**ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**--------\*\*\*\*\*--------**

****

BÁO CÁO BÀI TẬP LỚN

**MÔN HỌC: NHẬP MÔN KHOA HỌC DỮ LIỆU**

***Đề tài : Dự đoán giá nhà đất***

GVHD: PGS.TS Thân Quang Khoát

Mã lớp: 128735 - Nhóm 16

Nguyễn Văn Sáng 20194153

Nguyễn Hữu Hải 20194039

Nguyễn Minh Hiển 20194045

***Hà Nội, 1/2022***

**MỤC LỤC**

[1. Giới thiệu đề tài 3](#_Toc92742159)

[2. Crawl data và tiền xử lý dữ liệu 4](#_Toc92742160)

[2.1. Crawl data từ website Chotot.com 5](#_Toc92742161)

[2.2. Tiền xử lý dữ liệu 6](#_Toc92742162)

[3. Khám phá dữ liệu 10](#_Toc92742163)

[4. Áp dụng giải thuật học máy dự đoán giá nhà 17](#_Toc92742164)

[4.1. Giải thuật hồi quy tuyến tính 17](#_Toc92742165)

[4.2. Giải thuật k láng giềng gần nhất 18](#_Toc92742166)

[4.3. Phương pháp đánh giá giải thuật 22](#_Toc92742167)

[4.3.1. Root Mean Square Error 22](#_Toc92742168)

[4.3.2. R2 Score 22](#_Toc92742169)

[4.4. Thực nghiệm và đánh giá 23](#_Toc92742170)

[4.4.1. Thực nghiệm 23](#_Toc92742171)

[4.4.2. Đánh giá kết quả 24](#_Toc92742172)

[5. Xây dựng chương trình demo đơn giản 24](#_Toc92742173)

[6. Kết luận 26](#_Toc92742174)

[6.1. Kết luận chung 26](#_Toc92742175)

[6.2. Hướng phát triển tiếp theo 27](#_Toc92742176)

[7. Phân chia công việc 27](#_Toc92742177)

[8. Tài liệu tham khảo 27](#_Toc92742178)

1. Giới thiệu đề tài

Trong khoảng thời gian gần đây, bất động sản là đề tài thu hút nhiều sự quan tâm của mọi người. Bất động sản trong năm 2022 tiếp tục tăng mạnh, xuất hiện tình trạng bong bóng giá ảo, khiến những người có nhu cầu mua nhà gặp nhiều khó khăn trong việc chọn một ngôi nhà phù hợp với túi tiền của mình.

Đứng trước những vấn đề đó, nhóm em lựa chọn đề tài Dự đoán bất động sản, nhằm dự đoán giá của những ngôi nhà, nhằm đưa ra một con số cụ thể mang tính tham khảo. Bằng cách lấy dữ liệu từ trang website Chotot.com, nhóm em tiến hành huấn luyện mô hình, từ đó đưa ra những kết quả dự đoán với đầu vào là những thuộc tính cơ bản như Diện tích đất, Loại hình nhà ở, Địa chỉ, Số phòng ngủ,... và đầu ra là giá nhà.

Trước mắt, nhóm em sẽ triển khai project với mức độ của môn học. Còn trong tương lai, nếu có điều kiện, nhóm em sẽ cố gắng triển khai một mô hình có thể ứng dụng thực tế.

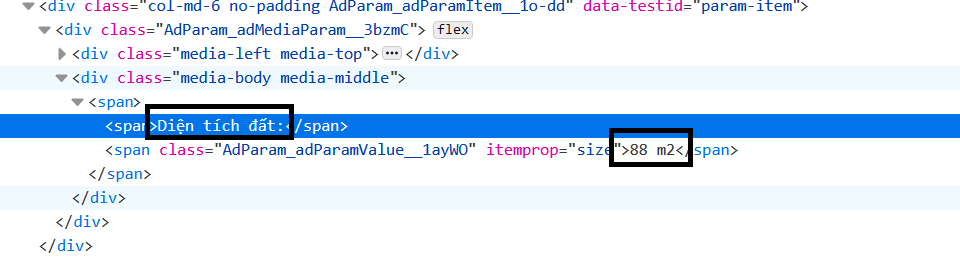
Tuy nhiên vì còn là sinh viên nên kiến thức thực tế còn hạn hẹp, do đó trong quá trình thực hiện đề tài môn học, nhóm em khó có thể tránh được những thiếu xót. Nhóm em rất mong được sự góp ý và chỉ dẫn của thầy cô và bạn bè để có thêm nhiều kiến thức chuyên môn. Nhóm em xin chân thành cảm ơn thầy Thân Quang Khoát và anh Trần Tùng Lâm với sự hướng dẫn và chỉ bảo nhiệt tình.

1. Crawl data và tiền xử lý dữ liệu

Dữ liệu có thể thu nhập từ các bên môi giới, người bán, người mua bất động sản, từ các website cung cấp dịch vụ mua bán bất động sản như: chotot.com, batdongsan.com.vn, mystay.vn…Nhóm hiện đang thu nhập dữ liệu từ trang chotot.vn

* 1. Crawl data từ website Chotot.com

Nhóm em sử dụng thư viện BeautifulSoup để crawl data từ trang web bằng cách bóc tách các thẻ html.



Kết quả, nhóm em đã thu thập được tổng cộng 129835 bản ghi, với 17 thuộc tính bao gồm Diện tích đất, Giá/m2, Tỉnh thành, Quận huyện, Số phòng ngủ, Số phòng vệ sinh, Sồ tầng, Giấy tờ pháp lý, Chiều ngang, Chiều dọc, Diện tích sử dụng, Kinh độ, Vĩ độ, Tình trạng nội thất, Loại hình nhà ở, Đặc điểm nhà đất, Hướng cửa chính.

Ảnh có chứa bàn

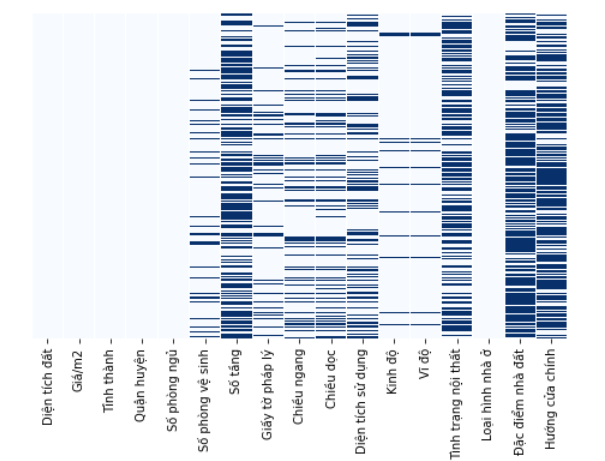
Mô tả được tạo tự động

Ảnh có chứa bàn

Mô tả được tạo tự động

* 1. Tiền xử lý dữ liệu

Bằng những quan sát ban đầu, nhóm em nhận ra những thuộc tính có số lượng giá trị NaN nhiều là Số tầng, Tình trạng nội thất, Đặc điểm nhà đất, Hướng cửa chính. Do đó, nhóm em quyết định loại bỏ những thuộc tính này khỏi tập dữ liệu.



Ban đầu nhóm em quyết định loại bỏ hẳn hai thuộc tính Quận huyện, Tỉnh thành vì đã có thuộc tính Kinh độ, Vĩ độ. Tuy nhiên, sau này nhóm em có cải tiến thêm, đó là thêm thuộc tính Mật độ dân số.

Bằng dữ liệu mật độ dân số dựa trên quận huyện của cả nước, nhóm em quyết định thay thế hai thuộc tính Quận huyện, Tỉnh thành bằng một thuộc tính mới là Mật độ dân số.

Nguồn thông tin: <https://vi.wikipedia.org/wiki/Danh_s%C3%A1ch_%C4%91%C6%A1n_v%E1%BB%8B_h%C3%A0nh_ch%C3%ADnh_c%E1%BA%A5p_huy%E1%BB%87n_c%E1%BB%A7a_Vi%E1%BB%87t_Nam>

Ảnh có chứa bàn

Mô tả được tạo tự động

Ảnh có chứa bàn

Mô tả được tạo tự động

Tiếp theo, nhóm em quyết định loại bỏ các bản ghi có những thuộc tính quan trọng bị NaN, đó là Giá/m2, Kinh độ, Loại hình nhà ở, Chiều ngang, Chiều dọc, Mật độ dân số và định dạng chuẩn lại dữ liệu.

Với thuộc tính Loại hình nhà ở, có tổng cộng 4 loại, Nhà ngõ, Nhà phố liền kề, Nhà mặt phố, Nhà biệt thự. Ban đầu nhóm em tách thuộc tính này ra làm 4 vector [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], [0, 0, 0, 1].

Ảnh có chứa bàn

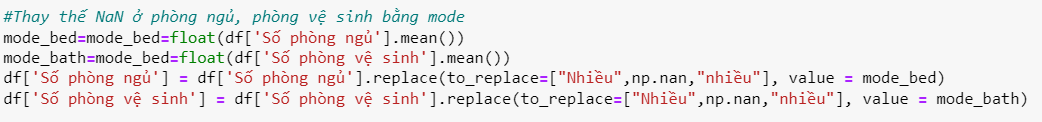
Mô tả được tạo tự động

Sau này nhóm em thử cải tiến dựa trên suy luận, tìm hiểu cũng như có phần “cảm tính” bằng cách gán giá trị cho từng loại hình nhờ ở, cụ thể, Nhà ngõ, Nhà phố liền kề, Nhà mặt phố, Nhà biệt thự lần lượt là 1, 2, 3, 4. Khi huấn luyện, cách thay thế thứ hai cho kết quả tốt hơn 0.4% nên nhóm em quyết định sử dụng cách thay thế thứ hai làm model chính.

Ảnh có chứa văn bản

Mô tả được tạo tự động

Cuối cùng, nhóm em thay thế những giá trị bị NaN của hai thuộc tính Số phòng ngủ, Số phòng vệ sinh bằng giá trị mean. Nhóm em có lần lượt thử nghiệm filling values với các cách như dùng KNN, Linear Regression, hoặc dùng giá trị mode nhưng đều không cho kết quả tốt như khi sử dụng giá trị mean.



Kết quả sau khi tiền xử lý dữ liệu, dataset còn lại 90817 bản ghi với 10 thuộc tính.

Ảnh có chứa văn bản, bàn

Mô tả được tạo tự động

Thống kê các bước nhóm đã thực hiện để tiền xử lý và làm sạch dữ liệu:

B1: Thêm trường dữ liệu “Giá” từ 2 trường “Giá/m2” và “Diện tích đất”.

B2: Thêm trường “Mật độ dân số” từ 2 trường “Quận huyện” và “Tỉnh thành”.

B3: Xóa bỏ các trường dữ liệu không cần thiết.

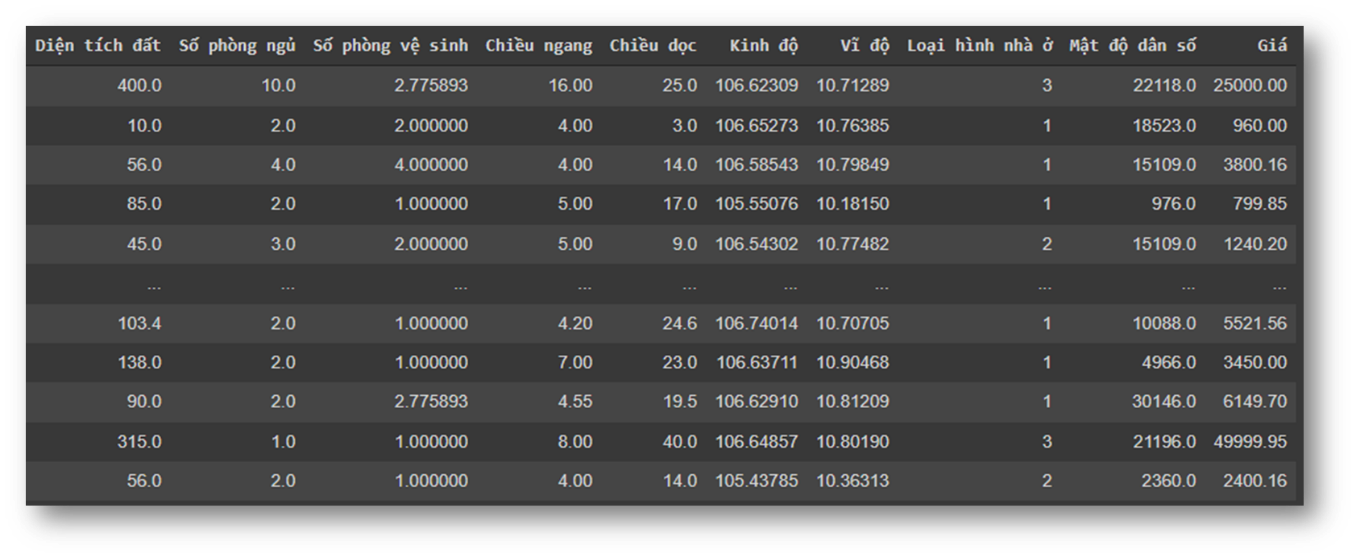
B4: Chuyển đổi trường “Loại hình nhà ở” thành số.

B5: Loại bỏ các bảng ghi có ít dữ liệu và có giá trị trùng nhau.

B6: Xóa bỏ các outlier của các trường dữ liệu.

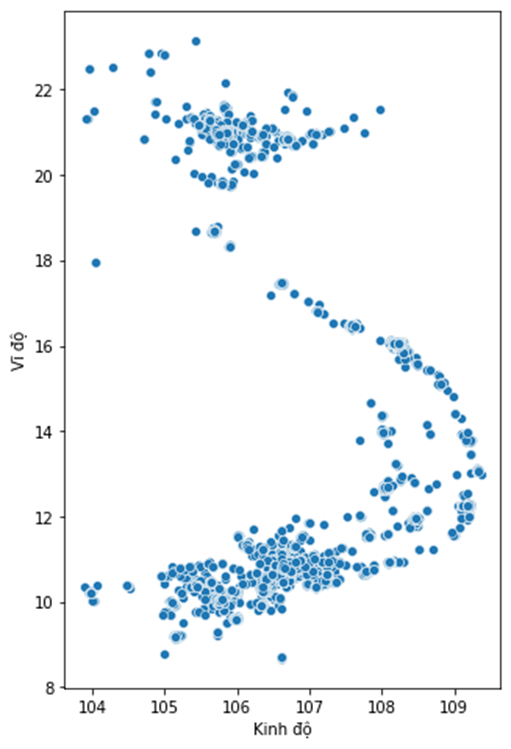
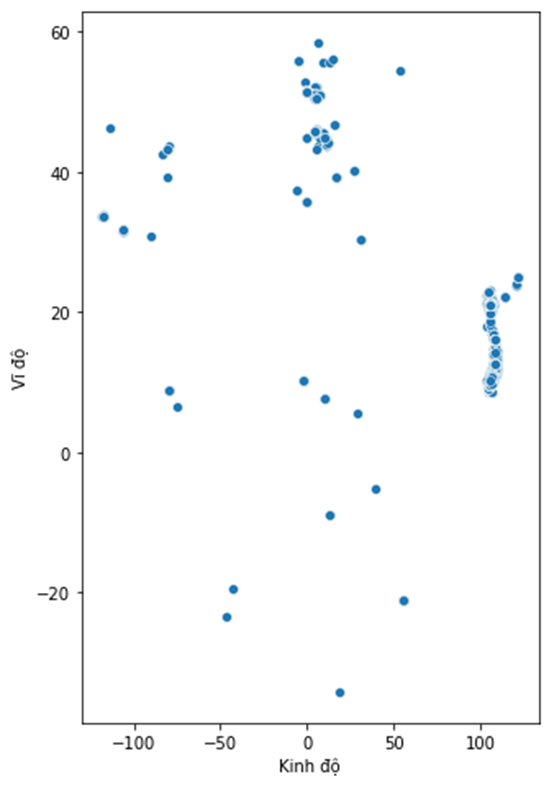
B7: Thay thế các giá trị bị khuyết của hai trường “Số phòng ngủ”, “Số phòng vệ sinh” bằng mean.

Dữ liệu sau khi tiền xử lý và làm sạch:

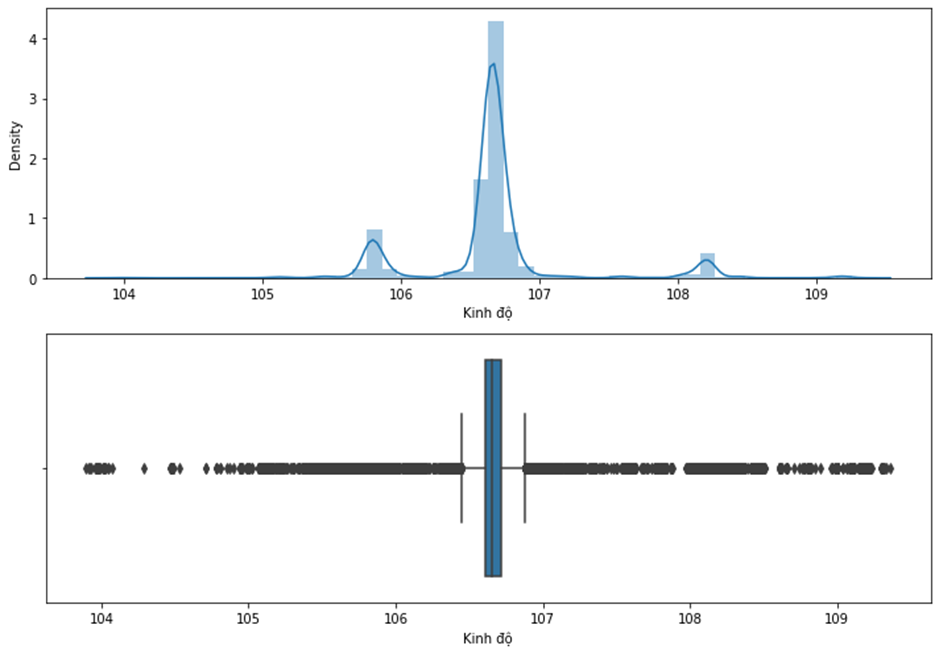


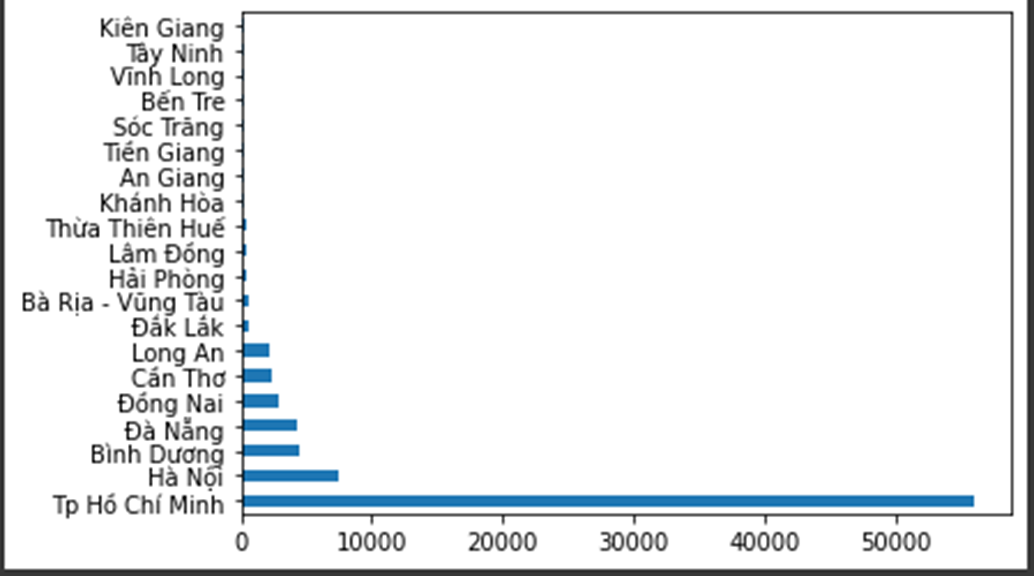
Chúng em sẽ giải thích dõ hơn các bước làm sạch ở phần khám phá dữ liệu.

1. Khám phá dữ liệu

Chúng em sẽ biểu diễn dữ liệu để có thể nhận xét được các thuộc tính đặc trưng.

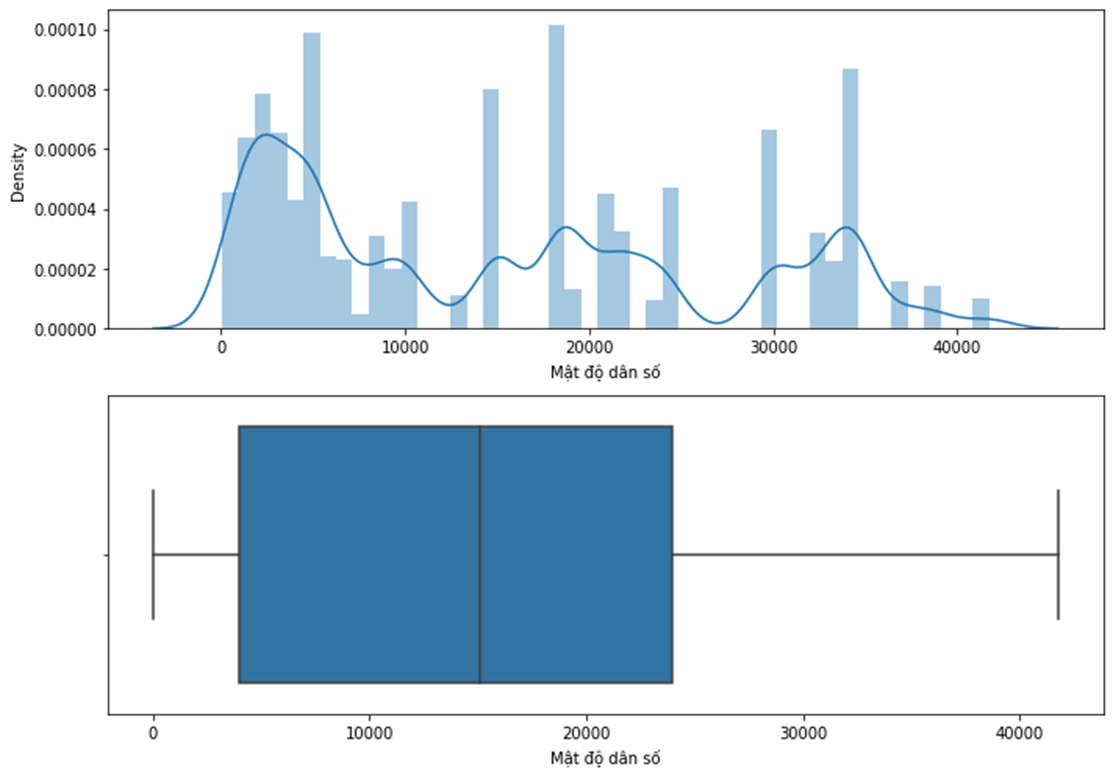
Trường kinh độ và vĩ độ có nhiều điểm vượt ra khỏi Việt Nam. Nhóm em thực hiện loại bỏ các điểm đó và tiến hành visual thì cho ra kết quả hình dạng biểu đồ của Việt Nam.

 *Biểu đồ phân phối Kinh độ và Vĩ độ*

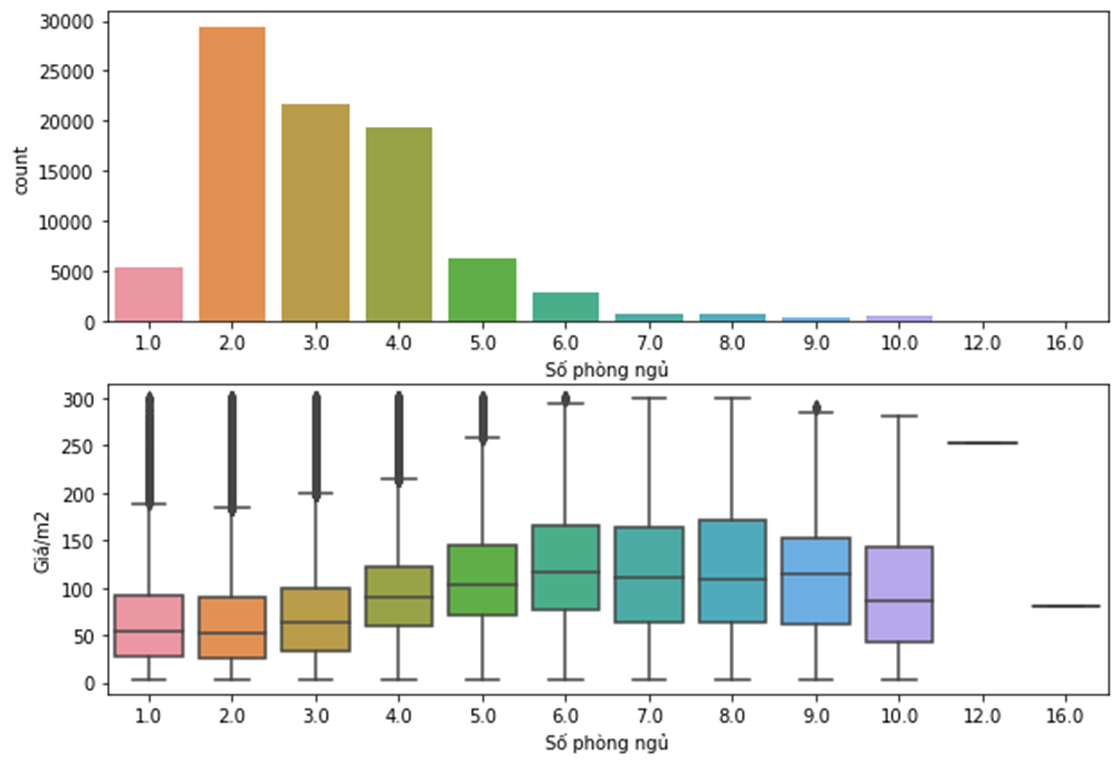
****

*Biểu đồ 20 tỉnh thành có số lượng bản ghi lớn nhất*

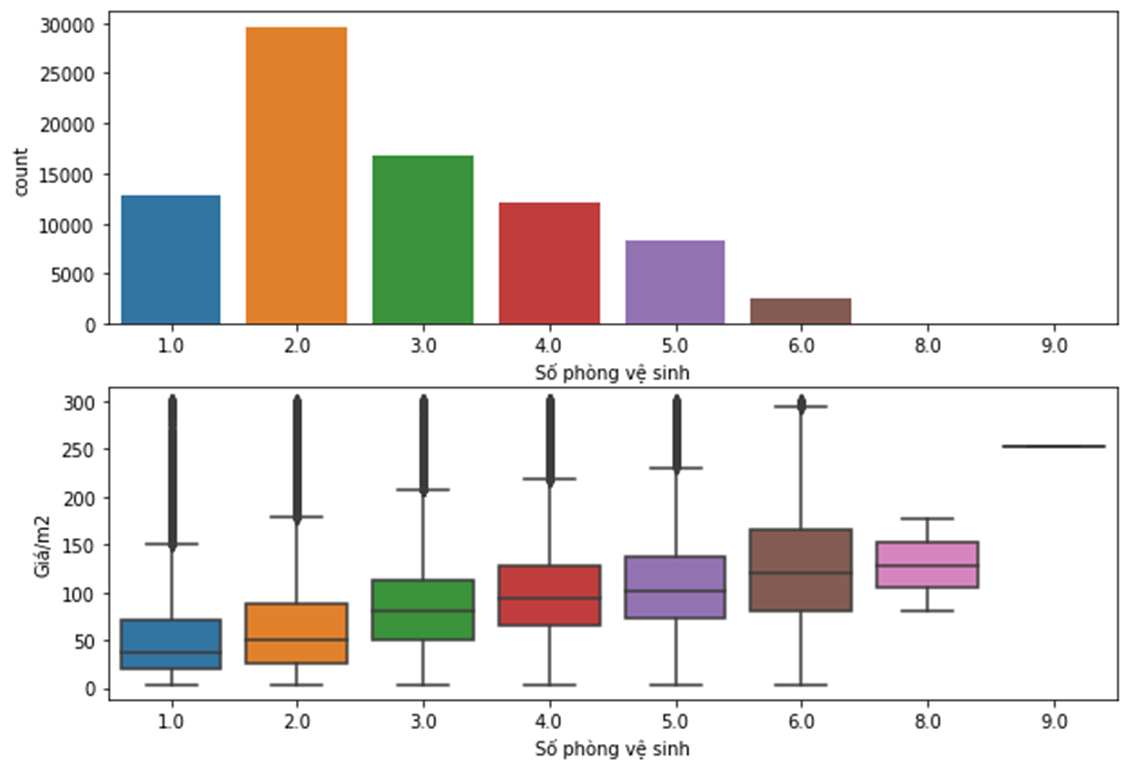
Nhận xét: Nhận thấy thành phố Hồ Chi Minh có số lượng bản ghi khá lớn so với các tỉnh còn lại. Do thành phố Hồ Chí Minh là thành phố trọng điểm Vì vậy mọi người sẽ tập trung ở thành phố này để sống và làm việc. Các tỉnh thành phố tiếp theo có thể nhìn được ở biểu đồ trên là những tỉnh thành phố cũng nổi tiếng, có thành phố trực thuộc trung ương, vì vậy cơ hội việc làm ở đây cũng rất cao, nên vấn đề nhà đất cũng rất “nóng”.



*Biều đồ mật độ dân số và biến động mật độ dân số*

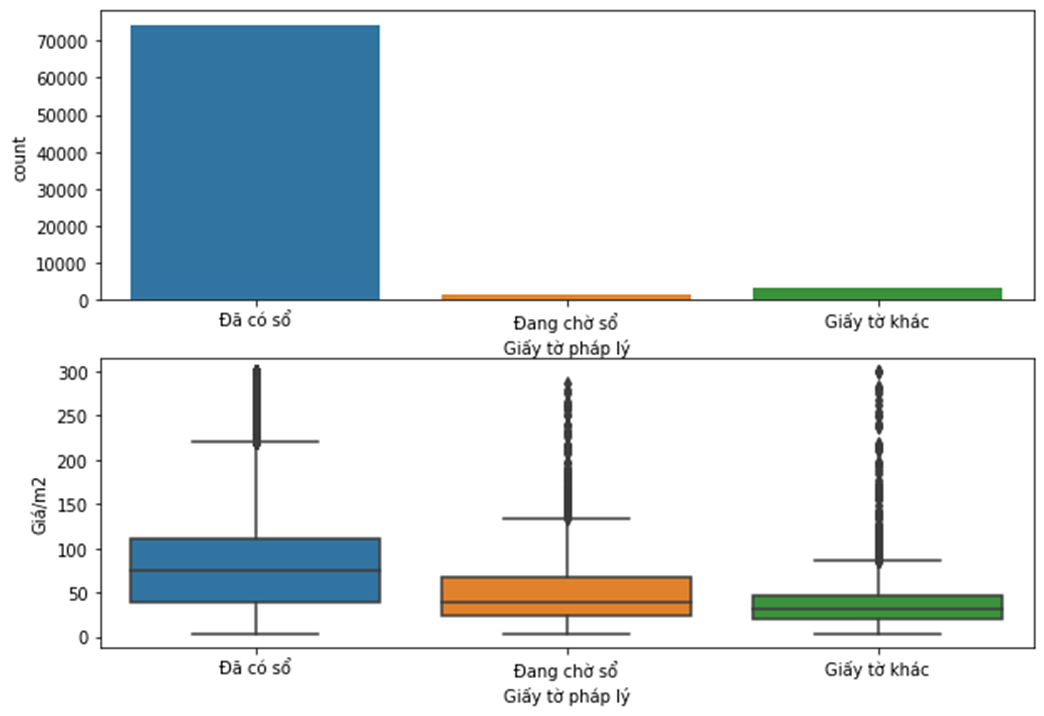


*Biểu đồ thể hiện số lượng phòng ngủ và sự tương quan số lượng phòng ngủ so với Giá/m2.*



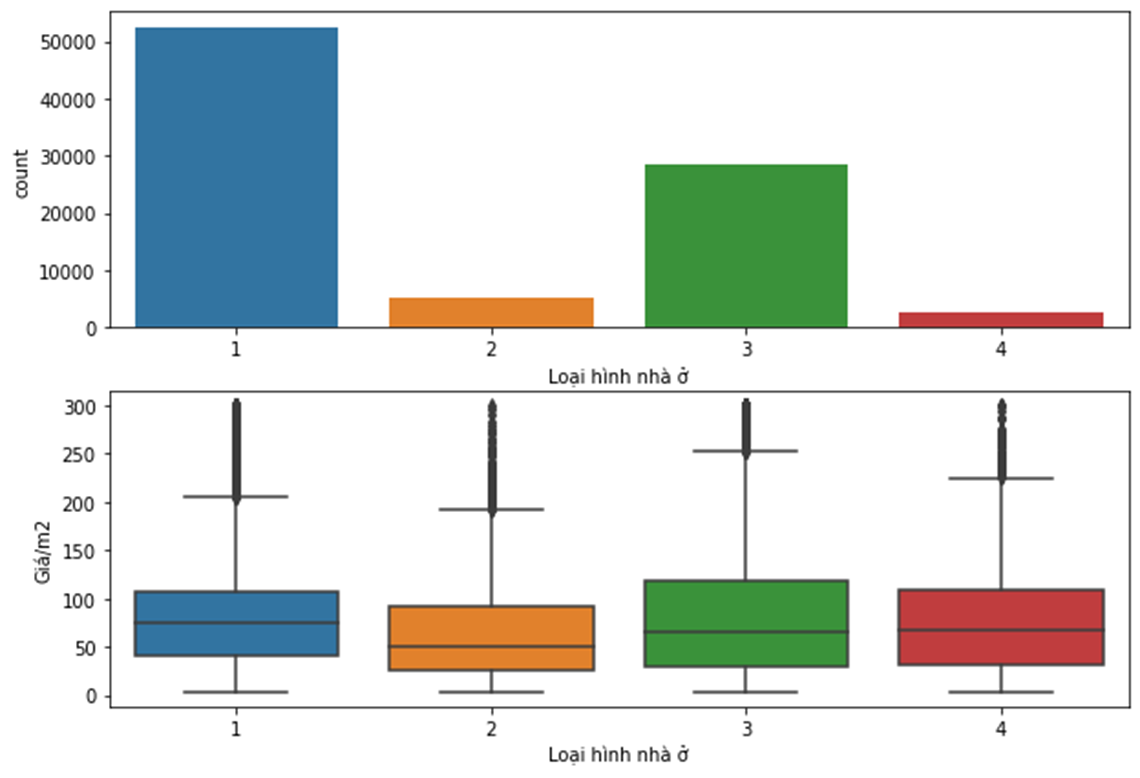
*Biểu đồ thể hiện số lượng phòng vệ sinh và sự tương quan số lượng phòng vệ sinh so với Giá/m2.*

Nhận xét: số lượng phòng ngủ và vệ sinh sẽ đa số trong khoảng 1-5 và giá tăng khi số lượng tăng kết quả biểu diễn này khá sát với thực tế.



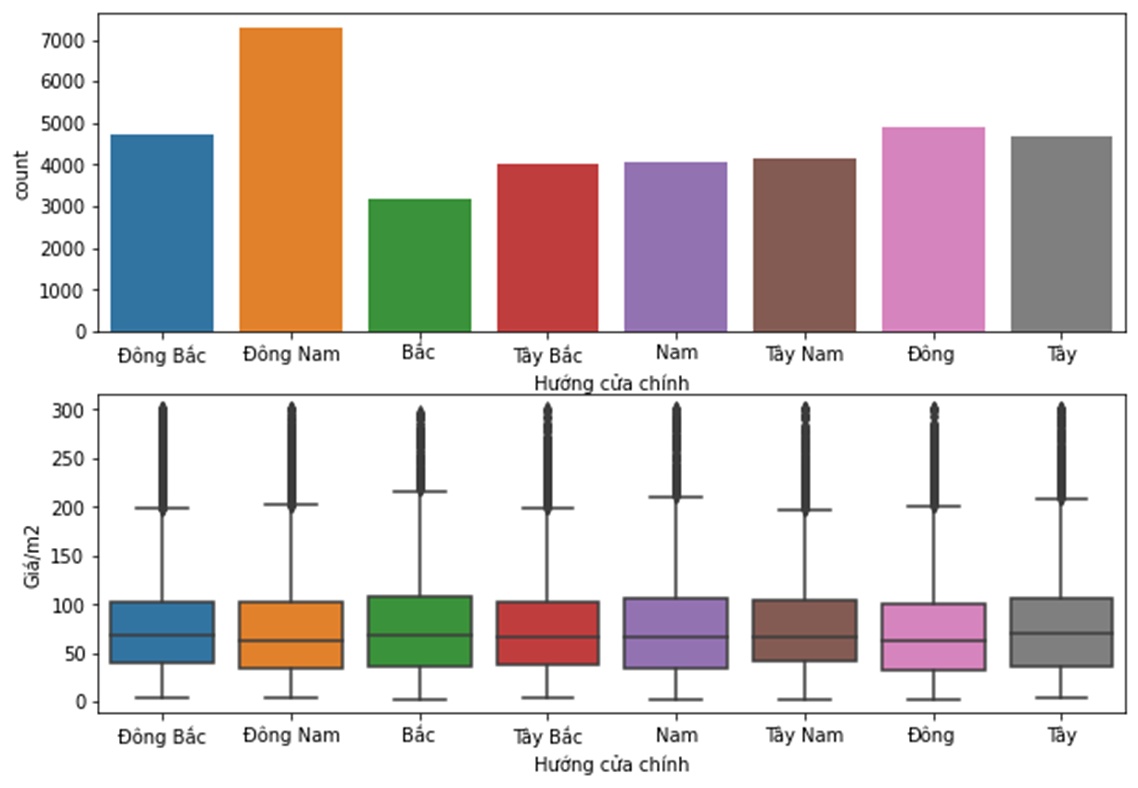
*Biểu đồ số lượng giấy tờ pháp lý và sự tương quan giấy tờ pháp lý và giá.*

Nhận xét: Số lượng nhà đã có sổ chiếm số lượng lớn nhất và giá cao nhất nằm ở các ngôi nhà đã có sổ.



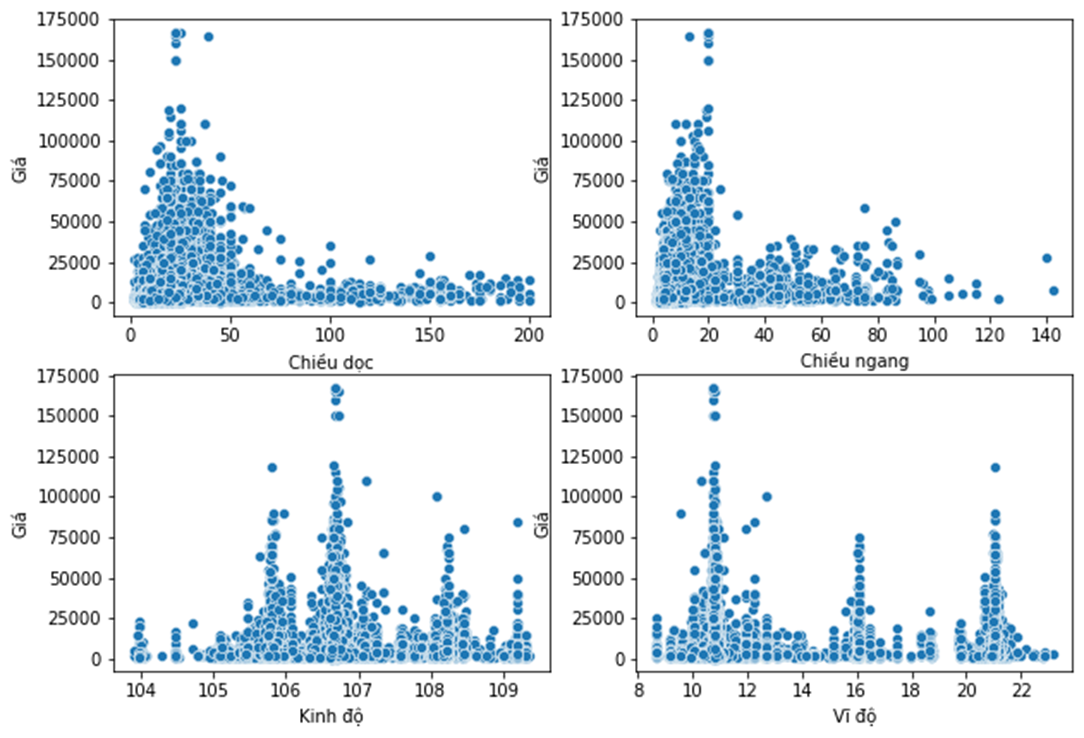
*Biểu đồ loại hình nhà ở*

Nhận xét: 1,2,3,4 sẽ biểu diễn lần lượt loại hình nhà ở: nhà ngõ, nhà phố liền kế, nhà mặt phố, nhà biệt thự. Nhận thấy nhà ngõ sẽ có số lượng bản ghi lớn nhất và sự phân bổ loại hình nhà ở theo giá/m2 còn phụ thuộc vào diện tích của từng loại hình nhà ở.



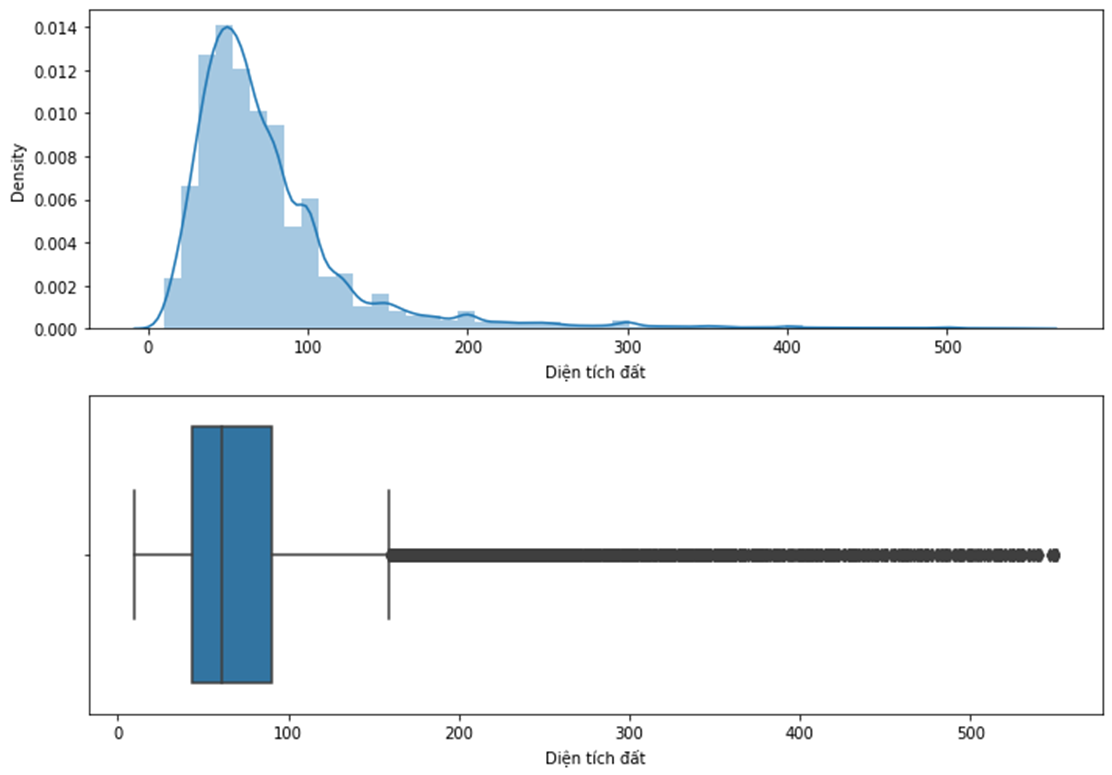
*Biểu đồ số lượng hướng cửa chính và sự tương quan giữa hướng cửa và giá*

Nhận xét: Ta thấy hướng Đông Nam có số lượng bản ghi lớn nhất và sự phân bố giá của hướng cửa chính đều giống nhau cho nên nhóm chúng em sẽ loại bỏ trường hướng cửa chính này.



*Biều đồ phân bổ số lượng và tương quan giữa các thuộc tính chiều dọc chiều ngang kinh độ vĩ độ và giá*

Nhận xét: Ta nhận thấy rõ rất nhiều giá trị ngoại lai cần phải lọc bỏ các outlier.



*Biểu đồ thế hiện sự phân bổ của trường diện tích đất*

Nhận xét: Đa số diện tích đất sẽ phân bố trong khoảng 50-100 m^2

* Từ các nhận xét này chúng em sẽ tiến hành các bước làm sạch dữ liệu như phần trình phía trên.

1. Áp dụng giải thuật học máy dự đoán giá nhà

Giá nhà đất là một dạng số thực, nên không thể giải quyết bài toán bằng bài toán phân loại được. Nhóm em sử dụng cả hai giải thuật hồi quy tuyến tính và k láng giềng gần nhất để so sánh và chọn ra giải thuật tốt nhất.

* 1. Giải thuật hồi quy tuyến tính

Giải thuật hồi quy tuyến tính cố gắng tính toán sự tương quan giữa dữ liệu đầu vào và kết quả tương ứng đầu ra. Ví dụ, sử dụng hồi quy tuyến tính để xét xem sự tương quan giữa các đặc điểm của căn nhà và giá nhà như thế nào từ các dữ liệu đầu vào, từ đó suy ra được một “công thức” gần đúng với các biến đầu vào là các đặc điểm của căn nhà, và đầu ra là giá của căn nhà đó. Từ đó chỉ cần biết được các đặc điểm của căn nhà, áp dụng “công thức” trên dự đoán được giá nhà là bao nhiêu.

Đầu vào của giải thuật hồi quy tuyến tính là một vectơ n chiều .

Đầu ra của giải thuật là một giá trị thực, được tính theo công thức đã xác định được từ quá trình đưa tập dữ liệu ban đầu vào giải thuật.

*Khi nào sử dụng giải thuật hồi quy tuyến tính*

Giải thuật hồi quy tuyến tính thực hiện các tính toán tuyến tính, đơn giản, dễ cài đặt và thuật toán dễ hiểu. Chi phí tính toán thấp (bởi vì các phép toán được thực hiện tuyến tính). Thời gian dự đoán nhanh, cho kết quả ngay, phù hợp với những dạng bài toán có độ phức tạp không quá cao và cần thời gian dự đoán ngay lập tức.

* 1. Giải thuật k láng giềng gần nhất

Giải thuật k-Nearest Neighbors (kNN) là một trong những thuật toán học có giám sát đơn giản nhất và hiệu quả trong một vài trường hợp đối với các bài toán học máy. Thuật toán này không học bất cứ một điều gì từ dữ liệu huấn luyện và mọi tính toán được thực hiện khi nó cần dự đoán kết của của dữ liệu mới. Giống như tên gọi của nó, kNN đi tìm đầu ra của một điểm dữ liệu mới bằng cách dựa trên thông tin của k hàng xóm gần nó nhất trong tập huấn luyện (dựa trên các hàm khoảng cách như Manhattan, Euclid, Cosine,…), không quan tâm đến việc có một trong vài điểm gần nhất này là nhiễu.

Thuật toán kNN có thể áp dụng được vào cả 2 loại bài toán học có giám sát là Phân lớp và Hồi quy:

- Bài toán phân lớp: Hàm mục tiêu phải có giá trị rời rạc, và đầu ra của cũng phải là một trong số các giá trị rời rạc đã xác định trước (một trong số các nhãn lớp).

- Bài toán hồi quy: Hàm mục tiêu phải có giá trị liên tục, và đầu ra là một giá trị thực.

Đối với bài toán Hồi quy, đầu ra của một điểm dữ liệu có thể sẽ bằng chính đầu ra của điểm dữ liệu gần nhất với nó trong tập huấn luyện (k = 1), hoặc trung bình có trọng số đầu ra của những điểm gần nó nhất, hoặc bằng một mối quan hệ dựa trên khoảng cách đối với các điểm gần nhất đó.

Đối với bài toán dự đoán giá nhà giải thuật KNN sẽ đi tìm ra k căn nhà (trong tổng số những căn nhà đã biết giá) có “khoảng cách” gần với căn nhà cần dự đoán giá nhất. Hay nói cách khác, từ những đặc điểm của các căn nhà đã biết giá và căn nhà cần dự đoán giá, thuật toán trên sẽ tính được k căn nhà có những đặc điểm giống với căn nhà cần dự đoán nhất, sau đó giá của căn nhà cần dự đoán sẽ được tính dựa trên giá của k căn nhà đó.

***Mô tả giải thuật kNN***

K láng giềng gần nhất (tên Tiếng Anh là K-Nearest Neighbor, viết tắt là KNN) là một giải thuật học máy có giám sát đơn giản, dễ cài đặt và có thể sử dụng để giải quyết cả hai dạng bài toán phân loại và hồi quy.

Ý tưởng của phương pháp học dựa trên láng giềng gần nhất đó là với một tập các ví dụ học, thuật toán đơn giản là lưu lại các ví dụ học, chưa xây dựng mô hình mô tả rõ ràng và tổng quát của hàm mục tiêu cần học. Đối với một ví dụ cần phân loại hay hồi quy, thuật toán xem xét ví dụ đó với các ví dụ học để gán giá trị của hàm mục tiêu (một nhãn hay một giá trị thực)

***Biểu diễn đầu vào của thuật toán***

Mỗi ví dụ x được biểu diễn là một vectơ n chiều trong không gian các vectơ X ∈ Rn.

x = (x1,x­1,…,xn), trong đó xi (∈R) là một số thực. Có thể áp dụng với hai kiểu bài toán: bài toán phân lớp (hàm mục tiêu có giá trị rời rạc, đầu ra của hệ thống là một trong số các giá trị rời rạc được xác định trước, là một trong các nhãn lớp) hoặc bài toán hồi quy (hàm mục tiêu có giá trị liên tục, đầu ra của hệ thống là một số thực không biết trước).

***Giải thuật phân lớp KNN***

Mỗi ví dụ học x được biểu diễn bởi 2 thành phần, mô tả của ví dụ: x=(x1,x2,…,xn), trong đó xi∈R, và giá trị đầu ra mong muốn: yx∈R (là một số thực). Trong giai đoạn học, thuật toán lưu lại các ví dụ học trong tập học D. Trong giai đoạn dự đoán, đối với mỗi ví dụ học x∈D, tính khoảng cách giữa x và z (ví dụ cần dự đoán), từ đó xác định được các láng giềng gần nhất của z, gồm k ví dụ học trong D gần nhất với z tính theo một hàm khoảng cách d.

***Các hàm tính khoảng cách***

*Các hàm tính khoảng cách hình học*

* Hàm Minkowski (p-norm)
* Hàm Mahattan (p=1)
* Hàm Euclid (p=2)

*Hàm khoảng cách Hamming*

Các thuộc tính đầu vào là kiểu nhị phân ({0, 1}), ví dụ x = (1, 0, 1, 0, 1)

Với

*Hàm tính độ tương tự Cosine*

Độ tương tự cosine được tính:

Suy ra khoảng cách giữa x và z là:

***Chuẩn hoá miền giá trị thuộc tính***

Lấy ví dụ với hàm tính khoảng cách Euclid ở trên. Giả sử mỗi ví dụ được biểu diễn bởi 3 thuộc tính là Age (tuổi), Income (thu nhập hàng tháng), Height (chiều cao đo theo mét).

Áp dụng công thức tính khoảng cách Euclid ta có khoảng cách giữa x và z là:

Thấy rằng giá trị khoảng cách bị quyết định chủ yếu bởi giá trị khoảng cách (sự khác biệt) giữa hai ví dụ đối với thuộc tính Income. Vì thuộc tính này có miền giá trị rất lớn so với các thuộc tính còn lại.

Do đó cần phải chuẩn hoá miền giá trị (đưa về cùng một khoảng giá trị). Khoảng giá trị [0, 1] thường được sử dụng. Đối với mỗi thuộc tính .

***Trọng số của các thuộc tính***

Lấy ví dụ với hàm khoảng cách Euclid, tất cả các thuộc tính có ảnh hưởng như nhau đối với giá trị khoảng cách.

Các thuộc tính khác nhau có thể (nên) có mức độ ảnh hưởng khác nhau đối với giá trị khoảng cách. Cần phải tích hợp các giá trị trọng số của các thuộc tính trong hàm tính khoảng cách.

với wi là trọng số của thuộc tính i

Làm sao để xác định các giá trị trọng số của các thuộc tính? Cần 2 yếu tố để xác định giá trị trọng số, yếu tố thứ nhất là dựa trên các tri thức cụ thể của bài toán (ví dụ: được chỉ định bởi các chuyên gia trong lĩnh vực của bài toán đang xét), yếu tố thứ hai là bằng một quá trình tối ưu hoá các giá trị trọng số (ví dụ: sử dụng một tập học để học một bộ các giá trị trọng số tối ưu).

***Sử dụng KNN khi nào***

Những bài toán có các đặc điểm sau sẽ phù hợp khi lựa chọn giải thuật kNN: các ví dụ được biểu diễn trong không gian vectơ Rn, số lượng các thuộc tính để biểu diễn ví dụ là không nhiều, một tập học có kích thước lớn.

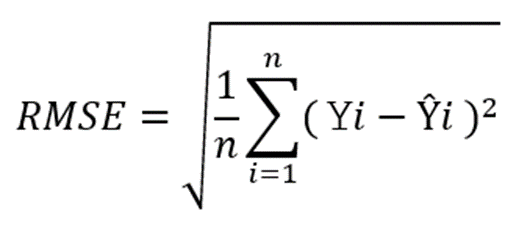
Ưu điểm của giải thuật kNN: giải thuật có chi phí thấp cho quá trình huấn luyện (chỉ lưu lại các ví dụ học), hoạt động tốt với các bài toán phân loại cho n lớp do đó không cần phải học n bộ phân loại cho n lớp, việc chọn k phù hợp (thường lớn hơn nhiều so với 1) có khả năng xử lý nhiễu cao bởi phân loại/dự đoán được thực hiện dựa trên k láng giềng gần nhất.

Nhược điểm của giải thuật kNN: phải lựa chọn hàm tính khoảng các thích hợp với từng bài toán, chi phí tính toán cao (thời gian, bộ nhớ) tại thời điểm phân loại/dự đoán, có thể cho kết quả kém/sai với các thuộc tính không liên quan.

* 1. Phương pháp đánh giá giải thuật

Để đánh giá và so sánh giữa hai giải thuật: hồi quy tuyến tính và K láng giềng gần nhất với nhau. Nhóm sử dụng hai phương pháp đánh giá là mean square error và r2 score của thư viện sklearn.

* + 1. Root Mean Square Error



Giá trị RMSE càng nhỏ thì giải thuật và bộ tham số chọn càng tốt và ngược lại.

* + 1. R2 Score



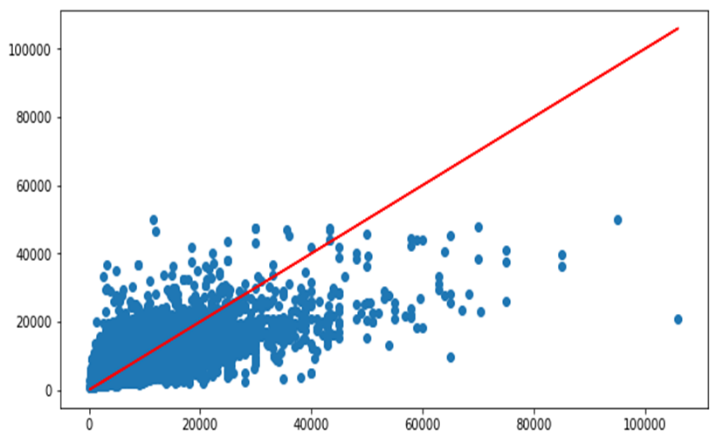
Với là giá trị thực tế,  là giá trị dự đoán

Giá trị R2 Score luôn có giá trị nhỏ hơn hoặc bằng 1. Giá trị này càng gần với 1 thì giải thuật và bộ tham số chọn càng tốt và ngược lại.

* 1. Thực nghiệm và đánh giá
     1. Thực nghiệm

Chia dữ liệu huấn luyện: Tập train gồm 68389 bản ghi và tập test gồm 17097 bản ghi.

Kết quả huấn luyện model Linear Regression (Scikit-Learn):

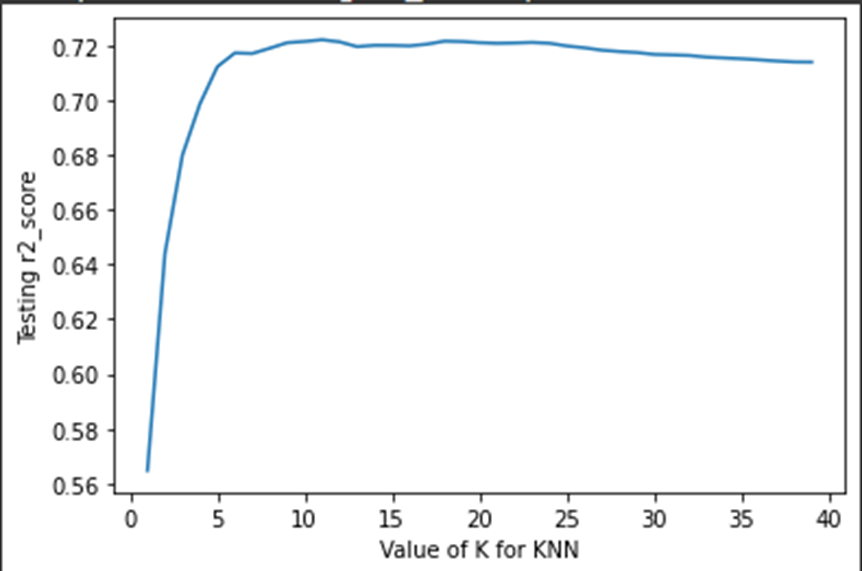
**** R2\_Score: 0.4363

RMSE: 4296.03

KNN: Từ quá trình train model ta thấy tối ưu nhất với hàm tính khoảng cách Minkowski có p = 1 và số láng giềng K = 12, đánh hệ số cao hơn cho các láng giềng gần

Kết quả: RMSE: 2968 và R2\_Score: 0.7220.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| weights | **k** | **p** | **RMSE** | **R2 Score** |
| distance | 7 | 1 | 2993.224 | 0.7172 |
| 8 | 1 | 2993.325 | 0.7170 |
| 9 | 1 | 2984.961 | 0.7189 |
| 10 | 1 | 2973.213 | 0.7209 |
| 11 | 1 | 2971.531 | 0.7214 |
| 12 | 1 | 2968.164 | 0.7220 |
| 13 | 1 | 2971.531 | 0.7212 |
| 14 | 1 | 2979.932 | 0.7195 |



* + 1. Đánh giá kết quả

Giải thuật KNN cho kết quả tốt hơn hẳn so với giải thuật hồi quy tuyến tính.

Trong giải thuật KNN, bộ tham số (weights= distance, p=1, k=12) cho kết quả tốt nhất với giá trị RMSE = 2968 và R2 Score = 0.7220.

=> KNN là mô hình đặc biệt thích hợp cho bài toán này.

Kết quả giữa hai giải thuật như vậy là bởi vì tập dữ liệu mặc dù đã được xử lý qua nhiều giai đoạn nhưng vẫn còn chứa dữ liệu bất thường và nhiễu.

=> Do đó giải thuật đơn giản như hồi quy tuyến tính khó có thể phù hợp được với bài toán này và với tập dữ liệu này.

1. Xây dựng chương trình demo đơn giản

Nhóm lựa chọn giải thật cho kết quả tốt nhất là KNN để xây dựng ứng dụng web dự đoán. Từ thông tin người dùng cung cấp về ngôi nhà qua thuật toán KNN trên sẽ trả ra kết quả dự đoán giá của ngôi nhà.

Nhóm em sử dụng thư viện Streamlit để demo chương trình.

Ảnh có chứa bàn

Mô tả được tạo tự động

Chương trình demo có nhược điểm là yêu cầu nhập input phần Kinh độ, Vĩ độ. Nhóm em có tìm hiểu rằng có thể sử dụng Google API để từ địa chỉ của ngôi nhà có thể lấy được tọa độ, tuy nhiên có thu phí. Trong tương lai nếu có cơ hội, nhóm em sẽ cố gắng tạo ra chương trình demo có tính thực tế hơn.

Ảnh có chứa bàn

Mô tả được tạo tự động

1. Kết luận
   1. Kết luận chung

Qua nghiên cứu và thực nghiệm, nhóm em đã đạt được một số kết quả như sau:

-Tìm hiểu về những đặc trưng của bất động sản, những thuộc tính quan trọng ảnh hướng tới giá bất động sản.

-Tiến hành thu thập dữ liệu thực tế để phục vụ huấn luyện mô hình.

-Nghiên cứu tổng quan về Khoa học dữ liệu, Học máy với các phương pháp cơ bản và ứng dụng thực tiễn sau này.

-Bước đầu nhóm em đã đưa ra được mô hình có thể dự đoán được giá bất động sản.

* 1. Hướng phát triển tiếp theo

- Nghiên cứu sâu về những đặc điểm của mô hình cũng như đặc trưng của Bất động sản.

- Tìm hiểu thêm những mô hình học máy cũng như những phương pháp xử lý dữ liệu.

- Mở rộng mô hình và tăng tính thực tế của bài toán.

1. Phân chia công việc

* Crawl: Nguyễn Minh Hiển, Nguyễn Hữu Hải
* Tiền xử lý: Nguyễn Văn Sáng
* Visualization, phân tích dữ liệu: Nguyễn Hữu Hải
* Dựng model: Nguyễn Minh Hiển, Nguyễn Văn Sáng
* Làm sản phẩm demo: Nguyễn Minh Hiển
* Làm slide, báo cáo: tương tự

1. Tài liệu tham khảo
2. Tài liệu học tập, slide môn học Nhập môn khoa học dữ liệu được các thầy cô cung cấp.
3. Sklearn R2 Score - https://scikitlearn.org/stable/modules/generated/sklearn.metrics.r2\_score.html
4. Sklearn Root Mean Square Error - https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean\_squared\_error.html
5. A Simple Introduction to K-Nearest Neighbors Algorithm - https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e
6. Linear Regression - <https://machinelearningcoban.com/2016/12/28/linearregression/>
7. Streamlit - [The fastest way to build and share data apps](https://streamlit.io/).