TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM

KHOA ĐÀO TẠO CHẤT LƯỢNG CAO

Chuyên Ngành: Công Nghệ Thông Tin

-------*-------**

MACHINE LEARNING

ĐỀ TÀI: ỨNG DỤNG THUẬT TOÁN HỌC GIÁM SÁTĐỂ ĐÁNH GIÁ BẤT ĐỘNG SẢN

Giảng viên:

Thầy Vũ Quang Huy

Sinh viên thực hiện:

Phạm Thu Thảo	**15110125**
Nguyễn Thị Phi Vân	**15110149**
Cao Nguyễn Vũ Toàn	**15110147**

Tp. HCM, tháng 12 năm 2018

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

Ngày.....Tháng.....Năm.....

Giáo Viên Hướng Dẫn

(Ký, ghi rõ họ và tên)

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

CHƯƠNG 1. ĐẶT VẤN ĐỀ - GIỚI THIỆU ĐỀ TÀI

1.1. Đặt vấn đề

Mỗi người chúng ta thường sẽ thực hiện giao dịch bất động sản ít nhất một lần trong đời. Số tiền dành cho mua nhà là không nhỏ, vì vậy việc người mua quan tâm không chỉ ở việc lựa chọn được một ngôi nhà ưng ý mà còn xem giá cả có hợp lý hay không. Bên cạnh đó, hiện nay nhu cầu sống của người dân ngày càng tăng cao theo thu nhập bình quân đầu người. Xuất phát từ ý nghĩ "An cư lạc nghiệp" mà có ý định đầu tư một căn căn hộ chất lượng, cao cấp của người dân đã dần trở nên phổ biến. Nhưng không phải bất kỳ ai cũng có thể định giá chính xác cho ngôi nhà hoặc căn hộ của mình. Đôi khi những sự nhầm lẫn hoặc sự thiếu kinh nghiệm cũng có thể dẫn đến dự đoán sai và dẫn đến thua lỗ là điều không thể tránh khỏi. Ngay cả những doanh nghiệp về bất động sản thì đây cũng là một vấn đế khá nan giải. Vậy làm sao, dựa vào đâu để đưa ra một báo giá chính xác nhất, gần với thực tế nhất, cũng chính là lo nhóm em chọn đề tài "Dự báo giá nhà đất".

1.2. Lý do chọn đề tài

Xuất phát từ ý định mang đến không chỉ người mua mà còn người bán một công cụ có thể báo giá chính xác nhất có thể, từ đó góp phần giúp thị trường bất động sản chuyên nghiệp hơn, giảm tối thiểu tổn thất thua lỗ nhất có thể.

1.3. Ý nghĩa

1.3.1. Ý nghĩa khoa học

- Tìm hiểu sâu hơn về cơ sở lý thuyết, làm rõ tính ứng dụng của thuật toán được áp dụng trong đề tài.
- Đề xuất phương pháp dự đoán giá nhà đất ổn định và chính xác bằng cách áp dụng thuật toán trong Machine Learning.

1.3.2. Ý nghĩa thực tiễn

- Kết quả sau khi nghiên cứu giúp người dùng dự đoán giá nhà đất tiện lợi và chính xác, tránh thua lỗ do thiếu kinh nghiệm hoặc sai sót.
- Định hướng sự phát triển của ngành bất động sản dựa trên kết quả nghiên cứu.

1.4. Mục tiêu

- Mang đến cho người dùng một công cụ báo giá gần thực tế nhất, đáng tin cậy.
- Dữ liệu cần cho việc dự đoán phải tinh gọn, không quá phức tạp. Nghĩa là bất kỳ ngôi nhà cũng đáp ứng đầy đủ những yêu cầu cho việc dự đoán.
- Mang ứng dụng phổ biến đến người quan tâm đến mua bán nhà nói riêng và bất động sản nói chung.
- Đáp ứng được nhu cầu phát triển đối với dữ liệu toàn cầu.

CHƯƠNG 2. LÝ THUYẾT – LINEAR REGRESSION

2.1. Linear regression

2.1.1. Linear là gì?

Vì ta cần tìm ra 1 đường thẳng, đường thẳng trong tiếng Anh là line, linear là "có thể biểu diễn bằng 1 đường thẳng"—trong toán học Việt Nam còn gọi là "tuyến tính" (tuyến trong tiếng Hán Việt nghĩa là đường thẳng—tuyến tính là có tính chất của đường thẳng).

2.1.2. Regression là gì?

Là một cái tên lấy từ bộ môn thống kê (statistics), một thuật toán/khái niệm mang từ bên statistics về dùng.
Trong ngành thống kê: regression là mối quan hệ giữa giá trị đầu ra (y) với các giá trị (biến) đầu vào (x, t …).

2.1.3. Linear regression là gì?

Linear regression là một phương pháp (thuật toán) thuộc loại đơn giản nhất trong Machine Learning, trong tiếng Việt gọi là *hồi quy tuyến tính*, với mục đích TÌM (vẽ) ra một đường thẳng, sao cho nó đi qua hoặc đi gần nhất với các điểm cho trước. Từ một tập dữ liệu cho trước, khi ta vẽ được một đường thẳng như vậy, ta có thể đoán xem các điểm khác sẽ nằm ở đâu.

Có thể nói mục tiêu của giải thuật hồi quy tuyến tính là dự đoán giá trị của một hoặc nhiều biến mục tiêu liên tục (continuous target variable) **y** dựa trên một véc-to đầu vào **x**.

Linear regression đóng một vai trò quan trọng trong lĩnh vực trí thông minh nhân tạo như Machine Learning . Thuật toán hồi quy tuyến tính là một trong những thuật toán học máy được giám sát cơ bản do tính đơn giản tương đối và các thuộc tính nổi tiếng của nó.

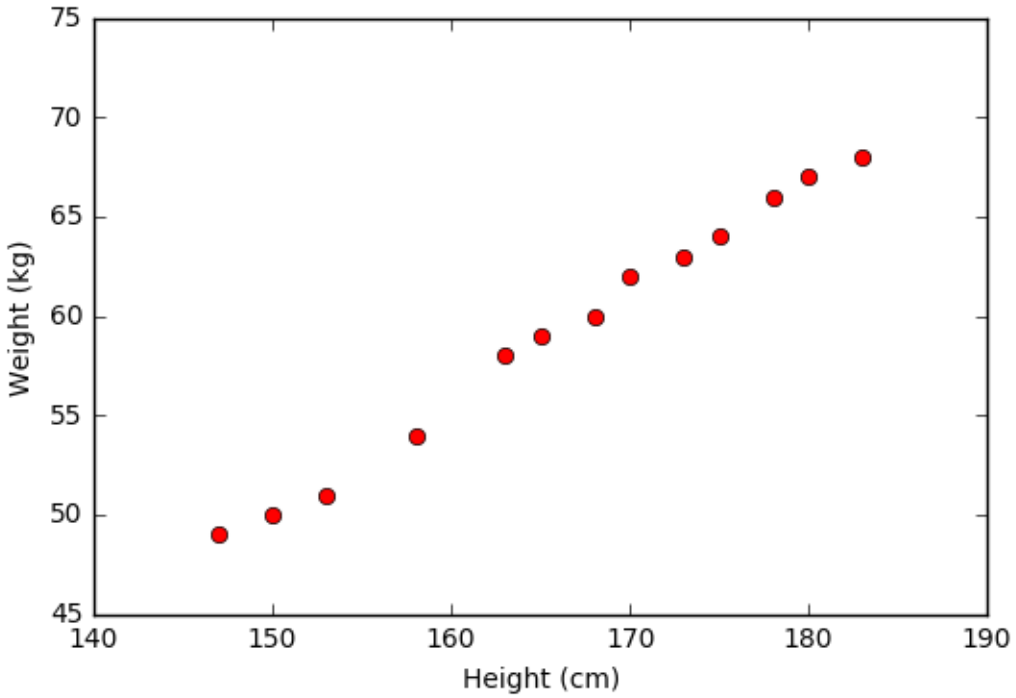
Xét ví dụ về chiều cao và cân nặng, ta có bảng dữ liệu như sau:

Chiều cao (cm)	Cân nặng (kg)	Chiều cao (cm)	Cân nặng (kg)
147	49	168	60
150	50	170	72
153	51	173	63
155	52	175	64
158	54	178	66
160	56	180	67
163	58	183	68
165	59		

Bảng 1. Bảng chiều cao, cân nặng tương ứng

Chúng ta có thể thấy là cân nặng sẽ tỉ lệ thuận với chiều cao (càng cao càng nặng), nên có thể sử dụng Linear Regression model cho việc dự đoán này.

Ta thể hiện lên đồ thị để thấy rõ hơn sự tuyến tính trong ví dụ này:



Hình 1. Hình thể hiện sự tương ứng của chiều cao và cân nặng [1]

Từ đồ thị này ta thấy rằng dữ liệu được sắp xếp gần như theo 1 đường thẳng, vậy mô hình Linear Regression nhiều khả nSăng sẽ cho kết quả tốt:

(cân nặng) = w_1*(chiều cao) + w_0

2.1.4. Các giả định trong phân tích hồi qui tuyến tính

Phân tích hồi qui tuyến tính không chỉ là việc mô tả các dữ liệu quan sát được trong mẫu (sample) nghiên cứu mà cần phải suy rộng cho mối liên hệ trong dân số (population). Vì vậy, trước khi trình bày và diễn dịch mô hình hồi qui tuyến tính cần phải dò tìm vi phạm các giả định. Nếu các giả định bị vi phạm thì các kết quả ước lượng không đáng tin cậy được.

Các giả định cần thiết trong hồi qui tuyến tính:

- x_i là biến số cố định, không có sai sót ngẫu nhiên trong đo lường.
- Phần dư (trị số quan sát trừ cho trị số ước đoán) phân phối theo luật phân phối chuẩn
- Phần dư có trị trung bình bằng 0 và phương sai không thay đổi cho mọi trị x_i
- Không có tương quan giữa các phần dư

2.1.5. Một vài tính chất của hồi quy tuyến tính

- Đường hồi quy luôn luôn đi qua trung bình của biến độc lập (x) cũng như trung bình của biến phụ thuộc (y)
- Đường hồi qui tối thiểu hóa tổng của "Diện tích các sai số". Đó là lý do tại sao phương pháp hồi quy tuyến tính được gọi là "Ordinary Least Square (OLS)"

- B1 giải thích sự thay đổi trong Y với sự thay đổi X bằng một đơn vị. Nói cách khác, nếu chúng ta tăng giá trị của X bởi một đơn vị thì nó sẽ là sự thay đổi giá trị của Y

2.1.6. Nhược điểm của Linear regression

- Hạn chế đầu tiên của Linear Regression là nó rất **nhạy cảm với nhiễu** (sensitive to noise)
- Hạn chế thứ hai của Linear Regression là nó **không biểu diễn được các mô hình phức tạp**

2.2. K-nearest neighbor

2.2.1. Định nghĩa

K-nearest neighbor (KNN) là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning. Khi training, thuật toán này *không học* một điều gì từ dữ liệu training (đây cũng là lý do thuật toán này được xếp vào loại lazy learning), mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới. K-nearest neighbor có thể áp dụng được vào cả hai loại của bài toán Supervised learning là Classification và Regression. KNN còn được gọi là một thuật toán Instance-based hay Memory-based learning.

2.2.2. Ý tưởng của thuật toán

Thuật toán có 2 đầu vào, một là tập các dữ liệu đã biết trước kiểu(loại) của từng dữ liệu (hay còn gọi là tập huấn luyện - training set), đầu vào thứ 2 là dữ liệu, chúng ta chưa biết kiểu(loại) dữ liệu đó. Đầu ra của thuật toán kNN là kiểu dữ liệu của đầu vào thứ 2.

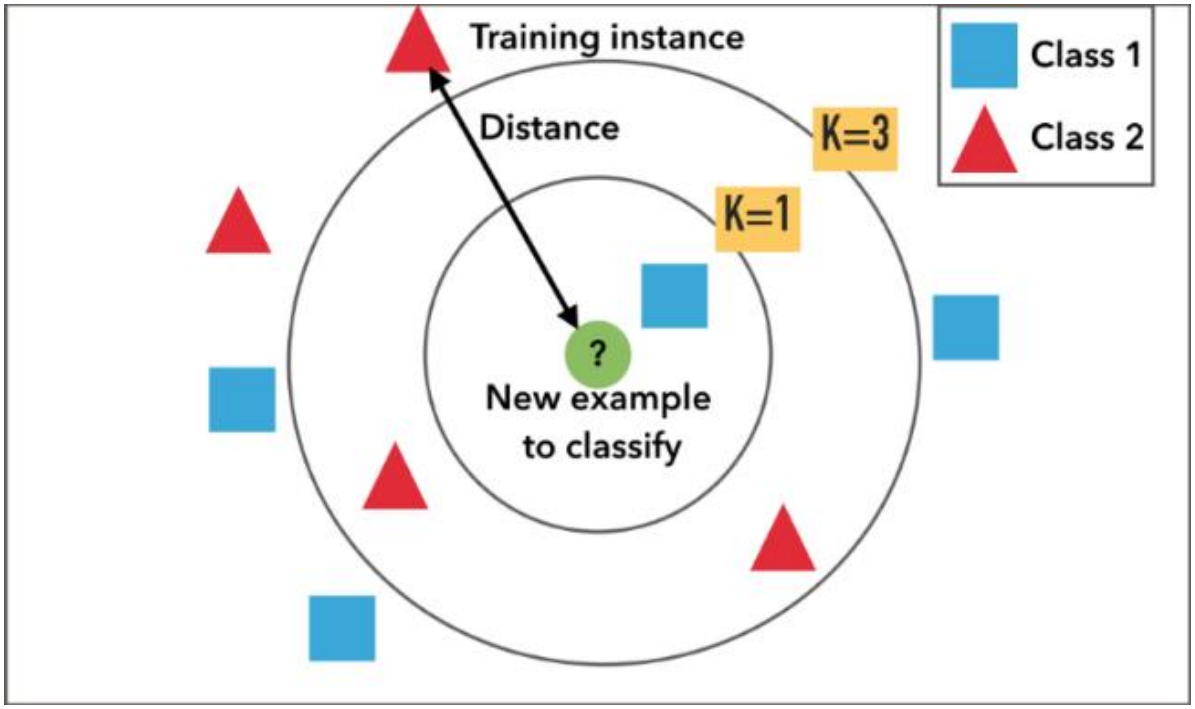
Vậy ta có thể hiểu KNN là thuật toán tìm label của một input dựa vào K điểm dữ liệu lân cận gần nó nhất (tức là điểm input).

2.2.3. Cách hoạt động của KNN

Dưới đây trình bày từng bước cách sử dụng KNN trong việc dự đoán với biến phụ thuộc định lượng

- Xác định tham số K (số láng giềng gần nhất)
- Tính khoảng cách (Distance) giữa Query point và tất cả training samples
- Sắp xếp khoảng cách và xác định K láng giềng gần nhất với Query point
- Lấy giá trị của biến phụ thuộc Y tương ứng của K láng giềng gần nhất
- Sử dụng giá trị trung bình (average) của biến phụ thuộc Y của K láng giềng gần nhất là giá trị dự đoán của Query point.

Ta có ví dụ đơn giản, như hình dưới đây:



Hình 2. Hình minh họa thuật toán K-nearest neighbor [2]

Nhìn vào hình, hình tròn màu xanh lá chính là đối tượng cần được phân loại, xung quanh nó có các hình vuông (Category = class 1) và hình tam giác (Category = class 2) khác.

- Khi k=1, category của đối tượng mà tương đồng với nó nhất sẽ được chọn, đó chính là hình vuông;
- Khi k=3, trong số 3 đối tượng tương đồng với nó nhất có đến 2 hình tam giác, do đó category tam giác sẽ được chọn.
- Với mỗi k láng giềng ta sẽ thu được nhãn phù hợp dữ liệu đầu vào

2.2.4. Ưu điểm của KNN

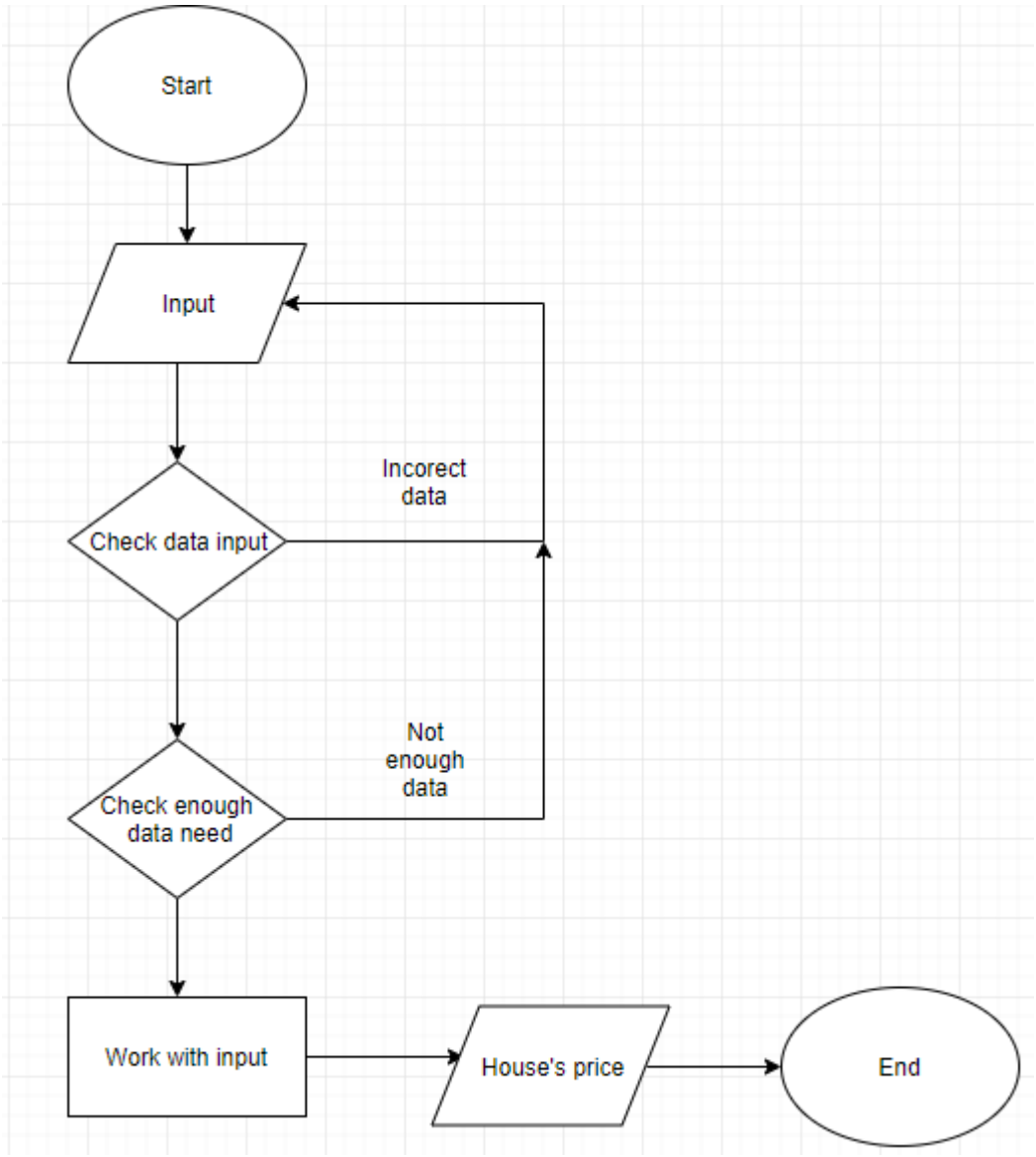
- Độ phức tạp tính toán của quá trình training là bằng 0. Do thuật toán này đơn giản gần như không phải học bất cứ điều gì.
- Việc dự đoán kết quả của dữ liệu mới rất đơn giản.
- Không cần giả sử gì về phân phối của các class.

2.2.5. Nhược điểm của KNN

- KNN rất nhạy cảm với nhiễu khi K nhỏ.
- Như đã nói, KNN là một thuật toán mà mọi tính toán đều nằm ở khâu test. Trong đó việc tính khoảng cách tới từng điểm dữ liệu trong training set sẽ tốn rất nhiều thời gian, đặc biệt là với các cơ sở dữ liệu có số chiều lớn và có nhiều điểm dữ liệu. Với K càng lớn thì độ phức tạp cũng sẽ tăng lên. Ngoài ra, việc lưu toàn bộ dữ liệu trong bộ nhớ cũng ảnh hưởng tới hiệu năng của KNN.

CHƯƠNG 3. MODEL

3.1. Flowchart

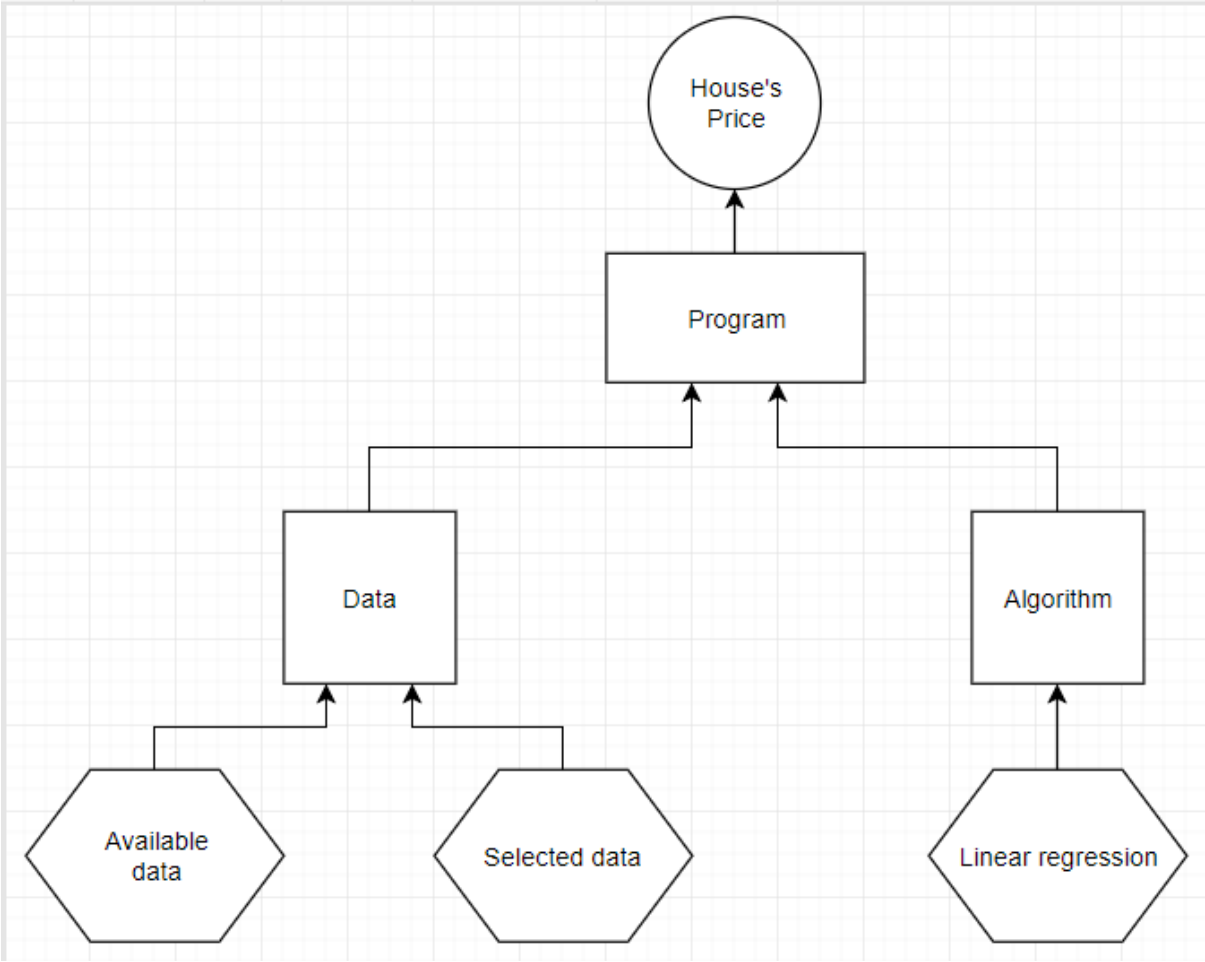


Hình 3. Hình Flowchart

Qua hình ảnh của flowchart ta có thể hình dung được các bước mà thuật toán này thực thi:

- Khi bắt đầu chương trình ta cần truyền vào dữ liệu cần thiết cho bài toán
- Để đảm bảo dữ liệu truyền vào là phù hợp ta phải kiểm tra trước khi thực thi
- Khi mọi dữ liệu đã được kiểm tra và hợp lệ, ta sẽ thực thi bài toán bằng giải thuật cùng với tập dữ liệu đã được chọn
- Xuất ra kết quả dự đoán cuối cùng và kết thúc chương trình.

3.2. Block diagram



Hình 4. Hình Block diagram

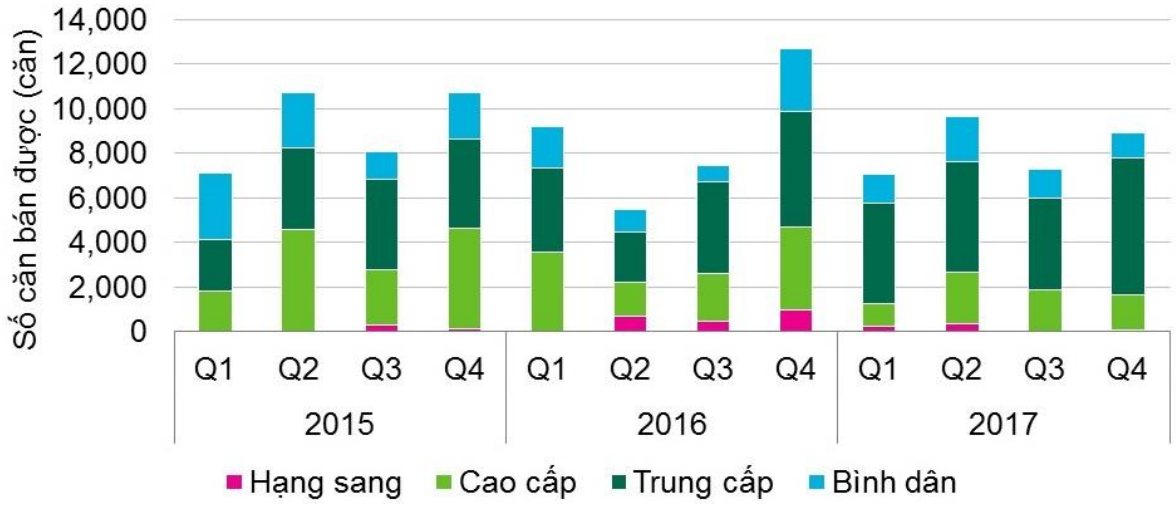
Qua blockdiagram ta hình dung được các thành phần tham cấu thành một phương pháp hỗ trợ dự đoán giá nhà đất:

- Dữ liệu: ta có thể sử dụng dữ liệu có sẵn nhưng phải qua quá trình chọn lọc vì một số dữ liệu không chính xác sẽ dẫn đến kết quả cuối cùng của bài toán sai hoàn toàn, và đây là điều không mong muốn.
- Giải thuật: giải thuật là phần quan trọng nhất, nó quyết định vấn đề có được giải quyết hay không và có chính xác không. Vì thế việc lựa chọn một giải thuật phù hợp là vô cùng quan trọng.

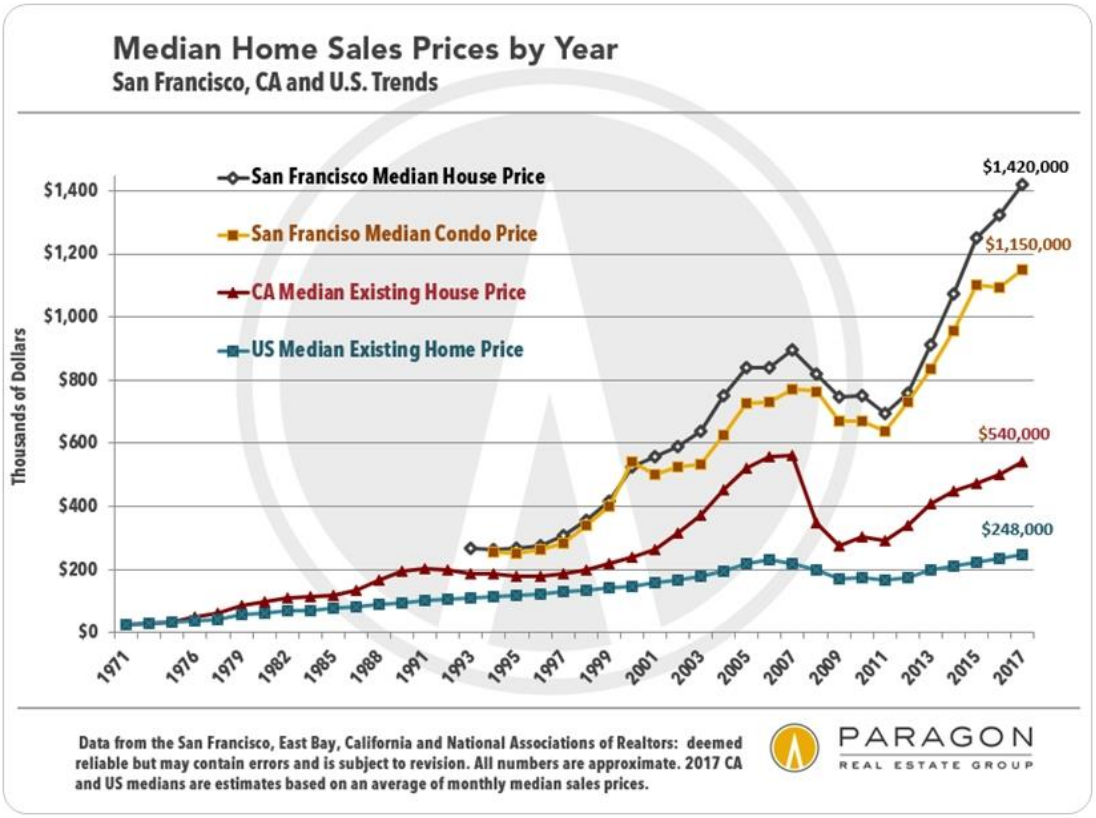
CHƯƠNG 4. GIẢI QUYẾT VẤN ĐỀ

4.1. Khảo sát

Để đánh giá chính xác giá của một căn nhà, người ta không chỉ đòi hỏi một sự hiểu biết chuyên môn về thị trường bất động sản (một thị trường rất biến động) mà còn đòi hỏi một sự hiểu biết thật sự tường tận về bản thân các thuộc tính của bất động sản đó (Mu et al., 2014). Những kiến thức này thường chỉ được lưu trữ bởi các đại lý kinh doanh bất động sản. Nếu chúng ta có thể nắm bắt kiến thức này bằng cách thu thập dữ liệu, sử dụng các dữ liệu mở, tận dụng sự giúp sức của các thuật toán, chương trình máy tính, các kiến thức này trở nên dễ tiếp cận hơn với các người dân bình thường, giúp đưa ra quyết định mà không cần dựa vào chuyên gia vì không may vị chuyên gia đó có thể tư vấn theo chiều hướng có lợi cho họ.



Hình 5. Biểu đồ tổng số lượng tiêu thụ căn hộ trong năm 2017 tại Hồ Chí Minh – Việt Nam [3]



Hình 6. Biểu đồ tổng số lượng tiêu thụ căn hộ trong năm 2017 tại một số nơi [2]

4.2. Phân tích, chọn lọc dữ liệu, mô hình dự đoán phù hợp

- *Tổng quan*
  - Cơ sở dữ liệu về thửa đất: Vị trí, toạ độ, hình thể. kích thước, diện tích, trích lục bản đồ thửa đất
  - Cơ sở dữ liệu pháp lý liên quan đến thửa đất (quyền sở hữu, sử dụng)
  - Cơ sở dữ liệu về nhà và các công trình xây dựng trên đất
  - Cơ sở dữ liệu về các căn hộ (địa chỉ và các hạn chế).
- *Các thông tin chi tiết*
  1. *Loại bất động sản:*
     Nêu cụ thể loại bất động sản: căn hộ chung cư, biệt thự, nhà liền kề hay nhà vườn..; văn phòng, cửa hàng, siêu thị, chợ, khách sạn, nhà nghỉ hay nhà trọ..; hạ tầng khu công nghiệp hay nhà xưởng, nhà máy, kho, bãi..; loại đất (đất ở, đất KCN, đất làm mặt bằng sản xuất kinh doanh...).
  2. *Vị trí bất động sản:*
     - Đối với nhà ở, công trình xây dựng..: nêu cụ thể số nhà, ngách, ngõ (hẻm), đường phố (thôn), phường (xã), quận (huyện), thành phố (tỉnh);
     - Đối với nhà chung cư: (ghi căn hộ số, tầng số, nhà chung cư số, khu đô thị, phường, quận, thành phố);

- Đối với bất động sản hình thành trong tương lai: (ghi: lô đất, biệt thự số, căn hộ số, chung cư số, dự án, khu đô thị, phường, quận, thành phố. Khuyến khích có sơ đồ vị trí kèm theo).

3. *Thông tin về quy hoạch:*

- Quy hoạch chi tiết của dự án hoặc mô hình.
- Thông tin liên quan đến bất động sản (tình hình giải phóng mặt bằng, hạ tầng kỹ thuật, hạ tầng xã hội… (nếu có).

4. *Quy mô, diện tích của bất động sản:*

- Đối với nhà ở, nhà chung cư: (ghi: tổng diện tích đất, diện tích xây dựng, tổng diện tích sàn, số tầng …).
- Đối với đất: (ghi: tổng diện tích đất, mật độ xây dựng, hệ số sử dụng đất, số tầng được phép xây dựng…).
- Đối với hạ tầng khu công nghiệp: (ghi: tổng diện tích đất, mật độ xây dựng, diện tích xây dựng, diện tích giao thông, đất cây xanh, diện tích công trình công cộng…).

5. *Đặc điểm, tính chất, công năng sử dụng, chất lượng của bất động sản:*

- Kết cấu công trình, móng, khung, tường, sàn, mái; công năng sử dụng bất động sản theo thiết kế ban đầu;
- Ảnh chụp toàn cảnh công trình tại thời điểm giới thiệu (nếu có); cấp, hạng công trình, năm xây dựng, chất lượng hiện tại…).

6. *Thực trạng hạ tầng kỹ thuật, hạ tầng xã hội (điện, cấp thoát nước, thông tin liên lạc, giao thông, nhà trẻ, trường học, chợ, bệnh viện…).*
7. *Tình trạng pháp lý của bất động sản, nêu tình hình hồ sơ, giấy tờ về quyền sở hữu, quyền sử dụng bất động sản và giấy tờ có liên quan đến việc tạo lập bất động sản; lịch sử về sở hữu, sử dụng bất động sản (nếu có).*
8. *Các hạn chế về quyền sở hữu, quyền sử dụng bất động sản (nếu có).*
9. *Giá bán, giá chuyển nhượng, giá cho thuê.*
10. *Quyền và lợi ích của người có liên quan.*
11. *Các thông tin liên quan đến chủ đầu tư, chủ sở hữu, chủ sử dụng bất động sản và các thông tin khác (nếu có).*

Theo như những loại dữ liệu ta thu thập được người ta sẽ đưa ra một số mô hình dự đoán.

Hiện nay, người ta chủ yếu ước lượng giá bất động sản dựa trên các phương pháp truyền thống như phương pháp so sánh trực tiếp, chiết trừ, thu nhập, thặng dư, hệ số điều chỉnh. Các phương pháp này chủ yếu nhờ sự phân tích và can thiệp của nhân viên định giá nên rất khó tránh khỏi sai lầm do chủ quan hoặc không minh bạch (Quỳnh và cs., 2015).

Ngoài các phương pháp truyền thống, trên thế giới đã và đang nghiên cứu và áp dụng rộng rãi các phương pháp có sử dụng đến các mô hình toán học để xác định giá trị bất động sản. Mới nhất là công trình (Król, 2015) sử dụng mô hình hodenic để mô hình hóa giá bất động sản ở Ba Lan.

Một cách tổng quát, trong mô hình hoderic, hàm giá của bất động sản phụ thuộc vào các thuộc tính của nó như vị trí so với trung tâm, gần đường, gần các khu tiện ích, diện tích nhà, số phòng ngủ, số tầng, số phòng tắm...
Các mô hình để xác định hàm giá có thể là các mô hình đơn giản như mô hình tuyến tính hay các mô hình phức tạp hơn như mô hình mũ, mô hình logarit,...

4.3. Định hướng giải quyết vấn đề

Mô hình tuyến tính là một mô hình đơn giản và được sử dụng nhiều trong bài toán xác định giá bất động sản. Trong các nghiên cứu về giá bất động sản có sử dụng đến mô hình tuyến tính chúng ta có thể kể đến các nghiên cứu của (Christian el al., 2009; Richard, 2009). Hồi quy tuyến tính xác định một đường thẳng hay một mặt phẳng qua các điểm dữ liệu trong không gian thuộc tính. Giả sử giá của bất động sản là y và các thuộc tính ảnh hưởng đến giá của nó như diện tích, độ rộng mặt tiền, độ rộng đường vào nhà, tình trạng pháp lý của khu đất, tiện ích của khu dân cư (điều kiện vệ sinh, điều kiện trường học, y tế), khoảng cách đến trung tâm phường, quận, thành phố...

4.4. Quá trình huấn luyện

4.4.1. Tập dữ liệu

– Dưới đây là hình ảnh minh họa của dữ liệu:

2	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotSl	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond
3	1461	20	RH	80	11622	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Feedr	Norm	1Fam	1Story	5	6
4	1462	20	RL	81	14267	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	6
5	1463	60	RL	74	13830	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	5	5
6	1464	60	RL	78	9978	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	6	6
7	1465	120	RL	43	5005	Pave	NA	IR1	HLS	AllPub	Inside	Gtl	StoneBr	Norm	Norm	TwnhsE	1Story	8	5
8	1466	60	RL	75	10000	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	Gilbert	Norm	Norm	1Fam	2Story	6	5
9	1467	20	RL	NA	7980	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	1Story	6	7
10	1468	60	RL	63	8402	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	6	5
11	1469	20	RL	85	10176	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	1Story	7	5
12	1470	20	RL	70	8400	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	4	5
13	1471	120	RH	26	5858	Pave	NA	IR1	Lvl	AllPub	FR2	Gtl	NAmes	Norm	Norm	TwnhsE	1Story	7	5
14	1472	160	RM	21	1680	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	BrDale	Norm	Norm	Twnhs	2Story	6	5
15	1473	160	RM	21	1680	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	BrDale	Norm	Norm	Twnhs	2Story	5	5
16	1474	160	RL	24	2280	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	NPkVill	Norm	Norm	Twnhs	2Story	6	6
17	1475	120	RL	24	2280	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	NPkVill	Norm	Norm	Twnhs	1Story	7	6
18	1476	60	RL	102	12858	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	2Story	9	5
19	1477	20	RL	94	12883	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NridgHt	Norm	Norm	1Fam	1Story	8	5
20	1478	20	RL	90	11520	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	PosN	Norm	1Fam	1Story	9	5
21	1479	20	RL	79	14122	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	1Story	8	5
22	1480	20	RL	110	14300	Pave	NA	Reg	HLS	AllPub	Inside	Mod	NridgHt	Norm	Norm	1Fam	1Story	9	5
23	1481	60	RL	105	13650	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NridgHt	Norm	Norm	1Fam	2Story	8	5

Hình 7. Hình minh họa tập dữ liệu

– Phân tích tập dữ liệu

STT	Tên trường	Kiểu dữ liệu	Ý nghĩa	Ghi chú
1	Id	Int64		
2	MSSubClass	Int64	Xác định các loại nhà có liên quan đến việc bán hàng.	

3	MSZoning	Object	Xác định việc phân loại quy hoạch chung của việc bán hàng.	
4	LotFrontage	Float64	Chân tuyến tính của đường phố kết nối với căn hộ	
5	LotArea	Int64	Kích thước lô	Tính bằng feet vuông
6	Street	Object	Loại đường vào căn hộ	
7	Alley	Object	Loại hẻm vào căn hộ	
8	LotShape	Object	Hình dạng chung	
9	LandContour	Object	Độ phẳng	
10	Utilities	Object	Loại tiện ích có sẵn	
11	LotConfig	Object	Cấu hình lô	
12	LandSlope	Object	Độ dốc	
13	Neighborhood	Object	Các vị trí thực tế giới hạn thành phố Ames	
14	Condition1	Object	Tiếp giáp với các điều kiện khác	
15	Condition2	Object	Tiếp giáp với các điều kiện khác	Nếu có nhiều hơn một
16	BldgType	Object	Loại nhà ở	
17	HouseStyle	Object	Phong cách nhà ở	
18	OverallQual	Int64	Đánh giá vật liệu tổng thể và hoàn thiện của ngôi nhà	Đánh giá theo mức độ: 1. Very poor 2. Poor 3. Fair 4. ... ... 10. Very Excellent
19	OverallCond	Int64	Đánh giá tình trạng chung của ngôi nhà	Đánh giá theo mức độ: 1. Very poor 2. Poor 3. Fair 4. ... ... 10. Very Excellent
20	YearBuilt	Int64	Ngày thi công ban đầu	
21	YearRemodAdd	Int64	Ngày tu sửa	Giống như ngày thi công nếu không tu sửa hoặc bổ sung
22	RoofStyle	Object	Loại mái	

23	RoofMatl	Object	Roof material	
24	Exterior1st	Object	Ngoại thất bao phủ bên ngoài căn hộ	
25	Exterior2nd	Object	Ngoại thất bao phủ bên ngoài căn hộ	Nếu có nhiều hơn một vật liệu
26	MasVnrType	Object	Loại gạch	
27	MasVnrArea	Float64	Diện tích gạch	Tính bằng feet vuông
28	ExterQual	Object	Đánh giá chất lượng vật liệu bên ngoài	
29	ExterCond	Object	Đánh giá tình trạng hiện tại của vật liệu bên ngoài	
30	Foundation	Object	Loại móng	
31	BsmtQual	Object	Đánh giá chiều cao của tầng hầm	Ví dụ: – Ex: Excellent (100+ inches) – Gd: Good (90 – 99 inches) – ...
32	BsmtCond	Object	Đánh giá tình trạng chung của tầng hầm	
33	BsmtExposure	Object	Đề cập đến lối đi hoặc tường vườn	
34	BsmtFinType1	Object	Đánh giá diện tích tầng hầm	Ví dụ: – GLQ: Good living quarters – ALQ: Average living quaters
35	BsmtFinSF1	Int64	Loại 1 feet vuông đã hoàn thành	
36	BsmtFinType2	Object	Đánh giá diện tích tầng hầm	Nếu có nhiều loai
37	BsmtFinSF2	Int64	Loại 1 feet vuông đã hoàn thành	
38	BsmtUnfSF	Int64	Feet vuông chưa hoàn thành của khu vực tầng hầm	
39	TotalBsmtSF	Int64	Tổng số feet vuông của diện tích tầng hầm	
40	Heating	Object	Loại sưởi ấm	
41	HeatingQC	Object	Chất lượng và điều kiện sưởi ấm	
42	CentralAir	Object	Điều hòa trung tâm	
43	Electrical	Object	Hệ thống điện	
44	1stFlrSF	Int64	Tầng 1 feet vuông	

45	2ndFlrSF	Int64	Tầng 2 feet vuông	
46	LowQualFinSF	Int64	Chất lượng thấp hoàn thành feet vuông	Tất cả các tầng
47	GrLivArea	Int64	Phần trên (mặt đất) khu vực sinh sống feet vuông	
48	BsmtFullBath	Int64	Tầng hầm, đầy đủ phòng tắm	Trường này được hiểu như phần sàn của phòng tắm
49	BsmtHalfBath	Int64	Tầng hầm, một nửa phòng tắm	Trường này được hiểu như phần sàn của phòng tắm
50	FullBath	Int64	Phòng tắm đầy đủ ở phần trên	Trường này được hiểu như phần che phía trên của phòng tắm
51	HalfBath	Int64	Phòng tắm một nữa phần trên	Trường này được hiểu như phần che phía trên của phòng tắm
52	Bedroom	Int64	Phòng ngủ	KHÔNG bao gồm phòng ngủ tầng hầm
53	Kitchen	Int64	Bếp	
54	KitchenQual	Object	Chất lượng bếp	
55	TotRmsAbvGrd	Int64	Tổng số phòng	Không bao gồm phòng tắm
56	Functional	Object	Chức năng gia đình	Giả sử điển hình trừ khi các khoản khấu trừ được bảo hành
57	Fireplaces	Int64	Số lò sưởi	
58	FireplaceQu	Object	Chất lượng lò sưởi	
59	GarageType	Object	Vị trí gara	
60	GarageYrBlt	Float64	Năm gara được xây dựng	
61	GarageFinish	Object	Hoàn thiện nội thất gara	Ví dụ: – Fin: Finished – RFn: Rough Finished – Unf: Unfinished – NA: No Garage
62	GarageCars	Int64	Kích thước gara	Theo công suất xe
63	GarageArea	Int64	Kích thước gara	Theo feet vuông
64	GarageQual	Object	Chất lượng gara	
65	GarageCond	Object	Điều kiện gara	

66	PavedDrive	Object	Đường lái xe	
67	WoodDeckSF	Int64	Diện tích sàn gỗ trong feet vuông	
68	OpenPorchSF	Int64	Khu vực hiên mở bằng feet vuông	
69	EnclosedPorch	Int64	Khu vực hiên nhà kèm theo trong feet vuông	
70	3SsnPorch	Int64	Diện tích hiên ba mùa bằng feet vuông	
71	ScreenPorch	Int64	Màn hình diện tích hiên bằng feet vuông	
72	PoolArea	Int64	Khu vực bể bơi tính bằng feet vuông	
73	PoolQC	Object	Chất lượng bể bơi	
74	Fence	Object	Chất lượng hàng rào	
75	MiscFeature	Object	Tính năng khác không được đề cập trong các danh mục khác	
76	MiscVal	Int64	Giá trị của tính năng khác	
77	MoSold	Int64	Tháng bán (MM)	Tháng
78	YrSold	Int64	Năm bán	Năm
79	SaleType	Object	Loại hình bán hàng	Ví dụ: – WD: Warranty Deed – Conventional – CWD: Warranty Deed – Cash – VWD: Warranty Deed – VA Loan – New: Home just constructed and sold – ...
80	SaleCondition	Object	Điều kiện bán hàng	
81	SalePrice	Int64	Giá nhà	

Bảng 2. Bảng phân tích tập dữ liệu được dùng trong thuật toán

4.4.2. Bắt đầu huấn luyện

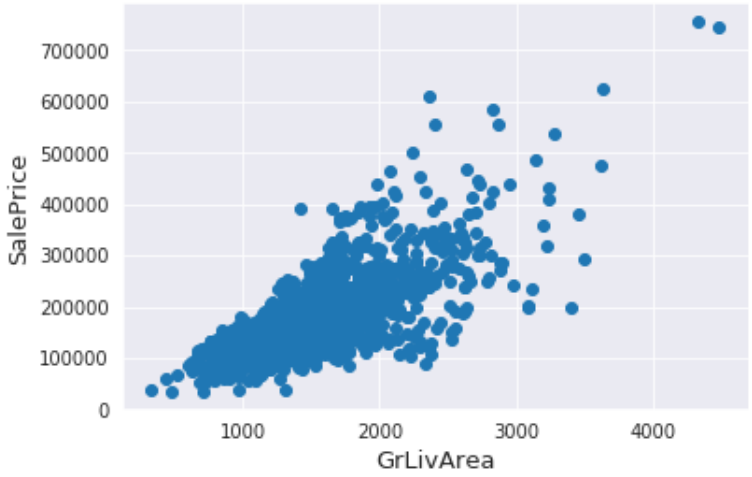
- Kích thước dữ liệu trước khi bỏ thuộc tính Id là: (1460, 81)

- Kích thước dữ liệu sau khi bỏ thuộc tính Id là: (1460, 80)

Hình 8. Hình biểu diễn tập dữ liệu trước khi loại bỏ trường Id

Ngoài ra trong đồ thị xuất hiện 2 điểm góc phải không hợp lý, ta sẽ dùng hàm xử lý để loại bỏ 2 điểm khả nghi này.

Ta được kết qua như sau:

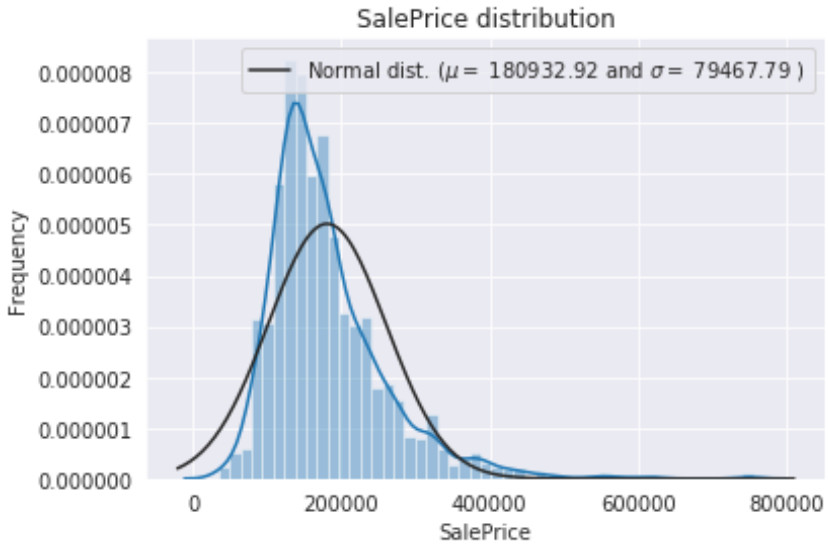


Hình 9. Hình biểu diễn tập dữ liệu sau khi loại bỏ trường Id và 2 điểm khả nghi

- Dự đoán khả năng mua của căn nhà trong tầm giá

  Ở đây ta dùng phương pháp K-nearest neighbor để giải quyết vấn đề này.

Giả sử ta thu thập được giá dự đoán của căn nhà từ bước trên, sau đó ta cần xác định mức giá này có phù hợp hay không, xác suất người mua sẽ mua là bao nhiêu.



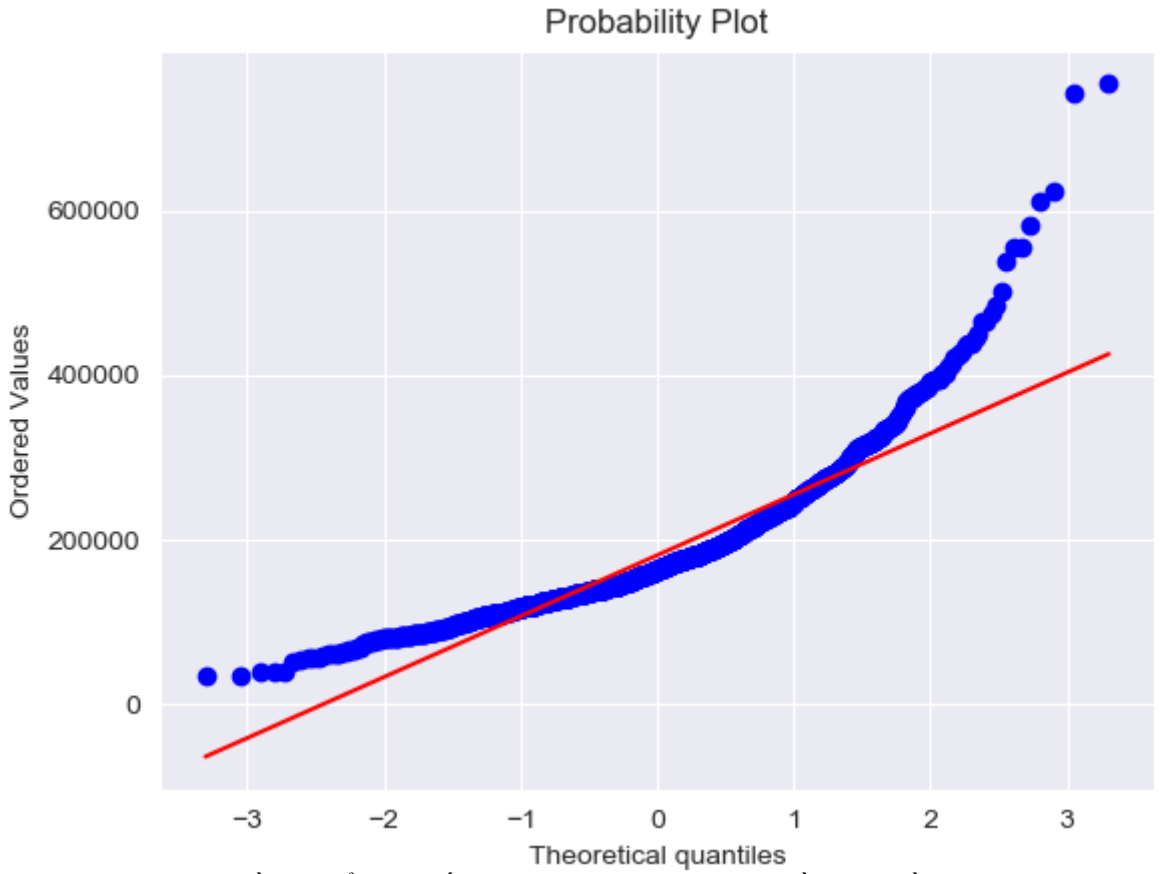
Hình 10. Hình thể hiện xác xuất được mua của căn nhà trong tầm giá tương ứng

Như trong đồ thị ta nhìn thấy 2 đường màu xanh và màu đen:

- Đường màu đen: thể hiện sự đánh giá chuẩn
- Đường màu xanh: biểu diễn kết quả của dự đoán, ở đây ta thấy sai số quá nhiều, và tất nhiên cần huấn luyện theo phương pháp khác hoặc trường dữ liệu khác.

- Tương tự, ta sẽ thể hiện qua giá trị căn nhà cùng với số lượng đã được mua:

Ở đây ta dùng thuật toán Linear regression để giải quyết bài toàn này:



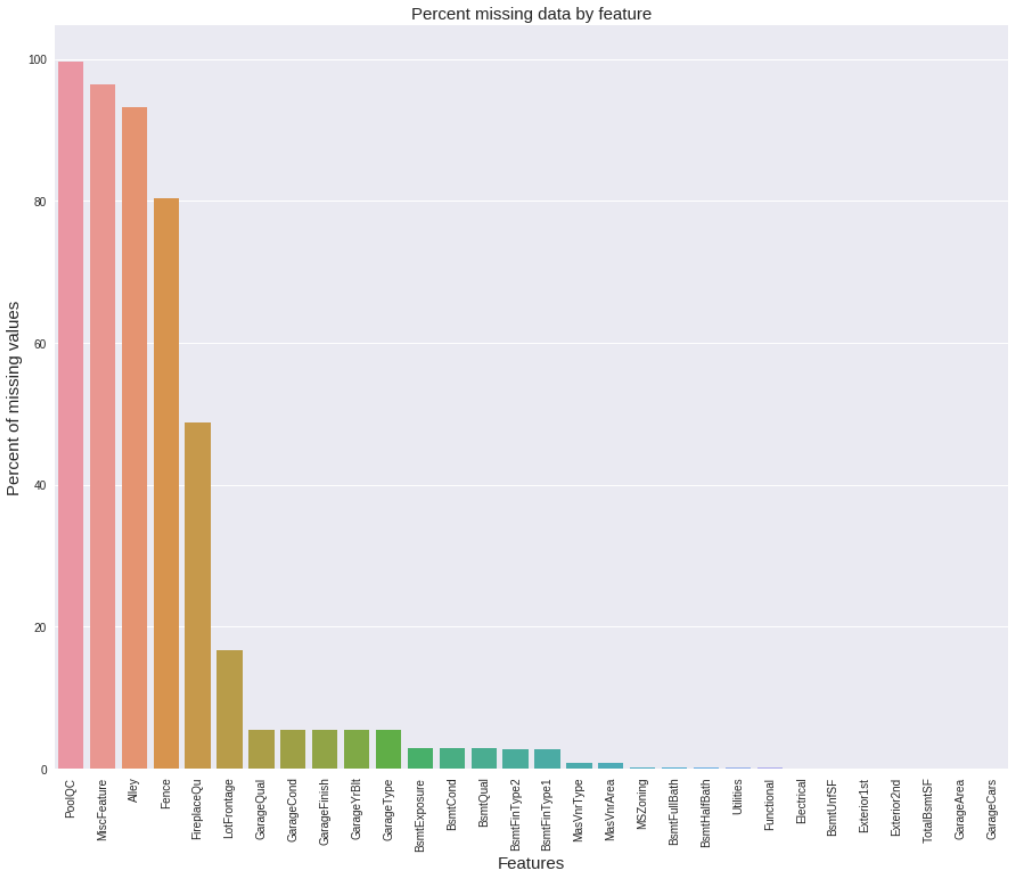
Hình 11. Hình đồ thị thể hiện số căn nhà được mua trong tầm giá bằng thuật toán Linear regression

Với hình ảnh được thể hiện qua độ thị, ta thấy:

- Đường màu đỏ: phân phối bình thường đối với thị trưởng thực tế về số căn nhà được mua trong tầm giá.
- Đường màu xanh: thực ra đây là những điểm dự đoán thông qua giải thuật Linear regression. Ta có thể thấy những điểm này khá gần nhau nên đã tạo thành một đường nối liền.

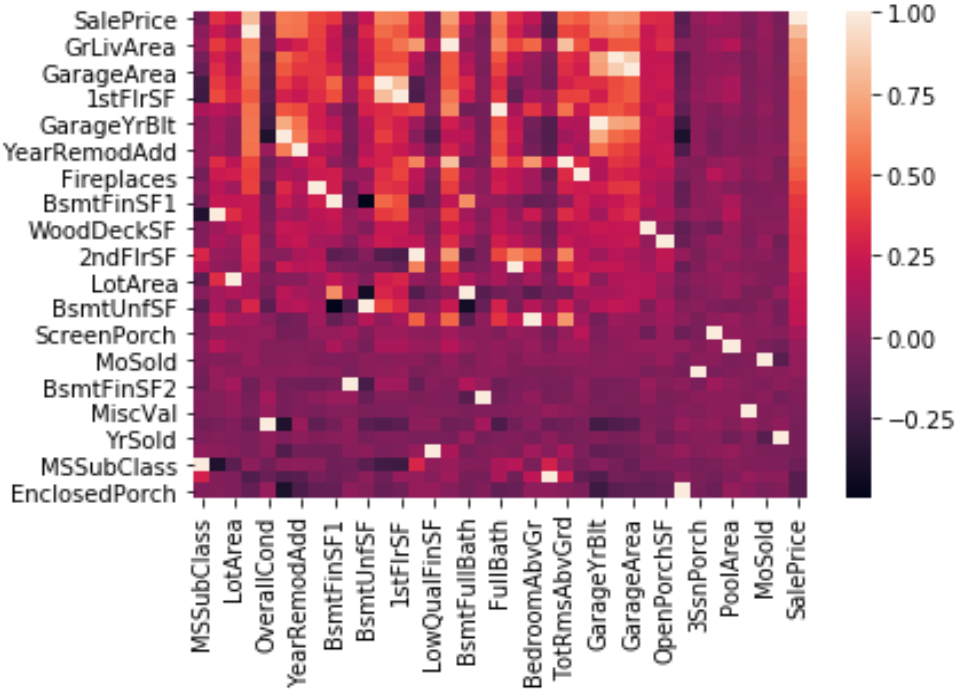
Tuy nhiên, ta vẫn thấy được sự chênh lệch giữa đường phân phối chuẩn và dự đoán.

- Trong quá trình huấn luyện, tập dữ liệu đã bị bó sót một số trường, và tỉ lệ của chúng như sau:



Hình 12. Hình minh họa tỉ lệ bị bỏ sót của một số trường dữ liệu

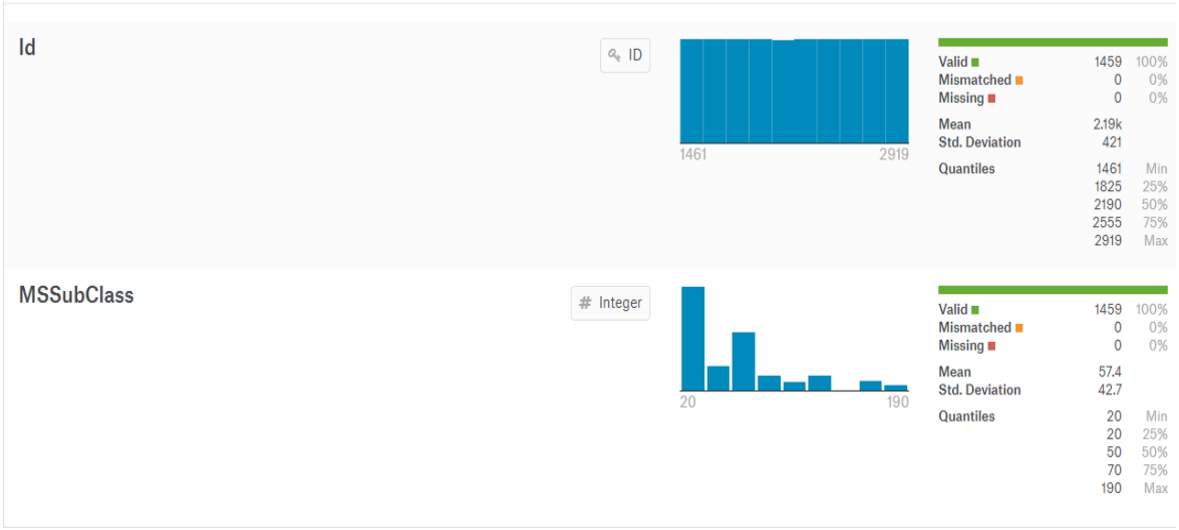
Ngoài ra ta hãy quan sát sự tương quan giữa các trường dữ liệu khác với Saleprice



Hình 13. Hình thể hiện sự tương quan giữa các trường dữ liệu khác với SalePrice

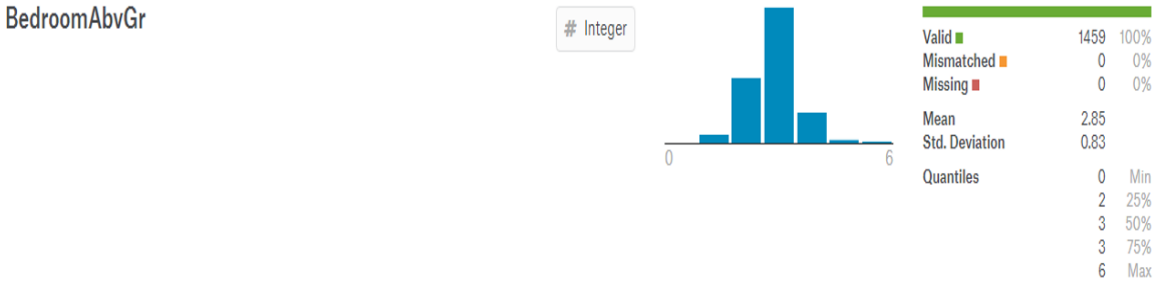
Sau đó ta sẽ làm đầy các trường bị bỏ sót bằng cách thêm giá trị "None" vào cho chúng. Vì thế, khi trải qua quá trình tính toán, những trường này sẽ không ảnh hưởng đến kết quả của bài toán vì lý do thiếu sót dữ liệu.

- Chúng ta tiếp tục quá trình huấn luyện, nhưng trước đó ta có một số liên quan đến việc thống kê các trường của tập dữ liệu như sau:
  - Theo Id, MSSubClass



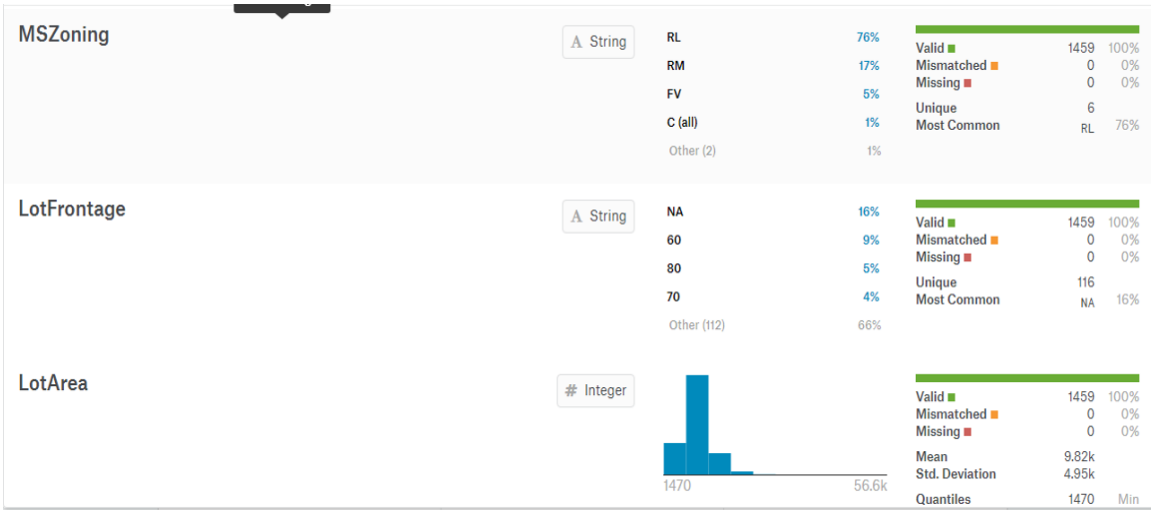
Hình 14. Hình thống kê, đánh giá tỉ lệ bị bỏ sót cụ thể tập dữ liệu theo Id, MSSubClass

  - Theo BedrooAbvGr



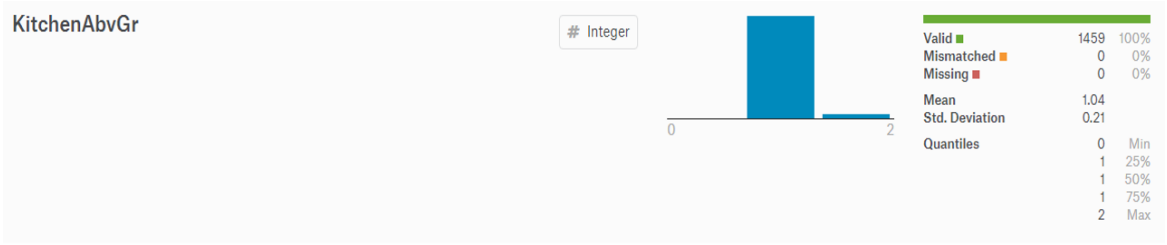
Hình 15. Hình thống kê, đánh giá tỉ lệ bị bỏ sót cụ thể tập dữ liệu theo BedrooAbvGr

- Theo MSZoning, LotFrontage, LotArea



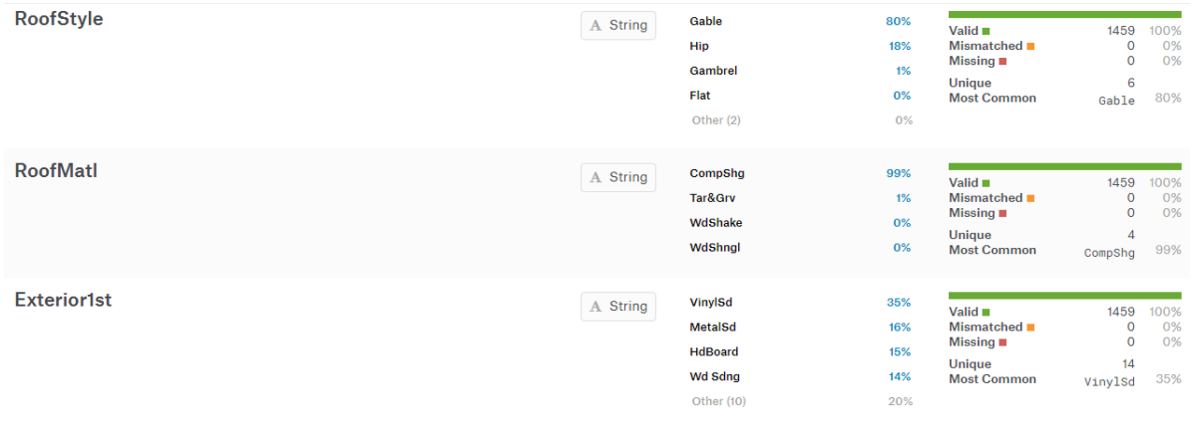
Hình 16. Hình thống kê, đánh giá tỉ lệ bị bỏ sót cụ thể tập dữ liệu theo MSZoning, LotFrontage, LotArea

- Theo KitchenAbvGr



Hình 17. Hình thống kê, đánh giá tỉ lệ bị bỏ sót cụ thể tập dữ liệu theo KitchenAbvGr

- Theo RoofStyle, RoofMatl, Exterior1st



Hình 18. Hình thống kê, đánh giá tỉ lệ bị bỏ sót cụ thể tập dữ liệu theo RoofStyle, RoofMatl, Exterior1st

Tiếp theo ta cần tính tổng diện tích cho tất cả các thành phần trong căn hộ, tất nhiên ta phải thêm một trường vào tập dữ liệu.

TotalSF = diện tích tầng hầm + diện tích tầng 1 + diện tích tầng 2

Và khi thực hiện bước này sẽ phát sinh ra sự chênh lệch so với trước đó, và đây là một số trường xảy ra chênh lệch cao nhất

	Skew
MiscVal	21.940
PoolArea	17.689
LotArea	13.109
LowQualFinSF	12.085
3SsnPorch	11.372
LandSlope	4.973
KitchenAbvGr	4.301
BsmtFinSF2	4.145
EnclosedPorch	4.002
ScreenPorch	3.945

Hình 19. Hình thể hiện sự chênh lệnh độ chính xác của các trường

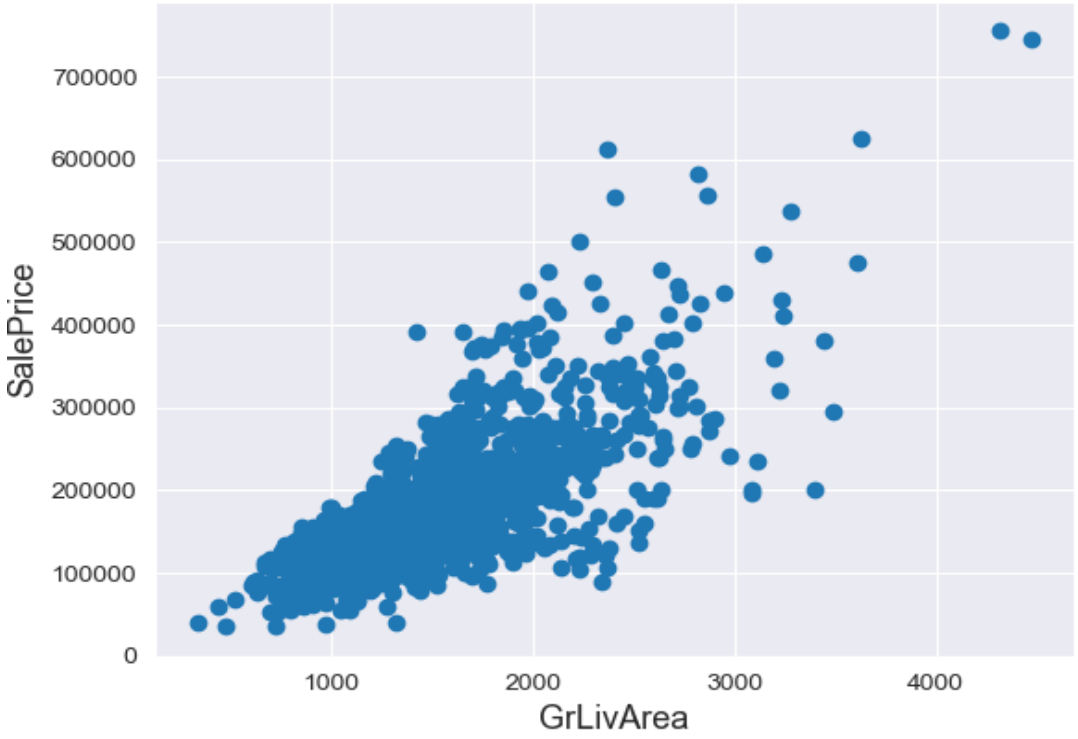
Ta nhận thấy độ chênh lệch của các trường trong tập dữ liệu là quá cao nên ta sẽ tiến hành xử lý đối với những trường hợp có độ lệch lớn hơn 0.75

Từ đây ta thu được một thuật toán tương đối hoàn chỉnh và tiến hành kiểm tra xem độ chính xác và mức độ hoạt động của thuật toán.

4.5. Kiểm tra mức độ chính xác của thuật toán

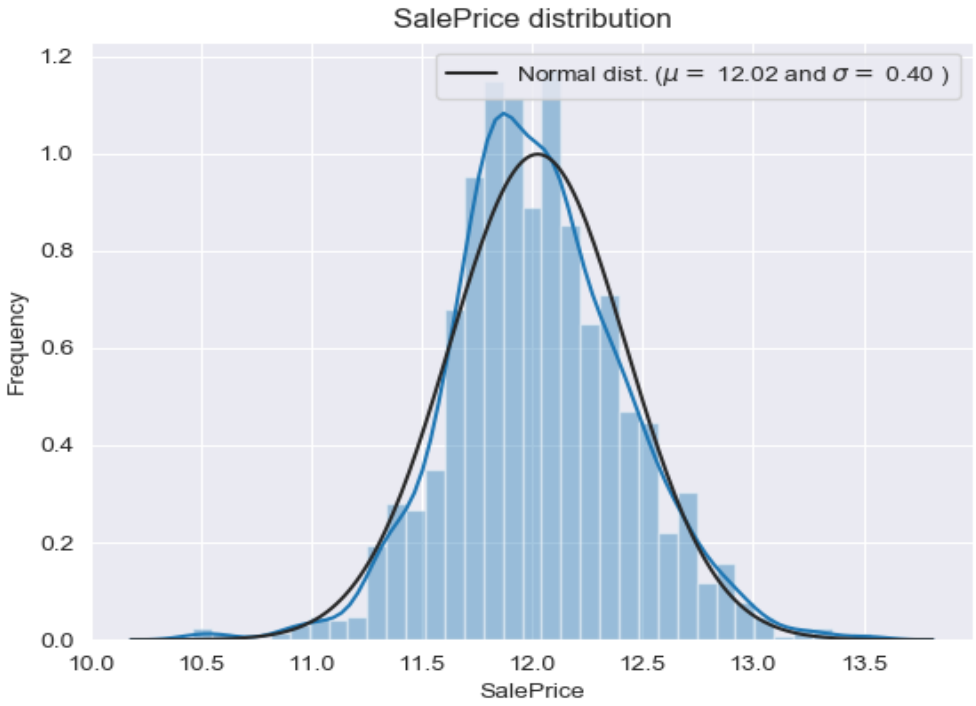
Sau quá trình huấn luyện ta sẽ kiểm tra mức độ chính xác của thuật toán qua các bước tương tự như phần huấn luyện như sẽ được cải tiến hơn.

- Kích thước dữ liệu sau khi bỏ thuộc tính Id, cũng như sau khi loại bỏ hai điểm khả nghi góc phải



Hình 20. Hình đồ thị Hình biểu diễn tập dữ liệu sau khi loại bỏ trường Id và 2 điểm khả nghi

- Dự đoán khả năng mua của căn nhà trong tầm giá

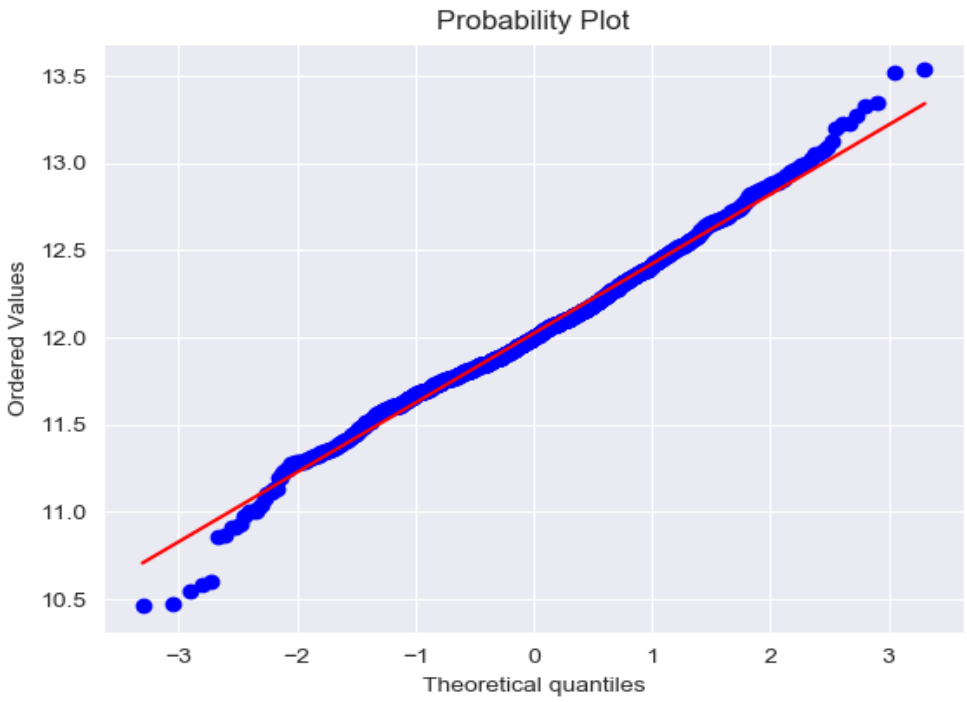


Hình 21. Hình thể hiện xác xuất được mua của căn nhà trong tầm giá tương ứng

Như đã giải thích trong phần huấn luyện, nếu đường màu đen là sự phân bố chuẩn, đường màu xanh là đường mà chúng ta dự đoán được từ tập dữ liệu thì có thể thấy được sự chính xác của thuật toán sau khi cải tiến là rất lớn.

Ta tiếp tục quá trình kiểm tra.

- Tương tự, ta sẽ thể hiện qua giá trị căn nhà cùng với số lượng đã được mua:



Hình 22. Hình đồ thị thể hiện số căn nhà được mua trong tầm giá bằng thuật toán Linear regression

Như ta thấy trong hình, đường phân phối chuẩn (màu đỏ) và các điểm dự đoán của chúng ta khá chính xác, độ chênh lệch không lớn. Cho thấy sự cải tiến trong thuật toán là chính xác.

Vì ta đã thực hiện phần thống kê tỉ lệ việc bị bỏ sót của các trường trong tập dữ liệu cũng như những bước làm đầy các trường bị bỏ sót trong phần huấn luyện nên ta không cần thực hiện lại trong phần kiểm tra mà ta sẽ đi đến phần tình toán quan trọng nhất của vấn đề - dự đoán giá nhà đất

Ta có thể thấy ở 2 bước trước đó mức độ chính xác của thuật toán là rất cao. Sau quá trình huấn luyện ta thu được tập kết quả giá nhà đất như sau:

,Id,SalePrice
0,1461,119790.9704891311
1,1462,159687.10411958414
2,1463,186993.4476352987
3,1464,196781.68329748328
4,1465,193150.18684102714
5,1466,172373.48120015772
6,1467,179188.92677630766
7,1468,162599.33044123874
8,1469,182859.75164827192
9,1470,122978.49040302366
10,1471,197234.16894754724
11,1472,95197.511293897
12,1473,94509.43973650345
13,1474,146413.80341463102
14,1475,112523.45279767194
15,1476,382672.4325992223
16,1477,248007.53830122255
17,1478,282844.35230961913
18,1479,284892.19684464915
19,1480,497593.81076979724
20,1481,318172.87992459506
21,1482,206581.67289241048
22,1483,178753.9527832576
23,1484,165833.71214966878

Hình 23. Hình minh họa kết quả dự đoán giá nhà đất

KẾT LUẬN

1. Kết quả đạt được

Qua quá trình tìm hiểu và thực hiện đề tài, chúng em thu được một số kết quả sau:

- Hiểu về thuật toán và nguyên lý hoạt động của thuật toán Linear regression và K-nearest neighbor.
- Ứng dụng thuật toán vào bài toán dự đoán giá nhà đất một cách phù hợp
- Hình thành phương pháp/ công cụ dự đoán giá nhà đất cho người dùng.

2. Đánh giá đề tài

- Ưu điểm
  - Đề tài cơ bản giải quyết được những mục tiêu ban đầu đề ra: phương pháp dự đoán giá nhà đất chính xác đáng tin cậy.
  - Có trải qua quá trình huấn luyện trên nhiều trường của tập dữ liệu nên độ chính xác khá cao.
  - Có thể dự đoán cùng lúc nhiểu căn hộ.
  - Tập dữ liệu dùng cho thuật toán trong đề tài khá lớn vì tập hợp gần như đầy đủ dữ liệu cho việc dự đoán, nhưng lại linh hoạt trong quá trình tính toán. Nghĩa là nếu trường dữ liệu nào không được làm đầy cũng không ảnh hưởng đến kết quả dự đoán sau cùng.
- Nhược điểm:
  - Do việc tập dữ liệu yêu cầu đầu là khá nhiều trường nên người dùng phải tập hợp tất cả lại thành file với đuôi mở rộng là .csv khi cho vào thuật toánì

3. Hướng phát triển

Trong những năm gần đây, thị trường bất động sản ngày càng sôi nổi hơn, phát triển hơn và có tiềm năng hơn, nên việc một phương pháp, công cụ giúp người dùng dự đoán giá nhà đất có độ chính xác cao.

Nên nhóm chúng em đề xuất một số hướng phát triển cho đề tài:

- Phát triển thành công cụ, chương trình, hay ứng dụng cụ thể mà người dùng có thể dễ dàng sử dụng
- Đầu vào tập dữ liệu cần linh hoạt hơn
- Khi xuất ra kết quả dự đoán cần trực quan hơn và có khả năng so sánh, giúp người dùng dễ quyết định hơn trong việc định giá căn hộ định bán hoặc định mua.

TÀI LIỆU THAM KHẢO

[1]. https://machinelearningcoban.com/2016/12/28/linearregression/
[2]. https://codelungtung.com/2018/08/22/image-classification-k-nearest-neighbor-classification/
[3]. http://nhasaipho.com/ty-le-tieu-thu-can-ho-tai-tp-hcm-gia-tang-trong-quy-iv2017/
[4]. https://www.bayareamarketreports.com/trend/3-recessions-2-bubbles-and-a-baby
[5]. https://www.kaggle.com/josh24990/simple-stacking-approach-top-12-score/data
[6]. https://www.kaggle.com/imaxplus/code-group6-ml?fbclid=IwAR2iZlSuJ3sDQazSIgCazsTygJ0pWT1JcGyz0zTcQLfpWzdleSvBZWI2cqo