**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH**

**NGUYỄN MINH KHANG**

**ỨNG DỤNG MÔ HÌNH HỌC MÁY TRONG VIỆC DỰ ĐOÁN GIÁ NHÀ**

**ĐỒ ÁN NGÀNH**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**TP. HỒ CHÍ MINH, 2025BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH**

**NGUYỄN MINH KHANG**

**ỨNG DỤNG MÔ HÌNH HỌC MÁY TRONG VIỆC DỰ ĐOÁN GIÁ NHÀ**

**Mã số sinh viên: 2251050037**

**ĐỒ ÁN NGÀNH**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**Giảng viên hướng dẫn: NGUYỄN THỊ MAI TRANG**

**TP. HỒ CHÍ MINH, 2025**

LỜI CẢM ƠN

Trước hết, em xin gửi lời cảm ơn chân thành đến cô **Nguyễn Thị Mai Trang**, giảng viên hướng dẫn, đã luôn tận tình định hướng và hỗ trợ em trong suốt quá trình thực hiện đề tài “Ứng dụng Machine Learning trong dự đoán giá nhà”. Sự chỉ bảo và những góp ý quý báu của cô là nguồn động lực và nền tảng quan trọng giúp em hoàn thành đồ án này.

Trong quá trình nghiên cứu, em đã học hỏi thêm nhiều kiến thức chuyên ngành về trí tuệ nhân tạo, phân tích dữ liệu cũng như phương pháp tiếp cận khoa học khi giải quyết vấn đề. Đây là cơ hội quý báu để em rèn luyện kỹ năng nghiên cứu, tư duy logic và tích lũy kinh nghiệm cho con đường học tập và làm việc sau này.

Mặc dù đã nỗ lực hết mình, song do hạn chế về thời gian và kinh nghiệm, đồ án khó tránh khỏi những thiếu sót. Em rất mong nhận được những nhận xét và góp ý từ quý thầy cô để tiếp tục hoàn thiện hơn trong tương lai.

**Em xin trân trọng cảm ơn!**

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

TÓM TẮT ĐỐ ÁN NGÀNH

Đồ án này tập trung vào việc xây dựng một mô hình dự đoán giá nhà bằng cách ứng dụng các mô hình học máy. Dữ liệu bất động sản được thu thập và xử lý với các đặc trưng như diện tích, vị trí, số phòng, loại nhà, pháp lý, v.v. Mô hình **Random Forest, Linear Regresstion** được sử dụng để dự đoán giá nhà dựa trên các thuộc tính đầu vào. Đồng thời, mô hình **Logistic Regression** được áp dụng để **phân loại loại hình nhà ở** (ví dụ: nhà phố, chung cư, biệt thự, v.v.). Kết quả cho thấy các mô hình đều mang lại độ chính xác cao và phù hợp với từng mục tiêu cụ thể. Hệ thống được tích hợp vào giao diện web thân thiện, giúp người dùng dễ dàng nhập thông tin và nhận được kết quả dự đoán tức thì, hỗ trợ hiệu quả trong việc định giá và ra quyết định mua bán bất động sản.

ABSTRACT

This project focuses on building a house price prediction model by applying machine learning models. Real estate data is collected and processed with features such as area, location, number of rooms, house type, legality, etc. Random Forest and Linear Regression models are used to predict house prices based on input attributes. At the same time, Logistic Regression models are applied to classify housing types (eg townhouses, apartments, villas, etc.). The results show that the models all provide high accuracy and are suitable for each specific goal. The system is integrated into a friendly web interface, helping users easily enter information and receive instant prediction results, effectively supporting real estate valuation and decision making.

**MỤC LỤC**

[LỜI CẢM ƠN 1](#_Toc208694359)

[NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN 2](#_Toc208694360)

[TÓM TẮT ĐỐ ÁN NGÀNH 3](#_Toc208694361)

[ABSTRACT 4](#_Toc208694362)

[DANH MỤC HÌNH VẼ 8](#_Toc208694363)

[MỞ ĐẦU 10](#_Toc208694364)

[Chương 1. GIỚI THIỆU ĐỀ TÀI 11](#_Toc208694365)

[1.1. Giới thiệu đề tài 11](#_Toc208694366)

[1.2. Lý do chọn đề tài 11](#_Toc208694367)

[1.3. Bố cục báo cáo 11](#_Toc208694368)

[Chương 2. CƠ SỞ LÝ THUYẾT 13](#_Toc208694369)

[2.1. Giới thiệu học máy 13](#_Toc208694370)

[2.1.1. Định nghĩa 13](#_Toc208694371)

[2.1.2. Phân loại mô hình học máy 13](#_Toc208694372)

[2.2. Thuật toán Random Forest 14](#_Toc208694373)

[2.2.1. Định nghĩa 14](#_Toc208694374)

[2.2.2. Ưu điểm 14](#_Toc208694375)

[2.2.3. Nhược điểm 14](#_Toc208694376)

[2.4. Flask – Framework xây dựng hệ thống web 15](#_Toc208694377)

[2.4.1. Định nghĩa 15](#_Toc208694378)

[2.4.2. Lý do chọn Flask 15](#_Toc208694379)

[2.4.3. Kiến trúc cơ bản 15](#_Toc208694380)

[2.5. HTML-CSS-BOOTSTRAP 16](#_Toc208694381)

[2.5.1. Định nghĩa 16](#_Toc208694382)

[2.5.2. Ưu điểm 16](#_Toc208694383)

[2.6. Github – Quản lý mã nguồn 17](#_Toc208694384)

[2.6.1. Định nghĩa 17](#_Toc208694385)

[2.6.2. Ưu điểm 17](#_Toc208694386)

[2.7. Pandas 17](#_Toc208694387)

[2.7.1. Định nghĩa 17](#_Toc208694388)

[2.8. Matplotlib & Seaborn 18](#_Toc208694389)

[2.8.1. Định nghĩa 18](#_Toc208694390)

[2.9. PowerBI 19](#_Toc208694391)

[2.9.1. Định nghĩa 19](#_Toc208694392)

[Chương 3. PHÂN TÍCH DỮ LIỆU VÀ HUẤN LUYỆN MÔ HÌNH. 20](#_Toc208694393)

[3.1. Tổng quan bộ dữ liệu 20](#_Toc208694394)

[3.1.1. Giới thiệu bài toán và mục tiêu dự án 20](#_Toc208694395)

[3.1.2. Lý do chọn mô hình 20](#_Toc208694396)

[3.2. Khai phá và phân tích bộ dữ liệu 21](#_Toc208694397)

[3.2.1. Nguồn gốc bộ dữ liệu 21](#_Toc208694398)

[3.2.2. Các bước tiền xử lý dữ liệu 22](#_Toc208694399)

[3.3. Trực quan hóa dữ liệu khi được làm sạch 30](#_Toc208694400)

[3.3.1. Biểu đồ phân bố giá nhà 30](#_Toc208694401)

[3.3.2. Biểu đồ phân bố giá nhà theo trung bình quận. 31](#_Toc208694402)

[3.3.3. Mối quan hệ giữa diện tích với giá nhà 31](#_Toc208694403)

[3.3.4. Mối quan hệ giữa giá nhà với số phòng ngủ 32](#_Toc208694404)

[3.4. Huấn luyện mô hình dự đoán giá nhà 32](#_Toc208694405)

[3.4.1. Huấn luyện mô hình Random Forest 32](#_Toc208694406)

[3.4.2. Huấn luyện mô hình Linear Regression 39](#_Toc208694407)

[3.5. So sánh độ chính xác của hai mô hình 44](#_Toc208694408)

[3.5.1. MAE, RMSE. 44](#_Toc208694409)

[3.5.2. R2 45](#_Toc208694410)

[3.6. Trực quan hóa dữ liệu với Power BI 46](#_Toc208694411)

[3.6.1. Tổng quan (phân bố bedroom, wc, district) 46](#_Toc208694412)

[3.6.2. Quan hệ Giá – Diện Tích – Quận 47](#_Toc208694413)

[Chương 4. HƯỚNG KẾT LUẬN VÀ PHÁT TRIỂN 49](#_Toc208694414)

[4.1. Kết quả đạt được và hạn chế 49](#_Toc208694415)

[4.1.1. Kết quả đạt được 49](#_Toc208694416)

[4.1.2. Một vài hạn chế 49](#_Toc208694417)

[4.2. Kết luận 49](#_Toc208694418)

[4.3. Hướng phát triển 50](#_Toc208694419)

[TÀI LIỆU THAM KHẢO 51](#_Toc208694420)

[PHỤ LỤC 52](#_Toc208694421)

DANH MỤC HÌNH VẼ

[Hinh 1. Ảnh về Machine Learning 13](file:///C:\Users\ADMIN\Downloads\BaoCaoDoAn.docx#_Toc208694422)

[Hinh 2. Ảnh về Random Forest 14](#_Toc208694423)

[Hinh 3. Ảnh về Python-Flask 15](#_Toc208694424)

[Hinh 4. Ảnh về HTML, CSS, Bootstrap 16](#_Toc208694425)

[Hinh 5. Ảnh về Github 17](file:///C:\Users\ADMIN\Downloads\BaoCaoDoAn.docx#_Toc208694426)

[Hinh 6. Ảnh về Pandas 18](#_Toc208694427)

[Hinh 7. Ảnh về Seaborn với Matplotlib 19](file:///C:\Users\ADMIN\Downloads\BaoCaoDoAn.docx#_Toc208694428)

[Hinh 8. Ảnh về Power BI 20](file:///C:\Users\ADMIN\Downloads\BaoCaoDoAn.docx#_Toc208694429)

[Hinh 9. Ảnh về bộ dữ liệu 21](#_Toc208694430)

[Hinh 10. Xem kiểu dữ liệu của các cột 22](#_Toc208694431)

[Hinh 11. Kết quả kiểu dữ liệu của các cột 23](#_Toc208694432)

[Hinh 12. Outlier cột Square 24](#_Toc208694433)

[Hinh 13. Xử lý Outlier cột Square 25](#_Toc208694434)

[Hinh 14. Outlier cột Price\_per\_square 25](#_Toc208694435)

[Hinh 15. Xử lý Outlier cột price\_per\_square 26](#_Toc208694436)

[Hinh 16. Outlier cột WC 26](#_Toc208694437)

[Hinh 17. Xử lý Outlier cột WC 27](#_Toc208694438)

[Hinh 18. Outlier cột Bedroom 27](#_Toc208694439)

[Hinh 19. Xử lý Outlier cột Bedroom 28](#_Toc208694440)

[Hinh 20. Biểu đồ phân bố giá nhà 30](#_Toc208694441)

[Hình 21. Biểu đồ phân bố giá nhà theo TB quận. 31](#_Toc208694442)

[Hinh 22. Mối quuan hệ giữa diện tích và giá nhà 31](#_Toc208694443)

[Hinh 23. Mối quan hệ giữa giá nhà với số phòng ngủ 32](#_Toc208694444)

[Hinh 24. Đặc trưng quan trọng cho phân khúc dưới 20 tỷ 35](#_Toc208694445)

[Hinh 25. Đặc trưng quan trong cho phân khúc trên 20 tỷ. 36](#_Toc208694446)

[Hinh 26. Độ chính xác cho phân khúc dưới 20 tỷ 36](#_Toc208694447)

[Hinh 27. Độ chính xác cho phân khúc trên 20 tỷ 37](#_Toc208694448)

[Hinh 28. Sai số dự đoán dưới 20 tỷ 38](#_Toc208694449)

[Hinh 29. Sai số dự đoán trên 20 tỷ 38](#_Toc208694450)

[Hinh 30. Scatter đánh giá mô hình dưới 20 tỷ 41](#_Toc208694451)

[Hinh 31. Scatter đánh giá mô hình trên 20 tỷ 42](#_Toc208694452)

[Hinh 32. Histplot đánh giá mô hình dưới 20 tỷ 42](file:///C:\Users\ADMIN\Downloads\BaoCaoDoAn.docx#_Toc208694453)

[Hinh 33. Hisplot đánh giá mô hình trên 20 tỷ 43](#_Toc208694454)

[Hinh 34. Độ chính xác MAE, RMSE 44](#_Toc208694455)

[Hinh 35. Độ chính xác R2 45](#_Toc208694456)

[Hinh 36. Tổng quan (phân bố bedroom, wc, district, loai nha) 46](#_Toc208694457)

[Hinh 37. Quan hệ Giá – Diện Tích – Quận 47](file:///C:\Users\ADMIN\Downloads\BaoCaoDoAn.docx#_Toc208694458)

MỞ ĐẦU

# GIỚI THIỆU ĐỀ TÀI

## Giới thiệu đề tài

Trong những năm gần đây, nhu cầu mua bán và đầu tư bất động sản tại các thành phố lớn như TP. Hồ Chí Minh, Hà Nội, Đà Nẵng,... ngày càng tăng cao. Tuy nhiên, việc định giá bất động sản vẫn còn nhiều bất cập như thiếu thông tin minh bạch, giá cả biến động liên tục và khó so sánh giữa các khu vực. Điều này gây khó khăn cho người mua, nhà đầu tư cũng như các nhà môi giới khi muốn xác định mức giá hợp lý của một bất động sản cụ thể.

Với sự phát triển mạnh mẽ của công nghệ, đặc biệt là lĩnh vực học máy (Machine Learning), việc xây dựng một hệ thống có khả năng **dự đoán giá nhà tự động** dựa trên các yếu tố như diện tích, vị trí, loại nhà, số phòng, pháp lý,... đã trở nên khả thi.

Đề tài này hướng đến việc phát triển một hệ thống hỗ trợ người dùng **dự đoán giá bất động sản**, phân loại loại nhà, đồng thời dự đoán xu hướng giá nhà trong tương lai gần (theo tháng). Người dùng có thể nhập các thông tin cụ thể để nhận được kết quả ước lượng giá từ mô hình đã huấn luyện. Ngoài ra, hệ thống còn hỗ trợ phân tích loại hình bất động sản và dự đoán xu hướng tăng/giảm giá theo thời gian nhằm nâng cao tính tiện ích và hỗ trợ ra quyết định.

## Lý do chọn đề tài

Hiện nay, giá nhà là một yếu tố quan trọng và ảnh hưởng lớn đến quyết định mua bán bất động sản. Tuy nhiên, việc định giá nhà lại thường mang tính chủ quan, phụ thuộc vào kinh nghiệm cá nhân hoặc thông tin rời rạc từ nhiều nguồn. Điều này dễ dẫn đến tình trạng định giá sai lệch, gây thiệt hại cho cả người mua lẫn người bán.

Do đó, việc áp dụng **các mô hình học máy** để hỗ trợ định giá nhà một cách **khách quan, tự động và có thể cập nhật theo thời gian thực** là một hướng tiếp cận đầy tiềm năng.

Bên cạnh đó, việc xây dựng hệ thống dự đoán xu hướng giá nhà trong tương lai cũng rất quan trọng đối với nhà đầu tư và người dân có nhu cầu mua nhà để ở. Hệ thống sử dụng mô hình dự đoán thời gian (ARIMA) và được triển khai trên nền tảng web giúp người dùng dễ tiếp cận, không cần cài đặt phức tạp, dễ dàng sử dụng ở mọi nơi, mọi lúc.

## Bố cục báo cáo

Báo cáo đồ án được chia thành 5 chương chính như sau:

* **Chương 1 – Giới thiệu đề tài:** Trình bày bối cảnh, lý do chọn đề tài và bố cục báo cáo.
* **Chương 2 – Cơ sở lý thuyết:** Trình bày các kiến thức nền về bất động sản, mô hình học máy, thuật toán Random Forest, Logistic Regression, Linear Regression.
* **Chương 3 – Phát triển hệ thống:** Phân tích yêu cầu, thiết kế hệ thống và quy trình xử lý dữ liệu, huấn luyện mô hình.
* **Chương 4 – Kết luận và hướng phát triển:** Đánh giá kết quả đạt được, những hạn chế còn tồn tại và đề xuất hướng cải tiến trong tương lai.

# CƠ SỞ LÝ THUYẾT

Nội dung chương này trình bày các kiến thức nền tảng liên quan đến đề tài, bao gồm tổng quan về học máy, các thuật toán được sử dụng (Random Forest, Logistic Regression), cũng như công nghệ triển khai hệ thống trên nền web.

## Giới thiệu học máy

### Định nghĩa

Học máy là một nhánh của trí tuệ nhân tạo (AI) cho phép máy tính học hỏi từ dữ liệu mà không cần được lập trình rõ ràng. Mục tiêu là xây dựng các mô hình có thể tự động học các mối quan hệ trong dữ liệu và đưa ra dự đoán hoặc phân loại chính xác.

### ****Phân loại mô hình học máy****

* **Học có giám sát (Supervised Learning)**: Sử dụng dữ liệu có nhãn (label), mô hình học cách ánh xạ đầu vào → đầu ra. Áp dụng cho dự đoán giá nhà, phân loại loại nhà, v.v.
* **Học không giám sát (Unsupervised Learning)**: Tìm hiểu cấu trúc ẩn trong dữ liệu không có nhãn (ví dụ: phân cụm khu vực).
* A blue banner with a robot and objects

  Description automatically generated with medium confidence**Học tăng cường (Reinforcement Learning)**: Máy học thông qua việc thử - sai, nhận phần thưởng hoặc hình phạt.

Hinh . Ảnh về Machine Learning

## Thuật toán Random Forest

### Định nghĩa

Random Forest là một thuật toán học có giám sát thuộc nhóm mô hình ensemble (tổ hợp), hoạt động bằng cách tạo ra nhiều cây quyết định (Decision Tree) và tổng hợp kết quả dự đoán của chúng.

### Ưu điểm

* Độ chính xác cao, đặc biệt với dữ liệu phức tạp
* Giảm hiện tượng overfitting
* Xử lý tốt dữ liệu không tuyến tính

### Nhược điểm

* Khó giải thích rõ ang từng quyết định.
* Tốc độ chậm hơn nếu số cây quá lớn.

A diagram of a tree

Description automatically generated

Hinh . Ảnh về Random Forest

## 2.4. ****Flask – Framework xây dựng hệ thống web****

### 2.4.1. Định nghĩa

Flask là một micro-framework Python nhẹ, được dùng để xây dựng các ứng dụng web đơn giản nhưng linh hoạt.

### ****2.4.2. Lý do chọn Flask****

* Dễ tích hợp với mô hình học máy
* Cấu trúc đơn giản, phù hợp với dự án nhỏ và vừa
* Tùy biến cao, dễ mở rộng

### ****2.4.3. Kiến trúc cơ bản****

* **Route (View)**: Xử lý request và trả về dữ liệu
* **Template (HTML)**: Hiển thị giao diện cho người dùng
* **Model (ML Models)**: Tích hợp các mô hình Machine Learning vào backend Flask để xử lý dữ liệu

A logo for a company

Description automatically generated

Hinh . Ảnh về Python-Flask

## 2.5. HTML-CSS-BOOTSTRAP

### 2.5.1. Định nghĩa

**HTML (HyperText Markup Language):**

Ngôn ngữ đánh dấu dùng để xây dựng cấu trúc cơ bản của trang web.

**CSS (Cascading Style Sheets):**

Dùng để tạo phong cách hiển thị như màu sắc, bố cục, font chữ,... cho giao diện HTML.

**Bootstrap:**

Là framework CSS phổ biến, giúp tạo giao diện hiện đại, responsive (tự động điều chỉnh theo kích thước màn hình).

### 2.5.2. Ưu điểm

* Tiết kiệm thời gian thiết kế
* Giao diện đẹp, chuyên nghiệp
* Hỗ trợ sẵn nhiều component như nút, form, bảng,…

A blue background with white text and blue symbols

Description automatically generated

Hinh . Ảnh về HTML, CSS, Bootstrap

## 2.6. ****Github – Quản lý mã nguồn****

### 2.6.1. Định nghĩa

* GitHub là nền tảng lưu trữ và quản lý mã nguồn bằng Git.
* Trong đề tài: Sử dụng để lưu trữ mã nguồn, theo dõi lịch sử thay đổi, chia sẻ với người khác.

### 2.6.2. Ưu điểm

* Quản lý lịch sử phiên bản
* Làm việc nhóm dễ dàng
* Miễn phí cho dự án cá nhân và sinh viên

A white cat in a circle with a purple background

Description automatically generated

*Hinh 5. Ảnh về Github*

## 2.7. Pandas

### 2.7.1. Định nghĩa

Pandas là một thư viện mã nguồn mở trong Python, hỗ trợ thao tác, xử lý và phân tích dữ liệu dạng bảng. Thư viện này cho phép làm việc hiệu quả với dữ liệu dạng bảng (DataFrame) như trong Excel hoặc cơ sở dữ liệu.

**Ứng dụng trong đề tài**:

* Đọc và xử lý dữ liệu bất động sản từ file CSV/Excel.
* Làm sạch dữ liệu, xử lý thiếu dữ liệu hoặc giá trị bất thường.
* Phân tích thống kê cơ bản và chuẩn bị dữ liệu để đưa vào mô hình học máy.

A logo of a panda and python

Description automatically generated

Hinh . Ảnh về Pandas

## 2.8. Matplotlib & Seaborn

### 2.8.1. Định nghĩa

**- Matplotlib:** Là thư viện vẽ đồ thị 2D trong Python, cung cấp các công cụ để tạo biểu đồ như biểu đồ cột, đường, tròn,…

**Ứng dụng trong đề tài**:

* Vẽ biểu đồ xu hướng giá bất động sản theo thời gian.
* Trực quan hóa phân phối dữ liệu như giá nhà, diện tích, số phòng.

- **Seaborn:** Là một thư viện trực quan hóa dữ liệu dựa trên Matplotlib, cung cấp giao diện đơn giản và đẹp hơn để thể hiện dữ liệu thống kê.

**Ứng dụng trong đề tài**:

* Vẽ biểu đồ mối quan hệ giữa các biến (diện tích, loại nhà, giá...).
* Vẽ heatmap thể hiện ma trận tương quan giữa các đặc trưng dữ liệu.

A group of logos with text

Description automatically generated

Hinh . Ảnh về Seaborn với Matplotlib

## 2.9. PowerBI

### 2.9.1. Định nghĩa

**Power BI:** Là công cụ phân tích và trực quan hóa dữ liệu mạnh mẽ do Microsoft phát triển. Công cụ này hỗ trợ kết nối dữ liệu từ nhiều nguồn và tạo dashboard tương tác.

**Ứng dụng trong đề tài**:

* Phân tích xu hướng giá nhà theo khu vực, loại nhà, thời gian.

Tạo báo cáo và dashboard trực quan phục vụ mục đích trình bày kết quả nghiên cứu.

A group of devices with graphs on screens

Description automatically generated

Hinh 9. Ảnh về PowerBI

Hinh . Ảnh về Power BI

# PHÂN TÍCH DỮ LIỆU VÀ HUẤN LUYỆN MÔ HÌNH.

## Tổng quan bộ dữ liệu

### Giới thiệu bài toán và mục tiêu dự án

Giá nhà đất là một vấn đề được quan tâm rộng rãi, ảnh hưởng trực tiếp đến nhu cầu an cư và đầu tư của hàng triệu người. Việc phân tích và dự đoán giá nhà có thể giúp người mua, người bán và nhà đầu tư đưa ra quyết định chính xác và hiệu quả hơn. Dự án này tập trung vào nội dung chính trong việc dự đoán giá nhà bằng mô hình Random Forest và Linear Regression.

### Lý do chọn mô hình

**Linear Regression**

* Là mô hình cơ bản, dễ hiểu, thường được dùng làm chuẩn so sánh (baseline).
* Giúp kiểm tra mối quan hệ tuyến tính giữa các yếu tố như diện tích, số phòng, vị trí… và giá nhà.
* Dễ diễn giải, phù hợp để rút ra insight từ dữ liệu.

**Random Forest**

* Có khả năng mô hình hóa quan hệ phi tuyến phức tạp, phù hợp vì giá nhà thường bị ảnh hưởng bởi nhiều yếu tố tương tác (vị trí, pháp lý, tiện ích, hạ tầng…).
* Ổn định, ít bị overfitting nhờ cơ chế lấy mẫu ngẫu nhiên và trung bình kết quả nhiều cây quyết định.
* Hoạt động tốt với dữ liệu có nhiều đặc trưng, kể cả dữ liệu nhiễu hoặc không tuyến tính.
* Cho phép đo lường tầm quan trọng của các biến, từ đó hiểu yếu tố nào ảnh hưởng mạnh nhất đến giá nhà.

## Khai phá và phân tích bộ dữ liệu

### Nguồn gốc bộ dữ liệu

- Bộ dữ liệu được cào từ trang <https://batdongsan.com.vn/nha-dat-ban-tp-hcm> và được lưu trữ dưới dạng file excel

-Bộ dự liệu bao gồm: 9852 dòng và 16 cột.

A white and black text on a white background

Description automatically generated

*Hinh 9. Ảnh về bộ dữ liệu*

* + Trong đó ý nghĩa của các cột là:
    - **title** (object): Tiêu đề tin đăng.
    - **price** (object): Giá của bất động sản (cần chuyển sang số để phân tích).
    - **square** (object): Diện tích (m²).
    - **price\_per\_square** (object): Giá trên mỗi mét vuông.
    - **bedroom** (float64): Số phòng ngủ (có missing values).
    - **wc** (float64): Số phòng vệ sinh (có missing values).
    - **location** (object): Địa chỉ/tên khu vực.
    - **Lat** (float64): Vĩ độ (latitude).
    - **Lon** (float64): Kinh độ (longitude).
    - **Loại nhà** (object): Loại hình bất động sản (nhà phố, chung cư, đất nền, …).
    - **Nội thất** (object): Tình trạng nội thất (có sẵn, đầy đủ, chưa có, …).
    - **Hướng nhà** (object): Hướng bất động sản (Đông, Tây, Nam, Bắc, …).
    - **Pháp lý** (object): Tình trạng pháp lý (sổ đỏ, sổ hồng, giấy tay, …).
    - **Mặt tiền** (object): Đặc điểm mặt tiền (chiều rộng mặt tiền, đường lớn/nhỏ, …).
    - **Số tầng** (object): Số tầng của bất động sản.
    - **Đường vào** (object): Độ rộng đường vào

### 3.2.2. Các bước tiền xử lý dữ liệu

#### 3.2.2.1. Chuyển các kiểu dữ liệu về dạng số

A white background with black dots

Description automatically generated

*Hinh 10. Xem kiểu dữ liệu của các cột*

Dựa vào hình trên ta có thể thấy các cột ta cần chuyển về dạng số bao gồm: price, square, price\_per\_square.

Sau khi chuyển các dữ liệu về dạng số ta có:

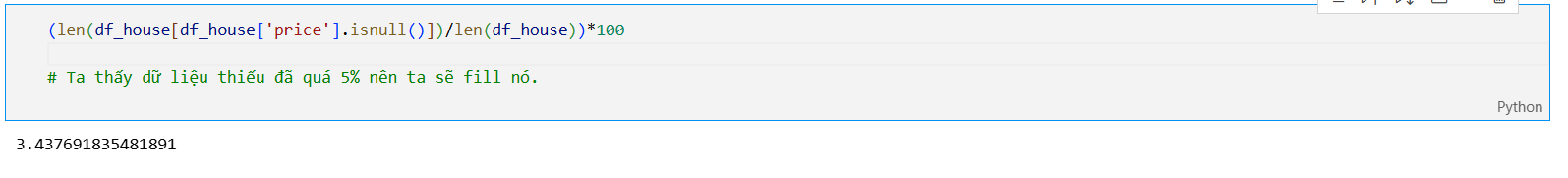
A white background with black dots

Description automatically generated

*Hinh 11. Kết quả kiểu dữ liệu của các cột*

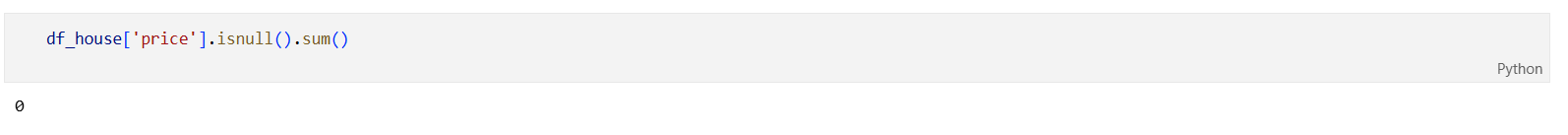
#### 3.2.2.2. Loại bỏ các giá trị NAN ở các cột.

##### 3.2.2.2.1. Price

****

Dựa vào hình trên ta thấy các giá trị bị thiếu ở cột Price chiếm 3.4% <5% nên ta sẽ loại bỏ các hàng có giá trị null ở cột price.

Kết quả sau khi xử lý:

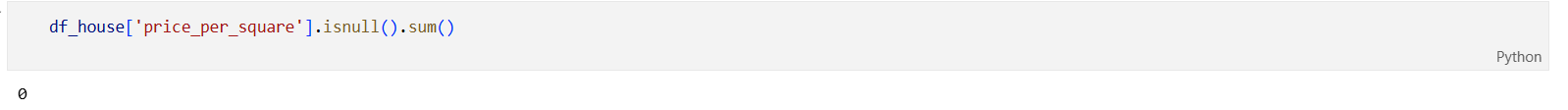


##### 3.2.2.2.2. price\_per\_square

Ta sẽ fill các giá trị Null ở trường price\_per\_square bằng cách lấy các giá trị Price/Square

****

Kết quả sau khi xử lý:

****

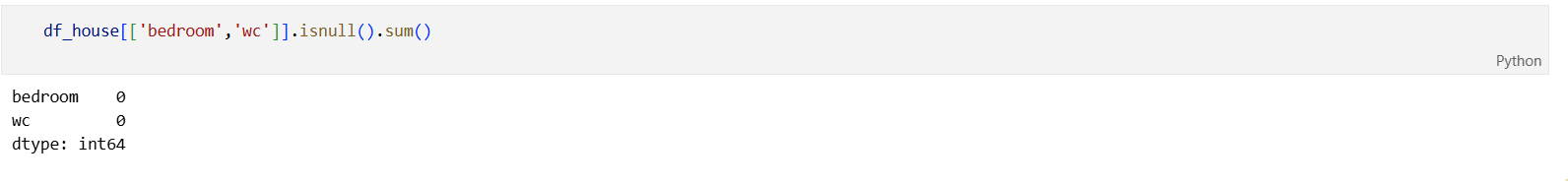
##### 3.2.2.2.3. wc, bedroom

Tương tự vì giá trị Wc, bedroom chiếm số lượng NaN khá lớn nên ta sẽ Fill bằng giá trị trung bình theo quận

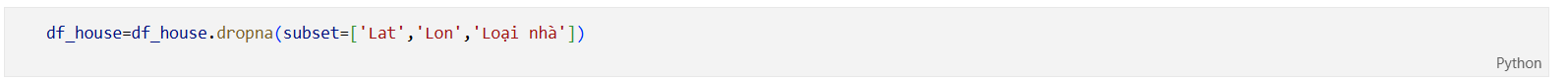
**A computer code on a white background

Description automatically generated**

Kết quả sau khi xử lý:

****3.2.2.2.4. Lat, Don, Loại nhà.

Vì các giá trị ở các cột này không thể fill theo median() nên ta sẽ loại bỏ các dòng chứa giá trị null.

****

#### 3.2.2.3. Giải quyết các giá trị bị Outlier

##### 3.2.2.3.1 Square

A white graph with black dots

Description automatically generated

*Hinh 12. Outlier cột Square*

Dựa vào hàm tính toán IQR và biểu đồ boxplot ta xác định được các giá trị nằm ngoài vùng 262.18 sẽ là các giá trị bị outlier.

* + Do đó ta sẽ tạo một DataFrame mới chỉ lấy các cột price <262.18.

Kết quả DataFrame mới ta thu được:

A blue rectangular object with white lines

Description automatically generated with medium confidence

*Hinh 13. Xử lý Outlier cột Square*

##### 3.2.2.3.2. price\_per\_square

A white rectangular object with black border

Description automatically generated

*Hinh 14. Outlier cột Price\_per\_square*

Dựa vào hàm tính toán IQR và biểu đồ boxplot ta xác định được các giá trị nằm ngoài vùng 325 sẽ là các giá trị bị outlier.

* + Do đó ta sẽ tạo một DataFrame mới chỉ lấy các cột price <325.

Kết quả DataFrame mới ta thu được:

A blue rectangular object with white text

Description automatically generated

*Hinh 15. Xử lý Outlier cột price\_per\_square*

##### 3.2.2.3.3. Wc.

**A graph with many dots

Description automatically generated with medium confidence**

*Hinh 16. Outlier cột WC*

Dựa vào hàm tính toán IQR và biểu đồ boxplot ta xác định được các giá trị nằm ngoài vùng 7 sẽ là các giá trị bị outlier.

* + Do đó ta sẽ tạo một DataFrame mới chỉ lấy các cột wc <7.

Kết quả DataFrame mới ta thu được:

A blue square with numbers and a white background

Description automatically generated

*Hinh 17. Xử lý Outlier cột WC*

##### 3.2.2.3.4 Bedroom

**A graph with dots and numbers

Description automatically generated**

*Hinh 18. Outlier cột Bedroom*

Dựa vào hàm tính toán IQR và biểu đồ boxplot ta xác định được các giá trị nằm ngoài vùng 5.5 sẽ là các giá trị bị outlier.

* + Do đó ta sẽ tạo một DataFrame mới chỉ lấy các cột bedroom <5.5.

Kết quả DataFrame mới ta thu được:

**A blue square with a black line

Description automatically generated with medium confidence**

*Hinh 19. Xử lý Outlier cột Bedroom*

#### 3.2.2.4. Tách cột Location và Loại Nhà.

Dữ liệu sau khi tách:

**A screenshot of a computer

Description automatically generated**

#### 3.2.2.5. Thiết lập lại chỉ mục để tránh xáo trộn dữ liệu

****

Sau khi thiết lập chỉ mục:

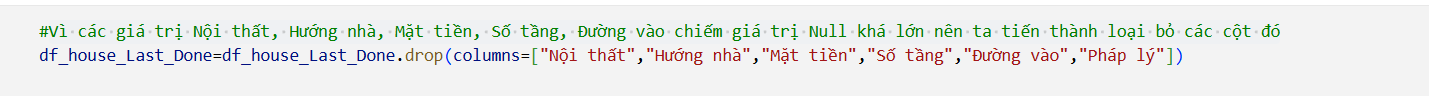
**A close-up of a computer screen

Description automatically generated**

##### 3.2.2.6. Loại bỏ các cột không cần thiết và thêm cột mới

###### 3.2.2.6.1. Loại bỏ các cột không cần thiết

Vì các giá trị Nội thất, Hướng nhà, Mặt tiền, Số tầng, Đường vào chiếm giá trị Null khá lớn nên ta tiến thành loại bỏ các cột đó

****

Dữ liệu sau khi loại bỏ:

**A screenshot of a computer

Description automatically generated**

###### 3.2.2.6.2. Thêm cột mới

Vì giá nhà thường hay biến động theo khoảng cách tới các quận/trung tâm quan trọng của TP.HCM nên ta tạo thêm các cột khoảng cách đến đó, ở đây ta xét các vị trí như Bưu Điện Tp.HCM(Quận 1), UBND(Quận 3), Trung tâm Bình Thạnh, Phú Mỹ Hưng(Quận 7), Sân bay Tân Sơn Nhất

Dữ liệu sau khi thêm:

**A screenshot of a computer

Description automatically generated**

## Trực quan hóa dữ liệu khi được làm sạch

### 3.3.1. Biểu đồ phân bố giá nhà

**A graph with a line graph

Description automatically generated with medium confidence**

*Hinh 20. Biểu đồ phân bố giá nhà*

**Nhận xét:**

* + Khoảng **0–1** trên trục thực chất là **1–10 tỷ VND**.
  + Phần lớn giá nhà nằm trong khoảng này, đặc biệt tập trung mạnh ở mức **3–8 tỷ VND**.
  + Từ **10–20 tỷ VND** (tương ứng 1–2 trên trục), số lượng vẫn còn đáng kể nhưng đã giảm.
  + Từ **30 tỷ VND trở lên**, số lượng rất ít, chỉ là các giá trị ngoại lai (outliers).

### 3.3.2. Biểu đồ phân bố giá nhà theo trung bình quận.

A graph of a graph showing the amount of the amount of money

Description automatically generated with medium confidence

*Hình 21. Biểu đồ phân bố giá nhà theo TB quận.*

**Nhận xét:**

Biểu đồ trên cho thấy sự chênh lệch rõ rệt về **giá nhà trung bình theo quận**:

* **Cao nhất** là Quận 1, với giá trung bình gần **20 tỷ VND**, tiếp theo là Quận 5, Quận 2 và Quận 10 (dao động quanh 16–18 tỷ VND).
* Các quận trung tâm khác như Phú Nhuận, Tân Bình, Bình Thạnh cũng duy trì mức giá cao (trên 15 tỷ VND).
* **Ngoại thành** như Nhà Bè, Bình Tân, Bình Chánh, Quận 12, Cần Giờ, Hóc Môn có mức giá trung bình thấp hơn (khoảng 7–10 tỷ VND).
* **Thấp nhất** là Củ Chi, chỉ khoảng 2 tỷ VND.

### 3.3.3. Mối quan hệ giữa diện tích với giá nhà

A graph showing a number of red dots

Description automatically generated

*Hinh 22. Mối quuan hệ giữa diện tích và giá nhà*

**Nhận xét:**

Biểu đồ scatter thể hiện mối quan hệ giữa **diện tích (m²)** và **giá nhà (VND)**:

* Có xu hướng **tương quan dương**: diện tích càng lớn thì giá nhà nhìn chung càng cao.
* Tuy nhiên, mối quan hệ không tuyến tính rõ rệt; cùng một diện tích, giá có thể dao động mạnh → điều này cho thấy còn phụ thuộc vào nhiều yếu tố khác (vị trí, tiện ích, pháp lý, …).
* Phần lớn nhà tập trung ở khoảng **30–100 m²** với giá dưới **20 tỷ VND**.
* Một số điểm ngoại lai thể hiện các căn nhà có giá rất cao so với diện tích, có thể là bất động sản đặc biệt (biệt thự, nhà phố trung tâm).

### 3.3.4. Mối quan hệ giữa giá nhà với số phòng ngủ

**A graph with green squares

Description automatically generated**

*Hinh 23. Mối quan hệ giữa giá nhà với số phòng ngủ*

**Nhận xét:**

Biểu đồ boxplot thể hiện **giá nhà theo số phòng ngủ**:

* **Xu hướng chung**: số phòng ngủ càng nhiều thì giá nhà trung bình có xu hướng cao hơn.
* Nhà **1–2 phòng ngủ** có giá trung bình thấp nhất (chủ yếu dưới **10 tỷ VND**), nhưng vẫn xuất hiện nhiều giá trị ngoại lai cao.
* Nhà **3–5 phòng ngủ** có giá trung bình cao hơn rõ rệt, khoảng **10–20 tỷ VND**, và cũng có nhiều ngoại lệ trên **30–40 tỷ VND**.
* Độ phân tán giá tăng dần theo số phòng ngủ → nhà nhiều phòng ngủ có sự chênh lệch giá rất lớn, phản ánh sự đa dạng về phân khúc (nhà phố, biệt thự, căn hộ cao cấp).

## Huấn luyện mô hình dự đoán giá nhà

### 3.4.1. Huấn luyện mô hình Random Forest

#### 3.4.1.1. Dùng LableEncoder cho cột distinct, loainha.

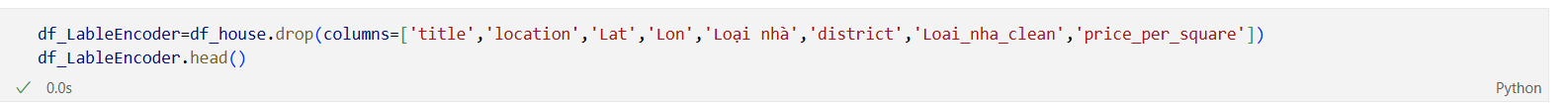
**- Label Encoder** sẽ chuyển từng giá trị chuỗi thành một con số (ví dụ: Nhà phố → 0, Chung cư → 1, Biệt thự → 2; hoặc Quận 1 → 0, Quận 2 → 1, …).

Dữ liệu sao khi Label Encoder:

A screenshot of a computer

Description automatically generated

#### 3.4.1.2. Loại bỏ các cột không cần thiết cho việc train model



Dữ liệu sau khi loại bỏ các cột:

A close-up of a number

Description automatically generated

* + Các dữ liệu đều đã được đưa về dạng số nên ta sẽ tiến hành huấn luyện mô hình.

#### 3.4.1.3 Tiến hành huấn luyện.

##### 3.4.1.3.1. Tạo biến đầu vào X và đầu ra y

Các giá trị cột X:

A screenshot of a computer

Description automatically generated

Giá trị cột y:

A white background with black dots

Description automatically generated

##### 3.4.1.3.2. Chia phân khúc giá nhà

Ở đây ta chia thành 2 model để dự đoán cho 2 phân khúc giá nhà:

* Phân khúc giá nhà dưới 20 tỷ.
* Phân khúc giá nhà trên 20 tỷ.

A close-up of a computer screen

Description automatically generated

#### 3.4.1.4. Kiểm tra độ chính xác của mô hình

A white background with black text

Description automatically generated

📊 Phân khúc LOW (<20 tỷ)

* **MAE** ≈ **87.5 triệu** → trung bình dự đoán lệch khoảng ~90 triệu.
* **RMSE** ≈ **127.6 triệu** → khi sai số nặng (outlier) có thể lệch trên 120 triệu.
* **R² = 0.93** → mô hình giải thích 93% biến động giá, khá tốt.

📌 **Kết luận:** Với các căn <20 tỷ, mô hình hoạt động **ổn định và sát giá**.

📊 Phân khúc HIGH (≥20 tỷ)

* **MAE** ≈ **186.8 triệu** → trung bình dự đoán lệch gần 190 triệu.
* **RMSE** ≈ **275.6 triệu** → khi sai số lớn có thể lệch gần 280 triệu.
* **R² = 0.75** → mô hình giải thích 75% biến động giá, thấp hơn nhiều so với phân khúc <20 tỷ.

📌 **Kết luận:** Với các căn ≥20 tỷ, mô hình **kém chính xác hơn**, sai số cao hơn do biến động giá lớn và dữ liệu hạn chế.

#### 3.4.1.5. Minh họa một số đặc trưng quan trọng, độ chính xác của mô hình, sai số dự đoán.

##### 3.4.4.1.5.1. Một số đặc trưng quan trọng

- Phân khúc dưới 20 tỷ:

A graph with a bar graph

Description automatically generated

*Hinh 24. Đặc trưng quan trọng cho phân khúc dưới 20 tỷ*

**Nhận xét:**

* + **Diện tích (square)** là yếu tố quan trọng vượt trội, ảnh hưởng lớn nhất đến giá.
  + **Vị trí (các quận như Bình Thạnh, Quận 3, Phú Nhuận, Quận 1, Quận 7)** cũng có vai trò quan trọng, phản ánh thực tế “vị trí quyết định giá trị”.
  + **Loại nhà, số phòng, tiện ích** có ảnh hưởng nhưng ở mức độ thấp hơn.  
    ➡️ Nhìn chung, phân khúc này nhấn mạnh **“diện tích và vị trí”** là hai yếu tố then chốt.

-Phân khúc trên 20 tỷ

A graph with red and white bars

Description automatically generated

*Hinh 25. Đặc trưng quan trong cho phân khúc trên 20 tỷ.*

**Nhận xét:**

* + **Diện tích (square)** vẫn giữ vai trò quan trọng nhất, nhưng khoảng cách so với các yếu tố khác không quá vượt trội như phân khúc dưới 20 tỷ.
  + **Vị trí (Quận 3, Quận 1, Quận 7, Bình Thạnh, Phú Nhuận)** nổi bật hơn, thể hiện rằng ở phân khúc cao cấp, vị trí trung tâm, đắc địa càng quyết định giá trị.
  + **Loại nhà** và **gần sân bay** có mức ảnh hưởng nhất định, trong khi số phòng và WC ít quan trọng.  
    ➡️ Với phân khúc này, **“diện tích + vị trí trung tâm”** là hai yếu tố chính, trong đó **vị trí có sức nặng hơn so với phân khúc < 20 tỷ**.

##### 3.4.4.1.5.2. Độ chính xác của mô hình.

- Phân khúc dưới 20 tỷ:

A graph showing a line of red dots

Description automatically generated

*Hinh 26. Độ chính xác cho phân khúc dưới 20 tỷ*

**Nhận xét:**

* + - Các điểm dữ liệu phân bố khá sát quanh đường chéo y=xy = xy=x, cho thấy mô hình dự đoán giá trong phân khúc này **khá chính xác**.
    - Sai số xuất hiện chủ yếu ở vùng giá cao hơn (gần 20 tỷ), khi một số điểm nằm lệch ra khỏi đường chéo, nhưng nhìn chung vẫn trong phạm vi chấp nhận được.
    - Điều này chứng tỏ mô hình **phù hợp và đáng tin cậy** để ước lượng giá bất động sản ở phân khúc thấp đến trung bình (< 20 tỷ).

- Phân khúc trên 20 tỷ:

A graph with red lines

Description automatically generated

*Hinh 27. Độ chính xác cho phân khúc trên 20 tỷ*

**Nhận xét:**

* + - Các điểm dữ liệu vẫn bám theo đường chéo y=xy = xy=x, nhưng mức độ phân tán cao hơn so với phân khúc < 20 tỷ.
    - Có hiện tượng **dự đoán thấp hơn thực tế** đối với nhiều bất động sản giá trị cao (outlier nằm phía dưới đường chéo).
    - Điều này cho thấy mô hình dự đoán ở phân khúc cao cấp **kém chính xác hơn**, khó nắm bắt do dữ liệu ít hơn và giá trị thường biến động mạnh bởi các yếu tố đặc thù (vị trí đắc địa, kiến trúc, thương hiệu dự án…).

##### 3.4.4.1.5.3. Sai số dự đoán

- Phân khúc dưới 20 tỷ:

A graph of a normal distribution with Eiffel Tower in the background

Description automatically generated

*Hinh 28. Sai số dự đoán dưới 20 tỷ*

**Nhận xét:**

* + - Phân phối sai số (Residual = Actual – Predicted) có dạng gần đối xứng quanh giá trị 0, cho thấy mô hình dự đoán **khá cân bằng**, không bị lệch nhiều về một phía.
    - Phần lớn sai số tập trung quanh 0 (tần suất cao nhất), chứng tỏ đa số dự đoán khá sát với giá trị thực tế.
    - Tuy nhiên vẫn tồn tại một số outlier với sai số lớn ở hai phía, đặc biệt là phía dương (giá thực cao hơn giá dự đoán), nghĩa là mô hình có xu hướng **dự đoán thấp hơn giá thực** cho một số bất động sản.

-Phân khúc trên 20 tỷ:

A green and red graph

Description automatically generated

*Hinh 29. Sai số dự đoán trên 20 tỷ*

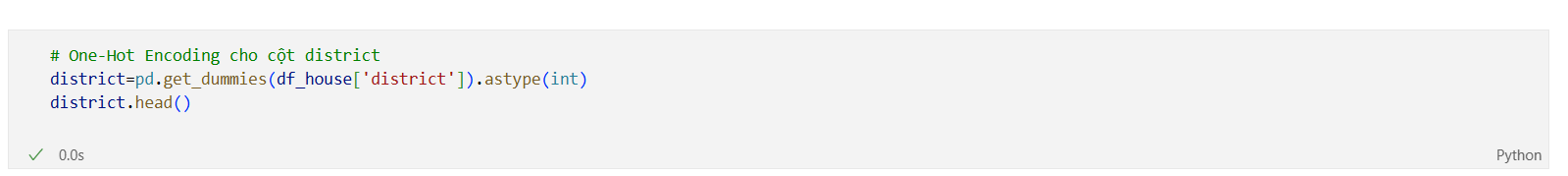
**Nhận xét:**

* + - Sai số vẫn phân bố quanh 0 nhưng **bị lệch nhẹ về phía dương**, nghĩa là mô hình có xu hướng **dự đoán thấp hơn thực tế** đối với nhiều bất động sản cao cấp.
    - Đỉnh phân phối tập trung gần 0, tuy nhiên mức tản mát rộng hơn so với phân khúc < 20 tỷ, thể hiện sai số lớn hơn.
    - Xuất hiện một số outlier với sai số dương khá cao (giá thực cao hơn nhiều so với dự đoán), cho thấy mô hình chưa bắt được các yếu tố đặc thù của thị trường cao cấp (như vị trí đắc địa, kiến trúc độc đáo, thương hiệu…).

### 3.4.2. Huấn luyện mô hình Linear Regression

#### 3.4.2.1. One-Hot Encoding cho cột district và loại nhà

##### 3.4.2.1.1. District



Dữ liệu sau khi one-hot encoding:

A white sheet with black text

Description automatically generated

##### 3.4.2.1.2. Loại nhà

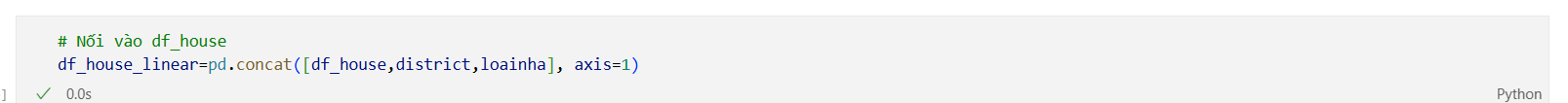


Dữ liệu sau khi one-hot encoding:

A screenshot of a computer

Description automatically generated

#### 3.4.2.2. Nối 2 bảng district và loại nhà vào dataframe



Dữ liệu sau khi nối:

A close-up of a white paper

Description automatically generated

#### 3.4.2.3. Tiến hành huấn luyện mô hình

##### 3.4.2.3.1. Tạo biến đầu vào X và đầu ra y

Giá trị của X:

A table with numbers and letters

Description automatically generated

Giá trị của y:

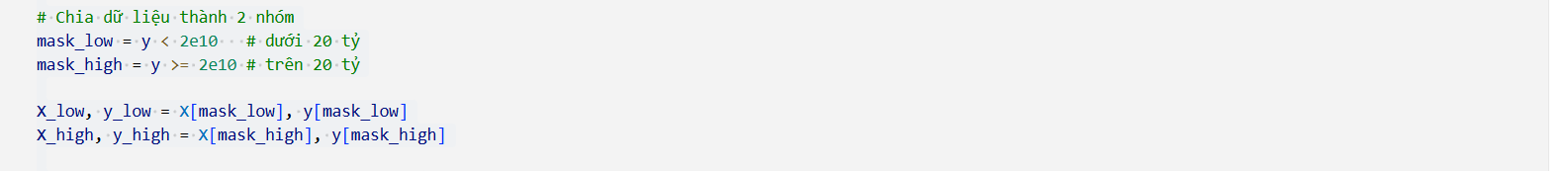
A white background with black and white clouds

Description automatically generated with medium confidence

##### 3.4.2.3.2. Chia phân khúc giá nhà.

Ở đây ta chia thành 2 model để dự đoán cho 2 phân khúc giá nhà:

* Phân khúc giá nhà dưới 20 tỷ.
* Phân khúc giá nhà trên 20 tỷ.



#### 3.4.2.4. Kiểm tra độ chính xác của mô hình

A white background with black and white clouds

Description automatically generated

**Nhận xét:**

* + - **Phân khúc < 20 tỷ**: Mô hình cho kết quả chưa tốt (R² ≈ 0.60, MAPE ≈ 28.6%), sai số còn khá cao → mô hình chỉ giải thích được một phần biến động giá.
    - **Phân khúc ≥ 20 tỷ**: Kết quả cũng chưa khả quan (R² rất thấp ≈ 0.10), nhưng MAPE thấp hơn (≈ 12.6%) → dự đoán có độ chính xác tương đối khi xét theo tỷ lệ %, tuy nhiên khả năng khái quát còn hạn chế.

#### 3.4.2.5. Biểu đồ đánh giá mô hình

A graph showing a number of red dots

Description automatically generated

*Hinh 30. Scatter đánh giá mô hình dưới 20 tỷ*

**Nhận xét:**

* **Nhìn chung mô hình dự đoán bám khá sát với thực tế**: các điểm dữ liệu phân bố quanh đường chéo đỏ y=xy=xy=x.
* Tuy nhiên, khi giá trị thực tế tăng dần, **độ phân tán ngày càng lớn** (heteroscedasticity) → mô hình dự đoán các căn giá cao hơn trong nhóm này thường **bị lệch xuống dưới** (underestimate).
* Có xu hướng **dự đoán thấp hơn thực tế** ở các căn giá trị cao trong nhóm (< 20 tỷ), điều này phù hợp với đặc điểm của hồi quy tuyến tính khi có phân phối lệch.

A graph with red dots

Description automatically generated

*Hinh 31. Scatter đánh giá mô hình trên 20 tỷ*

**Nhận xét:**

* Các điểm dữ liệu cũng nằm gần đường chéo, nhưng **tập trung thành cụm** quanh khoảng 20–30 tỷ.
* Một số điểm lệch khá xa đường chuẩn y=xy=xy=x.
* Với các căn giá trị càng cao, dự đoán càng **thiếu ổn định** và **thường thấp hơn thực tế**.
* A graph with a line graph

  Description automatically generated with medium confidenceĐiều này phù hợp với R² thấp hơn ở phân khúc này mà bạn tính được trước đó (~0.76).

*Hinh 32. Histplot đánh giá mô hình dưới 20 tỷ*

**Nhận xét:**

* Phân phối giá dự đoán bám khá sát giá thực tế trong khoảng 5–15 tỷ.
* Tuy nhiên vẫn có một số điểm dự đoán lệch cao hơn thực tế, kéo dài tới 30–40 tỷ.
* Xu hướng chung là mô hình hoạt động tốt ở vùng giá phổ biến, nhưng kém ổn định với các trường hợp biên.

A graph with a line graph

Description automatically generated with medium confidence

*Hinh 33. Hisplot đánh giá mô hình trên 20 tỷ*

**Nhận xét:**

* Phân phối giá dự đoán (cam) dịch sang trái so với thực tế (xanh), cho thấy mô hình thường **dự đoán thấp hơn** giá thật.
* Các giá trị thực tế trải rộng đến 40–45 tỷ, nhưng dự đoán tập trung nhiều quanh mức 22–30 tỷ → mô hình **chưa bao quát tốt các căn giá trị cao**.
* Sai số rõ rệt hơn so với phân khúc <20 tỷ, phù hợp với việc R² ở phân khúc này thấp (~0.76).

## So sánh độ chính xác của hai mô hình

### MAE, RMSE.

**A graph of different colored squares

Description automatically generated with medium confidence**

*Hinh 34. Độ chính xác MAE, RMSE*

**Nhận xét:**

**- Phân khúc < 20 tỷ:**

* **Linear Regression** có MAE ≈ 2.2 tỷ và RMSE ≈ 3.1 tỷ → sai số khá lớn.
* **Random Forest** thấp hơn nhiều (MAE ≈ 0.9 tỷ, RMSE ≈ 1.3 tỷ) → dự đoán ổn định và bám sát giá thực tế hơn.

**- Phân khúc ≥ 20 tỷ:**

* **Linear Regression** tiếp tục có sai số rất cao (MAE ≈ 3.4 tỷ, RMSE ≈ 4.8 tỷ).
* **Random Forest** tốt hơn (MAE ≈ 1.9 tỷ, RMSE ≈ 2.7 tỷ), nhưng sai số vẫn lớn hơn so với phân khúc <20 tỷ.

### ****R2****

A screenshot of a graph

Description automatically generated

*Hinh 35. Độ chính xác R2*

**Nhận xét:**  
 - **Phân khúc < 20 tỷ:**

* Linear Regression đạt R² ≈ 0.6 → chỉ giải thích được khoảng 60% biến động giá.
* Random Forest đạt R² ≈ 0.9 → giải thích đến ~90% biến động, chứng tỏ mô hình phù hợp hơn nhiều.

- **Phân khúc ≥ 20 tỷ:**

* Linear Regression gần như thất bại (R² ≈ 0.1), tức là mô hình không nắm bắt được mối quan hệ dữ liệu.
* Random Forest đạt R² ≈ 0.75 → vẫn còn sai số nhưng thể hiện rõ khả năng khái quát tốt hơn Linear Regression.

## 3.6. Trực quan hóa dữ liệu với Power BI

### 3.6.1. Tổng quan (phân bố bedroom, wc, district)

A screenshot of a computer

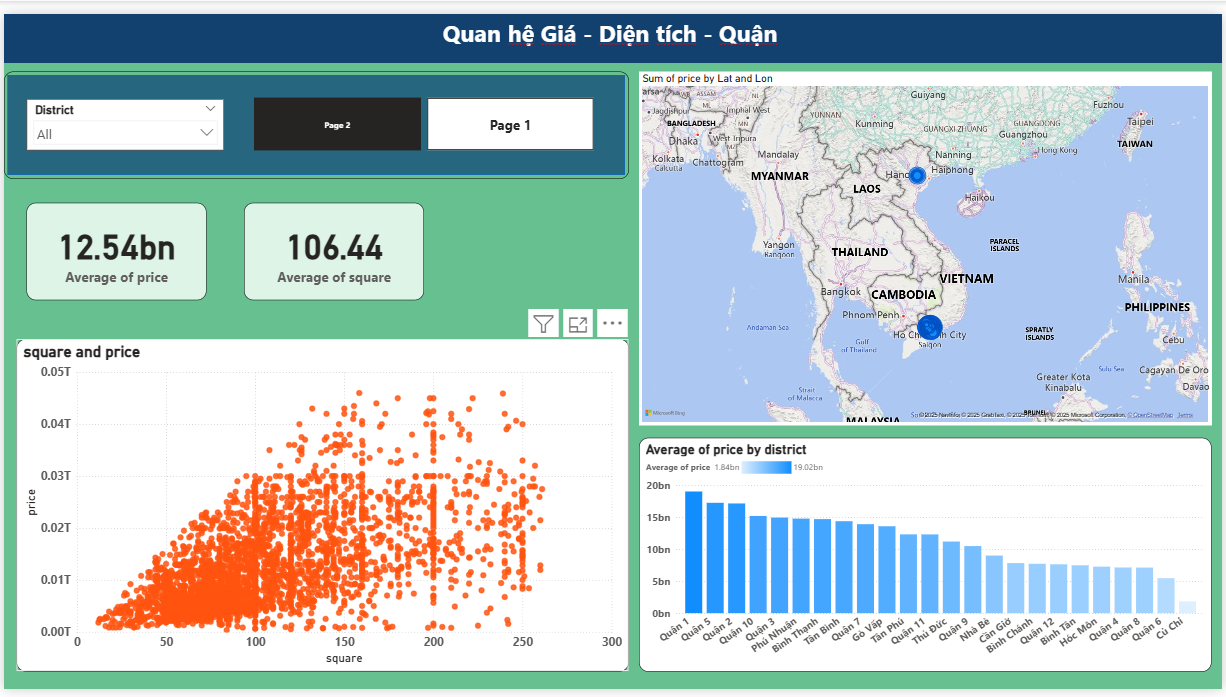
Description automatically generated

*Hinh 36. Tổng quan (phân bố bedroom, wc, district, loai nha)*

**Nhận xét tổng quan phân bố (bedroom, wc, district)**

1. **Số lượng bất động sản**
   * Tổng cộng có **3.798 bất động sản** trong tập dữ liệu,
2. **Phân bố số phòng ngủ (bedroom)**
   * Phổ biến nhất là bất động sản có **3 phòng ngủ (1.388 căn)**, chiếm tỷ trọng lớn nhất.
   * Tiếp theo là loại **4 phòng ngủ (1.096 căn)** và **2 phòng ngủ (782 căn)**.
   * Loại ít gặp nhất là **1 phòng ngủ (193 căn)**.  
     => Xu hướng cho thấy các bất động sản từ 2–4 phòng ngủ chiếm đa số, phù hợp với nhu cầu nhà ở gia đình.
3. **Phân bố số phòng vệ sinh (wc)**
   * **2 WC** chiếm ưu thế với khoảng **1.750 căn**, gần **một nửa tổng số**.
   * Các mức 3–4 WC cũng có số lượng đáng kể (khoảng 600 căn mỗi loại).
   * Trên 5 WC chỉ xuất hiện lẻ tẻ (dưới 200 căn).  
     => Điều này phản ánh sự cân đối giữa diện tích, số phòng và tiện ích.
4. **Phân bố theo quận/huyện (district)**
   * **Quận 9 và Quận 7** là nơi tập trung nhiều bất động sản nhất (trên 600 căn).
   * Các khu vực khác như **Bình Thạnh, Thủ Đức, Nhà Bè** cũng có số lượng khá cao (300–400 căn).
   * Một số quận trung tâm (Quận 1, Quận 3, Quận 5…) và vùng ven (Cần Giờ, Hóc Môn) có rất ít bất động sản (<100 căn).  
     => Xu hướng cho thấy nguồn cung tập trung nhiều ở các quận phát triển mới (Q9, Q7, Thủ Đức) thay vì khu vực trung tâm.

### 3.6.2. Quan hệ Giá – Diện Tích – Quận



*Hinh 37. Quan hệ Giá – Diện Tích – Quận*

**Nhận xét quan hệ Giá – Diện tích – Quận**

1. **Trung bình giá và diện tích**
   * Giá trung bình của bất động sản trong tập dữ liệu là **~12,54 tỷ đồng**.
   * Diện tích trung bình là **~106,44 m²**.  
     => Đây là mức phổ biến cho phân khúc nhà ở trung – cao cấp tại TP.HCM.
2. **Mối quan hệ giữa diện tích và giá (scatter plot)**
   * Biểu đồ phân tán cho thấy giá có xu hướng **tăng theo diện tích**, nhưng không hoàn toàn tuyến tính.
   * Nhiều bất động sản có diện tích lớn nhưng giá không tăng mạnh, phản ánh yếu tố **vị trí/quận** và **chất lượng nhà** cũng ảnh hưởng đến giá.
   * Dữ liệu xuất hiện một số **outlier** (giá cao bất thường dù diện tích không quá lớn).
3. **Phân bố giá trung bình theo quận**
   * Các quận trung tâm như **Quận 1 (~19,02 tỷ), Quận 3, Quận 10** có giá trung bình cao nhất.
   * Các quận vùng ven như **Củ Chi, Hóc Môn, Quận 6** có mức giá trung bình thấp hơn rõ rệt (dưới 10 tỷ).
   * Điều này phản ánh **chênh lệch giá giữa trung tâm và vùng ven**, phù hợp với thực tế thị trường.
4. **Phân bố không gian (bản đồ)**
   * Bất động sản tập trung chủ yếu ở khu vực **TP.HCM**.
   * Xuất hiện thêm một số điểm ở Hà Nội, tuy nhiên số lượng rất ít so với TP.HCM.

# HƯỚNG KẾT LUẬN VÀ PHÁT TRIỂN

## Kết quả đạt được và hạn chế

### Kết quả đạt được

* Xây dựng thành công mô hình **Linear Regression** cho phép mô tả quan hệ tuyến tính giữa diện tích, vị trí quận và giá nhà. Mô hình này cho thấy xu hướng chung: giá tăng theo diện tích và các quận trung tâm có giá cao hơn rõ rệt.
* Phát triển mô hình **Random Forest** giúp nắm bắt được các quan hệ phi tuyến phức tạp hơn, đồng thời khắc phục được hạn chế của Linear Regression trong việc xử lý outlier và sự khác biệt mạnh giữa các quận.
* Đánh giá mô hình bằng các chỉ số MAE, RMSE, R² cho thấy Random Forest đạt độ chính xác cao hơn, đặc biệt ở phân khúc giá trị cao (≥ 20 tỷ).
* So sánh hai mô hình đã cung cấp cái nhìn rõ ràng hơn về ưu – nhược điểm: Linear Regression đơn giản, dễ giải thích; Random Forest mạnh mẽ hơn về độ chính xác và khả năng mô tả dữ liệu phi tuyến.
* Kết quả phân tích giúp nhận diện được các yếu tố quan trọng nhất ảnh hưởng đến giá nhà, bao gồm diện tích và vị trí quận.

### Một vài hạn chế

* **Linear Regression** bị ảnh hưởng nhiều bởi outlier, làm giảm độ chính xác ở những bất động sản giá trị cao.
* **Random Forest** tuy cho kết quả tốt hơn nhưng tiêu tốn tài nguyên tính toán nhiều hơn, thời gian huấn luyện lâu hơn.
* Cả hai mô hình mới chỉ khai thác các đặc trưng cơ bản (diện tích, vị trí, số phòng), chưa tích hợp các yếu tố khác như năm xây dựng, tiện ích xung quanh, tình trạng pháp lý…
* Dữ liệu chưa thật sự đồng đều giữa các quận (một số quận ít mẫu), làm giảm tính tổng quát của mô hình.

## Kết luận

Trong 10 tuần thực hiện, đề tài đã xây dựng và triển khai hai mô hình dự đoán giá nhà là **Linear Regression** và **Random Forest**. Kết quả cho thấy:

* Linear Regression giúp đưa ra cái nhìn khái quát, dễ giải thích.
* Random Forest mang lại dự đoán chính xác hơn, đặc biệt trong bối cảnh thị trường phức tạp và nhiều yếu tố phi tuyến.

Các kết quả này chứng minh khả năng áp dụng mô hình học máy vào việc hỗ trợ định giá bất động sản, góp phần hỗ trợ người mua, người bán và nhà đầu tư trong việc ra quyết định.

## Hướng phát triển

- Mở rộng dữ liệu huấn luyện với nhiều đặc trưng hơn (năm xây dựng, pháp lý, tiện ích xung quanh, mật độ dân cư…).

- Tích hợp thêm các mô hình mạnh hơn như **Gradient Boosting, XGBoost, LightGBM** để so sánh.

- Phát triển hệ thống dự đoán giá trực tuyến: người dùng nhập diện tích, vị trí, số phòng → hệ thống trả về giá dự đoán.

- Tối ưu hóa Random Forest bằng kỹ thuật **hyperparameter tuning** để đạt hiệu suất cao hơn.

- Kết hợp với bản đồ trực quan (Mapbox/Leaflet) để hiển thị giá dự đoán theo từng khu vực.

- Xây dựng dashboard theo dõi hiệu năng mô hình và cập nhật dự đoán theo thời gian thực khi có dữ liệu mới.

TÀI LIỆU THAM KHẢO

[1] Scikit-learn, "Random Forests — scikit-learn 1.5.0 documentation," [Online]. Available: https://scikit-learn.org/stable/modules/ensemble.html#random-forests. [Accessed 15 08 2025].

[2] Scikit-learn, "Logistic Regression — scikit-learn 1.5.0 documentation," [Online]. Available: https://scikit-learn.org/stable/modules/linear\_model.html#logistic-regression. [Accessed 15 08 2025].

[3] Analytics Vidhya, "A Comprehensive Guide to Random Forest," [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/. [Accessed 16 08 2025].

[4] IBM, "What is Logistic Regression?," [Online]. Available: https://www.ibm.com/topics/logistic-regression. [Accessed 16 08 2025].

[5] Kaggle, "House Prices - Advanced Regression Techniques," [Online]. Available: https://www.kaggle.com/c/house-prices-advanced-regression-techniques. [Accessed 17 08 2025].

[6] Towards Data Science, "Predicting House Prices with Machine Learning," [Online]. Available: https://towardsdatascience.com/predicting-house-prices-with-machine-learning. [Accessed 17 08 2025].

PHỤ LỤC