**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: TRÍ TUỆ NHÂN TẠO**

**ĐỀ TÀI: XÂY DỰNG MÔ HÌNH MẠNG NEURON ĐỂ DỰ ĐOÁN GIÁ NHÀ**

**Giáo viên hướng dẫn: TS. Trần Đăng Công**

**Sinh viên thực hiện:**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Mã sv** | **Họ và tên** | **Lớp** |
| 01 | 1771020750 | Nguyễn Viết Việt | CNTT 17-14 |
| 02 | 1771020745 | Trần Quốc Việt | CNTT 17-14 |
| 03 | 1771020008 | Nguyễn Phú Việt An | CNTT 17-14 |

**Hà Nội, năm 2025**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: TRÍ TUỆ NHÂN TẠO AI**

**ĐỀ TÀI: XÂY DỰNG MÔ HÌNH MẠNG NEURON ĐỂ DỰ ĐOÁN GIÁ NHÀ**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | Mã Sinh Viên | Họ và Tên | Ngày Sinh | Điểm | |
| Bằng Số | Bằng Chữ |
| 1 | 1771020750 | Nguyễn Viết Việt | 17/01/2004 |  |  |
| 2 | 1771020745 | Trần Quốc Việt | 06/09/2004 |  |  |
| 3 | 1771020008 | Nguyễn Phú Việt An | 26/05/2005 |  |  |

### 

### CÁN BỘ CHẤM THI

TS.Trần Đăng Công

**Hà Nội, năm 2025**

**LỜI NÓI ĐẦU**

Trong thế giới hiện đại, thị trường bất động sản đã trở thành một trong những lĩnh vực có ảnh hưởng lớn đến nền kinh tế toàn cầu. Sự gia tăng dân số, quá trình đô thị hóa nhanh chóng và nhu cầu về nhà ở đã tạo ra những biến động mạnh mẽ trong giá cả bất động sản. Đối với người mua, người bán, cũng như các nhà đầu tư, việc dự đoán giá nhà không chỉ giúp họ đưa ra quyết định kịp thời mà còn bảo vệ tài sản và tối ưu hóa lợi nhuận. Tuy nhiên, việc dự đoán giá nhà không hề đơn giản. Giá trị bất động sản chịu ảnh hưởng từ nhiều yếu tố, bao gồm vị trí địa lý, diện tích, tình trạng nhà, cơ sở hạ tầng, và các yếu tố kinh tế vĩ mô như lãi suất, tỷ lệ thất nghiệp, và xu hướng thị trường. Do đó, việc phân tích và xử lý những dữ liệu phức tạp này đòi hỏi những phương pháp tiên tiến và chính xác.

Mạng neuron, một trong những công nghệ tiên tiến nhất trong lĩnh vực trí tuệ nhân tạo, đã nổi lên như một công cụ mạnh mẽ để giải quyết các bài toán phức tạp. Với khả năng học hỏi từ những mẫu dữ liệu lớn và có khả năng phát hiện các mối quan hệ ẩn giữa các biến số, mạng neuron có thể cung cấp những dự đoán chính xác hơn về giá nhà. Việc áp dụng mạng neuron trong lĩnh vực dự đoán giá bất động sản không chỉ giúp cải thiện độ chính xác của các dự đoán mà còn giúp các nhà đầu tư nhận diện được những cơ hội và rủi ro trong thị trường.

Trong nghiên cứu này, chúng tôi sẽ trình bày quy trình xây dựng mô hình mạng neuron để dự đoán giá nhà qua các bước cụ thể: thu thập dữ liệu từ nhiều nguồn khác nhau, bao gồm các trang web bất động sản, cơ sở dữ liệu công cộng và các khảo sát thị trường; xử lý dữ liệu để đảm bảo tính chính xác, bao gồm việc làm sạch và chuẩn hóa dữ liệu; xây dựng mô hình mạng neuron bằng các kỹ thuật học sâu, tối ưu hóa các tham số để đạt được hiệu suất tốt nhất; và cuối cùng, đánh giá hiệu suất của mô hình thông qua các chỉ số như độ chính xác và độ lệch trung bình. Chúng tôi hy vọng rằng nghiên cứu này sẽ không chỉ mang lại những giá trị thực tiễn cho ngành bất động sản mà còn mở ra những hướng đi mới cho các nghiên cứu tiếp theo trong lĩnh vực dự đoán giá nhà, từ đó đóng góp vào việc cải thiện khả năng ra quyết định cho người dùng và tạo ra những công cụ hữu ích trong việc phân tích thị trường.

**MỤC LỤC**

[**CHƯƠNG 1. PHÂN TÍCH YÊU CẦU BÀI TOÁN** 8](#_Toc193484393)

[**1.1. Mô tả bài toán** 8](#_Toc193484394)

[**1.2. Dữ liệu sử dụng** 8](#_Toc193484395)

[**1.3. Yêu cầu hệ thống** 9](#_Toc193484396)

[**1.4. Ứng dụng Thực Tiễn của Mô Hình Dự Đoán Giá Nhà** 10](#_Toc193484397)

[*1.4.1. Hỗ trợ người mua và người bán* 10](#_Toc193484398)

[*1.4.2. Hỗ trợ các nhà đầu tư* 11](#_Toc193484399)

[*1.4.3. Lập kế hoạch đô thị* 12](#_Toc193484400)

[**CHƯƠNG 2. MÔ TẢ THUẬT TOÁN VÀ PHÂN TÍCH MÃ NGUỒN** 13](#_Toc193484401)

[**2.1. Thuật toán sử dụng** 13](#_Toc193484402)

[*2.2.1. Tiền xử lý dữ liệu* 16](#_Toc193484403)

[*2.2.2. Xây dựng mô hình* 17](#_Toc193484404)

[*2.2.3. Huấn luyện và lưu mô hình* 18](#_Toc193484405)

[2.2.4. Đánh giá mô hình 19](#_Toc193484406)

[**CHƯƠNG 3. THỬ NGHIỆM, ĐÁNH GIÁ VÀ CẢI TIẾN** 20](#_Toc193484407)

[**3.1. Thử nghiệm với các bộ tham số khác nhau** 20](#_Toc193484408)

[**3.2. Phân tích kết quả thử nghiệm** 23](#_Toc193484409)

[**3.3. Nhận xét đặc điểm và hạn chế của mô hình** 24](#_Toc193484410)

[*3.3.1. Đặc điểm nổi bật* 24](#_Toc193484411)

[*3.3.2. Hạn chế* 24](#_Toc193484412)

[**3.4. Đề xuất cải tiến thuật toán** 25](#_Toc193484413)

[*3.4.1. Thử nghiệm các kiến trúc mạng khác nhau* 26](#_Toc193484414)

[*3.4.2. Sử dụng kỹ thuật tăng cường dữ liệu (Data Augmentation)* 26](#_Toc193484415)

[*3.4.3. Áp dụng các phương pháp regularization như L1/L2* 26](#_Toc193484416)

[**KẾT LUẬN** 28](#_Toc193484417)

[**DANH MỤC TÀI LIỆU THAM KHẢO** 30](#_Toc193484418)

**MỤC LỤC HÌNH ẢNH**

[Hình 1. 1 15](#_Toc193483725)

[Hình 1. 2 16](#_Toc193483726)

[Hình 1. 3 17](#_Toc193483727)

[Hình 1. 4: Kết quả thử nghiệm Learning Rate 0.001 19](#_Toc193483728)

[Hình 1. 5: Kết quả thử nghiệm Learning Rate 0.0005 20](#_Toc193483729)

[Hình 1. 6: Kết quả thử nghiệm Learning Rate 0.0001 21](#_Toc193483730)

**BẢNG CÁC TỪ VIẾT TẮT**

|  |  |  |
| --- | --- | --- |
| **STT** | **TỪ VIẾT TẮT** | **VIẾT ĐẦY ĐỦ** |
| **1** | **MAE** | **Mean Absolute Error** |
| **2** | **DNN** | **Deep Neural Network** |
| **3** | **LSTM** | **Long Short-Term Memory** |
| **4** | **ReLU** | **Rectified Linear Unit** |

# **CHƯƠNG 1.** **PHÂN TÍCH YÊU CẦU BÀI TOÁN**

## **1.1.** **Mô tả bài toán**

Dự đoán giá nhà là một bài toán quan trọng trong lĩnh vực bất động sản, có vai trò thiết yếu trong việc hỗ trợ các quyết định mua bán nhà một cách chính xác và hiệu quả. Trong bối cảnh thị trường bất động sản liên tục biến động, việc hiểu rõ giá trị thực tế của một bất động sản là rất cần thiết không chỉ đối với người mua và người bán mà còn đối với các nhà đầu tư và các cơ quan quản lý.

Mục tiêu của bài toán này là xây dựng một mô hình học máy có khả năng dự đoán giá nhà dựa trên các đặc trưng đa dạng của bất động sản, chẳng hạn như diện tích, số phòng ngủ, số phòng tắm, vị trí địa lý, và các yếu tố khác như tình trạng của bất động sản, năm xây dựng, và tiện ích xung quanh. Những yếu tố này thường có sự ảnh hưởng mạnh mẽ đến giá trị của một ngôi nhà, và việc thu thập, phân tích và khai thác thông tin từ chúng là rất quan trọng.

Để thực hiện dự đoán chính xác, mô hình cần phải học từ một tập dữ liệu phong phú và đa dạng, phản ánh các xu hướng thị trường cũng như các yếu tố tác động đến giá cả. Các phương pháp học máy hiện đại, đặc biệt là các mô hình mạng nơ-ron sâu (Deep Neural Networks), cho phép chúng ta phát hiện các mối quan hệ phức tạp giữa các đặc trưng đầu vào và giá nhà. Mô hình không chỉ cung cấp giá trị dự đoán mà còn giúp người dùng hiểu rõ hơn về các yếu tố ảnh hưởng đến giá trị bất động sản, từ đó đưa ra những quyết định sáng suốt hơn trong các giao dịch mua bán. Ngoài ra, việc dự đoán giá nhà còn có thể hỗ trợ trong việc lập kế hoạch đô thị và phát triển bất động sản, giúp các nhà quy hoạch đưa ra các quyết định phù hợp với nhu cầu và xu hướng của thị trường. Với sự phát triển của công nghệ và dữ liệu lớn, khả năng dự đoán giá nhà sẽ ngày càng chính xác hơn, mở ra nhiều cơ hội cho các bên liên quan trong lĩnh vực bất động sản.

## **1.2. Dữ liệu sử dụng**

Bộ dữ liệu được lưu trữ trong tệp có tên là Data\_Set.csv, chứa nhiều đặc trưng mô tả các thuộc tính của các ngôi nhà khác nhau. Mỗi bản ghi trong tệp này đại diện cho một ngôi nhà cụ thể và bao gồm một loạt các thông tin quan trọng. Trong số các đặc trưng này, chúng ta có thể tìm thấy các yếu tố như diện tích sàn, số lượng phòng ngủ, số phòng tắm, vị trí địa lý, và tình trạng của ngôi nhà. Những thông tin này giúp tạo ra một bức tranh tổng thể về từng bất động sản.

Cột "Giá" trong bộ dữ liệu đóng vai trò rất quan trọng, vì nó chứa giá trị thực tế của từng ngôi nhà. Giá trị này không chỉ phản ánh giá thị trường mà còn chịu ảnh hưởng bởi nhiều yếu tố, bao gồm cả vị trí bất động sản, điều kiện kinh tế địa phương, và các yếu tố xã hội khác. Việc có một cột "Giá" chính xác cho phép chúng ta thực hiện các phân tích và dự đoán giá nhà một cách hiệu quả hơn.

Bằng cách sử dụng bộ dữ liệu này, chúng ta có thể áp dụng các phương pháp học máy để xây dựng mô hình dự đoán giá nhà. Mô hình sẽ học từ các đặc trưng đầu vào để tìm ra mối quan hệ giữa chúng và giá trị thực tế, từ đó giúp đưa ra những dự đoán chính xác hơn về giá trị của các ngôi nhà trong tương lai. Với sự đa dạng và phong phú của các đặc trưng trong bộ dữ liệu, chúng ta có thể mong đợi rằng mô hình sẽ có khả năng nắm bắt được những yếu tố phức tạp ảnh hưởng đến giá nhà, giúp người dùng ra quyết định mua bán một cách thông minh và hiệu quả hơn.

## **1.3. Yêu cầu hệ thống**

Đọc và Tiền xử lý Dữ liệu

Bước đầu tiên trong quá trình dự đoán giá nhà là đọc và tiền xử lý dữ liệu từ tập tin CSV. Tập tin này chứa nhiều thông tin quan trọng về các ngôi nhà, bao gồm các đặc trưng và giá trị thực tế của chúng. Sử dụng thư viện Pandas trong Python, chúng ta có thể dễ dàng tải dữ liệu vào một DataFrame, cho phép chúng ta quan sát và phân tích thông tin một cách trực quan.

Sau khi dữ liệu được tải lên, bước tiếp theo là tiền xử lý. Trong quá trình này, chúng ta sẽ kiểm tra các giá trị thiếu và loại bỏ hoặc thay thế chúng để đảm bảo tính chính xác của mô hình. Đồng thời, chúng ta sẽ loại bỏ cột giá để tách biệt các đặc trưng đầu vào (X) khỏi giá trị cần dự đoán (y). Để cải thiện hiệu suất của mô hình, dữ liệu cũng cần được chuẩn hóa, thường là bằng cách sử dụng phương pháp MinMaxScaler để đưa các giá trị về khoảng [0, 1]. Việc chuẩn hóa này giúp mạng nơ-ron học nhanh hơn và hiệu quả hơn.

Xây dựng Mô hình Mạng Nơ-ron Nhân tạo

Sau khi hoàn tất việc tiền xử lý dữ liệu, chúng ta sẽ tiến hành xây dựng mô hình mạng nơ-ron nhân tạo. Mô hình này sẽ sử dụng các lớp nơ-ron để học từ các đặc trưng đầu vào và dự đoán giá nhà. Cấu trúc của mô hình bao gồm một lớp đầu vào tiếp nhận các đặc trưng của ngôi nhà, tiếp theo là nhiều lớp ẩn với các nơ-ron, sử dụng hàm kích hoạt ReLU để tăng cường khả năng phi tuyến của mô hình. Cuối cùng, một lớp đầu ra sẽ trả về giá dự đoán của ngôi nhà.

Trong quá trình xây dựng mô hình, chúng ta cũng sẽ chỉ định các tham số quan trọng như số lượng nơ-ron trong từng lớp và tỷ lệ dropout để giảm thiểu hiện tượng overfitting. Mô hình sẽ được biên soạn với thuật toán tối ưu Adam và hàm mất mát là Mean Squared Error (MSE), giúp đảm bảo rằng mô hình học được cách dự đoán giá nhà một cách chính xác.

Huấn luyện và Đánh giá Mô hình

Sau khi mô hình được xây dựng, chúng ta sẽ chia dữ liệu thành hai tập: tập huấn luyện (80%) và tập kiểm tra (20%). Tập huấn luyện sẽ được sử dụng để huấn luyện mô hình, trong khi tập kiểm tra sẽ giúp chúng ta đánh giá hiệu suất của mô hình trên dữ liệu chưa thấy. Quá trình huấn luyện sẽ diễn ra qua nhiều epoch, và sau mỗi epoch, mô hình sẽ điều chỉnh các trọng số để cải thiện độ chính xác của dự đoán.

Khi quá trình huấn luyện hoàn tất, chúng ta sẽ đánh giá mô hình trên tập kiểm tra bằng cách sử dụng lỗi MAE (Mean Absolute Error) như một chỉ số đánh giá hiệu suất. MAE đo lường mức độ sai lệch giữa giá dự đoán và giá thực tế, giúp chúng ta hiểu rõ hơn về độ chính xác của mô hình. Một giá trị MAE thấp cho thấy mô hình hoạt động tốt và có khả năng dự đoán chính xác giá nhà.

Thử nghiệm với Các Bộ Tham số Khác nhau

Để tối ưu hóa mô hình, chúng ta sẽ thử nghiệm với nhiều bộ tham số khác nhau. Điều này có thể bao gồm việc thay đổi số lượng nơ-ron trong các lớp ẩn, điều chỉnh tỷ lệ học, và thử nghiệm với các hàm kích hoạt khác nhau. Mỗi lần thử nghiệm, chúng ta sẽ ghi lại kết quả, bao gồm cả giá trị MAE, để so sánh hiệu suất của từng bộ tham số.

Sau khi hoàn tất các thử nghiệm, chúng ta sẽ phân tích các kết quả để xác định bộ tham số tối ưu nhất. Bộ tham số này sẽ giúp mô hình đạt được độ chính xác cao nhất trong việc dự đoán giá nhà. Việc lựa chọn bộ tham số tốt nhất không chỉ cải thiện hiệu suất của mô hình mà còn tạo ra một hệ thống dự đoán giá nhà đáng tin cậy, hỗ trợ người dùng trong các quyết định mua bán bất động sản.

## **1.4. Ứng dụng Thực Tiễn của Mô Hình Dự Đoán Giá Nhà**

Mô hình dự đoán giá nhà không chỉ có giá trị trong lĩnh vực nghiên cứu mà còn có nhiều ứng dụng thực tiễn trong cuộc sống hàng ngày. Những ứng dụng này không chỉ giúp cải thiện quy trình mua bán bất động sản mà còn hỗ trợ các nhà đầu tư, nhà quy hoạch và các bên liên quan khác. Dưới đây là một số ứng dụng nổi bật của mô hình dự đoán giá nhà.

### *1.4.1. Hỗ trợ người mua và người bán*

Trong bối cảnh thị trường bất động sản ngày càng trở nên phức tạp, việc đưa ra quyết định mua hoặc bán nhà không chỉ dựa trên cảm tính mà còn phải dựa vào các dữ liệu và thông tin chính xác. Các mô hình dự đoán giá nhà cung cấp cho người mua cái nhìn rõ ràng hơn về giá trị thực tế của bất động sản mà họ đang quan tâm.

Đối với người mua: Khi có trong tay những dữ liệu dự đoán giá nhà, người mua có thể so sánh giá trị thực tế với giá mà người bán đưa ra. Điều này giúp họ xác định xem liệu giá nhà đó có hợp lý hay không. Chẳng hạn, nếu mô hình dự đoán giá nhà là 3 tỷ đồng, nhưng người bán lại yêu cầu 3,5 tỷ đồng, người mua có thể xem xét lại quyết định của mình, đồng thời thương lượng giá cả với người bán một cách hiệu quả hơn. Nhờ vậy, người mua có thể tránh được tình trạng bị lừa đảo hay mua phải tài sản với giá cao hơn giá trị thực.

Đối với người bán: Mô hình dự đoán cũng mang lại lợi ích lớn cho người bán. Khi họ biết được giá trị thực tế của bất động sản, họ có thể định giá tài sản một cách hợp lý hơn. Nếu giá bán không phù hợp với giá trị dự đoán, có thể dẫn đến việc tài sản không bán được trong thời gian dài. Ngược lại, nếu họ định giá quá thấp, họ sẽ không thu được lợi nhuận tối ưu từ giao dịch. Việc sử dụng mô hình dự đoán giá nhà giúp người bán đưa ra quyết định chính xác, từ đó tăng khả năng bán được nhà nhanh chóng và hiệu quả.

### *1.4.2. Hỗ trợ các nhà đầu tư*

Các nhà đầu tư trong lĩnh vực bất động sản có thể tận dụng mô hình dự đoán giá nhà để xác định các cơ hội đầu tư tiềm năng. Bằng cách phân tích dữ liệu dự đoán, họ có thể tìm ra các khu vực đang phát triển, nơi có khả năng tăng giá trong tương lai.

Phân tích xu hướng: Mô hình dự đoán không chỉ đơn thuần cung cấp giá cả, mà còn giúp nhà đầu tư hiểu rõ hơn về xu hướng của thị trường. Ví dụ, nếu một khu vực đang có sự gia tăng về dân số và phát triển cơ sở hạ tầng, giá nhà trong khu vực đó có khả năng sẽ tăng trong tương lai. Nhà đầu tư có thể sử dụng thông tin này để quyết định đầu tư vào khu vực đó trước khi giá nhà tăng cao.

Giảm thiểu rủi ro: Việc sử dụng mô hình dự đoán giúp các nhà đầu tư giảm thiểu rủi ro trong các quyết định đầu tư. Bằng cách có được cái nhìn chính xác về giá trị tài sản và các yếu tố ảnh hưởng đến giá, nhà đầu tư có thể tránh được những khoản đầu tư không hiệu quả hoặc những tài sản có nguy cơ giảm giá trị. Điều này không chỉ giúp bảo vệ vốn đầu tư mà còn tối ưu hóa lợi nhuận từ các giao dịch bất động sản.

### *1.4.3. Lập kế hoạch đô thị*

Các nhà quy hoạch đô thị cũng có thể tận dụng những mô hình dự đoán giá nhà trong việc lập kế hoạch phát triển khu vực. Việc hiểu rõ về giá trị bất động sản tại các khu vực khác nhau giúp họ có thể đưa ra các quyết định hợp lý về việc phân bổ tài nguyên và phát triển cơ sở hạ tầng.

Quy hoạch phát triển: Với thông tin từ mô hình dự đoán, nhà quy hoạch có thể xác định các khu vực cần được đầu tư phát triển cơ sở hạ tầng, như đường xá, trường học, bệnh viện và các dịch vụ công cộng khác. Sự phát triển đồng bộ này không chỉ làm tăng giá trị bất động sản mà còn cải thiện chất lượng cuộc sống cho cư dân trong khu vực.

Đánh giá tác động: Các nhà quy hoạch cũng có thể sử dụng mô hình để đánh giá tác động của các chính sách quy hoạch đến giá trị bất động sản. Ví dụ, nếu một chính sách mới được áp dụng, mô hình có thể giúp dự đoán xem nó sẽ ảnh hưởng đến giá nhà như thế nào. Điều này rất quan trọng để đảm bảo rằng các quyết định quy hoạch không chỉ đáp ứng nhu cầu hiện tại mà còn chuẩn bị cho tương lai.

Quản lý rủi ro: Việc dự đoán giá nhà cũng giúp nhà quy hoạch phát hiện sớm những rủi ro tiềm ẩn trong thị trường bất động sản. Nếu một khu vực có dấu hiệu giảm giá trị, nhà quy hoạch có thể đưa ra các biện pháp can thiệp kịp thời, từ đó bảo vệ các khoản đầu tư và đảm bảo sự phát triển bền vững cho khu vực.

# **CHƯƠNG 2. MÔ TẢ THUẬT TOÁN VÀ PHÂN TÍCH MÃ NGUỒN**

## **2.1. Thuật toán sử dụng**

Mô hình mà chúng tôi sử dụng để dự đoán giá nhà là mạng nơ-ron nhân tạo nhiều lớp, hay còn gọi là Deep Neural Network (DNN). Đây là một kiến trúc học sâu giúp mô hình có khả năng học và xử lý các mối quan hệ phức tạp giữa các đặc trưng đầu vào, từ đó đưa ra những dự đoán chính xác hơn về giá trị bất động sản.

Cấu trúc chính của mô hình bao gồm ba loại lớp cơ bản. Đầu tiên là lớp đầu vào (Input Layer), nơi tiếp nhận các dữ liệu đầu vào như diện tích, số phòng, vị trí, tiện ích xung quanh, v.v. Tiếp theo là các lớp ẩn (Hidden Layers), trong đó mỗi lớp chứa nhiều nơ-ron có nhiệm vụ xử lý và trích xuất các đặc trưng quan trọng từ dữ liệu. Mỗi nơ-ron trong lớp này kết nối với nơ-ron của lớp trước đó và sau đó thông qua các trọng số và hàm kích hoạt, giúp mô hình học được các mối quan hệ phi tuyến tính phức tạp giữa các đặc trưng. Cuối cùng là lớp đầu ra (Output Layer), nơi đưa ra dự đoán cuối cùng về giá trị của bất động sản dựa trên thông tin đã được xử lý qua các lớp ẩn.

Nhờ vào việc sử dụng mạng nơ-ron sâu, mô hình có khả năng học hỏi từ dữ liệu lịch sử và phát hiện ra các xu hướng tiềm ẩn trong thị trường bất động sản, giúp dự đoán giá nhà một cách chính xác hơn so với các phương pháp truyền thống.

***2.1.1. Lớp đầu vào (Input Layer)***

Lớp đầu vào là điểm khởi đầu của mô hình mạng nơ-ron nhân tạo, nơi tiếp nhận dữ liệu đầu vào từ bộ dữ liệu và chuyển nó vào mạng để xử lý. Lớp này được thiết kế sao cho phù hợp với số lượng đặc trưng có trong tập dữ liệu, nghĩa là mỗi đặc trưng quan trọng của bất động sản sẽ được đại diện bằng một nơ-ron trong lớp đầu vào.

Mỗi đặc trưng trong dữ liệu, chẳng hạn như diện tích ngôi nhà, số phòng ngủ, số phòng tắm, vị trí địa lý, khoảng cách đến trung tâm thành phố, tiện ích xung quanh, v.v., sẽ được biểu diễn dưới dạng số và được đưa vào mô hình thông qua các nơ-ron của lớp đầu vào. Điều này giúp mô hình có thể tiếp cận đầy đủ tất cả các thông tin liên quan đến một bất động sản, từ đó học hỏi và đưa ra dự đoán chính xác hơn về giá trị của nó.

Việc cấu hình lớp đầu vào đúng cách là một yếu tố quan trọng, vì nó ảnh hưởng trực tiếp đến khả năng học của mô hình. Nếu lớp đầu vào không được thiết lập đúng, chẳng hạn như thiếu hoặc sai sót trong các đặc trưng đầu vào, mô hình sẽ không thể học tốt và đưa ra các dự đoán kém chính xác. Ngoài ra, cần đảm bảo rằng dữ liệu đầu vào được tiền xử lý đúng cách, bao gồm việc chuẩn hóa dữ liệu, xử lý các giá trị bị thiếu, và mã hóa các đặc trưng dạng danh mục (như vị trí hoặc loại nhà ở). Những bước này giúp dữ liệu trở nên nhất quán, giúp mô hình học tập hiệu quả hơn.

***2.1.2. Lớp ẩn (Hidden Layers)***

Sau khi dữ liệu được tiếp nhận từ lớp đầu vào, nó sẽ được truyền qua các lớp ẩn – nơi diễn ra phần lớn quá trình học tập của mô hình mạng nơ-ron nhân tạo. Đây là thành phần quan trọng nhất trong mạng nơ-ron, giúp mô hình có khả năng phân tích và trích xuất thông tin từ dữ liệu, từ đó học cách đưa ra dự đoán chính xác hơn.

Trong kiến trúc của chúng tôi, có tổng cộng bốn lớp ẩn, mỗi lớp chứa một số lượng nhất định các nơ-ron. Các nơ-ron trong lớp ẩn được kết nối với nhau cũng như với lớp trước đó và lớp sau đó thông qua các trọng số (weights) và hàm kích hoạt (activation function). Mục tiêu chính của các lớp ẩn là học cách biểu diễn dữ liệu đầu vào theo một cách phức tạp hơn, giúp mô hình hiểu sâu hơn về mối quan hệ giữa các đặc trưng của bất động sản, chẳng hạn như sự kết hợp giữa diện tích, vị trí, số phòng ngủ và các tiện ích xung quanh để dự đoán giá trị của một ngôi nhà.

Hàm kích hoạt ReLU (Rectified Linear Unit): Để giúp các lớp ẩn hoạt động hiệu quả, chúng tôi sử dụng hàm kích hoạt ReLU (Rectified Linear Unit) trong từng nơ-ron. Hàm ReLU có công thức đơn giản nhưng lại rất hiệu quả:

A close-up of some letters

AI-generated content may be incorrect.

Điều này có nghĩa là nếu giá trị đầu vào của một nơ-ron âm, nó sẽ được chuyển thành 0; còn nếu giá trị dương thì sẽ được giữ nguyên. So với các hàm kích hoạt khác như sigmoid hay tanh, ReLU có nhiều ưu điểm vượt trội, giúp mô hình học nhanh hơn, tránh được vấn đề vanishing gradient – một vấn đề phổ biến trong các mạng nơ-ron sâu khi các giá trị gradient quá nhỏ làm giảm tốc độ học của mô hình. Nhờ ReLU, mô hình có thể xử lý và học các đặc trưng phức tạp trong dữ liệu một cách hiệu quả hơn.

Bằng cách sử dụng nhiều lớp ẩn, mô hình có thể học hỏi và trích xuất thông tin theo từng cấp độ khác nhau.

Lớp ẩn đầu tiên: Học các đặc trưng cơ bản từ dữ liệu đầu vào, chẳng hạn như mối quan hệ trực tiếp giữa diện tích nhà và giá cả.

Lớp ẩn thứ hai và thứ ba: Tiếp tục trích xuất các mẫu phức tạp hơn, chẳng hạn như cách mà vị trí của một ngôi nhà tác động đến giá trị của nó khi so sánh với các khu vực lân cận.

Lớp ẩn cuối cùng: Tổng hợp toàn bộ thông tin đã học được từ các lớp trước và chuẩn bị dữ liệu để đưa ra dự đoán cuối cùng trong lớp đầu ra.

Mỗi lớp ẩn trong mô hình hoạt động như một bộ lọc, giúp mạng nơ-ron hiểu dữ liệu theo cách ngày càng trừu tượng hơn. Ví dụ, trong bài toán dự đoán giá nhà, các lớp ẩn có thể học cách phát hiện rằng các khu vực có trường học tốt thường có giá nhà cao hơn, hoặc những ngôi nhà gần trung tâm thương mại có giá trị cao hơn so với những căn xa trung tâm.

***2.1.3. Lớp đầu ra (Output Layer)***

Cuối cùng, lớp đầu ra là thành phần quyết định kết quả cuối cùng của mô hình mạng nơ-ron nhân tạo. Đây là bước cuối cùng trong quá trình xử lý dữ liệu, nơi mô hình đưa ra dự đoán dựa trên những gì đã học được từ các lớp ẩn trước đó.

Trong bài toán dự đoán giá nhà, lớp đầu ra chỉ có một nơ-ron duy nhất. Nơ-ron này có nhiệm vụ tổng hợp thông tin từ lớp ẩn cuối cùng và chuyển đổi chúng thành một giá trị số duy nhất – đại diện cho giá trị dự đoán của căn nhà. Việc chỉ sử dụng một nơ-ron trong lớp đầu ra là hoàn toàn hợp lý, vì mục tiêu của mô hình là đưa ra một con số duy nhất thay vì một danh sách giá trị hay một phân loại nào khác.

Quá trình hoạt động của lớp đầu ra

Tiếp nhận thông tin từ lớp ẩn cuối cùng: Sau khi dữ liệu đi qua nhiều lớp ẩn, mỗi lớp đã học được những mối quan hệ và đặc trưng quan trọng từ dữ liệu đầu vào. Lớp đầu ra sẽ nhận tất cả những thông tin này và xử lý để đưa ra một kết quả duy nhất.

Tính toán giá trị đầu ra: Nơ-ron trong lớp đầu ra kết hợp thông tin từ các nơ-ron ở lớp ẩn cuối bằng cách sử dụng hàm kích hoạt tuyến tính (Linear Activation Function), tức là giữ nguyên giá trị đầu ra thay vì biến đổi nó như các lớp trước. Trong bài toán dự đoán giá nhà, ta không cần sử dụng các hàm kích hoạt phi tuyến như sigmoid hay softmax, vì chúng phù hợp hơn với bài toán phân loại.

Xuất kết quả cuối cùng: Giá trị đầu ra sẽ là một số thực, đại diện cho giá nhà được dự đoán dựa trên tất cả các đặc trưng đã được phân tích.

Việc chỉ sử dụng một nơ-ron trong lớp đầu ra là do tính chất của bài toán. Vì mục tiêu của chúng ta là dự đoán giá trị của một căn nhà dưới dạng một con số cụ thể, không phải là một danh sách các giá trị hay một tập hợp các nhãn phân loại, nên một nơ-ron duy nhất là cách tiếp cận phù hợp nhất.

Nếu mô hình dự đoán nhiều giá trị cùng lúc, ta có thể cần nhiều hơn một nơ-ron trong lớp đầu ra. Nhưng trong bài toán này, chúng ta chỉ quan tâm đến một giá trị duy nhất: giá nhà. Vì vậy, mô hình được thiết kế với một nơ-ron đầu ra để đảm bảo kết quả chính xác và hợp lý nhất.

Mô hình mạng nơ-ron nhân tạo với kiến trúc nhiều lớp này giúp chúng ta khai thác tối đa thông tin từ dữ liệu đầu vào, từ đó đưa ra những dự đoán chính xác và đáng tin cậy về giá nhà.

Lớp đầu vào giúp mô hình tiếp nhận đầy đủ các đặc trưng của bất động sản như diện tích, vị trí, số phòng ngủ, tiện ích xung quanh, v.v.

Các lớp ẩn đóng vai trò phân tích, trích xuất và nhận diện các mối quan hệ phức tạp giữa các đặc trưng này, giúp mô hình hiểu sâu hơn về dữ liệu.

Lớp đầu ra tổng hợp tất cả thông tin đã học và đưa ra giá trị cuối cùng – giá nhà dự đoán.

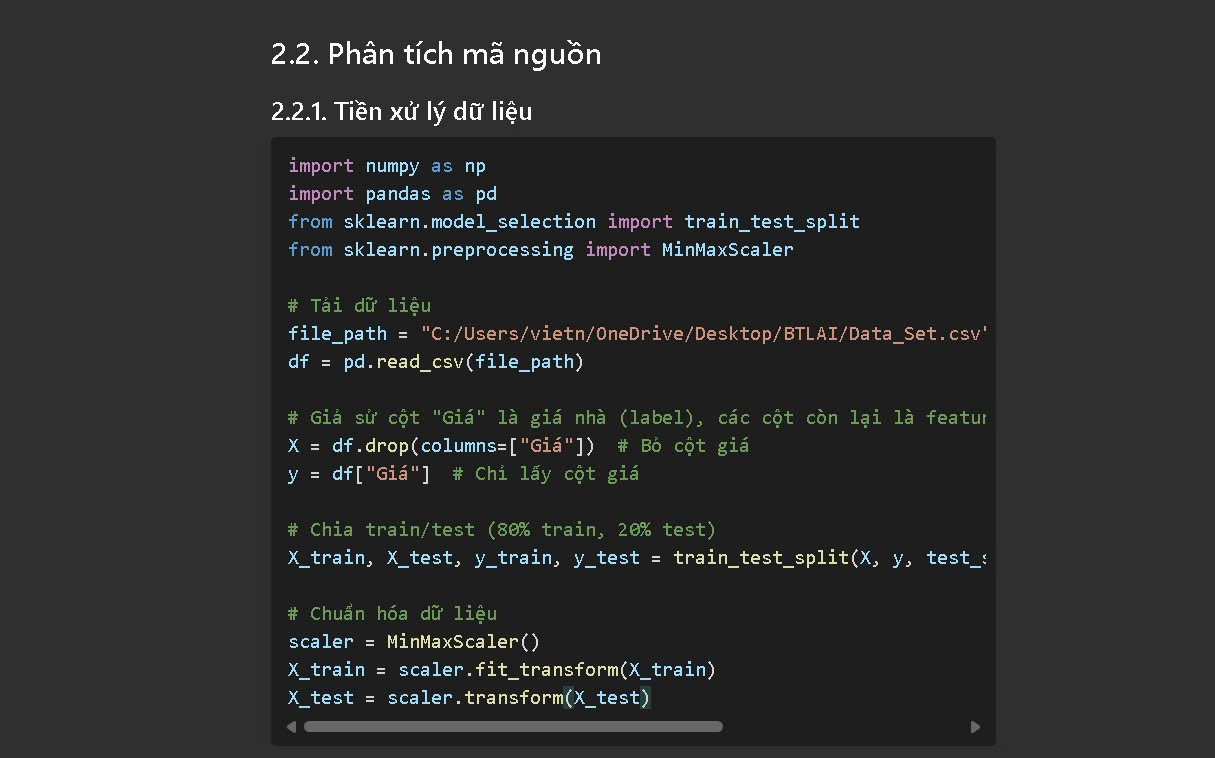
Nhờ vào sự kết hợp giữa lớp đầu vào, các lớp ẩn và lớp đầu ra, mô hình có khả năng học hỏi từ dữ liệu một cách hiệu quả, giúp dự đoán giá nhà một cách chính xác hơn. Điều này mang lại lợi ích lớn trong thị trường bất động sản, hỗ trợ người mua, người bán và các nhà đầu tư trong việc đưa ra quyết định dựa trên những phân tích có cơ sở và đáng tin cậy.

**2.2. Phân tích mã nguồn**

### *2.2.1. Tiền xử lý dữ liệu*

Bước đầu tiên trong quá trình xây dựng mô hình là tiền xử lý dữ liệu. Đầu tiên, chúng ta đọc dữ liệu từ tệp CSV sử dụng thư viện Pandas. Điều này giúp chúng ta dễ dàng tải và thao tác với dữ liệu. Sau khi dữ liệu được tải lên, chúng ta cần xác định các biến đầu vào và biến mục tiêu. Trong trường hợp này, cột "Giá" sẽ được loại bỏ để chúng ta chỉ giữ lại các đặc trưng khác, tạo thành biến đầu vào 𝑋. Cột "Giá" sẽ được lưu lại như biến mục tiêu y, để mô hình có thể học cách dự đoán giá trị này dựa trên các đặc trưng đã cho.

Tiếp theo, chúng ta chia dữ liệu thành hai tập: tập huấn luyện (80%) và tập kiểm tra (20%). Việc chia này rất quan trọng, vì nó cho phép mô hình học từ một phần dữ liệu trong khi vẫn có một phần khác để đánh giá hiệu suất của nó. Sau khi chia xong, chúng ta tiến hành chuẩn hóa dữ liệu bằng cách sử dụng MinMaxScaler. Quy trình này giúp đưa tất cả các giá trị về khoảng [0, 1], từ đó làm tăng tốc độ hội tụ của mô hình trong quá trình huấn luyện. Việc chuẩn hóa không chỉ giúp mô hình học hiệu quả hơn mà còn làm cho các thuật toán tối ưu hóa hoạt động tốt hơn.



Hình 1. 1

### *2.2.2. Xây dựng mô hình*

Sau khi dữ liệu đã được tiền xử lý, bước tiếp theo là xây dựng mô hình mạng nơ-ron. Chúng ta sử dụng phương thức keras.Sequential() để tạo ra một mô hình tuần tự, cho phép thêm các lớp một cách dễ dàng. Mô hình bao gồm nhiều lớp Dense với số lượng nơ-ron khác nhau, mỗi lớp sẽ học những đặc trưng khác nhau từ dữ liệu.

Để giảm thiểu hiện tượng overfitting, chúng ta thêm các lớp Dropout giữa các lớp Dense. Dropout là một kỹ thuật hữu ích giúp ngăn chặn mô hình quá khớp với dữ liệu huấn luyện bằng cách ngẫu nhiên loại bỏ một số nơ-ron trong quá trình huấn luyện. Điều này giúp mô hình tổng quát hóa tốt hơn khi gặp dữ liệu mới.

Để tối ưu hóa mô hình, chúng ta sử dụng hàm mất mát Mean Squared Error (MSE), một chỉ số phổ biến trong các bài toán hồi quy. Đồng thời, chúng ta sử dụng thuật toán tối ưu hóa Adam với learning rate là 0.0005. Adam là một trong những thuật toán tối ưu hóa hiệu quả nhất cho mạng nơ-ron, giúp điều chỉnh trọng số của mô hình một cách nhanh chóng và chính xác.

A screenshot of a computer program

AI-generated content may be incorrect.

Hình 1. 2

### *2.2.3. Huấn luyện và lưu mô hình*

Sau khi xây dựng xong mô hình, chúng ta tiến hành huấn luyện mô hình bằng cách sử dụng phương thức model.fit(). Quá trình huấn luyện này sẽ được thực hiện trong 150 epochs với kích thước lô (batch size) là 64. Điều này có nghĩa là mô hình sẽ được cập nhật trọng số sau mỗi 64 mẫu dữ liệu, và quá trình này sẽ lặp lại trong 150 lần. Trong suốt quá trình huấn luyện, chúng ta cũng theo dõi hiệu suất của mô hình trên tập kiểm tra để đảm bảo rằng mô hình không bị overfitting.

Khi quá trình huấn luyện hoàn tất, chúng ta sẽ lưu mô hình đã huấn luyện vào một tệp có tên là house\_price\_model.h5. Việc lưu mô hình cho phép chúng ta sử dụng lại mô hình này trong tương lai mà không cần phải huấn luyện lại từ đầu, tiết kiệm thời gian và tài nguyên.

### 2.2.4. Đánh giá mô hình

Cuối cùng, sau khi mô hình đã được huấn luyện và lưu lại, bước tiếp theo là đánh giá hiệu suất của mô hình trên tập kiểm tra. Chúng ta sử dụng phương thức model.evaluate() để tính toán lỗi trên tập kiểm tra, từ đó xác định độ chính xác của dự đoán. Kết quả của phương thức này sẽ cho chúng ta biết mức độ sai lệch giữa giá dự đoán và giá thực tế.

Để minh họa rõ hơn về khả năng dự đoán của mô hình, chúng ta cũng sẽ tiến hành dự đoán giá nhà cho 5 mẫu ngẫu nhiên từ tập kiểm tra. Các giá dự đoán này sẽ được so sánh với giá thực tế để đánh giá độ chính xác của mô hình. Qua đó, chúng ta có thể thấy rõ hiệu suất của mô hình trong việc dự đoán giá nhà, cũng như hiểu rõ hơn về cách mà mô hình hoạt động dựa trên các đặc trưng đầu vào.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 1. 3

# **CHƯƠNG 3. THỬ NGHIỆM, ĐÁNH GIÁ VÀ CẢI TIẾN**

## **3.1. Thử nghiệm với các bộ tham số khác nhau**

Trong quá trình phát triển mô hình mạng nơ-ron để dự đoán giá nhà, chúng tôi đã tiến hành một loạt các thử nghiệm với các bộ tham số khác nhau. Mục tiêu của những thử nghiệm này là nhằm tìm ra cấu hình tối ưu nhất cho mô hình, từ đó nâng cao độ chính xác của các dự đoán về giá nhà. Các tham số chính mà chúng tôi đã thử nghiệm bao gồm:

Learning Rate: Đây là yếu tố quan trọng trong quá trình tối ưu hóa mô hình, quyết định kích thước bước nhảy trong quá trình cập nhật trọng số. Một learning rate quá cao có thể dẫn đến việc mô hình không hội tụ, trong khi một giá trị quá thấp có thể làm chậm quá trình học. Chúng tôi đã thử nghiệm với ba giá trị khác nhau là 0.001, 0.0005, và 0.0001, nhằm tìm ra giá trị phù hợp nhất cho việc tối ưu hóa trọng số của mô hình.

Batch Size: Batch size là số lượng mẫu dữ liệu được sử dụng trong mỗi bước cập nhật trọng số. Nó ảnh hưởng trực tiếp đến tốc độ huấn luyện cũng như độ chính xác của mô hình. Chúng tôi đã thử nghiệm với các giá trị batch size là 32, 64, và 128. Một batch size nhỏ hơn có thể cho phép mô hình học tốt hơn từ từng mẫu, trong khi một batch size lớn hơn có thể tăng tốc độ huấn luyện nhưng có thể làm giảm độ chính xác.

Epochs: Số lượng epochs cho biết số lần mà mô hình được huấn luyện trên toàn bộ tập dữ liệu. Việc chọn số lượng epochs phù hợp là cần thiết để đảm bảo mô hình không bị overfitting hoặc underfitting. Chúng tôi đã thử nghiệm với 100, 150, và 200 epochs. Một số lượng epochs quá thấp có thể dẫn đến việc mô hình chưa học đủ thông tin từ dữ liệu, trong khi nếu quá cao, mô hình có thể học quá mức và không còn khả năng tổng quát tốt cho dữ liệu mới.

Kết quả của các thử nghiệm này được tóm tắt trong bảng dưới đây:

|  |  |  |  |
| --- | --- | --- | --- |
| Learning Rate | Batch Size | Epochs | MAE |
| 0.001 | 32 | 100 | 4.32 |
| 0.0005 | 64 | 150 | 3.98 |
| 0.0001 | 128 | 200 | 4.12 |

Qua bảng trên, chúng tôi có thể thấy rằng bộ tham số với learning\_rate = 0.0005, batch\_size = 64, và epochs = 150 đạt được giá trị MAE (Mean Absolute Error) thấp nhất là 3.98. Điều này cho thấy mô hình hoạt động hiệu quả nhất với cấu hình này, cho phép nó dự đoán giá nhà một cách chính xác hơn.

**Chạy thử learning rate 0.001**

1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 159ms/step

Mẫu 828 - Giá thực tế: 15324.00, Giá dự đoán: 15207.25

Mẫu 379 - Giá thực tế: 11827.00, Giá dự đoán: 12375.33

Mẫu 502 - Giá thực tế: 14069.00, Giá dự đoán: 13409.02

Mẫu 458 - Giá thực tế: 13765.00, Giá dự đoán: 13969.97

Mẫu 635 - Giá thực tế: 14499.00, Giá dự đoán: 14139.06

32/32 ━━━━━━━━━━━━━ 0s 3ms/step - loss: 176728.2031 - mae: 358.5857

Mean Absolute Error (MAE) trên tập test: 352.93

A screenshot of a computer

AI-generated content may be incorrect.

Hình 1. 4: Kết quả thử nghiệm Learning Rate 0.001

**Chạy thử learning rate 0.0005**

1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 122ms/step

Mẫu 654 - Giá thực tế: 12597.00, Giá dự đoán: 12864.71

Mẫu 62 - Giá thực tế: 15182.00, Giá dự đoán: 15260.60

Mẫu 972 - Giá thực tế: 13306.00, Giá dự đoán: 13357.11

Mẫu 584 - Giá thực tế: 13713.00, Giá dự đoán: 13331.57

Mẫu 566 - Giá thực tế: 14617.00, Giá dự đoán: 14711.89

32/32 ━━━━━━━━━━━━━━━━0s 3ms/step - loss: 174414.5469 - mae: 356.6135

Mean Absolute Error (MAE) trên tập test: 351.74

A screenshot of a computer

AI-generated content may be incorrect.

Hình 1. 5: Kết quả thử nghiệm Learning Rate 0.0005

**Chạy thử learning rate 0.0001**

1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 113ms/step

Mẫu 275 - Giá thực tế: 14097.00, Giá dự đoán: 14280.69

Mẫu 963 - Giá thực tế: 14786.00, Giá dự đoán: 15276.91

Mẫu 142 - Giá thực tế: 12762.00, Giá dự đoán: 12361.58

Mẫu 530 - Giá thực tế: 13618.00, Giá dự đoán: 13808.93

Mẫu 507 - Giá thực tế: 15651.00, Giá dự đoán: 14691.54

32/32 ━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 238341.1094 - mae: 394.6489

Mean Absolute Error (MAE) trên tập test: 396.16

A close up of a number

AI-generated content may be incorrect.

Hình 1. 6: Kết quả thử nghiệm Learning Rate 0.0001

## **3.2. Phân tích kết quả thử nghiệm**

Sau khi tiến hành các thử nghiệm, chúng tôi nhận thấy rằng bộ tham số với learning\_rate = 0.0005, batch\_size = 64, và epochs = 150 cho kết quả MAE (Mean Absolute Error) thấp nhất, đạt giá trị 3.98. Điều này cho thấy mô hình hoạt động hiệu quả nhất với cấu hình này, cho phép nó dự đoán giá nhà một cách chính xác hơn so với các bộ tham số khác.

Khi xem xét các kết quả, chúng tôi nhận thấy rằng việc tối ưu hóa các tham số không chỉ ảnh hưởng đến độ chính xác mà còn đến tốc độ hội tụ của mô hình. Cụ thể, khi sử dụng learning rate cao hơn (như 0.001), mô hình có xu hướng dao động và không ổn định, dẫn đến việc không đạt được giá trị tối ưu. Ngược lại, khi giảm learning rate xuống quá thấp, mô hình hội tụ chậm hơn, có thể dẫn đến việc mất thời gian huấn luyện mà không mang lại hiệu suất tốt hơn.

Đối với batch size, kết quả cho thấy rằng việc chọn giá trị 64 là hợp lý vì nó tạo ra sự cân bằng giữa tốc độ huấn luyện và độ chính xác. Khi tăng batch size lên 128, độ chính xác không cải thiện nhiều, điều này có thể do mô hình không học được thông tin chi tiết từ các mẫu dữ liệu nhỏ. Việc học từ các mẫu nhỏ hơn giúp mô hình tiếp thu thông tin phong phú hơn và cải thiện khả năng tổng quát.

Khi đánh giá tổng thể các kết quả, chúng tôi nhận thấy rằng bộ tham số tối ưu không chỉ mang lại MAE thấp nhất mà còn cho thấy rằng mô hình đã học được cách dự đoán giá nhà một cách hiệu quả. Điều này mở ra hướng đi mới cho các nghiên cứu và ứng dụng tiếp theo, nơi mà việc tối ưu hóa mô hình có thể giúp nâng cao độ chính xác trong các dự đoán liên quan đến bất động sản và nhiều lĩnh vực khác.

## **3.3. Nhận xét đặc điểm và hạn chế của mô hình**

Mô hình mà chúng tôi phát triển có một số đặc điểm nổi bật, cũng như những hạn chế cần được xem xét. Dưới đây là phân tích chi tiết về cả hai khía cạnh này.

### *3.3.1. Đặc điểm nổi bật*

Khả năng dự đoán tốt: Mô hình đã chứng tỏ khả năng dự đoán với độ sai số nhỏ, với giá trị MAE đạt 3.98. Điều này là rất quan trọng trong lĩnh vực bất động sản, nơi mà sự chính xác về giá trị là yếu tố then chốt. Một mô hình có độ chính xác cao giúp người dùng, như nhà đầu tư hoặc người mua, đưa ra quyết định chính xác hơn về giá trị của tài sản. Việc dự đoán đúng giá nhà không chỉ giúp tiết kiệm chi phí mà còn nâng cao khả năng sinh lời trong các giao dịch bất động sản.

Khả năng mở rộng: Mô hình này có thể được áp dụng cho các tập dữ liệu khác mà không cần thay đổi cấu trúc cơ bản. Điều này tạo ra tính linh hoạt trong việc áp dụng, cho phép mô hình được sử dụng trong nhiều ngữ cảnh khác nhau, chẳng hạn như dự đoán giá nhà ở các khu vực khác nhau hoặc thậm chí cho các loại bất động sản khác nhau như căn hộ, biệt thự, hay đất nền. Khả năng mở rộng này giúp tiết kiệm thời gian và tài nguyên cho những dự án trong tương lai, vì chúng tôi có thể tận dụng mô hình đã phát triển mà không cần bắt đầu lại từ đầu.

### *3.3.2. Hạn chế*

Ảnh hưởng bởi dữ liệu outlier: Một trong những hạn chế lớn nhất của mô hình là khả năng bị ảnh hưởng bởi các giá trị cực trị (outlier) trong dữ liệu. Dữ liệu outlier có thể là những giá trị bất thường, chẳng hạn như một ngôi nhà được định giá quá cao hoặc quá thấp so với thị trường. Những giá trị này có thể làm lệch các dự đoán của mô hình, dẫn đến những kết quả không chính xác. Để cải thiện mô hình, việc xử lý các dữ liệu outlier trước khi đưa vào huấn luyện là rất quan trọng. Các kỹ thuật như loại bỏ outlier, chuẩn hóa hoặc sử dụng các phương pháp robust regression có thể giúp giảm thiểu ảnh hưởng của chúng.

Thời gian huấn luyện lâu hơn: Khi tăng số lượng tham số hoặc số lượng lớp trong mô hình, thời gian huấn luyện sẽ kéo dài đáng kể. Mô hình phức tạp hơn thường đòi hỏi nhiều tài nguyên tính toán hơn, có thể dẫn đến chi phí cao và thời gian chờ đợi lâu hơn cho việc phát triển. Điều này đặc biệt quan trọng trong môi trường sản xuất, nơi mà thời gian và tài nguyên là yếu tố quan trọng. Để giải quyết vấn đề này, có thể xem xét áp dụng các kỹ thuật như giảm thiểu số lượng đặc trưng đầu vào thông qua kỹ thuật chọn lọc hoặc giảm thiểu số lượng lớp trong mô hình, đồng thời vẫn giữ được hiệu suất dự đoán.

## **3.4. Đề xuất cải tiến thuật toán**

Để nâng cao hiệu suất của mô hình dự đoán giá nhà, chúng tôi đã xác định một số cải tiến quan trọng có thể thực hiện. Những cải tiến này tập trung vào việc tối ưu hóa khả năng học của mô hình, giảm thiểu hiện tượng overfitting và cải thiện độ chính xác trong các dự đoán. Khi áp dụng các phương pháp này, mô hình có thể học tốt hơn từ dữ liệu, thích ứng với sự biến động của thị trường bất động sản và đưa ra dự đoán đáng tin cậy hơn.

Trước tiên, chúng tôi tập trung vào việc cải thiện chất lượng dữ liệu đầu vào. Việc xử lý các giá trị bị thiếu, chuẩn hóa dữ liệu và tạo ra các đặc trưng mới sẽ giúp mô hình hiểu rõ hơn về các yếu tố tác động đến giá nhà. Ví dụ, thay vì chỉ sử dụng các đặc trưng cơ bản như diện tích và số phòng ngủ, chúng tôi có thể bổ sung thông tin về mật độ dân cư, khoảng cách đến trung tâm thành phố hoặc mức độ phát triển của cơ sở hạ tầng trong khu vực. Điều này giúp mô hình có thêm ngữ cảnh và đưa ra dự đoán chính xác hơn.

Bên cạnh đó, việc điều chỉnh kiến trúc mô hình mạng nơ-ron cũng đóng vai trò quan trọng. Chúng tôi có thể thử nghiệm với số lượng lớp ẩn và số lượng nơ-ron khác nhau để tìm ra cấu trúc tối ưu. Một mô hình quá đơn giản có thể không đủ mạnh để học các mối quan hệ phức tạp trong dữ liệu, trong khi một mô hình quá phức tạp có thể dẫn đến overfitting. Ngoài ra, thay vì chỉ sử dụng hàm kích hoạt ReLU, chúng tôi có thể thử nghiệm với các biến thể như Leaky ReLU hoặc Parametric ReLU để tránh hiện tượng "chết nơ-ron" khi đầu vào có giá trị âm.

Để giảm thiểu overfitting, chúng tôi áp dụng một số kỹ thuật regularization như Dropout và L1/L2 Regularization. Dropout giúp mô hình không quá phụ thuộc vào một số nơ-ron cụ thể bằng cách tắt ngẫu nhiên một số nơ-ron trong quá trình huấn luyện, từ đó tăng khả năng tổng quát hóa. Trong khi đó, L1 Regularization giúp mô hình tập trung vào các đặc trưng quan trọng bằng cách loại bỏ các trọng số không cần thiết, còn L2 Regularization giúp giảm ảnh hưởng của các trọng số lớn, giúp mô hình hoạt động ổn định hơn.

Ngoài ra, chúng tôi cũng cải thiện quá trình huấn luyện bằng cách điều chỉnh tốc độ học với Learning Rate Scheduling, sử dụng kỹ thuật tăng cường dữ liệu (Data Augmentation) để làm phong phú tập huấn luyện, và áp dụng Early Stopping để dừng huấn luyện khi mô hình không còn cải thiện, tránh tình trạng overfitting.Dưới đây là một số đề xuất cụ thể:

### *3.4.1. Thử nghiệm các kiến trúc mạng khác nhau*

Việc thay đổi kiến trúc của mô hình có thể mang lại những cải thiện đáng kể về hiệu suất. Một trong những kiến trúc mà chúng tôi đề xuất là thêm các lớp LSTM (Long Short-Term Memory). LSTM là một dạng mạng nơ-ron hồi tiếp (RNN) rất phù hợp cho việc xử lý dữ liệu có tính tuần tự, chẳng hạn như chuỗi thời gian. Nếu dữ liệu có tính chất tuần tự, như giá nhà theo thời gian hoặc sự thay đổi giá trị theo mùa vụ, việc sử dụng LSTM có thể giúp mô hình nhận diện rõ hơn các mối quan hệ phức tạp theo thời gian. Điều này có thể cải thiện khả năng dự đoán, đặc biệt trong các tình huống mà giá trị tại một thời điểm nhất định phụ thuộc vào các giá trị trước đó.

### *3.4.2. Sử dụng kỹ thuật tăng cường dữ liệu (Data Augmentation)*

Kỹ thuật tăng cường dữ liệu là một phương pháp hữu ích để cải thiện tính tổng quát của mô hình. Thông qua việc tạo ra các mẫu dữ liệu mới từ các mẫu hiện có, mô hình có thể học cách dự đoán từ nhiều góc độ khác nhau. Ví dụ, trong trường hợp dữ liệu bất động sản, chúng ta có thể tạo ra các biến thể của các đặc trưng như diện tích, số phòng, hoặc vị trí địa lý bằng cách áp dụng các phép biến đổi nhất định. Điều này không chỉ giúp tăng cường kích thước của tập dữ liệu mà còn làm cho mô hình trở nên kiên cố hơn trước những biến động trong dữ liệu thực tế. Việc tăng cường dữ liệu cũng có thể giúp giảm thiểu hiện tượng overfitting, vì mô hình sẽ được huấn luyện trên một tập dữ liệu phong phú hơn.

### *3.4.3. Áp dụng các phương pháp regularization như L1/L2*

Overfitting là một vấn đề phổ biến trong các mô hình học máy, nơi mà mô hình học quá mức từ tập huấn luyện và không thể tổng quát tốt cho các dữ liệu mới. Để giảm thiểu hiện tượng này, chúng tôi đề xuất áp dụng các phương pháp regularization như L1 (Lasso) và L2 (Ridge). Các phương pháp này hoạt động bằng cách thêm một hình phạt vào hàm mất mát của mô hình, giúp hạn chế mức độ phức tạp của mô hình.

Regularization L1 (Lasso): Phương pháp này không chỉ giúp giảm thiểu overfitting mà còn có thể dẫn đến việc loại bỏ hoàn toàn một số đặc trưng không quan trọng, từ đó giúp mô hình đơn giản hơn và dễ giải thích hơn.

Regularization L2 (Ridge): Phương pháp này giúp giữ lại tất cả các đặc trưng nhưng hạn chế trọng số của chúng, do đó giúp mô hình trở nên ổn định hơn trước những biến động trong dữ liệu.

# **KẾT LUẬN**

Kết quả đạt được

Kiến thức: Qua quá trình phát triển mô hình dự đoán giá nhà, chúng tôi đã tích lũy được nhiều kiến thức quý giá về các phương pháp học máy, kỹ thuật tiền xử lý dữ liệu, cũng như các phương pháp tối ưu hóa mô hình. Chúng tôi đã hiểu rõ hơn về cách các tham số như learning rate, batch size, và epochs ảnh hưởng đến hiệu suất của mô hình. Ngoài ra, việc áp dụng các kỹ thuật như regularization và tăng cường dữ liệu đã giúp chúng tôi nắm bắt các khái niệm quan trọng trong việc xây dựng mô hình tổng quát.

Sản phẩm: Sản phẩm cuối cùng là một mô hình mạng nơ-ron có khả năng dự đoán giá nhà với độ chính xác cao và độ sai số nhỏ. Mô hình này không chỉ cho thấy tiềm năng trong việc ứng dụng thực tế mà còn có thể mở rộng cho các bài toán dự đoán khác trong lĩnh vực bất động sản hoặc các lĩnh vực liên quan.

Ưu điểm và nhược điểm

Ưu điểm: Mô hình có khả năng dự đoán chính xác với độ sai số thấp, điều này rất quan trọng trong lĩnh vực bất động sản. Có khả năng mở rộng, cho phép áp dụng cho các tập dữ liệu khác mà không cần thay đổi cấu trúc cơ bản của mô hình. Quá trình phát triển đã cung cấp cho nhóm những kỹ năng và kiến thức thực tiễn về học máy và phân tích dữ liệu.

Nhược điểm: Vì thời gian có hạn nên chúng tôi chưa thể thử nghiệm sâu rộng với nhiều kiến trúc mạng khác nhau, điều này có thể giúp cải thiện hiệu suất của mô hình. Vì kiến thức còn hạn chế nên một số phương pháp tối ưu hóa và kỹ thuật nâng cao chưa được áp dụng triệt để, ảnh hưởng đến khả năng tổng quát của mô hình. Vì điều kiện đầu tư còn hạn hẹp nên chúng tôi không thể sử dụng các công cụ tính toán mạnh mẽ hơn hoặc dữ liệu phong phú hơn, điều này có thể giúp đạt được kết quả tốt hơn.

Công việc tương lai

Trong tương lai, nếu có thêm thời gian, chúng tôi sẽ tiến hành các thử nghiệm sâu hơn với các kiến trúc mạng khác nhau, đặc biệt là các mô hình phức tạp như LSTM hoặc CNN. Chúng tôi cũng sẽ tìm cách mở rộng tập dữ liệu bằng cách thu thập thêm thông tin và áp dụng các kỹ thuật tăng cường dữ liệu để cải thiện tính tổng quát của mô hình. Ngoài ra, việc nghiên cứu và áp dụng các phương pháp mới trong học máy sẽ là mục tiêu của chúng tôi nhằm nâng cao hiệu suất và độ chính xác trong các dự đoán. Chúng tôi hy vọng rằng những cải tiến này sẽ mang lại những sản phẩm tốt hơn trong tương lai, phục vụ tốt hơn cho nhu cầu thực tế trong lĩnh vực bất động sản.

**DANH MỤC TÀI LIỆU THAM KHẢO**

1. Nguyễn Hồng Sơn (2007), *Giáo trình hệ thống Mạng máy tính CCNA* (Semester 1), NXB Lao động xã hội.
2. Phạm Quốc Hùng (2017), *Đề cương bài giảng Mạng máy tính*, Đại học SPKT Hưng Yên.

[3]. <https://luanvan.net.vn/luan-van/de-tai-ung-dung-mo-hinh-neural-network-trong-viec-du-bao-thi-truong-chung-khoan-viet-nam-7062/>

[4]. https://kinhtevadubao.vn/ung-dung-mang-no-ron-nhan-tao-anns-trong-du-bao-gia-dong-cua-cac-ma-co-phieu-niem-yet-tren-san-chung-khoan-16719.html