TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN



Khảo Sát Dữ Liệu Ames Housing

Môn Học: Phương Pháp Nghiên Cứu Khoa Học

Giảng viên: Đỗ Như Tài

Nhóm

Phạm Tấn Khương – 3122410191

Hoàng Vũ - 3122560089

Huỳnh Thanh Bình - 3122410033

Nguyễn Minh Tú - 3120411167

|  |  |  |
| --- | --- | --- |
| Hoàng Vũ |  | 100% |
| Huỳnh Thanh Bình |  | 100% |
| Phạm Tấn Khương |  | 100% |
| Nguyễn Minh Tú |  | 100% |

## 

**Mục Lục**

[1. Giới thiệu 4](#_Toc192197929)

[2. Tóm tắt 4](#_Toc192197930)

[3. Related Work (Công trình liên quan) 5](#_Toc192197931)

[4. Phương pháp nghiên cứu 7](#_Toc192197932)

[4.1. Xác định tài liệu nghiên cứu 7](#_Toc192197933)

[4.2. Sàng lọc tài liệu 8](#_Toc192197934)

[4.3 Quan Sát Bất Ngờ 10](#_Toc192197935)

[4.4 Kết Luận 10](#_Toc192197936)

[5. Review of Ames Housing (Đánh giá bộ dữ liệu Ames Housing) 11](#_Toc192197937)

[5.1 Tổng quan về dữ liệu: 11](#_Toc192197938)

[5.2 Ưu điểm: 11](#_Toc192197939)

[5.3 Nhược điểm: 11](#_Toc192197940)

[5.4 Khả năng ứng dụng: 11](#_Toc192197941)

[6. Analysis and Discussion(Phân tích và thảo luận) 11](#_Toc192197942)

[6.1. Nguồn gốc của Tập dữ liệu 12](#_Toc192197943)

[6.2. Vị trí của Tập dữ liệu 12](#_Toc192197944)

[6.3. Từ điển Dữ liệu 12](#_Toc192197945)

[6.4. Các Bài toán Liên quan 12](#_Toc192197946)

[6.5. Kết quả Đạt được và Độ đo Hiệu suất 12](#_Toc192197947)

[6.6. Các Bài Khảo sát và Nghiên cứu Liên quan 13](#_Toc192197948)

[6.7. Hạn chế và Hướng Nghiên cứu Tương lai 13](#_Toc192197949)

[7) Conclusion(kết luận) 13](#_Toc192197950)

## 1. Giới thiệu

Trong những năm gần đây, bộ dữ liệu Ames Housing đã thu hút sự quan tâm đáng kể của cộng đồng khoa học dữ liệu như một lựa chọn thay thế hiện đại cho bộ dữ liệu Boston Housing cổ điển. Được phát triển nhằm phản ánh điều kiện thị trường bất động sản hiện đại, [bộ dữ liệu này có nguồn gốc từ Ames, Iowa](https://www.tandfonline.com/doi/abs/10.1080/10691898.2011.11889627) – một thị trấn thuộc vùng Trung Tây của Hoa Kỳ với một mẫu số liệu ổn định và đại diện cho dữ liệu bất động sản dân cư. [Bộ dữ liệu](https://www.kaggle.com/datasets/prevek18/ames-housing-dataset), được chia sẻ trên Kaggle và do Shashank Necrothapa biên soạn, bao gồm các hồ sơ bán nhà từ năm 2006 đến 2010, với nhiều biến số phản ánh các đặc điểm vật lý (như diện tích đất, diện tích nhà ở, chất lượng xây dựng) và các yếu tố bối cảnh (như khu vực, phân vùng quy hoạch).

Mục tiêu chính của báo cáo này là khảo sát cấu trúc và các ứng dụng thực tiễn của bộ dữ liệu Ames Housing trong các bài toán phân tích hồi quy và dự án học máy. Với [hơn 2900 giao dịch](https://www.tmwr.org/ames) và [hơn 80 thuộc tính](https://hackmd.io/@jirvingphd/ames-data-dictionary), bộ dữ liệu cung cấp một môi trường phong phú để người nghiên cứu khám phá các thách thức trong xử lý dữ liệu, chẳng hạn như xử lý giá trị thiếu, biến đổi các biến số và giảm thiểu vấn đề đa cộng tuyến. Hơn nữa, bộ dữ liệu đã được sử dụng rộng rãi trong mục đích giáo dục, giúp sinh viên và nhà nghiên cứu xây dựng, đánh giá các mô hình dự đoán giá nhà từ những mô hình đơn giản như hồi quy tuyến tính cho đến các phương pháp phức tạp như các thuật toán tập hợp.

Thông qua việc sử dụng bộ dữ liệu này, người thực hành có thể nắm bắt được các yếu tố quan trọng ảnh hưởng đến giá trị bất động sản và phát triển các mô hình dự đoán mạnh mẽ, áp dụng được trong các tình huống thực tế. Báo cáo này sẽ trình bày chi tiết về nguồn gốc của bộ dữ liệu, các bước xử lý dữ liệu phổ biến cũng như các chiến lược mô hình hóa và đánh giá hiệu năng, từ đó nhấn mạnh tầm quan trọng của Ames Housing trong các ứng dụng hiện đại của khoa học dữ liệu.

## 2. Tóm tắt

Nghiên cứu này áp dụng tiêu chuẩn PRISMA để tổng hợp và đánh giá các bài báo liên quan đến dự đoán giá nhà sử dụng tập dữ liệu Ames Housing, được giới thiệu từ năm 2011 bởi Dean De Cock. Quy trình nghiên cứu bao gồm ba bước chính: (1) xác định tài liệu nghiên cứu thông qua các chiến lược tìm kiếm nâng cao trên Google Scholar với các từ khóa chuyên ngành; (2) sàng lọc và loại bỏ các bài báo không phù hợp hoặc trùng lặp thông qua việc đọc tiêu đề và tóm tắt; (3) phân tích chi tiết các bài báo còn lại, tập trung vào hướng đi nghiên cứu, phương pháp xử lý dữ liệu và cách giải quyết các bài toán dự đoán giá nhà.

Kết quả phân tích cho thấy tập dữ liệu Ames Housing, với [2.930 giao dịch](https://www.tmwr.org/ames) và [81 biến](https://www.tmwr.org/ames), đã được ứng dụng rộng rãi trong các nghiên cứu sử dụng nhiều mô hình học máy khác nhau như hồi quy tuyến tính, các thuật toán ensemble (XGBoost, LightGBM, CatBoost) và cả các mô hình mạng nơ-ron. Các bài báo nghiên cứu cũng nhấn mạnh các thách thức liên quan đến dữ liệu thiếu, đa cộng tuyến và việc lựa chọn đặc trưng phù hợp. Ngoài ra, quá trình đánh giá còn nhận thấy sự khác biệt về khả năng truy cập văn bản giữa các bài báo được xuất bản trên tạp chí truyền thống (có yêu cầu đăng ký hoặc trả phí) và các preprint trên arXiv (miễn phí).

Nhìn chung, Ames Housing là một nguồn dữ liệu phong phú và giá trị cho các bài toán dự đoán giá bất động sản, vừa phục vụ mục đích giảng dạy, vừa hỗ trợ các nghiên cứu thực tiễn trong lĩnh vực học máy. Các kết quả nghiên cứu tổng hợp từ các bài báo cho thấy rằng việc sử dụng tập dữ liệu này không chỉ giúp cải thiện độ chính xác của các mô hình dự đoán mà còn mở ra nhiều hướng nghiên cứu tiềm năng trong tương lai.

## 3. Related Work (Công trình liên quan)

1. **Các nghiên cứu liên quan trên Kaggle và các nguồn khác**

Tập dữ liệu Ames Housing đã được sử dụng trong nhiều nghiên cứu, bài viết, và notebook trên Kaggle, cũng như trong các nghiên cứu học thuật. Dưới đây là một số hướng nghiên cứu liên quan:

1. **Dự đoán giá nhà (House Price Prediction)**

* [Cuộc thi trên Kaggle](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques): Tập dữ liệu Ames Housing là nền tảng của cuộc thi "House Prices - Advanced Regression Techniques" trên Kaggle, nơi người tham gia xây dựng các mô hình để dự đoán SalePrice. Các phương pháp phổ biến bao gồm:
* [Hồi quy tuyến tính (Linear Regression)](https://www.kaggle.com/code/attackgnome/house-prices-advanced-regression-techniques): Một số người dùng trên Kaggle đã áp dụng hồi quy tuyến tính cơ bản để xác định các yếu tố quan trọng ảnh hưởng đến giá nhà, như OverallQual (chất lượng tổng thể) và GrLivArea (diện tích tầng trệt). Một notebook trên Kaggle đã chỉ ra rằng OverallQual có tương quan 0.79 với SalePrice, là yếu tố dự đoán mạnh.
* [Gradient Boosting](https://www.kaggle.com/code/davidanimaddo/ames-housing-prices-regression-xgboost): Các mô hình như XGBoost và LightGBM thường được sử dụng để đạt kết quả tốt hơn. Một số người tham gia đạt RMSE (Root Mean Squared Error) khoảng 0.11-0.12 trên tập kiểm tra bằng cách sử dụng XGBoost với các kỹ thuật xử lý giá trị thiếu và mã hóa biến phân loại.
* [Mạng nơ-ron (Neural Networks)](https://www.kaggle.com/code/erikbruin/house-prices-lasso-xgboost-and-a-detailed-eda): Một số nghiên cứu thử nghiệm mạng nơ-ron sâu, nhưng hiệu quả thường không vượt trội so với Gradient Boosting do tập dữ liệu có quy mô vừa phải.

1. **Phân tích dữ liệu khám phá (Exploratory Data Analysis - EDA)**

Nhiều notebook trên Kaggle tập trung vào EDA để hiểu rõ hơn về dữ liệu:

* Phân phối của SalePrice: Một phân tích phổ biến trên Kaggle cho thấy SalePrice có phân phối lệch phải (right-skewed). Người dùng thường áp dụng biến đổi logarit (log transformation) để làm cho phân phối gần với phân phối chuẩn hơn, giúp cải thiện hiệu suất của các mô hình hồi quy. [Dẫn chứng](https://nycdatascience.com/blog/python/eda-and-machine-learning-ames-housing-price-prediction-project)
* Tương quan giữa các biến: Ma trận tương quan (correlation matrix) thường được vẽ để xác định các biến có mối quan hệ mạnh với SalePrice. Ngoài OverallQual và GrLivArea, các biến như TotalBsmtSF (diện tích tầng hầm) và GarageCars (số lượng xe có thể đỗ trong gara) cũng cho thấy tương quan cao. [Dẫn chứng](https://rpubs.com/ryanward2102/1020667)
* Xử lý giá trị ngoại lai (Outliers): Một số người dùng đã phát hiện các giá trị ngoại lai, chẳng hạn như các ngôi nhà có diện tích lớn (GrLivArea > 4000) nhưng giá bán thấp bất thường, và loại bỏ chúng để cải thiện mô hình. [Dẫn chứng](https://feaz-book.com/outliers)

1. **Kỹ thuật xử lý đặc trưng (Feature Engineering)**

Tập dữ liệu Ames Housing có hơn 80 biến, nên kỹ thuật xử lý đặc trưng đóng vai trò quan trọng:

* Tạo đặc trưng mới: Một notebook trên Kaggle đã tạo đặc trưng TotalSF (tổng diện tích) bằng cách cộng GrLivArea và TotalBsmtSF, giúp cải thiện độ chính xác của mô hình. [Dẫn chứng](https://machinelearningmastery.com/feature-relationships-101/)
* Mã hóa biến phân loại: Các biến như Neighborhood, BldgType, và KitchenQual thường được mã hóa bằng kỹ thuật one-hot encoding hoặc target encoding để phản ánh sự khác biệt về giá giữa các danh mục. [Dẫn chứng](https://machinelearningmastery.com/classifying_variables/)
* Xử lý giá trị thiếu: Các cột như PoolQC, Alley, và Fence có tỷ lệ giá trị thiếu cao (trên 80%). Người dùng trên Kaggle thường điền giá trị "None" cho các cột này (vì giá trị thiếu thường có ý nghĩa, ví dụ: không có hồ bơi), hoặc loại bỏ các cột không cần thiết. [Dẫn chứng](https://cran.r-project.org/web/packages/AmesHousing/AmesHousing.pdf)
* Tuy nhiên, một số người dùng trên Kaggle nhận xét rằng Ames Housing có nhiều giá trị bị thiếu hơn Boston Housing, đòi hỏi kỹ thuật xử lý dữ liệu phức tạp hơn.

1. **Ứng dụng thực tế**

* Ngoài việc học tập và nghiên cứu, các kỹ thuật từ cuộc thi Kaggle đã được áp dụng trong ngành bất động sản. Ví dụ, các mô hình dự đoán giá nhà dựa trên Ames Housing có thể được sử dụng để định giá tự động (automated valuation models - AVMs) hoặc hỗ trợ các công ty bất động sản trong việc định giá tài sản.

1. **Bộ dữ liệu tương tự Ames Housing:**
   * [Bộ dữ liệu Boston Housing](https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html) (Harrison & Rubinfeld, 1978) là một tiền thân nổi tiếng, nhưng chỉ có 13 biến và 506 mẫu, ít chi tiết hơn so với Ames Housing (79 biến, 2.930 mẫu).
   * [Bộ dữ liệu California Housing](https://inria.github.io/scikit-learn-mooc/python_scripts/datasets_california_housing.html) cũng được sử dụng rộng rãi, nhưng tập trung vào dữ liệu tổng hợp cấp khu vực hơn là chi tiết từng ngôi nhà như Ames Housing.
2. **Một số nghiên cứu tiêu biểu có thể kể đến:**

* *"*[*Data Analysis on the Ames Housing Dataset"*](https://nycdatascience.com/blog/student-works/data-analysis-on-the-ames-housing-dataset) (NYC Data Science Academy): Trình bày các đặc điểm của dữ liệu và cách tiếp cận trong xây dựng mô hình dự đoán giá nhà.
* *"*[*Dự báo giá nhà bằng hồi quy Lasso và Ridge*](https://rpubs.com/NgKhanh2203/1280516)*"* (RPubs): Ứng dụng các kỹ thuật hồi quy để tối ưu hóa việc dự đoán giá nhà.
* *"Predicting Housing Prices Using Machine Learning"* (Kaggle): Nghiên cứu sử dụng nhiều mô hình học máy khác nhau để tối ưu dự đoán giá nhà dựa trên tập dữ liệu Ames.

## 4. Phương pháp nghiên cứu

Việc áp dụng một phương pháp nghiên cứu tiêu chuẩn không chỉ giúp đảm bảo chất lượng nghiên cứu mà còn tạo điều kiện để các nhà nghiên cứu khác có thể tái lập quy trình đánh giá. Do đó, nghiên cứu này sử dụng tiêu chuẩn PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) để thực hiện quá trình đánh giá. Vì nghiên cứu chỉ tập trung vào một cơ sở dữ liệu duy nhất, nên chỉ cần thực hiện một số điều chỉnh nhỏ, được trình bày trong các bước sau đây.

### 4.1. Xác định tài liệu nghiên cứu

Là một nghiên cứu tổng hợp, việc thu thập và phân loại các bài báo liên quan đến dự đoán giá nhà ở tại Ames, Iowa là bước đầu tiên. Nghiên cứu này tập trung vào việc xác định các xu hướng hiện tại, khoảng trống nghiên cứu và những từ khóa quan trọng thường xuyên xuất hiện trong lĩnh vực dự đoán giá nhà.

Sử dụng Google Scholar làm công cụ tìm kiếm ban đầu đã trả về hàng triệu kết quả, tuy nhiên, phần lớn trong số đó không liên quan hoặc không phù hợp với mục tiêu nghiên cứu. Các bài tổng quan trong lĩnh vực này thường thu thập tài liệu từ một cơ sở dữ liệu của Kaggle.

Quá trình tìm kiếm được thực hiện bằng cách sử dụng chiến lược tìm kiếm nâng cao, tập trung vào tiêu đề, tóm tắt và các từ khóa do tác giả chỉ định. Chỉ những bài báo từ năm 202 trở đi mới được xem xét để đảm bảo tính cập nhật của nghiên cứu. Các từ khóa được lựa chọn kỹ lưỡng để tối ưu hóa kết quả tìm kiếm, bao gồm:

* Các thuật ngữ về phương pháp dự đoán giá nhà: “Ames housing dataset”, “Dự đoán giá nhà”, “Giá trị tài sản”, “Hồi quy tuyến tính”, “Học máy”, “Mạng nơ-ron nhân tạo”, “Học sâu”...
* Các thuật ngữ về lĩnh vực ứng dụng: “Thị trường nhà ở”, “Giao dịch bất động sản”, “Thông tin nhà đất Ames”, “Đặc điểm nhà ở”, “Chỉ số kinh tế địa phương”...

### 4.2. Sàng lọc tài liệu

Mục tiêu của bước này là loại bỏ các bài báo trùng lặp và không phù hợp. Do giới hạn về số lượng toán tử trong truy vấn tìm kiếm.

Sau đó, nhóm nghiên cứu tiếp tục đánh giá mức độ phù hợp của các bài báo còn lại bằng cách đọc tiêu đề và tóm tắt. Kết quả cho thấy nhiều bài báo không liên quan trực tiếp đến dự đoán giá nhà ở Ames. Một số từ khóa có thể mang nhiều ý nghĩa khác nhau, chẳng hạn:

* Từ “định giá” có thể liên quan đến thẩm định tài sản tài chính thay vì bất động sản.
* Từ “hồi quy” có thể xuất hiện trong các nghiên cứu về kinh tế vĩ mô thay vì dự đoán giá nhà.
* Từ “đặc điểm nhà ở” có thể được sử dụng trong các nghiên cứu nhân khẩu học thay vì phân tích dữ liệu Ames Housing.

Ngoài ra, các bài báo không sử dụng mô hình học máy để phân tích dữ liệu Ames Housing cũng bị loại trừ. Sau quá trình sàng lọc, chỉ những nghiên cứu có liên quan nhất mới được giữ lại để tiếp tục phân tích sâu hơn.

Bảng Phân Tích Các bài báo về dữ liệu Ames Housing

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Tác Giả | Năm | Hướng Đi | Input | Output | Cách Giải Quyết | | Dean De Cock | 2011 | Giới thiệu tập dữ liệu cho mục đích giảng dạy hồi quy | Dữ liệu Ames Housing (2.930 quan sát, 79 biến) | Công cụ giáo dục cho dự án hồi quy | Thảo luận cách sử dụng trong khóa học, gợi ý xử lý quan sát bất thường. | | Nor Hamizah Zulkifley et al. | 2020 | Khảo sát kỹ thuật máy học cho dự đoán giá nhà | Dữ liệu Ames và các tập khác | So sánh hiệu suất mô hình | Phân tích văn học, so sánh các mô hình như ANN, SVR, XGBoost trên các thuộc tính. | | Sureyya Akyüz et al. | 2023 | Đề xuất mô hình lai dự đoán giá nhà | Dữ liệu Ames và Istanbul | Giá dự đoán, chỉ số RMSE, MAE | Kết hợp hồi quy tuyến tính, phân cụm, láng giềng gần, SVR, tối ưu hóa hiệu suất. | | Hemlata Sharma et al. | 2024 | Thuật toán dự đoán giá nhà tối ưu với XGBoost | Dữ liệu Ames (2.930 bản ghi, 82 biến) | Giá nhà dự đoán | Sử dụng XGBoost, tối ưu hóa tham số, cung cấp mã và dữ liệu trên GitHub và Kaggle. | | Muhammad Arbab Arshad et al. | 2024 | So sánh tầm quan trọng dự đoán và nhân quả | Dữ liệu Ames Housing  (2.900 quan sát, 79 biến) | Mối quan hệ nhân quả, độ chính xác | Áp dụng CatBoost và LightGBM, phân tích SHAP values, so sánh hiệu quả mô hình. | |

|  |  |
| --- | --- |
| Chế Độ Lọc | Đặc Tả |
| Năm xuất bản | 2011 trở đi |
| Loại Tài Liệu | Bài Báo Khoa Học |
| Ngôn Ngữ | Tiếng Anh |
| Trạng Thái Xuất Bản | Bản cuối cùng hoặc bản thảo |
| Khả Năng Truy Cập Các Văn Bản | Có Khả Năng Truy Cập Miễn Phí |

Phân Tích Chi Tiết

* Bài Báo 1 (De Cock, 2011): Bài báo này được xuất bản trong Journal of Statistics Education, một tạp chí được bình duyệt, vào năm 2011. Liên kết dẫn đến trang tóm tắt trên [Taylor & Francis](https://www.tandfonline.com/doi/abs/10.1080/10691898.2011.11889627), nơi văn bản đầy đủ yêu cầu đăng ký hoặc trả phí, phù hợp với mô hình xuất bản truyền thống của các nhà xuất bản học thuật lớn.
* Bài Báo 2 (Zulkifley et al., 2020): Bài báo này được xuất bản trong International Journal of Modern Education and Computer Science, một tạp chí được bình duyệt, vào năm 2020. Liên kết [MECS Press](https://www.mecs-press.org/ijmecs/ijmecs-v12-n6/v12n6-4.html) cung cấp văn bản đầy đủ miễn phí trực tuyến, điều này không phổ biến với các tạp chí thương mại nhưng có thể phản ánh chính sách mở của nhà xuất bản này.
* Bài Báo 3 (Akyüz et al., 2023): Bài báo này được xuất bản trong Computational Economics, một tạp chí được bình duyệt, vào năm 2023. Liên kết [Springer](https://link.springer.com/article/10.1007/s10614-022-10298-8) dẫn đến trang bài báo, nơi văn bản đầy đủ yêu cầu đăng ký hoặc trả phí, phù hợp với mô hình xuất bản của Springer, trừ khi bài báo được chỉ định là mở (open access), điều này không được ghi nhận ở đây.
* Bài Báo 4 (Sharma et al., 2024): Bài báo này là bản thảo trước (preprint) trên arXiv, với mã số arXiv:2402.04082, được đăng vào năm 2024. Như là một preprint, nó chưa được bình duyệt và văn bản đầy đủ có thể truy cập miễn phí trực tuyến trên [arXiv](https://arXiv.org/abs/2402.04082), phù hợp với chính sách của arXiv.
* Bài Báo 5 (Arshad và Li, 2024): Tương tự, bài báo này là bản thảo trước trên arXiv, với mã số arXiv:2409.02130, được đăng vào năm 2024. Văn bản đầy đủ cũng có thể truy cập miễn phí trực tuyến trên [arXiv](https://arXiv.org/abs/2409.02130), phản ánh tính chất mở của nền tảng này.

### 4.3 Quan Sát Bất Ngờ

Một chi tiết đáng chú ý là sự khác biệt trong khả năng truy cập văn bản giữa các bài báo xuất bản trong tạp chí (thường yêu cầu đăng ký) và các preprint trên arXiv (luôn miễn phí). Điều này có thể ảnh hưởng đến việc tiếp cận nghiên cứu, đặc biệt đối với các nhà nghiên cứu không có quyền truy cập vào các cơ sở dữ liệu trả phí.

### 4.4 Kết Luận

Dựa trên phân tích, có 3 bài báo được xuất bản trong tạp chí được bình duyệt (De Cock, 2011; Zulkifley et al., 2020; Akyüz et al., 2023), trong đó 2 bài yêu cầu đăng ký để truy cập văn bản đầy đủ, trong khi 1 bài có văn bản đầy đủ miễn phí. Hai bài còn lại là preprint trên arXiv (Sharma et al., 2024; Arshad và Li, 2024), cả hai đều có văn bản đầy đủ miễn phí trực tuyến. Người dùng có thể tham khảo các liên kết được cung cấp để truy cập, với lưu ý rằng một số có thể yêu cầu đăng ký hoặc trả phí.

Key Citations

* [Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project Journal of Statistics Education](https://www.tandfonline.com/doi/abs/10.1080/10691898.2011.11889627)
* [House Price Prediction using a Machine Learning Model: A Survey of Literature International Journal of Modern Education and Computer Science](https://www.mecs-press.org/ijmecs/ijmecs-v12-n6/v12n6-4.html)
* [A Novel Hybrid House Price Prediction Model Computational Economics Springer](https://link.springer.com/article/10.1007/s10614-022-10298-8)
* [An Optimal House Price Prediction Algorithm: XGBoost arXiv Preprint](https://arXiv.org/abs/2402.04082)
* [From Predictive Importance to Causality: Which Machine Learning Model Reflects Reality? arXiv Preprint](https://arXiv.org/abs/2409.02130)

## 5. Review of Ames Housing (Đánh giá bộ dữ liệu Ames Housing)

Nội dung:

### 5.1 Tổng quan về dữ liệu:

* + Ames Housing chứa thông tin về [2.930 giao dịch](https://www.tmwr.org/ames) nhà ở tại Ames, Iowa từ 2006-2010, [với 79 biến độc lập và 1 biến phụ thuộc](https://hackmd.io/@jirvingphd/ames-data-dictionary) (SalePrice).
  + Các biến bao gồm cả định lượng (quantitative) như diện tích (GrLivArea), số phòng ngủ (BedroomAbvGr), và định tính (qualitative) như chất lượng tổng thể (OverallQual), tình trạng nhà (SaleCondition).
  1. **Ưu điểm:**
  + Độ chi tiết: Với 79 biến, bộ dữ liệu cho phép phân tích sâu về nhiều khía cạnh của ngôi nhà, từ cấu trúc vật lý (diện tích, số tầng) đến yếu tố thẩm mỹ (chất lượng hoàn thiện).
  + Tính thực tế: Dữ liệu dựa trên giao dịch thực tế, phản ánh thị trường bất động sản tại một khu vực cụ thể.
  + Ứng dụng học máy: Phù hợp với các bài toán hồi quy, phân tích tương quan, và thử nghiệm các mô hình từ đơn giản đến phức tạp.

### 5.3 Nhược điểm:

* + Dữ liệu thiếu: Một số biến như PoolQC (chất lượng hồ bơi), MiscFeature (tính năng khác) có tỷ lệ thiếu dữ liệu cao (trên 90%), đòi hỏi kỹ thuật xử lý dữ liệu phức tạp.
  + Phạm vi giới hạn: Chỉ đại diện cho Ames, Iowa trong giai đoạn 2006-2010, không phản ánh xu hướng thị trường hiện tại (2025) hoặc các khu vực khác.
  + Phân phối lệch: Biến mục tiêu SalePrice có phân phối lệch phải, cần biến đổi để phù hợp với một số mô hình.

### 5.4 Khả năng ứng dụng:

* + Dự đoán giá nhà bằng các mô hình như XGBoost, Random Forest.
  + Phân tích các yếu tố quan trọng ảnh hưởng đến giá trị bất động sản (ví dụ: diện tích, chất lượng).
  + Dùng làm tài liệu giảng dạy trong các khóa học về phân tích dữ liệu và học máy.

**6. Analysis and Discussion(Phân tích và thảo luận)**

### **6.1. Nguồn gốc của Tập dữ liệu**

Tập dữ liệu **Ames Housing** được giới thiệu bởi **Dean De Cock** vào năm 2011 như một giải pháp thay thế cho tập dữ liệu **Boston Housing**, vốn thường được sử dụng trong các bài toán hồi quy. Tập dữ liệu này được trình bày trong bài báo: **"**[Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project](https://www.tandfonline.com/doi/abs/10.1080/10691898.2011.11889627)**"** được xuất bản trên Tạp chí Giáo dục Thống kê.

### **6.2. Vị trí của Tập dữ liệu**

Tập dữ liệu có sẵn công khai và có thể được truy cập từ nhiều nguồn, bao gồm:

* Kaggle (<https://www.kaggle.com/datasets/prevek18/ames-housing-dataset>)
* UCI Machine Learning Repository
* OpenML
* Bài báo gốc của De Cock

### **6.3. Từ điển Dữ liệu**

Tập dữ liệu chứa **2.930 quan sát** và **81 thuộc tính** mô tả các đặc điểm khác nhau của nhà ở tại Ames, Iowa. Một số biến quan trọng bao gồm:

* **SalePrice** (biến mục tiêu): Giá bán của ngôi nhà (USD).
* **LotArea**: Tổng diện tích lô đất (square feet).
* **OverallQual**: Chất lượng tổng thể của vật liệu và hoàn thiện (thang điểm 1-10).
* **GrLivArea**: Diện tích sử dụng trên mặt đất (square feet).
* **YearBuilt**: Năm xây dựng ngôi nhà.
* **GarageCars**: Số lượng chỗ đậu xe trong gara.
* **TotalBsmtSF**: Tổng diện tích tầng hầm.

Để xem đầy đủ từ điển dữ liệu, bạn có thể tham khảo bài báo gốc hoặc các nguồn như Kaggle.

### **6.4. Các Bài toán Liên quan**

Tập dữ liệu Ames Housing được sử dụng rộng rãi trong **học máy** và **phân tích thống kê** cho nhiều bài toán khác nhau, bao gồm:

* **Hồi quy:** Dự đoán giá nhà dựa trên các đặc điểm cấu trúc.
* **Lựa chọn đặc trưng:** Xác định những yếu tố ảnh hưởng lớn nhất đến giá nhà.
* **Tiền xử lý dữ liệu:** Xử lý giá trị khuyết và biến phân loại.
* **Phân tích khám phá dữ liệu (EDA):** Tìm hiểu xu hướng và mối tương quan giữa các biến.
* **Phân tích không gian địa lý:** Xem xét sự biến động giá nhà theo từng khu vực ở Ames, Iowa.

### **6.5. Kết quả Đạt được và Độ đo Hiệu suất**

Nhiều mô hình học máy đã được áp dụng cho tập dữ liệu này, sử dụng các độ đo như:

* **Lỗi trung bình tuyệt đối (MAE)**
* **Lỗi bình phương trung bình (MSE)**
* **Lỗi căn bậc hai trung bình (RMSE)**
* **Hệ số xác định (R²)**

Kết quả điển hình:

* **Hồi quy tuyến tính:** RMSE ~30.000-35.000 USD. [Dẫn chứng](https://rpubs.com/ryanward2102/1020667)
* **Random Forest:** RMSE ~24.000-28.000 USD. [Dẫn chứng](https://nycdatascience.com/blog/student-works/analyzing-data-to-understand-housing-prices-in-ames/)
* **Gradient Boosting (XGBoost, LightGBM, CatBoost):** RMSE ~19.000-24.000 USD. [Dẫn chứng](https://arxiv.org/abs/2402.04082)

### **6.6. Các Bài Khảo sát và Nghiên cứu Liên quan**

Nhiều nghiên cứu và bài viết đã phân tích tập dữ liệu Ames Housing, bao gồm:

* Các bài báo khoa học về **kỹ thuật đặc trưng và lựa chọn mô hình** trong dự đoán giá nhà. [Dẫn chứng](https://tapchi.vnua.edu.vn/wp-content/uploads/2021/08/tap-chi-so-9.1.5.pdf)
* Các cuộc thi trên Kaggle sử dụng tập dữ liệu này để tối ưu hóa mô hình hồi quy. [Dẫn chứng](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques)
* Các tài liệu mã nguồn mở và notebook Jupyter trình bày phân tích khám phá và mô hình dự đoán. [Dẫn chứng](https://www.kaggle.com/code/ngvnminh/d-o-n-gi-nh)

### **6.7. Hạn chế và Hướng Nghiên cứu Tương lai**

Dù được sử dụng rộng rãi, tập dữ liệu Ames Housing vẫn có một số hạn chế:

* **Dữ liệu chỉ áp dụng cho thành phố Ames, Iowa**, hạn chế khả năng tổng quát hóa.
* **Biến phân loại cần được xử lý kỹ lưỡng**, có thể làm phức tạp quá trình mô hình hóa.
* **Một số biến có tỷ lệ giá trị khuyết cao**, đòi hỏi các chiến lược bổ sung dữ liệu phù hợp.

Các hướng nghiên cứu tương lai có thể bao gồm:

* **Mô hình học sâu (Neural Networks)** để cải thiện độ chính xác.
* **Phân tích chuỗi thời gian** để xem xét xu hướng thị trường nhà ở theo thời gian.
* **Tích hợp dữ liệu bên ngoài** (chỉ số kinh tế, dữ liệu khí hậu) để nâng cao chất lượng dự đoán.

**7) Conclusion(kết luận)**

Tập dữ liệu Ames Housing là một bộ dữ liệu phong phú và đa dạng, phù hợp cho các bài toán **dự đoán giá bất động sản** và nghiên cứu học máy. Nó cung cấp một tập dữ liệu thực tế với nhiều đặc trưng quan trọng, giúp cải thiện khả năng thực hành của các nhà khoa học dữ liệu và sinh viên.