**BÁO CÁO THỰC TẬP**

Nguyễn Thành Trung

T9/2023

MỤC LỤC

[1. Tổng quan về Machine Learning 3](#_Toc146897677)

[1.1. Khái niệm 3](#_Toc146897678)

[1.2. Ứng dụng 3](#_Toc146897679)

[1.3. Quy trình của Machine Learning 4](#_Toc146897680)

[1.3.1. Thu thập dữ liệu 4](#_Toc146897681)

[1.3.2. Tiền xử lý dữ liệu 4](#_Toc146897682)

[1.3.3. Huấn luyện mô hình 5](#_Toc146897683)

[1.3.4. Đánh giá mô hình 5](#_Toc146897684)

[1.3.5. Triển khai mô hình 5](#_Toc146897685)

[1.4. Phân loại Machine Learning 5](#_Toc146897686)

[1.4.1. Supervised Learning 5](#_Toc146897687)

[1.4.2. Unsupervised Learning 6](#_Toc146897688)

[1.4.3. Semi-supervised Learning 7](#_Toc146897689)

[1.4.4. Reinforcement Learning 8](#_Toc146897690)

[1.5. Deep Learning 9](#_Toc146897691)

[2. Bài toán ứng dụng 10](#_Toc146897692)

[2.1. Mô tả bài toán 10](#_Toc146897693)

[2.2.Tiền xử lý dữ liệu 11](#_Toc146897694)

[2.2.1. Trực quan hóa dữ liệu 11](#_Toc146897695)

[2.2.2. Lựa chọn đặc trưng 22](#_Toc146897696)

[2.2.3. Xử lý outlier 23](#_Toc146897697)

[2.2.4. Chuẩn hóa dữ liệu 27](#_Toc146897698)

[2.2.5. Chia dữ liệu 28](#_Toc146897699)

[2.3. Huấn luyện và đánh giá mô hình 28](#_Toc146897700)

[2.3.1. Mô hình Linear Regression (LR) 28](#_Toc146897701)

[2.3.2. Mô hình KNN Regression 30](#_Toc146897702)

[2.3.3. Mô hình Support Vector Regression (SVR) 31](#_Toc146897703)

[2.3.4. Mô hình Random Forest Regression và Gradient Boosting Regression 34](#_Toc146897704)

[2.3.5. Kết luận 37](#_Toc146897705)

[3. Tổng kết 37](#_Toc146897706)

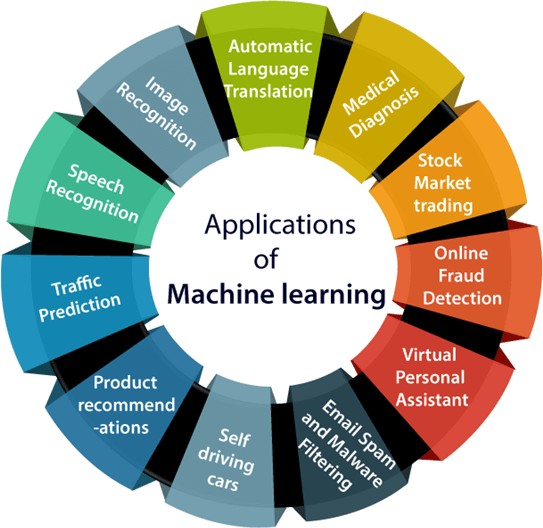
[4.Tài liệu tham khảo 37](#_Toc146897707)

# 1. Tổng quan về Machine Learning

## 1.1. Khái niệm

Machine Learning là một lĩnh vực con của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép hệ thống học tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Các thuật toán Machine Learning xây dựng mô hình dựa trên dữ liệu mẫu gọi là dữ liệu huấn luyện để đưa ra dự đoán hoặc quyết định mà không cần được lập trình rõ ràng.

## 1.2. Ứng dụng

Machine Learning đang được ứng dụng trong lĩnh vực thị giác máy tính, nhận diện giọng nói, lọc thư điện tử, hệ thống gợi ý, tránh gian lận, dự đoán,…

Hình 1.1. Ứng dụng của Machine Learning

## 1.3. Quy trình của Machine Learning

Hình 1.2. Quy trình của Machine Learning

Quy trình của Machine Learning sẽ bao gồm 5 bước là: thu thập dữ liệu, tiền xử lý dữ liệu, huấn luyện mô hình, đánh giá mô hình và cuối cùng là triển khai mô hình.

### 1.3.1. Thu thập dữ liệu

Quá trình thu thập dữ liệu là bước đầu tiên trong quy trình của Machine Learning, việc thu thập dữ liệu sẽ ảnh hưởng trực tiếp đến hiệu suất của mô hình. Bước thu thập dữ liệu này phụ thuộc vào một số điều kiện dưới đây:

* Số lượng dữ liệu càng lớn thì sẽ càng xây dựng được mô hình hiệu quả. Muốn có cái nhìn đúng đắn nhất về số lượng dữ liệu chúng ta cần xem xét các mô hình đã được huấn luyện và được đánh giá hiệu năng trên các dữ liệu mới.
* Dữ liệu cần phải liên quan đến bài toán, ít mất mát và trùng lặp, có thể thu thập từ nhiều nguồn và tích hợp với nhau, chất lượng dữ liệu và sự đa dạng của dữ liệu rất quan trọng.
* Phụ thuộc vào độ phức tạp của vấn đề và thuật toán sử dụng. VD: Nhận dạng ảnh và xử lý ngôn ngữ tự nhiên cần tập dữ liệu lớn bởi độ phức tạp vấn đề.
* Phụ thuộc vào độ phức tạp của mô hình. VD: Các mô hình phức tạp muốn đạt hiệu suất cao thì cần phải có nhiều dữ liệu để tính toán (các mô hình ensemble cần nhiều dữ liệu vì sử dụng đa mô hình kết hợp với nhau).
* Phụ thuộc vào chất lượng và tính chính xác của dữ liệu. VD: Giả sử có nhiều nhiễu và thông tin không đúng trong dữ liệu, ta cần phải tăng kích thước bộ dữ liệu để cải thiện độ chính xác cho mô hình.

### 1.3.2. Tiền xử lý dữ liệu

Sau quá trình thu thập dữ liệu ta cần phải tiền xử lý dữ liệu nhằm làm sạch dữ liệu trước khi huấn luyện với các mô hình, điều này giúp cải thiện hiệu năng của mô hình Machine Learning. Bước tiền xử lý dữ liệu bao gồm:

* Tìm hiểu thông tin về bộ dữ liệu, số lượng cá thể, số đặc trưng, phân phối của dữ liệu, mối quan hệ giữa các biến độc lập với nhau và với biến phụ thuộc
* Trực quan hóa dữ liệu
* Kiểm tra vấn đề dữ liệu bị thiếu, dữ liệu mất cân bằng, ngoại lệ trong dữ liệu
* Mã hóa đặc trưng phân loại sang đặc trưng định lượng
* Chuẩn hóa đặc trưng định lượng
* Lựa chọn đặc trưng để chọn ra đặc trưng có giá trị nhất cho mô hình
* Chia dữ liệu thành các tập huấn luyện và kiểm thử

### 1.3.3. Huấn luyện mô hình

Dữ liệu sau khi tiền xử lý sẽ được đưa vào huấn luyện với các mô hình Machine Learning. Trong bước này ta sẽ áp dụng các thuật toán phù hợp tùy thuộc vào loại dữ liệu và bài toán cần giải quyết, cuối cùng chọn ra thuật toán tối ưu để xây dựng được mô hình tốt. Quá trình huấn luyện sẽ bao gồm các công việc như sau:

* Lựa chọn thuật toán phù hợp cho bài toán để huấn luyện mô hình
* Sau quá trình kiểm thử nếu mô hình không tốt ta có thể cải thiện bằng cách thay đổi các siêu tham số của mô hình. Mỗi mô hình Machine Learning đều có những tham số đặc trưng để thay đổi, quá trình này là bước tinh chỉnh tham số của mô hình trong giai đoạn huấn luyện.

### 1.3.4. Đánh giá mô hình

Bước này sẽ bao gồm các công việc như tính toán, đánh giá độ chính xác của mô hình cuối cùng trên tập dữ liệu thực tế, xem xét tính quan trọng của thuộc tính trong mô hình cùng với chi phí vận hành để từ đó quyết định xây dựng lại và cải thiện mô hình với các bước trên hay triển khai mô hình.

### 1.3.5. Triển khai mô hình

Đưa mô hình vào thực tế, đánh giá lại mô hình liên tục để xây dựng và cải thiện mô hình.

## 1.4. Phân loại Machine Learning

Machine learning có thể được chia thành 4 phương thức học chính:

### 1.4.1. Supervised Learning

**A diagram of a diagram

Description automatically generated**

Hình 1.3. Mô hình Supervised Learning

Mô hình Supervised Learning là một phương thức học mà mô hình được huấn luyện bằng cách sử dụng dữ liệu đào tạo được gán nhãn (dữ liệu được gán nhãn là dữ liệu đầu vào đã được gán với đầu ra chính xác đã biết) và từ đó mô hình dự đoán cho dữ liệu không nhãn.

Supervised Learning được chia làm 2 loại là mô hình hồi quy và mô hình phân loại:

* **Mô hình hồi quy:** được sử dụng để dự đoán giá trị của biến phụ thuộc dựa trên các biến độc lập, trong đó biến phụ thuộc có giá trị liên tục.

VD: Dự đoán lượng mưa dựa trên nhiệt độ, áp suất không khí, thời gian, sức gió,…

Các thuật toán chính trong mô hình hồi quy là hồi quy tuyến tính (Linear Regression) và hồi quy đa thức (Polynomial Regression).

* **Mô hình phân loại:** được sử dụng để dự đoán lớp của biến phụ thuộc dựa trên các biến độc lập, trong đó biến phụ thuộc là giá trị rời rạc.

VD: Trong vấn đề phân loại thư rác thì thư được phân loại spam/not spam

Một số thuật toán phân loại được sử dụng là Logistic Regression, K-nearest-neighbor, Support Vector Machine,…

### 1.4.2. Unsupervised Learning

**A diagram of a process

Description automatically generated**

Hình 1.4. Mô hình Unsupervised Learning

Mô hình Unsupervised Learning được huấn luyện với dữ liệu không được gán nhãn, các thuật toán tự mô hình hóa cấu trúc ẩn trong dữ liệu để tìm ra tính chất hay đặc tính của dữ liệu.

Các thuật toán Unsupervised Learning dựa vào các mục đích khác nhau để phân loại, bao gồm:

* **Phân cụm (Clustering):** là các thuật toán gom dữ liệu vào các nhóm khác nhau sao cho mỗi điểm dữ liệu tương đồng với các điểm dữ liệu khác trong cùng nhóm, sử dụng các thuật toán như K-means clustering, Mean-shift clustering,…
* **Giảm chiều dữ liệu (Dimensionality Reduction):** là phương pháp biến đổi dữ liệu từ không gian chiều cao thành không gian chiều thấp đồng thời giữ lại một số thuộc tính có ý nghĩa của dữ liệu gốc giúp mô hình không bị quá phức tạp, sử dụng các thuật toán như Singular Value Decompostion (SVD), Principal Component Analysis (PCA),…

### 1.4.3. Semi-supervised Learning

Là một phương thức học nằm giữa supervised learning và unsupervised learning, sử dụng lượng nhỏ dữ liệu được gán nhãn và lượng lớn dữ liệu không nhãn để huấn luyện mô hình.

Một số kĩ thuật semi-supervised learning:

* **Self-training**: kĩ thuật tự đào tạo là quy trình lặp dự đoán pseudo-labeling cho dữ liệu không nhãn từ mô hình đã huấn luyện với dữ liệu được gán nhãn. Các pseudo label có độ tin cậy cao sẽ được bổ sung vào bộ dữ liệu gán nhãn để tiếp tục huấn luyện mô hình.

A diagram of a self-training method

Description automatically generated

Hình 1.5. Kĩ thuật Self-training

* **Co-training**: kĩ thuật được cải tiến từ self-training, hoạt động bằng cách đào tạo đồng thời hai mô hình phân loại khác nhau dựa trên hai bộ đặc trưng khác nhau và độc lập với nhau của dữ liệu. Các pseudo label có độ tin cậy cao nhất của mô hình 1 sẽ được bổ sung vào bộ dữ liệu huấn luyện có nhãn của mô hình 2 để tiếp tục huấn luyện và tương tự với các pseudo label có độ tin cậy cao của mô hình 2, quá trình này lặp lại nhiều lần để cải thiện hiệu suất mô hình.

A diagram of a training method

Description automatically generated

Hình 1.6. Kĩ thuật Co-training

Một số thuật toán được sử dụng trong semi-supervised learning là FixMatch, MixMatch,…

### 1.4.4. Reinforcement Learning

**A diagram of a diagram of a business

Description automatically generated**

Hình 1.7. Mô hình Reinforcement Learning

Reinforcement Learning là một phương thức học trong đó 1 tác nhân (agent) học cách phản hồi trong một môi trường (environment) bằng cách thực hiện các hành động (action) và xem phần thưởng (reward) của các hành động đó.

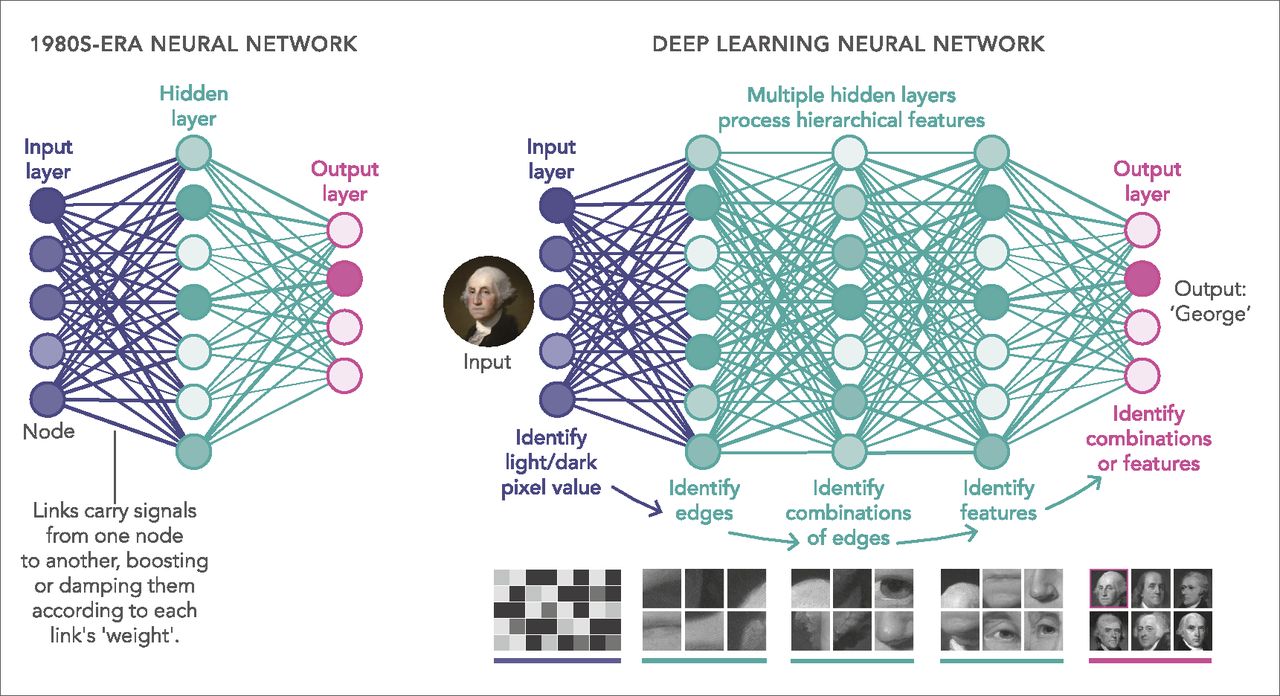
Trong reinforcement learning, agent học tự động từ phản hồi mà không có dữ liệu huấn luyện và học từ kinh nghiệm khi tương tác với environment.

Reinforcement Learning gồm 4 thành phần chính:

* **Policy**: cách agent hành động (action) tại 1 thời điểm nhất định. Mục tiêu của agent trong reinforcement learning là tìm policy sao cho tối đa hóa phần thưởng̣ (reward).
* **Reward signal**: Ở mỗi state (trạng thái) của environment, environment gửi reward signal đến agent. Agent nhận được reward signal tùy theo action, reward signal có thể ảnh hưởng đến policy vì bad action dẫn đến low reward nên agent cần thay đổi policy để chọn action khác trong tương lai.
* **Value function**: hàm giá trị cho biết thông tin về mức độ tốt của state và action cũng như reward mà agent có thể có.
* **Model**: Model giúp dự đoán về cách agent hành động. VD: Ở mỗi state của environment agent thực hiện một action thì model sẽ dự đoán state tiếp theo và reward.

Một số thuật toán sử dụng trong Reinforcement Learning là Q-learning, Deep Q- Networks, SARSA, DDPG,…

## 1.5. Deep Learning



Hình 1.8. Deep Learning

Deep Learning là một lĩnh vực con của Machine Learning. Deep Learning là phương thức học máy phức tạp hơn với mạng nơ ron (Neural Network), mô phỏng giống với cách bộ não con người tư duy và kết luận mà không cần đến nhiều sự can thiệp của con người trong quá trình học như Machine Learning.

Deep learning được ứng dụng trong thị giác máy tính, phân tích giọng nói, sinh văn bản, hệ thống lái xe tự động,...

# 2. Bài toán ứng dụng

## 2.1. Mô tả bài toán

Khí thải CO2 từ phương tiện giao thông là một trong những nguyên nhân chính gây ô nhiễm môi trường, việc kiểm soát lượng khí thải CO2 từ phương tiện do đó đã gợi lên mối quan tâm toàn cầu trong nghiên cứu ứng dụng trong lĩnh vực Machine Learning. Việc ước tính và trực quan mức tiêu thụ nhiên liệu và khí thải là rất quan trọng đối với định lượng chi phí năng lượng và ô nhiễm do giao thông vận tải gây ra.Với mong muốn xây dựng một mô hình có thể dự đoán lượng khí thải CO2 từ phương tiện và biết được những tác nhân quan trọng trong việc gây ra lượng khí thải CO2, báo cáo sẽ tập trung vào phân tích dữ liệu và thử nghiệm các thuật toán để từ đó chọn ra mô hình dự đoán chính xác nhất lượng khí thải CO2.

Bộ dữ liệu trong bài toán này được tổng hợp ở Canada trong vòng 7 năm từ 2014-2021, ghi lại chi tiết về lượng khí thải CO2 của 7385 các phương tiện và các thông tin kĩ thuật về phương tiện đó.

Đây là bài toán hồi quy với đầu ra cần dự đoán là CO2 Emission(g/km) có giá trị liên tục. Các đặc trưng của dữ liệu trong bài toán được đề cập ở bảng 2.1 dưới đây:

Bảng 2.1. Các đặc trưng của bài toán

|  |  |
| --- | --- |
| **Đặc trưng** | **Ý nghĩa** |
| Make | Hãng xe |
| Model | Mẫu xe |
| Vehicle Class | Phân khúc xe |
| Engine Size (L) | Dung tích động cơ |
| Cylinders | Xilanh động cơ |
| Tranmission | Hộp số truyền động |
| Fuel Type | Loại nhiên liệu |
| Fuel Consumption City (L/100km) | Nhiên liệu tiêu thụ trong đô thị (L/100km) |
| Fuel Consumption Hwy (L/100km) | Nhiên liệu tiêu thụ ngoài đô thị (L/100km) |
| Fuel Consumption Comb (L/100km) | Nhiên liệu tiêu thụ hỗn hợp (L/100km) |
| Fuel Consumption Comb (mpg) | Nhiên liệu tiêu thụ hỗn hợp (dặm trên gallon) |
| CO2 Emission (g/km) | Lượng khí thải CO2 ước tính (g/km) |

## 2.2.Tiền xử lý dữ liệu

Trong phần tiền xử lý dữ liệu này, báo cáo tập trung thực hiện các bước trực quan dữ liệu, xử lý outlier, lựa chọn đặc trưng, mã hóa dữ liệu và chia bộ dữ liệu thành ba tập khác nhau dùng để huấn luyện, tinh chỉnh mô hình và đánh giá.

### 2.2.1. Trực quan hóa dữ liệu

Bước trực quan hóa này sẽ bao gồm trực quan phân bố của đặc trưng phân loại, phân phối của đặc trưng định lượng, phân bố lượng CO2 trung bình ở mỗi đặc trưng và các kết quả được giải thích ở cuối các hình trực quan:

* Các đặc trưng phân loại: Make, Model, Vehicle Class, Transmission, Fuel Type
* Các đặc trưng định lượng: Engine Size(L), Cylinders, Fuel Consumption City (L/100km), Fuel Consumption Hwy (L/100km), Fuel Consumption Comb (L/100km), Fuel Consumption Comb (mpg), CO2 Emission (g/km)

Trước khi trực quan các đặc trưng phân loại ở trên, tác giả đã xem xét hai đặc trưng Transmission và Fuel Type có thể được xử lý trước:

* Đặc trưng Transmission gồm 27 giá trị duy nhất được xác định bởi hai thành phần là kí hiệu của mỗi loại hộp số bao gồm 5 giá trị riêng biệt (A, AM, AS, AV, M) và cấp của hộp số nên tác giả có cơ sở để nhóm các giá trị trên thành 5 giá trị mới là Automatic with Select Shift, Automatic, Manual, Automated Manual, Continuously Variable.
* Đặc trưng Fuel Type gồm 5 giá trị được kí hiệu (Z, X, D, E, N) nên tác giả muốn làm rõ 5 loại này bằng việc thay thế bằng ý nghĩa thực tế của chúng thành Regular Gasoline, Premium Gasoline, Ethanol(E85), Diesel, Natural Gas.
* Phân bố dữ liệu của các đặc trưng phân loại được trực quan bằng biểu đồ cột với trục x biểu thị các giá trị duy nhất của mỗi đặc trưng và trục y biểu thị số lượng xe được thể hiện ở các hình từ 2.1 đến 2.5 dưới đây:

A graph of a number of people

Description automatically generated with medium confidence

Hình 2.1. Phân bố của Make

**A graph of different colored bars

Description automatically generated with medium confidence**

Hình 2.2. Phân bố của Model

**A graph of a bar graph

Description automatically generated with medium confidence**

Hình 2.3. Phân bố của Vehicle Class

**A graph of a bar graph

Description automatically generated with medium confidence**

Hình 2.4. Phân bố của Transmission

**A graph with different colored bars

Description automatically generated**

Hình 2.5. Phân bố của Fuel Type

Kết quả trực quan cho các đặc trưng phân loại cho thấy được Ford là hãng xe phổ biến nhất (Hình 2.1), mẫu xe F-150 FFV 4x4 và F-150 FFV phổ biến nhất trong bộ dữ liệu (Hình 2.2), SUV-SMALL là phân khúc xe được lựa chọn nhiều nhất (Hình 2.3), Automatic with Select Shift xuất hiện nhiều nhất ở các phương tiện (Hình 2.4), Regular Gasoline là loại nhiên liệu được sử dụng nhiều nhất ở các phương tiện (Hình 2.5). Phân bố của đặc trưng Fuel Type có sự bất thường với giá trị phân loại Natural Gas có một dữ liệu duy nhất và ngoài thực tế loại nhiên liệu này hầu như không được được sử dụng trên phương tiện bởi vì sự bất lợi của nó nên ta sẽ loại bỏ dữ liệu này.

Để giải thích cho những quan sát trên thì tác giả đã tìm hiểu thực tế và thấy được rằng:

* Ford là hãng xe nổi tiếng và phổ biến nhất ở các nước Bắc Mỹ như USA và Canada và mẫu xe F-150 FFV 4x4 theo khảo sát là mẫu xe yêu thích ở Canada.
* Phân khúc SUV-SMALL được lựa chọn phổ biến ở Canada do việc thường xuyên di chuyển đường dài của người dân cùng với địa hình và khí hậu lạnh, chi phí nhiên liệu rẻ.
* Hộp số tự động Automatic xuất hiện ở phần lớn phương tiện có thể thấy được từ sự cải tiến của nó so với hộp số sàn - dễ sử dụng, thuận tiện và tiết kiệm nhiên liệu và nhiều lợi ích khác.
* Regular Gasoline hay xăng vẫn là loại nhiên liệu phổ biến bởi nó tạo ra nhiều năng lượng hơn so với các loại nhiên liệu khác và dễ sản xuất cũng như giá thành rẻ.
* Phân phối dữ liệu của các đặc trưng định lượng được trực quan bằng biểu đồ tần suất hay biểu đồ phân phối tần suất (Histogram) với trục x biểu thị phân bố của đặc trưng và trục y biểu thị mật độ thể hiện ở các hình từ 2.6 đến 2.12 dưới đây:

**A graph of a graph

Description automatically generated with medium confidence**

Hình 2.6. Phân phối của Engine Size

**A graph with a blue line

Description automatically generated**

Hình 2.7. Phân phối của Cylinders

**A graph of a curve

Description automatically generated with medium confidence**

Hình 2.8. Phân phối của Fuel Consumption City (L/100km)

**A graph with a line going up

Description automatically generated**

Hình 2.9. Phân phối của Fuel Consumption Hwy (L/100km)

**A graph of a graph

Description automatically generated with medium confidence**

Hình 2.10. Phân phối của Fuel Consumption Comb (L/100km)

**A graph with a line going up

Description automatically generated**

Hình 2.11. Phân phối của Fuel Consumption Comb (L/100km)

**A graph with a line going up

Description automatically generated**

Hình 2.12. Phân phối của CO2 Emission(g/km)

* Kết quả trực quan phân phối đặc trưng định lượng ở các hình từ 2.6 đến 2.12 cho thấy rằng dữ liệu không phân bố đều ở mỗi đặc trưng và hầu hết các đặc trưng định lượng có phân phối positive skew nghĩa là phân phối lệch phải, qua đó phân phối đặc trưng định lượng không phải phân phối chuẩn.
* Dựa vào hình dạng không phải phân phối chuẩn của đặc trưng định lượng, ta có thể xác định được sự xuất hiện của các dữ liệu bất thường tập trung ở phía phải của phân phối dữ liệu. Ta sẽ xem xét và xử lý các điểm dữ liệu bất thường này ở các phần sau.
* Phân bố lượng khí thải CO2 trung bình của các đặc trưng định lượng được trực quan bằng biểu đồ cột với trục x là các giá trị duy nhất và trục y biểu thị lượng khí thải CO2 ở các hình từ 2.13 đến 2.18 dưới đây:

**A graph of a graph of cylinders

Description automatically generated with medium confidence**

Hình 2.13. Phân bố lượng khí thải CO2 trung bình của Cylinders

**A graph of different colored bars

Description automatically generated**

Hình 2.14. Phân bố lượng khí thải CO2 trung bình của Engine Size(L)

**A graph of a graph

Description automatically generated with medium confidence**

Hình 2.15. Phân bố lượng khí thải CO2 trung bình của Fuel Consumption City (L/100km)

**A graph of a graph

Description automatically generated with medium confidence**

Hình 2.16. Phân bố lượng khí thải CO2 trung bình của Fuel Consumption Hwy (L/100km)

**A graph of a graph

Description automatically generated with medium confidence**

Hình 2.17. Phân bố lượng khí thải CO2 trung bình của Fuel Consumption Comb (L/100km)

**A graph of a bar graph

Description automatically generated with medium confidence**

Hình 2.18. Phân bố lượng khí thải CO2 trung bình của Fuel Consumption Comb (mpg)

Kết quả trực quan từ hình 2.13 đến 2.18 và việc tìm hiểu thực tế cho thấy rằng:

* Những xe có số lượng Cylinders (xilanh) càng nhiều và Engine Size (L) (dung tích động cơ) càng lớn thì sẽ càng thải nhiều CO2 bởi vì thực tế những xe này sẽ có công suất lớn vì vậy tiêu thụ nhiên liệu nhanh hơn dẫn đến thải nhiều khí thải CO2.
* Fuel Consumption City (L/100km) và Fuel Consumption Hwy (L/100km) càng cao thì sẽ càng thải ra nhiều CO2 dẫn đến Fuel Consumption Comb (L/100km) tăng cao cũng sẽ thải nhiều CO2 hơn. (Fuel Consumption Comb = 0.55 \* Fuel Consumption City + 0.45 \* Fuel Consumption Hwy)
* Fuel Consumption Comb (mpg) càng thấp thì sẽ thải nhiều CO2 bởi vì mpg là đơn vị dặm trên gallon nên xe càng có số dặm đi được trên gallon ít chứng tỏ xe tiêu thụ nhiên liệu nhiều dẫn đến lượng khí thải CO2 trung bình cao.
* Các đặc trưng này có mối liên kết chặt chẽ với nhau (Cylinder, Engine Size(L), Fuel Consumption Comb (L/100km), Fuel Consumption Hwy (L/100km), Fuel Consumption City (L/100km)) nên ta xác định được sẽ có hiện tượng các biến độc lập định lượng có mối tương quan cao với nhau.
* Phân bố lượng khí thải CO2 trung bình của các đặc trưng phân loại được trực quan bằng biểu đồ cột với trục x là các giá trị duy nhất và trục y biểu thị lượng khí thải CO2 trung bình ở các hình từ 2.19 đến 2.22 dưới đây:

**A graph with different colored bars

Description automatically generated with medium confidence**

Hình 2.19. Phân bố lượng khí thải CO2 trung bình của Make

**A chart of different colored rectangles

Description automatically generated**

Hình 2.20. Phân bố lượng khí thải CO2 trung bình của Vehicle Class

**A green red and orange squares

Description automatically generated**

Hình 2.21. Phân bố lượng khí thải CO2 trung bình của Transmission

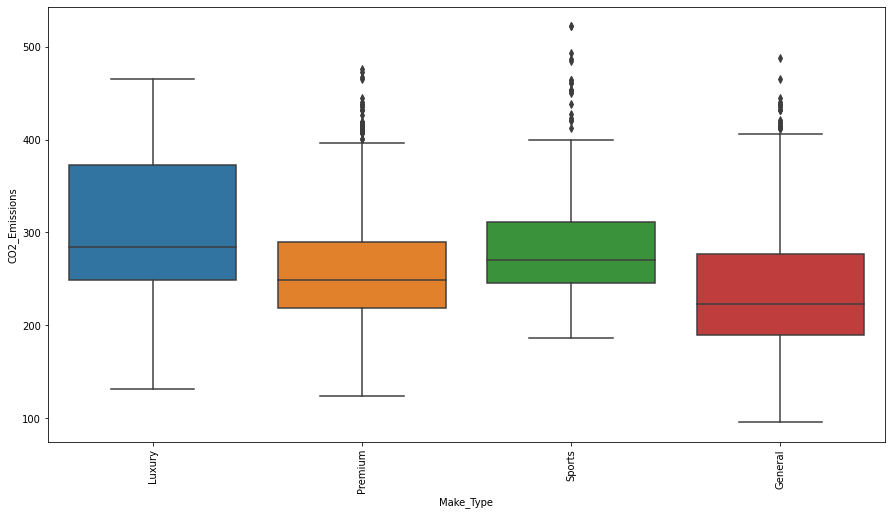
**A green red and orange squares

Description automatically generated** Hình 2.22. Phân bố lượng khí thải CO2 trung bình của Fuel Type

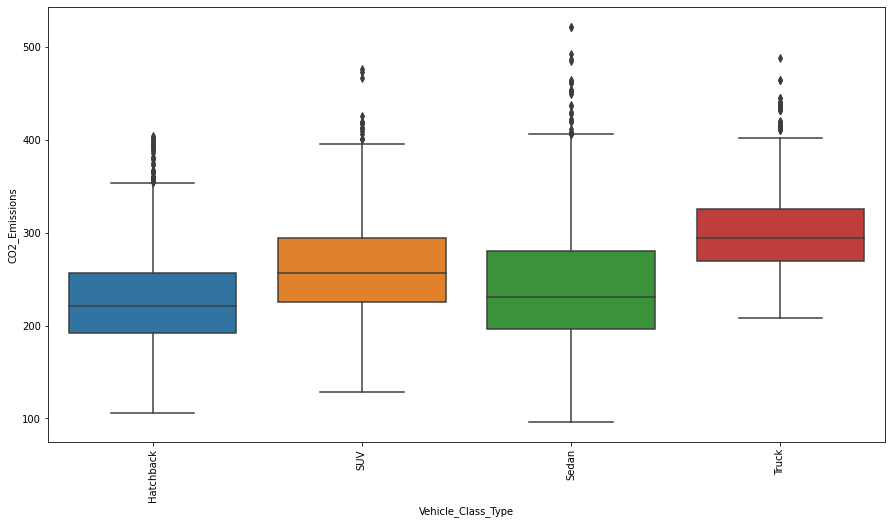
Sau khi trực quan lượng CO2 trung bình của các giá trị duy nhất ở mỗi đặc trưng phân loại cùng với tìm hiểu thực tế bên ngoài, tác giả xem xét hai đặc trưng Make và Vehicle Class được thể hiện trong hình 2.19 và 2.20 có thể xử lý được bằng cách nhóm các giá trị phân loại tương đồng với nhau:

* Đặc trưng Make gồm 42 giá trị nhóm về 4 giá trị mới là:
* General: bao gồm các hãng xe CHEVROLET, FIAT, FORD, KIA, HONDA, HYUNDAI, MITSUBISHI, NISSAN, RAM, SCION, SUBARU, TOYOTA, VOLKSWAGEN
* Premium: bao gồm các hãng xe ALFA ROMEO, AUDI, BMW, BUICK, CADILLAC, CHRYSLER, DOGDE, GMC, INFINITI, JEEP, LAND ROVER, LEXUS, MERCEDES-BENZ, MINI, SMART, VOLVO
* Luxury: bao gồm các hãng xe ACURA, BENTLEY, LINCOLN, ROLLS-ROYCE, GENESIS
* Sport: bao gồm các hãng xe BUGATTI, PORSCHE, MASERATI, ASTON MARTIN, LAMBORGHINI, JAGUAR, SRT
* Đặc trưng Vehicle Class gồm 16 giá trị nhóm về 4 giá trị mới là:
* Hatchback: bao gồm các phân khúc COMPACT, MINICOMPACT, SUBCOMPACT
* Sedan: bao gồm các phân khúc MID-SIZE, TWO-SEATER, FULL-SIZE, STATION WAGON – SMALL, STATION WAGON – MID-SIZE
* SUV: bao gồm các phân khúc SUV-SMALL, SUV – STANDARD, MINIVAN
* Truck: bao gồm các phân khúc VAN – CARGO, VAN – PASSENGER, PICKUP TRUCK – STANDARD, SPECIAL PURPOSE VEHICLE, PICKUP TRUCK – SMALL

Lượng CO2 trung bình của các giá trị mới trong đặc trưng Make và Vehicle Class được trực quan bằng các hình 2.23 và 2.24 ở dưới đây:



Hình 2.23. Phân bố lượng khí thải CO2 trung bình của Make sau khi nhóm giá trị



Hình 2.24. Phân bố lượng khí thải CO2 trung bình của Vehicle Class sau khi nhóm giá trị

Kết quả trực quan lượng khí thải CO2 trung bình của các giá trị duy nhất trong mỗi đặc trưng phân loại ở trên cho thấy được rằng:

* Trong đặc trưng Make các xe General thải ra ít CO2 trung bình hơn các xe Luxury và Sport, giải thích cho điều này là vì các xe Luxury và Sport có công suất lớn nhờ vào kích thước động cơ lớn cùng nhiều xilanh dẫn đến tiêu thụ nhiều nhiên liệu hơn và từ đó thải ra CO2 trung bình nhiều hơn. (Hình 2.23)
* Trong đặc trưng Vehicle Class các phân khúc xe Truck và SUV thải ra nhiều khí thải nhất cho thấy kích thước xe lớn sẽ cần kích thước động cơ lớn và từ đó tiêu thụ nhiên liệu nhiều hơn và thải ra nhiều CO2 hơn. (Hình 2.24)
* Các xe sử dụng hộp số Automatic sẽ tiêu thụ nhiên liệu nhiều hơn bởi vì các xe này cần nhiều năng lượng để sang số và chúng cần một bộ biến đổi sử dụng chất lỏng để truyền năng lượng từ động cơ sang hộp số và từ đó thải ra nhiều CO2 hơn các loại hộp số khác.
* Có sự bất thường trong đặc trưng Fuel Type khi loại nhiên liệu Ethanol thực tế thải ra ít CO2. Sau khi phân tích bộ dữ liệu thì tác giả thấy được rằng các xe sử dụng nhiên liệu Ethanol này chiếm tỷ lệ lớn các xe có tác nhân thải ra nhiều CO2 nhất như kích thước động cơ lớn, nhiều xilanh, điều này giải thích cho việc Ethanol có lượng CO2 trung bình cao hơn các loại nhiên liệu khác trong bộ dữ liệu.

### 2.2.2. Lựa chọn đặc trưng

Sau khi trực quan dữ liệu báo cáo tiếp tục với phần chọn lọc đặc trưng cho bài toán. Các đặc trưng được chọn lọc nhằm đảm bảo các thuật toán thực hiện hiệu quả. Mô hình xét đến tính tương quan của đăc trưng với nhau, đảm bảo không có tương quan quá cao giữa các đặc trưng.

Tính tương quan (correlation) giữa các biến độc lập định lượng và với biến phụ thuộc được trực quan hóa với bằng biểu đồ heatmap trong hình 2.25 và biểu đồ phân tán mô tả mối quan hệ giữa các đặc trưng định lượng trong hình 2.26.

**A grey and blue squares

Description automatically generated**

Hình 2.25. Biểu đồ heatmap thể hiện tương quan giữa các đặc trưng định lượng

A graph of blue lines

Description automatically generated with medium confidence

Hình 2.26. Biểu đồ phân tán thể hiện mối quan hệ giữa các đặc trưng định lượng

Dựa vào tính tương quan và mối quan hệ giữa các đặc trưng định lượng được thể hiện ở các biểu đồ 2.25 và 2.26 cùng với ý nghĩa của đặc trưng ta thấy được rằng:

* Fuel Consumption Comb (L/100km) = 0.55 \* Fuel Consumption City (L/100km) + 0.45 \* Fuel Consumption Hwy (L/100km) nên ta sẽ loại bỏ hai đặc trưng Fuel Consumption City (L/100km) và Fuel Consumption Hwy (L/100km).
* Đặc trưng Fuel Consumption Comb (mpg) cũng được loại bỏ bởi cách tính theo đơn vị khác của Fuel Consumption Comb (L/100km) nên sẽ không cần thiết.
* Hầu hết các đặc trưng định lượng tương quan cao với biến phụ thuộc cho thấy mối quan hệ tuyến tính của biến độc lập với đầu ra cần dự đoán.

### 2.2.3. Xử lý outlier

Dựa vào phần trực quan hóa dữ liệu, ta đã biết được rằng có những điểm dữ liệu bất thường trong phân phối của đặc trưng định lượng nên phần này sẽ tiếp tục xử lý các bất thường đó. Phân phối của dữ liệu không phải phân phối chuẩn nên cách phát hiện các điểm dữ liệu bất thường trong bài toán là sử dụng phương pháp IQR, những điểm nằm ngoài khoảng (Q1- 1.5IQR) và (Q3+1.5IQR) được cho là những điểm bất thường.

Phương pháp IQR được trực quan bằng biểu đồ boxplot cho thấy có các điểm bất thường ở các đặc trưng định lượng nhưng sau khi xem xét các điểm dữ liệu này với dữ liệu thực tế tác giả thấy được các điểm bất thường này không phải là lỗi đo lường mà là do biến thiên tự nhiên, có tồn tại ngoài thực tế với số lượng ít nên sẽ xem xét không loại bỏ mà thay thế các điểm dữ liệu bất thường (các giá trị lớn hơn ngưỡng (Q3+1.5IQR)) bằng ngưỡng (Q3+1.5IQR). Sau khi xử lý outlier ta thấy phân phối dữ liệu có sự thay đổi, dữ liệu phân bố đều trên toàn bộ tập dữ liệu.

Kết quả trước và sau khi xử lý bất thường ở các đặc trưng định lượng được thể hiện ở các hình từ 2.27 đến 2.30. Phân phối dữ liệu sau khi xử lý bất thường được thể hiện ở các hình từ 2.31 đến 2.34 dưới đây:

**A screenshot of a graph

Description automatically generated**

Hình 2.27. Xử lý bất thường trong đặc trưng Cylinders

**A screenshot of a graph

Description automatically generated**

Hình 2.28. Xử lý bất thường trong đặc trưng Engine\_Size(L)

**A screenshot of a graph

Description automatically generated**

Hình 2.29. Xử lý bất thường trong đặc trưng Fuel Consumption Comb (L/100km)

**A screenshot of a graph

Description automatically generated**

Hình 2.30. Xử lý bất thường trong biến phụ thuộc CO2 Emission(g/km)

**A graph with a line and a curve

Description automatically generated**

Hình 2.31. Phân phối của đặc trưng Cylinders

**A graph with a line going up

Description automatically generated**

Hình 2.31. Phân phối của đặc trưng Fuel Consumption Comb (L/100km)

**A graph of a graph

Description automatically generated**

Hình 2.31. Phân phối của đặc trưng CO2 Emission (g/km)

**A graph with blue bars and a dotted line

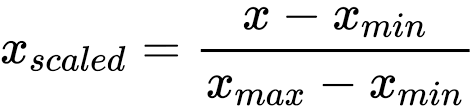
Description automatically generated**

Hình 2.31. Phân phối của đặc trưng Engine Size(L)

### 2.2.4. Chuẩn hóa dữ liệu

Nhằm đảm bảo cho thuật toán tính toán hiệu quả và chính xác, các dữ liệu đầu vào cần được chuẩn hóa.

Đối với những dữ liệu định lượng, phương pháp chuẩn hóa là đặt giá trị lớn nhất và nhỏ nhất là các cận của khoảng và đưa những giá trị trong dữ liệu về cùng một khoảng. Trong bài toán các đặc trưng định lượng được chuẩn hóa về giá trị mới [0,1].

****

Các giá trị nằm ngoài khoảng dữ liệu huấn luyện khi triển khai thực tế vẫn sẽ được tính toán theo min max của tập huấn luyện bởi vì các dữ liệu thực tế này phải có cùng tỷ lệ chuẩn hóa với dữ liệu huấn luyện.

Đối với các đặc trưng phân loại, bài toán lựa chọn hai cách tiếp cận là Dummy Encoding và Frequency Encoding. Các giá trị phân loại được đưa về vector chứa 0 và 1 thể hiện sự xuất hiện của lớp của mỗi mẫu dữ liệu, biểu diễn bằng các cột là các đặc trưng mới trong dữ liệu thông qua get\_dummies, là công cụ trong thư viện scikit-learn.

Với đặc trưng Model có 2053 giá trị duy nhất thì việc sử dụng biến dummies cho từng giá trị phân loại sẽ tăng số chiều của dữ liệu lên rất nhiều làm giảm hiệu suất mô hình nên phải chọn cách xử lý khác. Tác giả sử dụng phương pháp Frequency Encoding để mã hóa đặc trưng này, Frequency Encoding trả về tần suất xuất hiện của từng giá trị phân loại trong toàn bộ tập dữ liệu. Mô hình sẽ hiểu rằng giá trị phân loại có tần suất nhỏ ít đáng tin cậy hơn giá trị phân loại có tần suất cao, phương pháp này có ưu điểm là không tăng số chiều của dữ liệu và dễ thực hiện mà vẫn giữ thông tin của dữ liệu, hoạt động tốt với các thuật toán tree-based nhưng phương pháp mã hóa này giả định mối tương quan giữa tần suất và biến đầu ra.

Khi triển khai thực tế, những giá trị phân loại nằm trong lớp chưa xác định so với dữ liệu được huấn luyện, được đưa về lớp tương đồng với lớp có sẵn, hoặc sử dụng lớp “Unknown”, đã xác định trong mô hình. Tuy nhiên, việc thực hiện lại và cập nhật bước chuẩn hóa cũng được xem xét nếu dữ liệu huấn luyện không còn phù hợp với thực tế.

### 2.2.5. Chia dữ liệu

Dữ liệu được chia thành 3 tập khác nhau là tập train để huấn luyện mô hình, tập validation để tinh chỉnh tham số mô hình và tập test để đánh giá mô hình cuối cùng.

## 2.3. Huấn luyện và đánh giá mô hình

Các thuật toán chính được thử nghiệm trong bài toán là Linear Regression, KNN Regression, Support Vector Regression tuy nhiên các thuật toán ensemble như Random Forest Regression, Gradient Boosting Regression cũng được thử nghiệm để đánh giá mô hình một cách khách quan.

Mô hình được đánh giá qua các metric sau:

* R-squared : thể hiện biến độc lập giải thích bao nhiêu phần trăm variance của biến phụ thuộc, đo lường mức độ khớp của mô hình dự đoán với dữ liệu thực tế, giả định tất cả biến độc lập đều ảnh hưởng đến kết quả mô hình và càng gần 1 càng tốt, có thể âm.

A mathematical equation with black text

Description automatically generated

* Mean squared error: trung bình bình phương giá trị chêch lệch giữa dự đoán và thực tế, giá trị này càng gần 0 mô hình càng tốt.

A mathematical equation with numbers and symbols

Description automatically generated

### 2.3.1. Mô hình Linear Regression (LR)

Hồi quy tuyến tính là một thuật toán phân tích mối quan hệ giữa biến phụ thuộc với một hay nhiều biến độc lập, giả định mối quan hệ tuyến tính giữa các biến để tìm ra một hàm phù hợp để mô tả mối quan hệ này.

Trong báo cáo này mô hình Linear Regression sử dụng phương pháp Gradient Descent (GD) nhằm tìm các trọng số của mô hình bằng việc cực tiểu hóa loss function, các tham số quan trọng trong mô hình là learning rate (tốc độ học) và iteration (số bước lặp).

Tham số learning rate trong mô hình ảnh hưởng đến tốc độ hội tụ của thuật toán, learning rate lớn mô hình hội tụ nhanh hơn nhưng có thể đi qua điểm cực tiểu nếu quá lớn, learning rate nhỏ thuật toán hội tụ lâu hơn và có thể chưa đạt đến điểm cực tiểu khi hết số iteration. Điều kiện dừng có thể là hết số iteration hoặc loss function nhỏ hơn một threshhold nào đó.

Trong báo cáo này, mô hình sẽ sử dụng các tham số ban đầu được xác định sẵn là learning rate bằng 0.01 và iteration bằng 500 rồi tinh chỉnh tham số learning rate trong các giá trị [0.001, 0.01, 0.07, 0.1, 0.4, 0.6, 0.8] và tham số iteration trong các giá trị [1000, 2000, 6000, 8000] để tìm ra cặp tham số tốt nhất. Kết quả đánh giá của mô hình Linear Regression sử dụng phương pháp Gradient Descent được trình bày ở bảng 2.2 dưới đây:

Bảng 2.2. Kết quả đánh giá mô hình Linear Regression (LR)

với phương pháp Gradient Descent (GD)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| learning rate | iteration | R2 train | R2 val | MSE train | MSE val |
| 0.01 | 500 | 0.83 | 0.82 | 530.66 | 584.18 |
| 0.6 | 500 | 0.98 | 0.98 | 48.61 | 55.35 |
| 0.6 | 8000 | 0.99 | 0.98 | 29.9 | 41.85 |

Dựa vào kết quả đánh giá ta thấy được khi tăng learning rate và iteration mô hình cải thiện rất nhiều độ chính xác. Mô hình được chọn với tham số được tinh chỉnh: learning rate là 0.6 và iteration là 8000.

Kết quả mô hình cuối cùng trên tập test ở bảng 2.3 dưới đây:

Bảng 2.3. Kết quả đánh giá mô hình Linear Regression trên tập test

|  |  |  |
| --- | --- | --- |
| Model | R2 test | MSE test |
| Linear Regression with learning rate = 0.6, iteration = 8000 | 0.98 | 39.59 |

Kết quả mô hình rất tốt do việc mô hình Linear Regression hoạt động tốt trên dữ liệu tuyến tính, mô hình sẽ hoạt động không tốt trên dữ liệu phi tuyến tính vì mô hình giả định mối quan hệ tuyến tính giữa biến độc lập và phụ thuộc. Mô hình Linear Regression còn cho biết ảnh hưởng của mỗi đặc trưng đến kết quả dự đoán lượng khí thải CO2 nhờ vào các trọng số của mô hình, trọng số càng lớn càng cho thấy tầm ảnh hưởng của đặc trưng lên lượng khí thải CO2, được biểu diễn ở hình 2.32 dưới đây:

**A screenshot of a computer

Description automatically generated**

Hình 2.32. Khả năng ảnh hưởng của đặc trưng, tính toán bởi Linear Regression

Đặc trưng fuel\_cons\_comb là lượng tiêu thụ nhiên liệu hỗn hợp, đặc trưng quan trọng nhất đúng như dự đoán. Ta quan sát thấy rằng loại nhiên liệu Ethanol(E85) có ảnh hưởng thứ hai và có trọng số âm chứng tỏ sử dụng loại nhiên liệu này giúp giảm lượng khí thải CO2.

### 2.3.2. Mô hình KNN Regression

KNN Regression là thuật toán phi tham số dự đoán quan sát mới bằng cách lấy trung bình các quan sát trong một vùng lân cận dựa vào khoảng cách.

Thuật toán sẽ dự đoán quan sát mới trực tiếp từ dữ liệu huấn luyện bằng cách tính khoảng cách từ điểm cần dự đoán đến tất cả điểm trong tập huấn luyện, tìm k điểm có khoảng cách gần nhất và tính trung bình giá trị của k điểm đó.

Mô hình này có tham số k là quan trọng với hiệu suất của mô hình. Nếu k nhỏ mô hình có xu hướng dự đoán giá trị điểm gần nhất, mô hình sẽ cố khớp với dữ liệu huấn luyện và nhạy cảm với outlier dẫn đến low bias nhưng high variance, mô hình sẽ bị overfitting. Nếu k lớn mô hình giảm ảnh hưởng của outlier đến dự đoán, mô hình sẽ có high bias và low variance dẫn đến underfitting. Vì vậy ta cần chọn tham số k phù hợp cho mô hình để cân bằng variance và bias.

Mô hình được tinh chỉnh tham số trên tập validation trong quá trình huấn luyện để tìm k tốt nhất cho mô hình, là tham số cho MSE nhỏ nhất, tham số k là các giá trị ̣từ 1 đến 30 cho kết quả như hình 2.33 dưới đây:

A graph with a blue line

Description automatically generated

Hình 2.33. Training MSE và Validation MSE

Quá trình tinh chỉnh k đã cho kết quả k có giá trị bằng 1 là tốt nhất, mô hình không bị overfitting hay underfitting và có MSE nhỏ nhất, thường thì k nhận giá trị 1 sẽ gây ra tình trạng overfitting nhưng trong trường hợp bộ dữ liệu của bài toán thì lại cho kết quả tốt. Giải thích cho điều này là bởi vì dữ liệu không thực sự độc lập, các biến độc lập có sự tương quan với nhau dẫn đến kết quả trên mặc dù đã chia dữ liệu train và validation.

Kết quả đánh giá mô hình cuối cùng trên tập test ở bảng 2.4 dưới đây:

Bảng 2.4. Kết quả đánh giá mô hình KNN Regression trên tập test

|  |  |  |
| --- | --- | --- |
| Model | R2 test | MSE test |
| KNN Regression with k=1 | 0.98 | 45.13 |

Nhược điểm của mô hình KNN Regression là không cho biết ảnh hưởng của các đặc trưng lên biến phụ thuộc. KNN Regression hoạt động tốt trên cả dữ liệu phi tuyến tính và tuyến tính vì mô hình không giả định mối quan hệ tuyến tính giữa biến độc lập và phụ thuộc. Mô hình hoạt động kém trong trường hợp k nhỏ vì nhạy cảm với outlier và tốc độ tính toán khoảng cách có thể sẽ làm giảm hiệu suất của mô hình nếu dữ liệu nhiều chiều hoặc tập dữ liệu lớn.

### 2.3.3. Mô hình Support Vector Regression (SVR)

SVR là thuật toán sử dụng nền tảng của SVM để phù hợp với bài toán hồi quy. Trong thuật toán SVM, mô hình cố gắng tìm một siêu phẳng phù hợp nhất để phân tách hai lớp sao cho margin là lớn nhất, margin ở đây là khoảng cách từ siêu phẳng đến các điểm gần nhất ở mỗi lớp.

Thuật toán SVR được tổng quát hóa từ SVM bằng cách thêm vào epsilon-insensitive region xung quanh hàm dự đoán gọi là epsilon-tube. Việc thêm epsilon-tube này dẫn đến bài toán tối ưu là tìm siêu phẳng tốt nhất trong epsilon tube đồng thời giảm thiểu lỗi.

SVR xây dựng bài toán tối ưu hóa bằng cách xác định epsilon-insensitive loss function cực tiểu và tìm ra tube chứa nhiều điểm dữ liệu nhất. Mục tiêu là xây dựng một hàm dự đoán từ loss function được biểu diễn dưới dạng các support vector là các điểm nằm ngoài epsilon-tube. Bài toán tối ưu SVR tìm ra siêu phẳng thường được giải quyết bằng việc giải bài toán đối ngẫu Lagrange cùng hệ điều kiện KKT (\*).

Thuật toán SVR hoạt động tốt với cả dữ liệu tuyến tính và phi tuyến tính nhờ vào phương pháp Kernel, biến đổi dữ liệu gốc phi tuyến tính sang không gian mới mà ở đó dữ liệu gần như tuyến tính. SVR không phù hợp với bộ dữ liệu lớn và nhiều chiều bởi vì mô hình nhiều tham số sẽ dẫn đến mô hình phức tạp làm giảm hiệu suất tổng thể.

Trong báo cáo này, mô hình SVR được sử dụng là Linear SVR, các tham số quan trọng mô hình Linear SVR bao gồm hai tham số epsilon và C là tham số hiệu chỉnh.

Tham số epsilon xác định độ rộng của epsilon-tube, epsilon nhỏ sẽ có nhiều support vector tham gia vào quá trình xây dựng hàm dự đoán hơn dẫn đến low bias nhưng high variance, mô hình dễ bị overfitting. Tham số epsilon lớn sẽ có ít support vector hơn dẫn đến high bias và low variance, mô hình bị underfitting.

Tham số C là tham số hiệu chỉnh để cân bằng giữa độ rộng của epsilon-tube và lỗi dự đoán. C nhỏ mô hình cho phép lỗi lớn và ít support vector hơn, C lớn mô hình cố gắng giảm thiểu lỗi và nhiều support vector hơn.

Trong báo cáo này, mô hình sẽ tinh chỉnh hai tham số C và epsilon với các giá trị khác nhau và chọn ra mô hình có lỗi nhỏ nhất, trong đó % epsilon là phần trăm dữ liệu nằm trong epsilon-tube, còn lại là các điểm support vector. Kết quả đánh giá mô hình SVR được trình bày ở bảng 2.5 dưới đây:

Bảng 2.5. Kết quả đánh giá mô hình SVR

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| C | epsilon | R2 train | R2 val | MSE train | MSE val | % epsilon train | % epsilon val |
| 1 | 0.01 | 0.97 | 0.97 | 74.37 | 83.19 | 49.68 | 48.34 |
| 10 | 0.01 | 0.99 | 0.99 | 33.44 | 26.17 | 51.11 | 50.03 |
| 100 | 0.01 | 0.98 | 0.99 | 36.03 | 28.06 | 50.97 | 49.49 |
| 100 | 1 | 0.98 | 0.98 | 35.63 | 27.7 | 63.27 | 62.69 |
| 100 | 10 | 0.98 | 0.98 | 40.28 | 33.97 | 98.24 | 98.44 |
| 1 | 10 | 0.95 | 0.95 | 142.27 | 129.9 | 82.26 | 83.07 |
| 0.1 | 10 | 0.83 | 0.84 | 570.42 | 489.78 | 62.8 | 63.57 |

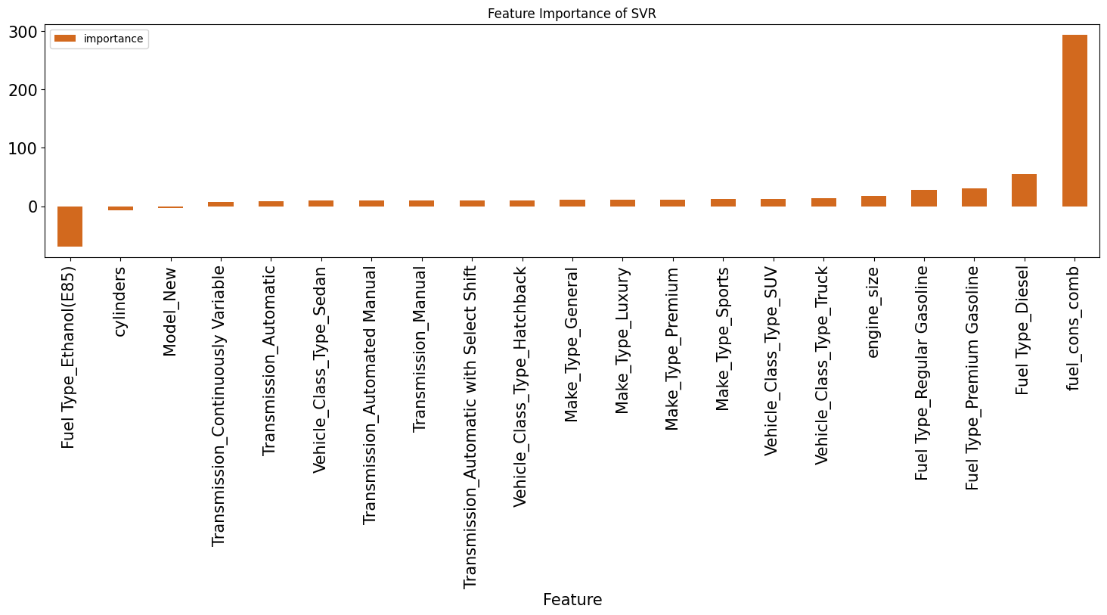
Qua quá trình tinh chỉnh tham số ta thấy được rằng khi epsilon nhỏ và C tăng lên thì MSE giảm xuống nhưng khi C rất lớn đến một mức nào đó mô hình không giảm MSE nữa mà có xu hướng tăng. Khi tăng C đến một giá trị rất lớn ta phải tăng đồng thời epsilon để giảm lỗi mô hình. Trong trường hợp epsilon lớn mà tham số C nhỏ thì sẽ có nhiều điểm dữ liệu trong epsilon- tube hơn nhưng MSE tăng lên.

Chọn ra mô hình có MSE thấp nhất với C: 10 và epsilon: 0.01. Kết quả đánh giá của mô hình trên tập test được trình bày ở bảng 2.6 dưới đây:

Bảng 2.6. Kết quả đánh giá mô hình SVR trên tập test

|  |  |  |  |
| --- | --- | --- | --- |
| Model | R2 test | MSE test | % epsilon test |
| SVR with C=10 and epsilon =0.01 | 0.98 | 34.49 | 48.95 |

Mô hình SVR tính toán ảnh hưởng của đặc trưng lên biến phụ thuộc được biểu diễn ở hình 2.34 dưới đây:



Hình 2.34. Khả năng ảnh hưởng của đặc trưng, tính toán bởi SVR

Ta thấy được mô hình SVR cho kết quả ảnh hưởng của đặc trưng tương tự như mô hình Linear Regression với đặc trưng fuel\_cons\_comb ảnh hưởng nhiều nhất đến lượng khí thải CO2 và loại nhiên liệu Ethanol(E85) giúp làm giảm lượng khí thải CO2.

### 2.3.4. Mô hình Random Forest Regression và Gradient Boosting Regression

Ngoài ba thuật toán trình bày ở trên, bài toán tiếp cận thêm một số thuật toán khác bao gồm hai thuật toán ensemble là Random Forest Regression và Gradient Boosting Regression.

Hai thuật toán được sử dụng đều dựa trên mô hình cây là Decision Tree Regression, với hai cách tiệp cận khác nhau.

Thuật toán Decision Tree Regression xây dựng mô hình dưới dạng cấu trúc cây, chia dữ liệu thành các tập con ngày càng nhỏ, kết quả là một cây có các decision node (nút quyết định) và leaves node (các nút lá). Các nút không phải lá đưa ra quyết định dựa trên giá trị input thông qua câu lệnh logic true false và dựa vào các quyết định này tạo ra đầu ra trong các nút lá bằng cách lấy trung bình giá trị mẫu trong nút.

Việc lựa chọn điều kiện dừng của thuật toán có thể ảnh hưởng đến tính hiệu quả và độ chính xác của mô hình ví dụ như là ngưỡng tiêu chí chia ở mỗi nút, độ sâu của cây, số lượng mẫu tối thiểu ở mỗi leaf node,…

* **Random Forest Regression**:

Thuật toán xây dựng nhiều Decision Tree độc lập với nhau sử dụng các tập dữ liệu con ngẫu nhiên từ dữ liệu gốc cùng với các bộ đặc trưng ngẫu nhiên và lấy trung bình giá trị dự đoán của các Decision Tree làm dự đoán cuối cùng.

Một vài tham số có thể thay đổi để tăng hiệu suất mô hình Random Forest Regression như giới hạn độ sâu của mỗi cây bởi cây càng sâu mô hình càng phức tạp dẫn đến overfitting. Mô hình có thể tăng số lượng Decision Tree để tăng độ chính xác của mô hình tuy nhiên mô hình sẽ huấn luyện lâu hơn. Số lượng đặc trưng ở mỗi lần chia có thể được xem xét, nếu các biến độc lập tương quan cao với nhau ta có thể giảm số lượng đặc trưng ở mỗi lần chia còn biến độc lập không tương quan và mô hình có độ chính xác thấp ta có thể tăng số lượng đặc trưng.

Random Forest Regression có thể hoạt động với cả đặc trưng định lượng và phân loại, xử lý tốt với dữ liệu bị thiếu và hiện tượng các biến độc lập tương quan với nhau cũng như tính toán nhanh trên bộ dữ liệu lớn có nhiều chiều. Tuy nhiên thuật toán có hiệu suất không cao với dữ liệu mất cân bằng và có thể mất nhiều thời gian dự đoán dữ liệu thực tế khi mà số lượng Decision Tree trong mô hình rất lớn.

Trong báo cáo này, mô hình Random Forest Regression được tinh chỉnh tham số để với tham số độ sâu của Decision Tree là max\_depth từ 1 đến 10, số lượng đặc trưng ở mỗi lần chia của cây max\_features trong các giá trị [3, 5, 10, 15, 20], số lượng Decision Tree n\_estimators nằm trong các giá trị [100, 200, 500, 750, 1000]. Mô hình Random Forest Regression cuối cùng với tham số tốt nhất là max\_depth: 9, max\_features: 15, n\_estimators: 200. Kết quả đánh giá cho mô hình Random Forest Regression trên tập test được biểu diễn ở bảng 2.6 dưới đây:

Bảng 2.6. Kết quả đánh giá mô hình Random Forest Regression trên tập test

|  |  |  |
| --- | --- | --- |
| Model | R2 test | MSE test |
| Random Forest Regression with max\_depth = 9, max\_features = 15, n\_estimators = 200 | 0.99 | 9.07 |

* **Gradient Boosting Regression:**

Thuật toán xây dựng các mô hình một cách tuần tự và mô hình sau cố gắng giảm thiểu lỗi của mô hình trước đó, xây dựng mô hình mới dựa trên lỗi hoặc residual của mô hình trước đó.

Mô hình cũng được tinh chỉnh với các tham số của Decision Tree và tham số hiệu chỉnh learning rate. Learning rate nằm trong giá trị từ 0 đến 1, tham số giới hạn ảnh hưởng của mỗi Decision Tree lên giá trị dự đoán cuối cùng, mỗi Decision Tree thêm vào sẽ cải thiện mô hình giúp giảm lỗi, tuy nhiên mô hình Gradient Boosting Regression học rất nhanh nên càng thêm thêm nhiều Decision Tree làm cho mô hình dẫn đến overfitting và learning rate được điều chỉnh để làm giảm tốc độ học của mô hình giúp tránh được overfitting. Tham số learning rate có mối quan hệ chặt chẽ với số lượng Decision Tree trong mô hình, nếu learning rate nhỏ thì phải tăng số lượng Decision Tree để cải thiện hiệu năng của mô hình.

Trong báo cáo này, mô hình Gradient Boosting Regression được tinh chỉnh với hai tham số là learning\_rate và n\_estimators, learning rate nằm trong các giá trị [0.1, 0.4, 0.7, 1] và n\_estimators nằm trong các giá trị [50, 100, 200, 500]. Mô hình tìm ra tham số tốt nhất là learning\_rate: 0.1 và n\_estimators: 500. Kết quả đánh giá cho mô hình Gradient Boosting Regression được biểu diễn ở bảng 2.7 dưới đây:

Bảng 2.7. Kết quả đánh giá mô hình Gradient Boosting Regression trên tập test

|  |  |  |
| --- | --- | --- |
| Model | R2 test | MSE test |
| Gradient Boosting with learning\_rate = 0.1, n\_estimators = 500 | 0.99 | 10.48 |

Mô hình Random Forest Regression và Gradient Boosting Regression tính toán khả năng ảnh hưởng của mỗi đặc trưng lên biến phụ thuộc của bài toán, được biểu diễn ở hình 2.35 và 2.36 dưới đây:

A graph with text and numbers

Description automatically generated

Hình 2.35. Khả năng ảnh hưởng của đặc trưng, tính toán bởi Random Forest Regression

A screen shot of a graph

Description automatically generated  
Hình 2.36. Khả năng ảnh hưởng của đặc trưng, tính toán bởi Gradient Boosting Regression

Ta thấy rằng giống với hai mô hình Linear Regression và SVR đã trình bày ở các phần 2.3.1 và 2.3.3, Random Forest Regression và Gradient Boosting Regression đều cho thấy đặc trưng fuel\_cons\_comb là lượng tiêu thụ nhiên liệu hỗn hợp có ảnh hưởng lớn nhất đến dự đoán lượng khí thải CO2.

### 2.3.5. Kết luận

Kết quả đánh giá của 5 mô hình trên tập test được tổng kết ở bảng 2.8 dưới đây:

Bảng 2.8. Kết quả đánh giá 5 mô hình cuối cùng trên tập test

|  |  |  |
| --- | --- | --- |
| **Model** | **R2 test** | **MSE test** |
| Linear Regression | 0.98 | 39.59 |
| KNN Regression | 0.98 | 45.13 |
| Support Vector Regression | 0.98 | 34.49 |
| Random Forest Regression | 0.99 | 9.07 |
| Gradient Boosting Regression | 0.99 | 10.48 |

Thông qua việc phân tích dữ liệu và quá trình huấn luyện, các mô hình đã xác định rằng đặc trưng lượng tiêu thụ nhiên liệu hỗn hợp có ảnh hưởng lớn nhất đến khả năng dự đoán lượng khí thải CO2 . Mô hình Random Forest Regression đã cho thấy khả năng xử lý tốt với dữ liệu của bài toán nhờ vào khả năng kết hợp nhiều mô hình, Random Forest Regression cũng có hiệu suất dự đoán tốt hơn so với các thuật toán khác được thử nghiệm. Bài toán hướng đến việc dự đoán lượng khí thải chính xác nhất có thể vậy nên mô hình với tiêu chí MSE thấp hơn được lựa chọn. Kết quả của nghiên cứu cung cấp một cơ sở cho việc đưa ra dự đoán lượng khí thải để kiểm soát lượng khí thải tốt hơn qua đó làm giảm ô nhiễm môi trường, trong đó lượng tiêu thụ nhiên liệu có thể được quan tâm để đưa ra những giải pháp nhằm giảm thiểu lượng khí thải CO2.

# 3. Tổng kết

Báo cáo trên cung cấp một cách nhìn tổng quan về Machine Learning và quy trình của một mô hình Machine Learning, được áp dụng vào bài toán dự đoán lượng khí thải CO2 từ phương tiện. Đầu tiên, báo cáo đề cập đến các bước tiền xử lý dữ liệu sau đó trình bày về các thuật toán được sử dụng để xây dựng mô hình, bao gồm Linear Regression, KNN Regression, Support Vector Regression và hai thuật toán ensemble là Random Forest Regression cùng với Gradient Boosting Regression. Các kết quả và hiệu suất của mô hình được trình bày bằng bảng và biểu đồ.

# 4.Tài liệu tham khảo

(\*) DIPLOMARBEIT Support Vector Machines for Regression Estimation and their Application to Chaotic Time Series Prediction <https://publik.tuwien.ac.at/files/pub-et_9604.pdf>