**Dự đoán giá nhà bằng cách sử dụng học máy**

# **Mục tiêu và Giới thiệu [1]**

Nghiên cứu này trình bày một thuật toán học máy nhằm mục đích **dự đoán giá trị nhà ở** trên thị trường bất động sản.

Việc dự đoán giá nhà trong thị trường bất động sản được xem là **thách thức** do tính phức tạp và biến động của nó.

Bài báo đề xuất một mô hình dự đoán giá nhà sử dụng các phương pháp học máy để đưa ra các ước tính giá chính xác và sâu sắc

Mục tiêu chính là xây dựng một kỹ thuật dự đoán đáng tin cậy và chính xác để ước tính giá trị bất động sản dân cư bằng cách sử dụng nhiều thuộc tính của tài sản.

Các kỹ thuật học máy có thể phân tích các tập dữ liệu lớn và xác định các mô hình tinh tế và mối tương quan phi tuyến tính ảnh hưởng đến giá cả, điều mà các kỹ thuật đánh giá truyền thống thường bỏ qua

Mô hình này nhằm mục đích hỗ trợ các chuyên gia bất động sản, chủ nhà và nhà đầu tư đưa ra quyết định sáng suốt trong thị trường năng động....

# **Dữ liệu và Phương pháp luận**

Nghiên cứu sử dụng một **tập dữ liệu toàn diện** bao gồm các chi tiết như vị trí, kích thước, số lượng phòng ngủ và phòng tắm, cùng với các chi tiết liên quan khác cho mỗi căn nhà.... Tập dữ liệu được tải về từ Kaggle và gồm hai tập con là 'train.csv' và 'test.csv'.

Quá trình phát triển mô hình bao gồm nhiều giai đoạn:

**Chuẩn bị dữ liệu ban đầu:** Sửa các lỗi như giá trị bị thiếu (missing values) và ngoại lai (outliers), đồng thời chuyển đổi các biến phân loại sang định dạng phù hợp để phân tích.... Các ngoại lai cho tập dữ liệu huấn luyện được trực quan hóa bằng biểu đồ hộp và râu. Sau khi loại bỏ các ngoại lai, kích thước của tập huấn luyện thay đổi từ (1460, 81) thành (1101, 81) [2]

**Tiền xử lý dữ liệu:** Tập dữ liệu huấn luyện và kiểm tra được nối lại, sau đó kiểm tra và xử lý các giá trị bị thiếu. Tổng cộng có 9 cột có giá trị bị thiếu được điền bằng 2 và 0. Các tập dữ liệu mới được tạo ra bằng cách sử dụng các kỹ thuật như mã hóa nhãn (label encoding), biến giả (dummy variables), phân tích thành phần chính (PCA), chọn các cột có tương quan lớn hơn 0.05 với cột SalePrice, và chọn lọc đặc trưng (feature selection).

**Chọn lọc đặc trưng (Feature Selection):** Các thuộc tính quan trọng nhất được xác định bằng kỹ thuật chọn lọc đặc trưng và được đưa vào mô hình dự đoán....

Tập dữ liệu được chia thành các tập con huấn luyện và kiểm tra để đảm bảo đánh giá mô hình mạnh mẽ

# **Triển khai Mô hình**

Tổng cộng **14 mô hình học máy** được triển khai để dự đoán giá nhà.... Kiến trúc cơ sở cho việc triển khai mô hình tuân theo khái niệm Học Chuyển Giao (Transfer Learning).

*Các mô hình được triển khai bao gồm...:*

1. Hồi quy Tuyến tính (Linear Regression)
2. Hồi quy Ridge (Ridge Regression)
3. Hồi quy Lasso (Lasso Regression)
4. Hồi quy Đa thức (Polynomial Regression - Bậc 2)
5. ElasticNet
6. Bộ hồi quy Vector Hỗ trợ (Support Vector Regressor - SVR)
7. Cây Quyết định (Decision Tree)
8. Rừng Ngẫu nhiên (Random Forest Regressor)
9. Bộ hồi quy Tăng cường Gradient (Gradient Boosting Regressor)
10. Bộ hồi quy Cây Extra (Extra Trees Regressor)
11. XGBoost (Extreme Gradient Boosting Regressor)
12. Bộ hồi quy K-Neighbors (KNeighbors Regressor)
13. Regularization + Deep Learning
14. Just Deep Sense Learning (hoặc Dense Deep Learning2)

# **Đánh giá Mô hình**

Các số liệu đánh giá như **sai số trung bình bình phương (mean squared error - MSE), sai số tuyệt đối trung bình (mean absolute error - MAE) và R-squared** được sử dụng để đánh giá hiệu suất mô hình.... Hàm evaluate() được định nghĩa để tính toán các số liệu này cho từng mô hình và từng tập dữ liệu đã xử lý khác nhau (df le, df dum, df pca dum, df cor dum, df fs dum).

Sơ bộ nghiên cứu cho thấy thuật toán **Gradient Boosting có khả năng dự đoán mạnh hơn**, tạo ra sai số dự đoán thấp nhất và điểm R-squared cao nhất....

Các kỹ thuật Ensemble, đặc biệt là **Random Forests và Gradient Boosting, liên tục thể hiện khả năng dự đoán mạnh mẽ**, vượt trội so với các mô hình đơn giản hơn như Hồi quy Tuyến tính với sai số dự đoán thấp hơn và giá trị R-squared cao hơn.

**Mạng nơ-ron (Neural networks)** cũng cho thấy sự thành công, đặc biệt trong việc xử lý các tập dữ liệu phức tạp.

Mô hình gradient boosting được tối ưu hóa thêm thông qua điều chỉnh siêu tham số (hyperparameter tuning), làm nổi bật khả năng nắm bắt các mối tương quan phức tạp trong dữ liệu và tạo ra dự đoán giá nhà chính xác....

# **Kết luận**

Bài toán tập trung vào việc **dự đoán giá nhà ở và kiểm tra hiệu quả của 14 mô hình máy học khác nhau**. Những mô hình này bao gồm một loạt các phương pháp, từ các phương pháp truyền thống như **hồi quy tuyến tính và cây quyết định đến các kỹ thuật tinh vi như hồi quy vector hỗ trợ, boosting gradient và học sâu.**

Nghiên cứu kết luận rằng các mô hình hồi quy như Linear, Ridge, Lasso, Polynomial và ElasticNet cùng với **Gradient Boosting Regressor, XGBoost và SVR** hoạt động tốt hơn các mô hình khác, làm nổi bật hiệu quả của chúng trong việc nắm bắt các mô hình phức tạp trên thị trường nhà ở....

Nghiên cứu nhấn mạnh **vai trò quan trọng của kỹ thuật đặc trưng (feature engineering) và tiền xử lý dữ liệu** trong việc cải thiện độ chính xác dự đoán của tất cả các mô hình....

Kết quả nghiên cứu cung cấp cái nhìn sâu sắc có giá trị cho ngành bất động sản và chứng minh tiềm năng của việc dự báo giá nhà chính xác và đáng tin cậy hơn.... Việc sử dụng các công nghệ tiên tiến này có thể cung cấp cho các bên liên quan các công cụ ra quyết định tốt hơn trong thị trường năng động và thay đổi này.

* Nghiên cứu này đóng góp vào lĩnh vực dự đoán giá nhà bằng cách kiểm tra và so sánh một loạt các mô hình học máy, đồng thời nhấn mạnh tầm quan trọng của việc chuẩn bị dữ liệu và lựa chọn mô hình phù hợp....

**MAE** cho bạn biết mức sai số “thật” trung bình (dễ diễn giải nhất).

**MSE** phạt nặng sai số lớn, thường dùng khi bạn muốn ưu tiên giảm những lỗi lớn hơn.

| **Tên DataFrame** | **Tiền xử lý áp dụng** |
| --- | --- |
| **df\_dum** | – Dữ liệu gốc (all features) sau khi chuyển toàn bộ các biến phân loại (categorical) thành dummy/one-hot encoding. |
| **df\_cor\_dum** | – Bắt đầu từ **df\_dum**, lọc bỏ các cặp biến có hệ số tương quan cao (loại bỏ multicollinearity) trước khi huấn luyện. |
| **df\_fs\_dum** | – Bắt đầu từ **df\_dum**, áp dụng bước **Feature Selection** (ví dụ: loại bỏ biến ít quan trọng theo thống kê hoặc model-based), rồi giữ lại và one-hot encoding những biến được chọn. |
| **df\_le** | – Dữ liệu gốc nhưng thay vì one-hot, các biến phân loại được **Label Encoding** (gán mỗi giá trị một số nguyên). |
| **df\_pca\_dum** | – Bắt đầu từ **df\_dum**, sau khi one-hot hóa, áp dụng **PCA** lên không gian đặc trưng (thường chỉ numeric hoặc cả numeric + dummy) để giảm chiều, rồi dùng các thành phần chính này làm input. |