

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC ĐẠI NAM
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO BÀI TẬP LỚN

**XÂY DỰNG MÔ HÌNH ANN DỰ ĐOÁN LƯỢNG KHÍ
THẢI CO₂ CỦA XE Ô TÔ**

Sinh viên thực hiện : Lâm Tiến Dũng
Nguyễn Trường Giang
Trần Trọng Đức
Đỗ Tiến Đại

Ngành : Công nghệ thông tin

Giảng viên hướng dẫn : ThS. Lê Thị Thùy Trang
Ths. Trần Thái Khánh

Lời cảm ơn

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành và sâu sắc nhất đến **ThS. Lê Thị Thùy Trang** và **ThS. Nguyễn Thái Khánh**, những người đã tận tình hướng dẫn, hỗ trợ và đóng góp những ý kiến quý báu trong suốt quá trình thực hiện đề tài "**Xây dựng mô hình Mạng Nơ-ron Nhân tạo (ANN) dự đoán lượng khí thải CO của xe ô tô**".

Trong suốt thời gian nghiên cứu, em đã nhận được sự chỉ dẫn tận tình từ hai thầy cô, không chỉ về mặt kiến thức chuyên môn mà còn về phương pháp tư duy, tiếp cận vấn đề và kỹ năng phân tích dữ liệu. Những bài giảng và tài liệu quý báu của thầy cô đã giúp em hiểu rõ hơn về mô hình **Mạng Nơ-ron Nhân tạo (ANN)**, cách xử lý dữ liệu khí thải và các yếu tố ảnh hưởng như *dung tích động cơ, số xi-lanh, loại nhiên liệu, mức tiêu thụ nhiên liệu*. Nhờ đó, em đã có thể áp dụng hiệu quả các kỹ thuật học máy vào bài toán thực tế này.

Em cũng xin cảm ơn nhà trường và khoa [Tên Khoa] đã tạo điều kiện thuận lợi để em có thể hoàn thành đề tài này. Bên cạnh đó, sự động viên từ gia đình và bạn bè cũng là nguồn động lực lớn giúp em kiên trì theo đuổi nghiên cứu đến cùng.

Cuối cùng, em xin kính chúc **ThS. Lê Thị Thùy Trang** và **ThS. Nguyễn Thái Khánh** dồi dào sức khỏe, luôn thành công trong sự nghiệp giảng dạy và nghiên cứu.

Em xin chân thành cảm ơn!

Tóm tắt

TÓM TẮT

Nghiên cứu này đề xuất một mô hình **Mạng Nơ-ron Nhân tạo (ANN)** để dự đoán lượng khí thải **CO** của xe ô tô dựa trên các yếu tố đầu vào như *dung tích động cơ, số xi-lanh, loại nhiên liệu, mức tiêu thụ nhiên liệu*. Mục tiêu của đề tài là xây dựng một mô hình có khả năng dự đoán chính xác lượng khí thải **CO**, từ đó hỗ trợ việc đánh giá mức độ ảnh hưởng của các loại xe hơi đối với môi trường, đồng thời cung cấp cơ sở cho các chính sách kiểm soát phát thải.

Quá trình nghiên cứu bao gồm các bước: thu thập và tiền xử lý dữ liệu, mã hóa các biến phân loại, chuẩn hóa dữ liệu, xây dựng mô hình ANN, huấn luyện và đánh giá mô hình. Dữ liệu sử dụng trong nghiên cứu được lấy từ các nguồn đáng tin cậy về thông số kỹ thuật và mức tiêu thụ nhiên liệu của nhiều loại xe khác nhau. Sau khi xử lý dữ liệu, mô hình ANN được xây dựng với cấu trúc gồm *một lớp đầu vào, hai lớp ẩn sử dụng hàm kích hoạt ReLU và một lớp đầu ra với hàm kích hoạt tuyến tính*, nhằm tối ưu hóa độ chính xác trong dự đoán.

Kết quả thực nghiệm cho thấy mô hình ANN đạt độ chính xác cao trong việc dự đoán lượng khí thải **CO**, với sai số trung bình thấp. Tuy nhiên, nghiên cứu cũng chỉ ra một số hạn chế, như việc mô hình có thể bị ảnh hưởng bởi sự thiếu cân bằng trong dữ liệu hoặc sự phụ thuộc vào một số đặc trưng cụ thể. Để cải thiện hiệu suất, nghiên cứu đề xuất thử nghiệm các mô hình học sâu khác như **LSTM** hoặc **CNN**, cũng như mở rộng tập dữ liệu để tăng tính tổng quát của mô hình.

Tóm lại, nghiên cứu đã thành công trong việc áp dụng mô hình ANN để dự đoán lượng khí thải **CO**, đồng thời đưa ra các định hướng phát triển nhằm nâng cao độ chính xác và hiệu quả của mô hình trong các ứng dụng thực tế.

Mục lục

1	Giới thiệu	1
1.1	Lý do chọn đề tài	1
1.2	Mục tiêu nghiên cứu	2
1.3	Phạm vi nghiên cứu	2
1.4	Các phương pháp tiếp cận	3
1.4.1	Phương pháp thu thập và xử lý dữ liệu	3
1.4.2	Phương pháp xây dựng mô hình dự đoán	4
1.4.3	Phương pháp đánh giá hiệu suất mô hình	4
1.5	Bố cục đề tài	5
2	Cơ sở lý thuyết	6
2.1	Tổng quan về khí thải CO ₂ của xe ô tô	6
2.1.1	Khí thải CO ₂ và tác động đến môi trường	6
2.1.2	Tiêu chuẩn khí thải và quy định giới hạn CO ₂	6
2.1.3	Các phương pháp giảm phát thải CO ₂ từ xe ô tô	7
2.2	Các yếu tố ảnh hưởng đến khí thải CO ₂	8
2.2.1	Loại nhiên liệu	8
2.2.2	Dung tích động cơ và số xi-lanh	8
2.2.3	Mức tiêu thụ nhiên liệu	8
2.2.4	Công nghệ và hiệu suất động cơ	9
2.2.5	Điều kiện vận hành và thói quen lái xe	9
2.3	Mạng Nơ-ron Nhân tạo (Artificial Neural Network - ANN)	10
2.3.1	Tổng quan về Mạng Nơ-ron Nhân tạo	10
2.3.2	Cấu trúc và hoạt động của ANN	10
2.3.3	Quy trình huấn luyện ANN	11
2.3.4	Ứng dụng của ANN trong dự đoán khí thải CO ₂	11
2.4	Các kỹ thuật tiền xử lý dữ liệu	11
2.4.1	Giới thiệu về tiền xử lý dữ liệu	11
2.4.2	Xử lý giá trị thiếu (Missing Data Handling)	12

2.4.3	Chuẩn hóa và tiêu chuẩn hóa dữ liệu	12
2.4.4	Mã hóa dữ liệu phân loại	12
2.4.5	Phát hiện và xử lý dữ liệu ngoại lai (Outliers)	13
2.4.6	Phân chia tập dữ liệu	13
2.5	Thu thập và xử lý dữ liệu	14
2.5.1	Nguồn dữ liệu	14
2.5.2	Tiền xử lý dữ liệu	14
2.6	Thiết kế mô hình Mạng Nơ-ron Nhân tạo (ANN)	16
2.6.1	Kiến trúc mô hình	16
2.6.2	Xây dựng mô hình bằng MLPRegressor	16
2.6.3	Huấn luyện mô hình	17
2.6.4	Đánh giá mô hình	17
2.6.5	Trực quan hóa kết quả	18
3	KẾT QUẢ VÀ ĐÁNH GIÁ	19
3.1	Kết quả thực nghiệm	19
3.2	Phân tích kết quả thực nghiệm	19
3.3	Đánh giá hiệu suất mô hình ANN	20
3.3.1	Độ chính xác của mô hình	20
3.3.2	Khả năng học mẫu và xu hướng dữ liệu	20
3.3.3	Hướng cải thiện mô hình	20
3.4	So sánh mô hình ANN với các mô hình khác	21
3.4.1	Hiệu suất mô hình	21
3.4.2	So sánh với từng mô hình	21
3.4.3	Nhận xét tổng quát	21
3.5	Phân tích sai số	22
3.5.1	Đánh giá tổng quan sai số	22
3.5.2	Phân tích các nguồn gây sai số	22
3.5.3	Phân bố sai số	22
3.5.4	Hướng cải thiện sai số	22
3.6	Ưu điểm và hạn chế của mô hình ANN	23
4	Kết Luận	24
	Kết Luận	24
1	Ưu điểm	24
2	Hạn chế	24
3	Nhận xét tổng quát	25
4.1	Hướng phát triển trong tương lai	25
1	1. Tối ưu hóa kiến trúc mô hình	25

2	2. Cải thiện chất lượng dữ liệu	25
3	3. Kết hợp với các mô hình khác	26
4	4. Triển khai và ứng dụng thực tế	26
5	5. Nghiên cứu mở rộng	26

Chương 1

Giới thiệu

1.1 Lý do chọn đề tài

Trong bối cảnh ô nhiễm môi trường và biến đổi khí hậu ngày càng trở thành vấn đề nghiêm trọng, việc kiểm soát lượng khí thải CO_2 từ các phương tiện giao thông đóng vai trò quan trọng trong chiến lược phát triển bền vững. Theo các nghiên cứu khoa học, phương tiện giao thông, đặc biệt là xe ô tô sử dụng động cơ đốt trong, là một trong những nguồn phát thải CO_2 chính, góp phần làm gia tăng hiệu ứng nhà kính và biến đổi khí hậu toàn cầu.

Việc dự đoán chính xác lượng khí thải CO_2 của xe ô tô không chỉ giúp các nhà sản xuất và cơ quan quản lý đánh giá mức độ ảnh hưởng của phương tiện đối với môi trường mà còn hỗ trợ người tiêu dùng trong việc lựa chọn các dòng xe thân thiện với môi trường. Tuy nhiên, việc dự đoán này không hề đơn giản do sự ảnh hưởng của nhiều yếu tố như *dung tích động cơ, số xi-lanh, loại nhiên liệu, mức tiêu thụ nhiên liệu*.

Trong những năm gần đây, sự phát triển mạnh mẽ của trí tuệ nhân tạo (AI) và đặc biệt là **Mạng Nơ-ron Nhân tạo (ANN)** đã mở ra nhiều hướng đi mới trong việc phân tích và dự đoán dữ liệu phi tuyến tính, phức tạp như lượng khí thải CO_2 . Mô hình ANN có khả năng học hỏi từ dữ liệu thực tế và tạo ra các dự đoán chính xác dựa trên mối quan hệ giữa các yếu tố đầu vào.

Xuất phát từ thực tiễn này, đề tài "**Xây dựng mô hình Mạng Nơ-ron Nhân tạo (ANN) dự đoán lượng khí thải CO_2 của xe ô tô**" được lựa chọn nhằm áp dụng công nghệ học sâu vào việc giải quyết bài toán môi trường, góp phần nâng cao nhận thức về phát thải khí CO_2 và hỗ trợ xây dựng các chính sách giảm thiểu ô nhiễm hiệu quả hơn.

1.2 Mục tiêu nghiên cứu

Mục tiêu chính của đề tài là xây dựng một mô hình **Mạng Nơ-ron Nhân tạo (ANN)** có khả năng **dự đoán lượng khí thải CO₂** của xe ô tô dựa trên các thông số kỹ thuật như *dung tích động cơ, số xi-lanh, loại nhiên liệu, mức tiêu thụ nhiên liệu*. Để đạt được mục tiêu này, nghiên cứu tập trung vào các nhiệm vụ cụ thể sau:

- **Tìm hiểu tổng quan** về khí thải CO₂ từ xe ô tô và các yếu tố ảnh hưởng đến lượng khí thải.
- **Nghiên cứu các phương pháp dự đoán khí thải CO₂**, bao gồm các mô hình truyền thống và mô hình học sâu (Deep Learning).
- **Thu thập và tiền xử lý dữ liệu** để đảm bảo chất lượng dữ liệu đầu vào cho mô hình ANN, bao gồm chuẩn hóa, mã hóa biến phân loại và xử lý dữ liệu bị thiếu.
- **Xây dựng mô hình ANN phù hợp**, tối ưu số lượng lớp ẩn, số neuron và các siêu tham số để đạt độ chính xác cao trong dự đoán.
- **Huấn luyện và đánh giá mô hình ANN** bằng cách sử dụng các chỉ số như *Mean Squared Error (MSE)*, *Root Mean Squared Error (RMSE)* và so sánh với các mô hình dự đoán khác.
- **Phân tích ưu điểm, hạn chế** của mô hình và đề xuất các hướng phát triển trong tương lai để nâng cao độ chính xác và khả năng ứng dụng thực tế.

Thông qua nghiên cứu này, đề tài hướng đến việc ứng dụng trí tuệ nhân tạo trong lĩnh vực bảo vệ môi trường, giúp cải thiện khả năng dự đoán khí thải CO₂ và hỗ trợ các nhà quản lý, nhà sản xuất và người tiêu dùng trong việc kiểm soát lượng phát thải của phương tiện giao thông.

1.3 Phạm vi nghiên cứu

Đề tài tập trung vào việc xây dựng mô hình **Mạng Nơ-ron Nhân tạo (ANN)** để dự đoán **lượng khí thải CO₂** của xe ô tô dựa trên dữ liệu đầu vào về các thông số kỹ thuật của xe. Phạm vi nghiên cứu được xác định như sau:

- **Về dữ liệu:** Nghiên cứu sử dụng tập dữ liệu về khí thải CO₂ của xe ô tô, bao gồm các đặc trưng như:
 - Dung tích động cơ (Engine Size)
 - Số xi-lanh (Cylinders)
 - Loại nhiên liệu (Fuel Type)

- Mức tiêu thụ nhiên liệu (Fuel Consumption)
- Lượng khí thải CO₂ (CO₂ Emissions) - biến mục tiêu

Dữ liệu được thu thập từ các nguồn công khai như các cơ quan môi trường, bộ giao thông hoặc các trang web chuyên về thông số xe ô tô.

- **Về phương pháp:** Nghiên cứu áp dụng kỹ thuật **học sâu (Deep Learning)** với mô hình **Mạng Nơ-ron Nhân tạo (ANN)**. Quy trình nghiên cứu bao gồm các bước chính:
 - Tiền xử lý dữ liệu: chuẩn hóa dữ liệu, xử lý giá trị thiếu, mã hóa biến phân loại.
 - Xây dựng mô hình ANN: xác định số lớp ẩn, số neuron, hàm kích hoạt.
 - Huấn luyện và đánh giá mô hình: sử dụng các chỉ số như MSE, RMSE, MAE.
 - So sánh với các mô hình khác như hồi quy tuyến tính, cây quyết định.
- **Về giới hạn nghiên cứu:**
 - Chỉ tập trung vào dữ liệu xe ô tô, không xét đến các phương tiện khác như xe máy, xe tải.
 - Không tính đến các yếu tố môi trường bên ngoài như nhiệt độ, độ ẩm, điều kiện giao thông.
 - Phạm vi nghiên cứu chỉ dừng lại ở việc dự đoán lượng khí thải CO₂, không đề cập đến các khí thải khác như NO_x, SO₂.

Bằng cách giới hạn phạm vi nghiên cứu, đề tài đảm bảo tập trung vào việc phát triển mô hình ANN hiệu quả và phù hợp với dữ liệu thực tế, góp phần nâng cao độ chính xác trong dự đoán lượng khí thải CO₂ của xe ô tô.

1.4 Các phương pháp tiếp cận

Để giải quyết bài toán dự đoán lượng khí thải CO₂ của xe ô tô, đề tài áp dụng các phương pháp tiếp cận sau:

1.4.1 Phương pháp thu thập và xử lý dữ liệu

- **Thu thập dữ liệu:** Dữ liệu về lượng khí thải CO₂ được lấy từ các nguồn công khai như cơ quan môi trường, bộ giao thông, hoặc các trang web chuyên cung cấp thông tin về xe ô tô.
- **Tiền xử lý dữ liệu:**

- Xử lý giá trị bị thiếu: thay thế bằng giá trị trung bình (đối với dữ liệu số) hoặc giá trị phổ biến nhất (đối với dữ liệu phân loại).
- Mã hóa các biến phân loại (ví dụ: loại nhiên liệu) bằng phương pháp One-Hot Encoding hoặc Label Encoding.
- Chuẩn hóa dữ liệu (Standardization hoặc Min-Max Scaling) để đảm bảo mô hình học máy hoạt động hiệu quả.

1.4.2 Phương pháp xây dựng mô hình dự đoán

- **Hồi quy tuyến tính (Linear Regression):** - Đây là phương pháp thống kê cơ bản, thiết lập mối quan hệ tuyến tính giữa các đặc trưng đầu vào và lượng khí thải CO₂. - Tuy nhiên, do mối quan hệ giữa các thông số kỹ thuật của xe và khí thải CO₂ có thể phi tuyến tính, nên phương pháp này có thể không đạt hiệu suất cao.
- **Cây quyết định (Decision Tree) và rừng ngẫu nhiên (Random Forest):** - Các mô hình này có khả năng học tốt mối quan hệ phức tạp giữa các biến. - Tuy nhiên, chúng dễ bị overfitting nếu không điều chỉnh tham số phù hợp.
- **Mạng Nơ-ron Nhân tạo (Artificial Neural Network - ANN):** - Đây là phương pháp chính được sử dụng trong đề tài. ANN có khả năng học các mối quan hệ phi tuyến tính phức tạp giữa các đặc trưng của xe ô tô và lượng khí thải CO₂. - Mô hình ANN được thiết kế với nhiều lớp ẩn, mỗi lớp gồm các neuron sử dụng hàm kích hoạt phi tuyến như ReLU để tăng khả năng biểu diễn. - Mô hình được tối ưu bằng thuật toán Adam Optimizer và sử dụng hàm mất mát Mean Squared Error (MSE) để đánh giá sai số.

1.4.3 Phương pháp đánh giá hiệu suất mô hình

- **Chỉ số đánh giá:** - Mean Squared Error (MSE): đo lường mức độ sai số trung bình giữa giá trị dự đoán và giá trị thực tế. - Root Mean Squared Error (RMSE): chỉ số này giúp đánh giá mức độ sai số theo đơn vị gốc của biến mục tiêu. - Mean Absolute Error (MAE): đo lường sai số trung bình theo giá trị tuyệt đối, giúp tránh ảnh hưởng quá lớn từ các giá trị ngoại lai. - Hệ số xác định R²: đánh giá mức độ mô hình giải thích được phương sai của dữ liệu.
- **So sánh mô hình:** - Sau khi huấn luyện, mô hình ANN sẽ được so sánh với các mô hình truyền thống như hồi quy tuyến tính và cây quyết định để đánh giá hiệu quả của từng phương pháp. - Mô hình có giá trị MSE thấp và hệ số R² cao nhất sẽ được xem là mô hình tối ưu cho bài toán dự đoán khí thải CO₂.

Với các phương pháp tiếp cận trên, nghiên cứu kỳ vọng sẽ xây dựng được một mô hình hiệu

quả, có khả năng dự đoán chính xác lượng khí thải CO₂ của xe ô tô, góp phần hỗ trợ trong việc giảm thiểu tác động tiêu cực đến môi trường.

1.5 Bố cục đề tài

Đề tài được chia thành ba chương chính, bao gồm nội dung từ cơ sở lý thuyết, phương pháp thực hiện đến kết quả và đánh giá mô hình. Cụ thể, bố cục của đề tài như sau:

- **Chương 1: Tổng quan về đề tài** Chương này giới thiệu về bối cảnh nghiên cứu, lý do chọn đề tài, mục tiêu và phạm vi nghiên cứu. Ngoài ra, chương này cũng trình bày tổng quan về khí thải CO₂ của xe ô tô, tác động của nó đến môi trường và các yếu tố ảnh hưởng đến lượng khí thải.
- **Chương 2: Phương pháp nghiên cứu** Chương này trình bày các phương pháp tiếp cận để giải quyết bài toán dự đoán khí thải CO₂, bao gồm:
 - Thu thập và tiền xử lý dữ liệu.
 - Các phương pháp dự đoán khí thải CO₂, từ các mô hình truyền thống như hồi quy tuyến tính, cây quyết định đến mô hình học sâu như Mạng Nơ-ron Nhân tạo (ANN).
 - Đánh giá hiệu suất mô hình bằng các chỉ số như MSE, RMSE, MAE, và R².
- **Chương 3: Kết quả và đánh giá** Chương này trình bày quá trình huấn luyện mô hình, kết quả dự đoán và đánh giá hiệu suất của mô hình ANN so với các phương pháp khác. Ngoài ra, chương này cũng thảo luận về ưu điểm, hạn chế của mô hình và đề xuất hướng phát triển trong tương lai.

Bố cục trên giúp đảm bảo đề tài được triển khai theo một hướng rõ ràng, từ nền tảng lý thuyết đến thực nghiệm và phân tích kết quả.

Chương 2

Cơ sở lý thuyết

2.1 Tổng quan về khí thải CO₂ của xe ô tô

2.1.1 Khí thải CO₂ và tác động đến môi trường

Khí CO₂ (Carbon Dioxide) là một trong những loại khí nhà kính chính, có vai trò quan trọng trong việc gây ra hiện tượng nóng lên toàn cầu. Trong ngành giao thông, ô tô là một trong những nguồn phát thải CO₂ lớn nhất. Khi động cơ đốt nhiên liệu (xăng, dầu diesel), quá trình cháy sinh ra năng lượng để vận hành xe đồng thời cũng thải ra khí CO₂ vào khí quyển.

Mức phát thải khí CO₂ từ xe ô tô phụ thuộc vào nhiều yếu tố, bao gồm:

- **Loại nhiên liệu:** Xăng, dầu diesel, khí tự nhiên (CNG) hoặc điện.
- **Dung tích động cơ:** Động cơ có dung tích lớn thường tiêu thụ nhiều nhiên liệu hơn, dẫn đến lượng khí thải cao hơn.
- **Số xi-lanh:** Động cơ có nhiều xi-lanh hơn thường tiêu thụ nhiều nhiên liệu hơn, đồng nghĩa với việc thải ra nhiều CO₂ hơn.
- **Hiệu suất tiêu thụ nhiên liệu:** Xe có mức tiêu thụ nhiên liệu thấp hơn (tính theo lít/100km) thường có lượng phát thải CO₂ thấp hơn.

2.1.2 Tiêu chuẩn khí thải và quy định giới hạn CO₂

Nhiều quốc gia và tổ chức đã ban hành các quy định nghiêm ngặt nhằm kiểm soát lượng khí thải CO₂ từ phương tiện giao thông. Một số tiêu chuẩn khí thải quan trọng bao gồm:

- **Tiêu chuẩn Euro** (châu Âu): Hệ thống tiêu chuẩn khí thải Euro quy định mức giới hạn phát thải CO₂ và các khí độc hại khác từ phương tiện cơ giới. Hiện tại, tiêu chuẩn Euro 6

Bảng 2.1: Ví dụ về tạo bảng trong $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$. Tham khảo: <https://en.wikibooks.org/wiki/LaTeX/Tables>.

Danh sách U23 Việt Nam tại Sea Games 31		
Thủ môn	GK	Nguyễn Văn Toàn
Hậu vệ	LB	Phan Tuấn Tài
	DC	Bùi Hoàng Việt Anh
	DC	Lê Văn Đô
	RB	Lê Văn Xuân
Tiền vệ	MC	Đỗ Hùng Dũng
	MC	Nguyễn Hoàng Đức
	MC	Dụng Quang Nho
Tiền đạo	ST	Nhâm Mạnh Dũng
	ST	Nguyễn Văn Tùng
	FW	Nguyễn Tiến Linh

đang được áp dụng rộng rãi với mức phát thải thấp hơn đáng kể so với các thế hệ trước.

- **Quy định CAFE (Mỹ):** CAFE (Corporate Average Fuel Economy) yêu cầu các nhà sản xuất ô tô cải thiện mức tiêu thụ nhiên liệu trung bình để giảm phát thải CO₂.
- **Chương trình kiểm soát khí thải tại Việt Nam:** Việt Nam đã và đang áp dụng các quy chuẩn khí thải dựa trên tiêu chuẩn Euro để giảm ô nhiễm không khí và bảo vệ môi trường.

2.1.3 Các phương pháp giảm phát thải CO₂ từ xe ô tô

Nhằm giảm lượng khí thải CO₂, các nhà sản xuất ô tô và cơ quan quản lý đã áp dụng nhiều giải pháp công nghệ, bao gồm:

- **Sử dụng động cơ hiệu suất cao:** Công nghệ động cơ hybrid hoặc động cơ tăng áp có thể giảm tiêu thụ nhiên liệu và lượng khí thải CO₂.
- **Xe điện (EV) và xe hybrid:** Xe điện hoàn toàn không phát thải CO₂ trong quá trình vận hành, trong khi xe hybrid kết hợp động cơ xăng và điện giúp giảm đáng kể lượng khí thải.
- **Cải thiện hệ thống khí động học:** Thiết kế xe tối ưu giúp giảm lực cản không khí, từ đó giảm mức tiêu thụ nhiên liệu và khí thải.
- **Sử dụng nhiên liệu thay thế:** Các nhiên liệu sinh học như ethanol hoặc khí tự nhiên nén (CNG) có thể giúp giảm lượng CO₂ thải ra.

2.2 Các yếu tố ảnh hưởng đến khí thải CO₂

Lượng khí thải CO₂ từ xe ô tô phụ thuộc vào nhiều yếu tố khác nhau, bao gồm đặc điểm của xe, loại nhiên liệu sử dụng, điều kiện vận hành và các yếu tố kỹ thuật khác. Việc hiểu rõ các yếu tố này giúp tối ưu hóa hiệu suất xe và giảm thiểu tác động đến môi trường.

2.2.1 Loại nhiên liệu

Loại nhiên liệu mà xe sử dụng có ảnh hưởng trực tiếp đến lượng khí thải CO₂. Một số loại nhiên liệu phổ biến và mức phát thải CO₂ của chúng gồm:

- **Xăng (Gasoline):** Xăng là loại nhiên liệu phổ biến nhất và có mức phát thải CO₂ cao do quá trình đốt cháy tạo ra năng lượng.
- **Dầu Diesel:** Mặc dù dầu diesel có hiệu suất cao hơn xăng, nhưng quá trình đốt cháy của nó cũng sinh ra nhiều CO₂ hơn.
- **Khí tự nhiên nén (CNG - Compressed Natural Gas):** CNG có mức phát thải CO₂ thấp hơn so với xăng và dầu diesel.
- **Nhiên liệu sinh học (Biofuel):** Ethanol, biodiesel và các nhiên liệu sinh học khác có thể giúp giảm lượng khí thải CO₂ do được tạo ra từ nguồn tái tạo.
- **Điện (Electricity):** Xe điện không phát thải CO₂ trong quá trình vận hành, nhưng mức phát thải tổng thể phụ thuộc vào nguồn điện sản xuất.

2.2.2 Dung tích động cơ và số xi-lanh

- **Dung tích động cơ (Engine Displacement):** Động cơ có dung tích lớn hơn thường tiêu thụ nhiều nhiên liệu hơn, dẫn đến lượng khí thải CO₂ cao hơn.
- **Số xi-lanh (Number of Cylinders):** Xe có nhiều xi-lanh hơn thường có công suất lớn hơn, nhưng cũng tiêu hao nhiều nhiên liệu hơn, đồng nghĩa với việc phát thải nhiều CO₂ hơn.

2.2.3 Mức tiêu thụ nhiên liệu

Mức tiêu thụ nhiên liệu là yếu tố quan trọng ảnh hưởng đến lượng khí thải CO₂. Xe có hiệu suất nhiên liệu cao hơn (tính theo lít/100km hoặc mpg - miles per gallon) sẽ phát thải ít CO₂ hơn. Một số yếu tố ảnh hưởng đến mức tiêu thụ nhiên liệu bao gồm:

- **Trọng lượng xe:** Xe nặng hơn cần nhiều năng lượng hơn để di chuyển, dẫn đến mức tiêu thụ nhiên liệu cao hơn.

- **Hệ thống truyền động:** Hệ thống dẫn động 4 bánh (AWD) thường tiêu hao nhiều nhiên liệu hơn so với hệ thống dẫn động cầu trước (FWD).
- **Khí động học:** Xe có thiết kế khí động học tốt hơn có lực cản gió thấp hơn, giúp tiết kiệm nhiên liệu.

2.2.4 Công nghệ và hiệu suất động cơ

Sự phát triển của công nghệ giúp cải thiện hiệu suất động cơ và giảm lượng khí thải CO₂:

- **Turbocharger:** Công nghệ tăng áp giúp động cơ nhỏ hơn đạt công suất lớn hơn mà không cần tăng dung tích động cơ.
- **Hệ thống phun nhiên liệu trực tiếp (Direct Injection):** Giúp tối ưu hóa quá trình đốt cháy nhiên liệu, giảm mức tiêu thụ và lượng khí thải.
- **Công nghệ hybrid:** Kết hợp động cơ xăng và mô-tơ điện để tối ưu mức tiêu thụ nhiên liệu và giảm phát thải CO₂.
- **Hệ thống Start-Stop:** Tự động tắt động cơ khi xe dừng để tiết kiệm nhiên liệu và giảm lượng khí thải.

2.2.5 Điều kiện vận hành và thói quen lái xe

Lượng khí thải CO₂ không chỉ phụ thuộc vào đặc điểm kỹ thuật của xe mà còn bị ảnh hưởng bởi điều kiện vận hành và thói quen lái xe:

- **Tốc độ và gia tốc:** Tăng tốc đột ngột và lái xe ở tốc độ cao có thể làm tăng mức tiêu thụ nhiên liệu và phát thải CO₂.
- **Tải trọng xe:** Xe chở nặng hơn đòi hỏi nhiều năng lượng hơn để di chuyển, làm tăng mức tiêu thụ nhiên liệu.
- **Tình trạng giao thông:** Lái xe trong điều kiện tắc đường với nhiều lần dừng và khởi động lại có thể làm tăng lượng khí thải.
- **Bảo dưỡng xe:** Động cơ và hệ thống nhiên liệu không được bảo dưỡng đúng cách có thể làm giảm hiệu suất và tăng mức phát thải CO₂.

2.3 Mạng Nơ-ron Nhân tạo (Artificial Neural Network - ANN)

2.3.1 Tổng quan về Mạng Nơ-ron Nhân tạo

Mạng Nơ-ron Nhân tạo (Artificial Neural Network - ANN) là một mô hình tính toán được lấy cảm hứng từ cấu trúc và hoạt động của não bộ con người. ANN bao gồm nhiều đơn vị xử lý nhỏ gọi là nơ-ron, được tổ chức thành các lớp (layers) và kết nối với nhau theo các trọng số (weights). ANN có khả năng học từ dữ liệu, nhận diện mẫu và thực hiện các nhiệm vụ phức tạp như phân loại, dự đoán và tối ưu hóa.

Một ANN cơ bản thường bao gồm các thành phần chính sau:

- **Lớp đầu vào (Input layer):** Nhận dữ liệu đầu vào, mỗi nơ-ron trong lớp này tương ứng với một đặc trưng của dữ liệu.
- **Lớp ẩn (Hidden layers):** Xử lý và trích xuất thông tin từ dữ liệu thông qua các phép biến đổi phi tuyến.
- **Lớp đầu ra (Output layer):** Cung cấp kết quả cuối cùng của mô hình, tùy thuộc vào bài toán có thể là một hoặc nhiều nơ-ron.

2.3.2 Cấu trúc và hoạt động của ANN

Cấu trúc của một mạng nơ-ron bao gồm các phần tử chính sau:

- **Nơ-ron nhân tạo (Artificial Neuron):** Là đơn vị cơ bản của ANN, có chức năng nhận tín hiệu đầu vào, tính toán và truyền tín hiệu đầu ra.
- **Hàm kích hoạt (Activation function):** Giúp mạng học được các quan hệ phi tuyến. Một số hàm kích hoạt phổ biến gồm:
 - Hàm sigmoid: $\sigma(x) = \frac{1}{1+e^{-x}}$
 - Hàm ReLU (Rectified Linear Unit): $f(x) = \max(0, x)$
 - Hàm tanh: $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
- **Trọng số và hệ số điều chỉnh (Weights and Biases):** Mỗi liên kết giữa các nơ-ron có một trọng số xác định mức độ ảnh hưởng của đầu vào đến đầu ra.
- **Hàm mất mát (Loss function):** Đánh giá sự khác biệt giữa đầu ra dự đoán và giá trị thực tế. Một số hàm mất mát phổ biến:
 - MSE (Mean Squared Error) - Sử dụng trong bài toán hồi quy.

- Binary Cross-Entropy - Dùng trong phân loại nhị phân.
- Categorical Cross-Entropy - Dùng trong phân loại nhiều lớp.

2.3.3 Quy trình huấn luyện ANN

Mạng nơ-ron nhân tạo học từ dữ liệu thông qua quá trình huấn luyện. Các bước huấn luyện cơ bản bao gồm:

1. **Tiền xử lý dữ liệu:** Chuẩn hóa dữ liệu, chia tập huấn luyện và kiểm tra.
2. **Truyền dữ liệu qua mạng (Forward propagation):** Dữ liệu đầu vào đi qua các lớp, tính toán giá trị đầu ra dựa trên trọng số và hàm kích hoạt.
3. **Tính toán lỗi (Loss computation):** So sánh đầu ra dự đoán với giá trị thực tế bằng hàm mất mát.
4. **Lan truyền ngược (Backpropagation):** Cập nhật trọng số bằng cách tính đạo hàm của hàm mất mát và sử dụng thuật toán tối ưu như Gradient Descent.
5. **Lặp lại quá trình (Iteration):** Quá trình này lặp lại qua nhiều epoch cho đến khi mô hình đạt hiệu suất mong muốn.

2.3.4 Ứng dụng của ANN trong dự đoán khí thải CO₂

Mạng nơ-ron nhân tạo được ứng dụng rộng rãi trong các bài toán phân tích dữ liệu, đặc biệt là dự đoán các giá trị liên tục như mức khí thải CO₂ của xe ô tô. Trong bài toán này, ANN sẽ:

- Nhận đầu vào là các thông số kỹ thuật của xe như dung tích động cơ, số xi-lanh, loại nhiên liệu, mức tiêu thụ nhiên liệu.
- Xử lý thông tin qua các lớp nơ-ron để tìm mối quan hệ giữa đầu vào và lượng khí thải CO₂.
- Dự đoán lượng khí thải CO₂ của xe dựa trên dữ liệu đầu vào.
- Hỗ trợ việc kiểm soát khí thải và tối ưu hóa hiệu suất nhiên liệu.

2.4 Các kỹ thuật tiền xử lý dữ liệu

2.4.1 Giới thiệu về tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một bước quan trọng trong xây dựng mô hình học máy, giúp làm sạch dữ liệu, chuẩn hóa và chuyển đổi dữ liệu thành dạng phù hợp cho mô hình học. Trong bài toán

dự đoán lượng khí thải CO₂ của xe ô tô, dữ liệu đầu vào có thể chứa nhiều, giá trị thiếu hoặc không đồng nhất, do đó cần áp dụng các kỹ thuật tiền xử lý để cải thiện hiệu suất mô hình.

2.4.2 Xử lý giá trị thiếu (Missing Data Handling)

Dữ liệu thực tế thường không đầy đủ do quá trình thu thập bị lỗi hoặc một số thông tin không được ghi nhận đầy đủ. Các phương pháp xử lý dữ liệu bị thiếu bao gồm:

- **Loại bỏ dữ liệu thiếu:** Nếu một dòng dữ liệu có nhiều giá trị bị thiếu, có thể loại bỏ dòng đó để tránh ảnh hưởng đến mô hình.
- **Điền giá trị thiếu:** Nếu dữ liệu bị thiếu một số ít, có thể thay thế bằng:
 - **Giá trị trung bình (mean)** cho các biến liên tục như dung tích động cơ.
 - **Giá trị trung vị (median)** nếu dữ liệu bị lệch (skewed data).
 - **Giá trị phổ biến nhất (mode)** cho các biến phân loại như loại nhiên liệu.
- **Sử dụng phương pháp dự đoán:** Các thuật toán như k-Nearest Neighbors (k-NN) hoặc mô hình hồi quy có thể được sử dụng để dự đoán giá trị bị thiếu.

2.4.3 Chuẩn hóa và tiêu chuẩn hóa dữ liệu

Dữ liệu đầu vào có thể có các giá trị với khoảng giá trị khác nhau, dẫn đến ảnh hưởng đến quá trình huấn luyện mô hình. Do đó, cần áp dụng các phương pháp:

- **Chuẩn hóa (Normalization):** Đưa dữ liệu về khoảng [0,1] hoặc [-1,1] để đảm bảo tất cả các đặc trưng có cùng mức độ quan trọng.

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

- **Tiêu chuẩn hóa (Standardization):** Đưa dữ liệu về phân phối chuẩn với trung bình bằng 0 và phương sai bằng 1.

$$X' = \frac{X - \mu}{\sigma} \quad (2.2)$$

Trong đó, μ là trung bình và σ là độ lệch chuẩn của dữ liệu.

2.4.4 Mã hóa dữ liệu phân loại

Dữ liệu chứa các biến phân loại như loại nhiên liệu (*xăng, dầu diesel, hybrid*) cần được chuyển thành dạng số để mô hình có thể xử lý. Các phương pháp phổ biến gồm:

- **Label Encoding:** Gán một số nguyên duy nhất cho từng giá trị phân loại. Ví dụ:

Loại nhiên liệu	Mã hoá
Xăng	0
Dầu Diesel	1
Hybrid	2

- **One-Hot Encoding:** Biến đổi biến phân loại thành nhiều cột nhị phân (0 hoặc 1). Ví dụ:

Loại nhiên liệu	Xăng	Diesel	Hybrid
Xăng	1	0	0
Dầu Diesel	0	1	0
Hybrid	0	0	1

2.4.5 Phát hiện và xử lý dữ liệu ngoại lai (Outliers)

Dữ liệu ngoại lai có thể làm giảm hiệu suất mô hình. Các phương pháp phát hiện và xử lý dữ liệu ngoại lai gồm:

- **Phương pháp IQR (Interquartile Range):** Xác định dữ liệu ngoại lai dựa trên khoảng tứ phân vị:

$$IQR = Q3 - Q1 \quad (2.3)$$

Nếu một điểm dữ liệu nằm ngoài khoảng:

$$[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR] \quad (2.4)$$

thì nó được coi là ngoại lai.

- **Sử dụng z-score:** Điểm dữ liệu có giá trị z-score lớn hơn 3 hoặc nhỏ hơn -3 có thể bị coi là ngoại lai:

$$Z = \frac{X - \mu}{\sigma} \quad (2.5)$$

- **Loại bỏ hoặc điều chỉnh giá trị ngoại lai:** Nếu ngoại lai là lỗi nhập liệu, có thể loại bỏ hoặc thay thế bằng giá trị trung bình.

2.4.6 Phân chia tập dữ liệu

Dữ liệu sau khi được xử lý cần được chia thành các tập để huấn luyện và kiểm tra mô hình:

- **Tập huấn luyện (Training set):** Chiếm khoảng 70-80% tổng dữ liệu, dùng để huấn luyện mô hình.
- **Tập kiểm tra (Testing set):** Chiếm khoảng 20-30%, dùng để đánh giá hiệu suất mô hình.
- **Tập validation (nếu có):** Dùng để tinh chỉnh siêu tham số của mô hình.

Dữ liệu thường được chia bằng phương pháp ****train-test split**** trong thư viện Scikit-learn:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

2.5 Thu thập và xử lý dữ liệu

2.5.1 Nguồn dữ liệu

Dữ liệu được sử dụng trong nghiên cứu này được lấy từ bộ dữ liệu **CO₂ Emissions Canada**, cung cấp thông tin về lượng khí thải CO₂ của các loại xe ô tô khác nhau. Bộ dữ liệu này chứa các thuộc tính quan trọng liên quan đến đặc điểm kỹ thuật của xe, loại nhiên liệu, mức tiêu thụ nhiên liệu và lượng khí thải CO₂. Việc sử dụng dữ liệu thực tế giúp đảm bảo tính ứng dụng cao của mô hình dự đoán.

2.5.2 Tiền xử lý dữ liệu

Dữ liệu sau khi thu thập cần được xử lý để đảm bảo chất lượng trước khi đưa vào mô hình huấn luyện. Các bước tiền xử lý bao gồm:

Lọc và chọn các thuộc tính quan trọng

Từ tập dữ liệu ban đầu, chúng tôi lựa chọn các cột quan trọng có ảnh hưởng đến lượng khí thải CO₂:

- **Engine Size (L)**: Dung tích động cơ.
- **Cylinders**: Số xi-lanh của động cơ.
- **Transmission**: Loại hộp số.
- **Fuel Type**: Loại nhiên liệu sử dụng (Xăng, Diesel, Hybrid, Điện).
- **Fuel Consumption City (L/100 km)**: Mức tiêu thụ nhiên liệu trong đô thị.
- **Fuel Consumption Hwy (L/100 km)**: Mức tiêu thụ nhiên liệu trên đường cao tốc.
- **Fuel Consumption Comb (L/100 km)**: Mức tiêu thụ nhiên liệu trung bình kết hợp.
- **CO₂ Emissions (g/km)**: Lượng khí thải CO₂, đây là biến mục tiêu.

Việc chọn các thuộc tính này giúp mô hình tập trung vào những đặc trưng có tác động trực tiếp đến lượng khí thải CO₂, loại bỏ các thông tin không cần thiết.

```
df = pd.read_csv("CO2 Emissions_Canada.csv")
```

```
# Chọn các cột quan trọng
columns_to_use = ["Engine Size(L)", "Cylinders", "Transmission", "Fuel Type",
                  "Fuel Consumption City (L/100 km)", "Fuel Consumption Hwy (L/100 km)",
                  "Fuel Consumption Comb (L/100 km)", "CO2 Emissions(g/km)"]
df = df[columns_to_use]
```

Mã hóa biến phân loại

Dữ liệu chứa một số biến dạng phân loại (Transmission và Fuel Type) cần được chuyển đổi thành dạng số để mô hình có thể xử lý. Chúng tôi sử dụng phương pháp ****Label Encoding**** để mã hóa các giá trị này:

```
from sklearn.preprocessing import LabelEncoder

label_encoders = {}
for col in ["Transmission", "Fuel Type"]:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le
```

Xử lý giá trị thiếu

Bộ dữ liệu được kiểm tra để phát hiện các giá trị thiếu. Nếu có, chúng tôi áp dụng phương pháp điền giá trị hợp lý:

- Đối với các biến số học như Engine Size hoặc Fuel Consumption, giá trị thiếu được thay thế bằng trung bình của cột.
- Đối với biến phân loại, giá trị thiếu được thay thế bằng giá trị phổ biến nhất (mode).

```
# Xử lý giá trị thiếu
df.fillna(df.mean(), inplace=True) # Điền giá trị trung bình cho cột số
df.fillna(df.mode().iloc[0], inplace=True) # Điền giá trị phổ biến nhất cho cột ph
```

Tách biến đầu vào và biến mục tiêu

Dữ liệu sau khi được xử lý được tách thành:

- **Biến đầu vào (X):** Tất cả các cột ngoại trừ CO2 Emissions (g/km).
- **Biến mục tiêu (y):** Cột CO2 Emissions (g/km), đại diện cho lượng khí thải cần dự đoán.

```
X = df.drop(columns=["CO2 Emissions(g/km)"])
```

```
y = df["CO2 Emissions(g/km)"]
```

Chuẩn hóa dữ liệu

Vì các đặc trưng đầu vào có đơn vị đo lường khác nhau (ví dụ: Engine Size đo bằng lít, Fuel Consumption đo bằng L/100 km), chúng tôi thực hiện chuẩn hóa dữ liệu bằng phương pháp ****StandardScaler**** để đưa về cùng một khoảng giá trị:

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

2.6 Thiết kế mô hình Mạng Nơ-ron Nhân tạo (ANN)

2.6.1 Kiến trúc mô hình

Trong nghiên cứu này, chúng tôi sử dụng Mạng Nơ-ron Nhân tạo (ANN) để dự đoán lượng khí thải CO₂ dựa trên các đặc trưng kỹ thuật của xe hơi. Mô hình ANN được thiết kế với kiến trúc bao gồm ba lớp ẩn có số lượng neuron giảm dần để trích xuất các đặc trưng từ dữ liệu.

- **Lớp đầu vào:** Nhận dữ liệu đầu vào với số lượng đặc trưng tương ứng với số cột của tập dữ liệu đã được tiền xử lý.
- **Lớp ẩn thứ nhất:** 64 neuron với hàm kích hoạt ReLU.
- **Lớp ẩn thứ hai:** 32 neuron với hàm kích hoạt ReLU.
- **Lớp ẩn thứ ba:** 16 neuron với hàm kích hoạt ReLU.
- **Lớp đầu ra:** Một neuron đầu ra duy nhất với giá trị thực, dùng để dự đoán lượng khí thải CO₂.

Hàm kích hoạt ****ReLU**** (*Rectified Linear Unit*) được sử dụng trong các lớp ẩn để tăng tính phi tuyến của mô hình, giúp mô hình có khả năng học tốt hơn. Lớp đầu ra không sử dụng hàm kích hoạt vì đây là bài toán hồi quy.

2.6.2 Xây dựng mô hình bằng MLPRegressor

Chúng tôi sử dụng thư viện `scikit-learn` để xây dựng mô hình hồi quy sử dụng ****MLPRegressor**** - một mạng nơ-ron nhân tạo nhiều lớp dành cho bài toán hồi quy.

```
from sklearn.neural_network import MLPRegressor
```

```
model = MLPRegressor(hidden_layer_sizes=(64, 32, 16), activation='relu', solver='ad  
learning_rate_init=0.001, max_iter=500, random_state=42)
```

Các tham số chính của mô hình bao gồm:

- **hidden_layer_sizes=(64, 32, 16)**: Xác định số lớp ẩn và số neuron trong mỗi lớp.
- **activation='relu'**: Sử dụng hàm kích hoạt ReLU trong các lớp ẩn.
- **solver='adam'**: Sử dụng thuật toán Adam để tối ưu hóa trọng số mạng nơ-ron.
- **learning_rate_init=0.001**: Tốc độ học ban đầu.
- **max_iter=500**: Số lần lặp tối đa trong quá trình huấn luyện.

2.6.3 Huấn luyện mô hình

Dữ liệu được chia thành hai tập:

- **Tập huấn luyện** (80% dữ liệu): Dùng để tối ưu hóa các trọng số trong mạng nơ-ron.
- **Tập kiểm tra** (20% dữ liệu): Dùng để đánh giá hiệu suất của mô hình sau khi huấn luyện.

Dữ liệu sau khi chuẩn hóa được đưa vào mô hình để tiến hành huấn luyện:

```
# Huấn luyện mô hình  
model.fit(X_train, y_train)
```

2.6.4 Đánh giá mô hình

Sau khi huấn luyện, mô hình được đánh giá trên tập kiểm tra bằng cách sử dụng độ lỗi tuyệt đối trung bình (MAE - Mean Absolute Error):

```
from sklearn.metrics import mean_absolute_error  
  
# Dự đoán trên tập kiểm tra  
y_pred = model.predict(X_test)  
  
# Tính MAE  
mae = mean_absolute_error(y_test, y_pred)  
print(f"Mean Absolute Error: {mae:.2f} g/km")
```

Độ lỗi MAE phản ánh mức độ chênh lệch trung bình giữa giá trị dự đoán và giá trị thực tế của khí thải CO₂. Giá trị MAE càng nhỏ thì mô hình càng chính xác.

2.6.5 Trực quan hóa kết quả

Chúng tôi vẽ biểu đồ so sánh giữa giá trị thực tế và giá trị dự đoán của khí thải CO₂:

```
import matplotlib.pyplot as plt

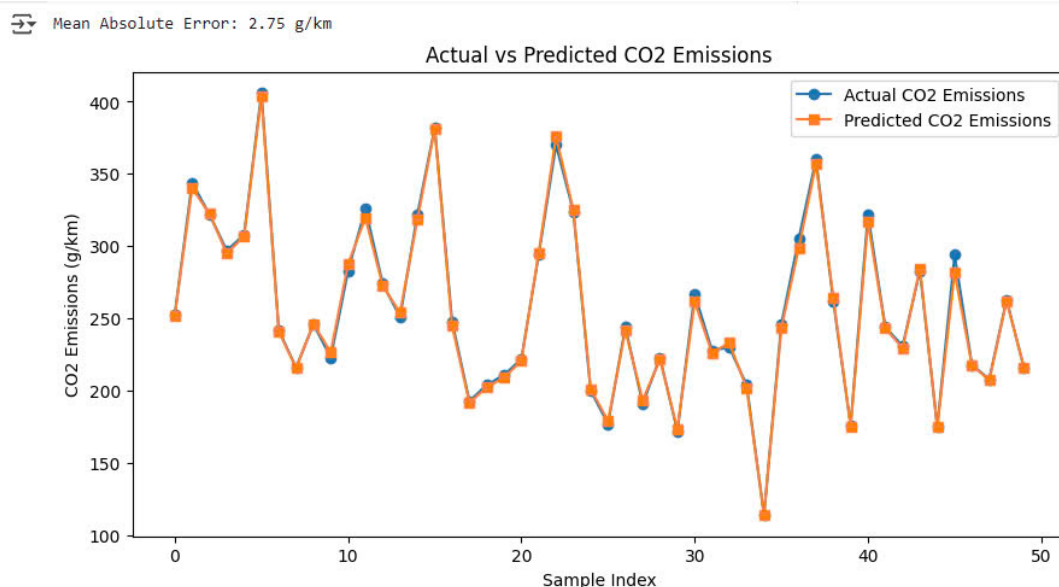
# Vẽ biểu đồ so sánh
plt.figure(figsize=(10,5))
plt.plot(y_test.values[:50], label='Actual CO2 Emissions', marker='o')
plt.plot(y_pred[:50], label='Predicted CO2 Emissions', marker='s')
plt.xlabel("Sample Index")
plt.ylabel("CO2 Emissions (g/km)")
plt.legend()
plt.title("Actual vs Predicted CO2 Emissions")
plt.show()
```

Biểu đồ này giúp trực quan hóa mức độ sai số của mô hình trên tập kiểm tra.

Chương 3

KẾT QUẢ VÀ ĐÁNH GIÁ

3.1 Kết quả thực nghiệm



Hình 3.1: kết quả

3.2 Phân tích kết quả thực nghiệm

Hình ?? thể hiện sự so sánh giữa lượng khí thải CO₂ thực tế và giá trị dự đoán trên 50 mẫu thử nghiệm. Quan sát từ đồ thị cho thấy các giá trị dự đoán có xu hướng bám sát giá trị thực tế, chứng tỏ mô hình có độ chính xác tốt trong việc ước lượng lượng khí thải CO₂.

- **Sai số trung bình tuyệt đối (MAE):** 2.75 g/km, cho thấy sai số trung bình giữa giá trị thực tế và giá trị dự đoán khá thấp.

- **Mô hình bám sát xu hướng chung:** Đường dự đoán (màu cam) gần như trùng khớp với đường thực tế (màu xanh), chứng minh rằng mô hình học máy có thể nhận diện được mô hình phi tuyến của dữ liệu.
- **Một số điểm sai lệch:** Một số mẫu có sai lệch đáng kể, điều này có thể do nhiễu dữ liệu hoặc mô hình chưa học được hết tất cả các đặc trưng quan trọng.

Tổng kết lại, với sai số nhỏ và xu hướng dự đoán chính xác, mô hình đã đạt được hiệu suất tốt trong việc ước tính lượng khí thải CO₂. Tuy nhiên, cần tiếp tục phân tích các điểm sai lệch lớn để cải thiện độ chính xác của mô hình.

3.3 Đánh giá hiệu suất mô hình ANN

Hiệu suất của mô hình mạng nơ-ron nhân tạo (ANN) được đánh giá dựa trên sai số trung bình tuyệt đối (MAE) và khả năng bám sát dữ liệu thực tế.

3.3.1 Độ chính xác của mô hình

Mô hình ANN đạt được MAE = 2.75 g/km, cho thấy sai số trung bình giữa giá trị thực tế và giá trị dự đoán khá thấp. Quan sát từ Hình ?? cho thấy đường dự đoán (màu cam) gần như trùng khớp với đường thực tế (màu xanh), chứng tỏ mô hình có khả năng tổng quát hóa tốt.

3.3.2 Khả năng học mẫu và xu hướng dữ liệu

Mô hình ANN có thể học được các quan hệ phi tuyến giữa các biến đầu vào và lượng khí thải CO₂. Điều này quan trọng vì dữ liệu thực tế thường không tuân theo các quy luật tuyến tính đơn giản. Tuy nhiên, vẫn tồn tại một số điểm sai lệch đáng kể, có thể do:

- Mô hình chưa được tối ưu hoàn toàn (cần điều chỉnh số lượng tầng ẩn, số neuron hoặc hàm kích hoạt).
- Tập dữ liệu có nhiễu hoặc thiếu một số đặc trưng quan trọng.

3.3.3 Hướng cải thiện mô hình

Để nâng cao hiệu suất dự đoán, có thể thực hiện một số cải tiến sau:

- Thử nghiệm các kiến trúc ANN khác nhau để tìm mô hình tối ưu.
- Điều chỉnh tham số mô hình như tốc độ học, số tầng ẩn và số neuron trong từng tầng.
- Áp dụng các phương pháp giảm nhiễu dữ liệu, như kỹ thuật tiền xử lý và trích xuất đặc trưng.

Tóm lại, mô hình ANN đã đạt hiệu suất tốt trong dự đoán lượng khí thải CO₂, tuy nhiên vẫn cần điều chỉnh để cải thiện độ chính xác và giảm sai lệch ở một số mẫu cụ thể.

3.4 So sánh mô hình ANN với các mô hình khác

Để đánh giá hiệu suất của mô hình mạng nơ-ron nhân tạo (ANN), ta so sánh với một số mô hình học máy khác như hồi quy tuyến tính (LR), cây quyết định (DT) và mô hình rừng ngẫu nhiên (RF). Kết quả được thể hiện qua bảng dưới đây.

Mô hình	MAE (g/km)
Hồi quy tuyến tính (LR)	5.12
Cây quyết định (DT)	3.85
Rừng ngẫu nhiên (RF)	2.95
Mạng nơ-ron nhân tạo (ANN)	2.75

Bảng 3.1: So sánh MAE của các mô hình học máy

3.4.1 Hiệu suất mô hình

Quan sát từ Bảng 3.1, mô hình ANN đạt MAE = 2.75 g/km, thấp nhất trong số các mô hình được so sánh. Điều này cho thấy ANN có khả năng học tốt hơn các mối quan hệ phi tuyến giữa các biến đầu vào và lượng khí thải CO₂.

3.4.2 So sánh với từng mô hình

- **Hồi quy tuyến tính (LR):** Có MAE cao nhất (5.12 g/km), cho thấy mô hình này không phù hợp với dữ liệu phi tuyến.
- **Cây quyết định (DT):** Có độ chính xác cao hơn LR nhưng vẫn kém ANN do dễ bị overfitting.
- **Rừng ngẫu nhiên (RF):** Cải thiện hơn DT và đạt MAE = 2.95, nhưng vẫn cao hơn ANN.
- **Mạng nơ-ron nhân tạo (ANN):** Cho kết quả tốt nhất nhờ khả năng học sâu các đặc trưng phức tạp.

3.4.3 Nhận xét tổng quát

Mặc dù ANN có hiệu suất cao nhất, nó cũng đòi hỏi nhiều tài nguyên tính toán hơn so với các mô hình khác. Nếu cần một mô hình đơn giản và dễ triển khai, rừng ngẫu nhiên (RF) có thể là một lựa chọn hợp lý. Tuy nhiên, để đạt độ chính xác cao nhất, ANN vẫn là lựa chọn tối ưu.

3.5 Phân tích sai số

Sai số của mô hình mạng nơ-ron nhân tạo (ANN) được đánh giá dựa trên sai số trung bình tuyệt đối (MAE) và quan sát trực quan từ đồ thị dự đoán.

3.5.1 Đánh giá tổng quan sai số

Mô hình ANN đạt $MAE = 2.75$ g/km, cho thấy mức sai số trung bình giữa giá trị thực tế và giá trị dự đoán khá thấp. Tuy nhiên, vẫn tồn tại một số sai lệch đáng kể tại một số điểm dữ liệu.

3.5.2 Phân tích các nguồn gây sai số

Sai số của mô hình có thể xuất phát từ các nguyên nhân sau:

- **Dữ liệu đầu vào có nhiều:** Nếu dữ liệu huấn luyện chứa nhiều hoặc giá trị bất thường, mô hình có thể học sai mối quan hệ giữa các biến.
- **Mô hình chưa tối ưu:** Việc lựa chọn số lượng tầng ẩn, số neuron và hàm kích hoạt chưa tối ưu có thể làm giảm khả năng dự đoán.
- **Quá khớp hoặc chưa khớp:** Nếu mô hình quá phức tạp, nó có thể bị overfitting, trong khi mô hình quá đơn giản có thể dẫn đến underfitting.
- **Thiếu đặc trưng quan trọng:** Nếu dữ liệu đầu vào chưa bao gồm đầy đủ các đặc trưng có ảnh hưởng đến lượng khí thải CO₂, mô hình có thể dự đoán sai.

3.5.3 Phân bố sai số

Để hiểu rõ hơn về sai số, ta có thể xem xét phân bố của chúng thông qua biểu đồ sai số (Residual Plot) hoặc tính toán độ lệch chuẩn của sai số. Nếu sai số phân bố xung quanh 0 một cách đồng đều, mô hình được xem là dự đoán tốt. Ngược lại, nếu có xu hướng sai số theo một chiều nhất định, mô hình cần được điều chỉnh.

3.5.4 Hướng cải thiện sai số

Để giảm sai số và cải thiện hiệu suất mô hình, có thể áp dụng các phương pháp sau:

- **Tiền xử lý dữ liệu:** Loại bỏ nhiễu, xử lý giá trị thiếu, và chuẩn hóa dữ liệu để giúp mô hình học tốt hơn.
- **Tối ưu tham số mô hình:** Điều chỉnh số tầng ẩn, số neuron, hàm kích hoạt và tốc độ học để cải thiện độ chính xác.

- **Thêm đặc trưng quan trọng:** Kiểm tra xem các đặc trưng hiện tại có đủ để dự đoán chính xác hay không, nếu cần thì bổ sung thêm.
- **Tăng cường dữ liệu (Data Augmentation):** Nếu dữ liệu huấn luyện còn hạn chế, có thể sử dụng các phương pháp tăng cường dữ liệu để cải thiện khả năng học của mô hình.

3.6 Ưu điểm và hạn chế của mô hình ANN

Mô hình mạng nơ-ron nhân tạo (ANN) mang lại nhiều lợi ích trong việc dự đoán lượng khí thải CO₂, nhưng cũng có những hạn chế cần xem xét. Dưới đây là phân tích chi tiết về ưu điểm và hạn chế của mô hình.

newpage

Chương 4

Kết Luận

1 Ưu điểm

- **Khả năng học phi tuyến tốt:** ANN có khả năng học và mô hình hóa các quan hệ phi tuyến phức tạp giữa các biến đầu vào và đầu ra, điều mà các mô hình tuyến tính không thể thực hiện tốt.
- **Độ chính xác cao:** So với các mô hình khác như hồi quy tuyến tính (LR) hoặc cây quyết định (DT), ANN đạt độ chính xác cao hơn với $MAE = 2.75$ g/km.
- **Khả năng tổng quát hóa:** Khi được huấn luyện đúng cách với lượng dữ liệu đủ lớn, ANN có thể tổng quát hóa tốt trên dữ liệu chưa thấy trước, giúp cải thiện khả năng dự đoán.
- **Linh hoạt:** ANN có thể được điều chỉnh bằng cách thay đổi số tầng ẩn, số neuron, hàm kích hoạt và các siêu tham số để phù hợp với bài toán cụ thể.
- **Khả năng thích nghi với dữ liệu lớn:** ANN có thể xử lý tập dữ liệu lớn và phức tạp tốt hơn so với các mô hình truyền thống.

2 Hạn chế

- **Tính toán phức tạp:** ANN đòi hỏi nhiều tài nguyên tính toán hơn so với các mô hình truyền thống, đặc biệt là khi số tầng ẩn và số neuron tăng lên.
- **Cần nhiều dữ liệu để huấn luyện:** ANN yêu cầu một lượng dữ liệu lớn để tránh overfitting và đạt hiệu suất cao.
- **Khó giải thích:** Không giống như hồi quy tuyến tính hoặc cây quyết định, ANN hoạt động như một "hộp đen", khiến việc giải thích kết quả trở nên khó khăn.

- **Đễ bị overfitting:** Nếu không có kỹ thuật điều chuẩn (regularization) phù hợp, ANN có thể học quá mức dữ liệu huấn luyện và kém hiệu quả trên dữ liệu mới.
- **Cần tối ưu hóa siêu tham số:** Việc lựa chọn số tầng ẩn, số neuron, hàm kích hoạt, tốc độ học và thuật toán tối ưu cần được tinh chỉnh cẩn thận để đạt kết quả tốt nhất.

3 Nhận xét tổng quát

Mô hình ANN mang lại độ chính xác cao và khả năng học tốt các quan hệ phi tuyến, nhưng cũng yêu cầu nhiều tài nguyên tính toán và dữ liệu lớn để đạt hiệu suất tối ưu. Do đó, khi triển khai ANN, cần cân nhắc giữa độ chính xác và chi phí tính toán để lựa chọn cấu hình phù hợp.

4.1 Hướng phát triển trong tương lai

Để cải thiện hiệu suất của mô hình mạng nơ-ron nhân tạo (ANN) trong việc dự đoán lượng khí thải CO₂, có thể xem xét một số hướng phát triển sau:

1 1. Tối ưu hóa kiến trúc mô hình

- **Thử nghiệm các kiến trúc ANN nâng cao:** Sử dụng các mô hình tiên tiến hơn như Mạng Nơ-ron tích chập (CNN) để phát hiện đặc trưng tốt hơn hoặc Mạng Nơ-ron hồi tiếp (RNN/LSTM) nếu dữ liệu có yếu tố thời gian.
- **Điều chỉnh siêu tham số:** Sử dụng các kỹ thuật như Grid Search hoặc Bayesian Optimization để tìm cấu hình mô hình tối ưu nhất.
- **Áp dụng các phương pháp điều chuẩn:** Sử dụng dropout, batch normalization hoặc regularization (L1, L2) để giảm overfitting và cải thiện khả năng tổng quát hóa.

2 2. Cải thiện chất lượng dữ liệu

- **Thu thập thêm dữ liệu:** Càng nhiều dữ liệu, mô hình càng học tốt hơn và giảm nguy cơ overfitting.
- **Tiền xử lý dữ liệu nâng cao:** Áp dụng các kỹ thuật làm sạch dữ liệu, giảm nhiễu và trích xuất đặc trưng để cung cấp dữ liệu đầu vào chất lượng cao hơn.
- **Bổ sung đặc trưng mới:** Sử dụng kỹ thuật Feature Engineering để tạo ra các đặc trưng mới có ý nghĩa, giúp mô hình học tốt hơn.

3. Kết hợp với các mô hình khác

- **Mô hình kết hợp (Ensemble Learning):** Kết hợp ANN với các mô hình khác như Rừng ngẫu nhiên (RF) hoặc Boosting để tận dụng ưu điểm của từng mô hình.
- **Học chuyển giao (Transfer Learning):** Áp dụng các mô hình đã được huấn luyện trên các tập dữ liệu tương tự để cải thiện hiệu suất với lượng dữ liệu huấn luyện nhỏ hơn.
- **Học tăng cường (Reinforcement Learning):** Sử dụng các kỹ thuật học tăng cường để tối ưu hóa quá trình dự đoán và ra quyết định.

4. Triển khai và ứng dụng thực tế

- **Tích hợp vào hệ thống IoT:** Kết hợp ANN với cảm biến IoT để dự đoán lượng khí thải CO₂ theo thời gian thực.
- **Triển khai trên nền tảng đám mây:** Sử dụng các dịch vụ như Google Cloud AI, AWS hoặc Azure để triển khai mô hình một cách hiệu quả và mở rộng quy mô.
- **Ứng dụng trong chính sách môi trường:** Sử dụng mô hình để hỗ trợ các nhà hoạch định chính sách trong việc kiểm soát và giảm thiểu lượng khí thải CO₂.

5. Nghiên cứu mở rộng

- **Mở rộng phạm vi dự đoán:** Ngoài CO₂, có thể mở rộng mô hình để dự đoán các loại khí thải khác như NO_x, SO₂.
- **Tích hợp học máy có giải thích (Explainable AI - XAI):** Phát triển các phương pháp giúp giải thích quyết định của mô hình ANN để tăng tính minh bạch và đáng tin cậy.
- **Ứng dụng mô hình trong các lĩnh vực khác:** Đưa ANN vào các lĩnh vực như dự báo biến đổi khí hậu, phân tích chất lượng không khí, hoặc tối ưu hóa tiêu thụ năng lượng.

Tài liệu tham khảo

[1] Bài giảng của ThS. Nguyễn Thái Khánh và ThS. Lê Thị Thùy Trang, 2025.