**BỘ GIÁO DỤC ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

**A picture containing font, graphics, logo, graphic design

Description automatically generated------------------------------**

**BÁO CÁO MÔN HỌC**

**XÁC SUẤT THỐNG KÊ VÀ PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI**

**PHÂN TÍCH DỮ LIỆU DỰ ĐOÁN GIÁ NHÀ TẠI THỦ ĐÔ HÀ NỘI**

**NHÓM SINH VIÊN : NHÓM 2**

**SINH VIÊN : 1. Trần Văn Lâm**

**2. Nguyễn Vũ Yến Nhi**

**3. Nguyễn Minh Sơn**

**4. Nguyễn Thị Yến**

**LỚP : CNTT 17-01**

**NGƯỜI HƯỚNG DẪN: ThS Phạm Thị Tố Nga**

**HÀ NỘI – 2025**

**BỘ GIÁO DỤC ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

**A picture containing font, graphics, logo, graphic design

Description automatically generated------------------------------**

**BÁO CÁO MÔN HỌC**

**XÁC SUẤT THỐNG KÊ VÀ PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI**

**PHÂN TÍCH DỮ LIỆU DỰ ĐOÁN GIÁ NHÀ TẠI THỦ ĐÔ HÀ NỘI**

**NHÓM SINH VIÊN : NHÓM 2**

**SINH VIÊN : 1. Trần Văn Lâm**

**2. Nguyễn Vũ Yến Nhi**

**3. Nguyễn Minh Sơn**

**4. Nguyễn Thị Yến**

**LỚP : CNTT 17-01**

|  |  |
| --- | --- |
| **GV CHẤM THI 1** | **GV CHẤM THI 2** |
|  |  |

**HÀ NỘI - 2025**

# PHÂN CÔNG NHIỆM VỤ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **MSV** | **Họ và tên** | **Nhiệm vụ** | **Nhận xét của GV** | **Điểm đánh giá** |
| 1 | 1771020412 | Trần Văn Lâm | Xây dựng mô hình và phân tích kết quả |  |  |
| 2 | 1771020523 | Nguyễn Vũ Yến Nhi | Làm sạch dữ liệu |  |  |
| 3 | 1771020599 | Nguyễn Minh Sơn | Thu thập dữ liệu |  |  |
| 4 | 1771020777 | Nguyễn Thị Yến | EDA |  |  |

# LỜI CAM ĐOAN

Em xin cam đoan rằng đề tài nghiên cứu "Phân tích dữ liệu dự đoán giá nhà tại Thủ đô Hà Nội" là kết quả nghiên cứu độc lập của nhóm chúng em. Đề tài được thực hiện trên cơ sở tìm hiểu lý thuyết, thu thập dữ liệu thực tế, áp dụng các phương pháp phân tích khoa học để đưa ra những kết luận có tính khách quan và chính xác.

Trong quá trình thực hiện đề tài, chúng em đã sử dụng các tài liệu, số liệu từ nhiều nguồn khác nhau, bao gồm báo cáo thị trường, dữ liệu bất động sản và các công trình nghiên cứu liên quan. Những tài liệu này đều được tham khảo một cách có chọn lọc và trích dẫn đầy đủ theo đúng quy định. Mọi nội dung trong bài đều thể hiện sự trung thực và trách nhiệm của chúng tôi đối với nghiên cứu này.

Chúng em cam kết không sao chép, đạo văn hay sử dụng trái phép bất kỳ nội dung nào từ các nguồn tài liệu khác. Nếu phát hiện có bất kỳ vi phạm nào liên quan đến tính trung thực và bản quyền của nghiên cứu, chúng em xin hoàn toàn chịu trách nhiệm trước nhà trường và các cơ quan có thẩm quyền.

Bằng lời cam đoan này, chúng em khẳng định tính chính xác, trung thực của nghiên cứu và hy vọng rằng đề tài sẽ mang lại những giá trị thực tiễn hữu ích cho việc phân tích và dự đoán giá nhà tại Hà Nội.

# LỜI CẢM ƠN

Xác suất thống kê và Phân tích dữ liệu là một môn học đại cương có tầm quan trọng đối với sinh viên nói chung và sinh viên nhóm ngành Công nghệ thông tin nói riêng. Do đó, việc dành cho môn học này một khối lượng thời gian nhất định và thực hành là điều tất yếu để giúp sinh viên có cơ sở vững chắc về kiến thức và kỹ năng cần thiết cho các môn học chuyên ngành cũng như công việc sau này. Sự phát triển và ra đời của toán tin nói chung và phần mềm Python, ngôn ngữ Python nói riêng đã hỗ trợ rất nhiều trong quá trình học tập và nghiên cứu bộ môn Xác suất thống kê và Phân tích dữ liệu. Việc phân tích và xử lý số liệu đã được rút ngắn và có hiệu quả cao hơn. Vì vậy mà việc tìm hiểu Python và ngôn ngữ Python trong việc thực hành môn học Xác suất thống kê và phân tích dữ liệu rất quan trọng và có tính cấp thiết. Ở bài tập lớn này, nhóm thực hiện nội dung: Ứng dụng hồi quy tuyến tính để xử lý và phân tích dữ liệu. Trong suốt quá trình thực hiện bài tập, nhóm đã nhận được rất nhiều sự quan tâm, ủng hộ và giúp đỡ tận tình của thầy cô và bạn bè.

Ngoài ra, nhóm cũng xin gửi lời tri ân chân thành nhất đến cô Phạm Thị Tố Nga –giảng viên giảng dạy bộ môn Xác suất thống kê và Phân tích dữ liệu của nhóm và là người hướng dẫn cho đề tài này. Nhờ sự hết lòng chỉ bảo mà nhóm đã hoàn thành bài tập đúng tiến độ và giải quyết tốt những vướng mắc gặp phải. Sự hướng dẫn của cô đã là kim chỉ nam cho mọi hành động của nhóm và phát huy được tối đa mối quan hệ hỗ trợ giữa cô và trò trong môi trường giáo dục. Lời cuối, xin một lần nữa gửi lời biết ơn sâu sắc đến các cá nhân, các thầy cô đã dành thời gian chỉ dẫn cho nhóm. Đây chính là niềm tin, là động lực to lớn để nhóm có thể hoàn thành đề tài này.

# LỜI NÓI ĐẦU

Trong bối cảnh nền kinh tế phát triển mạnh mẽ, thị trường bất động sản tại Thủ đô Hà Nội ngày càng trở nên sôi động và có nhiều biến động. Giá nhà không chỉ bị ảnh hưởng bởi các yếu tố kinh tế vĩ mô như lạm phát, lãi suất ngân hàng, chính sách quy hoạch đô thị, mà còn chịu tác động từ các yếu tố vi mô như vị trí địa lý, diện tích, tiện ích xung quanh, chất lượng cơ sở hạ tầng,… Việc dự đoán giá nhà một cách chính xác có ý nghĩa quan trọng đối với nhiều đối tượng, từ nhà đầu tư, doanh nghiệp bất động sản đến những cá nhân có nhu cầu mua bán nhà ở.

Với sự phát triển mạnh mẽ của khoa học dữ liệu và trí tuệ nhân tạo, các phương pháp phân tích dữ liệu ngày càng được ứng dụng rộng rãi trong việc dự đoán giá bất động sản. Trong đề tài "Phân tích dữ liệu dự đoán giá nhà tại Thủ đô Hà Nội", nhóm nghiên cứu sử dụng mô hình hồi quy đa biến để xây dựng mô hình dự báo giá nhà dựa trên nhiều yếu tố đầu vào khác nhau. Hồi quy đa biến là một phương pháp thống kê mạnh mẽ, cho phép đánh giá tác động của nhiều biến số độc lập lên biến phụ thuộc (giá nhà), từ đó đưa ra những dự đoán có độ chính xác cao.

Đề tài này không chỉ giúp hiểu rõ hơn về các yếu tố ảnh hưởng đến giá nhà tại Hà Nội mà còn cung cấp một công cụ hỗ trợ hữu ích cho các cá nhân và tổ chức trong việc ra quyết định liên quan đến bất động sản. Đồng thời, việc áp dụng phân tích dữ liệu vào thực tiễn cũng góp phần thúc đẩy sự phát triển của lĩnh vực khoa học dữ liệu tại Việt Nam.

# DANH MỤC VIẾT TẮT

|  |  |
| --- | --- |
| **Từ viết tắt** | **Viết đầy đủ** |
| ERD | Entity Relationship Diagram |
| NXB | Nhà Xuất Bản |
| EDA | Exploratory Data Analysis |
| MSE | Mean Squared Error |

# DANH MỤC HÌNH ẢNH

[Hình 2.1. Quy trình phân tích dữ liệu 10](#_Toc193218220)

[Hình 3.1. Dữ liệu dự đoán giá nhà ở thủ đô Hà Nội 11](#_Toc193218221)

[Hình 3.2. Dữ liệu ban đầu 12](#_Toc193218222)

[Hình 3.3. Các giá trị bị thiếu 12](#_Toc193218223)

[Hình 3.4. Các giá trị bị lỗi 13](#_Toc193218224)

[Hình 3.5. Dữ liệu sau khi làm sạch 13](#_Toc193218225)

[Hình 3.6. Các chỉ số thống kê cơ bản 14](#_Toc193218226)

[Hình 3.7. Kết quả cho biến Diện\_Tích\_Sàn 15](#_Toc193218227)

[Hình 3.8. Kết quả cho biến Số\_Phòng\_Ngủ 15](#_Toc193218228)

[Hình 3.9. Kết quả cho biến Số\_Phòng\_Tắm 16](#_Toc193218229)

[Hình 3.10. Kết quả cho biến Năm\_Xây\_Dựng 16](#_Toc193218230)

[Hình 3.11. Kết quả cho biến Diện\_Tích\_Đất 16](#_Toc193218231)

[Hình 3.12. Kết quả cho biến Số\_Chỗ\_Đậu\_Xe 17](#_Toc193218232)

[Hình 3.13. Kết quả cho biến Chất\_Lượng\_Khu\_Vực 17](#_Toc193218233)

[Hình 3.14. Kết quả cho biến Giá\_nhà 17](#_Toc193218234)

[Hình 3.15. Biểu đồ phân phối giá nhà 18](#_Toc193218235)

[Hình 3.16. Biểu đồ cột cho số phòng ngủ 19](#_Toc193218236)

[Hình 3.17. Ma trận tương quan 20](#_Toc193218237)

# MỤC LỤC

[PHÂN CÔNG NHIỆM VỤ i](#_Toc193218313)

[LỜI CAM ĐOAN ii](#_Toc193218314)

[LỜI CẢM ƠN iii](#_Toc193218315)

[LỜI NÓI ĐẦU iv](#_Toc193218316)

[DANH MỤC VIẾT TẮT v](#_Toc193218317)

[DANH MỤC HÌNH ẢNH vi](#_Toc193218318)

[MỤC LỤC vii](#_Toc193218319)

[CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI 1](#_Toc193218320)

[1.1 Giới thiệu chung về đề tài: 1](#_Toc193218321)

[1.2 Lý do chọn đề tài: 1](#_Toc193218322)

[1.3 Mục tiêu: 2](#_Toc193218323)

[1.4 Phạm vi nghiên cứu của đề tài: 2](#_Toc193218324)

[1.4.1 Dữ liệu sử dụng: 2](#_Toc193218325)

[1.4.2 Phương pháp phân tích: 2](#_Toc193218326)

[1.4.3 Phạm vi địa lý: 2](#_Toc193218327)

[1.4.4 Giới hạn nghiên cứu: 2](#_Toc193218328)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 3](#_Toc193218329)

[2.1 Trình bày về các khái niệm thống kê: 3](#_Toc193218330)

[2.1.1 Tổng thể và mẫu nghiên cứu: 3](#_Toc193218331)

[2.1.2 Các đặc trưng của mẫu và tổng thể: 5](#_Toc193218332)

[2.2 Trình bày các khái niệm về phân tích dữ liệu: 7](#_Toc193218333)

[2.2.1 Khái niệm về dữ liệu và phân tích dữ liệu: 8](#_Toc193218334)

[2.2.2 Quy trình phân tích dữ liệu: 8](#_Toc193218335)

[CHƯƠNG 3. phân tích dữ liệu dự đoán giá nhà ở thủ đô hà nội 11](#_Toc193218336)

[3.1 Thu thập dữ liệu: 11](#_Toc193218337)

[3.2 Làm sạch dữ liệu: 11](#_Toc193218338)

[3.2.1 Dữ liệu ban đầu: 12](#_Toc193218339)

[3.2.2 Kiểm tra và xử lí các giá trị bị thiếu: 12](#_Toc193218340)

[3.2.3 Kiểm tra và xử lí các giá trị bị lỗi: 13](#_Toc193218341)

[3.2.4 In ra bộ dữ liệu mới sau khi làm sạch: 13](#_Toc193218342)

[3.3 Khám phá dữ liệu: 14](#_Toc193218343)

[3.3.1 In ra các chỉ số thống kê cơ bản: 14](#_Toc193218344)

[3.3.2 In ra các chỉ số thống kê cho từng biến: 14](#_Toc193218345)

[3.3.3 Trực quan hóa phân phối giá nhà bằng biểu đồ histplot.: 18](#_Toc193218346)

[3.3.4 Thể hiện số lượng phòng ngủ bằng biểu đồ cột: 18](#_Toc193218347)

[3.3.5 Tính toán và biểu diễn mảng tương quan giữa các biến: 19](#_Toc193218348)

[3.4 Xây dựng mô hình: 20](#_Toc193218349)

[3.4.1 Chuẩn bị dữ liệu: 20](#_Toc193218350)

[3.4.2 Chia dữ liệu: 21](#_Toc193218351)

[3.4.3 Xây dựng mô hình: 21](#_Toc193218352)

[3.4.4 Đánh giá mô hình: 22](#_Toc193218353)

[3.4.5 Thử nghiệm mô hình Random Forest: 23](#_Toc193218354)

[3.5 Phân tích kết quả: 23](#_Toc193218355)

[3.6 Đưa ra kết luận và quyết định: 24](#_Toc193218356)

[CHƯƠNG 4. KẾT LUẬN 25](#_Toc193218357)

[4.1 Kết quả đạt được: 25](#_Toc193218358)

[4.1.1 Kiến thức: 25](#_Toc193218359)

[4.1.2 Sản phẩm: 25](#_Toc193218360)

[4.2 Ưu điểm và nhược điểm: 25](#_Toc193218361)

[4.2.1 Ưu điểm: 25](#_Toc193218362)

[4.2.2 Nhược điểm: 25](#_Toc193218363)

[4.3 Công việc tương lai: 26](#_Toc193218364)

[TÀI LIỆU THAM KHẢO 28](#_Toc193218365)

# GIỚI THIỆU ĐỀ TÀI

## Giới thiệu chung về đề tài:

Đề tài "Phân tích dữ liệu dự đoán giá nhà ở Thủ đô Hà Nội" sử dụng thuật toán Hồi quy tuyến tính nhằm nghiên cứu các yếu tố ảnh hưởng đến giá bất động sản và xây dựng mô hình dự đoán giá nhà dựa trên dữ liệu thực tế. Thị trường bất động sản Hà Nội có sự biến động lớn do nhiều yếu tố như vị trí địa lý, diện tích, số phòng ngủ, pháp lý và các đặc điểm khác. Vì vậy, việc áp dụng mô hình phân tích dữ liệu để dự đoán giá nhà có thể mang lại giá trị thực tiễn cao, giúp người mua, người bán và các nhà đầu tư đưa ra quyết định chính xác hơn.

Dữ liệu được sử dụng trong nghiên cứu này được thu thập từ thị trường bất động sản Hà Nội, bao gồm các thông tin quan trọng như địa chỉ, quận/huyện, loại hình nhà ở, số tầng, số phòng ngủ, diện tích, kích thước dài - rộng, tình trạng pháp lý và giá/m². Thuật toán Hồi quy tuyến tính sẽ được áp dụng để xác định mối quan hệ giữa các biến số này với giá nhà, từ đó tạo ra mô hình dự đoán hợp lý.

## Lý do chọn đề tài:

*Tính thực tiễn cao:* Dự đoán giá nhà là vấn đề quan trọng đối với nhiều đối tượng, từ người mua, người bán cho đến các nhà đầu tư bất động sản. Một mô hình dự đoán chính xác sẽ giúp họ đưa ra quyết định hợp lý hơn.

*Ứng dụng khoa học dữ liệu vào thực tế:* Hồi quy tuyến tính là một phương pháp phổ biến, đơn giản nhưng hiệu quả trong dự đoán giá trị liên tục. Việc áp dụng thuật toán này vào phân tích dữ liệu bất động sản giúp minh chứng khả năng ứng dụng của khoa học dữ liệu trong đời sống.

*Dữ liệu phong phú và đáng tin cậy:* Bộ dữ liệu thu thập từ thị trường bất động sản Hà Nội có quy mô lớn, bao gồm nhiều yếu tố quan trọng ảnh hưởng đến giá nhà, tạo điều kiện thuận lợi cho việc xây dựng mô hình dự đoán chính xác.

*Hỗ trợ ra quyết định đầu tư*: Với một mô hình dự đoán tốt, người mua có thể xác định giá trị hợp lý của một bất động sản trước khi giao dịch, trong khi nhà đầu tư có thể đánh giá xu hướng thị trường để đưa ra chiến lược phù hợp.

## Mục tiêu:

* Xây dựng mô hình dự đoán giá nhà ở Hà Nội dựa trên thuật toán Hồi quy tuyến tính.
* Xác định các yếu tố quan trọng ảnh hưởng đến giá bất động sản.
* Đánh giá độ chính xác của mô hình và cải thiện bằng cách điều chỉnh các biến đầu vào.
* Ứng dụng mô hình vào thực tế nhằm hỗ trợ người mua, người bán và nhà đầu tư ra quyết định hợp lý.

## Phạm vi nghiên cứu của đề tài:

### Dữ liệu sử dụng:

Bộ dữ liệu được thu thập từ thị trường bất động sản Hà Nội, bao gồm các thông tin như ngày đăng tin, địa chỉ, quận/huyện, loại hình nhà ở, giấy tờ pháp lý, số tầng, số phòng ngủ, diện tích, chiều dài, chiều rộng và giá/m².

Dữ liệu gốc: [kaggle.com](https://www.kaggle.com/datasets/ladcva/vietnam-housing-dataset-hanoi?utm_source=chatgpt.com)

### Phương pháp phân tích:

Sử dụng thuật toán Hồi quy tuyến tính để xây dựng mô hình dự đoán giá nhà dựa trên các biến số trong bộ dữ liệu.

### Phạm vi địa lý:

Tập trung vào các bất động sản trong khu vực Hà Nội, Việt Nam.

### Giới hạn nghiên cứu:

Mô hình chỉ xem xét các yếu tố nội tại của bất động sản như vị trí, diện tích, số phòng ngủ, loại hình nhà ở, giấy tờ pháp lý, số tầng, chiều dài, chiều rộng và giá/m². Các yếu tố bên ngoài như tình hình kinh tế, lãi suất ngân hàng, cơ sở hạ tầng xung quanh, tiện ích công cộng và xu hướng thị trường chung không được đưa vào phân tích trong phạm vi của đề tài này.

# CƠ SỞ LÝ THUYẾT

## Trình bày về các khái niệm thống kê:

Thống kê là ngành khoa học chuyên thu thập, tổ chức, phân tích và diễn giải dữ liệu. Thống kê giúp con người hiểu rõ hơn về dữ liệu, rút ra kết luận và đưa ra quyết định dựa trên thông tin thu thập được.[1]

Thống kê có thể chia thành hai nhánh chính:

* Thống kê mô tả (Descriptive Statistics)
* Dùng để tóm tắt và trình bày dữ liệu theo cách dễ hiểu.
* Bao gồm các phương pháp như bảng tần số, biểu đồ, trung bình, phương sai, độ lệch chuẩn, v.v.
* Thống kê suy diễn (Inferential Statistics)
* Dựa trên dữ liệu mẫu để đưa ra suy luận về tổng thể.
* Sử dụng các phương pháp như kiểm định giả thuyết, ước lượng tham số, hồi quy, v.v.

### Tổng thể và mẫu nghiên cứu:

#### Tổng thể:

Tổng thể là tập hợp tất cả các đối tượng hoặc phần tử có chung một đặc điểm nào đó mà ta quan tâm nghiên cứu.[1]

Phân loại tổng thể:

* Tổng thể hữu hạn (Finite Population): Có số phần tử đếm được, ví dụ:
* Số lượng sinh viên trong một trường đại học.
* Số lượng xe ô tô đang lưu thông trong một thành phố.
* Tổng thể vô hạn (Infinite Population): Không thể xác định số lượng chính xác do số phần tử quá lớn hoặc liên tục thay đổi, ví dụ:
* Tất cả khách hàng có thể mua sản phẩm của một công ty trong tương lai.
* Số vi khuẩn trong một giọt nước biển.

#### Mẫu nghiên cứu:

Mẫu là một tập hợp con của tổng thể được chọn để nghiên cứu. Vì việc khảo sát toàn bộ tổng thể thường không khả thi (do tốn kém thời gian và chi phí), ta cần lấy mẫu để phân tích và suy luận về tổng thể.[2]

*Ví dụ:*

* Một nhóm 500 sinh viên được chọn từ tổng số 20.000 sinh viên trong một trường đại học.
* 100 sản phẩm được chọn ngẫu nhiên từ tổng số 10.000 sản phẩm để kiểm tra chất lượng.

*Phương pháp chọn mẫu phổ biến:*

* Lấy mẫu ngẫu nhiên đơn giản (Simple Random Sampling):
* Chọn ngẫu nhiên các phần tử từ tổng thể sao cho mỗi phần tử có cơ hội như nhau để được chọn.
* Lấy mẫu hệ thống (Systematic Sampling):
* Chọn các phần tử theo một quy tắc nhất định, ví dụ chọn mỗi phần tử thứ kkk trong danh sách.
* Ví dụ: Chọn mỗi sinh viên thứ 10 trong danh sách để khảo sát.
* Lấy mẫu phân tầng (Stratified Sampling):
* Chia tổng thể thành các nhóm (tầng) theo đặc điểm nào đó, sau đó chọn mẫu từ từng tầng theo tỉ lệ nhất định.
* Ví dụ: Chia sinh viên theo năm học (năm nhất, năm hai, năm ba, năm tư) rồi chọn mẫu từ mỗi nhóm.
* Lấy mẫu cụm (Cluster Sampling):
* Chia tổng thể thành các nhóm nhỏ (cụm), sau đó chọn ngẫu nhiên một số cụm để khảo sát toàn bộ phần tử trong đó.
* Ví dụ: Chia toàn bộ học sinh trong một tỉnh thành các trường, sau đó chọn ngẫu nhiên một số trường để khảo sát toàn bộ học sinh trong trường đó.

### Các đặc trưng của mẫu và tổng thể:

#### Trung bình tổng thể:

Trung bình tổng thể là giá trị đại diện cho toàn bộ dữ liệu, tính bằng cách lấy tổng tất cả các giá trị chia cho số lượng phần tử.

* Trung bình của tổng thể (μ\muμ):

A mathematical equation with black text

AI-generated content may be incorrect.

Trong đó:

* Xi là các giá trị trong tổng thể.
* N là số lượng phần tử trong tổng thể.

Trung bình của mẫu ():

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.

Trong đó:

* n là số lượng phần tử trong mẫu.

Ví dụ:

Giả sử có một mẫu gồm 5 học sinh với điểm số: 7, 8, 9, 10, 6.

Trung bình mẫu là:

A number line with black numbers

AI-generated content may be incorrect.

#### Số trung vị (Median)

Trung vị là giá trị nằm ở giữa của tập dữ liệu sau khi đã sắp xếp theo thứ tự tăng dần.

* Nếu số phần tử lẻ, trung vị là phần tử ở vị trí giữa.
* Nếu số phần tử chẵn, trung vị là trung bình của hai phần tử giữa.

Ví dụ:

* Dữ liệu: 6, 7, 8, 9, 10 → Trung vị = 8 (phần tử giữa).

#### Mốt (Mode):

Mốt là giá trị xuất hiện nhiều nhất trong tập dữ liệu.

Ví dụ:

* Dữ liệu: 6, 7, 7, 8, 9, 10 → Mốt = 7 (xuất hiện 2 lần).
* Dữ liệu: 6 ,7 ,8 , 9, 10 → Không có mốt.
* Dữ liệu: 6, 7, 7, 8, 8, 9, 10 → Có hai mốt là 7 và 8 (dữ liệu đa mốt).

#### Phương sai (Variance):

Phương sai đo mức độ phân tán của dữ liệu so với trung bình.

* Phương sai tổng thể (σ2):

A black and white math equation

AI-generated content may be incorrect.

* Phương sai mẫu (S2):

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.

#### Độ lệch chuẩn (Standard Deviation):

Phương sai đo mức độ phân tán của dữ liệu so với trung bình.

* Phương sai tổng thể (σ2):

A black and white math equation

AI-generated content may be incorrect.

* Phương sai mẫu (S2):

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.

#### Độ lệch chuẩn (Standard Deviation):

Độ lệch chuẩn là căn bậc hai của phương sai, đo mức độ phân tán dữ liệu

A group of black and white symbols

AI-generated content may be incorrect.

Với ví dụ trên:



#### Biên độ (Range):

Biên độ đo khoảng cách giữa giá trị lớn nhất và nhỏ nhất của dữ liệu.

Range=Max−Min

Ví dụ:

Dữ liệu: 6, 7, 8, 9, 10 → Biên độ = 10 −6 = 4.

#### Độ lệch (Skewness):

Độ lệch đo mức độ bất đối xứng của dữ liệu.

* Skewness > 0: Phân phối lệch phải.
* Skewness < 0: Phân phối lệch trái.
* Skewness = 0: Phân phối đối xứng.

## Trình bày các khái niệm về phân tích dữ liệu:

Phân tích dữ liệu là quá trình làm việc với dữ liệu để tìm ra thông tin hữu ích, hỗ trợ ra quyết định. Theo Montgomery & Runger (2010), quá trình này bao gồm việc thu thập, làm sạch, khám phá dữ liệu (Exploratory Data Analysis - EDA), xây dựng mô hình và phân tích kết quả.[3]

### Khái niệm về dữ liệu và phân tích dữ liệu:

#### Khái niệm về dữ liệu:

Dữ liệu là tập hợp các giá trị được thu thập từ thực nghiệm, khảo sát hoặc đo lường. Theo Triola (2017), dữ liệu có thể chia thành:[3]

* Dữ liệu định lượng (Quantitative Data): Dữ liệu có thể đo lường, như cân nặng, điểm số, doanh thu.
* Dữ liệu định tính (Qualitative Data): Dữ liệu mô tả đặc điểm, như màu sắc, loại sản phẩm.

#### Khái niệm về phân tích dữ liệu:

Phân tích dữ liệu là quá trình xử lý và diễn giải dữ liệu để rút ra thông tin hữu ích. Theo Wackerly et al. (2014), phân tích dữ liệu gồm:

* Thống kê mô tả: Tóm tắt dữ liệu bằng bảng, biểu đồ và các đặc trưng trung tâm.
* Thống kê suy diễn: Dự đoán hoặc kiểm định giả thuyết từ dữ liệu mẫu.

### Quy trình phân tích dữ liệu:

Theo Montgomery & Runger (2010), quy trình phân tích dữ liệu gồm các bước sau:

**Bước 1: Thu thập dữ liệu**

Trước khi phân tích, cần thu thập dữ liệu từ các nguồn đáng tin cậy. Theo Triola (2017), có hai phương pháp chính:

* Dữ liệu sơ cấp: Thu thập trực tiếp từ khảo sát, thí nghiệm hoặc cảm biến.
* Dữ liệu thứ cấp: Dữ liệu có sẵn từ báo cáo, nghiên cứu trước đó.

**Ví dụ:** Khi nghiên cứu sự ảnh hưởng của thời gian học đến điểm số, dữ liệu có thể thu thập từ khảo sát sinh viên hoặc điểm thi từ hệ thống trường học.

**Bước 2: Làm sạch dữ liệu**

Dữ liệu thường chứa lỗi hoặc giá trị thiếu, cần xử lý trước khi phân tích. Theo Wackerly et al. (2014), các bước làm sạch gồm:

* Loại bỏ giá trị trùng lặp: Xóa bản ghi bị lặp lại.
* Xử lý dữ liệu bị thiếu: Điền giá trị trung bình, sử dụng mô hình dự đoán hoặc loại bỏ dữ liệu lỗi.
* Chuẩn hóa dữ liệu: Đưa về cùng đơn vị đo hoặc định dạng phù hợp.

**Ví dụ:** Nếu bộ dữ liệu điểm số của sinh viên có một số giá trị bị thiếu, có thể dùng trung bình cộng để thay thế.

**Bước 3: Khám phá dữ liệu (Exploratory Data Analysis - EDA)**

EDA là bước quan trọng giúp hiểu rõ cấu trúc dữ liệu trước khi xây dựng mô hình. Theo Montgomery & Runger (2010), EDA bao gồm:

* Thống kê mô tả: Tính toán các giá trị trung bình, phương sai, độ lệch chuẩn.
* Trực quan hóa dữ liệu: Dùng biểu đồ (histogram, scatter plot, boxplot) để tìm hiểu xu hướng và bất thường trong dữ liệu.

**Ví dụ:** Khi nghiên cứu điểm số sinh viên, có thể vẽ biểu đồ histogram để xem phân bố điểm có bình thường hay không.

**Bước 4: Xây dựng mô hình phân tích**

Sau khi hiểu dữ liệu, ta có thể áp dụng mô hình thống kê để tìm ra mối quan hệ giữa các biến. Theo Wackerly et al. (2014), các mô hình phổ biến gồm:

* + Hồi quy tuyến tính (Linear Regression): Dự đoán giá trị dựa trên biến đầu vào.
  + Phân tích phương sai (ANOVA): Kiểm định sự khác biệt giữa các nhóm dữ liệu.
  + Kiểm định giả thuyết (Hypothesis Testing): Đánh giá xem một giả thuyết có hợp lý hay không.

**Ví dụ:** Nếu muốn kiểm tra thời gian học có ảnh hưởng đến điểm số, ta có thể sử dụng hồi quy tuyến tính để phân tích.

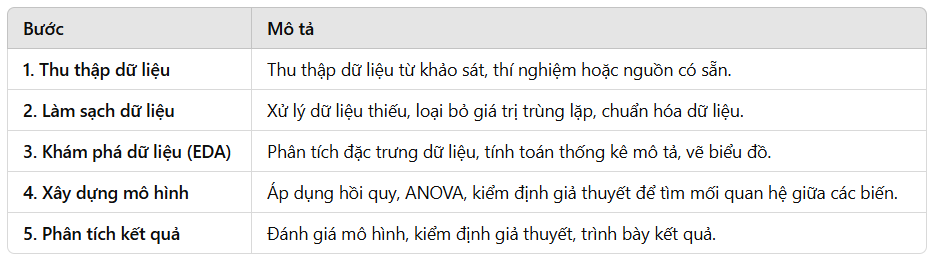
**Bước 5: Phân tích kết quả và đưa ra kết luận**

Sau khi chạy mô hình, cần diễn giải kết quả để đưa ra quyết định. Theo Triola (2017), bước này gồm:

* + Kiểm tra độ chính xác của mô hình: Dùng R² hoặc p-value để đánh giá mô hình có ý nghĩa thống kê hay không.
  + So sánh với giả thuyết ban đầu: Xem liệu kết quả có phù hợp với kỳ vọng không.
  + Báo cáo và trình bày kết quả: Tóm tắt phát hiện dưới dạng bảng, biểu đồ hoặc văn bản.

**Ví dụ:** Nếu kết quả hồi quy tuyến tính cho thấy thời gian học có mối quan hệ mạnh với điểm số (p-value < 0.05), ta có thể kết luận rằng học lâu hơn giúp cải thiện điểm.

Tóm tắt quy trình phân tích dữ liệu:



Hình 2.1. Quy trình phân tích dữ liệu

# phân tích dữ liệu dự đoán giá nhà ở thủ đô hà nội

## Thu thập dữ liệu:

A screenshot of a computer

AI-generated content may be incorrect.

Hình 3.1. Dữ liệu dự đoán giá nhà ở thủ đô Hà Nội

Bộ dữ liệu được thu thập từ thị trường bất động sản Hà Nội, bao gồm các thông tin như ngày đăng tin, địa chỉ, quận/huyện, loại hình nhà ở, giấy tờ pháp lý, số tầng, số phòng ngủ, diện tích, chiều dài, chiều rộng và giá/m².

Dữ liệu gốc: [kaggle.com](https://www.kaggle.com/datasets/ladcva/vietnam-housing-dataset-hanoi?utm_source=chatgpt.com)

## Làm sạch dữ liệu:

Việc làm sạch dữ liệu là bước quan trọng trong quy trình xử lý dữ liệu, giúp loại bỏ các giá trị thiếu, sai lệch hoặc không hợp lệ. Đầu tiên, dữ liệu được đọc từ file csv và sau đó in ra dữ liệu file dữ liệu ban đầu. Sau đó thực hiện các bước để phát hiện lỗi và xử lí lỗi bị thiếu, bị sai định dạng,…

### Dữ liệu ban đầu:

A screenshot of a computer

AI-generated content may be incorrect.

Hình 3.2. Dữ liệu ban đầu

### Kiểm tra và xử lí các giá trị bị thiếu:

# Kiểm tra các giá trị thiếu

print("\nCác giá trị thiếu trong dữ liệu:")

print(data.isnull().sum())

# Xử lý các giá trị thiếu (ví dụ: điền bằng giá trị trung bình)

data.fillna(data.mean(), inplace=True)

Để kiểm tra các giá trị bị thiếu sử dụng data.isnull()để trả về một DataFrame mới cùng kích thước với Data, nếu một ô có giá trị là NaN thì sẽ hiển thị là True, nếu không có giá trị thiếu thì sẽ hiển thị False.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 3.3. Các giá trị bị thiếu

Để xử lí giá trị bị thiếu dùng data.mean()tính trung bình của từng cột, điền các giá trị NaN bằng giá trị của từng cột tương ứng .

### Kiểm tra và xử lí các giá trị bị lỗi:

# Kiểm tra các giá trị lỗi (ví dụ: giá trị âm không hợp lệ)

print("\nCác giá trị lỗi trong dữ liệu:")

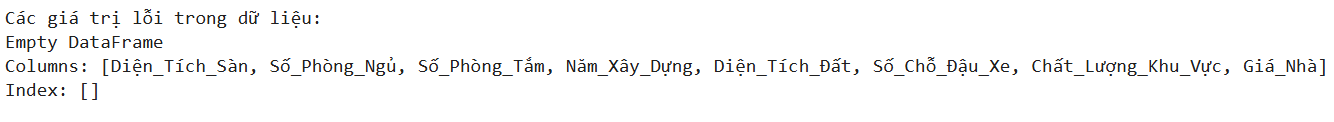
print(data[(data < 0).any(axis=1)])

# Xử lý các giá trị lỗi (ví dụ: thay thế bằng giá trị trung bình)

data[data < 0] = np.nan

data.fillna(data.mean(), inplace=True)

Để kiểm tra các giá trị bị lỗi cần có điều kiện data < 0 để kiểm tra xem giá trị nào trong data nhỏ hơn 0. Kết quả trả về True nếu giá trị nhỏ hơn 0, False nếu giá trị lớn hơn hoặc bằng 0.



Hình 3.4. Các giá trị bị lỗi

data[data < 0] = np.nan: Thay thế tất cả các giá trị âm bằng NaN và tính giá trị trung bình của từng cột, điền tất cả các ô NaN trong data bằng giá trị trung bình của cột tương ứng.

### In ra bộ dữ liệu mới sau khi làm sạch:

# In ra bộ dữ liệu mới sau khi đã làm sạch

print("\nDữ liệu sau khi làm sạch:")

print(data.head())

A screenshot of a computer screen

AI-generated content may be incorrect.

Hình 3.5. Dữ liệu sau khi làm sạch

## Khám phá dữ liệu:

### In ra các chỉ số thống kê cơ bản:

# In ra các chỉ số thống kê cơ bản

print("\nCác chỉ số thống kê cơ bản:")

print(data.describe())

Khi gọi “(data.describe())” Pandas sẽ tính toán các thống kê sau cho từng cột dữ liệu số:

* count: Số lượng giá trị hợp lệ (không bị thiếu NaN).
* mean: Giá trị trung bình của cột.
* std: Độ lệch chuẩn.
* min: Giá trị nhỏ nhất.
* 25% (Quartile 1 - Q1): Phân vị thứ nhất (dưới 25% dữ liệu).
* 50% (Quartile 2 - Q2 / Median): Trung vị (giá trị chính giữa của dữ liệu).
* 75% (Quartile 3 - Q3): Phân vị thứ ba (dưới 75% dữ liệu).
* max: Giá trị lớn nhất.

Kết quả các chỉ số thống kê cơ bản:

A screenshot of a computer code

AI-generated content may be incorrect.

Hình 3.6. Các chỉ số thống kê cơ bản

### In ra các chỉ số thống kê cho từng biến:

# In ra các chỉ số thống kê cho từng biến

for column in data.columns:

    print(f"\nThống kê cho biến {column}:")

    print(data[column].describe())

Vòng lặp for sẽ duyệt từng cột và thực hiện: In ra tiêu đề "Thống kê cho biến {column}", {column} là tên cột.

Gọi data[column].describe() để hiển thị thống kê chỉ cho cột đó.

#### Thống kê cho biến Diện\_Tích\_Sàn:

A screenshot of a computer

AI-generated content may be incorrect.

Hình 3.7. Kết quả cho biến Diện\_Tích\_Sàn

#### Thống kê cho biến Số\_Phòng\_Ngủ:

A white background with black text

AI-generated content may be incorrect.

Hình 3.8. Kết quả cho biến Số\_Phòng\_Ngủ

#### Thống kê cho biến Số\_Phòng\_Tắm:

A white background with black text

AI-generated content may be incorrect.

Hình 3.9. Kết quả cho biến Số\_Phòng\_Tắm

#### Thống kê cho biến Năm\_Xây\_Dựng:

A screenshot of a computer code

AI-generated content may be incorrect.

Hình 3.10. Kết quả cho biến Năm\_Xây\_Dựng

#### Thống kê cho biến Diện\_Tích\_Đất:

A white background with black text

AI-generated content may be incorrect.

Hình 3.11. Kết quả cho biến Diện\_Tích\_Đất

#### Thống kê cho biến Số\_Chỗ\_Đậu\_Xe:

A white background with black text

AI-generated content may be incorrect.

Hình 3.12. Kết quả cho biến Số\_Chỗ\_Đậu\_Xe

#### Thống kê cho biến Chất\_Lượng\_Khu\_Vực:

A screenshot of a computer code

AI-generated content may be incorrect.

Hình 3.13. Kết quả cho biến Chất\_Lượng\_Khu\_Vực

#### Thống kê cho biến Giá\_Nhà:

A white background with black text

AI-generated content may be incorrect.

Hình 3.14. Kết quả cho biến Giá\_nhà

### Trực quan hóa phân phối giá nhà bằng biểu đồ histplot.:

# Vẽ biểu đồ phân phối của giá nhà

plt.figure(figsize=(10, 6))

sns.histplot(data['Giá\_Nhà'], kde=True)

plt.title('Phân phối giá nhà')

plt.xlabel('Giá nhà')

plt.ylabel('Tần suất')

plt.show()

* sns.histplot(): Vẽ biểu đồ histogram (biểu đồ tần suất) dựa trên dữ liệu giá nhà.
* data['Giá\_Nhà']: Cột Giá\_Nhà từ tập dữ liệu data, chứa giá nhà của các căn hộ.
* kde=True: Thêm đường KDE (Kernel Density Estimation) – một đường cong mượt thể hiện phân phối của dữ liệu.

Hiển thị biểu đồ:

A graph with blue lines

AI-generated content may be incorrect.

Hình 3.15. Biểu đồ phân phối giá nhà

### Thể hiện số lượng phòng ngủ bằng biểu đồ cột:

# Vẽ biểu đồ cột cho số phòng ngủ

plt.figure(figsize=(10, 6))

sns.countplot(x='Số\_Phòng\_Ngủ', data=data)

plt.title('Số lượng phòng ngủ')

plt.xlabel('Số phòng ngủ')

plt.ylabel('Số lượng')

plt.show()

“sns.countplot(x='Số\_Phòng\_Ngủ', data=data)”

* *sns.countplot(...)* là hàm của thư viện seaborn dùng để vẽ biểu đồ cột (count plot).
* *x='Số\_Phòng\_Ngủ'* chỉ định cột trong data dùng làm trục x (số phòng ngủ).
* *data=data* chỉ định dataframe chứa dữ liệu

Hiển thị kết quả:

A graph with blue rectangular bars

AI-generated content may be incorrect.

Hình 3.16. Biểu đồ cột cho số phòng ngủ

### Tính toán và biểu diễn mảng tương quan giữa các biến:

# Vẽ biểu đồ tương quan giữa các biến

plt.figure(figsize=(12, 8))

sns.heatmap(data.corr(), annot=True, cmap='coolwarm')

plt.title('Ma trận tương quan')

plt.show()

*“sns.heatmap(data.corr(), annot=True, cmap='coolwarm')”*

* sns.heatmap(): Hàm này trong thư viện Seaborn dùng để vẽ bản đồ nhiệt (heatmap), thường sử dụng để biểu diễn ma trận tương quan.
* data.corr() tính toán ma trận tương quan giữa các cột số trong data. Hệ số tương quan có giá trị từ -1 đến 1:
* Gần 1: Hai biến có mối quan hệ tuyến tính dương mạnh (khi một biến tăng, biến kia cũng tăng).
* Gần -1: Hai biến có mối quan hệ tuyến tính âm mạnh (khi một biến tăng, biến kia giảm).
* Gần 0: Hai biến không có mối quan hệ tương quan tuyến tính đáng kể.
* annot=True: Hiển thị giá trị số trên mỗi ô của heatmap.
* cmap='coolwarm': Chỉ định bảng màu cho heatmap. coolwarm là một bảng màu có màu xanh dương cho giá trị âm, màu trắng cho giá trị gần 0 và màu đỏ cho giá trị dương.

Hiển thị kết quả:

A screenshot of a graph

AI-generated content may be incorrect.

Hình 3.17. Ma trận tương quan

## Xây dựng mô hình:

### Chuẩn bị dữ liệu:

X = data.drop('Giá\_Nhà', axis=1)

y = data['Giá\_Nhà']

* X chứa tất cả các đặc trưng (features), dùng để dự đoán.
* y chứa giá trị mục tiêu (Giá\_Nhà), là biến cần dự đoán.

Hàm .drop() được dùng để loại bỏ cột trong Giá\_Nhà, axis=1 là loại bỏ cột.

data['Giá\_Nhà']: Trích xuất cột 'Giá\_Nhà' từ DataFrame data. Y chứa toàn bộ giá trị của cột Giá\_Nhà

# Chuẩn hóa dữ liệu

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

* Tạo một đối tượng StandardScaler, dùng để tính toán trung bình và độ lệch chuẩn của dữ liệu.
* .fit\_transform(X):
* .fit(X): Tính toán giá trị trung bình (μ) và độ lệch chuẩn (σ) của từng cột trong X.
* .transform(X): Sử dụng các giá trị μ và σ đã tính được để chuẩn hóa dữ liệu.
* Kết quả: X\_scaled là một mảng NumPy chứa các giá trị của X đã được chuẩn hóa.

### Chia dữ liệu:

# Chia dữ liệu thành tập huấn luyện và tập kiểm tra

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_scaled, y, test\_size=0.2, random\_state=42)

* Hàm train\_test\_split() dùng để chia dữ liệu thành hai phần:
* Tập huấn luyện (X\_train, y\_train): Dùng để huấn luyện mô hình.
* Tập kiểm tra (X\_test, y\_test): Dùng để kiểm tra hiệu suất của mô hình sau khi huấn luyện.
* Tham số:
* X\_scaled: Đây là tập dữ liệu đầu vào (features) đã được chuẩn hóa.
* y là giá trị cần dự đoán.
* test\_size=0.2: Xác định tỷ lệ dữ liệu dành cho tập kiểm tra.
* random\_state=42: đảm bảo rằng mỗi lần chạy code, dữ liệu được chia theo cùng một cách.

### Xây dựng mô hình:

#### Xác định đầu vào và đầu ra của mô hình:

Trong bài toán dự đoán giá nhà tại thủ đô Hà Nội, cần xác định:

* Biến đầu ra (Y – Dependent Variable): Giá/m2 (đơn vị: triệu/m2).
* Biến đầu vào (X – Independent Variables): Các yếu tố có thể ảnh hưởng đến giá nhà, bao gồm:
* Diện tích (m2)
* Số phòng ngủ
* Số tầng
* Chiều dài và chiều rộng
* Quận (biến phân loại, cần mã hóa)
* Loại hình nhà ở (biến phân loại. cần mã hóa)

#### Xây dựng mô hình hồi quy tuyến tính:

# Xây dựng mô hình hồi quy tuyến tính

model = LinearRegression()

model.fit(X\_train, y\_train)

# Dự đoán trên tập kiểm tra

y\_pred = model.predict(X\_test)

.fit(X\_train, y\_train) giúp mô hình học từ dữ liệu huấn luyện (X\_train, y\_train).

.predict(X\_test) dùng để dự đoán giá trị y (biến phụ thuộc) dựa trên X\_test (tập kiểm tra).

### Đánh giá mô hình:

# Đánh giá mô hình

mse = mean\_squared\_error(y\_test, y\_pred)

r2 = r2\_score(y\_test, y\_pred)

print(f"\nMean Squared Error: {mse}")

print(f"R-squared: {r2}")

Hàm mean\_squared\_error: Tính trung bình bình phương sai số (MSE) – một chỉ số đo lường mức độ sai lệch giữa giá trị thực tế (y\_test) và giá trị dự đoán (y\_pred).

* Đánh giá sai số trung bình bình phương, càng nhỏ càng tốt.

Hàm r2\_score: Tính hệ số xác định R², đo lường mức độ phù hợp của mô hình với dữ liệu.

* Đánh giá độ phù hợp của mô hình, càng gần 1 càng tốt.

### Thử nghiệm mô hình Random Forest:

# Thử nghiệm mô hình Random Forest

rf\_model = RandomForestRegressor(n\_estimators=100, random\_state=42)

rf\_model.fit(X\_train, y\_train)

y\_pred\_rf = rf\_model.predict(X\_test)

# Đánh giá mô hình Random Forest

mse\_rf = mean\_squared\_error(y\_test, y\_pred\_rf)

r2\_rf = r2\_score(y\_test, y\_pred\_rf)

print(f"\nRandom Forest - Mean Squared Error: {mse\_rf}")

print(f"Random Forest - R-squared: {r2\_rf}")

* RandomForestRegressor: Đây là mô hình học máy thuộc nhóm Random Forest, được sử dụng để giải quyết bài toán hồi quy (Regression).
  + n\_estimators=100: Số lượng cây quyết định (decision trees) trong rừng. Ở đây, mô hình sử dụng 100 cây.
  + random\_state=42: Đảm bảo kết quả có thể lặp lại bằng cách đặt hạt giống ngẫu nhiên cố định.
* Mô hình sẽ xây dựng 100 cây quyết định (do n\_estimators=100). Mỗi cây học một phần dữ liệu khác nhau (bootstrap sampling).
* Cuối cùng, mô hình sẽ tổng hợp kết quả dự đoán từ tất cả các cây (trung bình kết quả của tất cả cây).

Kết quả đánh giá mô hình Random Forest:

A black text on a white background

AI-generated content may be incorrect.

## Phân tích kết quả:

# Tóm tắt kết quả

if r2 > 0.7:

    print("Mô hình dự đoán tốt.")

else:

    print("Mô hình cần cải thiện.")

# Đề xuất các cách cải thiện

print("\nĐề xuất cải thiện:")

print("1. Thử nghiệm các thuật toán khác như Random Forest, Gradient Boosting.")

print("2. Thu thập thêm dữ liệu để cải thiện độ chính xác.")

print("3. Xử lý tốt hơn các biến đầu vào, loại bỏ các biến không cần thiết.")

Đánh giá mô hình dựa trên R2:

* Biến r2 chứa giá trị R² của mô hình (đã tính trước đó).
* So sánh với 0.7:
* Nếu R² > 0.7, nghĩa là mô hình có độ chính xác tốt, có thể sử dụng.
* Nếu R² ≤ 0.7, mô hình chưa đạt yêu cầu, cần cải thiện.

## Đưa ra kết luận và quyết định:

***3.6.1. Kết luận về mô hình:***

* Mô hình Random Forest hoạt động tốt hơn Linear Regression:

Với bộ dữ liệu này, mô hình Random Forest cho kết quả dự đoán chính xác hơn so với mô hình Linear Regression, thể hiện qua các chỉ số MSE và R2. Điều này cho thấy rằng mối quan hệ giữa các biến đầu vào và giá nhà có thể là phi tuyến tính, và Random Forest có khả năng nắm bắt các mối quan hệ phức tạp này tốt hơn.

Giá trị R-squared của mô hình Random Forest cho thấy mô hình có khả năng giải thích một phần đáng kể sự biến động của giá nhà. Tuy nhiên, vẫn còn một phần sự biến động chưa được mô hình giải thích, cho thấy cần cải thiện thêm.

* Cần cải thiện thêm:

Mặc dù mô hình Random Forest cho kết quả tương đối tốt, vẫn cần tiếp tục cải thiện để tăng độ chính xác.

Việc thu thập thêm dữ liệu, đặc biệt là các biến có thể ảnh hưởng lớn đến giá nhà (ví dụ: vị trí, tiện ích xung quanh), sẽ giúp cải thiện đáng kể hiệu suất của mô hình.

Việc tinh chỉnh các siêu tham số của mô hình Random Forest có thể cải thiện kết quả.

***3.6.2. Quyết định và ứng dụng:***

* Đối với nhà đầu tư:

Mô hình này có thể được sử dụng như một công cụ hỗ trợ để dự đoán giá nhà và đưa ra quyết định đầu tư.

Tuy nhiên, cần kết hợp với các yếu tố khác như tình hình thị trường, vị trí, tiềm năng phát triển của khu vực để đưa ra quyết định cuối cùng.

Có thể sử dụng mô hình để đánh giá nhanh giá trị của một bất động sản.

* Đối với người mua nhà:

Mô hình có thể giúp người mua nhà tham khảo giá trị thị trường của các căn nhà và đưa ra quyết định mua bán hợp lý.

Nên so sánh kết quả dự đoán với giá cả thực tế trên thị trường và tham khảo ý kiến của các chuyên gia bất động sản.

* Đối với người bán nhà:

Mô hình giúp người bán định giá căn nhà một cách khách quan và phù hợp với thị trường.

Cần lưu ý rằng thị trường bất động sản luôn thay đổi, vì vậy cần cập nhật thông tin thường xuyên.

* Cải thiện mô hình:

Tiếp tục thu thập và làm sạch dữ liệu, đặc biệt là các biến liên quan đến vị trí và tiện ích.

Thử nghiệm các thuật toán khác như Gradient Boosting, XGBoost để so sánh hiệu suất.

Thực hiện tinh chỉnh siêu tham số cho mô hình Random Forest để tối ưu hóa hiệu suất.

Xây dựng thêm các biến mới từ các biến đã có để tăng thêm thông tin cho mô hình.

# KẾT LUẬN

## Kết quả đạt được:

### Kiến thức:

Nhóm đã nắm vững các khái niệm cơ bản và nâng cao về thống kê và phân tích dữ liệu. Sự hiểu biết sâu sắc về hồi quy tuyến tính đã giúp nhóm có khả năng phân tích và diễn giải mối quan hệ giữa các yếu tố ảnh hưởng đến giá nhà.

Thông qua việc tìm hiểu và ứng dụng Python cùng các thư viện như Pandas, NumPy và Scikit-learn, nhóm đã phát triển kỹ năng lập trình và phân tích dữ liệu, từ đó có thể thực hiện các phép toán phức tạp một cách hiệu quả

### Sản phẩm:

Nhóm đã xây dựng thành công mô hình hồi quy tuyến tính để dự đoán giá nhà dựa trên các yếu tố như diện tích, số phòng ngủ, số tầng, và các đặc điểm pháp lý. Mô hình đã cho kết quả dự đoán với độ chính xác tương đối cao, thể hiện qua các chỉ số như Mean Squared Error (MSE) và R-squared (R²).

Các kết quả này không chỉ phục vụ cho mục tiêu học tập mà còn có giá trị thực tiễn cho người mua, người bán và các nhà đầu tư bất động sản tại Hà Nội.

## Ưu điểm và nhược điểm:

### Ưu điểm:

Khả năng dự đoán: Mô hình cho thấy khả năng dự đoán giá nhà tương đối chính xác, giúp người dùng đưa ra quyết định thông minh hơn trong giao dịch bất động sản.

Vận dụng công nghệ: Nhóm đã ứng dụng thành công các công nghệ và phương pháp phân tích dữ liệu hiện đại, cải thiện quy trình xử lý và phân tích thông tin.

Đội ngũ hợp tác: Sự hợp tác hiệu quả giữa các thành viên trong nhóm đã tạo điều kiện thuận lợi cho việc chia sẻ kiến thức và kỹ năng, nâng cao chất lượng công việc.

### Nhược điểm:

Thời gian hạn chế: Thời gian thực hiện đề tài còn hạn chế, dẫn đến một số bước trong quy trình phân tích chưa được thực hiện một cách tối ưu.

Kiến thức còn hạn chế: Một số thành viên trong nhóm còn thiếu kinh nghiệm trong việc áp dụng các thuật toán học máy phức tạp, điều này có thể ảnh hưởng đến khả năng cải thiện và phát triển mô hình trong tương lai.

Dữ liệu chưa đầy đủ: Bộ dữ liệu thu thập được không đầy đủ và có thể thiếu sót ở một số thông tin quan trọng, điều này ảnh hưởng đến độ tin cậy của kết quả dự đoán.

## Công việc tương lai:

* Nâng cao và cải thiện mô hình: Nhóm sẽ tiếp tục nghiên cứu và áp dụng các thuật toán học máy tiên tiến hơn như Random Forest, Gradient Boosting và Neural Networks để cải thiện độ chính xác và khả năng dự đoán của mô hình. Việc này sẽ giúp nhóm hiểu rõ hơn về cách tối ưu hóa các tham số và lựa chọn mô hình phù hợp nhất cho dữ liệu bất động sản.
* Tìm hiểu thêm công nghệ: Em sẽ tìm hiểu thêm về các công nghệ như TensorFlow và Keras để có thể bổ sung thêm các chức năng như dự đoán giá nhà theo thời gian thực hoặc xây dựng các mô hình học sâu (deep learning). Những công nghệ này sẽ giúp mở rộng khả năng phân tích và cung cấp dự đoán chính xác hơn cho người dùng.
* Phát triển thêm chức năng: Nếu tìm được nguồn đầu tư cho dự án, nhóm sẽ phát triển thêm các chức năng như:
* Công cụ phân tích thị trường: Giúp người dùng theo dõi xu hướng giá nhà theo thời gian và so sánh với các khu vực khác.
* Ứng dụng di động: Tạo một ứng dụng di động cho phép người dùng dễ dàng truy cập vào thông tin và dự đoán giá nhà ngay trên điện thoại.
* Tính năng cảnh báo giá: Cung cấp thông báo cho người dùng khi giá nhà trong khu vực của họ đạt đến mức mà họ mong muốn.
* Mở rộng nghiên cứu: Nhóm sẽ nghiên cứu sâu hơn về các yếu tố ảnh hưởng đến giá nhà, bao gồm các yếu tố kinh tế, chính trị và xã hội. Việc này sẽ giúp mô hình trở nên toàn diện hơn và có thể đưa ra dự đoán chính xác hơn trong bối cảnh biến động của thị trường.
* Xây dựng cộng đồng người dùng: Nhóm sẽ tìm cách xây dựng một cộng đồng người dùng để nhận phản hồi và ý kiến từ những người sử dụng mô hình, từ đó cải thiện chất lượng dịch vụ và phát triển thêm các tính năng hữu ích.

# TÀI LIỆU THAM KHẢO

[1] “[TailieuVNU.com] Giáo trình Thống Kê Và Ứng Dụng - Đặng Hùng Thắng ( NXB Giáo Dục 1999)”.

[2] “3-lc3bd-thuye1babft-xc3a1c-sue1baa5t-vc3a0-the1bb91ng-kc3aa-toc3a1n-pgs-ts-nguye1bb85n-cao-vc483n-662-trang”.

[3] “PhanTichDuLieuVoiR\_NguyenVanTuan”.