ĐẠI HỌC QUỐC GIA TP. HCM

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO TỔNG KẾT

ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ SINH VIÊN NĂM 2025

Tên đề tài tiếng Việt:

Nghiên cứu tối ưu mô hình kết hợp ứng dụng vào giải quyết bài toán dự báo giá nhà tại Hà Nội

Tên đề tài tiếng Anh:

Research to optimize the combined model and apply in solving the problem of forecasting house prices in Ha Noi

Khoa/ Bộ môn: Hệ thống thông tin

Thời gian thực hiện: 6 tháng

Cán bộ hướng dẫn: TS. Nguyễn Thanh Bình

Tham gia thực hiện

TT	Họ và tên, MSSV	Chịu trách nhiệm	Điện thoại	Email
1	Nguyễn An Đức 22520268	Chủ nhiệm	0934894238	22520268@gm.uit.edu.vn
2	Hà Nhật Thái 22521316	Tham gia	0388874855	22521316@gm.uit.edu.vn

Thành phố Hồ Chí Minh – Tháng 05/2025

Ngày nhận hồ sơ	
Mã số đề tài	
(Do CQ quản lý ghi)	

ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO TỔNG KẾT

Tên đề tài tiếng Việt:

Nghiên cứu tối ưu mô hình kết hợp ứng dụng vào giải quyết bài toán dự báo giá nhà tại Hà Nội

Tên đề tài tiếng Anh:

Research to optimize the combined model and apply in solving the problem of forecasting house prices in Ha Noi

Ngày 26 tháng 05 năm 2025	Ngày 26 tháng 05 năm 2025
Cán bộ hướng dẫn	Sinh viên chủ nhiệm đề tài
(Họ tên và chữ ký)	(Họ tên và chữ ký)
	Nguyễn An Đức

THÔNG TIN KẾT QUẢ NGHIÊN CỨU

1. Thông tin chung:

- Tên đề tài: Nghiên cứu tối ưu mô hình kết hợp ứng dụng vào giải quyết bài toán dự báo giá nhà tại Hà Nội
- Chủ nhiệm: Nguyễn An Đức
- Thành viên tham gia: Hà Nhật Thái
- Cơ quan chủ trì: Trường Đại học Công nghệ Thông tin.
- Thời gian thực hiện: 6 tháng

2. Mục tiêu:

- Đưa ra mô hình kết hợp có độ chính xác và tin cậy cao nhất và áp dụng mô hình kết hợp để cung cấp dự đoán chính xác và đáng tin cậy về giá bất động sản dựa trên các dữ liệu kết hợp đa dạng.
- Ứng dụng mô hình kết hợp đề xuất để thực hiện dự đoán giá nhà với bộ dữ liệu tại Hà Nội.

3. Tính mới và sáng tạo: Tối ưu mô hình kết hợp (ensemble) nhằm nâng cao độ chính xác trong dự báo giá nhà tại Hà Nội, thông qua việc khai thác thế mạnh của từng thuật toán học máy.

4. Tóm tắt kết quả nghiên cứu:

- Xây dựng thành công bộ dữ liệu tại Hà Nội cho bài toán.
- Áp dụng mô hình kết hợp cho dự báo giá nhà, khẳng định tiềm năng vượt trội của việc kết hợp các mô hình học máy tiên tiến trong lĩnh vực dự báo giá bất động sản.
- Tài liệu báo cáo về nghiên cứu tối ưu mô hình kết hợp ứng dụng vào giải quyết bài toán dự báo giá nhà tại Hà Nội.
- Ứng dụng dự báo giá nhà cung cấp công cụ hữu ích cho các bên liên quan trong thị trường bất động sản, bao gồm người mua, người bán, nhà đầu tư và các nhà hoạch định

chính sách, hỗ trợ quá trình ra quyết định trong các hoạt động mua bán, đầu tư bất động sản, góp phần nâng cao tính minh bạch và hiệu quả của thị trường.

- Đóng gói mô hình thành một website hoàn chỉnh có thể triển khai nhanh chóng.

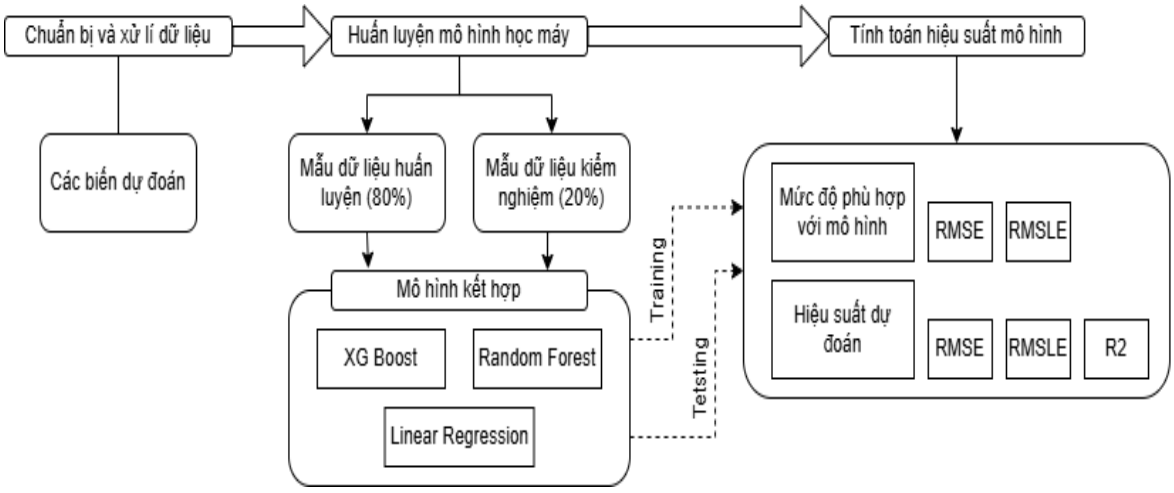
5. Tên sản phẩm: Hệ thống dự báo giá nhà tại Hà Nội sử dụng mô hình kết hợp tối ưu

6. Hiệu quả, phương thức chuyển giao kết quả nghiên cứu và khả năng áp dụng:

- Nghiên cứu triển khai kiến trúc ensemble, phối hợp các thuật toán học máy tiên tiến để khai thác đồng thời đặc trưng của thị trường bất động sản Hà Nội. Nhờ cơ chế kết hợp có trọng số và quy trình tối ưu siêu tham số tự động (Bayesian Optimization), hệ thống đạt mức cải thiện sai số RMSE và RMSLE so với các mô hình đơn lẻ. Đồng thời bảo đảm tính minh bạch, giúp người sử dụng hiểu rõ mức độ đóng góp của từng yếu tố vị trí, diện tích,... trong quá trình định giá, qua đó củng cố độ tin cậy khoa học của kết quả dự báo.

- Phương thức chuyển giao được thiết kế linh hoạt nhằm tối đa hóa khả năng ứng dụng. Toàn bộ mã nguồn, pipeline huấn luyện và trọng số mô hình được đóng gói kèm hướng dẫn triển khai chi tiết cho phép các cá nhân, tổ chức nghiên cứu, tái huấn luyện hoặc mở rộng tính năng.

7. Hình ảnh, sơ đồ minh họa chính :



Cơ quan chủ trì	Chủ nhiệm đề tài
(Ký, họ và tên, đóng dấu)	(Ký, họ và tên)
	Nguyễn An Đức

LỜI CẢM ƠN

Lời đầu tiên, nhóm nghiên cứu xin gửi lời cảm ơn sâu sắc tới Phòng Thí nghiệm Khoa Hệ thống Thông tin đã tạo điều kiện thuận lợi để chúng tôi triển khai và hoàn thành đề tài này.

Nhóm cũng trân trọng cảm ơn Thầy Nguyễn Thanh Bình, người đã tận tâm hướng dẫn, định hướng và động viên chúng tôi trong suốt quá trình thực hiện nghiên cứu. Sự nhiệt huyết cùng kiến thức chuyên môn của Thầy là nguồn cảm hứng giúp chúng tôi mạnh dạn tìm tòi, học hỏi những điều mới, với mục tiêu quan trọng nhất là phát triển bản thân.

Bên cạnh đó, tập thể quý Thầy Cô đã không ngừng chia sẻ tri thức, xây dựng nền tảng vững chắc và mở ra nhiều cơ hội thử thách, phát triển. Nhờ những đóng góp quý báu ấy, chúng tôi có đủ tự tin để khởi hành trên những hành trình mới trong tương lai.

Dù đã nỗ lực hết mình, bài báo cáo chắc chắn vẫn còn những thiếu sót. Rất mong quý Thầy Cô thông cảm, đóng góp ý kiến để nhóm hoàn thiện nghiên cứu một cách tốt nhất.

Một lần nữa, xin chân thành cảm ơn quý Thầy Cô.

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

DANH MỤC CÁC TỪ VIẾT TẮT

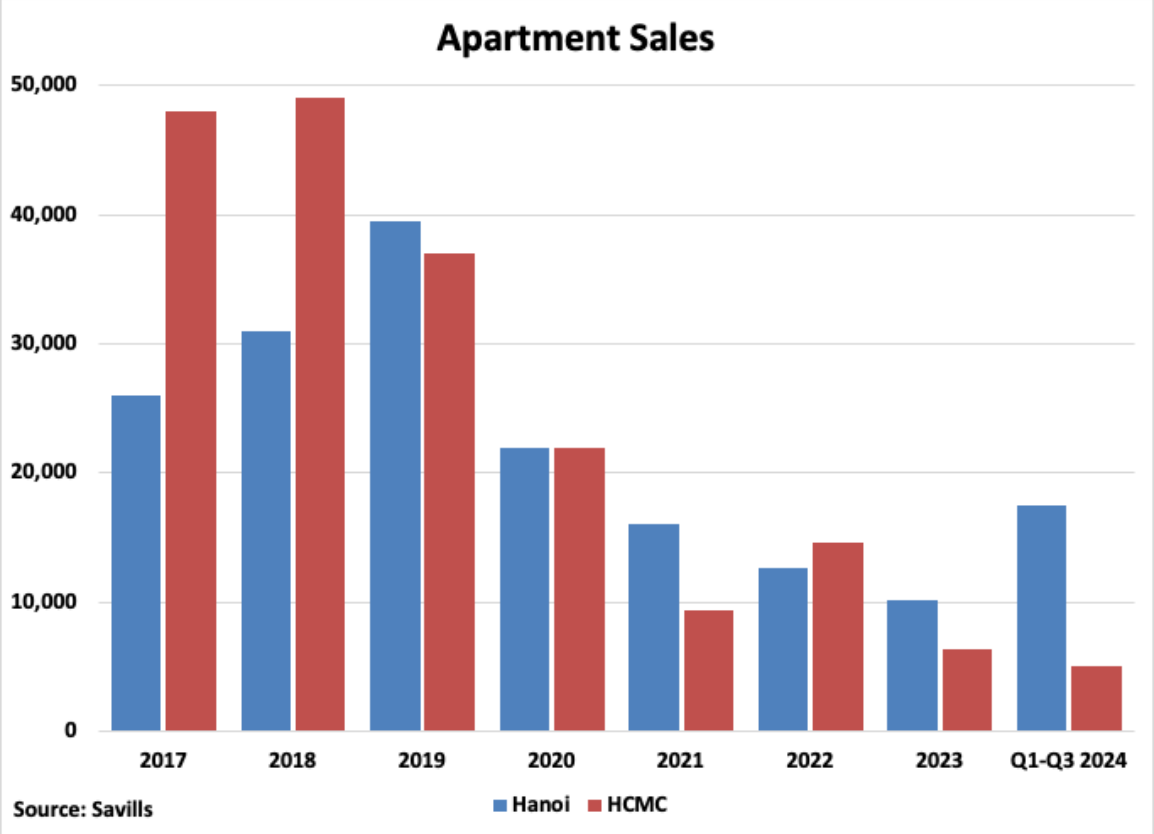
STT	Từ viết tắt	Ý nghĩa
1	XGBoost	Extreme Gradient Boosting
2	RMSE	Root Mean Squared Error
3	RMSLE	Root Mean Squared Logarithmic Error
4	LR	Linear Regression
5	RF	Random Forest
6	ANN	Artificial Neural Network
7	DNN	Deep Neural Network
8	LightGBM	Light Gradient Boosting Machine
9	MAPE	Mean Absolute Percentage Error
10	KNN	K-Nearest Neighbors
11	SVM	Support Vector Machine
12	MSE	Mean Squared Error
13	R^2	Coefficient of Determination

TÓM TẮT

Trong bối cảnh thị trường bất động sản Hà Nội ngày càng sôi động và phức tạp, việc dự báo giá nhà không chỉ là một bài toán kinh tế thuần túy mà còn là chìa khóa để hoạch định chính sách đô thị, tối ưu hóa danh mục đầu tư và hỗ trợ người dân đưa ra quyết định tài chính sáng suốt. Nghiên cứu này khai thác kho dữ liệu lớn từ Batdongsan.com – một trong những sàn giao dịch trực tuyến uy tín nhất Việt Nam – nhằm xây dựng mô hình dự báo giá nhà dựa trên chiến lược học tập tổ hợp (Ensemble Learning). Cụ thể, chúng tôi phát triển một mô hình xếp chồng (Stacking) kết hợp ba bộ học cơ sở có thế mạnh bổ trợ cho nhau: hồi quy tuyến tính (Linear Regression) đại diện cho mối quan hệ tuyến tính nền tảng; rừng ngẫu nhiên (Random Forest) với khả năng nắm bắt quan hệ phi tuyến và giảm phương sai; và Extreme Gradient Boosting (XGBoost) – thuật toán tăng cường cây quyết định tối ưu hóa tốc độ hội tụ và độ chính xác. Tập biến đầu vào được thiết kế toàn diện, bao quát cả yếu tố vật lý lẫn biến kinh tế – xã hội. Trước khi huấn luyện, dữ liệu được làm sạch, mã hóa, chuẩn hóa và chia tách nhằm bảo đảm khả năng tổng quát hóa. Hiệu quả của mô hình tổ hợp được định lượng bằng ba thước đo then chốt: Root Mean Squared Error (RMSE) nhấn mạnh sai lệch lớn, Root Mean Squared Logarithmic Error (RMSLE) đặc biệt hữu dụng khi giá nhà phân bố lệch phải và trải rộng nhiều bậc độ lớn, và hệ số xác định (R^2) biểu thị tỷ lệ phương sai được giải thích. Việc so sánh ba chỉ số này trên cả tập huấn luyện và kiểm thử giúp đánh giá toàn diện độ chính xác, độ ổn định và tiềm năng khái quát hóa của mô hình, từ đó mang lại cái nhìn sâu sắc hơn về cơ chế hình thành giá bất động sản tại Hà Nội. Những phát hiện này không những hỗ trợ bên mua – bán ra quyết định hiệu quả mà còn gợi ý hàm ý chính sách cho cơ quan quản lý trong việc điều tiết cung cầu và phát triển hạ tầng bền vững.

CHƯƠNG 1: MỞ ĐẦU

Trong lĩnh vực bất động sản, giá nhà đất là một yếu tố then chốt ảnh hưởng đến quyết định mua bán của người tiêu dùng và tác động trực tiếp đến nền kinh tế quốc gia. Việc dự đoán chính xác giá nhà đất không chỉ giúp người mua tìm được những ngôi nhà phù hợp mà còn hỗ trợ chính phủ trong việc điều chỉnh chính sách bất động sản một cách hợp lý. Đặc biệt, Hà Nội đang vươn lên thành hạt nhân tăng trưởng của toàn vùng Bắc Bộ. Trong quý I/2024, giá bình quân căn hộ đã đạt 2.210 USD/m², tăng 10,1 % so với cùng kỳ 2023 – tương đương 5,9 % sau điều chỉnh lạm phát. Đây là mức tăng tiếp nối sau khi giá nhà đã tăng 16,1% trong quý 1/2023 **[1]**.



Hình 1: Biểu đồ lượng giao dịch căn hộ tại Hà Nội và TP.HCM, 2017 – Q3 2024
(Nguồn: Savills).

Đà leo thang này tạo ra ba hệ quả rõ nét: thu hẹp ngân sách tiêu dùng hộ gia đình, nâng cao rào cản đối với người mua nhà lần đầu và nới rộng rủi ro cho vay thế

chấp. Sức nóng thị trường còn được bồi thêm bởi các cú hích hạ tầng như Vành đai 4, tuyến metro Nhổn – Ga Hà Nội khiến triển vọng giá ngắn hạn càng khó đoán định. . Sự biến động liên tục của thị trường bất động sản tại Hà Nội đã làm nổi bật tầm quan trọng của việc dự đoán giá nhà, khiến các mô hình dự báo trở nên cần thiết hơn bao giờ hết

Trong bối cảnh đó, nhu cầu xây dựng các mô hình dự báo giá nhà trở nên cấp thiết hơn bao giờ hết. Dự đoán giá nhà vẫn là một lĩnh vực nghiên cứu năng động và đang phát triển với những ý nghĩa thực tiễn quan trọng. Quá trình chuyển đổi từ các mô hình thống kê truyền thống sang thuật toán học máy đã đánh dấu một bước tiến đáng kể, với những nỗ lực không ngừng nhằm giải quyết các thách thức liên quan đến chất lượng dữ liệu, khả năng diễn giải mô hình và việc kết hợp thông tin không gian và thời gian. Khi nghiên cứu tiếp tục, độ chính xác và độ tin cậy của dự đoán giá nhà dự kiến sẽ được cải thiện, mang lại lợi ích cho người mua, người bán, nhà đầu tư và các nhà hoạch định chính sách. Dự báo chính xác không chỉ giúp người mua tối ưu quyết định mà còn cung cấp cơ sở khoa học để Nhà nước điều tiết tín dụng, hoạch định chính sách bất động sản và quản trị rủi ro vĩ mô. Sự bất ổn hiện tại vì thế nhấn mạnh vai trò trung tâm của dữ liệu và phân tích định lượng trong quản lý thị trường nhà ở Hà Nội.

CHƯƠNG 2: CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Ngày nay, cùng với sự phát triển mạnh mẽ của máy học và trí tuệ nhân tạo, có rất nhiều các nghiên cứu đã cho ra đời các mô hình dự đoán giá nhà và đã thu được một số kết quả khả quan. Trong đó có hai nhóm tiếp cận chính, mỗi nhóm dựa trên các phương pháp và mô hình khác nhau. Nhóm đầu tiên là các phương pháp hồi quy (Regression Methods). Nhóm này sử dụng các mô hình hồi quy tuyến tính và phi tuyến như hồi quy đa thức (Polynomial Regression), rừng ngẫu nhiên (Random Forest), hồi quy vectơ hỗ trợ (Support Vector Regression - SVR),... để dự đoán giá nhà dựa trên các dữ liệu đầu vào là thông tin thuộc tính của căn nhà như như vị trí, diện tích, và số phòng,… nhưng mô hình hồi quy giả định rằng sai số có phân phối chuẩn, với trung bình bằng 0 và phương sai không đổi. Tuy nhiên, nếu phân phối của sai số không tuân theo chuẩn này, kết quả dự đoán sẽ không chính xác **[2]**. Nhóm thứ hai là các mô hình học sâu (Deep Learning Models). Các mô hình này sử dụng mạng thần kinh (Neural Network) để dự đoán giá nhà, bao gồm Mạng thần kinh nhân tạo (Artificial Neural Network - ANN) và Mạng thần kinh sâu (Deep Neural Network - DNN). Mô hình ANN đã được thử nghiệm và cho thấy chỉ số lỗi (Root Mean Square Error - RMSE) trung bình thấp hơn so với một số mô hình hồi quy, nhưng chi phí tính toán cao và dễ gặp lỗi khi dữ liệu không đủ lớn hoặc không đồng đều **[3]**.

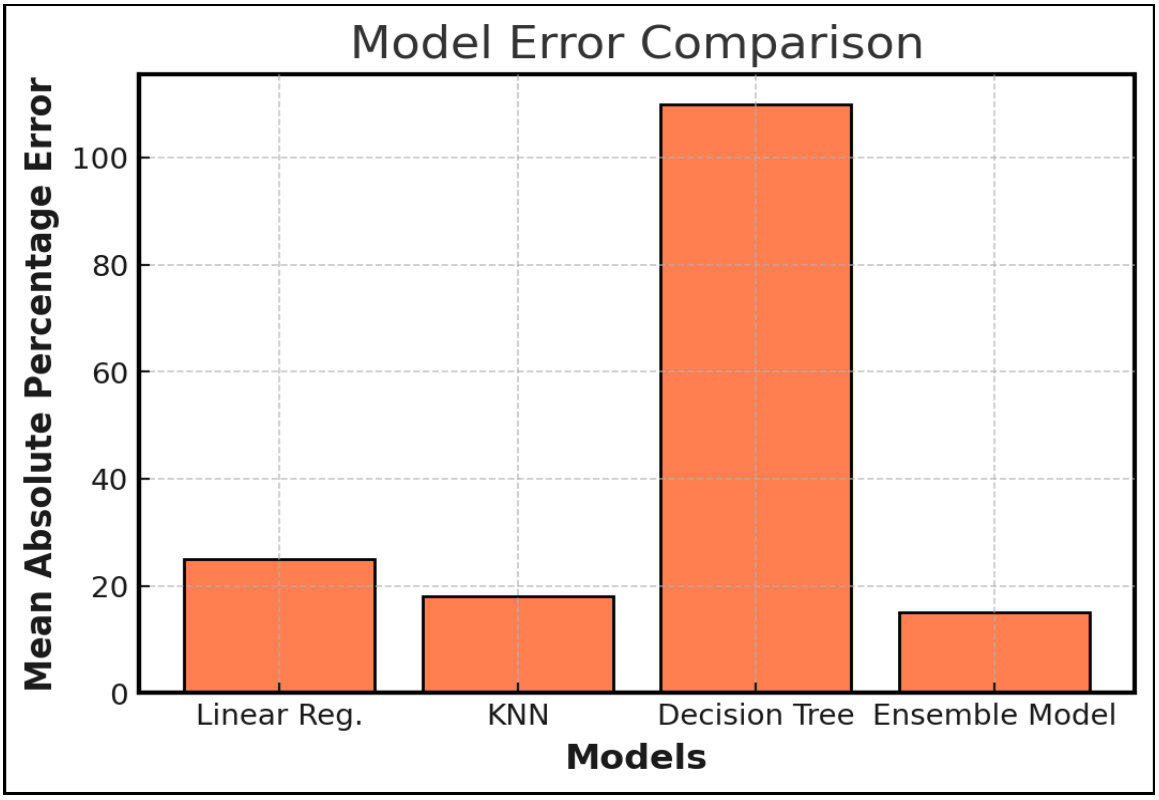
Chính vì nhược điểm của từng mô hình riêng lẻ, việc sử dụng mô hình kết hợp (Ensemble Models) là một phương pháp cần thiết để tối ưu hóa hiệu quả dự đoán giá nhà. Các mô hình kết hợp là một phương pháp trong lĩnh vực học máy nhằm tạo ra một mô hình mạnh hơn và ổn định hơn bằng cách kết hợp nhiều thuật toán đơn lại với nhau. Những mô hình này cố gắng đạt được kết quả chính xác hơn bằng cách kết hợp dự đoán của các thuật toán khác nhau. Các kỹ thuật phổ biến trong mô hình kết hợp bao gồm bao đóng (Bagging), tăng cường (Boosting), xếp chồng (Stacking) và trung bình (Averaging) giúp tăng tính ổn định của mô hình bằng cách giảm phương sai và tăng độ chính xác bằng cách giảm độ lệch **[4]** . Trong các nghiên cứu gần đây, mô hình kết hợp đã được chứng minh là một phương pháp hiệu quả trong

nhiều ứng dụng khác nhau. Chẳng hạn, trong nghiên cứu của Sibindi R, Mwangi RW, Waititu AG, các tác giả đã sử dụng mô hình kết hợp của Light Gradient Boosting Machine (LightGBM) và Extreme Gradient Boosting (XGBoost), so sánh các chỉ số hiệu suất như MSE, MAE, MAPE của mô hình kết hợp này với các mô hình riêng lẻ khác như: Adaboost, GBM, LBM, XGBoost. Cụ thể **Bảng 1** bên dưới là bảng so sánh chỉ số hiệu số của các mô hình này, kết quả cho thấy mô hình kết hợp của LightGBM và XGBoost được tối ưu hóa có kết quả hiệu suất tốt hơn với MSE, MAE và MAPE thấp hơn so với các thuật toán máy học cơ bản riêng lẻ theo kết quả của nghiên cứu trong tài liệu **[5]**.

Bảng 1: Bảng so sánh chỉ số hiệu suất giữa các mô hình học máy.

Algorithm	MSE	MAE	MAPE	Time Complexity (seconds)
Adaboost	0.564	0.588	0.395	1.791
LGBM	0.198	0.290	0.161	1.043
GBM	0.466	0.517	0.328	4.914
XGBoost	0.201	0.295	0.163	7.005
LGBM-XGBoost	**0.193**	**0.285**	**0.156**	**58.631**

Trong một nghiên cứu khác **[6]** , mô hình kết hợp trung bình có trọng số hoạt động tốt hơn đáng kể so với các mô hình riêng lẻ. Mô hình huấn luyện đạt độ chính xác 84%. Việc so sánh sai số phần trăm tuyệt đối trung bình (Mean Absolute Percentage Error - MAPE) của các mô hình được sử dụng: hồi quy tuyến tính (Linear Regression), K-Nearest Neighbors (KNN), cây quyết định (Decision Tree) và mô hình kết hợp được đưa ra dưới dạng biểu đồ thanh (**Hình 2**). Khi so sánh các mô hình khác nhau, chúng tôi thấy rằng mô hình kết hợp hoạt động tốt nhất với giá trị lỗi phần trăm tuyệt đối trung bình thấp nhất.



Hình 2: Bảng so sánh MAPE giữa các mô hình.

Những minh chứng này cho thấy việc sử dụng mô hình kết hợp không chỉ cải thiện độ chính xác mà còn giảm phương sai. Việc sử dụng kết hợp nhiều mô hình có thể tìm hiểu các tính năng của tập dữ liệu từ nhiều chiều khác nhau và có khả năng chuyển giao và khái quát hóa mạnh mẽ. Bằng cách sử dụng phương pháp mô hình kết hợp sẽ mang lại độ chính xác cao, khả năng khải quát tốt hơn và tận dụng sự đa dạng từ các mô hình riêng lẻ kết hợp lại với nhau. Phương pháp mô hình kết hợp đã thu hút sự chú ý ngày càng tăng trong lĩnh vực khai thác dữ liệu do hiệu suất tuyệt vời của nó trong phân tích dự đoán. Hơn nữa, sự khác biệt giữa các mô hình dự đoán cơ sở được tích hợp vào mô hình kết hợp càng lớn thì hiệu suất mà mô hình kết hợp đạt được càng tốt **[7]**. Trong nghiên cứu này chúng tôi đi vào khảo sát và ứng dụng các mô hình và cách thức kết hợp mô hình khác nhau để đánh giá hiệu quả trên tập dữ liệu vào bài toán dự báo giá bất động sản tại Hà Nội.

CHƯƠNG 3: GIẢI PHÁP ĐỀ XUẤT

3.1. Giới thiệu về đề tài

Để nâng cao khả năng dự đoán giá nhà, chúng tôi sẽ tiến hành tối ưu hóa hiệu suất bằng cách ứng dụng ba thuật toán hàng đầu đã được chứng minh về độ chính xác vượt trội trong các nghiên cứu dưới đây. Mục tiêu của chúng tôi là xác định mô hình tối ưu nhất cho tập dữ liệu hiện có, đảm bảo kết quả dự báo chính xác và tin cậy:

○ Nghiên cứu của Chowhaan và M. Jagan đã đánh giá hiệu suất và hiệu quả của các mô hình như XGBoost, rừng ngẫu nhiên (Random Forest), hồi quy tuyến tính, hồi quy Lasso (Lasso Regression) và máy vectơ hỗ trợ (Support Vector Machine - SVM) **[8]** . Kết luận cho thấy XGBoost có hiệu suất vượt trội nhờ khả năng xử lý các tập dữ liệu nhiều chiều, nắm bắt các mối quan hệ phức tạp và quản lý hiệu quả các tương tác tính năng (**Bảng 2**).

Bảng 2: Bảng đánh giá kết quả hiệu suất của các mô hình.

S.No	Model	Score	RMSE
1	Linear Regression	0.790384	64.898435
2	Lasso Regression	0.803637	62.813243
3	Support Vector Mạchine (SVM)	0.206380	126.278064
4	Random Forest	0.903507	44.032172
5	**XGBoost**	**0.886607**	**47.732530**

○ Nghiên cứu của Shahasane và Aditi đã sử dụng nhiều thuật toán hồi quy khác nhau để dự đoán giá nhà, như hồi quy tuyến tính, hồi quy Lasso và cây quyết định **[2]**. Sau khi áp dụng tất cả các thuật toán này vào tập dữ liệu giá nhà của thành phố

Bangalore, Ấn Độ, việc so sánh độ chính xác sẽ được thể hiện ở **Bảng 3**. Kết quả chỉ ra rằng độ chính xác tối đa là 84,77% được đưa ra bởi thuật toán hồi quy tuyến tính.

Bảng 3: Bảng so sánh độ chính xác của các mô hình.

Model	Best_Score
Decision_Tree	0.731685
Lasso	0.726745
Linear_Regression	**0.847796**

○ Năm thuật toán riêng biệt trong cụm hồi quy bao gồm: rừng ngẫu nhiên, cây quyết định, KNN, logistic, vectơ hỗ trợ là những phương pháp được sử dụng trong nghiên cứu của Tanamal và Rinabi **[9]**. Sau khi so sánh các tiêu chí lỗi, rừng ngẫu nhiên nổi lên như một thuật toán thời thượng với điểm chính xác cao nhất là 88% và các giá trị lỗi thấp nhất (**Bảng 4**).

Bảng 4: Bảng so sánh độ chính xác giữa các thuật toán khác nhau.

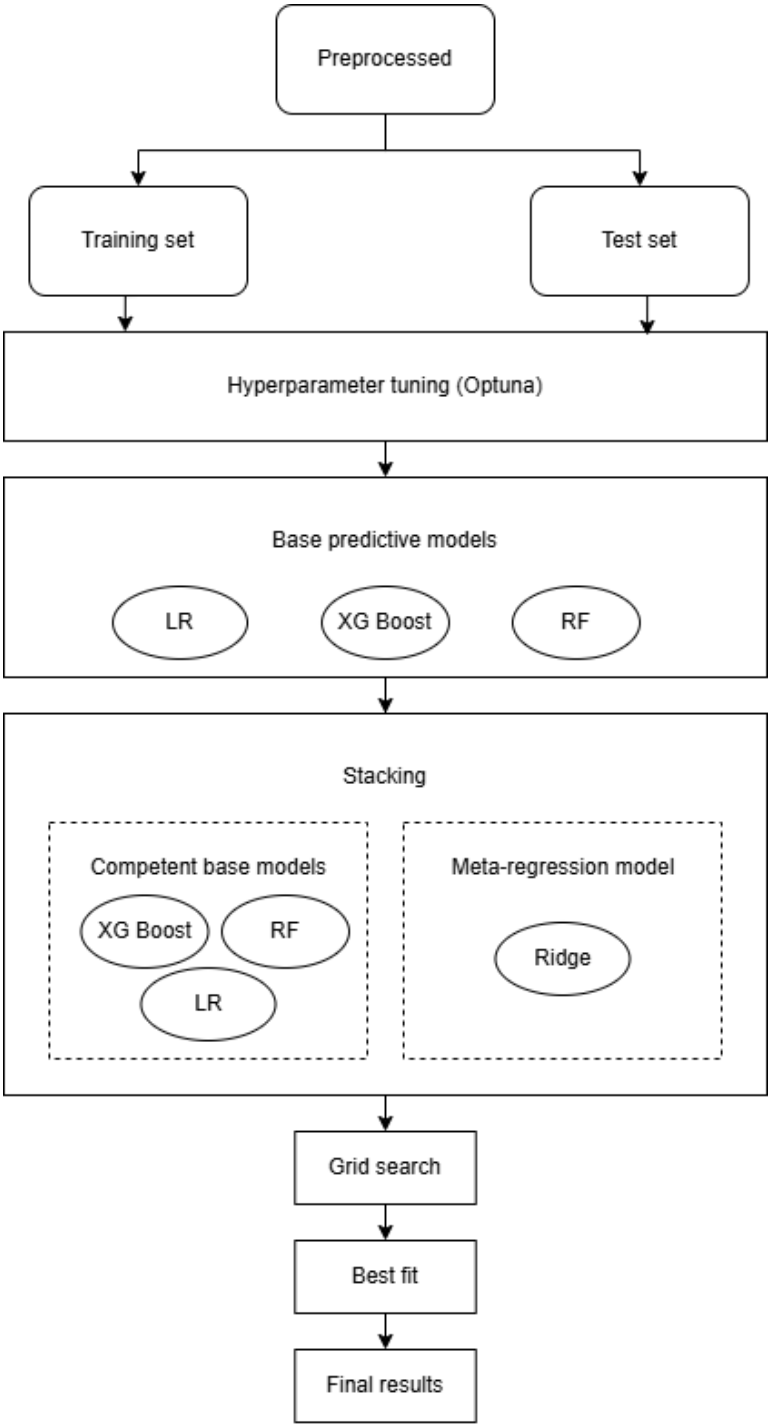
Model	F1 Score	Accuracy
K-Nearest Neighbour	0.70	0.70
Logistic	0.65	0.65
Support Vector Model	0.65	0.65
Decision Tree	0.75	0.75
Random Forest	**0.88**	**0.88**

Dựa vào kết quả của các nghiên cứu đã nêu trên, để nâng cao độ tin cậy trong dự đoán giá nhà, chúng tôi sẽ sử dụng ba mô hình được cho là hiệu quả nhất trong dự đoán giá nhà: XGBoost, hồi quy tuyến tính và rừng ngẫu nhiên làm tiền đề cho nghiên cứu mô hình kết hợp này. Bằng cách triển khai phương pháp tiếp cận tổng hợp trên các cặp thuật toán, chúng tôi sẽ khai thác điểm mạnh và khắc phục điểm yếu của từng thuật toán, từ đó nâng cao độ tin cậy của các dự đoán.

- Rừng ngẫu nhiên và XGBoost.
- Rừng ngẫu nhiên và hồi quy tuyến tính.
- XGBoost và hồi quy tuyến tính.
- Rừng ngẫu nhiên, XGBoost và hồi quy tuyến tính.

Phương pháp xây dựng mô hình kết hợp chúng tôi sẽ thực nghiệm cho bốn cặp kết hợp khác nhau từ ba thuật toán là xếp chồng. Xếp chồng tập trung vào việc kết hợp một số bộ phân loại được tạo bằng các thuật toán học khác nhau trên một tập dữ liệu duy nhất được tạo thành cặp vectơ đặc trưng và phân loại của chúng **[4]**. Để dự đoán các vấn đề trong các lĩnh vực khác nhau, cấu trúc dễ thích ứng hơn và mô

hình kết hợp dựa vào xếp chồng ổn định thể hiện những lợi thế đáng kể. Phương pháp kết hợp dựa trên xếp chồng chắc chắn có thể có hiệu quả trong lĩnh vực bất động sản nhờ những kết quả nổi bật của nó trong các lĩnh vực khác **[4].** Để đánh giá khả năng dự đoán của mô hình học tập kết hợp (Ensemble Learning), hai tiêu chí chính được sử dụng là hệ số xác định (R^2) và sai số bình phương trung bình gốc (RMSE). RMSE là quy tắc tính điểm bậc hai để đo giá trị trung bình độ lớn của sai số, là căn bậc hai của giá trị trung bình của bình phương chênh lệch giữa giá dự đoán và giá trị thực tế. Chỉ số RMSE có thể nằm trong khoảng từ 0 đến ∞ và không quan tâm đến hướng của lỗi. Điểm số RMSE càng thấp thì mô hình càng tốt. Ngoài ra, giá trị R^2 nằm trong khoảng từ 0 đến 1. Giá trị R^2 càng gần 1 thì mô hình càng phù hợp với dữ liệu thực nghiệm trong bài toán hồi quy. Ngược lại, R^2 càng gần 0 thì mô hình càng kém phù hợp với tập dữ liệu đó.



Hình 3: Kiến trúc mô hình đề xuất

Trong nghiên cứu này, chúng tôi đề xuất một kiến trúc mô hình học máy **(Hình 3)** dựa trên kỹ thuật Stacking, kết hợp ba mô hình nền là Random Forest, XG Boost, Linear Regression, với một mô hình meta là Ridge Regression. Mô hình được xây dựng trong môi trường Python với thư viện Scikit-learn, tích hợp Optuna cho tối ưu siêu tham số.

3.2. Tối ưu hóa các mô hình nền

Trong nghiên cứu này, hai trong ba mô hình nền là Random Forest và XGBoost đã được tối ưu hóa siêu tham số bằng thuật toán tối ưu hóa Bayesian, sử dụng thư viện Optuna. Đối với mô hình XGBoost, quá trình tối ưu được thực hiện trên không gian siêu tham số rộng bao gồm: số lượng cây (n_estimators), độ sâu tối đa của cây (max_depth), tốc độ học (learning_rate), tỷ lệ mẫu ngẫu nhiên (subsample), tỷ lệ đặc trưng sử dụng khi xây dựng mỗi cây (colsample_bytree), tham số phạt cho phân vùng không hiệu quả (gamma), và các hệ số điều chuẩn L1 (reg_alpha) và L2 (reg_lambda). Cấu hình tối ưu thu được là: n_estimators = 536, max_depth = 10, learning_rate ≈ 0.0101, subsample ≈ 0.6001, colsample_bytree ≈ 0.9765, gamma≈2.3278, reg_alpha≈1.3690, reg_lambda≈2.0116, và min_child_weight ≈ 1.2579. Mô hình đạt được sai số log trung bình căn bậc hai (Root Mean Squared Logarithmic Error – RMSLE) là 0.24199, được đánh giá thông qua quy trình cross-validation với 5 lần gập (5-fold cross-validation).

Tương tự, mô hình Random Forest cũng được tối ưu hóa với các siêu tham số bao gồm: số lượng cây (n_estimators), độ sâu tối đa của cây (max_depth), tỷ lệ đặc trưng sử dụng (max_features), cũng như các ràng buộc về phân chia node như số lượng mẫu tối thiểu để chia (min_samples_split) và số lượng mẫu tối thiểu ở lá (min_samples_leaf). Ngoài ra, lựa chọn có hay không sử dụng kỹ thuật lấy mẫu bootstrap (bootstrap) cũng được đưa vào quá trình tìm kiếm. Kết quả tối ưu hóa cho Random Forest cho thấy cấu hình hiệu quả nhất là: n_estimators = 1091, max_depth = 27, max_features ≈ 0.9827, min_samples_split = 2, min_samples_leaf = 1, và bootstrap = False.

Tất cả các mô hình được đánh giá bằng điểm số RMSLE thông qua 5-fold cross-validation nhằm đảm bảo tính tổng quát và độ ổn định của kết quả trên các tập dữ liệu chưa thấy.

3.3. Huấn luyện mô hình cơ sở

Sau khi tìm được các tham số tốt nhất, ba mô hình riêng biệt được huấn luyện lại trên tập huấn luyện với toàn bộ dữ liệu. Cụ thể:

- XGBoost được cấu hình để chạy trên GPU (gpu_hist) nhằm tối ưu tốc độ huấn luyện.
- Random Forest sử dụng đa luồng CPU.
- Linear Regression được huấn luyện như một baseline có tính tuyến tính toàn cục.

Mỗi mô hình được đánh giá bằng các thước đo chuẩn như RMSE, RMSLE và hệ số xác định R^2. Các biểu đồ scatter giữa giá trị thực và dự đoán cho từng mô hình cho thấy độ lệch và vùng giá sai khác giúp hiểu sâu hơn về đặc tính từng mô hình.

3.4. Kết hợp bằng kĩ thuật Stacking

Mô hình Stacking Regressor được xây dựng như một giải pháp mạnh mẽ để tổng hợp và tối ưu hóa khả năng dự đoán, bằng cách kết hợp thông minh ba mô hình cơ sở riêng biệt: Random Forest (rf_pipe), XGBoost (xgb_pipe), và một mô hình Hồi quy tuyến tính (lin_pipe). Mỗi mô hình này, được đóng gói gọn gàng trong các pipeline tiền xử lý và học máy của riêng chúng, hoạt động như những base learners, đưa ra những dự đoán ban đầu dựa trên các đặc trưng đã được xử lý.

Để biến những dự đoán riêng lẻ này thành một kết quả tổng hợp có độ chính xác cao, một mô hình RidgeCV được chọn làm final_estimator – hay còn gọi là meta-learner. RidgeCV nổi bật nhờ khả năng tự động lựa chọn hệ số điều chuẩn α tối ưu thông qua quá trình cross-validation nội bộ. Điều này không chỉ giúp kiểm soát hiệu quả hiện tượng quá khớp mà còn giảm thiểu tác động của đa cộng tuyến giữa các đầu ra của mô hình nền, từ đó nâng cao đáng kể sự ổn định và tin cậy của mô hình tổng hợp. Vai trò của RidgeCV là học cách gán trọng số phù hợp hoặc kết hợp một cách tuyến tính các dự đoán từ Random Forest, XGBoost và mô hình tuyến tính, để đưa ra dự đoán cuối cùng.

Quá trình huấn luyện của Stacking Regressor được thực hiện một cách tỉ mỉ thông qua K-Fold cross-validation với 10 lần chia (n_splits=10). Để đảm bảo tính ngẫu nhiên và khả năng tái lập của kết quả, dữ liệu được xáo trộn (shuffle=True) trước khi chia thành các folds, và một random_state cố định được đặt là 42. Cơ chế này đảm bảo rằng mô hình meta học được từ các dự đoán được tạo ra trên dữ liệu mà các mô hình cơ sở chưa từng nhìn thấy trong quá trình huấn luyện của chính chúng, mô phỏng chân thực hơn hiệu suất của mô hình trên dữ liệu mới. Hơn nữa,

việc sử dụng tham số n_jobs=-1 cho phép tận dụng toàn bộ số lõi CPU có sẵn, giúp tăng tốc đáng kể quá trình huấn luyện.

Một yếu tố then chốt trong cấu hình này là việc đặt tham số passthrough là False. Điều này có nghĩa là mô hình meta (RidgeCV) chỉ nhận đầu vào là các dự đoán từ ba mô hình cơ sở, mà không trực tiếp tiếp cận dữ liệu gốc. Cách tiếp cận này có hai lợi ích chính: thứ nhất, nó ngăn chặn hiệu quả hiện tượng rò rỉ thông tin từ dữ liệu gốc vào mô hình meta, đảm bảo tính khách quan của quá trình học; thứ hai, nó buộc mô hình meta phải tập trung hoàn toàn vào việc học cách kết hợp và điều chỉnh các quan điểm khác nhau từ các mô hình nền. Điều này giúp mô hình meta phát huy tối đa vai trò tổng hợp, tạo ra một dự đoán cuối cùng mạnh mẽ và đáng tin cậy hơn.

Với sự kết hợp chặt chẽ giữa các mô hình mạnh mẽ, chiến lược huấn luyện kỹ lưỡng và cơ chế bảo vệ khỏi rò rỉ thông tin, mô hình Stacking này được kỳ vọng sẽ mang lại hiệu suất dự đoán vượt trội, tận dụng tối đa điểm mạnh của từng thuật toán cơ sở và giảm thiểu nhược điểm riêng lẻ của chúng.

CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1. Môi trường thực nghiệm

Toàn bộ quá trình nghiên cứu đã được thực hiện trên nền tảng Kaggle Notebooks. Để đảm bảo hiệu suất tính toán tối ưu và hỗ trợ các mô hình phức tạp, môi trường thực nghiệm được cấu hình với tài nguyên phần cứng mạnh mẽ. Cụ thể, hệ thống được trang bị bộ vi xử lý Intel(R) Xeon(R) CPU 2.00GHz, với kiến trúc 4 lõi và 8 luồng, cùng với 30 GB RAM. Đặc biệt, quá trình huấn luyện mô hình được tăng tốc đáng kể nhờ sự hiện diện của hai bộ xử lý đồ họa NVIDIA T4 Tensor Core GPU, cung cấp tổng cộng 32 GB VRAM. Sự kết hợp đồng bộ giữa cấu hình phần cứng hiệu năng cao và các thư viện phần mềm được tối ưu hóa đã rút ngắn đáng kể thời gian huấn luyện mô hình, đặc biệt có lợi cho việc triển khai và đánh giá các mô hình ensemble phức tạp. Điều này không chỉ đảm bảo tính khả thi mà còn nâng cao hiệu quả tổng thể của nghiên cứu.

4.2. Thu thập dataset

Trong khuôn khổ đề tài này, Bộ dữ liệu này được thu thập để dự đoán giá nhà tại Hà Nội bằng cách sử dụng một kỹ thuật được gọi là thu thập dữ liệu từ web. Kỹ thuật này liên quan đến việc tự động lấy thông tin từ các trang web. Để tạo điều kiện thuận lợi cho quá trình này, chúng tôi đã sử dụng một thư viện trong Python có tên là Selenium. Selenium là một công cụ Python hỗ trợ phân tích cú pháp các tài liệu HTML và XML. Nó phân tích cú pháp cấu trúc HTML của một trang web và cho phép chúng tôi tìm kiếm các thẻ HTML, thuộc tính hoặc chuỗi văn bản cụ thể trong đó. Sau khi xác định được, chúng tôi có thể trích xuất dữ liệu mong muốn từ các phần tử này. Trong nghiên cứu này, chúng tôi đã sử dụng Selenium để điều hướng qua các trang của batdongsan.com, một trang web niêm yết bất động sản.

Sau khi cào dữ liệu xong, chúng tôi thu thập được bộ dữ liệu gồm 42113 dòng và các đặc trưng của bộ dữ liệu bao gồm những thông tin sau đây:

Bảng 5: Bảng các đặc trưng và loại dữ liệu

Đặc trưng	Mô tả đặc trưng	Loại dữ liệu
ID	Mã số duy nhất của căn nhà	Số
Tiêu đề tin đăng	Tóm tắt chính về căn nhà	Phi số
Địa chỉ	Địa chỉ của căn nhà bao gồm: quận/huyện và phường/xã	Phi số
Diện tích	Tổng diện tích của căn nhà (m2)	Số
Giá nhà	Tổng giá trị của căn nhà (triệu hoặc tỷ đồng)	Số
Giá/m2	Giá trị mỗi mét vuông của căn nhà (triệu hoặc tỷ đồng/m2)	Số
Loại hình nhà ở	Phân loại bất động sản: nhà mặt phố, nhà trong ngõ, chung cư,...	Phi số
Phòng ngủ	Số lượng phòng ngủ	Số
Phòng tắm	Số lượng phòng tắm	Số
Ban công	Căn nhà có hay không có không gian ban công	Phi số
Mặt tiền	Chiều rộng, chiều dài mặt tiền của căn nhà	Số
Giấy tờ pháp lý	Các loại giấy tờ chứng nhận quyền sở hữu như sổ đỏ, sổ hồng	Phi số
Ngày đăng tin	Thời gian mà căn nhà được đăng bán	Phi số

4.3. Làm sạch dữ liệu

Bộ dữ liệu sử dụng trong quá trình huấn luyện sẽ được làm sạch qua nhiều bước như:

- Loại bỏ các cột không cần thiết như: “link”, “id”, “city”, “title1”, “title2”

```
columns_to_drop = ['links', 'title1', 'title2', 'title3', 'city', 'index_', 'id']

# Bỏ các cột
df = df.drop(columns=columns_to_drop)
```

- Xử lý các khoảng trắng dư thừa ở cột “area”

```
# 1. Loại bỏ từ "m²" và khoảng trắng
df['area'] = df['area'].str.replace(' m²', '', regex=False)

# 2. Thay dấu phẩy bằng dấu chấm
df['area'] = df['area'].str.replace(',', '.', regex=False)
```

- Chuẩn hóa định dạng ngày tháng cột “date” (chuyển về yyyy/mm/dd)

```
# Chuyển cột date thành datetime
df['date'] = pd.to_datetime(df['date'], format='%d/%m/%Y')
df['year'] = df['date'].dt.year
df['month'] = df['date'].dt.month
df['day'] = df['date'].dt.day
df = df.drop(columns=['date'])
```

- Một số người dùng đã nhầm lẫn khi nhập giá trị giữa cột price và price_per_m2. Cụ thể, tổng giá có thể bị nhập vào cột price_per_m2 và giá trên mét vuông lại được nhập vào cột price. Để sửa lỗi này, dữ liệu được quét qua từng hàng. Nếu cột price chứa đơn vị “triệu/m²” và cột price_per_m2 chứa “tỷ”, các giá trị này sẽ được hoán đổi cho đúng vị trí:

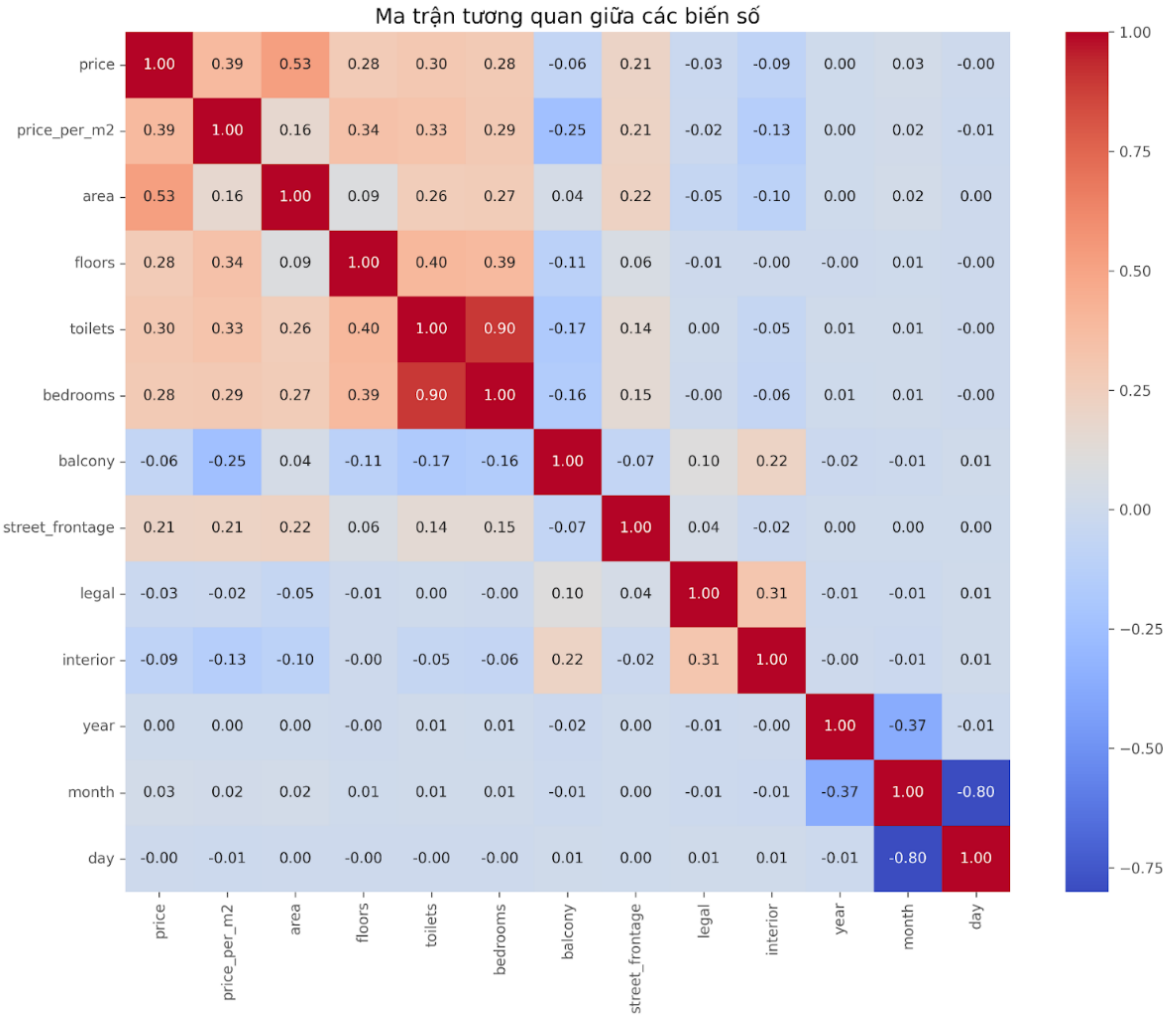
```
# Phát hiện nhầm lẫn và đổi chỗ dữ liệu
for idx in df.index:
    price_val = df.loc[idx, 'price']
    price_per_m2_val = df.loc[idx, 'price_per_m2']

    # Nếu cột price chứa "triệu/m²" và cột price_per_m2 chứa "tỷ", thì đổi chỗ
    if 'triệu/m²' in price_val and 'tỷ' in price_per_m2_val:
        df.loc[idx, 'price'] = price_per_m2_val
        df.loc[idx, 'price_per_m2'] = price_val
```

- Các giá trị “Thoả thuận” trong cột price gây cản trở phân tích do không phải dạng số. Để khắc phục, một hàm “clean_price” được sử dụng để chuyển đổi các giá trị “Thoả thuận” thành NaN:

```
def clean_price(value):
    # Nếu giá trị là "Thoả thuận" trả về NaN
    if 'Thoả thuận' in str(value) or pd.isna(value):
        return np.nan
```

4.4. Khai phá dữ liệu (EDA)

Trước khi triển khai bất kỳ mô hình nào, điều quan trọng là phải xác thực độ chính xác và tính phù hợp của tập dữ liệu để phân tích. Để thực hiện điều này, chúng tôi đã tiến hành Phân tích dữ liệu thăm dò (EDA) kỹ lưỡng, đi sâu vào các tính năng, thuộc tính và mối quan hệ của tập dữ liệu. Chỉ ra sự hiện diện và cường độ tương quan giữa các biến này. Hệ số tương quan, một chỉ số số nằm trong khoảng từ +1 đến -1, làm sáng tỏ mức độ liên kết giữa hai biến: hệ số dương biểu thị mối quan hệ tích cực, hệ số âm biểu thị mối quan hệ tiêu cực, và hệ số 0 biểu thị sự độc lập giữa các biến.



Hình 4: Ma trận tương quan giữa các biến số

Ma trận tương quan được trình bày ở trên cung cấp cái nhìn tổng quan về mối quan hệ tuyến tính giữa các biến số chính trong tập dữ liệu của chúng ta. Các giá trị trong ma trận là hệ số tương quan Pearson, nằm trong khoảng từ -1 đến 1, cho biết cường độ và chiều hướng của mối quan hệ: giá trị gần 1 thể hiện tương quan dương

mạnh (biến này tăng, biến kia cũng tăng), giá trị gần -1 thể hiện tương quan âm mạnh (biến này tăng, biến kia giảm), và giá trị gần 0 cho thấy không có hoặc có rất ít tương quan tuyến tính. Màu sắc trong biểu đồ cũng trực quan hóa cường độ và chiều hướng này, với các tông màu đỏ/cam biểu thị tương quan dương và các tông màu xanh biểu thị tương quan âm.

Kết quả phân tích cho thấy một số mối quan hệ nổi bật. Đáng chú ý, các biến liên quan đến đặc điểm của bất động sản như giá (price), giá trên mét vuông (price_per_m2), diện tích (area), số tầng (floors), số nhà vệ sinh (toilets) và số phòng ngủ (bedrooms) có mối tương quan dương rất mạnh mẽ với nhau (thể hiện qua các ô màu đỏ đậm và giá trị gần 1.00). Điều này khẳng định rằng những bất động sản có quy mô lớn hơn, nhiều tiện ích hơn thường đi kèm với giá trị cao hơn, phù hợp với kỳ vọng chung trên thị trường bất động sản. Chẳng hạn, một sự gia tăng về diện tích hoặc số lượng phòng ngủ thường đi kèm với mức giá tăng đáng kể.

Ngược lại, một số biến thể hiện mối tương quan yếu hoặc không đáng kể với các đặc điểm chính của bất động sản. Các biến về thời gian như năm (year), tháng (month) và ngày (day) hầu như không có tương quan tuyến tính với giá hoặc các thuộc tính vật lý khác của bất động sản. Điều này cho thấy thời điểm giao dịch cụ thể không phải là yếu tố dự báo mạnh mẽ cho giá trị bất động sản trong tập dữ liệu này. Ngoài ra, biến balcony (ban công) cho thấy một mối tương quan âm nhẹ đến trung bình với giá và giá trên mét vuông, một điểm cần được khám phá sâu hơn để hiểu rõ bản chất của mối quan hệ này trong bối cảnh dữ liệu cụ thể.

Tóm lại, ma trận tương quan này cung cấp cái nhìn cơ bản nhưng quan trọng về cấu trúc dữ liệu, làm nổi bật các biến có ảnh hưởng lớn đến giá bất động sản và xác định các biến có mối quan hệ yếu hơn hoặc không có mối quan hệ tuyến tính rõ ràng. Những thông tin này là nền tảng quý giá cho các phân tích sâu hơn, bao gồm việc xây dựng mô hình dự đoán giá và đánh giá các yếu tố ảnh hưởng cụ thể đến giá trị bất động sản.

4.5. Chuẩn hóa dữ liệu

Trích xuất và chuẩn hóa biến thể loại nhà: từ cột “title2” tách phần trước từ khóa “tại” thành cột house_type, sau đó quy về một số nhãn chuẩn (“Căn hộ chung

cư", "Nhà riêng"…), rồi chuyển sang biểu diễn one-hot để đưa trực tiếp vào mô hình.

Chuyển đổi đơn vị và kiểu số: cột "area" được loại bỏ ký tự "m²", thay dấu phẩy thành dấu chấm và ép kiểu float; "street_frontage" cũng tương tự, đồng thời điền giá trị 0 cho những bất động sản không có mặt tiền.

Mã hóa nhị phân: các cột như "interior", "legal", "balcony" chỉ định 1 nếu có, 0 nếu không, giúp phân biệt nhanh tính chất nội thất, tính pháp lý và ban công.

Tính toán và nội suy giá: sau khi chuẩn hóa các giá trị "triệu"/"tỷ" thành số thực, ta tính ngược "price" hoặc "price_per_m2" nếu thiếu, rồi áp dụng nội suy tuyến tính trên trục "area" cho các bản ghi thiếu cả hai thông số, cuối cùng điền median cho mọi NaN còn lại.

Tóm lại, chúng tôi đã hoàn thành chuyển đổi tập dữ liệu thô thành một tập dữ liệu (**Hình 5**) có cấu trúc đồng nhất và đáng tin cậy, tối ưu hóa cho phân tích và xây dựng mô hình hồi quy. Hiệu quả của việc làm sạch được thể hiện rõ rệt qua việc chuẩn hóa định dạng. Điều này đảm bảo tính nhất quán và khả năng tính toán chính xác cho các biến định lượng. Bên cạnh đó, việc xử lý giá trị thiếu đã được thực hiện một cách có hệ thống. Một điểm nổi bật khác là khả năng khắc phục lỗi nhập liệu bằng cách phát hiện và hoán đổi các giá trị bị nhầm lẫn giữa các cột. Điều này giải quyết vấn đề sai lệch dữ liệu gốc, đảm bảo thông tin giá cả được gán đúng cách, từ đó nâng cao độ chính xác của tập dữ liệu. Cuối cùng, việc chuyển đổi dữ liệu của các cột sang định dạng nhị phân đã tăng cường khả năng sử dụng dữ liệu, giúp các mô hình học máy dễ dàng tiếp nhận và xử lý. Nhờ các bước tiền xử lý toàn diện này, tập dữ liệu sau làm sạch đã đạt được độ tin cậy và cấu trúc tối ưu, là nền tảng vững chắc cho các giai đoạn phân tích chuyên sâu tiếp theo.

price	price_per_m2	area	floors	toilets	bedroom	balcony	street_frontage	legal	interior	house_type_C	house	house	house_t	house	year	month	day	district_Ba
54	370.11	145	6	2	3	1	0	1	0	TRUE	FALSE	FALSE	FALSE	FALSE	2024	11	5	FALSE
44	312.5	141	5	4	5	1	20	1	1	FALSE	TRUE	FALSE	FALSE	FALSE	2024	11	27	FALSE
29	381.33	75	6	8	8	1	5	1	1	FALSE	FALSE	FALSE	TRUE	FALSE	2024	12	17	FALSE
23	389.16	60	5	2	2	1	0	1	1	TRUE	FALSE	FALSE	FALSE	FALSE	2024	12	10	FALSE
21	278.67	75	6	2	3	0	0	1	1	TRUE	FALSE	FALSE	FALSE	FALSE	2024	9	25	FALSE
19	246.67	75	5	4	9	1	5	1	0	FALSE	TRUE	FALSE	FALSE	FALSE	2024	10	9	FALSE
12	132.96	90	5	4	3	0	0	1	0	TRUE	FALSE	FALSE	FALSE	FALSE	2024	11	13	FALSE
10	236.43	43	4	1	1	0	0	1	0	TRUE	FALSE	FALSE	FALSE	FALSE	2024	10	23	FALSE
9.5	220.93	43	6	1	1	0	0	1	0	TRUE	FALSE	FALSE	FALSE	FALSE	2024	10	23	FALSE
9.5	84.07	113	6	3	3	1	0	1	1	TRUE	FALSE	FALSE	FALSE	FALSE	2024	10	12	FALSE
8.1	92	88	5	4	2	0	0	1	1	TRUE	FALSE	FALSE	FALSE	FALSE	2024	10	12	FALSE
8.1	125.78	64	5	2	2	1	0	1	1	TRUE	FALSE	FALSE	FALSE	FALSE	2024	11	22	FALSE
6.1	67.78	90	5	2	2	1	0	1	1	TRUE	FALSE	FALSE	FALSE	FALSE	2024	10	1	FALSE
5.8	127.78	45	6	1	1	1	0	1	1	TRUE	FALSE	FALSE	FALSE	FALSE	2024	11	14	FALSE
5.6	129.3	43	5	1	1	0	0	1	0	TRUE	FALSE	FALSE	FALSE	FALSE	2024	10	23	FALSE
4.8	70	68	5	2	2	1	0	1	1	TRUE	FALSE	FALSE	FALSE	FALSE	2024	9	24	FALSE
3.8	60	64	1	2	3	0	0	1	1	TRUE	FALSE	FALSE	FALSE	FALSE	2024	10	3	FALSE
2.5	78.62	32	4	1	1	1	0	1	1	TRUE	FALSE	FALSE	FALSE	FALSE	2024	10	16	FALSE
980	27.22	36	1	1	1	0	0	1	1	TRUE	FALSE	FALSE	FALSE	FALSE	2025	3	8	FALSE

Hình 5: Tập dữ liệu dùng huấn luyện mô hình

Sau khi hoàn thành quá trình tiền xử lý dữ liệu, chúng tôi chia tập dữ liệu cho 80% cho tập train và 20% cho tập test để tiến hành huấn luyện mô hình:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)
```

4.6. Phương pháp đánh giá

Trong nghiên cứu về mô hình hồi quy (regression), chúng tôi đánh giá hiệu suất của mô hình bằng cách sử dụng các chỉ số khác biệt so với phân loại đa nhãn. Hồi quy tập trung vào việc dự đoán một hoặc nhiều biến số liên tục thay vì các nhãn rời rạc. Do đó, các chỉ số đánh giá hiệu suất cũng cần phản ánh khả năng của mô hình trong việc ước tính các giá trị số một cách chính xác. Các chỉ số như MAE, MSE và RMSE được sử dụng để so sánh hiệu suất trên cùng một tập dữ liệu. Việc so sánh định lượng giữa các phương pháp được thực hiện thông qua hệ thống chỉ số đánh giá toàn diện. Các chỉ số sai số tuyệt đối (MAE), sai số bình phương (MSE) và sai số bình phương trung bình căn bậc hai (RMSE) đóng vai trò then chốt trong việc xác định mô hình tối ưu, với nguyên tắc cơ bản: mô hình có giá trị các chỉ số này càng thấp càng thể hiện độ chính xác cao hơn.

Hệ số xác định R^2 cung cấp thước đo định lượng về mức độ phù hợp của mô hình với tập dữ liệu, trong đó giá trị tiệm cận 1 phản ánh khả năng giải thích phương sai biến phụ thuộc ngày càng hoàn thiện.

Các chỉ số sai số (MAE, MSE, RMSE) cung cấp thông tin về độ lớn của các sai lệch giữa dự đoán và thực tế. Phân tích sâu phân bố sai số thông qua các chỉ số MAE, MSE và RMSE cho phép nhận diện đặc điểm lỗi đặc thù của từng mô hình, bao gồm cả xu hướng và biên độ sai lệch. Việc phân tích sự phân bố của các sai số này cũng rất quan trọng để hiểu rõ hơn về các loại lỗi mà mô hình mắc phải.

Việc lựa chọn chỉ số ưu tiên cần được xem xét trong bối cảnh ứng dụng cụ thể. Đối với các bài toán nhạy cảm với sai số lớn, RMSE nên được ưu tiên do đặc tính phạt nặng các sai số cá biệt. Ngược lại, MAE phù hợp hơn khi yêu cầu diễn giải trực quan về độ lệch trung bình. R^2 đặc biệt giá trị khi cần đánh giá tổng thể mức độ phù hợp của mô hình với cấu trúc dữ liệu nền tảng.

Cách tiếp cận đa chỉ số này cho phép đánh giá toàn diện cả về độ chính xác điểm (thông qua MAE), độ nhạy với sai số lớn (qua RMSE) và khả năng giải thích tổng thể (bằng R^2), từ đó đưa ra quyết định lựa chọn mô hình tối ưu cho từng tình huống ứng dụng cụ thể.

Dưới đây, chúng tôi trình bày định nghĩa, cách tính toán, ý nghĩa của từng chỉ số trong bối cảnh hồi quy, lý do lựa chọn, và cách áp dụng chúng trong việc đánh giá mô hình.

Tổng quan về các phương pháp đánh giá đã được chúng tôi sử dụng:

- Mean Absolute Error - Sai số tuyệt đối trung bình: Trung bình của các sai số tuyệt đối giữa giá trị dự đoán và giá trị thực tế.
- Mean Squared Error - Sai số bình phương trung bình: Trung bình của bình phương các sai số giữa giá trị dự đoán và giá trị thực tế.
- Root Mean Squared Error - Căn bậc hai của sai số bình phương trung bình: Căn bậc hai của MSE.
- R2 (Coefficient of Determination) - Hệ số xác định: Tỷ lệ phương sai trong biến phụ thuộc có thể dự đoán được từ biến độc lập.

Định nghĩa, cách tính toán, ý nghĩa và lý do lựa chọn của từng chỉ số trong bối cảnh hồi quy:

- Mean Absolute Error (MAE):

- Định nghĩa: MAE đo lường độ lớn trung bình của các sai số giữa các giá trị dự đoán và giá trị thực tế. Nó cho biết trung bình các dự đoán của mô hình sai lệch bao nhiêu so với giá trị thực tế.
- Cách tính toán:

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right|$$

- Trong đó:
  - n là số lượng mẫu.
  - yi là giá trị thực tế của mẫu thứ i.
  - y^i là giá trị dự đoán của mẫu thứ i.
  - |yi−y^i| là sai số tuyệt đối của mẫu thứ i.
- Ý nghĩa: MAE dễ hiểu và diễn giải. Nó cho biết mức độ sai lệch trung bình của các dự đoán theo đơn vị của biến mục tiêu. MAE ít nhạy cảm hơn với các giá trị ngoại lệ so với MSE.
- Lý do lựa chọn: MAE hữu ích khi muốn có một thước đo sai số dễ diễn giải và khi các giá trị ngoại lệ không được coi là quá quan trọng.

- Mean Squared Error (MSE):
  - Định nghĩa: MSE đo lường trung bình của bình phương các sai số giữa các giá trị dự đoán và giá trị thực tế.
  - Cách tính toán:

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2$$

  - Trong đó:
    - n là số lượng mẫu.
    - yi là giá trị thực tế của mẫu thứ i.
    - y^i là giá trị dự đoán của mẫu thứ i.
    - $(yi−y^i)^2$ là bình phương sai số của mẫu thứ i.
  - Ý nghĩa: MSE các sai số lớn hơn nhiều so với các sai số nhỏ do việc bình phương. Điều này làm cho MSE nhạy cảm hơn với các giá trị ngoại lệ. Một

giá trị MSE nhỏ cho thấy mô hình có hiệu suất tốt hơn.

- Lý do lựa chọn: MSE thường được sử dụng trong tối ưu hóa mô hình vì nó có tính khả vi (differentiable). Nó cũng hữu ích khi muốn mạnh các dự đoán sai lệch lớn.

- Root Mean Squared Error (RMSE):
  - Định nghĩa: RMSE là căn bậc hai của MSE. Nó đo lường độ lệch chuẩn của các sai số dự đoán (phần dư).
  - Cách tính toán:

$$\mathbf{RMSE} = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2}$$

  - Trong đó các ký hiệu có ý nghĩa tương tự như trong công thức MSE.
  - Ý nghĩa: RMSE có cùng đơn vị với biến mục tiêu, điều này làm cho nó dễ diễn giải hơn MSE. Nó vẫn nhạy cảm với các giá trị ngoại lệ do MSE được sử dụng trong tính toán. Một giá trị RMSE nhỏ cho thấy mô hình có hiệu suất tốt hơn.
  - Lý do lựa chọn: RMSE là một chỉ số phổ biến để đánh giá hiệu suất hồi quy vì nó cung cấp một thước đo sai số có thể diễn giải được và vẫn nhạy cảm với các sai số lớn.

- R-squared (Coefficient of Determination):
  - Định nghĩa: R-squared đại diện cho tỷ lệ phương sai trong biến phụ thuộc có thể được giải thích bởi mô hình hồi quy. Nó cho biết mức độ phù hợp của mô hình với dữ liệu.
  - Cách tính toán:

$$\mathbf{R^2} = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

  - Trong đó:

- SSres (Sum of Squares Residual) là tổng bình phương các sai số (MSE nhân với n).
- SStot (Total Sum of Squares) là tổng bình phương độ lệch của các giá trị thực tế so với giá trị trung bình của chúng (y^-).
- y^- là giá trị trung bình của các giá trị thực tế.

- Ý nghĩa: R-squared có giá trị nằm trong khoảng từ 0 đến 1.
  - R^2=0 có nghĩa là mô hình không giải thích được bất kỳ phương sai nào trong biến phụ thuộc.
  - R^2=1 có nghĩa là mô hình giải thích hoàn toàn phương sai trong biến phụ thuộc. Giá trị R-squared càng cao thì mô hình càng phù hợp với dữ liệu.
- Lý do lựa chọn: R-squared cung cấp một cái nhìn tổng quan về mức độ phù hợp của mô hình với dữ liệu và khả năng giải thích phương sai của nó. Tuy nhiên, R-squared không cho biết liệu các hệ số hồi quy có ý nghĩa thống kê hay không và có thể bị ảnh hưởng bởi việc thêm các biến không liên quan vào mô hình (có thể dẫn đến R-squared tăng lên một cách giả tạo).

Cách áp dụng các chỉ số trong việc đánh giá mô hình hồi quy: Khi đánh giá một mô hình hồi quy, chúng ta thường xem xét đồng thời nhiều chỉ số để có cái nhìn toàn diện về hiệu suất của mô hình.

Tóm lại, việc đánh giá mô hình hồi quy đòi hỏi việc sử dụng kết hợp các chỉ số khác nhau để hiểu rõ về khả năng dự đoán chính xác và mức độ phù hợp của mô hình với dữ liệu.

4.7. Kết quả thực nghiệm

Để xác định mô hình có hiệu suất tối ưu, chúng tôi đã đánh giá bốn mô hình ensemble khác nhau dựa trên ba tiêu chí chính: Sai số trung bình gốc bình phương (RMSE), Sai số logarit trung bình gốc bình phương (RMSLE), và hệ số xác định (R^2). Các kết quả được tổng hợp trong **Bảng 6** dưới đây:

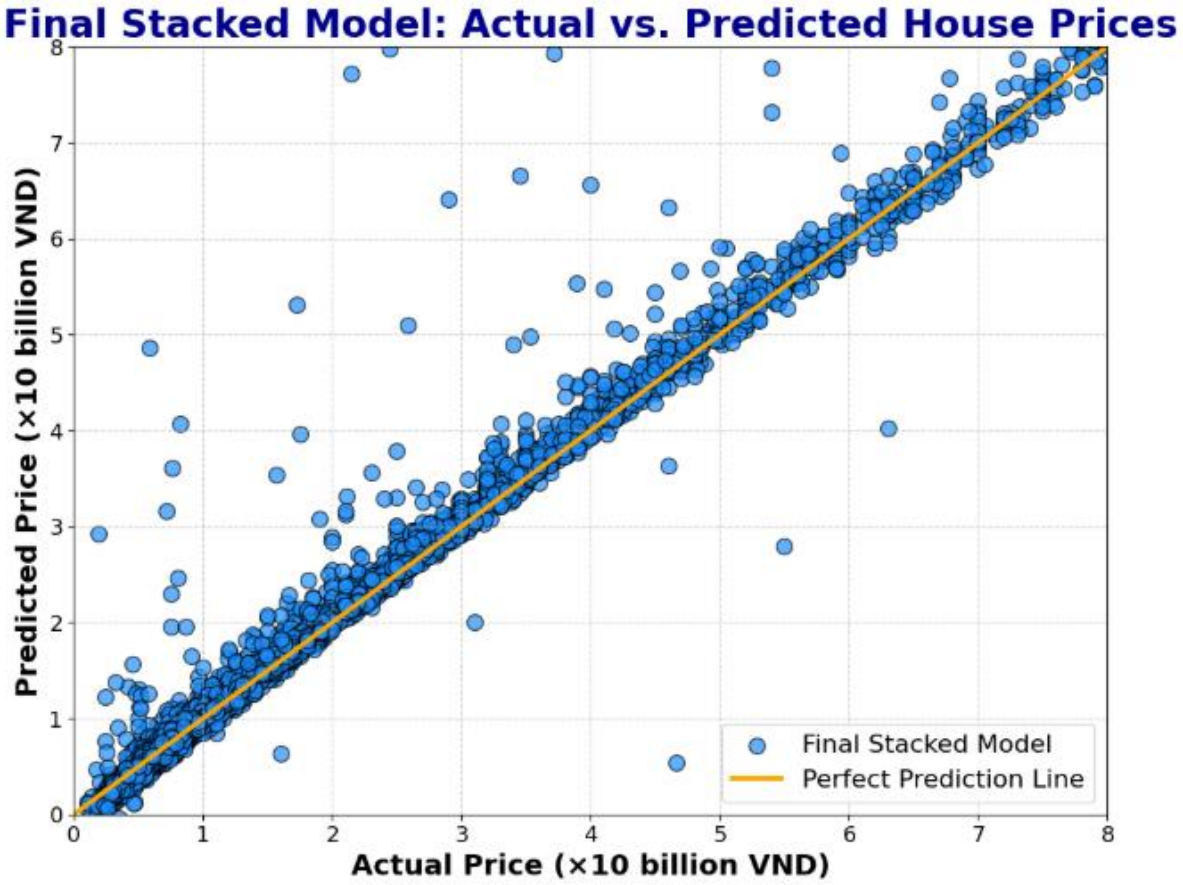
Bảng 6: Bảng đánh giá bốn mô hình ensemble

Model	RMSE	RMSLE	R^2
Random Forest	21.289	0.176	0.8094

XGBoost	21.030	0.203	0.8140
Linear Regression	37.237	0.879	0.4168
XGBoost + Linear Regression + Random Forest	19.271	0.195	0.8443
Random Forest + Linear Regression	20.391	0.245	0.8251
XGBoost + Random Forest	19.336	0.184	0.8428
XGBoost + Linear Regression	20.164	0.224	0.8290

Phân tích Bảng 6 cho thấy rõ ràng ưu thế của phương pháp kết hợp mô hình (ensemble learning) so với các phương pháp đơn lẻ. Các mô hình kết hợp tối ưu đạt được sự cải thiện toàn diện trên cả ba chỉ số đánh giá. Đặc biệt, khi so với Random Forest - mô hình đơn lẻ có RMSLE tốt nhất (0.176), mô hình kết hợp vẫn duy trì được sai số logarit ở mức tương đương (0.195) trong khi cải thiện đáng kể các chỉ số khác. Điều này cho thấy mô hình kết hợp không chỉ kế thừa ưu điểm của từng mô hình thành phần mà còn khắc phục được các nhược điểm riêng lẻ. Mô hình kết hợp giữa XGBoost, Linear Regression và Random Forest đạt hiệu suất tốt nhất trong số các mô hình kết hợp được thử nghiệm. Cụ thể, mô hình này có giá trị RMSE thấp nhất là 19.271, cho thấy sai số dự đoán trung bình thấp nhất. Đồng thời, nó cũng đạt được giá trị R2 cao nhất là 0.8443, nghĩa là 84.43% phương sai của biến phụ thuộc có thể được giải thích bởi các biến độc lập trong mô hình này.

Mặc dù mô hình XGBoost + Random Forest có giá trị RMSLE thấp nhất (0.184) trong các mô hình essemble, nhưng xét trên tổng thể, đặc biệt là với RMSE và R2, mô hình ba thành phần (XGBoost + Linear Regression + Random Forest) vẫn thể hiện sự vượt trội về khả năng dự đoán chính xác và giải thích dữ liệu. Các mô hình còn lại như Random Forest + Linear Regression và XGBoost + Linear Regression đều có hiệu suất thấp hơn rõ rệt so với mô hình ba thành phần và XGBoost + Random Forest.



Hình 6: Biểu đồ thể hiện sự tương quan giữa giá nhà thực tế và giá dự đoán bởi mô hình kết hợp cuối cùng

Do đó, dựa trên các tiêu chí đánh giá, mô hình XGBoost + Linear Regression + Random Forest được chọn là mô hình ensemble có hiệu suất tối ưu nhất cho bài toán này, mang lại khả năng dự đoán chính xác và độ tin cậy cao nhất.

Hình 6 (giá nhà thực tế so với giá nhà dự đoán) là một công cụ trực quan mạnh mẽ để đánh giá hiệu suất của mô hình Stacked cuối cùng trong việc dự đoán giá nhà. Trên biểu đồ này, trục hoành biểu thị giá nhà thực tế và trục tung biểu thị giá nhà dự đoán, cả hai đều được chuẩn hóa theo đơn vị "×10 tỷ VND". Đường chéo màu cam, được gọi là "Perfect Prediction Line", đóng vai trò là đường tham chiếu lý tưởng, nơi các giá trị dự đoán hoàn toàn trùng khớp với giá trị thực tế.

Sự phân bố của các điểm dữ liệu (chấm tròn màu xanh lam) là yếu tố then chốt để đánh giá mô hình. Quan sát tổng thể cho thấy phần lớn các điểm dữ liệu nằm rất sát

và tập trung dày đặc xung quanh đường "Perfect Prediction Line". Điều này minh chứng cho khả năng dự đoán mạnh mẽ và độ chính xác cao của mô hình Stacked. Việc các điểm gần như bám sát đường chéo trong một phạm vi rộng của giá trị cho thấy mô hình đã học được một mối quan hệ tuyến tính mạnh mẽ và hiệu quả giữa các đặc điểm đầu vào và giá nhà, giảm thiểu đáng kể sai số dự đoán.

Đặc biệt, ở phân khúc giá nhà thấp đến trung bình (khoảng từ 0 đến 5 ×10 tỷ VND), các điểm dữ liệu gần như hòa vào đường chéo, cho thấy độ chính xác gần như tuyệt đối của mô hình trong việc dự đoán giá các bất động sản phổ biến. Tuy nhiên, khi chuyển sang các giá trị giá cao hơn, mặc dù phần lớn các điểm vẫn bám sát đường chéo, chúng ta có thể quan sát thấy một số điểm bắt đầu có sự phân tán rộng hơn một chút. Sự phân tán này gợi ý rằng mô hình có thể đối mặt với thách thức nhỏ hơn trong việc dự đoán cực kỳ chính xác các bất động sản có giá trị đặc biệt cao hoặc có những đặc điểm ít phổ biến trong tập huấn luyện. Dù vậy, mức độ lệch lạc này vẫn nằm trong giới hạn chấp nhận được, và mô hình vẫn duy trì khả năng dự đoán đáng tin cậy.

Hơn nữa, việc không có xu hướng sai lệch rõ ràng cho thấy các sai số của mô hình là ngẫu nhiên, không có sự thiên vị dự đoán quá cao hoặc quá thấp đối với bất kỳ khoảng giá nào. Điều này củng cố tính tổng quát và độ tin cậy của mô hình khi áp dụng vào dữ liệu mới.

Tóm lại, biểu đồ này trực quan hóa một cách ấn tượng khả năng của mô hình Stacked cuối cùng. Nó không chỉ xác nhận hiệu suất định lượng cao (như đã phản ánh qua RMSE và R2) mà còn cho thấy mô hình hoạt động ổn định trên nhiều phân khúc giá, với độ chính xác đặc biệt cao ở phân khúc giá phổ biến và khả năng dự đoán đáng tin cậy ở các giá trị ngoại lai hơn.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Nghiên cứu này đã thành công trong việc xây dựng một bộ dữ liệu chất lượng cao và toàn diện về giá nhà tại Hà Nội. Quá trình thu thập, làm sạch và tiền xử lý dữ liệu cẩn thận, bao gồm các đặc trưng quan trọng như vị trí, diện tích, số phòng ngủ, và các tiện ích xung quanh, đã cung cấp nền tảng vững chắc cho việc huấn luyện và đánh giá mô hình. Điều này nhấn mạnh tầm quan trọng của một bộ dữ liệu đáng tin cậy trong việc đạt được kết quả dự đoán chính xác và có ý nghĩa trong các bài toán học máy ứng dụng thực tế.

Dựa trên bộ dữ liệu mà chúng tôi đã xây dựng được, nghiên cứu này đã đề xuất và đánh giá một mô hình kết hợp hiệu quả cho bài toán dự đoán giá nhà tại Hà Nội, tận dụng sức mạnh của ba thuật toán học máy mạnh mẽ: XGBoost, Random Forest, và Linear Regression. Bằng cách kết hợp khả năng nắm bắt các mối quan hệ phi tuyến phức tạp của XGBoost và RF với tính đơn giản và hiệu quả của LR trong việc mô hình hóa các mối quan hệ tuyến tính, mô hình đề xuất đã mang lại những kết quả dự đoán giá nhà đầy hứa hẹn.

Kết quả thực nghiệm trên bộ dữ liệu giá nhà thực tế tại Hà Nội đã chứng minh tính ưu việt của mô hình kết hợp so với việc sử dụng riêng lẻ từng thuật toán. Cụ thể: RMSE giảm từ 21.030 (XGBoost) xuống 19.271, tương ứng cải thiện 8.36%. RMSLE giảm từ 0.203 (XGBoost) xuống 0.195, cải thiện 3.94%. R^2 tăng từ 0.8140 (XGBoost) lên 0.8443, tương đương cải thiện 3.72%. So với Random Forest riêng lẻ, mô hình kết hợp giảm RMSE 9.48% (từ 21.289 xuống 19.271) và tăng R^2 4.31% (từ 0.8094 lên 0.8443). Điều này cho thấy việc kết hợp các mô hình khác nhau có khả năng khai thác được những khía cạnh khác nhau của dữ liệu, dẫn đến dự đoán chính xác và ổn định hơn.

Nghiên cứu cũng đã làm nổi bật vai trò của các yếu tố ảnh hưởng đến giá nhà tại Hà Nội, với phân tích tầm quan trọng của đặc trưng cho có tác động đáng kể đến giá trị bất động sản. Thông tin này có giá trị cho cả người mua, người bán và các nhà đầu

tư trong việc đưa ra quyết định sáng suốt.

5.2. Hướng phát triển

Dựa trên những kết quả đạt được, nghiên cứu này mở ra nhiều hướng phát triển tiềm năng để tiếp tục nâng cao hiệu quả và tính ứng dụng của mô hình dự đoán giá nhà tại Hà Nội:

- Tối ưu hoá kỹ thuật kết hợp: Nghiên cứu sâu hơn về các phương pháp kết hợp mô hình khác nhau, nghĩ ra hạn chế như sử dụng các kỹ thuật xếp chồng hoặc hòa trộn với các số quan trọng được học một cách tối ưu thay vì các phương pháp kết hợp đơn giản. Thử nghiệm các kết cấu kiến trúc phức tạp hơn có thể mang lại hiệu suất cao hơn.

- Xử lý đặc trưng nâng cao: Nghiên cứu và áp dụng các kỹ thuật xử lý đặc trưng phức tạp hơn, bao gồm tạo các đặc trưng tương tác giữa các biến hiện có, trích xuất đặc trưng từ dữ liệu không gian, và xử lý các đặc trưng thời gian.

- Tích hợp dữ liệu không gian và thời gian: Mở rộng mô hình để xem xét yếu tố không gian và yếu tố thời gian để dự đoán giá nhà một cách động hơn và chính xác hơn theo diễn biến thị trường.

- Ứng dụng các kỹ thuật Explainable AI (XAI): Tập trung vào việc giải thích kết quả dự đoán của mô hình kết hợp. Sử dụng các phương pháp XAI như SHAP (Shapley Additive Explanations) hoặc LIME (Local Interpretable Model-agnostic Explanations) để hiểu rõ hơn về đóng góp của từng đặc trưng và từng mô hình cơ sở vào kết quả dự đoán cuối cùng, tăng cường độ tin cậy và khả năng diễn giải của mô hình.

- Xây dựng hệ thống dự đoán trực tuyến: Phát triển một hệ thống hoặc ứng dụng web cho phép người dùng nhập thông tin về bất động sản và nhận được dự đoán giá dựa trên mô hình đã được huấn luyện. Hệ thống này có thể bao gồm các tính năng trực quan hóa kết quả và giải thích dự đoán.

- Đánh giá độ ổn định và khả năng khái quát hóa: Nghiên cứu sâu hơn về độ ổn định của mô hình trong các điều kiện thị trường khác nhau và đánh giá khả năng khái quát hóa của mô hình trên các khu vực địa lý khác hoặc các loại hình bất động sản khác.

Những hướng phát triển này sẽ góp phần làm cho mô hình dự đoán giá nhà tại Hà Nội trở nên mạnh mẽ, chính xác và hữu ích hơn trong thực tế, hỗ trợ tốt hơn cho các hoạt động giao dịch và đầu tư bất động sản.

TÀI LIỆU THAM KHẢO

[1]. Guide, G. P. (2024). Vietnam's Residential Property Market Analysis 2024.

[2]. Shahasane, A., Gosavi, M., Bhagat, A., Mishra, N., & Nerurkar, A. (2023). House Price Prediction Using Machine Learning.

[3]. Mostofi, F., Toğan, V., & Başağa, H. B. (2022). Real-estate price prediction with deep neural network and principal component analysis. Organization, Technology and Management in Construction: an International Journal, 14(1), 2741-2759.

[4]. Zhao, H., & Wang, K. (2023). Predicting Real Estate Price Using Stacking-Based Ensemble Learning. American Journal of Information Science and Technology, 7(2), 70-75.

[5]. Sibindi, R., Mwangi, R. W., & Waititu, A. G. (2023). A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. Engineering Reports, 5(4), e12599.

[6]. Kulkarni, S., Shajit, S., Mohite, A., Swati Sinha, D., & Student. (2021). House Price Prediction Using Ensemble Learning. 9, 2320–2882.

[7]. Renju, K., & Freni, S. (2024). An Ensemble Approach for Predicting The Price of Residential Property. International Journal of Information Technology, Research and Applications, 3(2), 27-38.

[8]. Chowhaan, M. J., Nitish, D., Akash, G., Sreevidya, N., & Shaik, S. (2023). Machine learning approach for house price prediction. Asian Journal of Research in Computer Science, 16(2), 54-61.

[9]. Tanamal, R., Rasyid Jr, N. M. K. S., Wiradinata, T., Soekamto, Y. S., & Saputri, T. R. D. (2023). House price prediction model using random forest in surabaya city.