

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**  
**TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**



**BÁO CÁO THỰC TẬP DOANH NGHIỆP**

**ĐỀ TÀI**

**PHÂN TÍCH DỮ LIỆU VÀ DỰ ĐOÁN GIÁ NHÀ BẰNG**  
**PHƯƠNG PHÁP HỒI QUY TUYẾN TÍNH**

**Giảng viên hướng dẫn**

**: TS. Nguyễn Mạnh Cường**

**Lớp**

**: KHMT02-K16**

**Sinh viên thực hiện**

**: Trần Gia Hoàng - 2021605995**

**Hà Nội – 2025**

# MỤC LỤC

MỤC LỤC .....	1
DANH MỤC HÌNH ẢNH.....	5
DANH MỤC BẢNG BIỂU .....	7
LỜI CẢM ƠN .....	8
LỜI NÓI ĐẦU .....	9
CHƯƠNG 1: TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU VÀ BÀI TOÁN DỰ ĐOÁN GIÁ NHÀ .....	12
1.1. Tổng quan về phân tích dữ liệu .....	12
1.1.1. Khái niệm về phân tích dữ liệu.....	12
1.1.2. Quy trình phân tích dữ liệu.....	12
1.1.3. Ưu điểm của phân tích dữ liệu.....	14
1.1.4. Nhược điểm của phân tích dữ liệu .....	14
1.1.5. Một số miền ứng dụng của phân tích dữ liệu .....	15
1.2. Tổng quan về bài toán dự báo.....	15
1.2.1. Lịch sử bài toán dự báo .....	15
1.2.2. Tình hình phát triển của bài toán dự báo ở Việt Nam .....	17
1.2.3. Tình hình phát triển của bài toán dự báo trên Thế giới.....	19

1.3. Phát biểu bài toán .....	19
1.3.1. Giới thiệu chung.....	19
1.3.2. Phương pháp tiếp cận .....	21
1.3.3. Xác định đầu vào, đầu ra của bài toán .....	22
1.3.4. Ý nghĩa thực tiễn của bài toán .....	22
1.3.5. Thuận lợi và khó khăn.....	25
<b>CHƯƠNG 2: CÁC KỸ THUẬT GIẢI QUYẾT BÀI TOÁN.....</b>	<b>27</b>
2.1. Phương hướng giải quyết .....	27
2.2. Mô hình cây quyết định (Decision tree).....	29
2.2.1 Giới thiệu .....	29
2.2.2 Đặc điểm.....	30
2.2.3 Ưu điểm và nhược điểm .....	33
2.3. Mô hình hồi quy Logistic .....	34
2.3.1. Giới thiệu .....	34
2.3.2. Đặc điểm.....	35
2.3.3. Phân loại hồi quy Logistic .....	36
2.3.4. Ưu điểm và nhược điểm .....	36
2.3.5. Ứng dụng .....	37
2.4. Mô hình hồi quy tuyến tính .....	37
2.4.1. Giới thiệu .....	37
2.4.2. Các thành phần của hồi quy tuyến tính .....	38
2.4.3. Đánh giá và diễn giải kết quả .....	39
2.4.4. Ứng dụng của mô hình hồi quy tuyến tính.....	40
2.4.5. Ưu điểm và nhược điểm .....	41

2.5. Lựa chọn mô hình giải quyết bài toán .....	42
<b>CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ .....</b>	<b>44</b>
3.1. Môi trường thực nghiệm .....	44
3.2. Dữ liệu thực nghiệm .....	45
3.3. Quy Trình thực nghiệm.....	46
3.3.1. Đặt mục tiêu.....	47
3.3.2. Tiền xử lý dữ liệu.....	47
3.3.3. Phân tích mô tả.....	50
3.3.4. Phân tích hồi quy .....	56
3.4. Đánh giá.....	60
3.5. Kết luận .....	61
<b>CHƯƠNG 4: CHƯƠNG TRÌNH DEMO .....</b>	<b>62</b>
4.1. Giới thiệu về Framework sử dụng (Framework Streamlit) .....	62
4.1.1. Giới thiệu về Streamlit .....	62
4.1.2. Vai trò của Streamlit .....	62
4.1.3. Ưu và nhược điểm.....	63
4.1.4. Ứng dụng của Streamlit.....	64
4.1.5. Cài đặt Framework Streamlit.....	64
4.2. Chuẩn bị tài nguyên .....	64
4.3. Xây dựng mô hình và Demo chương trình .....	65
4.3.1 Xây dựng mô hình.....	65
4.3.2 Demo chương trình .....	72
<b>KẾT LUẬN.....</b>	<b>80</b>

TÀI LIỆU THAM KHẢO .....	82
--------------------------	----

## DANH MỤC HÌNH ẢNH

Hình 1.1. Quy trình phân tích dữ liệu .....	12
Hình 2.1. Ví dụ về mô hình cây quyết định .....	30
Hình 2.2. Thuật toán hồi quy Logistic .....	35
Hình 2.3. Ví dụ thuật toán hồi quy tuyến tính .....	38
Hình 3.1. Google Colab .....	44
Hình 3.2. Ngôn ngữ lập trình Python .....	44
Hình 3.3. Google Drive .....	45
Hình 3.4. 15 dòng đầu của bộ dữ liệu gốc .....	46
Hình 3.5. Quy trình thực nghiệm đề tài phân tích dữ liệu .....	47
Hình 3.6. Kiểm tra dữ liệu bị khuyết .....	48
Hình 3.7. Xử lý dữ liệu bị khuyết .....	49
Hình 3.8. Xử lý dữ liệu bị trùng .....	49
Hình 3.9. Kết quả xử lý dữ liệu bị trùng .....	49
Hình 3.10. Xử lý giá trị ngoại lai .....	50
Hình 3.11. chuyển hóa cột dữ liệu .....	50
Hình 3.12. Tóm lược dữ liệu .....	51
Hình 3.13. Bảng tóm lược dữ liệu .....	52
Hình 3.14. Biểu đồ Hist Plot .....	52
Hình 3.15. Biểu đồ Histogram cho các cột .....	53
Hình 3.16. Biểu đồ Box Plot .....	55
Hình 3.17. Biểu đồ Box Plot .....	56

Hình 3.18. Chia tập dữ liệu huấn luyện và kiểm tra .....	57
Hình 3.19. Chuẩn hóa dữ liệu .....	57
Hình 3.20. Huấn luyện bằng mô hình Hồi quy tuyến tính .....	58
Hình 3.21. Dự đoán và đánh giá mô hình .....	58
Hình 3.22. Kết quả so sánh dữ liệu dự đoán với dữ liệu thực tế .....	59
Hình 3.23. Biểu đồ so sánh giữa dữ liệu dự đoán với dữ liệu thực tế .....	60
Hình 4.1. Cài đặt Streamlit .....	64
Hình 4.2. Lưu mô hình chuẩn hóa min-max .....	65
Hình 4.3. Lưu mô hình đã huấn luyện .....	65
Hình 4.5. Giao diện chương trình .....	73
Hình 4.6. Giao diện nhập dữ liệu .....	74
Hình 4.7. Mã nguồn nhận dữ liệu .....	75
Hình 4.8. Mô tả quá trình dự đoán .....	75
Hình 4.9. Quá trình lưu kết quả dự đoán .....	76
Hình 4.10. Hình kết quả chạy chương trình .....	76
Hình 4.11. Ảnh minh họa cho kết quả dự đoán .....	77
Hình 4.12. Biểu đồ lịch sử dự đoán giá nhà .....	78
Hình 4.13. Mã nguồn Reset dữ liệu .....	79
Hình 4.14. Giao diện chương trình sau khi Reset .....	79

## **DANH MỤC BẢNG BIỂU**

Bảng 4.1.Bảng đặc tả use case “Nhập dữ liệu” .....	66
Bảng 4.2.Bảng đặc tả use case ‘Dự Đoán’ .....	69
Bảng 4.3.Bảng đặc tả use case ‘Reset’ .....	71



## **LỜI CẢM ƠN**

Lời đầu tiên cho phép em gửi lời cảm ơn sâu sắc tới các thầy cô trong khoa Công nghệ thông tin – Trường Đại học Công Nghiệp Hà Nội, những người đã tận tụy chỉ bảo, dạy dỗ và truyền đạt cho em những kiến thức, những bài học quý báu và bổ ích. Đặc biệt em xin được bày tỏ lời cảm ơn chân thành tới giảng viên hướng dẫn TS. Nguyễn Mạnh Cường, người đã trực tiếp hướng dẫn, tận tình giải đáp thắc mắc và chỉ bảo trong suốt quá trình học tập, nghiên cứu và hoàn thành đồ án. Sau cùng, em xin gửi tình cảm sâu sắc tới gia đình và bạn bè vì đã luôn ở bên cạnh khuyến khích, động viên và giúp đỡ cả về vật chất cũng như tinh thần cho em trong suốt quá trình học tập để em hoàn thành tốt việc học tập của bản thân.

Trong quá trình nghiên cứu và thực hiện đề tài, do năng lực, kiến thức và trình độ bản thân em vẫn còn hạn hẹp nên không thể tránh khỏi những thiếu sót và em rất mong nhận được sự thông cảm và góp ý từ quý thầy cô cũng như các bạn đọc để nghiên cứu này có thể hoàn thiện hơn.

*Em xin trân trọng cảm ơn!*

**Sinh viên thực hiện**

*Trần Gia Hoàng*

## LỜI NÓI ĐẦU

Trong bối cảnh kinh tế thị trường ngày càng phát triển, bất động sản đã trở thành một trong những lĩnh vực có tác động mạnh mẽ đến nền kinh tế của mọi quốc gia. Giá nhà đất là một trong những yếu tố được nhiều người quan tâm, từ các nhà đầu tư bất động sản cho đến những người mua nhà lần đầu. Tuy nhiên, sự biến động của giá nhà đất phụ thuộc vào rất nhiều yếu tố khác nhau như vị trí địa lý, diện tích, tiện ích xung quanh, và tình hình kinh tế vĩ mô, làm cho việc dự đoán giá nhà trở thành một thách thức.

Trong lĩnh vực khoa học dữ liệu, các phương pháp phân tích và dự đoán giá trị tài sản đã được áp dụng rộng rãi để hỗ trợ việc đưa ra các quyết định chiến lược. Một trong những phương pháp được sử dụng phổ biến là hồi quy tuyến tính, một kỹ thuật dựa trên mối quan hệ giữa các biến đầu vào và đầu ra, nhằm dự đoán giá trị của biến phụ thuộc dựa trên các biến độc lập khác.

Mục tiêu của đề tài này là xây dựng một mô hình hồi quy tuyến tính dựa trên dữ liệu đầu vào và sử dụng các phương pháp thống kê và kỹ thuật phân tích dữ liệu để xác định các biến có liên quan và xây dựng một mô hình dự báo chính xác.

Qua đề tài "Phân tích dữ liệu và dự đoán giá nhà bằng phương pháp hồi quy tuyến tính", em đã thực hiện nghiên cứu và ứng dụng các kỹ thuật phân tích dữ liệu hiện đại để xây dựng mô hình dự đoán giá nhà một cách chính xác. Đề tài không chỉ nhằm mục đích tìm hiểu và vận dụng kiến thức đã học, mà còn hy vọng góp phần nhỏ vào việc cung cấp thông tin hữu ích cho các nhà phát triển, chủ đầu tư và các bên liên quan trong việc đưa ra quyết định kinh doanh và lập kế hoạch tương lai cũng như mở ra cơ hội cho việc nghiên cứu và ứng dụng các phương pháp dự báo trong lĩnh vực bất động sản.

Nội dung của quyển báo cáo sẽ gồm các chương như sau:

## **Chương 1: Tổng quan về phân tích dữ liệu và bài toán dự đoán giá nhà**

Chương này cung cấp cái nhìn tổng quan về phân tích dữ liệu, giới thiệu về bài toán dự đoán giá nhà, các đầu vào, đầu ra và đặc điểm của bài toán. Từ đó làm rõ bối cảnh và tầm quan trọng của đề tài nghiên cứu.

## **Chương 2: Các kỹ thuật giải quyết bài toán**

Nội dung của chương này tập trung vào việc phân tích các phương pháp hiện có để giải quyết bài toán dự đoán giá nhà cùng các ưu nhược điểm của chúng. Sau đó, phân tích chi tiết về phương pháp đề xuất để giải quyết bài toán dự đoán giá nhà.

## **Chương 3: Kết quả thực nghiệm**

Tại chương này, em sẽ trình bày về quá trình thực nghiệm và kết quả đạt được với phương pháp được đề xuất ở chương 2. Cuối cùng là đánh giá kết quả thực nghiệm đạt được sau khi ứng dụng phương pháp được đề xuất để giải quyết bài toán dự đoán giá nhà.

## **Chương 4: Chương trình demo**

Chương này mô tả chi tiết quá trình triển khai và xây dựng hệ thống dự đoán giá nhà bao gồm các bước thiết kế, lập trình và triển khai, thử nghiệm hệ thống.

## **Phần kết luận:**

Cuối cùng, trong phần kết luận của nghiên cứu, em sẽ tổng hợp các kết quả đạt được, các hướng phát triển và hướng mở rộng đề tài nghiên cứu cho tương lai.

Thông qua việc thực hiện nghiên cứu đề tài, em đã có cơ hội được tiếp cận nhiều hơn, tìm hiểu sâu hơn về lĩnh vực phân tích dữ liệu và dự báo, đặc biệt là áp dụng chúng vào thị trường bất động sản thực tế.

# CHƯƠNG 1: TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU VÀ BÀI TOÁN DỰ ĐOÁN GIÁ NHÀ

## 1.1. Tổng quan về phân tích dữ liệu

### 1.1.1. Khái niệm về phân tích dữ liệu

Phân tích dữ liệu là một quá trình đa bước, bao gồm việc kiểm tra chất lượng dữ liệu, làm sạch dữ liệu sai lệch hoặc thiếu sót, chuyển đổi dữ liệu sang định dạng phù hợp và cuối cùng là xây dựng các mô hình để khám phá những thông tin hữu ích. Thông tin thu được sẽ được sử dụng để đưa ra các kết luận có cơ sở và hỗ trợ quyết định.

Phân tích dữ liệu có nhiều khía cạnh và cách tiếp cận, bao gồm các kỹ thuật đa dạng dưới nhiều tên khác nhau và được sử dụng trong các lĩnh vực kinh doanh, khoa học và khoa học xã hội khác nhau. Trong thế giới kinh doanh ngày nay, phân tích dữ liệu đóng vai trò giúp đưa ra quyết định khoa học hơn và giúp doanh nghiệp hoạt động hiệu quả hơn.

### 1.1.2. Quy trình phân tích dữ liệu



Hình 1.1. Quy trình phân tích dữ liệu

Quá trình phân tích dữ liệu thường bao gồm các bước:

- Xác thực mục tiêu và thu thập dữ liệu:
  - + Xác thực mục tiêu: là những kết quả cụ thể mà ta muốn đạt được thông qua việc xử lý và phân tích dữ liệu. Mục tiêu này xác định hướng đi và phạm vi của quá trình phân tích, giúp ta tập trung vào việc thu thập thông tin quan trọng và thực hiện các phân để đáp ứng các yêu cầu hoặc nhu cầu cụ thể.
  - + Thu thập dữ liệu : là quá trình tích hợp thông tin từ nhiều nguồn vào một hệ thống thống nhất, nhằm phục vụ phân tích, nghiên cứu, quản lý, kinh doanh và ra quyết định trong các lĩnh vực khoa học, xã hội và kinh doanh. Dữ liệu thu thập được cho phép trả lời các câu hỏi liên quan và đánh giá kết quả.
- Tiền xử lý dữ liệu: là quá trình sử dụng các kỹ thuật làm cho dữ liệu ban đầu trở nên đầy đủ hơn do dữ liệu đầu vào thường không hoàn chỉnh, có thể bị nhiễu, bị thiếu khuyết, giá trị không chính xác. Tiền xử lý dữ liệu bao gồm 1 số bước như tóm lược dữ liệu, làm sạch dữ liệu, chuyển đổi dữ liệu, rời rạc hóa dữ liệu, ...
- Phân tích dữ liệu là bước cốt yếu trong việc khám phá mối quan hệ giữa dữ liệu đầu vào và đầu ra. Các kỹ thuật thường được áp dụng bao gồm hồi quy tuyến tính, phân tích mô tả, hồi quy logistic, cây quyết định và random forest.
- Kết luận và dự đoán: Dựa trên phân tích và thông tin từ dữ liệu đầu vào, chúng ta có thể đưa ra kết luận, hiểu rõ hơn về tình hình và có thể đưa ra dự đoán cho tương lai.

### **1.1.3. Ưu điểm của phân tích dữ liệu**

- Ra quyết định chính xác hơn: Giúp các tổ chức và cá nhân đưa ra quyết định dựa trên cơ sở dữ liệu, giảm thiểu rủi ro và tăng độ chính xác.
- Phát hiện xu hướng và mẫu: Nhận diện được các xu hướng trong dữ liệu giúp đưa ra dự đoán và chuẩn bị cho các thay đổi hoặc cơ hội tiềm năng.
- Tối ưu hóa quy trình: Cải thiện hiệu quả hoạt động bằng cách phát hiện các bước không cần thiết trong quy trình và tìm cách tối ưu hóa.
- Cá nhân hóa trải nghiệm khách hàng: Dựa trên dữ liệu, doanh nghiệp có thể tạo ra các chiến lược marketing và dịch vụ được tùy chỉnh theo nhu cầu của từng khách hàng.
- Tăng năng suất và lợi nhuận: Phân tích dữ liệu giúp phát hiện các cơ hội kinh doanh và quản lý tài nguyên hiệu quả hơn, từ đó tăng doanh thu và giảm chi phí.

### **1.1.4. Nhược điểm của phân tích dữ liệu**

- Tốn thời gian và nguồn lực: Quá trình thu thập, làm sạch và phân tích dữ liệu có thể tốn nhiều thời gian và đòi hỏi nguồn lực lớn, đặc biệt khi làm việc với lượng dữ liệu lớn.
- Yêu cầu kỹ năng chuyên môn: Để thực hiện phân tích dữ liệu chất lượng, cần có đội ngũ nhân sự am hiểu về thống kê, phân tích dữ liệu và các công cụ liên quan. Điều này có thể làm tăng chi phí đào tạo và tuyển dụng.
- Dễ dẫn đến sai lệch: Nếu dữ liệu không đầy đủ hoặc không đại diện, kết quả phân tích có thể bị sai lệch, dẫn đến quyết định không chính xác. Cũng vậy, sự thiên vị trong dữ liệu có thể làm ảnh hưởng đến kết quả.

- Vấn đề bảo mật và riêng tư: Xử lý dữ liệu cá nhân hoặc nhạy cảm đòi hỏi tuân thủ các quy định về bảo mật thông tin. Nếu không cẩn trọng, việc vi phạm quy định này có thể dẫn đến các vấn đề pháp lý và tổn thất uy tín.

- Khó khăn trong tích hợp: Đôi khi, việc tích hợp dữ liệu từ nhiều nguồn khác nhau để phân tích có thể gặp khó khăn do định dạng dữ liệu không nhất quán hoặc khác nhau về chất lượng.

### **1.1.5. Một số miền ứng dụng của phân tích dữ liệu**

- Kinh doanh và tiếp thị
- Y tế và chăm sóc sức khỏe
- Tài chính và ngân hàng
- Giáo dục
- Thể thao
- Giao thông vận tải
- Nông nghiệp
- Công nghiệp

## **1.2. Tổng quan về bài toán dự báo**

### **1.2.1. Lịch sử bài toán dự báo**

Bài toán dự báo có một lịch sử lâu đời và đã phát triển qua nhiều giai đoạn. Dưới đây là một cái nhìn tổng quan về lịch sử hình thành của bài toán dự báo:

- Thời kỳ tiền Công nghiệp (Trước thế kỷ 18): Trong giai đoạn này, con người thường dự báo dựa trên kinh nghiệm và tri thức truyền đạt qua thế hệ. Dự báo chủ yếu dựa trên sự quan sát của thiên văn học, thời tiết, và các hiện tượng tự nhiên.



- Cách mạng Công nghiệp và thống kê (Thế kỷ 18 - 19): Trong thời kỳ này, việc sử dụng số liệu và thống kê để dự báo đã trở nên phổ biến hơn. Những ý tưởng về xác suất và phân phối bắt đầu được áp dụng vào việc dự báo.

- Thế kỷ 20 và Kỹ thuật số hoá: Sự phát triển của máy tính và kỹ thuật số hoá đã mở ra những cơ hội mới trong việc dự báo. Các phương pháp thống kê, mô hình hóa toán học, và kỹ thuật machine learning bắt đầu được sử dụng rộng rãi để dự báo trong nhiều lĩnh vực.

- Thống kê Bayes và Kỹ thuật Machine learning (Thế kỷ 20 - 21): Thống kê Bayes và các kỹ thuật machine learning như học máy, học sâu, và học tăng cường đã thúc đẩy khả năng dự báo thông qua việc xử lý dữ liệu phức tạp và tìm ra các mẫu ẩn.

- Dự báo trong thời đại số hóa (Hiện nay): Với sự gia tăng mạnh mẽ về khả năng tính toán, khối lượng dữ liệu khổng lồ, và sự phát triển của trí tuệ nhân tạo, bài toán dự báo đang trở nên càng quan trọng và phức tạp hơn. Các công nghệ mới như big data analytics, deep learning, và dự báo dựa trên mạng xã hội đang mở ra nhiều cơ hội và thách thức mới trong lĩnh vực này.

- Trong suốt quá trình phát triển, bài toán dự báo đã chuyển từ việc dự đoán dựa trên sự quan sát đơn thuần đến việc sử dụng các phương pháp phức tạp để xác định mối quan hệ phức hợp và xu hướng từ dữ liệu. Lịch sử hình thành này thể hiện sự tiến bộ và tầm quan trọng của bài toán dự báo trong việc hỗ trợ quyết định và phát triển trong nhiều lĩnh vực.

Bài toán dự báo là một trong những thách thức quan trọng trong lĩnh vực phân tích dữ liệu, nơi chúng ta cố gắng dự đoán giá trị của một biến mục tiêu trong tương lai dựa trên dữ liệu lịch sử và các yếu tố ảnh hưởng. Mục tiêu chính của bài toán dự báo là xây dựng một mô hình có khả năng hiểu và ứng

dùng các mẫu, xu hướng và quy luật từ dữ liệu để thực hiện việc dự đoán một cách chính xác và đáng tin cậy.

Trong thời đại số hóa hiện nay, bài toán dự báo đối mặt với những thách thức và cơ hội mới. Sự phát triển của big data analytics cho phép thu thập và xử lý khối lượng lớn dữ liệu từ nhiều nguồn khác nhau, tạo ra cơ hội để tìm ra các mẫu và thông tin quan trọng từ dữ liệu này. Dự báo dựa trên mạng xã hội là một lĩnh vực mới nổi trong đó dữ liệu từ các nền tảng mạng xã hội được sử dụng để dự báo hành vi, xu hướng và tương tác của người dùng.

Tổng quan về lịch sử hình thành của bài toán dự báo cho thấy sự tiến bộ và quan trọng của nó trong việc hỗ trợ quyết định và phát triển trong nhiều lĩnh vực. Bài toán dự báo đã phát triển từ việc dự đoán dựa trên kinh nghiệm và quan sát đơn thuần đến việc sử dụng các phương pháp và công nghệ phức tạp để tìm hiểu và dự đoán dựa trên dữ liệu lịch sử.

### **1.2.2. Tình hình phát triển của bài toán dự báo ở Việt Nam**

Bài toán dự báo có sự ảnh hưởng to lớn tại cả Việt Nam. Dự báo giúp cải thiện quản lý, định hình chiến lược, và tối ưu hóa tài nguyên trong nhiều lĩnh vực. Có một số điểm đáng chú ý về tình hình phân tích dữ liệu tại Việt Nam:

- Phát triển đang ở giai đoạn đầu: Trong một số lĩnh vực, bài toán dự báo tại Việt Nam đang ở giai đoạn đầu của sự phát triển. Việc áp dụng các phương pháp phân tích dữ liệu và dự báo mới còn đang được tìm hiểu và thí nghiệm.
- Ứng dụng trong nông nghiệp và kinh tế: Tại Việt Nam, dự báo có ứng dụng quan trọng trong nông nghiệp, nhằm dự đoán thời tiết, mùa màng, và nhu cầu năng lượng. Nó cũng được áp dụng trong kinh tế, dự báo tăng trưởng GDP, lạm phát, và tỷ giá.

- Thách thức từ dữ liệu: Một thách thức cho việc dự báo tại Việt Nam là khả năng thu thập và quản lý dữ liệu chất lượng. Dữ liệu thường không đầy đủ và có thể gặp vấn đề về tính nhất quán và độ tin cậy.

Dưới đây là một cái nhìn tổng quan về phát triển của bài toán dự báo ở Việt Nam:

- Thời kỳ tiền Công nghiệp và Cách mạng Công nghiệp: Trước thế kỷ 18 và trong giai đoạn Cách mạng Công nghiệp, dự báo ở Việt Nam cũng dựa trên các quan sát và tri thức truyền đạt qua thế hệ, tương tự như các nước khác. Những thông tin về thời tiết, thiên văn học và các hiện tượng tự nhiên được sử dụng để dự báo.

- Thế kỷ 20 và Kỹ thuật số hoá: Với sự phát triển của máy tính và kỹ thuật số hoá, Việt Nam đã bắt kịp xu hướng sử dụng các phương pháp thống kê và mô hình hóa toán học để dự báo. Các ngành công nghiệp như tài chính, thương mại 16 và sản xuất đã áp dụng các phương pháp này để dự báo xu hướng thị trường, tiêu thụ và sản xuất.

- Thống kê Bayes và Kỹ thuật Machine learning: Thống kê Bayes và các kỹ thuật machine learning như học máy và học sâu đã được áp dụng rộng rãi trong bài toán dự báo ở Việt Nam. Các công ty và tổ chức nghiên cứu đã sử dụng các phương pháp này để dự báo trong lĩnh vực tài chính, thương mại, y tế và nông nghiệp.

- Sự phát triển của big data và dự báo dựa trên mạng xã hội: Việt Nam cũng đã nhận thấy tiềm năng của big data và dữ liệu từ mạng xã hội trong bài toán dự báo. Việc thu thập và phân tích dữ liệu từ các nguồn khác nhau như mạng xã hội, thiết bị cảm biến và các hệ thống thông tin công nghệ cao đang mở ra nhiều cơ hội mới trong việc dự báo tình hình và xu hướng.

### **1.2.3. Tình hình phát triển của bài toán dự báo trên Thế giới**

- Phát triển mạnh: Tại các quốc gia phát triển, bài toán dự báo đã được phát triển mạnh và có sự ứng dụng rộng rãi trong nhiều lĩnh vực như tài chính, thương mại điện tử, y tế, và năng lượng.
- Sự kết hợp của công nghệ mới: Các quốc gia nước ngoài thường kết hợp sự phát triển của công nghệ mới như trí tuệ nhân tạo, học máy, và big data analytics để cải thiện hiệu suất của bài toán dự báo.
- Tổng hợp dữ liệu: Một ưu điểm của các quốc gia phát triển là có khả năng tổng hợp dữ liệu từ nhiều nguồn khác nhau, tạo nền tảng cho việc dự báo chính xác hơn và đa dạng hơn.

## **1.3. Phát biểu bài toán**

### **1.3.1. Giới thiệu chung**

Trong xã hội hiện nay với đầy biến động của thị trường bất động sản hiện nay, việc đánh giá và dự đoán giá nhà là một thách thức lớn đối với các doanh nghiệp, và nhà phát triển bất động sản. Đây không chỉ là yếu tố quan trọng trong việc quản lý và lập chiến lược kinh doanh, mà còn là thành phần then chốt trong việc định hình vị thế và sức hút của thị trường. Khả năng dự đoán chính xác giá nhà giúp các bên liên quan nắm bắt thời cơ, giảm thiểu rủi ro và tối ưu hóa lợi nhuận, đồng thời cung cấp thông tin quý giá cho người mua và nhà đầu tư.

Sự biến động của thị trường bất động sản phụ thuộc vào nhiều yếu tố khác nhau. Để giải quyết vấn đề này, phân tích số liệu đã trở thành một công cụ quan trọng nhằm mục đích đo lường sự biến động của giá nhà. Bài toán phân tích này đòi hỏi sự kết hợp chặt chẽ giữa kiến thức, kinh nghiệm về thị trường bất động sản và sự sâu sắc đối với các phương pháp phân tích dữ liệu

hiện đại. Khả năng dự đoán giá nhà không chỉ giúp cho các doanh nghiệp nắm bắt xu hướng và đưa ra quyết định đầu tư đúng đắn, mà còn hỗ trợ trong việc lập kế hoạch tài chính và định hướng chiến lược kinh doanh hiệu quả.

Dữ liệu của bài toán cần được thu thập một cách tỉ mỉ và chọn lọc, từ các yếu tố cơ bản như vị trí địa lý, diện tích, số phòng ngủ, số phòng tắm, năm xây dựng, giá bán, đến những yếu tố xung quanh như số khoảng cách tới thành phố, số lượng bất động sản trong khu vực. Sự kết hợp thông tin chi tiết này giúp xây dựng cơ sở dữ liệu đa chiều, cung cấp cái nhìn toàn diện về sự biến động của thị trường bất động sản theo thời gian.

Bên cạnh đó, việc áp dụng các mô hình học máy như hồi quy tuyến tính, Random Forest, mạng nơ-ron hay cây quyết định giúp chúng ta hiểu rõ hơn về mối liên quan giữa các yếu tố và giá bán của ngôi nhà. Những mô hình này không chỉ giúp dự đoán giá bán dựa trên những yếu tố của ngôi nhà mà còn hỗ trợ các doanh nghiệp và các nhà đầu tư nắm bắt xu hướng và quyết định chiến lược tương lai.

Tuy nhiên, đối mặt với sự đa dạng và biến đổi liên tục của thị trường bất động sản, chúng ta sẽ phải đối diện với những thách thức như dữ liệu không đầy đủ, nhiễu loạn hoặc sai lệch, biến động giá cả theo thời gian cũng như sự đa dạng về yếu tố ảnh hưởng đến giá bán khiến khả năng nắm bắt yếu tố ảnh hưởng của mô hình bị suy giảm. Nhưng với kỹ thuật phân tích tiên tiến hiện nay cùng với sự sáng tạo và đổi mới liên tục, chúng ta hoàn toàn có thể tiếp cận và giải quyết bài toán một cách trọn vẹn và hiệu quả.

### **1.2.1. Mục tiêu của bài toán**

Mục tiêu của bài toán phân tích dữ liệu và dự đoán giá nhà là phân tích các yếu tố ảnh hưởng, nghiên cứu các yếu tố có thể ảnh hưởng tới giá trị của ngôi nhà như: vị trí địa lý, số phòng ngủ, số phòng tắm, năm xây dựng, tiện ích đi kèm, và nhiều yếu tố khác liên quan đến bất động sản, từ đó xây dựng một mô hình học máy có khả năng dự đoán giá trị của ngôi nhà.

Bài toán này có thể được sử dụng trong các nghiên cứu xã hội, kinh tế và chính trị để nghiên cứu các yếu tố có thể dẫn đến sự ảnh hưởng tới giá trị của một ngôi nhà. Từ đó có thể giúp người dùng đưa ra các quyết định, chiến lược kinh doanh thông minh và tối ưu.

Bài toán phân tích dữ liệu và dự đoán giá nhà cũng được sử dụng rộng rãi để chọn lọc và cung cấp thông tin quan trọng về sự ảnh hưởng của những yếu tố đến giá trị của ngôi nhà. Giúp các nhà phát triển, nhà đầu tư và các doanh nghiệp đưa ra quyết định về đầu tư và phát triển các dự án bất động sản, xác định chiến lược kinh doanh, nâng cao tiềm năng lợi nhuận.

### **1.3.2. Phương pháp tiếp cận**

Để phân tích và dự báo giá nhà, các phương pháp học máy như hồi quy tuyến tính thường được sử dụng để xác định mối quan hệ giữa giá nhà và các yếu tố ảnh hưởng. Các bước thực hiện bao gồm:

- Thu thập dữ liệu: Bước đầu tiên là thu thập các dữ liệu liên quan đến giá nhà và các yếu tố ảnh hưởng như diện tích, số phòng, vị trí, tiện ích xung quanh, tình hình kinh tế, và các yếu tố khác.

- Tiền xử lý dữ liệu: Sau khi thu thập dữ liệu, cần tiến hành tiền xử lý, bao gồm làm sạch dữ liệu, loại bỏ các giá trị bị thiếu hoặc ngoại lệ, chuẩn hóa dữ liệu và chuyển đổi dữ liệu về dạng phù hợp cho phân tích.
- Phân tích mô tả: Sử dụng phương pháp phân tích mô tả để xác định được mối quan hệ giữa giá nhà và các biến độc lập, từ đó chọn ra mô hình phù hợp nhất.
- Xây dựng mô hình: Dựa trên dữ liệu đã tiền xử lý, ta xây dựng mô hình hồi quy tuyến tính để dự báo giá nhà, sử dụng các thuật toán tối ưu nhằm điều chỉnh mô hình dự báo chính xác.
- Đánh giá và tinh chỉnh mô hình: Sau khi mô hình được xây dựng, cần đánh giá hiệu quả dự báo của mô hình bằng cách sử dụng các chỉ số như R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Adjusted R-squared. Nếu cần thiết, có thể điều chỉnh lại các tham số hoặc thêm biến độc lập để cải thiện kết quả.
- Dự báo giá nhà: Sau khi tinh chỉnh mô hình, mô hình này có thể được áp dụng để dự báo giá nhà trong tương lai bằng cách cung cấp các dữ liệu mới và nhận được kết quả dự báo.

### **1.3.3. Xác định đầu vào, đầu ra của bài toán**

- Đầu vào: Tổng cộng có 20640 bản ghi về thông tin của từng ngôi nhà được thu thập từ cuộc điều tra dân số thực tế ở California năm 1990.
- Đầu ra: Là giá trị của ngôi nhà, là biến mục tiêu mà mô hình sẽ được huấn luyện dựa trên các biến độc lập của ngôi nhà để dự đoán. Giá trị được biểu diễn dưới dạng số liên tục. Ngoài ra còn có thông số để đánh giá độ chính xác của mô hình.

### **1.3.4. Ý nghĩa thực tiễn của bài toán**

Bài toán dự báo giá nhà có tầm quan trọng lớn trong thực tế. Đối với các chủ đầu tư và nhà phát triển, việc dự đoán giá nhà giúp họ hiểu rõ xu hướng thị trường, từ đó xây dựng chiến lược kinh doanh, lựa chọn dự án phù hợp và đưa ra quyết định về giá cả, quảng cáo và tiếp thị. Đối với các nhà môi giới, việc dự đoán giá nhà hỗ trợ tối ưu hóa hoạt động và tạo ra các chiến lược tiếp cận khách hàng hiệu quả hơn.

– Hỗ trợ người mua và người bán bất động sản

+ Người mua: Giúp người mua nhà có được ước lượng hợp lý về giá trị căn nhà họ đang cân nhắc, từ đó giúp họ đưa ra quyết định mua bán có lợi nhất.

+ Người bán: Cung cấp cho người bán một mức giá tham khảo phù hợp với thị trường, từ đó có thể định giá bán hợp lý, tối ưu hóa lợi nhuận và tăng tính cạnh tranh của tài sản trên thị trường.

– Hỗ trợ các tổ chức tài chính và ngân hàng

+ Đánh giá giá trị tài sản thế chấp: Khi một khách hàng yêu cầu vay thế chấp bằng bất động sản, các tổ chức tài chính có thể sử dụng mô hình này để định giá tài sản, giúp đánh giá mức độ rủi ro khi cho vay.

+ Dự báo và phân tích rủi ro: Mô hình dự đoán giá nhà hỗ trợ ngân hàng và các tổ chức tài chính xác định các khu vực hoặc loại hình bất động sản tiềm ẩn rủi ro cao trong trường hợp thị trường biến động.

– Hỗ trợ các nhà đầu tư và nhà phát triển bất động sản

+ Đánh giá tiềm năng sinh lời: Các nhà đầu tư bất động sản có thể sử dụng mô hình để ước tính giá trị hiện tại và tiềm năng tăng giá trong



tương lai của bất động sản ở các khu vực khác nhau, giúp đưa ra quyết định đầu tư chiến lược.

- + **Lập kế hoạch phát triển:** Các nhà phát triển có thể sử dụng thông tin từ mô hình để lựa chọn khu vực đầu tư và xây dựng phù hợp, từ đó tối đa hóa lợi nhuận.
- **Hỗ trợ các cơ quan quản lý và quy hoạch đô thị**
  - + **Định hướng phát triển đô thị:** Các cơ quan chính phủ có thể sử dụng dữ liệu dự báo để hiểu được xu hướng giá bất động sản, từ đó quy hoạch và phát triển cơ sở hạ tầng hợp lý, đáp ứng nhu cầu của thị trường.
  - + **Điều tiết thị trường bất động sản:** Việc hiểu rõ các yếu tố ảnh hưởng đến giá nhà giúp cơ quan quản lý đưa ra chính sách điều tiết phù hợp để ngăn ngừa “bong bóng bất động sản” và giữ cho thị trường phát triển bền vững.
- **Ứng dụng trong lĩnh vực bảo hiểm**
  - + **Ước tính phí bảo hiểm:** Giá trị bất động sản là một yếu tố quan trọng trong việc xác định mức phí bảo hiểm nhà ở. Các công ty bảo hiểm có thể dựa vào mô hình dự báo giá nhà để ước lượng phí bảo hiểm hợp lý cho khách hàng.
- **Giáo dục và nghiên cứu thị trường bất động sản**
  - + **Phân tích thị trường:** Mô hình dự đoán giá nhà cung cấp cái nhìn sâu sắc về các yếu tố ảnh hưởng đến giá bất động sản, từ đó tạo cơ sở cho các nghiên cứu và giáo dục trong lĩnh vực kinh tế và tài chính.

- + Dự báo xu hướng: Phân tích giá nhà qua thời gian giúp các nhà kinh tế, nhà nghiên cứu và các bên liên quan dự báo xu hướng thị trường, phục vụ cho các nghiên cứu chuyên sâu hoặc hoạch định chiến lược dài hạn.

### **1.3.5. Thuận lợi và khó khăn**

Bài toán phân tích dự báo giá nhà mở ra nhiều cơ hội phát triển nhưng cũng đối diện với không ít thách thức.

#### **– Thuận lợi**

- + Dự báo thị trường: Dự báo giá nhà giúp các nhà phát triển và đầu tư đưa ra quyết định chiến lược thông minh hơn.

- + Tối ưu hóa lợi nhuận: Dự báo chính xác giúp quản lý giá bán và đầu tư hiệu quả, từ đó tối ưu hóa lợi nhuận.

- + Tăng cường cạnh tranh: Hiểu rõ xu hướng thị trường giúp các doanh nghiệp nắm bắt cơ hội cạnh tranh một cách tốt hơn.

- + Tối ưu chiến dịch quảng cáo: Dự báo giá nhà giúp tối ưu hóa các chiến dịch tiếp thị và quảng cáo, giúp tiếp cận đúng đối tượng khách hàng tiềm năng.

#### **– Khó khăn**

- + Biến động thị trường: Thị trường bất động sản dễ biến động do ảnh hưởng của nhiều yếu tố như chính trị, kinh tế, khiến việc dự báo trở nên khó khăn.

- + Chất lượng dữ liệu: Dữ liệu đầu vào có chất lượng không cao hoặc thiếu sót có thể dẫn đến dự báo sai lệch.

- + Phân tích phức tạp: Việc xây dựng và duy trì mô hình dự báo đòi hỏi kỹ năng chuyên môn cao và cần thời gian để điều chỉnh.
- + Nguồn lực và thời gian: Quá trình thu thập, xử lý và phân tích dữ liệu yêu cầu nhiều nguồn lực và thời gian, đặc biệt đối với các kịch bản dự báo phức tạp.
- + Rủi ro không xác định: Thị trường bất động sản chứa nhiều yếu tố không lường trước, dẫn đến dự báo có thể mang theo rủi ro cao.

## CHƯƠNG 2: CÁC KỸ THUẬT GIẢI QUYẾT BÀI TOÁN

### 2.1. Phương hướng giải quyết

Phương hướng giải quyết trong Phân tích dữ liệu và dự đoán giá nhà bằng phương pháp hồi quy tuyến tính có thể được thực hiện theo một chuỗi các bước cẩn thận và chi tiết sau đây:

- Thu thập và làm sạch dữ liệu
  - Thu thập dữ liệu: Tìm kiếm các tập dữ liệu bất động sản từ các nguồn đáng tin cậy, có thể bao gồm các trang web bất động sản, cơ sở dữ liệu mở, hoặc dữ liệu từ các công ty bất động sản.
  - Làm sạch dữ liệu: Kiểm tra và xử lý các giá trị bị thiếu, các dữ liệu không hợp lệ, hoặc các giá trị ngoại lai có thể ảnh hưởng đến chất lượng dự đoán. Các phương pháp làm sạch dữ liệu có thể bao gồm:
    - + Loại bỏ hoặc thay thế các giá trị bị thiếu.
    - + Phát hiện và xử lý ngoại lai bằng Z-score hoặc IQR.
    - + Chuẩn hóa hoặc mã hóa lại các biến phân loại.
- Khám phá dữ liệu (Exploratory Data Analysis - EDA)
  - Phân tích thống kê cơ bản: Xem xét các chỉ số như trung bình, trung vị, độ lệch chuẩn để hiểu rõ dữ liệu.
  - Trực quan hóa dữ liệu: Sử dụng các biểu đồ như scatter plot, histogram, box plot, heatmap để phát hiện mối quan hệ giữa các đặc điểm và giá nhà.
  - Phát hiện các mối tương quan: Xác định mối quan hệ giữa các biến độc lập và biến mục tiêu (giá nhà) để tìm ra những yếu tố có ảnh hưởng lớn nhất.

- Xử lý và biến đổi dữ liệu (Feature Engineering)
  - Mã hóa biến phân loại: Dùng các kỹ thuật như one-hot encoding hoặc label encoding để chuyển các biến phân loại thành dữ liệu số.
  - Tạo ra các đặc điểm mới: Tạo ra các đặc điểm mới từ các dữ liệu hiện có, ví dụ như mật độ dân số, khoảng cách đến các tiện ích công cộng, v.v.
  - Chuẩn hóa và tiêu chuẩn hóa dữ liệu: Nếu cần thiết, chuẩn hóa dữ liệu để các đặc điểm có phạm vi giá trị giống nhau, giúp cải thiện hiệu suất của mô hình.
- Chia tập dữ liệu
  - Chia dữ liệu thành tập huấn luyện và tập kiểm tra: Thông thường, tập dữ liệu được chia theo tỷ lệ như 80-20 hoặc 70-30. Có thể sử dụng `train_test_split` của thư viện `scikit-learn` để thực hiện việc này.
- Xây dựng và đánh giá mô hình
  - Xây dựng và sử dụng mô hình hồi quy tuyến tính để dự đoán giá nhà
  - Sử dụng các chỉ số đánh giá như: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared ( $R^2$ ),... để đánh giá độ chính xác và hiệu suất của mô hình.
  - So sánh kết quả với các phương pháp dự đoán khác để xác định hiệu quả của hệ thống.
  - Thử nghiệm trên các mẫu dữ liệu thực tế để kiểm chứng khả năng ứng dụng của mô hình trong điều kiện thực tế.
- Phát triển hệ thống ứng dụng

- Xây dựng một hệ thống phần mềm dự đoán giá nhà có khả năng sử dụng trong thực tế.
- Tích hợp giao diện thân thiện với người dùng, hỗ trợ dự đoán nhanh chóng và chính xác.

## **2.2. Mô hình cây quyết định (Decision tree)**

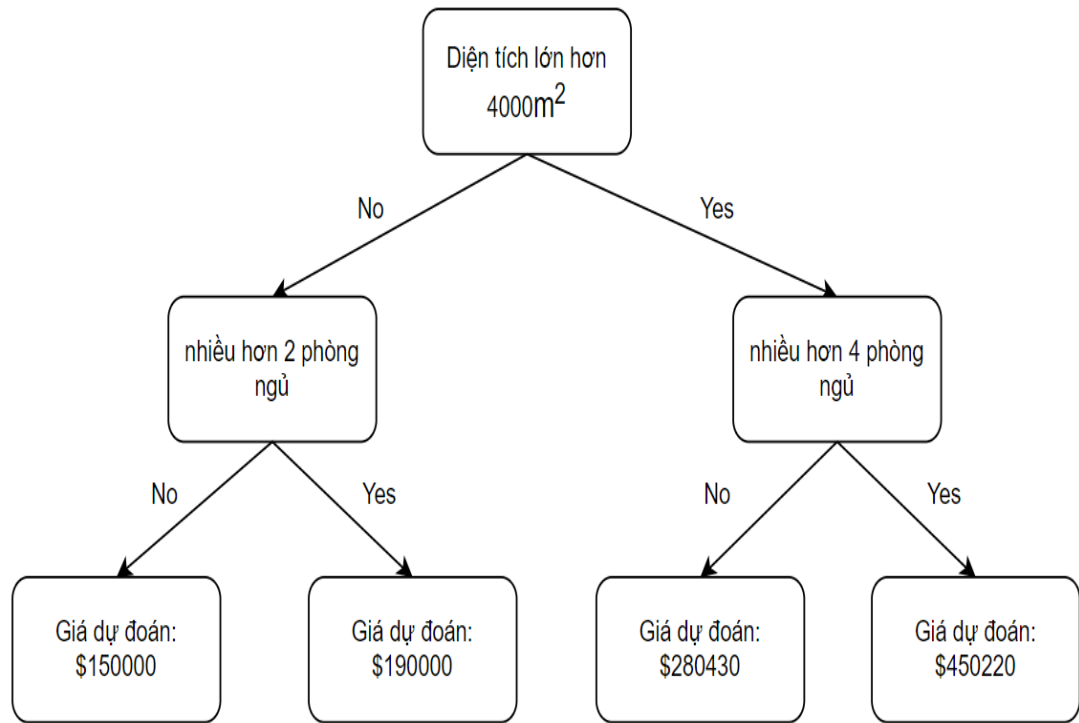
### **2.2.1 Giới thiệu**

Cây quyết định là một phương pháp phân loại và dự đoán phổ biến trong lĩnh vực học máy (Machine Learning) và khai phá dữ liệu (Data Mining). Thuật toán hoạt động bằng cách học các quy tắc quyết định đơn giản được suy ra từ các đặc điểm dữ liệu. Sau đó, quy tắc này có thể được sử dụng để dự đoán giá trị của mục tiêu cho mẫu dữ liệu mới.

Cây quyết định được biểu diễn dưới dạng cấu trúc cây, trong đó mỗi nút bên trong biểu diễn một thuộc tính, mỗi nhánh biểu diễn một quy tắc quyết định và mỗi nút lá biểu diễn một dự đoán. Thuật toán hoạt động bằng cách đệ quy chia dữ liệu thành các tập con ngày càng nhỏ hơn dựa trên các giá trị thuộc tính. Tại mỗi nút, thuật toán chọn thuộc tính chia dữ liệu tốt nhất thành các nhóm có giá trị mục tiêu khác nhau.

Tóm lại, cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (Classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các dữ liệu chưa biết.

Ta sẽ xét một ví dụ về cây quyết định. Giả sử dựa theo diện tích, số phòng ngủ, và giá bán của ngôi nhà. Mục tiêu là dựa trên các đặc điểm này và dự đoán giá nhà của một mẫu dữ liệu mới?



Hình 2.1. Ví dụ về mô hình cây quyết định

Dựa theo mô hình trên, ta có thể thấy rằng nếu căn nhà có diện tích lớn hơn 4000m<sup>2</sup> và có nhiều hơn 4 phòng ngủ thì căn nhà sẽ được bán với giá cao hơn so với căn nhà có diện tích nhỏ hơn 4000m<sup>2</sup> và có nhiều hơn 2 phòng ngủ.

### 2.2.2 Đặc điểm

- Các thành phần của cây quyết định:

- + Node gốc: Là nút trên cùng trong cây, biểu diễn toàn bộ tập dữ liệu. Đây là khởi đầu của quá trình ra quyết định.
- + Node nội bộ: Một nút tượng trưng cho sự lựa chọn liên quan đến một thuộc tính đầu vào. Phân nhánh các nút nội bộ kết nối chúng với các nút lá hoặc các nút nội bộ khác.
- + Node lá / Node đầu cuối: Một nút không có bất kỳ nút con nào biểu thị nhãn lớp hoặc giá trị số.

- + Node cha: Nút chia thành một hoặc nhiều nút con.
- + Node con: Các nút xuất hiện khi nút cha bị tách

#### - Thuật toán ID3

- + Thuật toán ID3 là một trong những thuật toán đầu tiên và được sử dụng nhiều nhất, được Ross Quinlan tạo ra vào năm 1986. Thuật toán ID3 xây dựng một cây quyết định từ một tập dữ liệu nhất định bằng phương pháp tham lam, từ trên xuống dưới.
- + Thuật toán hoạt động bằng cách sử dụng phương pháp tham lam chọn thuộc tính tối đa hóa mức tăng thông tin tại mỗi node. ID3 tính toán Entropy và mức tăng thông tin cho mỗi thuộc tính và chọn thuộc tính có mức tăng thông tin cao nhất để phân tách.
- + Thuật toán ID3 sử dụng Entropy để đo lường sự không chắc chắn hoặc rối loạn trong một tập dữ liệu và Information Gain để đánh giá và lựa chọn các thuộc tính từ đó tối ưu hóa việc phân chia dữ liệu và xây dựng mô hình cây quyết định hiệu quả hơn, dẫn đến những dự đoán chính xác hơn.
- + Entropy ( $H(D)$ ) là thước đo mức độ tinh khiết trong tập dữ liệu, được tính theo công thức:

$$H(D) = - \sum_{i=1}^c P_i \log_2 P_i$$

Trong đó:

- $H(D)$  : Entropy của tập dữ liệu D.
- $c$  : Số lượng lớp (Classes) trong tập dữ liệu.
- $P_i$  : Xác suất của lớp  $i$  trong tập dữ liệu D, được tính bằng cách chia số lượng mẫu trong lớp  $i$  cho tổng số mẫu trong D.



- + Information gain (  $IG(A)$  ) là độ thu lợi về thông tin, độ quan trọng về 1 thuộc tính về mặt thông tin, được tính theo công thức:

$$\text{Gain}(S, A) = H(S) - H(S|A)$$

Trong đó:

- $H(S)$  : Entropy của tập dữ liệu  $S$  trước khi phân chia
- $H(S|A)$  : Entropy của tập dữ liệu  $S$  sau khi phân chia theo thuộc tính  $A$ .

#### - Thuật toán C4.5

Thuật toán C4.5 là thuật toán được Ross Quinlan cải tiến từ thuật toán ID3. Trong các ứng dụng học máy và khai thác dữ liệu, đây là một phương pháp được ưa chuộng để tạo cây quyết định. C4.5 được tạo ra để khắc phục một số nhược điểm của ID3, bao gồm khả năng xử lý các điểm liên tục, thực hiện cắt tỉa cây để giảm xu hướng overfitting với tập huấn luyện và sử dụng tỉ lệ tăng thông tin ( Information gain ratio) thay vì tăng thông tin ( Information gain) như ID3, giúp giảm thiểu thiên lệch với thuộc tính có nhiều giá trị.

#### - Một số thuật toán khác

Bên cạnh ID3, C4.5, mô hình cây quyết định còn một số thuật toán khác như:

- + Thuật toán CHAID: tạo cây quyết định bằng cách sử dụng kiểm định Chi-square để xác định phân chia tốt nhất, phù hợp nhất cho các dữ liệu có phân phối không đồng đều hoặc không liên tục.

- + Thuật toán CART: sử dụng tạp chất Gini để phân loại. Khi chọn một thuộc tính để phân chia, thuật toán sẽ tính toán tạp chất Gini cho mỗi phân chia có thể và chọn phân chia có tạp chất thấp nhất.

- + MARS

- + Conditional Inference Trees

### **2.2.3 Ưu điểm và nhược điểm**

Cây quyết định là một thuật toán phổ biến trong lĩnh vực học máy và khai thác dữ liệu, được sử dụng rộng rãi cho các bài toán phân loại và dự đoán. Cây quyết định cũng có những ưu điểm và nhược điểm cần cân nhắc trước khi lựa chọn và triển khai:

- Ưu điểm:

- + Mô hình dễ dàng triển khai và có thể xây dựng nhanh chóng ngay cả với những người không chuyên sâu về học máy.

- + Mô hình sinh ra các quy tắc dễ hiểu cho người sử dụng và người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.

- + Có thể xử lý cả dữ liệu phân loại và dữ liệu liên tục, linh hoạt trong nhiều ứng dụng khác nhau.

- + Dễ dàng nắm bắt các quan hệ phi tuyến tính giữa các biến trong dữ liệu.

- + Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê

- + Có khả năng làm việc với dữ liệu lớn.

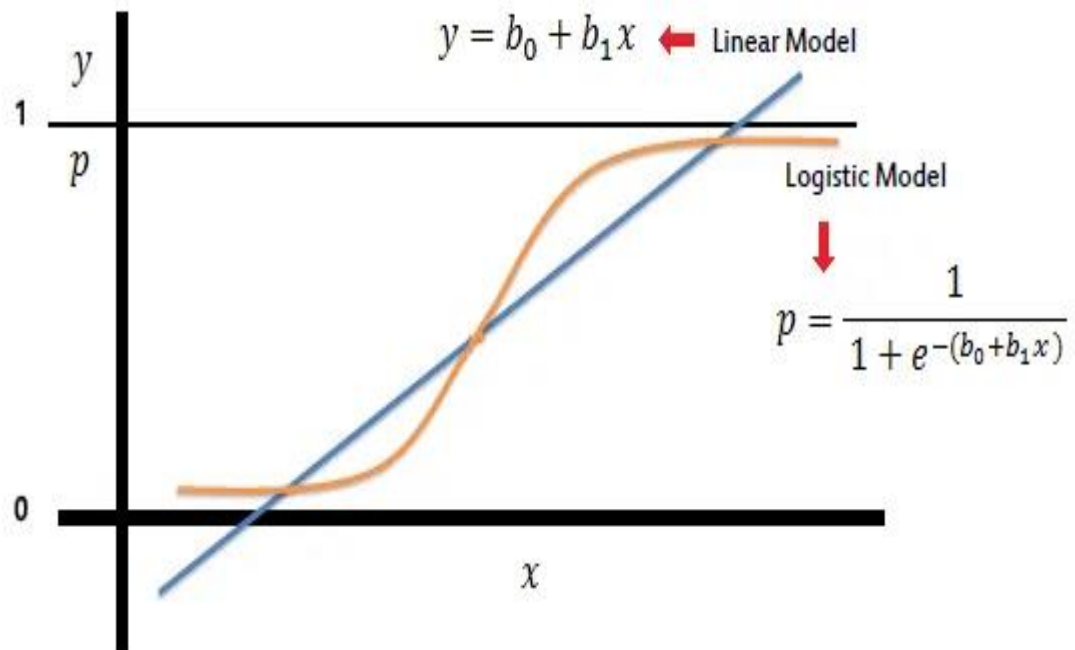
- Nhược điểm:

- + Cây quyết định có xu hướng tạo ra cây rất phức tạp khi cố gắng phân chia để đạt độ chính xác tối đa, dẫn đến overfitting, khiến mô hình không tổng quát tốt khi áp dụng trên dữ liệu mới.
- + Không ổn định và nhạy cảm với các thay đổi của dữ liệu, dẫn đến sự thay đổi đáng kể trong cấu trúc cây.
- + Tốn kém khi tính toán với dữ liệu lớn.

## **2.3. Mô hình hồi quy Logistic**

### **2.3.1. Giới thiệu**

Hồi quy logistic là một thuật toán học máy được sử dụng rộng rãi để giải quyết các bài toán phân loại nhị phân. Mô hình này dự đoán xác suất một sự kiện xảy ra, thường được biểu diễn dưới dạng một nhãn nhị phân (ví dụ: 0 hoặc 1, có hoặc không). Không giống như hồi quy tuyến tính dự đoán giá trị liên tục, hồi quy logistic đưa ra dự đoán về xác suất thuộc về một lớp nhất định.



Hình 2.2. Thuật toán hồi quy Logistic

### 2.3.2. Đặc điểm

- Hàm sigmoid: Hàm sigmoid là yếu tố cốt lõi của hồi quy logistic. Hàm này ánh xạ các giá trị đầu vào vào khoảng từ 0 đến 1, đại diện cho xác suất.
- Ngưỡng quyết định: Để đưa ra dự đoán cuối cùng, ta thường đặt một ngưỡng (thường là 0.5). Nếu xác suất dự đoán lớn hơn ngưỡng, ta kết luận dữ liệu thuộc về lớp dương, ngược lại thuộc về lớp âm.
- Tối ưu hóa: Mô hình tìm kiếm các tham số của hàm sigmoid bằng cách tối ưu hóa một hàm mất mát, thường là hàm cross-entropy.
- Phân loại nhị phân: Hồi quy logistic chủ yếu được sử dụng để giải quyết các bài toán phân loại nhị phân, tức là chỉ có hai lớp kết quả. Tuy nhiên, có

thể mở rộng để xử lý các bài toán đa lớp bằng các kỹ thuật như one-vs-rest hoặc one-vs-one.

### **2.3.3. Phân loại hồi quy Logistic**

- Phân loại hồi quy logistic là quá trình sử dụng mô hình hồi quy logistic để gán nhãn cho dữ liệu mới. Quá trình này bao gồm:

- Tính toán xác suất: Đưa dữ liệu mới vào mô hình để tính toán xác suất thuộc về từng lớp.

- So sánh với ngưỡng: So sánh xác suất tính được với ngưỡng đã đặt trước.

- Phân loại: Gán nhãn cho dữ liệu dựa trên kết quả so sánh.

Ví dụ: Trong bài toán dự đoán khách hàng có mua sản phẩm hay không, ta tính toán xác suất khách hàng đó mua sản phẩm. Nếu xác suất lớn hơn 0.5, ta kết luận khách hàng sẽ mua sản phẩm.

### **2.3.4. Ưu điểm và nhược điểm**

- Ưu điểm:

- Dễ hiểu: Mô hình tương đối dễ hiểu và giải thích.

- Hiệu quả: Có thể xử lý lượng lớn dữ liệu một cách hiệu quả.

- Ít nhạy cảm với outliers: Mô hình tương đối ổn định với các giá trị ngoại lai.

- Đa dạng ứng dụng: Được sử dụng rộng rãi trong nhiều lĩnh vực như y tế, tài chính, marketing.

- Nhược điểm :

- Giả định tuyến tính: Mô hình giả định mối quan hệ tuyến tính giữa các biến độc lập và biến phụ thuộc.

- Khó xử lý dữ liệu phi tuyến: Với dữ liệu có mối quan hệ phi tuyến phức tạp, hồi quy logistic có thể không đạt hiệu quả cao.
- Khó xử lý dữ liệu nhiều lớp: Để xử lý các bài toán phân loại đa lớp, cần sử dụng các kỹ thuật mở rộng.

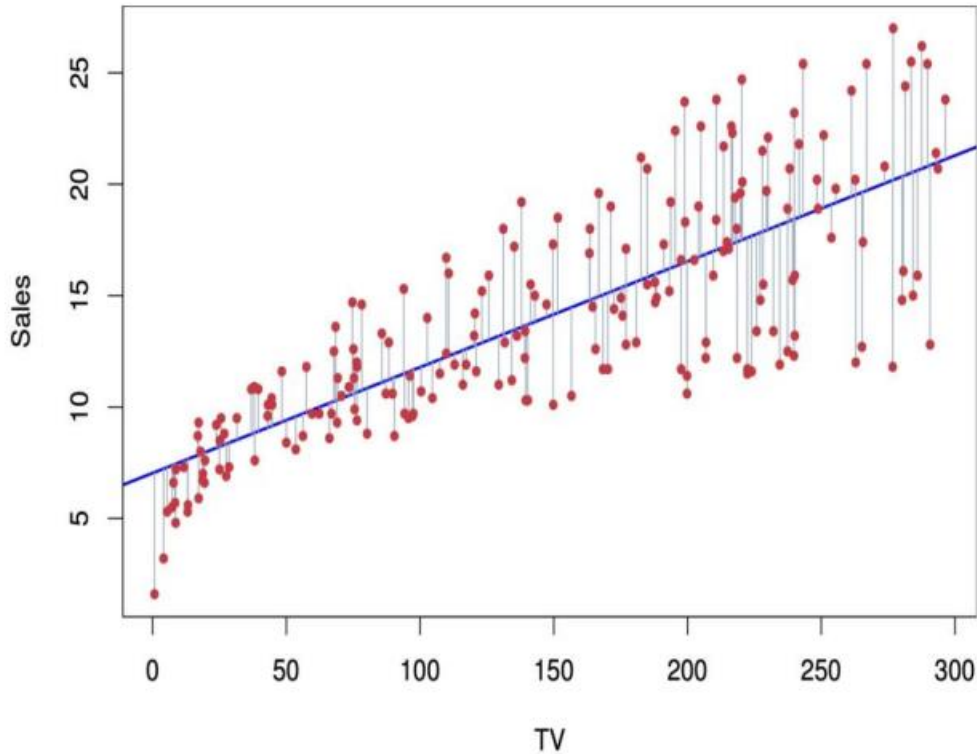
### **2.3.5. Ứng dụng**

- Phân loại email: Phân biệt email spam và không spam.
- Phát hiện gian lận: Phát hiện các giao dịch thẻ tín dụng gian lận.
- Phân loại văn bản: Phân loại các bài viết vào các chủ đề khác nhau.
- Dự đoán bệnh: Dự đoán khả năng một người mắc bệnh dựa trên các triệu chứng.
- Phân tích cảm xúc: Phân loại các bình luận trên mạng xã hội thành tích cực hoặc tiêu cực

## **2.4. Mô hình hồi quy tuyến tính**

### **2.4.1. Giới thiệu**

Hồi quy tuyến tính là một thuật toán học có giám sát (supervised learning). Trong học máy, nó là một phương pháp thống kê dùng để ước lượng mối quan hệ giữa các biến độc lập (input features) và biến phụ thuộc (output target). Hồi quy tuyến tính giả định rằng sự tương quan giữa các biến là tuyến tính, từ đó tìm ra hàm tuyến tính tốt nhất để biểu diễn mối quan hệ này. Thuật toán này dự báo giá trị của biến đầu ra từ các giá trị của các biến đầu vào.



Hình 2.3. Ví dụ thuật toán hồi quy tuyến tính

Hồi quy tuyến tính là một phương pháp thống kê mạnh mẽ và phổ biến được sử dụng để mô hình hóa mối quan hệ tuyến tính giữa một biến phụ thuộc và một hoặc nhiều biến độc lập. Phương pháp này có khả năng dự đoán và giải thích sự biến đổi của biến phụ thuộc dựa trên biến độc lập.

#### 2.4.2. Các thành phần của hồi quy tuyến tính

- Hồi quy tuyến tính dựa trên giả định rằng mối quan hệ giữa biến phụ thuộc và biến độc lập có thể được mô tả bằng một hàm tuyến tính.
- Công thức của mô hình hồi quy tuyến tính bội:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

Trong đó:

- Y là biến phụ thuộc mà chúng ta muốn dự đoán hoặc giải thích.

- $X_1, X_2, \dots, X_n$  là các biến độc lập được sử dụng để dự đoán  $Y$ .
  - $\beta_0, \beta_1, \beta_2, \dots, \beta_n$  là hệ số hồi quy ứng với từng biến độc lập, biểu thị độ ảnh hưởng của chúng lên biến phụ thuộc.
  - $\varepsilon$  là sai số ngẫu nhiên, biểu thị các yếu tố không thể dự đoán được trong mô hình ( $\varepsilon$  thực tế không tính được).
- Công thức hồi quy tuyến tính đơn giản: Trong hồi quy tuyến tính đơn giản, chúng ta xem xét mối quan hệ tuyến tính giữa một biến phụ thuộc  $Y$  và một biến độc lập  $X$ . Công thức hồi quy tuyến tính đơn giản có dạng:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Trong đó:

- $Y$  là biến phụ thuộc (cũng được gọi là biến mục tiêu hoặc biến phản hồi).
- $X$  là biến độc lập (cũng được gọi là biến giải thích hoặc biến đầu vào).
- $\beta_0$  và  $\beta_1$  là các hệ số hồi quy (cũng được gọi là hệ số góc và hệ số chặn).
- $\varepsilon$  là sai số ngẫu nhiên (cũng được gọi là sai số hồi quy).

### 2.4.3. Đánh giá và diễn giải kết quả

Để đánh giá hiệu suất của mô hình hồi quy tuyến tính, các thước đo thường được sử dụng bao gồm độ chính xác (accuracy), độ chính xác điều chỉnh (adjusted accuracy), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), R-squared, và F-test. Các thước đo này giúp đánh giá mức độ phù hợp của mô hình và khả năng giải thích biến động của biến phụ thuộc.

Diễn giải kết quả của mô hình hồi quy tuyến tính đòi hỏi sự hiểu biết về ngữ nghĩa của các biến độc lập và hệ số hồi quy. Các hệ số hồi quy  $\beta_1, \beta_2, \dots, \beta_n$  cho biết độ ảnh hưởng tương ứng của các biến độc lập lên biến phụ thuộc. Nếu hệ số dương, có thể kết luận rằng có một mối quan hệ tương quan dương



giữa biến độc lập và biến phụ thuộc. Nếu hệ số âm, có thể kết luận rằng có một mối quan hệ tương quan âm giữa biến độc lập và biến phụ thuộc

#### **2.4.4. Ứng dụng của mô hình hồi quy tuyến tính**

Mô hình hồi quy tuyến tính có rất nhiều ứng dụng thực tế trong các lĩnh vực khác nhau, nhờ vào tính đơn giản, dễ hiểu và khả năng giải thích rõ ràng mối quan hệ giữa các biến. Dưới đây là các ứng dụng tiêu biểu của mô hình này:

- **Kinh doanh & Marketing:**
  - + **Phân khúc khách hàng:** Xác định các nhóm khách hàng có đặc điểm tương đồng dựa trên hành vi mua hàng, sở thích, nhân khẩu học.
  - + **Phân tích hiệu quả quảng cáo:** Đo lường tác động của các chiến dịch quảng cáo trên doanh số bán hàng, nhận thức thương hiệu.
  - + **Dự báo nhu cầu:** Dự đoán nhu cầu sản phẩm/dịch vụ trong tương lai dựa trên các yếu tố như mùa vụ, xu hướng thị trường, hoạt động marketing.
  - + **Định giá sản phẩm:** Xác định giá bán tối ưu cho sản phẩm/dịch vụ dựa trên chi phí sản xuất, giá của đối thủ cạnh tranh, giá trị nhận thức của khách hàng.
- **Y tế & Dược:**
  - + **Dự đoán nguy cơ bệnh tật:** Xác định khả năng mắc bệnh dựa trên các yếu tố nguy cơ như tuổi tác, giới tính, lối sống, tiền sử gia đình.
  - + **Phân tích hiệu quả thuốc:** Đánh giá tác dụng của thuốc trên bệnh nhân dựa trên liều lượng, thời gian điều trị, đặc điểm bệnh nhân.
  - + **Dự báo chi phí y tế:** Ước tính chi phí điều trị dựa trên loại bệnh, độ phức tạp, thời gian điều trị.

- Khoa học & Công nghệ:
  - + Dự báo thời tiết: Dự đoán nhiệt độ, lượng mưa, gió dựa trên dữ liệu lịch sử, áp suất khí quyển, độ ẩm.
  - + Phân tích dữ liệu thí nghiệm: Xác định mối quan hệ giữa các biến trong thí nghiệm, đánh giá tác động của các yếu tố điều khiển.
  - + Dự đoán lỗi phần mềm: Xác định khả năng xảy ra lỗi phần mềm dựa trên các yếu tố như kích thước mã nguồn, độ phức tạp, lịch sử lỗi.
- Xã hội & Nhân văn:
  - + Phân tích dữ liệu kinh tế: Nghiên cứu mối quan hệ giữa các chỉ số kinh tế như GDP, lạm phát, tỷ lệ thất nghiệp.
  - + Phân tích dữ liệu xã hội: Nghiên cứu mối quan hệ giữa các yếu tố xã hội như thu nhập, giáo dục, tình trạng hôn nhân với các chỉ số sức khỏe, hạnh phúc.
  - + Dự đoán xu hướng xã hội: Dự đoán xu hướng thay đổi trong xã hội dựa trên dữ liệu về dân số, văn hóa, công nghệ.

#### **2.4.5. Ưu điểm và nhược điểm**

Hồi quy tuyến tính là một phương pháp quan trọng trong lĩnh vực thống kê và học máy để mô hình hóa mối quan hệ tuyến tính giữa các biến đầu vào và biến mục tiêu. Dưới đây là các ưu điểm và hạn chế của hồi quy tuyến tính:

- Ưu điểm:
  - + Đơn giản và dễ hiểu: Hồi quy tuyến tính là một phương pháp đơn giản và dễ hiểu để mô hình hóa mối quan hệ đơn tuyến tính giữa biến phụ thuộc và biến độc lập

- + Tính linh hoạt: Hồi quy tuyến tính có thể được áp dụng cho nhiều biến độc lập và có thể dễ dàng mở rộng để xem xét tác động của nhiều biến độc lập lên biến phụ thuộc
- + Dễ thực hiện: Có nhiều phương pháp ước lượng trong hồi quy tuyến tính, bao gồm phương pháp bình phương tối thiểu (OLS) được sử dụng rộng rãi và có thể tính toán dễ dàng.
- + Tính khả diễn giải: Hồi quy tuyến tính cung cấp các hệ số hồi quy có ý nghĩa thống kê và khả năng giải thích tương đối cho tác động của các biến độc lập lên biến phụ thuộc
- Nhược điểm:
  - + Giả định về tuyến tính: Hồi quy tuyến tính giả định mối quan hệ giữa biến phụ thuộc và biến độc lập là tuyến tính. Trong trường hợp không tuyến tính, mô hình hồi quy tuyến tính có thể không phù hợp và không cho kết quả chính xác.
  - + Nhạy cảm với nhiễu và quan sát ngoại lai: Hồi quy tuyến tính có thể bị ảnh hưởng bởi nhiễu và quan sát ngoại lai trong dữ liệu, và có thể dẫn đến ước lượng không chính xác và không ổn định.
  - + Không xử lý được tương quan và đa cộng tuyến: Hồi quy tuyến tính không xử lý được vấn đề tương quan cao giữa các biến độc lập hoặc đa cộng tuyến, khiến cho ước lượng hệ số hồi quy trở nên không chính xác và không đáng tin cậy.
  - + Giới hạn trong mô hình hóa phức tạp: Hồi quy tuyến tính có giới hạn trong việc mô hình hóa các mối quan hệ phi tuyến và phức tạp. Các mô hình tuyến tính không thể mô hình hóa các mẫu không tuyến tính phức tạp như đường cong, đường cong S, hay tương tác phi tuyến giữa các biến.

## 2.5. Lựa chọn mô hình giải quyết bài toán

Phương pháp phân tích hồi quy tuyến tính (Linear Regression) là sự kết hợp của tính đơn giản, khả năng ước lượng mối quan hệ tuyến tính, khả năng dự báo cùng với khả năng phân tích định lượng. Phương pháp này sẽ giúp ta đạt được mục tiêu nghiên cứu và trả lời những câu hỏi quan trọng. Vì vậy em lựa chọn phương pháp hồi quy tuyến tính để giải quyết đề bài.

## CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

### 3.1. Môi trường thực nghiệm

- Google Colab là một công cụ mạnh mẽ để phát triển, thử nghiệm và triển khai các mô hình học máy, bao gồm cả việc huấn luyện và dự đoán với mô hình hồi quy tuyến tính.

Google Colaboratory



*Hình 3.1. Google Colab*

- Ngôn ngữ lập trình: Python 3.7 và các thư viện hỗ trợ khác như Sklearn, Numpy, Pandas, Matplotlib và Seaborn.



*Hình 3.2. Ngôn ngữ lập trình Python*

- Lưu trữ: Google Drive, dịch vụ lưu trữ đám mây, cho phép người dùng lưu trữ và truy cập các tệp tin trực tuyến.



*Hình 3.3. Google Drive*

### **3.2. Dữ liệu thực nghiệm**

Bộ dữ liệu sử dụng trong nghiên cứu này được thu thập từ cuộc điều tra dân số thực tế ở California năm 1990. Dữ liệu liên quan được thu thập trong một quận nhất định của California và một số số liệu thống kê tóm tắt về chúng dựa trên cuộc điều tra dân số.

Bộ dữ liệu được phân tích ở đây là file dataset (.csv) có tên cụ thể là housing.csv, chứa 20640 bản ghi thông tin về giá nhà California.

Thông tin cụ thể như sau:

- Tên bộ dữ liệu: California Housing Prices
- Nguồn: <https://www.kaggle.com/datasets/camnugent/california-housing-prices/data>

- Dữ liệu 15 dòng đầu của dataset:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
1	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	83.252	452600.0	NEAR BAY
2	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	83.014	358500.0	NEAR BAY
3	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	72.574	352100.0	NEAR BAY
4	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	56.431	341300.0	NEAR BAY
5	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	38.462	342200.0	NEAR BAY
6	-122.25	37.85	52.0	919.0	213.0	413.0	193.0	40.368	269700.0	NEAR BAY
7	-122.25	37.84	52.0	2535.0	489.0	1094.0	514.0	36.591	299200.0	NEAR BAY
8	-122.25	37.84	52.0	3104.0	687.0	1157.0	647.0	3.12	241400.0	NEAR BAY
9	-122.26	37.84	42.0	2555.0	665.0	1206.0	595.0	20.804	226700.0	NEAR BAY
10	-122.25	37.84	52.0	3549.0	707.0	1551.0	714.0	36.912	261100.0	NEAR BAY
11	-122.26	37.85	52.0	2202.0	434.0	910.0	402.0	32.031	281500.0	NEAR BAY
12	-122.26	37.85	52.0	3503.0	752.0	1504.0	734.0	32.705	241800.0	NEAR BAY
13	-122.26	37.85	52.0	2491.0	474.0	1098.0	468.0	3.075	213500.0	NEAR BAY
14	-122.26	37.84	52.0	696.0	191.0	345.0	174.0	26.736	191300.0	NEAR BAY
15	-122.26	37.85	52.0	2643.0	626.0	1212.0	620.0	19.167	159200.0	NEAR BAY
16	-122.26	37.85	52.0	1136.0	182.0	507.0	254.0	2.125	140000.0	NEAR BAY

Hình 3.4. 15 dòng đầu của bộ dữ liệu gốc

- Thông tin cụ thể các cột của dataset như sau:

**longitude:** Kinh độ của vị trí địa lý.

**latitude:** Vĩ độ của vị trí địa lý.

**housing\_median\_age:** Tuổi trung bình của nhà ở trong khu vực.

**total\_rooms:** Tổng số phòng trong khu vực nhà ở.

**total\_bedrooms:** Tổng số phòng ngủ trong khu vực nhà ở.

**population:** Dân số trong khu vực nhà ở.

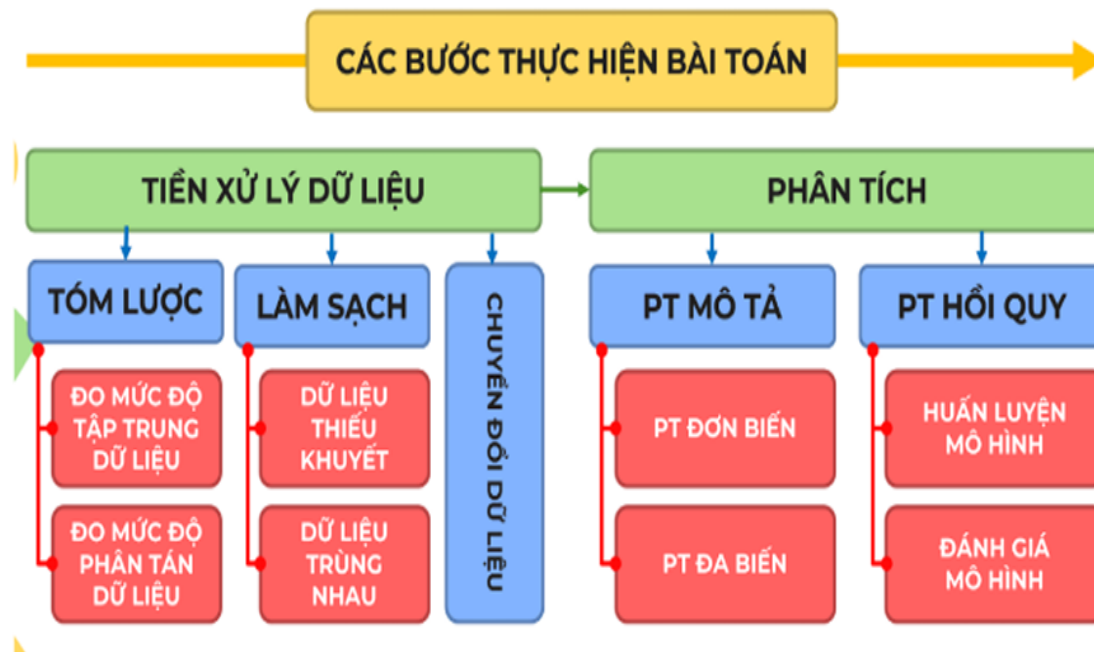
**households:** Số hộ gia đình trong khu vực nhà ở.

**median\_income:** Thu nhập trung bình của các hộ gia đình trong khu vực

**median\_house\_value:** Giá trị trung bình của các ngôi nhà trong khu vực

**ocean\_proximity:** Khoảng cách đến đại dương .

### 3.3. Quy Trình thực nghiệm



Hình 3.5. Quy trình thực nghiệm đề tài phân tích dữ liệu

### 3.3.1. Đặt mục tiêu

- Phân tích mô tả để thể hiện mối quan hệ giữa các giá trị của dữ liệu, từ đó đánh giá được tương quan của giá nhà .
- Phân tích hồi quy để dự báo giá nhà dựa theo mô hình hồi quy tuyến tính.

### 3.3.2. Tiền xử lý dữ liệu

- Làm sạch dữ liệu:

Làm sạch dữ liệu là quá trình loại bỏ các sai sót, lỗi, nhiễu và thông tin không chính xác hoặc không cần thiết khỏi tập dữ liệu ban đầu để đảm bảo dữ liệu đáng tin cậy và phù hợp cho việc phân tích và xử lý tiếp theo. Quá trình



làm sạch dữ liệu thường là một phần quan trọng trong tiền xử lý dữ liệu trước khi bắt đầu phân tích mô tả và cả phân tích hồi quy.

Một số tác vụ chính trong quá trình làm sạch dữ liệu bao gồm:

- Loại bỏ dữ liệu trùng lặp: Loại bỏ các bản ghi bị trùng lặp trong tập dữ liệu để tránh ảnh hưởng đến kết quả phân tích.
- Xử lý dữ liệu thiếu: Điền vào các giá trị thiếu hoặc quyết định loại bỏ chúng dựa trên ngữ cảnh và mục tiêu của phân tích.
- Xử lý giá trị ngoại lai : Xử lý giá trị ngoại lai bao gồm việc phát hiện và xử lý các điểm dữ liệu bất thường, nằm xa phân phối dữ liệu chính, bằng các phương pháp như loại bỏ, thay thế hoặc biến đổi
- Sửa lỗi và sai sót: Điều tra và sửa các lỗi cú pháp, sai sót chính tả hoặc sai sót logic trong dữ liệu.
- Chọn lọc đặc trưng: Xác định và lựa chọn các đặc trưng quan trọng nhất để sử dụng trong phân tích hoặc mô hình hóa.
- Kiểm tra dữ liệu khuyết :

	tỉ lệ khuyết
longitude	0.000000
latitude	0.000000
housing_median_age	0.000000
total_rooms	0.000000
total_bedrooms	1.002907
population	0.000000
households	0.000000
median_income	0.000000
median_house_value	0.000000
ocean_proximity	0.000000

Hình 3.6. Kiểm tra dữ liệu bị khuyết

- Ta sẽ điền khuyết bằng median :

```
# Xử lý giá trị khuyết với cột total_bedrooms
median = df['total_bedrooms'].median()
df['total_bedrooms'] = df['total_bedrooms'].fillna(median)
# Kiểm tra lại giá trị khuyết
df_missing = pd.DataFrame(df.isnull().sum(), columns=['Giá trị khuyết'])
df_missing
```

*Hình 3.7. Xử lý dữ liệu bị khuyết*

- Xử lý giá trị trùng lặp :

```
# xử lý data trùng lặp
row_duplicated = df.duplicated().sum()
print('\nSố lượng data bị trùng : ',row_duplicated)
# Xóa Data trùng lặp
df = df.drop_duplicates()
```

*Hình 3.8. Xử lý dữ liệu bị trùng*

Số lượng data bị trùng : 0

*Hình 3.9. Kết quả xử lý dữ liệu bị trùng*

- Xử lý giá trị ngoại lai bằng phương pháp IQR

```

# Xử lý giá trị ngoại lai
def xu_ly_gia_tri_ngoai_lai(df, column):

    # Tính toán IQR (Interquartile Range)
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1

    # Xác định giới hạn trên và dưới để loại bỏ giá trị ngoại lai
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # Thay thế các giá trị ngoại lai bằng giới hạn trên và dưới
    df[column] = df[column].clip(lower=lower_bound, upper=upper_bound)
    return df[column]

# Xử lý giá trị ngoại lai cho các cột liên quan đến lương
columns_to_process = ['longitude', 'latitude', 'housing_median_age', 'total_rooms', 'total_bedrooms', 'population', 'households', 'median_income', 'median_house_value']

for column in columns_to_process:
    df[column] = xu_ly_gia_tri_ngoai_lai(df, column)

```

Hình 3.10. Xử lý giá trị ngoại lai

**Chuyển đổi dữ liệu:** Chuyển đổi dữ liệu trong phân tích dữ liệu là quá trình thay đổi cách thức biểu diễn, xử lý hoặc áp dụng các phép toán trên dữ liệu ban đầu để tạo ra dữ liệu mới có ý nghĩa hoặc thuận tiện hơn cho mục đích phân tích. Nó có vai trò quan trọng trong việc biểu diễn trực quan hơn dataset, thuận tiện hơn trong việc phân tích dữ liệu.

Trong project này, ta thấy có cột duy nhất đang ở dạng Object, tức phi số, vì vậy, ta sẽ chuyển hóa cột dữ liệu này bằng dummy.

```

# Kiểm tra số nhãn trong dữ liệu dạng chữ
print('\nKiểm tra số nhãn')
print(df['ocean_proximity'], ' chứa ', len(df['ocean_proximity'].unique()), ' nhãn')
# Mã hóa các nhãn này thành dạng one-hot encoding, các nhãn khác sẽ bị bỏ qua
df_encode = pd.get_dummies(df['ocean_proximity']).astype(int)

# Kết hợp với dữ liệu gốc và loại bỏ cột ocean_proximity
df_new = pd.concat([df.drop(['ocean_proximity'], axis=1), df_encode], axis=1)

```

Hình 3.11. chuyển hóa cột dữ liệu

### 3.3.3. Phân tích mô tả

Phân tích mô tả trong phân tích dữ liệu là quá trình tóm tắt, mô tả và hiểu sâu về các đặc điểm, mẫu thái và thông tin quan trọng của tập dữ liệu. Với mục tiêu đó, ta sẽ tiến hành phân tích mô tả cho bộ dữ liệu của project theo cả 2 hướng phân tích đơn biến (trên từng biến) và phân tích đa biến (trên nhiều biến) bằng cách biểu diễn dưới các biểu đồ khác nhau.

- Tóm lược dữ liệu:

+ Tóm lược dữ liệu trong phân tích dữ liệu là quá trình tổng hợp, trích xuất và trình bày các thông tin quan trọng và chính xác từ tập dữ liệu ban đầu. Mục tiêu của việc tóm lược dữ liệu là giúp người đọc hoặc người xem nắm bắt được những điểm quan trọng và khái quát của dữ liệu mà không cần phải đọc hoặc xem toàn bộ dữ liệu gốc. Tóm lược dữ liệu bao gồm 2 loại đo: Đo mức độ tập trung dữ liệu (mean, median, mode, ...) và Đo mức độ phân tán dữ liệu (quartile, interquartile, standard deviation, ...).

+ Ta sẽ tiến hành tổng hợp các thông tin về độ tập trung và phân tán của dữ liệu. Những thông số này chỉ tương thích với các cột dữ liệu dạng thông số, vậy nên sẽ chỉ có” longitude, latitude, housing\_median\_age, total\_rooms, total\_bedrooms, population, households, median\_income, median\_house\_value” là được phân tích. Dưới đây là kết quả tóm lược dữ liệu bao gồm các thuộc tính count, mean, std, min, 25%, 50%, 75%, max, mode, median của các dữ liệu trên:

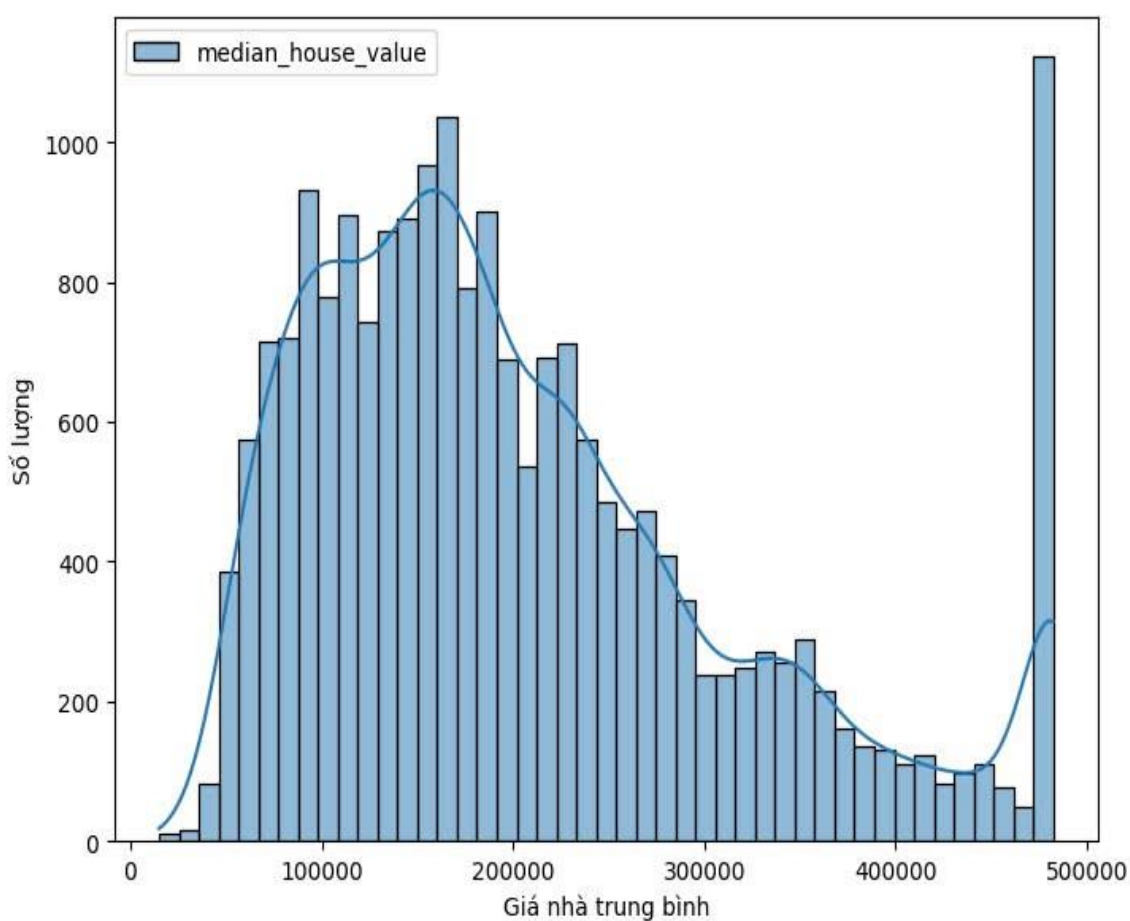
```
# loại các cột <1H OCEAN INLAND ISLAND NEAR BAY NEAR OCEAN để lập bảng mô tả
df_mo_ta = df.drop(columns = ['<1H OCEAN', 'INLAND', 'ISLAND', 'NEAR BAY', 'NEAR OCEAN'])
# Lập bảng mô tả
Bang_mo_ta = df_mo_ta.describe()
Bang_mo_ta
```

Hình 3.12. Tóm lược dữ liệu

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000
mean	-119.569704	35.631861	28.639486	2441.692472	501.182086	1336.959012	469.020107	3.801010	205981.224976
std	2.003532	2.135952	12.585558	1397.790038	284.133641	765.550830	265.507540	1.657658	113217.350152
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.000000
25%	-121.800000	33.930000	18.000000	1447.750000	297.000000	787.000000	280.000000	2.563400	119600.000000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800	179700.000000
75%	-118.010000	37.710000	37.000000	3148.000000	643.250000	1725.000000	605.000000	4.743250	264725.000000
max	-114.310000	41.950000	52.000000	5698.375000	1162.625000	3132.000000	1092.500000	8.013025	482412.500000

Hình 3.13. Bảng tóm lược dữ liệu

Biểu đồ Hist Plot biểu diễn giá nhà trung bình :

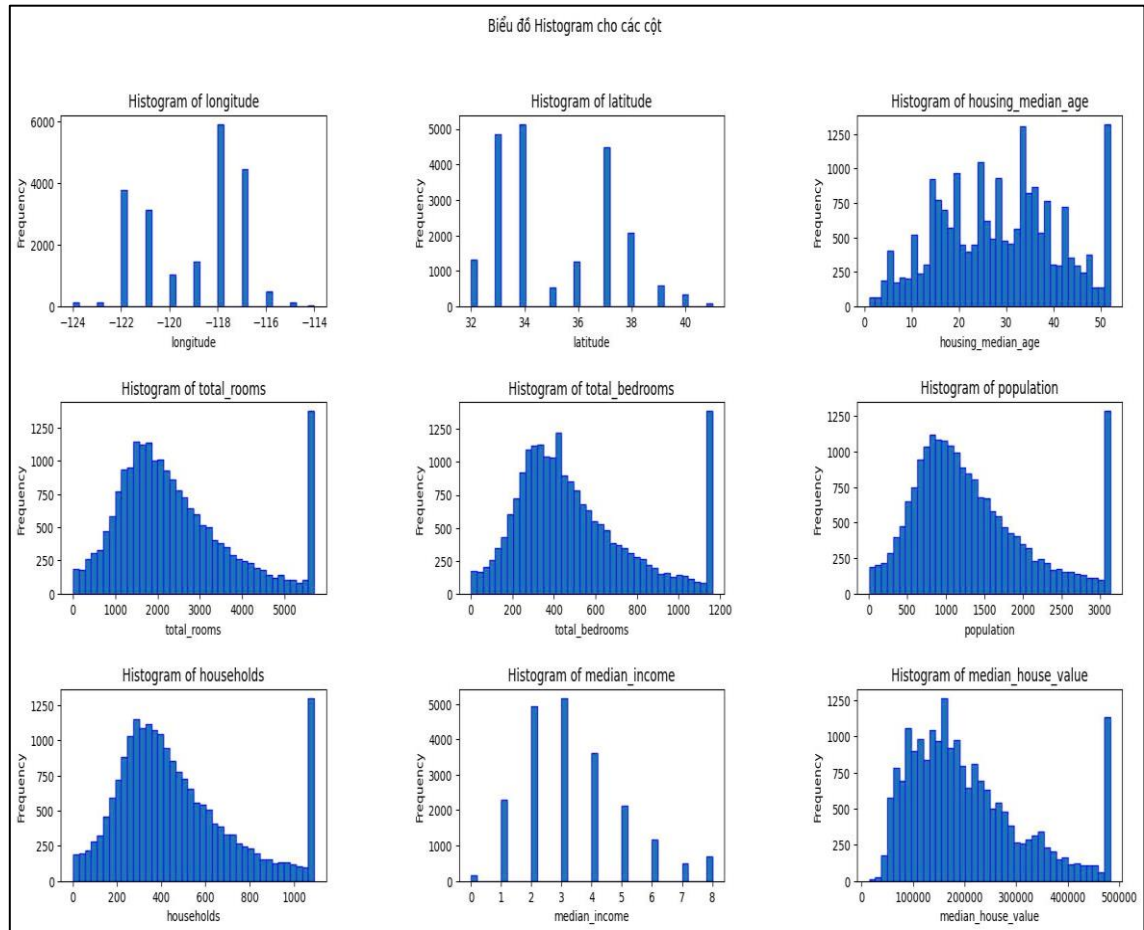


Hình 3.14. Biểu đồ Hist Plot

Biểu đồ hist plot thể hiện phân bố giá trị trung bình của nhà ở. Dữ liệu có xu hướng tập trung nhiều nhất ở khoảng giá trị từ 100,000 đến 200,000. Phân bố không hoàn toàn đối xứng, có vẻ hơi lệch phải, với một số lượng đáng

kể các giá trị cao hơn ở phía trên 400,000 và một số ít giá trị rất thấp. Đường cong mật độ giúp trực quan hóa xu hướng này rõ hơn.

Biểu đồ Histogram cho các cột:



Hình 3.15. Biểu đồ Histogram cho các cột

Nhận xét:

- **Histogram of longitude:** Phân bố không đồng đều, cho thấy sự tập trung của dữ liệu ở một vài khu vực kinh độ cụ thể, có thể là do mật độ dân số hoặc

các yếu tố địa lý khác. Có một đỉnh chính ở khoảng -118, cho thấy một phần lớn dữ liệu tập trung tại khu vực kinh độ đó.

- Histogram of latitude: Tương tự như kinh độ, phân bố không đều, tập trung ở một số khu vực vĩ độ nhất định. Cho thấy một sự phân bố không đồng đều về mặt địa lý. Các đỉnh cao cho thấy khu vực có nhiều dữ liệu hơn.

- Histogram of housing\_median\_age: Phân bố gần như đối xứng, có một phần nhỏ lệch phải, tập trung nhiều nhất ở giữa (khoảng 25-35 tuổi). Điều này cho thấy nhiều nhà ở có tuổi đời trung bình.

- Histogram of total\_rooms: Phân bố lệch phải rõ rệt, nhiều nhà ở có ít phòng hơn, và có một số lượng nhỏ nhà ở có số lượng phòng rất lớn. Chứng tỏ sự bất bình đẳng về quy mô nhà ở.

- Histogram of total\_bedrooms: Tương tự như total\_rooms, phân bố lệch phải, cho thấy nhiều nhà ở có ít phòng ngủ hơn và ít nhà ở có số lượng phòng ngủ rất lớn.

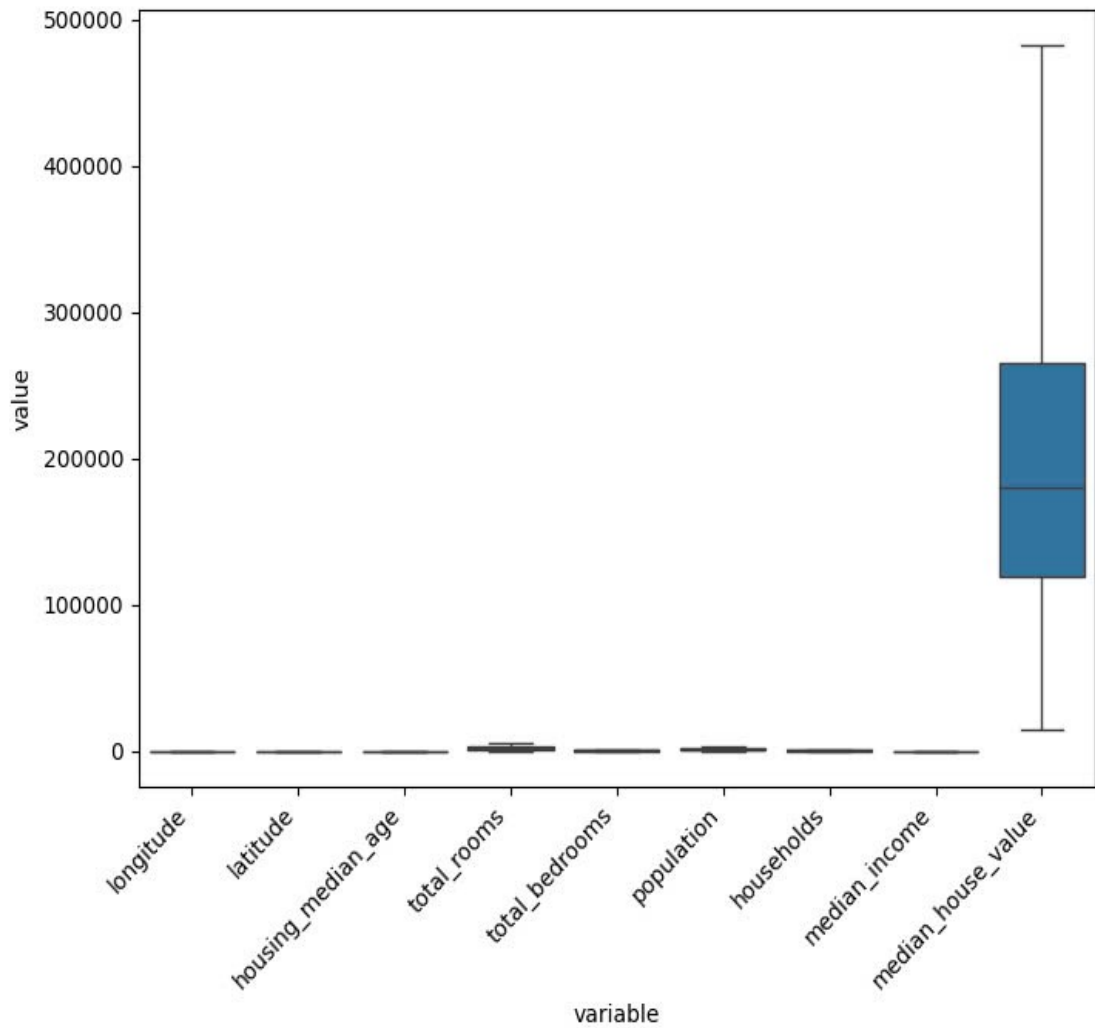
- Histogram of population: Phân bố lệch phải. Hầu hết các khu vực có dân số tương đối nhỏ, với một số ít khu vực có dân số lớn hơn nhiều.

- Histogram of households: Cũng có phân bố lệch phải, tương tự total\_rooms và total\_bedrooms, cho thấy số hộ gia đình nhỏ chiếm đa số, và số hộ gia đình lớn hơn khá ít.

- Histogram of median\_income: Phân bố đa đỉnh, cho thấy một số mức thu nhập trung bình phổ biến trong bộ dữ liệu. Điều này cho thấy sự phân tầng rõ rệt về thu nhập.

- Histogram of median\_house\_value: Phân bố lệch phải, cho thấy nhiều nhà ở có giá trị trung bình tương đối thấp, và một số ít nhà ở có giá trị rất cao. Đây là một phân bố điển hình của giá nhà ở, thường lệch phải do sự tồn tại của một số lượng nhỏ nhà ở có giá trị rất cao.

Biểu đồ Box Plot :

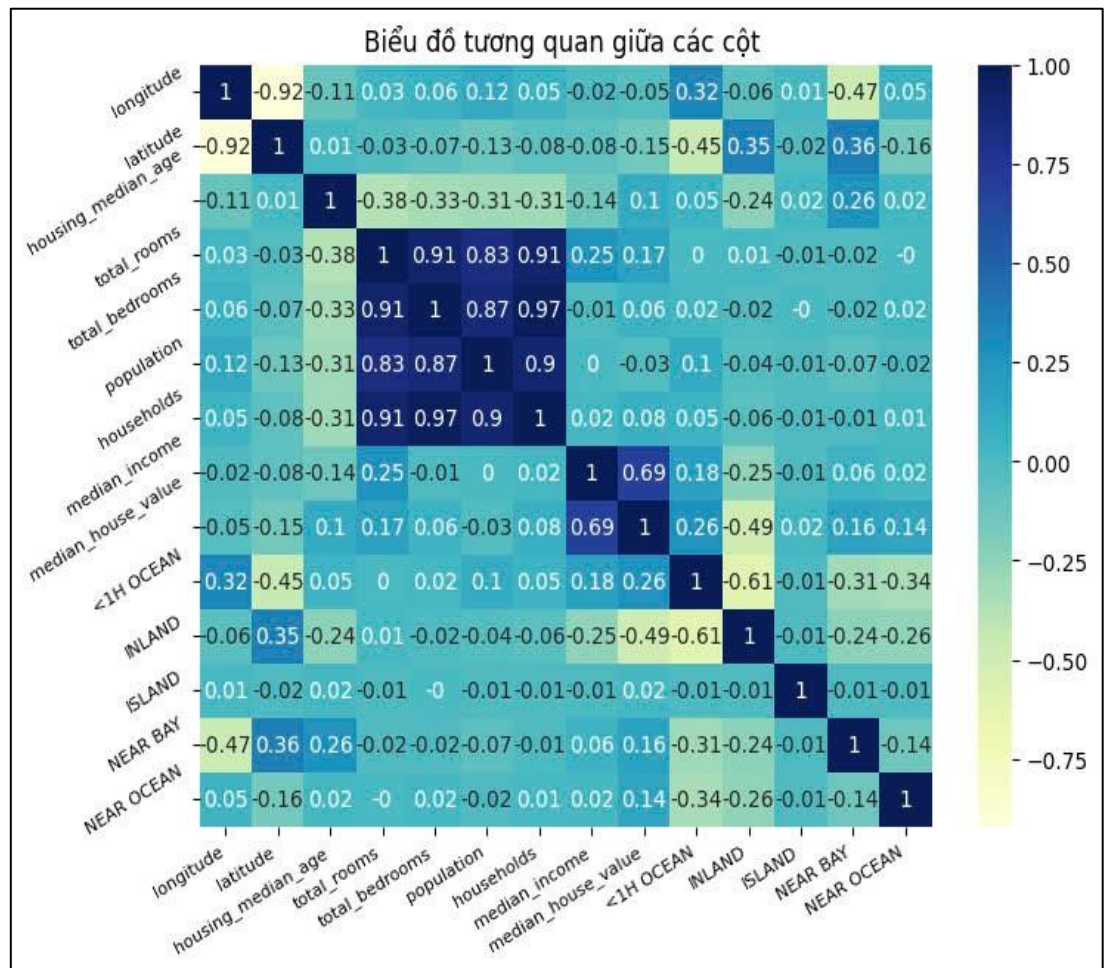


Hình 3.16. Biểu đồ Box Plot

Dựa theo sự phân phối trên, ta thấy các cột có mối quan hệ rất yếu so với cột median\_house\_value.

Biểu đồ tương quan các cột :





Hình 3.17. Biểu đồ Box Plot

Heatmap thể hiện tương quan mạnh giữa `total_rooms`, `total_bedrooms`, `population`, `households`; tương quan dương giữa `median_income` và `median_house_value`; và ảnh hưởng của vị trí địa lý đến giá nhà. Một số biến có tương quan yếu hoặc không tương quan.

### 3.3.4. Phân tích hồi quy

Ta tiến hành chia tập dữ liệu thành 2 phần , 70% dùng để huấn luyện (`train_size=0.7`) và 30% còn lại dùng để kiểm tra (`test_size=0.3`).

```

▶ # chia dữ liệu thành các tập huấn luyện (70 %) và test(30 %)
x_train, x_test, y_train, y_test = train_test_split(X, Y, train_size=0.7, test_size=0.3, random_state=42)
# Kiểm tra kích thước của dữ liệu
print(x_train.shape)
print(x_test.shape)
print(y_train.shape)
print(y_test.shape)

```

```

↵ (14447, 13)
  (6192, 13)
  (14447, 1)
  (6192, 1)

```

*Hình 3.18. Chia tập dữ liệu huấn luyện và kiểm tra*

Tiếp đó, tiến hành chuẩn hóa dữ liệu cho các biến độc lập

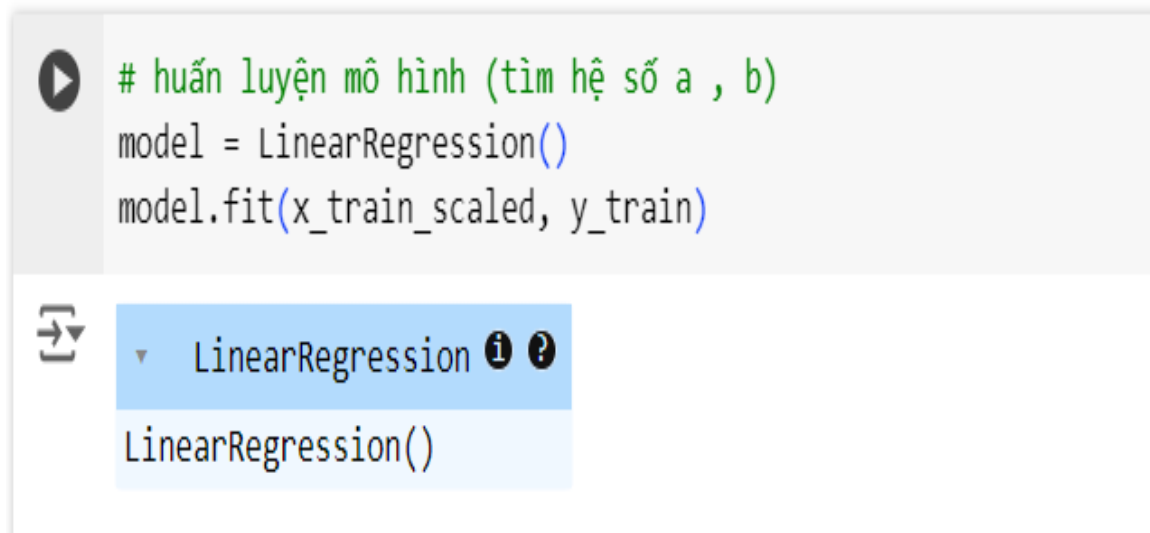
```

▶ # chuẩn hóa Min Max
minmax_scale = MinMaxScaler()
x_train_scaled = minmax_scale.fit_transform(x_train) # Chuẩn hóa tập huấn luyện
x_test_scaled = minmax_scale.transform(x_test) # Chuẩn hóa tập kiểm tra

```

*Hình 3.19. Chuẩn hóa dữ liệu*

Ta tiến hành huấn luyện bằng mô hình “Hồi quy tuyến tính” được thể hiện thông qua `LinearRegression()`.



Hình 3.20. Huấn luyện bằng mô hình Hồi quy tuyến tính

Cuối cùng, ta dự đoán kết quả trên tập `X_train` và `X_test` sau đó sử dụng hàm `rate()` để tiến hành đánh giá mô hình.



Hình 3.21. Dự đoán và đánh giá mô hình

Hàm rate sẽ cho chúng ta thấy được 2 độ đo là MSE (Mean Square Error) và R2\_score được thể hiện trên dữ liệu huấn luyện (train) và dữ liệu thử nghiệm (test). Sau đây là kết quả:

-----Kết quả thử nghiệm trên dữ liệu huấn luyện -----

Mean Squared Error (MSE): 4238473157.105766

R-squared (R2): 0.671074490160668

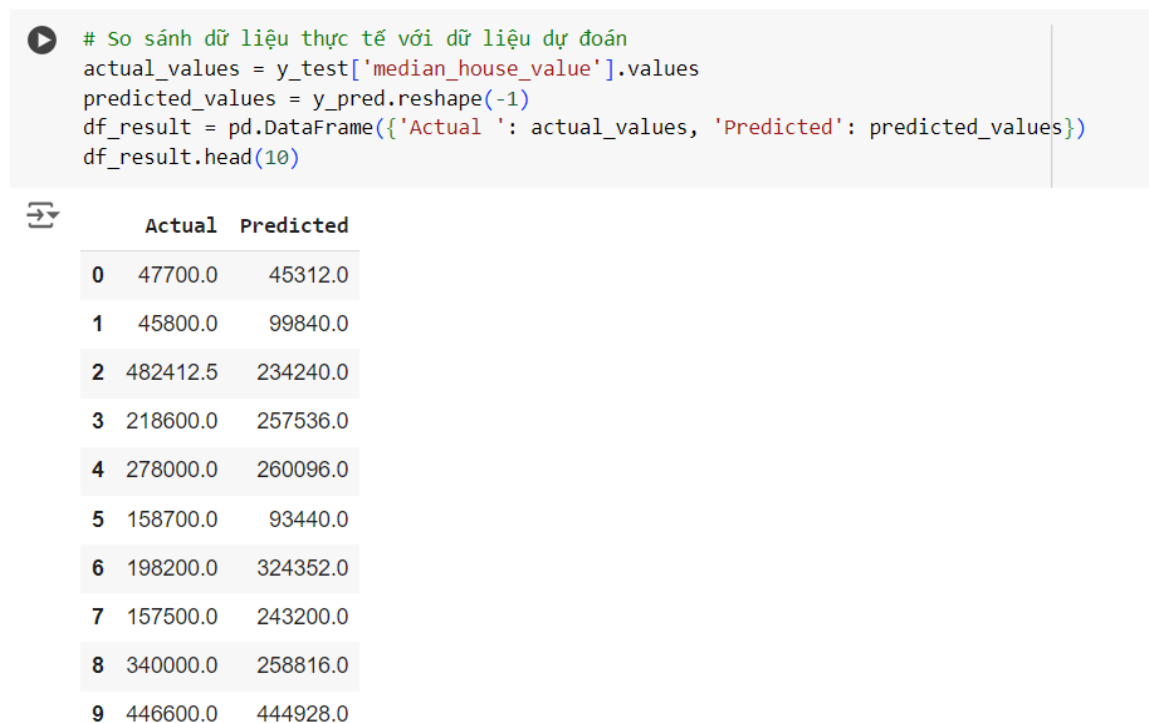
-----Kết quả thử nghiệm trên dữ liệu kiểm tra -----

Mean Squared Error (MSE): 4448074117.059552

R-squared (R2): 0.6485968222151116

(4238473157.105766, 0.671074490160668, 4448074117.059552,  
0.6485968222151116)

Ta sẽ so sánh dữ liệu sau khi dự đoán với dữ liệu thực tế:

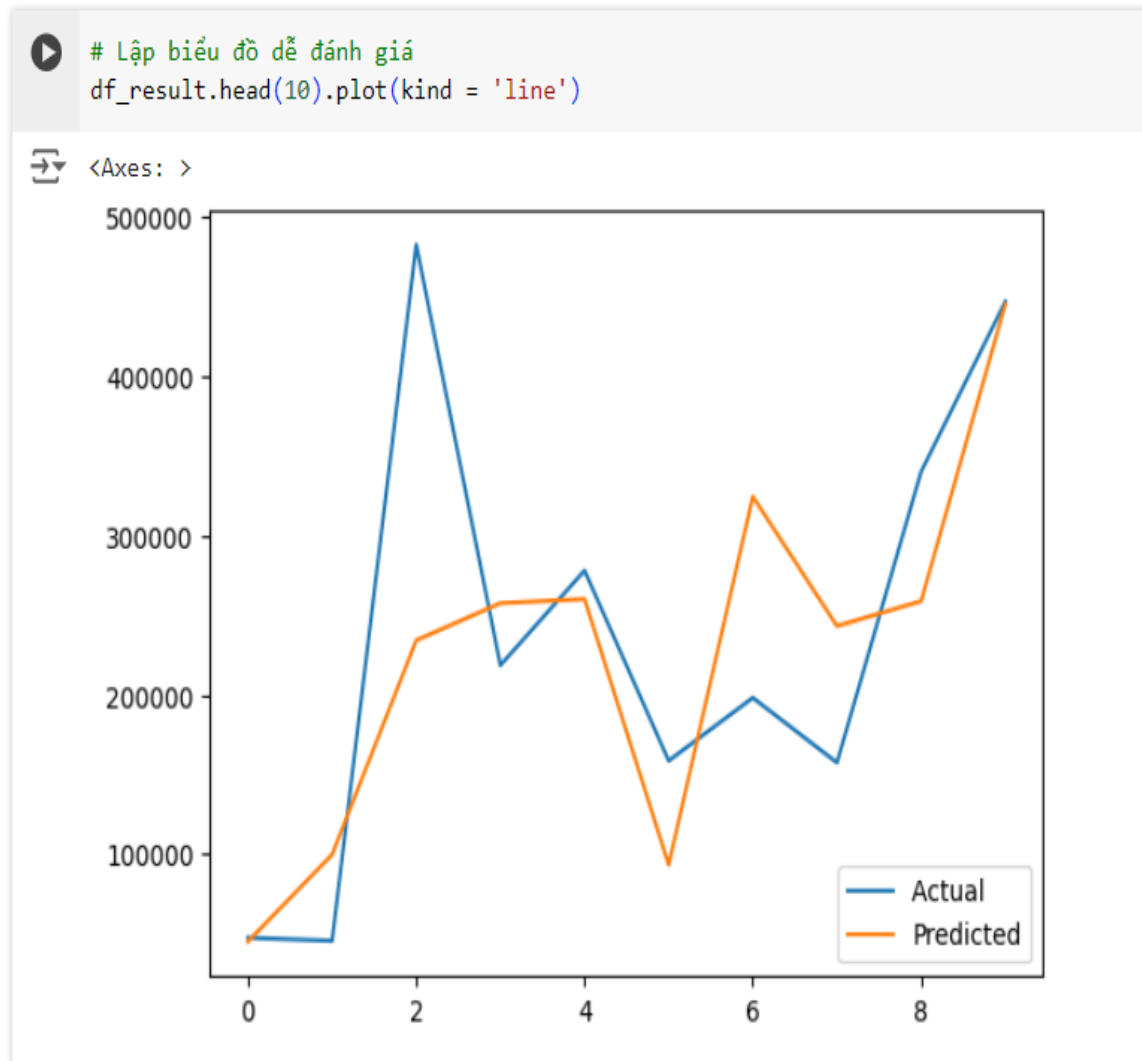


```
# So sánh dữ liệu thực tế với dữ liệu dự đoán
actual_values = y_test['median_house_value'].values
predicted_values = y_pred.reshape(-1)
df_result = pd.DataFrame({'Actual': actual_values, 'Predicted': predicted_values})
df_result.head(10)
```

	Actual	Predicted
0	47700.0	45312.0
1	45800.0	99840.0
2	482412.5	234240.0
3	218600.0	257536.0
4	278000.0	260096.0
5	158700.0	93440.0
6	198200.0	324352.0
7	157500.0	243200.0
8	340000.0	258816.0
9	446600.0	444928.0

Hình 3.22. Kết quả so sánh dữ liệu dự đoán với dữ liệu thực tế

Cuối cùng, ta thực hiện lập biểu đồ để trực quan hóa kết quả đánh giá mô hình để dễ dàng theo dõi mà cải thiện hiệu suất mô hình. Lưu mô hình cho lần sử dụng sau.



Hình 3.23. Biểu đồ so sánh giữa dữ liệu dự đoán với dữ liệu thực tế

### 3.4. Đánh giá

Phân phân tích mô tả đã phân tích được bộ dữ liệu ra các biểu đồ phù hợp và cho ta cái nhìn tổng quan về các yếu tố ảnh hưởng đến giá nhà tại California, chẳng hạn như mối liên hệ giữa vị trí địa lý, số phòng, và thu nhập trung bình. Tuy nhiên, đối với bài toán dự báo, mô hình phân tích hồi quy

tuyến tính đang không đạt hiệu quả cao với điểm  $R^2$  chỉ ở mức trung bình. Điều này cho thấy mô hình chưa thể nắm bắt tốt mối quan hệ giữa các đặc điểm và biến mục tiêu ('median\_house\_value'), có thể do mối quan hệ giữa các biến phức tạp hoặc dữ liệu chưa được xử lý tối ưu.

### **3.5. Kết luận**

Chương 3 đã trình bày phân thực nghiệm và đánh giá của dự án thông qua đầy đủ các bước, từ tiền xử lý dữ liệu, phân tích mô tả, đến xây dựng và đánh giá mô hình dự báo. Qua đó, đã đưa ra được các nhận xét về hiệu suất mô hình hồi quy tuyến tính, chỉ ra những hạn chế và các yếu tố cần cải thiện. Đồng thời, chương cũng đề xuất các hướng phát triển tiềm năng để nâng cao độ chính xác và hiệu quả của dự án trong tương lai.

## CHƯƠNG 4: CHƯƠNG TRÌNH DEMO

### 4.1. Giới thiệu về Framework sử dụng (Framework Streamlit)

#### 4.1.1. Giới thiệu về Streamlit

Streamlit là một framework mã nguồn mở (open-source) được viết bằng Python, chuyên dụng cho việc xây dựng và triển khai các ứng dụng web cho các dự án Machine Learning và Data Science. Điểm nổi bật của Streamlit nằm ở sự đơn giản và trực quan, cho phép người dùng tạo ra các ứng dụng web tương tác chỉ với vài dòng code Python mà không cần kiến thức chuyên sâu về phát triển web front-end (HTML, CSS, JavaScript).

Streamlit hoạt động bằng cách chạy một script Python và tự động tạo ra một ứng dụng web tương ứng. Mỗi khi script được thay đổi và lưu lại, ứng dụng web sẽ tự động cập nhật theo thời gian thực (hot-reloading), giúp quá trình phát triển và thử nghiệm diễn ra nhanh chóng và hiệu quả.

#### 4.1.2. Vai trò của Streamlit

Streamlit đóng vai trò như một cầu nối giữa code Python và giao diện người dùng web, giúp các nhà khoa học dữ liệu và kỹ sư machine learning dễ dàng chia sẻ công việc của mình với người khác. Một số trường hợp sử dụng phổ biến của Streamlit bao gồm:

- Prototype nhanh: Xây dựng nhanh chóng các bản demo và prototype cho các mô hình Machine Learning và ứng dụng Data Science.
- Chia sẻ kết quả phân tích dữ liệu: Trình bày kết quả phân tích dữ liệu một cách trực quan và tương tác thông qua các biểu đồ, bảng biểu và widget.

- Xây dựng dashboard: Tạo các dashboard theo dõi dữ liệu và hiệu suất, hiển thị các chỉ số quan trọng một cách dễ hiểu.
- Giảng dạy và học tập: Minh họa các khái niệm Machine Learning và Data Science thông qua các ứng dụng web tương tác.
- Xây dựng các ứng dụng web nhỏ và vừa: Phát triển các ứng dụng web tập trung vào xử lý và hiển thị dữ liệu.

#### **4.1.3. Ưu và nhược điểm**

##### **– Ưu điểm:**

- + Dễ sử dụng: Cú pháp đơn giản, dễ học và dễ sử dụng, ngay cả với người mới bắt đầu.
- + Nhanh chóng: Phát triển ứng dụng nhanh chóng nhờ tính năng hot-reloading và các component có sẵn.
- + Tương tác: Cung cấp nhiều widget tương tác giúp người dùng dễ dàng thao tác với ứng dụng.
- + Triển khai dễ dàng: Hỗ trợ triển khai lên Streamlit Cloud và các nền tảng khác một cách thuận tiện.
- + Miễn phí và mã nguồn mở: Có thể sử dụng miễn phí cho cả dự án cá nhân và thương mại.
- + Cộng đồng hỗ trợ mạnh mẽ: Cộng đồng người dùng đông đảo và nhiệt tình.

##### **– Nhược điểm:**

- + Khả năng tùy chỉnh giao diện hạn chế: So với các framework web truyền thống, khả năng tùy chỉnh giao diện của Streamlit còn hạn chế.
- + Phụ thuộc vào server: Ứng dụng Streamlit cần một server để chạy.



- + Chưa phù hợp cho các ứng dụng web phức tạp: Streamlit phù hợp hơn cho các ứng dụng web nhỏ và vừa, không phải là lựa chọn tốt nhất cho các ứng dụng web quy mô lớn và phức tạp.
- + Hạn chế về xử lý logic phía server: Streamlit tập trung vào việc hiển thị và tương tác, việc xử lý logic phức tạp phía server có thể gặp khó khăn.

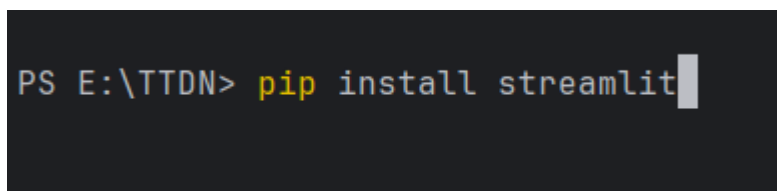
#### 4.1.4. Ứng dụng của Streamlit

Streamlit được ứng dụng rộng rãi trong nhiều lĩnh vực, đặc biệt là trong Machine Learning và Data Science. Dưới đây là một số ví dụ cụ thể:

- Phân tích dữ liệu thăm dò (Exploratory Data Analysis - EDA): Xây dựng các ứng dụng web tương tác để khám phá và phân tích dữ liệu.
- Trực quan hóa dữ liệu: Tạo các biểu đồ và đồ thị tương tác để hiển thị dữ liệu một cách trực quan.
- Xây dựng mô hình Machine Learning: Huấn luyện và đánh giá các mô hình Machine Learning, đồng thời triển khai chúng dưới dạng ứng dụng web.
- Theo dõi và giám sát mô hình: Xây dựng dashboard để theo dõi hiệu suất của mô hình Machine Learning trong thời gian thực.

#### 4.1.5. Cài đặt Framework Streamlit

Để cài đặt Streamlit trên pycharm, ta vào terminal gõ lệnh:



```
PS E:\TTDN> pip install streamlit
```

Hình 4.1. Cài đặt Streamlit

## 4.2. Chuẩn bị tài nguyên

Sau khi chuẩn hóa dữ liệu theo Min-Max, ta sẽ lưu mô hình chuẩn hóa lại để sử dụng cho mục đích chuẩn hóa dữ liệu đầu vào cho chương trình.

```
# chuẩn hóa Min Max
minmax_scale = MinMaxScaler()
x_train_scaled = minmax_scale.fit_transform(x_train) # Chuẩn hóa tập huấn luyện
x_test_scaled = minmax_scale.transform(x_test) # Chuẩn hóa tập kiểm tra
```

*Hình 4.2. Lưu mô hình chuẩn hóa min-max*

Sau khi thực hiện huấn luyện mô hình, ta sẽ lưu lại mô hình đã huấn luyện nhằm sử dụng để dự đoán giá nhà cho dữ liệu đầu vào của chương trình.

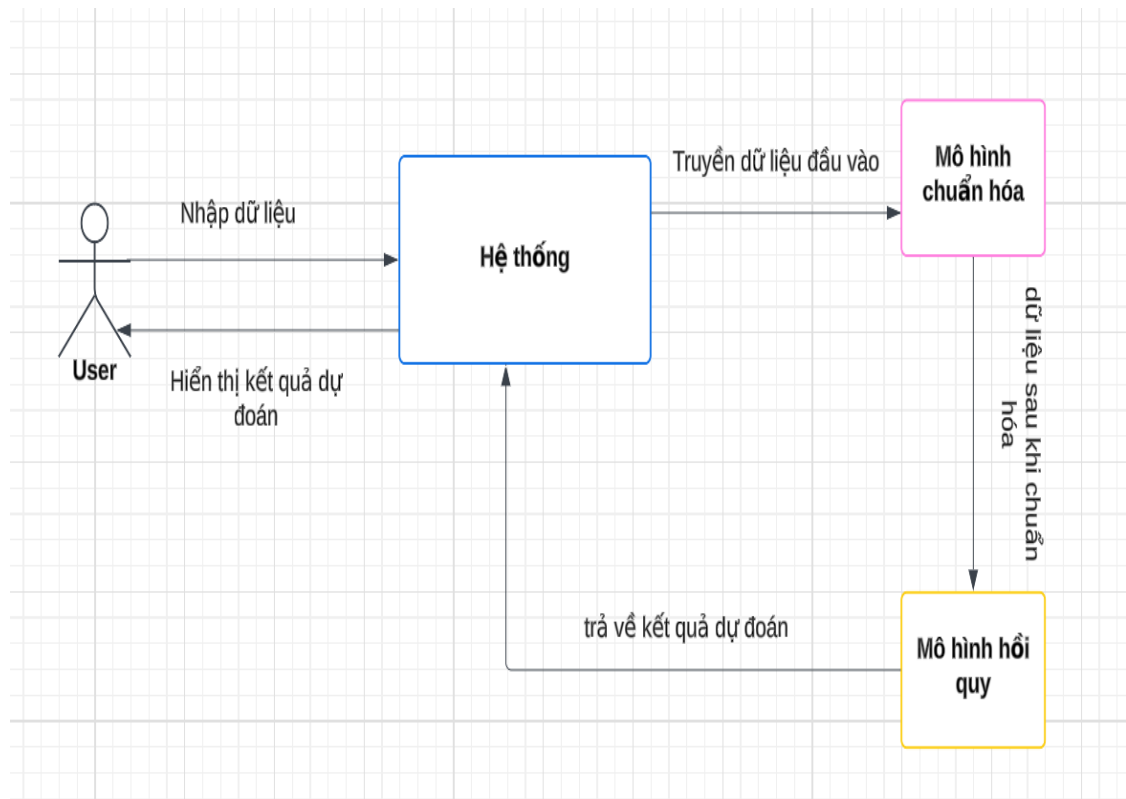
```
# Lưu mô hình vào file
with open('/content/drive/MyDrive/TTDN/linear.pkl', 'wb') as file:
    pickle.dump(model, file)
```

*Hình 4.3. Lưu mô hình đã huấn luyện*

## **4.3. Xây dựng mô hình và Demo chương trình**

### **4.3.1 Xây dựng mô hình**

- Xây dựng mô hình hệ thống:



Hình 4.4 Xây dựng mô hình hệ thống

– Đặc tả use case:

### 1. Đặc tả use case “Nhập dữ liệu”

Bảng 4.1. Bảng đặc tả use case “Nhập dữ liệu”

Tên use case	Nhập dữ liệu
Mục tiêu	Cho phép người dùng nhập dữ liệu để hệ thống thực hiện dự đoán.
Tác nhân	User (Người dùng)
Tiền điều kiện	<ul style="list-style-type: none"> <li>– Người dùng đã đăng nhập vào hệ thống (nếu cần).</li> <li>– Hệ thống sẵn sàng hoạt động.</li> </ul>

<p>Các bước chính</p>	<ol style="list-style-type: none"> <li>1. Người dùng khởi tạo yêu cầu nhập dữ liệu: Người dùng thực hiện hành động để bắt đầu quá trình nhập dữ liệu, ví dụ: nhấp vào nút "Dự đoán", chọn một chức năng dự đoán, hoặc truy cập một trang nhập dữ liệu.</li> <li>2. Hệ thống hiển thị giao diện nhập dữ liệu: Hệ thống hiển thị form hoặc giao diện cho phép người dùng nhập dữ liệu cần thiết cho dự đoán. Giao diện này nên rõ ràng, dễ hiểu và hướng dẫn người dùng nhập đúng định dạng dữ liệu.</li> <li>3. Người dùng nhập dữ liệu: Người dùng nhập dữ liệu vào các trường tương ứng trên giao diện.</li> <li>4. Hệ thống xác thực dữ liệu: Sau khi người dùng nhập dữ liệu, hệ thống tiến hành xác thực dữ liệu để đảm bảo tính hợp lệ. Ví dụ: kiểm tra kiểu dữ liệu, phạm vi giá trị, dữ liệu bắt buộc, v.v.</li> <li>5. Xử lý lỗi (nếu có): Nếu dữ liệu không hợp lệ, hệ thống hiển thị thông báo lỗi cho người dùng và yêu cầu người dùng sửa lại dữ liệu. Thông báo lỗi cần cụ thể và dễ hiểu để giúp người dùng sửa lỗi.</li> </ol>
-----------------------	---

	<p>6. Người dùng xác nhận dữ liệu: Sau khi nhập và kiểm tra dữ liệu, người dùng xác nhận dữ liệu đã nhập, ví dụ: bằng cách nhấp vào nút "Gửi", "Dự đoán" hoặc tương tự.</p> <p>7. Hệ thống lưu trữ dữ liệu: Hệ thống lưu trữ dữ liệu đã được xác thực để chuẩn bị cho quá trình chuẩn hóa và dự đoán.</p>
Hậu điều kiện	<ul style="list-style-type: none"> <li>– Dữ liệu đã được nhập và lưu trữ trong hệ thống.</li> <li>– Hệ thống sẵn sàng để chuẩn hóa và xử lý dữ liệu với mô hình chuẩn hóa và mô hình hồi quy.</li> </ul>
Luồng thay thế	<ul style="list-style-type: none"> <li>– Dữ liệu không hợp lệ: Nếu dữ liệu không hợp lệ ở bước 4, người dùng sẽ được yêu cầu nhập lại dữ liệu. Use case quay lại bước 3.</li> <li>– Người dùng hủy bỏ thao tác: Người dùng có thể hủy bỏ thao tác nhập dữ liệu bất cứ lúc nào. Use case kết thúc và không có dữ liệu nào được lưu trữ.</li> </ul>

## 2. Đặc tả use case “Dự Đoán”

Bảng 4.2. Bảng đặc tả use case ‘Dự Đoán’

Tên use case	Dự Đoán
Mục tiêu	Thực hiện dự đoán dựa trên dữ liệu người dùng đã nhập và hiển thị kết quả.
Tác nhân	User (Người dùng), Hệ thống
Tiền điều kiện	<ul style="list-style-type: none"><li>– Dữ liệu cần dự đoán đã được người dùng nhập và lưu trữ trong hệ thống (Use Case "Nhập Dữ Liệu Dự Đoán" đã hoàn thành).</li><li>– Mô hình chuẩn hóa và mô hình hồi quy đã được huấn luyện và sẵn sàng sử dụng.</li></ul>
Các bước chính	<ol style="list-style-type: none"><li><b>1.</b> Hệ thống truy xuất dữ liệu: Hệ thống truy xuất dữ liệu đã được người dùng nhập và lưu trữ trước đó.</li><li><b>2.</b> Chuẩn hóa dữ liệu: Hệ thống sử dụng mô hình chuẩn hóa để chuẩn hóa dữ liệu đầu vào. Bước này biến đổi dữ liệu về một định dạng phù hợp cho mô hình hồi quy.</li><li><b>3.</b> Dự đoán bằng mô hình hồi quy: Hệ thống sử dụng mô hình hồi quy đã được huấn luyện để thực hiện dự đoán dựa trên dữ liệu đã chuẩn hóa.</li></ol>

	<p><b>4.</b> Hệ thống xử lý kết quả dự đoán: Hệ thống có thể cần xử lý kết quả dự đoán trước khi hiển thị, ví dụ: làm tròn số, chuyển đổi đơn vị, v.v.</p> <p><b>5.</b> Hiển thị kết quả dự đoán: Hệ thống hiển thị kết quả dự đoán cho người dùng theo cách rõ ràng và dễ hiểu.</p>
Hậu điều kiện	Kết quả dự đoán được hiển thị cho người dùng.
Luồng thay thế	<ul style="list-style-type: none"> <li>– Lỗi trong quá trình chuẩn hóa: Nếu xảy ra lỗi trong quá trình chuẩn hóa dữ liệu (bước 2), hệ thống hiển thị thông báo lỗi cho người dùng.</li> <li>– Lỗi trong quá trình dự đoán: Nếu xảy ra lỗi trong quá trình dự đoán (bước 3), hệ thống hiển thị thông báo lỗi cho người dùng.</li> <li>– Không thể hiển thị kết quả: Nếu xảy ra lỗi trong quá trình hiển thị kết quả (bước 5), hệ thống hiển thị thông báo lỗi cho người dùng.</li> </ul>

### 3. Đặc tả use case ‘Reset’

*Bảng 4.3. Bảng đặc tả use case 'Reset'*

Tên use case	Reset (Đặt lại)
Mục tiêu	Cho phép người dùng xóa dữ liệu đã nhập và đặt lại hệ thống về trạng thái ban đầu để chuẩn bị cho một phiên dự đoán mới.
Tác nhân	User (Người dùng).
Tiền điều kiện	Người dùng đang ở trong giao diện nhập dữ liệu hoặc xem kết quả dự đoán.
Các bước chính	<ol style="list-style-type: none"><li><b>1.</b> Người dùng khởi tạo yêu cầu reset: Người dùng thực hiện hành động để bắt đầu quá trình reset, ví dụ: nhấp vào nút "Reset", "Đặt lại", hoặc tương tự.</li><li><b>2.</b> Hệ thống xác nhận yêu cầu reset (tùy chọn): Hệ thống có thể hiển thị hộp thoại xác nhận để đảm bảo người dùng muốn xóa dữ liệu hiện tại. Điều này giúp tránh việc người dùng vô tình xóa dữ liệu.</li><li><b>3.</b> Hệ thống xóa dữ liệu đã nhập: Hệ thống xóa dữ liệu người dùng đã nhập và lưu trữ trước đó, bao gồm cả dữ liệu đã chuẩn hóa và kết quả dự đoán (nếu có).</li></ol>



	<p><b>4.</b> Hệ thống đặt lại giao diện: Hệ thống đặt lại giao diện nhập dữ liệu về trạng thái ban đầu, xóa các giá trị đã nhập và hiển thị các trường trống.</p>
Hậu điều kiện	<ul style="list-style-type: none"> <li>– Dữ liệu đã nhập và kết quả dự đoán (nếu có) đã được xóa.</li> <li>– Giao diện nhập dữ liệu được đặt lại về trạng thái ban đầu.</li> <li>– Hệ thống sẵn sàng cho một phiên dự đoán mới.</li> </ul>
Luồng thay thế	<p>Người dùng hủy thao tác reset: Nếu người dùng hủy thao tác reset ở bước 2 (nếu có bước xác nhận), hệ thống sẽ trở lại trạng thái trước khi yêu cầu reset được khởi tạo. Không có dữ liệu nào bị xóa.</p>

#### 4.3.2 Demo chương trình

Giao diện khi chạy chương trình :

total\_rooms  
0

total\_bedrooms  
0

population  
0.00

households  
0.00

median\_income  
0.00

ocean\_proximity  
Choose an option

Predict

Reset

Deploy

## House Price Forecast

This app predicts House Price using a Linear Regression model.

*Hình 4.5. Giao diện chương trình*

Tiến hành nhập dữ liệu trên giao diện chương trình

The image shows a web interface for data input. It features six input fields, each with a label and a value, and two buttons at the bottom. The fields are:   
1. **total\_rooms**: A numeric input field with the value '0'.   
2. **total\_bedrooms**: A numeric input field with the value '0'.   
3. **population**: A numeric input field with the value '0.00'.   
4. **households**: A numeric input field with the value '0.00'.   
5. **median\_income**: A numeric input field with the value '0.00'.   
6. **ocean\_proximity**: A dropdown menu with the text 'Choose an option' and a downward arrow.   
At the bottom, there are two buttons: 'Predict' and 'Reset'.

*Hình 4.6. Giao diện nhập dữ liệu*

Sau đó ấn nút “Predict”, dữ liệu sẽ được truyền vào các biến tương ứng.

```
# Input fields for user to enter feature values
longitude = st.sidebar.number_input('longitude', value=0.0)
latitude = st.sidebar.number_input('latitude', value=0.0)
housing_median_age = st.sidebar.number_input('housing_median_age', value=0)
total_rooms = st.sidebar.number_input('total_rooms', value=0)
total_bedrooms = st.sidebar.number_input('total_bedrooms', value=0)
population = st.sidebar.number_input('population', value=0.0)
households = st.sidebar.number_input('households', value=0.0)
median_income = st.sidebar.number_input('median_income', value=0.0)
ocean_proximity = st.sidebar.selectbox('ocean_proximity', ('<1H OCEAN', 'INLAND', 'ISLAND', 'NEAR BAY', 'NEAR OCEAN'))
```

*Hình 4.7. Mã nguồn nhận dữ liệu*

Sau khi nhận được dữ liệu, dữ liệu sẽ được đưa vào mô hình chuẩn hóa để chuẩn hóa dữ liệu và thông qua mô hình hồi quy để trả về kết quả dự báo.

```
# Define a function to make predictions
def predict_house(longitude, latitude, housing_median_age, total_rooms, total_bedrooms, population, households, median_income, ocean_proximity):
    # Tạo mảng 1 chiều từ input_data
    input_data = np.array([longitude, latitude, housing_median_age, total_rooms, total_bedrooms, population, households, median_income, 0, 0, 0, 0, 0])
    if ocean_proximity == '<1H OCEAN':
        input_data[8] = 1
    elif ocean_proximity == 'INLAND':
        input_data[9] = 1
    elif ocean_proximity == 'ISLAND':
        input_data[10] = 1
    elif ocean_proximity == 'NEAR BAY':
        input_data[11] = 1
    elif ocean_proximity == 'NEAR OCEAN':
        input_data[12] = 1

    # Chuẩn hóa dữ liệu input_data bằng mô hình chuẩn hóa Min-Max
    input_data_normalized = loaded_minmax_scale.transform(input_data.reshape(1, -1))

    # Dự đoán giá trị bằng mô hình Linear Regressor đã nạp
    predicted_house = model.predict(input_data_normalized)

    return predicted_house[0]
```

*Hình 4.8. Mô tả quá trình dự đoán*

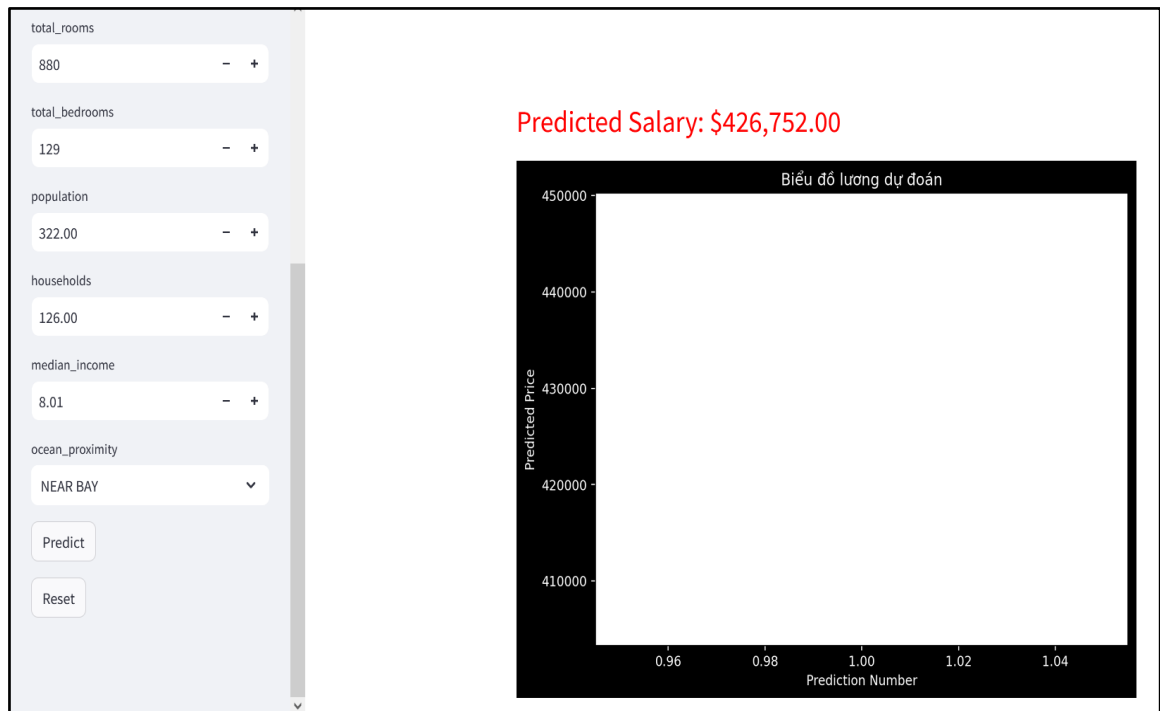
Kết quả dự đoán nhận được sẽ được lưu vào danh sách dự đoán giúp người dùng có thể xem lại kết quả dự đoán trước đó mà không cần nhập lại dữ liệu.

```
#Tạo 1 list để lưu dữ liệu dự đoán sau mỗi lần ấn nút predict bằng session state
if "predicted_houses" not in st.session_state:
    st.session_state.predicted_houses = [] # tạo list để lưu trữ các giá trị dự đoán

if st.sidebar.button('Predict'):
    predicted_house = predict_house(longitude, latitude, housing_median_age, total_rooms, total_bedrooms, population, households, median_income, ocean_proximity )
    predicted_house = float(predicted_house)
    st.session_state.predicted_houses.append(predicted_house) # Lưu trữ kết quả dự báo vào mảng
    st.markdown(f'<p class="red-text">Predicted Price house: ${(predicted_house):.2f}</p>', unsafe_allow_html=True)
```

Hình 4.9. Quá trình lưu kết quả dự đoán

Kết quả chạy trình :



Hình 4.10. Hình kết quả chạy chương trình

Danh sách dự đoán từ các lần nhấn trước đó:

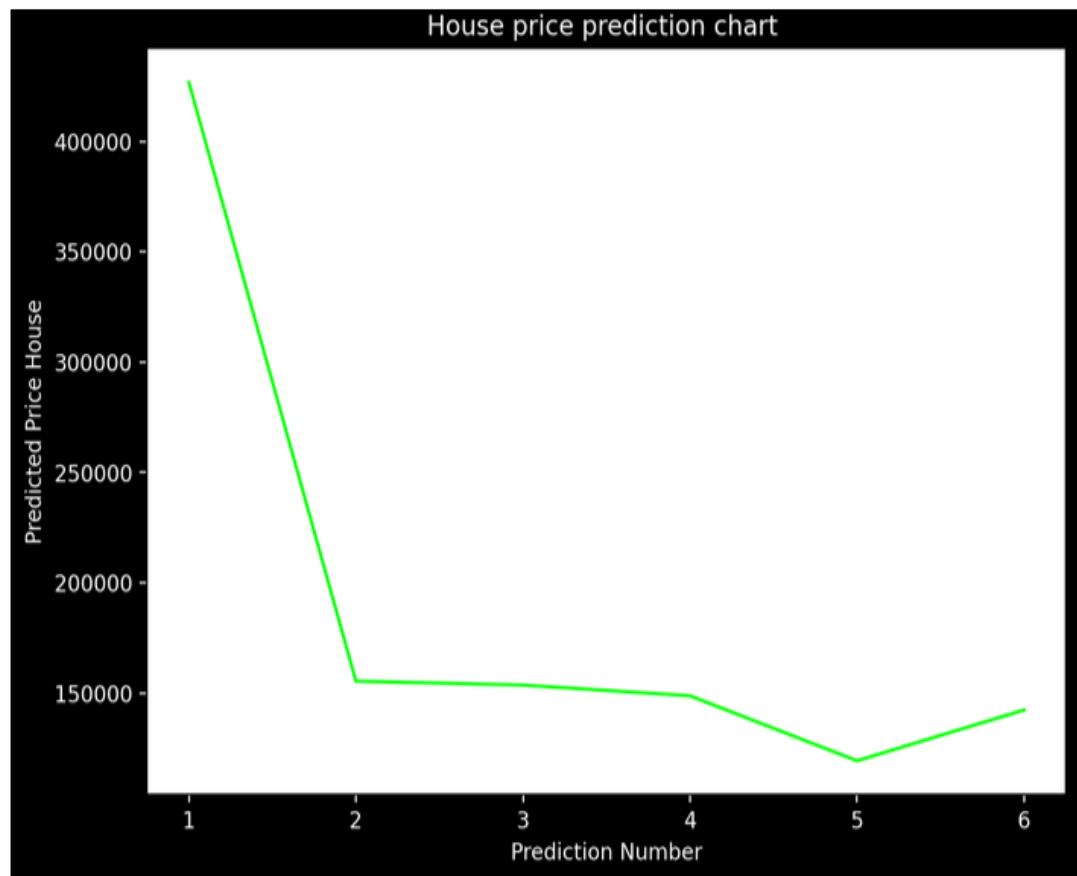
Dự đoán 1: \$426,752.00



This app predicts House Price using a Linear Regression model.

*Hình 4.11. Ảnh minh họa cho kết quả dự đoán*

Do các kết quả dự đoán trả về được lưu trữ trong 1 danh sách nên chương trình có thể hiển thị kết quả dự đoán của hiện tại và các lần nhập dữ liệu trước đó .



Danh sách dự đoán từ các lần nhấn trước đó:

Dự đoán 1: \$426,752.00

Dự đoán 2: \$155,392.00

Dự đoán 3: \$153,600.00

Dự đoán 4: \$148,736.00

Dự đoán 5: \$119,296.00

Dự đoán 6: \$142,336.00

Hình 4.12. Biểu đồ lịch sử dự đoán giá nhà

Nếu ta muốn xóa đi kết quả dự đoán trước đó, ta cần nhấn nút 'Reset' để cài đặt lại giá trị khởi tạo cho các biến dữ liệu tương ứng và xóa đi dữ liệu dự báo trước đó.

```

if st.sidebar.button('Reset'):
    # Đặt lại tất cả các giá trị về mặc định
    st.session_state['longitude'] = 0.0
    st.session_state['latitude'] = 0.0
    st.session_state['housing_median_age'] = 0.0
    st.session_state['total_rooms'] = 0
    st.session_state['total_bedrooms'] = 0
    st.session_state['population'] = 0.0
    st.session_state['households'] = 0.0
    st.session_state['median_income'] = 0.0
    st.session_state['ocean_proximity'] = '<1H OCEAN'

    # Xóa danh sách predicted_prices
    st.session_state.predicted_houses = []

```

Hình 4.13. Mã nguồn Reset dữ liệu

Kết quả chạy chương trình sau khi nhấn nút ‘Reset’:



Hình 4.14. Giao diện chương trình sau khi Reset



## KẾT LUẬN

Trong quá trình thực hiện đồ án "Dự đoán giá nhà bằng phương pháp hồi quy tuyến tính", em đã tập trung nghiên cứu và xây dựng một mô hình đơn giản nhưng hiệu quả, nhằm giải quyết bài toán dự báo giá nhà dựa trên các yếu tố đầu vào như thu nhập trung bình, số phòng, và vị trí địa lý. Mục tiêu của đồ án là phát triển một hệ thống có khả năng đưa ra dự báo giá nhà một cách chính xác, qua đó hỗ trợ các cá nhân và tổ chức trong việc ra quyết định liên quan đến bất động sản.

Qua các thí nghiệm, mô hình hồi quy tuyến tính đã cho thấy khả năng dự báo giá nhà ở mức độ cơ bản. Kết quả đánh giá với điểm  $R^2$  đạt 0.67 cho thấy mô hình có khả năng giải thích khoảng 67% sự biến thiên của giá trị thực tế. Tuy nhiên, các sai số như MSE và MAE vẫn còn cao, điều này chỉ ra rằng mô hình hồi quy tuyến tính chưa thể nắm bắt hoàn toàn mối quan hệ phức tạp giữa các đặc trưng và biến mục tiêu. Bên cạnh đó, phân tích dữ liệu cũng chỉ ra rằng một số đặc trưng quan trọng có thể chưa được khai thác tối đa, dẫn đến hạn chế trong việc cải thiện độ chính xác của dự báo.

Dù vậy, mô hình hồi quy tuyến tính vẫn là một giải pháp khởi đầu tốt nhờ tính đơn giản, dễ triển khai và khả năng diễn giải. Điều này đặc biệt hữu ích trong việc kiểm tra tính phù hợp của dữ liệu và xác định các đặc trưng quan trọng.

Trong tương lai, em sẽ tập trung vào việc mở rộng và làm sạch dữ liệu, đồng thời tích hợp các đặc trưng mới có ý nghĩa thực tiễn cao hơn. Bên cạnh đó, việc thử nghiệm các mô hình học máy nâng cao như Random Forest, Gradient Boosting hay mô hình học sâu (Deep Learning) cũng sẽ được nghiên cứu nhằm cải thiện khả năng dự báo. Với những bước phát triển này, em hy

vọng sẽ nâng cao đáng kể độ chính xác và hiệu quả của hệ thống, giúp nó đáp ứng tốt hơn các yêu cầu thực tế.

## TÀI LIỆU THAM KHẢO

- [1]. RICKY NGUYEN : PHÂN TÍCH DỮ LIỆU LỚN [Bài giảng] . Đại Học Công Nghiệp Hà Nội.
- [2]. Hồi quy tuyến tính là gì ? URL : <https://trituenhantao.io/machine-learning-co-ban/bai-3-linear-regression-hoi-quy-tuyen-tinh/> . Lần truy cập gần nhất 20/2/2025.
- [3]. Hồi quy Logistic ? URL : <https://trituenhantao.io/machine-learning-co-ban/bai-6-logistic-regression-hoi-quy-logistic/> . Lần truy cập gần nhất 20/2/2025.
- [4]. Hà Anh Tuấn, Bùi Văn Đắc, Vũ Văn Hùng (2023). Phân tích dự báo giá nhà sử dụng mô hình hồi quy tuyến tính. Đại Học Công Nghiệp Hà Nội