TRƯỜNG ĐẠI HỌC FPT

Môn học: ADY201m

BÁO CÁO ĐỀ TÀI

PHÂN TÍCH VÀ DỰ ĐOÁN GIÁ BẤT ĐỘNG SẢN

Thành viên nhóm:

SE201725	Nguyễn Gia Huy
SE203377	Trần Cảnh Nhật Minh
SE204049	Nguyễn Xuân Bình
SE201652	Đặng Thái Nguyên

Tháng 2 năm 2026

Mục lục

1. Giới thiệu

Thị trường bất động sản giữ vai trò trụ cột trong nền kinh tế nhưng thường chịu tác động phức tạp từ đa chiều yếu tố như vị trí, pháp lý, tiện ích..., khiến việc xác định giá trị thực trở thành thách thức lớn. Dự án này tiếp cận vấn đề bằng cách ứng dụng khoa học dữ liệu để khai thác nguồn thông tin phong phú từ các tin đăng bất động sản. Mục tiêu cốt lõi là định lượng mức độ ảnh hưởng của các thuộc tính then chốt và xây dựng mô hình dự đoán giá tin cậy. Kết quả nghiên cứu được kỳ vọng sẽ cung cấp cơ sở tham chiếu khách quan, hỗ trợ ra quyết định hiệu quả cho các bên tham gia thị trường và góp phần nâng cao tính minh bạch chung.

1.1. Mục tiêu Phân tích và Dự đoán

Phân tích: Tập trung khám phá mối quan hệ giữa giá và các thuộc tính (diện tích, vị trí, tiện ích, pháp lý). Nhóm đánh giá xu hướng phân bố giá theo khu vực và loại hình bất động sản, đồng thời nhận diện các rủi ro pháp lý tiềm ẩn từ dữ liệu tin đăng.

Dự đoán: Xây dựng mô hình dự báo giá dựa trên thông tin đầu vào của người dùng. Kết quả cung cấp mức giá tham khảo để so sánh với giá chào bán, hỗ trợ người mua thương lượng hiệu quả hơn.

2. Thu thập dữ liệu

Để đảm bảo cơ sở dữ liệu phong phú và đa chiều cho mô hình phân tích, nhóm nghiên cứu đã tiến hành tổng hợp dữ liệu từ hai nguồn chính: thông tin giao dịch thị trường và dữ liệu không gian về hạ tầng đô thị.

2.1. Dữ liệu bất động sản

Phạm vi thu thập dữ liệu tập trung vào 4 thị trường sôi động nhất cả nước bao gồm: **Hà Nội, Đà Nẵng, TP. Hồ Chí Minh và Bình Dương**. Việc lựa chọn các khu vực này giúp đảm bảo cỡ mẫu đủ lớn và phản ánh đúng xu hướng chung của thị trường. Nhóm sử dụng các kỹ thuật thu thập dữ liệu tự động (Web Crawling) trên nền tảng *batdongsan.com.vn* để trích xuất các thuộc tính quan trọng như giá bán, diện tích, kết cấu nhà và tình trạng pháp lý.

2.2. Dữ liệu địa lý và tiện ích

Bên cạnh thông tin tin đăng, yếu tố vị trí được định lượng thông qua dữ liệu bản đồ mở **OpenStreetMap (OSM)**. Nhóm ứng dụng các thuật toán tính toán không gian để thống kê mật độ tiện ích thiết yếu (trường học, y tế, thương mại) trong phạm vi bán kính xác định quanh từng bất động sản, từ đó làm giàu thêm thông tin đầu vào cho mô hình dự báo.

3. Tổng quan về dữ liệu thô

Sau quá trình thu thập, bộ dữ liệu thô bao gồm tổng cộng **39 trường dữ liệu**. Do số lượng biến lớn, nhóm phân loại chúng thành 4 nhóm chính để thuận tiện cho việc quản lý trước khi xử lý:

Nhóm dữ liệu	SL Biến	Mô tả các biến đại diện
1. Metadata tin đăng	10 biến	Các thông tin quản lý tin: *post_id, url, published_date, seller_name...* giúp định danh và lọc tin rác.
2. Vị trí địa lý	9 biến	Các cấp hành chính (*province, district, ward...*) và tọa độ (*latitude, longitude*).
3. Đặc điểm Bất động sản	14 biến	Thông tin tài sản: *price_total, area, num_floors, num_bedrooms, direction, furniture, legal_status...*
4. Tiện ích ngoại khu	6 biến	Số lượng tiện ích từ OSM: *num_schools_1km, num_hospitals_2km, num_markets_1km...*

Bảng 1: Tổng quan cấu trúc dữ liệu thô (39 cột)

Việc thu thập số lượng lớn các trường dữ liệu ban đầu giúp nhóm có cái nhìn toàn diện, tuy nhiên không phải tất cả đều đóng góp vào bài toán dự đoán giá. Quá trình chọn lọc sẽ được trình bày ở phần tiếp theo.

4. Phương pháp xử lý dữ liệu

Dữ liệu thô thu thập từ môi trường trực tuyến thường chứa nhiều nhiễu, định dạng không đồng nhất và các giá trị khuyết thiếu. Để xây dựng cơ sở dữ liệu tin cậy cho mô hình dự báo, nhóm nghiên cứu đã thực hiện quy trình tiền xử lý nghiêm ngặt nhằm chuẩn hóa và làm sạch thông tin.

4.1. Làm sạch và Chuẩn hóa dữ liệu

Quy trình xử lý tập trung giải quyết các vấn đề về định dạng và tính nhất quán của dữ liệu đầu vào:

- **Chuẩn hóa định dạng:** Các thuộc tính định lượng (giá, diện tích, kích thước) được trích xuất và chuyển đổi từ dạng văn bản sang định dạng số học để phục vụ tính toán.
- **Đồng bộ hóa dữ liệu:** Các thông tin về địa danh hành chính (Quận/Huyện) được quy hoạch về một danh mục thống nhất, loại bỏ các biến thể tên gọi khác nhau do quá trình nhập liệu thủ công.
- **Khử trùng lặp:** Loại bỏ các bản ghi tin đăng trùng lặp để đảm bảo tính duy nhất và độ chính xác của các quan sát trong tập dữ liệu.

4.2. Xử lý dữ liệu thiếu và Chọn lọc đặc trưng

Từ tập dữ liệu ban đầu, nhóm tiến hành sàng lọc để tối ưu hóa đầu vào cho mô hình:

- **Xử lý giá trị khuyết thiếu:** Chiến lược điền khuyết được áp dụng linh hoạt dựa trên đặc thù của từng loại hình bất động sản.
- **Trích chọn đặc trưng:** Loại bỏ các biến không mang giá trị dự báo hoặc gây nhiễu. Kết quả cuối cùng giữ lại **17 thuộc tính đặc trưng** quan trọng nhất được trình bày trong bảng dưới đây.

Nhóm biến	STT	Tên biến	Kiểu	Ý nghĩa sử dụng
1. Mục tiêu & Tham chiếu	1	`price_total`	Int	Biến mục tiêu: Tổng số tiền người bán mong muốn (VND)
	2	`price_per_m2`	Float	Đơn giá tham chiếu (VND/m^2)
2. Vị trí địa lý	3	`province`	Category	Tỉnh/Thành phố.
	4	`district`	Category	Quận/Huyện.
	5	`latitude`	Float	Vĩ độ.
	6	`longitude`	Float	Kinh độ.
3. Đặc điểm Bất động sản	7	`area`	Float	Diện tích sử dụng (m^2).
	8	`category`	Category	Loại hình (Chung cư...).
	9	`legal_status`	Category	Pháp lý (Sổ đỏ...).
	10	`frontage`	Float	Độ rộng mặt tiền (m).
	11	`road_width`	Float	Độ rộng đường vào (m).
	12	`num_bedrooms`	Int	Số phòng ngủ.
	13	`num_toilets`	Int	Số phòng vệ sinh.
	14	`num_floors`	Int	Số tầng.
4. Tiện ích ngoại khu	15	`num_schools_1km`	Int	Mật độ trường (bán kính 1km).
	16	`num_hospitals_2km`	Int	Mật độ y tế (bán kính 2km).
	17	`num_markets_1km`	Int	Mật độ chợ (bán kính 1km).

Bảng 2: Danh sách 17 biến đặc trưng phân theo nhóm logic

5. Kế hoạch thực hiện

5.1. Phương pháp làm việc (Mô hình OSEMN)

Dự án tuân thủ quy trình OSEMN: **O**btain (Thu thập) → **S**crub (Làm sạch) → **E**xplore (Khám phá) → **M**odel (Mô hình hóa) → i**N**terpret (Diễn giải).

Hình 1: Quy trình phân tích dữ liệu OSEMN

5.2. Phân công nhiệm vụ chi tiết

Dựa trên quy trình OSEMN, nhóm chia dự án thành 5 giai đoạn chính. Mỗi giai đoạn bao gồm các đầu việc cụ thể được phân công cho các thành viên như sau:

Giai đoạn	STT	Nội dung công việc	Thành viên
1. Obtain	1	Xác định mục tiêu và tiêu chí dữ liệu.	Cả nhóm
	2	Thu thập dữ liệu (Web Crawling, API).	Cả nhóm
2. Scrub	3	Kiểm tra, xử lý dữ liệu nhiễu và khuyết thiếu.	Huy, Minh
	4	Chuẩn hóa và gộp dataset thống nhất.	Huy, Minh
3. Explore	5	Trực quan hóa (Biểu đồ, Heatmap).	Huy, Nguyên
	6	Phân tích xu hướng và chọn đặc trưng.	Huy, Bình, Nguyên
4. Model	7	Kỹ thuật đặc trưng: Mã hóa, Chia tập.	Bình
	8	Huấn luyện mô hình Machine Learning.	Huy, Bình
	9	Tinh chỉnh tham số (Model Tuning).	Bình
5. iNterpret	10	Đánh giá kết quả mô hình.	Nguyên
	11	Tổng kết, kết luận và viết báo cáo.	Cả nhóm

Bảng 3: Bảng phân công nhiệm vụ theo 5 giai đoạn OSEMN