**ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue circle with white text

AI-generated content may be incorrect.

**PHƯƠNG PHÁP NGHIÊN CỨU KHOA HỌC**

**TRONG CÔNG NGHỆ THÔNG TIN**

**[BÀI TẬP 3:] KHẢO SÁT BÀI BÁO CHO DỮ LIỆU AMES HOUSING**

Sinh viên thực hiện:

**Văn Hoàng Như Ý - 3122410493**

**Nguyễn Hoàng Thanh Phương - 3122410329**

**Lê Hồng Sơn** - **3121410423**

**Đỗ Hữu Lộc - 3123410201**

Giảng viên: **Đỗ Như Tài**

Thành phố Hồ Chí Minh, ngày 06 tháng 03 năm 2025

A black background with a black square

AI-generated content may be incorrect.

**MỤC LỤC**

[**Bảng phân công công việc** 1](#_Toc192173008)

[**1. Giới thiệu chung về bộ dữ liệu Ames Housing** 2](#_Toc192173009)

[**2. Nguồn gốc và tài liệu tham khảo** 2](#_Toc192173010)

[**3. Dữ liệu và từ điển dữ liệu** 2](#_Toc192173011)

[**4. Các bài toán liên quan** 3](#_Toc192173012)

[**5. Kết quả đạt được và độ đo đánh giá** 4](#_Toc192173013)

[**6. Bài khảo sát về bộ dữ liệu** 4](#_Toc192173014)

[**7. Kết luận** 5](#_Toc192173015)

# **Bảng phân công công việc**

|  |  |  |  |
| --- | --- | --- | --- |
| **MSSV** | **Họ tên** | **Công việc** | **Mức độ hoàn thành** |
| 3122410493 | Văn Hoàng Như Ý | Phần 1 & 2 | 100% |
| 3122410329 | Nguyễn Hoàng Thanh Phương | Phần 3 & 4 | 100% |
| 3123410201 | Đỗ Hữu Lộc | Phần 5 & 6 | 100% |
| 3121410423 | Lê Hồng Sơn | Phần 7 & slide | 100% |

**KHẢO SÁT BÀI BÁO CHO DỮ LIỆU AMES HOUSING**

# **1. Giới thiệu chung về bộ dữ liệu Ames Housing**

Bộ dữ liệu Ames Housing mô tả thông tin về các giao dịch bất động sản tại thành phố Ames, bang Iowa, Hoa Kỳ, trong giai đoạn từ năm 2006 đến 2010. Bộ dữ liệu này được biên soạn bởi Dean De Cock nhằm phục vụ cho mục đích giáo dục trong khoa học dữ liệu. Nó được coi là một phiên bản hiện đại và mở rộng của bộ dữ liệu Boston Housing nổi tiếng.

Bộ dữ liệu Ames Housing thường được sử dụng như một lựa chọn thay thế cho bộ dữ liệu Boston Housing trong các bài toán dự đoán giá nhà. Dữ liệu cung cấp thông tin chi tiết hơn và phạm vi rộng hơn, giúp các nhà khoa học dữ liệu có nhiều cơ hội thực hành với các phương pháp phân tích và mô hình hóa tiên tiến.

# **2. Nguồn gốc và tài liệu tham khảo**

Bộ dữ liệu Ames Housing được giới thiệu trong bài báo khoa học "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project" của tác giả De Cock. Bài báo này cung cấp một cái nhìn tổng quan về dữ liệu, cấu trúc và các yếu tố liên quan.

* Link tham khảo bài báo: <https://github.com/wblakecannon/ames/blob/master/data/housing.csv>
* Bộ dữ liệu có thể được tải tại: [https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques](%20https:/www.kaggle.com/competitions/house-prices-advanced-regression-techniques)

# **3. Dữ liệu và từ điển dữ liệu**

Bộ dữ liệu chứa hơn 2.930 bản ghi với hơn 70 cột dữ liệu mô tả đặc điểm của từng căn nhà , như diện tích, số phòng, năm xây dựng, vị trí và nhiều yếu tố khác. Các cột quan trọng bao gồm:

* **SalePrice:** Giá bán của ngôi nhà (đơn vị USD), là biến mục tiêu trong các bài toán dự đoán giá nhà.
* **LotArea:** Diện tích đất (square feet), một yếu tố ảnh hưởng lớn đến giá trị bất động sản.
* **OverallQual:** Chất lượng tổng thể của ngôi nhà trên thang điểm từ 1 đến 10, với 10 là chất lượng cao nhất.
* **YearBuilt:** Năm xây dựng của ngôi nhà, ảnh hưởng đến tình trạng bảo trì và giá trị bất động sản.
* **GrLivArea:** Tổng diện tích sử dụng của ngôi nhà, bao gồm không gian sinh hoạt chính.
* **Neighborhood:** Khu vực mà ngôi nhà tọa lạc, ảnh hưởng đến giá trị và sự hấp dẫn của bất động sản.
* **GarageCars:** Số lượng chỗ đậu xe trong garage.
* **TotalBsmtSF:** Tổng diện tích tầng hầm.

Từ điển dữ liệu chi tiết có thể được tham khảo tại: <https://hackmd.io/@jirvingphd/ames-data-dictionary>

Với 80 biến giải thích, bao gồm:

* 23 biến định danh (nominal)
* 23 biến thứ bậc (ordinal)
* 14 biến rời rạc (discrete)
* 20 biến liên tục (continuous)

Các biến này mô tả gần như mọi khía cạnh có thể có của các ngôi nhà dân dụng

# **4. Các bài toán liên quan**

Bộ dữ liệu này thường được sử dụng trong các bài toán sau:

* **Dự đoán giá nhà (Regression)**: Xây dựng mô hình hồi quy tuyến tính và phi tuyến tính để dự đoán giá bán dựa trên các đặc điểm của ngôi nhà.

<https://nycdatascience.com/blog/python/eda-and-machine-learning-ames-housing-price-prediction-project/>

* **Phân tích dữ liệu thăm dò (EDA)**: Khám phá xu hướng dữ liệu, mối quan hệ giữa các đặc trưng và ảnh hưởng của từng yếu tố đến giá nhà.

<https://www.kaggle.com/code/abdelhai/eda-data-cleaning-ames-housing-dataset>

* **Tiền xử lý dữ liệu**: Xử lý dữ liệu bị thiếu, chuẩn hóa dữ liệu, mã hóa biến phân loại và xử lý các ngoại lệ.

<https://www.kaggle.com/code/bilalgondal786/tutorial-feature-engineering-housing-data>

**Machine Learning**: Áp dụng các thuật toán như Linear Regression, Random Forest, Gradient Boosting, XGBoost để dự đoán giá nhà và so sánh hiệu suất giữa các mô hình.

<https://github.com/sinhasagar507/Ames-house-price-prediction>

* **Feature Engineering**: Xây dựng các đặc trưng mới giúp tăng độ chính xác của mô hình dự đoán.

<https://en.wikipedia.org/wiki/Feature_engineering>

# **5. Kết quả đạt được và độ đo đánh giá**

Các mô hình dự đoán giá nhà dựa trên bộ dữ liệu Ames Housing được đánh giá bằng các độ đo phổ biến:

* **Root Mean Squared Error (RMSE)**: Đánh giá mức độ sai số trung bình theo đơn vị USD.
* **Mean Absolute Error (MAE)**: Sai số trung bình tuyệt đối, đo lường mức độ chênh lệch giữa giá trị dự đoán và thực tế.
* **R-squared (R²)**: Đánh giá mức độ giải thích của mô hình đối với dữ liệu thực tế.

Ví dụ, trong một dự án sử dụng bộ dữ liệu này, các mô hình hồi quy tuyến tính và hồi quy Ridge đã được triển khai để dự đoán giá nhà.

Một số kết quả tham khảo từ các nghiên cứu:

* Hồi quy tuyến tính cơ bản: RMSE ~30.000 USD
* Random Forest: RMSE ~22.000 USD
* XGBoost: RMSE ~20.000 USD
* Các mô hình phức tạp hơn có thể đạt RMSE dưới 20.000 USD với kỹ thuật tối ưu hóa phù hợp.

Link tài liệu tham khảo:

<https://rpubs.com/NgKhanh2203/1280516>

<https://github.com/Experimenters1/house-price-prediction>

<https://www.kaggle.com/code/davidanimaddo/ames-housing-prices-regression-xgboost>

# **6. Bài khảo sát về bộ dữ liệu**

Hiện tại không có bài khảo sát (survey) chính thức về bộ dữ liệu Ames Housing. Tuy nhiên, có nhiều bài viết và nghiên cứu sử dụng dữ liệu này để minh họa các phương pháp phân tích và mô hình hóa trong khoa học dữ liệu. Ví dụ, một dự án đã thực hiện phân tích dữ liệu thăm dò, kỹ thuật xây dựng đặc trưng và triển khai mô hình hồi quy tuyến tính để dự đoán giá trị bất động sản.

Một số tài liệu có thể tham khảo:

* Các bài phân tích EDA trên Kaggle.

VD: <https://www.kaggle.com/code/leeclemmer/exploratory-data-analysis-of-housing-in-ames-iowa>

* Các dự án ứng dụng trên GitHub.

VD: <https://github.com/jreynolds999/Ames-Housing-Data-Analysis>

* Các khóa học về học máy sử dụng bộ dữ liệu này làm ví dụ thực hành.

VD: <https://inria.github.io/scikit-learn-mooc/python_scripts/datasets_ames_housing.html>

# **7. Kết luận**

Bộ dữ liệu Ames Housing là một nguồn dữ liệu quan trọng và hữu ích cho các nghiên cứu về dự đoán giá bất động sản. Với số lượng đặc trưng phong phú và dữ liệu sạch, nó được xem là một bộ dữ liệu thay thế lý tưởng cho bộ dữ liệu Boston Housing nổi tiếng.

Việc khai thác Ames Housing giúp rèn luyện kỹ năng xử lý dữ liệu, xây dựng mô hình, và đánh giá hiệu suất mô hình trong thực tế. Hơn nữa, nó cung cấp một môi trường lý tưởng để thử nghiệm các kỹ thuật phân tích dữ liệu, từ đơn giản đến phức tạp, góp phần nâng cao năng lực nghiên cứu và ứng dụng thực tế trong lĩnh vực khoa học dữ liệu.