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

Ngày nhận hồ sơ	
(Do CQ quản lý ghi)	

THUYẾT MINH
ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ CẤP SINH VIÊN 2024

THÔNG TIN CHUNG

A1. Tên đề tài

- Tên tiếng Việt (IN HOA): NGHIÊN CỨU TỐI ƯU MÔ HÌNH KẾT HỢP ỨNG DỤNG VÀO GIẢI QUYẾT BÀI TOÁN DỰ BÁO GIÁ NHÀ TẠI HÀ NỘI.
- Tên tiếng Anh (IN HOA): RESEARCH TO OPTIMIZE THE COMBINED MODEL AND APPLY IN SOLVING THE PROBLEM OF FORECASTING HOUSE PRICES IN HANOI.

A2. Thời gian thực hiện

..06.. tháng (kể từ khi được duyệt).

A3. Tổng kinh phí

(Lưu ý tính nhất quán giữa mục này và mục B8. Tổng hợp kinh phí đề nghị cấp)

Tổng kinh phí: …**6**.. triệu đồng, gồm

- Kinh phí từ Trường Đại học Công nghệ Thông tin: ..**6**.. triệu đồng

A4. Chủ nhiệm

Họ và tên: Nguyễn An Đức

Ngày, tháng, năm sinh: 06/11/2004 Giới tính (Nam/Nữ): Nam

Số CCCD: 052204006963 ; Ngày cấp: 06/02/2023 ; Nơi cấp: Bình Định

Mã số sinh viên: 22520268

Số điện thoại liên lạc: 0934894238

Đơn vị (Khoa): Hệ thống Thông Tin

Số tài khoản: 1030902679 Ngân hàng: Vietcombank

A5. Thành viên đề tài

TT	Họ tên	MSSV	Khoa
1	Nguyễn An Đức	22520268	Hệ thống Thông tin
2	Hà Nhật Thái	22521316	Hệ thống Thông tin

Nhóm nghiên cứu được sự hướng dẫn bởi TS. Nguyễn Thanh Bình, giảng viên khoa Hệ thống Thông tin, trường Đại học Công nghệ Thông tin, Đại học Quốc Gia Thành phố Hồ Chí Minh.

MÔ TẢ NGHIÊN CỨU

B1. Giới thiệu về đề tài

B1.1. Tổng quan tình hình nghiên cứu

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. Theo thống kê thì giá nhà trung bình tại Hà Nội đã tăng đáng kể 10,1% (tương đương 5,9% sau khi điều chỉnh lạm phát) lên 2.210 USD/m2 trong quý 1/2024 so với cùng kỳ năm trước. Đây là mức tăng tiếp nối sau khi giá nhà đã tăng 16,1% trong quý 1/2023. 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 **[1]**.

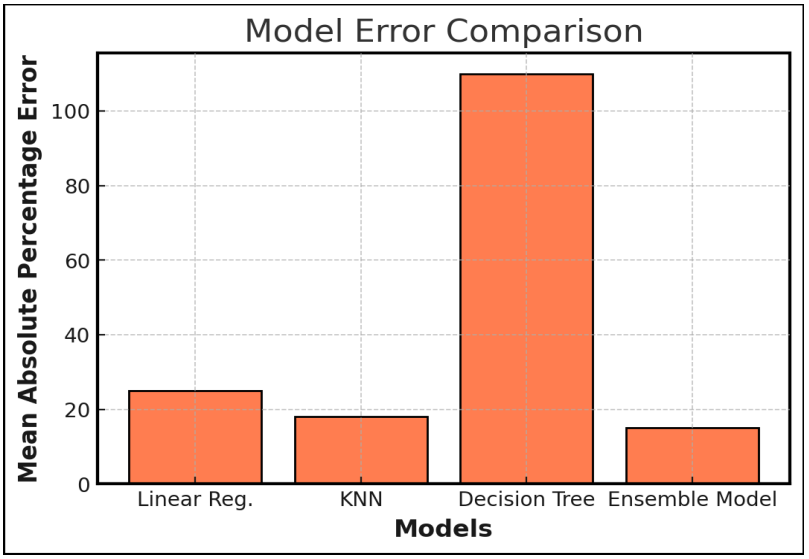
Ngày nay, cùng với sự phát triển mạnh mẽ của máy học và trí tuệ nhân tạo, có rất nhiều các nghiên cứu đã cho ra đời các mô hình dự đoán giá nhà và đã thu được một số kết quả khả quan. Trong đó có hai nhóm tiếp cận chính, mỗi nhóm dựa trên các phương pháp và mô hình khác nhau. Nhóm đầu tiên là các phương 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 **Hình 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]**.

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**

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

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.

B1.2. Tính ứng dụng

- Việc áp dụng mô hình kết hợp trong dự báo giá nhà tại Hà Nội hứa hẹn mang lại những kết quả vượt trội nhờ khả năng tận dụng ưu điểm và khắc phục nhược điểm của từng mô hình. Mô hình kết hợp này không chỉ nâng cao độ chính xác mà còn khắc phục những nhược điểm của mô hình riêng lẻ, giúp người mua nhà, doanh nghiệp bất động sản, và các tổ chức tài chính có được công cụ hỗ trợ quyết định hiệu quả và chính xác. Với quy trình xử lý dữ liệu và xây dựng mô hình khoa học, đề tài này có thể sẽ đóng góp tích cực vào lĩnh vực dự báo giá bất động sản không chỉ ở môi trường thực nghiệm cụ thể là Hà Nội mà còn ở bất kì thành phố nào trên thế giới.

- Kết quả nghiên cứu cơ bản này cũng sẽ làm cơ sở cho các nghiên cứu tiếp theo và mở ra nhiều cơ hội ứng dụng thực tiễn trong nhiều lĩnh vực, từ quy hoạch đô thị, tài chính, giáo dục đến công nghệ và dịch vụ khách hàng.

B1.3. Tổng quan phương pháp

- Thu thập dữ liệu:

+ Tập dữ liệu trong bài này được thu thập để dự đoán giá nhà ở Hà Nội bằng cách sử dụng một kỹ thuật được gọi là công cụ quét Web (Web Scraping). Nó giúp phân tích cú pháp các tài liệu HTML và XML. Nó phân tích cấu trúc HTML của một trang Web và cho phép chúng tôi tìm kiếm các thẻ, thuộc tính hoặc chuỗi văn bản HTML cụ thể trong đó. Sau khi xác định, chúng tôi có thể trích xuất dữ liệu mong muốn từ các phần tử này. Trong nghiên cứu này, chúng tôi đã sử dụng công cụ quét Web để phân tích qua các trang của Batdongsan.com, một trang Web niêm yết bất động sản tại Việt Nam. Chúng tôi đã xem các phần tử HTML chứa thông tin được thể hiện cụ thể trong **Hình 3** bao gồm: quận/huyện, phường/xã, loại hình nhà ở, giấy tờ pháp lý, số phòng ngủ, diện tích (m^2), chiều dài (m), chiều rộng (m), giá nhà (VND), giá/m^2 (đơn vị: nghìn đồng), phòng ngủ và phòng tắm.

Đặc trưng	Mô tả đặc trưng	Loại dữ liệu
Vị trí	Địa chỉ của căn nhà bao gồm: quận/huyện và phường/xã	Phi số
Loại hình nhà ở	Kiểu nhà ở như nhà phố, nhà trong hẻm	Phi số
Giấy tờ pháp lý	Các loại giấy tờ chứng nhận quyền sở hữu như sổ đỏ, sổ hồng	Phi số
Diện tích	Tổng diện tích của căn nhà (m^2)	Số
Chiều dài	Chiều dài của căn nhà (m)	Số
Chiều rộng	Chiều rộng của căn nhà (m)	Số
Giá nhà	Tổng giá trị của căn nhà (VND)	Số
Giá/m^2	Giá trị mỗi mét vuông của căn nhà (VND/m^2)	Số
Phòng ngủ	Số lượng phòng ngủ	Số
Phòng tắm	Số lượng phòng tắm	Số

Hình 3: Bảng các đặc trưng và loại dữ liệu.

+ Nhờ vào công cụ quét Web, chúng tôi có thể thu thập được một lượng lớn dữ liệu trên trang batdongsan.com một cách hiệu quả. Lưu trữ dữ liệu này vào cơ sở dữ liệu MySQL để sử dụng. Tập dữ liệu này làm cơ sở cho chúng tôi về dự đoán giá nhà tại Hà Nội.

- Tiền xử lý dữ liệu:

+ Ban đầu, các thông tin phi số sẽ được chuyển đổi thành tính năng số bằng cách sử dụng các kỹ thuật như One hot encoding, Label encoding. Việc giải quyết các dữ liệu còn thiếu là một bước quan trọng trong giai đoạn tiền xử lý. Các dữ liệu bị rỗng từ trong tập dữ liệu sẽ được thay thế bằng giá trị trung bình của các cột tương ứng. Ngoài ra, các kỹ thuật tiền xử lý khác đã được sử dụng để nâng cao chất lượng của tập dữ liệu nhằm nâng cao hiệu suất của các bước lập mô hình tiếp theo. Các kỹ thuật này có thể bao gồm chuẩn hóa để đảm bảo tất cả các tính năng đều ở quy mô tương đương, phát hiện và loại bỏ ngoại lệ để giảm thiểu tác động của các giá trị cực đoan, kỹ thuật tính năng để tạo các tính năng thông tin mới và kỹ thuật giảm kích thước như phân tích thành phần chính để giảm độ phức tạp của tập dữ liệu trong khi vẫn giữ được thông tin quan trọng. Mỗi kỹ thuật tiền xử lý này đóng một vai trò quan trọng trong việc chuẩn bị tập dữ liệu cho mô hình dự đoán chính xác và đáng tin cậy. Trước khi triển khai bất kỳ mô hình nào, phải xác thực tính chính xác và phù hợp của tập dữ liệu để phân tích. Để thực hiện điều này, chúng tôi đã tiến hành phân tích dữ liệu khám phá kỹ lưỡng, đi sâu vào các tính năng, thuộc tính của tập dữ liệu và mối quan hệ của chúng. Chỉ ra sự hiện diện và cường độ của mối tương quan giữa các biến này. Hệ số tương quan, một chỉ số bằng số nằm trong khoảng từ +1 đến -1, làm sáng tỏ mức độ liên kết giữa các 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.

- Mô hình kết hợp:

+ Sau khi hoàn thành quá trình khai phá dữ liệu, chúng tôi tiến hành tối ưu hóa dựa trên tính tương thích của từng mô hình thuật toán. Tập dữ liệu này sẽ được sử dụng để thực nghiệm trên ba thuật toán có độ chính xác cao nhất trong việc dự đoán giá nhà theo các nghiên cứu sau đây.

- 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 (**Hình 4**).

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**

Hình 4: Bảng đánh giá kết quả hiệu suất của các mô hình.

- 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 **Hình 5**. Kết quả chỉ ra rằng độ chính xác tối đa là 84,77% được đưa ra bởi thuật toán hồi quy tuyến tính.

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

Hình 5: Bảng so sánh độ chính xác của các mô hình.

- 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 (**Hình 6**).

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**

Hình 6: Bảng so sánh độ chính xác giữa các thuật toán khác nhau.

+ 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.

B1.4. Các thách thức

- Về dữ liệu:

+ Dữ liệu bất động sản có thể thiếu sót hoặc không đồng nhất. Thông tin như địa chỉ, loại hình nhà ở và diện tích có thể không đầy đủ hoặc không chính xác.

+ Dữ liệu chỉ thu thập từ một nguồn là trang batdongsan.com có thể làm nguồn dữ liệu không đa dạng, phong phú sẽ làm ảnh hưởng đến tính chính xác của mô hình.

- Về phương pháp:

+ Để mô hình hoạt động tốt, dữ liệu đầu vào cần được chuẩn bị kỹ lưỡng, bao gồm việc chuẩn hóa, giảm chiều dữ liệu, xử lý dữ liệu thiếu, và lựa chọn các đặc trưng quan trọng.

+ Trong quá trình tối ưu hóa sử dụng lượng dữ liệu hiện có, chúng tôi đối mặt với những hạn chế của các thuật toán máy học truyền thống. Ngoài ra, sự kết hợp của nhiều thuật toán lại với nhau sẽ làm tăng độ phức tạp về thời gian vì phải huấn luyện nhiều mô hình cùng một lúc.

- Về tính ứng dụng thực tiễn:

+ Thị trường bất động sản có thể thay đổi nhanh chóng do các yếu tố như chính sách quy hoạch, sự phát triển hạ tầng, hoặc thay đổi về nhu cầu của người mua. Mô hình cần được cập nhật liên tục để phản ánh đúng tình hình thực tế.

+ Ngoài các yếu tố cơ bản như diện tích, vị trí, và tiện ích, giá nhà còn bị ảnh hưởng bởi nhiều yếu tố khác như môi trường sống, an ninh, hay xu hướng đầu tư. Việc tích hợp các yếu tố này vào mô hình là một thách thức lớn.

B2. Mục tiêu, nội dung, kế hoạch nghiên cứu

B2.1 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.
- Xây dựng ứng dụng để triển khai mô hình dự báo giá nhà một cách trực quan dưới dạng ứng dụng tra cứu.

B2.2 Nội dung và phương pháp nghiên cứu

Nội dung 1: Khảo sát các công trình liên quan và tìm hiểu về bộ dữ liệu giá nhà tại Hà Nội.

Nội dung 1a: Khảo sát các công trình liên quan.

- Phương pháp nghiên cứu:

+ Khảo sát các mô hình học máy riêng lẻ trong dự báo giá nhà bao gồm dự báo giá nhà dựa trên phương pháp hồi quy, phương pháp học sâu,…

+ Rút ra các ưu điểm và nhược điểm của từng phương pháp, công nghệ.

- Kết quả dự kiến: Xác định những ưu điểm, nhược điểm và hạn chế của các phương pháp và công nghệ hiện đại. Từ đó định hướng đi nghiên cứu của đề tài.

Nội dung 1b: Xây dựng bộ dữ liệu giá bất động sản tại Hà Nội.

- Phương pháp nghiên cứu:

+ Tìm hiểu các bộ dữ liệu tốt trong các nghiên cứu trước đó.

+ Thu thập dữ liệu từ các trang Web bất động sản bằng công cụ quét Web bằng cách sử dụng thư viện trong Python là Selenium.

+ Tiền xử lý dữ liệu: xử lý giá trị thiếu, chuẩn hóa, phân loại.

- Kết quả dự kiến:

+ Bộ dữ liệu đa dạng và dồi dào với độ tin cậy và độ chính xác cao.

Nội dung 2: Tìm hiểu về mô hình kết hợp và các phương pháp kết hợp.

- Phương pháp nghiên cứu:

+ Tham khảo cơ sở lý thuyết, thực nghiệm của nghiên cứu, các video mô tả, các ứng dụng liên quan đến mô hình kết hợp và phương pháp xây dựng các mô hình kết hợp.

+ Xây dựng sơ bộ mô hình kết hợp.

- Kết quả dự kiến:

+ Nắm vững kiến thức về mô hình kết hợp và phương pháp kết hợp, tổng quan về kiến trúc và các đặc điểm. Hiểu về các thành phần, cách hoạt động cho mô hình kết hợp.

+ Nắm vững cách xây dựng mô hình kết hợp và phương pháp kết hợp xếp chồng cho XGBoost, hồi quy tuyến tính và rừng ngẫu nhiên.

Nội dung 3: Thực nghiệm và đánh giá kết quả.

- Phương pháp nghiên cứu:

+ Áp dụng phương pháp xếp chồng và tinh chỉnh các siêu tham số để phù hợp nhất cho việc huấn luyện mô hình.

+ Huấn luyện mô hình kết hợp dựa trên nội dung 1b.

+ Đánh giá mô hình: Kiểm tra hiệu suất mô hình trên tập kiểm tra bằng các chỉ số RMSE, R^2.

+ Triển khai và theo dõi: Triển khai mô hình vào môi trường thực tế, theo dõi hiệu suất, cập nhật định kỳ, và thu thập phản hồi để cải thiện.

- Kết quả dự kiến:

+ Đánh giá hiệu suất, độ chính xác, độ tin cậy của mô hình.

+ Triển khai cụ thể mô hình kết hợp có độ tin cậy và độ chính xác cao nhất.

Nội dung 4: Xây dựng ứng dụng Website dự báo giá nhà tại Hà Nội và kiểm thử

- Phương pháp:

+ Tìm hiểu cách xây dựng một ứng dụng Website.

+ Thực hiện xây dựng ứng dụng kết hợp với mô hình kết hợp.

+ Sửa lỗi và tối ưu hóa ứng dụng dựa trên phản hồi từ việc kiểm thử.

- Kết quả dự kiến:

+ Ứng dụng Website dự báo giá nhà ở Hà Nội.

B2.3 Kế hoạch nghiên cứu

- Giai đoạn 1: (Tuần 01 - 04)

+ Khảo sát bài toán dự đoán giá bất động sản từ các nghiên cứu trước đó.

+ Tìm hiểu tổng quan về các thuật toán máy học hướng truyền thống.

+ Tìm hiểu về bộ dữ liệu giá nhà tại Hà Nội.

+ Xây dựng một bộ dữ liệu cho bài toán dự đoán giá nhà bằng phương pháp học máy.

- Giai đoạn 2: (Tuần 05 - 10)

+ Cài đặt các mô hình riêng lẻ XGBoost, hồi quy tuyến tính và rừng ngẫu nhiên.

+ Cài đặt các mô hình kết hợp XGBoost, hồi quy tuyến tính và rừng ngẫu nhiên theo hướng tiếp cận đã đề xuất.

+ Thực nghiệm các thuật toán đã được cài đặt được trên bộ dữ liệu được xây dựng và so sánh từng mô hình.

- Giai đoạn 3: (Tuần 11 - 12)

+ Phân tích kết quả thực nghiệm, so sánh hiệu quả các mô hình kết hợp và chọn ra mô hình kết hợp tốt nhất.

- Giai đoạn 4: (Tuần 13 – 15)

+ Thiết kế giao diện cho Website.

+ Xây dựng Server API.

- Giai đoạn 5: (Tuần 16 – 20)

+ Xây dựng chức năng dự báo giá nhà.

+ Nghiên cứu Server cho hệ thống.

- Giai đoạn 6: (Tuần 21 – 22)

+ Hoàn thiện tất cả các chức năng.

+ Triển khai hệ thống lên môi trường.

+ Kiểm thử.

- Giai đoạn 7: (Tuần 23 – 24):

+ Xem và chỉnh sửa lại đề tài.

+ Viết báo cáo, nghiệm thu.

B3. 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.

B4. 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.

Ngày 15 tháng 8 năm 2024
Giảng viên hướng dẫn
(Ký và ghi rõ họ tên)

Nguyễn Thanh Bình

Ngày 15 tháng 8 năm 2024
Chủ nhiệm đề tài
(Ký và ghi rõ họ tên)

Nguyễn An Đức