

UỶ BAN NHÂN DÂN

THÀNH PHỐ HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

****

**BÁO CÁO TIỂU LUẬN**

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

**Thông tin sinh viên:**

Nguyễn Hữu Lộc – 3122410213

Văn Tuấn Kiệt - 3122410202

Mai Phúc Lâm - 3122410207

Nguyễn Đức Duy Lâm - 3122410208

**Giảng viên:** Đỗ Như Tài

**Thành phố Hồ Chí Minh, 2/2025**

Mục lục

**[1. Paper của cơ sở dữ liệu ở đâu? 3](#_Toc2707)**

**[2. Dữ liệu ở đâu? 3](#_Toc29862)**

**[3. Từ điển dữ liệu? 4](#_Toc17623)**

BÀI 3

Link bài báo: <https://www.researchgate.net/publication/369437029_Price_Prediction_of_Ames_Housing_Through_Advanced_Regression_Techniques>

Trả lời câu hỏi dưới

* Paper của cơ sở dữ liệu ở đâu?
* Dữ liệu ở đâu?
* Từ điển dữ liệu?
* Các bài toán liên quan?
* Kết quả đạt được là gì? Độ đo?
* Có bài survey về nó không?
* ....
* Dữ liệu có được tiền xử lý trước không? Nếu có, tác giả đã thực hiện những bước nào?
* Kết quả được so sánh với các nghiên cứu trước như thế nào?
* Có thử nghiệm trên các tập dữ liệu khác không, hay chỉ trên Ames Housing?
* **Kết quả nghiên cứu có thể áp dụng vào thực tế không?**
* **Có đề xuất nào để cải thiện mô hình không?**
* Nghiên cứu này có thể mở rộng hoặc ứng dụng vào bộ dữ liệu khác không?

### 1. Paper của cơ sở dữ liệu ở đâu?

* Bài báo gốc về cơ sở dữ liệu Ames Housing được viết bởi **Dean De Cock** và có tiêu đề: **"Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project"**.
* Bài báo này được xuất bản trên tạp chí **Journal of Statistics Education**, số 19(3), năm 2011.
* Có thể tìm bài báo thông qua một số trang web như [researchgate](https://www.researchgate.net/publication/369437029_Price_Prediction_of_Ames_Housing_Through_Advanced_Regression_Techniques) hoặc trang [google schoolar](https://www.tandfonline.com/doi/abs/10.1080/10691898.2011.11889627)

### 2. Dữ liệu ở đâu?

* Cơ sở dữ liệu Ames Housing được cung cấp công khai và phổ biến nhất trên **Kaggle**. Bạn có thể truy cập tại: <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data>
* Tổng quan về dữ liệu: Cơ sở dữ liệu được thu thập bởi Dean De Cock từ các giao dịch bất động sản tại Ames, Iowa, Hoa Kỳ, trong khoảng thời gian từ năm 2006 đến 2010. Nó được thiết kế như một thay thế hiện đại cho tập dữ liệu Boston Housing nổi tiếng. Dùng để dự đoán giá bán nhà (SalePrice) dựa trên các đặc trưng liên quan đến bất động sản, đồng thời là một tài nguyên học máy phổ biến cho các bài toán hồi quy.

### 3. Từ điển dữ liệu?

* Từ điển dữ liệu (data dictionary) đi kèm với tập dữ liệu trên Kaggle, nằm trong file **data\_description.txt**.
* Nó mô tả chi tiết 79 biến (features) và 1 biến mục tiêu (target) là **SalePrice**. Các biến bao gồm:
  + **Nominal**: 23 biến (ví dụ: Neighborhood - khu vực lân cận, RoofStyle - kiểu mái nhà, Exterior1st - vật liệu phủ ngoài).
  + **Ordinal**: 23 biến (ví dụ: OverallQual - chất lượng tổng thể, KitchenQual - chất lượng bếp, BsmtCond - tình trạng tần hầm).
  + **Discrete**: 14 biến (ví dụ: YearBuilt - năm xây dựng, BedroomAbvGr - số phòng ngủ).
  + **Continuous**: 20 biến (ví dụ: LotArea - diện tích lô đất, TotalBsmtSF - tổng diện tích tầng hầm).
* Từ điển giải thích ý nghĩa của từng biến, giá trị có thể có và cách chúng được mã hóa. Ví dụ:

+ SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.

+ MSSubClass: The building class

+ MSZoning: The general zoning classification