**TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM**

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ MÔN HỌC MÔN**

**NHẬP MÔN HỌC MÁY**

**MACHINE LEARNING**

**FINAL ESSAY**

*Người hướng dẫn*: **Thầy Lê Anh Cường**

*Người thực hiện*: **Trần Trung Hiếu – 52000888**

**Nguyễn Tiến Dũng – 52000883**

**Trần Quang Luân – 52100254**

Lớp **: 503044**

Khoá  **: 24**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

**TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM**

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ MÔN HỌC MÔN**

**NHẬP MÔN HỌC MÁY**

**MACHINE LEARNING**

**FINAL ESSAY**

*Người hướng dẫn*: **Thầy Lê Anh Cường**

*Người thực hiện*: **Trần Trung Hiếu – 52000888**

**Nguyễn Tiến Dũng – 52000883**

**Trần Quang Luân – 52100254**

Lớp **: 503044**

Khoá  **: 24**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

LỜI CẢM ƠN

Chúng em xin chân thành gửi lời cảm ơn này đến Thầy **Lê Anh Cường** giảng viên phụ trách giảng dạy bộ môn nhập môn học máy. Nhờ có sự tận tình giảng dạy, truyền đạt kiến thức của quý cô mà chúng em mới đủ kiến thức để hoàn thành báo cáo cuối kỳ này.

Song song với đó, chúng em cũng xin gửi lời cảm ơn đến Khoa Công Nghệ Thông Tin, trường Đại học Tôn Đức Thắng vì đã tạo điều kiện cho chúng em học tập, nghiên cứu trong suốt quá trình học tập môn học này nói riêng và cả quá trình học tại môi trường Đại học nói chung. Một lần nữa chúng em xin gửi lời cảm ơn chân thành đến mọi người và chúc tất cả thật nhiều sức khỏe.

*TP. Hồ Chí Minh, ngày 23 tháng 12 năm 2023*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*Nguyễn Tiến Dũng*

**ĐỒ ÁN / BÁO CÁO ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của Thầy **Lê Anh Cường**. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án môn học còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án môn học của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. HCM, ngày 23 tháng 12 năm 2023*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*Nguyễn Tiến Dũng*

TÓM TẮT

Cùng với sự phát triển của các lĩnh vực kinh tế, xã hội, nhu cầu ứng dụng công nghệ thông tin ngày càng cao và không ngừng biến đổi. Học máy là một lĩnh vực của trí tuệ nhân tạo, tập trung vào việc phát triển các thuật toán cho phép máy tính học từ dữ liệu mà không cần được lập trình rõ ràng. Học máy có thể được sử dụng để giải quyết nhiều loại vấn đề, bao gồm phân loại, dự đoán, phân cụm và xử lý ngôn ngữ tự nhiên.

Môn học nhập môn học máy cung cấp cho sinh viên kiến thức cơ bản về học máy. Môn học này bao gồm các chủ đề như các khái niệm cơ bản về học máy, tiền xử lý dữ liệu, xây dựng mô hình học máy và ứng dụng học máy.

Kiến thức và kỹ năng cần có để theo học môn học này bao gồm kiến thức cơ bản về toán và thống kê, và khả năng lập trình bằng Python hoặc một ngôn ngữ lập trình khác.

Mục tiêu của môn học này là giúp sinh viên hiểu được các khái niệm cơ bản về học máy và cách áp dụng học máy để giải quyết các vấn đề thực tế.

Học máy là một lĩnh vực đang phát triển nhanh chóng và có nhiều ứng dụng tiềm năng trong nhiều lĩnh vực khác nhau. Môn học nhập môn học máy cung cấp cho sinh viên nền tảng vững chắc để theo đuổi lĩnh vực này.

MỤC LỤC

[LỜI CẢM ƠN 1](#_Toc121002887)

[TÓM TẮT](#_Toc121002887) 3

[CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN 5](#_Toc121002888)

[1.1 Giới thiệu môn học 5](#_Toc121002889)

[1.2 Đề tài nghiên cứu 6](#_Toc121002890)

[PHẦN 2: ĐỀ TÀI BÁO CÁO 7](#_Toc121002891)

[2.1 Câu hỏi 1 7](#_Toc121002892)

[2.2 Câu hỏi 2,3,4 1](#_Toc121002892)1

[2.3 Câu hỏi 5 2](#_Toc121002892)4

DANH MỤC HÌNH ẢNH

[Hình 1: Đồ thị biểu diễn giá nhà theo quý mỗi năm 8](#_Toc121002889)

[Hình 2: Đồ thị biểu diễn giá nhà theo khu vực 9](#_Toc121002889)

[Hình 3: Đồ thị biểu diễn giá nhà theo phòng 9](#_Toc121002889)

[Hình 4: Đồ thị biểu diễn giá nhà theo loại nhà 1](#_Toc121002889)0

CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN

1.1 Giới thiệu môn học

Học máy là một lĩnh vực của trí tuệ nhân tạo, tập trung vào việc phát triển các thuật toán cho phép máy tính học từ dữ liệu mà không cần được lập trình rõ ràng. Học máy có thể được sử dụng để giải quyết nhiều loại vấn đề, bao gồm phân loại, dự đoán, phân cụm và xử lý ngôn ngữ tự nhiên.

Môn nhập môn học máy cung cấp cho sinh viên kiến thức cơ bản về học máy. Môn học này bao gồm các chủ đề sau:

- Các khái niệm cơ bản về học máy: Định nghĩa học máy, các loại học máy, các thuật toán học máy phổ biến.

- Tiền xử lý dữ liệu: Làm sạch dữ liệu, chuyển đổi dữ liệu, chuẩn hóa dữ liệu.

- Xây dựng mô hình học máy: Chọn mô hình, đào tạo mô hình, đánh giá mô hình.

- Ứng dụng học máy: Một số ứng dụng của học máy trong thực tế.

Học máy có thể được ứng dụng trong nhiều lĩnh vực khác nhau, bao gồm:

- Phân loại: Phân loại hình ảnh, phân loại văn bản, phân loại âm thanh, v.v.

- Dự đoán: Dự đoán giá cổ phiếu, dự đoán doanh thu, dự đoán điểm thi, v.v.

- Phân cụm: Phân cụm khách hàng, phân cụm sản phẩm, phân cụm dữ liệu, v.v.

- Xử lý ngôn ngữ tự nhiên: Phân tích cảm xúc, dịch tự động, nhận dạng ngôn ngữ, v.v.

Môn nhập môn học máy là một môn học quan trọng đối với sinh viên theo học các ngành liên quan đến trí tuệ nhân tạo. Môn học này cung cấp cho sinh viên kiến thức và kỹ năng cần thiết để áp dụng học máy để giải quyết các vấn đề thực tế.

1.2 Đề tài nghiên cứu

Đưa ra một bài toán dự đoán có thể giải quyết bằng học máy (machine leanring) với các yêu cầu sau:

- Số Feature/Attribute gồm nhiều kiểu: categorical và numerical;

- Dữ liệu phải chưa được học, thực tập trên lớp và trong bài tập về nhà;

1. Phân tích thống kê trên dữ liệu, vẽ các đồ thị để hiểu bài toán, hiểu dữ liệu. Tìm hiểu các đặc trưng và đánh gía vai trò của các đặc trưng đối với mục tiêu bài toán.
2. Ứng dụng các mô hình học máy cơ bản để giải quyết bài toán, bao gồm cả các mô hình thuộc Ensemble Learing.
3. Sử dụng Feed Forward Neural Network và Reccurent Neural Network (hoặc mô thuộc loại này) để giải quyết bài toán.
4. Áp dụng các kỹ thuật tránh Overfiting trên các mô hình của câu (2) và câu (3) để giải quyết bài toán.
5. Sau khi huấn luyện xong mô hình thì muốn cải thiện độ chính xác, ta sẽ làm gì để giải quyết nó? Phân tích các trường hợp sai, đề ra giải pháp và thực hiện nó, sau đó đánh giá xem có cải tiến so với trước không.

CHƯƠNG 2: ĐỀ TÀI BÁO CÁO

**Câu hỏi 1:** Phân tích thống kê trên dữ liệu, vẽ các đồ thị để hiểu bài toán, hiểu dữ liệu. Tìm hiểu các đặc trưng và đánh gía vai trò của các đặc trưng đối với mục tiêu bài toán

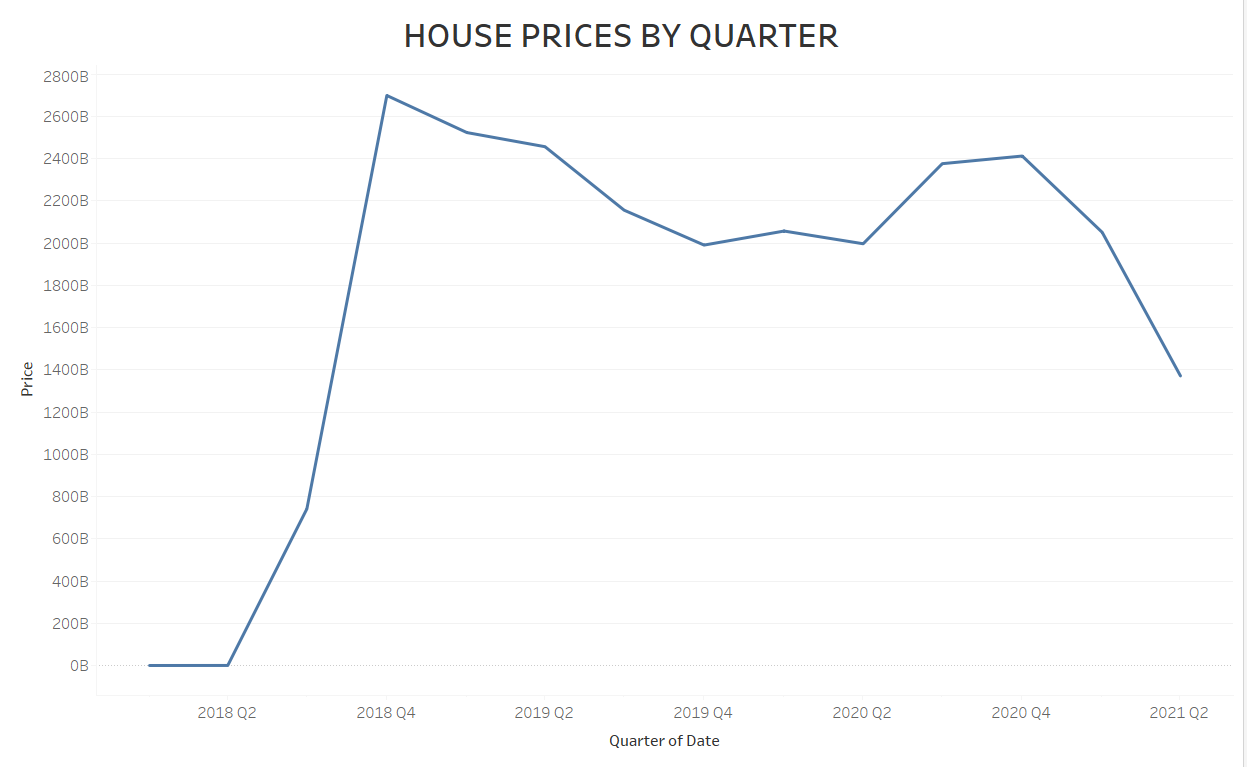
**Trả lời:**

Bộ dữ liệu bao gồm danh sách các đối tượng duy nhất của các cổng thông tin phổ biến để bán bất động sản ở Nga. Hơn 540 nghìn đối tượng. Bộ dữ liệu chứa 540000 đối tượng bất động sản ở Nga. Thị trường bất động sản Nga có lịch sử tương đối ngắn. Vào thời Xô Viết, mọi tài sản đều thuộc sở hữu nhà nước; người dân chỉ có quyền sử dụng chúng với những căn hộ được phân bổ dựa trên nơi làm việc của mỗi người. Kết quả là, các lựa chọn di chuyển khá hạn chế. Tuy nhiên, sau khi Liên Xô sụp đổ, thị trường bất động sản ở Nga nổi lên và người Muscovite có thể tư nhân hóa và sau đó lần đầu tiên bán và mua bất động sản. Ngày nay, bất động sản ở Nga đang bùng nổ. Nó mang lại nhiều cơ hội thú vị và lợi nhuận cao cho lối sống và đầu tư. Thị trường bất động sản đã ở giai đoạn tăng trưởng trong vài năm, điều đó có nghĩa là bạn vẫn có thể tìm thấy những bất động sản ở mức giá rất hấp dẫn nhưng có cơ hội tăng giá trị trong tương lai.

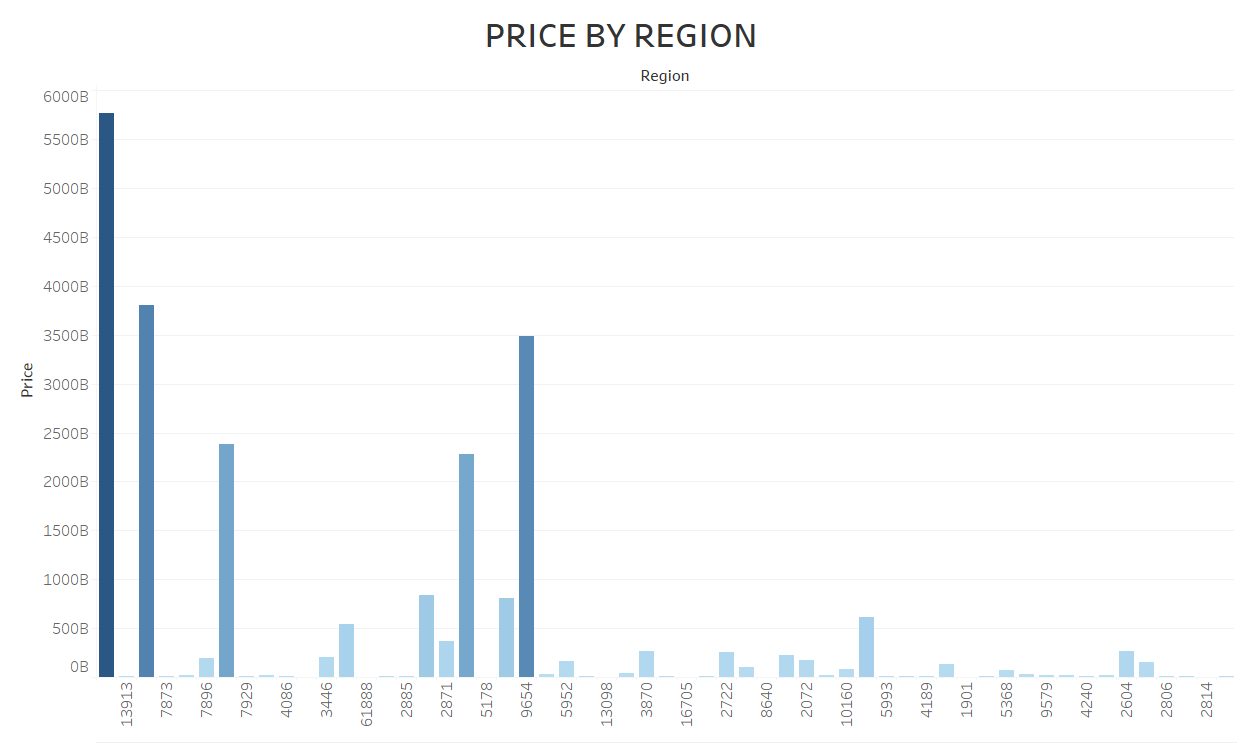
Tập dữ liệu có 13 trường dữ liệu:

|  |  |
| --- | --- |
| price | Giá cả |
| date | ngày công bố thông báo |
| time | thời điểm quảng cáo được xuất bản |
| geo\_lat | Vĩ độ |
| geo\_lon | Kinh độ |
| region | Khu vực của Nga |
| building\_type | Loại mặt tiền |
| level | Tầng căn hộ |
| levels | Số tầng |
| rooms | số lượng phòng khách |
| area | tổng diện tích căn hộ |
| kitchen\_area | Khu vực bếp |
| object\_type | Loại căn hộ |

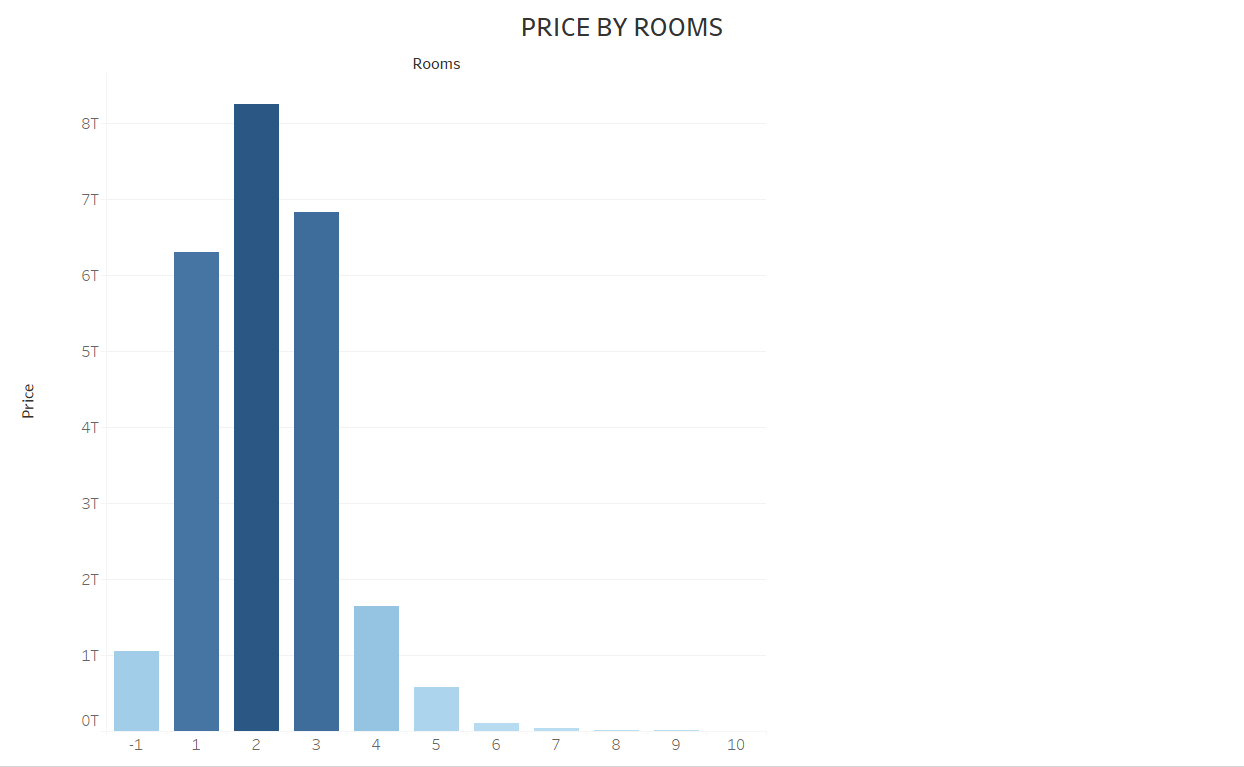
Bằng cách sử dụng tập dữ liệu này, chúng tôi cung cấp các thuật toán sử dụng nhiều chức năng khác nhau để dự đoán giá bất động sản. Các đối thủ cạnh tranh sẽ dựa vào một tập dữ liệu khổng lồ bao gồm dữ liệu nhà ở và các mô hình kinh tế vĩ mô. Một mô hình dự báo chính xác mang lại niềm tin cao hơn cho khách hàng trong nền kinh tế đầy biến động.



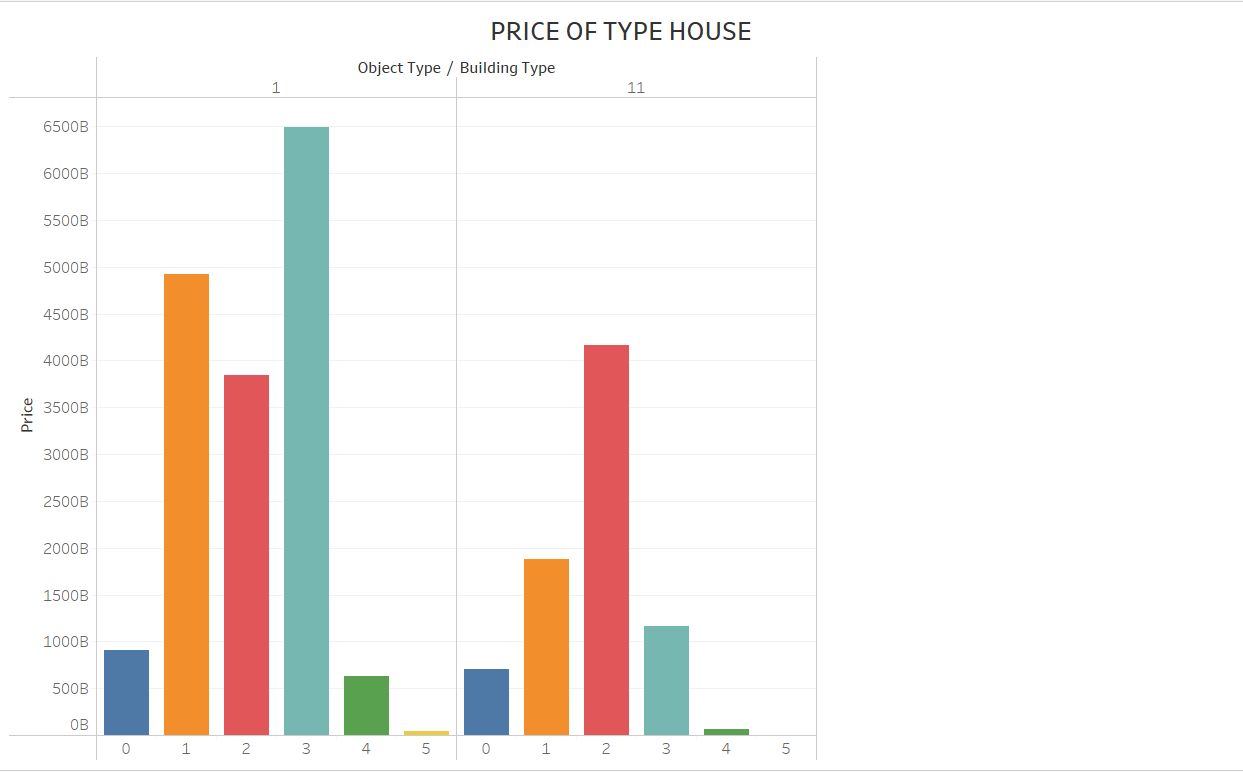
Hình 1: Đồ thị biểu diễn giá nhà theo quý mỗi năm



Hình 2: Đồ thị biểu diễn giá nhà theo khu vực



Hình 3: Đồ thị biểu diễn giá nhà theo phòng



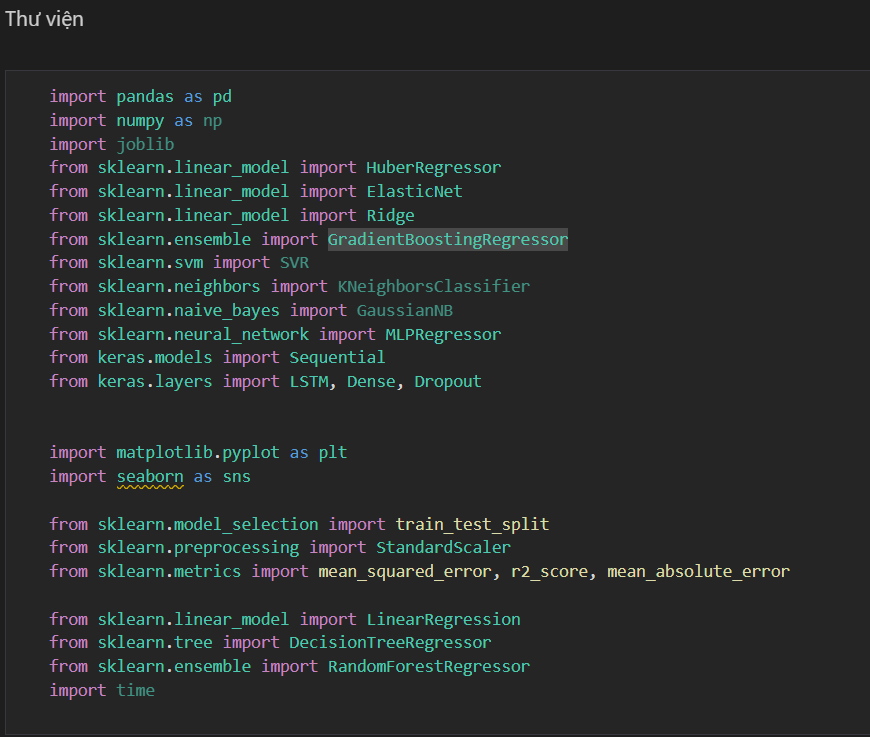
Hình 4: Đồ thị biểu diễn giá nhà theo loại nhà

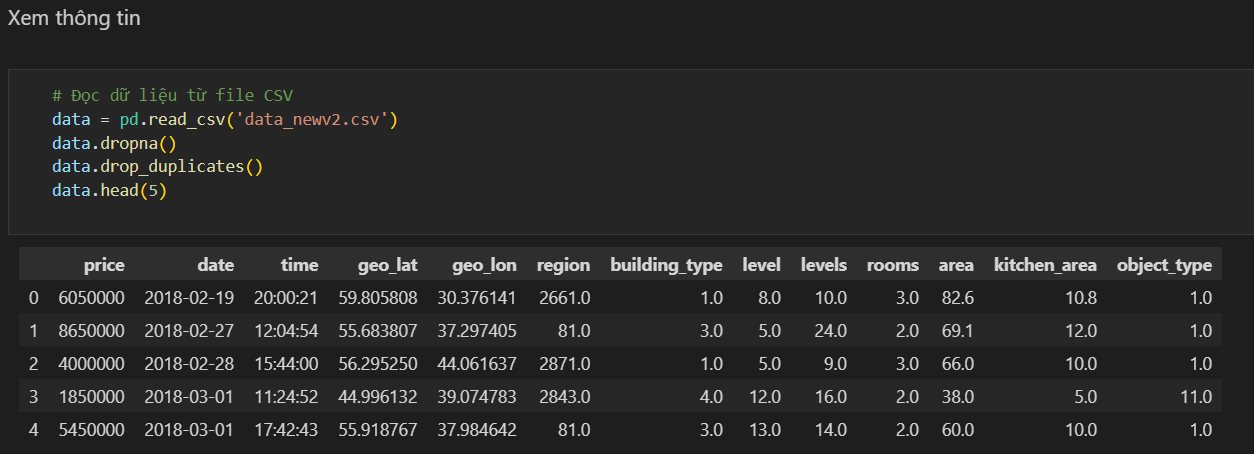
**Câu hỏi 2:** Ứng dụng các mô hình học máy cơ bản để giải quyết bài toán, bao gồm cả các mô hình thuộc Ensemble Learing.

**Câu hỏi 3:** Sử dụng Feed Forward Neural Network và Reccurent Neural Network (hoặc mô thuộc loại này) để giải quyết bài toán.

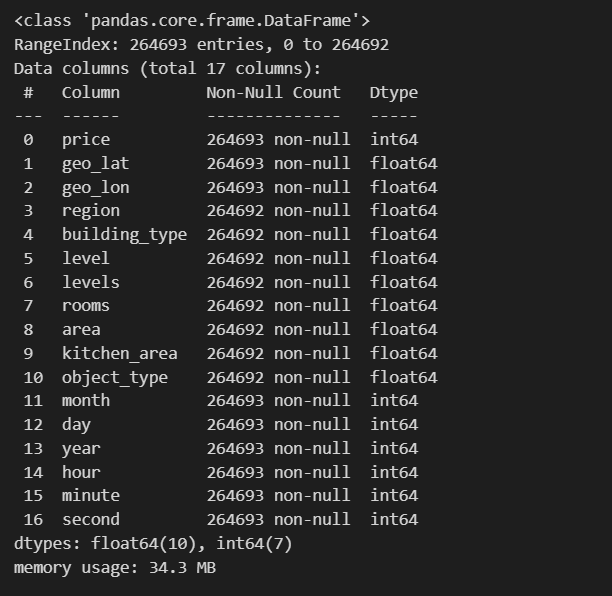
**Câu hỏi 4:** Áp dụng các kỹ thuật tránh Overfiting trên các mô hình của câu (2) và câu (3) để giải quyết bài toán.

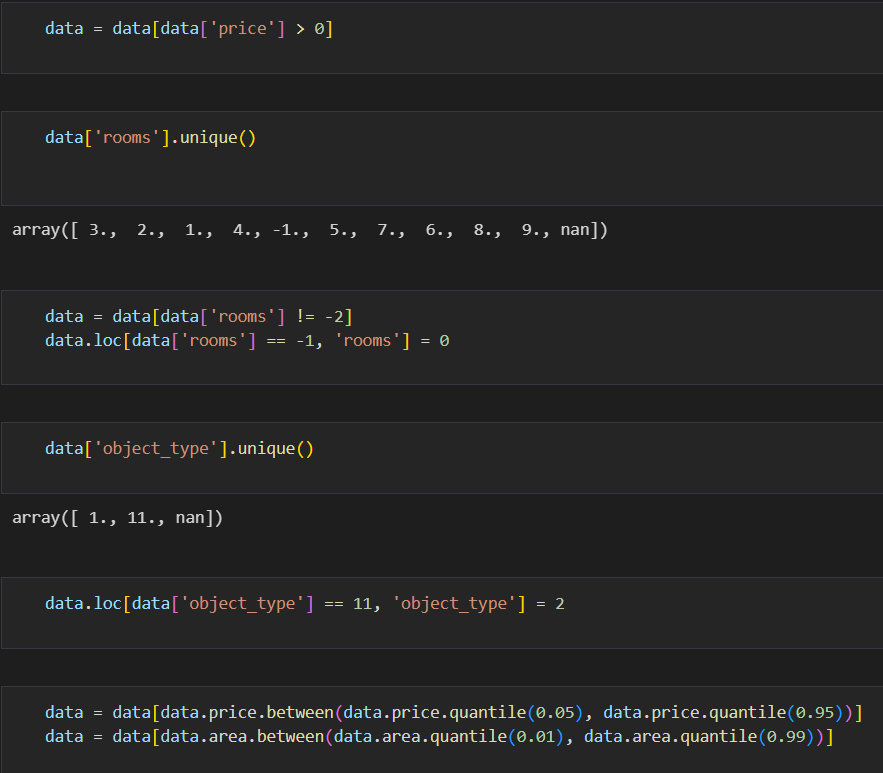
**Trả lời:**

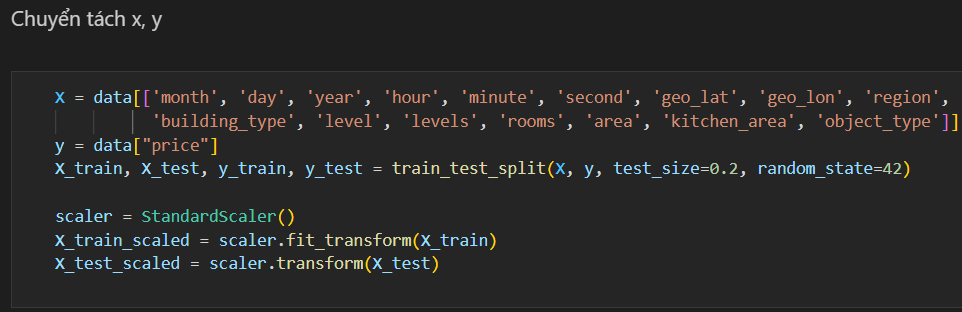


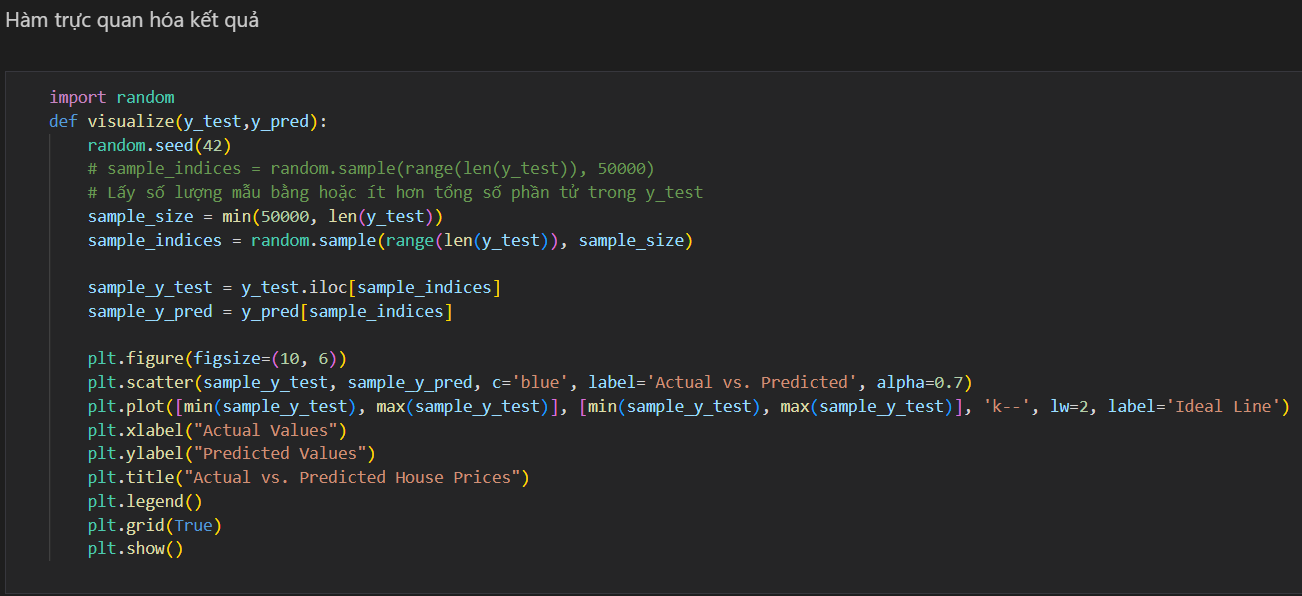


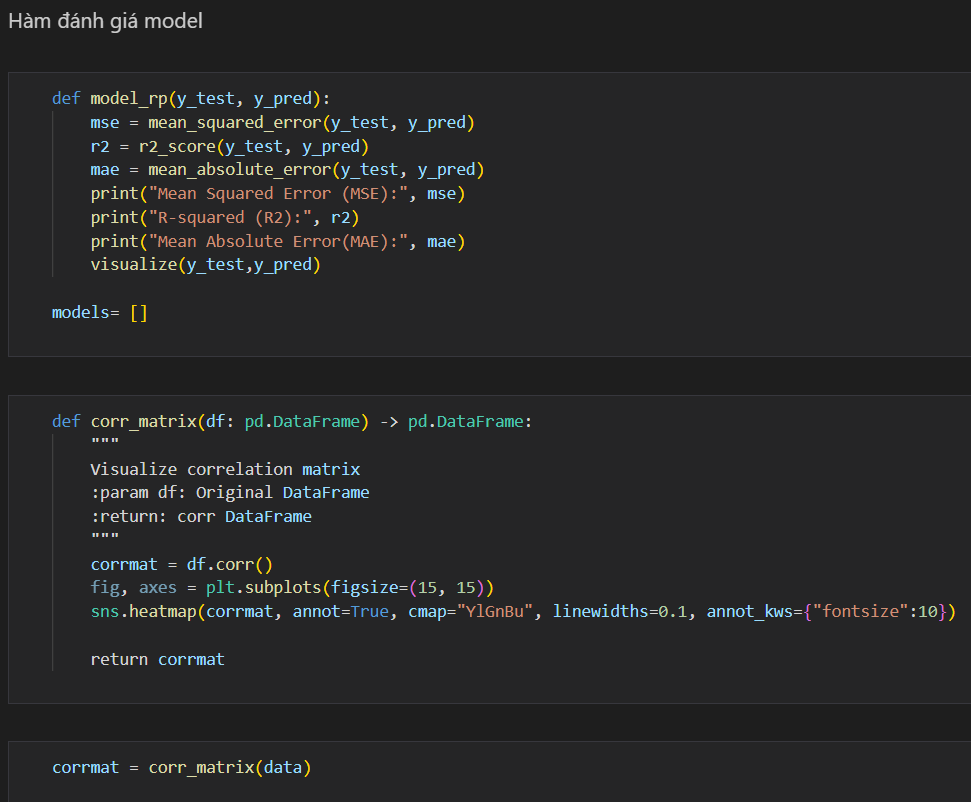


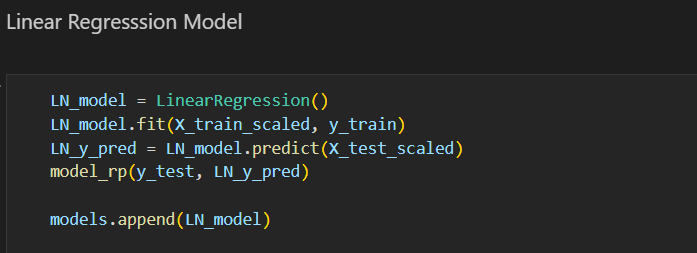


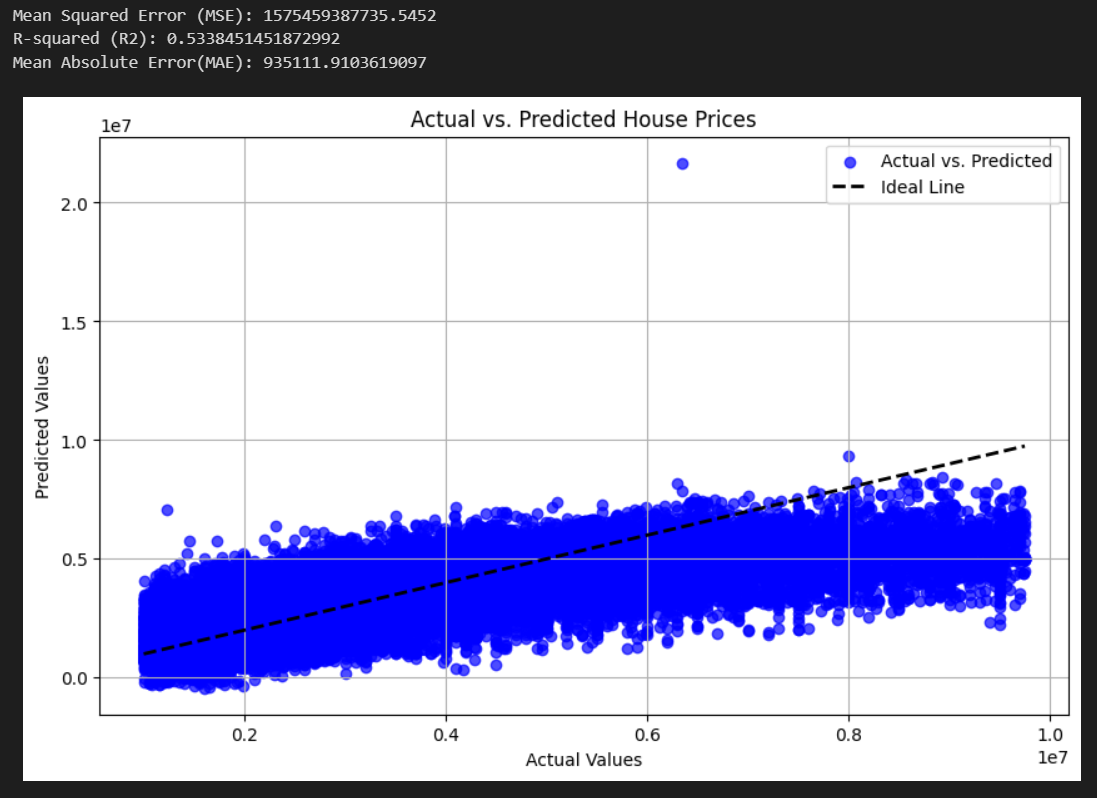


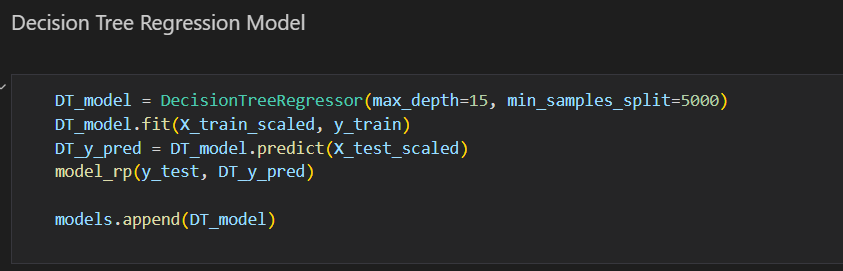


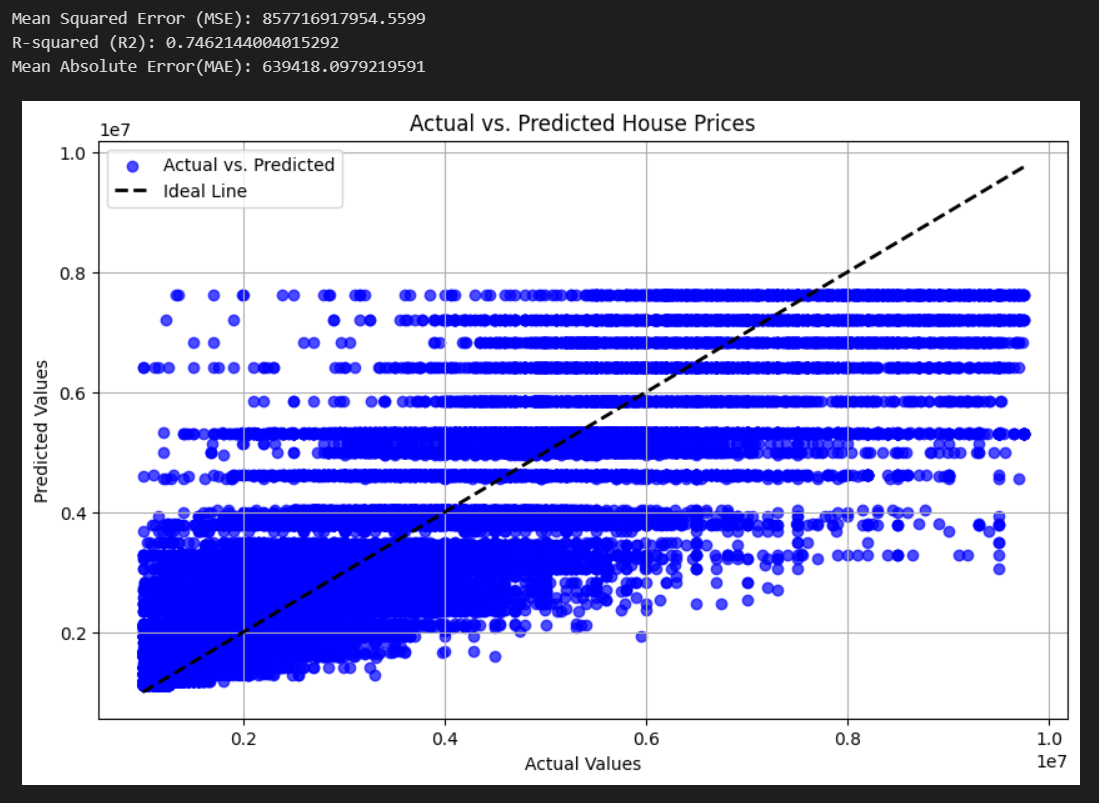


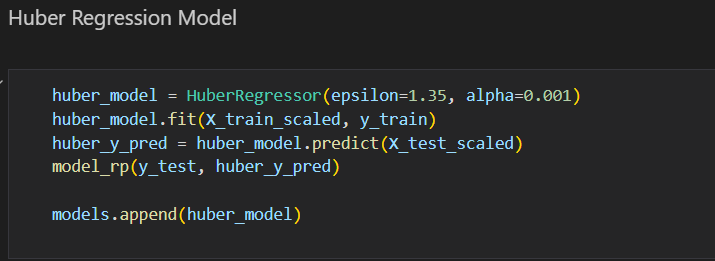


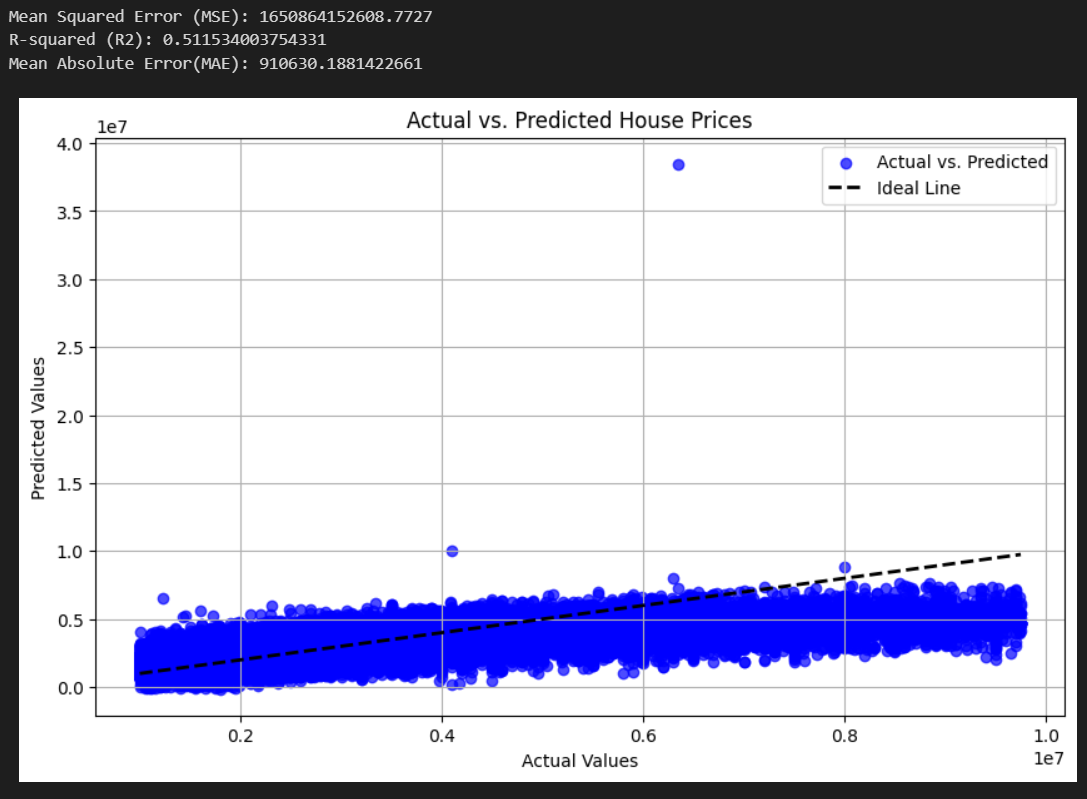


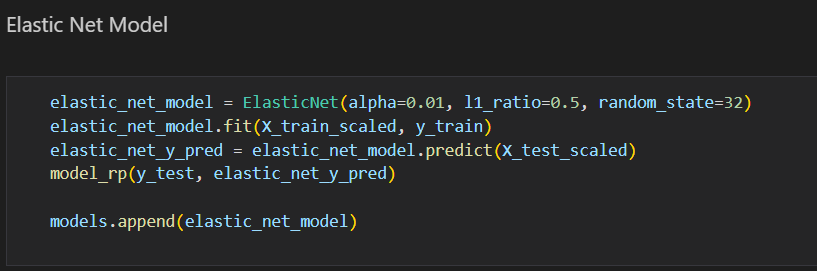


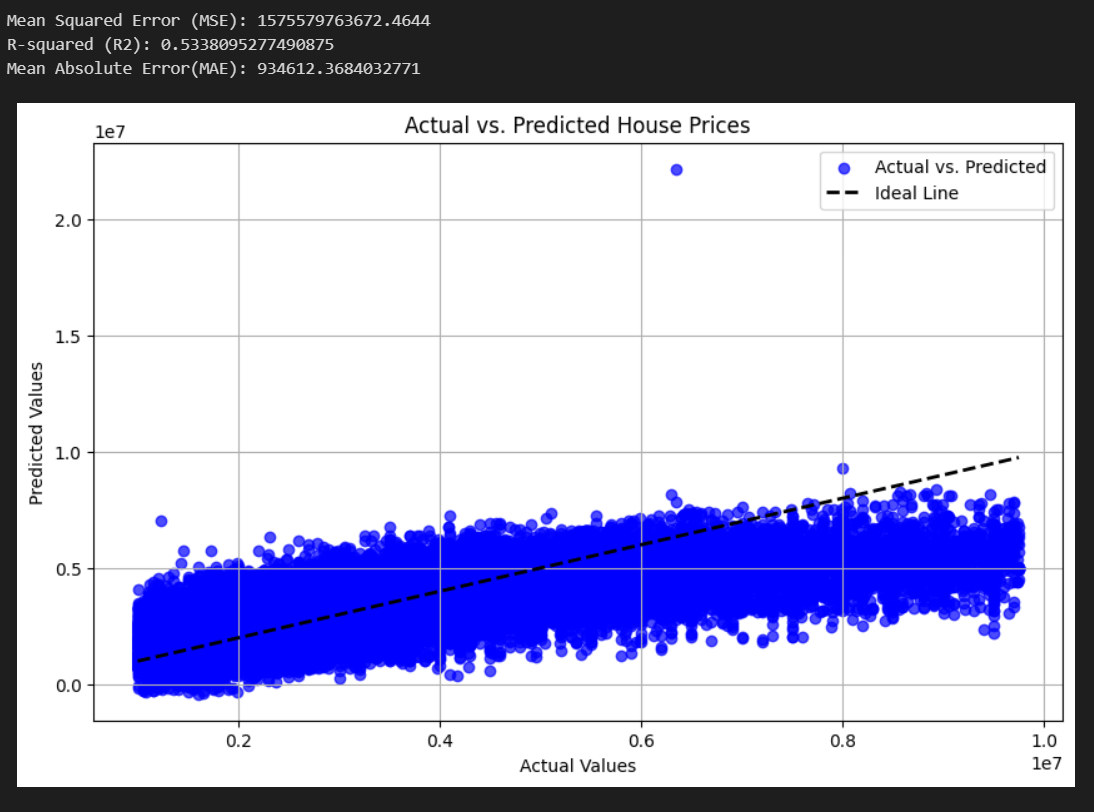


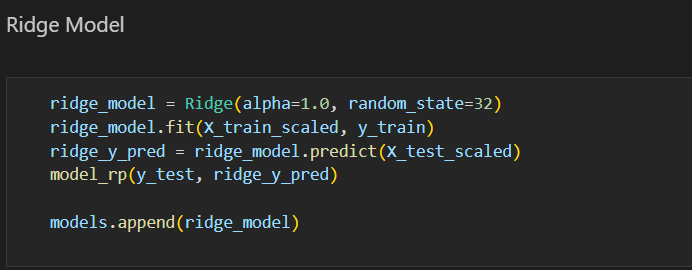


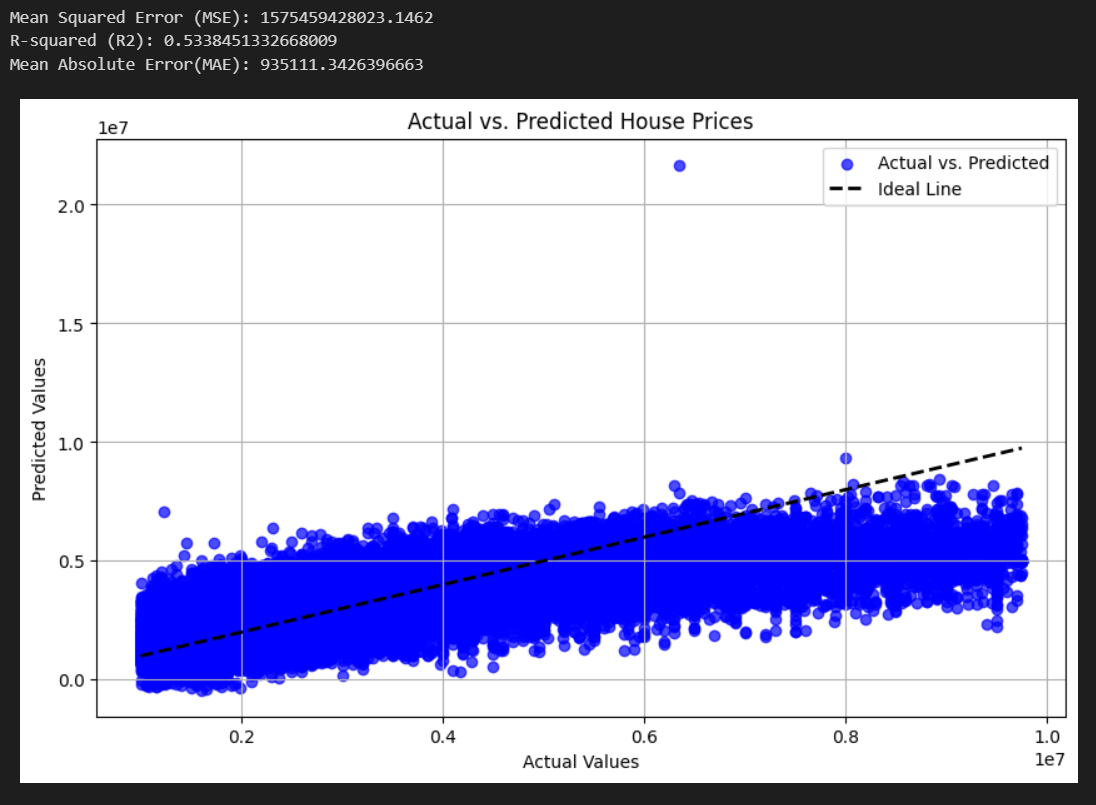


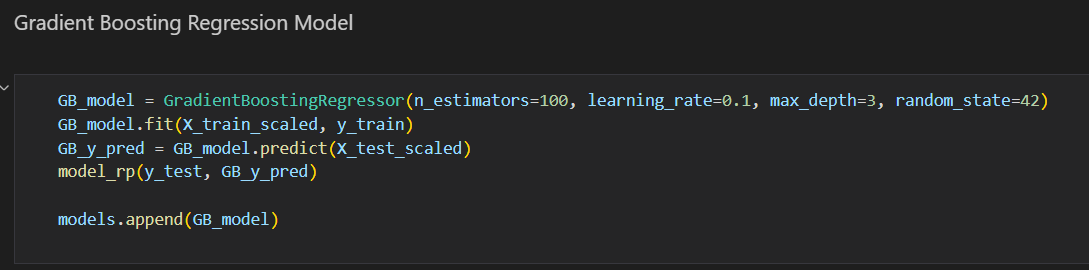


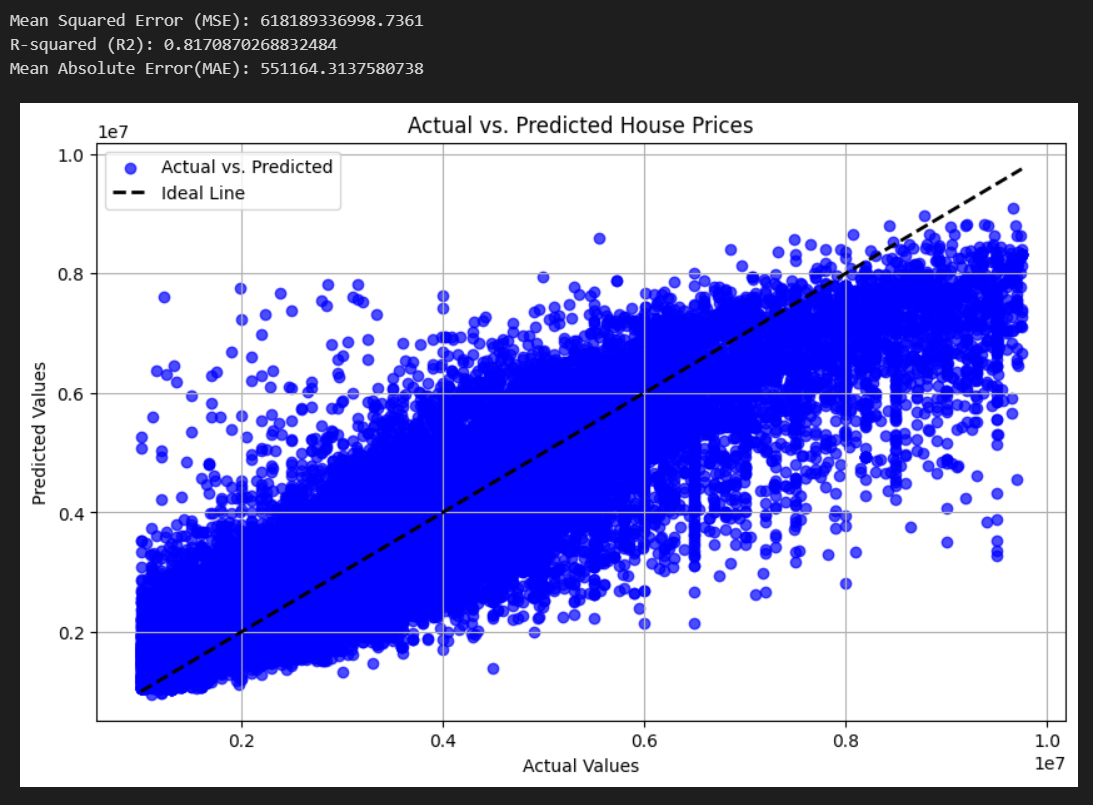


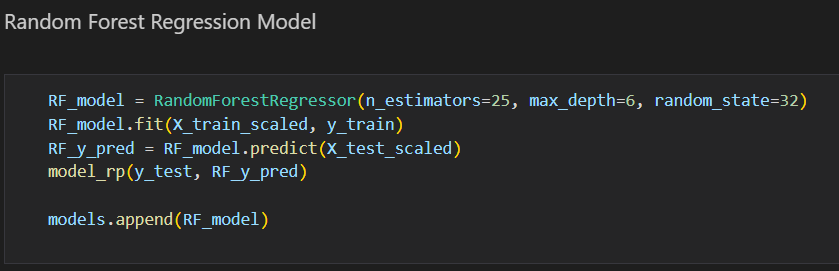


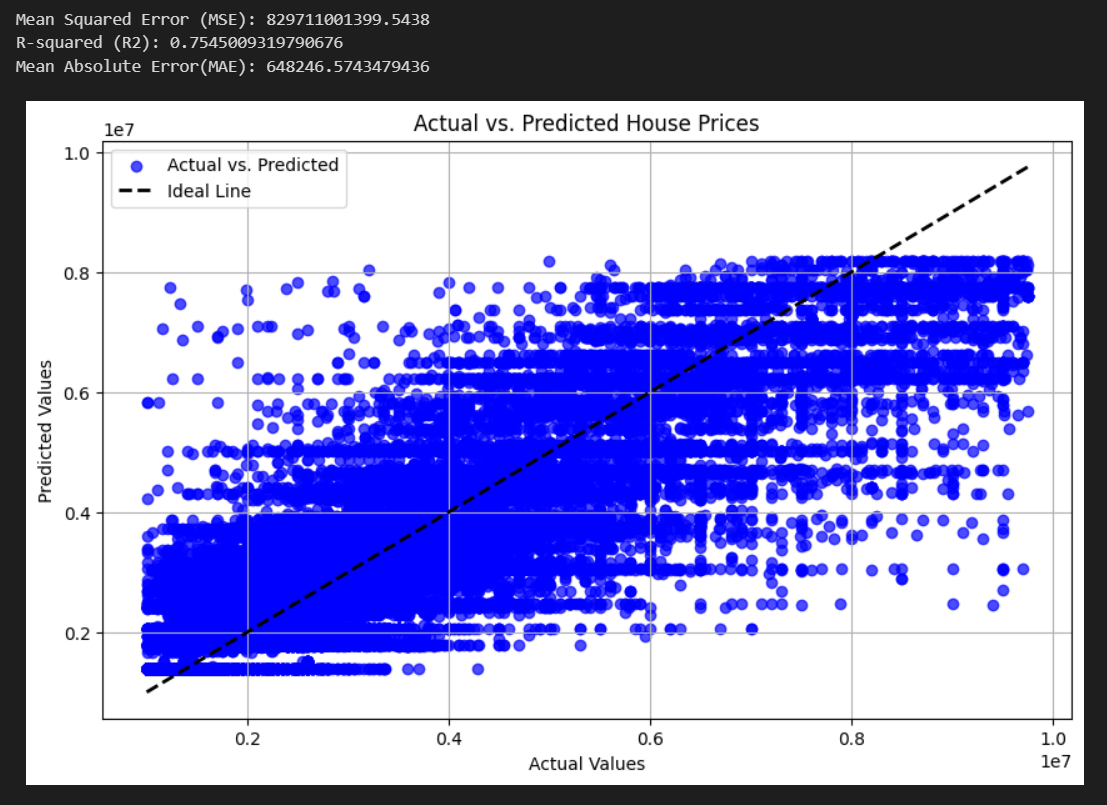


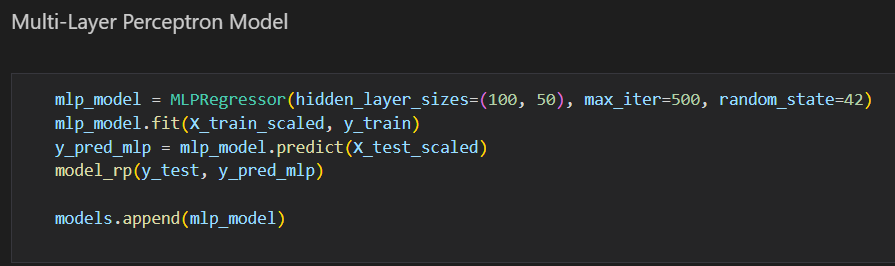


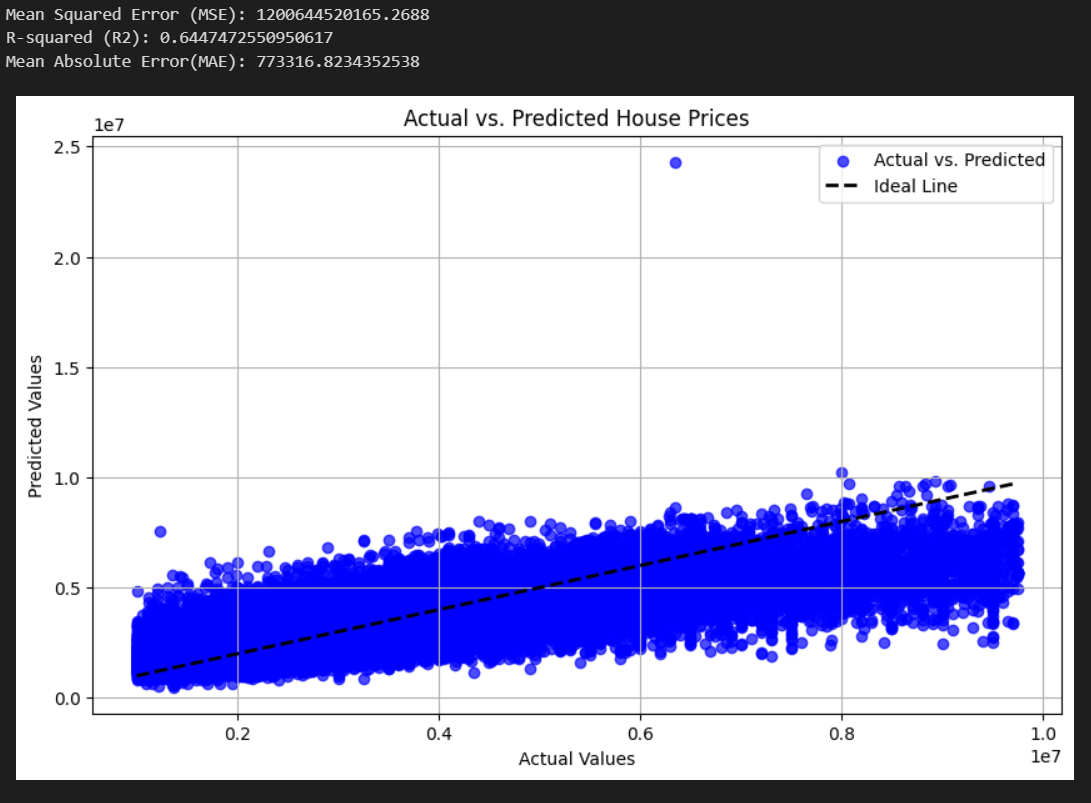


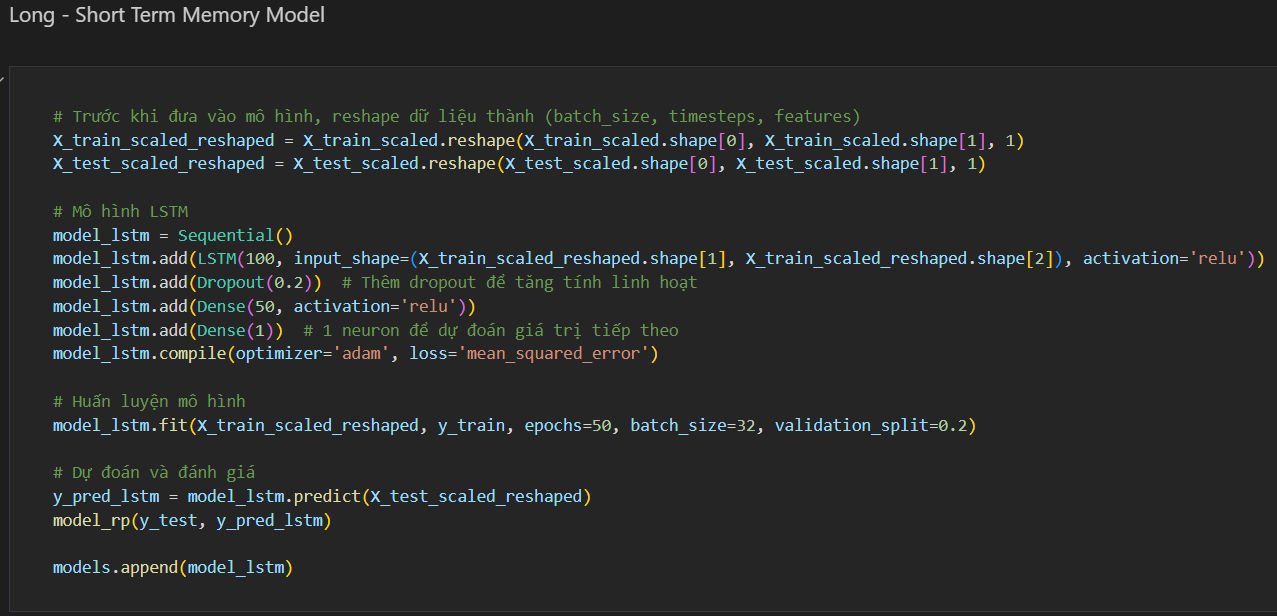


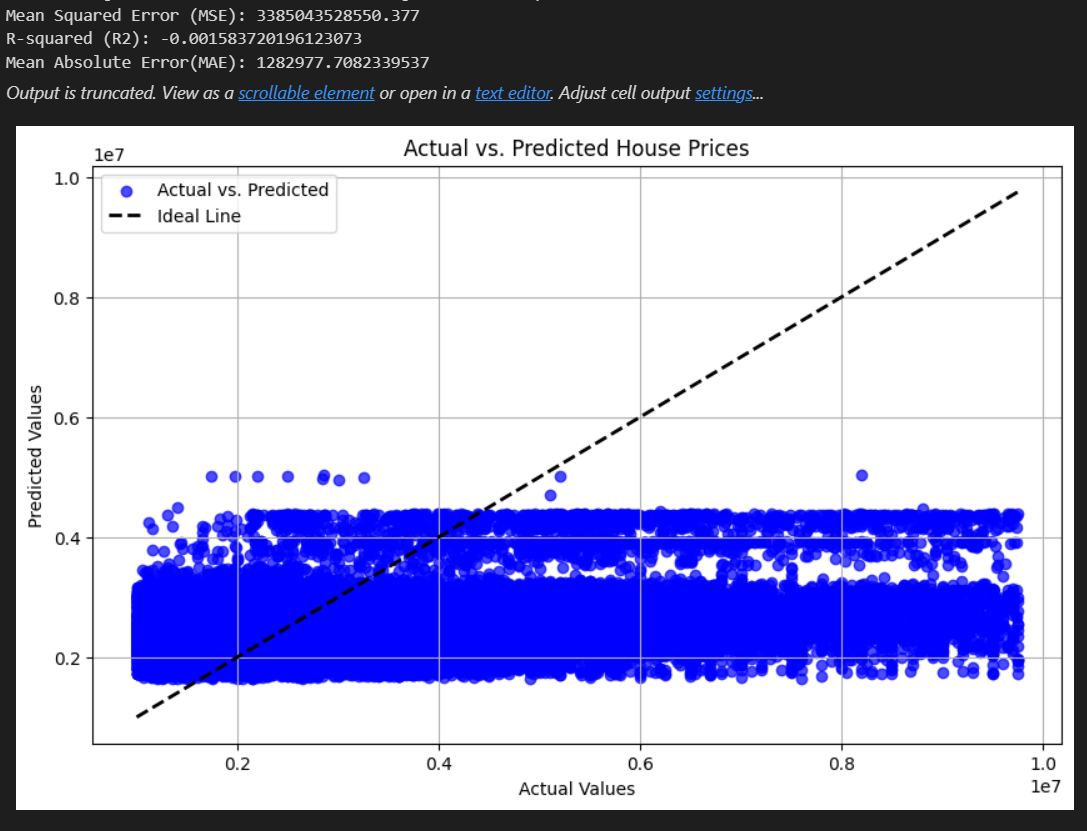


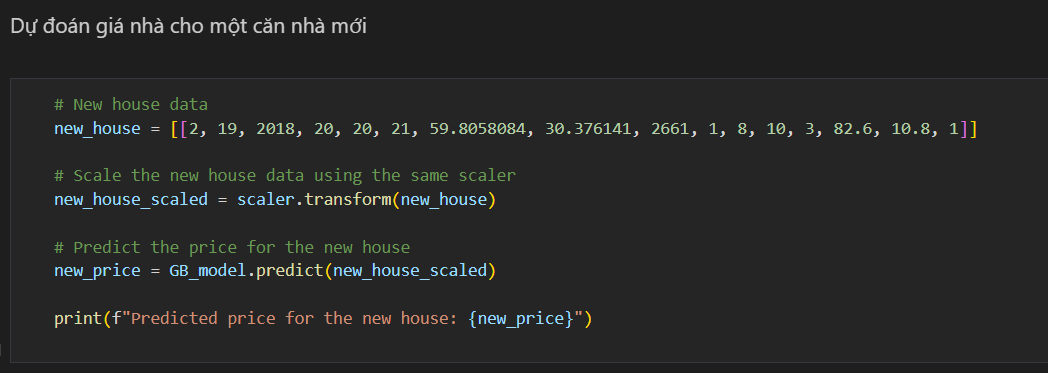












**Câu hỏi 5:** Sau khi huấn luyện xong mô hình thì muốn cải thiện độ chính xác, ta sẽ làm gì để giải quyết nó? Phân tích các trường hợp sai, đề ra giải pháp và thực hiện nó, sau đó đánh giá xem có cải tiến so với trước không.

**Trả lời:**

Sau khi huấn luyện xong mô hình trong machine learning và bạn muốn cải thiện độ chính xác, có một số phương pháp bạn có thể thực hiện để giải quyết vấn đề này. Dưới đây là một số gợi ý:

Tăng cường dữ liệu (Data Augmentation): Nếu bạn có một lượng nhỏ dữ liệu huấn luyện, bạn có thể tạo thêm dữ liệu bằng cách biến đổi, xoay, phóng to hay thu nhỏ ảnh, từ đó tạo thêm các mẫu huấn luyện mới.

Tối ưu hóa tham số (Hyperparameter Tuning): Thử nghiệm và tối ưu hóa các tham số của mô hình như learning rate, số lượng lớp, số lượng nơ-ron, hàm kích hoạt, v.v. Thay đổi các tham số này có thể ảnh hưởng lớn đến hiệu suất của mô hình.

Kiểm tra và chuẩn hóa dữ liệu (Data Inspection and Normalization): Đảm bảo dữ liệu của bạn đã được kiểm tra và chuẩn hóa đúng cách. Loại bỏ các giá trị ngoại lệ (outliers) và đảm bảo rằng dữ liệu đầu vào đang ở định dạng phù hợp.

Sử dụng mô hình phức tạp hơn hoặc thay đổi kiến trúc mô hình: Nếu mô hình hiện tại của bạn quá đơn giản, bạn có thể xem xét việc sử dụng mô hình phức tạp hơn như mạng nơ-ron sâu (deep neural network) hoặc thay đổi kiến trúc của mô hình để nó có thể học được đặc trưng phức tạp hơn từ dữ liệu.

Thử nghiệm các mô hình khác nhau: Thử nghiệm với các mô hình khác nhau để xem xét xem mô hình nào hoạt động tốt nhất cho dữ liệu của bạn. Đôi khi, việc chọn mô hình phù hợp có thể làm tăng độ chính xác.

Regularization: Sử dụng các kỹ thuật như dropout, L1 hoặc L2 regularization để giảm overfitting và tăng khả năng tổng quát hóa của mô hình.

Kiểm tra và đánh giá hiệu suất: Sử dụng các kỹ thuật kiểm tra và đánh giá hiệu suất mô hình như cross-validation để đảm bảo rằng mô hình không chỉ hoạt động tốt trên dữ liệu huấn luyện mà còn trên dữ liệu kiểm thử.

Ensemble Learning: Kết hợp nhiều mô hình khác nhau có thể tạo ra một mô hình mạnh mẽ hơn bằng cách sử dụng các kỹ thuật như bagging hoặc boosting.