**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

-------------------

A close up of a sign

Description automatically generated

**BÁO CÁO BÀI TẬP LỚN**

**MÔN HỌC: NHẬP MÔN KHOA HỌC DỮ LIỆU**

ĐỀ TÀI

**Dự đoán giá nhà đất**

**GVHD:**  PGS. TS. Thân Quang Khoát

**Nhóm: 5**

|  |  |
| --- | --- |
| Thành viên nhóm | MSSV |
| Dương Tiến Đạt | 20204946 |
| Lương Phúc Quang | 20215125 |
| Hoàng Văn Duy | 20204959 |
| Nguỵ Khắc Phi Long | 20210966 |

***Hà Nội, ngày 16 tháng 5 năm 2025***

Nội dung

[LỜI MỞ ĐẦU 4](#_Toc198664011)

[I. Giới thiệu vấn đề 5](#_Toc198664012)

[1) Bài toán đặt ra 5](#_Toc198664013)

[2) Phương hướng xử lý 5](#_Toc198664014)

[II. THU THẬP DỮ LIỆU 6](#_Toc198664015)

[1) Nguồn dữ liệu 6](#_Toc198664016)

[1.1 Nguồn thu thập 6](#_Toc198664017)

[1.2 Lý do lựa chọn 6](#_Toc198664018)

[1.3 Thách thức ban đầu 7](#_Toc198664019)

[2) Mục tiêu thu thập dữ liệu 7](#_Toc198664020)

[2.1 Mục tiêu chính 7](#_Toc198664021)

[2.2 Thông tin thu thập 7](#_Toc198664022)

[2.3 Yêu cầu dữ liệu 8](#_Toc198664023)

[3) Quy trình thu thập 8](#_Toc198664024)

[3.1 Tổng quan 8](#_Toc198664025)

[3.2 Chi tiết quy trình 8](#_Toc198664026)

[4) Kết quả đạt được 9](#_Toc198664027)

[ Thu thập dữ liệu đơn vị hành chính Việt Nam 10](#_Toc198664028)

[III. Tiền xử lý dữ liệu 11](#_Toc198664029)

[1) Phân tích sơ bộ tổng quan 11](#_Toc198664030)

[1.1 Thống kê số lượng thiếu của các đặc trưng 11](#_Toc198664031)

[1.2 Thống kê số lượng giá trị thiếu trên mỗi bản ghi 11](#_Toc198664032)

[1.3 Kiểm tra kiểu dữ liệu 12](#_Toc198664033)

[1.4 Phân tích đặc trưng số (Numerical) 13](#_Toc198664034)

[1.5 Phân tích đặc trưng phân loại (Categorical) 14](#_Toc198664035)

[2) Phân tích khám phá chuyên sâu 15](#_Toc198664036)

[2.1 Tổng quan dữ liệu sau xử lý 15](#_Toc198664037)

[2.2 Phân tích biến mục tiêu (Price) 16](#_Toc198664038)

[2.3 Phân bố các biến liên tục 17](#_Toc198664039)

[2.4 Phân bố nhãn biến phân loại 18](#_Toc198664040)

[3) Các bước xử lý chi tiết 20](#_Toc198664041)

[3.1 Tiền xử lý cơ bản 20](#_Toc198664042)

[3.2 Kỹ thuật trích xuất đặc trưng 20](#_Toc198664043)

[3.3 Xử lý phân phối lệch 21](#_Toc198664044)

[3.4 Xử lý giá trị thiếu 22](#_Toc198664045)

[3.5 Xử lý ngoại lệ 22](#_Toc198664046)

[4) Kết quả 22](#_Toc198664047)

[IV. Mô hình học máy 24](#_Toc198664048)

[1) Giải thuật hồi quy tuyến tính 24](#_Toc198664049)

[2) Giải thuật k láng giềng gần nhất 24](#_Toc198664050)

[3) XGBoost 27](#_Toc198664051)

[**A.** **Tổng quan về XGBoost** 27](#_Toc198664052)

[**B.** **Nguyên lý hoạt động của XGBoost** 27](#_Toc198664053)

[**Cơ sở lý thuyết** 27](#_Toc198664054)

[**Hàm mất mát và tối ưu hóa** 28](#_Toc198664055)

[**Quá trình xây dựng cây quyết định** 28](#_Toc198664056)

[**C.** **Ứng dụng XGBoost trong Project Dự đoán Giá Bất động sản** 28](#_Toc198664057)

[**Cấu hình mô hình** 28](#_Toc198664058)

[**Quy trình xử lý dữ liệu:** 29](#_Toc198664059)

[V. Đánh giá mô hình 29](#_Toc198664060)

[VI. Kết luận và khả năng mở rộng 32](#_Toc198664061)

LỜI MỞ ĐẦU

Trong bối cảnh thị trường bất động sản ngày càng phát triển mạnh mẽ, việc định giá bất động sản một cách chính xác và nhanh chóng đóng vai trò quan trọng đối với cả người mua, người bán và các nhà đầu tư. Tuy nhiên, giá trị của một bất động sản phụ thuộc vào nhiều yếu tố như vị trí địa lý, diện tích, tiện ích xung quanh, số phòng, thời điểm đăng bán, v.v., khiến quá trình định giá thủ công trở nên tốn thời gian và dễ sai lệch.

Với sự phát triển của trí tuệ nhân tạo và học máy, việc áp dụng các mô hình dự đoán vào bài toán định giá nhà đất đang dần trở thành xu hướng phổ biến. Dự án này nhằm xây dựng một hệ thống dự đoán giá nhà đất tại Việt Nam dựa trên dữ liệu được thu thập từ các nền tảng bất động sản trực tuyến. Thông qua việc xử lý dữ liệu, phân tích đặc trưng và huấn luyện mô hình, hệ thống có thể hỗ trợ người dùng ước lượng giá trị của một bất động sản một cách khách quan và hiệu quả hơn.

# Giới thiệu vấn đề

## Bài toán đặt ra

Giá bất động sản là một yếu tố then chốt trong các quyết định mua bán, đầu tư, và định giá tài sản. Tuy nhiên, giá cả thường biến động mạnh và phụ thuộc vào nhiều yếu tố như vị trí, diện tích, số phòng, tình trạng pháp lý, hạ tầng xung quanh, v.v.

Trong bài toán này, nhóm đặt mục tiêu xây dựng một mô hình **dự đoán giá bất động sản** dựa trên các đặc trưng đầu vào được lấy từ các trang web thương mại điện tử bất động sản. Đây là một bài toán **hồi quy (regression)** với biến mục tiêu là **giá bán bất động sản (Price)**.

## Phương hướng xử lý

Dữ liệu được **thu thập** từ chuyên mục **“Bán nhà riêng”** trên trang Batdongsan.com.vn bằng công cụ crawl dữ liệu tự động. Mỗi bản ghi bao gồm các thông tin như tiêu đề tin đăng, địa chỉ, diện tích, số phòng ngủ, số toilet, ngày đăng và giá bán.

Sau khi thu thập dữ liệu thô, nhóm tiến hành các bước xử lý và xây dựng mô hình như sau:

* **Tiền xử lý dữ liệu**: xử lý giá trị thiếu, chuẩn hóa các đặc trưng số, mã hóa các đặc trưng phân loại.
* **Phân tích dữ liệu**: kiểm tra phân phối của các biến, phát hiện giá trị ngoại lệ, đánh giá mối tương quan giữa các đặc trưng và biến mục tiêu.
* **Xử lý dữ liệu nâng cao**: biến đổi logarit cho biến Price nhằm giảm độ lệch phân phối, cắt ngưỡng hoặc loại bỏ các ngoại lệ bất thường.
* **Tạo và chọn đặc trưng**: sinh thêm các đặc trưng mới từ dữ liệu gốc và chọn ra các đặc trưng quan trọng nhất cho mô hình.
* **Huấn luyện mô hình**: thử nghiệm nhiều thuật toán học máy như Linear Regression, K-Nearest Neighbors (KNN) và XGBoost.
* **Đánh giá mô hình**: sử dụng các chỉ số như MAE, RMSE và R² để đánh giá độ chính xác và khả năng tổng quát hóa của mô hình.

Các bước trên được triển khai theo quy trình tuần tự, với mục tiêu cuối cùng là xây dựng một mô hình dự đoán giá bất động sản chính xác, có thể hỗ trợ người dùng ước lượng giá trị tài sản một cách nhanh chóng và có cơ sở.

# THU THẬP DỮ LIỆU

## Nguồn dữ liệu

### Nguồn thu thập

* **Tên website:** batdongsan.com.vn
* **Loại dữ liệu:** Bất động sản nhà riêng
* **Phạm vi:** Toàn quốc
* **Địa chỉ URL:** <https://batdongsan.com.vn/ban-nha-rieng>

Ảnh có chứa văn bản, ảnh chụp màn hình, Website, Trang web

Nội dung do AI tạo ra có thể không chính xác.

### Lý do lựa chọn

Batdongsan.com.vn là một trong những nền tảng bất động sản lớn và uy tín nhất tại Việt Nam, được đánh giá bởi các tổ chức như **Comscore** và **SimilarWeb** là kênh thông tin bất động sản hàng đầu tại Đông Nam Á. Website có nhiều đặc điểm nổi bật khiến nó trở thành nguồn dữ liệu phù hợp cho bài toán dự đoán giá nhà:

* **Lượng truy cập lớn và ổn định**: Gần 10 triệu lượt truy cập/tháng, hơn 1 triệu tin rao được cập nhật thường xuyên.
* **Dữ liệu phong phú, đa dạng loại hình bất động sản**: Bao gồm nhà riêng, nhà phố, đất nền, căn hộ, biệt thự,...
* **Cập nhật liên tục**: Các tin rao được làm mới và kiểm duyệt theo chu kỳ, đảm bảo tính thời sự.
* **Cấu trúc dữ liệu rõ ràng**: HTML và định dạng thông tin ổn định, thuận lợi cho việc tự động trích xuất bằng các công cụ như Selenium.
* **Uy tín và độ tin cậy cao**: Là đối tác của nhiều chủ đầu tư và môi giới trong nước, thông tin trên website có độ xác thực tương đối cao so với nhiều nguồn không chính thống.

### Thách thức ban đầu

Trong quá trình thu thập dữ liệu từ Batdongsan.com.vn, nhóm gặp một số khó khăn kỹ thuật đáng kể:

* **Không có API công khai**, buộc phải thu thập dữ liệu trực tiếp từ giao diện người dùng (UI).
* **Cơ chế chống bot mạnh**: Website sử dụng CAPTCHA và phát hiện hành vi tự động, dễ chặn các công cụ crawl thông thường như Selenium.
* **Cấu trúc website phức tạp**: Thông tin chi tiết nằm sâu trong các trang con, phân tán và được tải động (dynamic rendering).
* **Yêu cầu xử lý đồng thời**: Để tối ưu thời gian, hệ thống cần thực hiện thu thập đa luồng (multi-threading) và đảm bảo tính ổn định, đồng bộ trong quá trình crawl.

## Mục tiêu thu thập dữ liệu

### Mục tiêu chính

* Thu thập dữ liệu chuyên mục *Bán nhà riêng* trên Batdongsan.com.vn
* Lưu trữ dữ liệu ở định dạng CSV, với cấu trúc rõ ràng và đầy đủ thông tin
* Phục vụ phân tích đặc điểm và xu hướng giá nhà riêng tại Việt Nam
* Làm tiền đề cho các nghiên cứu sâu hơn về các yếu tố ảnh hưởng đến giá bất động sản

### Thông tin thu thập

Dữ liệu được thu thập từ từng bài đăng trong chuyên mục **Bán nhà riêng**, bao gồm:

* **Thông tin tổng quan** (tiêu đề, địa điểm, diện tích, mức giá, thời gian đăng) lấy từ trang phân trang danh sách.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

* **Thông tin chi tiết** (số phòng ngủ, số phòng tắm/vệ sinh, số tầng, hướng nhà, hướng ban công, đường vào, mặt tiền, pháp lý, nội thất) thu thập khi truy cập vào trang xem chi tiết của từng bất động sản.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, hàng

Nội dung do AI tạo ra có thể không chính xác.

### Yêu cầu dữ liệu

* Đầy đủ, chính xác, cập nhật thường xuyên
* Định dạng thống nhất, dễ xử lý (CSV)
* Thu thập từ nguồn uy tín, đảm bảo độ tin cậy cao (batdongsan.com.vn)

## Quy trình thu thập

### Tổng quan

Ban đầu, nhóm gặp phải hạn chế do hệ thống CAPTCHA chặn khi cố gắng thu thập dữ liệu trực tiếp trên từng trang chi tiết. Để khắc phục, nhóm đã xây dựng quy trình thu thập dữ liệu thành hai bước riêng biệt:

* **Bước 1:** Thu thập danh sách URL bài đăng từ các trang phân trang.
* **Bước 2:** Truy cập từng URL chi tiết để lấy thông tin đầy đủ.

Phương pháp này vừa giúp giảm nguy cơ bị chặn, vừa tăng hiệu quả thu thập dữ liệu.

### Chi tiết quy trình

**Bước 1: Crawl danh sách URL**

* **Công cụ sử dụng:**
* **Selenium** và **selenium-stealth** để điều khiển trình duyệt và ngụy trang nhằm tránh bị phát hiện là bot.
* **ThreadPoolExecutor** để thực hiện đa luồng, tăng tốc độ thu thập.
* **WebDriverWait** để chờ các phần tử trên trang tải hoàn chỉnh.
* **Lock** để đồng bộ việc ghi dữ liệu vào file, tránh lỗi khi ghi đồng thời.
* **Quy trình thực hiện:**
* Khởi tạo trình duyệt Chrome ở chế độ headless (không hiển thị giao diện) để tối ưu tài nguyên.
* Cấu hình trình duyệt với các biện pháp ngụy trang như thay đổi user-agent, xóa cookies,...
* Truy cập tuần tự các trang phân trang trong chuyên mục “Bán nhà riêng”.
* Chờ trang tải xong danh sách bài đăng bằng WebDriverWait.
* Trích xuất các thông tin cơ bản như: tiêu đề, địa điểm, mức giá, ngày đăng, URL chi tiết.
* Áp dụng cơ chế retry (thử lại) khi gặp lỗi kết nối hoặc tải trang, kèm theo ghi log chi tiết.
* Ghi danh sách URL bài đăng thu thập được vào file urls.csv để sử dụng ở bước tiếp theo.

**Bước 2: Crawl chi tiết bài đăng**

* **Công cụ:**
* Sử dụng lại các công cụ ở bước 1: Selenium, selenium-stealth, ThreadPoolExecutor, WebDriverWait, Lock.
* Thêm phần xử lý mapping dữ liệu để chuyển đổi các biểu tượng, icon đặc thù trên trang thành dạng chữ dễ đọc, dễ phân tích.
* **Quy trình:**
* Đọc danh sách URL từ file **urls.csv**, chia nhỏ các URL để xử lý theo từng luồng nhằm tối ưu hiệu suất và giảm tải bộ nhớ.
* Truy cập từng URL chi tiết một cách tuần tự hoặc đa luồng, sử dụng Selenium để điều khiển trình duyệt.
* Chờ trang chi tiết tải đầy đủ, sau đó trích xuất các thông tin chi tiết bao gồm số phòng ngủ, số phòng tắm, số tầng, hướng nhà, hướng ban công, đường vào, mặt tiền, pháp lý, nội thất, mô tả chi tiết, v.v.
* Thực hiện mapping biểu tượng và icon (ví dụ icon pháp lý, nội thất) thành dữ liệu dạng chữ theo quy ước để dễ xử lý về sau.
* Kiểm tra tính hợp lệ và đầy đủ của dữ liệu thu thập được.
* Ghi dữ liệu chi tiết vào file **details.csv**, đảm bảo định dạng chuẩn CSV, đồng bộ khi ghi để tránh lỗi.

## Kết quả đạt được

Dữ liệu thu thập được tổng cộng có 49.304 bài đăng trong vòng, quá trình thu thập mất 15 giờ với tỷ lệ thành công 100%, dữ liệu bao gồm các thông tin sau:

* **Tiêu đề:** Tóm tắt nội dung chính của bài đăng, thể hiện loại bất động sản và điểm nổi bật.
* **Địa điểm:** Vị trí cụ thể của bất động sản (tỉnh/thành, khu vực), quan trọng để phân tích theo vùng.
* **Mức giá:** Giá bán, yếu tố then chốt cho dự đoán và phân tích thị trường.
* **Diện tích:** Diện tích sử dụng hoặc đất, ảnh hưởng trực tiếp đến giá và nhu cầu.
* **Số phòng ngủ:** Số lượng phòng ngủ, phản ánh sự phù hợp với nhu cầu sinh hoạt.
* **Số phòng tắm, vệ sinh:** Cho biết mức độ tiện nghi và hiện đại của căn nhà.
* **Số tầng:** Thể hiện cấu trúc, quy mô sử dụng hoặc khai thác của nhà.
* **Hướng nhà:** Ảnh hưởng đến phong thủy, ánh sáng và thông gió.
* **Hướng ban công:** Tác động đến môi trường sống và phong thủy.
* **Đường vào:** Đặc điểm lối vào (rộng, hẹp…), ảnh hưởng đến giá trị và khả năng tiếp cận.
* **Mặt tiền:** Bề ngang mặt trước ngôi nhà, quan trọng với thẩm mỹ và kinh doanh.
* **Pháp lý:** Loại giấy tờ sở hữu (sổ đỏ, sổ hồng...), đảm bảo tính pháp lý.
* **Nội thất:** Mức độ trang bị nội thất đi kèm, ảnh hưởng đến giá trị sử dụng.
* **Ngày đăng:** Thời điểm bài viết được đăng, phục vụ phân tích thời gian thực.
* **URL:** Liên kết bài viết gốc để kiểm tra và xác minh khi cần.

## Thu thập dữ liệu đơn vị hành chính Việt Nam

Nhóm tiến hành thu thập thêm dữ liệu về **đơn vị hành chính Việt Nam** nhằm mở rộng các đặc trưng liên quan đến địa điểm trong phân tích bất động sản. Các thông tin được thu thập bao gồm:

* Dân số (người)
* Diện tích (km²)
* Mật độ dân số (người/km²)
* Số đơn vị hành chính cấp xã

**Nguồn dữ liệu:** [Wikipedia – Danh sách đơn vị hành chính cấp huyện của Việt Nam](https://vi.wikipedia.org/wiki/Danh_s%C3%A1ch_%C4%91%C6%A1n_v%E1%BB%8B_h%C3%A0nh_ch%C3%ADnh_c%E1%BA%A5p_huy%E1%BB%87n_c%E1%BB%A7a_Vi%E1%BB%87t_Nam)

**Công cụ sử dụng:**

* **Scrapy framework:** dùng để thu thập và trích xuất thông tin từ bảng dữ liệu HTML.

**Định dạng lưu trữ**: Dữ liệu được lưu dưới dạng **JSON**, thuận tiện cho xử lý và tích hợp với bộ dữ liệu bất động sản.

**Đánh giá quá trình thu thập**:

* Không bị CAPTCHA hay giới hạn truy cập
* Cấu trúc HTML rõ ràng, dễ parse
* Không cần Selenium hoặc kỹ thuật ngụy trang

Nhờ đó, toàn bộ quá trình được thực hiện đơn giản, nhanh chóng và hiệu quả.

# Tiền xử lý dữ liệu

## Phân tích sơ bộ tổng quan

### Thống kê số lượng thiếu của các đặc trưng

Trước tiên, ta tiến hành kiểm tra các đặc trưng và tỷ lệ giá trị thiếu (missing values) của các đặc trưng trong tập dữ liệu thu thập được từ trang web bất động sản. Kết quả được thể hiện trong **Hình 1**, dưới dạng phần trăm số giá trị bị thiếu tương ứng với từng cột.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Hình .Tỷ lệ giá trị thiếu theo từng đặc trưng

Từ bảng trên, có thể thấy rằng một số đặc trưng có tỷ lệ thiếu rất cao, đặc biệt là **Hướng ban công (87.40%)** và **Hướng nhà (78.82%)**.

Do tỷ lệ thiếu quá lớn, việc xử lý hoặc bổ sung dữ liệu cho hai đặc trưng này không khả thi hoặc không hiệu quả. Do đó, nhóm quyết định **loại bỏ hai đặc trưng "Hướng nhà" và "Hướng ban công" khỏi tập dữ liệu** để đảm bảo chất lượng đầu vào cho mô hình học máy.

Các đặc trưng còn lại sẽ được xử lý thiếu dữ liệu theo từng trường hợp cụ thể hoặc giữ nguyên nếu tỷ lệ thiếu không đáng kể.

### Thống kê số lượng giá trị thiếu trên mỗi bản ghi

Sau khi đã loại bỏ hai đặc trưng **"Hướng nhà"** và **"Hướng ban công"** do tỷ lệ thiếu quá cao, nhóm tiếp tục kiểm tra số lượng giá trị bị thiếu trên từng bản ghi.

**Bảng 1** thể hiện số lượng bản ghi tương ứng với số lượng giá trị bị thiếu, cùng với tỷ lệ phần trăm trên tổng số dữ liệu:

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Hình . Thống kê số lượng giá trị thiếu trên mỗi bản ghi

**Nhận xét:**

* Một số lượng lớn các bản ghi có từ 3 giá trị bị thiếu trở lên, chiếm tỷ lệ đáng kể trong toàn bộ tập dữ liệu. Điều này có thể ảnh hưởng tiêu cực đến hiệu quả của các mô hình học máy nếu giữ nguyên các bản ghi này.
* Để đảm bảo độ tin cậy và giảm nhiễu, nhóm quyết định **loại bỏ các bản ghi có từ 3 giá trị bị thiếu trở lên**.
* Sau khi lọc bỏ các bản ghi thiếu nhiều dữ liệu, tập dữ liệu còn lại khoảng **40.000 bản ghi** – vẫn đủ lớn và đảm bảo tính đại diện để thực hiện các phân tích thống kê cũng như huấn luyện mô hình dự đoán một cách hiệu quả.

### Kiểm tra kiểu dữ liệu

Tiếp theo, ta tiến hành **kiểm tra kiểu dữ liệu của các đặc trưng** trong tập dữ liệu nhằm đảm bảo rằng mỗi đặc trưng có kiểu dữ liệu phù hợp với bản chất thông tin mà nó biểu diễn.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Hình . Kiểu dữ liệu các đặc trưng

**Nhận xét:**

* Các cột như **Mức giá**, **Diện tích**, **Số phòng ngủ**, **Số phòng tắm, vệ sinh**, **Số tầng**, **Mặt tiền** thực chất là các đặc trưng dạng số. Vì vậy, ta cần **chuyển đổi về kiểu số (int hoặc float)** để phục vụ cho các bước xử lý tiếp theo như chuẩn hóa, trực quan hóa, và đưa vào mô hình dự đoán.
* Ngược lại, các cột như **Địa điểm**, **Pháp lý**, **Nội thất** là các đặc trưng dạng phân loại hoặc mô tả, nên giữ nguyên kiểu object. Tuy nhiên, các cột này cần được:
  + **Chuẩn hóa giá trị**: loại bỏ các ký tự không cần thiết, viết hoa/thường đồng nhất,...
  + **Xử lý lỗi chính tả hoặc biểu diễn không nhất quán**: để tránh tạo ra quá nhiều giá trị rời rạc trong quá trình mã hóa sau này.

### Phân tích đặc trưng số (Numerical)

Trong bước này, ta kiểm tra các đặc trưng số trong tập dữ liệu để đảm bảo đơn vị đo lường đã được chuẩn hóa và đồng nhất, tránh sai lệch trong phân tích và xây dựng mô hình.

Các đặc trưng số gồm: **Mức giá**, **Diện tích**, **Số phòng ngủ**, **Số phòng tắm/vệ sinh**, **Số tầng**, **Đường vào**, và **Mặt tiền**.

Với các đặc trưng không có đơn vị như "số phòng ngủ", số phòng tắm/vệ sinh, số tầng (là giá trị rời rạc, nguyên), không cần kiểm tra đơn vị.

Các đặc trưng có đơn vị như mức giá, diện tích, đường vào, mặt tiền cần được chuẩn hóa do có thể biểu diễn không đồng nhất (ví dụ: "3 tỷ", "2.5 triệu/m²", "120 m²", "12.5 mét").

* **Phân tích và xử lý cột Mức giá**

Trước tiên, ta thống kê số lượng các đơn vị xuất hiện trong cột Mức giá như sau:

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Nội dung do AI tạo ra có thể không chính xác.

**Nhận xét:**

* Đơn vị tỷ chiếm đa số, phù hợp cho phân tích giá nhà.
* Đơn vị triệu có thể quy đổi sang tỷ.
* Đơn vị triệu/m² biểu thị giá theo diện tích, có thể tính giá tổng bằng cách nhân với diện tích.
* Các đơn vị như Thỏa thuận, nghìn, tỷ/m², triệu/tháng không phù hợp và sẽ bị loại bỏ.
* **Phân tích các đặc trưng số còn lại**

Các đặc trưng số còn lại trong tập dữ liệu đã có đơn vị đồng nhất và hợp lý, do đó không cần thực hiện các bước chuẩn hóa bổ sung.

### Phân tích đặc trưng phân loại (Categorical)

Các đặc trưng phân loại trong dữ liệu bao gồm: Địa điểm, Hướng nhà, Pháp lý, và Nội thất.

* Đặc trưng Địa điểm

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, số

Nội dung do AI tạo ra có thể không chính xác.

Trong dữ liệu gốc, thông tin địa điểm được lưu dưới dạng chuỗi văn bản theo cấu trúc "Quận/Huyện, Tỉnh/Thành phố". Để phục vụ cho việc phân tích và xử lý dữ liệu hiệu quả hơn, nhóm đã tách trường này thành hai cột riêng biệt:

* Quận/Huyện
* Tỉnh/Thành phố

Việc tách riêng các cấp hành chính giúp thuận tiện hơn trong việc phân tích, thống kê theo từng khu vực, cũng như thực hiện các thao tác nhóm dữ liệu theo địa phương.

* Đặc trưng Pháp lý

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, màu đen

Nội dung do AI tạo ra có thể không chính xác.

Nhìn vào dữ liệu phân tích, có thể thấy cột Pháp lý chứa nhiều cách ghi khác nhau, từ các dạng chuẩn như *“Sổ đỏ”*, *“Sổ hồng”* đến các mô tả dài dòng, thiếu nhất quán như *“Giấy viết tay”*, *“Pháp lý: thuế phí chủ lo”*, v.v. Điều này gây khó khăn cho quá trình xử lý và phân tích.

**Kế hoạch xử lý:**

* Chuẩn hóa và phân loại thành 2 nhóm chính:
  + **Có pháp lý**: Sổ đỏ, Sổ hồng, Hợp đồng mua bán.
  + **Không pháp lý**: Giấy viết tay, các mô tả không đề cập cụ thể đến quyền sở hữu.
* Cần rà soát kỹ từ khóa để đảm bảo phân loại chính xác vào các nhóm.
* Đặc trưng Nội thất

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, màu đen

Nội dung do AI tạo ra có thể không chính xác.

Dữ liệu cho thấy cột **Nội thất** có nhiều cách ghi khác nhau, từ các giá trị chuẩn như *“Đầy đủ”*, *“Cơ bản”*, *“Không nội thất”* đến các mô tả chi tiết dài dòng, thiếu nhất quán.

**Kế hoạch xử lý:**

* Chuẩn hóa và phân loại thành 2 nhóm:
  + **Có nội thất**: Bao gồm *Đầy đủ*, *Cơ bản* và các mô tả cho thấy có trang bị nội thất.
  + **Không nội thất**: Bao gồm *Không nội thất* hoặc các mô tả tương đương.
* Cần xác định từ khóa đại diện chính xác để phân loại chính xác vào các nhóm.

## Phân tích khám phá chuyên sâu

### Tổng quan dữ liệu sau xử lý

Sau khi thực hiện làm sạch và chuẩn hóa dữ liệu theo kết quả của **Phân tích sơ bộ tổng quan (bước 1)**, tập dữ liệu còn lại 36.632 bản ghi với 11 đặc trưng. Các đặc trưng cũ đã được chuyển đổi sang tên tiếng Anh để thuận tiện cho việc phân tích chuyên sâu.

Cụ thể, các đặc trưng gồm:

|  |  |  |
| --- | --- | --- |
| **Tên biến** | **Kiểu dữ liệu** | **Mô tả ngắn** |
| Price | Float64 | Giá bất động sản (số thực) |
| Area | Float64 | Diện tích (số thực) |
| Bedrooms | Int64 | Số phòng ngủ (số nguyên) |
| Bathrooms | Int64 | Số phòng tắm (số nguyên) |
| Floors | Int64 | Số tầng (số nguyên) |
| AccessWidth | Float64 | Chiều rộng lối đi (số thực) |
| FacadeWidth | Float64 | Chiều rộng mặt tiền (số thực) |
| LegalStatus | string | Tình trạng pháp lý (chuỗi ký tự) |
| Furnishing | string | Tình trạng nội thất (chuỗi ký tự) |
| District | string | Quận/huyện (chuỗi ký tự) |
| Province | string | Tỉnh/thành phố (chuỗi ký tự) |

* Thống kê lại tỉ lệ thiếu:

Ảnh có chứa văn bản, ảnh chụp màn hình, số, biểu đồ

Nội dung do AI tạo ra có thể không chính xác.

Dựa vào biểu đồ tỷ lệ giá trị thiếu, ta thấy:

**AccessWidth** (Độ rộng hẻm) và **FacadeWidth** (Chiều ngang mặt tiền) là hai cột có tỷ lệ thiếu cao nhất, lần lượt là 30.66% và 24.63%.

→ Có thể cân nhắc loại bỏ hoặc thay thế tùy theo mức độ quan trọng của feature trong bài toán dự đoán.

Các cột **Bathrooms**, **Floors**, **Bedrooms** có tỷ lệ thiếu từ 5–10%, cần xem xét điền giá trị thay thế hợp lý hoặc sử dụng các kỹ thuật xử lý thiếu dữ liệu.

Các cột **Price, Area, LegalStatus, Furnishing, District, Province** không có giá trị thiếu.

### Phân tích biến mục tiêu (Price)

Ảnh có chứa hàng, Sơ đồ, biểu đồ, ảnh chụp màn hình

Nội dung do AI tạo ra có thể không chính xác.

Hình . Biểu đồ histogram và plotbox của Price

Biến Price có phân phối lệch phảirõ rệt, khi phần lớn bất động sản có giá dưới 20 tỷ đồng, trong khi các bất động sản giá trị cao xuất hiện ít và phân bố thưa dần. Đây là đặc trưng phổ biến trong dữ liệu bất động sản, phản ánh thị trường tập trung chủ yếu ở phân khúc trung bình và thấp.

Biểu đồ boxplot cho thấy nhiều ngoại lệ vượt quá 50 tỷ đồng, có thể là các bất động sản đặc biệt (biệt thự, dự án lớn) hoặc lỗi nhập liệu. Những giá trị này cần được kiểm tra và xử lý kỹ lưỡng để tránh làm sai lệch mô hình dự báo.

Dựa trên phân tích, nhóm quyết định:

* Loại bỏ hoặc cắt ngưỡng trên (capping) với các giá trị bất thường.
* Áp dụng biến đổi logarit (log-transform) cho biến Price để giảm độ lệch và giúp mô hình học được mối quan hệ tuyến tính tốt hơn.

### Phân bố các biến liên tục

* Dưới đây là hàng loạt các biểu đồ histogram và plotbox của các biến liên tục

Ảnh có chứa văn bản, biểu đồ, ảnh chụp màn hình, Song song

Nội dung do AI tạo ra có thể không chính xác.

**Nhận xét:**

Hầu hết các biến liên tục đều có phân phối lệch phải mạnh, với phần lớn giá trị tập trung ở mức thấp và tồn tại một số lượng nhỏ giá trị ngoại lệ rất lớn.

Những giá trị này có thể xuất phát từ một số bất động sản cao cấp như biệt thự hoặc dự án đặc biệt, tuy nhiên chúng chỉ chiếm tỷ lệ rất nhỏ và không mang tính đại diện cho thị trường mục tiêu. Nếu giữ lại, các giá trị này có thể gây nhiễu và làm sai lệch quá trình huấn luyện mô hình.

**Phương án xử lý:**

* **Áp dụng các biến đổi logarit hoặc Box-Cox** để giảm độ lệch của phân phối, giúp dữ liệu phù hợp hơn với giả định của các thuật toán học máy.
* **Chuẩn hóa dữ liệu** và sử dụng **phương pháp IQR** để phát hiện và xử lý ngoại lệ một cách có hệ thống.
* **Xử lý giá trị thiếu** bằng các phương pháp thống kê phù hợp như trung vị (median), trung bình (mean) hoặc mode.

Các bước xử lý này nhằm đảm bảo chất lượng dữ liệu đầu vào, tăng độ ổn định và hiệu quả của các mô hình dự đoán sau này.

### Phân bố nhãn biến phân loại

* Phân bố theo tình trạng pháp lý (LegalStatus)

Ảnh có chứa ảnh chụp màn hình, văn bản, biểu đồ, Sơ đồ

Nội dung do AI tạo ra có thể không chính xác.

Dữ liệu khảo sát cho thấy phần lớn các bất động sản đều có tình trạng pháp lý hợp lệ (giá trị yes), chiếm khoảng 35.000 bản ghi trong tổng số. Ngược lại, số lượng bất động sản không có pháp lý rõ ràng (no) rất ít, chỉ khoảng 1.500 bản ghi.

**Nhận xét:**

* Thị trường bất động sản được khảo sát chủ yếu bao gồm các sản phẩm có pháp lý minh bạch, phản ánh xu hướng ưu tiên các sản phẩm hợp pháp từ phía người bán hoặc các sàn giao dịch.
* Điều này cũng có thể xuất phát từ nhu cầu người mua muốn đảm bảo tính pháp lý rõ ràng khi giao dịch.
* Tuy nhiên, khả năng dữ liệu bị “lọc” khi loại bỏ những sản phẩm không rõ pháp lý cũng cần được lưu ý, điều này có thể ảnh hưởng đến tính toàn diện của bộ dữ liệu.
* Phân bố theo tình trạng Nội thất (Furnishing)

Ảnh có chứa ảnh chụp màn hình, văn bản, biểu đồ, Sơ đồ

Nội dung do AI tạo ra có thể không chính xác.

* Phân bố theo tỉnh thành (Province)

Ảnh có chứa ảnh chụp màn hình, hàng, Sơ đồ, biểu đồ

Nội dung do AI tạo ra có thể không chính xác.

Phân bố dữ liệu theo tỉnh thành cho thấy phần lớn các bản ghi tập trung ở hai đô thị lớn là Hà Nội và TP. Hồ Chí Minh, trong khi các tỉnh thành khác chỉ chiếm tỷ lệ nhỏ. Sự phân bố không đều này dẫn đến độ lệch về địa lý trong tập dữ liệu.

* Nhằm khắc phục hiện tượng mất cân bằng trong phân bố dữ liệu giữa các tỉnh thành, nhóm nghiên cứu đã bổ sung một số đặc trưng phản ánh mức độ phát triển kinh tế – xã hội của từng khu vực. Trong đó, hai biến quan trọng gồm:
* **GDP\_USD**: Tổng sản phẩm trên địa bàn (GDP) của mỗi tỉnh thành, tính theo đơn vị USD.
* **Distribute**: Mật độ dân số trung bình của các đơn vị hành chính cấp quận/huyện thuộc từng tỉnh.

Hai đặc trưng này đóng vai trò phản ánh trình độ phát triển tương đối của các tỉnh thành và giúp mô hình học máy tiếp cận dữ liệu một cách cân bằng và hiệu quả hơn.

## Các bước xử lý chi tiết

Toàn bộ quá trình làm sạch và chuẩn hóa dữ liệu được chia thành 5 bước chính, nhằm đảm bảo chất lượng dữ liệu đầu vào trước khi đưa vào phân tích và xây dựng mô hình dự đoán giá bất động sản.

### Tiền xử lý cơ bản

* Chuẩn hóa cấu trúc dữ liệu

**Loại bỏ thông tin không cần thiết:** Các cột như Tiêu đề, URL, Ngày đăng, Hướng nhà, Hướng ban công được loại bỏ để giảm nhiễu và tập trung vào các thuộc tính có giá trị phân tích cao.

**Chuẩn hóa tên cột:** Đổi toàn bộ tên cột sang tiếng Anh, đồng bộ định dạng, tránh lỗi trong quá trình xử lý và lập trình.

* Chuẩn hóa giá trị số

**Xử lý định dạng số:** Loại bỏ ký tự đặc biệt, chuẩn hóa dấu phân cách thập phân nhằm đảm bảo dữ liệu có thể chuyển đổi sang kiểu số một cách chính xác.

**Chuẩn hóa đơn vị tiền tệ:** Chuyển toàn bộ giá trị về đơn vị tỷ đồng, xử lý các trường hợp đặc biệt như thỏa thuận, nghìn, triệu/tháng, tỷ/m².

* Chuẩn hóa thông tin địa điểm

**Tách thông tin địa điểm:** Phân tách địa chỉ thành hai phần: District (quận/huyện) và Province (tỉnh/thành phố).

**Chuẩn hóa tên địa phương:** Chuyển về chữ thường, loại bỏ khoảng trắng và ký tự thừa.

**Xử lý trường hợp đặc biệt:** Ánh xạ lại các địa danh cũ (đổi tên, sáp nhập), đảm bảo tính nhất quán trong toàn bộ tập dữ liệu.

### Kỹ thuật trích xuất đặc trưng

* **Bổ sung thông tin địa lý và kinh tế**

Dữ liệu được merge (gộp) với thông tin thu thập từ Wikipedia nhằm bổ sung các đặc trưng địa phương chi tiết đến cấp quận/huyện, bao gồm:

* **Phân bố dân cư:** Dùng để ước lượng mật độ dân số.
* **GDP địa phương (GDP)**: Thu thập dữ liệu GRDP của các tỉnh/thành để phản ánh tình hình kinh tế địa phương.

### Xử lý phân phối lệch

Để xử lý các biến có phân phối **lệch phải mạnh**, phương pháp **log-transform** được áp dụng nhằm đưa phân phối về gần chuẩn, từ đó giúp mô hình học máy hoạt động hiệu quả hơn và **giảm ảnh hưởng của các giá trị ngoại lệ**.

* **Phương pháp sử dụng**: Hàm log1p(x) (tương đương log(1 + x)), đảm bảo xử lý được cả giá trị 0.
* **Các biến được áp dụng**: Price, Area, Bedrooms, Bathrooms, Floors, AccessWidth, FacadeWidth.

Kết quả:

Ảnh có chứa văn bản, biểu đồ, Song song, Sơ đồ

Nội dung do AI tạo ra có thể không chính xác.

**Kết quả:**

* Sau khi biến đổi, các biến **Price**, **Area**, **Floors** có phân phối gần chuẩn hơn.
* Các biến **Bedrooms**, **Bathrooms**, **FacadeWidth** và **AccessWidth** tuy chưa đạt chuẩn hoàn toàn nhưng **giảm rõ rệt độ lệch** và dễ xử lý hơn trong bước modeling tiếp theo.

### Xử lý giá trị thiếu

Việc xử lý giá trị thiếu được thực hiện dựa trên bản chất và phân phối của từng biến:

* Đối với các biến phân loại rời rạc như **Bedrooms**, **Bathrooms**, và **Floors**, phương pháp **mode** (giá trị xuất hiện nhiều nhất) được lựa chọn. Phương pháp này giúp giữ nguyên tính thực tế của dữ liệu, tránh tạo ra các giá trị trung gian không hợp lý (ví dụ như 2.5 phòng ngủ).
* Đối với các biến liên tục như **AccessWidth** và **FacadeWidth**, phương pháp **median** (trung vị) được sử dụng do các biến này có phân phối lệch phải rõ rệt. Sử dụng trung vị giúp giảm thiểu ảnh hưởng của các giá trị ngoại lệ, đồng thời giữ được sự ổn định và tính đại diện của dữ liệu.

### Xử lý ngoại lệ

Việc xử lý ngoại lệ được thực hiện dựa trên đặc điểm phân phối và bản chất của từng biến nhằm loại bỏ các điểm dữ liệu bất thường có thể ảnh hưởng tiêu cực đến hiệu quả của mô hình, đồng thời giữ lại các giá trị hợp lý về mặt thực tế.

* **Dữ liệu Giá, Diện tích và Số tầng (Price, Area, Floors):**
* **Phương pháp áp dụng**: Sử dụng **Z-score** với ngưỡng ±3.
* **Lý do lựa chọn**: Các biến này có phân phối gần chuẩn sau khi log-transform, nên phương pháp Z-score phù hợp để phát hiện các giá trị lệch khỏi trung bình. Việc sử dụng ngưỡng ±3 giúp loại bỏ các điểm thực sự bất thường mà vẫn giữ lại các giá trị hợp lệ, đảm bảo tính ổn định của dữ liệu.
* **Dữ liệu Số phòng và Kích thước (Bedrooms, Bathrooms, AccessWidth, FacadeWidth):**
* **Phương pháp áp dụng**: Sử dụng **Interquartile Range (IQR)** với hệ số 1.5.
* **Lý do lựa chọn**: Các biến này có phân phối lệch phải rõ rệt và chứa nhiều giá trị ngoại lệ hợp lý. Phương pháp IQR giúp phát hiện các giá trị quá xa so với vùng phân bố chính, đồng thời hạn chế loại bỏ các dữ liệu vẫn hợp lý về mặt thực tế.

## Kết quả

Sau khi thực hiện toàn bộ các bước làm sạch và chuẩn hóa dữ liệu, bao gồm tiền xử lý cơ bản, trích xuất đặc trưng, xử lý phân phối lệch, giá trị thiếu và ngoại lệ:

* **Số lượng bản ghi**:
  + Ban đầu: **49.304** bản ghi.
  + Sau xử lý: **26.629** bản ghi còn lại.
  + **Tỷ lệ giữ lại ~54%** dữ liệu, đảm bảo chất lượng và loại bỏ các bản ghi không phù hợp cho mô hình hóa.
* **Phân phối các đặc trưng số**:

Ảnh có chứa văn bản, biểu đồ, Kế hoạch, Bản vẽ kỹ thuật

Nội dung do AI tạo ra có thể không chính xác.

**Ảnh có chứa văn bản, biểu đồ, Kế hoạch, Bản vẽ kỹ thuật

Nội dung do AI tạo ra có thể không chính xác.**

* + Các biến như **Price, Area, Floors** sau khi log-transform và xử lý ngoại lệ bằng Z-score đã có phân phối gần chuẩn.
  + Các biến **Bedrooms, Bathrooms, FacadeWidth, AccessWidth** tuy không hoàn toàn đạt chuẩn nhưng độ lệch đã được giảm đáng kể, phù hợp hơn cho việc đưa vào mô hình học máy.
  + **Biểu đồ phân phối** minh họa rõ sự cải thiện về mặt hình dạng phân phối và loại bỏ phần lớn giá trị ngoại lệ.
* **Các đặc trưng phân loại**:
  + Các đặc trưng dạng phân loại (categorical) như **District, Province** đã được chuẩn hóa về mặt định dạng và nội dung (chuyển về tiếng Anh, chữ thường, loại bỏ khoảng trắng/thừa, ánh xạ tên cũ).
  + Các biến phân loại rời rạc như **Bedrooms, Bathrooms, Floors** cũng được xử lý giá trị thiếu bằng mode để giữ tính hợp lý và nhất quán trong dữ liệu.
* **Đánh giá chất lượng dữ liệu sau xử lý**:
  + **Không còn giá trị thiếu** trong các biến quan trọng.
  + **Giá trị ngoại lệ đã được loại bỏ hợp lý**, giúp cải thiện chất lượng đầu vào và tránh gây nhiễu khi huấn luyện mô hình.
  + **Cấu trúc và định dạng dữ liệu đồng nhất**, sẵn sàng phục vụ cho bước phân tích, trực quan hóa và xây dựng mô hình dự đoán.

# Mô hình học máy

Giá nhà đất là một dạng số thực, nên không thể giải quyết bài toán bằng bài toán phân loại được. Nhóm em sử dụng cả hai giải thuật hồi quy tuyến tính k láng giềng gần nhất, và XGBoost. Sau đó so sánh kết quả để chọn ra mô hình tốt nhất

## Giải thuật hồi quy tuyến tính

Giải thuật hồi quy tuyến tính cố gắng tính toán sự tương quan giữa dữ liệu đầu vào và kết quả tương ứng đầu ra. Ví dụ, sử dụng hồi quy tuyến tính để xét xem sự tương quan giữa các đặc điểm của căn nhà và giá nhà như thế nào từ các dữ liệu đầu vào, từ đó suy ra được một “công thức” gần đúng với các biến đầu vào là các đặc điểm của căn nhà, và đầu ra là giá của căn nhà đó. Từ đó chỉ cần biết được các đặc điểm của căn nhà, áp dụng “công thức” trên dự đoán được giá nhà là bao nhiêu.

A black background with a black square

AI-generated content may be incorrect.

Đầu vào của giải thuật hồi quy tuyến tính là một vectơ n chiều A black background with a black square

AI-generated content may be incorrect..

Đầu ra của giải thuật là một giá trị thực, được tính theo công thức đã xác định được từ quá trình đưa tập dữ liệu ban đầu vào giải thuật.

**Khi nào sử dụng giải thuật hồi quy tuyến tính**

Giải thuật hồi quy tuyến tính thực hiện các tính toán tuyến tính, đơn giản, dễ cài đặt và thuật toán dễ hiểu. Chi phí tính toán thấp (bởi vì các phép toán được thực hiện tuyến tính). Thời gian dự đoán nhanh, cho kết quả ngay, phù hợp với những dạng bài toán có độ phức tạp không quá cao và cần thời gian dự đoán ngay lập tức.

## Giải thuật k láng giềng gần nhất

Giải thuật k láng giềng gần nhất (K-Nearest Neighbor hay KNN) tìm ra k căn nhà (trong tổng số những căn nhà đã biết giá) có “khoảng cách” gần với căn nhà cần dự đoán giá nhất. Hay nói cách khác, từ những đặc điểm của các căn nhà đã biết giá và căn nhà cần dự đoán giá, thuật toán trên sẽ tính được k căn nhà có những đặc điểm giống với căn nhà cần dự đoán nhất, sau đó giá của căn nhà cần dự đoán sẽ được tính dựa trên giá của k căn nhà đó.

Mô tả giải thuật kNN

K láng giềng gần nhất (tên Tiếng Anh là K-Nearest Neighbor, viết tắt là KNN) là một giải thuật học máy có giám sát đơn giản, dễ cài đặt và có thể sử dụng để giải quyết cả hai dạng bài toán phân loại và hồi quy.

Ý tưởng của phương pháp học dựa trên láng giềng gần nhất đó là với một tập các ví dụ học, thuật toán đơn giản là lưu lại các ví dụ học, chưa xây dựng mô hình mô tả rõ ràng và tổng quát của hàm mục tiêu cần học. Đối với một ví dụ cần phân loại hay hồi quy, thuật toán xem xét ví dụ đó với các ví dụ học để gán giá trị của hàm mục tiêu (một nhãn hay một giá trị thực)

***Biểu diễn đầu vào của thuật toán***

Mỗi ví dụ x được biểu diễn là một vectơ n chiều trong không gian các vectơ X ∈ Rn.

x = (x1,x1,…,xn), trong đó xi (∈R) là một số thực. Có thể áp dụng với hai kiểu bài toán: ![A black background with a black square

AI-generated content may be incorrect.]()bài toán phân lớp (hàm mục tiêu có giá trị rời rạc, đầu ra của hệ thống là một trong số các giá trị rời rạc được xác định trước, là một trong các nhãn lớp) hoặc bài toán hồi quy (hàm mục tiêu có giá trị liên tục, đầu ra của hệ thống là một số thực không biết trước).

***Giải thuật phân lớp KNN***

Mỗi ví dụ học x được biểu diễn bởi 2 thành phần, mô tả của ví dụ: x=(x1,x2,…,xn), trong đó xi∈R, và giá trị đầu ra mong muốn: yx∈R (là một số thực). Trong giai đoạn học, thuật toán lưu lại các ví dụ học trong tập học D. Trong giai đoạn dự đoán, đối với mỗi ví dụ học x∈D, tính khoảng cách giữa x và z (ví dụ cần dự đoán), từ đó xác định được các láng giềng gần nhất của z, gồm k ví dụ học trong D gần nhất với z tính theo một hàm khoảng cách d.

***Các hàm tính khoảng cách***

*Các hàm tính khoảng cách hình học*

* Hàm Minkowski (p-norm)

A black background with a black square

AI-generated content may be incorrect.

* Hàm Mahattan (p=1)

A black background with a black square

AI-generated content may be incorrect.

* Hàm Euclid (p=2)

A black background with a black square

AI-generated content may be incorrect.

* *Hàm khoảng cách Hamming*

Các thuộc tính đầu vào là kiểu nhị phân ({0, 1}), ví dụ x = (0, 1, 1, 0, 1)

A black background with a black square

AI-generated content may be incorrect.

Với A black background with a black square

AI-generated content may be incorrect.

Hàm tính độ tương tự Cosine

Độ tương tự cosine được tính:

A black background with a black square

AI-generated content may be incorrect.

Suy ra khoảng cách giữa x và z là: A black background with a black square

AI-generated content may be incorrect.

***Chuẩn hoá miền giá trị thuộc tính***

Lấy ví dụ với hàm tính khoảng cách Euclid ở trên. Giả sử mỗi ví dụ được biểu diễn bởi 3 thuộc tính là Age (tuổi), Income (thu nhập hàng tháng), Height (chiều cao đo theo mét).

Ảnh có chứa màu đen, bóng tối

Nội dung do AI tạo ra có thể không chính xác.

Ảnh có chứa màu đen, bóng tối

Nội dung do AI tạo ra có thể không chính xác.

Áp dụng công thức tính khoảng cách Euclid ta có khoảng cách giữa x và z là:

Ảnh có chứa màu đen, bóng tối

Nội dung do AI tạo ra có thể không chính xác.

Thấy rằng giá trị khoảng cách bị quyết định chủ yếu bởi giá trị khoảng cách (sự khác biệt) giữa hai ví dụ đối với thuộc tính Income. Vì thuộc tính này có miền giá trị rất lớn so với các thuộc tính còn lại.

Do đó cần phải chuẩn hoá miền giá trị (đưa về cùng một khoảng giá trị). Khoảng giá trị [0, 1] thường được sử dụng. Đối với mỗi thuộc tính

A black background with a black square

AI-generated content may be incorrect..

**Trọng số của các thuộc tính**

Lấy ví dụ với hàm khoảng cách Euclid, tất cả các thuộc tính có ảnh hưởng như nhau đối với giá trị khoảng cách.

Các thuộc tính khác nhau có thể (nên) có mức độ ảnh hưởng khác nhau đối với giá trị khoảng cách. Cần phải tích hợp các giá trị trọng số của các thuộc tính trong hàm tính khoảng cách.

A black background with a black square

AI-generated content may be incorrect.với wi là trọng số của thuộc tính i

Làm sao để xác định các giá trị trọng số của các thuộc tính? Cần 2 yếu tố để xác định giá trị trọng số, yếu tố thứ nhất là dựa trên các tri thức cụ thể của bài toán (ví dụ: được chỉ định bởi các chuyên gia trong lĩnh vực của bài toán đang xét), yếu tố thứ hai là bằng một quá trình tối ưu hoá các giá trị trọng số (ví dụ: sử dụng một tập học để học một bộ các giá trị trọng số tối ưu).

***Sử dụng KNN khi nào***

Những bài toán có các đặc điểm sau sẽ phù hợp khi lựa chọn giải thuật kNN: các ví dụ được biểu diễn trong không gian vectơ Rn, số lượng các thuộc tính để biểu diễn ví dụ là không nhiều, một tập học có kích thước lớn.

Ưu điểm của giải thuật kNN: giải thuật có chi phí thấp cho quá trình huấn luyện (chỉ lưu lại các ví dụ học), hoạt động tốt với các bài toán phân loại cho n lớp do đó không cần phải học n bộ phân loại cho n lớp, việc chọn k phù hợp (thường lớn hơn nhiều so với 1) có khả năng xử lý nhiễu cao bởi phân loại/dự đoán được thực hiện dựa trên k láng giềng gần nhất.

Nhược điểm của giải thuật kNN: phải lựa chọn hàm tính khoảng các thích hợp với từng bài toán, chi phí tính toán cao (thời gian, bộ nhớ) tại thời điểm phân loại/dự đoán, có thể cho kết quả kém/sai với các thuộc tính không liên quan.

## XGBoost

1. **Tổng quan về XGBoost**

XGBoost (eXtreme Gradient Boosting) là một thuật toán học máy tiên tiến được phát triển bởi Tianqi Chen vào năm 2016. Đây là một phiên bản cải tiến của thuật toán Gradient Boosting, được thiết kế để tối ưu hóa hiệu suất và tốc độ tính toán. XGBoost đã chứng minh được hiệu quả vượt trội trong nhiều cuộc thi machine learning và các ứng dụng thực tế, đặc biệt là trong lĩnh vực dự đoán giá bất động sản.

1. **Nguyên lý hoạt động của XGBoost**

**Cơ sở lý thuyết**

XGBoost hoạt động dựa trên nguyên lý boosting, trong đó các mô hình yếu (weak learners) được kết hợp tuần tự để tạo thành một mô hình mạnh. Mỗi mô hình mới được huấn luyện để sửa lỗi của các mô hình trước đó. Điểm khác biệt chính của XGBoost so với các thuật toán boosting khác là việc sử dụng gradient descent để tối ưu hóa hàm mất mát.

**Hàm mất mát và tối ưu hóa**

Trong bài toán hồi quy, XGBoost sử dụng hàm mất mát bình phương sai số:

L(y, ŷ) = (y - ŷ)²

Trong đó:

* y là giá trị thực tế
* ŷ là giá trị dự đoán

Để tối ưu hóa hàm mất mát này, XGBoost sử dụng phương pháp gradient descent, tính toán gradient (đạo hàm bậc nhất) và hessian (đạo hàm bậc hai) của hàm mất mát:

* Gradient: g = ∂L/∂ŷ = -2(y - ŷ)
* Hessian: h = ∂²L/∂ŷ² = 2

**Quá trình xây dựng cây quyết định**

XGBoost xây dựng cây quyết định một cách tham lam (greedy) bằng cách tối ưu hóa độ lợi thông tin (Gain) cho mỗi cách tách node. Độ lợi thông tin được tính theo công thức:

Gain = [GL²/(HL + λ) + GR²/(HR + λ) - (GL + GR)²/(HL + HR + λ)] - γ

Trong đó:

* GL, HL: tổng gradient và hessian của node con trái
* GR, HR: tổng gradient và hessian của node con phải
* λ: tham số L2 regularization
* γ: tham số điều chỉnh việc tách node

1. **Ứng dụng XGBoost trong Project Dự đoán Giá Bất động sản**

**Cấu hình mô hình**

Trong project này, XGBoost được triển khai với cấu hình tối ưu sau:

XGBRegressor(

    n\_estimators=200,                 # Số lượng cây

    max\_depth=6,                        # Độ sâu tối đa của mỗi cây

    learning\_rate=0.1,                  # Tốc độ học

    subsample=0.8,                      # Tỷ lệ mẫu cho mỗi cây

    colsample\_bytree=0.8,          # Tỷ lệ features cho mỗi cây

    min\_child\_weight=1,            # Trọng số tối thiểu trong node con

    gamma=0,                             # Tham số điều chỉnh việc tách node

    reg\_alpha=0.1,                      # L1 regularization

    reg\_lambda=1,                      # L2 regularization

    objective='reg:squarederror'  # Hàm mất mát

)

**Quy trình xử lý dữ liệu:**

* Tiền xử lý: xử lý dữ liệu thiếu, mã hóa biến phân loại, chuẩn hóa dữ liệu số
* Chia tập dữ liệu: 80% huấn luyện, 20% kiểm tra
* Huấn luyện mô hình với 200 cây quyết định
* Đánh giá hiệu suất thông qua các metrics: MSE, MAPE, MAE, R²

# Đánh giá mô hình

Nhóm sử dụng các metrics để đánh giá điểm của một mô hình. Các metrics được chọn là:

**MAE (Mean Absolute Error – Sai số tuyệt đối trung bình)**

* **Cơ chế hoạt động**: MAE đo lường mức độ sai lệch trung bình tuyệt đối giữa giá trị thực tế và giá trị dự đoán. Nói cách khác, nó cho biết mô hình dự đoán sai bao nhiêu đơn vị, trung bình trên toàn bộ dữ liệu.
* **Công thức**:

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.

**RMSE (Root Mean Squared Error – Căn bậc hai của sai số bình phương trung bình)**

* **Cơ chế hoạt động**: RMSE đánh giá mức độ sai số bằng cách lấy trung bình bình phương sai số, sau đó căn bậc hai. RMSE nhấn mạnh những sai số lớn hơn do việc bình phương sai số.
* **Công thức**:

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.

**R² Score (Hệ số xác định – R-squared)**

* **Cơ chế hoạt động**: R² cho biết mô hình giải thích được bao nhiêu phần trăm phương sai của dữ liệu thực. Giá trị R² nằm trong khoảng từ 0 đến 1 (đôi khi có thể nhỏ hơn 0 nếu mô hình quá tệ).
* **Công thức**:

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.

**MAPE (Mean Absolute Percentage Error – Sai số phần trăm tuyệt đối trung bình)**

* **Cơ chế hoạt động**: MAPE đo sai số trung bình theo phần trăm so với giá trị thực tế. Đây là thước đo phổ biến để đánh giá mức độ chính xác tương đối.
* **Công thức**:

A number and symbols on a white background

AI-generated content may be incorrect.

Áp dụng các phương pháp đánh giá mô hình vào 3 mô hình nhóm đã chọn:

**Hồi quy tuyến tính (Linear Regression)**

Thực hiện đưa dữ liệu train vào giải thuật hồi quy tuyến tính sau đó dự đoán trên tập test thu được kết quả:

* RMSE: 3.38 (tỷ đồng)
* R2 Score: 0.517 (51.7%)
* MAE: 2.50 (tỷ đồng)
* MAPE: 29.75%

**XGBoost**

Mô hình XGBoost cho kết quả tốt nhất trong các mô hình được thử nghiệm:

* RMSE: 2.20 (tỷ đồng)
* R2 Score: 0.796 (79.6%)
* MAE: 1.51 (tỷ đồng)
* MAPE: 18.03%

**K láng giềng gần nhất (KNN)**

Mô hình KNN với k=5 và weights='distance' cho kết quả:

* RMSE: 2.93 (tỷ đồng)
* R2 Score: 0.636 (63.6%)
* MAE: 1.95 (tỷ đồng)
* MAPE: 23.25%

**Kết luận**

Qua thực nghiệm với 3 mô hình khác nhau, có thể thấy:

**XGBoost vượt trội toàn diện**:  
Mô hình XGBoost cho kết quả tốt nhất ở cả bốn chỉ số đánh giá. Cụ thể, RMSE và MAE đều thấp nhất, cho thấy sai số tuyệt đối nhỏ hơn các mô hình còn lại. Đồng thời, MAPE chỉ 18.03% – mức sai số tương đối khá thấp – và R² đạt 79.6%, cho thấy mô hình có khả năng giải thích gần 80% biến động của dữ liệu đầu ra. Điều này phản ánh tính linh hoạt và khả năng học phi tuyến tính tốt của XGBoost trong bài toán phức tạp như định giá nhà đất.

**Hồi quy tuyến tính thể hiện hạn chế rõ rệt**:  
Với RMSE và MAE cao nhất (3.38 và 2.50 tỷ đồng), cùng với R² chỉ đạt 51.7%, mô hình tuyến tính tỏ ra không phù hợp trong việc mô hình hóa mối quan hệ phức tạp giữa các đặc trưng đầu vào và giá trị đầu ra. MAPE gần 30% cho thấy sai số dự đoán tương đối lớn so với giá trị thực tế.

**KNN ở mức trung bình, nhưng vẫn thua XGBoost**:  
KNN cho kết quả khá hơn hồi quy tuyến tính nhưng vẫn kém hơn XGBoost trên cả bốn mặt. Đặc biệt, RMSE là 2.93 tỷ và MAE là 1.95 tỷ – cao hơn nhiều so với XGBoost. Ngoài ra, R² = 63.6% cho thấy mô hình này có cải thiện nhất định nhưng chưa khai thác được đầy đủ cấu trúc dữ liệu.