Mục lục

[**1.** **Introduction** 2](#_Toc192779214)

[**2.** **Methodology** 2](#_Toc192779215)

[**3.** **Machine Learning Algorithms for House Price Prediction** 3](#_Toc192779216)

[**3.1.** **Lý thuyết giá trị thuộc tính (Lancaster, 1966)** 3](#_Toc192779217)

[**3.2.** **Mô hình giá hedonistic (Rosen, 1974)** 3](#_Toc192779218)

[**3.3.** **Sự chuyển dịch sang học máy** 4](#_Toc192779219)

[**3.4.** **Mối liên hệ giữa lý thuyết truyền thống và học máy** 4](#_Toc192779220)

[**4.** **Dataset and Feature Engineering** 5](#_Toc192779221)

[**4.1.** **Datasets** 5](#_Toc192779222)

[**4.2.** **Feature Engineering** 5](#_Toc192779223)

[**5.** **Machine Learning Techniques and Performance** 6](#_Toc192779224)

[**6.** **Research Gaps and Future Directions** 8](#_Toc192779225)

[**7.** **Conclusion** 8](#_Toc192779226)

[**References** 9](#_Toc192779227)

1. **Introduction**

Dự đoán giá nhà là một bài toán quan trọng trong lĩnh vực bất động sản và khoa học dữ liệu, đóng vai trò hỗ trợ ra quyết định cho người mua, người bán và các nhà hoạch định chính sách. Truyền thống, các mô hình hồi quy tuyến tính và mô hình giá hedonistic đã được sử dụng để ước lượng giá nhà dựa trên các đặc điểm vật lý và vị trí [2,3]. Tuy nhiên, những phương pháp này thường không đủ linh hoạt để xử lý dữ liệu phi tuyến hoặc tương tác phức tạp giữa các biến.

Sự phát triển của học máy (machine learning) đã mở ra nhiều hướng tiếp cận mới, cho phép xây dựng các mô hình dự đoán chính xác và mạnh mẽ hơn. Nhiều thuật toán như random forest, gradient boosting, SVM, KNN và neural networks đã được áp dụng trong các nghiên cứu gần đây [4–10], cho thấy hiệu quả vượt trội trong định giá bất động sản.

Bài khảo sát này nhằm tổng hợp và phân tích các kỹ thuật học máy được sử dụng trong dự đoán giá nhà, tập trung vào mô hình, dữ liệu đầu vào, kỹ thuật xử lý đặc trưng và đánh giá hiệu suất mô hình. Mục tiêu là cung cấp một cái nhìn toàn diện và cập nhật cho lĩnh vực đang phát triển này.

1. **Methodology**

Bài khảo sát này được thực hiện dựa trên việc tổng hợp và phân tích các nghiên cứu học thuật liên quan đến dự đoán giá nhà bằng thuật toán học máy. Nguồn tài liệu được lựa chọn từ các cơ sở dữ liệu khoa học uy tín như Google Scholar, IEEE Xplore, ScienceDirect và SpringerLink.

Các từ khóa chính được sử dụng trong quá trình tìm kiếm bao gồm:  
“house price prediction”, “machine learning”, “real estate valuation”, “hedonic pricing”, “housing dataset”.

Tiêu chí lựa chọn bao gồm:

* Nghiên cứu có ứng dụng thuật toán học máy vào dự đoán giá nhà;
* Có trình bày dữ liệu đầu vào và độ đo hiệu suất mô hình;
* Công bố trong các hội thảo, tạp chí học thuật uy tín;
* Ưu tiên các bài viết gần đây.

Tổng cộng 10 nghiên cứu tiêu biểu được chọn lọc làm cơ sở phân tích, bao gồm các công trình áp dụng các mô hình như Linear Regression, Random Forest, SVM, Gradient Boosting, KNN và Neural Networks. Những tài liệu này cung cấp một nền tảng đa dạng để đánh giá và so sánh các phương pháp tiếp cận trong lĩnh vực dự đoán giá bất động sản.

1. **Machine Learning Algorithms for House Price Prediction**

Việc định giá bất động sản là một chủ đề đã được nghiên cứu sâu rộng trong kinh tế học và bất động sản học từ nhiều thập kỷ trước. Các mô hình dự đoán giá nhà hiện đại dựa trên nền tảng của một số lý thuyết kinh tế cốt lõi, trong đó nổi bật nhất là Lý thuyết giá trị thuộc tính (Attribute Theory) của Lancaster [2] và Mô hình giá hedonistic (Hedonic Pricing Model) của Rosen [3].

* 1. **Lý thuyết giá trị thuộc tính (Lancaster, 1966)**

Theo Lancaster [2], giá trị của một hàng hóa không chỉ đến từ bản thân hàng hóa đó mà từ các đặc tính cấu thành (attributes) mà hàng hóa mang lại. Trong trường hợp bất động sản, một căn nhà không chỉ là một đơn vị sản phẩm duy nhất mà là một tổ hợp các đặc tính như diện tích, số phòng ngủ, vị trí địa lý, năm xây dựng, tiện ích giao thông, và môi trường sống xung quanh. Mỗi đặc điểm này đóng góp một phần vào giá trị tổng thể mà người mua sẵn sàng trả.

Lý thuyết này là nền tảng cho việc tách giá trị của một căn nhà thành các yếu tố thành phần – một nguyên tắc cốt lõi trong các mô hình định giá hiện nay.

* 1. **Mô hình giá hedonistic (Rosen, 1974)**

Tiếp nối ý tưởng của Lancaster, Rosen [3] phát triển Hedonic Pricing Model (HPM), mô tả giá của một sản phẩm như một hàm của các đặc tính định lượng và định tính. Trong thị trường bất động sản, giá nhà được mô hình hóa là một hàm tuyến tính hoặc phi tuyến của các đặc trưng như diện tích, chất lượng xây dựng, tiện nghi, vị trí, và các yếu tố môi trường.

Mô hình hedonistic được áp dụng rộng rãi trong các nghiên cứu kinh tế và định giá tài sản vì khả năng giải thích rõ ràng tác động của từng biến đặc trưng lên giá nhà. Tuy nhiên, khi số lượng biến lớn và quan hệ giữa các biến trở nên phức tạp, HPM truyền thống (thường dưới dạng hồi quy tuyến tính) không còn hiệu quả.

* 1. **Sự chuyển dịch sang học máy**

Mặc dù mô hình hedonistic là một khung lý thuyết mạnh mẽ, song các kỹ thuật hồi quy tuyến tính cổ điển gặp nhiều hạn chế:

* Không xử lý tốt các mối quan hệ phi tuyến giữa các đặc trưng;
* Nhạy cảm với outliers;
* Khó tối ưu khi có quá nhiều biến tương tác hoặc tương quan đa cộng tuyến.

Do đó, các nghiên cứu hiện đại đã chuyển sang áp dụng các thuật toán học máy, với khả năng học từ dữ liệu mà không cần giả định trước dạng thức mô hình. Các kỹ thuật như Random Forest, Gradient Boosting, SVM và Neural Networks không chỉ vượt trội về độ chính xác mà còn có khả năng phát hiện các mối quan hệ phức tạp trong dữ liệu mà mô hình tuyến tính không thể mô tả được.

Trong bối cảnh này, các mô hình học máy có thể được xem như một sự mở rộng tự nhiên của Hedonic Pricing Model, nhưng với khả năng xử lý phi tuyến và dữ liệu lớn.

* 1. **Mối liên hệ giữa lý thuyết truyền thống và học máy**

Một điểm quan trọng cần nhấn mạnh là: học máy không phủ nhận lý thuyết kinh tế truyền thống, mà ngược lại, nó kế thừa nền tảng đó và mở rộng phạm vi mô hình hóa bằng công cụ hiện đại. Việc hiểu rõ cơ sở lý thuyết của các đặc tính bất động sản vẫn là bước thiết yếu để lựa chọn đặc trưng đầu vào và đánh giá tính hợp lý của mô hình học máy.

1. **Dataset and Feature Engineering**

Dữ liệu là yếu tố cốt lõi trong bất kỳ mô hình học máy nào, và đối với bài toán dự đoán giá nhà, chất lượng của dữ liệu cùng với kỹ thuật xử lý đặc trưng (feature engineering) có ảnh hưởng trực tiếp đến độ chính xác của mô hình. Trong các nghiên cứu được khảo sát, nhiều bộ dữ liệu đã được sử dụng với sự đa dạng về nguồn gốc, quy mô và đặc trưng.

* 1. **Datasets**

Ames Housing Dataset: chứa hơn 2900 mẫu với hơn 70 đặc trưng về đặc điểm vật lý, vị trí và chất lượng xây dựng của căn nhà.

Melbourne Housing Dataset [10]: tập dữ liệu thực tế từ thành phố Melbourne, Australia, bao gồm thông tin vị trí, giá trị, ngày bán, v.v.

Korean Real Estate Dataset [7]: dữ liệu quy mô lớn phục vụ định giá bất động sản đại trà tại Hàn Quốc.

Vietnam Housing Dataset [4]: tập dữ liệu thực địa tại Việt Nam với nhiều biến phân loại và định lượng.

Slovenia Housing Dataset [6], India Housing Dataset [5]: được sử dụng để đánh giá khả năng mô hình hóa tại các thị trường khác nhau.

* 1. **Feature Engineering**

Feature Engineering là một bước then chốt trong quy trình xây dựng mô hình học máy, đặc biệt trong bài toán dự đoán giá nhà. Quá trình này giúp chuẩn hóa, biến đổi và lựa chọn các đặc trưng phù hợp nhằm cải thiện độ chính xác của mô hình và khả năng tổng quát hóa trên dữ liệu mới.

Việc xử lý đặc trưng không chỉ giúp cải thiện độ chính xác mô hình mà còn làm tăng khả năng diễn giải của mô hình, đặc biệt trong các phương pháp hồi quy tuyến tính. Những mô hình mạnh như Random Forest hay XGBoost vẫn phụ thuộc đáng kể vào chất lượng đặc trưng đầu vào.

**Table 1 - Datasets Used in Reviewed Studies**

|  |  |  |
| --- | --- | --- |
| Kỹ thuật | Mô tả | Nghiên cứu áp dụng |
| Xử lý missing values | Điền trung bình, median, hoặc loại bỏ dòng thiếu dữ liệu | [4], [6], [7] |
| Encoding biến phân loại | Label Encoding, One-Hot Encoding | [4], [8] |
| Scaling và Normalization | Chuẩn hóa dữ liệu liên tục để tránh bias | [4], [10] |
| Log transformation | Biến đổi log cho các biến skewed như SalePrice hoặc GrLivArea | [6], [8] |
| Feature Selection | Chọn đặc trưng dựa trên tương quan, importance từ tree models | [4], [7] |

Một số nghiên cứu còn sử dụng kỹ thuật tạo đặc trưng mới (feature construction) hoặc giảm chiều dữ liệu (PCA), đặc biệt trong các mô hình như ANN hoặc XGBoost [4], [9].

1. **Machine Learning Techniques and Performance**

Các mô hình học máy (machine learning – ML) được áp dụng ngày càng phổ biến trong lĩnh vực dự đoán giá nhà nhờ khả năng học từ dữ liệu và mô hình hóa các mối quan hệ phi tuyến phức tạp. Mỗi thuật toán có cơ chế hoạt động riêng, kèm theo những ưu điểm và hạn chế phụ thuộc vào bản chất của dữ liệu. Các thuật toán sẽ được đánh giá thông qua các độ đo như MAE, RMSE và R².

Từ các nghiên cứu khảo sát, có thể thấy rằng Random Forest, XGBoost và LightGBM là những mô hình cho kết quả dự đoán tốt nhất trên hầu hết tập dữ liệu, với R² thường từ 0.75–0.88 và RMSE thấp hơn so với các mô hình truyền thống. Ho et al. (2021) [8] và Truong et al. (2020) [4] đều khẳng định boosting models có ưu thế vượt trội trong hiệu suất.

Mặc dù Linear Regression vẫn được sử dụng như một baseline đơn giản và dễ giải thích, nhưng hiệu suất thường kém hơn đáng kể, đặc biệt trong dữ liệu phi tuyến hoặc đa chiều [5], [6]. ANN và SVM được ghi nhận là hiệu quả trong dữ liệu lớn hoặc phức tạp, nhưng thường đòi hỏi nhiều công đoạn tuning và thời gian huấn luyện hơn [4], [9], [10].

Ngoài ra, một số nghiên cứu đã thử nghiệm mô hình kết hợp (hybrid/stacked), tuy chưa phổ biến nhưng được xem là xu hướng tiềm năng trong tương lai [4].

**Table 2 - Performance Comparison**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Nghiên cứu | Mô hình | |  | | --- | |  |   MAE | RMSE | R² |
| Truong et al. [4] | Linear Regression | ~11,500 | ~14,500 | ~0.72 |
|  | Random Forest | ~9,000 | ~11,500 | ~0.81 |
|  | |  | | --- | |  |   XGBoost | ~8,000 | ~10,200 | 0.88 |
|  | ANN | ~8,500 | ~11,000 | ~0.85 |
|  | SVM | ~9,600 | ~12,000 | ~0.79 |
|  | KNN | ~10,000 | ~13,000 | ~0.76 |
| Ho et al. (2021) [8] | Linear Regression | ~22,000 | ~28,500 | ~0.70 |
|  | Random Forest | ~19,000 | ~25,000 | ~0.78 |
|  | XGBoost | ~18,500 | ~24,000 | ~0.80 |
|  | LightGBM | ~18,000 | ~23,500 | 0.82 |
| Phan (2018) [10] | SVM | ~27,000 | ~35,000 | ~0.74 |
|  | KNN | ~29,000 | ~37,000 | ~0.70 |

1. **Research Gaps and Future Directions**

Mặc dù các mô hình học máy đã cho thấy hiệu quả cao trong dự đoán giá nhà, nhưng vẫn tồn tại một số khoảng trống nghiên cứu cần được khắc phục. Thứ nhất, phần lớn nghiên cứu hiện nay vẫn chỉ sử dụng dữ liệu có cấu trúc, trong khi các yếu tố như hình ảnh, mô tả văn bản hoặc dữ liệu địa lý vẫn chưa được khai thác đầy đủ. Thứ hai, các mô hình phi tuyến như Random Forest hay XGBoost thường khó giải thích, đòi hỏi sự tích hợp các kỹ thuật giải thích mô hình như SHAP hoặc LIME. Bên cạnh đó, việc thiếu một bộ benchmark chuẩn cũng khiến cho việc so sánh giữa các nghiên cứu trở nên khó khăn. Cuối cùng, yếu tố thời gian chưa được đưa vào dự báo giá nhà, dù giá bất động sản luôn có xu hướng biến động theo thời gian.

Trong tương lai, các hướng nghiên cứu tiềm năng bao gồm: tích hợp dữ liệu phi cấu trúc, ứng dụng mô hình học sâu (Deep Learning), phát triển mô hình kết hợp (hybrid/ensemble) và cải thiện khả năng giải thích mô hình.

1. **Conclusion**

Bài khảo sát này đã tổng hợp và phân tích các kỹ thuật học máy được áp dụng trong dự đoán giá nhà, với trọng tâm là các mô hình phổ biến như Linear Regression, Random Forest, XGBoost, LightGBM, SVM, KNN và ANN. Qua các nghiên cứu được lựa chọn, có thể nhận thấy rằng các mô hình phi tuyến và mô hình ensemble như Random Forest, XGBoost và LightGBM thường cho kết quả dự đoán vượt trội, với độ chính xác cao và khả năng xử lý dữ liệu phức tạp tốt hơn so với các mô hình truyền thống.

Bên cạnh đó, chất lượng dữ liệu và kỹ thuật xử lý đặc trưng cũng đóng vai trò then chốt trong việc nâng cao hiệu suất dự đoán. Dữ liệu đầu vào càng được xử lý kỹ lưỡng – từ xử lý thiếu dữ liệu, chuẩn hóa đến lựa chọn đặc trưng – thì càng cải thiện độ chính xác của mô hình.

Tuy nhiên, bài toán dự đoán giá nhà vẫn còn tồn tại một số thách thức như thiếu dữ liệu phi cấu trúc, hạn chế trong khả năng diễn giải mô hình và chưa tích hợp yếu tố thời gian. Các nghiên cứu tương lai nên tập trung vào hướng phát triển mô hình lai (hybrid), tăng cường khả năng giải thích, và mở rộng ứng dụng sang dữ liệu thực tế phức tạp hơn.

Tổng thể, học máy đang đóng vai trò ngày càng quan trọng trong lĩnh vực bất động sản và có tiềm năng ứng dụng mạnh mẽ trong thực tiễn, nếu được kết hợp chặt chẽ giữa kỹ thuật mô hình, dữ liệu chất lượng và hiểu biết chuyên ngành.

# **References**

[1]. Kang, Y., Zhang, F., Peng, W., et al. (2021) Understanding house price appreciation using multi-source big geo-data and machine learning. Land Use Policy, 111, 104919.

[2]. Lancaster, K.J. (1966) A new approach to consumer theory. Journal of political economy, 74(2), 132-157.

[3]. Rosen, S. (1974) Hedonic prices and implicit markets: product differentiation in pure competition. Journal of political economy, 82(1), 34-55.

[4]. Truong, Q., Nguyen, M., Dang, H., et al. (2020) Housing price prediction via improved machine learning techniques. Procedia Computer Science, 174, 433-442.

[5]. Ghosalkar, N. N., Dhage, S. N. (2018) Real estate value prediction using linear regression. 2018 fourth international conference on computing communication control and automation (ICCUBEA). IEEE, 1-5.

[6]. Čeh, M., Kilibarda, M., Lisec, A., et al. (2018) Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. ISPRS international journal of geo-information, 7(5), 168.

[7]. Hong, J., Choi, H., Kim, W. (2020) A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. International Journal of Strategic Property Management.

[8]. Ho, W.K.O., Tang, B.S., Wong, S.W. (2021) Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1), 48-70.

[9]. Pai, P.F., Wang, W.C. (2020) Using machine learning models and actual transaction data for predicting real estate prices. Applied Sciences, 10(17), 5832.

[10]. Phan, T.D. (2018) Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. 2018 International conference on machine learning and data engineering (iCMLDE). IEEE, 35-42.