**TRƯỜNG ĐẠI HỌC MỞ HÀ NỘI**

**KHOA CÔNG NGHỆ THÔNG TIN**

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

****

**BÁO CÁO BÀI TẬP LỚN**

**MÔN HỌC: NHẬP MÔN KHAI PHÁ DỮ LIỆU**

**VÀ MÁY HỌC**

**Đề tài: Dự đoán giá nhà đất tại Hà Nội**

***Giảng viên hướng dẫn:* ThS. Vũ Xuân Hạnh**

***Nhóm/Sinh viên thực hiện: Nhóm 14***

**Nguyễn Minh Đức - 2110A05**

**Trần Thanh Hải - 2110A05**

**Đỗ Thị Ngọc Huyền – 2110A05**

**Nguyễn Mai Linh - 2110A05**

**Hà  Nội – 2024**

**Phân công nhiệm vụ**

| Họ và tên | Nhiệm vụ | Ghi chú |
| --- | --- | --- |
| Nguyễn Minh Đức | - Tìm hiểu các phương pháp xử lý dữ liệu, phân tích dữ liệu, các mô hình học máy, các phương pháp đánh giá mô hình học máy.  - Thu thập, tiền xử lý, vector hóa dữ liệu.  - Tích hợp mô hình học máy vào phần mềm.  - Tổng hợp viết báo cáo. |  |
| Trần Thanh Hải | - Tìm hiểu các lý thuyết Machine Learning, Data Mining, các phương pháp xử lý và đánh giá dữ liệu.  - Thu thập dữ liệu.  - Tổng hợp viết báo cáo. |  |
| Đỗ Thị Ngọc Huyền | - Tìm hiểu các lý thuyết Machine Learning, Data Mining, các phương pháp xử lý và đánh giá dữ liệu.  - Thu thập dữ liệu  - Tổng hợp viết báo cáo. |  |
| Nguyễn Mai Linh | - Tiền xử lý dữ liệu, vector hóa dữ liệu.  - Tìm hiểu các phương pháp xử lý dữ liệu, phân tích dữ liệu.  - Tổng hợp viết báo cáo. |  |

**MỤC LỤC**

[**1. Tổng quan về đề tài 4**](#_heading=h.jua3isomwc36)

[1.1. Đặt vấn đề 4](#_heading=h.t8bqxectoufk)

[1.2. Mục đích và xác định vấn đề 4](#_heading=h.a9s96p4wqlyv)

[**2. Cơ sở lý thuyết 5**](#_heading=h.w95awilpuyg)

[2.1. Tổng quan về kỹ thuật khai phá dữ liệu 5](#_heading=h.ud4jque6jom9)

[2.2. Thuật toán hồi quy tuyến tính 8](#_heading=h.hei7icbqujdg)

[**3. Xây dựng mô hình 9**](#_heading=h.uqe2vdk6vsl9)

[3.1. Cơ sở dữ liệu xây dựng mô hình 9](#_heading=h.qwn0lkvwmqcv)

[3.1.1. Thu thập dữ liệu 10](#_heading=h.jj5aso69fj4g)

[3.1.2. Tiền xử lý dữ liệu 10](#_heading=h.jqto6bihzz2x)

[3.1.3. Phân tích dữ liệu 19](#_heading=h.q9m6nmvr34kt)

[3.2. Xây dựng ứng dụng 22](#_heading=h.9ugil30rm6d)

[**4. Tài Liệu Tham Khảo 22**](#_heading=h.l8cn1312819f)

# 

# Tổng quan về đề tài

## Đặt vấn đề

Trong thị trường bất động sản ngày nay, việc đánh giá và dự đoán giá nhà đất đang đối mặt với những thách thức đáng kể. Mặc dù có nhiều yếu tố ảnh hưởng, từ vị trí địa lý đến tiện ích và yếu tố thị trường, nhưng việc đưa ra dự đoán chính xác vẫn là một nhiệm vụ khó khăn. Trong khi đó, sự phát triển của công nghệ dữ liệu và máy học đã mở ra những cơ hội mới trong việc cải thiện khả năng dự đoán giá nhà đất.

Tính đến hiện nay, việc áp dụng khai phá dữ liệu và máy học vào dự đoán giá nhà đất không chỉ là một cơ hội, mà còn là một nhu cầu thiết yếu. Bằng cách sử dụng các phương pháp và công nghệ tiên tiến, chúng ta có thể khai thác và phân tích một lượng lớn dữ liệu từ nhiều nguồn khác nhau, từ thông tin thị trường đến dữ liệu về môi trường xã hội và kinh tế. Điều này giúp chúng ta hiểu sâu hơn về các yếu tố ảnh hưởng đến giá nhà đất và xây dựng các mô hình dự đoán chính xác hơn.

Tuy nhiên, trong khi tiềm năng của việc áp dụng dữ liệu và máy học rất lớn, nhưng cũng không thiếu những thách thức. Việc xử lý và phân tích dữ liệu lớn có thể đòi hỏi cơ sở hạ tầng và kỹ năng chuyên môn mạnh mẽ. Ngoài ra, việc đảm bảo tính minh bạch và công bằng trong việc sử dụng các mô hình dự đoán cũng là một yếu tố quan trọng cần được xem xét.

Vì vậy, trong bối cảnh này, việc nghiên cứu và áp dụng các phương pháp khai phá dữ liệu và máy học để dự đoán giá nhà đất không chỉ là một cơ hội mà còn là một thách thức đối với cả nhà nghiên cứu và doanh nghiệp. Qua đó, chúng ta hy vọng có thể tạo ra những công cụ và phương pháp hiệu quả hơn để hỗ trợ quyết định đầu tư và phát triển bền vững của thị trường bất động sản trong tương lai.

## Mục đích và xác định vấn đề

**Mục đích của đề tài "Dự đoán giá nhà đất đã qua sử dụng" là:**

* Phát triển một mô hình dự đoán gần như chính xác giá trị của nhà đất dựa trên các yếu tố như địa chỉ, số phòng, sổ đỏ và các yếu tố khác.
* Cung cấp một công cụ hữu ích cho những người muốn mua/bán đất để đưa ra quyết định mua bán dựa trên dự đoán giá cả.
* Tăng cường minh bạch và công bằng trong quá trình mua bán nhà đất bằng cách cung cấp thông tin dự đoán giá cả có độ chính xác cao và đáng tin cậy.
* Khám phá và áp dụng các phương pháp và công nghệ mới trong lĩnh vực khoa học dữ liệu và học máy để xử lý và phân tích dữ liệu nhà đất một cách hiệu quả.
* Tóm lại, mục đích của đề tài là tạo ra một hệ thống dự đoán giá nhà đất đáng tin cậy và hữu ích, đồng thời đóng góp vào việc nâng cao hiệu quả hoạt động trong thị trường bất động sản và cung cấp giá trị cho cả người mua và người bán.

**Vấn đề cụ thể mà đề tài "Dự đoán giá nhà đất" sẽ tập trung giải quyết là:**

* Khả năng dự đoán chính xác giá trị của nhà đất trong thị trường bất động sản, giá trị của một nhà đất có thể phụ thuộc vào nhiều yếu tố
* Vấn đề đặt ra là làm thế nào để xây dựng một mô hình dự đoán có khả năng dự đoán chính xác giá trị của nhà đất dựa trên các thông tin này

**Đối tượng nghiên cứu**

* Dữ liệu về nhà đất**:**
* Đối tượng chính của nghiên cứu là dữ liệu liên quan đến các nhà đất trên địa bàn Hà Nội, gồm các dữ liệu về địa chỉ, loại hình nhà ở, giấy tờ pháp lý và các yếu tố khác

**Phương pháp thu thập dữ liệu**

* Dữ liệu được lấy trên trang web: kaggle.com và chotot

# Cơ sở lý thuyết

## Tổng quan về kỹ thuật khai phá dữ liệu

**Khái niệm**

* Khai phá dữ liệu (Data Mining) là quá trình tự động khám phá các mẫu, quy luật, thông tin ẩn hoặc tri thức có giá trị từ các tập dữ liệu lớn và phức tạp. Mục tiêu chính của khai phá dữ liệu là tìm ra các mối quan hệ, xu hướng và thông tin hữu ích từ dữ liệu để hỗ trợ ra quyết định, dự đoán và khám phá tri thức mới.
* KPDL thu hút sự chú ý của nền công nghiệp thông tin và xã hội trong những năm gần đây. Với sự phát triển của công nghệ thông tin, dữ liệu lưu trữ mỗi ngày trở thành một cơ sở dữ liệu rất lớn. Dựa vào khối lượng dữ liệu này, ta dùng những kỹ thuật KPDL để chuyển dữ liệu đó thành những thông tin có ích hoặc rút ra những tri thức mới từ dữ liệu thu thập được.Giáo sư Tom Mitchell định nghĩa Khai phá dữ liệu như sau: “Khai phá dữ liệu là việc sử dụng dữ liệu lịch sử để khám phá những quy tắc và cải thiện những quyết định trong tương lai”

**Ý nghĩa**

* Ý nghĩa của khai phá dữ liệu là tạo ra những giá trị và cơ hội mới từ các tập dữ liệu lớn và phức tạp. Dưới đây là một số ý nghĩa quan trọng của khai phá dữ liệu:
* Khám phá tri thức ẩn: Khai phá dữ liệu giúp khám phá những thông tin, mẫu, và quy luật ẩn trong dữ liệu mà không thể dễ dàng nhận biết bằng mắt thường. Điều này giúp hiểu rõ hơn về mối quan hệ và xu hướng trong dữ liệu, từ đó tạo ra những tri thức mới.
* Hỗ trợ ra quyết định: Các kết quả từ quá trình khai phá dữ liệu có thể được sử dụng để hỗ trợ ra quyết định trong các lĩnh vực khác nhau như kinh doanh, y tế, tài chính, marketing, v.v. Bằng cách hiểu rõ hơn về dữ liệu, người ra quyết định có thể đưa ra các quyết định thông minh và có căn cứ hơn.
* Dự đoán và phân tích xu hướng: Khai phá dữ liệu cho phép dự đoán và phân tích xu hướng trong dữ liệu, từ đó giúp dự đoán các sự kiện tương lai và hiểu rõ hơn về hành vi của người dùng, khách hàng, hoặc các biến khác trong một hệ thống.
* Tối ưu hóa quá trình kinh doanh và hoạt động: Bằng cách sử dụng thông tin từ quá trình khai phá dữ liệu, tổ chức có thể tối ưu hóa quá trình kinh doanh, tăng hiệu quả và giảm chi phí. Ví dụ, có thể tối ưu hóa chuỗi cung ứng, quản lý rủi ro hoặc tối ưu hóa chiến lược marketing.
* Tạo ra giá trị mới: Khai phá dữ liệu tạo ra giá trị mới từ dữ liệu, từ đó tạo ra cơ hội mới cho sự phát triển và cạnh tranh trong các lĩnh vực kinh doanh và xã hội.

🡺 Tóm lại, khai phá dữ liệu mang lại nhiều lợi ích quan trọng trong việc hiểu và tận dụng thông tin từ dữ liệu, từ đó cải thiện quyết định và hoạt động trong các lĩnh vực khác nhau.

**Các giai đoạn về quá trình khai phá dữ liệu**

1. **Tìm hiểu nghiệp vụ và dữ liệu**

Giai đoạn này ta cần xác định vấn đề cần giải quyết, tìm hiểu kiến thức về bài toán đang thực hiện bao gồm các tri thức của các chuyên gia trong lĩnh vực cần nghiên cứu từ đó xác định chính xác nguồn dữ liệu để thu thập đồng thời phải hiểu được cấu trúc dữ liệu, ý nghĩa và tầm quan trọng của nó để từ đó ta đưa ra bài toán cụ thể để giải quyết vấn đề.

1. **Chuẩn bị dữ liệu**

 Giai đoạn này ta dùng các kỹ thuật tiền xử lý dữ liệu để xử lý dữ liệu đã thu thập được sao cho các giải thuật KPDL có thể hiểu được. Tiền xử lý dữ liệu bao gồm:

+ Xử lý dữ liệu bị thiếu hoặc mất: Các giá trị bị thiếu hoặc mất sẽ được thay thế bằng các giá trị thích hợp hơn hoặc xóa những dữ liệu sai miền giá trị và giải quyết sự không nhất quán.

+ Khử sự trùng lặp dữ liệu : Loại bỏ những dữ liệu bị trùng.

+ Giảm nhiễu dữ liệu: Các dữ liệu bị nhiễu sẽ được điều chỉnh hoặc loại ra khỏi cơ sở dữ liệu.

+ Rời rạc hóa dữ liệu: Các dữ liệu số sẽ được rời rạc hóa ra dạng phù hợp cho khai phá dữ liệu.

+ Giảm chiều: Loại bớt các thuộc tính chứa ít thông tin để tiết kiệm thời gian và tài nguyên của máy tính.

1. **Mô hình hóa dữ liệu**

 Dùng các giải thuật của KPDL để tìm ra các qui luật của dữ liệu, quan trọng nhất trong giai đoạn này là tìm được giải thuật phù hợp để giải quyết vấn đề đã đặt ra.

1. **Hậu xử lý và đánh giá mô hình**

 Đây là giai đoạn biến đổi từ những luật rút ra được (của giai đoạn trước) từ tập huấn luyện sang dạng phù hợp với nghiệp vụ của bài toán đang nghiên cứu. Đồng thời cũng sẽ là giai đoạn đánh giá của các chuyên gia tư vấn dựa trên tập dữ liệu thử. Dựa vào nhận xét và hỗ trợ của các chuyên gia khi đó sẽ điều chỉnh kịp thời các mô hình của các giai đoạn trước. Các mô hình đạt yêu cầu với các chuyên gia sẽ được sử dụng.

1. **Triển khai mô hình**

 Các mô hình đạt yêu cầu sẽ được xây dựng thành chương trình ứng dụng thực tế nhằm hỗ trợ đưa ra quyết định theo yêu cầu của người dùng.

## Thuật toán hồi quy tuyến tính

**Khái niệm**

Trong khai phá dữ liệu, thuật toán hồi quy tuyến tính được sử dụng để mô hình hóa mối quan hệ tuyến tính giữa các biến độc lập và biến mục tiêu. Ý tưởng cơ bản của thuật toán này là tìm một đường tuyến tính (trong không gian hai chiều) hoặc siêu phẳng (trong không gian ba chiều) sao cho tổng bình phương của sai số giữa các giá trị dự đoán và các giá trị quan sát là nhỏ nhất.

Dưới đây là một số khái niệm quan trọng liên quan đến thuật toán hồi quy tuyến tính trong khai phá dữ liệu:

* Mục tiêu: Dự đoán giá trị của biến mục tiêu dựa trên các biến độc lập.
* Hàm mục tiêu: Trong hồi quy tuyến tính, hàm mục tiêu thường là một hàm tuyến tính của các biến độc lập. Ví dụ, nếu chúng ta có một biến độc lập, hàm mục tiêu có thể được biểu diễn như sau: γ=0+ 1, trong đó © là biến mục tiêu, ⎟là biến độc lập, 0 và 1 là các tham số của mô hình.
* Tham số của mô hình: Trong hồi quy tuyến tính, các tham số của mô hình (như 0 và 1 được ước lượng bằng phương pháp bình phương nhỏ nhất, trong đó cố gắng tối thiểu hóa tổng bình phương của sai số giữa các giá trị quan sát và giá trị dự đoán.
* Đánh giá mô hình: Mô hình hồi quy tuyến tính thường được đánh giá bằng cách sử dụng các độ đo như hệ số xác định(R2), sai số trung bình (MSE), hay các độ đo tương tự để đánh giá hiệu suất dự báo của mô hình.
* Ứng dụng: Thuật toán hồi quy tuyến tính được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau của khai phá dữ liệu, bao gồm kinh doanh, tài chính, y tế, khoa học xã hội và nhiều lĩnh vực khác để dự đoán và mô tả mối quan hệ giữa các biến.

**Cụ thể về thuật toán**

Dưới đây là một ví dụ cụ thể về việc sử dụng thuật toán hồi quy tuyến tính trong khai phá dữ liệu:

Ví dụ: Dự đoán giá nhà dựa trên diện tích

Giả sử bạn là một nhà phân tích dữ liệu và bạn muốn dự đoán giá của các căn nhà dựa trên diện tích của chúng. Trong trường hợp này, diện tích sẽ là biến độc lập và giá nhà sẽ là biến mục tiêu.

Thu thập dữ liệu: Bạn thu thập một tập dữ liệu gồm diện tích của các căn nhà và giá bán tương ứng.

Tiền xử lý dữ liệu: Bạn kiểm tra dữ liệu để đảm bảo tính toàn vẹn và xử lý bất kỳ giá trị thiếu nào. Nếu có giá trị ngoại lai hoặc dữ liệu không chính xác, bạn có thể quyết định loại bỏ chúng.

Áp dụng hồi quy tuyến tính: Bạn áp dụng thuật toán hồi quy tuyến tính vào tập dữ liệu của mình. Trong trường hợp này, mô hình hồi quy tuyến tính có thể được biểu diễn như sau:

Giá\_nhà = 0 + 1 . diện\_tích

Trong đó:

- Giá\_nhà là giá nhà dự đoán,

- diện\_tích là diện tích của căn nhà,

- 0 và 1 là các tham số của mô hình cần được ước lượng.

Ước lượng tham số: Sử dụng phương pháp của bình phương nhỏ nhất, bạn ước lượng các tham số 0 và 1 sao cho tổng bình phương của sai số là nhỏ nhất.

Đánh giá mô hình: Bạn đánh giá hiệu suất của mô hình bằng cách sử dụng các độ đo như hệ số xác định (R2), sai số trung bình (MSE), hay các độ đo tương tự.

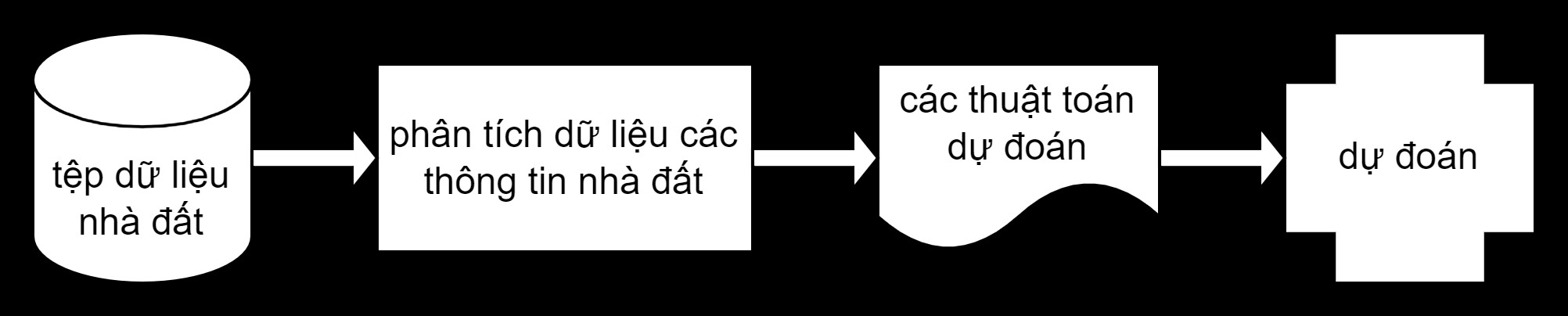
Sử dụng mô hình: Khi mô hình đã được xác định và đánh giá, bạn có thể sử dụng nó để dự đoán giá nhà cho các căn nhà mới dựa trên diện tích của chúng.

Thông qua việc sử dụng thuật toán hồi quy tuyến tính, bạn có thể tạo ra một mô hình đơn giản nhưng hiệu quả để dự đoán giá nhà dựa trên diện tích.

# Xây dựng mô hình

## Cơ sở dữ liệu xây dựng mô hình

Sau khi thu thập dữ liệu ta cần xây dựng cơ sở dữ liệu , lưu trữ các thông tin cần thiết cho bộ điều khiển theo mô hình sau:



### Thu thập dữ liệu

Dữ liệu được thu thập qua các kênh như Kaggle, chotot,.. dưới dạng file csv:



Dữ liệu bao gồm các thông số về nhà đất:

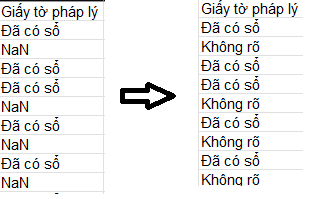
* + STT
  + Ngày
  + Địa chỉ
  + Quận
  + Huyện
  + Loại hình nhà ở
  + Giấy tờ pháp lý
  + Số tầng
  + Số phòng ngủ
  + Diện tích
  + Dài, rộng
  + Giá

### Tiền xử lý dữ liệu

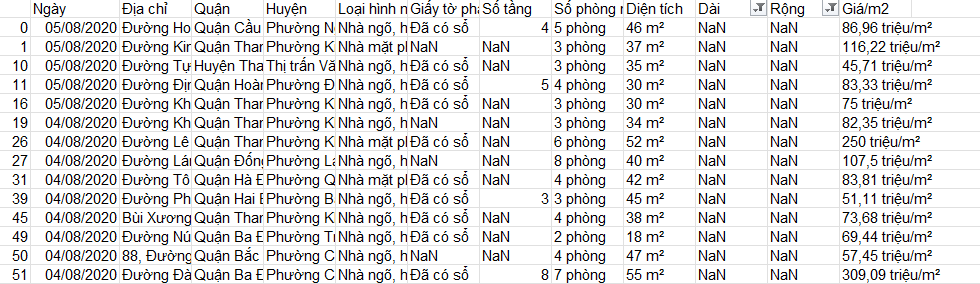
Để thực hiện mô hình khai phá luật kết hợp ta cần hiệu chỉnh lại dữ liệu và loại bỏ các thuộc tính không cần thiết :

Quá trình tiền xử lý dữ liệu:

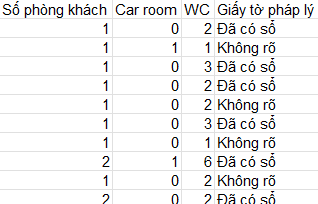
* Xử lý dữ liệu bị thiếu hoặc bị mất (NaN):
  + Các giá trị bị thiếu được thay thế bằng các giá trị phù hợp hơn (cột giấy tờ pháp lý dữ liệu bị thiếu được chuyển thành không rõ)



* + Nếu một dòng dữ liệu có 2 giá trị NaN ở thuộc tính “Rộng”,”Dài” sẽ bị loại bỏ



* Khử trùng lặp dữ liệu xử lý những dữ liệu bị trùng.
* Bổ sung thêm các thuộc tính :”Phòng khách”,”Car room”,”WC”(dữ liệu được bổ sung qua chotot)



* Giảm chiều dữ liệu: lược bỏ thuộc tính “STT”,”Ngày”
* Loại bỏ thuộc tính “Địa chỉ” (không thể vecto hóa) sử dụng thuộc tính “Quận” và “Huyện” làm địa chỉ.
* Lọc và loại bỏ những dòng dữ liệu gây nhiễu làm giảm hiệu suất:



* Chuẩn hóa dữ liệu:

| **Diện tích** | | **Phòng khách, ngủ, xe oto, WC** | |
| --- | --- | --- | --- |
| Dữ liệu gốc | Chuẩn hóa dữ liệu | Dữ liệu gốc | Chuẩn hóa dữ liệu |
| x m² ( string) | X( number) | x phòng (string) | x (number) |
| 46 m² | 46 | 5 phòng | 5 |

| **Dài, Rộng** | | **Giá/m²** | |
| --- | --- | --- | --- |
| Dữ liệu gốc | Chuẩn hóa dữ liệu | Dữ liệu gốc | Chuẩn hóa dữ liệu |
| x m ( string) | X( number) | Giá/m² (string) | Giá (number) |
| 46 m | 46 | 86,96 triệu/m² | 86,96 |

* Rời rạc hóa dữ liệu :

| **Quận** | | | | | |
| --- | --- | --- | --- | --- | --- |
| ID | Tên Quận | ID | Tên Quận | ID | Tên Quận |
| 1 | Quận Ba Đình | 11 | Quận Bắc Từ Liêm | 21 | Huyện Mỹ Đức |
| 2 | Quận Hoàn Kiếm | 12 | Quận Nam Từ Liêm | 22 | Huyện Phú Xuyên |
| 3 | Quận Tây Hồ | 13 | Thị xã Sơn Tây | 23 | Huyện Phúc Thọ |
| 4 | Quận Long Biên | 14 | Huyện Ba Vì | 24 | Huyện Quốc Oai |
| 5 | Quận Cầu Giấy | 15 | Huyện Chương Mỹ | 25 | Huyện Sóc Sơn |
| 6 | Quận Đống Đa | 16 | Huyện Đan Phượng | 26 | Huyện Thạch Thất |
| 7 | Quận Hai Bà Trưng | 17 | Huyện Đông Anh | 27 | Huyện Thanh Oai |
| 8 | Quận Hoàng Mai | 18 | Huyện Gia Lâm | 28 | Huyện Thanh Trì |
| 9 | Quận Thanh Xuân | 19 | Huyện Hoài Đức | 29 | Huyện Thường Tín |
| 10 | Quận Hà Đông | 20 | Huyện Mê Linh |  |  |

| **Phường** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Quận Ba Đình | | Quận Hoàn Kiếm | | Quận Tây Hồ | |
| ID | Tên phường | ID | Tên Phường | ID | Tên Phường |
| 1 | Cống Vị | 1 | Cửa Đông | 1 | Bưởi |
| 2 | Điện Biên | 2 | Cửa Nam | 2 | Nhật Tân |
| 3 | Đội Cấn | 3 | Chương Dương | 3 | Phú Thượng |
| 4 | Giảng Võ | 4 | Đồng Xuân | 4 | Quảng An |
| 5 | Kim Mã | 5 | Hàng Bạc | 5 | Tứ Liên |
| … | … | … | … | … | … |
| 14 | Ngọc Khánh | 17 | Trần Hưng Đạo | 8 | Yên Phụ |

| **Phường** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Quận Long Biên | | Quận Cầu Giấy | | Quận Đống Đa | |
| ID | Tên phường | ID | Tên Phường | ID | Tên Phường |
| 1 | Bồ Đề | 1 | Dịch Vọng | 1 | Cát Linh |
| 2 | Cự Khối | 2 | Dịch Vọng Hậu | 2 | Hàng Bột |
| 3 | Đức Giang | 3 | Mai Dịch | 3 | Kim Liên |
| 4 | Gia Thụy | 4 | Nghĩa Đô | 4 | Khâm Thiên |
| 5 | Giang Biên | 5 | Nghĩa Tân | 5 | Khương Thượng |
| … | … | … | … | … | … |
| 14 | Việt Hưng | 8 | Yên Hòa | 21 | Văn Miếu |

| **Phường** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Quận Hai Bà Trưng | | Quận Hoàng Mai | | Quận Thanh Xuân | |
| ID | Tên Phường | ID | Tên Phường | ID | Tên phường |
| 1 | Bạch Đằng | 1 | Đại Kim | 1 | Hạ Đình |
| 2 | Bách Khoa | 2 | Định Công | 2 | Kim Giang |
| 3 | Bạch Mai | 3 | Giáp Bát | 3 | Khương Đình |
| 4 | Bùi Thị Xuân | 4 | Hoàng Liệt | 4 | Khương Mai |
| 5 | Cầu Dền | 5 | Hoàng Văn Thụ | 5 | Khương Trung |
| … |  | … | … | … | … |
| 20 | Vĩnh Tuy | 14 | Yên Sở | 11 | Thượng Đình |

| **Phường** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Quận Hà Đông | | Quận Bắc Từ Liêm | | Quận Nam Từ Liêm | |
| ID | Tên Phường | ID | Tên Phường | ID | Tên phường |
| 1 | Biên Giang | 1 | Cổ Nhuế 1 | 1 | Cầu Diễn |
| 2 | Dương Nội | 2 | Cổ Nhuế 2 | 2 | Đại Mỗ |
| 3 | Đồng Mai | 3 | Đông Ngạc | 3 | Mễ Trì |
| 4 | Hà Cầu | 4 | Đức Thắng | 4 | Mỹ Đình 1 |
| 5 | Kiến Hưng | 5 | Liên Mạc | 5 | Mỹ Đình 2 |
| … | … | … | … | … | … |
| 17 | Yết Kiêu | 13 | Xuân Tảo | 10 | Xuân Phương |

| **Huyện** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Thị xã Sơn Tây | | Huyện Ba Vì | | Huyện Chương Mỹ | |
| ID | Tên Phường/Xã | ID | Tên Xã | ID | Tên T.trấn/Xã |
| 1 | Ngô Quyền | 1 | Phú Sơn | 1 | Xuân Mai |
| 2 | Phú Thịnh |  |  | 2 | Chúc Sơn |
| 3 | Cổ Đông |  |  | 3 | Hoàng Văn Thụ |
| 4 | Sơn Đông |  |  | 4 | Phụng Châu |

| **Huyện** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Huyện Đan Phượng | | Huyện Đông Anh | | Huyện Gia Lâm | |
| ID | Tên T.trấn/Xã | ID | Tên T.trấn/Xã | ID | Tên T.trấn/Xã |
| 1 | Phùng | 1 | Đông Anh | 1 | Trâu Quỳ |
| 2 | Đan Phượng | 2 | Bắc Hồng | 2 | Yên Viên |
| 3 | Phương Đình | 3 | Dục Tú | 3 | Cổ Bi |
| 4 | Tân Lập | 4 | Đại Mạch | 4 | Dương Quang |
| 5 | Thượng Mỗ | 5 | Đông Hội | 5 | Đa Tốn |
|  |  | … | … | … | … |
|  |  | 16 | Xuân Nộn | 10 | Yên Viên |

| **Huyện** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Huyện Hoài Đức | | Huyện Mê Linh | | Huyện Mỹ Đức | |
| ID | Tên T.trấn/Xã | ID | Tên T.trấn/Xã | ID | Tên T.trấn/Xã |
| 1 | Trạm Trôi | 1 | Quang Minh | 1 | Hợp Thanh |
| 2 | An Khánh | 2 | Đại Thịnh |  |  |
| 3 | An Thượng | 3 | Mê Linh | Huyện Phúc Thọ | |
| 4 | Di Trạch | 4 | Tam Đồng | ID | Tên T.trấn/Xã |
| 5 | Dương Liễu | 5 | Tiền Phong | 1 | Ngọc Tảo |
| … | … |  |  |  |  |
| 13 | Vân Côn |  |  |  |  |

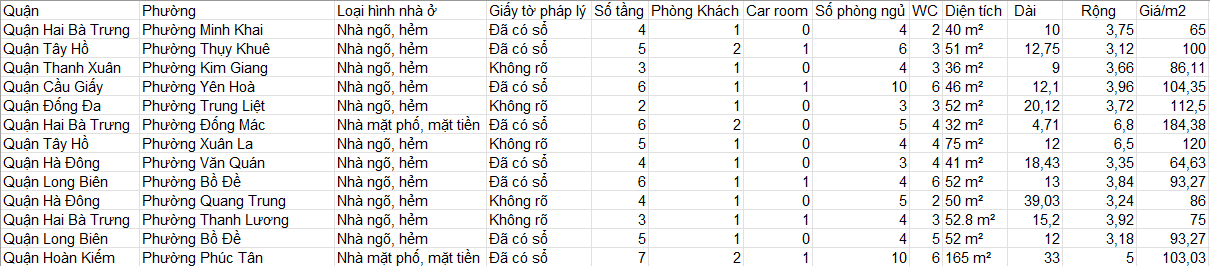
| **Huyện** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Huyện Quốc Oai | | Huyện Sóc Sơn | | Huyện Thạch Thất | |
| ID | Tên T.trấn/Xã | ID | Tên T.trấn/Xã | ID | Tên T.trấn/Xã |
| 1 | Quốc Oai | 1 | Sóc Sơn | 1 | Hương Ngải |
| 2 | Đại Thành | 2 | Minh Phú | 2 | Bình Phú |
| 3 | Đồng Quang | 3 | Phú Cường |  |  |
| 4 | Nghĩa Hương | 4 | Phú Minh |  |  |
| 5 | Ngọc Liệp | 5 | Phù Lỗ |  |  |
| 6 | Phú Cát | … | … |  |  |
| 7 | Sài Sơn | 10 | Tiên Dược |  |  |

| **Huyện** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Huyện Thanh Oai | | Huyện Thanh Trì | | Huyện Thường Tín | |
| ID | Tên T.trấn/Xã | ID | Tên T.trấn/Xã | ID | Tên T.trấn/Xã |
| 1 | Kim Bài | 1 | Văn Điển | 1 | Thường Tín |
| 2 | Bích Hòa | 2 | Đại Áng | 2 | Duyên Thái |
| 3 | Cự Khê | 3 | Đông Mỹ | 3 | Hà Hồi |
| 4 | Đỗ Đông | 4 | Hữu Hòa | 4 | Khánh Hà |
| 5 | Phương Trung | 5 | Liên Ninh | 5 | Nhị Khê |
| 6 | Tam Hưng | … | … | … | … |
|  |  | 14 | Tân Triều | 9 | Vân Tảo |

| **Loại hình nhà ở** | |
| --- | --- |
| 1 | Nhà biệt thự |
| 2 | Nhà mặt phố, mặt tiền |
| 3 | Nhà ngõ, hẻm |
| 4 | Nhà phố liền kề |
| 5 | Không rõ |

| **Giấy tờ pháp lý** | |
| --- | --- |
| 1 | Đã có sổ |
| 2 | Đang chờ sổ |
| 3 | Giấy tờ khác |
| 4 | Không rõ |

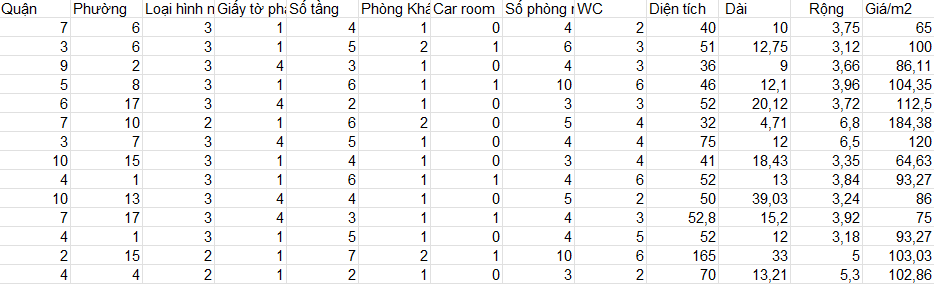
* Dữ liệu thu được sau khi chuẩn hóa dữ liệu:



* Bảng kiểu dữ liệu các thuộc tính

| **STT** | **Thuộc tính** | **Kiểu DL** | **Ví dụ** | **Mô tả** |
| --- | --- | --- | --- | --- |
| 1 | Quận | Numeric | 1,2,…29 | Địa chỉ nhà |
| 2 | Phường | Numeric | 1,2,3,... | Địa chỉ nhà |
| 3 | Loại hình nhà ở | Numeric | 1,2,3,4,5 | Kiểu nhà ở |
| 4 | Giấy tờ pháp lý | Numeric | 1,2,3,4 | Tình trạng pháp lý |
| 5 | Số tầng | Numeric | 1,2,3,... |  |
| 6 | Số phòng khách | Numeric | 1,2,3,... |  |
| 7 | Nhà để xe oto | Numeric | 0,1 |  |
| 8 | Phòng ngủ | Numeric | 1,2,3,… |  |
| 9 | WC | Numeric | 1,2,3,… | Phòng vệ sinh |
| 10 | Diện tích | Numeric | 20~400 | Diện tích ngôi nhà (m²) |
| 11 | Dài | Numeric |  | Chiều dài của căn nhà |
| 12 | Rộng | Numeric |  | Chiều rộng của nhà |
| 13 | Giá | Numeric | 4,… | Giá trên từng m² |

* Sau bước tiền xử lý dữ liệu và vecto hóa thu được như sau:

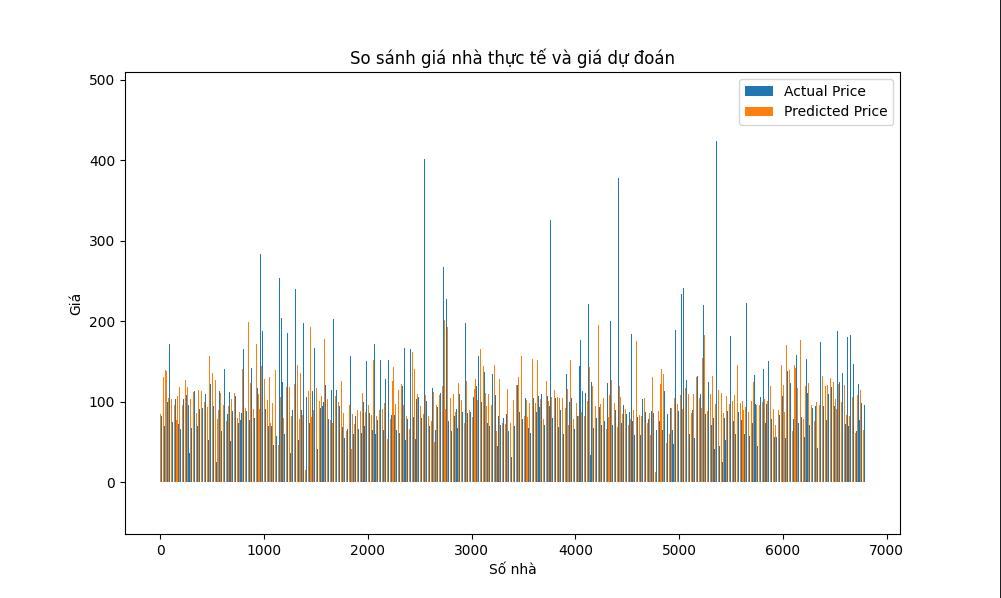
****

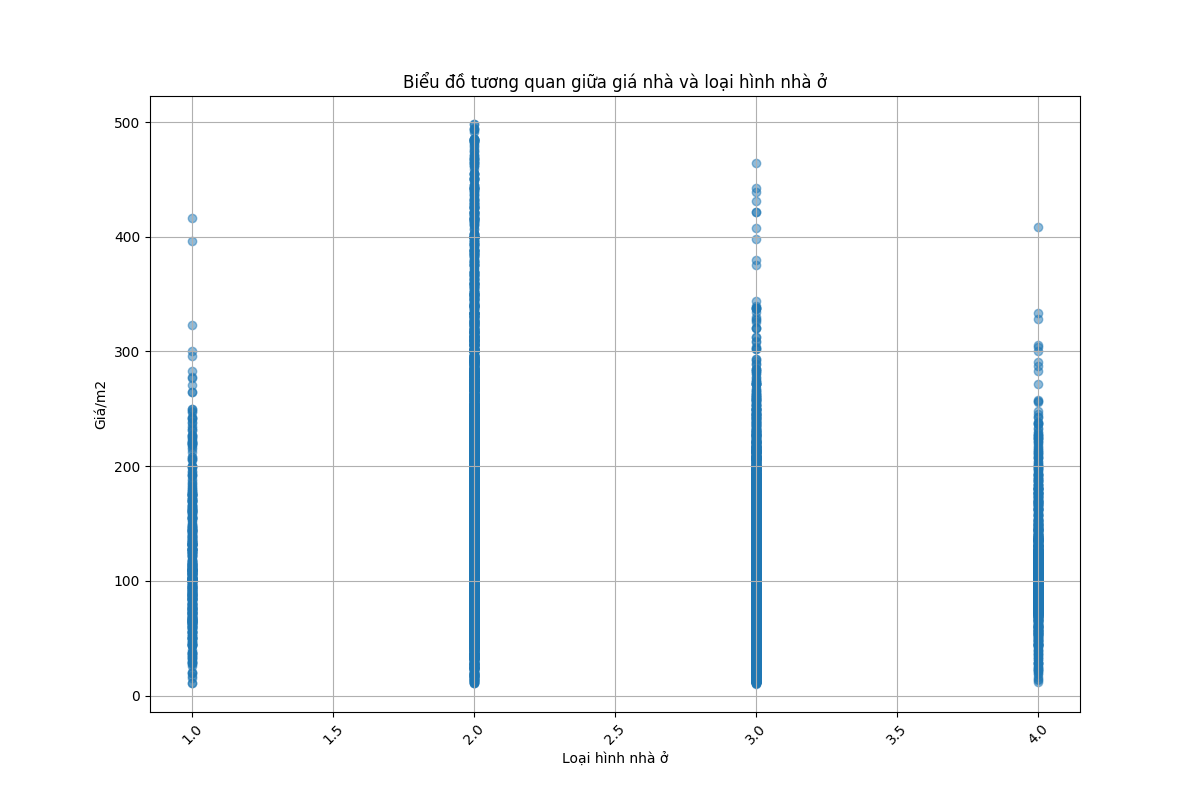
### Phân tích dữ liệu

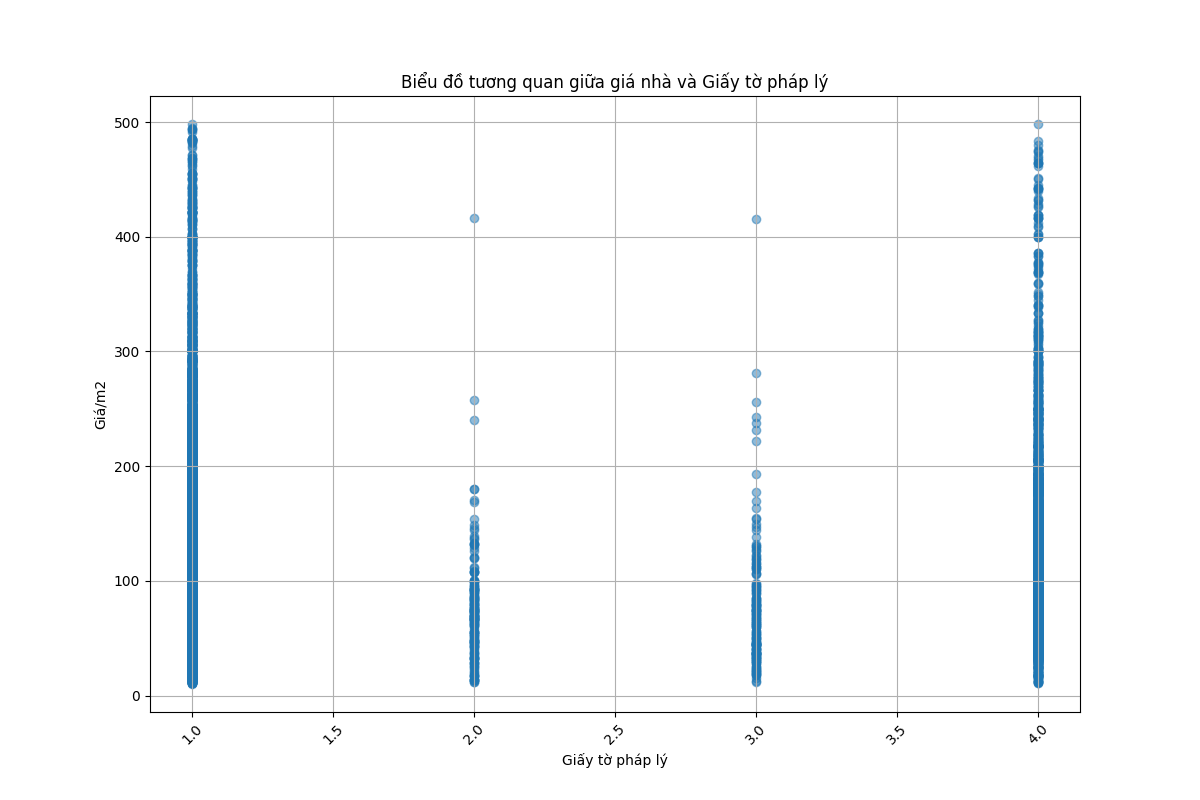
So sánh hiệu quả giữa 2 thuật toán Linear Regression và KNN (n\_neighbors = 5) trên tập training gồm 27149 dòng dữ liệu (80% của tập dữ liệu) thu được kết quả:

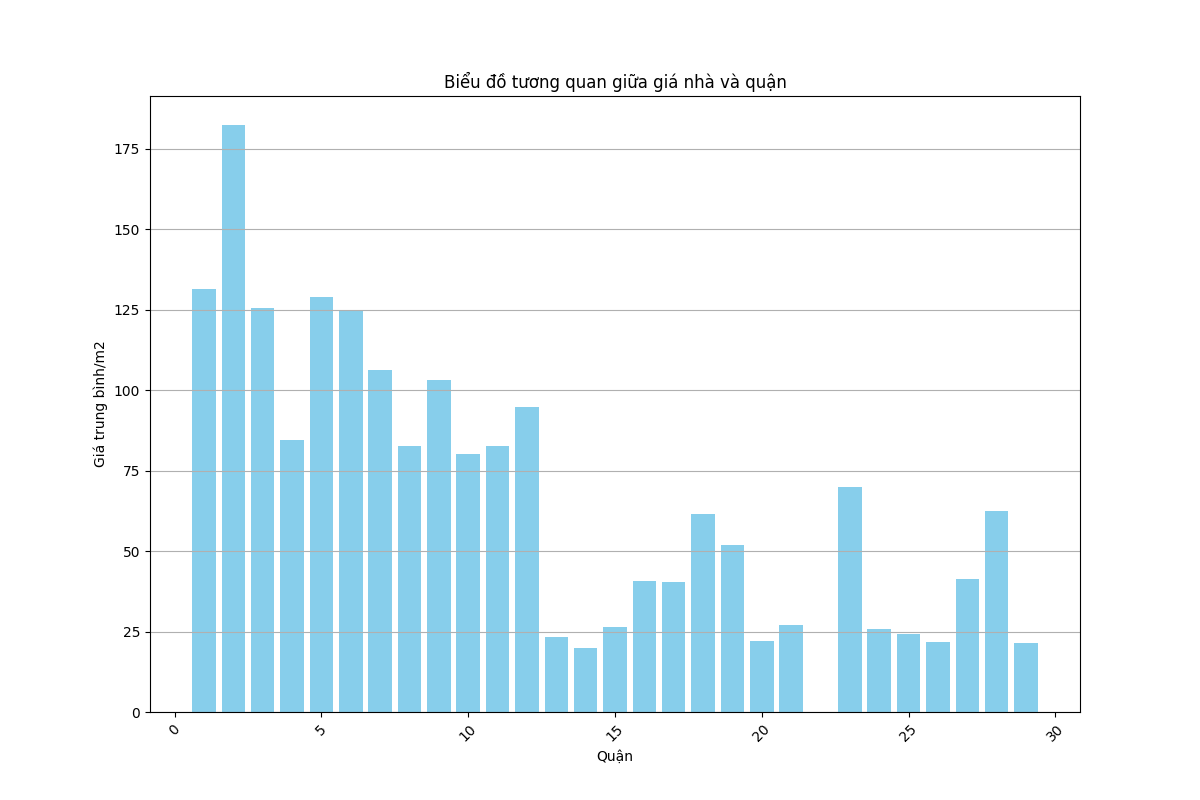
|  | Linear Regression | KNN |
| --- | --- | --- |
| Correlation coefficient | 50,49% | 46% |
| Mean absolute error | 32,268 | 16,941 |
| Mean absolute error with test | 24,021 | 26,048 |
| Time | 0.07s | 16.01s |

Linear Regression có hiệu suất tốt hơn khi có hệ số tương quan cao hơn, sai số tuyệt đối trên file test (6788 dòng dữ liệu 20% của tệp dữ liệu ban đầu) và thời gian chạy thấp hơn so với KNN. Một số đề tài dự đoán khác cũng sử dụng linear regression để thực hiện dự đoán giá nhà đất.



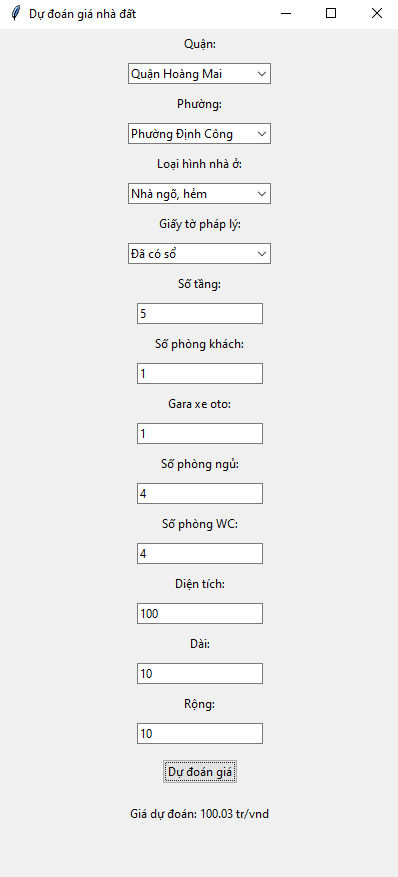






## Xây dựng ứng dụng

Để dự đoán giá nhà đất tại Hà Nội, nhóm thực hiện tạo một form cơ bản với ngôn ngữ python có phần giao diện gồm các combobox, text box đơn giản sau:



# Tài Liệu Tham Khảo

[1] Bài giảng môn Nhập môn Khai phá dữ liệu và Máy học, ThS. Vũ Xuân Hạnh, Khoa Công nghệ thông tin, Trường Đại Học Mở Hà Nội

[2] Machine Learning cơ bản, Vũ Hữu Tiệp

[3] [Vietnam Housing Dataset (Hanoi) (kaggle.com)](https://www.kaggle.com/datasets/ladcva/vietnam-housing-dataset-hanoi/data)

[4] [Bất Động Sản Hà Nội Cập Nhật Mới Tháng 04/2024 (nhatot.com)](https://www.nhatot.com/mua-ban-bat-dong-san-ha-noi)

[5] [GitHub - HungTrinhIT/FinalProject-Datascience: Đồ án cuối kì môn khoa học dữ liệu ứng dụng. Thu thập data bằng cách parsing HTML và sử dụng các mô hình học máy để giải quyết câu hỏi được đặt ra ban đầu.](https://github.com/HungTrinhIT/FinalProject-Datascience)