phi sản xuất bị giới hạn ở mức 16% càng làm cho thị trường bất động sản thêm phần ảm đạm, xua tan kỳ vọng thị trường phục hồi của các nhà đầu tư bất động sản. Năm 2012 – là năm chứng kiến nhiều doanh nghiệp phá sản nhất trong vòng 10 năm qua. Theo thống kê của VCCI ( Phòng thương mại và công nghiệp Việt Nam) cho biết, trung bình những năm qua, mỗi năm có khoảng 5,000-7,000 doanh nghiệp phá sản/giải thể, tức thấp hơn năm 2011 khoảng 8 lần. 2012 cũng là năm cạnh tranh khốc liệt, thậm chí là cuộc chiến sống còn của doanh nghiệp BĐS. Hàng loạt công ty dịch vụ môi giới và cho thuê BĐS đóng cửa. Các Chủ Đầu Tư thiếu năng lực về tài chính gần như tuyên bố phá sản hoặc phải rao bán dự án.

2.2 Những yếu tố ảnh hưởng đến thị trường bất động sản

Sự phát triển kinh tế: Phát triển kinh tế của một nước tất yếu làm gia tăng các nhu cầu sử dụng đất trong các lĩnh vực sản xuất, nhu cầu về văn phòng cho thuê, trung tâm thương mại… Hơn nữa, khi kinh tế phát triển, thu nhập của người dân sẽ tăng cao và do đó làm gia tăng nhu cầu về nhà ở đối với người dân. Việt Nam là một nước đang phát triển, luôn duy trì được tốc độ tăng trưởng GDP cao so với các nước khác trên Thế giới. Do ảnh hưởng của cuộc khủng hoảng kinh tế toàn cầu từ cuối năm 2007, tốc độ tăng trưởng GDP đã có sự sụt giảm so với những năm gần đây có phần sụt giảm.

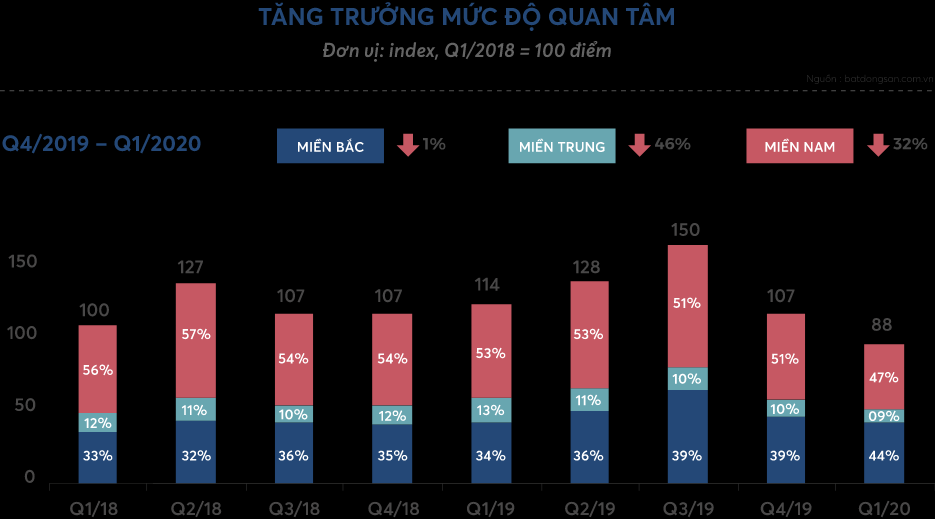
Dân số: Tăng dân số cũng đồng nghĩa với tăng các nhu cầu về các loại hàng hoá trên thị trường, để đáp ứng nhu cầu các ngành sản xuất phải mở rộng, phát triển về quy mô từ đó làm tăng nhu cầu về sử dụng đất. Theo báo cáo từ cơ quan thống kê, năm 2012 dân số Việt Nam ước đạt 88.78 triệu người, tăng 1.06% so với năm trước. Tốc độ tăng không cao so với vài năm trước và vẫn nằm trong tốp tăng thấp dưới 1.1% đạt được từ năm 2007 đến nay. Tuy nhiên, tỷ lệ dân số thành thị tại Việt Nam chỉ đạt 32.45%, còn thấp hơn nhiều so với các nước trong khu vực (châu Á đạt trên 44% và trên thế giới là trên 51%).. Tỷ lệ đô thị hóa dự kiến sẽ gia tăng trong các năm tiếp theo, nhu cầu nhà ở tại thành thị do vậy sẽ không ngừng gia tăng trong các năm tới.

**Dòng vốn FDI**: Dòng vốn FDI đổ vào Việt Nam liên tục gia tăng, trong đó vốn đổ vào cho thị trường bất động sản chiếm tỷ trọng khá cao. Điều đó chứng tỏ thị trường Bất động sản Việt Nam rất hấp dẫn với các nhà đầu tư nước ngoài. Đây là một trong những nguồn vốn góp phần hỗ trợ chính thúc đây sự tăng trưởng của ngành BĐS. Năm 2012, thu hút vốn FDI đạt 13 tỷ USD, trong đó riêng đầu tư vào ngành BĐS chiếm 14.2% tổng vốn đầu tư đạt 1.8 tỷ USD, đứng thứ 2 sau vốn FDI đầu tư vào công nghiệp chế biến, chế tạo.

**Pháp luật:** Pháp luật có thể được xem là yếu tố tác động quan trọng đến sự hình thành và phát triển của thị trường bất động sản biểu hiện qua những quy định pháp luật về quyền mua, bán, cho thuê, thế chấp, góp vốn liên doanh…. bằng bất động sản.

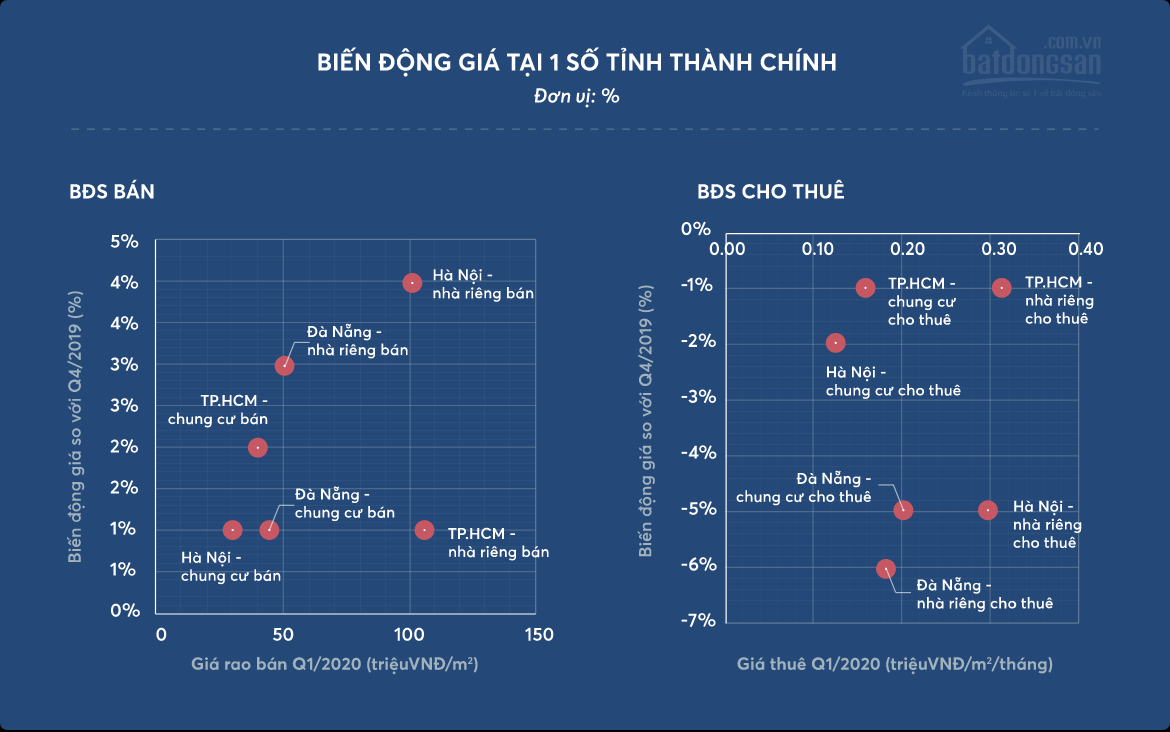
2.3 Báo cáo sự biến động của thị trường năm 2020

Trong Q1/2020, tổng lượng tin đăng toàn trang Batdongsan.com.vn giảm 28% trong khi mức độ quan tâm giảm 23% so với Q4/2019. Điều này phản ánh nguồn cung cũng như nguồn cầu của toàn thị trường suy giảm đáng kể, đặc biệt từ thời điểm đầu tháng 3/2020, do dịch bệnh Covid-19 bùng phát trở lại. Mức độ quan tâm của người tìm kiếm BĐS không suy giảm đồng đều tại tất cả các tỉnh thành trên cả nước. Mức độ quan tâm tại miền Bắc gần như không đổi, chỉ giảm nhẹ 1%, trong khi miền Nam giảm 32% và miền Trung giảm 46%.



Hình 1.3: Biến động của thị trường bất động sản

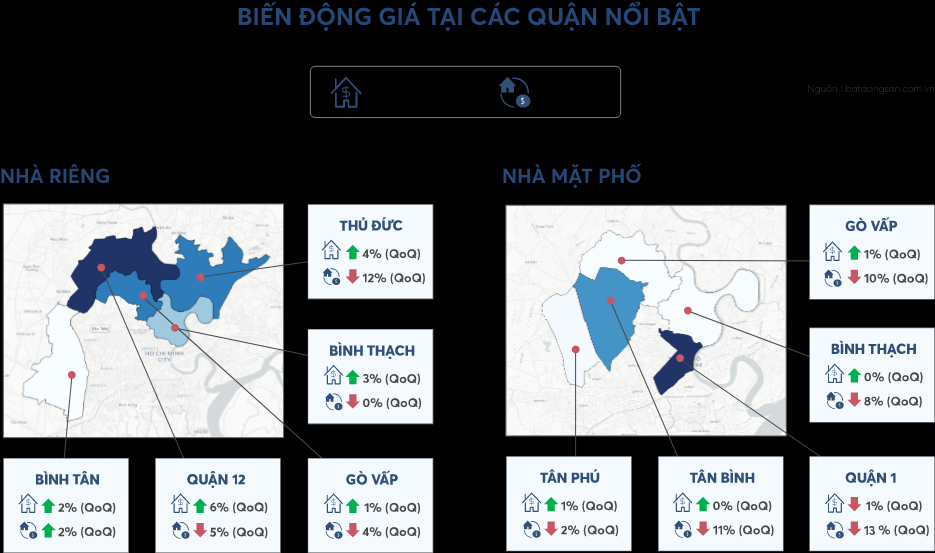
Nguồn: [<https://batdongsan.com.vn/Landingpage/bao-cao-thi-truong-bat-dong-san-quy-1-2020/index.html>]

Giá bất động sản phân hóa rõ rệt ở 2 thị trường BĐS bán và cho thuê. Trong khi giá BĐS bán tương đối ổn định thì giá bất động sản cho thuê suy giảm đáng kể so với Q1/2019. Giá rao bán nhà riêng tại Hà Nội tăng nhiều nhất, tăng gần 4% so với Q4/2019. Mặc dù lượng giao dịch giảm trên thị trường, giá rao bán trung bình của hầu hết các thành phố đều chứng kiến mức độ tăng nhẹ trong khoảng 1-3%. Mức tăng giá này là rất nhỏ so với năm 2019 khi giá của nhiều loại hình BĐS tại TP.HCM tăng trưởng 2 con số. Theo xu hướng ngược lại, giá thuê BĐS giảm ở tất cả các loại hình BĐS tại tất cả các thành phố lớn, do nhu cầu cho thuê tăng cao nhưng nhu cầu thuê giảm.

Hình 1.4: Biến động của thị trường bất động sản tại một số tỉnh thành.

Nguồn: [[https://batdongsan.com.vn](https://batdongsan.com.vn/Landingpage/bao-cao-thi-truong-bat-dong-san-quy-1-2020/index.html)]

**Thị trường HCM**

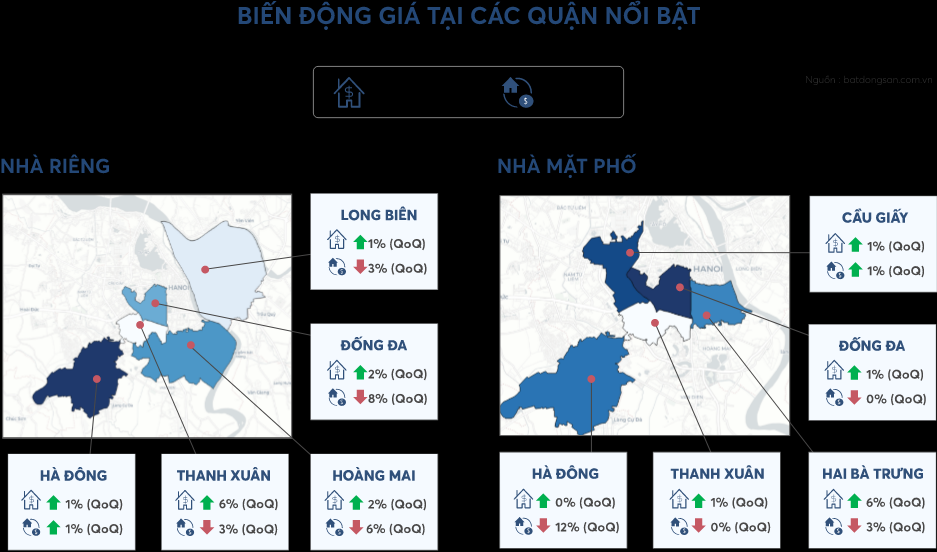
Tại TP.HCM, bất động sản nhà riêng ghi nhận xu hướng giảm với mức độ quan tâm giảm 26%, lượng tin đăng giảm 32% so với Q4/2019. Giá bán có xu hướng tăng nhẹ trong khi giá thuê giảm ở hầu hết các quận. Cụ thể, giá thuê các căn nhà riêng khu vực ngoài trung tâm sụt giảm mạnh như quận Thủ Đức (-12%), Quận 12 (-5%) và quận Gò Vấp (-4%). Đối với thị trường nhà mặt phố, xu hướng giảm còn được thể hiện rõ ràng hơn với mức độ quan tâm giảm 33%, lượng tin đăng giảm 37%, giá bán không tăng và giá thuê tại nhiều quận sụt giảm trung bình 11% so với Q4/2019. Đứng đầu trong các quận có giá thuê giảm mạnh nhất là Quận 1 (-13%), sau đó đến quận Tân Bình (-11%) và quận Gò Vấp (-10%).

Hình 1.5: Biến động của thị trường tại HCM.

Nguồn: [[https://batdongsan.com.vn](https://batdongsan.com.vn/Landingpage/bao-cao-thi-truong-bat-dong-san-quy-1-2020/index.html)]

**Thị trường Hà Nội**

Thị trường nhà đất tại Hà Nội ghi nhận mức giảm thấp hơn so với thị trường TP.HCM. Loại hình BĐS nhà riêng có mức độ quan tâm suy giảm nhẹ với 6%, lượng tin đăng giảm 31%, giá thuê giảm trung bình 5% so với Q4/2019. Các căn nhà riêng có mức giá thuê sụt giảm đáng kể tập trung tại các quận Đống Đa, Hoàng Mai, Thanh Xuân và Long Biên. Cùng chung xu hướng với loại hình nhà riêng, nhà mặt phố tại Hà Nội có mức độ quan tâm giảm 10%, lượng tin đăng giảm 21% và giá thuê trung bình giảm trung bình 6% so với Q4/2019. Trong số các quận nổi bật, quận có mức giá thuê sụt giảm mạnh nhất là quận Hà Đông, giảm 12%.



Hình 1.6: Biến động của thị trường tại Hà Nội.

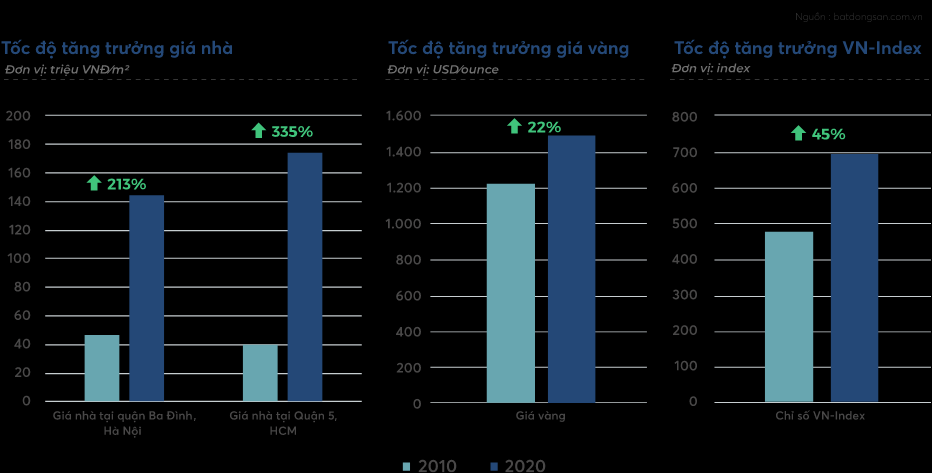
Nguồn: [[https://batdongsan.com.vn](https://batdongsan.com.vn/Landingpage/bao-cao-thi-truong-bat-dong-san-quy-1-2020/index.html)]

**Tầm nhìn thị trường 2020**

Không chỉ bất động sản mà hầu hết tất cả các ngành nghề đều đang chịu thiệt hại tùy mức độ, dưới tác động của đại dịch Covid-19. Trong Q1/2020, người tìm kiếm BĐS, nhà đầu tư đều có tâm lý e dè, nghe ngóng và tích trữ tiền mặt, vàng để đảm bảo cho những nhu cầu cơ cơ bản. Tuy nhiên, theo một khảo sát của Vnexpress với khoảng gần 130.000 người được hỏi, có đến 29% vẫn có nhu cầu đầu tư, mua nhà đất, điều này cho thấy nhu cầu đầu tư và sở hữu nhà ở luôn rất lớn tại thị trường Việt Nam.

Thời gian vừa qua, các chủ đầu tư, sàn giao dịch BĐS lớn tại Việt Nam cũng phải tạm hoãn hoặc hủy bỏ việc mở bán, dẫn đến nguồn cung sơ cấp hạn hữu nhưng họ cũng chuẩn bị kế hoạch trong Q3/2020 và Q4/2020. Những nguồn hàng này sau khi được giao dịch sơ cấp thành công có thể sẽ được giao dịch luôn ở cả thị trường thứ cấp.

Cũng theo dữ liệu nghiên cứu của chúng tôi đối với 3 loại tài sản phổ biến tại thị trường Việt Nam hiện nay, vàng và chứng khoán hiện đang có tốc độ tăng trưởng thấp hơn so với BĐS tại một số quận trung tâm, TP lớn tại Việt Nam. Mức độ tăng trưởng về giá có thể suy giảm, đi ngang trong ngắn hạn dưới tác động của dịch bệnh nhưng khó tác động đến thị trường BĐS trong dài hạn và theo nhận định của một số chuyên gia, thị trường BĐS có thể phục hồi và tăng trưởng giao dịch trở lại từ cuối quý 3, đầu Q4/2020.



Hình 1.7: Dự đoán trong tương lai.

Nguồn: [[https://batdongsan.com.vn](https://batdongsan.com.vn/Landingpage/bao-cao-thi-truong-bat-dong-san-quy-1-2020/index.html)]

CHƯƠNG 3 – CƠ SỞ LÝ THUYẾT

**3.1 Kỹ thuật Crawl data**

***3.1.1 Web Crawling là gì?***

***Web crawling*** là quá trình tự động trích xuất các thông tin từ các trang web và lưu trữ nó dưới một định dạng phù hợp. Chương trình mà thực hiện công việc này gọi là ***web crawler***.

Thông thường, khi muốn lấy một số thông tin từ các trang web, chúng ta sẽ dùng các API mà các trang đó cung cấp. Đây là cách đơn giản, tuy nhiên không phải trang web nào cũng cung cấp sẵn API cho chúng ta sử dụng. Do đó chúng ta cần một kĩ thuật để lấy các thông tin từ các trang web đó mà không thông qua API.

***3.1.2 Thư viện BeautifulSoup***

Thư viện [**BeautifulSoup**](https://www.crummy.com/software/BeautifulSoup/bs4/doc/) là một thư viện của Python cho phép chúng ta lấy dữ liệu từ HTML đơn giản và hiệu quả.



Hình 1.8: Thư viện BeautifulSoup.

Nguồn: [[https://www.digitalocean.com](https://www.digitalocean.com/community/tutorials/how-to-scrape-web-pages-with-beautiful-soup-and-python-3)]

***3.1.3 Demo***

Trước hết chúng ta cần lấy nội dung của trang web và parse nó như sau:

from bs4 import BeautifulSoup

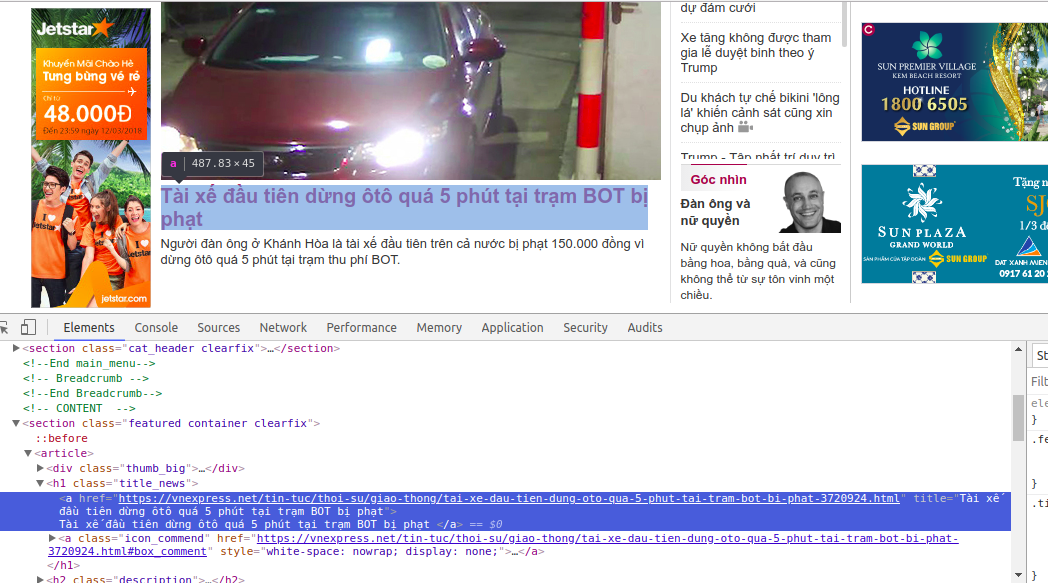
import urllib.request

url = 'https://vnexpress.net'

page = urllib.request.urlopen(url)

soup = BeautifulSoup(page, 'html.parser')

Tiếp theo chúng ta sẽ inspect và lấy thử một tin tức trước. Nhấn chuột phải vào tin tức đầu tiên của trang web và chọn inspect, bạn sẽ thấy thẻ <a> chứa tiêu đề của tin tức, được lồng bên trong thẻ <section class="featured container clearfix">.



Hình 1.9: Source code web cần lấy dữ liệu.

Nguồn: [[https://viblo.asia](https://viblo.asia/p/web-crawling-voi-thu-vien-beautifulsoup-1VgZvNGOZAw)]

Chúng ta có thể lấy ra tiêu đề và liên kết bằng cách dùng phương thức [find()](https://www.crummy.com/software/BeautifulSoup/bs4/doc/#find) như sau:

from bs4 import BeautifulSoup

import urllib.request

url = 'https://vnexpress.net'

page = urllib.request.urlopen(url)

soup = BeautifulSoup(page, 'html.parser')

new\_feed = soup.find('section', class\_='featured container clearfix').find('a')

title = new\_feed.get('title')

link = new\_feed.get('href')

print('Title: {} - Link: {}'.format(title, link))

Chạy đoạn script trên, bạn sẽ được kết quả:

Title: Tài xế đầu tiên dừng ôtô quá 5 phút tại trạm BOT bị phạt - Link: https://vnexpress.net/tin-tuc/thoi-su/giao-thong/tai-xe-dau-tien-dung-oto-qua-5-phut-tai-tram-bot-bi-phat-3720924.html

Thay vì chỉ lấy một tin như trên, bạn có thể lấy tất cả các tin tức bên cạnh tin tức vừa lấy như sau bằng cách dùng [find\_all()](https://www.crummy.com/software/BeautifulSoup/bs4/doc/#find-all):

from bs4 import BeautifulSoup

import urllib.request

url = 'https://vnexpress.net'

page = urllib.request.urlopen(url)

soup = BeautifulSoup(page, 'html.parser')

new\_feeds = soup.find(

'section', class\_='featured container clearfix').find\_all('a')

for feed in new\_feeds:

title = feed.get('title')

link = feed.get('href')

print('Title: {} - Link: {}'.format(title, link))

Chạy đoạn script trên, bạn sẽ được:

Title: Tài xế đầu tiên dừng ôtô quá 5 phút tại trạm BOT bị phạt - Link: https://vnexpress.net/tin-tuc/thoi-su/giao-thong/tai-xe-dau-tien-dung-oto-qua-5-phut-tai-tram-bot-bi-phat-3720924.html

Title: Chi 50 triệu sửa phòng con nhỏ, tôi phải thay tất cả đồ sau 3 năm - Link: https://giadinh.vnexpress.net/tin-tuc/nha-dep/khong-gian-song/chi-50-trieu-sua-phong-con-nho-toi-phai-thay-tat-ca-do-sau-3-nam-3720639.html

Title: 8 điểm phụ nữ hay soi đàn ông nhất - Link: https://giadinh.vnexpress.net/tin-tuc/to-am/8-diem-phu-nu-hay-soi-dan-ong-nhat-3719129.html

Title: Người Việt ngày càng uống nhiều bia - Link: https://kinhdoanh.vnexpress.net/tin-tuc/hang-hoa/nguoi-viet-ngay-cang-uong-nhieu-bia-3721014.html

Title: Người gốc Việt giả danh thanh tra tống tiền hơn chục tiệm nail ở Mỹ - Link: https://vnexpress.net/tin-tuc/the-gioi/nguoi-viet-5-chau/nguoi-goc-viet-gia-danh-thanh-tra-tong-tien-hon-chuc-tiem-nail-o-my-3720994.html

Title: Hiện tượng thác nước hiếm gặp ở ven biển Philippines - Link: https://vnexpress.net/tin-tuc/khoa-hoc/chuyen-la/hien-tuong-thac-nuoc-hiem-gap-o-ven-bien-philippines-3720972.html

Title: Bé trai 5 tuổi ở Trung Quốc trốn khỏi trường mẫu giáo - Link: https://vnexpress.net/tin-tuc/giao-duc/be-trai-5-tuoi-o-trung-quoc-tron-khoi-truong-mau-giao-3721003.html

Title: Trump ân xá thủy thủ bị án tù vì chụp ảnh trong tàu ngầm hạt nhân - Link: https://vnexpress.net/tin-tuc/the-gioi/quan-su/trump-an-xa-thuy-thu-bi-an-tu-vi-chup-anh-trong-tau-ngam-hat-nhan-3720977.html

Title: So sánh CR-V bản E, Santa Fe bản tiêu chuẩn và X-Trail 2.0SL? - Link: https://vnexpress.net/tin-tuc/oto-xe-may/tu-van/so-sanh-cr-v-ban-e-santa-fe-ban-tieu-chuan-va-x-trail-2-0sl-3720560.html

Title: Đạo diễn ‘The Shape of Water’ ly dị vợ sau 32 năm sống chung - Link: https://giaitri.vnexpress.net/tin-tuc/gioi-sao/quoc-te/dao-dien-the-shape-of-water-ly-di-vo-sau-32-nam-song-chung-3721001.html

Title: Góc nhìn - Link: /tin-tuc/goc-nhin/

Title: Đàn ông và nữ quyền - Link: https://vnexpress.net/tin-tuc/goc-nhin/dan-ong-va-nu-quyen-3720424.html

Trong số các kết quả trên, bạn sẽ thấy có một kết quả không như mong muốn là Title: Góc nhìn - Link: /tin-tuc/goc-nhin/. Đây không phải là một tin tức mà là tiêu đề của một đầu mục, và đường link của nó cũng không phải là dạng đầy đủ. Chúng ta có thể loại bỏ nó bằng cách dùng một hàm để [filter](https://www.crummy.com/software/BeautifulSoup/bs4/doc/#a-function) nó như sau:

from bs4 import BeautifulSoup

import urllib.request

import re

def not\_relative\_uri(href):

return re.compile('^https://').search(href) is not None

url = 'https://vnexpress.net'

page = urllib.request.urlopen(url)

soup = BeautifulSoup(page, 'html.parser')

new\_feeds = soup.find(

'section', class\_='featured container clearfix').find\_all(

'a', class\_='', href=not\_relative\_uri)

for feed in new\_feeds:

title = feed.get('title')

link = feed.get('href')

print('Title: {} - Link: {}'.format(title, link))

Chạy đoạn script trên, bạn sẽ được kết quả như mong muốn.

Bây giờ mình sẽ lưu nó dưới dạng CSV file. Python đã có sẵn module csv và chúng ta cứ dùng thôi:

from bs4 import BeautifulSoup

import urllib.request

import re

import csv

def not\_relative\_uri(href):

return re.compile('^https://').search(href) is not None

url = 'https://vnexpress.net'

page = urllib.request.urlopen(url)

soup = BeautifulSoup(page, 'html.parser')

new\_feeds = soup.find(

'section', class\_='featured container clearfix').find\_all(

'a', class\_='', href=not\_relative\_uri)

with open('data.csv', 'w') as csv\_file:

writer = csv.writer(csv\_file)

writer.writerow(['Title', 'Link'])

for feed in new\_feeds:

writer.writerow([feed.get('title'), feed.get('href')])

Như vậy là chúng ta đã thực hiện việc crawling đơn giản với vài dòng code. Thực tế việc crawling có thể phức tạp hơn nhiều. Một số điều lưu ý khi crawling:

* Không spam các trang web. Điều này có thể làm các trang web mà bạn đang crawling bị crash.
* Việc crawling tùy thuộc vào layout của từng trang web và layout đó có thể thay đổi theo thời gian. Do đó khi layout thay đổi, chúng ta cũng có thể sẽ phải thay đổi code.

**3.2 Các thư viện cần thiết**

**Pandas**

Thư viện pandas trong python là một thư viện mã nguồn mở, hỗ trợ đắc lực trong thao tác dữ liệu. Đây cũng là bộ công cụ phân tích và xử lý dữ liệu mạnh mẽ của ngôn ngữ lập trình python. Thư viện này được sử dụng rộng rãi trong cả nghiên cứu lẫn phát triển các ứng dụng về khoa học dữ liệu. Thư viện này sử dụng một cấu trúc dữ liệu riêng là Dataframe. Pandas cung cấp rất nhiều chức năng xử lý và làm việc trên cấu trúc dữ liệu này. Chính sự linh hoạt và hiệu quả đã khiến cho pandas được sử dụng rộng rãi.



Hình 1.10: Thư viện Pandas.

Nguồn: [[https://topdev.vn](https://topdev.vn/blog/pandas-python-tutorial/)]

**Re**

Regular Expression (RegEx) hay còn gọi là Biểu thức chính quy là một đoạn các ký tự đặc biệt theo những khuôn mẫu (pattern) nhất định, đại diện cho chuỗi hoặc một tập các chuỗi

**Matplotlib**

Một thư viện dung để biễu diễn data bằng đồ thị rất được nhiều người sử dụng

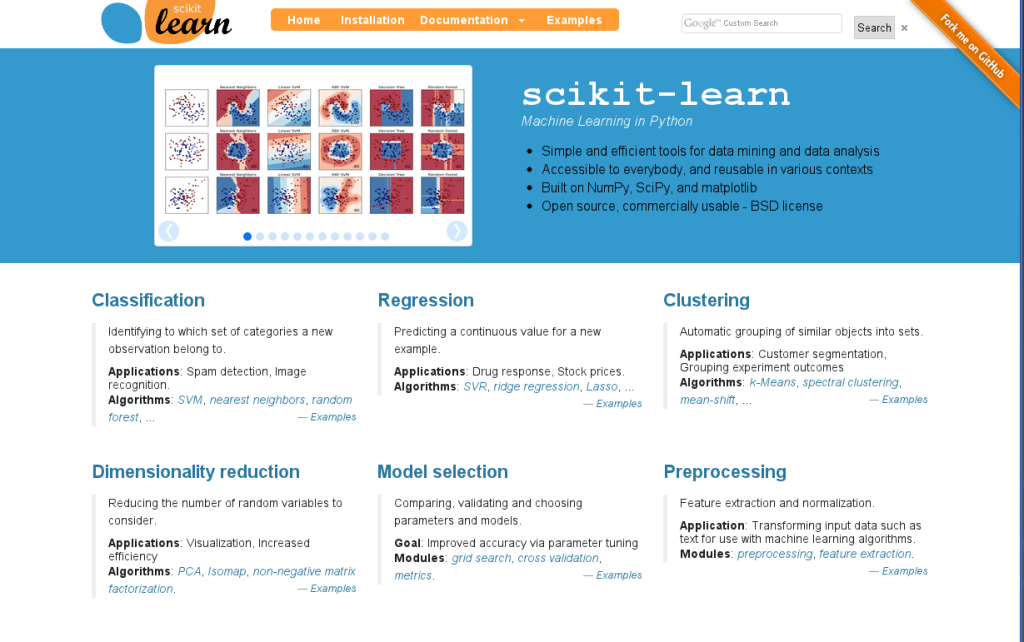
**Scikit-learn**

Thư viện này tích hợp rất nhiều thuật toán hiện đại và cố điển giúp bạn vừa học vừa tiến hành đưa ra các giải pháp hữu ích cho bài toán của bạn một cách đơn giản.

Scikit-learn ban đầu được đề xuất bởi David Cournapeau trong một dự án mùa hè của Google vào năm 2007.

Later Matthieu Brucher tham gia dự án trên và bắt đầu sử dụng nó làm một phần luận văn tiến sĩ của ông ấy. Vào năm 2010, INRIA bắt đầu tài trợ và phiên bản đầu tiên được xuất bản (v0.1 beta) vào cuối tháng 1 năm 2010.

Dự án vẫn đang được nghiên cứu bởi một đội ngũ hơn 30 nhà nghiên cứu đến từ các công ty lớn INRIA, Google, Tinyclues và Python Software Foundation.



Hình 1.11: Thư viện Scikit-learn.

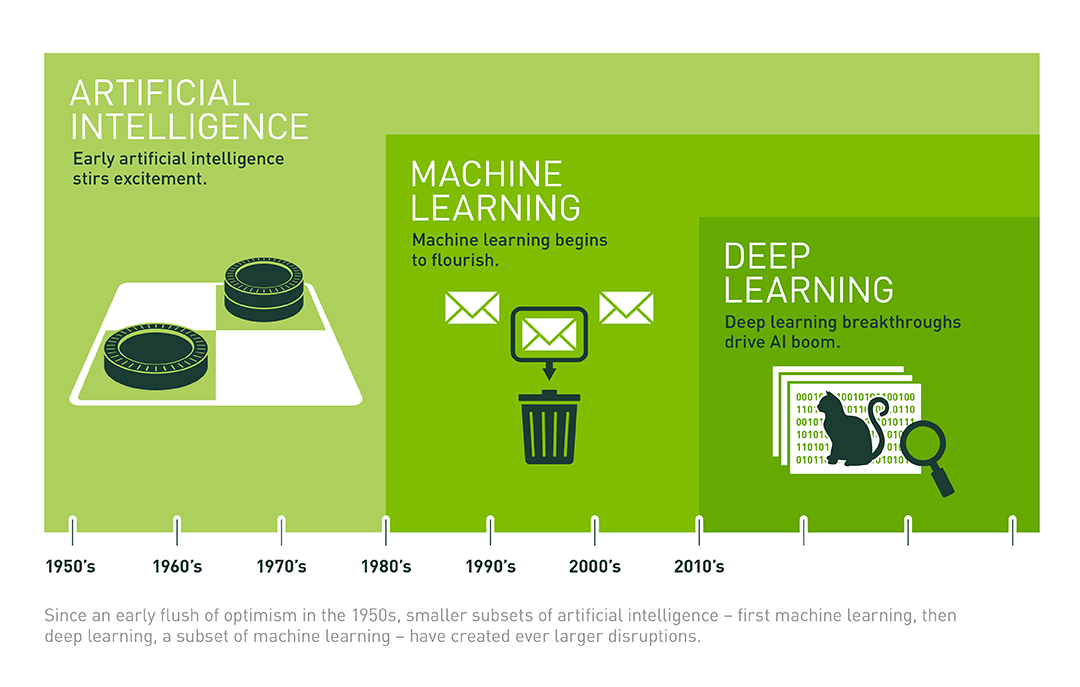
Nguồn: [[https://codelearn.io](https://codelearn.io/sharing/scikit-learn-trong-python-la-gi#:~:text=Scikit%2Dlearn%20l%C3%A0%20g%C3%AC%3F,%2C%20clustering%20%2C%20v%C3%A0%20dimensionality%20reduction%20.)]

**3.3 Giới thiệu Machine Learning**

Những năm gần đây, AI - Artificial Intelligence (Trí Tuệ Nhân Tạo), và cụ thể hơn là Machine Learning (Học Máy hoặc Máy Học) nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư (1 - động cơ hơi nước, 2 - năng lượng điện, 3 - công nghệ thông tin). Trí Tuệ Nhân Tạo đang len lỏi vào mọi lĩnh vực trong đời sống mà có thể chúng ta không nhận ra. Xe tự hành của Google và Tesla, hệ thống tự tag khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, máy chơi cờ vây AlphaGo của Google DeepMind, …, chỉ là một vài trong vô vàn những ứng dụng của AI/Machine Learning. (Xem thêm [Jarvis - trợ lý thông minh cho căn nhà của Mark Zuckerberg](https://www.facebook.com/zuck/posts/10103351073024591))

Machine Learning là một tập con của AI. Theo định nghĩa của Wikipedia, *Machine learning is the subfield of computer science that “gives computers the ability to learn without being explicitly programmed”*. Nói đơn giản, Machine Learning là một lĩnh vực nhỏ của Khoa Học Máy Tính, nó có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể. Bạn Nguyễn Xuân Khánh tại đại học Maryland đang viết một cuốn sách về Machine Learning bằng tiếng Việt khá thú vị, các bạn có thể tham khảo bài [Machine Learning là gì?](https://khanh-personal.gitbook.io/ml-book-vn/machine-learning-la-gi).

Những năm gần đây, khi mà khả năng tính toán của các máy tính được nâng lên một tầm cao mới và lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn, Machine Learning đã tiến thêm một bước dài và một lĩnh vực mới được ra đời gọi là Deep Learning (Học Sâu - *thực sự tôi không muốn dịch từ này ra tiếng Việt*). Deep Learning đã giúp máy tính thực thi những việc tưởng chừng như không thể vào 10 năm trước: phân loại cả ngàn vật thể khác nhau trong các bức ảnh, tự tạo chú thích cho ảnh, bắt chước giọng nói và chữ viết của con người, giao tiếp với con người, hay thậm chí cả sáng tác văn hay âm nhạc (Xem thêm [8 Inspirational Applications of Deep Learning](http://machinelearningmastery.com/inspirational-applications-deep-learning/))



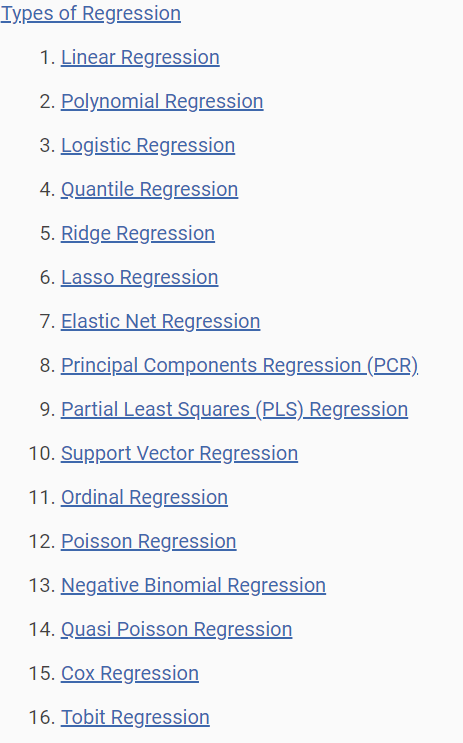
Hình 1.12: Machine Learning.

Nguồn: [<https://machinelearningcoban.com/>]

**3.4 Mô hình hồi quy trong Machine Learning**

Hồi quy là một trong những thuật toán cơ bản nhất của Machine Learning. Học có giám sát (Supervised Learning) được chia ra làm 2 dạng lớn là hồi quy (regression) và phân loại (classification) dựa trên tập dữ liệu mẫu - tập huấn luyện (training data).

Hiện nay có tổng cộng 16 loại mô hình hồi quy:



Hình 1.13: Các mô hình hồi quy.

Nguồn: [[https://www.listendata.com](https://www.listendata.com/2018/03/regression-analysis.html)]

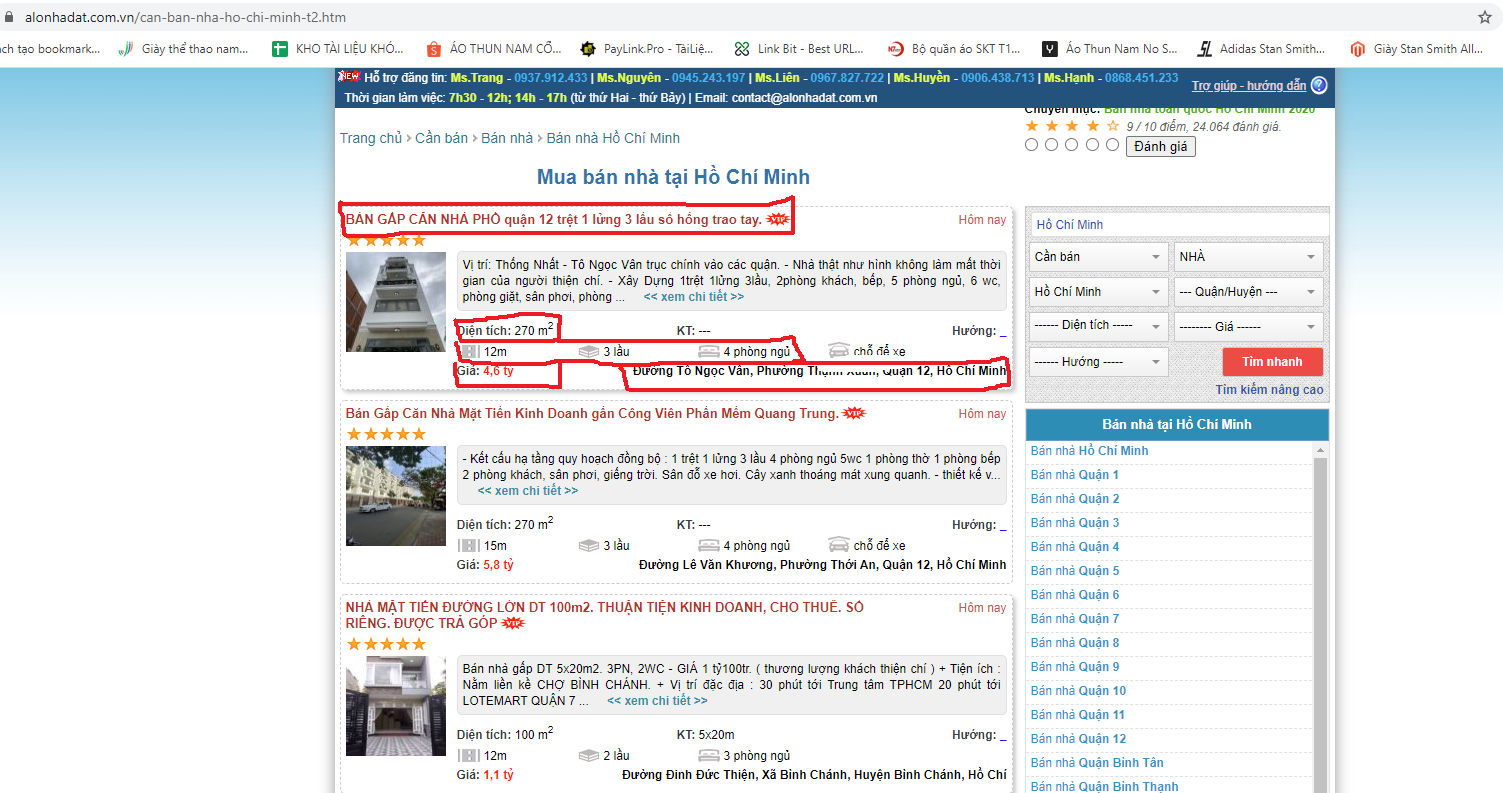
CHƯƠNG 4 – QÚA TRÌNH THỰC HIỆN

**4.1 Thu thập dữ liệu.**

Tại Việt Nam vẫn chưa tồn tại một chỉ số bất động sản như tại các nước phát triển. Không những vậy thông tin về thị trường này cũng không có và nếu có cũng không thể kiểm chứng được. Về phía cơ quan chính phủ - mà cụ thể là cơ quan thống kê quốc gia GSO cũng không đưa ra các báo cáo thống kê về thị trường nhà ở.

Nhằm vượt qua những trở ngại trên, đồ án này sử dụng một cách tiếp cận đơn giản sau: *thu thập và khai thác dữ liệu từ những tin nhắn rao bán bất động sản ở khu vực tp HCM được đăng trên*[*https://alonhadat.com.vn*](https://alonhadat.com.vn/)*- một website chuyên về giao dịch bất động sản.*

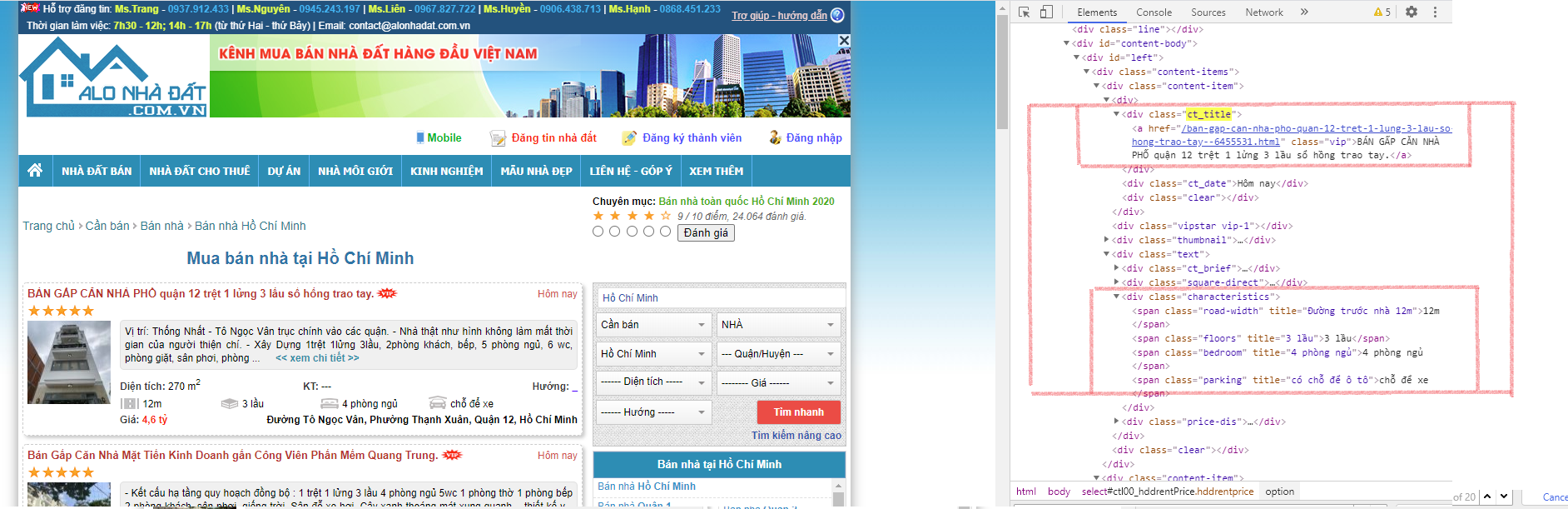
Những thông tin thu thập trên website này sẽ bao gồm: Tên bất động sản (name) ,Diện tích (square\_meter), Đường trước nhà (road-width), Giá (price), Số phòng ngủ (bedroom\_no), Số lầu (floors), Đường (street), Khu vực (Area) và Quận (District).



Hình 1.14: Trang web được chọn để thu thập dữ liệu.

Nguồn: [[https://alonhadat.com.vn](https://alonhadat.com.vn/can-ban-nha-ho-chi-minh-t2.htm)]

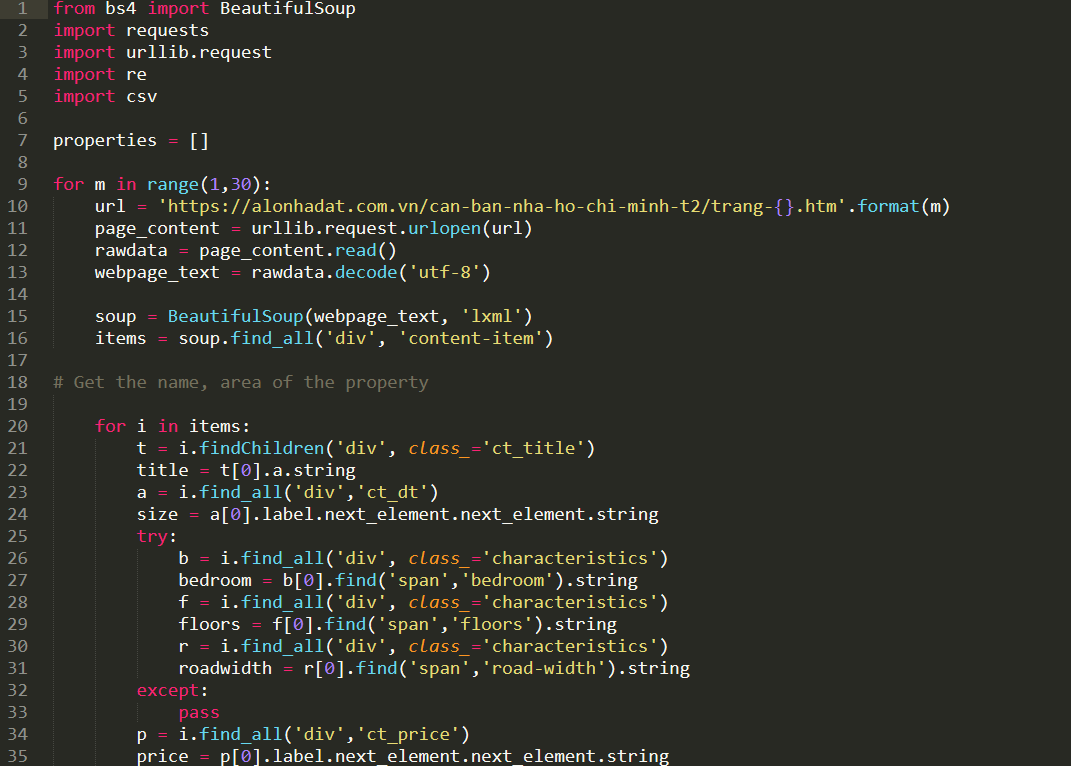
Trước hết muốn code lấy dữ liệu chúng ta phải xác định dữ liệu muốn lấy nằm ở những phần tử html nào trên trang web.



Hình 1.15: Xác định các phần tử html chứa dữ liệu.

Nguồn: [[https://alonhadat.com.vn](https://alonhadat.com.vn/can-ban-nha-ho-chi-minh-t2.htm)]

Thực hiện code để crawl dữ liệu:



Hình 1.16: Code crawl dữ liệu.

Tên bất động sản (title) sẽ nằm trong thẻ <Div> có class = ‘ct\_title’

Những thông tin còn lại sẽ nằm ở thẻ <Div> cha có class = ‘text’ được chứa trong thẻ <Div> con có class =’characteristic’. Chúng sẽ được chứa trong thẻ <Span>. Dùng hàm find() để tìm đúng các thông tin cần thiết.



Hình 1.17: Ghi dữ liệu vào file.

Các thông tin về Giá (price) và địa chỉ sẽ được làm tương tự.

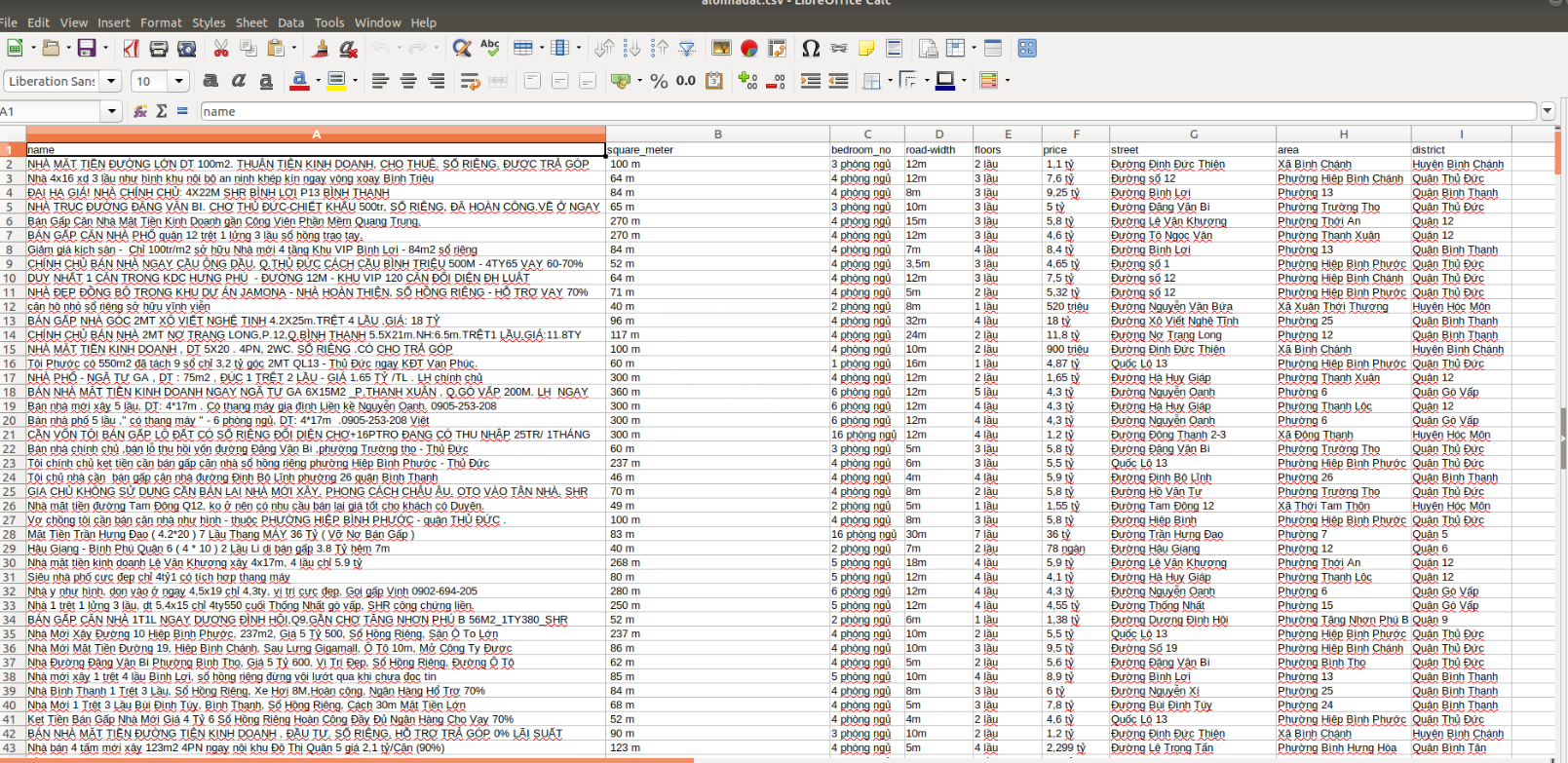
Sau đó, sẽ tạo một đối tượng property để lưu các thông tin.

Cuối cùng sẽ là bước ghi thông tin đã thu thập được trên website vào tệp ‘alonhadat.csv’, các thông tin sẽ tương ứng với các trường trong tệp csv.

Chạy chương trình crawl dữ liệu.



Hình 1.18: Chạy chương trình.

Dữ liệu thu được:

Hình 1.19: Dữ liệu đã thu được

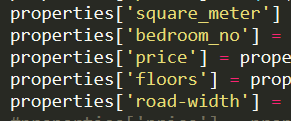
**4.2 Làm sạch và chuẩn bị dữ liệu**

Để chuẩn bị cho việc dự đoán, dữ liệu phải được làm sạch. Những dữ liệu không đúng, bị lỗi phải được bỏ đi thì việc dự đoán mới thực sự tốt.

Như dữ liệu ban đầu thì chúng ta có thế thấy 5 trường dữ liệu ở hình dưới có thể chuyển đổi và đưa vào mô hình hồi quy để dự đoán.

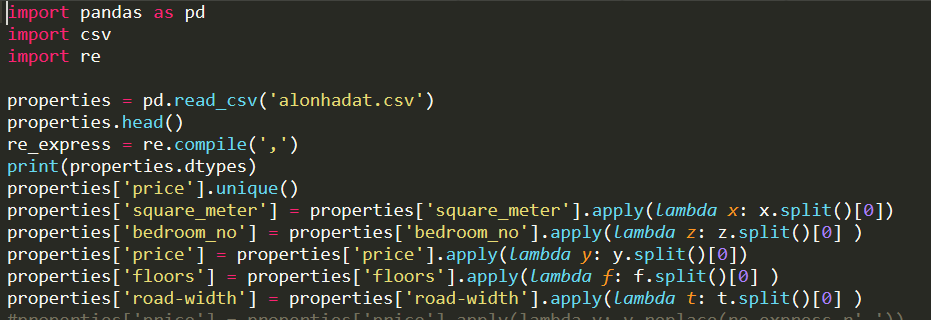


Hình 1.20: Trường dữ liệu có thể đem vào mô hình dự đoán



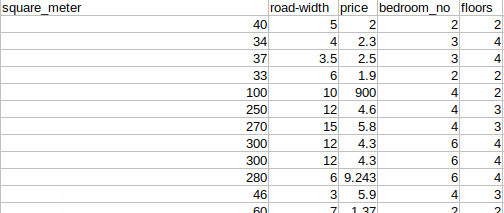
Hình 1.21: Dữ liệu đã thu được

Trước hết, phải bỏ đi các đơn vị, chữ hay kí tự ở đằng sau con số.



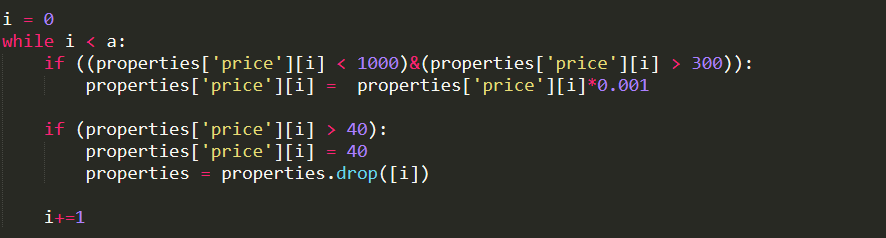
Hình 1.22: Code giữ lại phần số

Và kết quả thu được.



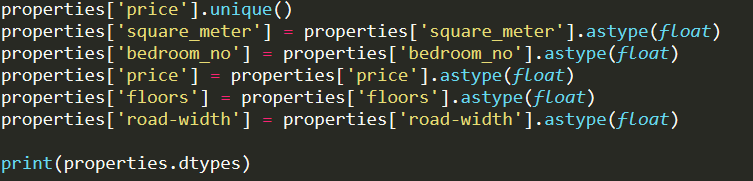
Hình 1.23: Kết quả đã giữ lại phần số

Tuy nhiên, vẫn còn một vấn đề ở trường price, đơn vị tỷ và triệu thể hiện trên con số vẫn chưa thể đưa về cùng một đơn vị. Chính vì vậy, sẽ nhân con số có đơn vị triệu với 0.001 để cũng đơn vị với tỷ.



Hình 1.24: Phần code chuyển đổi mức giá cùng đơn vị.

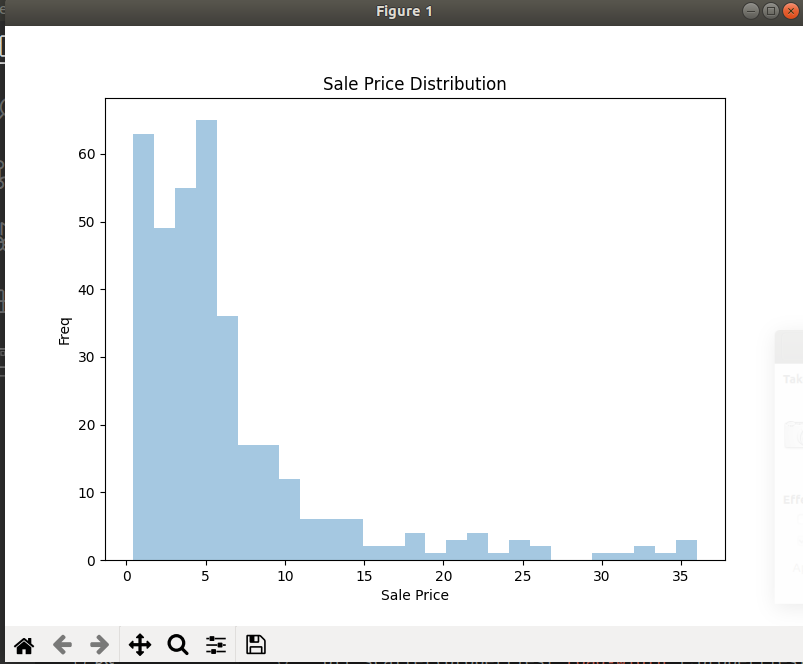
Kiểm tra kiểu dữ liệu của các trường thì thấy rằng 5 trường dữ liệu có số đang là kiểu Object. Tiếp theo sẽ tiến hành chuyển đổi sang kiểu dữ liệu Float.



Hình 1.25: Phần code chuyển đổi kiểu dữ liệu.

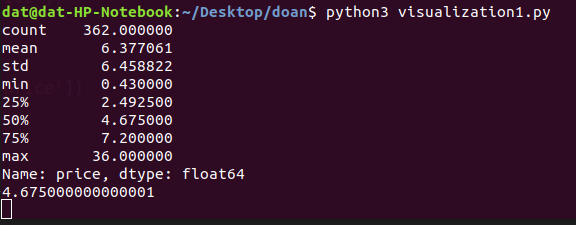
**4.3 Trực quan hóa dữ liệu**

Sau khi đã dọn dẹp dữ liệu xong thì chúng ta sẽ đưa dữ liệu thể hiện dưới dạng đồ thị. Dưới đây là đồ thì thể hiện giá bất động sản (đơn vị tỷ ) theo tuần suất xuất hiện.



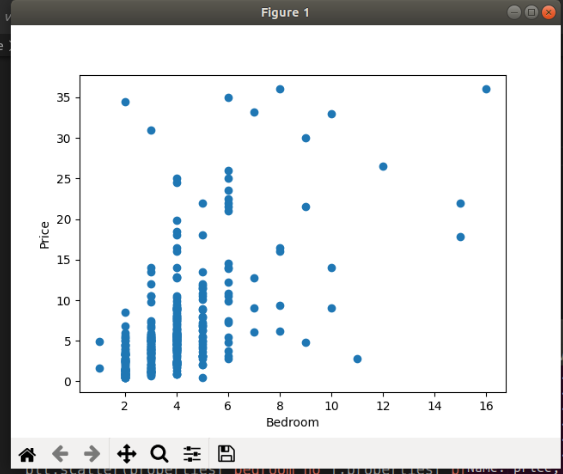
Hình 1.26: Đồ thị thể hiện tuần suất xuất hiện của từng mệnh giá.

Đây là kết quả sau khi tính phương sai (std-standard deviation) là 6.458822 và độ lệch chuẩn ( mean ) = 6.377061, trung vị ( median) là 4.675000000001 của giá nhà đất bằng hàm describe() và median().



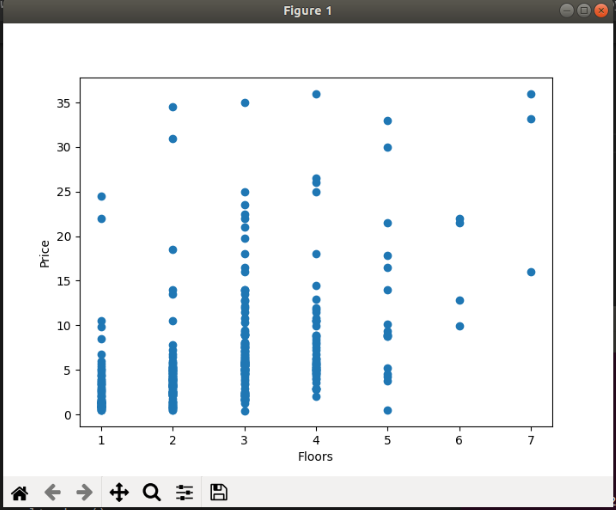
Hình 1.27: Kết quả tính trung vị, độ lệch chuẩn và phương sai.

Sự tương quan giữa giá nhà và số phòng ngủ có tỉ lệ trung hòa với nhau, số phòng ngủ không phản ảnh tốt mức giá như thế nào.



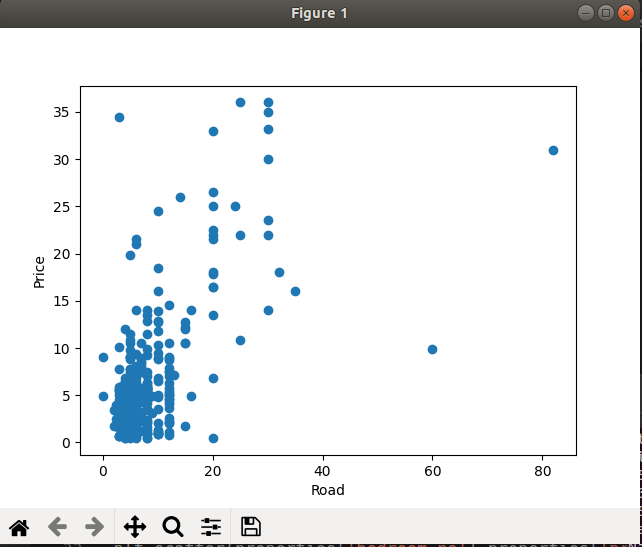
Hình 1.28: Đồ thị phân bố dữ liệu giữa giá nhà và số phòng ngủ.

Sự tương quan giữa giá nhà và số lầu theo tỉ lệ thuận, giá nhà bắt đầu tăng theo tỉ lệ với số lầu.



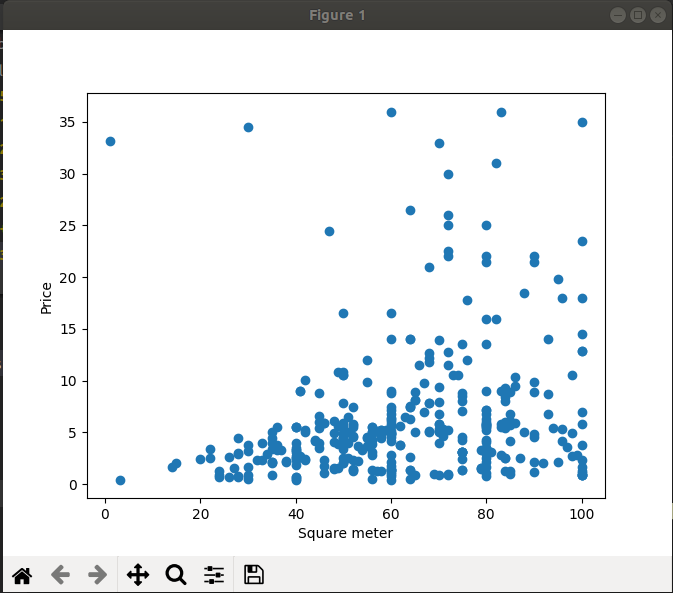
Hình 1.29: Đồ thị phân bố dữ liệu giữa giá nhà và số lầu.

Sự tương quan giữa giá nhà và chiều rộng đường trước nhà theo chiều thuận, tức là giá nhà tăng khi đường trước nhà càng rộng.



Hình 1.30: Đồ thị phân bố dữ liệu giữa giá nhà và chiều dài đường trước nhà.

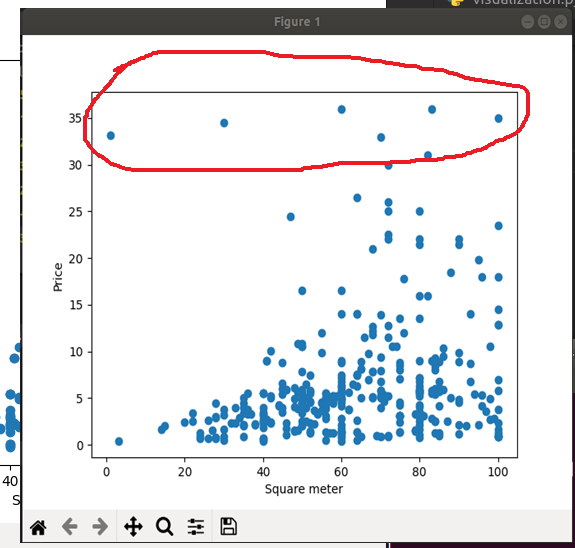
Sự tương quan giữa giá nhà và diện tích theo tỷ lệ thuận, giá nhà sẽ cao nếu diện tích càng lớn.



Hình 1.31: Đồ thị phân bố dữ liệu giữa giá nhà và diện tích.

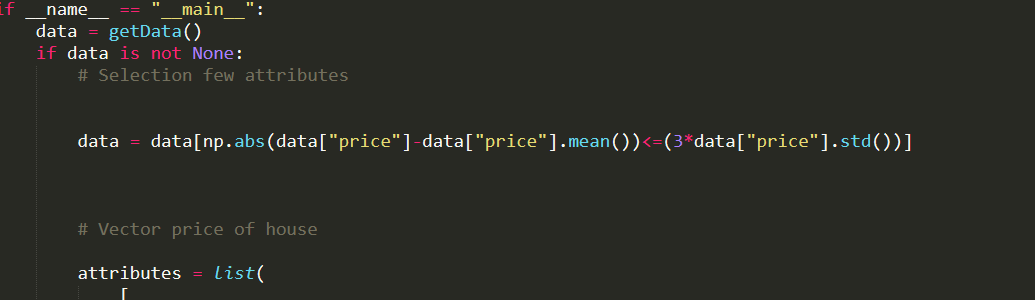
**4.4 Áp dụng mô hình hồi quy vào dự đoán giá nhà đất.**

Trước khi sử dụng các mô hình dự đoán thì chuyện cần làm tiếp theo đó chính là tiền xử lí dữ liệu. Các mô hình hồi quy rất nhạy với dữ liệu nhiễu. Vì vậy, cần phải loại bỏ bớt những dữ liệu đó.



Hình 1.32: Dữ liệu nhiễu.

Để loại bỏ những dữ liệu như thế, áp dụng độ lệch chuẩn (std) loại trừ mọi thứ có giá cao hơn hoặc thấp hơn 3 lần độ lệch chuẩn.



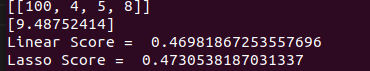
Hình 1.33: Code xử lý dữ liệu giá nhà đất nhiễu.

Sau khi xong bước tiền xử lý, chúng ta sẽ sử dụng hai mô hình hồi quy khá thông dụng là Linear Regression và Lasso Regression.



Hình 1.34: Triển khai mô hình dự đoán Linear Regression và Lasso Regression.

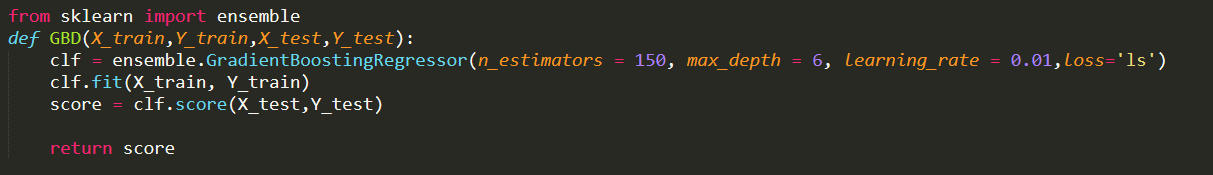
Và kết quả ( Score ) đạt được lần lượt là 0.47 và 0.473 với hai mô hình.



Hình 1.35: Kết quả mô hình dự đoán Linear Regression và Lasso Regression.

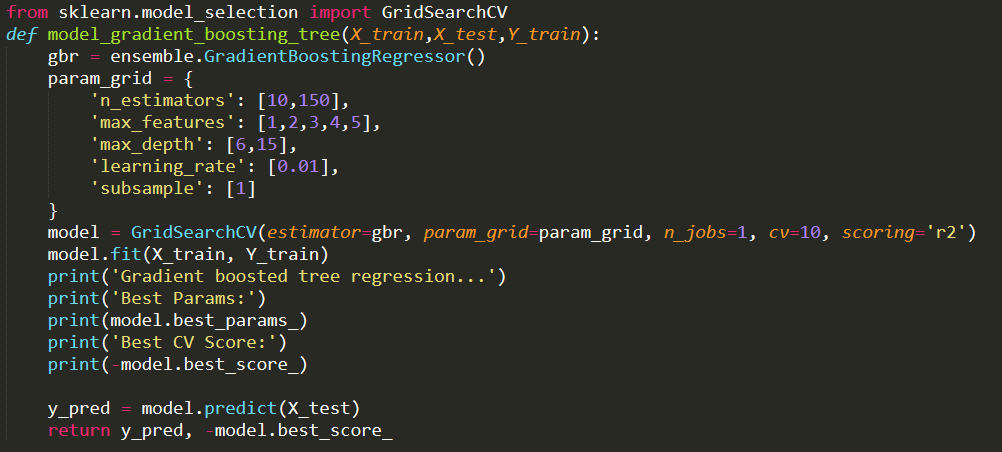
Kết quả không thực sự quá tốt vì nó không gần tới 1. Để có được kết quả tốt hơn, ở đây sẽ sử dụng một kỹ thuật và một mô hình khác.

Trước hết sẽ nói về mô hình gradient boosting regression, đây là mô hình hoàn toàn mới được nhiều chuyên gia sử dụng để nâng cao hiệu quả dự đoán.



Hình 1.36: Code triển khai mô hình gradient boosting regression.

Các tham số trong mô hình gradient boosting regression như n\_estimators, max\_depth, learning\_rate, loss có thể được chọn tùy mỗi người. Tuy nhiên việc lựa chọn này nó sẽ ảnh hưởng đến hiệu quả của mô hình dựa vào tập dữ liệu của bạn. Chính vì lẽ đó, một kỹ thuật để chọn ra những tham số tới ưu cho mô hình sẽ được áp dụng vào đề tài để nâng cao hiệu quả của mô hình. Đó là kỹ thuật **GridSearchCV**.



Hình 1.37: Code áp dụng kỹ thuật GridSearch vào

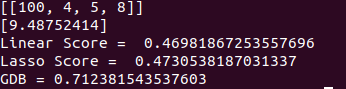
mô hình Gradient boosting regression .

Các tham số “best” tốt nhất được chọn sau khi áp dụng kỹ thuật GridSearchCV.



Hình 1.38: Các tham số tốt nhất để tối ưu hiệu quả dự đoán của mô hình Gradient boosting regression

Và cuối cùng là kết quả dự đoán mà mô hình Gradient boosting regression sau khi áp dụng kỹ thuật GridSearchCV.



Hình 1.39: Kết quả dự đoán giá nhà với mô hình Gradient boosting regression.

Kết quả có phần khả quan hơn so với hai mô hình Linear Regression và Lasso Regression.

CHƯƠNG 5 – KẾT QUẢ ĐẠT ĐƯỢC.

Sau khi đã hoàn thành xong đề tài Dự đoán sự thay đổi và giá nhà đất em đã có thể hiểu sâu hơn về thị trường bất động sản qua đó hình thành một nền kiến thức ổn trong lĩnh vực tài chính kinh tế trong môi trường công nghệ thông tin mà đặc biết là khoa học máy tính. Không những vậy, những kiến thức và những lúc thực hành về khai phá dữ liệu đã giúp bản thân em có thêm nhiều kỹ năng hơn về dữ liệu – một kỹ năng rất quan trọng trong ngành. Ngoài ra, kiến thức về Machine Learning mà em đã học khi thực hiện đề tài này cũng rất hay và bổ ích, nó cho em một cái nhìn rất hay về học máy.

Kết quả mà chúng em đạt được sau khi thực hiện đề tài cũng còn vài điểm hạn chế. Em vẫn chưa làm quá tốt việc khử dữ liệu nhiễu và kết quả dự đoán còn đôi chút hạn chế. Tuy nhiên, đây cũng đã là sự nỗ lực cố gắng, tâm huyết nhất mà chúng em đã mang vào đề tài. Em mong nhận được sự ủng hộ của các giảng viên.

**TÀI LIỆU THAM KHẢO**

**Tiếng Việt**

1. <https://viblo.asia/p/web-crawling-voi-thu-vien-beautifulsoup-1VgZvNGOZAw>
2. <http://scholar.vimaru.edu.vn/sites/default/files/thinhnv/files/dm_-_chapter_2_-_preprocessing_0.pdf>
3. <https://topdev.vn/blog/pandas-python-tutorial/>
4. <https://github.com/MBKraus/Predicting_real_estate_prices_using_scikit-learn/blob/master/Models.py>
5. <https://machinelearningcoban.com/2016/12/26/introduce/#:~:text=Machine%20Learning%20l%C3%A0%20m%E1%BB%99t%20t%E1%BA%ADp%20con%20c%E1%BB%A7a%20AI.&text=N%C3%B3i%20%C4%91%C6%A1n%20gi%E1%BA%A3n%2C%20Machine%20Learning,%C4%91%C6%B0%E1%BB%A3c%20l%E1%BA%ADp%20tr%C3%ACnh%20c%E1%BB%A5%20th%E1%BB%83.>
6. <https://www.wikihow.vn/T%C3%ACm-Mean,-Median,-v%C3%A0-Mode>
7. <https://techblog.vn/cach-tiep-can-trong-bai-toan-khai-pha-va-phan-tich-du-lieu-trong-machine-learning>

**Tiếng Anh**

1. <https://github.com/MBKraus/Predicting_real_estate_prices_using_scikit-learn/blob/master/Models.py>
2. <https://www.listendata.com/2018/03/regression-analysis.html>

**PHỤ LỤC**