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% |

## 

## 3. 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.

### 3.1. 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”...

### 3.2. 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.

## 3.3 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.

## 3.4 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)