BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN MÔN HỌC: MÁY HỌC VÀ CÔNG CỤ

ĐỀ TÀI: CO2 EMISSIONS

Giảng viên hướng dẫn:

Ts. Nguyễn Tấn Trần Minh Khang

Ths. Quan Chí Khánh An

Sinh viên thực hiện:

Nguyễn Phúc Bình	21520638
Tô Thế Kiệt	21522263

Thành phố Hồ Chí Minh, 19 tháng 12 năm 2023

LỜI CẢM ƠN

Lời đầu tiên chúng em xin chân thành gửi lời cảm ơn đến với giảng viên môn học là thầy Nguyễn Tấn Trần Minh Khang và thầy Quan Chí Khánh An. Cảm ơn thầy đã giảng dạy, truyền đạt rất nhiều kiến thức bổ ích về bộ môn và được trang bị thêm nhiều kỹ năng mới vô cùng cần thiết cho ngành học. Cảm ơn thầy đã hướng dẫn nhiệt tình, giải đáp thắc mắc cũng như hỗ trợ chúng em rất nhiều trong quá trình thực hiện đồ án.

Với khả năng và thời gian có hạn nên không thể tránh khỏi những thiếu sót, chúng em rất mong được sự quan tâm, giúp đỡ và thông cảm của thầy để chúng em hoàn thiện hơn về đồ án của nhóm mình.

Một lần nữa, chúng em xin chân thành cảm ơn!

Thành phố Hồ Chí Minh, tháng 12 năm 2023

Nhóm sinh viên thực hiện

Nội dung

Lời cảm ơn 2
Chương 1: Overview.....5
1. Hiện trạng: 5
2. Giới thiệu bài toán 5
3. Khả năng ứng dụng trong thực tế:.....9
Chương II. Background and Related work 10
1. Các mô hình sử dụng thuật toán giải quyết các vấn đề trong thực tế có liên quan tới đề tài đồ án.....10
1.1. Xây dựng mô hình cảnh báo, dự báo theo phương pháp học máy có giảm sát thử nghiệm dự báo xâm ngập mặn cho lưu vực sông Hậu.....10
1.2. Application of machine learning initiatives and intelligent perspectives for CO2 emission reduction in construction.....11
2. Khảo sát hướng đi.....13
2.1.1. Các phương pháp Machine Learning.....13
2.1.2. Phân tích tổng quan một lượt các thuật toán đã được học và thực hành trong bộ môn.....14
2.1.3. Tiến hành thử nghiệm tìm ra thuật toán tối ưu nhất trên bài toán 18
Chương 3: Approach.....26
1. Đặc điểm của dữ liệu 26
2. Giải quyết missing bằng fill data và xếp theo mức độ quan trọng.....28
Chương 4: Results.....30
1. Đánh giá kết quả đầu ra của thuật toán đã sử dụng trong đồ án.....30
Chương 5: Conclusion 33
Chương 6: Reference 33

TÓM TẮT ĐỒ ÁN

Đồ án thực hiện đề tài dự đoán khối lượng khí thải CO2 dựa theo các dữ liệu được cung cấp ở tập dữ liệu huấn luyện. Sau khi huấn luyện mô hình, kết quả thu được là lượng CO2 thải ra ở kinh độ độ, vĩ độ ở thời điểm các tuần trong năm 2022. Để đạt được kết quả dự đoán gần chính xác nhất mô hình sử dụng tập dữ liệu huấn luyện đạt chuẩn và được sàng lọc kĩ càng và chặt chẽ với gần 73 trường thuộc tính.

Mô hình sử dụng thuật toán **Random forest**. Đây là một thuật toán máy học được sử dụng cho cả bài toán phân loại và dự đoán. Nó là mô hình dựa trên nguyên tắc của "ensemble learning" (học tập từ đoàn tụ), trong đó có nhiều cây quyết định được tạo ra và kết hợp để tạo ra một mô hình mạnh mẽ và ổn định hơn.

Bên cạnh đó mô hình chia tập huấn luyện và tập kiểm thử theo tỉ lệ 7:3. Đây là sự lựa chọn phổ biến và hợp lý trong trường hợp của bài toán này theo sự đánh giá của nhóm. Việc chia tỉ lệ như vậy giúp dễ dàng kiểm thử độ chính xác của mô hình. Ngăn chặn Overfitting khi mô hình quá thích ứng với dữ liệu đào tạo và không thể tổng quát hóa tốt cho dữ liệu mới. Tập kiểm thử đại diện cho một tập dữ liệu độc lập, giúp đánh giá hiệu suất tổng thể của mô hình dựa trên dữ liệu thực tế. Hơn thế nữa, việc chia tỉ lệ còn giúp tối ưu hóa các tham số. Khi đào tạo mô hình người ta thường điều chỉnh các tham số để tối ưu hóa hiệu suất trên tập kiểm thử. Sau đó tập kiểm thử được sử dụng để xác nhận rằng mô hình không bị "over-tuned" (quá mức tinh chỉnh) và vẫn hoạt động tốt trên dữ liệu mới.

Nhóm thực hiện đề tài và giải quyết với bài toán mong muốn mô hình đạt được kết quả tốt nhất và hy vọng mô hình có thể tiệm cận với việc ứng dụng thực tế bởi lẽ việc dự đoán được lượng khí thải CO2 mang nhiều lợi ích quan trọng, đặc biệt là trong ngữ cảnh của biến đổi khí hậu và bảo vệ môi trường. Từ đó có thể hiệu ra rằng việc tác động của con người tới môi trường, vấn đề quản lý năng lượng và tài nguyên, phát triển chính sách bảo vệ môi trường quan trọng như thế nào trong bối cảnh đất nước, doanh nghiệp đẩy mạnh công nghiệp, đẩy mạnh sản xuất đa dạng hàng hóa.

Chương 1: Overview

1. Hiện trạng:

Hiện nay, thực tế cho thấy biến đổi khi hậu và mức lượng khí thai CO2 đang ở mức lo ngại và đang gây ảnh hưởng toàn cầu. Nghiên cứu mới cho thấy nồng độ CO2 trong khí quyển vào tháng 5 đã đạt ngưỡng 420 ppm, cao hơn 50% so với thời kỳ tiền công nghiệp. Đây là nồng độ cao chưa từng thấy trong khoảng 4 triệu năm qua, Cơ quan Khí quyển và Đại dương Quốc gia Mỹ (NOAA) hôm 03/6/2022 nhấn mạnh. Các phép đo được thực hiện tại Đài quan sát Mauna Loa ở Hawaii, nơi lý tưởng nhất nằm trên một ngọn núi lửa, cho phép nó thoát khỏi ảnh hưởng có thể có của ô nhiễm cục bộ. Trước cuộc "cách mạng công nghiệp", nồng độ carbon dioxide (CO2) đã duy trì ổn định ở mức 280 phần triệu (ppm) trong khoảng 6.000 năm văn minh nhân loại. Mức 420 ppm hiện tại có thể so sánh với nồng độ ước tính trên 400 ppm cách đây 4,1 đến 4,5 triệu năm. Vào thời điểm đó, mực nước biển cao hơn bây giờ từ 5 đến 25 m, đủ cao để nhấn chìm nhiều thành phố lớn hiện nay. [1]

Nồng độ CO2 quá cao trong nhà có thể cản trở suy nghĩ của chúng ta và thậm chí có thể gây nguy hiểm lớn hơn cho sức khỏe con người. Theo một nghiên cứu về 24 nhân viên, mức độ nhận thức thấp hơn 50% khi những người tham gia tiếp xúc với 1.400ppm CO2 so với 550ppm trong một ngày làm việc [2]

2. Giới thiệu bài toán

Bài toán CO2 Emissions, bài toán dự đoán nồng độ CO2 trong không khí. Dựa theo các dữ liệu được cung cấp trong tập huấn luyện. Sau khi huấn luyện mô hình, kết quả thu được là lượng CO2 thải ra ở kinh độ độ, vĩ độ ở thời điểm các tuần trong năm 2022. Để đạt được kết quả dự đoán gần chính xác nhất mô hình sử dụng tập dữ liệu huấn luyện đạt chuẩn và được sàng lọc kĩ càng và chặt chẽ với gần 73 trường thuộc tính.

Bảng các trường thuộc tính input. Valid và missing của nó trong tập huấn luyện:

STT	Name	Valid	Missing
1	latitude	100%	0%
2	longtitude	100%	0%
3	year	100%	0%
4	week_no	100%	0%
5	SulphurDioxide_SO2_column_number_density	82%	18%
6	SulphurDioxide_SO2_column_number_density_amf	82%	18%
7	SulphurDioxide_SO2_slant_column_number_density	82%	18%
8	SulphurDioxide_cloud_fraction	82%	18%
9	SulphurDioxide_sensor_azimuth_angle	82%	18%
10	SulphurDioxide_sensor_zenith_angle	82%	18%
11	SulphurDioxide_solar_azimuth_angle	82%	18%
12	SulphurDioxide_solar_zenith_angle	82%	18%
13	SulphurDioxide_SO2_column_number_density_15km	82%	18%
14	CarbonMonoxide_CO_column_number_density	97%	3%
15	CarbonMonoxide_H2O_column_number_density	97%	3%
16	CarbonMonoxide_cloud_height	97%	3%
17	CarbonMonoxide_sensor_altitude	97%	3%
18	CarbonMonoxide_sensor_azimuth_angle	97%	3%
19	CarbonMonoxide_sensor_zenith_angle	97%	3%
20	CarbonMonoxide_solar_azimuth_angle	97%	3%
21	CarbonMonoxide_solar_zenith_angle	97%	3%
22	NitrogenDioxide_NO2_column_number_density	77%	23%
23	NitrogenDioxide_tropospheric_NO2_column_number_density	77%	23%
24	NitrogenDioxide_stratospheric_NO2_column_number_density	77%	23%
25	NitrogenDioxide_NO2_slant_column_number_density	77%	23%
26	NitrogenDioxide_tropopause_pressure	77%	23%
27	NitrogenDioxide_absorbing_aerosol_index	77%	23%
28	NitrogenDioxide_cloud_fraction	77%	23%

29	NitrogenDioxide_sensor_altitude	77%	23%
30	NitrogenDioxide_sensor_azimuth_angle	77%	23%
31	NitrogenDioxide_sensor_zenith_angle	77%	23%
32	NitrogenDioxide_solar_azimuth_angle	77%	23%
33	NitrogenDioxide_solar_zenith_angle	77%	23%
34	Formaldehyde_tropospheric_HCHO_column_number_density	91%	9%
35	Formaldehyde_tropospheric_HCHO_column_number_density_amf	91%	9%
36	Formaldehyde_HCHO_slant_column_number_density	91%	9%
37	Formaldehyde_cloud_fraction	91%	9%
38	Formaldehyde_solar_zenith_angle	91%	9%
39	Formaldehyde_solar_azimuth_angle	91%	9%
40	Formaldehyde_sensor_zenith_angle	91%	9%
41	Formaldehyde_sensor_azimuth_angle	91%	9%
42	UvAerosolIndex_absorbing_aerosol_index	99%	1%
43	UvAerosolIndex_sensor_altitude	99%	1%
44	UvAerosolIndex_sensor_azimuth_angle	99%	1%
45	UvAerosolIndex_sensor_zenith_angle	99%	1%
46	UvAerosolIndex_solar_azimuth_angle	99%	1%
47	UvAerosolIndex_solar_zenith_angle	99%	1%
48	Ozone_O3_column_number_density	99%	1%
49	Ozone_O3_column_number_density_amf	99%	1%
50	Ozone_O3_slant_column_number_density	99%	1%
51	Ozone_O3_effective_temperature	99%	1%
52	Ozone_cloud_fraction	99%	1%
53	Ozone_sensor_azimuth_angle	99%	1%
54	Ozone_sensor_zenith_angle	99%	1%
55	Ozone_solar_azimuth_angle	99%	1%
56	Ozone_solar_zenith_angle	99%	1%
57	UvAerosolLayerHeight_aerosol_height	1%	99%
58	UvAerosolLayerHeight_aerosol_pressure	1%	99%

59	UvAerosolLayerHeight_aerosol_optical_depth	1%	99%
60	UvAerosolLayerHeight_sensor_zenith_angle	1%	99%
61	UvAerosolLayerHeight_sensor_azimuth_angle	1%	99%
62	UvAerosolLayerHeight_solar_azimuth_angle	1%	99%
63	UvAerosolLayerHeight_solar_zenith_angle	1%	99%
64	Cloud_cloud_fraction	99%	1%
65	Cloud_cloud_top_pressure	99%	1%
66	Cloud_cloud_top_height	99%	1%
67	Cloud_cloud_base_pressure	99%	1%
68	Cloud_cloud_base_height	99%	1%
69	Cloud_cloud_optical_depth	99%	1%
70	Cloud_surface_albedo	99%	1%
71	Cloud_sensor_azimuth_angle	99%	1%
72	Cloud_sensor_zenith_angle	99%	1%
73	Cloud_solar_azimuth_angle	99%	1%
74	Cloud_solar_zenith_angle	99%	1%
75	emission	100%	0%

Bảng 1: Các trường thuộc tính đầu vào

Dựa trên những dữ liệu đầu vào như trên, mô hình tiến hành xử lý và tính toán. Cuối cùng cho ra kết quả. Ví dụ được thể hiện dưới bảng sau:

ID_LAT_LON_YEAR_WEEK	Emission
ID_-0.510_29.290_2022_00	3.547523416
ID_-0.510_29.290_2022_01	4.01650905
ID_-0.510_29.290_2022_02	4.145595447
ID_-0.510_29.290_2022_03	4.13600204
ID_-0.510_29.290_2022_04	4.25215841
ID_-0.510_29.290_2022_05	4.247781777

Bảng 2: Dữ liệu đầu ra

ID_LAT_LON_YEAR_WEEK, đây là trường thuộc tính đầu tiên mà kết quả cho ra. Cụ thể phân tích như sau:

Tên	Mô tả
LAT	Vĩ độ - LATIDUTE
LON	Kinh độ - LONGIDUTE
YEAR	Năm, cụ thể ở bài toán này tính toán lượng CO2 năm 2022
WEEK	Tuần, một năm có sấp sỉ 52 tới 53 tuần tùy năm nhuận hay không nhuận
Emission	Nồng độ CO2

Bảng 3: Giải thích dữ liệu đầu ra

3. Khả năng ứng dụng trong thực tế:

Nhìn chung, trong đồ án nhóm đã giải quyết được yêu cầu đặt ra của bài toán và mô hình đã có thể tính toán và dự đoán một cách khá chỉn chu và điểm số đạt được qua các công cụ đo, đánh giá mô hình ở mức chấp nhận được.

Tuy nhiên, mô hình mới chỉ dừng ở phạm vi môn học và nghiên cứu nhỏ lẻ, vì vậy để mang ứng dụng vào thực tế thì tính khả thi và hữu dụng vẫn còn phải xem xét qua nhiều yếu tố khác như:

- **Yếu tố dữ liệu**: Hiện tại mô hình đang chạy tốt trên tập dữ liệu được lọc và cho sẵn như thế, nhưng câu hỏi đặt ra là liệu với lượng dữ liệu lớn hơn, nhiều trường thuộc tính đầu vào hơn thì liệu mô hình có còn giữ được kết quả như bây giờ hay sẽ có những kết quả khác?
- **Khả năng tích hợp**: Liệu mô hình có khả năng tích hợp với các hệ thống và quy trình trọng thực tế hay không?
- **Mức độ linh hoạt**: Mô hình phải có khả năng biến đổi linh hoạt, thích ứng với sự biến đổi và cập nhật trong môi trường hiện nay.
- **Độ tương thích công nghệ**: Mô hình phải đảm bảo tích hợp được vào các công nghệ mới và xu hướng công nghệ trong tương lai

Ngoài việc tự nghiên cứu và thử nghiệm, nhóm còn tham khảo các bài toán liên quan cũng như các trang thông tin, bài báo nước ngoài để khảo sát xem hiện tại trên thế giới đã có ai làm về vấn đề này chưa, họ sử dụng mô hình như thế nào và cách mô hình đó được đưa vào ứng dụng thực tế. Sau khi đánh giá và phân tích, nhóm đánh giá hiện tại mô hình được sử dụng trong đồ án môn học vẫn còn thiếu tính khả thi, chưa thể mang áp dụng vào thực tế.

Chương II. Background and Related work

1. Các mô hình sử dụng thuật toán giải quyết các vấn đề trong thực tế có liên quan tới đề tài đồ án.

1.1. Xây dựng mô hình cảnh báo, dự báo theo phương pháp học máy có giảm sát thử nghiệm dự báo xâm ngập mặn cho lưu vực sông Hậu.

- Mô hình học máy có giám sát *(supervised learning):*

  Mô hình học có giám sát là mô hình học trên dữ liệu có dãn nhãn, tức là mục tiêu của bài toán machine learning cần học đã được gán nhãn sẵn trong dữ liệu huấn luyện. Dữ liệu đầu vào của quá trình học bao gồm cả vector đầu vào chứa các thuộc tính của dữ liệu lẫn giá trị đầu ra mục tiêu (gọi là nhãn của dữ liệu). Nói cách khác, học máy có giám sát cho phép dự đoán đầu ra của một dữ liệu mới dựa trên các cặp (đầu vào, đầu ra) đã biết từ trước thu được từ bộ dữ liệu huấn luyện. Bộ dữ liệu huấn luyện bao gồm các cặp (data, label), tức (dữ liệu, nhãn). Chẳng hạn, bộ dữ liệu hoa tử đằng (Iris) chứa các thuộc tính là chiều dài và chiều rộng của cánh hoa và đài hoa, các thuộc tính này tạo thành dữ liệu đầu vào (data). Đồng thời, nó cũng chứa cả nhãn (class label) của mục tiêu dự đoán (dòng hoa là một trong ba loại: Setosa, versicolor và virginica) [3]
- Phương pháp dự báo chuỗi thời gian ARIMA [3]
- Tiền xử lý, lọc dữ liệu: Một trong những nhiệm vụ quan trọng trong việc ứng dụng mô hình học máy đó là vấn đề tiền xử lý, lọc dữ liệu. Việc làm này nhằm

loại bỏ những số liệu không thực sự tin cậy, đồng thời đưa chuỗi số liệu theo cùng một cấu trúc, định dạng và thời đoạn theo thời gian. Dưới đây là một số bước trong quá trình tiền xử lý, lọc dữ liệu cho mô hình học máy được thực hiện trong nền tảng Brightics AI: *Xác định thư mục chứa dữ liệu cho mô hình học máy; khai báo cấu hình các trường dữ liệu cho mô hình học máy; lựa chọn các thành phần dữ liệu cho mô hình học máy*. [3]

Đánh giá: *Mô hình trên sử dụng phương pháp dự báo chuỗi thời gian ARIMA và mô hình hồi quy tuyến tính. Sau khi nghiên cứu kĩ nhóm sẽ cân nhắc thử ứng dụng thuật toán vô giải quyết vấn đề của đồ án*

1.2. Application of machine learning initiatives and intelligent perspectives for CO2 emission reduction in construction

- Phương pháp: Phương pháp đánh giá này được thiết kế theo hai câu hỏi nghiên cứu:
  - AI và ML được được xử dụng trong xây dựng để giảm lượng khí thải CO2 theo những hướng nào?,
  - Những kỹ thuật AI và ML nào được sử dụng để nghiên cứu giảm lượng khí thải?
- Một đánh giá có hệ thống được tiến hành để tìm ra các bài viết có liên quan giải quyết những câu hỏi này. Phương pháp thu thập và lọc tài liệu của các bài báo là một cách tiếp cận rộng rãi, từ trên xuống, thu thập các bài viết sử dụng các cụm từ rộng trước tiên và sau đó loại trừ những bài viết có thể không liên quan đến nghiên cứu. Phương pháp này bao gồm những bước sau:

  (1) Phân loại các các câu hỏi và loại trừ

(2) Thu thập các bài viết liên quan thông qua tìm kiếm, sàng lọc có hệ thống các bài viết liên quan đến tiêu chí đưa vào và loại trừ

(3) Phân tích các bài viết quan trọng, các ấn phẩm học tập.

[4]

Sustainable materials design and production using ML techniques [4]

Authors	Sustainable Components	Soft Computing Method	Output
Lee et al. (2021)	Waffle slabs	NSGA-II algorithm	- Reduced CO_2 emissions and costs - Waffle forms specifications, particularly the ribs height and the distance between ribs, among the design parameters, had the highest impact on optimizing cost and CO_2 emissions - Regarding NSGA-II, the maximum generation was set to 40, and the crossover and mutation rates were 0.95 and 0.05, respectively
Zhang and Zhang (2021)	RC beams	Multi-objective genetic algorithm	- The results showed that a 6% additional construction cost could make up for a 14.7% emission reduction
Lanikova et al. (2014)	Structure design	Reliability-based structure optimization	- 8.9% of the cost and 11.1% of emissions were reduced in comparison to conventional design processes
Gan et al. (2019)	High-rise RC structure	Optimality criteria genetic algorithm	- CO_2 emissions and cost of materials decreased by 18–24%
Park et al. (2013)	Steel RC columns	Genetic algorithm-based optimization	- The proposed optimized design could decrease 30.3% of CO_2 emissions, 31.5% of the cost, and 7.8% of steel section and total concrete weights

Để xử lý được bài toán, rất nhiều chuyên gia trên thế giới đã tiến hành thử nghiệm với nhiều mô hình học máy với nhiều phương thức khác nhau cũng như chọn lọc những nguồn dữ liệu đáng tin cậy và độ khả thi cao nhất.

2. Khảo sát hướng đi

2.1.1. Các phương pháp Machine Learning

- **Supervised Learning**: Là một kỹ thuật học máy được sử dụng trong sản xuất để đào tạo các mô hình được sử dụng dữ liệu được nhãn dán. Trong sản xuất, Supervised learning có thể được sử dụng để phân loại lỗi, xác định các thông số sản xuất và dự đoán thời gian sử dụng hữu ích còn lại của thiết bị
- **Unsupervised Learning**: Là một kỹ thuật học máy được sử dụng trong sản xuất để đào tạo các mô hình sử dụng dữ liệu chưa được gán nhãn. Phương pháp này dựa vào dữ liệu được nhãn dán và các kỹ thuật phân cụm để xác định các mẫu và cấu trúc trong dữ liệu. Trong sản xuất, unsupervised learning có thể được sử dụng để xác định các nhóm sản phẩm tương tự hoặc xác định các khiếm khuyết khó phát hiện.
- **Deep learning**: là một tập hợp con của học máy sử dụng mạng lưới thần kinh để mô hình hóa các mẫu quan hệ phức tạp giữa các thông số sản xuất và chất lượng sản phẩm. Nó cũng có thể được sử dụng để phân tích lượng lớn dữ liệu từ cảm biến và các nguồn khác nhằm dự đoán lỗi thiết bị.
- **Reinfocrcement learning**: Là một kỹ thuật học máy được sử dụng trong sản xuất để dạy các mô hình đưa ra quyết định dựa trên phản hồi từ môi trường. Trong sản xuất, Reinforce learning có thể được sử dụng để tối ưu hóa quy trình sản xuất và cải thiện việc bảo trì thiết bị.

Sau quá trình tìm hiểu và đánh giá. Cũng như phân tích lượng data được cung cấp sẵn để tiến hành thực thi đồ án môn học. Nhóm chọn phương pháp Supervised Learning để tiến hành đồ án lần này.

2.1.2. Phân tích tổng quan một lượt các thuật toán đã được học và thực hành trong bộ môn.

Lưu ý: Thông số và đánh giá được sử dụng lại qua các bài toán thực hành trong bộ môn, đã nộp trên hệ thống moodle trước đó. Đây không phải đánh giá thuật toán cho đồ án, chỉ là sự tham khảo để quyết định cho hướng đi cuối cùng cho đồ án này.

- **Hồi quy SVR**

Ưu điểm	Nhược điểm
Hiệu suất tốt với dữ liệu phi tuyến tính	Khó điều chỉnh nếu sử dụng nhiều hàm kernel phức tạp
Điều chỉnh linh hoạt	Tính toán phức tạp
Khả năng xử lý dữ liệu hiệu quả	Không thể áp dụng trực tiếp cho nhiều lớp có nhiều đầu ra.

⇒ Nhìn chung, hồi quy SVR là một công cụ mạnh mẽ cho các bài toán hồi quy, đặc biệt là khi dữ liệu có tính phi tuyến hoặc chứa nhiễu. Tuy nhiên việc điều chỉnh siêu tham số và xử lý dữ liệu lớn có thể đòi hỏi một số kiến thức chuyên sâu và tài nguyên tính toán

- **Hồi quy rừng quyết định**

Ưu điểm	Nhược điểm
Khả năng xử lý dữ liệu phức tạp mà không cần nhiều tiền xử lý hay chuẩn hóa dữ liệu	Phụ thuộc vào số lượng cây quyết định. Việc sử dụng nhiều cây làm tăng thời gian tính toán
Khả năng xử lý dữ liệu thiếu bằng cách sử dụng giá trị trung bình hoặc các kỹ thuật khác để điền giá trị thiếu	Khó hiểu. Kết quả từ mô hình khó giải thích như một mô hình tuyến tính.
Tính linh hoạt và hiệu suất tốt, thích nghi tốt với dữ liệu.	Không thể xử lý mối quan hệ phức tạp giữa các đặc trưng.
Điều chỉnh dễ dàng. Có thể điều chỉnh các siêu tham số như số cây quyết định, độ sâu của cây…	

⇒ *Nhìn chung, Rừng quyết định là một phương pháp hữu ích cho hồi quy, đặc biệt khi đối diện với dữ liệu phức tạp và cần một phương pháp xử lý nhiễu.*

- **Cây quyết định:**

Ưu điểm	Nhược điểm
Dễ hiểu và giải thích, dễ dàng trực quan hóa cấu trúc cây quyết định và làm rõ cách mô hình đưa ra dự đoán	Dễ bị overfitting
Phù hợp cho dữ liệu có đặc trưng hỗn hợp	Khả năng chọn bias
Khả năng xử lý dữ liệu bị thiếu	Không hiệu quả với các mối quan hệ phức tạp
Tốc độ tính toán nhanh	Khả năng chịu ảnh hưởng của nhiễu

⇒ Nhìn chung, Cây quyết định là một công cụ mạnh, tuy nhiên vẫn phải cân nhắc để đưa ra những lựa chọn phụ thuộc khi train model

- **Hồi quy đa thức:**

Ưu điểm	Nhược điểm
Phù hợp cho mối quan hệ phi tuyến tính	Dễ xảy ra overfitting
Tính linh hoạt	Khó điều chỉnh
Dễ triển khai	Khả năng biểu diễn hạn chế
Dễ hiểu	Nhạy cảm với nhiễu

⇒ Nhìn chung, để dự đoán mối quan hệ giữa biến độc lập và biến phụ thuộc thì hồi quy đa thức khá phù hợp. Hữu ích trong mối quan hệ phi tuyến tính. Tuy nhiên cần cân nhắc bậc của đa thức và quản lý overfitting

- **K-NN**

Ưu điểm	Nhược điểm
Dễ hiểu và triển khai	Nhạy cảm với dữ liệu nhiễu
Không yêu cầu huấn luyện phức tạp	Không hiệu quả với dữ liệu có nhiều chiều
Đặc biệt tốt cho dữ liệu có cấu trúc đơn giản	

⇒ Nhìn chung, k-NN là một trong những phương pháp đơn giản và dễ triển khai trong machine learning. Có thể áp dụng trong các tập dữ liệu đơn giản. Tuy nhiên cần phải cần nhắc đối với bài toán cần giải quyết và đưa ra lựa chọn phù hợp

- **Logistic Regression:**

Ưu điểm	Nhược điểm
Dễ hiểu và diễn giải	Yêu cầu biến độc lập không phụ thuộc nhau
Hiệu suất tốt trên dữ liệu tuyến tính hoặc có mối quan hệ tuyến tính	Không phù hợp với dữ liệu phi tuyến
Phân phối xác suất	Dễ bị ảnh hưởng bởi dữ liệu nhiễu
Thích hợp cho bài toán phân loại nhị phân	Khi dữ liệu quá phức tạp hoặc không tuân theo đặc điểm tuyến tính, Logistic Regression có thể dễ dàng bị underfitting.

⇒ Logistic Regression là một mô hình quan trọng và được sử dụng rộng rãi trong machine learning. Nó đơn giản, dễ hiểu và hiệu quả trong nhiều trường hợp, đặc biệt là khi dữ liệu có tính tuyến tính. Tuy nhiên, cần lưu ý nhược điểm của nó, đặc biệt là khi đối mặt với dữ liệu phi tuyến hoặc có tương quan cao. Quyết định sử dụng Logistic Regression cần dựa trên đặc điểm cụ thể của dữ liệu và mục tiêu của bài toán.

Sau khi tiến hành phân tích, tìm hiểu và đánh giá từng thuật toán, nhóm nhận ra những ưu điểm và nhược điểm của từng thuật toán. Cuối cùng, nhóm tiến hành thử nghiệm từng chút một trên bài toán đối với những thuật

2.1.3. Tiến hành thử nghiệm tìm ra thuật toán tối ưu nhất trên bài toán

- **Thử nghiệm với thuật toán Linear Regression:**

Hồi quy tuyến tính được sử dụng để dự đoán giá trị của một biến dựa trên giá trị của biến khác. Biến muốn dự đoán gọi là biến phụ thuộc. Biến đang sử dụng để dự đoán giá trị của biến khác gọi là biến độc lập.

Mục tiêu tối ưu của Linear Regression là tìm giá trị B0 và B1 sao cho tổng bình phương của sai số là nhỏ nhất. Điều này thường được thực hiện bằng cách sử dụng phương pháp bình phương tối thiểu (Least Squares) [5]

Dưới đây là kết quả mô hình đạt được, điểm số của mô hình được thể hiện ở cột score, bên cạnh đó đi kèm còn có thời gian chạy với từng fill data khác nhau.

Fill Data	Score	Time
Mean	0.0244	0.415s
0	0.023	0.858s
-1	0.0241	0.84s

Trước khi tiến hành train cũng như giải bài toán, chúng ta cần tiến hành xử lý missing data, hay còn gọi cách khác là làm sạch dữ liệu. Xử lý missing data là một phần quan trọng, có nhiều công cụ mạnh hiện nay như pandas.

- **Thử nghiệm với thuật toán Decision tree**

Cây quyết định là mô hình machine learning được sử dụng trong các nhiệm vụ phân loại và dự đoán. Nó là một cấu trúc dữ liệu cây, trong đó mỗi nút đại diện cho một quyết định hoặc một bài kiểm tra trên một thuộc tính, mỗi cạnh kết nối giữa các nút đại diện cho kết quả của một bài kiểm tra và các nút con của nó.

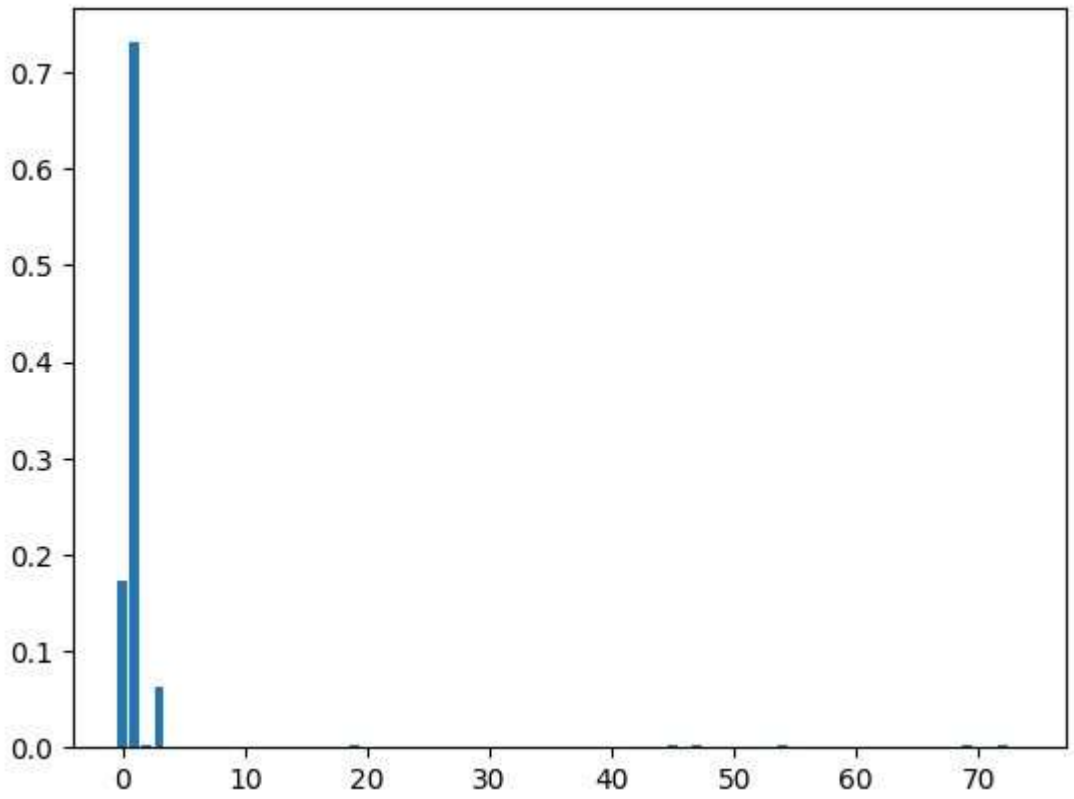
Mục tiêu của cây quyết định là phân loại dữ liệu dựa trên các thuộc tính. Quyết định về việc chia dữ liệu được thực hiện bằng cách lựa chọn thuộc tính và giá trị cụ thể của thuộc tính đó để tạo ra các nhánh.

Ưu điểm của cây quyết định là dễ hiểu và dễ giải thích. Cây quyết định tạo ra các quy tắc quyết định rõ ràng và dễ hiểu. Đặc biệt hơn vì cây quyết định không hoặc ít bị ảnh hưởng nhiều bởi các biến độc lập có tỷ lệ khác nhau.

Tiến hành thuật toán trong đồ án:

Dưới đây là kết quả mô hình đạt được, điểm số của mô hình được thể hiện ở cột score, bên cạnh đó đi kèm còn có thời gian chạy với từng fill data khác nhau.

Fill Data	Score	Time
Mean	0.956	7.59s
-1	0.957	8.007s
0	0.957	8.592S



Hình 1: Biểu đồ mức độ quan trọng của các featured

Sau khi tiến hành fill data, thấy mean cho score và time phù hợp, tiến hành lọc các featured theo mức độ quan trọng để cho ra kết quả tốt hơn.

Kết quả xảy ra khi áp dụng decision tree khi chưa lọc cột:

Chưa lọc cột	Score	Time
	0.9565	8s

Vì có nhiều trường giá trị không quá quan trọng tới mô hình, nhóm tiến hành lọc các trường giá trị có độ quan trọng nhỏ. Thử lần đầu tiên với độ quan trọng bé hơn 0.02

Số lượng featured bị xóa	Score	Time
71	0.9715	1s

Thử lần tiếp theo với độ quan trọng bé hơn 0.002

Số lượng featured bị xóa	Score	Time

64	0.9716	1.6s

Thử với độ quan trọng bé hơn 0.003

Số lượng featured bị xóa	Score	Time
70	0.967	1.12s

Thử với độ quan trọng bé hơn 0.0003

Số lượng featured bị xóa	Score	Time
46	0.9609	4.19s

Thử với độ quan trọng bé hơn 0.065

Số lượng featured bị xóa	Score	Time
72	0.900	0.31s

- **Thử nghiệm với thuật toán XGBoost**

XGBoost thuộc loại thuật toán Gradient Boosting, nơi nó xây dựng một chuỗi các cây quyết định theo cách tuần tự. Mỗi cây cố gắng sửa sai số của cây trước đó. XGBoost có Regularizarion để kiểm soát độ phức tạp của mô hình và giảm nguy cơ quá mức (Overfitting).

XGBoost sử dụng tốc độ học để kiểm soát mức độ cập nhật của các trọng số sau mỗi cây. Điều này giúp kiểm soát Overfitting.

Tiến hành thuật toán trong đồ án:

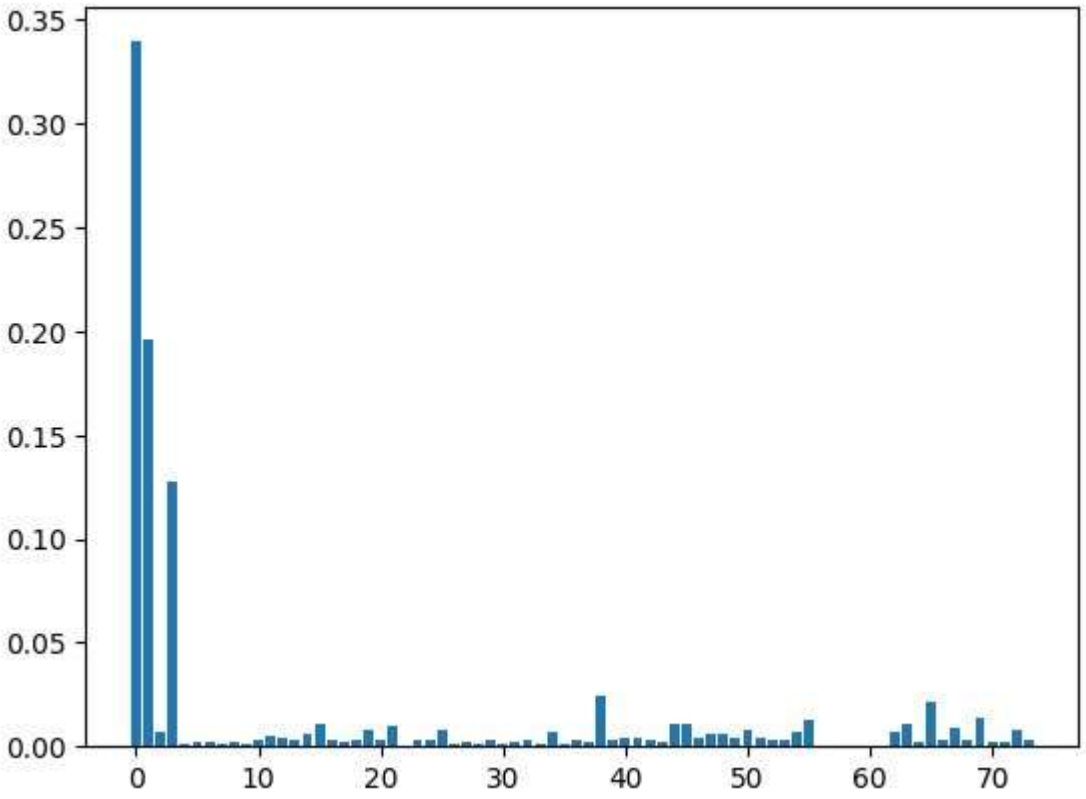
Dưới đây là kết quả mô hình đạt được, điểm số của mô hình được thể hiện ở cột score, bên cạnh đó đi kèm còn có thời gian chạy với từng fill data khác nhau.

Fill Data	Score	Time
0	0.966	10.17s
-1	0.97	4s
mean	0.967	4.8s

Sau khi tiến hành fill data, thấy fill = -1 cho score và time phù hợp, tiến hành lọc các featured theo mức độ quan trọng để cho ra kết quả tốt hơn.

Kết quả xảy ra khi áp dụng decision tree khi chưa lọc cột:

Chưa lọc cột	Score	Time
	0.97	10s



Hình 2: Biểu độ mức độ quan trọng của các featured

Vì có nhiều trường giá trị không quá quan trọng tới mô hình, nhóm tiến hành lọc các trường giá trị có độ quan trọng nhỏ. Thử lần đầu tiên với độ quan trọng bé hơn 0.0005

Số lượng featured bị xóa	Score	Time
7	0.9715	4.5s

Thử lần tiếp theo với độ quan trọng bé hơn 0.002

Số lượng featured bị xóa	Score	Time
66	0.969	1.6s

Thử với độ quan trọng bé hơn 0.003

Số lượng featured bị xóa	Score	Time
32	0.9718	5.333s

Thử với độ quan trọng bé hơn 0.005

Số lượng featured bị xóa	Score	Time
50	0.9725	1.88s

Thử với độ quan trọng bé hơn 0.01

Số lượng featured bị xóa	Score	Time
62	0.9726	1.32s

Thử với độ quan trọng bé hơn 0.02

Số lượng featured bị xóa	Score	Time
69	0.9667	0.719s

Thử với độ quan trọng bé hơn 0.05

Số lượng featured bị xóa	Score	Time
71	0.9748	0.7s

Thử với độ quan trọng bé hơn 0.13

Số lượng featured bị xóa	Score	Time
72	0.89	0.3s

- **Thử nghiệm với thuật toán Random Forest:**

Random Forest là một tập hợp các cây quyết định, mỗi cây được xây dựng độc lập từ một phần ngẫu nhiên của dữ liệu. Trong random forest, mỗi cây được huấn luyện trên một tập

con ngẫu nhiên của dữ liệu, được tạo ra thông qua bagging. Bagging giúp giảm việc overfitting và tăng tính ổn định của mô hình.

Đối với bài toán phân loại, Random Forest sử dụng phương pháp đa số để quyết định kết quả cuối cùng dựa trên dự đoán của từng cây. Đối với bài toán hồi quy, kết quả có thể được tính bằng cách lấy giá trị trung bình của các dự đoán.

Random Forest có khả năng xử lý dữ liệu nhiễu tốt hơn so với một số mô hình đơn lẻ, như cây quyết định. Nếu có dữ liệu thiếu, Random Forest vẫn có thể hoạt động hiệu quả mà không cần phải điền giá trị thiếu.

Tiến hành thuật toán trong đồ án:

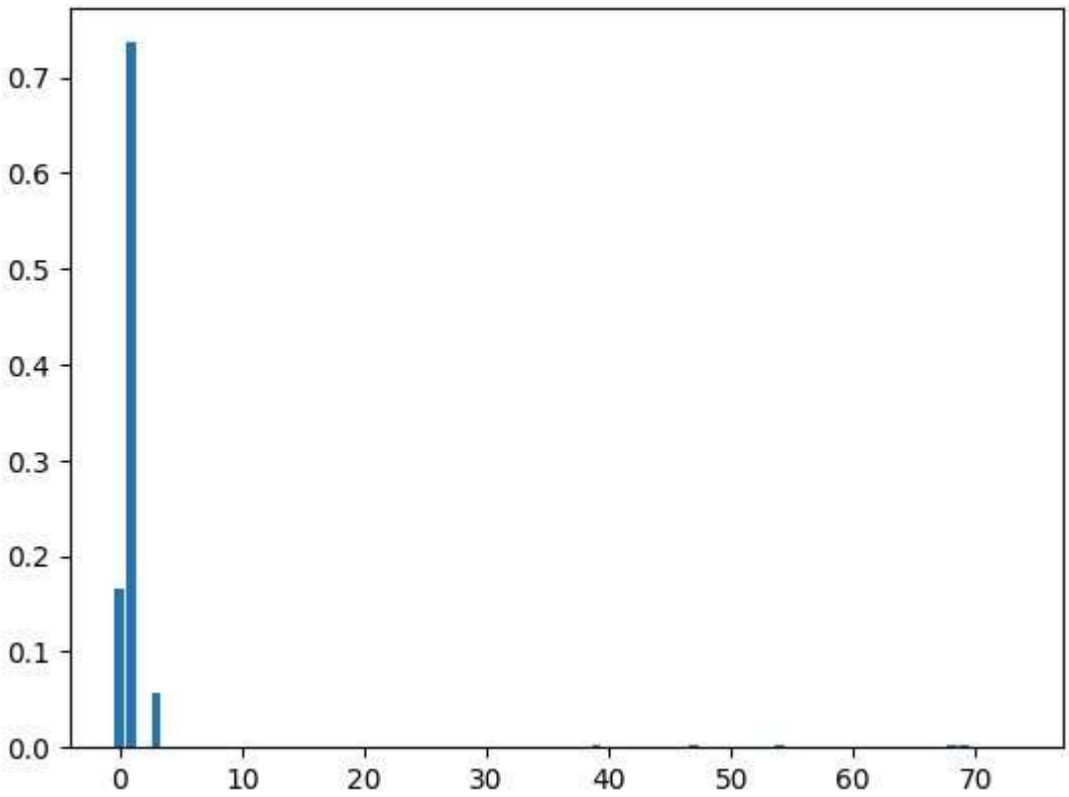
Dưới đây là kết quả mô hình đạt được, điểm số của mô hình được thể hiện ở cột score, bên cạnh đó đi kèm còn có thời gian chạy với từng fill data khác nhau.

Fill Data	Score	Time
-1	0.97	49s
0	0.9723	48.8s
Mean	0.971	49.2s

Sau khi tiến hành fill data, thấy fill = 0 cho score và time phù hợp, tiến hành lọc các feature theo mức độ quan trọng để cho ra kết quả tốt hơn.

Kết quả xảy ra khi áp dụng decision tree khi chưa lọc cột:

Chưa lọc cột	Score	Time
	0.971	49.2s



Vì có nhiều trường giá trị không quá quan trọng tới mô hình, nhóm tiến hành lọc các trường giá trị có độ quan trọng nhỏ. Thử lần đầu với độ quan trọng bé hơn 0.002

Số lượng featured bị xóa	Score	Time
71	0.977	1.097s

Thử với độ quan trọng bé hơn 0.003

Số lượng featured bị xóa	Score	Time
70	0.979	2.23s

Thử với độ quan trọng bé hơn 0.0003

Số lượng featured bị xóa	Score	Time
27	0.9714	34.62s

Thử với độ quan trọng bé hơn 0.056

Số lượng featured bị xóa	Score	Time
72	0.900	0.622s

Chương 3: Approach

1. Đặc điểm của dữ liệu

STT	Name	Valid	Missing
1	latitude	100%	0%
2	longtitude	100%	0%
3	year	100%	0%
4	week_no	100%	0%
5	SulphurDioxide_SO2_column_number_density	82%	18%
6	SulphurDioxide_SO2_column_number_density_amf	82%	18%
7	SulphurDioxide_SO2_slant_column_number_density	82%	18%
8	SulphurDioxide_cloud_fraction	82%	18%
9	SulphurDioxide_sensor_azimuth_angle	82%	18%
10	SulphurDioxide_sensor_zenith_angle	82%	18%
11	SulphurDioxide_solar_azimuth_angle	82%	18%
12	SulphurDioxide_solar_zenith_angle	82%	18%
13	SulphurDioxide_SO2_column_number_density_15km	82%	18%
14	CarbonMonoxide_CO_column_number_density	97%	3%
15	CarbonMonoxide_H2O_column_number_density	97%	3%
16	CarbonMonoxide_cloud_height	97%	3%
17	CarbonMonoxide_sensor_altitude	97%	3%
18	CarbonMonoxide_sensor_azimuth_angle	97%	3%
19	CarbonMonoxide_sensor_zenith_angle	97%	3%
20	CarbonMonoxide_solar_azimuth_angle	97%	3%
21	CarbonMonoxide_solar_zenith_angle	97%	3%
22	NitrogenDioxide_NO2_column_number_density	77%	23%
23	NitrogenDioxide_tropospheric_NO2_column_number_density	77%	23%

24	NitrogenDioxide_stratospheric_NO2_column_number_density	77%	23%
25	NitrogenDioxide_NO2_slant_column_number_density	77%	23%
26	NitrogenDioxide_tropopause_pressure	77%	23%
27	NitrogenDioxide_absorbing_aerosol_index	77%	23%
28	NitrogenDioxide_cloud_fraction	77%	23%
29	NitrogenDioxide_sensor_altitude	77%	23%
30	NitrogenDioxide_sensor_azimuth_angle	77%	23%
31	NitrogenDioxide_sensor_zenith_angle	77%	23%
32	NitrogenDioxide_solar_azimuth_angle	77%	23%
33	NitrogenDioxide_solar_zenith_angle	77%	23%
34	Formaldehyde_tropospheric_HCHO_column_number_density	91%	9%
35	Formaldehyde_tropospheric_HCHO_column_number_density_amf	91%	9%
36	Formaldehyde_HCHO_slant_column_number_density	91%	9%
37	Formaldehyde_cloud_fraction	91%	9%
38	Formaldehyde_solar_zenith_angle	91%	9%
39	Formaldehyde_solar_azimuth_angle	91%	9%
40	Formaldehyde_sensor_zenith_angle	91%	9%
41	Formaldehyde_sensor_azimuth_angle	91%	9%
42	UvAerosolIndex_absorbing_aerosol_index	99%	1%
43	UvAerosolIndex_sensor_altitude	99%	1%
44	UvAerosolIndex_sensor_azimuth_angle	99%	1%
45	UvAerosolIndex_sensor_zenith_angle	99%	1%
46	UvAerosolIndex_solar_azimuth_angle	99%	1%
47	UvAerosolIndex_solar_zenith_angle	99%	1%
48	Ozone_O3_column_number_density	99%	1%
49	Ozone_O3_column_number_density_amf	99%	1%
50	Ozone_O3_slant_column_number_density	99%	1%
51	Ozone_O3_effective_temperature	99%	1%
52	Ozone_cloud_fraction	99%	1%
53	Ozone_sensor_azimuth_angle	99%	1%

54	Ozone_sensor_zenith_angle	99%	1%
55	Ozone_solar_azimuth_angle	99%	1%
56	Ozone_solar_zenith_angle	99%	1%
57	UvAerosolLayerHeight_aerosol_height	1%	99%
58	UvAerosolLayerHeight_aerosol_pressure	1%	99%
59	UvAerosolLayerHeight_aerosol_optical_depth	1%	99%
60	UvAerosolLayerHeight_sensor_zenith_angle	1%	99%
61	UvAerosolLayerHeight_sensor_azimuth_angle	1%	99%
62	UvAerosolLayerHeight_solar_azimuth_angle	1%	99%
63	UvAerosolLayerHeight_solar_zenith_angle	1%	99%
64	Cloud_cloud_fraction	99%	1%
65	Cloud_cloud_top_pressure	99%	1%
66	Cloud_cloud_top_height	99%	1%
67	Cloud_cloud_base_pressure	99%	1%
68	Cloud_cloud_base_height	99%	1%
69	Cloud_cloud_optical_depth	99%	1%
70	Cloud_surface_albedo	99%	1%
71	Cloud_sensor_azimuth_angle	99%	1%
72	Cloud_sensor_zenith_angle	99%	1%
73	Cloud_solar_azimuth_angle	99%	1%
74	Cloud_solar_zenith_angle	99%	1%
75	emission	100%	0%

Nhìn bảng các trường thuộc tính của dữ liệu chúng ta có thể thấy có mức độ missing trong dữ liệu. Tức là các trường thuộc tính có giá trị bị thiếu. Trước khi tiến hành huấn luyện chúng ta cần phải làm sạch dữ liệu, xử lý các dữ liệu đó. Tuy nhiên sau khi tìm hiểu bộ dữ liệu trên thì nhóm đã nhận thấy rằng sự mất mát dữ liệu của bài toán là loại MNAR(Missing not at random)khi dữ liệu mất đi không hoàn toàn ngẫu nhiên.

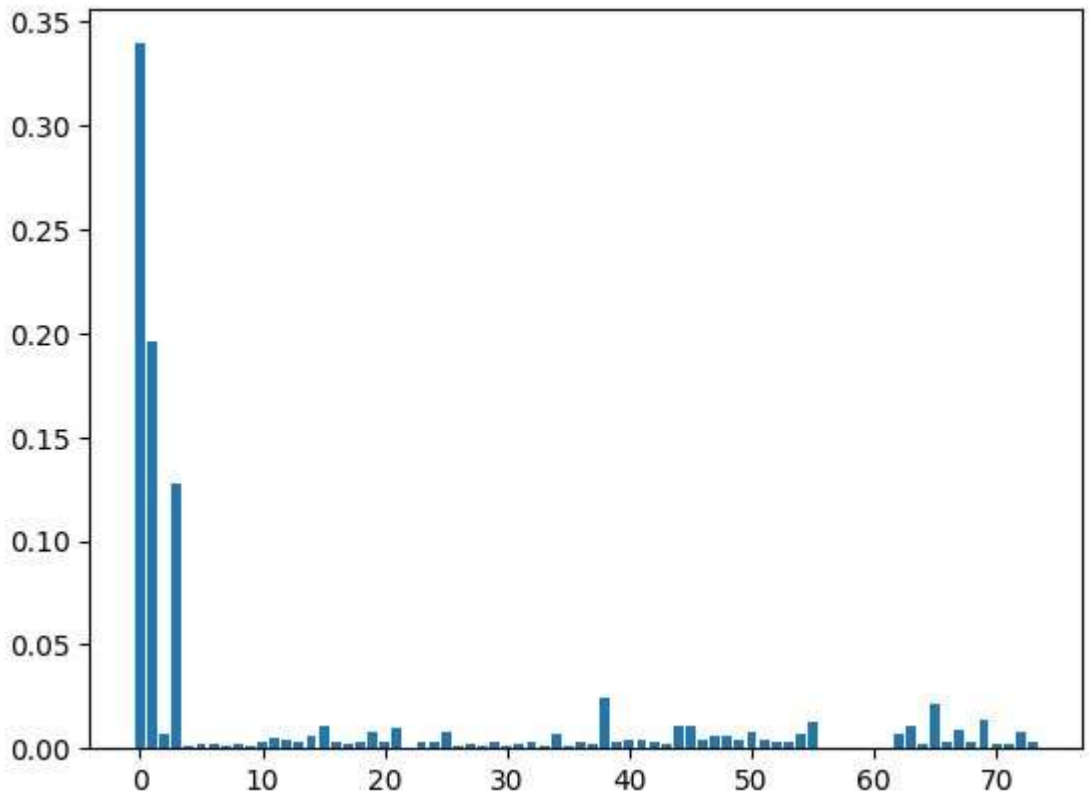
2. Giải quyết missing bằng fill data và xếp theo mức độ quan trọng.

Để giải quyết vấn đề này nhóm thực hiện bằng cách fill data và loại bỏ các trường thuộc tính theo mức độ quan trọng. Cụ thể như ví dụ sau:

Fill Data	Score	Time
Mean	0.0244	0.415s
0	0.023	0.858s
-1	0.0241	0.84s

Nhóm fill data lần lượt theo các giá trị mean, 0, -1. Sau đó dựa theo điểm số cũng như thời gian chạy để đánh giá data fill nào phù hợp, lựa chọn data đó.

Sau khi lựa chọn xong, nhóm tiến hành xuất danh sách điểm mức độ quan trọng của các trường thuộc tính, ví dụ như biểu đồ sau



Chúng ta có sấp sỉ 73 trường thuộc tính, sẽ có 73 điểm số đánh giá mức độ quan trọng của các trường thuộc tính đó. Nhìn vào biểu đồ nhóm sẽ đưa ra đánh giá và phân tích phù hợp. Sau đó sẽ tiến hành lọc những trường thuộc tính dưới điểm số bao nhiêu.

Cuối cùng sau khi lọc xong các trường thuộc tính có điểm mức độ quan trọng thấp, chúng ta tiến hành huấn luyện.

Chương 4: Results

1. Đánh giá kết quả đầu ra của thuật toán đã sử dụng trong đồ án

1.1. Thuật toán Linear Regression

- Sau khi thử nghiệm thuật toán này thì nhóm nhận ra rằng thuật toán này không phù hợp với lại tập dữ liệu dự đoán CO2 emissions vì tập dữ liệu này không tuyến tính và nhiều cột thuộc tính khác có mức độ ảnh hưởng đến kết quả khác nhau. Ngoài ra ở tập ở dữ liệu này, dữ liệu missing chiếm hầu hết các dữ liệu được nhập vào nên nếu fill dữ liệu vào thì sẽ tạo ra sự bất hợp lý cho kho dữ liệu train của thuật toán.
- Thuật toán có độ chính xác rất kém khi với score với độ đo R2 chỉ khoảng ở : 0.024 với việc chia tập train là bảy phần và tập test là ba phần trong tập dữ liệu train của đề bài.

1.2. Thuật toán Decision tree

- Sau khi thử nghiệm thuật toán này thì nhóm nhận ra rằng thuật toán này khá phù hợp với lại tập dữ liệu CO2 emissions vì với việc sử dụng thuật toán decision tree sẽ được chia ra các trường hợp đầu ra kết quả khác nhau dựa trên kết quả đầu vào của dữ liệu. Với trường hợp ở tập dữ liệu CO2 emission thì dữ liệu đầu vào khá phù thuộc vào các trường hợp dữ liệu khác nhau như “Vị trí địa lý”, “Thời gian”, và các trường thuộc tính khác liên quan của dữ liệu để từ đó có thể đưa ra kết quả đầu ra cho dự đoán của dữ liệu, chính nhờ thế mà kết quả dự đoán đầu ra của bài toán khá chính xác.

- Bài toán áp dụng Decision tree có độ chính xác cao với độ đo R2 có score: 0.957
- Tuy nhiên sau khi nhóm kiểm thử độ quan trọng của các trường dữ liệu ảnh hưởng đến model thì nhóm nhận thấy rằng khá nhiều trường dữ liệu với mức độ ảnh hưởng của nó đến dữ liệu đầu ra là khá thấp nên chính vì thế, nhóm đã sử dụng phương pháp tìm độ quan trọng và loại bỏ bớt các cột có độ quan trọng nhỏ hơn 0.02 và loại bỏ 71 cột dữ liệu thì độ chính xác đã của thuật toán đã tăng lên. Bài toán có độ chính xác tăng lên với độ đo R2 có score: 0. 9715.

Kết quả thuật toán thu được khá tốt tuy nhiên vẫn chưa phải tốt nhất khi Decision tree chỉ là một mô hình đơn lẻ.

1.3. Thuật toán XGBoost

- Thuật toán XGBoost thuộc loại thuật toán Gradient Boosting, nơi nó xây dựng một chuỗi các cây quyết định theo cách tuần tự. Chính vì thế mô hình này sẽ cải thiện độ chính xác của thuật toán một cách tốt hơn so với thuật toán Decision tree khi thuật toán này sẽ dựa trên các cây Decision tree đã được dựa và cải thiện sai số so với thông tin cây quyết định cho trước.
- Bài toán áp dụng XGBoost có độ chính xác cao với độ đo R2 có score tầm: 0.967
- Tuy nhiên sau khi nhóm kiểm thử độ quan trọng của các trường dữ liệu ảnh hưởng đến model thì nhóm nhận thấy rằng khá nhiều trường dữ liệu với mức độ ảnh hưởng của nó đến dữ liệu đầu ra là khá thấp nên chính vì thế, nhóm đã sử dụng phương pháp tìm độ quan trọng và loại bỏ bớt các cột có độ quan trọng nhỏ hơn 0.05 và loại bỏ 71 cột dữ liệu thì độ chính xác đã của thuật toán đã tăng lên. Bài toán có độ chính xác tăng lên với độ đo R2 có score: 0. 9748.

  Điểm độ đo R2 thu được ở mô hình này tốt hơn so với lại điểm của Decision tree có tăng lên nhưng vẫn không quá đáng kể so với mô hình Decision tree.

1.4. Thuật toán Random Forest

- Random Forest là một tập hợp các cây quyết định, mỗi cây được xây dựng độc lập từ một phần ngẫu nhiên của dữ liệu. Với thuật toán này nhóm sẽ chọn ra số cây mà mô hình sử dụng là 10 vì nếu sử dụng nhiều cây quá sẽ ảnh hưởng đến tốc độ train của model cũng như là xảy ra hiện tượng overfitting khi test với dữ liệu mới khi dự đoán. Chính vì dùng nhiều cây quyết định ngẫu nhiên hơn nên việc dự đoán của mô hình này sẽ tốt hơn so với việc sử dụng chỉ một cây đơn lẻ như mô hình Decision Tree để từ đó cải thiện độ chính xác của bài toán hơn.
- Bài toán áp dụng Random Forest có độ chính xác cao với độ tập train chiếm 7 phần và tập test chiếm 3 phần với độ đo R2 có score tầm: 0.9724.
- Tuy nhiên sau khi nhóm kiểm thử độ quan trọng của các trường dữ liệu ảnh hưởng đến model thì nhóm nhận thấy rằng khá nhiều trường dữ liệu với mức độ ảnh hưởng của nó đến dữ liệu đầu ra là khá thấp nên chính vì thế, nhóm đã sử dụng phương pháp tìm độ quan trọng và loại bỏ bớt các cột có độ quan trọng nhỏ hơn 0.003 và loại bỏ 70 cột dữ liệu thì độ chính xác đã của thuật toán đã tăng lên. Bài toán có độ chính xác tăng lên với độ đo R2 có score: 0. 979.
- Tuy nhiên với việc lấy ngẫu nhiên n cây quyết định thì việc train model ở từng thời điểm khác nhau sẽ có thể có sự chênh lệch về mức độ dự đoán của độ chính xác đó của bài toán.

  Điểm độ đo R2 thu được ở mô hình này tốt hơn khá nhiều so với Decision tree vì nó áp dụng mô hình đa cây quyết định so với Decision là đơn cây.

Sau khi tiến hành phân tích và dánh giá 4 mô hình, thử nghiệm các các thuật toán. Nhóm nhận thấy rằng mô hình thuật toán Random Forest là thuật toán phù hợp nhất khi áp dụng ở bộ dữ liệu train CO2 emissions của bài toán được đặt ra.

Chương 5: Conclusion

Qua quá trình làm, thử nghiệm qua rất nhiều thuật toán để giải quyết bài toán, thực hiện đánh giá các thuật toán và tham khảo nhiều nguồn thông tin trên mạng. Random Forest được nhóm đánh giá là thuật toán phù hợp cho bài toán.

Đối với bài toán phân loại, Random Forest sử dụng phương pháp đa số để quyết định kết quả cuối cùng dựa trên dự đoán của từng cây. Đối với bài toán hồi quy, kết quả có thể được tính bằng cách lấy giá trị trung bình của các dự đoán.

Random Forest có khả năng xử lý dữ liệu nhiễu tốt hơn so với một số mô hình đơn lẻ, như cây quyết định. Nếu có dữ liệu thiếu, Random Forest vẫn có thể hoạt động hiệu quả.

Tuy kết quả cho ra của mô hình cho ra có thể không đạt được độ chính xác ở mức hoàn hảo tuy nghiên dựa theo những số liệu đã được cung cấp và chứng minh, mô hình giải quyết bài toán CO2 Emission sử dụng random forest vẫn có thể hoạt động và cho ra kết quả tương đối.

Chương 6: Reference

[1] Đ. Dương, "Nồng độ CO2 trong khí quyển tăng cao kỷ lục," *VNEXPRESS,* p. 1, 2022.

[2] "CO và CO2: Mối nguy hiểm tiềm tàng," *mes-ionar,* p. 3, 2022.

[3] Cục Công nghệ Thông tin - Bộ Tài nguyên và Môi trường, "Xây dựng mô hình cảnh báo, dự báo theo phương pháp học máy có giám sát thử nghiệm dự báo xâm ngập mặn cho lưu vực sông Hậu," *Tài nguyên và môi trường.*

[4] "Application of machine learning initiatives and intelligent perspectives for CO2 emissions reduction in construction".

[5] “What is linear regression?,” *ibm.com.*