HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Phùng Thị Thu

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI - 2026

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Phùng Thị Thu

ÁP DỤNG MÔ HÌNH LONG SHORT - TERM MEMORY TRONG DỰ ĐOÁN GIÁ NHÀ TẠI VIỆT NAM

Chuyên ngành: Khoa học máy tính

Mã số: 8.48.01.01

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. ĐỖ THANH HÀ

HÀ NỘI - 2026

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi.

Các số liệu, kết quả nêu trong đề án tốt nghiệp là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

TÁC GIẢ



Phùng Thị Thu

LỜI CẢM ƠN

Tôi xin trân trọng bày tỏ lòng biết ơn sâu sắc đến tập thể các thầy, cô giáo Học viện Công nghệ Bưu Chính viễn thông đã tận tình giảng dạy, truyền đạt kiến thức và tạo điều kiện thuận lợi cho tôi trong suốt quá trình học tập và nghiên cứu tại Học viện. Đặc biệt, tôi xin gửi lời cảm ơn chân thành đến cô giáo Tiến sĩ Đỗ Thanh Hà, người đã trực tiếp hướng dẫn tôi hoàn thành đề án tốt nghiệp này. Xin được ghi nhận sự động viên và hỗ trợ từ gia đình, bạn bè và đồng nghiệp trong suốt thời gian thực hiện.

Mặc dù đã cố gắng nhiều song đề án tốt nghiệp khó tránh khỏi những hạn chế nhất định, tôi rất mong nhận được sự góp ý quý báu của quý thầy cô và các nhà nghiên cứu để công trình được hoàn thiện hơn.

Xin trân trọng cảm ơn!

HỌC VIÊN



Phùng Thị Thu

MỤC LỤC

DANH MỤC CHỮ VIẾT TẮT iii
DANH MỤC BẢNG iv
DANH MỤC HÌNH v
MỞ ĐẦU 1
1. Lý do chọn đề tài 1
2. Tổng quan về vấn đề nghiên cứu 2
3. Mục đích nghiên cứu 2
4. Đối tượng và phạm vi nghiên cứu 3
5. Phương pháp nghiên cứu 4
Chương 1: TỔNG QUAN CƠ SỞ LÝ THUYẾT 5
1.1. Bài toán dự đoán giá nhà và chuỗi thời gian 5
1.1.1. Bài toán dự đoán giá nhà 5
1.1.2. Chuỗi thời gian 5
1.2. Một số mô hình học máy truyền thống 6
1.2.1. Các mô hình hồi quy cơ bản 7
1.2.2. Mô hình AutoRegressive Integrated Moving Average (ARIMA) 8
1.2.3. Mô hình rừng ngẫu nhiên (Random Forest - RF) 9
1.3. Các mô hình dự đoán học sâu 9
1.3.1. Mạng Artificial Neural Networks (ANN) 10
1.3.2. Mạng Convolutional Neural Networks (CNN) 11
1.3.3. Mạng Recurrent Neural Networks (RNN) 12
1.3.4. Mô hình Long Short-Term Memory (LSTM) 13
1.4. Cơ sở dữ liệu giá nhà 15
1.4.1. Trên thế giới 15
1.4.2. Tại Việt Nam 16
1.5. Một số công trình nghiên cứu liên quan đến đề tài 18
1.5.1. Công trình nghiên cứu trong nước 18
1.5.2. Các công trình nghiên cứu quốc tế 19

Chương 2: THIẾT KẾ VÀ THỰC NGHIỆM MÔ HÌNH DỰ ĐOÁN GIÁ NHÀ BẰNG LONG SHORT-TERM MEMORY 21
2.1. Mô hình LSTM trong dự đoán giá nhà 21
2.2. Thu thập và xử lý dữ liệu 22
2.3. Xây dựng tập dữ liệu chuỗi thời gian 30
2.4. Thiết kế kiến trúc mô hình LSTM 31
2.5. Huấn luyện và đánh giá mô hình 33
2.6. Nhận xét, đánh giá chung 36
Chương 3. KẾT QUẢ THỰC NGHIỆM VÀ LUẬN BÀN 38
3.1. Kết quả thực nghiệm 38
3.2. Đánh giá kết quả 41
3.2.1. Hiệu quả mô hình 41
3.2.2. Ý nghĩa thực tiễn 41
3.2.3. Hạn chế 41
3.2.4. Khả năng mở rộng 42
KẾT LUẬN 45
DANH MỤC TÀI LIỆU THAM KHẢO 48
PHỤ LỤC 50

DANH MỤC CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
AI	Artificial Intelligenc	Trí tuệ nhân tạo
ANN	Artificial Neural Networks	Mạng nơ-ron nhân tạo
AR	Autoregressive	Mô hình thống kê
ARIMA	AutoRegressive Integrated Moving Average	Trung bình trượt tích phân tự hồi quy
CNN	Convolutional Neural Networks	Mạng nơ-ron tích chập
DL	Deep Learning	Mô hình học sâu
DNN	Deep Neural Network	Mạng nơ-ron sâu
GIS	Geographic Information System	Dữ liệu bản đồ số
LSTM	Long Short-Term Memory	Bộ nhớ ngắn dài hạn
MAE	Mean Absolute Error	Trung bình sai số tuyệt đối
MAPE	Mean Absolute Percentage Error	Sai số phần trăm trung bình tuyệt đối
MSE	Mean Squared Error	Sai số bình phương trung bình
RNN	Recurrent Neural Networks	Mạng nơ-ron hồi quy
RMSE	Root Mean Squared Error	Sai số bình phương trung bình căn bậc hai
RF	Random Forest	Rừng ngẫu nhiên

DANH MỤC BẢNG

Tên bảng	Trang
Bảng 1.1: Một số nguồn cung cấp dữ liệu giá nhà trên thế giới	16
Bảng 1.2: Một số trang web cung cấp dữ liệu giá nhà tại Việt Nam	17
Bảng 1.3: Nguồn dữ liệu mở về giá nhà tại Việt Nam của một số cơ quan nhà nước	17
Bảng 1.4: Một số tổ chức nghiên cứu cung cấp báo cáo thị trường giá nhà tại Việt Nam	18
Bảng 2.1: Quá trình lọc bỏ dữ liệu URL	24
Bảng 2.2: Mô tả các thuộc tính có trong dữ liệu JSON tương ứng với mỗi căn nhà	26
Bảng 2.3: Mô tả quá trình tiền xử lý file JSON	28
Bảng 2.4: Chuyển đổi định dạng dữ liệu thời gian	30
Bảng 2.5: Mô tả các tham số của mô hình LSTM và chi tiết tham số từng lớp	33
Bảng 2.6: Tham số cấu hình của mạng LSTM	33
Bảng 3.1: Kết quả của mô hình LSTM đánh giá trên bộ dữ liệu kiểm thử (Testing dataset)	38
Bảng 3.2: Tham số huấn luyện của các mô hình được so sánh	39
Bảng 3.3: So sánh hiệu suất của các mô hình	40

DANH MỤC HÌNH

Tên hình	Trang
Hình 1.1: Cấu trúc của mạng CNN	11
Hình 1.2: Mô hình RNN	12
Hình 1.3: Mô hình LSTM	13
Hình 2.1: Quy trình xây dựng mô hình dự đoán giá nhà bằng LSTM	22
Hình 2.2: Giao diện danh mục "Nhà riêng" khu vực Hà Nội trên trang batdongsan.com	23
Hình 2.3: Dữ liệu JSON mẫu về thông tin thu thập được từ một mẫu bất động sản	25
Hình 2.4: Biểu đồ Top 10 quận/huyện (cũ) tại Hà Nội có số lượng tin đăng mua bán nhà riêng cao nhất giai đoạn từ tháng 11/2024 đến tháng 10/2025	27
Hình 2.5: Biểu đồ Loss Over Epochs	34

MỞ ĐẦU

1. Lý do chọn đề tài

Trong những năm gần đây, thị trường bất động sản tại Việt Nam, đặc biệt là tại các đô thị lớn như Hà Nội, thành phố Hồ Chí Minh đã chứng kiến sự biến động mạnh mẽ cả về giá cả lẫn nhu cầu giao dịch. Việc dự đoán chính xác giá nhà không chỉ hỗ trợ các cá nhân và tổ chức trong việc ra quyết định mua bán, đầu tư mà còn có ý nghĩa quan trọng trong quy hoạch và hoạch định chính sách phát triển đô thị.

Tuy nhiên, giá nhà thường biến động phức tạp do chịu ảnh hưởng từ nhiều yếu tố như vị trí địa lý, thời điểm giao dịch, chính sách kinh tế và xu hướng thị trường, v.v. Do đó, việc xây dựng một mô hình dự đoán chính xác là một thách thức lớn đối với các nhà nghiên cứu và nhà phân tích dữ liệu.

Trong bối cảnh đó, các phương pháp truyền thống như hồi quy tuyến tính hay mô hình ARIMA bộc lộ hạn chế khi xử lý dữ liệu phi tuyến và phụ thuộc thời gian dài. Trong khi đó, các mô hình học sâu, đặc biệt là mạng Long Short-Term Memory (LSTM) đã chứng minh hiệu quả vượt trội trong việc xử lý chuỗi thời gian có tính dài hạn và nhiều yếu tố ảnh hưởng. Với khả năng ghi nhớ thông tin theo thời gian và học mối quan hệ phức tạp giữa các đặc trưng, mô hình LSTM ngày càng được ứng dụng rộng rãi trong các bài toán như dự báo tài chính, phân tích hành vi tiêu dùng và gần đây là trong lĩnh vực bất động sản. Hiện nay, việc ứng dụng LSTM vào dự đoán giá nhà tại Việt Nam vẫn còn khá mới mẻ và chưa có nhiều nghiên cứu chuyên sâu.

Xuất phát từ nhu cầu thực tiễn cũng như khoảng trống trong nghiên cứu hiện nay, đề tài ***"Áp dụng mô hình Long Short-Term Memory trong dự đoán giá nhà tại Việt Nam"*** được lựa chọn nhằm mục tiêu khai thác tiềm năng của mô hình LSTM trong việc dự báo chuỗi thời gian phức tạp, đồng thời đề xuất một hướng tiếp cận mới mang tính ứng dụng cao. Kết quả nghiên cứu kỳ vọng sẽ đóng góp vào việc nâng cao độ chính xác trong dự đoán giá nhà, đồng thời mở rộng khả năng ứng dụng của học sâu trong lĩnh vực bất động sản tại Việt Nam trong thời đại số hóa.

2. Tổng quan về vấn đề nghiên cứu

Trong những năm gần đây, thị trường bất động sản Việt Nam đã chứng kiến sự tăng trưởng mạnh mẽ, đặc biệt tại các thành phố lớn. Giá nhà liên tục biến động do ảnh hưởng của nhiều yếu tố như tốc độ đô thị hóa, chính sách tài chính tiền tệ, nhu cầu dân cư và xu hướng đầu tư. Việc dự đoán chính xác giá nhà không chỉ giúp người dân có kế hoạch tài chính phù hợp mà còn hỗ trợ các nhà đầu tư, doanh nghiệp và cơ quan quản lý đưa ra quyết định hiệu quả. Tuy nhiên, bài toán dự đoán giá nhà vẫn tồn tại nhiều thách thức. Dữ liệu bất động sản thường mang tính chuỗi thời gian, có độ trễ và phụ thuộc vào các yếu tố kinh tế xã hội phức tạp. Các phương pháp truyền thống như hồi quy tuyến tính, mô hình ARIMA hay cây quyết định thường không đủ khả năng xử lý mối quan hệ phi tuyến và phụ thuộc dài hạn trong dữ liệu.

Trong bối cảnh đó, các mô hình học sâu (Deep Learning) đã mở ra hướng tiếp cận mới. Đặc biệt, mô hình LSTM - một biến thể của mạng nơ-ron hồi quy (RNN) được thiết kế để ghi nhớ thông tin dài hạn trong chuỗi dữ liệu, giúp cải thiện độ chính xác trong dự đoán. LSTM đã được ứng dụng thành công trong nhiều lĩnh vực như tài chính, y tế, dịch vụ và đang dần được quan tâm trong lĩnh vực bất động sản.

Tại Việt Nam, việc ứng dụng LSTM vào dự đoán giá nhà vẫn còn khá mới mẻ, chưa có nhiều nghiên cứu chuyên sâu. Do đó, đề tài này hướng đến việc xây dựng mô hình LSTM phù hợp với dữ liệu thực tế tại Việt Nam, từ đó nâng cao hiệu quả và tiềm năng ứng dụng trong thực tiễn. Đề tài đóng góp một hướng tiếp cận khả thi trong việc ứng dụng trí tuệ nhân tạo - mô hình LSTM vào bài toán dự báo giá nhà tại Việt Nam, đồng thời mở ra tiềm năng cho các nghiên cứu và ứng dụng sâu rộng hơn trong lĩnh vực bất động sản và phân tích dữ liệu chuỗi thời gian.

3. Mục đích nghiên cứu

Mục tiêu tổng quát: Xây dựng và đánh giá hiệu quả của mô hình LSTM trong việc dự đoán giá nhà tại Việt Nam dựa trên dữ liệu chuỗi thời gian nhằm cải thiện độ chính xác so với các phương pháp truyền thống.

Mục tiêu cụ thể:

(i) Phân tích đặc điểm dữ liệu giá nhà tại Việt Nam và các yếu tố ảnh hưởng đến biến động giá;

(ii) Tiền xử lý dữ liệu và xây dựng tập dữ liệu phù hợp cho mô hình học sâu;

(iii) Thiết kế và huấn luyện mô hình LSTM để dự đoán giá nhà theo chuỗi thời gian;

(iv) So sánh hiệu suất của mô hình LSTM với các mô hình truyền thống như hồi quy tuyến tính, ARIMA;

(v) Đề xuất hướng ứng dụng mô hình vào thực tiễn như hỗ trợ ra quyết định đầu tư, định giá tài sản hoặc hoạch định chính sách.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu:

(i) Dữ liệu giá nhà theo chuỗi thời gian tại Việt Nam (bao gồm các thông tin như giá bán, diện tích, vị trí, thời điểm giao dịch và các yếu tố liên quan đến thị trường bất động sản);

(ii) Mô hình học sâu LSTM (có khả năng ghi nhớ thông tin dài hạn, giúp dự đoán xu hướng giá nhà dựa trên các yếu tố lịch sử);

(iii) Các yếu tố ảnh hưởng đến giá nhà (vị trí địa lý, biến động thị trường, thu nhập bình quân, tốc độ đô thị hóa, v.v.).

Phạm vi nghiên cứu:

(i) Không gian: Tập trung vào dữ liệu giá nhà đất tại thành phố Hà Nội, nơi có thị trường sôi động và dữ liệu tương đối đầy đủ;

(ii) Thời gian: Dữ liệu được thu thập trong vòng khoảng 12 tháng nhằm đảm bảo tính cập nhật và phản ánh xu hướng thị trường;

(iii) Kỹ thuật: Nghiên cứu tập trung vào việc xây dựng, huấn luyện và đánh giá mô hình LSTM. Không đi sâu vào các mô hình hybrid phức tạp hay các yếu tố pháp lý - chính sách của thị trường bất động sản;

(iv) Dữ liệu: Sử dụng dữ liệu thứ cấp từ các nguồn công khai như website rao bán nhà (Batdongsan.com.vn, Alonhadat.com.vn, v.v.), báo cáo thị trường (Bộ Xây dựng, CBRE Vietnam, Savills Vietnam, v.v.) hoặc cơ sở dữ liệu mở (Open Data TP.HCM, Open Data Hà Nội, v.v.). Không bao gồm dữ liệu cá nhân hoặc thông tin nội bộ từ doanh nghiệp.

5. Phương pháp nghiên cứu

Phương pháp lý thuyết: tổng hợp, phân tích tài liệu liên quan (thị trường bất động sản và các mô hình dự đoán giá nhà; học máy, học sâu; dữ liệu chuỗi thời gian, mạng nơ - ron hồi tiếp, LSTM, v.v).

Phương pháp thực nghiệm: Triển khai các thí nghiệm trên bộ dữ liệu cụ thể, đánh giá kết quả bằng các độ đo định lượng.

Chương 1: TỔNG QUAN CƠ SỞ LÝ THUYẾT

Chương đầu tiên giới thiệu những cơ sở lý thuyết phục vụ cho việc thiết kế mô hình dự đoán giá nhà bằng LSTM, gồm các nội dung sau: bài toán dự đoán giá nhà, chuỗi thời gian, các mô hình học máy truyền thống và học sâu, nguồn dữ liệu giá nhà hiện nay và cuối cùng là một số nghiên cứu liên quan đến việc ứng dụng mô hình LSTM trong bài toán dự đoán giá nhà. Qua đó khẳng định mô hình LSTM là hoàn toàn phù hợp với nghiên cứu này.

1.1. Bài toán dự đoán giá nhà và chuỗi thời gian

1.1.1. Bài toán dự đoán giá nhà

Dự đoán giá nhà là một bài toán kinh điển trong lĩnh vực trí tuệ nhân tạo và có ý nghĩa, mang lại giá trị thực tế cao trong lĩnh vực kinh tế - xã hội, đầu tư bất động sản, và công tác hoạch định các chính sách đô thị. Mục tiêu bài toán hướng tới là thiết kế một mô hình có khả năng ước lượng chính xác giá cho mỗi sản phẩm nhà đất trên cơ sở những đặc trưng như vị trí địa lý, hệ thống tiện ích xung quanh, diện tích, số phòng và đặc biệt là có yếu tố thời gian. Ở Việt Nam, thị trường bất động sản hiện nay biến động lớn và chịu ảnh hưởng bởi những yếu tố kinh tế - xã hội, nên việc áp dụng những mô hình dự đoán hiện đại trong bối cảnh hiện nay là hết sức cần thiết để nâng cao chất lượng, hiệu quả dự báo, từ đó hỗ trợ ra quyết định được chính xác.

1.1.2. Chuỗi thời gian

Chuỗi thời gian (time series) là 1 tập những giá trị theo dõi được sắp xếp theo trật tự thời gian như: giây, phút, giờ, ngày, tháng, v.v. Chuỗi thời gian có 4 đặc điểm chính đó là tính mùa vụ (seasonnality), tính vòng lặp (cyclicity), tính xu hướng (trend) và tính bất thường (irregularity) [1].

Trong khoa học dự báo thì dữ liệu trong chuỗi thời gian là 1 loại dữ liệu đặc biệt, nó được thu thập và sắp xếp có trình tự về thời gian. Mỗi điểm dữ liệu trong chuỗi sẽ đại diện cho 1 giá trị tại 1 thời điểm nhất định, chẳng hạn giá nhà trung bình theo tháng hoặc quý trong một khu vực. Một trong những điểm nổi bật của

chuỗi thời gian là giá trị hiện thời có thể chịu sự ảnh hưởng từ các giá trị trong trước đấy. Vì vậy công việc phân tích, dự báo chuỗi thời gian đòi hỏi mô hình cần khai thác được hiệu quả mối quan hệ này.

Quy trình dự báo được thực hiện như sau:

Đầu tiên là chuẩn bị dữ liệu, dữ diệu phải được làm sạch ngay sau khi thu thập, tiếp theo được mã hóa biến, sau đó được chuẩn hóa, có thể biến đổi, hoặc tạo thêm các biến mới. Đồng thời, bộ dữ liệu sẽ được phân chia làm 2 phần, một phần cho tập huấn luyện, 1 phần cho tập kiểm tra. Để đảm bảo hiệu quả của tính toán, thông thường tập kiểm tra chiếm khoảng từ 20-30%.

Tiếp theo là thực hiện huấn luyện mô hình, lấy các dữ liệu trong tập huấn luyện đưa vào huấn luyện, xây dựng những mô hình khác nhau.

Sau đó tiến hành kiểm tra và so sánh, trước tiên kiểm tra mức độ chính xác hoặc những chỉ số sai số ở trong phần huấn luyện và phần kiểm tra, từ đấy có sơ sở để tìm ra được mô hình tốt và tiềm năng hơn.

Tiếp tục tinh chỉnh, cải thiện hiệu suất của mô hình, những mô hình tốt tiếp tục được tinh chỉnh lại tham số, đồng thời lặp lại quy trình trình huấn luyện nhằm nâng cao hiệu quả của mô hình.

Cuối cùng là bước dự báo, mô hình nào được chọn sẽ là mô hình tốt và tối ưu nhất. Nó sẽ được áp dụng cho đối với tất cả dữ liệu ban đầu để cho ra kết quả dự báo tương lai.

1.2. Một số mô hình học máy truyền thống

Học máy - ML là cách thức máy tính khai thác những kinh nghiệm, quan sát hoặc các dữ liệu sẵn có nhằm cải thiện hiệu quả hoạt động trong tương lai, thay vì chỉ tuân thủ những quy tắc được lập trình trước đó. Đây là một lĩnh vực nghiên cứu trọng yếu của trí tuệ nhân tạo đối với rất nhiều ứng dụng thực tiễn đem lại hiệu quả cao. Những mô hình học máy sẽ cho phép máy tính nhận diện các mẫu đưa vào phân tích, đưa ra các dự đoán hiệu quả. Học máy hiện đang được ứng dụng rộng khắp trong

nhiều ngành nghề như y học, kinh tế tài chính, thị giác máy tính, v.v. điển hình là trong dự báo chuỗi thời gian, đem lại hiệu quả cao. Chẳng hạn, ML với bài toán về dự đoán giá nhà, nó giúp mô hình xác định các quan hệ giữa những yếu tố đầu vào (vị trí, thời điểm giao dịch, diện tích, số phòng, v.v.) và giá trị ngôi nhà. Từ đó, đưa ra dự đoán đối với các trường hợp mới. Các mô hình học máy truyền thống thường được sử dụng trong bài toán dự báo chuỗi thời gian:

1.2.1. Các mô hình hồi quy cơ bản

Hồi quy tuyến tính là 1 mô hình nền tảng và quan trọng trong nhóm học máy có giám sát, thường được dùng cho các trường hợp dự báo giá trị thay đổi liên tiếp. Ở mô hình này, giả định rằng giữa biến đầu vào đối với biến mục tiêu có tồn tại mối liên hệ tuyến tính, công thức tổng quát (1.1):

$$y = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_p x_p + \epsilon \tag{1.1}$$

Trong đó: w_j là những hệ số hồi quy, x_j là những đặc trưng, y là biến mục tiêu, ϵ là sai số ngẫu nhiên.

Mục tiêu cốt lõi của việc huấn luyện mô hình là làm giảm tổng sai lệch bình phương giữa giá trị dự đoán với giá trị thực tế, bằng cách ứng dụng các thuật toán tối ưu như giảm Gradient.

Tuy nhiên, mối liên hệ giữa giá nhà trong thực tế và những đặc trưng thường không tuân theo dạng tuyến tính đơn giản. Nhiều hiện tượng có tính phi tuyến hoặc tồn tại các tương tác phức tạp giữa các biến, khiến cấu trúc dữ liệu không thể được nắm bắt đầy đủ bởi mô hình tuyến tính. Vì vậy, cần phải mở rộng sang những mô hình hồi quy phi tuyến, đặc biệt là mô hình hồi quy đa thức để mô hình có khả năng thể học được những mối liên hệ phi tuyến và những tương tác bậc cao giữa những đặc trưng.

Hồi quy đa thức thêm vào những số hạng bậc của các biến đầu vào cũng như các tương tác giữa chúng để mở rộng không gian đặc trưng. Ví dụ, với hai biến đầu vào x_1 và x_2 và bậc đa thức $d=2$, những đặc trưng mới được tạo ra

gồm $x_1, x_2, x_1^2, x_2^2, x_1x_2$, làm mô hình có thể học được những đường cong phức tạp và những mối quan hệ tương tác, phát triển khả năng biểu diễn của mô hình.

Đối với mô hình hồi quy tuyến tính, những hệ số thông thường được xác định là bình phương tối thiểu. Đây là một kỹ thuật thống kê ra đời từ rất lâu, có hiệu quả cao, triển khai dễ dàng trên các phần mềm và trong tính toán. Đồng thời, để có thể phục vụ cho hoạt động dự đoán, mô hình cũng cung cấp một số công học rõ ràng, dễ hiểu.

1.2.2. Mô hình AutoRegressive Integrated Moving Average (ARIMA)

ARIMA là 1 mô hình kinh điển trong lĩnh vực dự báo chuỗi thời gian và đã được ứng dụng rộng rãi. ARIMA dựa trên các giá trị trong quá khứ của chuỗi để ước lượng và dự đoán kết quả ở tương lai. Mô hình này có khả năng mô tả chuỗi thời gian có xu hướng mùa vụ, nhưng cũng có thể gặp một số hạn chế đối với dữ liệu phi tuyến, hoặc những yếu tố ảnh hưởng phức tạp.

Mô hình ARIMA gồm 3 phần chính:

Tự hồi quy - AR: biểu diễn mối liên hệ tuyến tính giữa giá trị hiện thời với các giá trị trong quá khứ.

Trung bình trượt - MA: biểu thị giá trị hiện thời có mối liên hệ phụ thuộc với các sai số của các bước trước đấy.

Sai phân - I: giúp chuỗi trở nên ổn định bằng cách lấy sai phân từ chuỗi gốc.

Công thức (1.2) là phương trình của mô hình ARIMA(p,d,q):

$$\Delta Y_t = \Phi_1 \Delta_{t-1} + \Phi_2 \Delta Y_{t-2} + \cdots + \Phi_p \Delta Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \qquad (1.2)$$

Trong công thức đó:

ΔY_t là giá trị sai phân bậc d của chuỗi, tại thời điểm t

Φ_1, Φ_2, Φ_p đây là hệ số của thành phần tự hồi quy với độ trễ tương ứng

$\theta_1, \theta_2, \theta_p$ là hệ số của thành phần trung bình trượt ứng với sai số độ trễ tương ứng

ε_t là nhiễu trắng, tại thời điểm t với kỳ vọng = 0 và phương sai không đổi.

1.2.3. Mô hình rừng ngẫu nhiên (Random Forest - RF)

Rừng ngẫu nhiên là mô hình học kết hợp có thể áp dụng được cho cả 2 trường hợp hồi quy và phân loại. Mô hình này tổng hợp thành quả từ những cây quyết định để đưa ra kết quả dự đoán duy nhất, nên có tính linh hoạt và dễ sử dụng, đạt độ chuẩn xác lớn. Rừng ngẫu nhiên xây dựng 1 tập hợp các cây quyết định và mỗi cây sẽ được tiến hành huấn luyện trong 1 tập hợp con ngẫu nhiên của những đặc trưng và thay bằng phụ thuộc vào 1 cây độc nhất, RF khai thác "sức mạnh tập thể" để tạo ra những dự đoán đáng tin cậy hơn. Việc tăng lượng cây trong rừng sẽ giúp nâng cao sự chính xác và hạn chế hiện tượng quá khớp. Quy trình huấn luyện RF gồm 3 bước cơ bản sau: (1) Thực hiện lấy các mẫu ngẫu nhiên trong bộ dữ liệu huấn luyện để tạo thành bộ con. (2) Tiến hành lấy ngẫu nhiên ss biến ($ss \leq rr$) với rr là tổng số đặc trưng đầu, từ đó xây dựng cây quyết định dựa vào những biến đó cùng bộ dữ liệu con đã tạo trước đó. Quy trình được lặp lại nhiều lần với mục đích tạo thành nhiều cây quyết định. (3) Tính trung bình kết quả từ những cây quyết định và đưa ra những dự báo.

Trong khi đó, những mô hình truyền thống thường giả định rằng giữa các biến có quan hệ tuyến tính. Tuy nhiên, dữ liệu thực tế, nhất là dữ liệu giá nhà, thường mang tính phi tuyến, đồng thời chịu tác động bởi những yếu tố phức tạp như chính sách, tâm lý xã hội hay đặc thù địa phương. Ngoài ra, đối với những mô hình truyền thống còn yêu cầu dữ liệu cần thỏa mãn những giả định nghiêm ngặt như tính dừng, phân phối chuẩn của sai số và không có tự tương quan dư thừa, khiến việc áp dụng vào thực tiễn còn vướng nhiều khó khăn. Hiện nay, dữ liệu ngày càng nhiều và phức tạp về cấu trúc thì tính linh hoạt cũng như khả năng mở rộng của những mô hình này dần bộc lộ rõ hạn chế.

1.3. Các mô hình dự đoán học sâu

Khi dữ liệu ngày càng nhiều và độ phức tạp cao, mang tính phi tuyến hoặc có đặc trưng chuỗi thời gian như biến động giá nhà theo tháng, sự hạn chế của những mô hình học máy truyên thống sẽ bộc lộ. Chúng ít có thể khai thác những mối liên hệ tiềm ẩn sâu của tập dữ liệu, nhất là những phụ thuộc dài hạn hoặc các tương tác phi

tuyến giữa những đặc trưng nên xu thế các nghiên cứu đã chuyển sang các mô hình học sâu và sử dụng các mạng noron nhân tạo nhiều lớp để mô hình hóa những hàm ánh xạ phức tạp.

Điểm mạnh nổi bật của học sâu chính là khả năng tự động trích xuất đặc trưng từ dữ liệu đầu vào và không can thiệp thủ công. Các mô hình DL tiêu biểu như: ANN, CNN, RNN và đặc biệt là LSTM đã khẳng định hiệu quả vượt trội trong các trường hợp nhận dạng và xử lý hình ảnh, dự báo dữ liệu chuỗi thời gian cũng như trong xử lý ngôn ngữ tự nhiên.

1.3.1. Mạng Artificial Neural Networks (ANN)

Trong ANN, mỗi nơ-ron thực hiện 1 phép biến đổi tuyến tính có trọng số, sau đó dùng hàm kích hoạt nhằm tạo ra đầu ra. ANN có thể học những hàm ánh xạ phức tạp, song lại thường gặp hạn chế trong xử lý chuỗi dữ liệu có yếu tố phụ thuộc vào thời gian. Trong học máy và học sâu, ANN đóng vai trò nền tảng cho nhiều mô hình hiện đại hơn như CNN, RNN và LSTM.

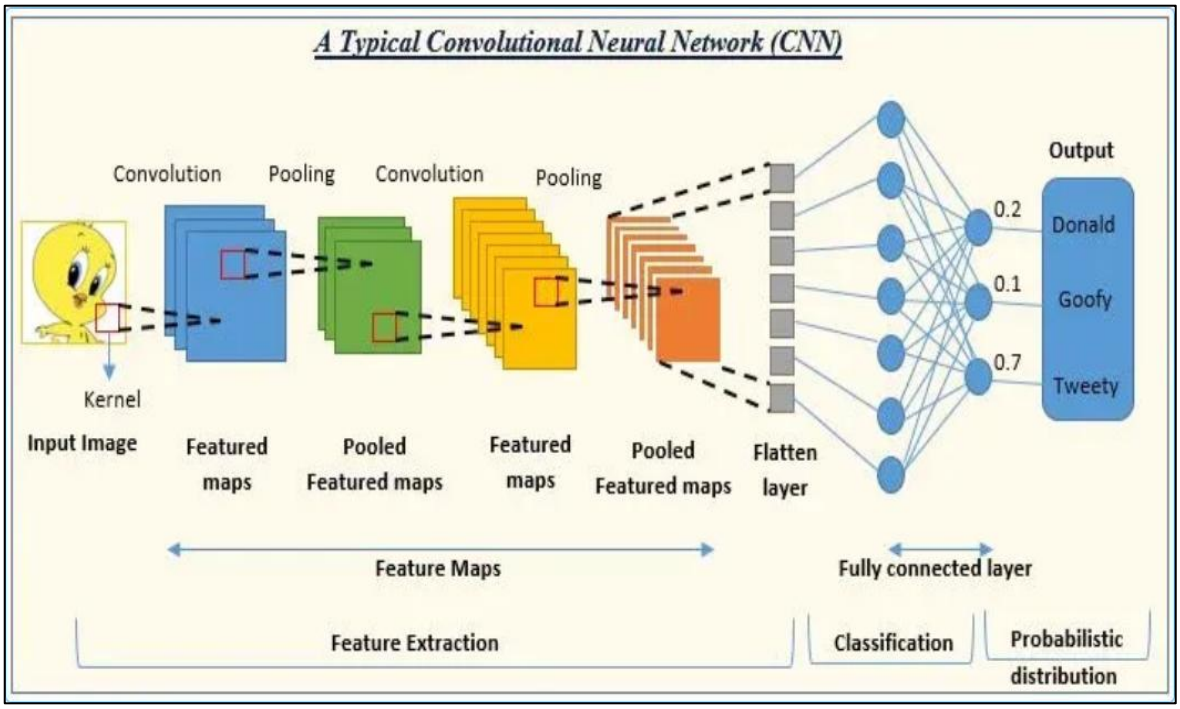
Mục tiêu cốt lõi của mạng noron nhân tạo là học được hàm ánh xạ từ dữ liệu vào đến đầu ra qua chu trình huấn luyện với bộ dữ liệu thực tế. ANN có thể học các liên hệ phi tuyến giữa dữ liệu đầu vào và đầu ra, điều này không thể hiện ở những mô hình truyền thống. Ví dụ, mối quan hệ giữa vị trí địa lý, diện tích, hệ thống tiện ích xung quanh ngôi nhà với giá nhà thường không là tuyến tính nhưng ANN có thể học được những mẫu phức tạp này.

Tuy nhiên, ANN truyền thống cũng tồn tại những điểm yếu, nhất là trong giải quyết kiểu dữ liệu chuỗi thời gian. Do thiếu cơ chế ghi nhớ trạng thái trước đó, ANN thường không khai thác được những mối liên quan hệ phụ thuộc theo thời gian tại tập dữ liệu. Đó chính là yếu tố rất quan trọng trong dự báo giá nhà theo thời gian, khi xu hướng thị trường và những yếu tố có tính mùa vụ có ảnh hưởng lớn. Vì vậy, nhằm giải quyết hạn chế này, các mô hình RNN, LSTM giúp mạng có thể ghi nhớ, xử lý thông tin theo chuỗi.

1.3.2. Mạng Convolutional Neural Networks (CNN)

CNN là 1 loại mạng nơ-ron chuyên biệt, có độ chính xác rất lớn trong giải quyết loại dữ liệu có cấu trúc dạng lưới, nhất là dữ liệu ảnh. Xuất phát từ cảm hứng về vỏ não thị giác con người, CNN có thể tự học và hoàn toàn thích nghi đối với các hệ thống phân cấp không gian của đặc trưng của dữ liệu đầu vào. Vì vậy, CNN trở thành kiến trúc nền tảng cho phần lớn các tác vụ thị giác máy tính hiện nay.

CNN được hình thành từ nhiều lớp Convolution xếp chồng lên nhau. Ngoài ra, mạng còn dùng một số hàm kích hoạt phi tuyến (chẳng hạn hàm *ReLU* hay hàm *tanh*) để kích hoạt trọng số ở các Node và sau khi đi qua những hàm này, mỗi lớp thu được các trọng số tại các Node và tạo ra những thông tin mờ, phục vụ cho những lớp kế tiếp [8].



Hình 1.1: Cấu trúc của mạng CNN

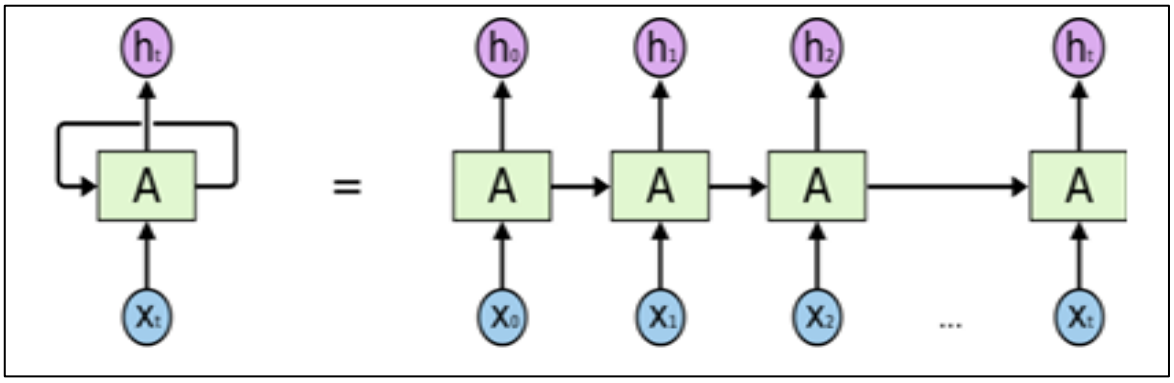
(Nguồn: https://vietnix.vn/cnn-la-gi)

Cách thức hoạt động của CNN khác các mạng nơ-ron truyền thống, CNN sử dụng phép toán đặc biệt có tên là tích chập, nó cho phép mạng học những đặc trưng trong phạm vi tiếp nhận cục bộ, đồng thời có thể duy trì mối quan hệ không gian giữa những pixel.

1.3.3. Mạng Recurrent Neural Networks (RNN)

RNN là 1 loại mạng noron đặc biệt được xây dựng nhằm xử lý bộ dữ liệu theo trình tự, trong đó thông tin tại thời điểm hiện thời có thể phụ thuộc vào những trạng thái đã diễn ra trước đó. Khác những mạng noron truyền thống, mạng RNN có khả năng ghi nhớ dữ liệu trong quá khứ qua cơ chế phản hồi nội tại, từ đó tạo làm cho mô hình học được các mối liên hệ theo thời gian của chuỗi dữ liệu.

Trong khi, các mạng noron truyền thống thường có 3 lớp chính (lớp đầu vào, các lớp ẩn và lớp đầu ra) với lớp đầu vào và đầu ra hoạt tách biệt nhau thì đối với chuỗi thời gian, 2 thành phần này lại có sự liên kết chặt chẽ. Chính thế, RNN được phát triển với ý nghĩa cốt lõi là sẽ lưu giữ những thông tin từ các bước trước đấy để cho ra dự báo chính xác hơn đối với thời điểm hiện tại [7].



Hình 1.2: Mô hình RNN

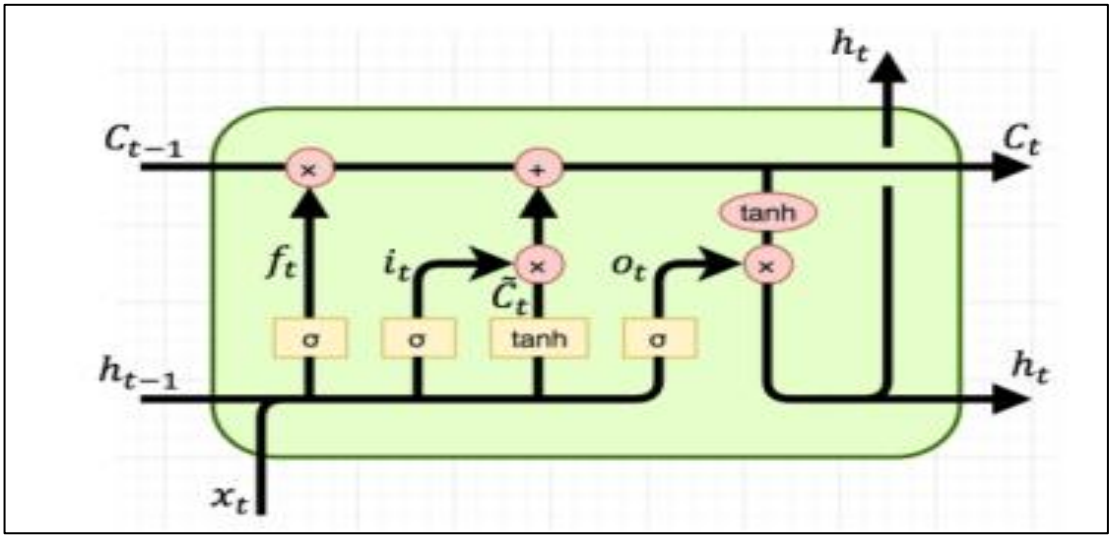
RNN truyền thống sẽ gặp một vấn đề lớn trong thời gian huấn luyện, đó là hiện tượng bị biến mất, hoặc bị bùng nổ gradient (vanishing/exploding gradient). Khi chuỗi dữ liệu quá dài, các gradient được truyền ngược qua nhiều bước thời gian có thể trở nên quá bé hoặc quá to, làm cho mô hình khó khăn để học được những phụ thuộc dài hạn một cách hiệu quả. Hệ quả là RNN chỉ có thể ghi nhớ thông tin trong ngắn hạn, làm giảm độ chính xác đối với những bài toán cần nhận diện xu hướng dài hạn, chẳng hạn như dự đoán giá nhà. Để khắc phục hạn chế này, các biến thể nâng cao của RNN đã được cải tiến, tiêu biểu là LSTM, GRU. Những mô hình này bổ sung cơ chế kiểm tra, rà soát dòng thông tin thông qua các cổng và cho phép mạng xác định dữ liệu nào

cần lưu trữ lại, cập nhật hay xóa bỏ. Nhờ vậy, LSTM và GRU có khả năng học được những liên hệ dài hạn từ các chuỗi thời gian, giúp nâng cao hiệu quả của các dự báo.

Thị trường bất động sản Việt Nam hiện nay có đặc điểm nổi bật là biến động theo chu kỳ và bị tác động bởi nhiều yếu tố theo thời gian, mô hình RNN cùng các biến thể của nó có giá trị quan trọng. Việc áp dụng RNN giúp mô hình khai thác được xu hướng giá, những biến động mang tính mùa vụ, cũng như những liên hệ tiềm ẩn giữa thời gian và giá trị của nhà đất. Hơn nữa, để có thể đạt hiệu quả tối ưu nhất, cũng cần lựa chọn kiến trúc phù hợp, trong đó LSTM thường được ưu tiên bởi khả năng về ghi nhớ dài hạn của nó và tính ổn định cao trong suốt thời gian huấn luyện.

1.3.4. Mô hình Long Short-Term Memory (LSTM)

LSTM do Hochreiter và Schmidhuber giới thiệu năm 1997, đây là 1 kiến trúc cải tiến của mạng nơ-ron hồi quy nhằm giải quyết hiện tượng biến mất gradient và các phụ thuộc dài hạn. Về cơ bản, RNN là 1 chuỗi những mô-đun lặp lại. Trong những mô hình mạng nơ-ron hồi quy thông thường, các đơn vị lặp lại có cấu trúc khá đơn giản, chủ yếu gồm 1 tầng *tanh*. Tương tự RNN và LSTM cũng có cấu trúc tuần tự nhưng những đơn vị lặp lại sẽ được thiết kế phức tạp hơn. Cụ thể, mô hình LSTM có 4 thành phần chính là Cell, Forget Gate, Input Gate, Output Gate, phối hợp cùng nhau theo cơ chế đặc biệt phục vụ xử lý và lưu trữ thông tin hiệu quả hơn [7].



Hình 1.3: Mô hình LSTM

Cổng quên (*Forget Gate*): Với chức năng xác định và loại bỏ dữ liệu không cần thiết từ những trạng thái trước đấy. Nó nhận dữ liệu x_t từ dữ liệu đầu vào, cùng trạng thái ẩn h_{t-1}, tiếp theo đưa qua 1 hàm kích hoạt σ để tạo thành giá trị $^{(0,1)}$, từ đó xác định phần dữ liệu nào cần giữ lại và phần nào cần loại bỏ.

$$f_t = \sigma(U_f * x_t + W_f * h_{t-1} + b_f) \quad (1.3)$$

Những dữ liệu mới sẽ được lưu giữ ở trạng thái hiện tại.

Cổng đầu vào (Input Gate) tiếp nhận dữ liệu từ trạng thái hiện thời cùng trạng thái ẩn của bước trước, rồi đưa qua hai hàm, đó là hàm kích hoạt σ và trả về giá trị $^{(0,1)}$ nhằm xác định những thành phần cần cập nhật. Đồng thời, hàm kích hoạt *tanh* sinh ra 1 véctơ mới C_t có giá trị $^{(-1,1)}$ bổ sung vào trạng thái.

$$i_t = \sigma(U_i * x_t + W_i * h_{t-1} + b_i) \quad (1.4)$$

$$\widetilde{C_t} = tanh(U_c * x_t + W_c * h_t + b_c) \quad (1.5)$$

Ô trạng thái (Cell state) C_t sẽ tiến hành cập nhật trạng thái bằng cách tiếp nhận những thông tin đã bị bỏ ở các ô trước đó, đồng thời những thông tin mới sẽ bổ sung thông qua cổng đầu vào $i_t * \widetilde{C_t}$

$$C_t = \sigma(f_t * C_{t-1} + i_t * \widetilde{C_t}) \quad (1.6)$$

Cổng đầu ra (Output gate) có vai trò lựa chọn các dữ liệu được đưa ra từ trạng thái hiện tại. Nó quyết định lượng thông tin trạng thái cần xuất bằng cách sử dụng các giá trị do hàm kích hoạt trả về. Sau đó thì trạng thái mới sẽ được truyền qua hàm *tanh* nhằm tạo ra 1 giá trị. Cuối cùng, 2 giá trị này được sẽ kết hợp lại để thành đầu ra cho trạng thái ẩn.

$$o_t = \sigma(U_0 * x_t + W_0 * h_{t-1} + b_0) \quad (1.7)$$

$$h_t = tanh(C_t) * o_t \quad (1.8)$$

Dựa trên cơ chế hoạt động đã trình bày, LSTM được xem là nổi trội hơn rất nhiều so với RNN. Nó có thể truy xuất dữ liệu từ 1 tập dữ liệu lớn hơn rất nhiều, nhờ đó rất phù hợp trong các trường hợp dự báo có yếu tố chuỗi thời gian dài hạn.

Bên cạnh đó, một trong các ưu điểm của LSTM kể đến là nó có thể mở rộng và phối kết hợp với những kiến trúc học sâu khác. LSTM có thể được xếp chồng thành nhiều lớp để tăng độ sâu của mô hình hoặc có thể kết hợp các mạng tích chập trong xử lý dữ liệu không gian, ví dụ bản đồ vị trí bất động sản. LSTM còn có thể tích hợp với attention mechanism để tăng khả năng tập trung vào các phần quan trọng của chuỗi dữ liệu. Ngoài ra, LSTM có một số điểm yếu như tốc độ huấn luyện chậm, cần số lượng dữ liệu đủ lớn để có khả năng nâng cao hiệu quả dự đoán.

Với bài toán dự đoán giá nhà, học sâu (Deep Learning) giúp mô hình biết những chiều hướng của thị trường, các biến động theo mùa và những yếu tố như kinh tế, xã hội sẽ tác động lớn đến giá nhà. Ngoài ra, học sâu còn có thể mở rộng và thích ứng tốt đối với dữ liệu lớn. Khi nguồn dữ liệu bất động sản càng lớn và phức tạp như dữ liệu giao dịch mua bán, dữ liệu định vị theo bản đồ, dữ liệu về kinh tế vĩ mô, các mô hình học sâu có khả năng tích hợp, xử lý đồng thời các nguồn thông tin để nâng cao chất lượng, hiệu quả dự báo. Mặc dù vậy, học sâu cũng yêu cầu lượng tài nguyên phục vụ tính toán lớn đáng kể, thời gian cần để huấn luyện dài, đồng thời các kỹ thuật tối ưu tương đối phức tạp. Vì vậy việc triển khai cần được cân nhắc kỹ lưỡng về cả kỹ thuật và chi phí để triển khai.

1.4. Cơ sở dữ liệu giá nhà

1.4.1. Trên thế giới

Cơ sở dữ liệu giá nhà toàn cầu có thể truy cập từ các tổ chức uy tín như: Ngân hàng Dự trữ Liên bang Dallas, BIS - Residential Property Prices, OECD Housing Prices, IMF - Global Housing Watch, Global Property Guide và Numbeo Property Investment Indexumbeo, v.v. Các nguồn này cung cấp dữ liệu theo quốc gia, chỉ số giá, tỷ lệ thu nhập, tỷ suất lợi nhuận cho thuê và xu hướng thị trường, cụ thể tại Bảng 1.1:

Bảng 1.1: Một số nguồn cung cấp dữ liệu giá nhà trên thế giới

Nguồn	Mô tả	Định dạng
Dallas Fed - International House Price Database	Dữ liệu giá nhà và thu nhập khả dụng theo quý từ năm 1975 đến nay, đã điều chỉnh mùa vụ	Bảng dữ liệu, biểu đồ
IMF - Global Housing Watch	Theo dõi giá nhà, tỉ lệ giá/thu nhập, giá/thuê và các chỉ số định giá theo quý	Bảng dữ liệu, biểu đồ
Numbeo - Property Prices Index by Country	Chỉ số giá nhà, tỷ suất lợi nhuận cho thuê, tỉ lệ giá/thu nhập, giá/thuê theo quốc gia	Bảng xếp thứ hạng, dữ liệu định kỳ
BIS - Residential Property Prices	Dữ liệu chi tiết và tổng hợp về giá nhà tại hơn 60 quốc gia, chia theo vùng và loại hình	Bảng dữ liệu, dashboard
Global Property Guide	So sánh giá nhà, lợi suất cho thuê, giá/m^2 tại hơn 80 quốc gia. Có dữ liệu lịch sử 1,5,10,15 năm	Bảng xếp hạng, biểu đồ
OECD Housing Prices	Chỉ số giá nhà thực và danh nghĩa, tỷ lệ giá/thuê, giá/thu nhập	Biểu đồ, bảng dữ liệu
Numbeo Property Investment Index	Chỉ số nhà, tỷ suất lợi nhuận cho thuê, giá/m^2, giá/thu nhập	Bảng xếp hạng, dữ liệu định kỳ

1.4.2. Tại Việt Nam

Dữ liệu trong lĩnh vực bất động sản ở Việt Nam hiện nay cũng đã có những nguồn cung cấp phong phú, đa dạng và số lượng nhiều hơn trước. Tuy nhiên, dữ liệu vẫn đang ở trong tình trạng phân tán và thiếu đồng bộ. Hiện nay, thông tin về giá nhà được cung cấp qua một số nền tảng thương mại như Batdongsan.com.vn, Homedy, Guland, Landviet, v.v. cũng như từ nguồn dữ liệu mở của nhà nước, các báo cáo thị trường do một số tổ chức cung cấp, cụ thể sau đây:

Bảng 1.2: Một số trang web cung cấp dữ liệu giá nhà tại Việt Nam

Nguồn dữ liệu	Mô tả	Định dạng
Batdongsan.com.vn	Biểu đồ giá nhà đất theo khu vực, loại hình, có dữ liệu lịch sử	Bảng HTML, biểu đồ giá
Guland.vn	Bản đồ giá nhà đất toàn quốc, tra cứu theo tọa độ, số nhà, sổ đỏ	Bảng HTML, bản đồ tương tác
Homedy.com	Bảng giá nhà đất 64 tỉnh/thành, chia theo quận/huyện, phường/xã, đường phố	Bảng HTML, có bản đồ
Landviet.com.vn	Bảng giá chi tiết theo địa phương, cập nhật hàng tháng	Bảng HTML
Nhatot.com	Phân tích giá nhà theo phân khúc, khu vực, xu hướng thị trường	Bảng HTML
Onehousing.vn	Tập trung vào căn hộ, dự án lớn; cung cấp thông tin giá, diện tích, tiện ích	Bảng HTML, Bảng giá dự án
Dothi.net	Tin rao mua bán và cho thuê đa dạng (nhà phố, chung cư, đất nền, v.v.)	Bảng HTML
123nhadatviet.vn	Tin rao mua bán và cho thuê đa dạng (nhà phố, chung cư, đất nền, v.v.)	Bảng HTML
Cafeland.vn	Ngoài tin rao còn có bảng giá dự án, phân tích thị trường	Bảng HTML, Bảng giá dự án
Alonhadat.com.vn	Tin rao mua bán và cho thuê đa dạng (nhà phố, chung cư, đất nền, v.v.)	Bảng HTML, có bộ lọc theo khu vực, loại hình

Bảng 1.3: Nguồn dữ liệu mở về giá nhà tại Việt Nam của một số quan nhà nước

Nguồn	Mô tả	Định dạng
Open Data TP.HCM	Bảng giá đất, quy hoạch, bản đồ hành chính	Excel, CSV
Open Data Hà Nội	Bảng giá đất và quy hoạch đất	Excel, CSV
Thư viện pháp luật	Quyết định của UBND về bảng giá đất từng tỉnh/thành phố	PDF, Word

Bảng 1.4: Một số tổ chức nghiên cứu cung cấp báo cáo thị trường giá nhà tại Việt Nam

Tổ chức	Loại dữ liệu
Bộ Xây dựng	Báo cáo định kỳ về chuyên đề bất động sản
Tổng cục Thống kê	Thống kê liên quan đến các chỉ số giá đất
Savills Vietnam	Báo cáo tình hình bất động sản, phân tích giá bán, giá thuê
CBRE Vietnam	Báo cáo giá căn hộ, văn phòng, nhà phố, đất nền
DKRA Vietnam	Báo cáo về thị trường nhà tại các tỉnh phía Nam

1.5. Một số công trình nghiên cứu liên quan đến đề tài

1.5.1. Công trình nghiên cứu trong nước

Luận văn "Dự đoán giá bất động sản tại Quận 7, Thành phố Hồ Chí Minh bằng mô hình LSTM" của Đỗ Hoàng Hiệp, Đại học Quốc gia Tp Hồ Chí Minh mang đến góc nhìn chuyên sâu về ứng dụng hiệu quả mô hình học sâu LSTM trong dự đoán giá nhà đất tại Quận 7. Trong đề tài, tác giả không chỉ mô tả các nguyên lý hoạt động của mô hình LSTM mà còn phân tích để làm rõ các yếu tố ảnh hưởng đến giá nhà đất trong khu vực. Nguồn dữ liệu cung cấp cho nghiên cứu được khai thác ở trang web *https://batdongsan.com.vn* [5].

Đề tài khoa học "Dự đoán giá nhà sử dụng dữ liệu Kaggle" của tác giả Vũ Đăng Hoàng Long & Nguyễn Huỳnh Đại Lợi, Huỳnh Long Nam, Trần Thanh Phúc, Nguyễn Đăng Trung Tiến - Đại học Khoa học tự nhiên Tp Hồ Chí Minh. Nhóm nghiên cứu đã sử dụng dữ liệu Ames Housing từ Kaggle để thiết kế các mô hình phục vụ dự đoán giá nhà. Nhóm đã thử nghiệm trên một số mô hình học máy, trong đó có LSTM. Công trình nghiên cứu mang đến một ví dụ thực nghiệm về việc áp dụng LSTM trong dự đoán giá nhà, mặc dù dữ liệu của Mỹ nhưng phương pháp có thể chuyển giao cho thị trường Việt Nam [6].

Bài viết "Công nghệ dự đoán giá nhà bằng Machine Learning giúp tối ưu lợi nhuận" tại HBR Việt Nam đã nêu khái quát về việc sử dụng học máy trong lĩnh vực dự đoán giá nhà, nhấn mạnh vai trò của những yếu tố như mức lãi suất, tình hình lạm phát, mức GDP, thị trường chứng khoán. Bài viết gợi ý hướng tích hợp dữ liệu kinh

tế vào LSTM để nâng cao hiệu quả dự đoán và độ chính xác trong bài toán dự báo giá nhà [2].

Đồ án tốt nghiệp "Mô hình hồi quy tuyến tính và ứng dụng trong dự báo giá nhà" - Đại học Bách khoa Hà Nội của Đỗ Văn Hùng. Trong Đồ án, tác giả sử dụng các mô hình hồi quy tuyến tính để cho ra các kết quả dự báo về giá nhà và tìm ra 9 thuộc tính có tác động, ảnh hưởng đáng kể đối với giá nhà gồm: Số tầng, số phòng, số phòng tắm, khu vực sống, số mặt tiền, diện tích đất, tình trạng nhà, chất lượng nhà, năm xây dựng [4].

Khóa luận tốt nghiệp "Ứng dụng một số mô hình học máy trong dự đoán giá bất động sản tại Hà Nội" của tác giả Trịnh Thị Thu Hòa - Học viện Ngân hàng đã khai thác thế mạnh và ứng dụng một số mô hình học máy để đưa ra dự báo giá nhà đất ở Hà Nội. Dữ liệu dùng trong nghiên cứu khóa luận được thu thập gồm các yếu tố nội tại và một số yếu tố pháp lý có tác động trực tiếp đến giá của bất động sản như: địa chỉ căn nhà, diện tích, nhà mặt tiền/nhà trong ngõ, kích thước ngang, kích thước dài, độ rộng đường/ngõ trước nhà, tình trạng pháp lý, số tầng, số lượng phòng ngủ, nhà bếp, số lượng phòng ăn, chỗ để xe ô tô, sân thượng [3].

1.5.2. Các công trình nghiên cứu quốc tế

Bài báo "House Price Prediction Using LSTM" của nhóm tác giả Xiaochen Chen, Lai Wei, Jiaxin Xu - Đại học Khoa học & Công nghệ Hồng Kông đã tập trung nghiên cứu áp dụng một số mô hình học máy trên dữ liệu giá nhà giai đoạn 2004 - 2016 tại 4 thành phố lớn: Bắc Kinh, Quảng Châu, Thâm Quyến và Thượng Hải. Với mục tiêu của nghiên cứu chính là dự đoán giá nhà trung bình trong hai tháng cuối năm 2016 bằng cách so sánh mô hình ARIMA truyền thống với LSTM. Nghiên cứu cho kết quả là mô hình LSTM đạt độ chính xác vượt trội hơn đối với dự báo dữ liệu kiểu chuỗi thời gian [18].

Bài báo "Comparison of Real Estate Price Prediction Based on LSTM and LGBM" - Đại học Nông nghiệp Trung Quốc của tác giả Songyi Shi. Bài báo so sánh hiệu quả dự đoán giá nhà tại Bắc Kinh giữa 2 mô hình LSTM và LightGBM, sử

dụng dữ liệu nhà ở cũ. Kết quả LSTM phù hợp, hiệu quả hơn đối với bài toán chuỗi thời gian đơn thuần ($R^2 = 0.91$), trong khi LightGBM hiệu quả hơn khi có nhiều yếu tố đầu vào [19].

Bài báo "Linear Regression for House Price Prediction" - NRI Institute of Technology, Ấn Độ của tác giả V. Tejaswini và cộng sự. Mặc dù tập trung vào hồi quy tuyến tính, nghiên cứu này cung cấp bối cảnh so sánh với những mô hình LSTM trong lĩnh vực dự đoán giá nhà [20].

Bài báo "Real Estate Price Prediction Using Deep Learning Techniques" đăng trên Springer Journal đã thực hiện so sánh một số mô hình như CNN, RNN và LSTM trong dự đoán giá của bất động sản với tập dữ liệu đã được thu thập được từ nhiều quốc gia. Công trình này cung cấp các đánh giá định lượng về độ chính xác, tính ổn định và triển vọng mở rộng của từng mô hình [21].

Bài báo "House Price Prediction Using LSTM Neural Network" - IEEE Conference Proceedings của nhiều nhóm nghiên cứu quốc tế. Tác giả đã trình bày về mô hình LSTM trong lĩnh vực dự đoán giá nhà đất trên cơ sở dữ liệu kiểu chuỗi thời gian và so sánh với những mô hình học máy truyền thống như mô hình ARIMA cùng những mô hình hồi quy tuyến tính. Bài báo cho thấy, đối với việc học những mối liên hệ phi tuyến và phụ thuộc dài hạn đối với dữ liệu về giá nhà thì mô hình LSTM cho hiệu quả vượt trội [22].

Tóm lại, Chương 1 đã hệ thống hóa cơ sở lý thuyết để phục vụ thiết kế, xây dựng mô hình dự đoán giá nhà bằng LSTM. Từ bài toán về dự đoán giá nhà, đặc điểm chuỗi thời gian và một số mô hình dự báo truyền thống cho đến học máy, học sâu đã được trình bày có hệ thống. Đặc biệt, phần phân tích về mô hình LSTM đã làm sáng tỏ căn cứ khoa học để lựa chọn phương pháp này cho nghiên cứu.

Chương 2: THIẾT KẾ VÀ THỰC NGHIỆM MÔ HÌNH DỰ ĐOÁN GIÁ NHÀ BẰNG LONG SHORT-TERM MEMORY

Chương 2 tập trung thiết kế và triển khai thực nghiệm mô hình LSTM, nội dung chương bao gồm: khảo sát mô hình LSTM trong dự báo giá nhà, thu thập và xử lý tập dữ liệu, chuyển đổi thành chuỗi thời gian, thiết kế, huấn luyện và đánh giá kết quả thực nghiệm của mô hình. Trên cơ sơ đó, đưa ra các nhận xét chung về ưu điểm, thế mạnh cũng như những yếu điểm, hạn chế của mô hình.

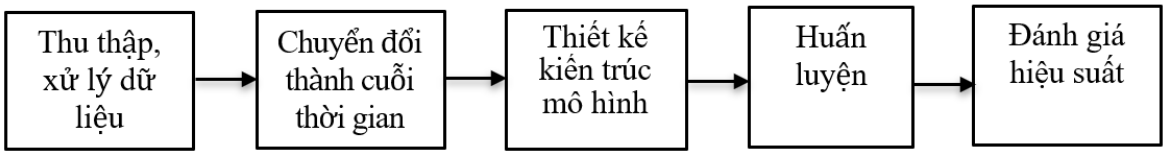
2.1. Mô hình LSTM trong dự đoán giá nhà

Khả năng học và nắm bắt những xu hướng dài hạn trong xử lý dữ liệu về giá nhà là ưu điểm nổi bật của LSTM. Thị trường nhà đất vốn mang tính chu kỳ và chịu sự tác động bởi nhiều yếu tố như: Những chính sách tín dụng, tốc độ đô thị hóa, mức lãi suất ngân hàng và tâm lý người dùng. Nó không chỉ tác động ngay lập tức mà còn để lại tác động kéo dài theo thời gian.

Mô hình LSTM vận hành dựa trên cơ chế những cổng điều khiển kết hợp với bộ nhớ trạng thái, cho phép ghi nhận và khai thác các ảnh hưởng đó để đưa ra những dự báo chính xác hơn đối với những mô hình truyền thống trước đây. Mặt khác, LSTM còn có thể giải quyết được dữ liệu phi tuyến đa chiều. Trong bài toán về dự đoán giá nhà thì dữ liệu đầu vào có thể gồm nhiều đặc trưng như là vị trí địa lý, diện tích nhà, số phòng ngủ, những tiện ích, thời điểm giao dịch, thậm chí cả những chỉ số kinh tế vĩ mô. Việc tích hợp những đặc trưng này vào trong mô hình LSTM giúp nó học được các quan hệ phức tạp giữa những yếu tố, từ đó giúp nâng mức độ chuẩn xác trong lĩnh vực dự báo.

Với những khả năng học phụ thuộc dài hạn, xử lý đối với dữ liệu phi tuyến và tích hợp đa đặc trưng, LSTM được xem là một công cụ vượt trội, rất phù hợp để dùng trong bài toán dự đoán giá nhà đất tại Việt Nam. Mô hình LSTM mở ra hướng tiếp cận mới cho những nghiên cứu và đẩy mạnh ứng dụng trong lĩnh vực dự đoán giá trị bất động sản để hỗ trợ ra các quyết định đầu tư, hoạch định các cơ chế, chính sách và quản lý thị trường được hiệu quả hơn.

Quy trình xây dựng mô hình gồm các bước như Hình 2.1: Trước tiên là thu thập, xử lý dữ liệu, chuyển đổi thành chuỗi thời gian, sau đó thiết kế kiến trúc cho mô hình, tiến hành huấn luyện, đánh giá hiệu quả của mô hình.



Hình 2.1: Quy trình xây dựng mô hình dự đoán giá nhà bằng LSTM

Quy trình này không những chỉ mang tính kỹ thuật, mà còn phản ánh tư duy hệ thống trong việc xử lý một bài toán phức tạp mà trong đó nguồn dữ liệu đầu vào mang tính đa chiều và có biến động theo thời gian. Mô hình học sâu LSTM chính là một lựa chọn tối ưu nhằm khai thác tối đa các đặc điểm chuỗi thời gian đối với dữ liệu giá nhà tại Việt Nam.

2.2. Thu thập và xử lý dữ liệu

Quá trình thu thập dữ liệu giá nhà được tiến hành bằng kỹ thuật cào dữ liệu tự động, trong đó Selenium được dùng để điều khiển trình duyệt và tải các trang web mục tiêu. Phạm vi thu thập URL được giới hạn trong mục **"Nhà riêng"** thuộc khu vực **Hà Nội** trên trang ***batdongsan.com***. Để đảm bảo tính đồng nhất và chất lượng cho bộ dữ liệu huấn luyện, quy trình thu thập dữ liệu được thiết kế chặt chẽ qua các bước sau:

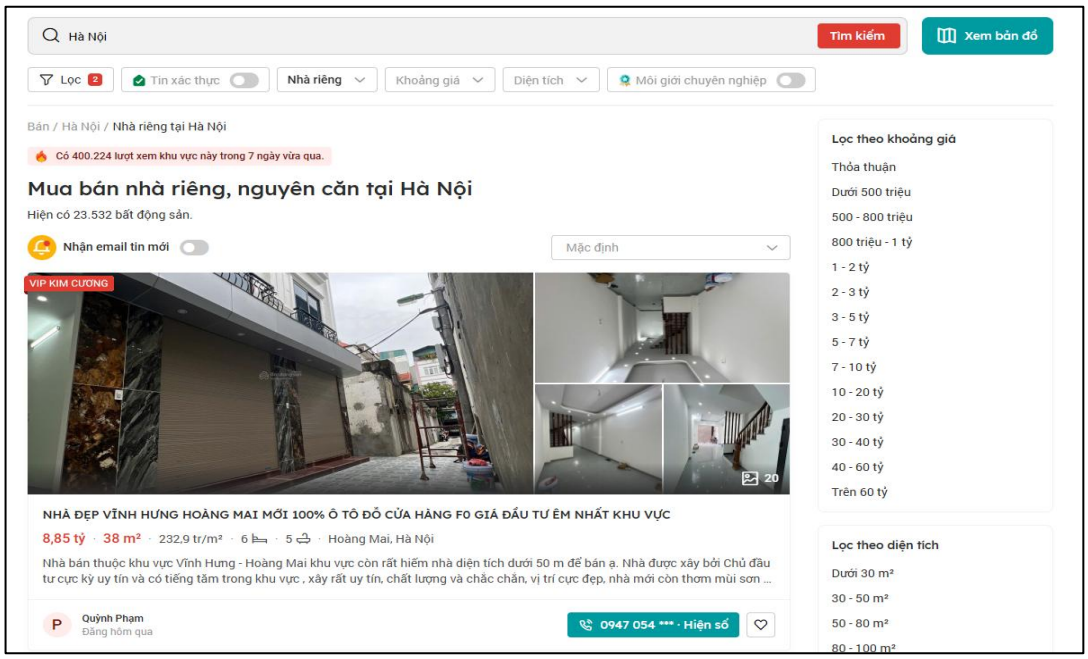
- *Xác định phạm vi và mục tiêu:* Việc thu thập danh sách đường dẫn (URL) được giới hạn nghiêm ngặt trong phân mục "Nhà riêng" tại khu vực Hà Nội trên website *batdongsan.com.vn*. Việc xác định phạm vi này giúp loại bỏ nhiễu từ các loại hình bất động sản khác (như chung cư, đất nền hay nhà kho) vốn có đặc điểm định giá hoàn toàn khác biệt.

- *Chiến lược khai thác dữ liệu phân trang:* Do hệ thống của website hiển thị kết quả theo dạng phân trang, với mật độ tối đa 20 tin đăng trên mỗi trang thuật toán cào dữ liệu được thiết lập để quét qua 30 trang kết quả đầu tiên. Theo tính toán, phạm

vi này cung cấp một lượng mẫu sơ bộ khoảng từ 500 đến 600 URL. Đây là số lượng mẫu vừa đủ để bao quát các biến động giá nhà tại nhiều quận huyện khác nhau trong cùng một thời điểm.

- *Cơ chế khử trùng lặp dữ liệu:* Sau khi thu thập được danh sách URL thô, hệ thống thực hiện bước kiểm tra tính duy nhất. Thay vì chỉ so sánh toàn bộ đường dẫn vốn có thể thay đổi nhẹ về cấu trúc theo thời gian, thuật toán tiến hành trích xuất mã định danh tin đăng (post_id). Nếu hai URL khác nhau nhưng có cùng một post_id, hệ thống sẽ chỉ giữ lại một đại diện duy nhất. Bước này cực kỳ quan trọng để tránh hiện tượng mô hình bị thiên kiến (bias) do học lặp đi lặp lại cùng một dữ liệu.

- *Kiểm soát chất lượng mẫu thông qua lọc thủ công:* Trong giai đoạn cuối, các bài đăng chứa thông tin rao bán theo lô hoặc "nhiều căn nhà trong một tin" đã được loại bỏ thủ công. Lý do là vì mục tiêu của đề án là dự đoán giá trị cụ thể của một căn nhà đơn lẻ. Các dữ liệu gộp này thường có đơn giá và diện tích gây nhiễu, không phản ánh đúng thực tế thị trường của phân khúc nhà riêng phổ thông, do đó việc loại bỏ chúng giúp tăng cường độ chính xác cho mô hình dự báo sau này.



Hình 2.2: Giao diện danh mục "Nhà riêng" khu vực Hà Nội trên trang batdongsan.com

(Nguồn: Trang batdongsan.com)

Bảng 2.1: Quá trình lọc bỏ dữ liệu URL

Bước thực hiện	Số lượng URL
Dữ liệu URL thô	529
Lọc bỏ các URL bị trùng	-50
Lọc bỏ URL chứa thông tin nhiều hơn một bất động sản (bán cả lô nhà thay vì chỉ một căn nhà)	-73
Lọc bỏ URL chứa thông tin bán đất nền	-355
Lọc bỏ URL không có thông tin về lịch sử giá nhà	-51
Dữ liệu URL còn lại (tương ứng với số bài đăng)	105

Sau khi truy cập vào từng bài đăng, nội dung HTML được phân tích cú pháp bằng BeautifulSoup để trích xuất các thông tin như mã tin, tên bài đăng, vị trí, giá cả, diện tích và các thông số khác. Các phần tử HTML chứa dữ liệu được định danh thông qua thẻ, lớp CSS hoặc ID, sau đó tổ chức thành cấu trúc từ điển để lưu trữ và xử lý.

Thông tin thu được từ mỗi URL sẽ được lưu dưới định dạng JSON vào các tệp riêng biệt. Mỗi 1 tệp tương ứng với 1 bản ghi duy nhất và được đặt tên theo thứ tự URL đã xử lý, ví dụ như property_data_1.json, property_data_2.json. Tất cả các tệp JSON được lưu vào một thư mục đầu ra được tạo sẵn, giúp quản lý dữ liệu được hệ thống, dễ dàng truy xuất cho những bước xử lý tiếp theo.

Kết quả thu về là 1 tập hợp các tệp JSON lưu trữ những thông tin cụ thể của giá nhà, bao gồm giá bán, diện tích, vị trí và các thông số khác, như minh họa ở Hình 2.3.

```
{
  "post_id": "44439349",
  "title": "Bán nhà riêng tại Đỗ Ngọc Du, 10,9 tỷ, 43m2, 3PN, 6WC, giá tốt, chính chủ ,thang máy 0528 533 ***",
  "location": "Đường Đỗ Ngọc Du Phường Đồng Nhân Hai Bà Trưng Hà Nội",
  "price_range": "10,9 tỷ",
  "price_per_square_meter": "~253,49 triệu/m²",
  "area": "43 m²",
  "floors": "6 tầng",
  "bedrooms": "3 phòng",
  "bathrooms": "6 phòng",
  "house_direction": "N/A",
  "balcony_direction": "N/A",
  "front_width": "N/A",
  "road_width": "N/A",
  "legal_status": "Sổ đỏ/ Sổ hồng",
  "description": "Nhà riêng tọa lạc tại Đường Đỗ Ngọc Du Phường Đồng Nhân Hai Bà Trưng Hà Nội với diện tích 43m² …",
  "price_change_over_time":
  [
    {
      "time": "T11/24",
      "price_per_square_meter_million_vnd": 210.0
    },
    …
    {
      "time": "T10/25",
      "price_per_square_meter_million_vnd": 265.0
    }
  ]
}
```

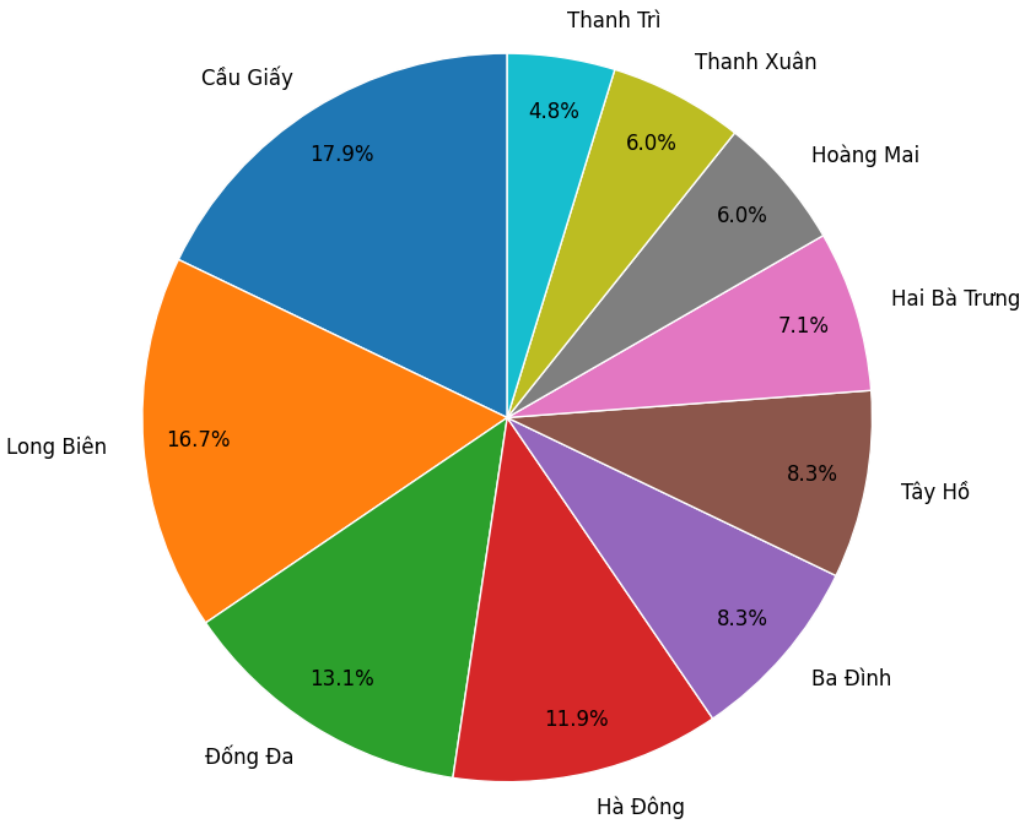

Hình 2.3: Dữ liệu JSON mẫu về thông tin thu thập được từ một mẫu bất động sản

Bảng 2.2: Mô tả các thuộc tính có trong dữ liệu JSON tương ứng với mỗi căn nhà

Tên trường	Kiểu dữ liệu	Mô tả
`post_id`	String	ID duy nhất của mỗi bài đăng (ví dụ: "44362440").
`title`	String	Tiêu đề của bài đăng (ví dụ: "Bán nhà riêng tại Đỗ Ngọc Du, 10,9 tỷ, 43m2, 3PN, 6WC...").
`location`	String	Địa chỉ bất động sản (ví dụ: "Đường Đỗ Ngọc Du_Phường Đồng Nhân_Hai Bà Trưng_Hà Nội").
`price_range`	String	Giá tổng thể của bất động sản (ví dụ: "10.9 tỷ").
`price_per_square_meter`	String	Giá mỗi mét vuông (ví dụ: "253.49 triệu/m²").
`area`	String	Diện tích bất động sản (ví dụ: "43 m²").
`floors`	String	Số tầng của nhà (nếu có). VD: 6 tầng
`bedrooms`	String	Số phòng ngủ (nếu có). VD: 3 phòng
`bathrooms`	String	Số phòng tắm (nếu có). VD: 6 phòng
`house_direction`	String	Hướng nhà (nếu có).
`balcony_direction`	String	Hướng ban công (nếu có).
`front_width`	String	Chiều rộng mặt tiền (nếu có).
`road_width`	String	Chiều rộng đường trước nhà (nếu có).
`legal_status`	String	Tình trạng pháp lý (ví dụ: "Sổ đỏ").
`description`	String	Mô tả chi tiết (bao gồm tiện ích, vị trí, tiềm năng đầu tư, v.v.).

Tên trường	Kiểu dữ liệu	Mô tả
`price_change_over_time`	List of dictionaries	Lịch sử thay đổi giá theo thời gian.
`time`	String	Thời điểm thay đổi giá (ví dụ:"T11/24", ...).
`price_per_square_meter_million_vnd`	Float	Giá mỗi mét vuông tại thời điểm đó (đơn vị: triệu đồng/m²).

Những dữ liệu trên có giá trị lớn trong việc thiết kế mô hình học máy để phân tích các chiều hướng biến động giá, dự đoán giá nhà, đồng thời nghiên cứu những yếu tố tác động đến giá nhà. Sau quy trình thu thập và làm sạch, ta thu được bộ dữ liệu với tỷ lệ tương ứng trong Hình 2.4:



Hình 2.4: Biểu đồ Top 10 quận/huyện (cũ) của Hà Nội có số lượng tin đăng mua bán nhà riêng cao nhất giai đoạn từ tháng 11/2024 đến tháng 10/2025

Biểu đồ tại Hình 2.4 thể hiện top 10 quận/huyện trong tổng số 24 quận/huyện (cũ) của Hà Nội có số lượng tin đăng mua bán nhà riêng cao nhất trong khoảng từ tháng 11/2024 đến tháng 10/2025. Đây chính là các nơi có giao dịch bất động sản sôi động nhất, phản ánh xu hướng tìm kiếm và đầu tư nhà ở trong thời điểm khảo sát.

Quận Cầu Giấy dẫn đầu với tỷ trọng 17.9%, tiếp theo là quận Long Biên với 16.7% và quận Đống Đa chiếm 13.1% , cho thấy sự lan tỏa mạnh mẽ của nhu cầu nhà ở ra khu vực phía Bắc của thành phố.

Tiền xử lý dữ liệu là thực hiện chuẩn hóa dữ liệu và làm sạch từng trường trong tập dữ liệu đảm bảo tính nhất quán cũng như khả năng khai thác trong phân tích hoặc huấn luyện những mô hình học máy. Bảng 2.3 mô tả dữ liệu trước và sau tiền xử lý.

Bảng 2.3: Mô tả quá trình tiền xử lý file JSON

Trường	Trước	Sau
post_id	"44444351"	Giữ nguyên
title	"Bán nhà riêng tại Đỗ Ngọc Du, 10,9 tỷ, 43m2, 3PN, 6WC,."	Giữ nguyên
location	"Đường Đỗ Ngọc Du, Phường Đồng Nhân, Hai Bà Trưng, Hà Nội"	Giữ nguyên
price_range	"10.9 tỷ"	1090000000
price_per_square_meter	"253.49 triệu/m²"	2534900000
area	"43 m²"	43
floors	"N/A"	""
bedrooms	"N/A"	""

Trường	Trước	Sau
bathrooms	"N/A"	""
legal_status	"Sổ đỏ"	Giữ nguyên
description	"Nhà riêng tọa lạc tại Đỗ Ngọc Du, Phường Đồng Nhân, Hai Bà Trưng, Hà Nội, với diện tích 43m² ..."	Giữ nguyên
price_change_over_time	[{"time": "T11/24", "price_per_square_meter_million_vnd": 210}, ...]	Giữ nguyên

Các trường liên quan đến diện tích (area) và giá cả (price_range, price_per_square_meter) thường được biểu diễn dưới dạng chuỗi văn bản với đơn vị đo lường không đồng nhất. Ví dụ: diện tích "43 m²" được chuẩn hóa thành “43” bằng cách loại bỏ đơn vị đo lường và chuyển dấu phẩy thành dấu chấm. Giá bán tổng (price_range) từ "10.9 tỷ" được chuyển đổi thành 10900000000 bằng cách nhân với hệ số 1 tỷ. Tương tự, giá/m² (price_per_square_meter) từ " triệu/m²" được chuyển đổi thành 86000000. Những thay đổi này giúp thống nhất đơn vị đo lường và dễ dàng thực hiện các phép tính toán sau này.

Đối với các trường không có thông tin, ví dụ như số tầng (floors), số phòng ngủ (bedrooms), và số phòng tắm (bathrooms) thường chứa giá trị "N/A". Trong quá trình tiền xử lý, các giá trị này phải chuyển đổi thành chuỗi rỗng ("") để tránh gây nhầm lẫn khi phân tích dữ liệu.

Lịch sử biến động giá (price_change_over_time) được giữ nguyên cấu trúc nhưng giá trị giá/m² tại mỗi thời điểm được chuẩn hóa về đơn vị cơ bản là đồng Việt Nam. Ví dụ: giá trị 253.49 triệu/m² được chuyển đổi thành 2534900000.

Các trường như post_id, title, location, và legal_status không yêu cầu thay đổi vì chúng đã ở định dạng phù hợp và không bị tác động trong quá trình phân tích.

2.3. Xây dựng tập dữ liệu chuỗi thời gian

Quá trình hình thành tập dữ liệu kiểu chuỗi thời gian trong đoạn mã trên được thực hiện bằng cách tổ chức và chuẩn hóa thông tin giá nhà theo các mốc thời gian cụ thể. Dữ liệu gốc đưa vào lưu dưới dạng các tệp JSON và trong đó mỗi 1 tệp chứa lịch sử thay đổi giá của một bài đăng. Mỗi bản ghi trong danh sách price_change_over_time bao gồm 2 trường chính: thời gian (time) và giá/m² (price_per_square_meter_million_vnd). Giá/m² ở đây đại diện cho giá bán phổ biến nhất theo mét vuông của những căn nhà trong từng khu vực cụ thể. Việc sử dụng giá trị này giúp phản ánh sát nhất mức giá của thị trường tại mỗi thời điểm và mỗi khu vực.

Để tính thống nhất cũng như thuận lợi cho việc xử lý, các chuỗi thời gian ban đầu (như "Q1/25" hoặc "T3/25") cần được chuyển đổi thành định dạng chuẩn YYYY-MM-DD. Quy trình xử lý thời gian bao gồm các bước sau:

Bảng 2.4: Chuyển đổi định dạng dữ liệu thời gian

Định dạng thời gian	Ví dụ	Chuyển đổi	Kết quả
Quý	"Q1/25"	Ngày đầu tiên của quý	"2025-01-01"
Tháng	"T3/25"	Ngày đầu tiên của tháng	"2025-03-01"

Việc chuyển đổi này giúp thống nhất định dạng thời gian, đảm bảo được dữ liệu sắp xếp và truy vấn theo thứ tự thời gian một cách chính xác. Sau khi chuẩn hóa, cột time trong DataFrame được chuyển đổi sang kiểu datetime, cho phép sắp xếp và phân tích dữ liệu đạt hiệu quả hơn.

Mỗi bản ghi trong chuỗi thời gian được tổ chức thành ba trường chính: mã bài đăng (post_id), thời gian (time), và giá/m² (price_per_square_meter). Mã bài đăng

đóng vai trò là định danh duy nhất để nhóm các thay đổi giá thuộc cùng một bất động sản. Thời gian đã được chuẩn hóa về định dạng YYYY-MM-DD, và giá/m² phản ánh giá bán phổ biến nhất theo mét vuông trong khu vực cụ thể mà bất động sản thuộc về. Điều này giúp dữ liệu không chỉ mang tính đại diện cho từng bài đăng cụ thể mà còn phản ánh xu hướng giá chung của khu vực.

Sau đó, dữ liệu được sắp xếp tăng dần của post_id và time, đảm bảo rằng các thay đổi giá của cùng một bất động sản được xuất hiện theo trình tự thời gian. Ngoài ra, để giảm thiểu sự chênh lệch về giá giữa những sản phẩm nhà đất khác nhau, giá/m² được chuẩn hóa bằng phương pháp Min-Max Scaling. Phương pháp này ánh xạ tất cả những giá trị vào khoảng từ 0 đến 1, giúp dễ dàng so sánh và phân tích được xu hướng biến động giá (tăng/giảm) theo thời gian, công thức (2.1):

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2.1)$$

Trong công thức đó: X là giá trị dữ liệu gốc; X_{max} là giá trị dữ liệu lớn nhất; X_{min} là giá trị dữ liệu nhỏ nhất; X_{scaled} là giá trị đã được chuẩn hóa.

Quá trình chuẩn hóa giúp hạn chế tác động của những giá trị bất thường (outliers) và làm cho cơ sở dữ liệu trở nên đồng nhất hơn, đặc biệt trong trường hợp các thuộc tính có vùng giá trị khác nhau.

Sau khi xử lý, bộ dữ liệu được chia thành 2 phần: 80% dành cho huấn luyện, 20% dành để kiểm tra, nhằm đảm bảo khách quan đối với việc đánh giá kết quả hoạt động của mô hình.

2.4. Thiết kế kiến trúc mô hình LSTM

Kiến trúc mô hình được tối ưu hóa theo hướng tinh gọn nhưng đảm bảo hiệu suất dự báo cao, tập trung khai thác đặc tính chuỗi thời gian của dữ liệu, cấu trúc chi tiết bao gồm:

- *Lớp mạng hồi quy LSTM (Input & Hidden Layer)*: (1) Cơ chế tiếp nhận dữ liệu: Mô hình sử dụng cửa sổ trượt với độ dài chuỗi (sequence_length) bằng 6. Điều

này có nghĩa là tại mỗi thời điểm dự báo, mạng sẽ phân tích dữ liệu biến động giá của 6 tháng liên tiếp trước đó để tìm ra quy luật và đưa ra ước tính cho tháng kế tiếp. (2) Cấu trúc đầu vào (Input Shape): Dữ liệu đầu vào được định hình dưới dạng tensor (6, 1), tương ứng với 6 bước thời gian và 1 đặc trưng duy nhất là chỉ số giá đã qua chuẩn hóa (price_scaled). Việc chuẩn hóa này giúp tăng tốc độ hội tụ và tránh hiện tượng gradient bị bùng nổ trong quá trình huấn luyện. (3) Xử lý phi tuyến: Lớp LSTM được cấu hình với 50 đơn vị (units), đóng vai trò là "bộ nhớ dài hạn" để trích xuất các đặc trưng phức tạp. Việc tích hợp hàm kích hoạt ReLU (Rectified Linear Unit) giúp mô hình phá vỡ tính tuyến tính đơn thuần, từ đó nhận diện được các mẫu biến động bất thường hoặc các xu hướng tiềm ẩn trong chuỗi dữ liệu bất động sản vốn có độ nhiễu cao.

- *Lớp kết nối đầy đủ (Output Dense Layer):* (1) Chức năng dự báo: Sau khi thông tin được tổng hợp qua lớp LSTM, một lớp Dense (Fully Connected) với một nơ-ron duy nhất được sử dụng để chuyển đổi các đặc trưng đã học thành một giá trị thực cụ thể. (2) Đầu ra liên tục: Lớp này không sử dụng hàm kích hoạt phi tuyến (hoặc có thể hiểu là sử dụng hàm kích hoạt tuyến tính ngầm định) nhằm cho phép mô hình dự báo chính xác các giá trị số thực liên tục. Kết quả đầu ra chính là mức giá dự kiến của bất động sản trong chu kỳ thời gian tiếp theo, đảm bảo tính trực quan và khả năng ứng dụng thực tế cao.

Mô hình được biên dịch với hàm tối ưu hóa Adam - một giải pháp tối ưu hóa thích nghi dựa trên các ước lượng mô men bậc nhất và mô men bậc hai của Gradient, cho tốc độ hội tụ nhanh, ổn định. MSE là hàm mất mát được chọn là rất phù hợp đối với bài toán hồi quy và đo chênh lệch bình phương giữa kết quả dự đoán với thực tế. Mô hình sau huấn luyện có thể học được mối liên hệ động giữa các giá trị trong quá khứ để dự đoán tương lai, hỗ trợ ra những quyết định đầu tư hoặc hoạch định về bất động sản.

Bảng 2.5: Mô tả các tham số của mô hình LSTM và chi tiết tham số từng lớp

Thành phần	Giá trị /Cấu hình
Kiến trúc mô hình	Sequential
Lớp đầu vào	Shape: `(6, 1)`
Lớp LSTM	50 unit, activation: `relu`
Lớp đầu ra	1 unit, activation: `linear`
Hàm tối ưu hóa	Adam
Hàm mất mát	MSE (Mean Squared Error)
Tổng số tham số	10,451
Tham số có thể huấn luyện (Trainable parameters)	10,451

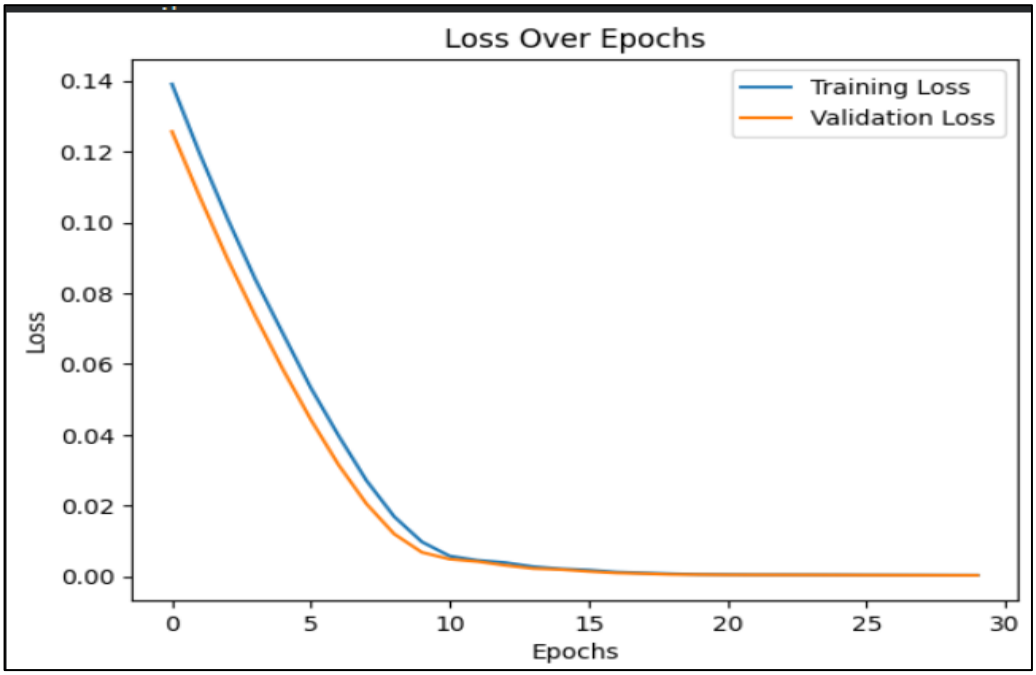
Bảng 2.6: Tham số cấu hình của mạng LSTM

Lớp (Loại)	Output Shape	Số tham số
lstm_1 (LSTM)	(None, 50)	10,400
dense_1 (Dense)	(None, 1)	51
Tổng cộng		10,451

2.5. Huấn luyện và đánh giá mô hình

Biểu đồ Loss Over Epochs tại Hình 2.5 cung cấp các thông tin quan trọng trong quá trình huấn luyện mô hình LSTM, trong đó hàm mất mát được dùng là MSE. MSE chính là lựa chọn phổ biến, rất phù hợp đối với bài toán hồi quy, đặc biệt khi mục tiêu là dự đoán 1 giá trị liên tục, như trong trường hợp này biến động

giá theo thời gian của giá nhà. Việc lựa chọn MSE xuất phát từ đặc tính phạt mạnh các sai số lớn nhờ việc bình phương khoảng cách giữa giá trị thực tế và kết quả dự đoán, từ đó thúc đẩy mô hình tập trung giảm thiểu những lỗi nghiêm trọng nhất.



Hình 2.5: Biểu đồ Loss Over Epochs

Trong thời gian đầu huấn luyện thì cả Training Loss và Validation Loss đều giảm mạnh, chứng tỏ mô hình học nhanh từ bộ dữ liệu, đồng thời nắm bắt được những mẫu phức tạp trong chuỗi thời gian. Sau khoảng 10 epoch, cả hai đường loss dần ổn định ở mức rất thấp và không có sự dao động đáng kể, điều này chứng tỏ mô hình đã đạt trạng thái hội tụ. Kết quả này khẳng định mô hình không những học tốt mà còn duy trì được tính ổn định trong suốt quá trình huấn luyện.

Để đánh giá chất lượng và mức độ hiệu quả mô hình, có thể sử dụng một số thước đo sau:

- Sai số trung bình tuyệt đối (MAE)

MEA [12] thước đo độ chính xác của mô hình dự đoán, hoặc phép đo và được tính bằng trung bình cộng của toàn bộ các sai số tuyệt đối giữa giá trị thực tế với giá trị dự đoán, công thức (2.2):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2.2)$$

Trong công thức đó thì *n* là số lượng mẫu, y_i là giá trị dữ liệu thực tế và $\hat{y}_i$ là giá trị dự đoán

Ưu điểm độ đo MAE là dễ hiểu và dễ diễn giải, không bị tác ảnh hưởng bởi các giá trị ngoại lai lớn do không bình phương sai số. MAE cung cấp 1 thước đo trực quan về mức độ chuẩn xác của mô hình dự báo.

Độ đo MAE thường được sử dụng trong những mô hình hồi quy và dự báo để đánh giá được độ chuẩn xác của dự đoán.

Ý nghĩa của MAE là thể hiện mức độ chuẩn xác của các phép đo, thể hiện mức chênh lệch giữa các lần đo. Nó giúp đánh giá độ tin cậy của các kết quả thực nghiệm; nếu số lượng lần đo càng nhiều thì giá trị trung bình càng chính xác, MAE càng nhỏ thì kết quả đo lường tiệm cận với giá trị thực.

- Sai số bình phương trung bình (MSE) và căn bậc hai của nó (RMSE)

+ MSE [12] là giá trị trung bình cộng của bình phương sai số giữa các giá trị thực tế và kết quả dự đoán, công thức (2.3):

$$MSE = \sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n} \quad (2.3)$$

Trong công thức đó thì *n* là số lượng mẫu, y_i là giá trị dữ liệu thực tế và $\hat{y}_i$ là giá trị dự đoán.

+ RMSE dùng để đưa sai số về cùng đơn vị với bộ dữ liệu gốc, được tính theo công thức (2.4):

$$\text{RMSE} = \sqrt{\text{MSE}} \quad (2.4)$$

Ý nghĩa: MSE làm nổi bật những sai số lớn do việc bình phương khiến chúng có trọng số cao hơn. Trong khi đó, RMSE dễ diễn giải hơn vì nó có cùng đơn vị với

bộ dữ liệu gốc, thường dùng để đánh giá mức độ chuẩn xác của mô hình. Với cả hai thước này, các giá trị này càng bé thì mô hình càng chính xác cao.

- Hệ số xác định R-squared (R^2)

R^2 [13] là một chỉ số mô tả mức độ mà một mô hình hồi quy tuyến tính có thể thể giải thích sự biến động của các biến phụ thuộc trên cơ sở các biến độc lập. R^2 chạy từ 0 đến 1, khi R^2 càng tiến gần đến 1 thì mô hình càng giải thích tốt hơn đối với biến động trong tập dữ liệu, công thức (2.5):

$$R^2 = 1 - \frac{ESS}{TSS} \quad (2.5)$$

Trong công thức này, ESS chính là tổng bình phương sai số dư; TSS chính là tổng bình phương sai số của toàn bộ tập mẫu.

Với công thức trên, tỷ lệ % biến động của các biến phụ thuộc được giải thích bởi những biến độc lập, trong mô hình hồi quy.

Ý nghĩa của $R^{2:}$

Nếu R^2 càng cao thì mô hình có khả năng giải thích tốt hơn, ngược lại nếu R^2 thấp, mô hình cần được cải thiện hoặc có thể không phù hợp so với tập dữ liệu.

R^2 dùng để so sánh mô hình, nếu mô hình có R^2 cao hơn thường dự đoán biến phụ thuộc chính xác hơn.

2.6. Nhận xét, đánh giá chung

Mô hình được xây dựng và thực nghiệm trên dữ liệu giá nhà đất tại Hà Nội từ tháng 11/2024 đến tháng 10/2025. Các khu vực được chọn bao gồm cả quận trung tâm và vùng ven, nhằm kiểm tra tính khái quát của mô hình. Kết quả thực nghiệm đã chứng minh mô hình LSTM có khả năng dự báo chính xác các xu hướng tăng/giảm giá theo thời gian, nhất là tại những vùng có tập dữ liệu ổn định và đầy đủ.

Việc ứng dụng mô hình LSTM trong dự báo giá nhà ở Việt Nam có thể mang lại rất nhiều lợi ích. Mô hình này có khả năng học những phụ thuộc dài hạn, xử lý dữ liệu phi tuyến và đa chiều, đồng thời nắm bắt chuẩn xác những xu hướng thị trường.

Mặc dù vậy, mô hình vẫn còn những yếu điểm nhất định như huấn luyện mô hình cần số lượng dữ liệu lớn, chất lượng cao và nguồn tài nguyên tính toán mạnh. Ngoài ra, mô hình mang tính "hộp đen", khiến việc giải thích kết quả cho người dùng cuối trở nên khó khăn. Hiệu quả của mô hình cũng chịu tác động đáng kể từ quá trình tiền xử lý, lựa chọn các đặc trưng.

Tổng kết, mô hình học sâu LSTM là một hướng tiếp cận tiềm năng trong dự báo giá nhà tại Việt Nam, với khả năng mở rộng và ứng dụng thực tiễn cao. Việc tiếp tục cải tiến mô hình, bổ sung thêm dữ liệu về kinh tế xã hội và phát triển các kỹ thuật giải thích mô hình sẽ là hướng đi quan trọng cho các nghiên cứu sau.

Chương 3. KẾT QUẢ THỰC NGHIỆM VÀ LUẬN BÀN

Chương 3 tập trung trình bày kết quả sau thực nghiệm của mô hình LSTM trong dự đoán giá nhà tại Việt Nam. Nội dung chương bao gồm kết quả triển khai mô hình và đánh giá hiệu quả thông qua các thước đo chuẩn, đồng thời so sánh với những mô hình truyền thống như ARIMA và hồi quy tuyến tính. Qua đó, làm rõ hiệu quả của mô hình LSTM trong bài toán dự báo giá nhà theo thời gian, cũng như phân tích hiệu quả, ý nghĩa thực tiễn, hạn chế và tiềm năng mở rộng.

3.1. Kết quả thực nghiệm

Sau khi hoàn thành xong toàn bộ quá trình thiết kế và huấn luyện mô hình LSTM trên tập dữ liệu giá nhà thực tế tại Việt Nam, mô hình đã được tiến hành thử nghiệm tại Hà Nội trong khoảng từ tháng 11/2024 đến tháng 10/2025. Dữ liệu đã được chuẩn hóa, chuyển đổi thành chuỗi thời gian theo từng quận, huyện (cũ) với mỗi điểm dữ liệu đại diện cho giá nhà trung bình theo tháng.

Bảng 3.1: Kết quả của mô hình LSTM đánh giá trên bộ dữ liệu kiểm thử (Testing dataset)

Độ đo	Kết quả
MAE	9.18 (triệu/m²)
MSE	161.11
RMSE	12.69
Coefficient of Determination (R^2)	0.9861

Kết quả thử nghiệm mô hình cho thấy hiệu suất rất cao: Với hệ số xác định R^2 đạt 0.9861, điều này cho thấy mô hình giải thích được gần 99% sự biến động của giá nhà, minh chứng cho độ tin cậy cực kỳ cao của thuật toán LSTM đối với bộ dữ liệu này. Với các giá trị độ đo MAE (sai số tuyệt đối trung bình) 9.18 (triệu/m2), MSE

(sai số bình phương trung bình) 161.11 và RMSE (căn bậc hai sai số bình phương trung bình) 12.69 là nhỏ, càng chứng tỏ mô hình có hiệu suất cao.

So sánh hiệu hiệu quả của các mô hình

Để xác định được hiệu quả của mô hình LSTM, nghiên cứu đã so sánh với những mô hình như ARIMA và hồi quy tuyến tính. Tất cả những mô hình này đều được huấn luyện trên cùng 1 bộ dữ liệu và cùng được đánh giá bằng các chỉ số tương tự. Chi tiết tham số huấn luyện có trong Bảng 3.2.

Bảng 3.2: Tham số huấn luyện của các mô hình được so sánh

Mô hình	**Tham số huấn luyện**	**Giải thích**
ARIMA	p = 1	Số lượng độ trễ (lags) được sử dụng trong AutoRegressive.
	d = 1	Số lần sai phân (differencing) để làm chuỗi dừng (stationary).
	q = 0	Số lượng trung bình trượt (moving average - MA) được sử dụng trong thành phần nhiễu.
Linear Regression	Không có tham số huấn luyện	

Kết quả tại Bảng 3.3 đánh giá hiệu quả của các mô hình LSTM, Linear Regression, ARIMA đối với bài toán dự đoán giá nhà theo chuỗi thời gian cho thấy sự khác biệt rõ ràng giữa các công cụ.

Bảng 3.3: So sánh hiệu suất của các mô hình

Mô hình	MAE	MSE	RMSE	R^2
Linear Regression	7.72	113.55	10.66	0.9902
LSTM	9.18	161.11	12.69	0.9861
ARIMA	117.14	28064.14	117.14	0.15

Mô hình Linear Regression đạt hiệu suất rất tốt với MAE = 7.72, MSE = 113.55, RMSE = 10.66 và R^2 = 0.9902. Kết quả này chứng tỏ rằng mô hình tuyến tính có khả năng giải thích gần như toàn bộ biến động trong dữ liệu. Tuy nhiên, Linear Regression không thể xử lý các phụ thuộc phi tuyến hoặc xu hướng thời gian phức tạp, khiến nó hạn chế trong việc nắm bắt những biến động đột ngột trong chuỗi dữ liệu.

LSTM cũng đã đạt hiệu suất ổn định với R^2 = 0.9861, nhưng sai số tuyệt đối (MAE = 9.18, MSE = 161.11, RMSE = 12.69) cao hơn so với Linear Regression. Điều này có thể do LSTM cố gắng học các mẫu phi tuyến và phụ thuộc thời gian phức tạp hơn - điều mà mô hình tuyến tính sẽ bỏ qua. Đối với trường hợp dự đoán giá nhà theo chuỗi thời gian, LSTM có ưu thế đặc biệt nhờ khả năng nắm bắt cả xu hướng dài hạn lẫn ngắn hạn, đồng thời nhận diện được những biến động bất thường trong dữ liệu.

Ngược lại, mô hình ARIMA hoàn toàn thất bại với R^2 = 0.15, MAE = 117.14, MSE = 28064.14 và RMSE = 117.14. Kết quả cho thấy ARIMA không phù hợp cho dự đoán giá nhà theo chuỗi thời gian.

Tóm lại, Linear Regression là lựa chọn hiệu quả nếu dữ liệu có cấu trúc tuyến tính rõ ràng, trong khi LSTM phù hợp hơn khi cần xử lý các phụ thuộc phi tuyến và xu hướng thời gian phức tạp. ARIMA nên được loại bỏ đối với bài toán này.

3.2. Đánh giá kết quả

3.2.1. Hiệu quả mô hình

Mô hình LSTM đã chứng minh được khả năng học các phụ thuộc dài hạn trong chuỗi thời gian, giải quyết hiệu quả đối với dữ liệu phi tuyến, đa chiều, đồng thời đáp ứng tốt với biến động thị trường. Việc đạt được sai số thấp và hệ số tương quan cao cho thấy mô hình hoàn toàn có thể dùng như một công cụ hỗ trợ để ra quyết định trong lĩnh vực bất động sản.

3.2.2. Ý nghĩa thực tiễn

Hiện nay, thị trường bất động sản tại Việt Nam đang từng bước chuyển đổi số, việc áp dụng mô hình LSTM mang đến nhiều giá thực tiễn. Đối với nhà đầu tư, mô hình có khả năng giúp hỗ trợ dự đoán chính xác hơn xu thế biến động giá bất động sản để đưa ra các quyết định đầu đúng đắn.

Với người mua nhà, đặc biệt là lần đầu tiếp cận thị trường, mô hình cung cấp công cụ hỗ trợ đáng tin cậy để đánh giá mức giá hợp lý của giá nhà tại một số khu vực cụ thể. Qua đó, mô hình có khả năng hỗ trợ người mua lựa chọn căn nhà phù hợp so với tầm tài chính và nhu cầu cá nhân.

Đối với cơ quan quản lý, mô hình có vai trò như 1 công cụ hữu ích trong theo dõi sự xu hướng trong thị trường bất động sản. Từ kết quả dự đoán giúp đánh giá hiệu quả các chính sách hiện hành, kịp thời điều chỉnh cho phù hợp. Ngoài ra, dữ liệu từ mô hình cũng hỗ trợ quá trình hoạch định quy hoạch đô thị, giúp phân bổ nguồn lực hợp lý, đồng thời đáp ứng tốt hơn nhu cầu người dân trong thời gian tới.

3.2.3. Hạn chế

Mô hình đã đạt được những kết quả đáng ghi nhận, tuy nhiên mô hình vẫn có một số hạn chế cần khắc phục. Những hạn chế này xuất phát từ cả chất lượng của dữ liệu đưa vào, phương pháp dự đoán cũng như khả năng giải thích kết quả.

Nguồn dữ liệu thu thập được còn khá giới hạn cả về phạm vi địa lý lẫn loại hình bất động sản. Hiện tại, mô hình chỉ tập trung vào nhà riêng tại khu vực Hà Nội,

trong khi chưa mở rộng sang những sản phẩm bất động sản khác, ví dụ như căn chung cư, biệt thự hay đất nền. Hơn nữa, sự phân bố dữ liệu không đồng đều và có sự thiên lệch rõ rệt về một số quận huyện cụ thể, dẫn đến nguy cơ tạo ra kết quả dự đoán thiếu tính đại diện. Bên cạnh đó, quy mô bộ dữ liệu còn khá khiêm tốn với chỉ hơn 100 mẫu. Với số lượng mẫu ít, mô hình có thể gặp khó khăn trong việc học các quy luật phức tạp hoặc đảm bảo độ tin cậy khi ứng dụng vào thực tế.

Về phương pháp, mô hình dự đoán được xây dựng chủ yếu dựa vào giá trị lịch sử (như dữ liệu giá từ các tháng trước) làm yếu tố đầu vào duy nhất để đưa ra dự báo. Tuy nhiên, còn nhiều yếu tố ảnh hưởng không nhỏ đến giá trị nhà đất nhưng chưa được tận dụng do thiếu dữ liệu. Ví dụ, một căn nhà nằm ở quận Hoàn Kiếm (gần trung tâm Hà Nội) thường có giá trị cao hơn đáng kể so với một căn nhà cùng diện tích ở vùng ven Hà Nội, chẳng hạn như huyện Ba Vì. Hay như một căn nhà gần trường học, bệnh viện hoặc công viên sẽ hấp dẫn hơn và có giá trị lớn hơn so với một căn nhà không có các tiện ích này. Tương tự, những yếu tố kinh tế vĩ mô như lãi suất ngân hàng tăng cao có thể khiến người mua hạn chế vay vốn, hoặc chính sách siết chặt cấp giấy phép xây dựng có thể làm giảm nguồn cung trên thị trường, qua đó tác động mạnh đến giá cả. Tuy nhiên, những thông tin về vị trí chi tiết, hạ tầng xung quanh hay biến động kinh tế vĩ mô thường không có đầy đủ trong dữ liệu hiện có. Điều này khiến mô hình không thể khai thác triệt để các yếu tố quan trọng để đưa ra kết quả dự báo chính xác. Như vậy, hạn chế không bắt nguồn từ năng lực của mô hình mà xuất phát từ việc dữ liệu đầu vào còn thiếu sót và chưa toàn diện.

Một hạn chế quan trọng khác là bản chất "hộp đen" của mô hình. Bản chất "hộp đen" của mô hình LSTM (Long Short-Term Memory) bắt nguồn từ cấu trúc, cơ chế hoạt động phức tạp của mạng nơ-ron sâu, khiến việc giải thích chi tiết quá trình xử lý và đưa ra quyết định gặp khó khăn.

3.2.4. Khả năng mở rộng

Để mở rộng khả năng áp dụng vào thực tế, mô hình có thể phát triển theo một số hướng sau:

1. Mở rộng phạm vi và quy mô dữ liệu

Hiện tại, dữ liệu chủ yếu tập trung vào nhà riêng tại khu vực Hà Nội, với số lượng mẫu hạn chế. Trong tương lai, phạm vi thu thập dữ liệu có thể mở rộng sang những loại bất động sản khác như: biệt thự, căn chung cư, đất nền, hoặc thậm chí là các loại tài sản thương mại. Đồng thời, việc mở rộng địa lý ra các tỉnh thành khác sẽ giúp mô hình trở nên đa dạng và có tính đại diện cao hơn. Ngoài ra, việc gia tăng số lượng mẫu dữ liệu bằng các phương pháp thu thập tự động từ nhiều trang bất động sản khác nhau sẽ giúp nâng cao độ chính xác và độ tin cậy của mô hình.

2. Tận dụng tối đa các yếu tố đầu vào

Mặc dù thông tin ban đầu về bất động sản cung cấp một số yếu tố khách quan trọng như vị trí địa lý, hệ thống tiện ích xung quanh, diện tích, số phòng ngủ, dữ liệu này vẫn chưa được khai thác hiệu quả do tình trạng thiếu hụt và không chuẩn hóa. Một số bài đăng chỉ cung cấp thông tin cơ bản như diện tích và vị trí, trong khi các thông tin chi tiết khác như số phòng ngủ, số phòng vệ sinh, hướng nhà hay diện tích mặt đường thường bị bỏ sót hoặc không đầy đủ. Điều này khiến mô hình có thể gặp khó khăn trong việc hiểu được cấu trúc và giá trị thực tế của thị trường. Bên cạnh đó, thông tin về vị trí địa lý thường mang tính quảng cáo, tập trung vào các yếu tố tích cực như "gần bệnh viện, trường học X" mà không đề cập đến những yếu tố tiêu cực có thể ảnh hưởng đến giá trị tài sản, ví dụ như khu vực dính quy hoạch, tranh chấp pháp lý hoặc gần các nguồn gây ô nhiễm. Việc thiếu thông tin tiêu cực này làm giảm khả năng mô hình nắm bắt bức tranh toàn diện của thị trường bất động sản. Ngoài ra, dữ liệu thu thập được từ các bài đăng thường không được chuẩn hóa, có thể dẫn đến sự không đồng nhất trong cách biểu diễn dữ liệu. Ví dụ, cách ghi chú về tiện ích xung quanh có thể khác nhau giữa các bài đăng, thậm chí có bài chỉ nêu chung chung, không có chi tiết cụ thể. Những vấn đề này không chỉ cản trở, làm khó quá trình tiền xử lý dữ liệu mà còn tác động đáng kể đến chất lượng của mô hình dự đoán.

Bên cạnh việc khai thác dữ liệu về vị trí căn nhà, việc tích hợp những yếu tố như lãi suất ngân hàng, tốc độ tăng trưởng GDP hay các chính sách nhà đất mới ban hành sẽ giúp mô hình nắm bắt sát hơn biến động của thị trường.

3. Cải thiện khả năng giải thích của mô hình

Do bản chất "hộp đen" của LSTM, việc giải thích kết quả dự đoán hiện tại còn gặp khá nhiều khó khăn. Vì vậy, để giải quyết điều này, có thể ứng dụng các kỹ thuật hỗ trợ giải thích như SHAP (SHapley Additive exPlanations) hoặc phân tích độ nhạy (sensitivity analysis). Các công cụ này có thể giúp làm rõ các mức độ ảnh hưởng của mỗi yếu tố đầu vào đến kết quả cuối cùng, qua đó tăng cường niềm tin của người dùng vào mô hình.

4. Phát triển ứng dụng thực tế và tích hợp hệ thống

Mô hình có thể được tích hợp vào những nền tảng trực tuyến, hoặc các ứng dụng di động hỗ trợ người dùng trong ra quyết định liên quan đến bất động sản. Ví dụ như xây dựng một hệ thống gợi ý giá nhà dựa trên các tiêu chí đầu vào mà người dùng cung cấp hoặc phát triển công cụ để phân tích xu hướng thị trường để phục vụ nhà đầu tư và các cơ quan quản lý.

Chương 3 đã thể hiện toàn diện, đầy đủ về kết quả thực nghiệm mô hình học sâu LSTM, từ hiệu suất dự đoán đến so sánh với những mô hình truyền thống. Kết quả cho thấy, LSTM có thể học và khai thác các mối quan hệ phi tuyến, xu hướng dài hạn cũng như những biến động phức tạp trong tập dữ liệu. Mặc dù vậy, mô hình vẫn còn những hạn chế có liên quan đến dữ liệu, yếu tố đầu vào và khả năng giải thích. Tuy nhiên, với tiềm năng mở rộng và ứng dụng thực tế trong lĩnh vực bất động sản, LSTM hứa hẹn trở thành công cụ hữu ích hỗ trợ nhà đầu tư, người mua nhà và các cơ quan quản lý nhà nước.

KẾT LUẬN

Kết quả của nghiên cứu

Trong đề tài **"*Áp dụng mô hình Long Short-Term Memory trong dự đoán giá nhà tại Việt Nam*"**, mô hình LSTM đã được nghiên cứu, thiết kế và triển khai nhằm giải quyết bài toán dự đoán giá nhà tại Việt Nam theo chuỗi thời gian. Thông qua quá trình xây dựng kiến trúc mô hình, xử lý dữ liệu thực tế và thực nghiệm tại khu vực Hà Nội, nghiên cứu đã chứng minh rằng LSTM là một công cụ hiệu quả trong việc khai thác các mối quan hệ dài hạn và phi tuyến trong dữ liệu bất động sản. Kết quả thực nghiệm cho thấy mô hình LSTM có khả năng dự đoán chính xác xu hướng giá nhà, phản ứng tốt với các biến động thị trường và vượt trội so với các mô hình truyền thống như hồi quy tuyến tính và ARIMA. Mô hình không chỉ học được các đặc điểm nội tại của chuỗi thời gian mà còn có thể tích hợp các đặc trưng bổ sung như vị trí địa lý, loại hình bất động sản và các yếu tố kinh tế xã hội.

Tuy nhiên, nghiên cứu cũng nhận diện một số thách thức trong việc áp dụng LSTM vào thực tế, bao gồm yêu cầu về dữ liệu lớn và chất lượng cao, thời gian huấn luyện dài và tính khó giải thích của mô hình. Những hạn chế này đặt ra yêu cầu tiếp tục cải tiến mô hình và mở rộng hướng nghiên cứu trong tương lai.

Kiến nghị đề xuất

Dựa trên kết quả nghiên cứu và thực nghiệm, một số kiến nghị được đề xuất nhằm nâng cao hiệu quả ứng dụng mô hình LSTM trong dự đoán giá nhà tại Việt Nam trong tương lai như sau:

Về dữ liệu

Cần mở rộng phạm vi thu thập để bổ sung thêm các loại hình bất động sản khác ngoài nhà riêng, chẳng hạn như căn hộ chung cư, biệt thự, đất nền hoặc tài sản thương mại. Hiện tại, dữ liệu chủ yếu tập trung ở khu vực Hà Nội, dẫn đến sự thiên lệch về mặt địa lý. Việc mở rộng phạm vi địa lý ra các tỉnh thành khác sẽ giúp mô hình trở nên đại diện và chính xác hơn. Đồng thời, mặc dù đã có sẵn các yếu tố như

vị trí địa lý, diện tích, số phòng, tiện ích xung quanh và các yếu tố kinh tế vĩ mô, nhưng chúng chưa được khai thác triệt để. Do đó, cần phân tích sâu hơn để tích hợp hiệu quả các yếu tố này vào mô hình, chẳng hạn bằng cách áp dụng kỹ thuật tạo biến tương tác (interaction features) giữa các yếu tố để tìm ra mối quan hệ tiềm ẩn. Ngoài ra, thị trường bất động sản luôn biến động, vì vậy cần xây dựng cơ chế tự động cập nhật dữ liệu thường xuyên thông qua các phương pháp như web scraping hoặc hợp tác với các nền tảng bất động sản để lấy dữ liệu giao dịch gần đây.

Về mô hình

Nên thử nghiệm các mô hình hybrid kết hợp LSTM với các thuật toán học máy khác như XGBoost, Random Forest hoặc Gradient Boosting để tận dụng ưu điểm của từng phương pháp. Điều này không chỉ giúp cải thiện độ chính xác mà còn tăng cường khả năng giải thích kết quả, đặc biệt trong bối cảnh bản chất "hộp đen" của LSTM gây khó khăn cho người dùng không chuyên.

Về ứng dụng thực tiễn

Cần phát triển các công cụ hỗ trợ ra quyết định dành cho nhà đầu tư, người mua nhà và cơ quan quản lý. Các công cụ này cần cung cấp thông tin chi tiết về xu hướng biến động giá theo thời gian và các yếu tố ảnh hưởng, giúp người dùng đưa ra quyết định chính xác hơn. Mô hình cũng có thể được tích hợp vào các trang web hoặc ứng dụng bất động sản phổ biến, giúp cung cấp ước lượng giá trị bất động sản theo thời gian thực. Điều này không chỉ mang lại lợi ích lớn cho người dùng cuối mà còn tăng tính cạnh tranh cho các nền tảng kinh doanh bất động sản. Giao diện của các công cụ này cần thân thiện, đặc biệt đối với người dùng không chuyên, những người cần hiểu rõ cơ sở của các dự đoán trước khi đưa ra quyết định.

Hướng nghiên cứu trong tương lai

Cần nghiên cứu việc tích hợp LSTM với các kỹ thuật tiên tiến như Attention Mechanism hoặc kiến trúc Transformer để nâng cao khả năng xử lý chuỗi thời gian và dữ liệu phi cấu trúc. Ngoài việc dự đoán giá nhà, mô hình có thể được mở rộng để giải quyết các bài toán khác trong lĩnh vực bất động sản, chẳng hạn như phân tích xu

hướng khu vực hoặc đánh giá rủi ro đầu tư. Với sự phát triển mạnh mẽ của thị trường bất động sản tại Việt Nam, việc áp dụng và cải tiến liên tục mô hình LSTM sẽ không chỉ mang lại giá trị học thuật mà còn góp phần hiện đại hóa ngành này thông qua chuyển đổi số. Những công cụ dự đoán thông minh này hứa hẹn sẽ trở thành một phần không thể thiếu trong tương lai, giúp các bên liên quan đưa ra quyết định chính xác và hiệu quả hơn./.

DANH MỤC TÀI LIỆU THAM KHẢO

*** Tài liệu tiếng Việt:**

[1] Lê Trung Can và Trần Phước Lộc (2024), "Dự báo chuỗi thời gian với một số mô hình học máy và ứng dụng", *Tạp chí Khoa học Đại học Cần Thơ*, Tập 60, Số chuyên đề: Khoa học tự nhiên (Toán-Lý)

[2] HBR Việt Nam (2023), *Công nghệ dự đoán giá nhà bằng Machine Learning giúp tối ưu lợi nhuận*, https://hbr.com.vn/du-doan-gia-nha-machine-learning, truy cập ngày 10/10/2025

[3] Trịnh Thị Thu Hòa (2022), *Khóa luận tốt nghiệp " Ứng dụng một số mô hình học máy trong dự đoán giá bất động sản tại Hà Nội",* Học viện Ngân hàng, https://vn-document.net/document/du-doan-gia-bat-dong-san-tai-ha-noi-bang-mo-hinh-hoc-may, truy nhập ngày 10/10/2025

[4] Đỗ Văn Hùng (2023), *Đồ án tốt nghiệp "Hồi quy tuyến tính trong dự báo giá nhà",* Đại học Bách Khoa Hà Nội, https://vn-document.net/document/do-an-hoi-quy-tuyen-tinh-du-bao-gia-nha, truy nhập ngày 10/10/2025.

[5] Đỗ Hoàng Hiệp (2021), *Luận văn tốt nghiệp "Dự đoán giá bất động sản tại Quận 7 thành phố Hồ Chí Minh bằng mô hình Long Short - Term Memory*, Đại học quốc gia thành phố Hồ Chí Minh, https://vn-document.net/document/du-doan-gia-bat-dong-san-tai-quan-7-tphcm-bang-mo-hinh-lstm, truy nhập ngày 10/10/2025

[6] Vũ Đăng Hoàng Long, Nguyễn Huỳnh Đại Lợi, Huỳnh Long Nam, Trần Thanh Phúc, Nguyễn Đăng Trung Tiến (2022), *Dự đoán giá nhà sử dụng dữ liệu Kaggle,* Trường Đại học Khoa học Tự nhiên TP.HCM, https://github.com/vlozg/houseprice_predict, truy cập ngày 10/10/2025.

[7] Nguyễn Hương Ly, Hoàng Thị Thu Hà (2022), "Kết hợp mô hình học máy và mô hình thống kê trong dự báo chuỗi thời gian: Trường hợp lạm pháp tại Việt Nam giai đoạn 2000-2021", *Tạp chí Khoa học kinh tế* - Số 10 (01), Đại học Đà Nẵng.

[8] Hoàng Nữ Thanh Tuyền, Đoàn Nguyễn Bảo An, Vương Thị Thảo Bình (2024), "Ứng dụng mô hình nơ-ron nhân tạo LSTM dự đoán giá cổ phiếu Việt Nam", *Working Papers Series,* Trường Đại học Ngoại thương.

[9] https://vietnix.vn/cnn-la-gi, truy cập ngày 10/10/2025.

[10] https://openscience.vn/chi-tiet-du-lieu/co-so-du-lieu-gia-nha-quoc-te-193, truy cập ngày 10/10/2025.

*** Tài liệu Tiếng Anh:**

[11] A. Kumar and R. Singh (2021), *Real estate price prediction using deep learning techniques* J. Big Data Analytics, Springer, doi: 10.1007/s41060-021-00250-9

[12] Hastie, Tibshirani & Friedman (2009), The Elements of Statistical Learning, Springer. [Online]. Available: https://books.google.com.vn/books/about/The_Elements_of_Statistical_Learning

[13] Montgomery, Peck & Vining (2021), Introduction to Linear Regression Analysis, Wiley. [Online]. Available: https://www.wiley.com/en-us/Introduction+to+Linear+Regression+Analysis

[14] S. Hochreiter and J. Schmidhuber (1997), "Long short-term memory" *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735

[15] J. Brownlee (2018), *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*, Machine Learning Mastery.

[16] Y. Zhang and S. Wang (2020), *House price prediction using LSTM neural network,* in Proc. IEEE Int. Conf. Artificial Intelligence, pp. 112–117. doi: 10.1109/ICAI50945.2020.00025

[17] https://www.kaggle.com/datasets/prevek18/ames-housing-dataset, truy nhập ngày 10/10/2025.

[18] https://arxiv.org/pdf/1709.084322, truy nhập ngày 10/10/2025.

[19]https://researchgate.net/publication/371141678_Comparison_of_Real_Estate_Price_Prediction_Based_on_LSTM_and_LGBM, truy nhập ngày 10/10/2025.

[20] https://ijrar.org/papers/IJRAR24A2880, truy nhập ngày 10/10/2025.

[21] https://link.springer.com/article/10.1007/s40745-024-00543-2e, truy nhập ngày 10/10/2025.

[22] https://ieeexplore.ieee.org/abstract/document/10235649, truy nhập ngày 10/10/2025.

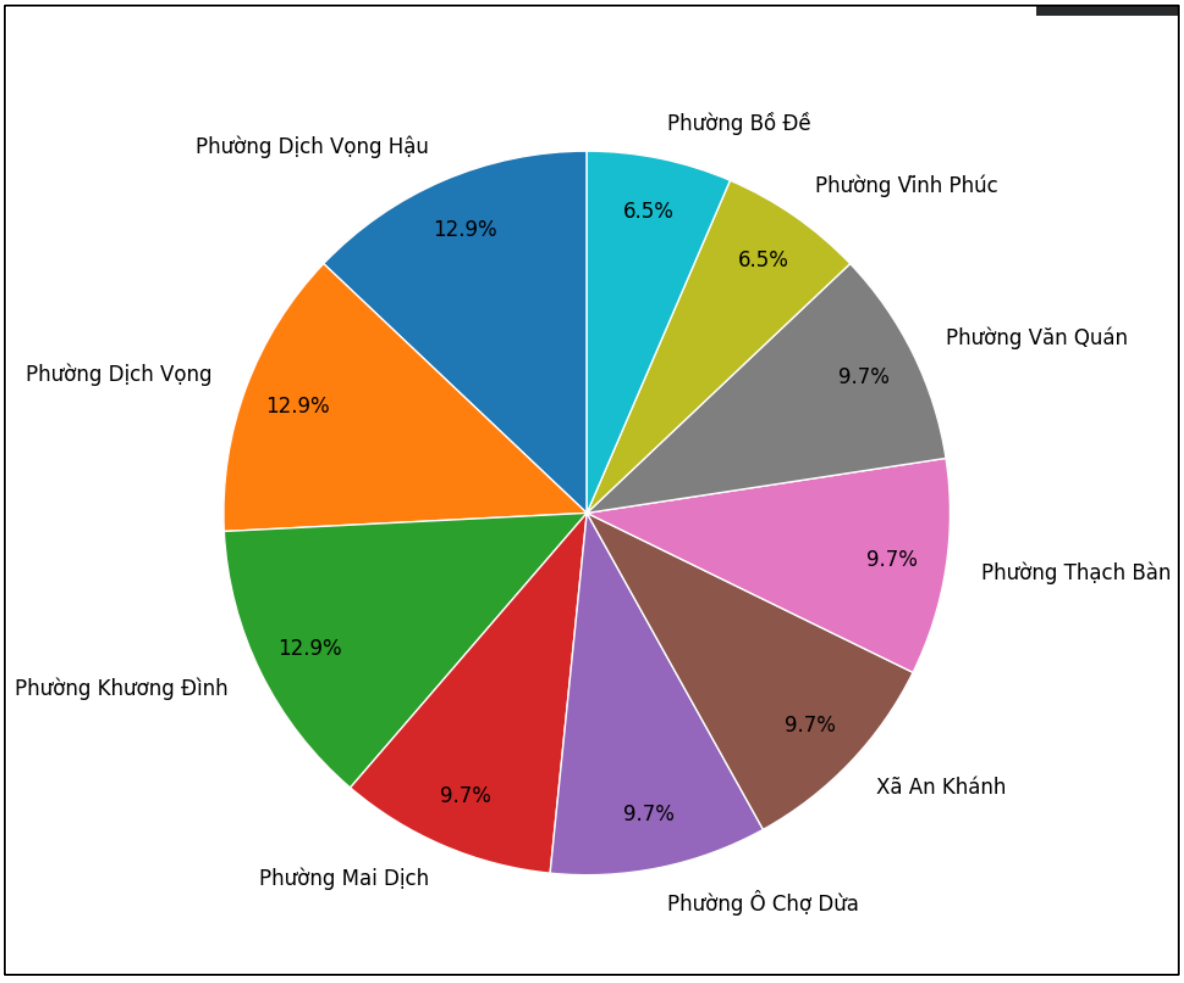
PHỤ LỤC

1. Danh sách thư viện chính được sử dụng và mục đích sử dụng

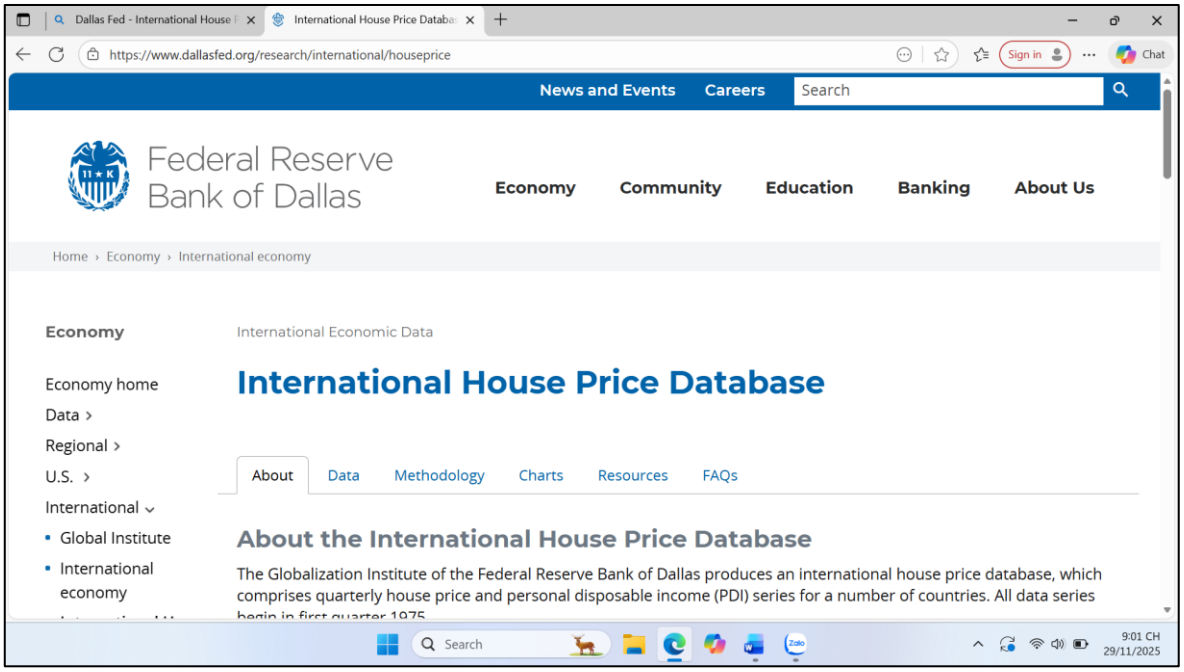
Bước thực hiện	Thư viện	Mục đích sử dụng
Thu thập dữ liệu	Selenium	Tự động hóa trình duyệt để thu thập dữ liệu từ các trang web.
	Requests	Gửi yêu cầu HTTP để lấy nội dung trang web.
	Time	Điều khiển thời gian chờ giữa các yêu cầu (tránh bị chặn bởi server).
	BeautifulSoup	Phân tích cú pháp HTML và trích xuất thông tin từ các bài đăng.
	Pandas	Lưu trữ và xử lý dữ liệu thu thập được dưới dạng DataFrame.
Tiền xử lý dữ liệu	Pandas	Làm sạch, lọc và chuẩn hóa dữ liệu (ví dụ: tách quận/phường từ cột location).
	NumPy	Xử lý tính toán số học trên dữ liệu (ví dụ: tính trung bình, tổng).
	re	Xử lý chuỗi văn bản (ví dụ: tách thông tin từ địa chỉ).
Phân tích dữ liệu	Scikit-learn	Xây dựng và đánh giá mô hình học máy
	Statsmodels	Phân tích thống kê và xây dựng mô hình ARIMA.

Bước thực hiện	Thư viện	Mục đích sử dụng
Trực quan hóa dữ liệu	Matplotlib	Vẽ biểu đồ số lượng nhà theo quận/phường, phân phối diện tích, giá nhà.

2. Top 10 phường/ xã (cũ) của Hà Nội có số lượng tin đăng mua bán nhà riêng cao nhất trong giai đoạn từ tháng 11/2024 đến tháng 10/2025



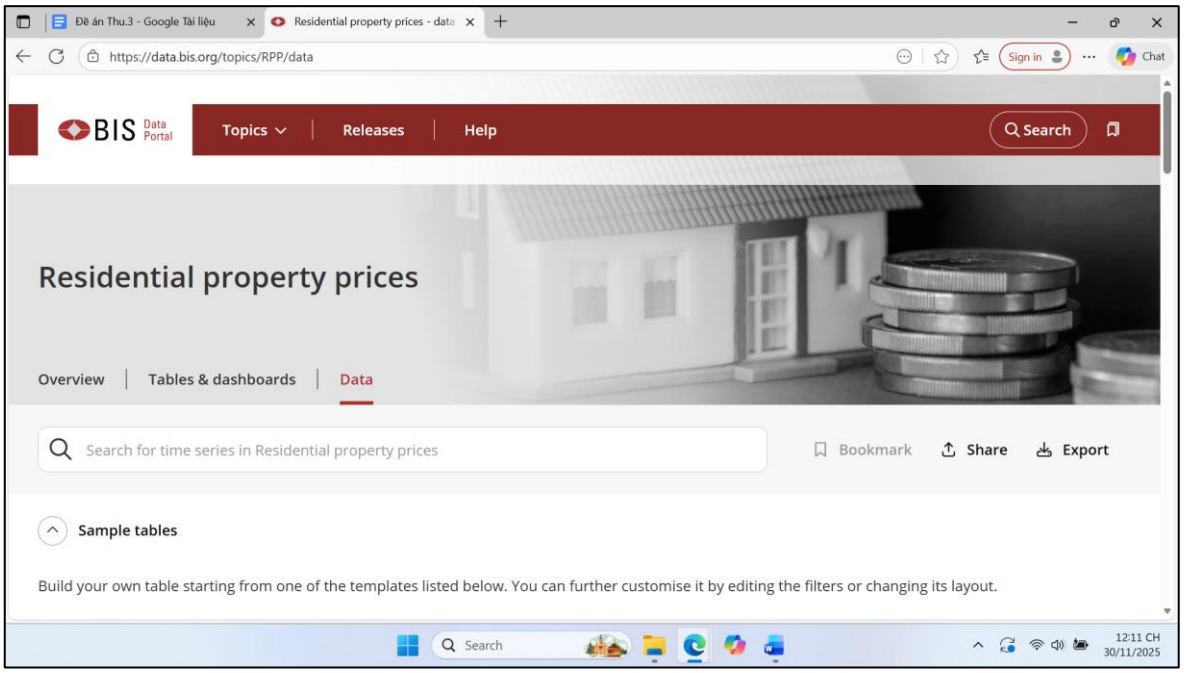
3. Giao diện một số trang web cung cấp dữ liệu giá nhà trên thế giới



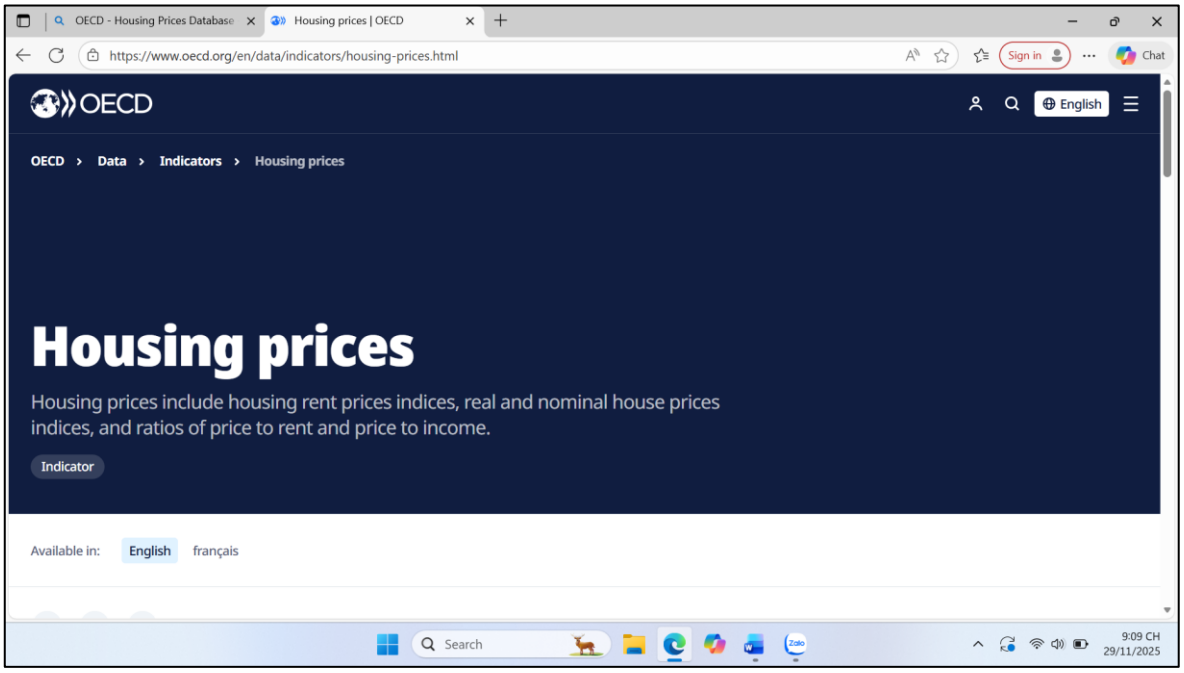
https://www.dallasfed.org/research/international/houseprice



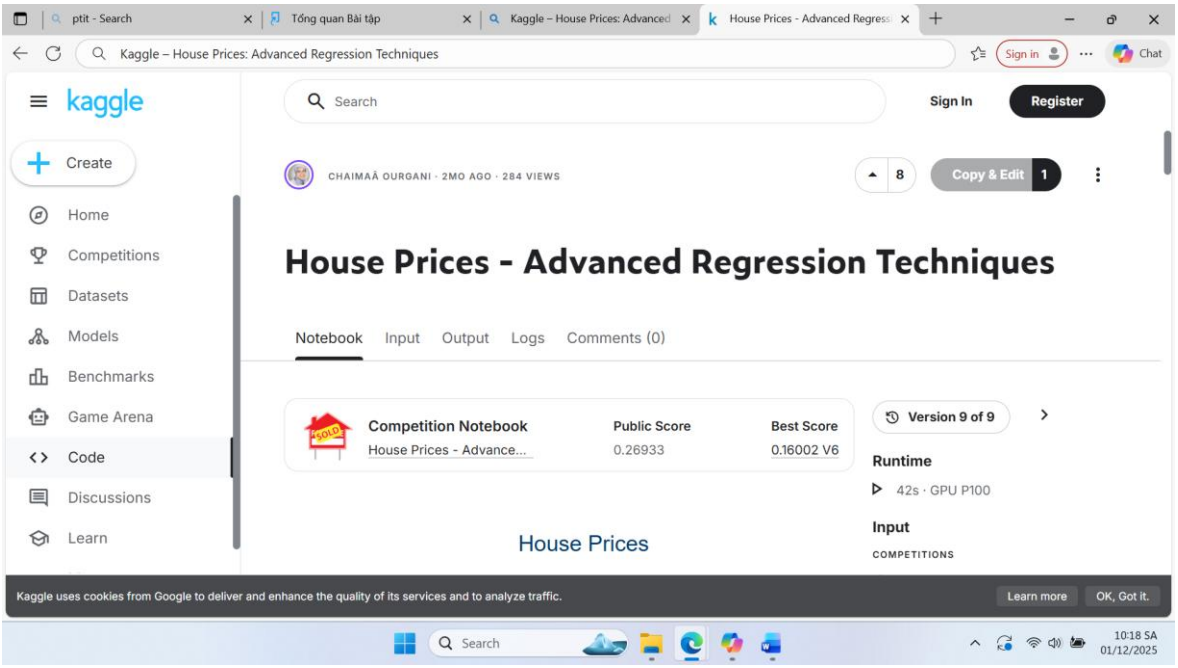
https://www.imf.org/external/research/housing/



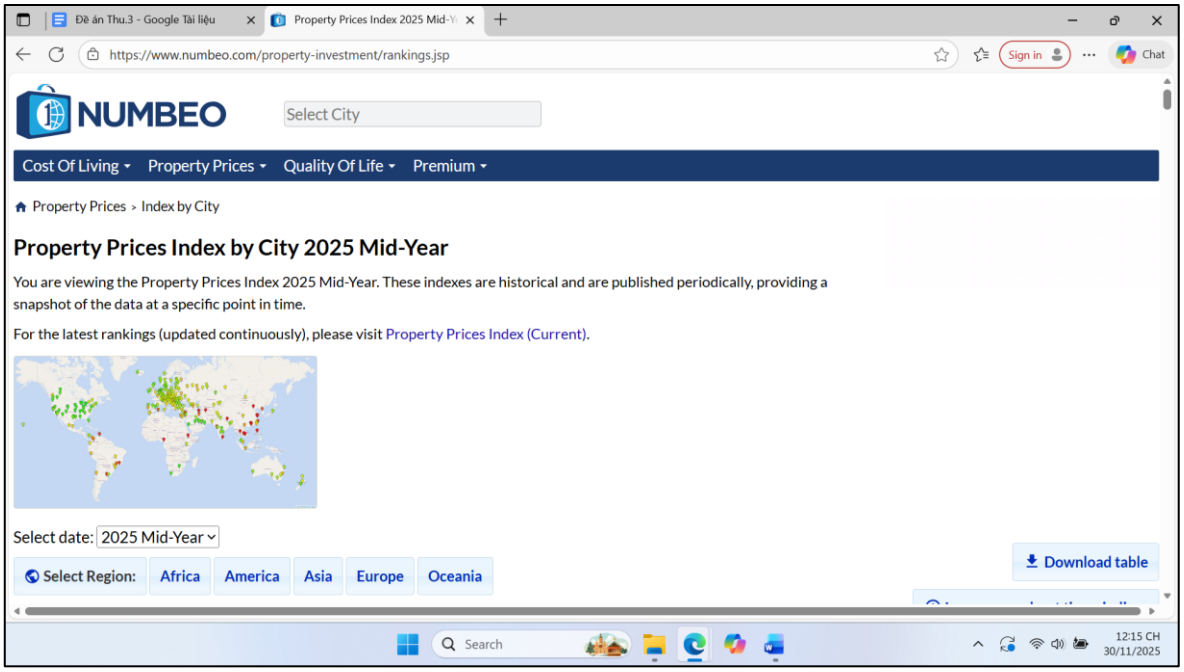
https://data.bis.org/topics/RPP/data



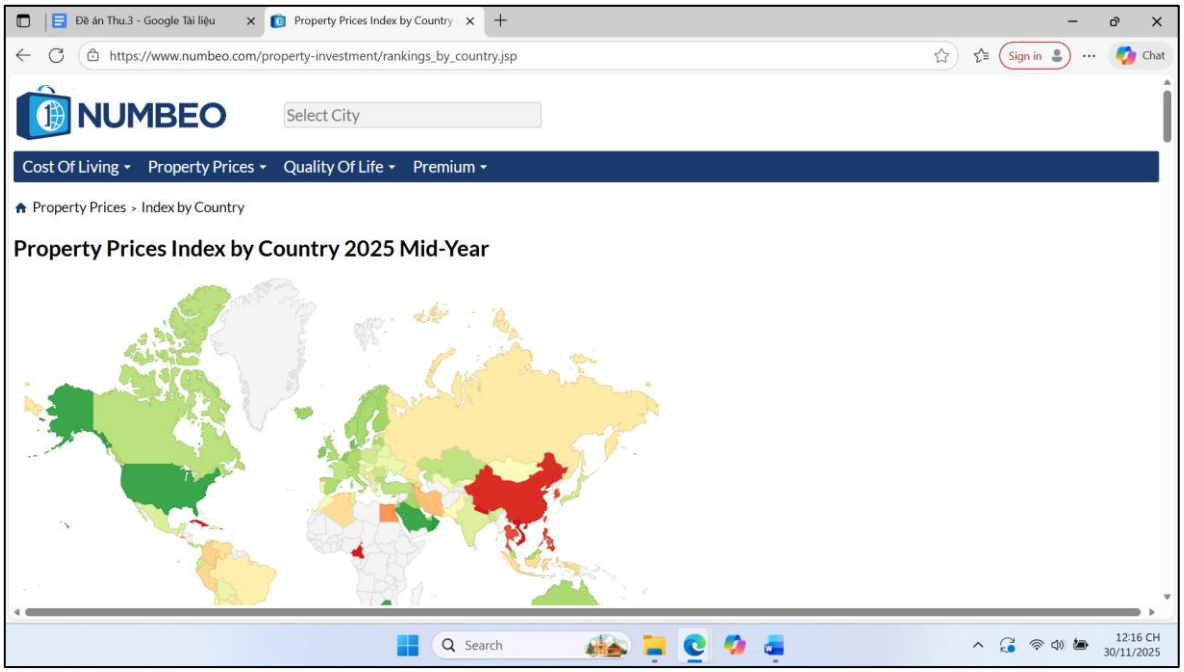
https://www.oecd.org/en/data/indicators/housing-prices.html



https://www.kaggle.com/code/chaimaourgani/house-prices-advanced-regression-techniques

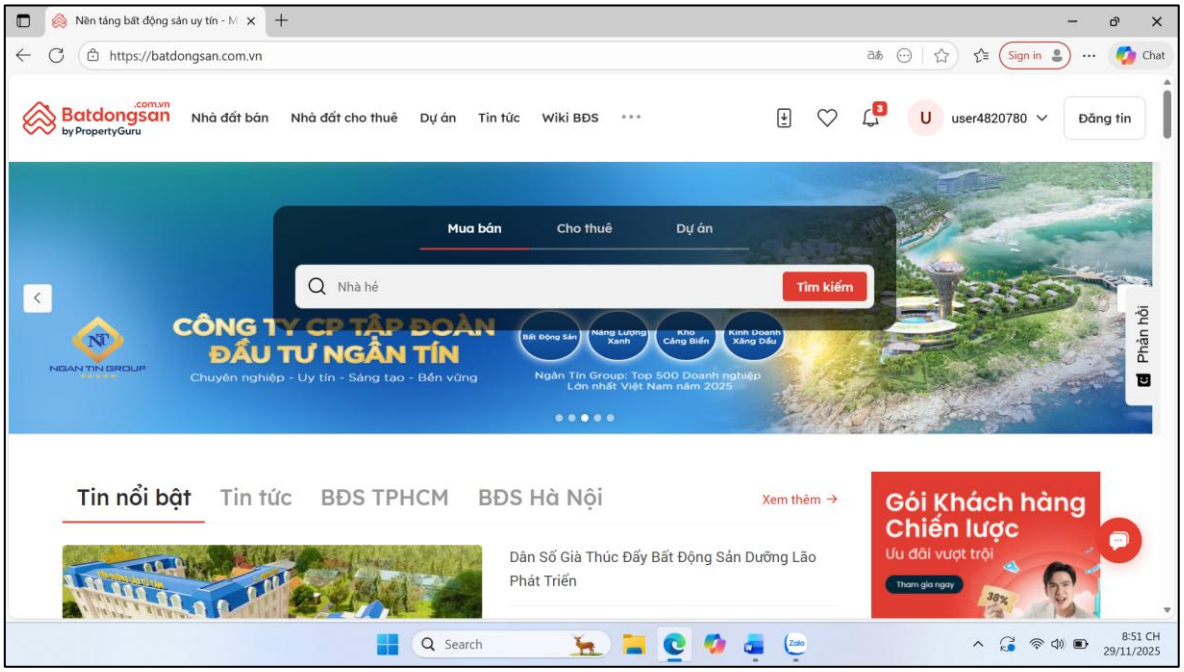


https://www.numbeo.com/property-investment/rankings.jsp

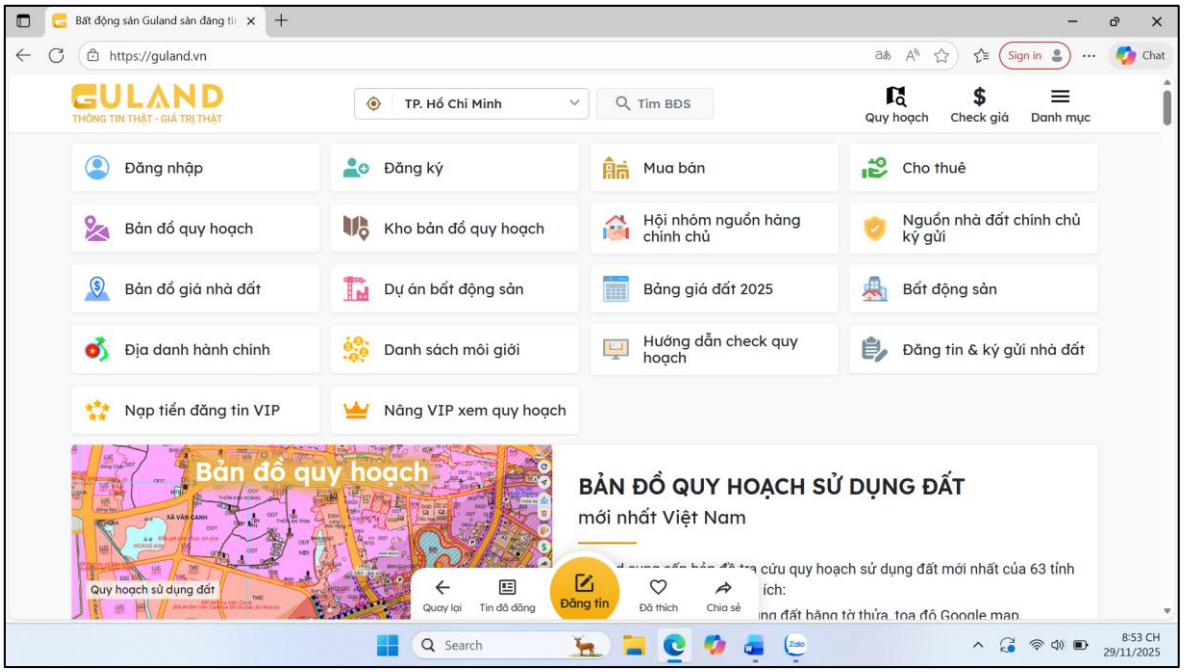


https://www.numbeo.com/property-investment/rankings_by_country.jsp

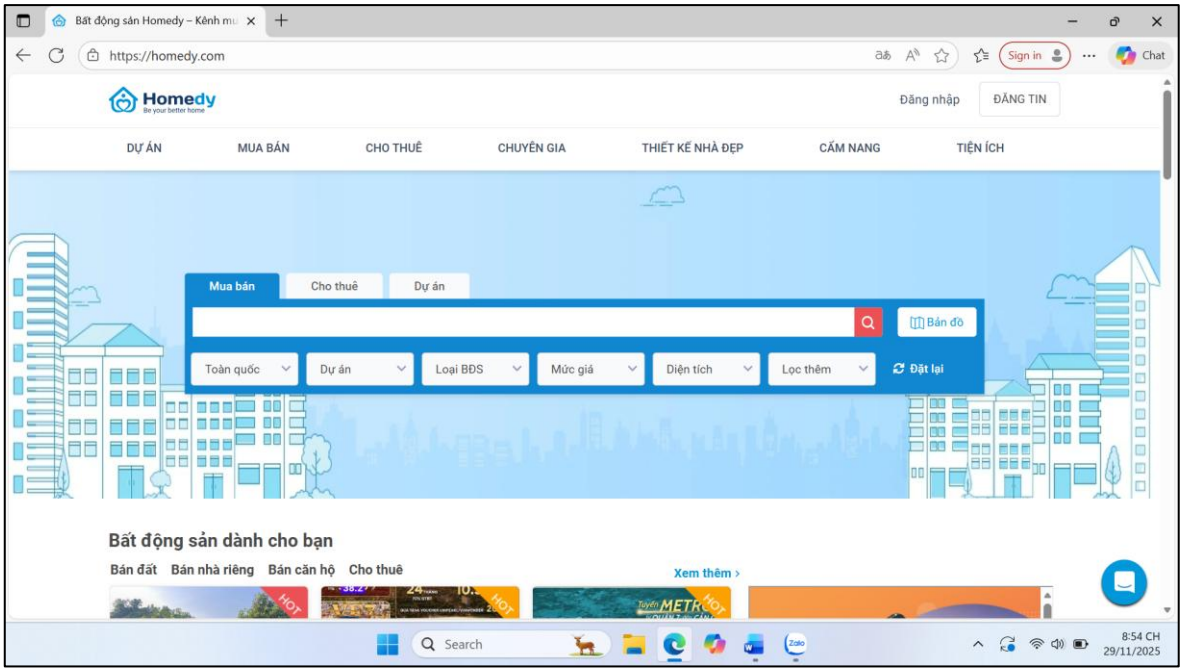
4. Giao diện một số trang web mua bán bất động sản tại Việt Nam



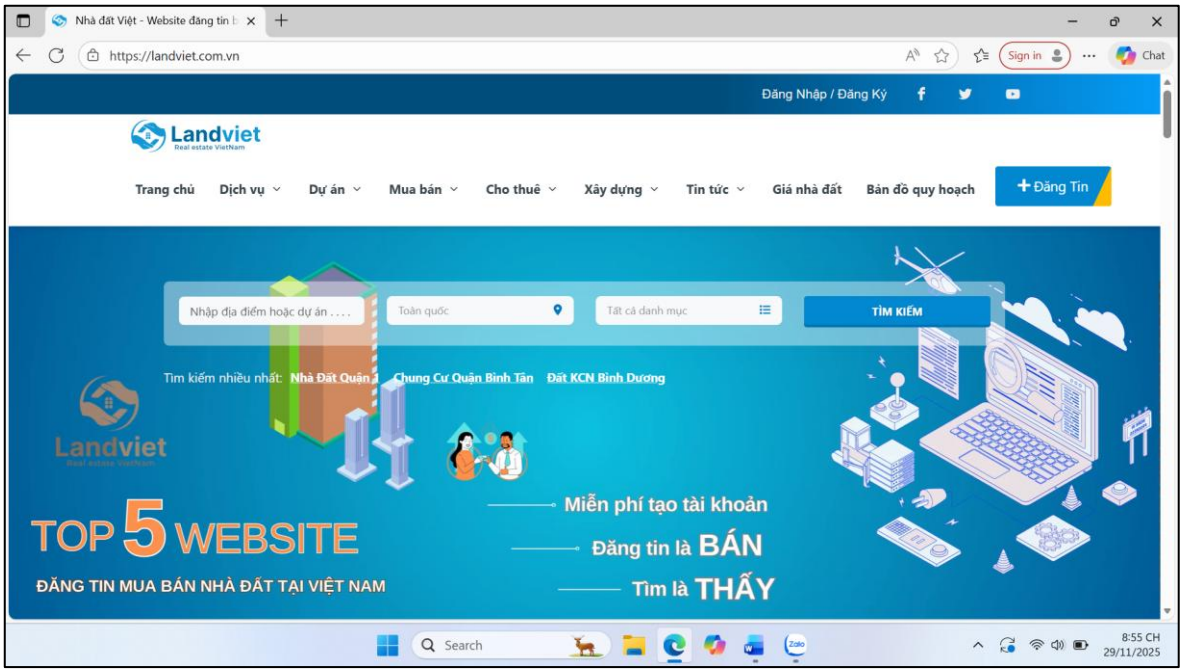
https://batdongsan.com.vn/



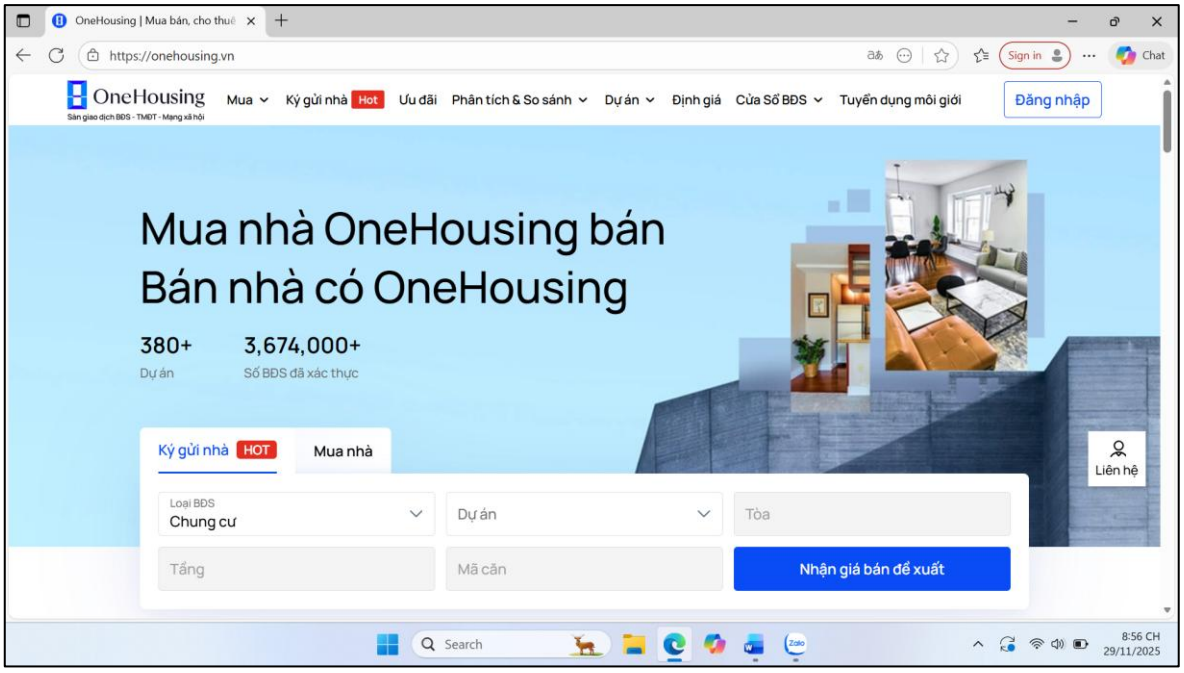
https://guland.vn/



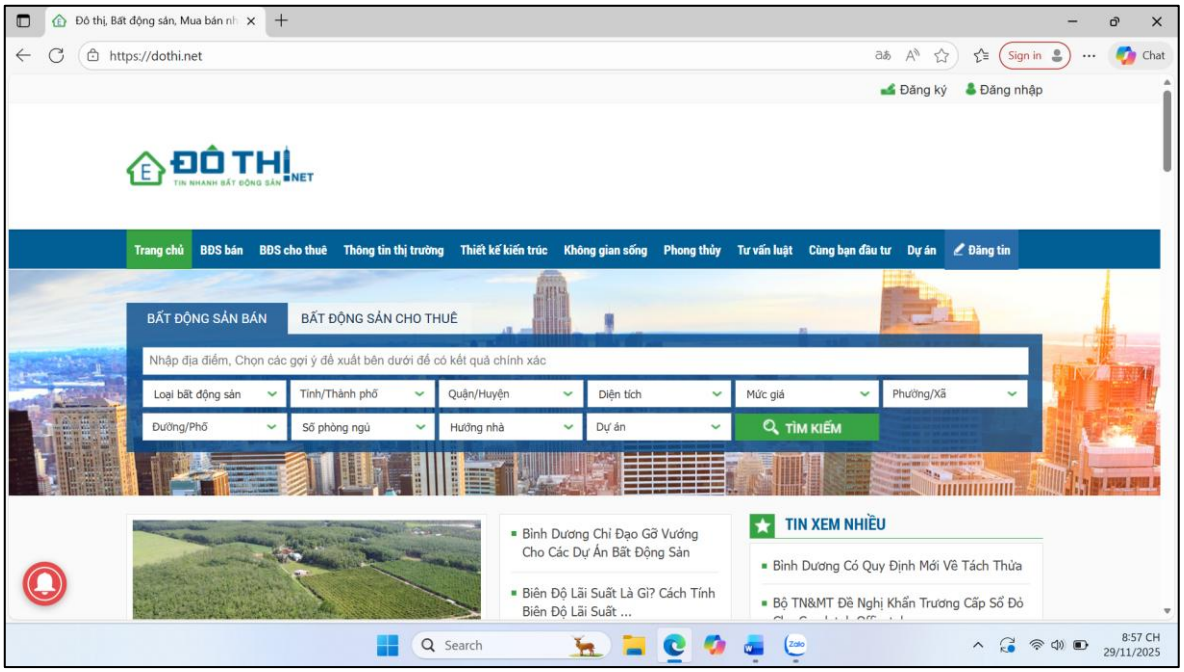
https://homedy.com/



https://landviet.com.vn/



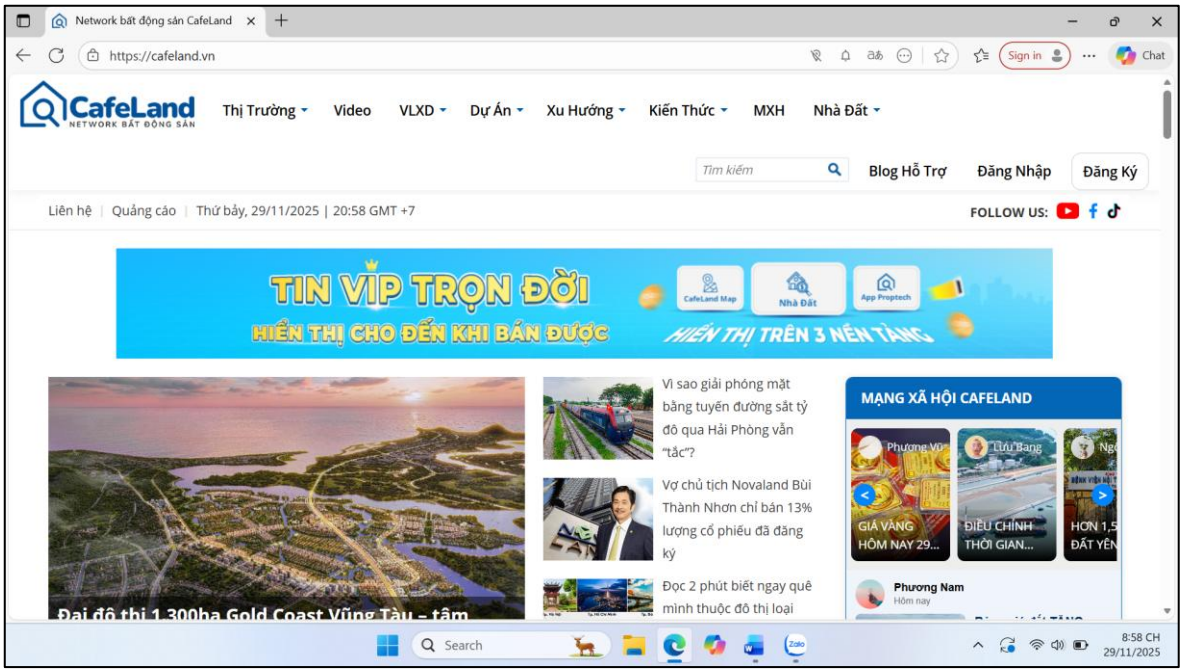
https://onehousing.vn/



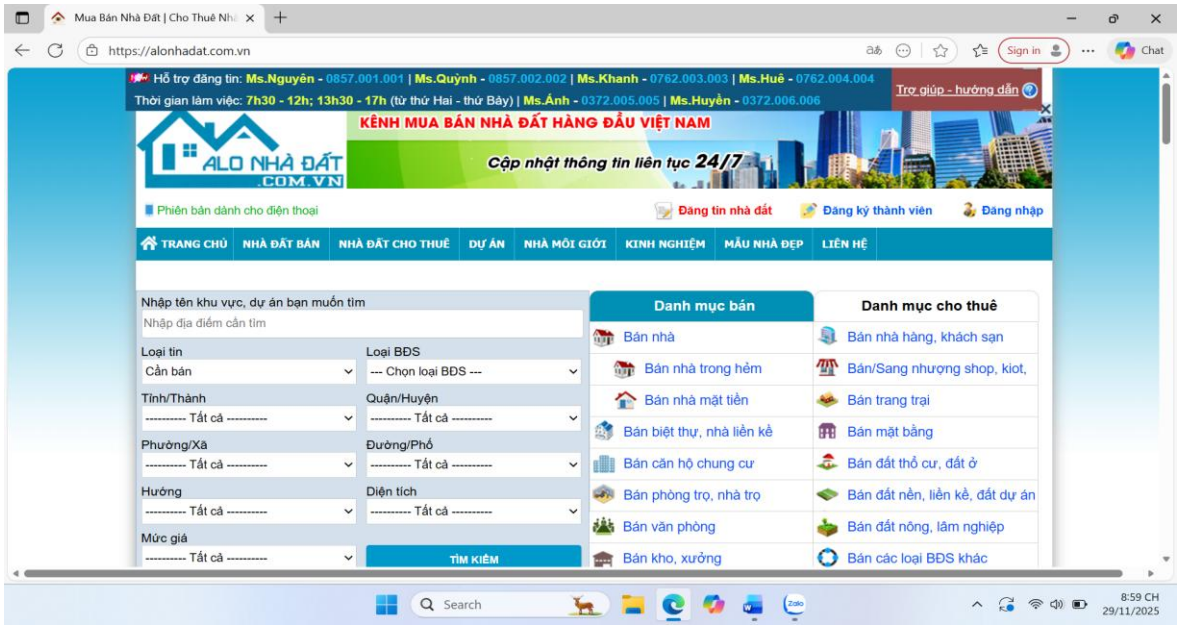
https://dothi.net/



https://123nhadatviet.com/



https://cafeland.vn/



https://alonhadat.com.vn/

10% Tính tương đồng nói chung

Tổng cộng của tất cả các kết quả trùng khớp, bao gồm cả các nguồn trùng lặp, cho mỗi c...

Đã lọc khỏi Báo cáo

- Mục lục tham khảo
- Văn bản được trích dẫn

Nguồn hàng đầu

6% Nguồn Internet
6% Ấn bản
3% Bài tập được nộp (bài của học sinh)

Học viên	Người hướng dẫn khoa học
Phùng Thị Thu	TS. Đỗ Thanh Hà

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BÁO CÁO GIẢI TRÌNH
SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Họ và tên học viên: **Phùng Thị Thu**

Chuyên ngành: Khoa học máy tính

Khóa: 2024 đợt I

Tên đề tài: Áp dụng mô hình Long Short-Term Memory trong dự đoán giá nhà tại Việt Nam

Người hướng dẫn khoa học: TS. Đỗ Thanh Hà

Ngày bảo vệ: 24/01/2026

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

TT	Ý kiến hội đồng	Sửa chữa của học viên
1	Làm rõ dữ liệu nào được khai thác	Đã tiếp thu, mô tả chi tiết làm rõ dữ liệu được khai thác tại mục 2.2 (trang 22, trang 23)
2	Khảo sát vấn đề cần giải quyết của bài toán	Đã tiếp thu, mô tả rõ hơn về khảo sát vấn đề cần giải quyết của bài toán tại mục 2.4 (trang 31, trang 32)
3	Làm rõ mô hình, giải thích kết quả thực nghiệm	Đã tiếp thu làm rõ mô hình, giải thích kết quả thực nghiệm tại mục 3.1 (trang 38, trang 40)
4	Trình bày học sâu quá nhiều, nên cô đọng lại và tập trung LSTM	Đã tiếp thu lược bớt các nội dung học sâu ANN, CNN, RNN và làm rõ hơn LSTM (từ trang 10 đến trang 13)
5	Rà soát, chỉnh sửa đề án về trình bày khuân dạng đoạn văn, hình vẽ, bảng biểu	Đã rà soát, chỉnh sửa tổng thể trong Đề án

Hà Nội, ngày 02 tháng 02 năm 2026

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN	THƯ KÝ HỘI ĐỒNG	NGƯỜI HƯỚNG DẪN KHOA HỌC	HỌC VIÊN
TS. Nguyễn Duy Phương	TS. Đỗ Thị Liên	TS. Đỗ Thanh Hà	Phùng Thị Thu

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BIÊN BẢN
HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 3299/QĐ-HV ngày 26 tháng 12 năm 2025 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi ..9h... giờ ..45... phút, ngày 24 tháng 01 năm 2025 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Phùng Thị Thu**

Tên đề án tốt nghiệp: **Áp dụng mô hình Long Short-Term Memory trong dự đoán giá nhà tại Việt Nam**

Ngành: **Khoa học máy tính** Mã số: **8.48.01.01**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: ..5...../ 05

TT	HỌ VÀ TÊN	TRÁCH NHIỆM TRONG HĐ	GHI CHÚ
1	**TS. Nguyễn Duy Phương**	Chủ tịch	
2	**TS. Đỗ Thị Liên**	Thư ký	
3	**PGS.TS. Nguyễn Long Giang**	Phản biện 1	
4	**PGS.TS. Đỗ Trung Tuấn**	Phản biện 2	
5	**PGS.TS. Nguyễn Hữu Quỳnh**	Uỷ viên	

Các nội dung thực hiện:

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên (Có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

- Vì sao chọn bài toán này?
- Vì sao chọn mô hình LSTM?
- Đầu vào cuối cùng đưa vào mô hình là gì?
- Thông tin về địa điểm được sử dụng như thế nào trong đề án?

7. Trả lời của học viên:

Học viên tiếp thu và chỉnh sửa quyển.

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên *(có văn bản kèm theo).*

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:
  1. Trưởng Ban kiểm phiếu: Đỗ Trung Tuấn
  2. Uỷ viên Ban kiểm phiếu: Đỗ Thị Liên
  3. Uỷ viên Ban kiểm phiếu: Nguyễn Long Giang
- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.
- Ban kiểm phiếu làm việc:
- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu *(có Biên bản họp Ban kiểm phiếu kèm theo)*
- **Điểm trung bình của đề án tốt nghiệp:** 7.5

Kết luận:

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:
   - Khảo sát vấn đề cần giải quyết của kế toán
   - Dữ liệu nào được khai thác
   - Làm rõ mô hình, giải thích kết quả thực nghiệm

2. Đề nghị Học viện công nhận (hoặc ~~không~~) và cấp bằng (hoặc ~~không~~) thạc sĩ cho học viên: Có

3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho NCS.... Không

Buổi làm việc kết thúc vào 10h30 cùng ngày.

Chủ tịch	**Thư ký**
Phương	
TS. Nguyễn Duy Phương	**TS. Đỗ Thị Liên**

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc
==========

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề án: Áp dụng mô hình LSTM trong dự đoán giá nhà tại Việt Nam

Chuyên ngành: Khoa học máy tính

Mã số: 8.48.01.01

Họ và tên học viên: Phùng Thị Thu

Họ và tên người nhận xét: PGS.TS. Nguyễn Long Giang

Cơ quan công tác: Viện Công nghệ thông tin, Viện Hàn Lâm KH&CN Việt Nam

Số điện thoại: 0904739189 Email: nlgiang@ioit.ac.vn

NỘI DUNG NHẬN XÉT

1. Cơ sở khoa học và thực tiễn, sự cần thiết lựa chọn đề tài

Đề án giải quyết bài toán dự báo giá nhà sử dụng mạng LSTM với đóng góp chính là thu thập, tiền xử lý, chuẩn hóa dữ liệu. Chủ đề nghiên cứu của đề án có ý nghĩa khoa học và thực tiễn.

2. Nội dung của đề án, các kết quả đã đạt được

Đề án đã đạt được kết quả chính như sau:

1) Tìm hiểu và hệ thống hóa các mô hình học sâu trong bài toán dự đoán
2) Thu thập, tiền xử lý, chuẩn hóa dữ liệu về giá nhà sang dữ liệu chuỗi thời gian nhằm xây dựng bộ dữ liệu huấn luyện cho mô hình LSTM. Huấn luyện, đánh giá mô hình LSTM với các mô hình học máy khác nhằm minh chứng cho tính hiệu quả của mô hình LSTM.

Về cơ bản, đóng góp chính của học viên nằm ở thu thập, tiền xử lý, chuẩn hóa và chuyển đổi dữ liệu thực cho mô hình LSTM.

3. Những vấn đề học viên cần chỉnh sửa, bổ sung

- Rà soát, chỉnh sửa đề án về trình bày, khuôn dang đoạn văn, hình vẽ, bảng biểu
- Phần trình bày các mô hình học sâu quá nhiều, nên cô đọng lại và chỉ trình bày mô hình LSTM là chính.
- Về quy trình thu thập, tiền xử lý, chuẩn hóa và chuyển đổi sang dữ liệu chuỗi thời gian, đề án cần đặc tả bộ dữ liệu chuỗi thời gian thu được đưa vào mô hình LSTM là gì ? (bao nhiêu mẫu, cấu trúc dữ liệu...) và tại sao lạu sử dụng được cho các mô

hình phân lớp khác? Các thông tin về khu vực (vị trí) được sử dung cho việc dự đoán giá nhà như thế nào ?

4. Kết luận

- Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.
- Đề nghị học viên chỉnh sửa/bổ sung luận văn theo góp ý.

Hà Nội, ngày 20 tháng 01 năm 2026

XÁC NHẬN CỦA CƠ QUAN

NGƯỜI NHẬN XÉT

Nguyễn Long Giang

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập Tự do Hạnh phúc

BẢN NHẬN XÉT
Đề án tốt nghiệp Thạc sỹ kĩ thuật

Tên đề tài đề án : **Áp dụng mô hình bộ nhớ ngắn hạn dài trong dự đoán giá nhà tại Việt Nam**

Chuyên ngành: Khoa học máy tính

Mã số: 8.48.01.01.

Tên học viên : **Phùng Thị Thu**

Họ và tên người nhận xét: Đỗ Trung Tuấn,

Đơn vị, Cơ quan công tác: Học viện Công nghệ Bưu chính Viễn thông

Ý KIẾN NHẬN XÉT

1. Tính thời sự, cấp thiết, ý nghĩa khoa học và thực tiễn của đề tài

Mô hình bộ nhớ ngắn hạn dài, là một dạng cải tiến của mạng thần kinh đệ qui RNN, được thiết kế đặc biệt để giải quyết vấn đề mất mát đạo hàm, tức vanishing gradient, giúp mô hình có khả năng học và ghi nhớ các phụ thuộc xa trong dữ liệu chuỗi.

Học viên sử dụng mô hình LSTM cho bài toán dự đoán giá nhà. Bài toán dự đoán là sử dụng các mô hình toán học và thuật toán học máy để ước tính các kết quả chưa biết trong tương lai hoặc các giá trị bị thiếu dựa trên dữ liệu lịch sử.

Đề án tốt nghiệp đáp ứng mã ngành đào tạo sau đại học, chuyên ngành khoa học máy tính, mã số 8.48.01.01.

Học viên Phùng Thị Thu được cô TS. Đỗ Thanh Hà hướng dẫn.

Kết quả đề án không trùng lặp với các đề án, đề tài nghiên cứu khoa học đã bảo vệ tại cơ sở đào tạo.

2. Về nội dung, chất lượng của đề án, các kết quả đã đạt được

Học viên đã có các tài liệu liên quan đến việc bảo vệ kết quả nghiên cứu và ứng dụng, gồm bản viết đề án tốt nghiệp và bản tóm tắt đề án, đề cương viết đề án đã được bộ môn chuyên ngành thẩm định và góp ý.

Nội dung bản tóm tắt đề án tốt nghiệp thể hiện các ý chính của bản viết đề án tốt nghiệp. Đề án tốt nghiệp của học viên chia thành các chương, liên quan đến các vấn đề sau :

- Bài toán dự đoán giá nhà và khái niệm liên quan đến chuỗi thời gian. Học viên đã đề cập nhiều mô hình học sâu như ANN, CNN, RNN và LSTM, và quyết định sử dụng mô hình LSTM;
- Học viên đã thiết kế mô hình thực nghiệm và đánh giá kết quả thu được.

3. Những vấn đề cần giải thích thêm

Liên quan đến kết quả của học viên, qua bản viết đề án tốt nghiệp, người nhận xét thấy :

- Có thể dự đoán bằng mô hình toán, thay vì sử dụng mô hình học sâu ?

4. Kết luận

Theo các kết quả đạt được của đề án tốt nghiệp về nghiên cứu và thực nghiệm, căn cứ vào yêu cầu của đề án tốt nghiệp thạc sĩ chuyên ngành, tôi đồng ý để học viên Phùng Thị Thu được bảo vệ luận văn trước Hội đồng chấm đề án tốt nghiệp thạc sĩ.

Hà nội. 22 tháng 01 năm 2026

Người nhận xét kí tên

Đỗ Trung Tuấn