

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI



BÁO CÁO THỰC NGHIỆM **HỌC PHẦN: PHÂN TÍCH DỮ LIỆU LỚN**

ĐỀ TÀI: PHÂN TÍCH DỰ BÁO GIÁ NHÀ Ở MỸ BÀNG
PHƯƠNG PHÁP HỒI QUY TUYẾN TÍNH

Giảng viên hướng dẫn : TS. Nguyễn Mạnh Cường

Lớp : 20241IT6077003 - K16

Nhóm thực hiện : Nhóm 18

Thành Viên :
1. Đoàn Đại Dương – 2021606493
2. Nguyễn Văn Hiến – 2021607091
3. Quán Xuân Dương – 2021606693

Hà Nội - 2024

MỤC LỤC

LỜI CẢM ƠN	1
LỜI MỞ ĐẦU	2
CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI	4
1.1. Tổng quan về phân tích dữ liệu	4
1.1.1. Phân tích dữ liệu là gì	4
1.1.2. Quy trình phân tích dữ liệu	4
1.2. Tổng quan về bài toán dự báo	6
1.2.1. Lịch sử bài toán dự báo	6
1.2.2. Tình hình phát triển của bài toán dự báo ở Việt Nam	8
1.2.3. Tình hình phát triển của bài toán dự báo ở thế giới	10
1.3. Giới thiệu Bài toán	11
1.4. Đầu vào và đầu ra của bài toán	12
1.5. Tầm quan trọng của bài toán	13
1.6. Ứng dụng	14
1.7. Cơ hội và hạn chế	15
CHƯƠNG 2: PHƯƠNG PHÁP KỸ THUẬT	17
2.1. Phương hướng tiếp cận bài toán	17
2.2. Mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN)	17
2.2.1. Giới thiệu	17
2.2.2. Cấu trúc của mạng nơ-ron nhân tạo	18
2.2.3. Phân loại	19
2.2.4. Quy trình hoạt động	20
2.2.5. Ưu điểm và nhược điểm	23
2.3. Random Forest	23
2.3.1. Giới thiệu	23

2.3.2. Đặc điểm	23
2.3.4. Ưu điểm và nhược điểm.....	26
2.4. Hồi quy tuyến tính	27
2.4.1. Giới thiệu	27
2.4.2. Các loại hồi quy tuyến tính	32
2.4.3. Ứng dụng.....	33
2.4.4. Ưu điểm và nhược điểm.....	33
2.5. Support Vector Machines – SVM.....	34
2.5.1. Giới thiệu	34
2.5.2. Cấu trúc	35
2.5.3. Quy trình hoạt động	36
2.5.4. Phân loại.....	38
2.5.5. Ưu điểm và nhược điểm của SVM	38
2.6. Cây quyết định (Decision Tree).....	38
2.6.1. Giới thiệu	38
2.6.2. Cấu trúc của cây quyết định.....	39
2.6.3. Quy trình hoạt động	40
2.6.4. Phân loại.....	41
2.6.5. Ưu điểm và nhược điểm của cây quyết định	41
2.7. Kết luận.....	42
CHƯƠNG 3 THỰC NGHIỆM VÀ ĐÁNH GIÁ.....	44
3.1. Dữ liệu thực nghiệm	44
3.2 Môi trường thực nghiệm	45
3.3. Quy trình thực nghiệm	45
3.3.1. Đặt mục tiêu	45
3.3.2. Tiền xử lý dữ liệu.....	45
3.3.3. Phân tích mô tả.....	48

3.4. Đánh giá và đề xuất.....	64
3.5. Kết luận.....	65
CHƯƠNG 4 : XÂY DỰNG SẢN PHẨM DEMO	66
4.1. Giới thiệu về Framework được sử dụng	66
4.2. Chuẩn bị tài nguyên xây dựng chương trình	67
KẾT LUẬN.....	74
TÀI LIỆU THAM KHẢO	75

DANH MỤC HÌNH ẢNH

Hình 1. 1: Quy trình phân tích dữ liệu	5
Hình 1. 2: Giới thiệu về bài toán.....	11
Hình 2. 1: Giới thiệu về ANN.....	17
Hình 2. 2: Cấu trúc cơ bản của mạng nơ-ron ANN	18
Hình 2. 3: Mô hình ANN cơ bản	19
Hình 2. 4: Quy trình hoạt động của mạng ANN	21
Hình 2. 5: Quy trình hoạt động của mô hình Random Forest	24
Hình 2. 6: Quy trình hoạt động mô hình hồi quy tuyến tính	28
Hình 2. 7: Ví dụ về hồi quy tuyến tính	33
Hình 2. 8: Quy trình hoạt động của mô hình SVM	36
Hình 2. 9: Quy trình hoạt động của Cây quyết định	40
Hình 3. 1: 15 dòng đầu của bộ dữ liệu gốc	44
Hình 3. 2: Quy trình thực nghiệm đề tài phân tích dữ liệu	45
Hình 3. 3: Thông tin tóm lược dữ liệu của cột dữ liệu dạng số	46
Hình 3. 4: Kiểm tra dữ liệu bị khuyết	47
Hình 3. 5: Kiểm tra kiểu dữ liệu cho từng cột	48
Hình 3. 6: Đếm số lượng các giá trị duy nhất tại cột “Address”	48
Hình 3. 7: Biểu đồ cột thể hiện	49
Hình 3. 8: Biểu đồ box chart.....	50
Hình 3. 9: Biểu đồ scatter với đường hồi quy màu đỏ.....	52
Hình 3. 10: Biểu đồ heatmap	54
Hình 3. 11: Mô tả chọn lọc đặc trưng	58
Hình 3. 12: Hàm đánh giá mô hình.....	59
Hình 3. 13: Mô tả kỹ thuật 10 – Fold Cross Validation	63
Hình 4. 1: Chuẩn hóa Min, Max	67

Hình 4. 2: Mô hình xây dựng chương trình	68
Hình 4. 3: Giao diện chương trình	70
Hình 4. 4: Quá trình nhập dữ liệu đầu vào.....	71
Hình 4. 5: Mô tả quá trình truyền dữ liệu	72
Hình 4. 6: Quá trình chuẩn hóa dữ liệu.....	72
Hình 4. 7: Kết quả của chương trình.....	73

DANH MỤC BẢNG BIỂU

Bảng 4. 1: Mô tả use case dự báo giá nhà.....	68
Bảng 4. 2: Mô tả use case reset dữ liệu.....	69

LỜI CẢM ƠN

Lời đầu tiên, chúng em xin gửi lời cảm ơn chân thành và sâu sắc tới các thầy cô trong khoa Công nghệ thông tin - Trường Đại học Công Nghiệp Hà Nội, những người đã tận tâm truyền đạt cho chúng em những kiến thức quý báu, những bài học bổ ích và thiết thực trong suốt quá trình học tập. Đặc biệt, chúng em xin bày tỏ sự tri ân và gửi lời cảm ơn sâu sắc tới giảng viên TS. Nguyễn Mạnh Cường, người đã trực tiếp hướng dẫn và chỉ bảo chúng em trong suốt quá trình nghiên cứu, giúp chúng em hoàn thành đề tài này. Những kiến thức quý giá và những lời khuyên chân thành của thầy đã là động lực lớn lao để chúng em có thể vượt qua khó khăn và hoàn thành nhiệm vụ.

Đề tài này không chỉ giúp chúng em rèn luyện kỹ năng tư duy phân tích và xử lý dữ liệu, mà còn củng cố khả năng trình bày thông tin một cách logic và rõ ràng. Chúng em hy vọng những kiến thức và kinh nghiệm thu được từ nghiên cứu này sẽ là nền tảng vững chắc giúp chúng em tiếp tục phát triển trong tương lai, không chỉ trong học tập mà còn trong công việc và cuộc sống.

Tuy nhiên, trong suốt quá trình nghiên cứu, do năng lực và kiến thức của chúng em còn hạn chế, không thể tránh khỏi những thiếu sót. Chúng em mong nhận được sự thông cảm và những ý kiến đóng góp quý báu từ quý thầy cô cũng như các bạn trong lớp để hoàn thiện hơn.

Chúng em xin trân trọng cảm ơn!

Nhóm sinh viên thực hiện

Đoàn Đại Dương

Nguyễn Văn Hiên

Quán Xuân Đường

LỜI MỞ ĐẦU

Trong bối cảnh thị trường bất động sản Mỹ luôn biến động và phức tạp, việc dự đoán giá nhà ở trở thành một vấn đề quan trọng đối với nhiều đối tượng, từ người mua nhà, nhà đầu tư đến các nhà hoạch định chính sách. Giá nhà ở không chỉ chịu tác động bởi các yếu tố kinh tế vĩ mô mà còn phụ thuộc vào nhiều yếu tố khác như vị trí địa lý, diện tích, số phòng ngủ, tuổi nhà, tình hình cơ sở hạ tầng,... Sự đa dạng và phức tạp của các yếu tố này đòi hỏi một phương pháp phân tích hiệu quả để đưa ra những dự báo chính xác.

Bài toán dự đoán giá nhà ở có ý nghĩa quan trọng trong ngành bất động sản, nơi các nhà đầu tư và người mua cần hiểu rõ cơ cấu giá, dự đoán biến động thị trường và xác định chiến lược tài chính hiệu quả. Nó cũng cung cấp cho người tiêu dùng thông tin hữu ích để đưa ra quyết định mua nhà sáng suốt. Phân tích giá nhà có thể dựa trên nhiều biến số như vị trí địa lý, diện tích, số lượng phòng ngủ, tuổi của ngôi nhà, các tiện ích xung quanh, và nhiều yếu tố khác.

Nội dung báo cáo của bài tập lớn gồm 3 chương như sau:

Chương 1: Tổng quan về đề tài

Bài toán Phân Tích Giá Nhà Ở Mỹ. Chương này giới thiệu tổng quan về vấn đề phân tích giá nhà ở Mỹ, đi sâu vào lý do chọn đề tài, và đặt câu hỏi nghiên cứu cụ thể. Qua đó, chương này tạo nền tảng cho việc hiểu rõ mục tiêu và ý nghĩa của nghiên cứu, đồng thời làm rõ tầm quan trọng của việc dự đoán giá nhà trong bối cảnh thị trường bất động sản đang ngày càng biến động và phức tạp.

Chương 2: Phương pháp kỹ thuật

Các Kỹ Thuật để Xử Lý Bài Toán. Chương này tập trung vào việc giới thiệu và thảo luận các kỹ thuật được sử dụng để xử lý bài toán phân tích giá nhà ở Mỹ, bao gồm chuẩn bị dữ liệu, biểu diễn mô hình hồi quy tuyến tính và các phương pháp đánh giá mô hình. Chương sẽ cung cấp cái nhìn tổng quan về lý thuyết và quy trình thực hiện cụ thể, hỗ trợ cho việc hiểu rõ quá trình phân tích dữ liệu và dự đoán giá nhà ở.

Chương 3: Kết quả thực nghiệm

Kết Quả Thực Nghiệm. Chương cuối cùng trình bày các kết quả thực nghiệm từ việc áp dụng các kỹ thuật và phương pháp giới thiệu ở chương 2 vào bài toán phân

tích giá nhà ở. Thông qua việc trình bày các dữ liệu và kết quả định lượng, chương này sẽ minh họa cách các kỹ thuật đã đề xuất có thể được áp dụng và đánh giá, đồng thời tổng kết những bài học và kiến thức thu được từ nghiên cứu.

Thông qua quá trình nghiên cứu về phân tích giá nhà ở Mỹ sử dụng phương pháp hồi quy tuyến tính, nhóm đã học được cách ứng dụng lý thuyết vào thực tế của thị trường bất động sản đầy phức tạp. Báo cáo này không chỉ giúp mở rộng kiến thức mà còn có giá trị thực tiễn cho những ai quan tâm đến việc dự đoán giá trong ngành bất động sản. Nhóm hy vọng rằng những kết quả và phương pháp nghiên cứu trong báo cáo này sẽ trở thành nguồn tài liệu hữu ích, đồng thời khơi dậy sự đam mê trong cộng đồng nghiên cứu, góp phần vào sự phát triển của ngành công nghiệp bất động sản, nơi tri thức mới sẽ dẫn dắt chúng ta tới những khám phá và tiến bộ đáng kể.

CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI

1.1. Tổng quan về phân tích dữ liệu

1.1.1. Phân tích dữ liệu là gì

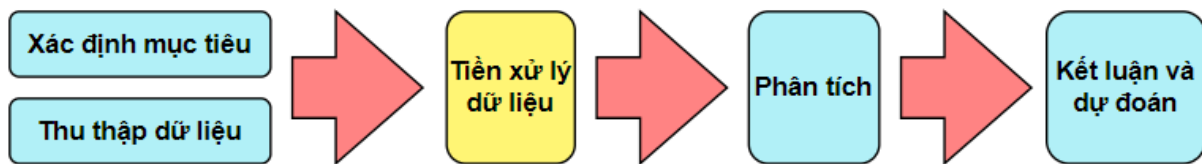
Phân tích dữ liệu là một lĩnh vực nghiên cứu và ứng dụng thực tế tập trung vào việc khai thác và trích xuất những hiểu biết giá trị từ dữ liệu. Đây là quá trình toàn diện, bao gồm các bước kiểm tra, làm sạch, chuyển đổi, và mô hình hóa dữ liệu, nhằm phát hiện thông tin hữu ích, đưa ra kết luận có giá trị và hỗ trợ việc ra các quyết định chiến lược.

Phân tích dữ liệu đã trở thành một công cụ không thể thiếu trong nhiều lĩnh vực và ngành nghề khác nhau, từ khoa học dữ liệu, kinh doanh, tài chính, y tế, marketing, hành chính công cho đến nghiên cứu khoa học và giáo dục. Thông qua quá trình phân tích, các tổ chức và cá nhân có thể khám phá các xu hướng ẩn trong dữ liệu, xác định mẫu số chung, nhận diện mối tương quan, và tìm ra những thông tin cốt lõi có khả năng ảnh hưởng lớn đến quyết định và chiến lược. Những hiểu biết này không chỉ cải thiện hiệu quả hoạt động mà còn tạo ra các giá trị mới và cơ hội phát triển trong cả ngắn hạn và dài hạn.

Để đạt được hiệu quả, phân tích dữ liệu đòi hỏi sự kết hợp của nhiều phương pháp và công cụ tiên tiến. Các thuật toán thống kê được sử dụng để mô tả và phân tích dữ liệu một cách định lượng. Kỹ thuật khai phá dữ liệu giúp khám phá các mẫu ẩn sâu và các mối quan hệ trong tập dữ liệu lớn. Học máy (machine learning) đóng vai trò quan trọng trong việc xây dựng các mô hình dự đoán, tự động hóa phân tích, và phát hiện các xu hướng tiềm ẩn. Bên cạnh đó, trực quan hóa dữ liệu là một yếu tố không thể thiếu, giúp biến các con số khô khan thành biểu đồ, đồ thị sinh động, dễ hiểu, từ đó truyền tải thông tin một cách rõ ràng và hiệu quả.

Ngoài ra, các mô hình hóa nâng cao như phân cụm (clustering), phân tích hồi quy (regression analysis), và các phương pháp tối ưu hóa khác cũng thường xuyên được áp dụng để đưa ra các dự đoán hoặc khuyến nghị chính xác hơn. Nhờ vào những công cụ và phương pháp này, dữ liệu thô trở thành nguồn thông tin có ý nghĩa sâu sắc, giúp tổ chức xây dựng kế hoạch chiến lược và đưa ra các quyết định đúng đắn, tối ưu hóa hiệu suất và giảm thiểu rủi ro trong hoạt động kinh doanh và nghiên cứu.

1.1.2. Quy trình phân tích dữ liệu.



Hình 1. 1: Quy trình phân tích dữ liệu

Quy trình phân tích dữ liệu thường bao gồm các bước chính:

Xác định mục tiêu và thu thập dữ liệu:

- Xác định mục tiêu là bước đầu tiên và đóng vai trò cốt lõi trong toàn bộ quy trình phân tích dữ liệu. Đây là việc xác định rõ ràng những kết quả cụ thể mà chúng ta muốn đạt được thông qua việc xử lý và phân tích dữ liệu. Mục tiêu này sẽ định hướng cho các hoạt động tiếp theo, giúp chúng ta tập trung vào việc thu thập và xử lý thông tin quan trọng, phù hợp với yêu cầu hoặc nhu cầu cụ thể.
- Thu thập dữ liệu là quá trình lấy dữ liệu từ nhiều nguồn khác nhau như cơ sở dữ liệu, tệp tin, trang web, thiết bị cảm biến, hoặc khảo sát. Dữ liệu này có thể tồn tại dưới nhiều dạng, bao gồm số liệu, văn bản, hình ảnh hoặc âm thanh. Đảm bảo tính đầy đủ, chính xác và hợp pháp của dữ liệu là yếu tố quan trọng trong bước này.

Tiền xử lý dữ liệu: Dữ liệu thô thường không hoàn chỉnh và có thể chứa nhiều lỗi như nhiễu, giá trị bị thiếu hoặc không chính xác. Tiền xử lý dữ liệu là bước cần thiết để làm sạch, chuẩn hóa và chuẩn bị dữ liệu cho các bước phân tích tiếp theo. Quá trình này bao gồm các hoạt động như tóm lược, làm sạch, tích hợp, chuyển đổi, rút gọn và rời rạc hóa dữ liệu. Kết quả của bước này là một tập dữ liệu nhất quán, đáng tin cậy và phù hợp với các phương pháp phân tích đã chọn.

Phân tích dữ liệu: Đây là bước trung tâm trong quy trình phân tích, nơi các kỹ thuật và công cụ phân tích được áp dụng để khám phá thông tin hữu ích từ dữ liệu. Quá trình này có thể sử dụng các phương pháp như phân tích mô tả, phân tích hồi quy, phân tích sự khác biệt, thống kê, học máy (machine learning) và khai phá dữ liệu (data mining). Các phương pháp này giúp xác định các mối quan hệ, xu hướng và thông tin ẩn trong dữ liệu, tạo cơ sở để đưa ra các quyết định chính xác.

Kết luận và dự đoán: Dựa trên kết quả phân tích, chúng ta có thể rút ra các kết luận có giá trị, hiểu rõ hơn về tình hình hiện tại và thậm chí đưa ra các dự đoán cho tương lai. Những kết luận này không chỉ cung cấp cái nhìn sâu sắc mà còn hỗ trợ việc lập kế hoạch chiến lược, tối ưu hóa hiệu suất và giải quyết các vấn đề phức tạp một cách hiệu quả.

Quy trình phân tích dữ liệu, khi được thực hiện đúng cách, giúp chuyển đổi dữ liệu thô thành thông tin có ý nghĩa, hỗ trợ mạnh mẽ trong việc ra quyết định và cải thiện hoạt động trong nhiều lĩnh vực.

1.2. Tổng quan về bài toán dự báo.

1.2.1. Lịch sử bài toán dự báo

Bài toán dự báo đã có lịch sử phát triển lâu đời, trải qua nhiều giai đoạn khác nhau, phản ánh sự tiến bộ trong tư duy và công nghệ của con người. Dưới đây là cái nhìn chi tiết về sự hình thành và phát triển của bài toán dự báo:

- *Thời kỳ tiền Công nghiệp (Trước thế kỷ 18):* Trong giai đoạn này, dự báo chủ yếu dựa trên kinh nghiệm thực tế và tri thức được truyền lại qua nhiều thế hệ. Con người quan sát các hiện tượng tự nhiên như chuyển động của các thiên thể, sự thay đổi thời tiết, hay chu kỳ mùa vụ để đưa ra dự đoán về tương lai. Các phương pháp dự báo mang tính trực giác, không có sự hỗ trợ từ khoa học hay các công cụ toán học. Tuy nhiên, điều này phản ánh nhu cầu bức thiết của con người trong việc chuẩn bị cho những sự kiện quan trọng như mùa màng, săn bắn hay di cư.
- *Cách mạng Công nghiệp và thống kê (Thế kỷ 18 - 19):* Cuộc cách mạng công nghiệp đánh dấu bước ngoặt lớn trong việc sử dụng dữ liệu và phương pháp khoa học để dự báo. Sự phát triển của các lý thuyết xác suất, thống kê và toán học cho phép con người hệ thống hóa các quan sát và biến chúng thành công cụ dự báo đáng tin cậy hơn. Các nhà khoa học bắt đầu ứng dụng các phương pháp như phân tích chuỗi thời gian để dự đoán xu hướng trong kinh doanh, khí tượng học, và tài chính. Đây cũng là thời kỳ mà dữ liệu được ghi chép và quản lý tốt hơn, mở đường cho các phân tích chính xác hơn.
- *Thế kỷ 20 và Kỹ thuật số hóa:* Với sự xuất hiện của máy tính, bài toán dự báo bước sang một kỷ nguyên mới. Máy tính cho phép xử lý dữ liệu nhanh chóng, hiệu quả và với quy mô lớn. Các mô hình toán học phức tạp như hồi quy tuyến

tính, phân tích chuỗi thời gian ARIMA và các phương pháp thống kê tiên tiến được áp dụng trong nhiều lĩnh vực như kinh tế, kỹ thuật, y học và quản lý. Đồng thời, sự phát triển của công nghệ thông tin và kỹ thuật số hóa đã thúc đẩy việc thu thập và lưu trữ dữ liệu, tạo cơ sở vững chắc cho các mô hình dự báo hiện đại.

- *Thống kê Bayes và Kỹ thuật Machine Learning (Thế kỷ 20 - 21)*: Các phương pháp hiện đại như thống kê Bayes và học máy (machine learning) đã mang lại những tiến bộ vượt bậc trong lĩnh vực dự báo. Thống kê Bayes cho phép dự báo dựa trên xác suất và cập nhật thông tin khi có dữ liệu mới. Học sâu (deep learning), học tăng cường (reinforcement learning) và các thuật toán tiên tiến khác đã mở ra khả năng dự đoán trong các môi trường phức tạp, từ nhận dạng giọng nói, hình ảnh đến dự đoán hành vi của con người. Những công nghệ này có khả năng xử lý lượng lớn dữ liệu và phát hiện các mẫu phức tạp mà con người khó nhận biết.
- *Dự báo trong thời đại số hóa (Hiện nay)*: Hiện tại, bài toán dự báo đang ở đỉnh cao của sự phát triển nhờ sự bùng nổ của dữ liệu lớn (big data) và trí tuệ nhân tạo (AI). Các công nghệ như phân tích dữ liệu lớn, học sâu và dự báo dựa trên mạng xã hội không chỉ cải thiện độ chính xác mà còn mở rộng phạm vi ứng dụng của dự báo. Dữ liệu từ các nền tảng như mạng xã hội, cảm biến IoT và hệ thống thông minh giúp dự báo hành vi người dùng, xu hướng thị trường và cả sự kiện bất ngờ.

Bài toán dự báo đã trải qua một hành trình dài, từ những dự đoán dựa trên kinh nghiệm đơn thuần đến việc sử dụng các công cụ và công nghệ hiện đại để xử lý dữ liệu phức tạp. Mục tiêu của dự báo là xây dựng các mô hình có khả năng hiểu và ứng dụng các mẫu, xu hướng và quy luật từ dữ liệu lịch sử để đưa ra các dự đoán chính xác và đáng tin cậy.

Trong thời đại số hóa, bài toán dự báo không chỉ là một thách thức mà còn là cơ hội để con người tối ưu hóa các quyết định và phát triển bền vững. Sự tiến bộ về công nghệ đã biến dự báo từ một công cụ hỗ trợ thành một phần không thể thiếu trong các lĩnh vực như kinh tế, y tế, khoa học và quản lý. Đây là minh chứng rõ nét cho tầm quan trọng và sức mạnh của dự báo trong việc thúc đẩy sự phát triển của xã hội.

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

Bài toán dự báo có một tầm ảnh hưởng sâu rộng và quan trọng trong các lĩnh vực phát triển kinh tế, xã hội và công nghệ tại Việt Nam. Dự báo không chỉ giúp cải thiện việc quản lý, tối ưu hóa tài nguyên mà còn là công cụ quyết định trong việc định hình chiến lược phát triển dài hạn. Dưới đây là một cái nhìn chi tiết về sự phát triển và ứng dụng của bài toán dự báo tại Việt Nam:

- *Phát triển đang ở giai đoạn đầu:* Tại Việt Nam, bài toán dự báo vẫn đang trong quá trình nghiên cứu, thử nghiệm và áp dụng. Mặc dù có sự quan tâm lớn đến việc ứng dụng các phương pháp phân tích dữ liệu hiện đại như học máy, trí tuệ nhân tạo, và phân tích dữ liệu lớn, nhưng việc triển khai thực tế ở nhiều lĩnh vực vẫn còn gặp phải khó khăn. Năng lực ứng dụng các công nghệ mới vẫn chưa đồng đều giữa các ngành nghề, và sự thiếu hụt về nguồn nhân lực chất lượng cao trong phân tích dữ liệu cũng là một thách thức lớn. Tuy nhiên, sự nhận thức và đầu tư vào công nghệ phân tích dự báo đang gia tăng mạnh mẽ, mở ra cơ hội phát triển lớn trong tương lai.
- *Ứng dụng trong nông nghiệp và kinh tế:* Nông nghiệp là một trong những lĩnh vực tiên phong trong việc ứng dụng bài toán dự báo tại Việt Nam. Dự báo giúp nông dân dự đoán các yếu tố như thời tiết, mùa vụ, sự biến động của dịch bệnh và nhu cầu tiêu thụ, từ đó điều chỉnh kế hoạch sản xuất kịp thời và hiệu quả. Các mô hình dự báo trong nông nghiệp giúp tối ưu hóa năng suất, giảm thiểu rủi ro và tăng trưởng bền vững. Ngoài ra, trong lĩnh vực kinh tế, việc dự báo vĩ mô như tăng trưởng GDP, lạm phát, và tỷ giá hối đoái đang ngày càng trở nên quan trọng trong việc xây dựng chính sách và chiến lược phát triển. Các công ty tài chính và ngân hàng tại Việt Nam đã bắt đầu sử dụng các mô hình dự báo thống kê để phân tích các xu hướng thị trường, giúp cải thiện quyết định đầu tư và phân bổ nguồn lực. Các mô hình dự báo này đang ngày càng đóng vai trò quan trọng trong việc định hướng chiến lược phát triển của các tổ chức.
- *Thách thức từ dữ liệu:* Một trong những thách thức lớn nhất mà bài toán dự báo tại Việt Nam phải đối mặt chính là chất lượng và tính nhất quán của dữ liệu. Dữ liệu thu thập được từ nhiều nguồn khác nhau đôi khi thiếu đầy đủ, không đồng nhất hoặc không chính xác. Điều này khiến việc xây dựng các mô hình dự báo trở nên khó khăn và kết quả dự báo không luôn đạt độ chính xác

như mong muốn. Để giải quyết vấn đề này, việc cải thiện cơ sở hạ tầng công nghệ thông tin, chuẩn hóa quy trình thu thập và xử lý dữ liệu là yêu cầu cấp thiết.

Tổng quan về sự phát triển của bài toán dự báo tại Việt Nam:

- *Thời kỳ tiền Công nghiệp và Cách mạng Công nghiệp:* Trước thế kỷ 18, bài toán dự báo tại Việt Nam chủ yếu dựa vào kinh nghiệm và tri thức dân gian truyền lại qua các thế hệ. Người dân Việt Nam đã sử dụng các quan sát về thiên nhiên, thời tiết và các hiện tượng tự nhiên để đưa ra các dự báo sơ bộ về mùa màng và các sự kiện quan trọng khác. Dự báo trong giai đoạn này mang tính chất thủ công và không có sự hỗ trợ của các phương pháp khoa học.
- *Thế kỷ 20 và Kỹ thuật số hóa:* Với sự phát triển của công nghệ và máy tính, Việt Nam đã bắt đầu tiếp cận các phương pháp dự báo thống kê trong thế kỷ 20. Các mô hình toán học như hồi quy, phân tích chuỗi thời gian đã được sử dụng để dự đoán các yếu tố vĩ mô như thị trường tiêu thụ, sản xuất, giá cả và các chỉ số tài chính. Sự phổ biến của máy tính và các công cụ phần mềm đã mở ra cơ hội mới cho việc áp dụng các phương pháp dự báo chính xác và khoa học hơn trong các ngành công nghiệp.
- *Thống kê Bayes và Kỹ thuật Machine Learning:* Những năm gần đây, các phương pháp dự báo tiên tiến như thống kê Bayes và học máy (machine learning) đã được áp dụng rộng rãi tại Việt Nam. Những phương pháp này giúp phân tích các dữ liệu phức tạp và tìm ra những mẫu ẩn sau các tập dữ liệu lớn. Các ngành như tài chính, y tế, sản xuất, và thương mại điện tử đang tận dụng các mô hình dự báo máy học để cải thiện độ chính xác trong dự đoán và tối ưu hóa chiến lược kinh doanh. Các tổ chức nghiên cứu tại Việt Nam cũng đã và đang triển khai các mô hình học sâu (deep learning) để dự báo hành vi người tiêu dùng, xu hướng thị trường và các yếu tố tác động đến nền kinh tế.
- *Sự phát triển của Big Data và dự báo dựa trên mạng xã hội:* Việt Nam cũng đã bắt đầu nhận thức rõ về tiềm năng của Big Data và dữ liệu từ các nền tảng mạng xã hội trong việc dự báo. Dữ liệu từ các nguồn như Facebook, Twitter, và các ứng dụng mạng xã hội khác cung cấp một lượng thông tin khổng lồ về hành vi người dùng, xu hướng tiêu dùng, và những sự kiện đang diễn ra trong xã hội. Các công ty công nghệ, ngân hàng và các tổ chức nghiên cứu đã bắt đầu áp dụng phương pháp phân tích dữ liệu lớn để dự báo các xu hướng thị

trường và hành vi người tiêu dùng. Dự báo dựa trên mạng xã hội hiện đang là một trong những lĩnh vực nghiên cứu thú vị và nhiều tiềm năng tại Việt Nam.

Bài toán dự báo tại Việt Nam đang có những bước tiến mạnh mẽ trong việc áp dụng các công nghệ mới, đặc biệt là trong các lĩnh vực nông nghiệp, kinh tế và công nghiệp. Mặc dù vẫn còn gặp phải một số thách thức về chất lượng dữ liệu và nguồn nhân lực, nhưng sự phát triển nhanh chóng của công nghệ và sự chuyển mình trong nhận thức về tầm quan trọng của dự báo sẽ tạo ra cơ hội lớn cho việc ứng dụng dự báo vào các hoạt động chiến lược trong tương lai. Bài toán dự báo sẽ tiếp tục đóng vai trò quan trọng trong việc giúp các tổ chức và doanh nghiệp tại Việt Nam đưa ra quyết định chính xác và phát triển bền vững.

1.2.3. Tình hình phát triển của bài toán dự báo ở thế giới

Phát triển mạnh mẽ và ứng dụng toàn diện: Ở các quốc gia phát triển, bài toán dự báo đã được triển khai và áp dụng một cách rộng rãi và hiệu quả trong nhiều lĩnh vực quan trọng như tài chính, thương mại điện tử, y tế, năng lượng và sản xuất. Những phương pháp dự báo hiện đại giúp các ngành này không chỉ cải thiện khả năng ra quyết định mà còn tối ưu hóa quy trình vận hành, dự đoán nhu cầu tiêu thụ, xác định các rủi ro tiềm ẩn và tối ưu hóa nguồn lực. Việc ứng dụng các mô hình dự báo không chỉ mang lại lợi ích tức thì mà còn tạo ra nền tảng vững chắc cho các chiến lược dài hạn, thúc đẩy tăng trưởng bền vững và đổi mới sáng tạo trong các lĩnh vực này.

Tích hợp công nghệ tiên tiến để nâng cao hiệu quả: Các quốc gia phát triển hiện đang áp dụng sự kết hợp mạnh mẽ giữa các công nghệ tiên tiến như trí tuệ nhân tạo (AI), học máy (machine learning), phân tích dữ liệu lớn (big data analytics), và các công nghệ đột phá khác để nâng cao hiệu suất dự báo. Việc tích hợp các công nghệ này giúp xử lý và phân tích khối lượng dữ liệu khổng lồ một cách nhanh chóng và hiệu quả, đồng thời cho phép mô hình dự báo tự động học và cải tiến theo thời gian. Công nghệ AI và học máy còn có khả năng nhận diện những mẫu dữ liệu phức tạp, phát hiện xu hướng ẩn mà các phương pháp phân tích truyền thống khó có thể nắm bắt, từ đó đưa ra các dự báo chính xác hơn trong nhiều lĩnh vực, từ dự báo tài chính đến dự báo nhu cầu khách hàng trong thương mại điện tử.

Khả năng tổng hợp và tích hợp dữ liệu đa dạng: Một trong những ưu thế nổi bật của các quốc gia phát triển là khả năng thu thập, tổng hợp và tích hợp dữ liệu từ

hiều nguồn khác nhau, chẳng hạn như hệ thống cảm biến, mạng xã hội, giao dịch tài chính, dữ liệu từ các thiết bị thông minh, và các nguồn thông tin mở. Khả năng này không chỉ giúp xây dựng các mô hình dự báo mạnh mẽ mà còn mang đến cái nhìn sâu sắc và toàn diện về các yếu tố tác động đến các dự báo. Việc tích hợp dữ liệu từ nhiều nguồn phong phú giúp tăng độ chính xác và đa chiều của dự báo, từ đó cung cấp những thông tin quan trọng giúp các doanh nghiệp, tổ chức và chính phủ đưa ra các quyết định chiến lược kịp thời và hiệu quả.

1.3. Giới thiệu Bài toán

Trong thời đại ngày nay, khi thị trường bất động sản phát triển mạnh mẽ, việc hiểu rõ các yếu tố ảnh hưởng đến giá nhà là vô cùng quan trọng để các nhà đầu tư, chủ sở hữu và người mua có thể đưa ra các quyết định sáng suốt. Trong một thế giới đầy biến động và đa dạng, giá cả của mỗi ngôi nhà không chỉ là một con số, mà còn là bức tranh phức tạp phản ánh nhiều yếu tố khác nhau, từ vị trí địa lý, cơ sở hạ tầng đến nhu cầu và xu hướng xã hội.



Hình 1. 2: Giới thiệu về bài toán

Khi công nghệ ngày càng tiên tiến và sự cạnh tranh trở nên khốc liệt, giá nhà ở không chỉ phản ánh chất lượng xây dựng và tiện ích, mà còn là kết quả của quá trình đổi mới và khả năng đáp ứng linh hoạt với nhu cầu thị trường. Các yếu tố như vị trí, cơ sở hạ tầng, tiện ích công cộng, và thậm chí là ảnh hưởng của các yếu tố xã hội, như xu hướng bảo vệ môi trường và phát triển bền vững, đều góp phần tạo nên sự đa dạng trong phân khúc giá bất động sản.

Giá nhà không chỉ là một con số trên hợp đồng, mà còn phản ánh một quá trình nghiên cứu và phát triển không ngừng trong ngành bất động sản. Các chủ đầu tư liên tục tìm cách cải tiến trong thiết kế, vật liệu xây dựng, và các công nghệ thông minh để tạo ra những ngôi nhà không chỉ đẹp mắt và tiện nghi mà còn đáp ứng mong muốn và yêu cầu ngày càng cao của khách hàng.

Mặc dù giá cả có thể là một trở ngại đối với nhiều người, nhưng nó cũng là nguồn thông tin quan trọng để hiểu rõ giá trị thực sự của mỗi ngôi nhà. Bằng cách phân tích giá, chúng ta có thể nhận thức được những yếu tố quyết định đến giá trị của từng căn nhà, từ những căn hộ tiết kiệm năng lượng cho đến những biệt thự sang trọng với thiết kế đẳng cấp.

Thực tế, phân tích giá không chỉ là công cụ hữu ích cho người mua và chủ đầu tư mà còn là động lực để các nhà phát triển bất động sản không ngừng cải thiện sản phẩm của mình. Nhờ đó, mỗi con số trên biểu đồ giá không chỉ là một dữ liệu khô khan mà còn là câu chuyện về sự sáng tạo và cam kết của ngành bất động sản đối với sự tiến bộ và hoàn thiện không ngừng.

Phân tích giá nhà không chỉ là công việc của các chuyên gia trong ngành, mà còn là một hành trình mà mỗi người mua nhà có thể tham gia để hiểu rõ hơn về lý do mỗi ngôi nhà có giá trị như vậy. Hãy cùng nhau bước vào thế giới phức tạp này, khám phá những câu chuyện đằng sau các con số và tìm hiểu về sự đa dạng và phong cách của thị trường nhà ở hiện nay.

1.4. Đầu vào và đầu ra của bài toán

Bước vào thế giới phức tạp của bài toán phân tích dự báo giá nhà ở Mỹ, chúng ta không chỉ đối mặt với những con số và dữ liệu, mà còn là sự thấu hiểu sâu sắc về các yếu tố đầu vào và kỳ vọng về đầu ra. Đầu vào của bài toán này không chỉ giới

hạn trong các đặc điểm vật lý của ngôi nhà, mà còn mở rộng đến các biến liên quan đến kinh tế, thị trường bất động sản, nhu cầu của người mua, và xu hướng xã hội.

Các yếu tố như diện tích, số lượng phòng ngủ, số tầng là những thông tin cơ bản về ngôi nhà, nhưng để có cái nhìn toàn diện, chúng ta cần xem xét cả các yếu tố khác như vị trí, chất lượng xây dựng, và uy tín của khu vực. Điều này đặt ra câu hỏi: Làm thế nào những yếu tố này tương tác với nhau và ảnh hưởng đến giá trị của mỗi căn nhà?

Ngoài ra, đầu vào còn bao gồm các yếu tố kinh tế như lãi suất ngân hàng, tỷ lệ lạm phát, và xu hướng thị trường lao động, vì đây là những yếu tố tác động mạnh mẽ đến khả năng chi trả của người mua và nhu cầu chung về bất động sản. Việc biến động lãi suất hoặc thay đổi trong nền kinh tế có thể tác động đến giá nhà như thế nào là một thách thức lớn trong quá trình dự đoán giá nhà ở.

Khi nhìn vào đầu ra của bài toán, đó không chỉ là con số cuối cùng mà còn là sự hiểu biết về giá trị thực sự của mỗi căn nhà. Làm thế nào giá cả được xác định, làm thế nào nó phản ánh nhu cầu và mong đợi của người mua, và làm thế nào nó thay đổi theo thời gian - tất cả đều là những khía cạnh quan trọng cần được nghiên cứu.

Tóm lại, đầu vào và đầu ra của bài toán phân tích giá nhà ở không chỉ là số liệu và con số, mà còn là câu chuyện phức tạp về sự tương tác giữa các yếu tố kinh tế, thị trường, và xã hội. Bằng cách hiểu rõ những yếu tố này, chúng ta có thể đưa ra dự đoán chính xác và có được nhận thức sâu sắc về giá trị thực sự của mỗi ngôi nhà trên thị trường bất động sản đa dạng ngày nay.

1.5. Tầm quan trọng của bài toán

Hiện nay, khi thị trường bất động sản tại Mỹ ngày càng đa dạng và cạnh tranh, bài toán phân tích giá nhà trở thành một khía cạnh quan trọng và không thể thiếu của ngành này. Việc hiểu rõ về cơ cấu giá, các yếu tố ảnh hưởng và mối quan hệ giữa chúng không chỉ là chìa khóa để định hình chiến lược đầu tư mà còn giúp các quyết định đó đảm bảo tính cạnh tranh và phản ánh đúng nhu cầu của thị trường.

Tầm quan trọng của bài toán phân tích giá nhà không chỉ nằm ở việc xác định giá trị cuối cùng của mỗi căn nhà, mà còn ở khả năng dự báo và hiểu biết sâu sắc về xu hướng thị trường. Điều này giúp các nhà đầu tư, chủ đầu tư và các bên liên quan

không chỉ đưa ra quyết định chiến lược về giá mà còn xác định được sự phù hợp với đối tượng khách hàng và nâng cao giá trị thương hiệu.

Ngoài ra, bài toán phân tích giá nhà còn đóng vai trò quan trọng trong việc tạo ra sự minh bạch và công bằng trên thị trường. Khi người mua nhà có thể hiểu rõ lý do một ngôi nhà có mức giá nhất định thông qua việc phân tích các yếu tố như vị trí, tiện ích, và chất lượng xây dựng, điều đó sẽ giúp xây dựng niềm tin và góp phần vào sự ổn định của thị trường.

Phân tích giá nhà ở Mỹ không chỉ giúp các doanh nghiệp trong ngành bất động sản mà còn mang lại lợi ích toàn diện, từ việc hỗ trợ người tiêu dùng trong quá trình lựa chọn căn nhà phù hợp đến việc định hình xu hướng và tiêu chuẩn ngành. Đây không chỉ là một khía cạnh kỹ thuật, mà còn là yếu tố quyết định đến sự thành công và phát triển bền vững của ngành bất động sản trước những thách thức phức tạp ngày nay.

1.6. Ứng dụng

Ứng dụng của bài toán phân tích giá nhà ở Mỹ trong thực tế rất đa dạng và có vai trò quan trọng trong quá trình quản lý, lập chiến lược đầu tư và hiểu biết về thị trường bất động sản. Trước tiên, phân tích giá giúp các nhà đầu tư và các công ty bất động sản xác định mức giá hợp lý cho từng bất động sản, tối ưu hóa lợi nhuận và đồng thời thu hút đúng nhóm khách hàng.

Một ứng dụng quan trọng khác của phân tích giá nhà là trong việc xây dựng chiến lược tiếp thị và quảng cáo. Việc hiểu rõ cơ cấu giá giúp các doanh nghiệp định hình thông điệp tiếp thị phù hợp, tập trung vào những giá trị và đặc điểm nổi bật của bất động sản như vị trí, tiện ích, và thiết kế nhằm thu hút sự quan tâm của khách hàng tiềm năng.

Ngoài ra, phân tích giá còn đóng vai trò trong quản lý chiến lược phân bổ và cung cấp nhà ở dựa trên dự báo giá và nhu cầu thị trường. Điều này giúp các công ty bất động sản cân bằng được cung cầu, tránh tình trạng dư thừa hoặc thiếu hụt nhà ở, và duy trì sự linh hoạt trong các dự án xây dựng và phân phối.

Đối với người mua nhà, phân tích giá giúp họ hiểu rõ hơn về giá trị thực sự và tính công bằng của một căn nhà, từ đó đưa ra quyết định mua sắm sáng suốt. Sự minh

bạch về cơ cấu giá cũng tạo ra một thị trường bất động sản lành mạnh và tin cậy, giúp người tiêu dùng tự tin hơn khi so sánh và lựa chọn giữa các bất động sản.

Cuối cùng, ứng dụng của bài toán phân tích giá nhà ở không chỉ đóng vai trò quan trọng trong chiến lược kinh doanh của các doanh nghiệp bất động sản mà còn góp phần hình thành một thị trường bất động sản linh hoạt và phát triển bền vững. Qua đó, bài toán này mang lại lợi ích không chỉ cho các công ty mà còn cho người tiêu dùng và cho sự ổn định, phát triển của ngành bất động sản trong dài hạn.

1.7. Cơ hội và hạn chế

Bài toán phân tích giá nhà ở Mỹ không chỉ mang lại nhiều cơ hội mà còn đối mặt với những hạn chế lớn, tạo nên một thách thức đa chiều cho ngành bất động sản trong việc nắm bắt và dự đoán giá trị của nhà ở.

Cơ Hội:

- *Tăng cường khả năng định giá chính xác:* Bài toán phân tích giá nhà giúp hiểu rõ hơn về mối liên hệ giữa các yếu tố như vị trí, diện tích, tiện ích và giá trị của bất động sản. Nhờ đó, các doanh nghiệp có thể tối ưu hóa chiến lược định giá, thu hút khách hàng bằng những sản phẩm phù hợp với nhu cầu của họ.

- *Phát triển chiến lược tiếp thị và thương hiệu:* Bài toán này là công cụ đắc lực để xây dựng chiến lược tiếp thị và định hình thương hiệu, nhờ vào sự hiểu biết sâu sắc về cơ cấu giá bất động sản. Điều này mở ra cơ hội để định vị thương hiệu, từ phân khúc nhà ở giá rẻ đến phân khúc cao cấp.

- *Cải thiện sự hiểu biết về thị trường:* Phân tích giá nhà ở cung cấp cái nhìn toàn cảnh và sâu sắc về thị trường. Điều này giúp các công ty bất động sản có vị thế cạnh tranh hơn, kịp thời điều chỉnh chiến lược kinh doanh, phù hợp với những biến động và xu hướng thị trường.

Hạn Chế:

- *Khó khăn trong dự đoán biến động giá:* Một thách thức lớn của phân tích giá nhà ở là dự đoán sự biến động giá qua thời gian. Các yếu tố như chi phí nguyên vật liệu xây dựng, lãi suất vay, hay thậm chí là biến động kinh tế có thể làm cho chiến lược định giá trở nên khó duy trì.

- *Ảnh hưởng của yếu tố khách quan:* Các yếu tố khách quan như tình hình kinh tế, chính sách nhà ở, và tình trạng môi trường có thể ảnh hưởng đến giá bất động sản

theo cách khó lường trước. Điều này làm cho việc dự đoán và ổn định giá nhà ở trở nên khó khăn.

- *Sự thay đổi của nhu cầu thị trường*: Bài toán cũng đối mặt với sự thay đổi liên tục của nhu cầu thị trường. Ví dụ, sự gia tăng nhu cầu nhà ở sinh thái, công nghệ thông minh, hay nhà ở theo phong cách đô thị đều làm cho thị trường nhà ở trở nên phức tạp và đa chiều hơn.

Tóm lại, bài toán phân tích giá nhà ở Mỹ mang lại cơ hội lớn cho các doanh nghiệp trong ngành bất động sản nhằm tối ưu hóa giá trị và xây dựng chiến lược kinh doanh hiệu quả. Tuy nhiên, bài toán này cũng đối mặt với những thách thức đáng kể, từ những biến động khó đoán của thị trường đến tác động của các yếu tố khách quan như chính sách kinh tế và môi trường. Những yếu tố này đòi hỏi ngành bất động sản phải không ngừng đổi mới, duy trì tính linh hoạt và sẵn sàng ứng phó với những thay đổi liên tục để giữ vững vị thế và đảm bảo sự phát triển bền vững trong một thị trường đầy biến động.

CHƯƠNG 2: PHƯƠNG PHÁP KỸ THUẬT

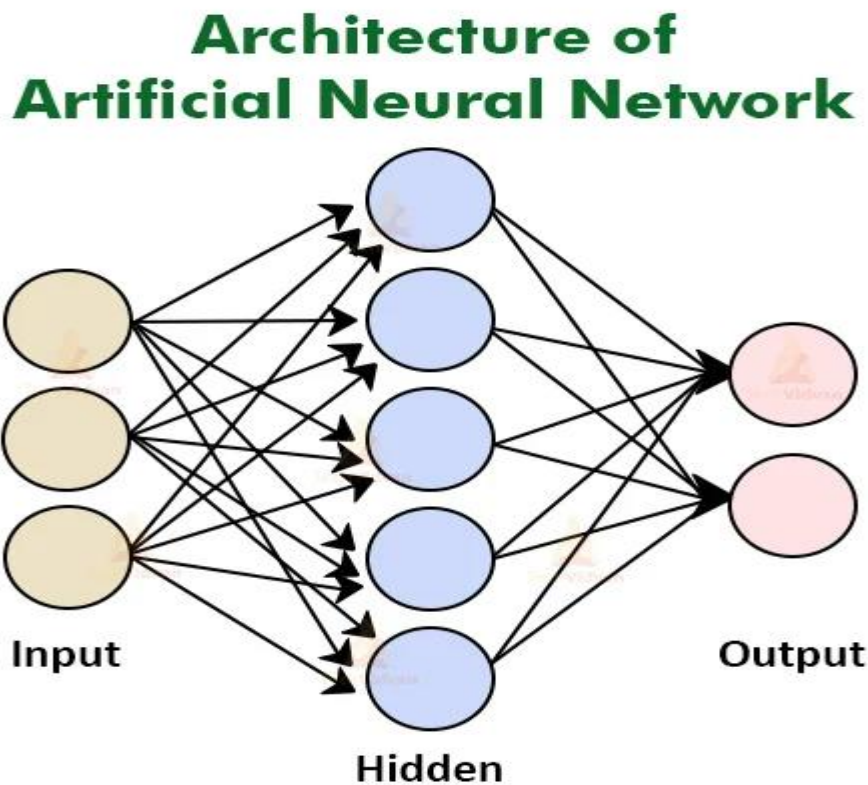
2.1. Phương hướng tiếp cận bài toán

Quá trình giải quyết bài toán bắt đầu bằng việc thu thập và tiền xử lý các bộ dữ liệu thực nghiệm liên quan tới giá nhà. Tiếp đến là nghiên cứu và áp dụng các kỹ thuật để dự đoán giá nhà sẽ được bán ra.

2.2. Mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN)

2.2.1. Giới thiệu

Về cơ bản, đây là một mô hình tính toán. Chúng được xây dựng dựa trên cấu trúc và chức năng của mạng lưới nơ-ron trong Sinh học (mặc dù cấu trúc của ANN sẽ bị ảnh hưởng bởi một luồng thông tin). ANN lấy ý tưởng từ cách hoạt động của bộ não con người để tạo ra các kết nối phù hợp, và gồm nhiều lớp nơ-ron, bao gồm lớp đầu vào, lớp ẩn và lớp đầu ra. Do đó, mạng nơ-ron này sẽ thay đổi, phụ thuộc vào các lớp đó.



Hình 2. 1: Giới thiệu về ANN

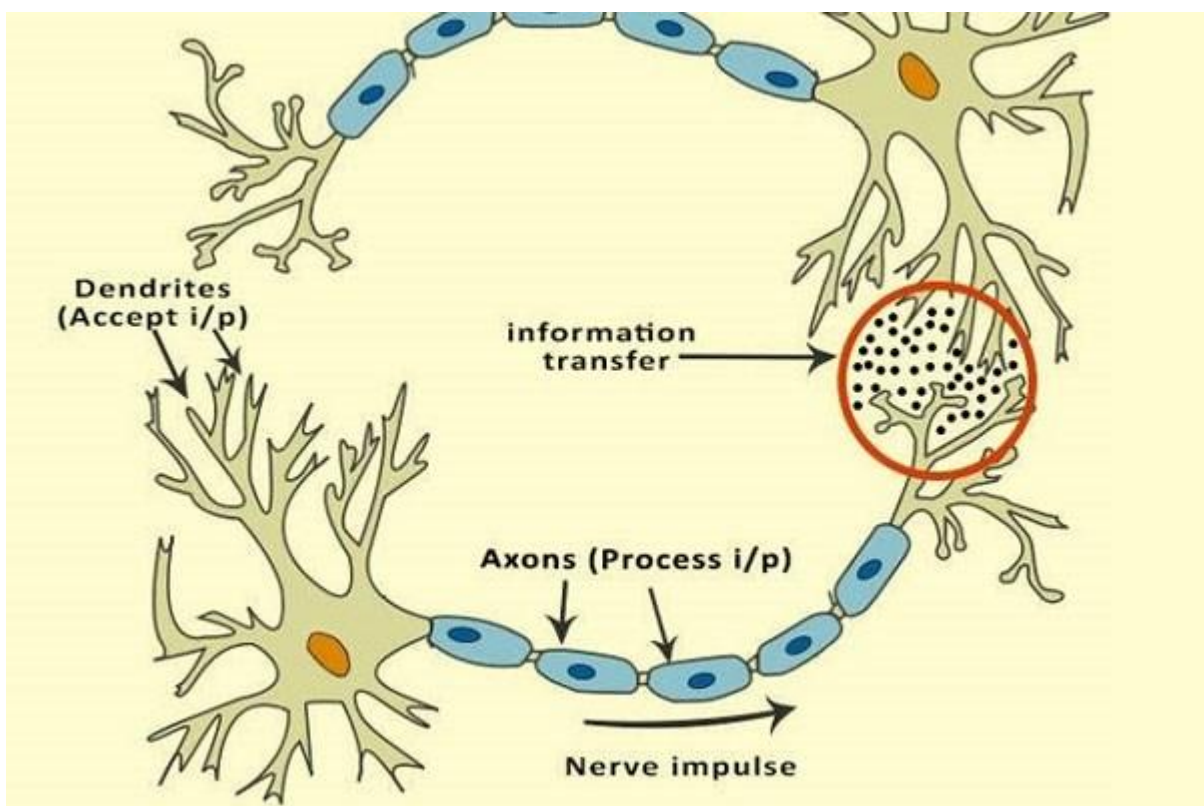
Trong bài toán dự đoán giá nhà, ANN có thể được sử dụng để hiểu và mô hình hóa mối quan hệ phức tạp giữa các đặc trưng của nhà và giá của nó.

2.2.2. Cấu trúc của mạng nơ-ron nhân tạo

Như trong phần giới thiệu, ANN được lấy ý tưởng từ cách hoạt động của bộ não con người - tạo ra các kết nối phù hợp. Do đó, ANN đã sử dụng các silicon và dây điện để làm nơ-ron và đuôi gai sống cho mình.

Trong cơ thể con người, 1 phần não đã bao gồm 86 tỷ tế bào thần kinh và chúng được kết nối với hàng nghìn tế bào khác thông qua Axons. Bởi vì con người có rất nhiều đầu vào thông tin khác nhau từ các giác quan, nên cơ thể cũng có nhiều đuôi gai để giúp truyền thông tin này.

Chúng sẽ tạo ra xung điện để di chuyển, truyền thông tin trong mạng lưới nơ-ron thần kinh này. Và điều này cũng tương tự cho mạng nơ-ron nhân tạo ANN - Khi cần xử lý các vấn đề khác nhau, nơ-ron sẽ gửi một thông điệp đến một nơ-ron khác.

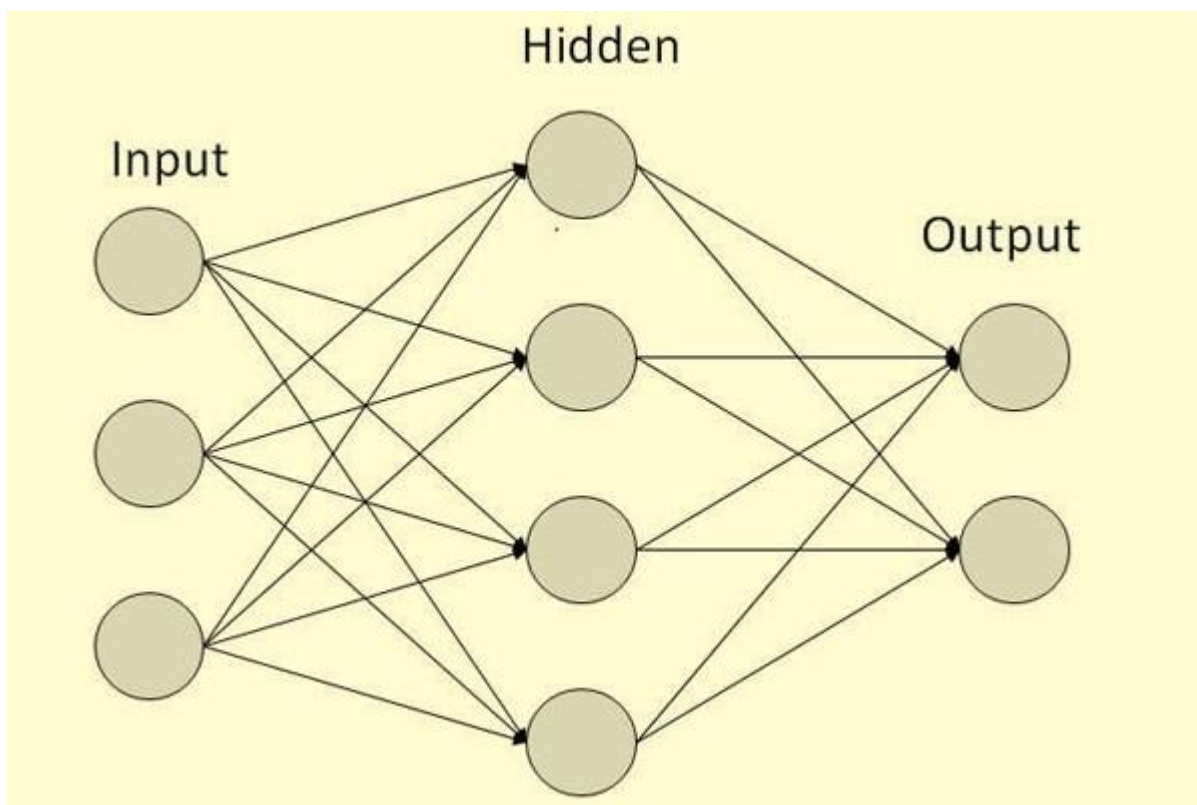


Hình 2. 2: Cấu trúc cơ bản của mạng nơ-ron ANN

Do đó, chúng ta có thể nói rằng ANN sẽ bao gồm nhiều nút bên trong, chúng bắt chước các tế bào thần kinh sinh học bên trong não người. Các mạng ANN sẽ kết nối các nơ-ron này bằng các liên kết và chúng có tương tác với nhau.

Các nút trong ANN được sử dụng để lấy dữ liệu đầu vào. Hơn nữa, việc thực hiện các thao tác trên dữ liệu cũng rất đơn giản. Sau khi thực hiện những thao tác với dữ liệu, các hoạt động này được chuyển cho các tế bào thần kinh khác. Đầu ra tại mỗi nút được gọi là giá trị kích hoạt hoặc giá trị nút của nó.

Mỗi liên kết trong mạng ANN đều có liên quan với trọng lượng. Ngoài ra, chúng có khả năng học hỏi. Điều đó sẽ diễn ra bằng cách thay đổi các giá trị trọng lượng. Dưới đây là một hình minh họa về một ANN đơn giản:



Hình 2. 3: Mô hình ANN cơ bản

2.2.3. Phân loại

Hiện nay đang có hai loại ANN là FeedForward và Feedback.

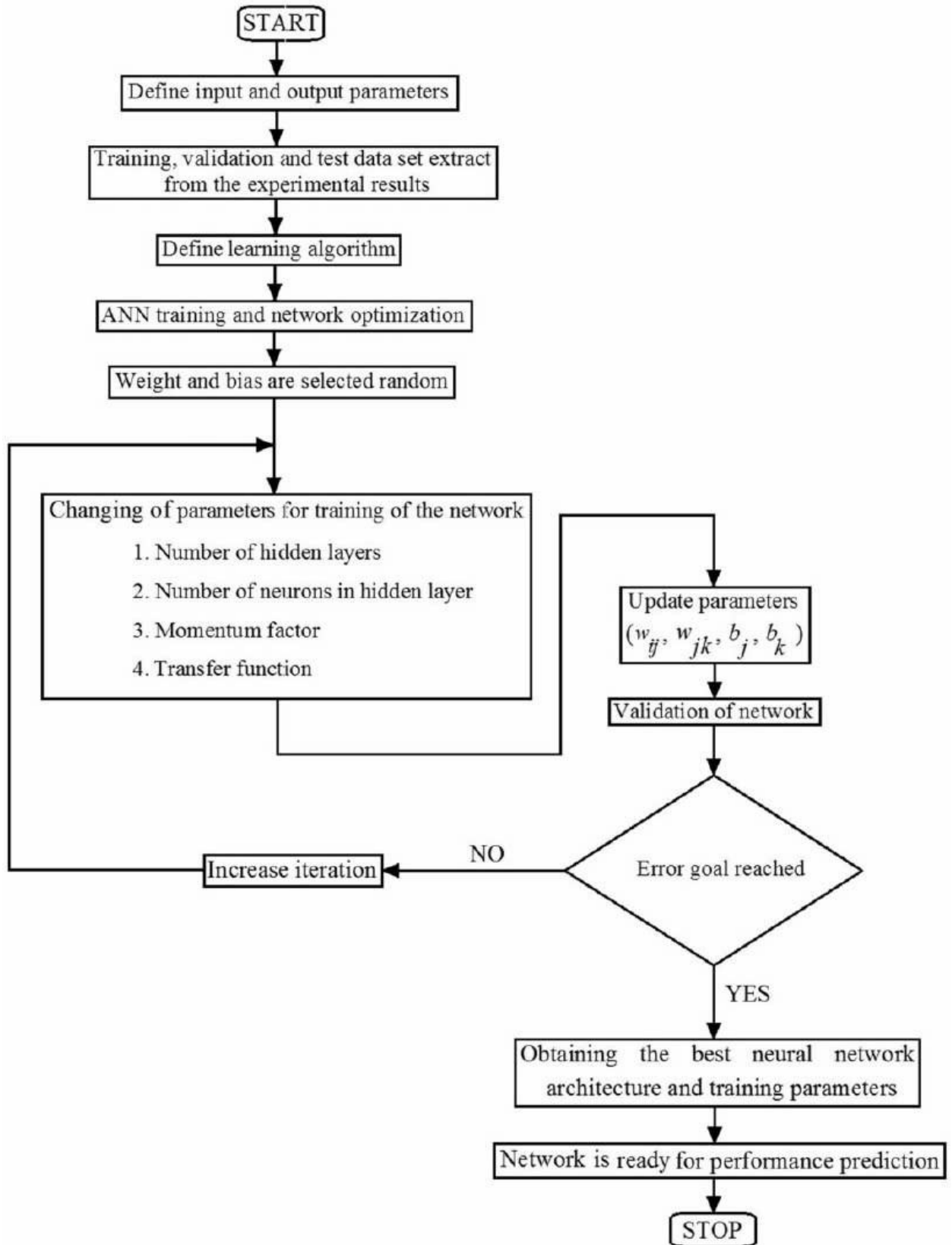
FeedForward ANN:

Mạng FeedForward ANN có luồng thông tin một chiều. Một đơn vị sẽ được sử dụng để gửi thông tin cho một đơn vị khác mà không nhận được bất kỳ thông tin nào. Ngoài ra, chúng sẽ không xuất hiện vòng phản hồi (gửi ngược thông tin về lại). Mô hình này thường được sử dụng để nhận dạng một mẫu cụ thể, vì chúng chứa các đầu vào và đầu ra cố định.

FeedBack ANN:

Trong mạng Nơron nhân tạo này, chúng sẽ cho phép các vòng lặp phản hồi. Chúng ta thường sử dụng mô hình này trong các bộ nhớ có thể giải quyết nội dung.

2.2.4. Quy trình hoạt động



Hình 2. 4: Quy trình hoạt động của mạng ANN

Quy trình hoạt động của mạng ANN bắt đầu bằng việc chuẩn bị dữ liệu đầu vào. Dữ liệu cần được thu thập từ các nguồn đáng tin cậy, sau đó làm sạch và chuẩn hóa để đảm bảo tính đồng nhất và khả năng xử lý hiệu quả bởi mạng. Các dữ liệu phân loại phải được mã hóa thành các giá trị số, giúp mạng có thể xử lý chúng dễ dàng hơn.

Tiếp theo, mạng sẽ xây dựng kiến trúc bao gồm các lớp khác nhau. Lớp đầu vào nhận dữ liệu đã chuẩn hóa từ bên ngoài, lớp ẩn sẽ xử lý và trích xuất các đặc trưng ẩn trong dữ liệu, và lớp đầu ra sẽ đưa ra kết quả dự đoán hoặc phân loại dựa trên các tín hiệu từ các lớp ẩn.

Khi dữ liệu đầu vào được cung cấp, quá trình lan truyền tiến sẽ xảy ra. Tại mỗi nơ-ron trong mạng, tín hiệu đầu vào sẽ được nhân với trọng số và cộng thêm hệ số bù. Sau đó, tín hiệu này sẽ trải qua một hàm kích hoạt để tạo ra đầu ra của nơ-ron đó. Tín hiệu tiếp tục truyền qua các lớp cho đến khi đạt đến lớp đầu ra, nơi mà kết quả cuối cùng được tính toán.

Sau khi có kết quả dự đoán, mạng sẽ tính toán sai số bằng cách đo lường sự khác biệt giữa kết quả dự đoán và giá trị thực tế. Một hàm mất mát sẽ được sử dụng để tính toán sai số này, tùy thuộc vào loại bài toán, có thể là Mean Squared Error (MSE) cho bài toán hồi quy hoặc Cross-Entropy cho bài toán phân loại.

Để giảm sai số và tối ưu hóa mô hình, quá trình lan truyền ngược sẽ được thực hiện. Tại đây, gradient (đạo hàm) của hàm mất mát đối với trọng số và hệ số bù sẽ được tính toán. Gradient này giúp xác định mức độ thay đổi của hàm mất mát khi trọng số thay đổi. Sau đó, thuật toán tối ưu như Gradient Descent sẽ được sử dụng để cập nhật các giá trị này, nhằm giảm sai số.

Mô hình sẽ được huấn luyện qua nhiều vòng lặp (epoch) với dữ liệu huấn luyện, và mục tiêu là giảm dần sai số trong mỗi vòng lặp. Các chỉ số hiệu suất như độ chính xác, MSE hoặc F1-score sẽ được theo dõi để đánh giá hiệu quả của quá trình huấn luyện.

Cuối cùng, sau khi mô hình đã được huấn luyện và kiểm tra, nó sẽ được triển khai để sử dụng trong các ứng dụng thực tế. Mô hình sẽ đưa ra các dự đoán hoặc phân loại cho dữ liệu mới, giúp giải quyết các vấn đề thực tiễn một cách hiệu quả.

2.2.5. Ưu điểm và nhược điểm

ANN được sử dụng rất rộng rãi trong các bài toán dự báo và dự đoán. Tuy nhiên nó cũng có ưu điểm và hạn chế riêng.

Ưu điểm:

- Khả năng học mẫu phức tạp
- Xử lý dữ liệu không đồng nhất
- Tính linh hoạt và mở rộng

Nhược điểm:

- Dễ bị quá khớp
- Đòi hỏi lượng dữ liệu lớn
- Độ phức tạp và đòi hỏi tính toán cao
- Khó hiểu và khó giải thích

2.3. Random Forest

2.3.1. Giới thiệu

Random Forest là một thuật toán học máy được sử dụng trong bài toán phân loại và hồi quy. Nó là một phương pháp kết hợp của nhiều cây quyết định (Decision trees) để tạo ra một mô hình dự đoán mạnh mẽ.

2.3.2. Đặc điểm

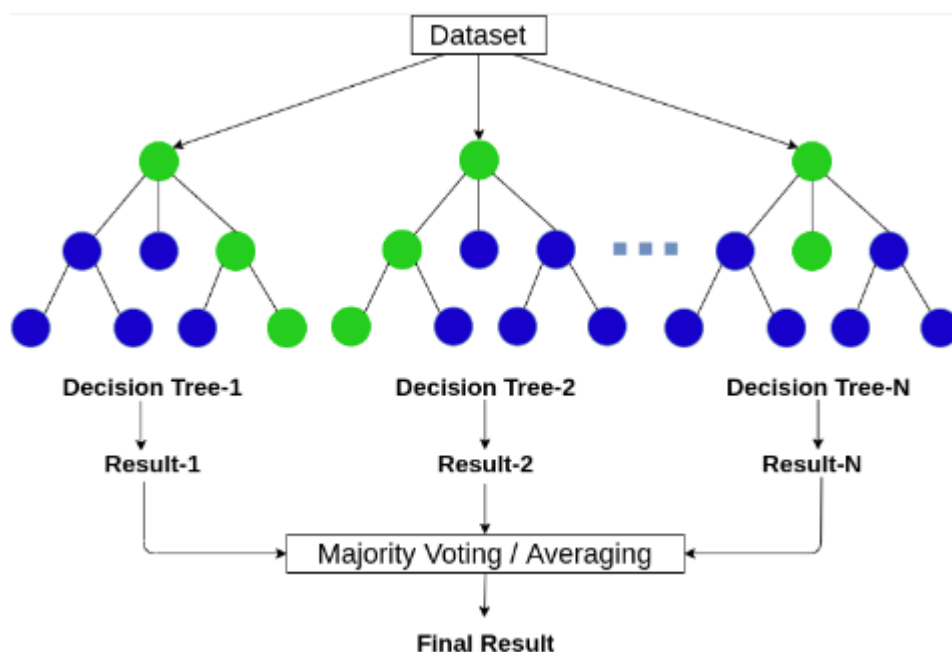
Random Forest là một thuật toán học máy mạnh mẽ và phổ biến. Điểm đáng chú ý của Random Forest là việc kết hợp nhiều cây quyết định độc lập để tạo ra một mô hình dự đoán mạnh mẽ. Mỗi cây trong tập hợp được huấn luyện trên một tập dữ liệu con được lấy mẫu ngẫu nhiên từ tập dữ liệu gốc, tạo ra sự đa dạng và giảm overfitting.

Khi có một dữ liệu mới cần dự đoán, Random Forest sẽ thu thập dự đoán từ mỗi cây và sử dụng đa số phiếu bầu để xác định kết quả cuối cùng. Điều này giúp tăng độ chính xác và ổn định của dự đoán. Ngoài ra, Random Forest cũng cung cấp thông tin về độ quan trọng của từng đặc trưng trong mô hình, giúp hiểu và phân tích tác động của các đặc trưng lên kết quả dự đoán.

Random Forest cũng có khả năng xử lý dữ liệu lớn và chống overfitting. Nó được áp dụng rộng rãi trong nhiều bài toán học máy, bao gồm phân loại, hồi quy và các bài toán dự đoán và phân tích dữ liệu. Với sự kết hợp của tính đa dạng, khả năng tổng quát hóa và khả năng xếp hạng đặc trưng, Random Forest là một công cụ quan trọng để xây dựng mô hình dự đoán mạnh mẽ và tin cậy.

2.3.3. Quy trình hoạt động

Random Forest là một thuật toán học máy dựa trên phương pháp "ensemble learning", trong đó nhiều mô hình (cây quyết định) được huấn luyện độc lập và kết quả của chúng được kết hợp lại để đưa ra dự đoán cuối cùng. Quy trình hoạt động của Random Forest có thể được chia thành các bước sau:



Hình 2. 5: Quy trình hoạt động của mô hình Random Forest

2.3.3.1. Chuẩn bị dữ liệu đầu vào

Dữ liệu huấn luyện: Mô hình Random Forest yêu cầu một tập dữ liệu huấn luyện đầy đủ với các đặc trưng và nhãn (hoặc kết quả) cho bài toán phân loại hoặc hồi quy.

Tiền xử lý dữ liệu:

- Làm sạch dữ liệu: Loại bỏ giá trị thiếu, ngoại lệ.

- Mã hóa các giá trị phân loại nếu cần (ví dụ: One-hot encoding).
- Chuẩn hóa hoặc chuẩn hóa (scaling) nếu cần thiết, mặc dù Random Forest không yêu cầu xử lý này quá chặt chẽ.

2.3.3.2. Xây dựng nhiều cây quyết định (Decision Trees)

Mô hình Random Forest xây dựng nhiều cây quyết định (n thường được gọi là số lượng ước tính trong "forest").

- *Bước 1 - Bootstrapping (Sampling có thay thế)*: Chọn ngẫu nhiên một phần dữ liệu huấn luyện (với sự thay thế), tạo ra các "bootstrap samples". Mỗi cây quyết định trong Random Forest sẽ được huấn luyện trên một tập con khác nhau của dữ liệu.
- *Bước 2 - Chọn ngẫu nhiên đặc trưng*: Khi xây dựng mỗi cây quyết định, tại mỗi nút chia, thuật toán Random Forest chỉ xét một tập hợp con ngẫu nhiên các đặc trưng, thay vì xét tất cả các đặc trưng, giúp giảm phương sai và làm giảm quá trình học quá mức (overfitting).

2.3.3.3. Huấn luyện các cây quyết định

Phân chia dữ liệu tại mỗi nút:

- Tại mỗi nút của cây quyết định, thuật toán lựa chọn đặc trưng và điểm chia sao cho giảm được "impurity" (tính không thuần nhất) của tập dữ liệu con. Các tiêu chí phổ biến để đo độ tinh khiết của các nút là **Gini impurity** (trong phân loại) hoặc **Mean Squared Error (MSE)** (trong hồi quy).
- Quá trình phân chia này tiếp tục cho đến khi cây đạt đến độ sâu tối đa hoặc không thể phân chia nữa.

Quá trình huấn luyện: Sau khi tất cả các cây quyết định được xây dựng từ các bootstrap samples và các đặc trưng ngẫu nhiên, mô hình đã hoàn tất việc huấn luyện.

2.3.3.4. Dự đoán từ mỗi cây quyết định

Khi mô hình Random Forest đã được huấn luyện, chúng ta sẽ sử dụng các cây quyết định để thực hiện dự đoán.

Dự đoán cho mỗi cây: Mỗi cây trong rừng sẽ đưa ra dự đoán của mình. Đối với bài toán phân loại, kết quả của mỗi cây là một nhãn (label). Đối với bài toán hồi quy, kết quả là giá trị số.

2.3.3.5. Kết hợp các dự đoán

Mục tiêu: Tính toán dự đoán cuối cùng của mô hình Random Forest bằng cách kết hợp kết quả của các cây quyết định.

Phân loại: Sử dụng majority voting (bỏ phiếu đa số) để đưa ra dự đoán. Tức là nhãn được dự đoán bởi nhiều cây quyết định nhất sẽ là kết quả cuối cùng.

Hồi quy: Tính toán trung bình giá trị dự đoán của tất cả các cây để đưa ra giá trị dự đoán cuối cùng.

2.3.3.6. Đánh giá mô hình

Đánh giá hiệu suất: Sau khi mô hình Random Forest đưa ra dự đoán, chúng ta sẽ đánh giá mô hình trên tập dữ liệu kiểm tra bằng các chỉ số như:

- **Độ chính xác (Accuracy):** Cho bài toán phân loại.
- **MSE (Mean Squared Error):** Cho bài toán hồi quy.
- **F1-score, Precision, Recall** cho các bài toán phân loại.

Nếu hiệu suất không đạt yêu cầu, có thể điều chỉnh các tham số như số lượng cây (`n_estimators`), độ sâu của cây (`max_depth`), và các tham số khác.

2.3.3.7. Triển khai mô hình

Áp dụng mô hình Random Forest vào thực tế để đưa ra dự đoán cho dữ liệu mới. Sau khi huấn luyện và đánh giá, mô hình Random Forest có thể được triển khai trên hệ thống sản xuất hoặc tích hợp vào các ứng dụng.

2.3.4. Ưu điểm và nhược điểm

Ưu điểm:

- *Độ chính xác cao:* Random Forest có khả năng xử lý các đặc trưng phức tạp và tìm ra các quan hệ phức tạp trong dữ liệu, do đó thường cho hiệu suất dự đoán cao.

- *Khả năng xử lý dữ liệu lớn:* Random Forest có thể xử lý được tập dữ liệu lớn và có khả năng chống lại overfitting (quá khớp) trong mô hình.
- *Khả năng xếp hạng đặc trưng:* Random Forest có thể đánh giá độ quan trọng của từng đặc trưng trong mô hình, giúp xác định xem đặc trưng nào có ảnh hưởng lớn nhất đến kết quả dự đoán.

Nhược điểm:

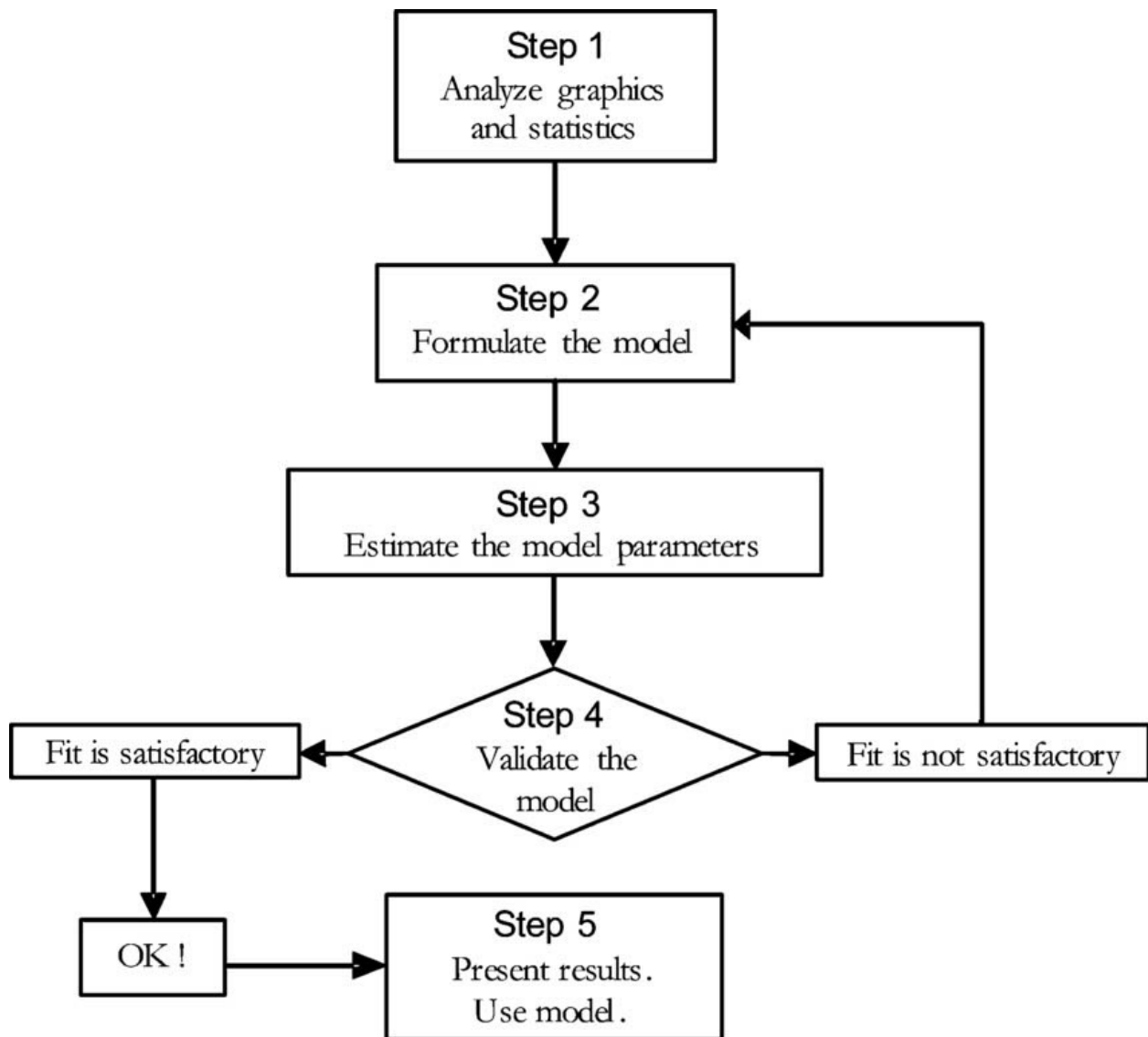
- *Khả năng diễn giải:* Một nhược điểm của Random Forest là khả năng diễn giải kém. Khi một mô hình Random Forest được xây dựng, nó cung cấp thông tin về độ quan trọng của các đặc trưng, nhưng không giải thích rõ ràng cách các đặc trưng ảnh hưởng đến kết quả dự đoán
- *Tính khả thi và tốn thời gian:* Random Forest có thể tốn nhiều thời gian để xây dựng, đặc biệt khi có một lượng lớn cây quyết định hoặc khi tập dữ liệu lớn. Việc xây dựng nhiều cây quyết định độc lập và kết hợp chúng yêu cầu tính toán phức tạp, tăng thời gian huấn luyện mô hình
- *Khả năng xử lý dữ liệu không đồng nhất:* Random Forest không xử lý tốt với các tập dữ liệu có tính không đồng nhất.

2.4. Hồi quy tuyến tính

2.4.1. Giới thiệu

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. Quy trình hoạt động



Hình 2. 6: Quy trình hoạt động mô hình hồi quy tuyến tính

Bước 1: Chuẩn bị Dữ liệu

Để mô hình hoạt động hiệu quả, dữ liệu đầu vào cần được chuẩn bị kỹ lưỡng. Quá trình chuẩn bị này bao gồm các bước quan trọng sau:

1. **Thu thập dữ liệu:** Các dữ liệu đầu vào bao gồm các đặc trưng (features) và biến mục tiêu (target variable) được thu thập từ các nguồn phù hợp, đảm bảo tính đầy đủ và chính xác.
2. **Làm sạch dữ liệu:** Xử lý các giá trị thiếu hoặc ngoại lệ trong dữ liệu là rất quan trọng. Các giá trị thiếu có thể được thay thế bằng giá trị trung bình, trung

vị hoặc theo các phương pháp phù hợp khác. Các ngoại lệ có thể bị loại bỏ hoặc điều chỉnh để tránh ảnh hưởng đến mô hình.

3. **Mã hóa dữ liệu phân loại:** Nếu dữ liệu chứa các biến phân loại, chúng cần được mã hóa thành các dạng số để mô hình có thể sử dụng. Các kỹ thuật như **One-Hot Encoding** hay **Label Encoding** thường được áp dụng.
4. **Chuẩn hóa dữ liệu (nếu cần thiết):** Mặc dù hồi quy tuyến tính không yêu cầu chuẩn hóa, nhưng nếu các đặc trưng có đơn vị đo lường khác nhau (ví dụ: chiều cao tính bằng mét, trọng lượng tính bằng kg), việc chuẩn hóa sẽ giúp mô hình không bị ảnh hưởng bởi sự chênh lệch về độ lớn giữa các đặc trưng, đồng thời giúp quá trình huấn luyện nhanh và hiệu quả hơn.

Bước 2: Xây dựng Mô hình Hồi quy Tuyến tính

Mô hình hồi quy tuyến tính giả định một mối quan hệ tuyến tính giữa các đặc trưng và biến mục tiêu. Mô hình có thể biểu diễn dưới dạng công thức:

$$y = w_1x_1 + w_2x_2 + \dots + w_nx_n + b$$

Trong đó:

- y là giá trị mục tiêu (output) mà mô hình cần dự đoán.
 x_1, x_2, \dots, x_n là các đặc trưng đầu vào.
- w_1, w_2, \dots, w_n là các trọng số (weights) mà mô hình cần tìm.
- b là hệ số bù (bias), giúp mô hình linh hoạt hơn trong việc điều chỉnh các dự đoán.

Bước 3: Huấn luyện Mô hình

Mô hình hồi quy tuyến tính được huấn luyện để tối ưu hóa các tham số w_1, w_2, \dots, w_n và b sao cho mô hình dự đoán chính xác nhất. Quá trình này diễn ra qua các bước sau:

1. Định nghĩa hàm mất mát (Loss function):

Để đánh giá sai số giữa các dự đoán của mô hình và giá trị thực tế, mô hình sử dụng **Mean Squared Error (MSE)**. MSE đo lường độ lệch bình phương giữa giá trị thực tế và giá trị dự đoán:

$$MSE = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2$$

Trong đó, m là số mẫu dữ liệu, y_i là giá trị thực tế, và \hat{y}_i là giá trị dự đoán.

2. Tối ưu hóa trọng số và hệ số bù:

Để giảm thiểu sai số, mô hình sử dụng phương pháp **Gradient Descent**, một thuật toán tối ưu hóa. Phương pháp này cập nhật các trọng số w_j và hệ số bù b theo hướng giảm dần của gradient của hàm mất mát:

$$w_j := w_j - \alpha \frac{\partial}{\partial w_j} MSE$$

$$b := b - \alpha \frac{\partial}{\partial b} MSE$$

Trong đó, α là tốc độ học (learning rate), quyết định mức độ điều chỉnh trong mỗi bước.

3. Lặp lại quá trình:

Quá trình tối ưu hóa được lặp lại qua nhiều vòng (epochs), với mục tiêu giảm thiểu giá trị MSE và giúp mô hình học được mối quan hệ tốt nhất giữa các đặc trưng và biến mục tiêu.

Bước 4: Dự đoán với Mô hình

Sau khi huấn luyện, mô hình có thể sử dụng các trọng số và hệ số bù đã được tối ưu để dự đoán giá trị y cho dữ liệu mới. Cách thức dự đoán vẫn giữ nguyên công thức đã định nghĩa ở bước 2:

$$\hat{y} = w_1 x_1 + w_2 x_2 + \dots + w_n x_n + b$$

Trong đó, \hat{y} là giá trị dự đoán từ mô hình dựa trên các đặc trưng đầu vào mới.

Bước 5: Đánh giá Mô hình

Sau khi mô hình đã hoàn thành quá trình huấn luyện và dự đoán, cần phải đánh giá hiệu quả của mô hình. Một số chỉ số thường được sử dụng để đánh giá mô hình hồi quy tuyến tính bao gồm:

1. R-squared (R^2):

R^2 là chỉ số quan trọng đo lường mức độ phù hợp của mô hình với dữ liệu. Nó phản ánh phần trăm biến động của biến mục tiêu mà mô hình có thể giải thích. Giá trị R^2 gần 1 cho thấy mô hình giải thích tốt biến động trong dữ liệu.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Trong đó, \bar{y} là giá trị trung bình của y .

2. Mean Squared Error (MSE):

MSE giúp đo lường độ sai lệch bình quân giữa giá trị thực tế và giá trị dự đoán của mô hình. MSE càng thấp chứng tỏ mô hình càng chính xác.

3. Mean Absolute Error (MAE):

MAE là sai số tuyệt đối trung bình, giúp đo lường mức độ chênh lệch giữa các giá trị dự đoán và thực tế mà không tính đến hướng sai lệch (dương hay âm). MAE càng nhỏ càng cho thấy mô hình hoạt động tốt.

Bước 6: Triển khai Mô hình

Sau khi mô hình đạt được kết quả đánh giá tốt, mô hình có thể được triển khai vào thực tế. Điều này có thể bao gồm việc tích hợp mô hình vào các ứng dụng phần mềm hoặc hệ thống phân tích dữ liệu, nơi mô hình sẽ dự đoán giá trị của biến mục tiêu cho các dữ liệu mới trong thời gian thực.

Mô hình hồi quy tuyến tính có thể được ứng dụng trong nhiều lĩnh vực, từ dự đoán giá trị tài sản, phân tích xu hướng thị trường, đến các bài toán trong quản lý chuỗi cung ứng và tối ưu hóa chiến lược kinh doanh.

Quy trình huấn luyện mô hình hồi quy tuyến tính bao gồm các bước chuẩn bị dữ liệu, xây dựng mô hình, huấn luyện mô hình, dự đoán và đánh giá. Mô hình này tìm kiếm một mối quan hệ tuyến tính giữa các đặc trưng và biến mục tiêu, tối ưu hóa các tham số (trọng số và hệ số bù) để đạt được dự đoán chính xác nhất, từ đó có thể ứng dụng trong các bài toán thực tế như dự báo tài chính, phân tích hành vi người tiêu dùng, và nhiều ứng dụng khác.

2.4.2. Các loại 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.
- ε là sai số ngẫu nhiên, biểu thị các yếu tố không thể dự đoán được trong mô hình (ε thực tế không tính được).

Công thức hồi quy tuyến tính đơn giản:

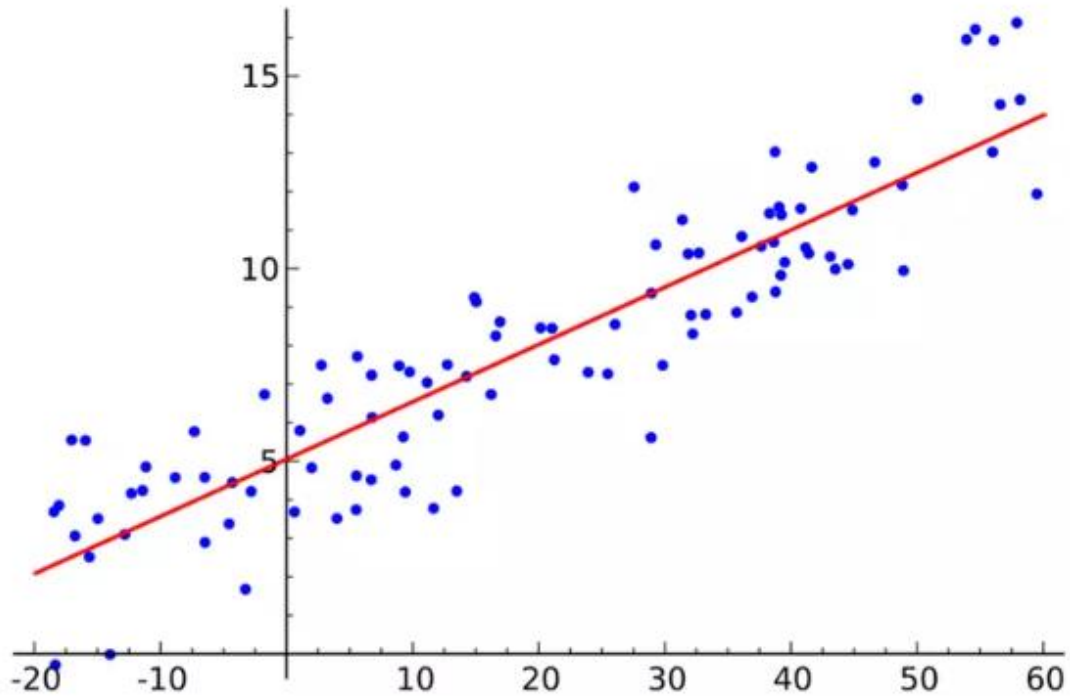
Trong hồi quy tuyến tính đơn giản, chúng ta xem xét mối quan hệ tuyến tính giữa một biến phụ thuộc Y và một biến độc lập X .

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).
- β_0 và β_1 là các hệ số hồi quy (cũng được gọi là hệ số góc và hệ số chặn).
- ε là sai số ngẫu nhiên (cũng được gọi là sai số hồi quy).



Hình 2. 7: Ví dụ về hồi quy tuyến tính

2.4.3. Ứng dụng

Linear Regression được ứng dụng rộng rãi trong nhiều lĩnh vực, như:

- *Dự báo giá cả*: dự đoán giá nhà, giá cổ phiếu, giá nhiên liệu dựa trên các yếu tố như vị trí, kích thước, chất lượng, lượng cung cầu, ...
- *Dự báo điểm số*: dự đoán điểm số của học sinh dựa trên thời gian học, nỗ lực, kỹ năng, trình độ giáo viên, ...
- *Dự báo sản phẩm*: dự đoán đầu ra sản xuất dựa trên thời gian, công suất, nguyên liệu, lao động, ...
- *Phân tích chuỗi thời gian*: dự đoán xu hướng và chu kỳ của các chuỗi dữ liệu, như bất động sản, thời tiết, xu hướng sản xuất, ...

2.4.4. Ưu điểm và nhược điểm

Ưu điểm:

- *Đơn giản và dễ hiểu*: Hồi quy tuyến tính là một phương pháp đơn giản và dễ hiểu để mô hình hoá 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. Support Vector Machines – SVM

2.5.1. Giới thiệu

Support Vector Machines (SVM) là một trong những phương pháp mạnh mẽ và phổ biến trong lĩnh vực học máy, đặc biệt trong các bài toán phân loại và hồi quy. SVM được phát triển dựa trên lý thuyết học máy và các khái niệm toán học về không

gian vector. Ý tưởng cơ bản của SVM là tìm ra một siêu phẳng tối ưu (hyperplane) phân chia các lớp dữ liệu trong không gian đa chiều sao cho khoảng cách giữa các điểm dữ liệu gần nhất và siêu phẳng là lớn nhất. Đây là một mô hình có khả năng học tốt từ dữ liệu có kích thước lớn và phức tạp.

Trong bài toán phân loại, SVM có thể giúp xác định được các lớp của dữ liệu dựa trên các đặc trưng của nó. Trong bài toán dự báo giá trị (hồi quy), SVM cũng có thể được áp dụng để dự đoán giá trị liên tục từ dữ liệu đầu vào.

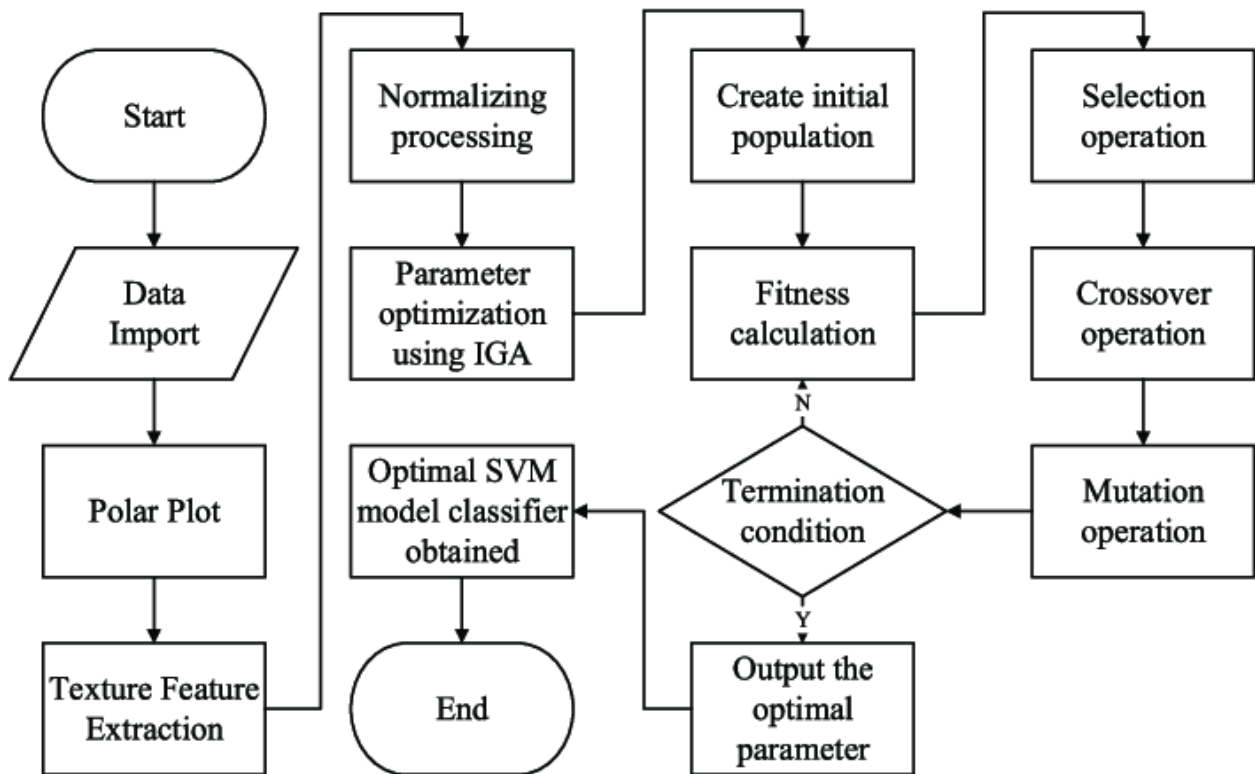
2.5.2. Cấu trúc

SVM hoạt động trong không gian đa chiều, nơi mỗi điểm dữ liệu là một vector trong không gian đó. Mục tiêu của SVM là tìm ra siêu phẳng phân chia tối ưu giữa các lớp dữ liệu sao cho khoảng cách giữa các điểm gần nhất thuộc hai lớp (gọi là margin) là lớn nhất. Việc này giúp đảm bảo rằng mô hình sẽ phân loại chính xác dữ liệu mới.

Cấu trúc cơ bản của SVM bao gồm các thành phần sau:

- *Hỗ trợ vector (Support Vectors)*: Đây là các điểm dữ liệu quan trọng nhất, chúng nằm gần siêu phẳng và đóng vai trò quan trọng trong việc xác định siêu phẳng tối ưu.
- *Siêu phẳng (Hyperplane)*: Đây là mặt phân chia dữ liệu trong không gian đa chiều.
- *Margin*: Khoảng cách giữa siêu phẳng và các điểm dữ liệu gần nhất của mỗi lớp.

2.5.3. Quy trình hoạt động



Hình 2. 8: Quy trình hoạt động của mô hình SVM

Quy trình hoạt động của mô hình SVM (Support Vector Machine) bắt đầu với việc chuẩn bị dữ liệu. Trước khi đưa dữ liệu vào mô hình, quá trình làm sạch và xử lý dữ liệu là rất quan trọng. Dữ liệu huấn luyện cần được làm sạch, loại bỏ các giá trị thiếu, xử lý các ngoại lệ và mã hóa các đặc trưng phân loại nếu cần thiết. Đồng thời, việc chuẩn hóa dữ liệu là cần thiết vì SVM phụ thuộc vào khoảng cách giữa các điểm dữ liệu. Các phương pháp chuẩn hóa thường được sử dụng là chuẩn hóa về phạm vi $[0,1]$ hoặc $[-1,1]$.

Sau khi chuẩn bị dữ liệu, SVM sẽ tìm siêu phẳng tối ưu để phân chia các lớp dữ liệu. Mục tiêu là tối đa hóa **margin**, tức là khoảng cách giữa siêu phẳng và các điểm dữ liệu gần nhất của mỗi lớp, được gọi là **Support Vectors**. Siêu phẳng tối ưu này sẽ tạo ra một ranh giới phân chia rõ ràng và hiệu quả giữa các lớp. Việc tìm kiếm siêu phẳng tối ưu giúp mô hình phân loại các điểm dữ liệu một cách chính xác và hiệu quả hơn.

Khi đã xác định được siêu phẳng, SVM tiếp tục quá trình tối ưu hóa. Việc này liên quan đến việc điều chỉnh các trọng số (weights) và hệ số bù (bias) để tối đa hóa margin. Mô hình sẽ sử dụng các phương pháp tối ưu hóa như **Gradient Descent** hoặc **Quadratic Programming** để tìm ra các giá trị của trọng số và bias sao cho siêu phẳng phân chia các lớp dữ liệu với khoảng cách lớn nhất. Các điểm gần siêu phẳng, gọi là Support Vectors, sẽ được sử dụng để thực hiện tối ưu hóa, giúp mô hình trở nên chính xác và hiệu quả hơn.

Đối với các bài toán phân loại không tuyến tính, SVM sử dụng **Kernel Trick** để chuyển dữ liệu vào không gian đặc trưng cao hơn, nơi có thể phân chia các lớp dữ liệu một cách tuyến tính. Việc áp dụng Kernel Trick giúp mô hình SVM giải quyết được các bài toán phân loại phức tạp mà các phương pháp tuyến tính không thể xử lý được. Các loại kernel phổ biến bao gồm **Linear Kernel**, **Polynomial Kernel**, và **Radial Basis Function (RBF) Kernel**, mỗi loại sẽ thích hợp với các loại bài toán khác nhau.

Khi mô hình đã được huấn luyện, nó có thể phân loại các điểm dữ liệu mới dựa trên siêu phẳng đã học. SVM tính toán giá trị của hàm phân loại và sử dụng giá trị này để xác định lớp của các điểm dữ liệu. Nếu giá trị hàm phân loại lớn hơn 0, điểm dữ liệu sẽ được phân loại vào lớp dương, ngược lại, nếu nhỏ hơn 0, nó sẽ được phân loại vào lớp âm. Quá trình này cho phép mô hình phân loại dữ liệu mới một cách chính xác và hiệu quả.

Cuối cùng, sau khi huấn luyện và phân loại, mô hình cần được đánh giá để xác định hiệu suất của nó. Các chỉ số đánh giá phổ biến trong phân loại bao gồm **accuracy** (độ chính xác), **precision** (độ chính xác), **recall** (độ nhạy) và điểm F1. Trong các bài toán hồi quy, có thể sử dụng các chỉ số như **Mean Squared Error (MSE)** hoặc **Mean Absolute Error (MAE)** để đánh giá chất lượng mô hình. Quá trình đánh giá này giúp xác định mức độ hiệu quả và phù hợp của mô hình SVM trong các bài toán thực tế.

Sau khi hoàn tất huấn luyện và đánh giá, mô hình có thể được triển khai vào các ứng dụng thực tế. Mô hình SVM sẽ phân loại hoặc dự đoán giá trị cho các điểm dữ liệu mới, từ đó phục vụ cho các mục đích như nhận dạng mẫu, phân tích dữ liệu và dự báo. Nhờ vào khả năng xử lý các bài toán phân loại có ranh giới phân chia rõ

ràng và khả năng sử dụng Kernel Trick để giải quyết các vấn đề phân loại không tuyến tính, SVM là một công cụ mạnh mẽ trong học máy.

2.5.4. Phân loại

SVM có thể được sử dụng trong hai bài toán chính:

- *Phân loại nhị phân (Binary Classification)*: Trong trường hợp này, SVM phân chia dữ liệu thành hai lớp.
- *Phân loại đa lớp (Multi-class Classification)*: Mặc dù SVM chủ yếu là một thuật toán phân loại nhị phân, nhưng với một số kỹ thuật như "One-vs-One" hoặc "One-vs-Rest", SVM cũng có thể được áp dụng cho bài toán phân loại đa lớp.

2.5.5. Ưu điểm và nhược điểm của SVM

Ưu điểm:

- *Khả năng phân loại chính xác*: SVM thường cho kết quả phân loại chính xác, đặc biệt trong các bài toán phân loại phức tạp.
- *Khả năng xử lý dữ liệu phi tuyến tính*: Nhờ vào việc sử dụng các hàm kernel, SVM có thể giải quyết tốt các bài toán phân loại phi tuyến tính.
- *Không dễ bị overfitting*: Đặc biệt khi sử dụng một margin rộng, SVM có thể hạn chế hiện tượng quá khớp (overfitting) và mang lại hiệu quả ổn định.

Nhược điểm:

- *Yêu cầu bộ nhớ lớn*: SVM yêu cầu lưu trữ tất cả các điểm dữ liệu trong bộ nhớ và có thể trở nên chậm khi làm việc với các tập dữ liệu lớn.
- *Khó khăn trong việc chọn kernel*: Việc lựa chọn kernel và các tham số điều chỉnh (như C, gamma) có thể là một quá trình phức tạp và tốn thời gian.
- *Độ phức tạp tính toán cao*: Đặc biệt đối với dữ liệu lớn hoặc không gian chiều cao, quá trình tối ưu hóa và tìm kiếm siêu phẳng tối ưu có thể yêu cầu tính toán phức tạp.

2.6. Cây quyết định (Decision Tree)

2.6.1. Giới thiệu

Cây quyết định là một thuật toán học máy giám sát (supervised learning) được sử dụng để giải quyết các bài toán phân loại và hồi quy. Cây quyết định hoạt động

bằng cách phân chia dữ liệu thành các nhánh dựa trên các câu hỏi hoặc điều kiện, giúp xác định giá trị đầu ra cuối cùng. Cây quyết định giống như một cây phân loại, với mỗi nút đại diện cho một thuộc tính, các nhánh biểu thị các lựa chọn của thuộc tính đó và các lá cuối cùng là các quyết định hoặc giá trị được dự đoán.

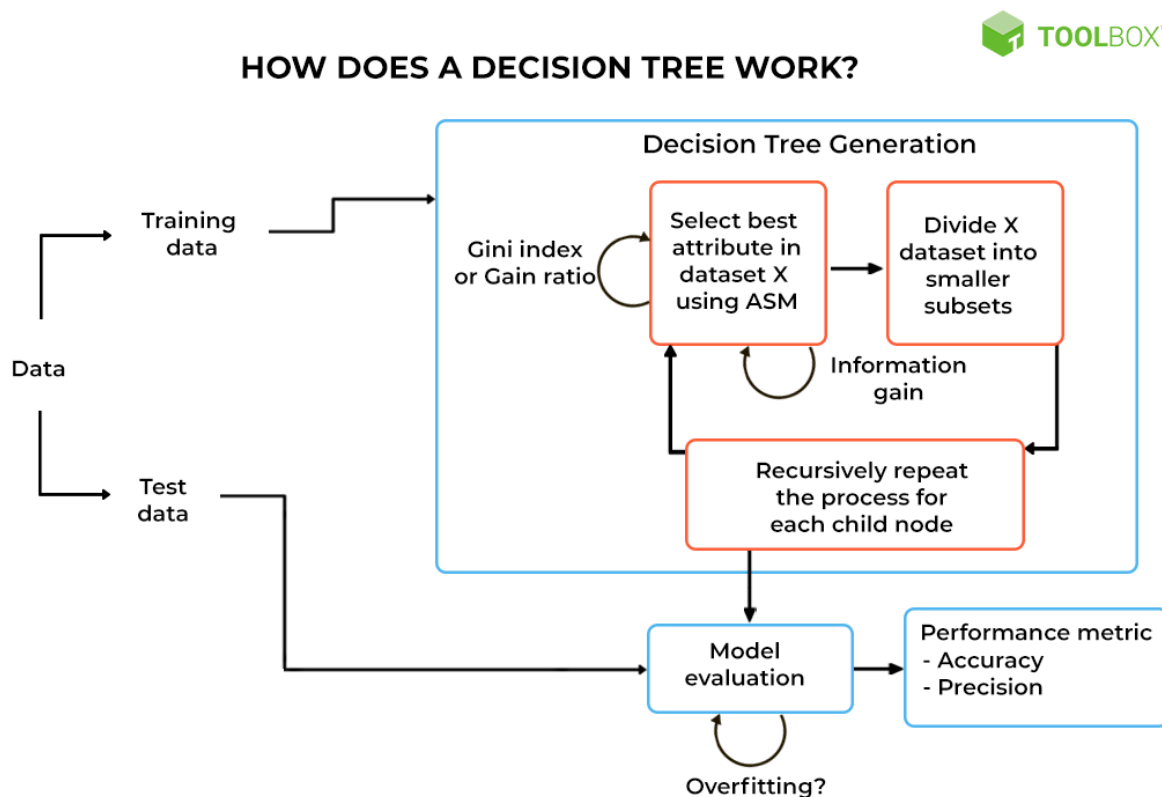
Cây quyết định đặc biệt có lợi khi làm việc với dữ liệu có các thuộc tính phân loại và có thể dễ dàng giải thích kết quả của mô hình. Chúng là một công cụ mạnh mẽ trong các hệ thống ra quyết định và phân tích dữ liệu.

2.6.2. Cấu trúc của cây quyết định

Cấu trúc của cây quyết định bao gồm các phần sau:

- *Nút gốc (Root Node)*: Đây là nút đầu tiên của cây, nơi quá trình phân chia dữ liệu bắt đầu.
- *Nút phân chia (Decision Nodes)*: Các nút này đại diện cho các quyết định cần đưa ra tại mỗi bước của quá trình phân chia dữ liệu. Mỗi nút sẽ kiểm tra một thuộc tính và phân chia dữ liệu dựa trên các giá trị của thuộc tính đó.
- *Nút lá (Leaf Nodes)*: Đây là các nút cuối cùng trong cây, nơi mà quá trình phân chia dữ liệu kết thúc. Mỗi nút lá đại diện cho một nhãn lớp (trong bài toán phân loại) hoặc một giá trị đầu ra (trong bài toán hồi quy).
- *Các nhánh (Branches)*: Các nhánh nối giữa các nút, đại diện cho các lựa chọn hoặc điều kiện phân chia dữ liệu.

2.6.3. Quy trình hoạt động



Hình 2. 9: Quy trình hoạt động của Cây quyết định

Quy trình hoạt động của mô hình **Decision Tree** (Cây Quyết Định) diễn ra qua một số bước cơ bản từ việc chuẩn bị dữ liệu đến việc huấn luyện và dự đoán.

Bước đầu tiên trong quy trình là chuẩn bị dữ liệu. Dữ liệu đầu vào cần được làm sạch và tiền xử lý để loại bỏ các giá trị thiếu, xử lý ngoại lệ và mã hóa các đặc trưng phân loại nếu cần. Việc chuẩn hóa dữ liệu không phải lúc nào cũng cần thiết đối với Decision Tree, vì cây quyết định không phụ thuộc vào khoảng cách giữa các điểm dữ liệu. Tuy nhiên, dữ liệu cần phải có đầy đủ các thông tin để mô hình có thể học hỏi và phân loại chính xác.

Khi dữ liệu đã được chuẩn bị, mô hình Decision Tree sẽ bắt đầu quá trình huấn luyện. Trong quá trình huấn luyện, Decision Tree phân chia dữ liệu thành các nhánh dựa trên các đặc trưng sao cho mỗi nhánh càng đồng nhất với một lớp mục tiêu càng tốt. Việc phân chia này được thực hiện dựa trên các tiêu chí như **Entropy** và **Gini Index**, nhằm giảm thiểu sự không đồng nhất trong mỗi nhánh. Cây quyết định tiếp

tục phân chia cho đến khi đạt được một mức độ đồng nhất đủ lớn hoặc một điều kiện dừng đã được xác định trước, như số lượng nhánh tối đa hoặc chiều sâu của cây.

Trong mỗi bước phân chia, Decision Tree sẽ tìm ra đặc trưng tối ưu để chia dữ liệu. Đặc trưng này được chọn dựa trên chỉ số **Information Gain** (Tăng Thông Tin) hoặc **Gini Index** (Chỉ Số Gini), hai tiêu chí phổ biến giúp mô hình đánh giá sự phù hợp của đặc trưng để phân loại dữ liệu. Cây quyết định sẽ tiếp tục phân chia dữ liệu tại mỗi điểm chia cho đến khi không còn đặc trưng nào có thể phân chia tốt hơn hoặc cây đạt độ sâu tối đa.

Sau khi mô hình đã được huấn luyện và cây quyết định hoàn tất, nó có thể được sử dụng để dự đoán các giá trị mới. Mỗi điểm dữ liệu mới sẽ được đưa qua cây quyết định, nơi các đặc trưng của nó sẽ được so sánh với các điều kiện trong các nhánh cây. Quy trình này sẽ tiếp tục cho đến khi cây đi đến một lá (leaf node), nơi mà lớp hoặc giá trị dự đoán sẽ được đưa ra.

Cuối cùng, mô hình Decision Tree cần được đánh giá để kiểm tra hiệu suất của nó. Các chỉ số như **accuracy** (độ chính xác), **precision** (độ chính xác), **recall** (độ nhạy) hoặc **F1-score** có thể được sử dụng để đánh giá chất lượng mô hình trong các bài toán phân loại. Đối với các bài toán hồi quy, các chỉ số như **Mean Squared Error (MSE)** hoặc **Mean Absolute Error (MAE)** có thể được áp dụng.

Tóm lại, mô hình Decision Tree hoạt động bằng cách xây dựng một cây phân chia dữ liệu qua các nhánh, nơi mỗi nhánh đại diện cho một quyết định dựa trên các đặc trưng của dữ liệu, nhằm phân loại hoặc dự đoán chính xác các điểm dữ liệu mới.

2.6.4. Phân loại

Cây quyết định có thể được sử dụng cho hai loại bài toán chính:

- *Phân loại (Classification)*: Dùng để phân chia dữ liệu thành các lớp hoặc nhóm. Ví dụ, phân loại email thành thư rác và không phải thư rác.
- *Hồi quy (Regression)*: Dùng để dự đoán một giá trị liên tục. Ví dụ, dự đoán giá trị bất động sản dựa trên các đặc điểm như diện tích, vị trí và năm xây dựng.

2.6.5. Ưu điểm và nhược điểm của cây quyết định

Ưu điểm:

- *Dễ hiểu và giải thích*: Cây quyết định rất dễ hình dung và có thể giải thích được, giúp người dùng dễ dàng hiểu cách mô hình ra quyết định.
- *Không yêu cầu chuẩn hóa dữ liệu*: Không giống như nhiều mô hình học máy khác, cây quyết định không yêu cầu dữ liệu phải được chuẩn hóa hoặc làm sạch quá nhiều.
- *Có thể xử lý dữ liệu dạng phân loại và liên tục*: Cây quyết định có thể xử lý cả dữ liệu dạng phân loại và dữ liệu liên tục một cách hiệu quả.
- *Khả năng xử lý dữ liệu thiếu*: Một số thuật toán cây quyết định có thể xử lý được dữ liệu thiếu (missing values) mà không cần phải loại bỏ chúng.

Nhược điểm:

- *Dễ bị overfitting*: Nếu cây quá sâu, mô hình có thể học quá mức vào dữ liệu huấn luyện, dẫn đến overfitting và kém hiệu quả với dữ liệu mới.
- *Khó khăn trong việc tổng quát hóa*: Cây quyết định không phải lúc nào cũng tổng quát tốt khi đối mặt với các mối quan hệ phức tạp giữa các đặc trưng của dữ liệu.
- *Không ổn định*: Một sự thay đổi nhỏ trong dữ liệu có thể dẫn đến một cây quyết định hoàn toàn khác nhau, khiến mô hình trở nên không ổn định.
- *Phức tạp trong việc tối ưu hóa*: Khi có nhiều thuộc tính hoặc lớp dữ liệu, cây quyết định có thể trở nên rất phức tạp, gây khó khăn trong việc tối ưu hóa mô hình.

2.7. Kết luận

Trong Chương 2, nhóm đã tiến hành nghiên cứu toàn diện các kỹ thuật hiện có nhằm giải quyết bài toán, đồng thời phân tích chi tiết ưu nhược điểm và tính phù hợp của từng phương pháp. Qua quá trình đánh giá kỹ lưỡng, nhóm đã lựa chọn hồi quy tuyến tính (Linear Regression) làm giải pháp chính nhờ sự tương thích cao với tập dữ liệu có mối quan hệ tuyến tính giữa các biến độc lập và biến phụ thuộc. Đây là phương pháp không chỉ dễ triển khai mà còn mang lại khả năng diễn giải rõ ràng, giúp nhóm hiểu được mức độ ảnh hưởng của từng yếu tố đầu vào đối với kết quả đầu ra. Điều này cung cấp cơ sở quan trọng cho việc phân tích dữ liệu và đưa ra các quyết định chiến lược.

Một trong những điểm mạnh nổi bật của hồi quy tuyến tính là tính đơn giản trong triển khai và chi phí tính toán thấp, rất phù hợp với các bài toán yêu cầu xử lý nhanh hoặc áp dụng trên các tập dữ liệu lớn. Hơn thế nữa, phương pháp này dễ dàng mở rộng và điều chỉnh, tạo điều kiện thuận lợi cho việc thử nghiệm và tối ưu hóa mô hình. Tuy nhiên, nhóm cũng nhận thức rõ các hạn chế của kỹ thuật này, đặc biệt khi đối mặt với dữ liệu không tuyến tính hoặc bị nhiễu nhiều. Để khắc phục, nhóm đã thực hiện các bước tiền xử lý dữ liệu cẩn thận, bao gồm loại bỏ các điểm ngoại lai (outliers) và đánh giá mức độ tương quan giữa các biến, đảm bảo dữ liệu đạt chất lượng cao trước khi áp dụng mô hình.

Nhìn chung, quyết định lựa chọn hồi quy tuyến tính không chỉ dựa trên sự phù hợp về mặt lý thuyết mà còn được củng cố bởi tính thực tiễn và hiệu quả triển khai. Đây là giải pháp hứa hẹn mang lại kết quả khả quan trong việc giải quyết bài toán, đồng thời tạo nền tảng vững chắc để nhóm mở rộng ứng dụng trong tương lai.

CHƯƠNG 3 THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1. Dữ liệu thực nghiệm

Trong project này, bộ dữ liệu được phân tích ở đây là file dataset (.csv) có tên là USA_Housing.csv chứa 5000 dòng thông tin về giá nhà tại nước Mỹ.

Cụ thể thông tin như sau:

- Tên bộ dữ liệu: USA_Housing
- Nguồn: [USA Housing](#)

Dữ liệu 15 dòng đầu của dataset:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
1	79545.45857431678	5.682861321615587	7.009188142792237	4.09	23086.800502686456	1059033.5578701235	Aurabury, NE 37010-5101
2	79248.64245482568	6.0028998082752425	6.730821019094919	3.09	40173.07217364482	1505890.91484695	Lake Kathleen, CA 48958
3	61287.067178656784	5.865889840310001	8.512727430375099	5.13	36882.15939970458	1058987.9878760849	Wauwatosa, WI 06482-3489
4	63345.24004622798	7.1882360945186425	5.586728664827653	3.26	34310.24283090706	1260616.8066294468	St Barnett FPO AP 44820
5	59982.197225708034	5.040554523106283	7.839387785120487	4.23	26354.109472103148	630943.4893385402	Raymond FPO AE 09386
6	80175.7541594853	4.9884077575337145	6.104512439428879	4.04	26748.428424689715	1068138.0743935304	443 Tracyport, KS 16077
7	64698.46342788773	6.025335906887153	8.147759585023431	3.41	60828.24908540716	1502055.8173744078	Nguyenburgh, CO 20247
8	78394.33927753085	6.9897797477182815	6.620477995185026	2.42	36516.358972493836	1573936.5644777215	William, TN 17778-6483
9	59927.66081334963	5.36212556960358	6.3931209805509015	2.3	29387.39600281585	798869.5328331633	SS Gilbert FPO AA 20957
10	81885.92718409566	4.423671789897876	8.167688003472351	6.1	40149.96574921337	1545154.8126419624	Box 0958 DPO AE 97025
11	80527.47208292288	8.09351268063935	5.042746799645982	4.1	47224.35984022191	1707045.722158058	700 Janetbury, NM 26854
12	50593.69549704281	4.496512793097035	7.467627404008019	4.49	34343.991885578806	663732.3968963273	Davisborough, PW 78603
13	39033.809236982364	7.671755372854428	7.250029317273495	3.1	39220.36146737246	1042814.0978200927	1 Huffmanland, NE 52457
14	73163.6634410467	6.919534825456555	5.9931879009455695	2.27	32326.123139488096	1291331.5184858206	North John, AR 26532-5136
15	69391.3801843616	5.344776176735725	8.406417714534253	4.37	35521.294033173246	1402818.2101658515	Box 4420 APO AP 08302

Hình 3. 1: 15 dòng đầu của bộ dữ liệu gốc

Thông tin cụ thể các cột của dataset như sau:

- **Avg. Area Income:** Thu nhập trung bình của cư dân trong thành phố nơi ngôi nhà tọa lạc.
- **Avg. Area House Age:** Tuổi trung bình của các ngôi nhà trong cùng thành phố.
- **Avg. Area Number of Rooms:** Số phòng trung bình của các ngôi nhà trong cùng thành phố.
- **Avg. Area Number of Bedrooms:** Số phòng ngủ trung bình của các ngôi nhà trong cùng thành phố.
- **Area Population:** Dân số của thành phố nơi ngôi nhà tọa lạc.
- **Price:** Giá bán của ngôi nhà.
- **Address:** Địa chỉ của ngôi nhà.

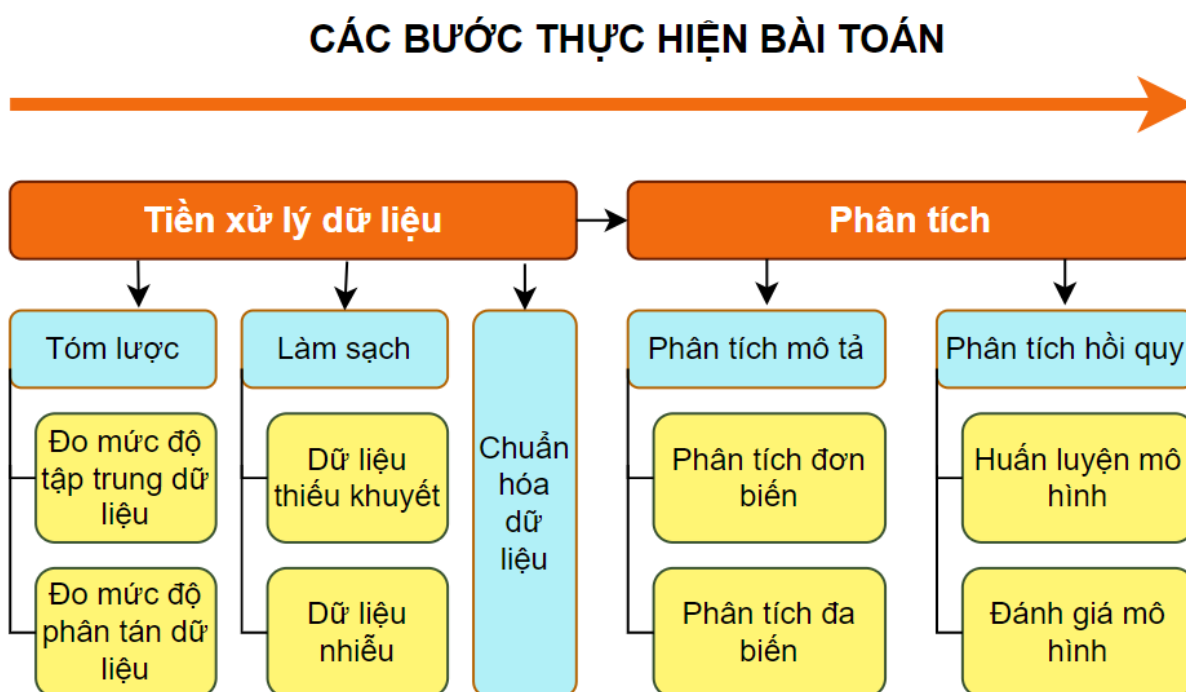
3.2 Môi trường thực nghiệm

Cấu hình máy: Chip Intel(R) Core i5, Ram 8GB, 64-bit.

Công cụ hỗ trợ: thuật toán được chạy trên Python 3.11, cmd trên Window 10.

Tập dữ liệu sử dụng để thực nghiệm: gồm 5000 thông tin về dữ liệu của các ngôi nhà tại nước Mỹ. Các phương pháp tham gia thực nghiệm: phương pháp hồi quy tuyến tính.

3.3. Quy trình thực nghiệm



Hình 3. 2: 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 cũng như chất lượng dữ liệu.

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

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, 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 "Avg. Area Income, Avg. Area House Age, Avg. Area Number of Rooms, Avg. Area Number of Bedrooms, Area Population, Price" 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:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+06
std	10657.991214	0.991456	1.005833	1.234137	9925.650114	3.531176e+05
min	17796.631190	2.644304	3.236194	2.000000	172.610686	1.593866e+04
25%	61480.562388	5.322283	6.299250	3.140000	29403.928702	9.975771e+05
50%	68804.286404	5.970429	7.002902	4.050000	36199.406689	1.232669e+06
75%	75783.338666	6.650808	7.665871	4.490000	42861.290769	1.471210e+06
max	107701.748378	9.519088	10.759588	6.500000	69621.713378	2.469066e+06

Hình 3. 3: Thông tin tóm lược dữ liệu của cột dữ liệu dạng số

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.
- *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.

Ta sẽ kiểm tra dữ liệu bị thiếu

```
Avg. Area Income      0
Avg. Area House Age    0
Avg. Area Number of Rooms  0
Avg. Area Number of Bedrooms 0
Area Population        0
Price                  0
Address                0
dtype: int64
```

Hình 3. 4: Kiểm tra dữ liệu bị khuyết

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 duy nhất cột “Address” đang ở dạng Object, ngoài ra vì nó chứa đến 5000 giá trị khác nhau, bằng đúng số lượng điểm dữ liệu trong bộ. Điều này cho thấy mỗi giá trị trong cột là duy nhất, không mang tính chất phân loại hay khái quát. Việc giữ lại cột này không mang lại ý nghĩa phân tích hay dự đoán, mà còn làm tăng độ phức tạp của mô hình. Vì vậy, loại bỏ cột là lựa chọn đúng đắn.

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 7 columns):
 #   Column                                  Non-Null Count  Dtype
---  -
 0   Avg. Area Income                       5000 non-null   float64
 1   Avg. Area House Age                    5000 non-null   float64
 2   Avg. Area Number of Rooms              5000 non-null   float64
 3   Avg. Area Number of Bedrooms           5000 non-null   float64
 4   Area Population                        5000 non-null   float64
 5   Price                                  5000 non-null   float64
 6   Address                                5000 non-null   object
dtypes: float64(6), object(1)
memory usage: 273.6+ KB
```

Hình 3. 5: Kiểm tra kiểu dữ liệu cho từng cột

```
print(len(df['Address'].unique()))

5000
```

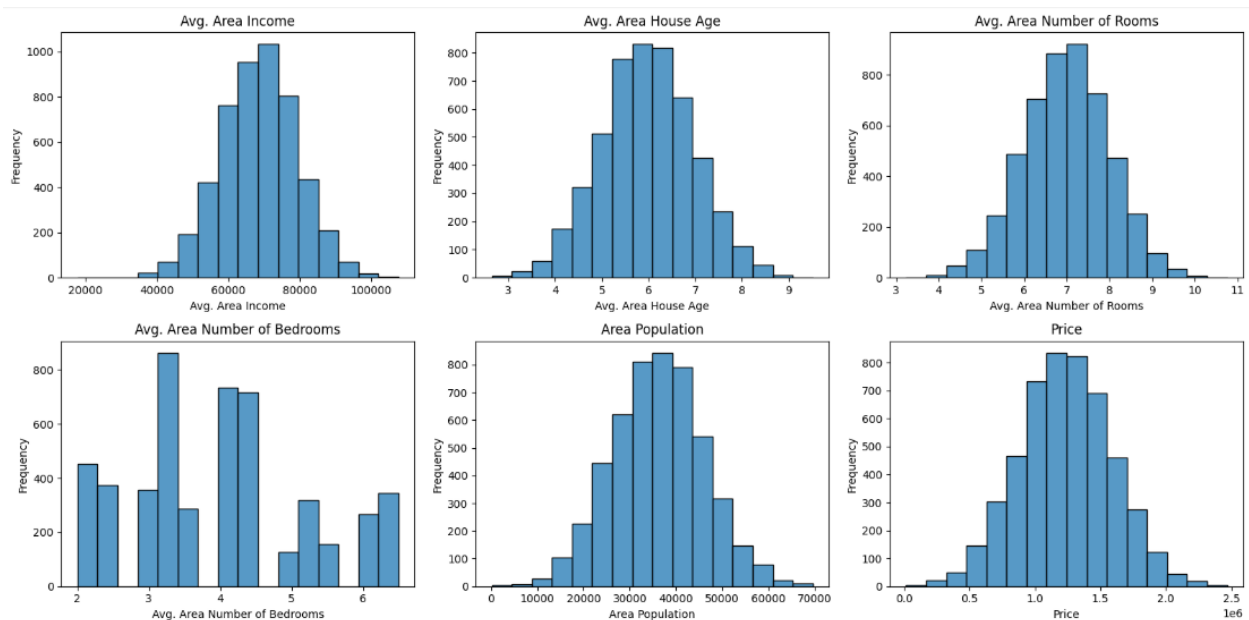
Hình 3. 6: Đếm số lượng các giá trị duy nhất tại cột “Address”

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, xu hướng 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.

3.3.3.1. Phân tích đơn biến

Biểu đồ 1: Biểu đồ phân phối của các thuộc tính (Histogram)



Hình 3. 7: Biểu đồ cột thể hiện

Nhận xét về đặc điểm của tập dữ liệu:

1. Avg. Area Income:

- Có dạng phân phối chuẩn với trung tâm nằm khoảng 60,000 - 80,000.
- Phù hợp với mô hình hồi quy vì phân phối đồng đều.

2. Avg. Area House Age:

- Có dạng phân phối chuẩn, tập trung ở khoảng 5 - 7 năm.
- Điều này cho thấy dữ liệu về tuổi nhà ổn định và có thể hỗ trợ tốt trong mô hình.

3. Avg. Area Number of Rooms:

- Có dạng phân phối chuẩn, tập trung quanh 6 - 8 phòng.
- Đây là dấu hiệu tốt, phù hợp với giả định tuyến tính trong hồi quy.

4. Avg. Area Number of Bedrooms:

- Phân phối rời rạc, ít đối xứng, với nhiều đỉnh ở 2, 3 và 4 phòng ngủ.
- Số lượng phòng ngủ có sự đa dạng, không hoàn toàn tuân theo phân phối chuẩn.

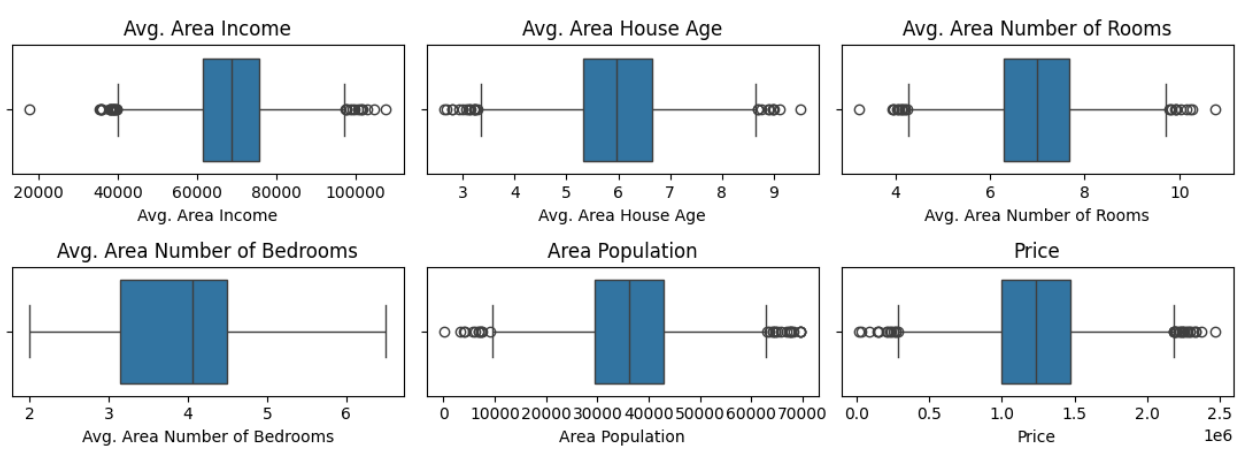
5. Area Population:

- Có dạng phân phối gần chuẩn, tập trung ở khoảng 20,000 - 50,000 người.
- Có thể dùng tốt trong hồi quy.

6. Price:

- Có dạng phân phối chuẩn, tập trung ở khoảng 1 triệu - 1.5 triệu đô la.
- Đây là biến mục tiêu (target), rất phù hợp để áp dụng mô hình hồi quy tuyến tính

Biểu đồ 2: Biểu đồ boxplot and whisker plot của các thuộc tính



Hình 3. 8: Biểu đồ box chart

Dựa trên biểu đồ ta thấy được đặc điểm của tập dữ liệu

1. Thu nhập trung bình khu vực (Avg. Area Income)

- Trung vị: \$60,000
- IQR: \$60,000 đến \$75,000
- Có nhiều giá trị ngoại lai (outliers) dưới \$40,000 và trên \$90,000
- **Nhận xét:** Thu nhập biến động lớn, cho thấy sự đa dạng về mức sống trong khu vực.

2. Tuổi nhà trung bình khu vực (Avg. Area House Age)

- Trung vị: 6 năm

- IQR: 5 đến 7 năm
- Một số giá trị ngoại lai dưới 4 năm và trên 8 năm
- **Nhận xét:** Tuổi nhà tương đối đồng đều, nhưng có một số nhà mới hoặc cũ đáng kể.

3. Số lượng phòng trung bình khu vực (Avg. Area Number of Rooms)

- Trung vị: 7 phòng
- IQR: 6 đến 8 phòng
- Một số giá trị ngoại lai dưới 5 phòng và trên 9 phòng
- **Nhận xét:** Số lượng phòng khá đồng đều, nhưng có sự đa dạng trong thiết kế nhà ở.

4. Số lượng phòng ngủ trung bình khu vực (Avg. Area Number of Bedrooms)

- Trung vị: 4 phòng ngủ, lệch về phía q3
- IQR: 3 đến 4 phòng ngủ
- Gần như không có giá trị ngoại lai
- **Nhận xét:** Số lượng phòng ngủ ít biến động và ổn định hơn so với các thuộc tính khác.

5. Dân số khu vực (Area Population)

- Trung vị: 35,000 người
- IQR: 30,000 đến 40,000 người
- Nhiều giá trị ngoại lai dưới 20,000 và trên 60,000
- **Nhận xét:** Dân số có sự biến động lớn, phản ánh sự phân bố không đồng đều.

6. Giá nhà (Price)

- Trung vị: \$1.2 triệu
- IQR: \$1 triệu đến \$1.5 triệu
- Nhiều giá trị ngoại lai dưới \$800,000 và trên \$2 triệu

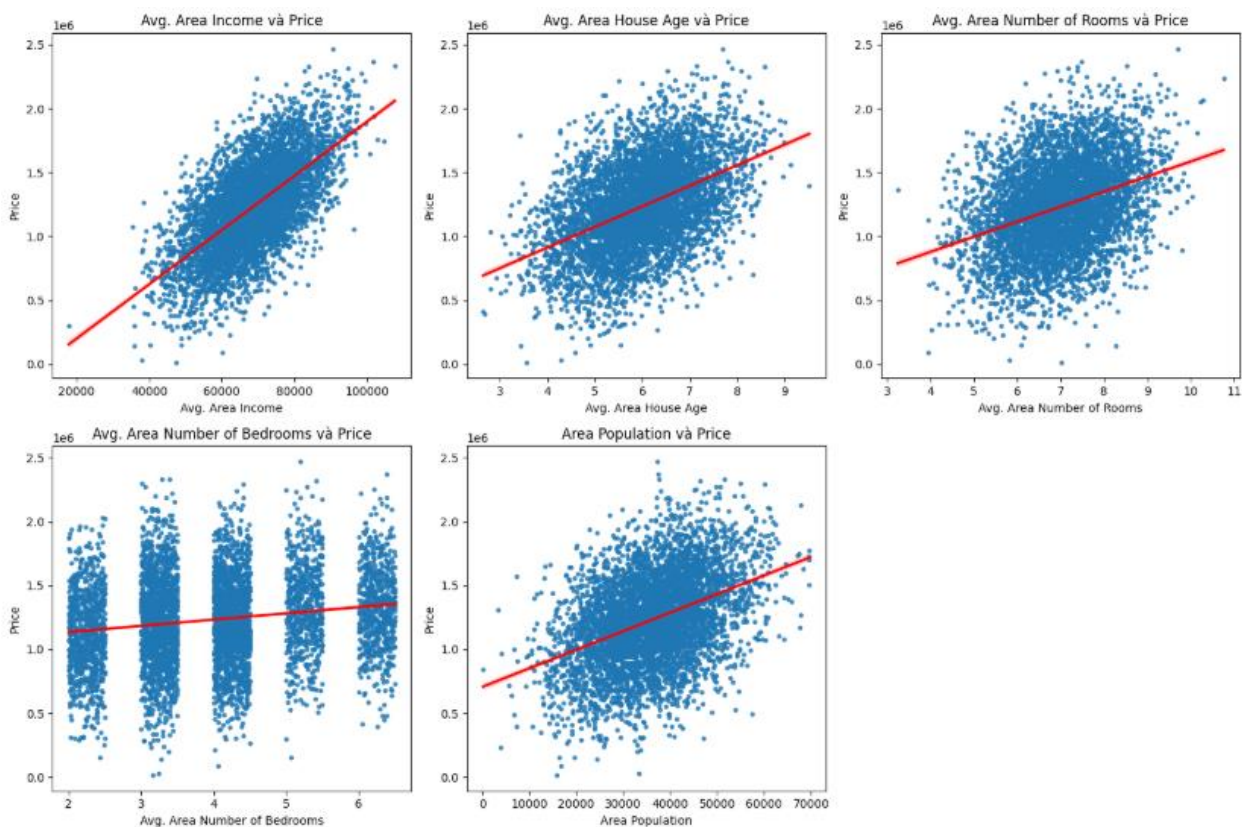
- **Nhận xét:** Giá nhà biến động lớn, cho thấy sự khác biệt đáng kể về giá nhà.

Mức độ phù hợp của bộ dữ liệu cho mô hình hồi quy tuyến tính

- **Đặc điểm phân phối:** Hầu hết các thuộc tính có phân phối tương đối đối xứng và nằm chính giữa khoảng IQR, ngoại trừ Số lượng phòng ngủ.
- **Sự hiện diện của outliers:** Các giá trị ngoại lai (outliers) trong tất cả các thuộc tính có thể ảnh hưởng đến hiệu suất của mô hình hồi quy tuyến tính, vì hồi quy tuyến tính nhạy cảm với outliers.
- **Phân phối chuẩn:** Dữ liệu nhìn chung có sự phân bố tương đối đều và tuân theo phân phối chuẩn, phù hợp cho mô hình hồi quy tuyến tính sau khi xử lý outliers.

3.3.3.2. Phân tích đa biến

Biểu đồ 3: Biểu đồ scatter thể hiện mối quan hệ giữa các biến độc lập và biến phụ thuộc



Hình 3. 9: Biểu đồ scatter với đường hồi quy màu đỏ

Dựa trên các biểu đồ tán xạ (scatter plots) với đường hồi quy màu đỏ, ta có thể phân tích mối quan hệ giữa các biến độc lập và biến phụ thuộc (giá nhà - **Price**) như sau:

1. Avg. Area Income (Thu nhập trung bình khu vực) và Price:

- Biểu đồ cho thấy một **mối quan hệ tuyến tính dương mạnh mẽ**.
- Khi **thu nhập trung bình khu vực** tăng, giá nhà cũng có xu hướng tăng.
- Điều này cho thấy rằng thu nhập khu vực là một yếu tố quan trọng ảnh hưởng đến giá nhà.

2. Avg. Area House Age (Tuổi trung bình của nhà) và Price:

- Biểu đồ cho thấy mối quan hệ tuyến tính dương nhưng yếu hơn so với "Avg. Area Income".
- Nhà có tuổi càng cao thì giá nhà cũng có xu hướng tăng, tuy nhiên, mức độ tương quan này khá thấp.

3. Avg. Area Number of Rooms (Số lượng phòng trung bình) và Price:

- Có một **mối quan hệ tuyến tính dương nhẹ** giữa số lượng phòng và giá nhà.
- Số lượng phòng tăng thì giá nhà cũng có xu hướng tăng, tuy nhiên mức độ phân tán cao.

4. Avg. Area Number of Bedrooms (Số phòng ngủ trung bình) và Price:

- Mối quan hệ rất yếu và gần như không rõ ràng.
- Đường hồi quy có độ dốc nhỏ, cho thấy rằng số phòng ngủ không phải là yếu tố quan trọng ảnh hưởng đến giá nhà.

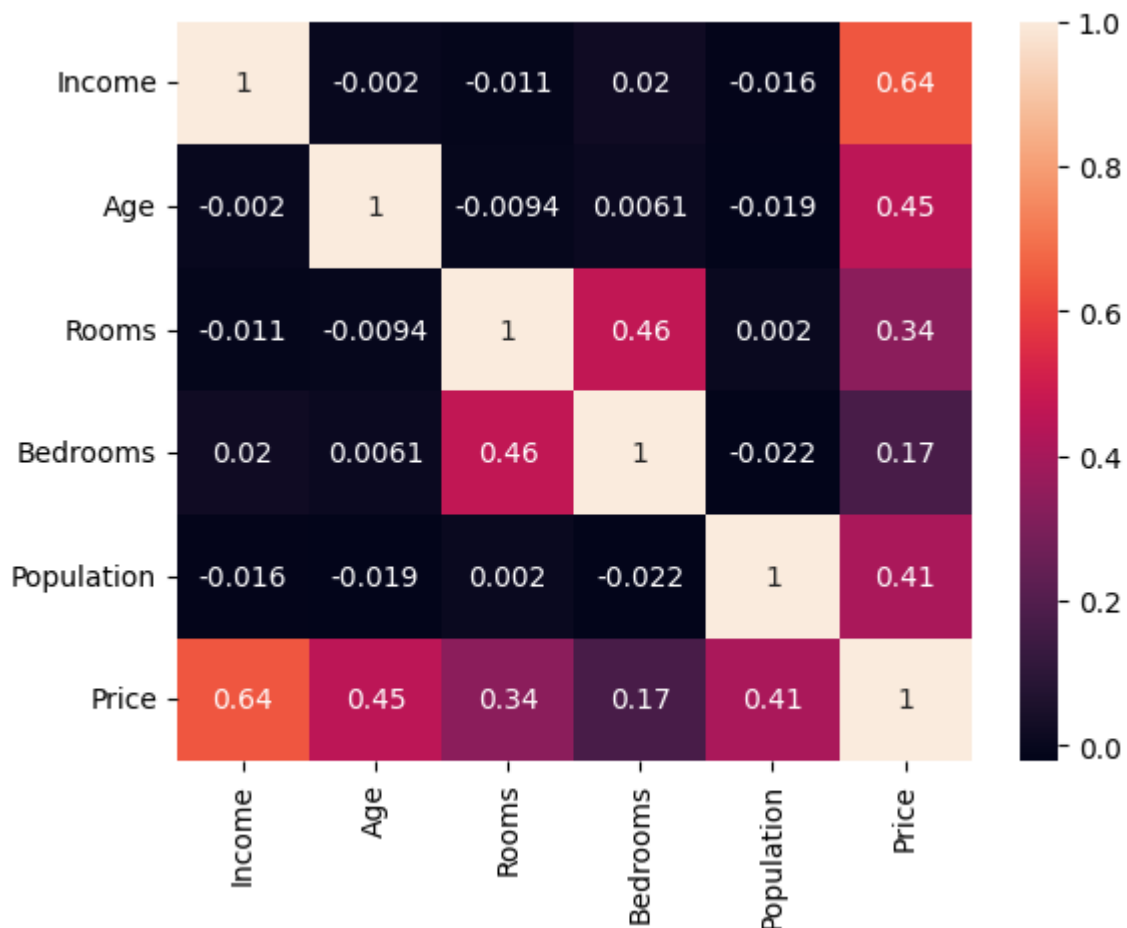
5. Area Population (Dân số khu vực) và Price:

- Có một mối quan hệ tuyến tính dương vừa phải.
- Khi dân số trong khu vực tăng, giá nhà có xu hướng tăng, tuy nhiên sự phân tán dữ liệu khá cao.

Từ đó, có thể kết luận:

- **Avg. Area Income** có tác động mạnh mẽ nhất đến **Price**, trong khi các biến khác như **Avg. Area House Age** và **Area Population** có mối quan hệ trung bình.
- **Avg. Area Number of Bedrooms** gần như không ảnh hưởng đáng kể đến giá nhà.
- Điều này cho thấy các yếu tố kinh tế như thu nhập khu vực và quy mô dân số ảnh hưởng mạnh hơn so với các yếu tố vật lý như số lượng phòng hay tuổi của nhà.

Biểu đồ 4: Biểu đồ thể hiện mối liên hệ giữa giá nhà và các yếu tố khác (heatmap)



Hình 3. 10: Biểu đồ heatmap

Biểu đồ **heatmap** trong bài toán dự đoán giá nhà (Price) được sử dụng để thể hiện **ma trận tương quan** giữa các biến độc lập (Income, Age, Rooms, Bedrooms, Population, Price).

Population) và biến mục tiêu (Price). Mức độ tương quan được biểu diễn qua các giá trị từ -1 đến 1, trong đó:

- **1**: Tương quan dương hoàn hảo (biến này tăng, biến kia tăng).
- **0**: Không có mối quan hệ tuyến tính giữa hai biến.
- **-1**: Tương quan âm hoàn hảo (biến này tăng, biến kia giảm).

Phân tích chi tiết các mối quan hệ

➤ Mối quan hệ với biến mục tiêu Price (Giá nhà)

Income (Thu nhập): 0.64

- Đây là biến có tương quan cao nhất với giá nhà.
- Điều này cho thấy thu nhập trung bình trong khu vực có ảnh hưởng mạnh đến giá nhà. Các khu vực có thu nhập cao thường có giá nhà cao hơn.
- **Ý nghĩa thực tế**: Giá nhà thường phản ánh mức sống và sức mua của dân cư.

Age (Tuổi nhà): 0.45

- Tuổi trung bình của các ngôi nhà cũng có mối tương quan dương với giá nhà.
- **Ý nghĩa thực tế**: Các ngôi nhà lâu đời trong những khu vực có giá trị lịch sử hoặc phát triển từ lâu thường có giá cao hơn.

Rooms (Số phòng): 0.34

- Số lượng phòng trong khu vực có mối tương quan dương trung bình với giá nhà.
- **Ý nghĩa thực tế**: Nhà có nhiều phòng thường lớn hơn, tiện nghi hơn và có giá cao hơn.

Population (Dân số): 0.41

- Mật độ dân số trong khu vực có mối tương quan dương trung bình với giá nhà.

- **Ý nghĩa thực tế:** Các khu vực đông dân cư có nhu cầu nhà ở cao hơn, dẫn đến giá nhà tăng.

Bedrooms (Số phòng ngủ): 0.17

- Đây là biến có tương quan thấp nhất với giá nhà.
- **Ý nghĩa thực tế:** Số phòng ngủ không ảnh hưởng mạnh đến giá nhà, có thể do nó thường đi kèm với số lượng phòng tổng thể (Rooms), khiến tác động riêng của Bedrooms bị giảm.

➤ **Mối quan hệ giữa các biến độc lập**

Rooms và Bedrooms: 0.46

- Có mối tương quan dương tương đối cao.
- **Ý nghĩa thực tế:** Số lượng phòng ngủ thường đi cùng với số lượng phòng tổng thể trong nhà.

Các biến còn lại (Income, Age, Population):

- Tương quan giữa các biến này rất thấp (gần 0), cho thấy chúng không có mối quan hệ tuyến tính rõ ràng với nhau.
- **Ý nghĩa thực tế:** Các yếu tố như thu nhập, tuổi nhà, dân số không bị phụ thuộc lẫn nhau.

➤ **Ý nghĩa tổng quan**

1. Tương quan với Price:

- Thu nhập (Income), tuổi nhà (Age), và dân số (Population) là các yếu tố chính ảnh hưởng đến giá nhà.
- Số phòng ngủ (Bedrooms) có ít ảnh hưởng hơn, do đó có thể xem xét giảm trọng số hoặc loại bỏ khỏi mô hình nếu cần đơn giản hóa.

2. Phân tích đa biến:

- Không có sự phụ thuộc mạnh giữa các biến độc lập (trừ Rooms và Bedrooms), cho thấy dữ liệu ít bị đa cộng tuyến (multicollinearity).

➤ **Ứng dụng trong bài toán dự đoán giá nhà**

Quan trọng nhất:

- Biến Income nên được ưu tiên trong việc xây dựng mô hình dự đoán, vì nó có mối tương quan mạnh nhất với giá nhà.
- Age và Population cũng cần được giữ lại vì chúng có mức tương quan đáng kể với Price.

Giảm số lượng biến:

- Có thể xem xét loại bỏ Bedrooms, vì nó có mối tương quan thấp nhất với Price và bị ảnh hưởng bởi Rooms.

➤ Kết luận

Biểu đồ heatmap cho thấy thu nhập (Income) là yếu tố quan trọng nhất trong việc dự đoán giá nhà, theo sau là tuổi nhà (Age) và dân số (Population). Các mối quan hệ này có thể được khai thác trong các mô hình dự đoán và giúp xác định giá trị nhà ở dựa trên các yếu tố liên quan.

3.3.4. Phân tích hồi quy

Với dự án hiện tại, mục tiêu được đặt ra là cần dự báo giá trị nhà trung bình dựa theo mô hình hồi quy tuyến tính. Từ đó, ta đặt ra biến mục tiêu để dự báo (Target) chính là 'Price' của bộ dữ liệu.

Bởi sự tương quan giữa một số biến độc lập như Rooms và Bedrooms với biến phụ thuộc chưa được tốt có mức tương quan yếu, đặc biệt Bedrooms có thể không đóng góp nhiều vào mô hình hồi quy, do vậy, ta cần thực hiện 'chọn lọc đặc trưng' để có thể tìm được sự phụ thuộc tốt nhất cho bài toán trên.


```

from sklearn.metrics import mean_squared_error, r2_score
import numpy as np
import copy
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import PolynomialFeatures

old = list(df.columns)
attributes = ['Income', 'Age', 'Rooms', 'Bedrooms', 'Population', 'Price']
new = ["x1", "x2", "x3", "x4", "x5", 'y']
df2 = df.copy()
df2 = df2.rename(columns=dict(zip(old, new)))

X = df2
y = df2['y']

from sklearn.feature_selection import SelectKBest, f_regression
list_feature = []
feature_X_new = X.shape[1]
print(feature_X_new)

for i in range(1, feature_X_new):
    k_best = SelectKBest(score_func=f_regression, k=i)
    X_select = k_best.fit_transform(X, y)
    best_feature_names = X.columns[k_best.get_support()]
    list_feature.append(best_feature_names)

```

```

def all_result(best_feature_names, X_new, y_new):
    x = X_new[best_feature_names]
    y = y_new

    x_train_2, x_test_2, y_train_1, y_test_1 = train_test_split(x, y, test_size=0.2, random_state = 42)

    scaler = MinMaxScaler()
    x_train_2 = scaler.fit_transform(x_train_2)
    x_test_2 = scaler.transform(x_test_2)

    re_2feature = LinearRegression()
    re_2feature.fit(x_train_2, y_train_1)

    y_test_predict = re_2feature.predict(x_test_2)
    Y_train_predict = re_2feature.predict(x_train_2)
    rate(y_test_1, y_test_predict, y_train_1, Y_train_predict)

    # Xác định phương trình hồi quy
    coefficients = re_2feature.coef_ # Hệ số của các đặc trưng
    intercept = re_2feature.intercept_ # Hệ số chặn
    feature_equation = " + ".join(
        [f"{coef:.2f} * {feature}" for coef, feature in zip(coefficients, best_feature_names)]
    )
    regression_equation = f"y = {intercept:.2f} + {feature_equation}"

    print(f"Phương trình hồi quy tuyến tính: {regression_equation}")

```

Hình 3. 11: Mô tả chọn lọc đặc trưng

Ta sẽ sử dụng hàm SelectKBest để chọn ra các đặc trưng tốt nhất dựa vào K, sau đó, ta sẽ lấy được tên của những features dựa vào k đó bằng hàm 'get_support'. Sau khi có được những features đó, ta gán chúng vào X, còn y chính là cột Price.

Ta tiến hành chia dữ liệu và sau đó chuẩn hóa dữ liệu của X

Tiếp đó, ta huấn luyện bằng mô hình 'Hồi quy tuyến tính' được thể hiện thông qua LinearRegression().

Cuối cùng, ta đánh giá mô hình dựa trên kết quả của hàm rate:

```
def rate(y_test, y_test_predict, y_train, y_train_predict):  
    # RMSE trên tập test  
    rmse_test = np.sqrt(mean_squared_error(y_test, y_test_predict))  
    # RMSE trên tập train  
    rmse_train = np.sqrt(mean_squared_error(y_train, y_train_predict))  
    # R2 Score trên tập train  
    r2_train = r2_score(y_train, y_train_predict)  
    # R2 Score trên tập test  
    r2_test = r2_score(y_test, y_test_predict)  
  
    # In kết quả  
    print(f"RMSE Train: {rmse_train:.2f}")  
    print(f"R2 Score Train: {r2_train:.2f}")  
    print(f"RMSE Test: {rmse_test:.2f}")  
    print(f"R2 Score Test: {r2_test:.2f}")
```

Hình 3. 12: Hàm 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 thẩm định (test).

Sau đây là kết quả:

```

x1: Income
x2: Age
x3: Rooms
x4: Bedrooms
x5: Population
y: Price

=====

RMSE Train: 271139.97
R2 Score Train: 0.41
RMSE Test: 272387.38
R2 Score Test: 0.40
Phương trình hồi quy tuyến tính:  $y = 152393.38 + 1843183.18 * x1$ 

=====

RMSE Train: 219197.17
R2 Score Train: 0.62
RMSE Test: 218363.68
R2 Score Test: 0.61
Phương trình hồi quy tuyến tính:  $y = -378762.98 + 1847547.13 * x1 + 1098222.59 * x2$ 

=====

RMSE Train: 158655.03
R2 Score Train: 0.80
RMSE Test: 158799.18
R2 Score Test: 0.80
Phương trình hồi quy tuyến tính:  $y = -947300.34 + 1874064.88 * x1 + 1114344.80 * x2 + 1053998.98 * x5$ 

=====

RMSE Train: 101308.08
R2 Score Train: 0.92
RMSE Test: 100367.93
R2 Score Test: 0.92
Phương trình hồi quy tuyến tính:  $y = -1413942.65 + 1882364.28 * x1 + 1125911.57 * x2 + 852330.76 * x3 + 1060175.67 * x5$ 

=====

```

RMSE Train: 101273.49

R2 Score Train: 0.92

RMSE Test: 100444.06

R2 Score Test: 0.92

Phương trình hồi quy tuyến tính: $y = -1413286.26 + 1881708.61 * x_1 + 1125667.50 * x_2 + 842611.01 * x_3 + 10981.70 * x_4 + 1060509.57 * x_5$

=====

Tổng hợp kết quả bằng kỹ thuật 10-Fold Cross Validation theo các bước:

Bước 1: Chia dữ liệu thành 10 phần

- Chia tập dữ liệu ngẫu nhiên thành **10 phần** bằng nhau (hoặc gần bằng nhau).
- Mỗi phần được gọi là một **fold**.

Bước 2: Huấn luyện và kiểm tra

- Lặp qua từng fold (từ 1 đến 10):
 1. Sử dụng 9 fold để huấn luyện mô hình.
 2. Sử dụng fold còn lại làm tập kiểm tra để đánh giá mô hình.

Bước 3: Tính toán hiệu suất

- Đo lường hiệu suất của mô hình trên tập kiểm tra (ví dụ: **Accuracy**, **Precision**, **Recall**, **RMSE**, hoặc các chỉ số khác phù hợp với bài toán).
- Lưu kết quả đánh giá của từng fold.

Bước 4: Tính kết quả trung bình

- Tính **trung bình** và **độ lệch chuẩn** của các chỉ số hiệu suất trên 10 fold.
- Đây là kết quả tổng quát của mô hình, phản ánh khả năng tổng quát hóa trên dữ liệu mới.

```

import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import KFold
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler
import time

# 1. Chọn thuộc tính và biến phụ thuộc
X = df[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',
        'Avg. Area Number of Bedrooms', 'Area Population']].values
y = df['Price'].values

# 2. Chuẩn hóa dữ liệu
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X)

# 3. Khởi tạo mô hình và thông số
model = LinearRegression()
kf = KFold(n_splits=10, shuffle=True, random_state=42) # 10-fold cross-validation

# 4. Lưu kết quả
train_times = []
test_times = []
rmse_results = []
r2_results = []

# 5. Đánh giá mô hình
fold_details = []
for train_index, test_index in kf.split(X_scaled):
    X_train, X_test = X_scaled[train_index], X_scaled[test_index]
    y_train, y_test = y[train_index], y[test_index]

    # Huấn luyện và kiểm tra thời gian
    start_train = time.time()
    model.fit(X_train, y_train)
    train_time = time.time() - start_train
    train_times.append(train_time)

    start_test = time.time()
    y_pred = model.predict(X_test)
    test_time = time.time() - start_test
    test_times.append(test_time)

    # Tính RMSE và R2
    rmse = np.sqrt(mean_squared_error(y_test, y_pred))
    r2 = r2_score(y_test, y_pred)
    rmse_results.append(rmse)
    r2_results.append(r2)

    # Lưu chi tiết từng fold
    fold_details.append({
        'Fold': len(fold_details) + 1,
        'Train Time': train_time,
        'Test Time': test_time,
        'RMSE': rmse,
        'R2': r2
    })

```

```

# 6. Tính trung bình kết quả
average_rmse = np.mean(rmse_results)
average_r2 = np.mean(r2_results)
average_train_time = np.mean(train_times)
average_test_time = np.mean(test_times)

# 7. Tạo bảng chi tiết
df_details = pd.DataFrame(fold_details)
print("\n===== Bảng chi tiết từng fold =====")
print(df_details)
print("=====")
print("\n")

# 8. In kết quả tổng hợp
print("===== Kết quả tổng hợp =====")
print(f"Trung bình RMSE: {average_rmse:.4f}")
print(f"Trung bình R2: {average_r2:.4f}")
print(f"Thời gian train trung bình: {average_train_time:.4f} giây")
print(f"Thời gian test trung bình: {average_test_time:.4f} giây")
print("=====")

```

Hình 3. 13: Mô tả kỹ thuật 10 – Fold Cross Validation

Kết quả sau khi tiến hành kỹ thuật 10 – Fold Cross Validation

```

===== Bảng chi tiết từng fold =====

```

	Fold	Train Time	Test Time	RMSE	R2
0	1	0.003823	0.000000	101609.987606	0.918164
1	2	0.002609	0.001005	99155.074954	0.917995
2	3	0.002079	0.000000	100406.350633	0.906619
3	4	0.002012	0.000000	101344.811049	0.921254
4	5	0.001524	0.001000	104257.727251	0.901544
5	6	0.003002	0.000000	102871.999281	0.919678
6	7	0.004004	0.000000	98505.938636	0.923458
7	8	0.002998	0.000000	103818.177336	0.915252
8	9	0.001521	0.000000	102639.675035	0.921360
9	10	0.002000	0.000000	97507.328119	0.927343

```

=====

===== Kết quả tổng hợp =====
Trung bình RMSE: 101211.7070

```

Trung bình R2: 0.9173

Thời gian train trung bình: 0.0026 giây

Thời gian test trung bình: 0.0002 giây

=====

3.4. Đánh giá và đề xuất

a. Đánh giá mô hình hồi quy tuyến tính

R^2 : 0.92

R^2 là chỉ số thể hiện phần trăm phương sai của biến phụ thuộc (giá nhà) được giải thích bởi các biến độc lập trong mô hình. Với giá trị $R^2 = 0.92$, mô hình giải thích được 92% sự biến động của giá trị mục tiêu, điều này cho thấy mô hình có khả năng dự đoán khá chính xác. Đây là một kết quả khá ấn tượng, chứng tỏ rằng các yếu tố được chọn trong mô hình có ảnh hưởng mạnh mẽ đến giá nhà.

$RMSE$: \$101,211

$RMSE$ đo lường sai số trung bình giữa các dự đoán và giá trị thực tế. Với giá trị $RMSE = \$101,211$, sai số này có thể được coi là khá lớn nếu xét trong bối cảnh giá nhà dao động từ \$15,938.66 đến \$2,469,066. Tuy nhiên, khi tính tỷ lệ sai số so với phạm vi giá trị của "Price", ta có thể tính được tỷ lệ sai số khoảng 4.1%. Cụ thể, tỷ lệ sai số có thể được tính như sau:

$$\text{Tỷ lệ sai số} = \frac{101,211}{2,453.128} \times 100 \approx 4.1\%$$

Tỷ lệ sai số này cho thấy rằng mức độ sai số không phải là quá lớn so với phạm vi giá trị tổng thể. Tuy nhiên, với các giá trị nhà có mức giá thấp, sai số này có thể chiếm tỷ lệ lớn hơn, gây khó khăn trong việc dự báo chính xác cho những bất động sản có giá trị thấp.

b. Đề xuất cải thiện

Để cải thiện hiệu suất mô hình và giảm thiểu sai số, có thể áp dụng một số phương pháp như sau:

Xử lý dữ liệu đầu vào

- **Chuẩn hóa dữ liệu:** Khi các biến độc lập có giá trị dao động lớn, chuẩn hóa chúng có thể giúp mô hình học hiệu quả hơn và cải thiện kết quả dự đoán.
- **Thêm biến phi tuyến:** Một phương pháp có thể cải thiện mô hình là thêm các biến phi tuyến như bình phương, căn bậc hai, hoặc tương tác (interaction terms) giữa các yếu tố. Những biến này giúp mô hình nắm bắt được các mối quan hệ phức tạp hơn giữa các đặc trưng và giá nhà.

Thử nghiệm mô hình khác

- **Random Forest hoặc Gradient Boosting:** Các mô hình như Random Forest hoặc Gradient Boosting (ví dụ: XGBoost, LightGBM) có thể xử lý tốt hơn với các mối quan hệ phi tuyến và phân phối phức tạp trong dữ liệu. So sánh hiệu suất của các mô hình này với hồi quy tuyến tính có thể giúp xác định mô hình phù hợp hơn cho bài toán dự đoán giá nhà.

Kiểm tra outliers (giá trị ngoại lai)

- Giá trị ngoại lai (outliers) có thể làm gia tăng giá trị $RMSE$ và làm giảm độ chính xác của mô hình. Sử dụng các biểu đồ như boxplot có thể giúp phát hiện và loại bỏ hoặc xử lý các giá trị ngoại lai này.

Đánh giá thêm các chỉ số hiệu suất

- Bên cạnh R^2 và $RMSE$, việc bổ sung các chỉ số khác như Mean Absolute Error (MAE) sẽ giúp cung cấp cái nhìn toàn diện hơn về sai số của mô hình và đánh giá hiệu suất dự báo một cách đầy đủ hơn.

3.5. Kết luận

Chương 3 đã trình bày phần thực nghiệm và đánh giá của đề tài thông qua đầy đủ các bước từ tiền xử lý dữ liệu cho tới phân tích mô tả và bài toán dự báo. Từ đó đưa ra được các đánh giá và đề xuất phù hợp để cải thiện kết quả của đề tài trong tương lai.

CHƯƠNG 4 : XÂY DỰNG SẢN PHẨM DEMO

4.1. Giới thiệu về Framework được sử dụng

Tkinter là một thư viện mạnh mẽ và dễ sử dụng trong Python, được tích hợp sẵn để xây dựng giao diện người dùng đồ họa (GUI - Graphical User Interface). Là một wrapper của **Tcl/Tk**, một hệ thống giao diện đồ họa lâu đời, Tkinter giúp các lập trình viên Python dễ dàng tạo ra các ứng dụng GUI mà không cần cài đặt thêm bất kỳ thư viện bên ngoài nào. Mặc dù Tkinter có thể không mạnh mẽ như các thư viện GUI phức tạp khác, nhưng với sự đơn giản và hiệu quả, nó đã trở thành một công cụ lý tưởng cho những ứng dụng GUI nhỏ gọn hoặc dành cho người mới bắt đầu.

Tkinter cung cấp một loạt các widget cơ bản, chẳng hạn như Label (nhãn), Button (nút bấm), Entry (trường nhập liệu), Text (vùng văn bản) và Canvas (vẽ đồ họa), giúp lập trình viên dễ dàng tạo ra giao diện người dùng linh hoạt và trực quan. Thư viện này hỗ trợ mô hình lập trình dựa trên sự kiện (event-driven programming), giúp gán các hành động cho từng widget, ví dụ như khi người dùng nhấn nút hoặc nhập liệu vào một trường. Điều này giúp tạo ra những ứng dụng GUI phản hồi nhanh chóng và dễ dàng tương tác.

Cấu trúc cơ bản của một ứng dụng Tkinter thường bao gồm ba bước chính: Tạo cửa sổ chính, thêm các widget vào cửa sổ và cuối cùng là chạy vòng lặp sự kiện. Vòng lặp này sẽ duy trì cửa sổ giao diện mở và chờ đợi các hành động từ người dùng. Một ví dụ điển hình là tạo ra một cửa sổ đơn giản với một nhãn và một nút. Khi người dùng nhấn nút, nội dung của nhãn sẽ thay đổi. Đây là cách thức cơ bản để xây dựng các ứng dụng GUI nhẹ nhàng nhưng hiệu quả với Tkinter.

Một trong những ưu điểm lớn của Tkinter là sự đơn giản và dễ học. Đối với những ai mới bắt đầu tìm hiểu lập trình GUI, Tkinter mang đến một cách tiếp cận trực quan, giúp người dùng nhanh chóng xây dựng và thử nghiệm ứng dụng của mình. Hơn nữa, vì là một thư viện chuẩn của Python, Tkinter không yêu cầu cài đặt thêm bất kỳ thư viện nào, giúp tiết kiệm thời gian và công sức cho người sử dụng. Tuy nhiên, mặc dù Tkinter rất mạnh mẽ trong việc xây dựng các ứng dụng đơn giản, nhưng khi yêu cầu tính năng phức tạp hơn hoặc giao diện đồ họa chi tiết, các thư viện như PyQt hay wxPython có thể là những lựa chọn tối ưu hơn.

Mặc dù giao diện mặc định của Tkinter có thể không bắt mắt như một số thư viện GUI khác, nhưng bạn hoàn toàn có thể tùy chỉnh giao diện của mình theo nhu cầu, từ màu sắc, font chữ, đến bố cục của các widget. Điều này mang đến sự linh hoạt và khả năng mở rộng khi phát triển các ứng dụng đồ họa với Python. Với Tkinter, việc tạo ra các ứng dụng GUI nhanh chóng và dễ dàng là một lợi thế lớn, tuy nhiên, đối với những ứng dụng đòi hỏi các tính năng đồ họa phức tạp, việc chuyển sang các công cụ mạnh mẽ hơn sẽ là một lựa chọn hợp lý.

Tóm lại, **Tkinter** là một thư viện lý tưởng cho việc phát triển các ứng dụng GUI đơn giản và hiệu quả trong Python. Sự dễ sử dụng, tính linh hoạt và khả năng tích hợp tốt với các ứng dụng Python giúp Tkinter trở thành công cụ rất phù hợp cho những ai muốn xây dựng các ứng dụng giao diện người dùng mà không gặp phải các phức tạp của các thư viện đồ họa phức tạp.

4.2. Chuẩn bị tài nguyên xây dựng chương trình

Khi thực hiện chuẩn hóa dữ liệu, ta sẽ lưu mô hình chuẩn hóa nhằm mục đích chuẩn hóa dữ liệu nhập vào

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(*arrays: X, y, test_size=0.3, random_state=42)

from sklearn.preprocessing import StandardScaler, MinMaxScaler

from sklearn.pipeline import Pipeline

pipeline = Pipeline([
    ('min_max_scaler', MinMaxScaler())
])

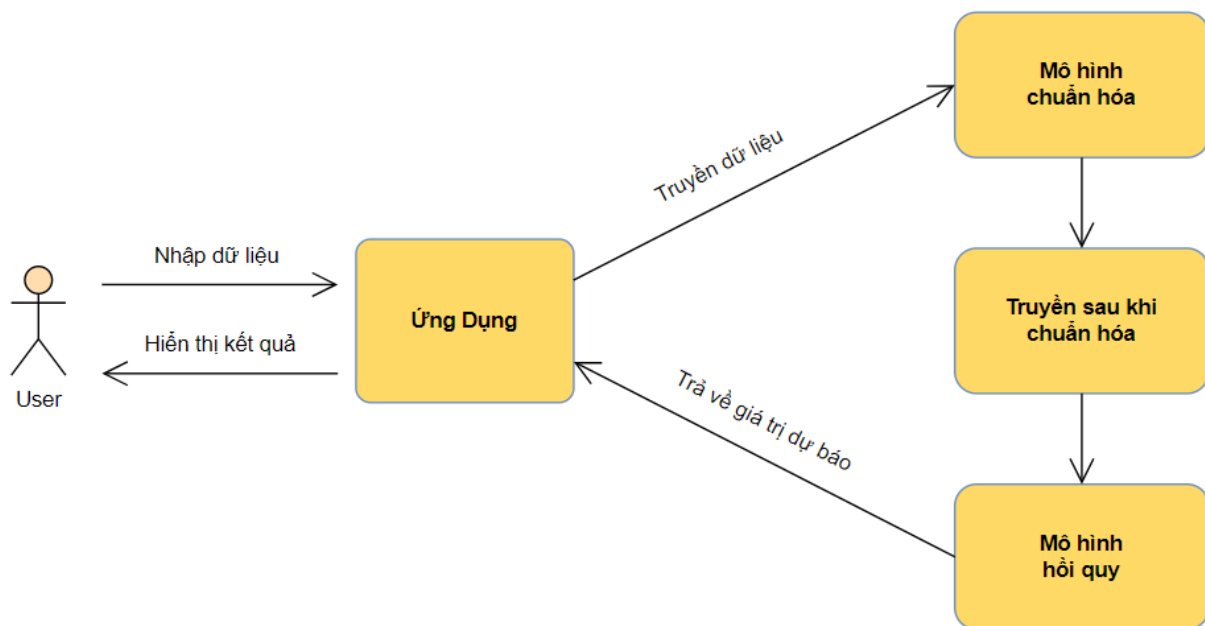
X_train = pipeline.fit_transform(X_train)
X_test = pipeline.transform(X_test)
```

Hình 4. 1: Chuẩn hóa Min, Max

4.3. Xây dựng mô hình và demo chương trình

4.3.1. Xây dựng mô hình

Mô hình xây dựng chương trình như sau:



Hình 4. 2: Mô hình xây dựng chương trình

- Mô tả use case dự báo giá nhà

Bảng 4. 1: Mô tả use case dự báo giá nhà

Mã use case	UC1
Tên use case	Dự báo giá nhà
Tóm tắt	Use case này cho phép người dùng dự báo giá nhà dựa vào các giá trị đã nhập
Actor	Người dùng
Tiền điều kiện	Người dùng đã nhập dữ liệu hợp lệ
Đảm bảo tối thiểu	Nếu người dùng ấn nút “Predict” mà không có tiền điều kiện, hệ thống sẽ thông báo lỗi

Đảm bảo thành công	Người dùng dự báo được giá trị dựa trên dữ liệu đã nhập
Kích hoạt	Tự động
Luồng sự kiện	Người dùng nhập thông tin giá nhà và kích vào nút “Predict”

- Mô tả use case reset dữ liệu

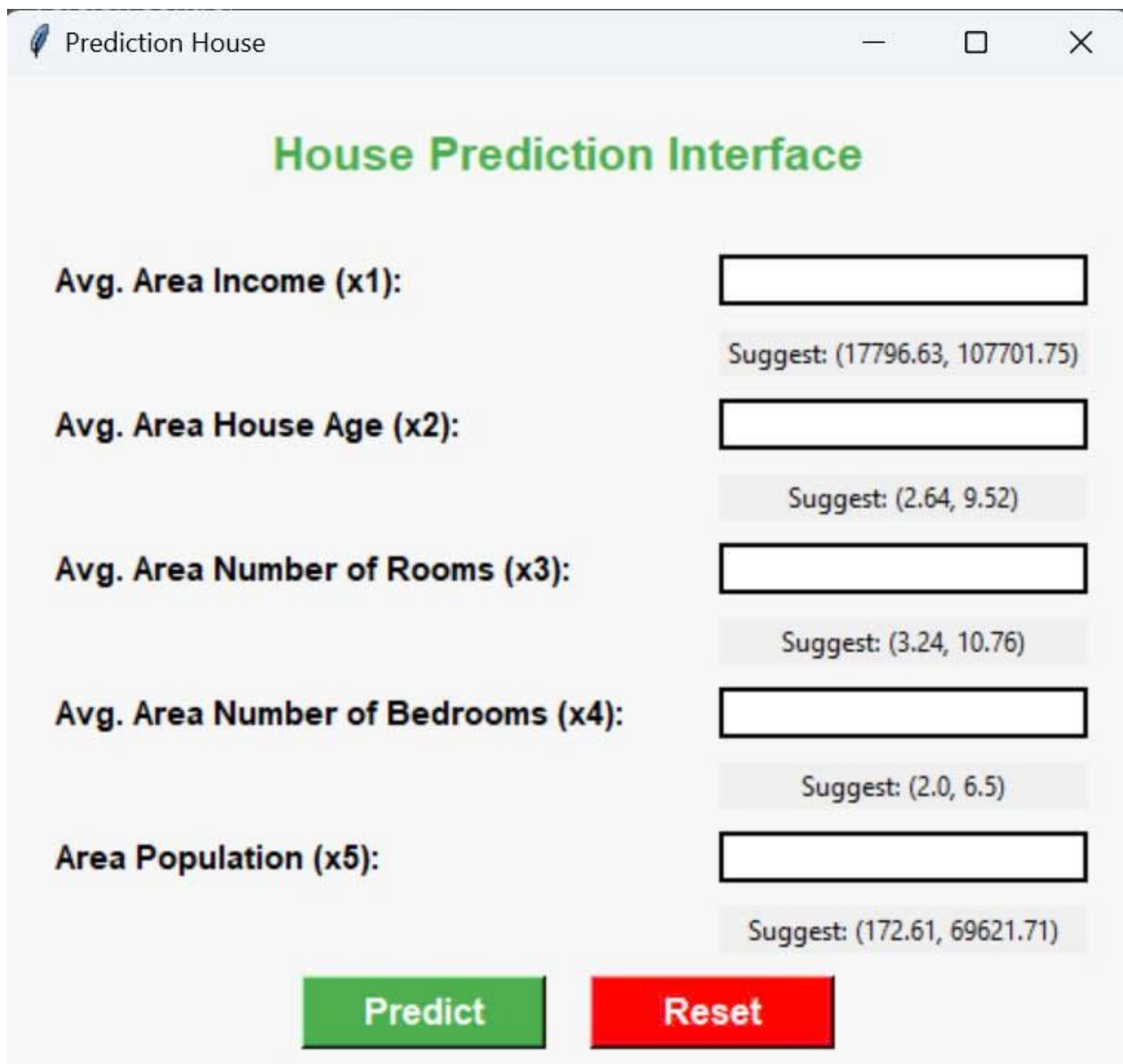
Bảng 4. 2: Mô tả use case reset dữ liệu

Mã use case	UC2
Tên use case	Reset dữ liệu
Tóm tắt	Use case này cho phép người xóa tất cả dữ liệu trong các trường nhập thông tin
Actor	Người dùng
Tiền điều kiện	Không có
Đảm bảo tối thiểu	Không có
Đảm bảo thành công	Xóa toàn bộ dữ liệu trong các trường nhập và nhận hiển thị kết quả
Kích hoạt	Tự động

Luồng sự kiện	Người dùng kích vào nút “Reset”
----------------------	---------------------------------

4.3.2 Demo chương trình

Giao diện khi mở chương trình



Prediction House

House Prediction Interface

Avg. Area Income (x1):
Suggest: (17796.63, 107701.75)

Avg. Area House Age (x2):
Suggest: (2.64, 9.52)

Avg. Area Number of Rooms (x3):
Suggest: (3.24, 10.76)

Avg. Area Number of Bedrooms (x4):
Suggest: (2.0, 6.5)

Area Population (x5):
Suggest: (172.61, 69621.71)

Predict **Reset**

Hình 4. 3: Giao diện chương trình

Prediction House

House Prediction Interface

Avg. Area Income (x1):
Suggest: (17796.63, 107701.75)

Avg. Area House Age (x2):
Suggest: (2.64, 9.52)

Avg. Area Number of Rooms (x3):
Suggest: (3.24, 10.76)

Avg. Area Number of Bedrooms (x4):
Suggest: (2.0, 6.5)

Area Population (x5):
Suggest: (172.61, 69621.71)

Hình 4. 4: Quá trình nhập dữ liệu đầu vào

Sau đó ấn nút predict, dữ liệu sẽ được truyền vào các biến tương ứng:

```

avg_income = float(entry_income.get().strip())
check1 = check_input(max_arr[0], min_arr[0], avg_income)
avg_income = min_max_normalization(avg_income, max_arr[0], min_arr[0],)
house_age = float(entry_house_age.get().strip())
check2 = check_input(max_arr[1], min_arr[1], house_age)
house_age = min_max_normalization(house_age, max_arr[1], min_arr[1])
num_rooms = float(entry_num_rooms.get().strip())
check3 = check_input(max_arr[2], min_arr[2], num_rooms)
num_rooms = min_max_normalization(num_rooms, max_arr[2], min_arr[2])
num_bedrooms = float(entry_num_bedrooms.get().strip())
check4 = check_input(max_arr[3], min_arr[3], num_bedrooms)
num_bedrooms = min_max_normalization(num_bedrooms, max_arr[3], min_arr[3])
area_population = float(entry_area_population.get().strip())
check5 = check_input(max_arr[4], min_arr[4], area_population)
area_population = min_max_normalization(area_population, max_arr[4], min_arr[4])

```

Hình 4. 5: Mô tả quá trình truyền dữ liệu

Sau đó, mô hình chuẩn hóa sẽ chuẩn hóa các biến trên và dự báo kết quả

```

def predict():
    check4 = check_input(max_arr[3], min_arr[3], num_bedrooms)
    num_bedrooms = min_max_normalization(num_bedrooms, max_arr[3], min_arr[3])
    area_population = float(entry_area_population.get().strip())
    check5 = check_input(max_arr[4], min_arr[4], area_population)
    area_population = min_max_normalization(area_population, max_arr[4], min_arr[4])
    print(check1, check2, check3, check4, check5)
    if check1 or check2 or check3 or check4 or check5:
        messagebox.showwarning( title: "Warning",
                                message: "Please enter values within the recommended range to ensure the accuracy of the results!")
    y = (-1413286.26 + 1881708.61 * avg_income + 1125667.50 * house_age + 842611.01 * num_rooms
        + 10981.70 * num_bedrooms + 1060509.57 * area_population)
    if y<0:
        messagebox.showinfo( title: 'Message',
                              message: 'Do giá trị nhập không thực tế trong ngữ cảnh dự đoán giá nhà,'
                              ' do đó được điều chỉnh về 0 để đúng với ý nghĩa thực tế!')
        y=0
    lbl_name_equation.config(text="Linear Regression Equation:")
    lbl_name_result.config(text="Price Prediction Result:")
    lbl_result.config(text=y)
    lbl_equation.config(
        text=f"y = {round(lin_reg.intercept_, 2)} + {round(coefs[0], 2)} * x1 + {round(coefs[1], 2)} * x2 "
        f"+ \n{round(coefs[2], 2)} * x3 + {round(coefs[3], 2)} * x4 + {round(coefs[4], 2)} * x5")
    except ValueError:
        messagebox.showerror( title: "Error", message: "Please enter valid numeric values!")

```

Hình 4. 6: Quá trình chuẩn hóa dữ liệu

Kết quả của chương trình:

Prediction House

— □ ×

House Prediction Interface

Avg. Area Income (x1):

60000

Suggest: (17796.63, 107701.75)

Avg. Area House Age (x2):

5

Suggest: (2.64, 9.52)

Avg. Area Number of Rooms (x3):

6

Suggest: (3.24, 10.76)

Avg. Area Number of Bedrooms (x4):

6

Suggest: (2.0, 6.5)

Area Population (x5):

30000

Suggest: (172.61, 69621.71)

Predict

Reset

Linear Regression Equation:

$$y = -1367019.46 + 1879413.7 * x1 + 1131983.4 * x2 + 844046.34 * x3 + 10624.93 * x4 + 1009411.14 * x5$$

Price Prediction Result:

\$630,524.01

Hình 4. 7: Kết quả của chương trình

KẾT LUẬN

Trong quá trình thực hiện đề tài "Phân tích dự báo giá nhà ở Mỹ bằng phương pháp hồi quy tuyến tính," nhóm chúng em đã nghiên cứu kỹ lưỡng và triển khai nhiều hoạt động cần thiết để đạt được mục tiêu đề ra. Những nỗ lực này bao gồm nhiều khía cạnh quan trọng sau:

Trước hết, nhóm đã tiến hành thu thập dữ liệu từ nhiều nguồn uy tín, chủ yếu tập trung vào các tập dữ liệu bất động sản tại Mỹ. Sau khi dữ liệu được thu thập, nhóm đã thực hiện tiền xử lý như loại bỏ nhiễu, chuẩn hóa các biến phân loại và mã hóa dữ liệu để đảm bảo tính nhất quán. Sau đó, nhóm đã áp dụng các kỹ thuật phân tích dữ liệu như phân tích mô tả theo đơn biến và đa biến để hiểu rõ các xu hướng trong tập dữ liệu.

Về khía cạnh áp dụng mô hình, nhóm đã sử dụng phương pháp hồi quy tuyến tính để phân tích mối quan hệ giữa các biến độc lập như thu nhập và tuổi nhà đối với giá nhà. Chúng em đã đánh giá hiệu suất mô hình qua các chỉ số như R^2 và RMSE, đồng thời sử dụng kỹ thuật 10 – Fold Cross Validation để đảm bảo tính đồng nhất trong kết quả.

Kết quả nghiên cứu cho thấy, mô hình đã đạt được độ chính xác đáng kể, phân tích rõ ràng những yếu tố ảnh hưởng lớn nhất đến giá nhà như thu nhập và dân cư khu vực. Tuy nhiên, chúng em nhận thấy vẫn còn nhiều khía cạnh chưa khai thác hết, chẳng hạn như chưa thử nghiệm với các mô hình học sâu khác như Random Forest hoặc mạng nơ-ron nhân tạo, và chưa tích hợp các biến kinh tế vĩ mô như lãi suất hoặc tỷ lệ lạm phát.

Trong tương lai, nhóm dự kiến sẽ nghiên cứu để tích hợp và so sánh với các mô hình học máy tiên tiến khác. Chúng em cũng sẽ tăng cường quy mô dữ liệu, đồng thời tăng khả năng trực quan hóa kết quả bằng các biểu đồ tương tác. Cuối cùng, mô hình sẽ được triển khai trong thực tế để hỗ trợ quyết định và dự báo cho ngành bất động sản.

Tổng kết, đề tài này đã giúp nhóm chúng em nâng cao năng lực nghiên cứu, phân tích dữ liệu và áp dụng mô hình học máy vào thực tiễn. Những kết quả này không chỉ mang ý nghĩa học thuật mà còn góp phần cung cấp tài liệu tham khảo quan trọng cho những nghiên cứu sau này.

TÀI LIỆU THAM KHẢO

[1] Sanford Weisberg: Applied Linear Regression. Nhà xuất bản John Wiley & Sons năm 2013

[2] Website học machine learning cơ bản, <https://machinelearningcoban.com/>, ngày truy cập gần nhất: 25/12/2024.

[3] Website tìm hiểu mạng nơ ron nhân tạo, <https://www.deeplearning.ai/>, ngày truy cập gần nhất: 25/12/2024.

[4] Website tìm hiểu các mô hình hồi quy tuyến tính, <https://people.duke.edu/~rnau/testing.htm>, ngày truy cập gần nhất: 25/12/2024

[5] Website học ngôn ngữ lập trình, <https://www.w3schools.com/>, ngày truy cập gần nhất: 25/12/2024