# Tham gia của các thành viên

|  |  |  |
| --- | --- | --- |
| **Thành viên** | **Phần phụ trách** | **Tỷ lệ tham gia** |
| Trần Quý Đạt | * Tìm hiểm thuật toán, mô hình * Thu thập dữ liệu * Làm slide, thuyết trình | 30% |
| Nguyễn Thị Lan | * Tìm hiểu thuật toán, mô hình. * Thu thập dữ liệu * Làm báo cáo, slide, thuyết trình | 30% |
| Đặng Thị Thanh Mai | * Tìm hiểu thuật toán, mô hình. * Thu thập dữ liệu * Làm báo cáo | 20% |
| Nguyễn Tiến Anh | * Tìm hiểu thuật toán, mô hình. * Xử lý và phân tích dữ liệu * Làm báo cáo | 20% |

# Lời cảm ơn

Trước tiên, chúng em xin gửi lời cảm ơn chân thành đến Học viện Công nghệ Bưu chính Viễn thông, đặc biệt là Khoa Công nghệ Thông tin 1, vì đã tạo điều kiện cho chúng em học tập và nghiên cứu thông qua môn học “Phát triển các hệ thống thông minh”. Môn học này đã mang lại cho chúng em những kiến thức quý báu và nền tảng vững chắc trong lĩnh vực công nghệ.

Chúng em xin bày tỏ lòng biết ơn sâu sắc đến giảng viên của bộ môn, người đã tận tâm giảng dạy, hướng dẫn và truyền đạt cho chúng em những kiến thức bổ ích trong suốt quá trình học tập. Nhờ sự chỉ bảo tận tình và sự hỗ trợ quý báu của thầy, chúng em đã hoàn thành được bài tập lớn cho môn học này một cách hiệu quả. Những kinh nghiệm và kiến thức mà chúng em học được từ thầy không chỉ hữu ích trong học tập hiện tại mà còn là hành trang quý giá cho sự nghiệp sau này.

Chúng em xin cảm ơn các bạn đã đồng hành và giúp đỡ nhau trong quá trình học tập.

Chúng em cũng xin gửi lời cảm ơn bố mẹ đã luôn ủng hộ và động viên trong suốt quá trình thực hiện bài tập này.

# Mục lục

[Tham gia của các thành viên i](#_Toc184768661)

[Lời cảm ơn ii](#_Toc184768662)

[Mục lục iii](#_Toc184768663)

[Các từ viết tắt vi](#_Toc184768664)

[Danh sách các hình vẽ viii](#_Toc184768665)

[Mở đầu 1](#_Toc184768666)

[1.Lí do chọn đề tài 1](#_Toc184768667)

[2.Cấu trúc báo cáo 2](#_Toc184768668)

[CHƯƠNG 1. 3](#_Toc184768669)

[Giới thiệu về học máy, học sâu và bài toán 3](#_Toc184768670)

[1.1. Học máy 3](#_Toc184768671)

[1.1.1. Khái niệm và lịch sử phát triển 3](#_Toc184768672)

[1.1.2. Các phương pháp học máy 4](#_Toc184768673)

[1.1.3. Ứng dụng của học máy 6](#_Toc184768674)

[1.1.4. Thách thức của học máy 7](#_Toc184768675)

[1.2 Học sâu 7](#_Toc184768676)

[1.2.1. Khái niệm và nguyên lý cơ bản 7](#_Toc184768677)

[1.2.2. Các kiến trúc học sâu phổ biến 8](#_Toc184768678)

[1.2.3. Ứng dụng của học sâu 11](#_Toc184768679)

[1.2.4. Thách thức của Học Sâu 11](#_Toc184768680)

[1.3. Mô hình dự đoán nồng độ khí CO2 tại Hà Nội 12](#_Toc184768681)

[1.3.1. Giới thiệu về mô hình dự đoán nồng độ khí CO2 tại Hà Nội 12](#_Toc184768682)

[1.3.2. Các phương pháp dự đoán nồng độ khí CO2 truyền thống 13](#_Toc184768683)

[1.3.3. Các phương pháp dự đoán nồng độ khí CO2 sử dụng học máy, học sâu 13](#_Toc184768684)

[1.3.4. Định nghĩa bài toán cụ thể trong dự án 14](#_Toc184768685)

[1.4. Giải pháp: thuật toán và phương pháp 15](#_Toc184768686)

[1.4.1. Chuẩn bị dữ liệu 15](#_Toc184768687)

[1.4.2. Thiết kế kiến trúc cho bài toán 18](#_Toc184768688)

[1.4.3. Triển khai và tinh chỉnh mô hình 20](#_Toc184768689)

[1.5. Kết luận 21](#_Toc184768690)

[Chương 2. 22](#_Toc184768691)

[Phân tích dữ liệu 22](#_Toc184768692)

[2.1. Bộ dữ liệu 22](#_Toc184768693)

[2.1.1. Giới thiệu bộ dữ liệu 22](#_Toc184768694)

[2.1.2. Nguồn dữ liệu 22](#_Toc184768695)

[2.1.3. Đặc điểm của dữ liệu 23](#_Toc184768696)

[2.2. Phân tích dữ liệu bùng nổ EDA 25](#_Toc184768697)

[2.2.1. Các thông số thống kê và trực quan hoá dữ liệu 25](#_Toc184768698)

[2.3 Kết luận 28](#_Toc184768699)

[CHƯƠNG 3. 30](#_Toc184768700)

[Thực nghiệm 30](#_Toc184768701)

[3.1. Mô hình 30](#_Toc184768702)

[3.1.1. Hồi quy tuyến tính 30](#_Toc184768703)

[3.1.2. Rừng ngẫu nhiên 31](#_Toc184768704)

[3.1.3. LSTM 32](#_Toc184768705)

[3.2. Huấn luyện mô hình 33](#_Toc184768706)

[3.2.1. Dữ liệu đầu vào 33](#_Toc184768707)

[3.2.2. Huấn luyện mô hình Hồi quy tuyến tính 33](#_Toc184768708)

[3.2.3. Huấn luyện mô hình Rừng ngẫu nhiên 34](#_Toc184768709)

[3.2.4. Huấn luyện mô hình LSTM 35](#_Toc184768710)

[3.2.5 Kết quả huấn luyện 35](#_Toc184768711)

[3.3. Đánh giá mô hình 38](#_Toc184768712)

[3.3.1. Đánh giá độ chính xác của mô hình 38](#_Toc184768713)

[3.3.2. Kết quả dự đoán 39](#_Toc184768714)

[3.3.3. Kết luận 39](#_Toc184768715)

[Kết luận 40](#_Toc184768716)

[1. Kết quả đã đạt được 40](#_Toc184768717)

[2. Phương hướng phát triển đề tài 40](#_Toc184768718)

[Tài liệu tham khảo 41](#_Toc184768719)

# Các từ viết tắt

|  |  |  |
| --- | --- | --- |
| **Viết tắt** | **Tiếng Anh** | **Giải thích** |
| CNN | Convolutional Neural Networks | Mạng nơron tích chập |
| ANN | Artificial Neural Network | Mạng nơron nhân tạo |
| ReLU | Rectified Linear Unit | Là 1 hàm kích hoạt |
| RNN | Recurrent Neural Networks | Mạng nơron hồi quy |
| GPU | Graphics Processing Unit | Bộ xử lý đồ hoạ |
| GAN | Generative Adversarial Networks | Mạng đối nghịch tạo sinh |
| PCA | Principal Component Analysis | Phân tích thành phần chính |
| t-SNE | t-distributed Stochastic Neighbor Embedding | Phép nhúng lân cận ngẫu nhiên phân bố t |
| EDA | Exploratory Data Analysis | Phân tích dữ liệu khám phá |
| GPT | Generative  Pre-training Transformer | Mô hình chuyển đổi tiền huấn luyện tạo sinh |
| t-SNE | t-distributed Stochastic Neighbor Embedding | Phép nhúng lân cận ngẫu nhiên phân bố t |
| LSTM | Long Short-Term Memory | Bộ nhớ ngắn hạn dài |
| GRU | Gated Recurrent Unit | Đơn vị hồi quy có tổng |
| RMSE | Root Mean Square Error | Sai số bìnnh phương trung bình |
| MAE | Mean Absolute Error | Sai số tuyệt đối trung bình |
| BERT | Bidirectional Encoder Representations from Transformers | Biểu diễn mã hoá hai chiều từ mô hình chuyển đổi |
| TPU | Tensor Processing Unit | Bộ xử lý Tensor |
| SVM |  |  |
| CSV | Comma-Separated Values | Tên một định dạng tập tin văn bản |
| MSE | Mean Squared Error | Sai số bình phương trung bình |
| IOT | Internet of things | Mạng lưới vạn vật kết nối |

# Danh sách các hình vẽ

[Hình 1. 1. Học máy 3](#_Toc184768720)

[Hình 1. 2. Học máy có giám sát 4](#_Toc184768721)

[Hình 1. 3. Học máy không giám sát 5](#_Toc184768722)

[Hình 1. 4. Học máy bán giám sát 6](#_Toc184768723)

[Hình 1. 5. Học sâu 8](#_Toc184768724)

[Hình 1. 6. Mạng nơ ron tích chập 9](#_Toc184768725)

[Hình 1. 7. Mạng nơ ron hồi quy 10](#_Toc184768726)

[Hình 1. 8. Số lượng giá trị mất mát trong dữ liệu 15](#_Toc184768727)

[Hình 1. 9. Biểu đồ thể hiện sự mất mát giá trị 16](#_Toc184768728)

[Hình 1. 10. Loại bỏ giá trị bất bình thường 16](#_Toc184768729)

[Hình 1. 11. Chuẩn hóa dữ liệu với Standard Scaler 17](#_Toc184768730)

[Hình 1. 12. Phân chia dữ liệu trong dự án 17](#_Toc184768731)

[Hình 2. 1. Cảm biến MQ-135 23](#_Toc184768732)

[Hình 2. 2. Đặc điểm bộ dữ liệu Accuweather 24](#_Toc184768733)

[Hình 2. 3. Đặc điểm bộ dữ liệu lấy từ thiết bị IoT 25](#_Toc184768734)

[Hình 2. 4. Biểu đồ về sự phân bố khí Co2 26](#_Toc184768735)

[Hình 2. 5. Biểu đồ hộp phân phối nồng độ CO2 27](#_Toc184768736)

[Hình 2. 6. Biểu đồ Violet phân phối nồng độ CO2 27](#_Toc184768737)

[Hình 2. 7. Biểu đồ Density Plot nồng độ CO2 28](#_Toc184768738)

[Hình 2. 8. Biểu đồ Log Plot phân phối nồng độ CO2 28](#_Toc184768739)

[Hình 2. 9. Biểu đồ QQ Plot phân phối nồng độ CO2 28](#_Toc184768740)

[Hình 3. 1.Mô hình hồi quy tuyến tính 31](#_Toc184768741)

[Hình 3. 2. Mô hình rừng ngẫu nhiên 31](#_Toc184768742)

[Hình 3. 3. Mô hình LSTM 32](#_Toc184768743)

[Hình 3. 4. Hệ số xác định 36](#_Toc184768744)

[Hình 3. 5. Kết quả huấn luyện mô hình Hồi quy tuyến tính 36](#_Toc184768745)

[Hình 3. 6. Kết quả huấn luyện mô hình Rừng ngẫu nhiên 36](#_Toc184768746)

[Hình 3. 7. Kết quả huấn luyện mô hình LSTM 37](#_Toc184768747)

[Hình 3. 8. Biểu đô Train Loss 38](#_Toc184768748)

[Hình 3. 9. Kết quả dự đoán 3 mô hình 39](#_Toc184768749)

# Mở đầu

## 1.Lí do chọn đề tài

Trong những năm gần đây, công nghệ thông minh và trí tuệ nhân tạo đã có những bước tiến vượt bậc, mở ra nhiều cơ hội mới cho các lĩnh vực khác nhau, đặc biệt là môi trường. Việc ứng dụng AI cùng các mô hình dự đoán giúp tự động hóa quy trình giám sát và dự báo chất lượng không khí, không chỉ hỗ trợ việc ra quyết định nhanh chóng mà còn cải thiện đáng kể độ chính xác trong các dự báo về khí hậu. Điều này đáp ứng nhu cầu ngày càng cao về bảo vệ sức khỏe cộng đồng và thúc đẩy sự phát triển bền vững trong công tác quản lý môi trường.

Dự đoán nồng độ khí CO2 là một khâu quan trọng trong công tác giám sát chất lượng không khí và đề xuất các biện pháp giảm thiểu ô nhiễm, đòi hỏi tính chính xác cao. Tuy nhiên, các phương pháp dự đoán truyền thống còn nhiều hạn chế, dễ dẫn đến sai lệch và độ chính xác không cao. Do đó, việc tìm kiếm giải pháp công nghệ để cải thiện độ chính xác của mô hình dự báo là vô cùng cần thiết.

Đề tài "Mô hình dự đoán nồng độ khí CO2 tại Hà Nội" được lựa chọn nhằm đáp ứng nhu cầu thực tiễn trong công tác quản lý và giám sát chất lượng không khí. Hà Nội là một trong những thành phố có mật độ dân số cao và mức độ đô thị hóa nhanh chóng, dẫn đến tình trạng ô nhiễm không khí đáng báo động. Việc dự đoán chính xác nồng độ CO2 không chỉ giúp các cơ quan chức năng kiểm soát ô nhiễm một cách chủ động mà còn nâng cao nhận thức của cộng đồng về vấn đề này.

Với sự phát triển mạnh mẽ của các phương pháp học máy và trí tuệ nhân tạo, áp dụng công nghệ này vào việc dự báo nồng độ CO2 là giải pháp khả thi và hiệu quả. Đề tài này nhằm xây dựng một mô hình dự đoán dựa trên AI, có khả năng phân tích và dự đoán chính xác các mức CO2 trong không khí, từ đó hỗ trợ quá trình giám sát chất lượng không khí hiện đại, giảm thiểu sai số và nâng cao tính chính xác.

Bên cạnh đó, việc thực hiện đề tài này còn giúp nhóm vận dụng các kiến thức đã học về trí tuệ nhân tạo, xử lý dữ liệu và học máy để giải quyết một vấn đề thực tiễn cụ thể. Điều này không chỉ củng cố kiến thức lý thuyết mà còn phát triển kỹ năng nghiên cứu và ứng dụng công nghệ. Hơn nữa, đề tài có tính ứng dụng cao, góp phần thúc đẩy xu hướng giám sát môi trường thông minh và mang lại nhiều lợi ích cho công tác bảo vệ môi trường trong tương lai.

## 2.Cấu trúc báo cáo

Báo cáo gồm có các chương:

* Chương 1. Giới thiệu về học máy – học sâu
* Chương 2. Giới thiệu đề tài
* Chương 3. Thực nghiệm

# CHƯƠNG 1.

# Giới thiệu về học máy, học sâu và bài toán

## 1.1. Học máy

### 1.1.1. Khái niệm và lịch sử phát triển

Học máy hay máy học là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Thuật ngữ "học máy" ra đời từ những năm 1950, với các công trình của Alan Turing và Arthur Samuel. Từ đó, học máy đã phát triển thành một lĩnh vực quan trọng, với nhiều ứng dụng trong cuộc sống hằng ngày.

**Khái niệm học máy**

Học máy là quá trình máy sử dụng rất nhiều dữ liệu để nó huấn luyện và kiểm chứng lại mô hình để tạo mô hình hoàn chỉnh cho người sử dụng. Học máy là máy tự huấn luyện mô hình và con người sử dụng mô hình đó.

Dựa theo cách học, học máy có thể phân thành 3 loại cơ bản : học máy giám sát, học máy không giám sát và học bán giám sát.

A person looking at a robot

Description automatically generated

Hình 1. 1. Học máy

**Lịch sử phát triển**

Học máy bắt đầu từ thập niên 1950 với các nền tảng lý thuyết của Alan Turing và các mô hình nơ-ron như Perceptron. Qua các thập niên 1960-1980, học máy phát triển mạnh về lý thuyết với các thuật toán như K-means, mạng nơron và thuật toán lan truyền ngược. Đến thập niên 1990, các phương pháp như máy vector hỗ trợ và cây quyết định được ứng dụng rộng rãi. Thập niên 2000 chứng kiến sự phát triển của học sâu nhờ vào dữ liệu lớn. Từ 2010 đến nay, học máy bùng nổ với các mô hình như GAN, AlphaGo và GPT, ứng dụng vào nhiều lĩnh vực từ y tế đến xử lý ngôn ngữ tự nhiên.

### 1.1.2. Các phương pháp học máy

**Học máy có giám sát**

Học máy có giám sát là loại học mà máy tính được đào tạo trên một tập dữ liệu đã được gán nhãn. Nói cách khác, trong bài toán này, chúng ta cung cấp cho máy tính cả dữ liệu đầu vào và đầu ra mong muốn. Mục đích của việc học có giám sát là dự đoán kết quả đầu ra cho dữ liệu mới chưa được gán nhãn.

A diagram of a person

Description automatically generated

Hình 1. 2. Học máy có giám sát

Học máy có giám sát lại được chia thành hai nhóm chính: Hồi quy và phân loại.

* Hồi quy: Khi kết quả đầu ra là một giá trị liên tục, chúng ta sử dụng thuật toán hồi quy.
* Có nhiều loại hồi quy, bao gồm hồi quy tuyến tính, hồi quy đa thức, hồi quy logistic (dùng cho phân loại nhị phân), và hồi quy Ridge/Lasso.
* Quy trình của thuật toán bao gồm chuẩn bị dữ liệu, huấn luyện mô hình, dự đoán giá trị cho dữ liệu mới, và đánh giá độ chính xác bằng các chỉ số như RMSE hoặc MAE.
* Hồi quy thường được sử dụng trong các lĩnh vực như: dự đoán giá nhà và doanh thu.
* Phân loại: Khi kết quả đầu ra là một giá trị rời rạc, chúng ta sử dụng thuật toán phân loại.
* Phân loại là quá trình phân nhóm dữ liệu đầu vào thành các nhãn cụ thể dựa trên các đặc trưng đã biết. Mô hình được huấn luyện bằng dữ liệu đã gán nhãn để dự đoán nhãn cho dữ liệu mới.
* Các thuật toán phổ biến bao gồm hồi quy logistic, cây quyết định, và máy vector hỗ trợ.
* Phân loại được ứng dụng trong nhận diện hình ảnh, phân loại email, và phân loại văn bản.

**Học máy không giám sát**

Học máy không giám sát xử lý vấn đề khi chúng ta không biết kết quả đầu ra mong muốn và không có nhãn cho dữ liệu. Mục tiêu của học không giám sát là khám phá cấu trúc ẩn trong dữ liệu và tìm ra một cách biểu diễn dữ liệu phù hợp.

A diagram of a diagram

Description automatically generated

Hình 1. 3. Học máy không giám sát

Học máy không giám sát chủ yếu được chia thành hai nhóm:

* Phân cụm: là một phương pháp để nhóm dữ liệu thành các cụm dựa trên sự tương đồng giữa các đặc trưng.
* Mục tiêu của phân cụm là tìm cấu trúc ẩn trong dữ liệu mà không cần nhãn.
* Các thuật toán phổ biến bao gồm K-means và DBSCAN.
* Phân cụm được ứng dụng trong phân loại khách hàng, phân loại tài liệu, …
* Giảm chiều dữ liệu: Giảm số lượng biến trong dữ liệu, giữ lại các thông tin quan trọng nhất.
* Mục tiêu: giảm độ phức tạp và cải thiện hiệu suất mô hình.
* Các thuật toán phổ biến bao gồm PCA và t-SNE.
* Phân chiều dữ liệu thường được sử dụng trong giảm chiều dữ liệu ảnh, tìm các yếu tố tiềm ẩn trong thông tin người dùng, …

**Học bán giám sát**

Học máy bán giám sát kết hợp cả hai phương pháp học có giám sát và học không giám sát, trong đó một phần dữ liệu được gán nhãn và phần còn lại không. Mục tiêu của học máy bán giám sát là tận dụng lợi thế của cả hai phương pháp để cải thiện độ chính xác của mô hình.

A diagram of a computer network

Description automatically generated

Hình 1. 4. Học máy bán giám sát

### 1.1.3. Ứng dụng của học máy

Học máy hiện đang được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau, từ y tế, tài chính đến giải trí.

### 1.1.4. Thách thức của học máy

Mặc dù mang lại nhiều lợi ích, học máy cũng đối mặt với một số thách thức quan trọng:

* **Chất lượng dữ liệu:** Hiệu quả của học máy phụ thuộc mạnh mẽ vào chất lượng của dữ liệu. Dữ liệu không đầy đủ, nhiễu, hoặc thiên vị có thể làm suy giảm độ chính xác của các mô hình. Việc thu thập và làm sạch dữ liệu là một quá trình tốn kém và phức tạp.
* **Tính giải thích của mô hình:** Các mô hình học máy phức tạp, đặc biệt là các mô hình học sâu, thường khó giải thích và minh bạch. Điều này gây ra thách thức khi sử dụng các mô hình trong các lĩnh vực yêu cầu sự minh bạch cao, như y tế hoặc pháp lý.

## 1.2 Học sâu

### 1.2.1. Khái niệm và nguyên lý cơ bản

**Khái niệm học sâu**

Học sâu là một lĩnh vực con của học máy, sử dụng các mạng nơron nhân tạo với nhiều lớp để tự động học và trích xuất đặc trưng phức tạp từ dữ liệu. Nó đặc biệt hiệu quả khi xử lý dữ liệu lớn, chẳng hạn như hình ảnh, âm thanh, văn bản, và được ứng dụng rộng rãi trong các lĩnh vực như nhận diện hình ảnh, xử lý ngôn ngữ tự nhiên và xe tự lái.

A close-up of a computer screen

Description automatically generated

Hình 1. 5. Học sâu

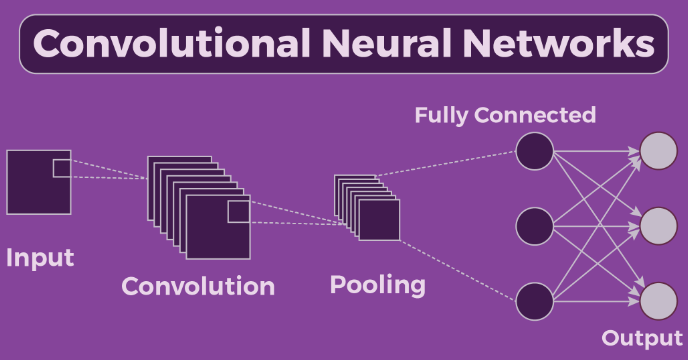
**Nguyên lý hoạt động của mạng nơ ron nhân tạo**

Mạng nơ-ron nhân tạo mô phỏng hoạt động của hệ thần kinh con người, với các tế bào nơ-ron nhân tạo kết nối với nhau qua các trọng số. Dữ liệu được lan truyền qua các lớp của mạng, sau đó điều chỉnh trọng số qua quá trình học dựa trên sai số giữa đầu ra dự đoán và giá trị thực tế.

### 1.2.2. Các kiến trúc học sâu phổ biến

Học sâu đã phát triển thành nhiều loại kiến trúc mạng nơ-ron khác nhau để phù hợp với các bài toán cụ thể.

**Mạng nơ-ron tích chập:** Mạng nơ-ron tích chập là một loại mô hình học sâu rất phổ biến, đặc biệt được sử dụng trong các bài toán xử lý ảnh và video như nhận dạng hình ảnh, phát hiện đối tượng, và phân loại video. Nó lấy cảm hứng từ cấu trúc của vỏ não thị giác của con người và có khả năng học và tự động trích xuất các đặc trưng từ dữ liệu đầu vào.



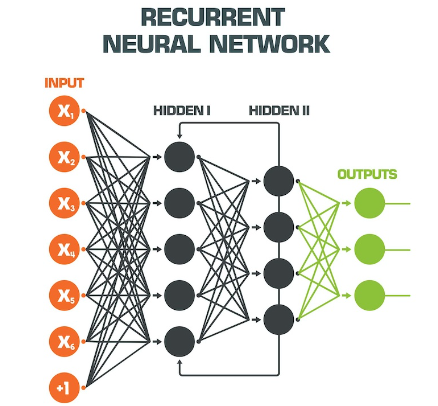
Hình 1. 6. Mạng nơ ron tích chập

Cấu trúc cơ bản của CNN:

* Lớp tích chập: Đây là lớp chính trong CNN để trích xuất các đặc trưng từ dữ liệu đầu vào. Lớp này thực hiện phép toán tích chập giữa một nhân (hay bộ lọc) với dữ liệu, tạo ra một bản đồ đặc trưng . Kernel là một ma trận giá trị nhỏ, trượt qua toàn bộ dữ liệu để tìm các đặc trưng như cạnh, góc, và kết cấu.
* Lớp kích hoạt: Sau lớp tích chập, các giá trị đặc trưng thường được đưa qua một hàm kích hoạt, phổ biến là hàm ReLU để tăng tính phi tuyến của mô hình. Hàm ReLU có dạng f (x) = max (0, x), giúp loại bỏ các giá trị âm và tăng cường khả năng học đặc trưng của mạng.
* Lớp lấy mẫu: Lớp này thường được sử dụng để giảm kích thước của bản đồ đặc trưng, giúp giảm bớt số lượng tham số và độ phức tạp tính toán của mô hình. Phổ biến nhất là max-pooling, lấy giá trị lớn nhất trong mỗi vùng nhỏ của bản đồ đặc trưng. Lớp này giúp CNN không nhạy cảm với các thay đổi nhỏ trong vị trí của đặc trưng.
* Lớp kết nối hoàn toàn:
* Cuối cùng, các bản đồ đặc trưng được chuyển thành một vectơ và đưa qua một hoặc nhiều lớp kết nối hoàn toàn. Đây là nơi mà CNN sẽ kết hợp các đặc trưng trích xuất và đưa ra dự đoán cuối cùng.
* Lớp này được sử dụng trong các mô hình phân loại, tạo ra xác suất cho từng lớp.
* Cách CNN học từ dữ liệu : Trong quá trình huấn luyện, CNN điều chỉnh các tham số của kernel để giảm sai số giữa dự đoán và nhãn thực. Quá trình học này được thực hiện thông qua lan truyền ngược) và thuật toán tối ưu hóa, thường là giảm độ dốc. Qua nhiều vòng lặp huấn luyện, CNN học được các đặc trưng phức tạp hơn, từ những đường nét cơ bản trong các lớp đầu đến các đặc trưng mang tính chất phân loại ở các lớp sâu hơn.
* Ứng dụng của CNN: Xử lý hình ảnh và thị giác máy tính**,** Xử lý ngôn ngữ tự nhiên,….

**Mạng nơ-ron hồi quy**

Mạng nơ-ron hồi quy là một loại mạng nơ-ron được thiết kế để xử lý dữ liệu chuỗi thời gian hoặc dữ liệu có sự phụ thuộc giữa các bước, như văn bản hoặc âm thanh. Điểm nổi bật của RNN là khả năng ghi nhớ thông tin từ các bước thời gian trước đó thông qua trạng thái ẩn cho phép mạng xử lý dữ liệu có thứ tự và phụ thuộc lẫn nhau. Tuy nhiên, RNN cổ điển gặp phải vấn đề mất mát dữ liệu khi chuỗi dữ liệu quá dài, làm giảm hiệu quả học tập. Để khắc phục, các phiên bản cải tiến như LSTM và GRU được phát triển, giúp ghi nhớ thông tin dài hạn tốt hơn.



Hình 1. 7. Mạng nơ ron hồi quy

RNN được ứng dụng rộng rãi trong xử lý ngôn ngữ tự nhiên, dự báo chuỗi thời gian, và nhận dạng giọng nói.

**Mạng chuyển đổi**

Mạng chuyển đổi là một kiến trúc mạng nơ-ron đặc biệt dành cho việc xử lý chuỗi dữ liệu, nổi bật với cơ chế tự chú ý giúp nắm bắt mối quan hệ giữa các phần tử trong chuỗi mà không cần xử lý tuần tự như RNN. Với khả năng tính toán song song, mạng chuyển đổi xử lý nhanh hơn và hiệu quả hơn trên chuỗi dài. Kiến trúc của nó gồm hai phần chính là bộ mã hoá và bộ giải mã, rất phù hợp cho các tác vụ như dịch máy, tóm tắt văn bản, và sinh văn bản. Các mô hình tiên tiến như BERT, GPT, và T5 đều dựa trên chuyển đổi.

### 1.2.3. Ứng dụng của học sâu

Học sâu đã có nhiều ứng dụng đột phá trong các lĩnh vực xử lý hình ảnh, xử lý ngôn ngữ tự nhiên và các hệ thống tự động hóa:

* **Nhận diện hình ảnh và video:** Mạng CNN đã cải thiện đáng kể độ chính xác trong các bài toán nhận diện hình ảnh và video, từ nhận diện khuôn mặt, phân loại đối tượng cho đến phát hiện các dấu hiệu bất thường trong ảnh y tế.
* **Xử lý ngôn ngữ tự nhiên:** Học sâu, đặc biệt là các mô hình chuyển đổi, đã đưa xử lý ngôn ngữ tự nhiên lên một tầm cao mới, từ dịch máy, phân tích cảm xúc, đến tạo ra văn bản tự động. Các mô hình như GPT-3 và BERT đang được ứng dụng trong nhiều lĩnh vực từ chatbot, trợ lý ảo đến phân tích dữ liệu văn bản.
* **Ô tô tự hành và robot:** Các hệ thống học sâu được ứng dụng trong ô tô tự hành để phân tích dữ liệu từ cảm biến, camera và radar, giúp xe nhận biết môi trường xung quanh, phát hiện chướng ngại vật và điều hướng tự động.

### 1.2.4. Thách thức của Học Sâu

Mặc dù đã đạt được nhiều thành tựu, học sâu vẫn còn nhiều thách thức về mặt kỹ thuật và ứng dụng thực tiễn.

* **Tính tốn kém về tài nguyên:** Học sâu yêu cầu lượng dữ liệu lớn và tài nguyên tính toán mạnh mẽ để huấn luyện các mô hình phức tạp. Điều này dẫn đến chi phí cao và đòi hỏi sự đầu tư lớn về phần cứng, đặc biệt là GPU hoặc TPU.
* **Vấn đề quá khớp:** Với mô hình phức tạp và nhiều tham số, học sâu dễ gặp phải tình trạng quá khớp, khi mô hình học quá mức từ dữ liệu huấn luyện mà không tổng quát tốt cho dữ liệu mới. Các phương pháp như regularization và dropout được sử dụng để giảm thiểu tình trạng này.

## 1.3. Mô hình dự đoán nồng độ khí CO2 tại Hà Nội

### 1.3.1. Giới thiệu về mô hình dự đoán nồng độ khí CO2 tại Hà Nội

Dự đoán nồng độ khí CO₂ là một bài toán quan trọng trong lĩnh vực môi trường và biến đổi khí hậu, nhằm mục đích xác định mức độ ô nhiễm không khí trong thời gian thực hoặc dự báo trong tương lai. Quá trình này không chỉ đơn thuần là ước lượng mức CO₂ mà còn yêu cầu phân tích các yếu tố ảnh hưởng phức tạp như thời tiết, hoạt động giao thông, và mật độ dân số. Việc hiểu rõ và dự đoán chính xác nồng độ CO₂ giúp các nhà quản lý đưa ra các biện pháp hiệu quả để cải thiện chất lượng không khí, bảo vệ sức khỏe cộng đồng và giảm thiểu tác động của biến đổi khí hậu.

Các mô hình dự đoán nồng độ khí CO₂ thường sử dụng các kỹ thuật học máy và học sâu để trích xuất các đặc trưng phức tạp từ dữ liệu môi trường, từ đó đưa ra dự đoán chính xác cho từng khu vực và thời gian cụ thể. Các yếu tố như nhiệt độ, độ ẩm, tốc độ gió, và hoạt động công nghiệp đều có ảnh hưởng đáng kể đến mức CO₂ trong không khí. Để đạt được độ chính xác cao, các mô hình dự đoán phải có khả năng phân tích mối quan hệ giữa các yếu tố này và nồng độ CO₂, đồng thời xử lý tốt các biến động đột ngột trong dữ liệu. Dự đoán khí CO₂ và các chất ô nhiễm khác đóng vai trò quan trọng trong lĩnh vực nghiên cứu về môi trường và biến đổi khí hậu, từ việc cung cấp cái nhìn về xu hướng ô nhiễm đến hỗ trợ các chính sách cải thiện chất lượng không khí.

### 1.3.2. Các phương pháp dự đoán nồng độ khí CO2 truyền thống

* **Ưu điểm:**
* Dễ dàng triển khai và có thể cung cấp kết quả nhanh chóng.
* Phân tích dễ hiểu, trực quan, và ít tốn tài nguyên.
* Phù hợp cho các mối quan hệ tuyến tính giữa các biến, khi dữ liệu không quá phức tạp.
* **Nhược điểm:**
* Không phù hợp cho các dữ liệu phi tuyến tính và phức tạp.
* Độ chính xác thấp nếu nồng độ CO₂ bị ảnh hưởng bởi nhiều yếu tố phức tạp.
* Khó giải quyết các vấn đề phức tạp khi chỉ có các biến số độc lập đơn giản.

### 1.3.3. Các phương pháp dự đoán nồng độ khí CO2 sử dụng học máy, học sâu

* **Ưu điểm của học máy trong dự đoán nồng độ khí CO2:**
* **Dễ triển khai:** Các thuật toán học máy như hồi quy tuyến tính, SVM, cây quyết định có thể dễ dàng triển khai và yêu cầu ít dữ liệu huấn luyện.
* **Hiệu quả với dữ liệu nhỏ:** Học máy hoạt động tốt khi có một bộ dữ liệu nhỏ hoặc không quá phức tạp**.**
* **Giải thích được mô hình:** Các mô hình học máy đơn giản như hồi quy hoặc cây quyết định dễ hiểu và giải thích, giúp xác định các yếu tố ảnh hưởng đến nồng độ CO2**.**
* **Yêu cầu tài nguyên tính toán thấp:** Các phương pháp học máy không đòi hỏi phần cứng mạnh mẽ và có thể triển khai trên các hệ thống tính toán nhẹ.
* **Ưu điểm của học sâu trong dự đoán nồng độ khí CO2:**
* **Xử lý dữ liệu lớn và phức tạp**: Các mô hình học sâu (như mạng nơ-ron sâu) có thể học từ những tập dữ liệu rất lớn và phức tạp, phù hợp với các đặc điểm phi tuyến tính và mối quan hệ phức tạp giữa các yếu tố ảnh hưởng đến CO2.
* **Tự động trích xuất đặc trưng**: Mạng nơ-ron có thể tự động trích xuất các đặc trưng từ dữ liệu mà không cần can thiệp của con người.
* **Hiệu quả cao với dữ liệu lớn**: Khi có đủ dữ liệu huấn luyện, học sâu thường cho kết quả rất chính xác và hiệu quả trong việc dự đoán nồng độ khí CO2.

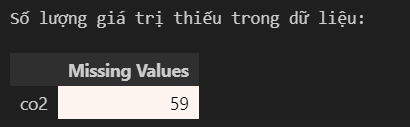
### 1.3.4. Định nghĩa bài toán cụ thể trong dự án

* **Đầu vào:** Đầu vào của bài toán là các dữ liệu môi trường bao gồm nhiệt độ, độ ẩm, áp suất khí quyển, tốc độ gió, mật độ dân số, hoạt động giao thông và các chỉ số môi trường khác được thu thập từ các trạm đo không khí và cơ quan quản lý môi trường tại Hà Nội và các dữ liệu thu thập từ các thiết bị điện tử IoT. Dữ liệu có thể bao gồm các thông số thu thập theo chuỗi thời gian hàng ngày hoặc theo giờ để nắm bắt sự biến đổi của nồng độ CO₂ trong không khí.
* **Đầu ra:** Đầu ra của hệ thống là dự đoán nồng độ CO₂ tại các khu vực khác nhau trong Hà Nội trong khoảng thời gian nhất định (theo giờ hoặc theo ngày). Mô hình dự đoán sẽ cho phép xác định mức độ ô nhiễm trong tương lai gần, giúp chính quyền địa phương đưa ra các biện pháp kiểm soát chất lượng không khí kịp thời.
* **Mục tiêu:** Mục tiêu của dự án là xây dựng một mô hình học máy hoặc học sâu có khả năng dự đoán chính xác nồng độ CO₂ dựa trên các yếu tố môi trường và hoạt động xã hội tại Hà Nội. Điều này bao gồm việc thu thập và xử lý dữ liệu môi trường, thiết kế và huấn luyện mô hình (có thể là mô hình hồi quy hoặc mô hình LSTM/CNN kết hợp), đánh giá hiệu suất của mô hình, và triển khai hệ thống dự đoán nồng độ CO₂ trong môi trường thực tế để hỗ trợ quản lý chất lượng không khí.

## 1.4. Giải pháp: thuật toán và phương pháp

### 1.4.1. Chuẩn bị dữ liệu

* **Thu thập dữ liệu:** Để xây dựng các một mô hình học máy, học sâu hiệu quả cho bài toán dự đoán nồng độ khí thải CO2, việc thu thập dữ liệu là bước đầu tiên và quan trọng. Các nguồn dữ liệu có thể sử dụng bao gồm:
* **Bộ dữ liệu công khai:** Sử dụng các bộ dữ liệu có sẵn từ trang web Accuweather.
* **Thu thập dữ liệu riêng**: Nếu bộ dữ liệu công khai không đáp ứng đủ yêu cầu về độ đa dạng và kích thước, nhóm có thể tự thu thập dữ liệu từ các nguồn như mạng xã hội, nơi sinh sống, các thiết bị điện tử IOT.
* **Tiền xử lý dữ liệu:** Tiền xử lý dữ liệu là bước quan trọng để đảm bảo dữ liệu đầu vào phù hợp với các mô hình. Các bước chính bao gồm:
* **Nội suy để lấp các giá trị bị khuyết**: Trong quá trình thu thập dữ liệu, có thể xuất hiện các giá trị bị thiếu. Để khắc phục điều này, chúng tôi sử dụng các phương pháp nội suy nhằm ước lượng và lấp đầy các giá trị bị thiếu, giúp dữ liệu hoàn chỉnh và hạn chế ảnh hưởng đến độ chính xác của mô hình.

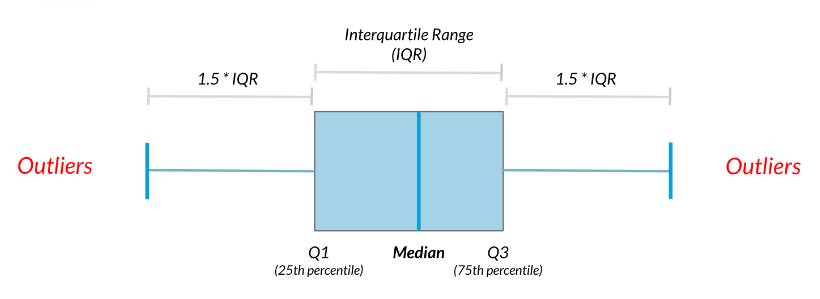


Hình 1. 8. Số lượng giá trị mất mát trong dữ liệu



Hình 1. 9. Biểu đồ thể hiện sự mất mát giá trị

* **Loại bỏ outlier**: Outliers là các giá trị bất thường hoặc sai lệch trong dữ liệu, có thể do lỗi đo lường hoặc các biến cố không thường xuyên. Để loại bỏ tác động tiêu cực của outliers, chúng tôi thực hiện kiểm tra và loại bỏ các giá trị này, giúp mô hình tập trung vào các dữ liệu chính xác và ổn định hơn.



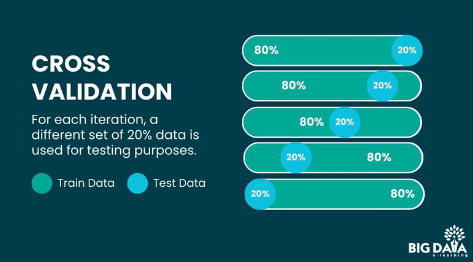
Hình 1. 10. Loại bỏ giá trị bất bình thường

* **Đưa dữ liệu về định dạng chuẩn**: Sau khi xử lý các giá trị bị khuyết và loại bỏ outliers, dữ liệu sẽ được định dạng lại sao cho phù hợp với yêu cầu của mô hình. Các bước này bao gồm chuyển đổi kiểu dữ liệu, định dạng lại cấu trúc, và sắp xếp dữ liệu theo trình tự hợp lý để dễ dàng xử lý và phân tích.
* **Chuẩn hóa dữ liệu với Standard Scaler**: Để đưa dữ liệu về phân phối chuẩn, chúng tôi sử dụng phương pháp Standard Scaler. Phương pháp này giúp chuẩn hóa các đặc trưng về cùng một thang đo, trung bình bằng 0 và độ lệch chuẩn bằng 1, giúp mô hình học tốt hơn mà không bị ảnh hưởng bởi các giá trị có thang đo khác nhau.



Hình 1. 11. Chuẩn hóa dữ liệu với Standard Scaler

* **Phân Chia Dữ Liệu:** Sau khi tiền xử lý, dữ liệu được chia thành các tập riêng biệt để huấn luyện và đánh giá mô hình:



Hình 1. 12. Phân chia dữ liệu trong dự án

* **Tập huấn luyện**: Chiếm khoảng 80% tổng dữ liệu, được sử dụng để huấn luyện mô hình.
* **Tập kiểm tra**: Chiếm khoảng 20% tổng dữ liệu, được sử dụng để đánh giá mô hình sau khi huấn luyện. Tập kiểm tra giúp kiểm tra độ chính xác và khả năng tổng quát hóa của mô hình.
* **Tập kiểm định**: Chiếm khoảng 20% trong tập huấn luyện, được sử dụng để đánh giá hiệu suất cuối cùng của mô hình sau khi đã hoàn tất quá trình huấn luyện và tinh chỉnh.

### 1.4.2. Thiết kế kiến trúc cho bài toán

Ở đây để bao quát và đánh giá được khách quan, chúng tôi sử dụng 3 mô hình cho bài toán này: Rừng ngẫu nhiên**,** LSTM, mô hình hồi quy tuyến tính

* **Thiết kế kiến trúc Hồi quy tuyến tính cho bài toán:**
* **Lựa chọn kiến trúc Hồi quy tuyến tính**: Hồi quy tuyến tính là mô hình đơn giản nhưng hiệu quả cho việc dự đoán khi nồng độ CO₂ có mối quan hệ tuyến tính với các yếu tố đầu vào. Mô hình được xây dựng bằng cách kết hợp với các phương pháp hồi quy Ridge và Lasso.
* **Cấu hình tham số**: Các tham số bao gồm:
* regressor: Sử dụng Ridge và Lasso để kiểm soát overfitting và tăng độ chính xác.
* regressor\_alpha: Giá trị alpha để điều chỉnh mức độ regularization, thử nghiệm với các giá trị từ 0.001 đến 10.
* **Điều chỉnh các siêu tham số:**
* **Pipeline**: Sử dụng Pipeline để tự động hóa quy trình chuẩn hóa dữ liệu với StandardScaler và huấn luyện mô hình với các thuật toán hồi quy (Ridge hoặc Lasso).
* **Grid Search**: Sử dụng GridSearchCV để tìm tham số tối ưu, bao gồm việc lựa chọn phương pháp hồi quy phù hợp và giá trị alpha tốt nhất.
* **Chiến lượcchuẩn hoá:**
* **Ridge và Lasso**: Cả hai phương pháp đều cung cấp cơ chế regularization để giảm thiểu overfitting. Ridge tập trung vào việc giảm giá trị của các hệ số, trong khi Lasso có thể loại bỏ hoàn toàn một số đặc trưng ít quan trọng.
* **Thiết kế kiến trúc Rừng ngẫu nhiên cho bài toán:**
* **Lựa chọn kiến trúc Random Forest**: Việc lựa chọn mô hình rừng ngẫu nhiên dựa trên khả năng kết hợp nhiều cây quyết định để cải thiện độ chính xác và khả năng tổng quát hóa trong dự đoán nồng độ CO₂.
* **Cấu hình tham số**: Các tham số của mô hình được tinh chỉnh qua Grid Search để tìm cấu hình tốt nhất:
* n\_estimators: Số lượng cây trong rừng, với các giá trị thử nghiệm là [50, 100, 200].
* max\_depth: Độ sâu tối đa của mỗi cây, giúp kiểm soát overfitting, giá trị trong khoảng [None, 5, 10].
* min\_samples\_split và min\_samples\_leaf: Điều kiện tối thiểu để phân tách một nút hoặc để giữ lại một nút lá, giúp giảm overfitting.
* **Điều chỉnh các siêu tham số:**
* **Grid Search**: Sử dụng GridSearchCV với các giá trị tham số được thử nghiệm để tìm ra bộ tham số tối ưu nhất cho mô hình, giúp nâng cao độ chính xác dự đoán.
* **Cross-Validation**: Thực hiện chia dữ liệu thành nhiều phần và huấn luyện trên các tập con để giảm thiểu khả năng overfitting.
* **Chiến lược chuẩn hoá**:
* **Time Series Split**: Sử dụng TimeSeriesSplit để đảm bảo tính thứ tự thời gian trong quá trình chia dữ liệu và đánh giá mô hình
* **Thiết kế kiến trúc LSTM cho bài toán**:
* **Lựa chọn kiến trúc LSTM**: Mô hình LSTM được lựa chọn để khai thác các mối quan hệ dài hạn trong dữ liệu chuỗi thời gian, thích hợp cho dự đoán nồng độ CO₂ với dữ liệu có tính tuần hoàn hoặc phụ thuộc theo thời gian.
* **Cấu trúc mô hình:**
* LSTM bao gồm các lớp tuần hoàn với số lượng num\_layers=2 và kích thước ẩn hidden\_size=50.
* Hai lớp fully connected (fc1 và fc2) để chuyển đổi kết quả từ lớp LSTM thành đầu ra dự đoán cuối cùng.
* **Điều chỉnh các siêu tham số:**
* **Hàm kích hoạt**: Sử dụng hàm kích hoạt ReLU trong lớp fully connected để tăng khả năng học phi tuyến của mô hình.
* **Batch Size và Learning Rate**: Các tham số như kích thước batch và learning rate được điều chỉnh để tối ưu hóa quá trình huấn luyện và đạt hiệu suất cao nhất.
* **Chiến lược chuẩn hoá:**
* **Gradient Clipping**: Hạn chế độ lớn của gradient để tránh tình trạng gradient bùng nổ, giữ cho quá trình huấn luyện ổn định.
* **Early Stopping**: Áp dụng dừng sớm khi độ chính xác trên tập kiểm tra không còn cải thiện, giúp giảm thiểu overfitting và tối ưu thời gian huấn luyện

### 1.4.3. Triển khai và tinh chỉnh mô hình

* **Tinh chỉnh**: Sau khi hoàn thành quá trình huấn luyện mô hình cơ bản, quá trình fine-tuning được thực hiện để tối ưu hóa hiệu suất của mô hình dự đoán nồng độ CO₂ thông qua các bước sau:
* **Điều chỉnh siêu tham số**: Thử nghiệm các giá trị khác nhau cho các tham số như learning rate, số lượng lớp, số lượng đơn vị ẩn, và số lượng filter (đối với mô hình Deep Learning) để tìm ra cấu hình tối ưu cho mô hình. Các tham số này có thể ảnh hưởng lớn đến khả năng hội tụ và độ chính xác của mô hình.
* **Thay đổi kiến trúc mạng**: Thêm hoặc bớt các lớp trong mô hình, điều chỉnh kích thước filter hoặc số lượng nút trong các lớp ẩn để cải thiện khả năng học và khả năng dự đoán của mô hình. Đặc biệt trong LSTM, việc thay đổi số lượng đơn vị ẩn và số lớp LSTM có thể giúp cải thiện hiệu quả mô hình cho dữ liệu chuỗi thời gian.
* **Áp dụng các kỹ thuật regularization**: Để giảm thiểu overfitting, các kỹ thuật như dropout và batch normalization được áp dụng. Điều này giúp mô hình giữ được tính tổng quát và giảm thiểu sai số khi dự đoán trên dữ liệu mới.
* **Kiểm thử trên dữ liệu thực tế**: Sau khi tinh chỉnh, mô hình được kiểm thử trên dữ liệu thực tế để đánh giá khả năng dự đoán nồng độ CO₂.
* **Đánh giá trên tập kiểm định:** Sử dụng tập dữ liệu kiểm định chưa từng được mô hình huấn luyện để đánh giá độ chính xác và khả năng tổng quát hóa của mô hình.
* **Kiểm tra tính khả chuyển:** Đảm bảo rằng mô hình hoạt động tốt trên các tập dữ liệu từ các nguồn khác nhau hoặc điều kiện môi trường khác nhau, chẳng hạn như khác biệt về thời tiết hoặc mùa.
* **Phản hồi và điều chỉnh**: Dựa trên kết quả kiểm thử, mô hình có thể tiếp tục điều chỉnh các tham số và kiến trúc để cải thiện hiệu suất trước khi triển khai trên dữ liệu thực tế.

## 1.5. Kết luận

Học máy và học sâu là hai lĩnh vực quan trọng trong trí tuệ nhân tạo, với những đóng góp lớn lao cho sự phát triển của công nghệ và ứng dụng thực tiễn. Học máy mang lại các giải pháp hiệu quả cho nhiều bài toán phân tích và dự đoán dữ liệu, trong khi học sâu tạo ra những bước đột phá trong các bài toán xử lý phức tạp hơn, như hình ảnh, âm thanh và văn bản. Tuy nhiên, cả hai đều đối mặt với những thách thức về dữ liệu, tính giải thích của mô hình và yêu cầu tài nguyên tính toán. Với sự phát triển không ngừng của công nghệ, học máy và học sâu hứa hẹn sẽ tiếp tục là những công cụ quan trọng trong việc xây dựng các hệ thống thông minh, nâng cao hiệu suất và cải thiện chất lượng cuộc sống.

# Chương 2.

# Phân tích dữ liệu

## 2.1. Bộ dữ liệu

### 2.1.1. Giới thiệu bộ dữ liệu

Dự án này sử dụng bộ dữ liệu gồm các thông số môi trường liên quan đến nồng độ CO2 tại Hà Nội, được thu thập từ nhiều nguồn đáng tin cậy. Bộ dữ liệu bao gồm các yếu tố chính như: nồng độ CO2, nhiệt độ, độ ẩm, áp suất khí quyển, tốc độ gió, mật độ dân cư và các thông số khác ảnh hưởng đến chất lượng không khí và nồng độ CO2 trong không khí.

Dữ liệu được thu thập theo chuỗi thời gian từ các trạm quan trắc môi trường, giúp đảm bảo tính nhất quán và độ chính xác. Mỗi mục dữ liệu được gán nhãn với thời gian cụ thể để hỗ trợ cho việc dự đoán nồng độ CO2 trong tương lai. Dữ liệu được chuẩn hóa để đảm bảo các giá trị nằm trong khoảng hợp lý, giúp mô hình học máy xử lý dễ dàng và nhanh chóng.

Bộ dữ liệu này bao phủ nhiều biến số và được kiểm soát chất lượng chặt chẽ, nhằm giảm thiểu các yếu tố ngoại cảnh như lỗi đo lường, sự thay đổi đột ngột của môi trường, và sai số trong quá trình thu thập. Dữ liệu cũng được làm sạch để loại bỏ các giá trị bất thường và thiếu, nhằm nâng cao hiệu suất và độ chính xác của mô hình dự đoán nồng độ CO2.

### 2.1.2. Nguồn dữ liệu

* Dữ liệu sử dụng cho dự án này được lấy từ nhiều nguồn khác nhau để đảm bảo tính phong phú và đa dạng. Một phần dữ liệu được thu thập từ trang web AccuWeather, cung cấp các thông tin thời tiết và nồng độ CO₂ tại khu vực Hà Nội.
* Dự án còn sử dụng cảm biến chất lượng không khí MQ-135 để thu thập dữ liệu trực tiếp từ môi trường sống, bao gồm các thông số như nhiệt độ, độ ẩm, và nồng độ CO₂ theo thời gian thực.
* Cảm biến chất lượng không khí MQ-135:
* Cảm biến MQ-135 là một loại cảm biến khí phổ biến trong các ứng dụng giám sát chất lượng không khí, đặc biệt là trong các môi trường cần phát hiện các khí ô nhiễm như ammonia, nitơ oxit, benzen, các hợp chất hữu cơ dễ bay hơi, carbon dioxide, và nhiều khí độc hại khác.
* Cảm biến MQ-135 sử dụng nguyên lý thay đổi điện trở của vật liệu nhạy cảm khi tiếp xúc với các khí mục tiêu. Cấu tạo của cảm biến bao gồm một điện cực nóng và một lớp vật liệu cảm biến. Khi có khí ô nhiễm tiếp xúc với bề mặt của vật liệu nhạy cảm, nó sẽ gây ra sự thay đổi trong điện trở của lớp vật liệu này. Sự thay đổi này sẽ được đo và chuyển thành tín hiệu điện tử, phản ánh mức độ ô nhiễm trong không khí.



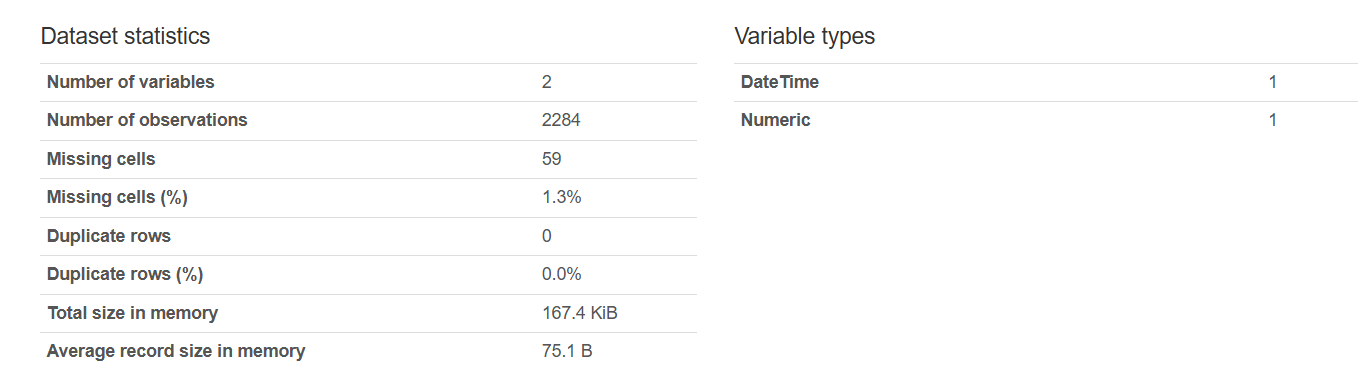
Hình 2. 1. Cảm biến MQ-135

### 2.1.3. Đặc điểm của dữ liệu

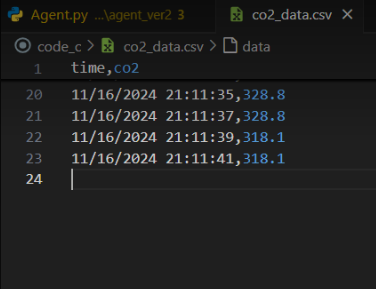
Bộ dữ liệu sử dụng trong dự án này có định dạng đồng nhất, nhằm tối ưu hóa cho quá trình huấn luyện mô hình. Mặc dù dữ liệu được thu thập từ nhiều nguồn khác nhau, tất cả các yếu tố đều được chuẩn hóa để đảm bảo tính nhất quán trong quá trình xử lý và phân tích. Các đặc trưng chính trong bộ dữ liệu bao gồm các chỉ số môi trường quan trọng như nồng độ CO₂, nhiệt độ, độ ẩm, áp suất khí quyển, và tốc độ gió.

Dữ liệu được lưu trữ dưới định dạng CSV, một trong những định dạng phổ biến và dễ sử dụng cho các ứng dụng học máy và xử lý dữ liệu lớn. Định dạng CSV không chỉ giúp tiết kiệm dung lượng lưu trữ nhờ cơ chế nén hiệu quả mà còn giúp mô hình dễ dàng xử lý các yếu tố đầu vào mà không gặp quá nhiều khó khăn về tính toán. Định dạng này đảm bảo rằng mô hình có thể xử lý dữ liệu nhanh chóng và hiệu quả, đồng thời vẫn duy trì được độ chính xác cao trong việc trích xuất các đặc trưng cần thiết.

Về kích thước và cấu trúc dữ liệu, mỗi bản ghi trong bộ dữ liệu đại diện cho các thông số được thu thập theo chuỗi thời gian (theo ngày hoặc giờ) tại Hà Nội. Bộ dữ liệu này đã được làm sạch, loại bỏ các giá trị thiếu và bất thường để đảm bảo chất lượng. Các thông số được chuẩn hóa để mô hình có thể nhận diện và học tốt hơn các xu hướng cũng như mối quan hệ phức tạp giữa các yếu tố, từ đó tăng độ chính xác trong dự đoán nồng độ CO₂.



Hình 2. 2. Đặc điểm bộ dữ liệu Accuweather



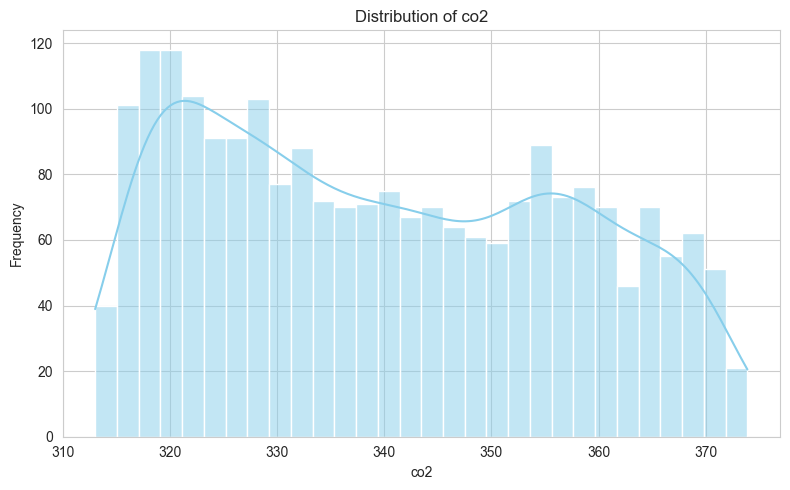
Hình 2. 3. Đặc điểm bộ dữ liệu lấy từ thiết bị IoT

## 2.2. Phân tích dữ liệu bùng nổ EDA

### 2.2.1. Các thông số thống kê và trực quan hoá dữ liệu

* Phân phối nồng độ khí CO2:

Nồng độ khí CO2 là một yếu tố quan trọng ảnh hưởng đến chất lượng không khí và sức khỏe con người. Nồng độ CO2 thể hiện mức độ ô nhiễm không khí và sự thay đổi trong môi trường. Mức CO2 quá cao có thể gây nguy hiểm cho sức khỏe, trong khi mức CO2 quá thấp có thể phản ánh một môi trường thiếu sự trao đổi khí hiệu quả. Phân tích nồng độ CO2 giúp phát hiện các khu vực có chất lượng không khí không tốt và đưa ra các biện pháp cải thiện phù hợp, từ đó đảm bảo môi trường sống lành mạnh hơn.



Hình 2. 4. Biểu đồ về sự phân bố khí Co2

* **Giá trị trung bình: 340.14**

Giá trị trung bình của nồng độ Co2 là 340.14, cho thấy rằng mức độ Co2 trong không khí ở mức trung bình khá ổn định, không quá cao hoặc quá thấp.

* **Trung vị: 338.30**

Giá trị trung vị của nồng độ Co2 là 338.30, gần với giá trị trung bình. Điều này cho thấy rằng phân phối nồng độ Co2 trong tập dữ liệu khá đồng đều và không có sự lệch lớn về phía giá trị cực đại hay cực tiểu.

* **Q1: 324.80**

Phân vị thứ nhất (Q1) là 324.80, cho thấy rằng 25% các mẫu có nồng độ Co2 dưới mức này. Đây là nhóm có mức Co2 thấp hơn trong tập dữ liệu.

* **Q3: 358.04**

Phân vị thứ ba (Q3) là 358.04, cho thấy rằng 75% các mẫu có nồng độ Co2 dưới mức này, và 25% các mẫu có nồng độ Co2 cao hơn. Điều này cho thấy có một số mẫu có nồng độ CO2 cao hơn mức trung bình nhưng không quá lệch về phía giá trị cực đại.

* **Giá trị nhỏ nhất: 313.00**

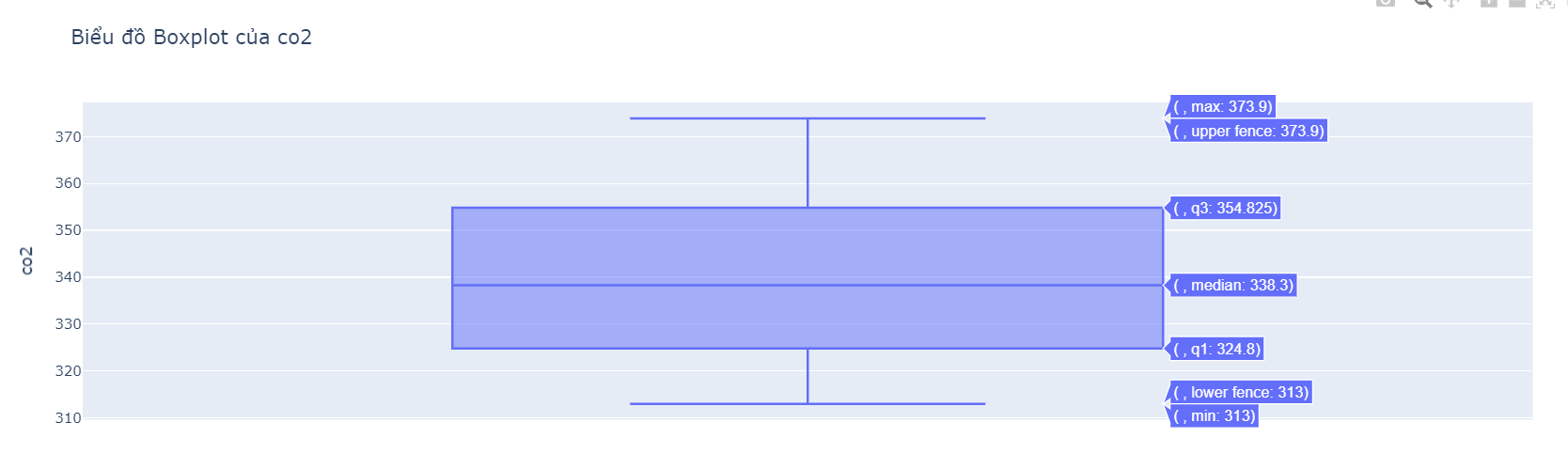
Nồng độ Co2 nhỏ nhất trong tập dữ liệu là 313.00, cho thấy có một số mẫu có nồng độ Co2 rất thấp. Những mẫu này phản ánh chất lượng không khí khá tốt ở ngày có nồng độ Co2 đó.

* **Giá trị lớn nhất: 373.90**

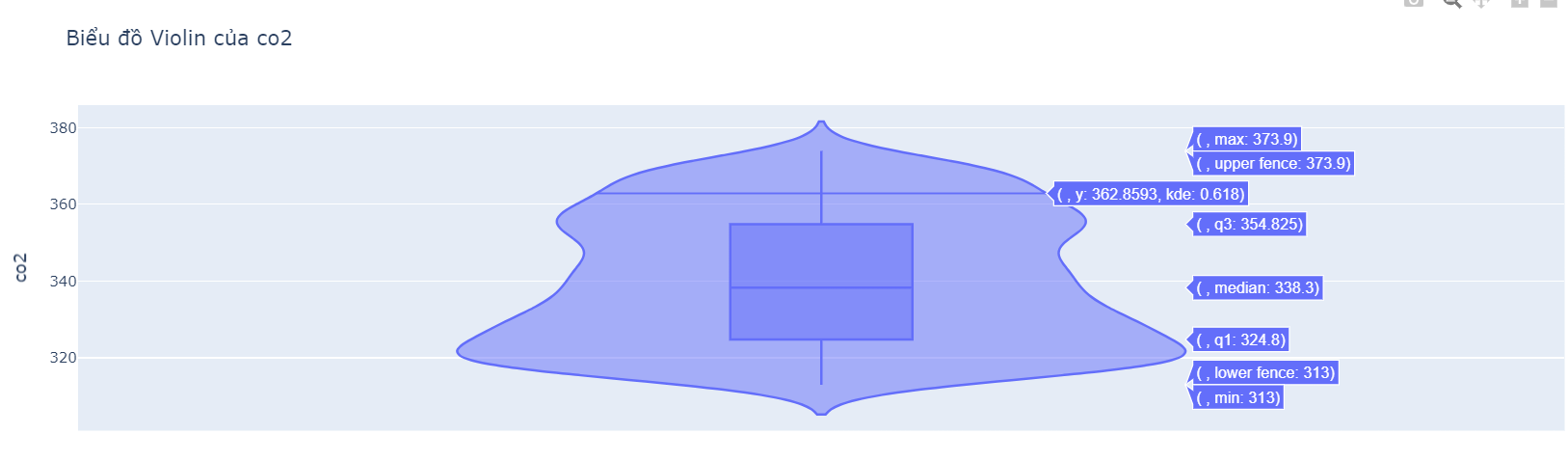
Nồng độ Co2 lớn nhất trong tập dữ liệu là 373.90, cho thấy có một số giá trị cao đáng kể phản ánh chất lượng không khí khá ô nhiễm ở những ngày có nồng độ Co2 này.

* **Độ lệch chuẩn: 17.00**

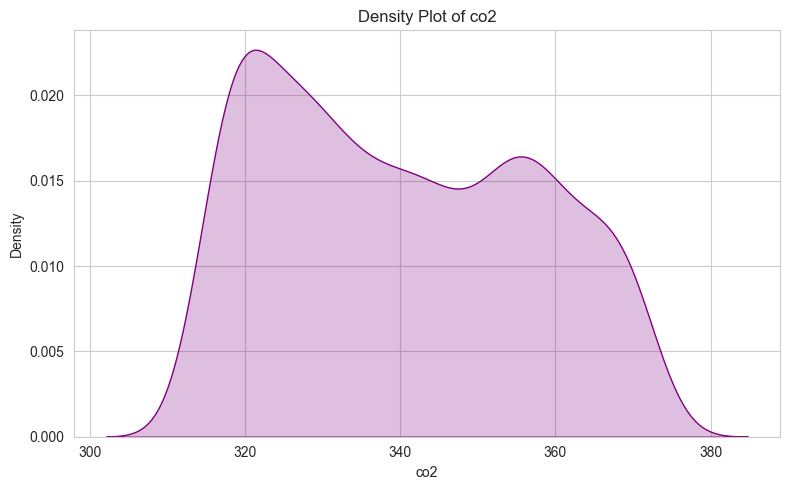
Độ lệch chuẩn của nồng độ Co2 là 17.00, cho thấy mức độ phân tán tương đối lớn giữa các giá trị. Mặc dù có sự biến động về nồng độ CO2, nhưng độ lệch chuẩn này không quá cao, cho thấy phần lớn các giá trị có nồng độ CO2 gần với giá trị trung bình.



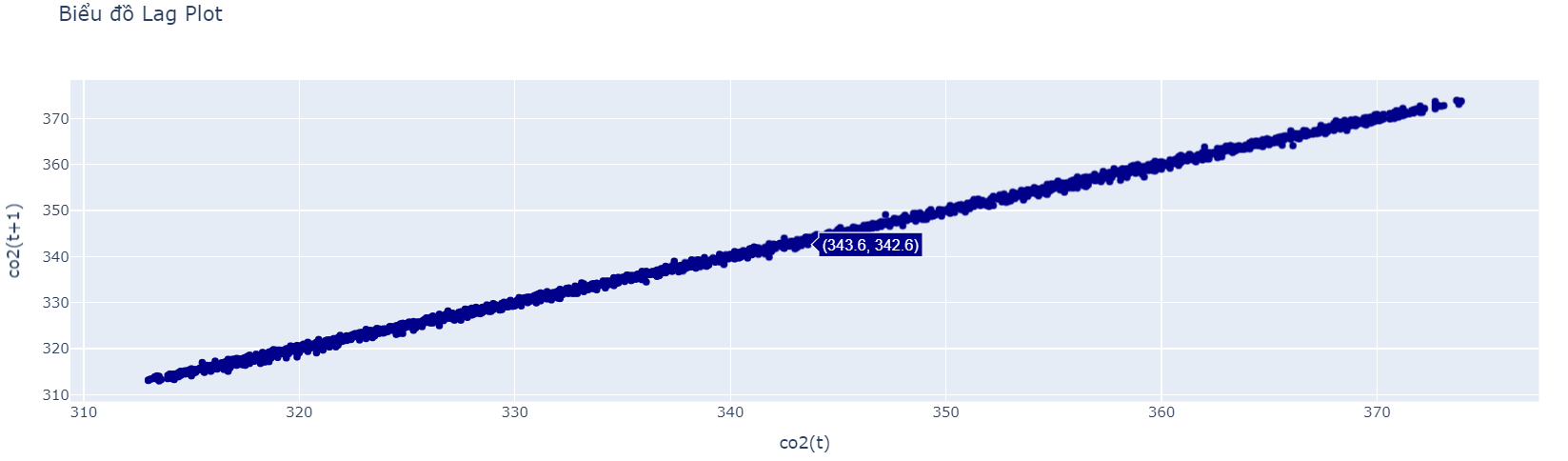
Hình 2. 5. Biểu đồ hộp phân phối nồng độ CO2



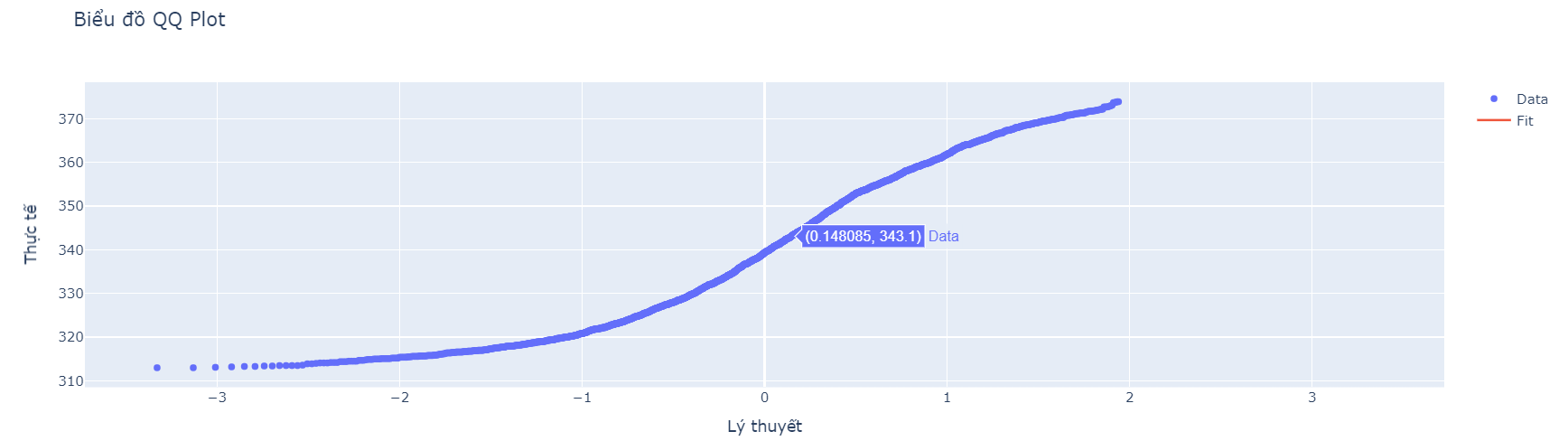
Hình 2. 6. Biểu đồ Violet phân phối nồng độ CO2



Hình 2. 7. Biểu đồ Density Plot nồng độ CO2



Hình 2. 8. Biểu đồ Log Plot phân phối nồng độ CO2



Hình 2. 9. Biểu đồ QQ Plot phân phối nồng độ CO2

## 2.3 Kết luận

Bộ dữ liệu nồng độ khí CO2 có sự phân bố đa dạng và đầy đủ, cung cấp các thông tin về mức CO2 trong không khí từ nhiều nguồn khác nhau. Nhờ vậy, mô hình có thể học được các đặc trưng của nồng độ CO2 ở nhiều mức độ khác nhau, hỗ trợ cho việc dự đoán chính xác hơn.

Bộ dữ liệu này có số lượng dữ liệu đầu vào khá lớn với nhiều giá trị nồng độ CO2 khác nhau, điều này giúp mô hình có nhiều tài nguyên để cải thiện độ chính xác. Ngoài ra, nồng độ CO2 trong dữ liệu có sự phân bố tương đối đồng đều, không có sự lệch lớn về phía giá trị cực đại hoặc cực tiểu, giúp việc phân tích và xây dựng mô hình trở nên chính xác hơn.

Tuy nhiên, vẫn có một số giá trị CO2 quá thấp hoặc quá cao, cũng như có sự phân tán nhất định trong độ lệch chuẩn. Điều này cần được xử lý để đảm bảo tính nhất quán và chất lượng dữ liệu đầu vào.

Tổng kết lại, bộ dữ liệu có cấu trúc tốt, với sự đa dạng và phân bố đồng đều, giúp mô hình có thể thực hiện phân tích nồng độ CO2 một cách hiệu quả. Tuy nhiên, cần thực hiện các bước tiền xử lý như chuẩn hóa và loại bỏ các giá trị ngoại lai để nâng cao chất lượng dữ liệu đầu vào, từ đó cải thiện độ chính xác của mô hình.

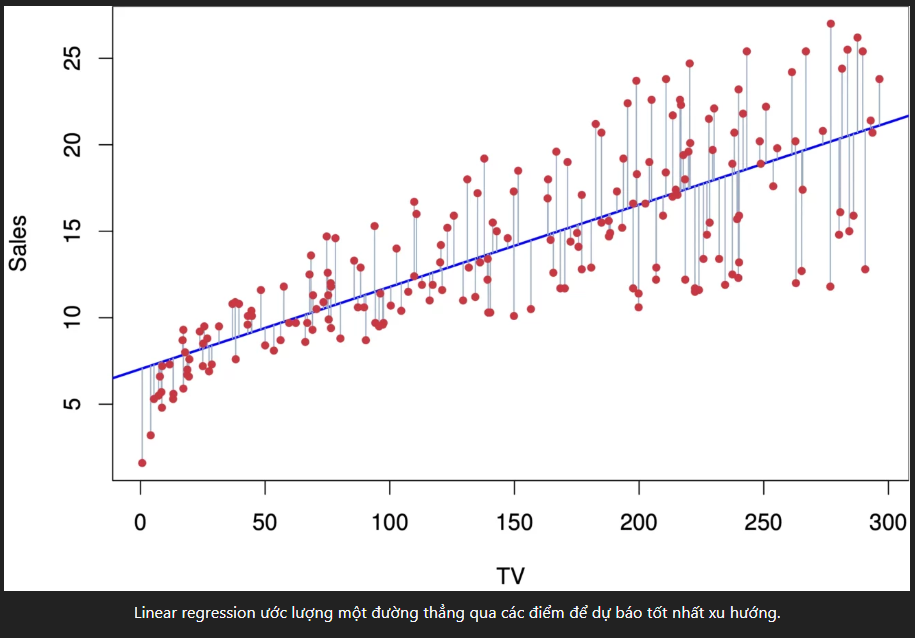
# CHƯƠNG 3.

# Thực nghiệm

## 3.1. Mô hình

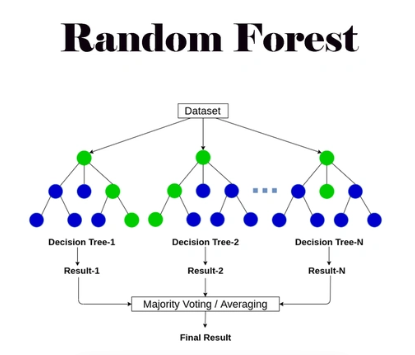
### 3.1.1. Hồi quy tuyến tính

* **Kiến trúc cơ bản:**
* Hồi quy tuyến tính là một mô hình đơn giản, với một hàm số tuyến tính để mô tả mối quan hệ giữa các biến đầu vào (đặc trưng) và biến đầu ra (biến mục tiêu).
* Mô hình có công thức: y=w1x1+w2x2+...+wnxn+by = w\_1x\_1 + w\_2x\_2 + ... + w\_nx\_n + by=w1​x1​+w2​x2​+...+wn​xn​+b, trong đó:
* yyy: Biến dự đoán.
* xix\_ixi​: Các biến đầu vào (đặc trưng).
* wiw\_iwi​: Trọng số (weight) của mỗi đặc trưng.
* bbb: Hệ số tự do (bias).
* **Cơ chế hoạt động:**
* Mô hình cố gắng tìm các trọng số wiw\_iwi​ và hệ số bbb sao cho tổng bình phương của sự sai khác giữa giá trị thực và giá trị dự đoán là nhỏ nhất.
* Quá trình tối ưu hóa thường sử dụng các thuật toán như Gradient Descent để tìm ra giá trị tối ưu cho các tham số.
* Sau khi huấn luyện, mô hình có thể đưa ra dự đoán cho các giá trị mới bằng cách sử dụng công thức tuyến tính đã học.

****

Hình 3. 1.Mô hình hồi quy tuyến tính

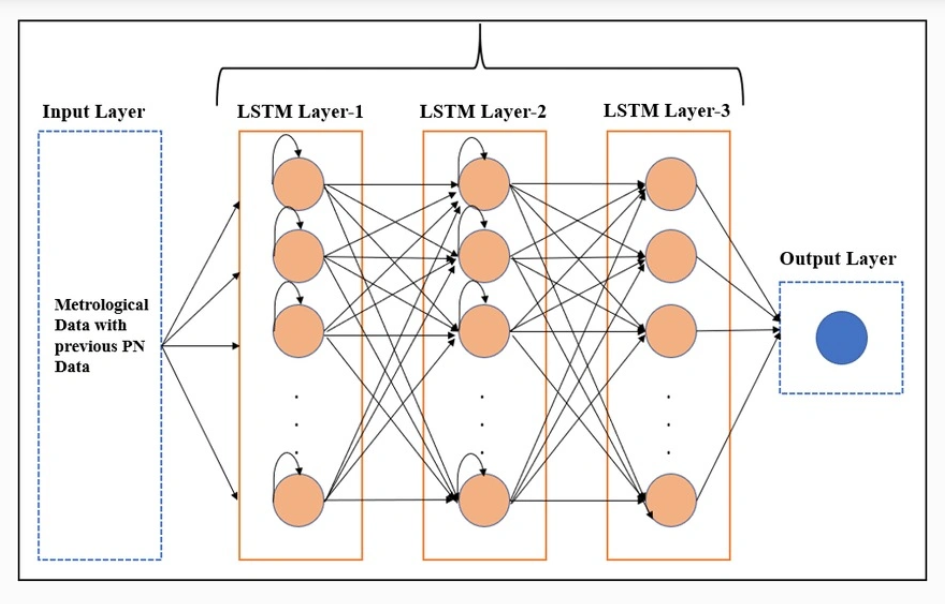
### 3.1.2. Rừng ngẫu nhiên

****

Hình 3. 2. Mô hình rừng ngẫu nhiên

* **Kiến trúc cơ bản:**
* Rừng ngẫu nhiên là một tập hợp các cây quyết định hoạt động độc lập. Mỗi cây được huấn luyện trên một tập dữ liệu con ngẫu nhiên.
* Mỗi nút trong từng cây chỉ xem xét một số lượng giới hạn các đặc trưng để phân chia, giúp tăng tính ngẫu nhiên.
* **Cơ chế hoạt động:**
* Khi có một đầu vào mới, mỗi cây trong rừng sẽ đưa ra một dự đoán riêng lẻ.
* Dự đoán cuối cùng của rừng ngẫu nhiên là kết quả trung bình (cho bài toán hồi quy) hoặc kết quả đa số (cho bài toán phân loại) của tất cả các cây.
* Nhờ sự kết hợp của nhiều cây, rừng ngẫu nhiên có khả năng giảm overfitting và mang lại độ chính xác cao hơn

### 3.1.3. LSTM

****

Hình 3. 3. Mô hình LSTM

* **Kiến trúc cơ bản:**
* LSTM là một dạng của RNN được thiết kế để khắc phục vấn đề gradient vanishing trong quá trình học.
* LSTM có cấu trúc gồm nhiều ô nhớ kết nối nối tiếp theo thời gian. Mỗi ô nhớ chứa ba cổng chính: cổng quên, cổng đầu vào, và cổng đầu ra.
* Các cổng này giúp LSTM điều chỉnh lượng thông tin cần giữ lại và lượng thông tin cần quên, nhằm lưu giữ được các mối quan hệ dài hạn trong dữ liệu chuỗi thời gian.
* **Cơ chế hoạt động:**
* **Cổng quên**: Xác định phần nào của thông tin từ bước trước đó cần được loại bỏ. Thông tin này được điều chỉnh dựa trên giá trị của đầu vào và trạng thái ẩn từ bước trước.
* **Cổng đầu vào**: Xác định lượng thông tin mới cần lưu trữ vào ô nhớ. Thông tin đầu vào được kết hợp với trạng thái ẩn trước đó và kích hoạt thông qua hàm sigmoid, sau đó được cập nhật vào trạng thái của ô nhớ.
* **Cổng đầu ra**: Xác định phần nào của thông tin từ ô nhớ sẽ được xuất ra làm trạng thái ẩn cho bước tiếp theo.
* **Trạng thái ô nhớ:** Là nơi lưu giữ thông tin chính của mỗi bước và được cập nhật qua từng bước thời gian

## 3.2. Huấn luyện mô hình

### 3.2.1. Dữ liệu đầu vào

Dữ liệu được lưu trữ trong một file CSV, bao gồm tổng cộng 2225 hàng và 2 cột: “Time” và “Co2”. Cột “Time” chứa thông tin về thời gian đo, còn cột “Co2” ghi lại giá trị nồng độ khí Co2 tương ứng tại từng thời điểm. Bộ dữ liệu không chỉ cung cấp các giá trị nồng độ Co2 mà còn giúp phân tích sự biến động của nồng độ Co2 theo thời gian.

### 3.2.2. Huấn luyện mô hình Hồi quy tuyến tính

* **Quá trình huấn luyện**: Quá trình huấn luyện mô hình hồi quy tuyến tính gồm các bước chính:
* **Khởi tạo mô hình**: Mô hình hồi quy tuyến tính được thiết lập với các tùy chọn Regularization như Ridge và Lasso để kiểm soát overfitting.
* **Chọn hàm mất**: MSE được sử dụng làm hàm mất để đánh giá độ lệch giữa giá trị dự đoán và giá trị thực, phù hợp với bài toán hồi quy.
* **Chọn thuật toán tối ưu**: Grid Search kết hợp với Cross-Validation được sử dụng để lựa chọn phương pháp hồi quy và giá trị alpha tối ưu.
* **Thiết lập siêu tham số**: Các tham số như alpha được tối ưu hóa để điều chỉnh mức độ regularization và giảm thiểu khả năng overfitting.
* **Hàm mất và thuật toán tối ưu**:
* **Hàm mất**: MSE giúp đo lường sai số bình phương giữa giá trị dự đoán và giá trị thực, từ đó tối ưu hóa quá trình hồi quy.
* **Thuật toán tối ưu**: Grid Search và Cross-Validation giúp tìm bộ tham số tốt nhất cho mô hình Linear Regression, đặc biệt là giá trị alpha trong Ridge và Lasso.

### 3.2.3. Huấn luyện mô hình Rừng ngẫu nhiên

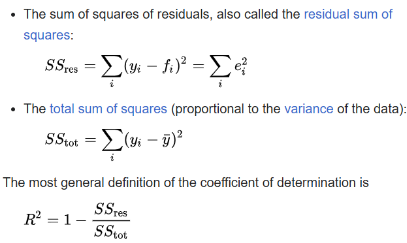
* **Quá trình huấn luyện**: Quá trình huấn luyện mô hình Random Forest bao gồm các bước sau:
* **Khởi tạo mô hình**: Mô hình rừng ngẫu nhiên được khởi tạo với một số lượng cây nhất định (n\_estimators) cùng với các tham số khác như max\_depth, min\_samples\_split, và min\_samples\_leaf để điều chỉnh cấu trúc cây.
* **Chọn hàm mất**: Với bài toán hồi quy, MSE thường được sử dụng để đo độ lệch giữa giá trị dự đoán và giá trị thực, giúp tối ưu hóa quá trình học của mô hình.
* **Chọn thuật toán tối ưu**: Grid Search được sử dụng để tìm các siêu tham số tối ưu cho mô hình, bao gồm số lượng cây và độ sâu của mỗi cây.
* **Thiết lập siêu tham số**: Các giá trị cho n\_estimators, max\_depth, min\_samples\_split, và min\_samples\_leaf được tinh chỉnh thông qua Grid Search, giúp mô hình đạt độ chính xác tối ưu trên tập kiểm tra.
* **Hàm mất và thuật toán tối ưu**:
* **Hàm mất**: MSE giúp mô hình tối ưu độ chính xác trong bài toán hồi quy bằng cách giảm thiểu sai số bình phương giữa dự đoán và thực tế.
* **Thuật toán tối ưu**: Grid Search và Cross-Validation giúp lựa chọn bộ tham số tốt nhất cho mô hình Random Forest, nâng cao hiệu suất và giảm thiểu overfitting

### 3.2.4. Huấn luyện mô hình LSTM

* **Quá trình huấn luyện**: Quá trình huấn luyện mô hình LSTM gồm các bước sau:
* **Khởi tạo mô hình**: Mô hình LSTM được khởi tạo với các lớp tuần hoàn, kích thước ẩn , và số lớp để tối ưu khả năng học chuỗi thời gian của mô hình.
* **Chọn hàm mất**: MSE được chọn làm hàm mất cho bài toán hồi quy, giúp đo lường độ lệch giữa dự đoán và thực tế.
* **Chọn thuật toán tối ưu**: Sử dụng thuật toán tối ưu Adam với khả năng điều chỉnh tốc độ học và động lượng để cải thiện quá trình hội tụ.
* **Thiết lập siêu tham số**: Thiết lập các tham số như kích thước, tỷ lệ học tập, và số lượng epoch dựa trên thử nghiệm để đạt hiệu suất tốt nhất cho mô hình.
* **Hàm mất và thuật toán tối ưu**:
* **Hàm mất**: MSE được dùng để giảm thiểu độ lệch bình phương giữa dự đoán và thực tế, phù hợp với bài toán dự đoán chuỗi thời gian.
* **Thuật toán tối ưu**: Adam giúp điều chỉnh tốc độ học dựa trên độ dốc và giá trị trung bình của gradient, giúp mô hình hội tụ nhanh hơn và tránh overfitting

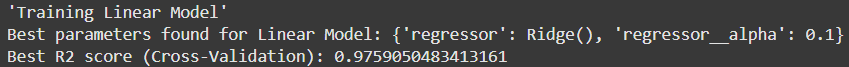
### 3.2.5 Kết quả huấn luyện

* **Đánh giá hiệu quả:** Sau mỗi epoch huấn luyện, mô hình được đánh giá trên tập kiểm tra bằng các chỉ số sau:
* **Hệ số xác định (R² Score):** Chỉ số R² là một thước đo đánh giá độ phù hợp của mô hình với dữ liệu. R² càng gần 1 cho thấy mô hình dự đoán càng chính xác và phù hợp với dữ liệu

****

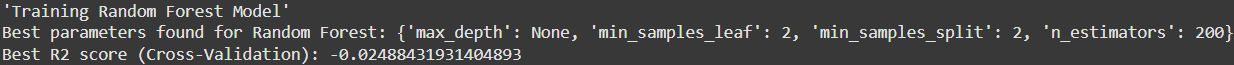
Hình 3. 4. Hệ số xác định

* **Mô hình Hồi quy tuyến tính**



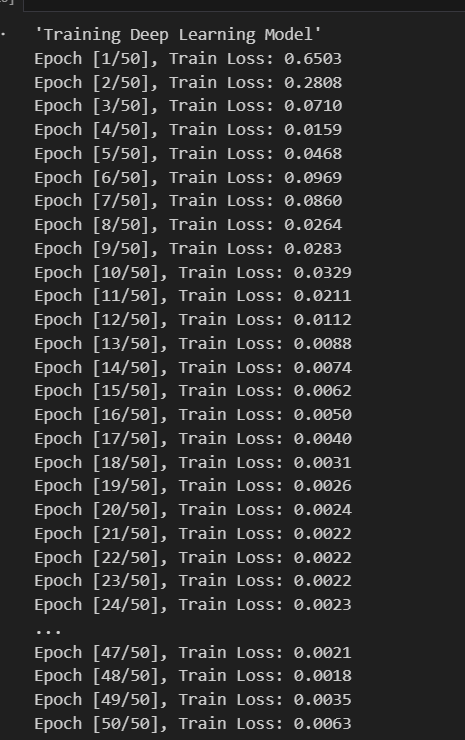
Hình 3. 5. Kết quả huấn luyện mô hình Hồi quy tuyến tính

* R² Score (Cross-Validation): 0.9759
* R² Score trên tập kiểm tra: 0.9907
* Đánh giá: Hồi quy tuyến tính có R² cao, gần với giá trị 1, cho thấy mô hình này rất phù hợp với dữ liệu và có khả năng dự đoán chính xác nồng độ CO₂. Mô hình này hoạt động tốt nhờ vào cấu trúc đơn giản và độ tuyến tính của dữ liệu
* **Mô hình Rừng ngẫu nhiên**



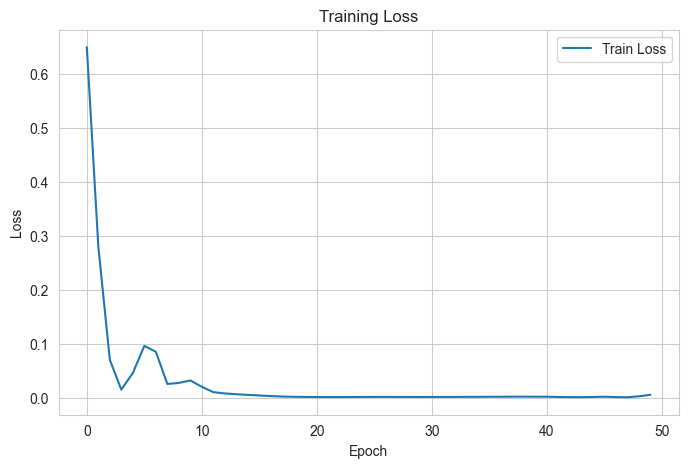
Hình 3. 6. Kết quả huấn luyện mô hình Rừng ngẫu nhiên

* R² Score (Cross-Validation): -0.0248
* R² Score trên tập kiểm tra: -1.0463
* Đánh giá: Rừng ngẫu nhiên có R² âm trên tập kiểm tra, điều này cho thấy mô hình không phù hợp với dữ liệu và có khả năng bị overfitting hoặc không đủ tính tổng quát. Kết quả này có thể do cấu trúc cây ngẫu nhiên không phù hợp với dữ liệu chuỗi thời gian trong bài toán dự đoán nồng độ CO2.
* **Mô hình LSTM**



Hình 3. 7. Kết quả huấn luyện mô hình LSTM

* Độ chính xác trên tập huấn luyện:
* Mô hình đạt được độ chính xác là 96.16% trên tập huấn luyện, đây là một giá trị tương đối tốt.
* Hàm mất mát trên tập huấn luyện:
* Giá trị hàm mất mát trên tập huấn luyện giảm dần từ 0.6503 ở epoch 1 xuống 0.0159 ở epoch 4. Đây là dấu hiệu tích cực cho thấy mô hình đang học hiệu quả và giảm thiểu sai số qua các bước lặp.



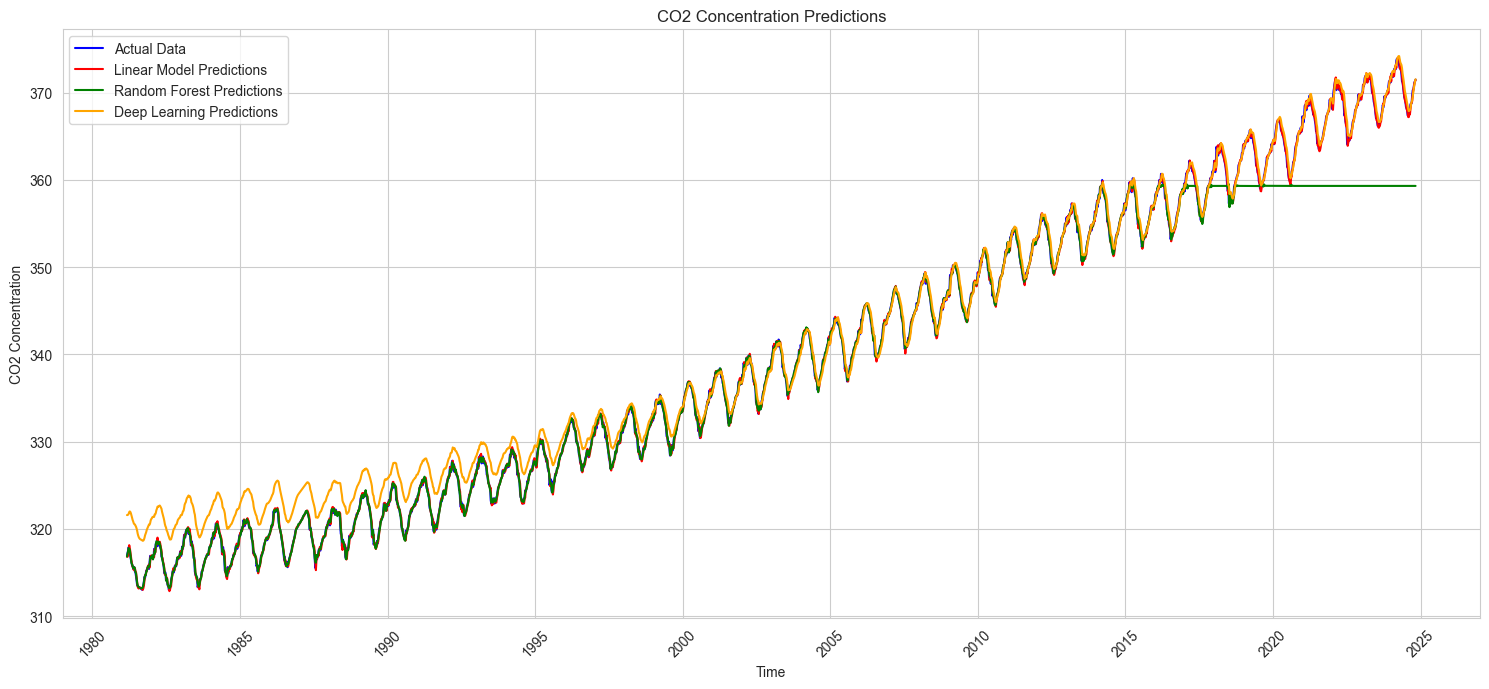
Hình 3. 8. Biểu đô Train Loss

## 3.3. Đánh giá mô hình

### 3.3.1. Đánh giá độ chính xác của mô hình

* Độ chính xác trên tập kiểm tra
* Mô hình hồi quy tuyến tính đạt độ chính xác 99.07%, độ chính xác cao nhất trong 3 mô hình (Rừng ngẫu nhiên (0.00%) – LSTM (98.00%)). Điều này cho thấy mô hình Hồi quy tuyến tính và LSTM có khả năng khái quát hóa khá tốt và không gặp tình trạng quá khớp nghiêm trọng. Với mô hình Rừng ngẫu nhiên không phù hợp đối với tập dữ liệu này.

### 3.3.2. Kết quả dự đoán



Hình 3. 9. Kết quả dự đoán 3 mô hình

### 3.3.3. Kết luận

Mô hình Hồi quy tuyến tính và LSTM hoạt động tốt trong việc dự đoán nồng độ khí Co2 với độ chính xác cao. Tuy nhiên, với mô hình Rừng ngẫu nhiên lại không phù hợp áp dụng trong bài toán dự đoán này.

# Kết luận

## 1. Kết quả đã đạt được

Xây dựng thành công 3 mô hình Hồi quy tuyến tính, Rừng ngẫu nhiên và LSTM, từ đó so sánh độ chính xác của các mô hình trong bài toán dự đoán nồng độ khí Co2 . Sau quá trình huấn luyện, mô hình Hồi quy tuyến tính đạt được độ chính xác cao nhất với 97.59% trên tập huấn luyện và 99.07% trên tập kiểm tra, xếp thứ hai là mô hình LSTM đạt độ chính xác 96.16% trên tập huấn luyện và 98% trên tập kiểm tra. Xếp cuối là mô hình Rừng ngẫu nhiên với 0% trên cả 2 tập huấn luyện và kiểm tra. Từ đó rút ra nhận xét mô hình Rừng ngẫu nhiên không phù hợp cho bài toán dự đoán này.

Hai mô hình Hồi quy tuyến tính và LSTM hoạt động tốt trong việc dự đoán nồng độ khí Co2, với sự chính xác cao và mức loss thấp. Quá trình huấn luyện được theo dõi cẩn thận và đảm bảo hiệu quả huấn luyện.

## 2. Phương hướng phát triển đề tài

Tối ưu hóa mô hình: Mặc dù mô hình đạt kết quả tốt, vẫn còn một số dao động trong độ chính xác kiểm thử và độ mất mát kiểm thử. Các phương pháp như tăng cường dữ liệu (data augmentation), sử dụng kỹ thuật chuẩn hóa (như loại bỏ ngẫu nhiên), hoặc tinh chỉnh tốc độ học có thể giúp mô hình ổn định hơn và giảm thiểu mất mát.

Mở rộng tập dữ liệu: Mô hình có thể được cải thiện thêm nếu có một tập dữ liệu đa dạng và phong phú hơn. Việc thu thập nhiều giá trị nồng độ Co2 hơn sẽ giúp mô hình trở nên học được các dữ liệu đa dạng hơn trong các điều kiện thực tế.

Ứng dụng thực tế: Đề tài có thể tiếp tục phát triển với việc tích hợp mô hình vào các hệ thống dự đoán nồng độ khí Co2 trong thực tế, ví dụ như hệ thống giám sát chất lượng không khí, hệ thống dự đoán Co2 trong sản xuất công nghiệp, hệ thống dự đoán khí Co2 trong dự đoán biến đổi khí hậu…

# Tài liệu tham khảo