Phân Tích và Xây Dựng Mô Hình Dự Đoán Nồng Độ CO Trong Không Khí

Phạm Đức Thể[1,2,3] and Đỗ Trọng Hợp[1,2,4]

[1] Đại Học Công Nghệ Thông Tin TP.HCM, Việt Nam
[2] Đại Học Quốc Gia TP.HCM, Việt Nam
[3] 19522253@gm.uit.edu.vn
[4] hopdt@uit.edu.vn

Tóm tắt nội dung Chất lượng không khí có ảnh hưởng không nhỏ đến sức khỏe con người. Ô nhiễm không khí dẫn đến một loạt các vấn đề sức khỏe, đặc biệt là ở trẻ em. Một trong những tác nhân ảnh hưởng đến chất lượng không khí là Carbon Monoxide (CO). Dự đoán nồng độ CO trong không khí nhằm đưa ra các cảnh báo sớm, kịp thời cho phép chính phủ và các tổ chức liên quan khác thực hiện các bước cần thiết để bảo vệ những người dễ bị tổn thương nhất, khỏi tiếp xúc với không khí có chất lượng nguy hiểm. Trong báo cáo này, chúng tôi thực hiện phân tích và xây dựng mô hình dự đoán nồng độ CO trong không khí trên bộ dữ liệu Air Quality. Quá trình xử lý missing values, chúng tôi tạo ra 2 bộ dữ liệu REMOVE và MEAN. Sau quá trình phân tích ANOVA, chúng tôi thu được các bộ dữ liệu: REMOVE gốc, REMOVE ANOVA đơn thuộc tính, REMOVE ANOVA tương tác 2 thuộc tính, MEAN gốc và MEAN ANOVA tương tác 2 thuộc tính. Trong phần thực nghiệm, chúng tôi sử dụng các thuật toán Machine Learning và Deep Learning để so sánh kết quả thông qua các độ đo: R^2, MSE, RMSE, MAE. Kết quả tốt nhất theo độ đo RMSE là 0.3789 sử dụng Support Vector Regression trên bộ dữ liệu REMOVE ANOVA đơn thuộc tính.

Keywords: Air Quality · Exploratory Data Analysis · Missing Values · ANOVA · Linear Regression · Decision Tree Regression · Random Forest Regression · Support Vector Regression · Neural Network.

1 Giới Thiệu

Ô nhiễm không khí là sự thay đổi lớn trong thành phần của không khí, chủ yếu do khói, bụi, hơi hoặc các khí lạ được đưa vào không khí, có sự tỏa mùi, làm giảm tầm nhìn xa, gây biến đổi khí hậu, gây bệnh cho con người và cũng có thể gây hại cho sinh vật khác như động vật và cây lương thực, nó có thể làm hỏng môi trường tự nhiên hoặc xây dựng. Hoạt động của con người và các quá trình tự nhiên có thể gây ra ô nhiễm không khí. Một chất gây ô nhiễm không khí là một chất trong không khí có thể gây hại cho con người và hệ sinh thái. Chất này có thể là các hạt rắn, giọt chất lỏng, hoặc khí. Các chất gây ô nhiễm không khí phổ biến như: Carbon Dioxide (CO_2), Sulfur Oxide (SOx), Oxide

Nitơ (NOx), Carbon Monoxide (CO), các hạt mịn (PM), Amonia (NH_3), Ozone (O_3), ... Trong đó, CO là một loại khí không màu, không mùi, độc nhưng không gây kích thích. Nó là sản phẩm của sự đốt cháy không đầy đủ của nhiên liệu như khí tự nhiên, than đá hoặc gỗ. Khói xả từ các phương tiện giao thông là một nguồn chính của CO. Nồng độ CO trong không khí có ảnh hưởng trực tiếp đến chất lượng không khí và sức khỏe của con người.

Dự đoán nồng độ CO trong không khí nhằm đánh giá mức độ ô nhiễm không khi do tác nhân khí CO gây ra, từ đó giúp đưa ra các cảnh báo kịp thời cho phép chính phủ và các tổ chức liên quan khác thực hiện các bước cần thiết để bảo vệ những người dễ bị tổn thương nhất, khỏi tiếp xúc với không khí có chất lượng nguy hiểm. Báo cáo này nhằm mục đích phân tích và xây dựng một mô hình có thể xem xét dữ liệu chất lượng không khí đã được ghi lại trước đó và dự đoán nồng đồ CO trong không khí.

Trong báo cáo này, trước tiên chúng tôi trình bày về bộ dữ liệu Air Quality được sử dụng cho bài toán và các bước tiền xử lý dữ liệu cơ bản như: xử lý kiểu dữ liệu và xử lý missing values. Từ góc độ xử lý missing values, chúng tôi xem xét 2 loại chiến lược xậy dựng bộ dữ liệu là REMOVE và MEAN. REMOVE – xóa tất cả các dòng dữ liệu bị missing values của thuộc tính CO_GT. MEAN – gán các missing values bằng giá trị trung bình của thuộc tính CO_GT (Phần 2). Sau đó, chúng tôi tiến hành phân tích khám phá bộ dữ liệu trong Phần 3. Hướng tiếp cận bài toán được mô tả chi tiết trong Phần 4. Quy trình phân tích ANOVA trên từng bộ dữ liệu được trình bày đầy đủ trong Phần 5. Trong Phần 6, chúng tôi tiến hành thực nghiệm và phân tích kết quả của các mô hình Machine Learning và Deep Learning trên các bộ dữ liệu khác nhau. Cuối cùng, chúng tôi rút ra kết luận ở Phần 7.

2 Bộ Dữ Liệu

2.1 Bộ dữ liệu gốc

Bộ dữ liệu chúng tôi sử dụng trong báo cáo này có tên là *Air Quality Data Set* được lấy từ UCI Machine Learning Repository, một trang web cung cấp dữ liệu miễn phí cho các dự án về machine learning [1, 2]. Bộ dữ liệu chứa 9,357 dòng dữ liệu và 15 thuộc tính là các phản hồi trung bình hàng giờ từ một loạt 5 cảm biến hóa học oxit kim loại được nhúng trong *Thiết bị đa cảm biến hóa học chất lượng không khí (Air Quality Chemical Multisensor Device)*. Thiết bị được đặt trên cánh đồng ở một khu vực bị ô nhiễm nghiêm trọng, trong một thành phố của Italian. Dữ liệu được ghi lại từ tháng 3 năm 2004 đến tháng 2 năm 2005 (một năm). Nồng độ trung bình hàng giờ của Ground Truth (GT) đối với CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) và Nitrogen Dioxide (NO2) và được cung cấp bởi một *máy phân tích tham chiếu (reference analyzer)* được chứng nhận đặt cùng vị trí. Bộ dữ liệu có nhiều *giá trị bị thiếu (missing values)*, các missing values này được gán với giá trị -200. Bộ dữ liệu này có thể được sử dụng riêng cho mục đích nghiên cứu. Mục đích thương mại được loại trừ hoàn toàn. Thông tin chi tiết về các thuộc tính của bộ dữ liệu được thể hiện trong Bảng 1

Index	Thuộc tính	Ý nghĩa
0	DATE	Ngày (DD/MM/YYYY).
1	TIME	Thời gian trong ngày (HH.MM.SS) (24 giờ).
2	CO(GT)	Nồng độ CO trung bình thực sự hàng giờ (mg/m^3).
3	PT08.S1(CO)	Phản hồi cảm biến trung bình hàng giờ (Thiếc oxit - nominally CO targeted).
4	NMHC(GT)	Nồng độ tổng thể của HydroCarbons Non Metanic trung bình thực sự hàng giờ ($microg/m^3$).
5	C6H6(GT)	Nồng độ Benzen trung bình thực sự hàng giờ ($microg/m^3$).
6	PT08.S2(NMHC)	Phản hồi cảm biến trung bình hàng giờ (Titania - nominally NMHC targeted).
7	NOx(GT)	Nồng độ NOx trung bình thực sự hàng giờ (ppb).
8	PT08.S3(NOx)	Phản hồi cảm biến trung bình hàng giờ (Oxit vonfram - nominally NOx targeted).
9	NO2(GT)	Nồng độ NO2 trung bình thực sự hàng giờ ($microg/m^3$).
10	PT08.S4(NO2)	Phản hồi cảm biến trung bình hàng giờ (Oxit vonfram - nominally NO2 targeted).
11	PT08.S5(O3)	Phản hồi cảm biến trung bình hàng giờ (Oxit indium - nominally O3 targeted).
12	T	Temperature - Nhiệt độ (°C).
13	RH	Relative Humidity - Độ ẩm tương đối (%).
14	AH	Absolute Humidity - Độ ẩm tuyệt đối.

Bảng 1: Thông tin chi tiết của các thuộc tính.

2.2 Tiền xử lý dữ liệu

Dữ liệu tốt đóng vai tối trò quan trọng trong việc tạo ra các mô hình dự báo có độ chính xác cao và tổng quát hóa. Để có được bộ dữ liệu tốt ứng với từng tác vụ thì ta có những cách xử lý dữ liệu khác nhau, trước tiên ta cần phải kiểm tra các insight đầu tiên và cơ bản của bộ dữ liệu, sau đó sẽ nêu ra các hướng giải quyết, làm sạch dữ liệu để có được dữ liệu tốt cho bài toán. Đối với bài toán dự đoán nồng độ CO trong không khí sử dụng bộ dữ liệu Air Quality Data Set, chúng tôi tiến hành đọc dữ liệu và hiển thị ra như Hình 1 và 2a để kiểm tra các vấn đề cơ bản của một bộ dữ liệu như: số lượng điểm dữ liệu, số lượng thuộc tính, kiểu dữ liệu và missing values. Từ Hình 1 và 2a, ta có 2 vấn đề cần xử lý là datatype (các thuộc tính bị sai kiểu dữ liệu) và missing values (dữ liệu bị thiếu).

Datatype Có 5 thuộc bị sai kiểu dữ liệu gồm: CO_GT, C6H6_GT, T, RH, AH. 5 thuộc tính này bị gán sai kiểu dữ liệu là object, ta cần chuyển chúng về với đúng kiểu dữ liệu là float64. Với thuộc tính DATE và TIME có kiểu dữ liệu là object chúng tôi tiến hành nối chúng lại và chuyển về kiểu dữ liệu datetime64. Kết quả của quá trình xử lý này được thể hiện ở Hình 2b.

Missing values Hình 3a thể hiện số lượng và tỉ lệ missing values của từng thuộc tính trong bộ dữ liệu, với những thuộc tính có tỉ lệ missing values trên

	DATE	TIME	CO_GT	PT08_S1_CO	NMHC_GT	C6H6_GT	PT08_S2_NMHC	NOx_GT	PT08_S3_NOx	NO2_GT	PT08_S4_NO2	PT08_S5_O3	T	RH	AH
0	10/03/2004	18.00.00	2,6	1360.0	150.0	11,9	1046.0	166.0	1056.0	113.0	1692.0	1268.0	13,6	48,9	0,7578
1	10/03/2004	19.00.00	2	1292.0	112.0	9,4	955.0	103.0	1174.0	92.0	1559.0	972.0	13,3	47,7	0,7255
2	10/03/2004	20.00.00	2,2	1402.0	88.0	9,0	939.0	131.0	1140.0	114.0	1555.0	1074.0	11,9	54,0	0,7502
3	10/03/2004	21.00.00	2,2	1376.0	80.0	9,2	948.0	172.0	1092.0	122.0	1584.0	1203.0	11,0	60,0	0,7867
4	10/03/2004	22.00.00	1,6	1272.0	51.0	6,5	836.0	131.0	1205.0	116.0	1490.0	1110.0	11,2	59,6	0,7888
5	10/03/2004	23.00.00	1,2	1197.0	38.0	4,7	750.0	89.0	1337.0	96.0	1393.0	949.0	11,2	59,2	0,7848
6	11/03/2004	00.00.00	1,2	1185.0	31.0	3,6	690.0	62.0	1462.0	77.0	1333.0	733.0	11,3	56,8	0,7603
7	11/03/2004	01.00.00	1	1136.0	31.0	3,3	672.0	62.0	1453.0	76.0	1333.0	730.0	10,7	60,0	0,7702
8	11/03/2004	02.00.00	0,9	1094.0	24.0	2,3	609.0	45.0	1579.0	60.0	1276.0	620.0	10,7	59,7	0,7648
9	11/03/2004	03.00.00	0,6	1010.0	19.0	1,7	561.0	NaN	1705.0	NaN	1235.0	501.0	10,3	60,2	0,7517

Hình 1: Thông tin 10 điểm dữ liệu đầu tiên của bộ dữ liệu.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9357 entries, 0 to 9356
Data columns (total 15 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   DATE           9357 non-null   object
 1   TIME           9357 non-null   object
 2   CO_GT          7674 non-null   object
 3   PT08_S1_CO     8991 non-null   float64
 4   NMHC_GT        914 non-null    float64
 5   C6H6_GT        8991 non-null   object
 6   PT08_S2_NMHC   8991 non-null   float64
 7   NOx_GT         7718 non-null   float64
 8   PT08_S3_NOx    8991 non-null   float64
 9   NO2_GT         7715 non-null   float64
 10  PT08_S4_NO2    8991 non-null   float64
 11  PT08_S5_O3     8991 non-null   float64
 12  T              8991 non-null   object
 13  RH             8991 non-null   object
 14  AH             8991 non-null   object
dtypes: float64(8), object(7)
memory usage: 1.1+ MB
```

(a) Datatype ban đầu

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9357 entries, 0 to 9356
Data columns (total 14 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   DATE_TIME      9357 non-null   datetime64[ns]
 1   CO_GT          7674 non-null   float64
 2   PT08_S1_CO     8991 non-null   float64
 3   NMHC_GT        914 non-null    float64
 4   C6H6_GT        8991 non-null   float64
 5   PT08_S2_NMHC   8991 non-null   float64
 6   NOx_GT         7718 non-null   float64
 7   PT08_S3_NOx    8991 non-null   float64
 8   NO2_GT         7715 non-null   float64
 9   PT08_S4_NO2    8991 non-null   float64
 10  PT08_S5_O3     8991 non-null   float64
 11  T              8991 non-null   float64
 12  RH             8991 non-null   float64
 13  AH             8991 non-null   float64
dtypes: datetime64[ns](1), float64(13)
memory usage: 1023.5 KB
```

(b) Datatype sau khi xử lý

Hình 2: Thông tin bộ dữ liệu

50% (NMHC_GT) chúng tôi sẽ loại bỏ trực tiếp thuộc tính đó khỏi bộ dữ liệu. Sau khi xóa bỏ thuộc tính NMHC_GT, ta còn lại số lượng và tỉ lệ missing values của các thuộc tính như Hình 3b. Ta thấy, thuộc CO_GT có số lượng missing values cao nhất (1683 $\sim$ 18%) mà thuộc tính này lại là thuộc tính mục tiêu cần dự đoán của bài toán. Nên chúng tôi đề xuất 2 chiến lược để xử lý như sau: (1) **REMOVE**: Xóa tất cả các dòng dữ liệu bị missing values của thuộc tính CO_GT. (2) **MEAN**: Điền các missing values bằng giá trị trung bình của thuộc tính CO_GT. Sau khi xử lý missing values của thuộc tính CO_GT xong, với các missing values ở các thuộc tính khác chúng tôi tiến hành điền các missing values bằng giá trị trung bình của từng thuộc tính. Cuối cùng chúng tôi thu được 2 bộ dữ liệu mới ứng với 2 chiến lược xử lý missing values là: Air Quality–REMOVE và Air Quality–MEAN (gọi tắt là REMOVE và MEAN).

	NMHC_GT	CO_GT	NO2_GT	NOx_GT	PT08_S1_CO	C6H6_GT	PT08_S2_NMHC	PT08_S3_NOx	PT08_S4_NO2	PT08_S5_O3	T	RH	AH	DATE_TIME
Total	8443.000	1683.000	1642.000	1639.000	366.000	366.000	366.000	366.000	366.000	366.000	366.000	366.000	366.000	0.0
Percent	90.232	17.987	17.548	17.516	3.912	3.912	3.912	3.912	3.912	3.912	3.912	3.912	3.912	0.0

(a) Missing values của tất cả các thuộc tính của bộ dữ liệu.

	CO_GT	NO2_GT	NOx_GT	PT08_S1_CO	C6H6_GT	PT08_S2_NMHC	PT08_S3_NOx	PT08_S4_NO2	PT08_S5_O3	T	RH	AH	DATE_TIME
Total	1683.000	1642.000	1639.000	366.000	366.000	366.000	366.000	366.000	366.000	366.000	366.000	366.000	0.0
Percent	17.987	17.548	17.516	3.912	3.912	3.912	3.912	3.912	3.912	3.912	3.912	3.912	0.0

(b) Missing values của các thuộc tính sau khi bỏ đi thuộc tính NMHC_GT.

Hình 3: Missing values

3 Exploratory Data Analysis

3.1 Thống kê mô tả

Thống kê mô tả được sử dụng để mô tả những đặc tính cơ bản của dữ liệu. Hình 4a và 4b lần lượt là thống kê mô tả chi tiết các thuộc tính của 2 bộ dữ liệu, cho biết các thông tin như: count – số lượng các điểm dữ liệu, mean – giá trị trung bình, std – độ lệch chuẩn, min – giá trị nhỏ nhất, 25% – tứ phân vị thứ nhất, 50% – tứ phân vị thứ hai (median – trung vị), 75% – tứ phân vị thứ ba, max – giá trị lớn nhất.

	CO_GT	PT08_S1_CO	C6H6_GT	PT08_S2_NMHC	NOx_GT	PT08_S3_NOx	NO2_GT	PT08_S4_NO2	PT08_S5_O3	T	RH	AH
count	7674.000000	7674.000000	7674.000000	7674.000000	7674.000000	7674.000000	7674.000000	7674.000000	7674.000000	7674.000000	7674.000000	7674.000000
mean	2.152750	1110.118439	10.267318	946.852033	254.861050	827.288598	114.718587	1445.247563	1042.626961	17.794058	49.067383	0.991197
std	1.453252	213.938253	7.279405	259.705310	209.480706	251.074837	46.919686	342.735535	396.774427	8.670728	17.072196	0.391282
min	0.100000	647.000000	0.200000	387.000000	2.000000	322.000000	2.000000	551.000000	221.000000	-1.900000	9.200000	0.184700
25%	1.100000	953.000000	4.800000	752.000000	107.000000	657.000000	82.000000	1215.250000	759.000000	11.500000	36.200000	0.714375
50%	1.800000	1087.000000	8.900000	934.000000	201.000000	807.000000	113.090000	1456.260000	1013.000000	17.550000	49.230000	0.983950
75%	2.900000	1235.000000	14.000000	1116.750000	326.000000	949.000000	141.000000	1659.000000	1287.000000	23.500000	61.800000	1.235200
max	11.900000	2040.000000	63.700000	2214.000000	1479.000000	2683.000000	340.000000	2775.000000	2523.000000	44.600000	88.700000	2.180600

(a) Air Quality–REMOVE.

	CO_GT	PT08_S1_CO	C6H6_GT	PT08_S2_NMHC	NOx_GT	PT08_S3_NOx	NO2_GT	PT08_S4_NO2	PT08_S5_O3	T	RH	AH
count	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000
mean	2.152255	1099.833043	10.082984	939.153244	246.897307	835.493464	113.091031	1456.264418	1022.906280	18.317914	49.234037	1.025705
std	1.316069	212.791672	7.302650	261.560236	193.426632	251.743954	43.920954	339.367559	390.612324	8.657639	16.974801	0.395836
min	0.100000	647.000000	0.100000	383.000000	2.000000	322.000000	2.000000	551.000000	221.000000	-1.900000	9.200000	0.184700
25%	1.200000	941.000000	4.600000	743.000000	112.000000	666.000000	86.000000	1242.000000	742.000000	12.000000	36.600000	0.746100
50%	2.150000	1075.000000	8.600000	923.000000	229.000000	818.000000	113.090000	1456.260000	983.000000	18.300000	49.230000	1.015400
75%	2.600000	1221.000000	13.600000	1105.000000	284.000000	960.000000	133.000000	1662.000000	1255.000000	24.100000	61.900000	1.296200
max	11.900000	2040.000000	63.700000	2214.000000	1479.000000	2683.000000	340.000000	2775.000000	2523.000000	44.600000	88.700000	2.231000

(b) Air Quality–MEAN.

Hình 4: Thống kê mô tả của các bộ dữ liệu.

3.2 Ma trận tương quan

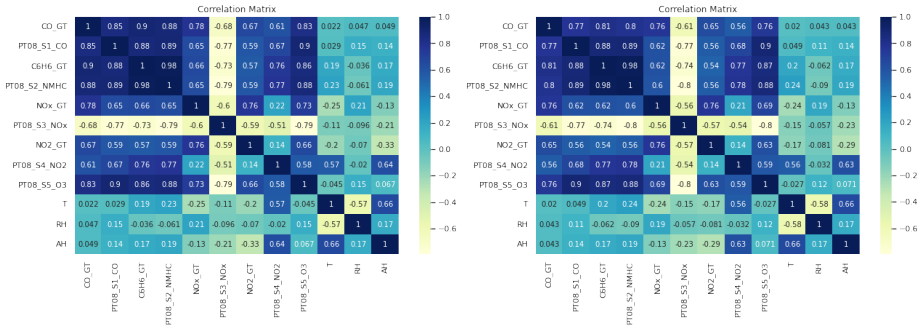
Hệ số tương quan (Correlation coefficient) [3] là chỉ số thống kê đo lường mức độ mạnh yếu của mối quan hệ giữa hai thuộc tính. Hệ số tương quan Pearson (r) có giá trị giao động trong khoảng liên tục từ -1 đến +1:

- r = 0: Hai biến không có tương quan tuyến tính.
- r = 1; r = -1: Hai biến có mối tương quan tuyến tính tuyệt đối.
- r < 0: Hệ số tương quan âm. Nghĩa là giá trị biến x tăng thì giá trị biến y giảm và ngược lại, giá trị biến y tăng thì giá trị biến x giảm.
- r > 0: Hệ số tương quan dương. Nghĩa là giá trị biến x tăng thì giá trị biến y tăng và ngược lại, giá trị biến y tăng thì giá trị biến x cũng tăng.

Hệ số tương quan pearson (r) chỉ có ý nghĩa khi và chỉ khi mức ý nghĩa quan sát nhỏ hơn mức ý nghĩa $\alpha = 5\%$.

- Nếu r thuộc khoảng từ 0.50 đến ± 1, thì nó được cho là tương quan mạnh.
- Nếu r thuộc khoảng từ 0.30 đến ± 0.49, thì nó được gọi là tương quan trung bình.
- Nếu r nằm dưới ± 0.29, thì nó được gọi là một mối tương quan yếu.

Ma trận tương quan (Correlation Matrix) là một bảng thể hiện hệ số tương quan giữa các biến khi ta có nhiều hơn 2 biến trong bộ dữ liệu. Mỗi ô trong bảng hiển thị mối tương quan giữa hai biến. Hình 5 thể hiện mối quan quan giữa các thuộc tính trong bộ dữ liệu. Dựa vào định nghĩa mức độ tương quan trên ta có: các thuộc tính tương quan yếu với thuộc tính CO_GT trên cả 2 bộ dữ liệu là: AH, RH, T; các thuộc tính tương quan mạnh với thuộc tính CO_GT trên cả 2 bộ dữ liệu là: PT08_S1_CO, C6H6_GT, PT08_S2_NMHC, PT08_S3_NOx, PT08_S4_NO2, NOx_GT, NO2_GT, PT08_S5_O3; không có thuộc tính nào tương quan trung bình với thuộc tính CO_GT trên cả 2 bộ dữ liệu. Hệ số tương quan của bộ dữ liệu REMOVE lớn hơn MEAN ở hầu hết các thuộc tính.

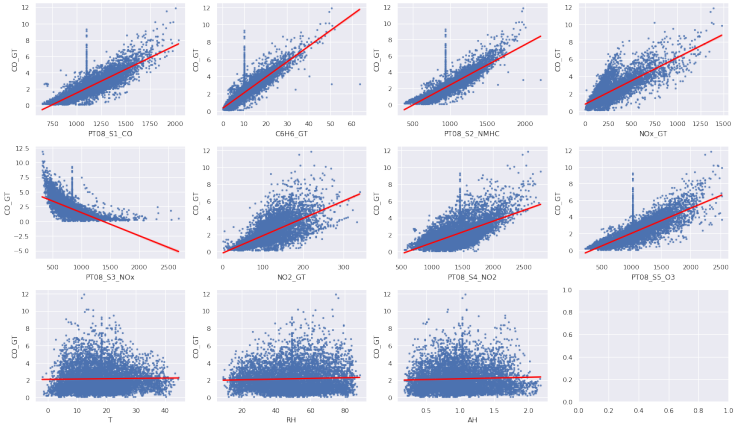


(a) Air Quality–REMOVE. (b) Air Quality–MEAN

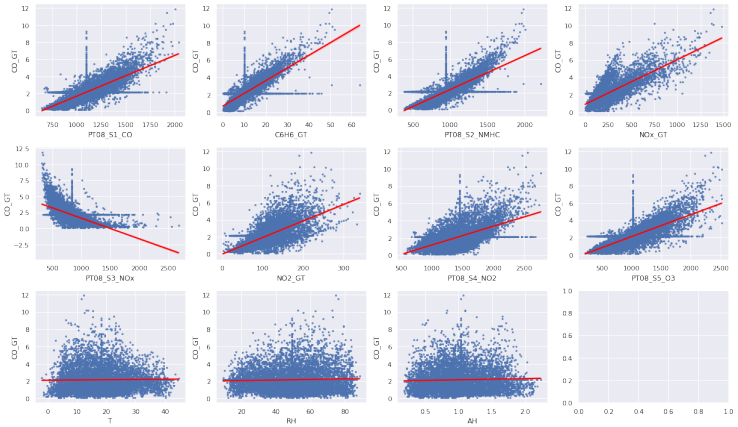
Hình 5: Correlation Matrix

3.3 Regression plot

Regression plot thể hiện mức độ phụ thuộc lẫn nhau của các biến khác nhau (mối tương quan giữa các biến), các đường hồi quy được sinh ra dựa trên hệ số tương quan. Hình 6 thể hiện mối tương quan giữa thuộc tính CO_GT với các thuốc tính còn lại trong bộ dữ liệu (trực quan hệ số tương quan của thuộc tính CO_GT với các thuộc tính khác ở Hình 5). Qua đây, ta có thể thấy rõ hơn về mối quan hệ tương quan tuyến tính của các thuộc tính với biến mục tiêu CO_GT. Các thuộc tính T, RH, AH hầu như không có mối quan hệ tương quan tuyến tính với biến mục tiêu, các thuộc tính còn lại đều có mối quan hệ tương quan tuyến tính với biến mục tiêu (đặc biệt thuộc tính PT08_S1_NOx có tương quan không đồng thuận với biến mục tiêu (tương quan âm - nghịch biến)).



(a) Air Quality–REMOVE.



(b) Air Quality–MEAN.

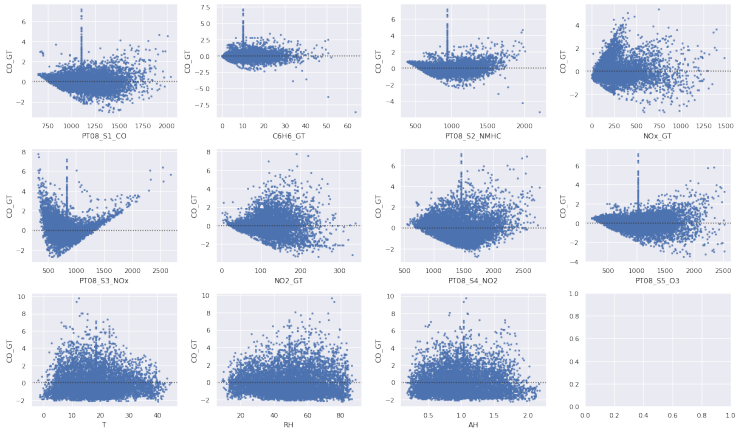
Hình 6: Regression plot.

3.4 Residual plot

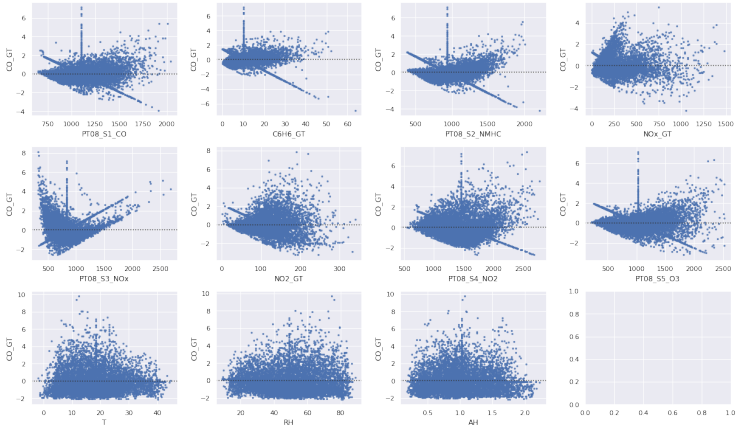
Residual (Phần dư) được tính bằng hiệu số giữa giá trị kỳ vọng và giá trị thực tế của biến phụ thuộc. Giá trị kỳ vọng được tính bằng cách thay thế các giá trị khác nhau của biến độc lập trong phương trình hồi quy đã phát triển.

$$residual = y_i - \hat{y}_i \tag{1}$$

Residual plot là biểu đồ được sử dụng để kiểm xem các giả định được đưa ra trong phân tích hồi quy có đúng hay không. Nó là một đồ thị được vẽ giữa các phần dư của một mô hình hồi quy cụ thể và biến độc lập. Hình 7 trình bày residual plot của thuộc tính CO_GT với các thuộc tính khác.



(a) Air Quality–REMOVE.

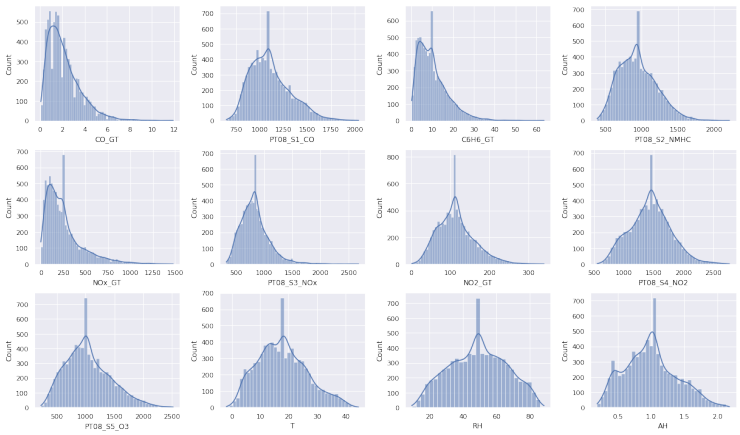


(b) Air Quality–MEAN.

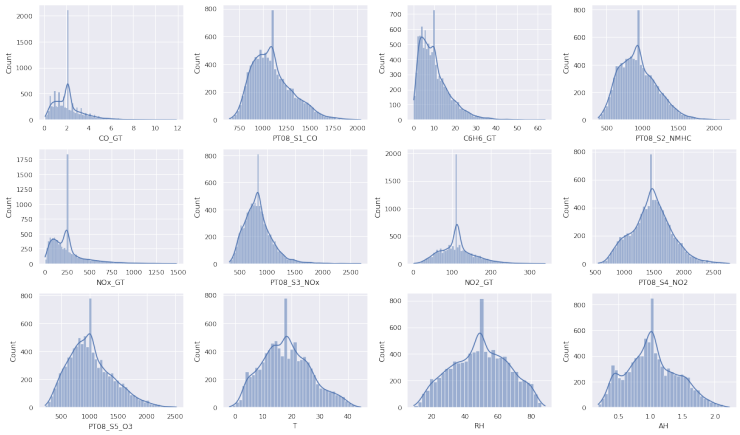
Hình 7: Residual plot.

3.5 Histogram plot

Histogram là một dạng biểu đồ biểu diễn phân phối tần suất của một biến, thấy bằng hình ảnh sự thay đổi, biến động của một tập hợp các dữ liệu theo những hình dạng nhất định.



(a) Air Quality–REMOVE.



(b) Air Quality–MEAN.

Hình 8: Histogram plot.

Hình 8a là histogram của bộ dữ liệu xử lý theo chiến lượt REMOVE. Ta thấy, hầu hết hình dạng biểu đồ của các thuộc tính đều bị lệch phải. Có các giá trị có số lượng cao hơn nhiều so với các giá trị còn lại là do quá trình xử lý missing values điền bằng mean.

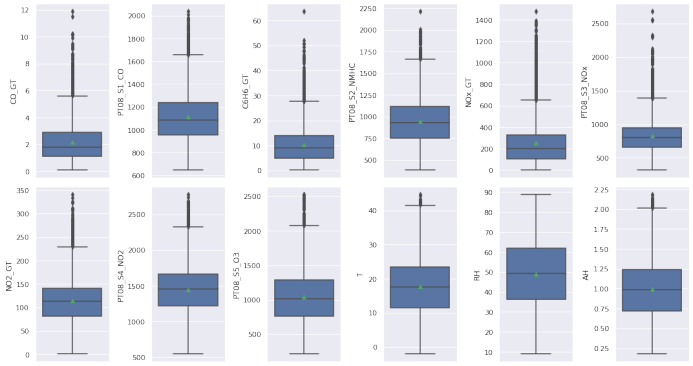
Hình 8b là histogram của bộ dữ liệu xử lý theo chiến lượt MEAN. Ta thấy, hầu hết hình dạng biểu đồ của các thuộc tính cũng đều bị lệch phải như Hình

8a. Vì xử lý missing values bằng cách điền giá trị mean nên số lượng các giá trị mean cao hơn nhiều so với bộ dữ liệu REMOVE, phân phối dữ liệu có những giá trị cao bất thường, điều này có thể ảnh hưởng không tốt đến khả năng dự đoán của mô hình.

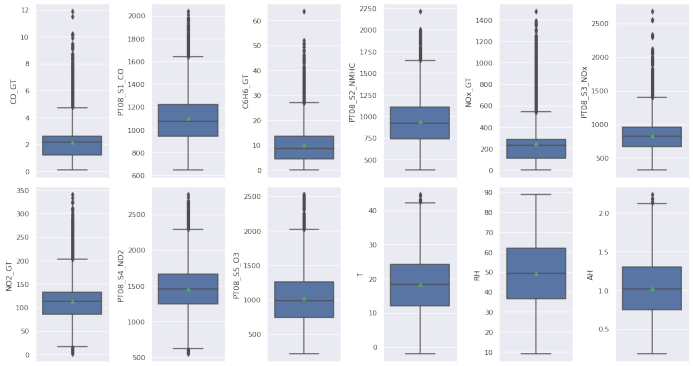
Qua các biểu đồ phân phối dữ liệu ở Hình 8, ta có thể thấy được rằng cách xử lý missing values bằng cách điền các giá trị mean này vẫn chưa được tốt. Hướng giải quyết có thể sẽ mang lại kết quả tốt hơn là tính mean theo ngày/giờ để điền vào các missing value.

3.6 Box plot

Box plot (Biểu đồ hộp) là một loại biểu đồ thể hiện các khuôn hình của dữ liệu định tính, là biểu đồ diễn tả 5 vị trí phân bố của dữ liệu, đó là: giá trị nhỏ nhất (min), tứ phân vị thứ nhất (Q1), trung vị (median), tứ phân vị thứ 3 (Q3) và giá trị lớn nhất (max) [4].



(a) Air Quality–REMOVE.



(b) Air Quality–MEAN.

Hình 9: Box plot.

Box plot giúp biểu diễn các đại lượng quan trọng của dãy số như giá trị nhỏ nhất (min), giá trị lớn nhất (max), tứ phân vị (quartile) [5], khoảng biến thiên tứ phân vị (Interquartile Range) một cách trực quan, dễ hiểu.

Hình 9 Box plot của 2 bộ dữ liệu REMOVE và MEAN. Ta thấy, hầu hết các thuộc tính đều có outlier [6] (trừ thuộc tính RH). Trong đó các thuộc tính CO_GT, PT08_S1_CO, C6H6_GT, NOx_GT, PT08_S3_NOx, NO2_GT có số lượng outlier lớn (nhiều).

4 Hướng Tiếp Cận

Từ bộ dữ liệu Air Quality gốc, sau quá trình xử lý missing values ta thu được 2 bộ dữ liệu mới là REMOVE và MEAN. Qua quá trình phân tích ANOVA chúng tôi chọn được các biến đơn thuộc tính và xét sự tương tác 2 thuộc tính của tất cả các thuộc tính có ảnh hưởng lớn đến mô hình hồi quy. Trước khi huấn luyện mô hình hồi quy, chúng tôi tiến hành chuẩn hóa dữ liệu áp dụng kỹ thuật *Feature Scaling* là *Standardization* (được trình bày rõ hơn trong Phần 4.1). Để tạo ra mô hình hồi quy cho bài toán dự đoán nồng độ CO trong không khí, chúng tôi tiến hành xây dựng các mô hình *Machine Learning (Máy Học)* và Deep *Learning (Học Sâu)* như: Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Machine, Neural Network. Để đánh giá mô hình chúng tôi sử dụng 4 độ đo chính là: R-Squared, Mean Squared Error, Root Mean Squared Error, Mean Absolute Error (chi tiết Phần 4.7).

4.1 Feature Scaling

Chuẩn hóa dữ liệu là một khái niệm chung đề cập đến hành động chuyển đổi các giá trị ban đầu của tập dữ liệu thành các giá trị mới. Các giá trị mới thường được mã hóa liên quan đến chính tập dữ liệu và được chia tỷ lệ theo một cách nào đó.

Data standardization là một loại kỹ thuật chuẩn hóa cụ thể. Nó đôi khi được gọi là z-score normalization. z-score còn gọi là điểm tiêu chuẩn, là giá trị được biến đổi cho mỗi điểm dữ liệu. Để chuẩn hóa tập dữ liệu bằng cách sử dụng standardization, chúng ta lấy mọi giá trị x_i bên trong tập dữ liệu và biến đổi nó thành giá trị z_i tương ứng bằng công thức sau:

$$z_i = \frac{x_i - \mu_k}{\sigma_k} \tag{2}$$

Trong đó: μ_k và σ_k lần lượt là trung bình và độ lệch chuẩn của từng thuộc tính thứ k. Quá trình standardization này chuyển giá trị trung bình của tập dữ liệu thành 0 và độ lệch chuẩn của nó thành 1.

[5] Tứ phân vị là đại lượng mô tả sự phân bố và sự phân tán của tập dữ liệu. Tứ phân vị có 3 giá trị, đó là tứ phân vị thứ nhất, thứ nhì, và thứ ba. 3 giá trị này chia một tập hợp dữ liệu (đã sắp xếp dữ liệu theo trật từ từ bé đến lớn) thành 4 phần có số lượng quan sát đều nhau.

[6] *Outlier (Ngoại lệ)* là điểm dữ liệu khác biệt đáng kể so với các điểm dữ liệu khác trong tập dữ liệu.

4.2 Linear Regression

Hồi Quy Tuyến Tính (Linear Regression) là một phương pháp phân tích quan hệ giữa biến phụ thuộc Y với một hay nhiều biến độc lập X. Mô hình hóa sử dụng hàm tuyến tính (bậc 1). Các tham số của mô hình (hay hàm số) được ước lượng từ dữ liệu. Hồi quy tuyến tính được sử dụng rộng rãi trong thực tế do tính chất đơn giản hóa của hồi quy. Nó cũng dễ ước lượng. Phương trình hồi quy tuyến tính có dạng như sau:

$$y = w_0 + w_1 \times x_1 + w_2 \times x_2 \cdots + w_n \times x_n \tag{3}$$

Trong đó: y là biến phụ thuộc, $x_1, x_2, \ldots x_n$ là các biến độc lập, $w_0, w_1, w_2, \ldots w_n$ là các hệ số (tham số) của mô hình. Bài toán đi tìm các hệ số tối ưu $\{w_0, w_1, w_2, \ldots w_n\}$ chính vì vậy được gọi là bài toán Linear Regression.

4.3 Decision Tree Regression

Trong lĩnh vực máy học, *cây quyết định (Decision Tree)* là một kiểu mô hình dự báo (predictive model), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/hiện tượng. Mỗi một nút trong (internal node) tương ứng với một biến; đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó. Kỹ thuật học máy dùng trong cây quyết định được gọi là học bằng cây quyết định, hay chỉ gọi với cái tên ngắn gọn là cây quyết định [5].

Dữ liệu được cho dưới dạng các bản ghi có dạng:

$$(x, y) = (x_1, x_2, x_3, \ldots, x_k, y)$$

Biến phụ thuộc y là biến mà chúng ta cần tìm hiểu, phân loại hay dự đoán. $x_1, x_2, x_3, \ldots$ là các biến sẽ giúp ta thực hiện công việc đó.

Cây hồi quy (Regression tree) ước lượng các hàm giá có giá trị là số thực thay vì được sử dụng cho các nhiệm vụ phân loại. (ví dụ: ước tính giá một ngôi nhà hoặc khoảng thời gian một bệnh nhân nằm viện)

4.4 Random Forest Regression

Rừng ngẫu nhiên (Random forests) hoặc *rừng quyết định ngẫu nhiên (random decision forests)* là một phương pháp học tập tổng hợp để *phân loại (classification)*, *hồi quy (regression)* và các tác vụ khác hoạt động bằng cách xây dựng vô số cây quyết định tại thời điểm huấn luyện. Kết quả đầu ra của mô hình Random Forest được tổng hợp từ kết quả của các cây quyết định mà nó tạo ra. Đối với các tác vụ hồi quy, giá trị trung bình hoặc dự đoán trung bình của các cây riêng lẻ được trả về. Để hiểu rõ hơn về thuật toán Rừng ngẫu nhiên, hãy cùng tìm hiểu các bước sau:

– Bước 1: Chọn số lượng cây quyết định muốn tạo, gọi là n.
– Bước 2: Xây dựng n cây quyết định, với mỗi cây:
  • Bước 2.1: Chọn K điểm dữ liệu ngẫu nhiên trong tập dữ liệu.
  • Bước 2.2: Xây dựng cây quyết định dựa trên K điểm dữ liệu được chọn.
– Bước 3: Đối với một điểm dữ liệu mới, ta thực hiện dự đoán trên tất cả cây quyết định xây dựng được. Kết quả đầu ra của điểm dữ liệu này có thể được lấy là trung bình cộng dự đoán của tất cả các cây quyết định.

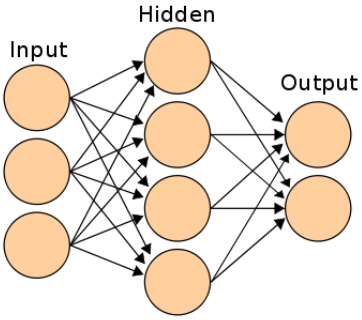
4.5 Support Vector Regression

Support Vector Machine (SVM) là một khái niệm trong thống kê và khoa học máy tính cho một tập hợp các phương pháp học có giám sát liên quan đến nhau để phân loại và phân tích hồi quy. SVM dạng chuẩn nhận dữ liệu vào và phân loại chúng vào hai lớp khác nhau. Do đó SVM là một thuật toán phân loại nhị phân. Với một bộ các ví dụ luyện tập thuộc hai thể loại cho trước, thuật toán luyện tập SVM xây dựng một mô hình SVM để phân loại các ví dụ khác vào hai thể loại đó. Một mô hình SVM là một cách biểu diễn các điểm trong không gian và lựa chọn ranh giới giữa hai thể loại sao cho khoảng cách từ các ví dụ luyện tập tới ranh giới là xa nhất có thể. Các ví dụ mới cũng được biểu diễn trong cùng một không gian và được thuật toán dự đoán thuộc một trong hai thể loại tùy vào ví dụ đó nằm ở phía nào của ranh giới [6].

Support Vector Regression (SVR) là một thuật toán Regression (hồi quy) dựa trên SVM. Thuật toán SVR (hay SVM) sẽ tìm một số vector đặc biệt (gọi là support vectors). Mô hình dự đoán kết quả đầu ra của những điểm dữ liệu mới dựa trên các vector đặc biệt này.

4.6 Artificial Neural Network

Mạng neural nhân tạo (Artificial Neural Network – ANN) hay thường gọi ngắn gọn là *mạng neural (neural network - NN)* là một mô hình toán học hay mô hình tính toán được xây dựng dựa trên các mạng neural sinh học. Nó gồm có một nhóm các neural nhân tạo (nút) nối với nhau, và xử lý thông tin bằng cách truyền theo các kết nối và tính giá trị mới tại các nút (cách tiếp cận connectionism đối với tính toán). Trong nhiều trường hợp, mạng neural nhân tạo là một hệ thống thích ứng (adaptive system) tự thay đổi cấu trúc của mình dựa trên các thông tin bên ngoài hay bên trong chảy qua mạng trong quá trình học. Trong thực tế sử dụng, nhiều mạng neural là các công cụ mô hình hóa dữ liệu thống kê phi tuyến. Chúng có thể được dùng để mô hình hóa các mối quan hệ phức tạp giữa dữ liệu vào và kết quả hoặc để tìm kiếm các dạng/mẫu trong dữ liệu [7].



Hình 10: Kiến trúc ANN

Các lớp trong một Neural Network:

- Input layer: Nhận dữ liệu đầu vào.
- Kết nối giữa các layer với nhau, gồm: input layer, các lớp hiden layer khác và output layer.
- Output layer: Đưa ra kết quả từ dữ liệu đầu vào. Dữ liệu kết quả có thể là:
  - Label: dạng categorical đối với bài toán phân lớp (classification).
  - Value: dạng numeric đối với bài toán hồi quy (regression) hay xếp hạng (ranking).

4.7 Độ đo đánh giá

R-Squared Trong thống kê, hệ số xác định, được ký hiệu là R^2 hoặc r^2 và được gọi là '*R squared (R bình phương)*' là tỷ lệ của phương sai trong biến phụ thuộc có thể dự đoán được từ (các) biến độc lập. Nó là một thống kê được sử dụng trong bối cảnh của các mô hình thống kê có mục đích chính là dự đoán các kết quả trong tương lai hoặc kiểm tra các giả thuyết, trên cơ sở các thông tin liên quan khác. Nó cung cấp một thước đo về mức độ nhân rộng của các kết quả quan sát được của mô hình, dựa trên tỷ lệ của tổng biến động của các kết quả được mô hình giải thích [8].

Cho một bộ dữ liệu có n giá trị $y_1, y_2, \ldots, y_n$ (được gọi chung là y_i hoặc dưới dạng vectơ $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$), mỗi giá trị được liên kết với một giá trị được điều chỉnh (hoặc được mô hình hóa hoặc được dự đoán) $f_1, f_2, \ldots, f_n$ (được gọi là f_i, hoặc đôi khi là $\hat{y}_i$, như một vectơ $\mathbf{f}$).

Giá trị trung bình của dữ liệu quan sát:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4}$$

Hàm đánh giá R^2 được định nghĩa là:

$$R^2 = 1 - \frac{SSE_{model}}{SSE_{baseline}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{5}$$

Trong đó:

- SSE_{model} là giá trị hàm lỗi Squared Sum của tập dữ liệu khi đánh giá trên mô hình đang xét.
- $SSE_{baseline}$ là giá trị hàm lỗi Squared Sum của tập dữ liệu khi đánh giá trên mô hình baseline[7] (đường cơ sở).

Giá trị của hệ số R^2 luôn nằm trong đoạn $(-\infty, 1]$:

- Nếu $R^2 < 0$: Mô hình tệ hơn mô hình đường cơ sở.
- Nếu $R^2 = 0$: Mô hình giống như mô hình cơ sở.
- Nếu $R^2 = 1$: Mô hình chính xác tuyệt đối.

R^2 càng lớn (càng gần 1) thì độ chính xác của mô hình với tập dữ liệu đang xét càng cao. Một mô hình được xem là tốt nếu $R^2 > 0.8$.

Nhược điểm:

- Khi sử dụng hàm đánh giá R Squared để so sánh hai mô hình với số lượng đặc trưng đầu vào khác nhau, thì mô hình với số lượng đặc trưng đầu vào lớn hơn (gần như luôn luôn) cho giá trị R Squared lớn hơn. Vì vậy, hàm đánh giá R squared sẽ (gần như luôn luôn) nói rằng, mô hình nhận nhiều đặc trưng đầu vào hơn là tốt hơn, cho dù có một số đặc trưng không tương quan với kết quả đầu ra.
- Khi dữ liệu quá ít, giá trị của hàm R Squared sẽ không ổn định và không đáng tin cậy.

Mean Squared Error (MSE) Trong thống kê học, sai số toàn phương trung bình, viết tắt MSE (Mean squared error) của một phép ước lượng là trung bình của bình phương các sai số, tức là sự khác biệt giữa các ước lượng và những gì được đánh giá. MSE là một hàm rủi ro, tương ứng với giá trị kỳ vọng của sự mất mát sai số bình phương hoặc mất mát bậc hai. Sự khác biệt xảy ra do ngẫu nhiên, hoặc vì các ước lượng không tính đến thông tin có thể cho ra một ước tính chính xác hơn. MSE đánh giá chất lượng của một ước lượng (ví dụ, một hàm toán học lập bản đồ mẫu dữ liệu của một tham số của dân số từ đó các dữ liệu được lấy mẫu) hoặc một yếu tố dự báo (ví dụ, một bản đồ chức năng có số liệu vào tùy ý để một mẫu của các giá trị của một số biến ngẫu nhiên). Định nghĩa của một MSE khác với những gì tương ứng cho dù là một trong những mô tả một ước lượng, hay một yếu tố dự báo [9].

Nếu $\hat{Y}$ là một vector của n giá trị dự báo, và Y là vector các giá trị quan sát được của biến dự đoán, thì MSE của phép dự báo có thể ước lượng theo công thức:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{6}$$

Tức là MSE là trung bình ($\frac{1}{n}\sum_{i=1}^{n}$) của bình phương các sai số ($(Y_i - \hat{Y})^2$). Đây là định lượng dễ dàng tính được cho một mẫu cụ thể (phụ thuộc mẫu).

[7] Mô hình này chỉ dự đoán một kết quả đầu ra duy nhất đó cho mọi điểm dữ liệu đầu vào, đó là giá trị trung bình của tất cả các kết quả đầu ra trong tập dữ liệu.

Root Mean Squared Error (RMSE) Là một biện pháp thường được sử dụng trong những khác biệt giữa các giá trị (các giá trị mẫu hoặc tổng thể) được dự đoán bởi một mô hình hay một ước lượng và các giá trị quan sát được. RMSE đại diện cho căn bậc hai của thời điểm mẫu thứ hai về sự khác biệt giữa các giá trị dự đoán và giá trị quan sát hoặc giá trị trung bình bậc hai của những khác biệt này. Các độ lệch này được gọi là phần dư khi các phép tính được thực hiện trên mẫu dữ liệu được sử dụng để ước tính và được gọi là lỗi (hoặc lỗi dự đoán) khi tính toán ngoài mẫu. RMSE phục vụ để tổng hợp cường độ của các lỗi trong các dự đoán trong nhiều thời điểm khác nhau thành một thước đo duy nhất về sức mạnh dự đoán. RMSE là thước đo độ chính xác, để so sánh các lỗi dự báo của các mô hình khác nhau cho một tập dữ liệu cụ thể chứ không phải giữa các bộ dữ liệu, vì nó phụ thuộc vào quy mô [10].

RMSE được tính như sau:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{7}$$

Trong đó: n là số lượng điểm dữ liệu trong tập dữ liệu, y_i là kết quả thực của điểm dữ liệu thứ i, $\hat{y}_i$ là kết quả dự đoán cởi mô hình của điểm dữ liệu thứ i.

RMSE được xem như là độ sai lệch trung bình giữa đầu ra dự đoán với đầu ra thực. RMSE có giá trị trong khoảng $[0, +\infty)$:

- RMSE = 0: Mô hình chính xác tuyệt đối.
- RMSE càng nhỏ, độ chính xác của mô hình càng cao.

RMSE thường được sử dụng nhiều vì: không phụ thuộc vào số lượng điểm dữ liệu và cùng đơn vị với kết quả đầu ra. Nhược điểm: Phụ thuộc vào miền giá trị đầu ra của dữ liệu.

Mean Absolute Error (MAE) Trong thống kê, MAE là một phương pháp đo lường sự khác biệt giữa hai biến liên tục. Giả sử rằng X và Y là hai biến liên tục thể hiện kết quả dự đoán của mô hình và kết quả thực tế. MAE được tính bằng tổng sai số tuyệt đối chia cho cỡ mẫu [11]:

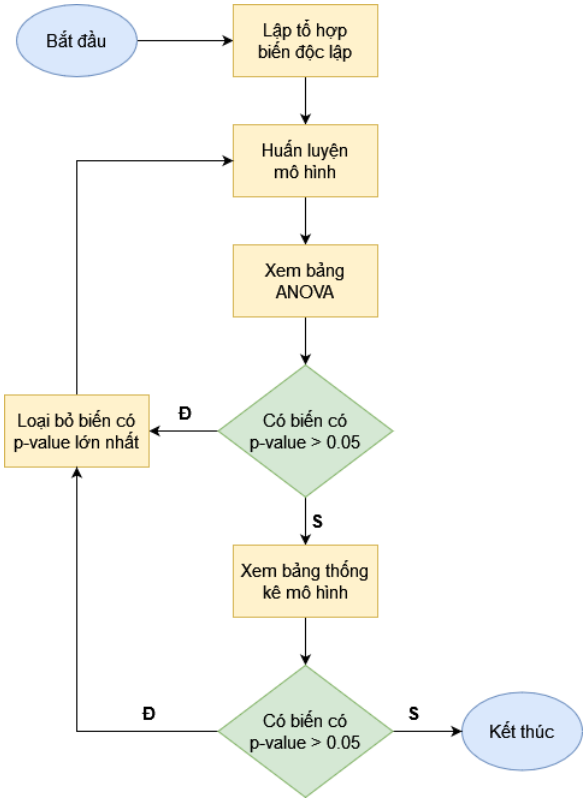
$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n} \tag{8}$$

Do đó, nó là một trung bình cộng của các sai số tuyệt đối $|e_i| = |y_i - x_i|$. MAE sử dụng cùng thang đo với dữ liệu được đo. Đây được gọi là thước đo độ chính xác phụ thuộc vào thang đo và do đó không thể được sử dụng để so sánh giữa các chuỗi sử dụng các thang đo khác nhau. Độ đo này thường được sử dụng để đánh giá sự sai khác giữa mô hình dự đoán và tập dữ liệu testing trong các bài toán hồi quy. Chỉ số này càng nhỏ thì mô hình học máy càng chính xác.

5 Phân Tích ANOVA

Đối với từng bộ dữ liệu, chúng tôi tiến hành xây dựng và chọn lọc trên 2 loại mô hình:

- Mô hình 1: Áp dụng quy trình ANOVA (Hình 11) để loại bỏ các thuộc tính ít ảnh hưởng đến mô hình ($p - value > 0.05$).
- Mô hình 2: Xét tương tác 2 thuộc tính của tất cả các thuộc tính trong bộ dữ liệu và tiến hành quy trình ANOVA (Hình 11) để loại bỏ các đơn thuộc tính và các thuộc tính tương tác ít ảnh hưởng đến mô hình ($p - value > 0.05$).



Hình 11: Quy trình ANOVA.

Quy trình xây dựng và chọc lọc các thuộc tính có ý nghĩa theo các bước sau:

- Bước 1: Lập tổ hợp thuộc tính (biến độc lập).
- Bước 2: Khớp mô hình hồi quy với tổ hợp thuộc tính đã xác định.
- Bước 3: Nếu có thuộc tính có $p - value > 0.05$ trong bảng ANOVA, thì ta sẽ loại bỏ thuộc tính có $p - value$ cao nhất và quay lại Bước 2.

– Bước 4: Nếu có thuộc tính có $p - value > 0.05$ trong bảng Fitting Linear Models, thì ta sẽ loại bỏ thuộc tính có $p - value$ cao nhất và quay lại Bước 2.
– Bước 5: Khi không còn thuộc tính nào có $p - value > 0.05$ thì ta sẽ kết thúc quy trình ANOVA và thu được tổ hợp các thuộc tính có ảnh hưởng lớn đến mô hình.

5.1 Bộ dữ liệu Air Quality–REMOVE.

ANOVA đơn thuộc tính Thực hiện phân tích ANOVA với tất cả các thuộc tính của bộ dữ liệu.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
PT08_S1_CO	1	11692	11692	48826.044	< 2e-16	***
C6H6_GT	1	1672	1672	6980.888	< 2e-16	***
PT08_S2_NMHC	1	15	15	64.576	1.07e-15	***
NOx_GT	1	781	781	3261.151	< 2e-16	***
PT08_S3_NOx	1	44	44	184.058	< 2e-16	***
NO2_GT	1	40	40	166.288	< 2e-16	***
PT08_S4_NO2	1	28	28	118.545	< 2e-16	***
PT08_S5_O3	1	29	29	123.125	< 2e-16	***
T	1	30	30	124.825	< 2e-16	***
RH	1	37	37	154.972	< 2e-16	***
AH	1	1	1	3.013	0.0827	.
Residuals	7662	1835	0			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Hình 12: Bảng ANOVA các thuộc tính mô hình 1.

Ở lần đầu tiên thực hiện phân tích ANOVA, có duy nhất 1 thuộc tính có $p - value > 0.05$ là AH với $p - value = 0.0827$, do vậy ta sẽ loại bỏ thuộc tính này và tiếp tục thực hiện phân tích. Sau khi loại bỏ hết các thuộc tính có $p - value > 0.05$ ta thu được kết quả như Hình 13.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
PT08_S1_CO	1	11692	11692	48813.22	< 2e-16	***
C6H6_GT	1	1672	1672	6979.06	< 2e-16	***
PT08_S2_NMHC	1	15	15	64.56	1.08e-15	***
NOx_GT	1	781	781	3260.30	< 2e-16	***
PT08_S3_NOx	1	44	44	184.01	< 2e-16	***
NO2_GT	1	40	40	166.24	< 2e-16	***
PT08_S4_NO2	1	28	28	118.51	< 2e-16	***
PT08_S5_O3	1	29	29	123.09	< 2e-16	***
T	1	30	30	124.79	< 2e-16	***
RH	1	37	37	154.93	< 2e-16	***
Residuals	7663	1836	0			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Hình 13: Bảng ANOVA các thuộc tính mô hình 1.

Sau đó, chúng tôi kiểm tra mức độ ý nghĩa của các hệ số theo bảng Fitting Linear Models và loại bỏ những thuộc tính không có quan hệ tuyến tính với CO_GT.

```
Call:
lm(formula = CO_GT ~ PT08_S1_CO + C6H6_GT + PT08_S2_NMHC + NOx_GT +
    PT08_S3_NOx + NO2_GT + PT08_S4_NO2 + PT08_S5_O3 + T + RH,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3698 -0.1946  0.0124  0.1918  4.3029

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.318e+00  1.397e-01  -9.437  < 2e-16 ***
PT08_S1_CO    1.371e-03  7.355e-05  18.642  < 2e-16 ***
C6H6_GT       8.795e-02  4.928e-03  17.846  < 2e-16 ***
PT08_S2_NMHC -7.818e-05  1.636e-04  -0.478 0.632713
NOx_GT        2.408e-03  5.951e-05  40.467  < 2e-16 ***
PT08_S3_NOx   1.573e-04  4.734e-05   3.322 0.000897 ***
NO2_GT        2.519e-03  2.335e-04  10.789  < 2e-16 ***
PT08_S4_NO2   1.056e-03  5.662e-05  18.647  < 2e-16 ***
PT08_S5_O3   -5.247e-04  4.088e-05 -12.835  < 2e-16 ***
T            -2.692e-02  1.610e-03 -16.721  < 2e-16 ***
RH           -8.410e-03  6.757e-04 -12.447  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4894 on 7663 degrees of freedom
Multiple R-squared:  0.8867,    Adjusted R-squared:  0.8866
F-statistic:  5999 on 10 and 7663 DF,  p-value: < 2.2e-16
```

Hình 14: Bảng Fitting Linear Models các thuộc tính mô hình 1.

Hình 14 ta thấy có 1 thuộc tính có $p - value > 0.05$, do vậy ta sẽ loại bỏ thuộc tính này và thực hiện lại phân tích ANOVA. Kết quả thu được như Hình 15.

```
             Df Sum Sq Mean Sq F value Pr(>F)
PT08_S1_CO    1  11692   11692 48818.1 <2e-16 ***
C6H6_GT       1   1672    1672  6979.8 <2e-16 ***
NOx_GT        1    795     795  3320.1 <2e-16 ***
PT08_S3_NOx   1     39      39   161.0 <2e-16 ***
NO2_GT        1     45      45   188.9 <2e-16 ***
PT08_S4_NO2   1     29      29   122.4 <2e-16 ***
PT08_S5_O3    1     30      30   124.5 <2e-16 ***
T             1     28      28   117.9 <2e-16 ***
RH            1     39      39   161.8 <2e-16 ***
Residuals  7664   1836       0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hình 15: Bảng ANOVA các thuộc tính mô hình 1.

Ta tiếp tục quay lại kiểm tra mức độ ý nghĩa của các hệ số theo bảng Fitting Linear Models. Vì không còn thuộc tính nào có $p - value > 0.05$ nữa (Hình 16). Nên ta sẽ kết thúc quá trình phân tích ANOVA.

```
Call:
lm(formula = CO_GT ~ PT08_S1_CO + C6H6_GT + NOx_GT + PT08_S3_NOx +
    NO2_GT + PT08_S4_NO2 + PT08_S5_O3 + T + RH, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3526 -0.1951  0.0126  0.1923  4.3041

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.368e+00  9.354e-02 -14.625  < 2e-16 ***
PT08_S1_CO   1.370e-03  7.350e-05  18.637  < 2e-16 ***
C6H6_GT      8.601e-02  2.810e-03  30.613  < 2e-16 ***
NOx_GT       2.409e-03  5.949e-05  40.492  < 2e-16 ***
PT08_S3_NOx  1.681e-04  4.157e-05   4.044 5.31e-05 ***
NO2_GT       2.502e-03  2.307e-04  10.843  < 2e-16 ***
PT08_S4_NO2  1.048e-03  5.430e-05  19.304  < 2e-16 ***
PT08_S5_O3  -5.286e-04  4.004e-05 -13.200  < 2e-16 ***
T           -2.687e-02  1.607e-03 -16.723  < 2e-16 ***
RH          -8.331e-03  6.550e-04 -12.720  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4894 on 7664 degrees of freedom
Multiple R-squared:  0.8867,    Adjusted R-squared:  0.8866
F-statistic:  6666 on 9 and 7664 DF,  p-value: < 2.2e-16
```

Hình 16: Bảng Fitting Linear Models các thuộc tính mô hình 1.

Mô hình hồi quy có dạng:

$$\begin{aligned} CO\_GT = & -1.368e^{+00} + 1.370e^{-03} \times PT08\_S1\_CO \\ & + 8.601e^{-02} \times C6H6\_GT + 2.409e^{-03} \times NOx\_GT \\ & + 1.681e^{-04} \times PT08\_S3\_NOx + 2.502e^{-03} \times NO2\_GT \\ & + 1.048e^{-03} \times PT08\_S4\_NO2 - 5.286e^{-04} \times PT08\_S5\_O3 \\ & - 2.687e^{-02} \times T - 8.331e^{-03} \times RH \end{aligned}$$

ANOVA tương tác 2 thuộc tính Thực hiện phân tích ANOVA với tất cả các thuộc tính và kết hợp với các tương tác 2 thuộc tính của tất cả các thuộc tính trong bộ dữ liệu. Kết quả thực hiện ANOVA lần đầu tiên được thể hiện ở Hình 17.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
PT08_S1_CO	1	11692	11692	60183.325	< 2e-16	***
C6H6_GT	1	1672	1672	8604.692	< 2e-16	***
PT08_S2_NMHC	1	15	15	79.597	< 2e-16	***
NOx_GT	1	781	781	4019.718	< 2e-16	***
PT08_S3_NOx	1	44	44	226.871	< 2e-16	***
NO2_GT	1	40	40	204.967	< 2e-16	***
PT08_S4_NO2	1	28	28	146.119	< 2e-16	***
PT08_S5_O3	1	29	29	151.765	< 2e-16	***
T	1	30	30	153.860	< 2e-16	***
RH	1	37	37	191.019	< 2e-16	***
AH	1	1	1	3.714	0.054011	.
I(PT08_S1_CO * C6H6_GT)	1	6	6	32.460	1.26e-08	***
I(PT08_S1_CO * PT08_S2_NMHC)	1	2	2	10.936	0.000947	***
I(PT08_S1_CO * NOx_GT)	1	39	39	199.233	< 2e-16	***
I(PT08_S1_CO * PT08_S3_NOx)	1	5	5	27.089	1.99e-07	***
I(PT08_S1_CO * NO2_GT)	1	18	18	93.315	< 2e-16	***
I(PT08_S1_CO * PT08_S4_NO2)	1	53	53	274.135	< 2e-16	***
I(PT08_S1_CO * PT08_S5_O3)	1	2	2	12.457	0.000419	***
I(PT08_S1_CO * T)	1	2	2	8.606	0.003361	**
I(PT08_S1_CO * RH)	1	0	0	0.021	0.883590	
I(PT08_S1_CO * AH)	1	0	0	0.480	0.488533	
I(C6H6_GT * PT08_S2_NMHC)	1	30	30	152.132	< 2e-16	***
I(C6H6_GT * NOx_GT)	1	0	0	0.032	0.858714	
I(C6H6_GT * PT08_S3_NOx)	1	1	1	3.992	0.045747	*
I(C6H6_GT * NO2_GT)	1	2	2	9.581	0.001974	**
I(C6H6_GT * PT08_S4_NO2)	1	11	11	58.721	2.04e-14	***
I(C6H6_GT * PT08_S5_O3)	1	10	10	53.966	2.25e-13	***
I(C6H6_GT * T)	1	0	0	2.188	0.139116	
I(C6H6_GT * RH)	1	4	4	19.197	1.19e-05	***
I(C6H6_GT * AH)	1	5	5	23.768	1.11e-06	***
I(PT08_S2_NMHC * NOx_GT)	1	5	5	27.544	1.58e-07	***
I(PT08_S2_NMHC * PT08_S3_NOx)	1	0	0	1.270	0.259766	
I(PT08_S2_NMHC * NO2_GT)	1	29	29	148.104	< 2e-16	***
I(PT08_S2_NMHC * PT08_S4_NO2)	1	5	5	24.433	7.86e-07	***
I(PT08_S2_NMHC * PT08_S5_O3)	1	2	2	10.600	0.001136	**
I(PT08_S2_NMHC * T)	1	4	4	19.943	8.09e-06	***
I(PT08_S2_NMHC * RH)	1	0	0	1.502	0.220447	
I(PT08_S2_NMHC * AH)	1	3	3	17.243	3.32e-05	***
I(NOx_GT * PT08_S3_NOx)	1	1	1	4.205	0.040330	*
I(NOx_GT * NO2_GT)	1	20	20	104.257	< 2e-16	***
I(NOx_GT * PT08_S4_NO2)	1	32	32	167.267	< 2e-16	***
I(NOx_GT * PT08_S5_O3)	1	1	1	3.402	0.065161	.
I(NOx_GT * T)	1	0	0	0.606	0.436145	
I(NOx_GT * RH)	1	3	3	16.177	5.83e-05	***
I(NOx_GT * AH)	1	1	1	5.816	0.015902	*
I(PT08_S3_NOx * NO2_GT)	1	1	1	4.151	0.041641	*
I(PT08_S3_NOx * PT08_S4_NO2)	1	2	2	7.861	0.005065	**
I(PT08_S3_NOx * PT08_S5_O3)	1	1	1	6.267	0.012319	*
I(PT08_S3_NOx * T)	1	5	5	25.390	4.79e-07	***
I(PT08_S3_NOx * RH)	1	0	0	0.029	0.865580	
I(PT08_S3_NOx * AH)	1	2	2	12.833	0.000343	***
I(NO2_GT * PT08_S4_NO2)	1	12	12	63.288	2.04e-15	***
I(NO2_GT * PT08_S5_O3)	1	1	1	4.378	0.036442	*
I(NO2_GT * T)	1	1	1	5.211	0.022475	*
I(NO2_GT * RH)	1	6	6	32.979	9.68e-09	***
I(NO2_GT * AH)	1	0	0	0.200	0.654343	
I(PT08_S4_NO2 * PT08_S5_O3)	1	18	18	91.679	< 2e-16	***
I(PT08_S4_NO2 * T)	1	2	2	9.546	0.002011	**
I(PT08_S4_NO2 * RH)	1	0	0	0.272	0.602290	
I(PT08_S4_NO2 * AH)	1	1	1	7.478	0.006261	**
I(PT08_S5_O3 * T)	1	0	0	0.046	0.830234	
I(PT08_S5_O3 * RH)	1	0	0	1.532	0.215878	
I(PT08_S5_O3 * AH)	1	1	1	6.830	0.008983	**
I(T * RH)	1	0	0	0.179	0.672363	
I(T * AH)	1	4	4	20.158	7.23e-06	***
I(RH * AH)	1	0	0	2.251	0.133551	
Residuals	7607	1478	0			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Hình 17: Bảng ANOVA các thuộc tính mô hình 2.

Ta có thể thấy rằng có rất nhiều thuộc tính có $p-value > 0.05$, vì vậy ta cần thực hiện ANOVA nhiều lần để tìm ra các thuộc tính có ý nghĩa. Sau khoảng 30 lần thực hiện ANOVA để loại bỏ từng thuộc tính có $p-value > 0.05$, ta thu được kết quả như Hình 18.

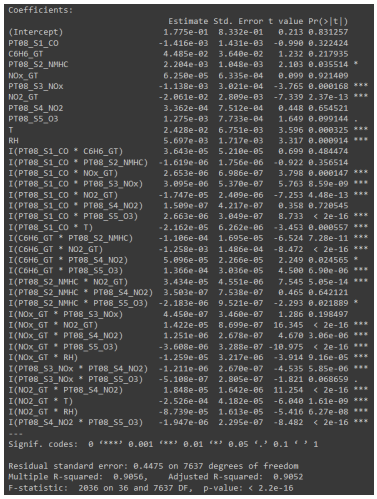
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
PT08_S1_CO	1	11692	11692	58388.607	< 2e-16	***
C6H6_GT	1	1672	1672	8348.092	< 2e-16	***
PT08_S2_NMHC	1	15	15	77.223	< 2e-16	***
NOx_GT	1	781	781	3899.846	< 2e-16	***
PT08_S3_NOx	1	44	44	220.106	< 2e-16	***
NO2_GT	1	40	40	198.855	< 2e-16	***
PT08_S4_NO2	1	28	28	141.762	< 2e-16	***
PT08_S5_O3	1	29	29	147.239	< 2e-16	***
T	1	30	30	149.272	< 2e-16	***
RH	1	37	37	185.323	< 2e-16	***
I(PT08_S1_CO * C6H6_GT)	1	6	6	31.242	2.36e-08	***
I(PT08_S1_CO * PT08_S2_NMHC)	1	2	2	11.841	0.000583	***
I(PT08_S1_CO * NOx_GT)	1	37	37	184.943	< 2e-16	***
I(PT08_S1_CO * PT08_S3_NOx)	1	6	6	29.328	6.30e-08	***
I(PT08_S1_CO * NO2_GT)	1	19	19	94.298	< 2e-16	***
I(PT08_S1_CO * PT08_S4_NO2)	1	52	52	260.034	< 2e-16	***
I(PT08_S1_CO * PT08_S5_O3)	1	3	3	12.773	0.000354	***
I(PT08_S1_CO * T)	1	3	3	12.860	0.000338	***
I(C6H6_GT * PT08_S2_NMHC)	1	30	30	147.690	< 2e-16	***
I(C6H6_GT * NO2_GT)	1	1	1	5.319	0.021116	*
I(C6H6_GT * PT08_S4_NO2)	1	13	13	64.174	1.31e-15	***
I(C6H6_GT * PT08_S5_O3)	1	10	10	51.350	8.45e-13	***
I(PT08_S2_NMHC * NO2_GT)	1	6	6	30.860	2.87e-08	***
I(PT08_S2_NMHC * PT08_S4_NO2)	1	4	4	18.517	1.70e-05	***
I(PT08_S2_NMHC * PT08_S5_O3)	1	4	4	19.759	8.91e-06	***
I(NOx_GT * PT08_S3_NOx)	1	3	3	13.056	0.000304	***
I(NOx_GT * NO2_GT)	1	22	22	107.486	< 2e-16	***
I(NOx_GT * PT08_S4_NO2)	1	30	30	151.411	< 2e-16	***
I(NOx_GT * PT08_S5_O3)	1	7	7	34.168	5.26e-09	***
I(NOx_GT * RH)	1	3	3	14.481	0.000143	***
I(PT08_S3_NOx * PT08_S4_NO2)	1	2	2	10.834	0.001001	**
I(PT08_S3_NOx * PT08_S5_O3)	1	3	3	13.128	0.000293	***
I(NO2_GT * PT08_S4_NO2)	1	18	18	89.646	< 2e-16	***
I(NO2_GT * T)	1	3	3	16.061	6.19e-05	***
I(NO2_GT * RH)	1	6	6	32.001	1.60e-08	***
I(PT08_S4_NO2 * PT08_S5_O3)	1	14	14	71.942	< 2e-16	***
Residuals	7637	1529	0			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Hình 18: Bảng ANOVA các thuộc tính mô hình 2.

Sau đó, chúng tôi kiểm tra mức độ ý nghĩa của các hệ số theo bảng Fitting Linear Models và loại bỏ những thuộc tính không có quan hệ tuyến tính với CO_GT.

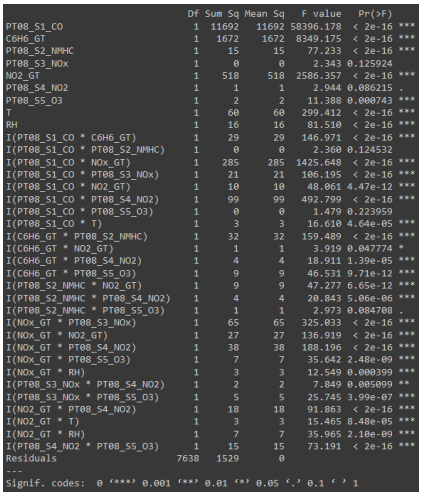
Hình 19a ta thấy có rất thuộc tính đơn và thuộc tính tương tác có $p-value > 0.05$, do vậy ta sẽ loại bỏ thuộc tính có $p-value$ cao nhất và thực hiện lại phân tích ANOVA. Kết quả thu được như Hình 19b.



```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.775e-01  8.332e-01   0.213 0.831257
PT08_S1_CO                    -1.416e-03  1.431e-03  -0.990 0.322424
C6H6_GT                        4.485e-02  3.640e-02   1.232 0.217935
PT08_S2_NMHC                   2.204e-03  1.048e-03   2.103 0.035514 *
NOx_GT                         6.250e-05  6.335e-04   0.099 0.921409
PT08_S3_NOx                   -1.138e-03  3.021e-04  -3.765 0.000168 ***
NO2_GT                        -2.061e-02  2.809e-03  -7.339 2.37e-13 ***
PT08_S4_NO2                    3.362e-04  7.512e-04   0.448 0.654521
PT08_S5_O3                     1.275e-03  7.733e-04   1.649 0.099144
T                              2.428e-02  6.751e-03   3.596 0.000325 ***
RH                             5.697e-03  1.717e-03   3.317 0.000914 ***
I(PT08_S1_CO * C6H6_GT)        3.643e-05  5.210e-05   0.699 0.484474
I(PT08_S1_CO * PT08_S2_NMHC)  -1.619e-06  1.756e-06  -0.922 0.356514
I(PT08_S1_CO * NOx_GT)         2.653e-06  6.986e-07   3.798 0.000147 ***
I(PT08_S1_CO * PT08_S3_NOx)    3.895e-06  5.370e-07   5.763 8.59e-09 ***
I(PT08_S1_CO * NO2_GT)        -1.747e-05  2.409e-06  -7.253 4.48e-13 ***
I(PT08_S1_CO * PT08_S4_NO2)    1.509e-07  4.217e-07   0.358 0.720545
I(PT08_S1_CO * PT08_S5_O3)     2.663e-06  3.049e-07   8.733  < 2e-16 ***
I(PT08_S1_CO * T)             -2.162e-05  6.261e-06  -3.453 0.000557 ***
I(C6H6_GT * PT08_S2_NMHC)     -1.106e-04  1.695e-05  -6.524 7.28e-11 ***
I(C6H6_GT * NO2_GT)           -1.258e-03  1.486e-04  -8.472  < 2e-16 ***
I(C6H6_GT * PT08_S4_NO2)       5.096e-05  2.266e-05   2.249 0.024565 *
I(C6H6_GT * PT08_S5_O3)        1.366e-04  3.036e-05   4.500 6.90e-06 ***
I(PT08_S2_NMHC * NO2_GT)       3.434e-05  4.551e-06   7.545 5.05e-14 ***
I(PT08_S2_NMHC * PT08_S4_NO2)  3.503e-07  7.538e-07   0.465 0.642121
I(PT08_S2_NMHC * PT08_S5_O3)  -2.183e-06  9.521e-07  -2.293 0.021889 *
I(NOx_GT * PT08_S3_NOx)        4.450e-07  3.460e-07   1.286 0.198497
I(NOx_GT * NO2_GT)             1.422e-05  8.699e-07  16.345  < 2e-16 ***
I(NOx_GT * PT08_S4_NO2)        1.251e-06  2.678e-07   4.670 3.06e-06 ***
I(NOx_GT * PT08_S5_O3)        -3.608e-06  3.288e-07 -10.975  < 2e-16 ***
I(NOx_GT * RH)                -1.259e-05  3.217e-06  -3.914 9.16e-05 ***
I(PT08_S3_NOx * PT08_S4_NO2)  -1.211e-06  2.670e-07  -4.535 5.85e-06 ***
I(PT08_S3_NOx * PT08_S5_O3)   -5.108e-07  2.805e-07  -1.821 0.068659 .
I(NO2_GT * PT08_S4_NO2)        1.848e-05  1.642e-06  11.254  < 2e-16 ***
I(NO2_GT * T)                 -2.526e-04  4.182e-05  -6.040 1.61e-09 ***
I(NO2_GT * RH)                -8.739e-05  1.613e-05  -5.416 6.27e-08 ***
I(PT08_S4_NO2 * PT08_S5_O3)   -1.947e-06  2.295e-07  -8.482  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4475 on 7637 degrees of freedom
Multiple R-squared:  0.9056,    Adjusted R-squared:  0.9052
F-statistic:  2036 on 36 and 7637 DF,  p-value: < 2.2e-16
```

(a) Bảng Fitting Linear Models

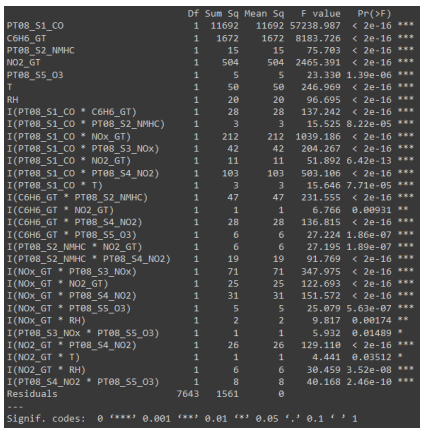


```
                               Df Sum Sq Mean Sq   F value   Pr(>F)
PT08_S1_CO                      1  11692   11692 58396.178  < 2e-16 ***
C6H6_GT                         1   1672    1672  8349.175  < 2e-16 ***
PT08_S2_NMHC                    1     15      15    77.233  < 2e-16 ***
PT08_S3_NOx                     1      0       0     2.343 0.125924
NO2_GT                          1    518     518  2586.357  < 2e-16 ***
PT08_S4_NO2                     1      1       1     2.944 0.086215 .
PT08_S5_O3                      1      2       2    11.388 0.000743 ***
T                               1     60      60   299.412  < 2e-16 ***
RH                              1     16      16    81.510  < 2e-16 ***
I(PT08_S1_CO * C6H6_GT)         1     29      29   146.971  < 2e-16 ***
I(PT08_S1_CO * PT08_S2_NMHC)    1      0       0     2.360 0.124532
I(PT08_S1_CO * NOx_GT)          1    285     285  1425.648  < 2e-16 ***
I(PT08_S1_CO * PT08_S3_NOx)     1     21      21   106.195  < 2e-16 ***
I(PT08_S1_CO * NO2_GT)          1     10      10    48.061 4.47e-12 ***
I(PT08_S1_CO * PT08_S4_NO2)     1     99      99   492.799  < 2e-16 ***
I(PT08_S1_CO * PT08_S5_O3)      1      0       0     1.479 0.223959
I(PT08_S1_CO * T)               1      3       3    16.610 4.64e-05 ***
I(C6H6_GT * PT08_S2_NMHC)       1     32      32   159.489  < 2e-16 ***
I(C6H6_GT * NO2_GT)             1      1       1     3.919 0.047774 *
I(C6H6_GT * PT08_S4_NO2)        1      4       4    18.911 1.39e-05 ***
I(C6H6_GT * PT08_S5_O3)         1      9       9    46.531 9.71e-12 ***
I(PT08_S2_NMHC * NO2_GT)        1      9       9    47.277 6.65e-12 ***
I(PT08_S2_NMHC * PT08_S4_NO2)   1      4       4    20.843 5.06e-06 ***
I(PT08_S2_NMHC * PT08_S5_O3)    1      1       1     2.973 0.084708 .
I(NOx_GT * PT08_S3_NOx)         1     65      65   325.033  < 2e-16 ***
I(NOx_GT * NO2_GT)              1     27      27   136.919  < 2e-16 ***
I(NOx_GT * PT08_S4_NO2)         1     38      38   188.196  < 2e-16 ***
I(NOx_GT * PT08_S5_O3)          1      7       7    35.642 2.48e-09 ***
I(NOx_GT * RH)                  1      3       3    12.549 0.000399 ***
I(PT08_S3_NOx * PT08_S4_NO2)    1      2       2     7.849 0.005099 **
I(PT08_S3_NOx * PT08_S5_O3)     1      5       5    25.745 3.99e-07 ***
I(NO2_GT * PT08_S4_NO2)         1     18      18    91.863  < 2e-16 ***
I(NO2_GT * T)                   1      3       3    15.465 8.48e-05 ***
I(NO2_GT * RH)                  1      7       7    35.965 2.10e-09 ***
I(PT08_S4_NO2 * PT08_S5_O3)     1     15      15    73.191  < 2e-16 ***
Residuals                    7638   1529       0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(b) Bảng ANOVA.

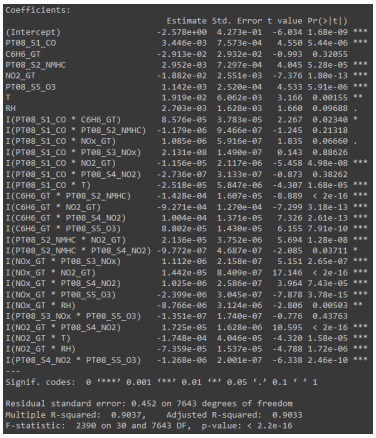
Hình 19: Bảng Fitting Linear Models và ANOVA các thuộc tính mô hình 2.

Tiếp tục thực hiện phân tích ANOVA đến khi loại bỏ hết các thuộc tính có $p-value > 0.05$, ta thu được kết quả như Hình 20a. Sau đó, chúng tôi kiểm tra mức độ ý nghĩa của các hệ số theo bảng Fitting Linear Models và loại bỏ những thuộc tính không có quan hệ tuyến tính với CO_GT. Hình 20b ta thấy vẫn còn vài thuộc tính đơn và thuộc tính tương tác có $p-value > 0.05$, do vậy ta sẽ loại bỏ thuộc tính có $p-value$ cao nhất và thực hiện lại phân tích ANOVA.



```
                               Df Sum Sq Mean Sq   F value   Pr(>F)
PT08_S1_CO                      1  11692   11692 57238.987  < 2e-16 ***
C6H6_GT                         1   1672    1672  8183.726  < 2e-16 ***
PT08_S2_NMHC                    1     15      15    75.703  < 2e-16 ***
NO2_GT                          1    504     504  2465.391  < 2e-16 ***
PT08_S5_O3                      1      5       5    23.330 1.39e-06 ***
T                               1     50      50   246.969  < 2e-16 ***
RH                              1     20      20    96.695  < 2e-16 ***
I(PT08_S1_CO * C6H6_GT)         1     28      28   137.242  < 2e-16 ***
I(PT08_S1_CO * PT08_S2_NMHC)    1      3       3    15.525 8.22e-05 ***
I(PT08_S1_CO * NOx_GT)          1    212     212  1039.186  < 2e-16 ***
I(PT08_S1_CO * PT08_S3_NOx)     1     42      42   204.267  < 2e-16 ***
I(PT08_S1_CO * NO2_GT)          1     11      11    51.892 6.42e-13 ***
I(PT08_S1_CO * PT08_S4_NO2)     1    103     103   503.106  < 2e-16 ***
I(PT08_S1_CO * T)               1      3       3    15.646 7.71e-05 ***
I(C6H6_GT * PT08_S2_NMHC)       1     47      47   231.555  < 2e-16 ***
I(C6H6_GT * NO2_GT)             1      1       1     6.766  0.00931 **
I(C6H6_GT * PT08_S4_NO2)        1     28      28   136.815  < 2e-16 ***
I(C6H6_GT * PT08_S5_O3)         1      6       6    27.224 1.86e-07 ***
I(PT08_S2_NMHC * NO2_GT)        1      6       6    27.195 1.89e-07 ***
I(PT08_S2_NMHC * PT08_S4_NO2)   1     19      19    91.769  < 2e-16 ***
I(NOx_GT * PT08_S3_NOx)         1     71      71   347.975  < 2e-16 ***
I(NOx_GT * NO2_GT)              1     25      25   122.693  < 2e-16 ***
I(NOx_GT * PT08_S4_NO2)         1     31      31   151.572  < 2e-16 ***
I(NOx_GT * PT08_S5_O3)          1      5       5    25.079 5.63e-07 ***
I(NOx_GT * RH)                  1      2       2     9.817  0.00174 **
I(PT08_S3_NOx * PT08_S5_O3)     1      1       1     5.932  0.01489 *
I(NO2_GT * PT08_S4_NO2)         1     26      26   129.110  < 2e-16 ***
I(NO2_GT * T)                   1      1       1     4.441  0.03512 *
I(NO2_GT * RH)                  1      6       6    30.459 3.52e-08 ***
I(PT08_S4_NO2 * PT08_S5_O3)     1      8       8    40.168 2.46e-10 ***
Residuals                    7643   1561       0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Bảng ANOVA



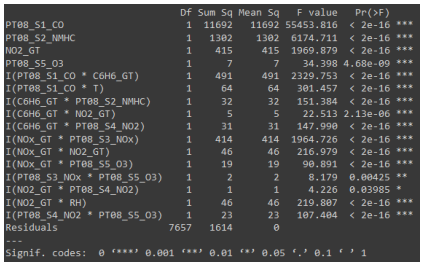
```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -2.578e+00  4.273e-01  -6.034 1.68e-09 ***
PT08_S1_CO                     3.446e-03  7.573e-04   4.550 5.44e-06 ***
C6H6_GT                       -2.913e-02  2.032e-02  -0.993  0.32055
PT08_S2_NMHC                   2.952e-03  7.297e-04   4.045 5.28e-05 ***
NO2_GT                        -1.882e-02  2.551e-03  -7.376 1.80e-13 ***
PT08_S5_O3                     1.142e-03  2.520e-04   4.533 5.91e-06 ***
T                              1.919e-02  6.062e-03   3.166  0.00155 **
RH                             2.703e-03  1.628e-03   1.660  0.09688 .
I(PT08_S1_CO * C6H6_GT)        8.576e-05  3.783e-05   2.267  0.02340 *
I(PT08_S1_CO * PT08_S2_NMHC) -1.179e-06  9.466e-07  -1.245  0.21318
I(PT08_S1_CO * NOx_GT)         1.085e-06  5.916e-07   1.835  0.06660 .
I(PT08_S1_CO * PT08_S3_NOx)    2.131e-08  1.490e-07   0.143  0.88626
I(PT08_S1_CO * NO2_GT)        -1.156e-05  2.117e-06  -5.458 4.98e-08 ***
I(PT08_S1_CO * PT08_S4_NO2)   -2.736e-07  3.133e-07  -0.873  0.38262
I(PT08_S1_CO * T)             -2.518e-05  5.847e-06  -4.307 1.68e-05 ***
I(C6H6_GT * PT08_S2_NMHC)     -1.428e-04  1.607e-05  -8.889  < 2e-16 ***
I(C6H6_GT * NO2_GT)           -9.271e-04  1.270e-04  -7.299 3.18e-13 ***
I(C6H6_GT * PT08_S4_NO2)       1.004e-04  1.371e-05   7.326 2.61e-13 ***
I(C6H6_GT * PT08_S5_O3)        8.802e-05  1.430e-05   6.155 7.91e-10 ***
I(PT08_S2_NMHC * NO2_GT)       2.136e-05  3.752e-06   5.694 1.28e-08 ***
I(PT08_S2_NMHC * PT08_S4_NO2) -9.772e-07  4.687e-07  -2.085  0.03711 *
I(NOx_GT * PT08_S3_NOx)        1.112e-06  2.158e-07   5.151 2.65e-07 ***
I(NOx_GT * NO2_GT)             1.442e-05  8.409e-07  17.146  < 2e-16 ***
I(NOx_GT * PT08_S4_NO2)        1.025e-06  2.586e-07   3.964 7.43e-05 ***
I(NOx_GT * PT08_S5_O3)        -2.399e-06  3.045e-07  -7.878 3.78e-15 ***
I(NOx_GT * RH)                -8.766e-06  3.124e-06  -2.806  0.00503 **
I(PT08_S3_NOx * PT08_S5_O3)   -1.351e-07  1.740e-07  -0.776  0.43763
I(NO2_GT * PT08_S4_NO2)        1.725e-05  1.628e-06  10.595  < 2e-16 ***
I(NO2_GT * T)                 -1.748e-04  4.046e-05  -4.320 1.58e-05 ***
I(NO2_GT * RH)                -7.359e-05  1.537e-05  -4.788 1.72e-06 ***
I(PT08_S4_NO2 * PT08_S5_O3)   -1.268e-06  2.001e-07  -6.338 2.46e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.452 on 7643 degrees of freedom
Multiple R-squared:  0.9037,    Adjusted R-squared:  0.9033
F-statistic:  2390 on 30 and 7643 DF,  p-value: < 2.2e-16
```

(b) Bảng Fitting Linear Models.

Hình 20: Bảng Fitting Linear Models và ANOVA các thuộc tính mô hình 2.

Thực hiện lặp đi lặp lại quá trình trên nhiều lần, ta thu được kết quả cuối cùng được thể hiện ở Hình 21. Kết thúc quá trình phân tích ANOVA.



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
PT08_S1_CO	1	11692	11692	55453.816	< 2e-16 ***
PT08_S2_NMHC	1	1302	1302	6174.711	< 2e-16 ***
NO2_GT	1	415	415	1969.879	< 2e-16 ***
PT08_S5_O3	1	7	7	34.398	4.68e-09 ***
I(PT08_S1_CO * C6H6_GT)	1	491	491	2329.753	< 2e-16 ***
I(PT08_S1_CO * T)	1	64	64	301.457	< 2e-16 ***
I(C6H6_GT * PT08_S2_NMHC)	1	32	32	151.384	< 2e-16 ***
I(C6H6_GT * NO2_GT)	1	5	5	22.513	2.13e-06 ***
I(C6H6_GT * PT08_S4_NO2)	1	31	31	147.990	< 2e-16 ***
I(NOx_GT * PT08_S3_NOx)	1	414	414	1964.726	< 2e-16 ***
I(NOx_GT * NO2_GT)	1	46	46	216.979	< 2e-16 ***
I(NOx_GT * PT08_S5_O3)	1	19	19	90.891	< 2e-16 ***
I(PT08_S3_NOx * PT08_S5_O3)	1	2	2	8.179	0.00425 **
I(NO2_GT * PT08_S4_NO2)	1	1	1	4.226	0.03985 *
I(NO2_GT * RH)	1	46	46	219.807	< 2e-16 ***
I(PT08_S4_NO2 * PT08_S5_O3)	1	23	23	107.404	< 2e-16 ***
Residuals	7657	1614	0		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(a) Bảng ANOVA

```
Call:
lm(formula = CO_GT ~ PT08_S1_CO + PT08_S2_NMHC + NO2_GT + PT08_S5_O3 +
    I(PT08_S1_CO * C6H6_GT) + I(PT08_S1_CO * T) + I(C6H6_GT *
    PT08_S2_NMHC) + I(C6H6_GT * NO2_GT) + I(C6H6_GT * PT08_S4_NO2) +
    I(NOx_GT * PT08_S3_NOx) + I(NOx_GT * NO2_GT) + I(NOx_GT *
    PT08_S5_O3) + I(PT08_S3_NOx * PT08_S5_O3) + I(NO2_GT * PT08_S4_NO2) +
    I(NO2_GT * RH) + I(PT08_S4_NO2 * PT08_S5_O3), data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.9322 -0.1860  0.0164  0.1927  3.3733

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -8.659e-01  9.337e-02  -9.274  < 2e-16 ***
PT08_S1_CO                   1.086e-03  1.200e-04   9.052  < 2e-16 ***
PT08_S2_NMHC                 1.422e-03  9.773e-05  14.548  < 2e-16 ***
NO2_GT                      -1.044e-02  7.501e-04 -13.915  < 2e-16 ***
PT08_S5_O3                   1.219e-03  1.316e-04   9.258  < 2e-16 ***
I(PT08_S1_CO * C6H6_GT)      6.646e-05  7.413e-06   8.966  < 2e-16 ***
I(PT08_S1_CO * T)           -2.660e-05  1.321e-06 -20.141  < 2e-16 ***
I(C6H6_GT * PT08_S2_NMHC)   -6.240e-05  7.555e-06  -8.258  < 2e-16 ***
I(C6H6_GT * NO2_GT)         -2.301e-04  4.218e-05  -5.454 5.07e-08 ***
I(C6H6_GT * PT08_S4_NO2)     5.755e-05  5.374e-06  10.709  < 2e-16 ***
I(NOx_GT * PT08_S3_NOx)      2.138e-06  1.274e-07  16.790  < 2e-16 ***
I(NOx_GT * NO2_GT)           1.305e-05  6.845e-07  19.060  < 2e-16 ***
I(NOx_GT * PT08_S5_O3)      -8.011e-07  8.359e-08  -9.583  < 2e-16 ***
I(PT08_S3_NOx * PT08_S5_O3) -4.460e-07  5.941e-08  -7.507 6.75e-14 ***
I(NO2_GT * PT08_S4_NO2)      9.659e-06  7.074e-07  13.654  < 2e-16 ***
I(NO2_GT * RH)              -7.139e-05  4.708e-06 -15.163  < 2e-16 ***
I(PT08_S4_NO2 * PT08_S5_O3) -8.390e-07  8.095e-08 -10.364  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4592 on 7657 degrees of freedom
Multiple R-squared:  0.9004,    Adjusted R-squared:  0.9002
F-statistic:  4325 on 16 and 7657 DF,  p-value: < 2.2e-16
```

(b) Bảng Fitting Linear Models.

Hình 21: Bảng Fitting Linear Models và ANOVA các thuộc tính mô hình 2.

Mô hình hồi quy có dạng:

$$
\begin{aligned}
CO\_GT = &-8.659e^{-01} + 1.086e^{-03} \times PT08\_S1\_CO \\
&+ 1.422e^{-03} \times PT08\_S2\_NMHC - 1.044e^{-02} \times NO2\_GT \\
&+ 1.219e^{-03} \times PT08\_S5\_O3 \\
&+ 6.646e^{-05} \times I(PT08\_S1\_CO * C6H6\_GT) \\
&- 2.660e^{-05} \times I(PT08\_S1\_CO * T) \\
&- 6.240e^{-05} \times I(C6H6\_GT * PT08\_S2\_NMHC) \\
&- 2.301e^{-04} \times I(C6H6\_GT * NO2\_GT) \\
&+ 5.755e^{-05} \times I(C6H6\_GT * PT08\_S4\_NO2) \\
&+ 2.138e^{-06} \times I(NOx\_GT * PT08\_S3\_NOx) \\
&+ 1.305e^{-05} \times I(NOx\_GT * NO2\_GT) \\
&- 8.011e^{-07} \times I(NOx\_GT * PT08\_S5\_O3) \\
&- 4.460e^{-07} \times I(PT08\_S3\_NOx * PT08\_S5\_O3) \\
&+ 9.659e^{-06} \times I(NO2\_GT * PT08\_S4\_NO2) \\
&- 7.139e^{-05} \times I(NO2\_GT * RH) \\
&- 8.390e^{-07} \times I(PT08\_S4\_NO2 * PT08\_S5\_O3)
\end{aligned}
$$

5.2 Bộ dữ liệu Air Quality–MEAN.

ANOVA đơn thuộc tính Thực hiện phân tích ANOVA với tất cả các thuộc tính của bộ dữ liệu.

```
               Df Sum Sq Mean Sq  F value   Pr(>F)
PT08_S1_CO      1   9695    9695 27528.69  < 2e-16 ***
C6H6_GT         1   1228    1228  3487.31  < 2e-16 ***
PT08_S2_NMHC    1     35      35   100.46  < 2e-16 ***
NOx_GT          1   1579    1579  4483.09  < 2e-16 ***
PT08_S3_NOx     1     81      81   229.74  < 2e-16 ***
NO2_GT          1     38      38   109.31  < 2e-16 ***
PT08_S4_NO2     1    167     167   475.39  < 2e-16 ***
PT08_S5_O3      1     18      18    51.42 8.00e-13 ***
T               1     17      17    47.04 7.41e-12 ***
RH              1     48      48   136.65  < 2e-16 ***
AH              1      7       7    20.05 7.62e-06 ***
Residuals    9345   3291       0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Bảng ANOVA

```
Call:
lm(formula = CO_GT ~ PT08_S1_CO + C6H6_GT + PT08_S2_NMHC + NOx_GT +
    PT08_S3_NOx + NO2_GT + PT08_S4_NO2 + PT08_S5_O3 + T + RH +
    AH, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0688 -0.2509 -0.0163  0.2585  3.3539

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.667e-01  1.586e-01  -5.466 4.73e-08 ***
PT08_S1_CO    9.248e-04  8.209e-05  11.266  < 2e-16 ***
C6H6_GT       7.125e-02  5.376e-03  13.253  < 2e-16 ***
PT08_S2_NMHC -1.142e-03  1.874e-04  -6.095 1.14e-09 ***
NOx_GT        3.257e-03  6.641e-05  49.038  < 2e-16 ***
PT08_S3_NOx   1.624e-04  5.747e-05   2.827  0.00471 **
NO2_GT        3.007e-03  2.691e-04  11.177  < 2e-16 ***
PT08_S4_NO2   1.511e-03  6.473e-05  23.346  < 2e-16 ***
PT08_S5_O3   -3.533e-04  4.580e-05  -7.713 1.35e-14 ***
T            -1.439e-02  2.696e-03  -5.338 9.62e-08 ***
RH           -5.439e-03  1.032e-03  -5.269 1.40e-07 ***
AH           -2.256e-01  5.039e-02  -4.478 7.62e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5934 on 9345 degrees of freedom
Multiple R-squared:  0.7969,    Adjusted R-squared:  0.7967
F-statistic:  3334 on 11 and 9345 DF,  p-value: < 2.2e-16
```

(b) Bảng Fitting Linear Models.

Hình 22: Bảng Fitting Linear Models và ANOVA các thuộc tính mô hình 1.

Ở lần đầu tiên thực hiện phân tích ANOVA, không có thuộc tính nào có $p-value > 0.05$ (Hình 22a). Do đó, ta sẽ thực hiện kiểm tra mức độ ý nghĩa của các hệ số theo bảng Fitting Linear Models (Hình 22b). Kết quả thu được là vẫn không có thuộc tính nào có $p-value > 0.05$. Vậy, quá trình phân tích ANOVA không bỏ được thuộc tính nào, vì thế mô hình thu được là mô hình trên bộ dữ liệu gốc.

Mô hình hồi quy có dạng:

$$\begin{aligned} CO\_GT = &-8.667e^{-01} + 9.248e^{-04} \times PT08\_S1\_CO \\ &+ 7.125e^{-02} \times C6H6\_GT - 1.142e^{-03} PT08\_S2\_NMHC \\ &+ 3.257e^{-03} \times NOx\_GT + 1.624e^{-04} \times PT08\_S3\_NOx \\ &+ 3.007e^{-03} \times NO2\_GT + 1.511e^{-03} \times PT08\_S4\_NO2 \\ &- 3.533e^{-04} \times PT08\_S5\_O3 - 1.439e^{-02} \times T \\ &- 5.439e^{-03} \times RH - 2.256e^{-01} \times AH \end{aligned}$$

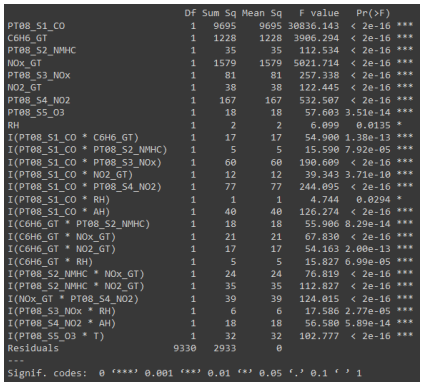
ANOVA tương tác 2 thuộc tính Thực hiện phân tích ANOVA với tất cả các thuộc tính và kết hợp với các tương tác 2 thuộc tính của tất cả các thuộc tính trong bộ dữ liệu. Kết quả thực hiện ANOVA lần đầu tiên được thể hiện ở Hình 23.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
PT08_S1_CO	1	9695	9695	31846.132	< 2e-16	***
C6H6_GT	1	1228	1228	4034.238	< 2e-16	***
PT08_S2_NMHC	1	35	35	116.220	< 2e-16	***
NOx_GT	1	1579	1579	5186.193	< 2e-16	***
PT08_S3_NOx	1	81	81	265.767	< 2e-16	***
NO2_GT	1	38	38	126.456	< 2e-16	***
PT08_S4_NO2	1	167	167	549.949	< 2e-16	***
PT08_S5_O3	1	18	18	59.490	1.36e-14	***
T	1	17	17	54.415	1.76e-13	***
RH	1	48	48	158.085	< 2e-16	***
AH	1	7	7	23.197	1.49e-06	***
I(PT08_S1_CO * C6H6_GT)	1	13	13	44.236	3.08e-11	***
I(PT08_S1_CO * PT08_S2_NMHC)	1	6	6	20.676	5.51e-06	***
I(PT08_S1_CO * NOx_GT)	1	0	0	0.098	0.753710	
I(PT08_S1_CO * PT08_S3_NOx)	1	40	40	130.241	< 2e-16	***
I(PT08_S1_CO * NO2_GT)	1	16	16	51.586	7.38e-13	***
I(PT08_S1_CO * PT08_S4_NO2)	1	58	58	191.522	< 2e-16	***
I(PT08_S1_CO * PT08_S5_O3)	1	0	0	0.099	0.752875	
I(PT08_S1_CO * T)	1	1	1	2.582	0.108092	
I(PT08_S1_CO * RH)	1	17	17	56.718	5.50e-14	***
I(PT08_S1_CO * AH)	1	3	3	8.312	0.003948	**
I(C6H6_GT * PT08_S2_NMHC)	1	16	16	52.897	3.80e-13	***
I(C6H6_GT * NOx_GT)	1	41	41	133.936	< 2e-16	***
I(C6H6_GT * PT08_S3_NOx)	1	2	2	5.882	0.015313	*
I(C6H6_GT * NO2_GT)	1	5	5	17.579	2.78e-05	***
I(C6H6_GT * PT08_S4_NO2)	1	2	2	6.052	0.013908	*
I(C6H6_GT * PT08_S5_O3)	1	43	43	140.136	< 2e-16	***
I(C6H6_GT * T)	1	3	3	8.580	0.003407	**
I(C6H6_GT * RH)	1	2	2	8.144	0.004330	**
I(C6H6_GT * AH)	1	3	3	10.365	0.001289	**
I(PT08_S2_NMHC * NOx_GT)	1	13	13	41.332	1.35e-10	***
I(PT08_S2_NMHC * PT08_S3_NOx)	1	5	5	16.428	5.10e-05	***
I(PT08_S2_NMHC * NO2_GT)	1	20	20	65.656	6.05e-16	***
I(PT08_S2_NMHC * PT08_S4_NO2)	1	0	0	0.223	0.636990	
I(PT08_S2_NMHC * PT08_S5_O3)	1	1	1	3.073	0.079635	.
I(PT08_S2_NMHC * T)	1	0	0	0.223	0.636673	
I(PT08_S2_NMHC * RH)	1	0	0	0.063	0.801244	
I(PT08_S2_NMHC * AH)	1	2	2	5.741	0.016594	*
I(NOx_GT * PT08_S3_NOx)	1	1	1	2.603	0.106663	
I(NOx_GT * NO2_GT)	1	7	7	22.588	2.04e-06	***
I(NOx_GT * PT08_S4_NO2)	1	64	64	209.802	< 2e-16	***
I(NOx_GT * PT08_S5_O3)	1	1	1	3.439	0.063724	.
I(NOx_GT * T)	1	0	0	0.047	0.828803	
I(NOx_GT * RH)	1	0	0	0.283	0.594534	
I(NOx_GT * AH)	1	3	3	10.742	0.001051	**
I(PT08_S3_NOx * NO2_GT)	1	4	4	14.763	0.000123	***
I(PT08_S3_NOx * PT08_S4_NO2)	1	2	2	7.374	0.006630	**
I(PT08_S3_NOx * PT08_S5_O3)	1	1	1	2.055	0.151712	
I(PT08_S3_NOx * T)	1	8	8	26.202	3.14e-07	***
I(PT08_S3_NOx * RH)	1	3	3	9.640	0.001910	**
I(PT08_S3_NOx * AH)	1	2	2	6.764	0.009317	**
I(NO2_GT * PT08_S4_NO2)	1	1	1	3.426	0.064199	.
I(NO2_GT * PT08_S5_O3)	1	1	1	4.733	0.029619	*
I(NO2_GT * T)	1	0	0	0.114	0.735802	
I(NO2_GT * RH)	1	2	2	5.308	0.021244	*
I(NO2_GT * AH)	1	8	8	27.722	1.43e-07	***
I(PT08_S4_NO2 * PT08_S5_O3)	1	23	23	75.106	< 2e-16	***
I(PT08_S4_NO2 * T)	1	4	4	13.767	0.000208	***
I(PT08_S4_NO2 * RH)	1	6	6	20.910	4.88e-06	***
I(PT08_S4_NO2 * AH)	1	2	2	6.450	0.011109	*
I(PT08_S5_O3 * T)	1	0	0	1.007	0.315656	
I(PT08_S5_O3 * RH)	1	0	0	0.089	0.764871	
I(PT08_S5_O3 * AH)	1	0	0	0.787	0.375103	
I(T * RH)	1	1	1	1.975	0.159950	
I(T * AH)	1	6	6	20.325	6.61e-06	***
I(RH * AH)	1	0	0	0.215	0.642969	
Residuals	9290	2828	0			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Hình 23: Bảng ANOVA các thuộc tính mô hình 2.

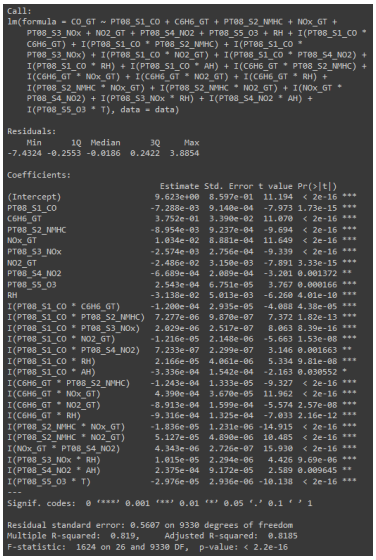
Ta có thể thấy rằng có rất nhiều thuộc tính có $p - value > 0.05$, vì vậy ta cần thực hiện ANOVA nhiều lần để tìm ra các thuộc tính có ý nghĩa. Thực hiện các bước tương tư như bộ dữ liệu ở Phần 5.1 ta thu được kết quả cuối cùng được thể hiện ở Hình 24.



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
PT08_S1_CO	1	9695	9695	30836.143	< 2e-16 ***
C6H6_GT	1	1228	1228	3906.294	< 2e-16 ***
PT08_S2_NMHC	1	35	35	112.534	< 2e-16 ***
NOx_GT	1	1579	1579	5021.714	< 2e-16 ***
PT08_S3_NOx	1	81	81	257.338	< 2e-16 ***
NO2_GT	1	38	38	122.445	< 2e-16 ***
PT08_S4_NO2	1	167	167	532.507	< 2e-16 ***
PT08_S5_O3	1	18	18	57.603	3.51e-14 ***
RH	1	2	2	6.099	0.0135 *
I(PT08_S1_CO * C6H6_GT)	1	17	17	54.900	1.38e-13 ***
I(PT08_S1_CO * PT08_S2_NMHC)	1	5	5	15.590	7.92e-05 ***
I(PT08_S1_CO * PT08_S3_NOx)	1	60	60	190.609	< 2e-16 ***
I(PT08_S1_CO * NO2_GT)	1	12	12	39.343	3.71e-10 ***
I(PT08_S1_CO * PT08_S4_NO2)	1	77	77	244.095	< 2e-16 ***
I(PT08_S1_CO * RH)	1	1	1	4.744	0.0294 *
I(PT08_S1_CO * AH)	1	40	40	126.274	< 2e-16 ***
I(C6H6_GT * PT08_S2_NMHC)	1	18	18	55.906	8.29e-14 ***
I(C6H6_GT * NOx_GT)	1	21	21	67.830	< 2e-16 ***
I(C6H6_GT * NO2_GT)	1	17	17	54.163	2.00e-13 ***
I(C6H6_GT * RH)	1	5	5	15.827	6.99e-05 ***
I(PT08_S2_NMHC * NOx_GT)	1	24	24	76.819	< 2e-16 ***
I(PT08_S2_NMHC * NO2_GT)	1	35	35	112.827	< 2e-16 ***
I(NOx_GT * PT08_S4_NO2)	1	39	39	124.015	< 2e-16 ***
I(PT08_S3_NOx * RH)	1	6	6	17.586	2.77e-05 ***
I(PT08_S4_NO2 * AH)	1	18	18	56.580	5.89e-14 ***
I(PT08_S5_O3 * T)	1	32	32	102.777	< 2e-16 ***
Residuals	9330	2933	0		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(a) Bảng ANOVA



```
Call:
lm(formula = CO_GT ~ PT08_S1_CO + C6H6_GT + PT08_S2_NMHC + NOx_GT +
    PT08_S3_NOx + NO2_GT + PT08_S4_NO2 + PT08_S5_O3 + RH + I(PT08_S1_CO *
    C6H6_GT) + I(PT08_S1_CO * PT08_S2_NMHC) + I(PT08_S1_CO *
    PT08_S3_NOx) + I(PT08_S1_CO * NO2_GT) + I(PT08_S1_CO * PT08_S4_NO2) +
    I(PT08_S1_CO * RH) + I(PT08_S1_CO * AH) + I(C6H6_GT * PT08_S2_NMHC) +
    I(C6H6_GT * NOx_GT) + I(C6H6_GT * NO2_GT) + I(C6H6_GT * RH) +
    I(PT08_S2_NMHC * NOx_GT) + I(PT08_S2_NMHC * NO2_GT) + I(NOx_GT *
    PT08_S4_NO2) + I(PT08_S3_NOx * RH) + I(PT08_S4_NO2 * AH) +
    I(PT08_S5_O3 * T), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4324 -0.2553 -0.0186  0.2422  3.8854

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  9.623e+00  8.597e-01  11.194  < 2e-16 ***
PT08_S1_CO                  -7.288e-03  9.140e-04  -7.973 1.73e-15 ***
C6H6_GT                      3.752e-01  3.390e-02  11.070  < 2e-16 ***
PT08_S2_NMHC                -8.954e-03  9.237e-04  -9.694  < 2e-16 ***
NOx_GT                       1.034e-02  8.881e-04  11.649  < 2e-16 ***
PT08_S3_NOx                 -2.574e-03  2.756e-04  -9.339  < 2e-16 ***
NO2_GT                      -2.486e-02  3.150e-03  -7.891 3.33e-15 ***
PT08_S4_NO2                 -6.689e-04  2.089e-04  -3.201 0.001372 **
PT08_S5_O3                   2.543e-04  6.751e-05   3.767 0.000166 ***
RH                          -3.138e-02  5.013e-03  -6.260 4.01e-10 ***
I(PT08_S1_CO * C6H6_GT)     -1.200e-04  2.935e-05  -4.088 4.38e-05 ***
I(PT08_S1_CO * PT08_S2_NMHC) 7.277e-06  9.870e-07   7.372 1.82e-13 ***
I(PT08_S1_CO * PT08_S3_NOx)  2.029e-06  2.517e-07   8.063 8.39e-16 ***
I(PT08_S1_CO * NO2_GT)      -1.216e-05  2.148e-06  -5.663 1.53e-08 ***
I(PT08_S1_CO * PT08_S4_NO2)  7.233e-07  2.299e-07   3.146 0.001663 **
I(PT08_S1_CO * RH)           2.166e-05  4.061e-06   5.334 9.81e-08 ***
I(PT08_S1_CO * AH)          -3.336e-04  1.542e-04  -2.163 0.030552 *
I(C6H6_GT * PT08_S2_NMHC)   -1.243e-04  1.333e-05  -9.327  < 2e-16 ***
I(C6H6_GT * NOx_GT)          4.390e-04  3.670e-05  11.962  < 2e-16 ***
I(C6H6_GT * NO2_GT)         -8.913e-04  1.599e-04  -5.574 2.57e-08 ***
I(C6H6_GT * RH)             -9.316e-04  1.325e-04  -7.033 2.16e-12 ***
I(PT08_S2_NMHC * NOx_GT)    -1.836e-05  1.231e-06 -14.915  < 2e-16 ***
I(PT08_S2_NMHC * NO2_GT)     5.127e-05  4.890e-06  10.485  < 2e-16 ***
I(NOx_GT * PT08_S4_NO2)      4.343e-06  2.726e-07  15.930  < 2e-16 ***
I(PT08_S3_NOx * RH)          1.015e-05  2.294e-06   4.426 9.69e-06 ***
I(PT08_S4_NO2 * AH)          2.375e-04  9.172e-05   2.589 0.009645 **
I(PT08_S5_O3 * T)           -2.976e-05  2.936e-06 -10.138  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5607 on 9330 degrees of freedom
Multiple R-squared:  0.819,     Adjusted R-squared:  0.8185
F-statistic:  1624 on 26 and 9330 DF,  p-value: < 2.2e-16
```

(b) Bảng Fitting Linear Models.

Hình 24: Bảng Fitting Linear Models và ANOVA các thuộc tính mô hình 2.

Mô hình hồi quy có dạng:

$$
\begin{aligned}
CO\_GT = {} & 9.623e^{00} - 7.288e^{-03} \times PT08\_S1\_CO + 3.752e^{-01} \times C6H6\_GT \\
& - 8.954e^{-03} \times PT08\_S2\_NMHC + 1.034e^{-02} \times NOx\_GT \\
& - 2.574e^{-03} \times PT08\_S3\_NOx - 2.486e^{-02} \times NO2\_GT \\
& - 6.689e^{-04} \times PT08\_S4\_NO2 + 2.543e^{-04} \times PT08\_S5\_O3 \\
& - 3.138e^{-02} \times RH - 1.200e^{-04} \times I(PT08\_S1\_CO * C6H6\_GT) \\
& + 7.277e^{-06} \times I(PT08\_S1\_CO * PT08\_S2\_NMHC) \\
& + 2.029e^{-06} \times I(PT08\_S1\_CO * PT08\_S3\_NOx) \\
& - 1.216e^{-05} \times I(PT08\_S1\_CO * NO2\_GT) \\
& + 7.233e^{-07} \times I(PT08\_S1\_CO * PT08\_S4\_NO2) \\
& + 2.166e^{-05} \times I(PT08\_S1\_CO * RH) \\
& - 3.336e^{-04} \times I(PT08\_S1\_CO * AH) \\
& - 1.243e^{-04} \times I(C6H6\_GT * PT08\_S2\_NMHC) \\
& + 4.390e^{-04} \times I(C6H6\_GT * NOx\_GT) \\
& - 8.913e^{-04} \times I(C6H6\_GT * NO2\_GT) \\
& - 9.316e^{-04} \times I(C6H6\_GT * RH) \\
& - 1.836e^{-05} \times I(PT08\_S2\_NMHC * NOx\_GT) \\
& + 5.127e^{-05} \times I(PT08\_S2\_NMHC * NO2\_GT) \\
& + 4.343e^{-06} \times I(NOx\_GT * PT08\_S4\_NO2) \\
& + 1.015e^{-05} \times I(PT08\_S3\_NOx * RH) \\
& + 2.375e^{-04} \times I(PT08\_S4\_NO2 * AH) \\
& - 2.976e^{-05} \times I(PT08\_S5\_O3 * T)
\end{aligned}
$$

6 Thực Nghiệm và Phân Tích Kết Quả

Qua quá trình phân tích ANOVA trên 2 bộ dữ liệu REMOVE và MEAN, với bộ dữ liệu REMOVE ta có 3 loại: REMOVE gốc, REMOVE đơn thuộc tính và REMOVE tương tác 2 thuộc tính; với bộ dữ liệu MEAN ta có 2 loại: MEAN gốc và MEAN tương tác 2 thuộc tính. Ta tiến hành cài đặt các mô hình, thuật toán đã trình bày trong Phần 4 trên các bộ dữ liệu và đánh giá kết quả thông qua các độ đo $R^2, MSE, RMSE, MAE$ để so sánh kết quả đạt được.

Với từng bộ dữ liệu, chúng tôi chia 80% cho tập training và 20% cho tập testing. Kích thước của các tập dữ liệu được trình bày tại Bảng 2:

		Train shape	Test shape
REMOVE	Gốc	(6139, 11)	(1535, 11)
	ANOVA đơn thuộc tính	(6139, 9)	(1535, 9)
	ANOVA tương tác 2 thuộc tính	(6139, 16)	(1535, 16)
MEAN	Gốc	(7485, 11)	(1872, 11)
	ANOVA tương tác 2 thuộc tính	(7485, 26)	(1872, 26)

Bảng 2: Kích thước của các tập dữ liệu.

Để tránh tính ngẫu nhiên, chúng tôi thực hiện 5 lần chạy và ghi lại kết quả, sau đó tính kết quả trung bình của 5 lần chạy. Kết quả trung bình được thể hiện tại Bảng 3.

			Train				Test			
			R^2	MSE	RMSE	MAE	R^2	MSE	RMSE	MAE
REMOVE	Gốc	LR	0.8863	0.2407	0.4907	0.3098	0.8880	0.2333	0.4830	0.3096
		DT	0.9355	0.1365	0.3695	0.2502	0.8956	0.2176	0.4664	0.3070
		RF	**0.9875**	**0.0265**	**0.1629**	**0.0999**	0.9243	0.1577	0.3971	0.2490
		SVR	0.9451	0.1162	0.3409	0.1959	**0.9310**	**0.1438**	**0.3792**	**0.2285**
		NN	0.9421	0.1226	0.3500	0.2304	0.9255	0.1552	0.3940	0.2571
	ANOVA đơn thuộc tính	LR	0.8863	0.2409	0.4908	0.3095	0.8880	0.2333	0.4830	0.3091
		DT	0.9354	0.1368	0.3698	0.2497	0.8977	0.2131	0.4616	0.3020
		RF	**0.9873**	**0.0269**	**0.1640**	**0.1004**	0.9241	0.1580	0.3975	0.2497
		SVR	0.9447	0.1171	0.3422	0.1968	**0.9311**	**0.1435**	**0.3789**	**0.2275**
		NN	0.9398	0.1274	0.3569	0.2363	0.9233	0.1598	0.3997	0.2625
	ANOVA tương tác 2 thuộc tính	LR	0.8994	0.2131	0.4616	0.2972	0.9036	0.2009	0.4482	0.2929
		DT	0.9368	0.1338	0.3658	0.2484	0.8945	0.2198	0.4688	0.3046
		RF	**0.9869**	**0.0278**	**0.1668**	**0.1025**	0.9190	0.1687	0.4108	0.2578
		SVR	0.9453	0.1158	0.3403	0.2003	**0.9271**	**0.1519**	**0.3897**	**0.2379**
		NN	0.9462	0.1141	0.3376	0.2260	0.9241	0.1581	0.3976	0.2655
MEAN	Gốc	LR	0.7921	0.3535	0.5946	0.4025	0.8136	0.3461	0.5883	0.3964
		DT	0.8851	0.1955	0.4421	0.3093	0.8174	0.3390	0.5823	0.3828
		RF	**0.9776**	**0.0381**	**0.1952**	**0.1261**	**0.8738**	**0.2343**	**0.4841**	0.3144
		SVR	0.8820	0.2007	0.4480	0.2624	0.8641	0.2523	0.5023	**0.2997**
		NN	0.8871	0.1920	0.4381	0.2999	0.8706	0.2402	0.4901	0.3305
	ANOVA tương tác 2 thuộc tính	LR	0.8139	0.3165	0.5626	0.3860	0.8364	0.3037	0.5511	0.3775
		DT	0.8844	0.1966	0.4434	0.3041	0.8054	0.3612	0.6010	0.3800
		RF	**0.9768**	**0.0394**	**0.1986**	**0.1297**	**0.8688**	**0.2437**	**0.4936**	0.3197
		SVR	0.8810	0.2024	0.4498	0.2661	0.8604	0.2593	0.5092	**0.3060**
		NN	0.8943	0.1798	0.4239	0.2899	0.8658	0.2491	0.4990	0.3372

Bảng 3: Kết quả trung bình trên 5 lần chạy.

Từ kết quả tại Bảng 3 ta có thể thấy rằng:

- **Về độ đo:** Để dễ so sánh hiệu suất của các mô hình với các bộ dữ liệu khác nhau, chúng tôi thống nhất chọn độ đo RMSE để so sánh hiệu suất giữa các mô hình và bộ dữ liệu khác nhau. Kết quả được đánh giá trên tập test.
- **Về thuật toán:** Linear Regression (LR) cho kết quả tệ nhất trong 5 mô hình; Decision Tree Regression (DT) cho kết quả tốt hơn mô hình LR nhưng lại bị vấn đề overfitting[8] nhẹ (chênh lệch độ đo RMSE giữa train và test trên các bộ dữ liệu khác nhau là ~ 0.1); Random Forest Regression (RF) mô hình này rất tệ vì bị hiện tượng overfitting khá nghiêm trọng (chênh lệch độ đo RMSE giữa train và test trên các bộ dữ liệu khác nhau là ~ 0.25). Support Vector Regression (SVR) cho kết quả tốt và ổn định nhất. Neural Network (NN) cho kết quả khá tốt chỉ xếp sau mô hình SVM, đặt biệt ở bộ dữ liệu MEAN kết quả của NN còn tốt hơn SVR.
- **Về bộ dữ liệu:** Bộ dữ liệu REMOVE cho kết quả trung bình cao hơn bộ dữ liệu MEAN (chênh lệch RMSE train ~ 0.03 và RMSE test ~ 0.1).
- **Bộ dữ liệu và thuật toán:**
  - REMOVE: LR cho kết quả tốt hơn khi thực hiện ANOVA tương tác 2 thuộc tính (RMSE giảm từ 0.4830 còn 0.4482). DT cho kết quả tốt hơn khi thực hiện ANOVA đơn thuộc tính nhưng kết quả khi ANOVA tương tác 2 thuộc tính lại không tốt bằng dữ liệu gốc (RMSE lần lượt là 0.4664, 0.4616 và 0.4688). RF đạt kết quả tốt nhất trên bộ dữ liệu gốc và hiệu suất giảm dần khi thực hiện ANOVA đơn thuộc tính và ANOVA tương tác 2 thuộc tính (RMSE lần lượt là 0.3971, 0.3975 và 0.4108). SVR tương tự như DT, cho kết quả tốt hơn khi thực hiện ANOVA đơn thuộc tính nhưng kết quả khi ANOVA tương tác 2 thuộc tính lại không tốt bằng dữ liệu gốc (RMSE lần lượt là 0.3792, 0.3789 và 0.3897). NN cho kết quả trên giữ liệu gốc là tốt nhất và hiệu suất giảm dần trên dữ liệu ANOVA tương tác 2 thuộc tính và ANOVA đơn thuộc tính (RMSE lần lượt là 0.3940, 0.3976 và 0.3997). Nhìn chung SVR là mô hình cho kết quả tốt, ổn định nhất trên tất cả các bộ dữ liệu và mô hình đạt kết quả tốt nhất trên bộ dữ liệu ANOVA đơn thuộc tính, nên mô hình được chọn là mô hình sử dụng thuật toán SVR trên bộ dữ liệu REMOVE ANOVA đơn thuộc tính.
  - MEAN: LR cho kết quả tốt hơn khi thực hiện ANOVA tương tác 2 thuộc tính (RMSE giảm từ 0.5883 còn 0.5511). DT cho kết quả trên bộ dữ liệu gốc tốt hơn khi thực hiện ANOVA tương tác 2 thuộc tính (RMSE lần lượt là 0.5823 và 0.6010). RF cho kết quả trên bộ dữ liệu gốc tốt hơn khi thực hiện ANOVA tương tác 2 thuộc tính (RMSE lần lượt là 0.4841 và 0.4936). SVR cho kết quả trên bộ dữ liệu gốc tốt hơn khi thực hiện ANOVA tương tác 2 thuộc tính (RMSE lần lượt là 0.5023 và 0.5092). NN cho kết quả trên bộ dữ liệu gốc tốt hơn khi thực hiện ANOVA tương tác 2 thuộc tính (RMSE lần lượt là 0.4901 và 0.4990). Nhìn chung mô hình sử dụng thuật toán RF và bộ dữ liệu gốc là cho kết quả tốt nhất,

[8] Overfitting là hiện tượng mô hình quá khớp với dữ liệu training, nhưng kết quả trên tập testing lại thấp.

nhưng vì mô hình bị hiện tượng overfitting khá nghiệm trong nên ta sẽ không chọn mô hình này, ta chọn mô hình ổn định hơn là NN trên bộ dữ liệu MEAN gốc.

Qua các kết quả phân tích ở trên, ta thấy:

- Dữ liệu xử lý missing values theo chiến lược REMOVE cho kết quả cao hơn dữ liệu xử lý missing values bằng chiến lược MEAN ở tất cả các mô hình. Điều này chứng tỏ việc xử lý dữ liệu missing values theo chiến lược REMOVE là tốt hơn chiến lược MEAN. Thật vậy, trong thực tế việc xử lý missing values theo chiến lược MEAN là không tốt, dễ tạo ra các nhiễu loạn làm ảnh hưởng lớn đến hiệu suất dự đoán của mô hình. Ngoài cách xử lý dữ liệu missing values theo chiến lược REMOVE thì chúng tôi cũng đề xuất một cách xử lý dữ liệu missing values khác là lấy mean theo ngày/giờ của từng thuộc tính.
- Bộ dữ liệu ban đầu đã có được hiệu suất rất tốt, nên quá trình thực hiện phân tích ANOVA để loại bỏ các thuộc tính ít ảnh hưởng đến đầu ra hoặc xem xét các tương tác của các thuộc tính nhằm tạo ra các bộ dữ liệu mới không thực sự quá hiệu quả để cải thiện hiệu suất dự đoán của các mô hình. Tuy nhiên, chúng ta có thể dùng ANOVA đơn thuộc tính để loại bỏ bớt các thuộc tính ít ảnh hưởng đến mô hình nhưng vẫn đảm bảo mô hình giữ nguyên hiệu suất dự đoán hoặc cao hơn đôi chút, từ đó có thể giảm kích thước của bộ dữ liệu, giảm được chi phí tài nguyên và tính toán.
- Mô hình cuối cùng tốt nhất mà chúng tôi đạt được là mô hình sử dụng thuật toán Support Vector Regression kết hợp với bộ dữ liệu REMOVE ANOVA đơn thuộc tính (RMSE = 0.3789).

7 Kết Luận

Trong báo cáo này, chúng tôi đã thực hiện tìm hiểu, phân tích và xây dựng mô hình dự đoán nồng độ CO trong không khí dựa trên bộ dữ liệu Air Quality. Với bộ dữ liệu Air Quality ban đầu có nhiều missing values, nên chúng tôi đã tiến hành các phương pháp xử lý missing values và cho ra 2 bộ dữ liệu mới là: Air Quality-REMOVE và Air Quality-MEAN. Thực hiện quá trình phân tích ANOVA trên 2 bộ dữ liệu REMOVE và MEAN, chúng tôi thu được các bộ dữ liệu sau: REMOVE gốc, REMOVE ANOVA đơn thuộc tính, REMOVE ANOVA tương tác 2 thuộc tính, MEAN gốc và MEAN ANOVA tương tác 2 thuộc tính. Tiếp theo, chúng tôi tiến hành thực nghiệm các bộ dữ liệu với các mô hình sử dụng các thuật toán: Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression và Neural Network. Kết quả tốt nhất mà chúng tối đạt được là mô hình Support Vector Regression được huấn luyện trên bộ dữ liệu REMOVE ANOVA đơn thuộc tính, với độ đo RMSE = 0.3789. Bộ dữ liệu gốc đã có được hiệu suất tốt nên quá trình phân tích ANOVA không thực sự giúp cải thiện đáng kể kết quả dự đoán của mô hình. Tuy nhiên, chúng ta có thể dùng ANOVA đơn thuộc tính để loại bỏ bớt các thuộc tính ít ảnh hưởng đến mô hình nhưng vẫn đảm bảo mô hình giữ nguyên hiệu suất dự

đoán hoặc cao hơn đôi chút, từ đó có thể giảm kích thước của bộ dữ liệu, giảm được chi phí tài nguyên và tính toán.

Hướng phát triển trong tương lai:

- **Bộ dữ liệu:** Xử lý các missing values tốt hơn nữa, như đề suất ở phần trước là thử xử lý các missing values bằng cách điền bằng các giá trị mean của từng thuộc tính theo ngày/giờ. Ngoài ra, chúng ta có thể tiến hành thu thập thêm dữ liệu từ thực tế thông qua các cảm biến (sensor).
- **Mô hình:** Áp dụng các kỹ thuật, mô hình Deep Learning như: RNN, LSTM, ... và các mô hình Time Series như: ARIMA, ... để cải thiện kết quả dự đoán tốt hơn nữa.

Tài liệu

1. Saverio, D. V., Massera, E., Piga, M., Martinotto, L., Francia, G. D.: On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators B: Chemical, Vol. 129(2), 750–757.
2. Saverio, D. V.: ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development. Accessed: 6 July 2022 [Online]. Available https://archive.ics.uci.edu/ml/datasets/Air+Quality, Air Quality Data Set.
3. Tran Quang Quy: Department of Computer Sciences & Technology. Accessed: 6 July 2022 [Online]. Available https://rpubs.com/tranquangquy_ictu/769561, Correlation Coefficient.
4. Toploigiai. Accessed: 6 July 2022 [Online]. Available https://toploigiai.vn/cach-nhan-xet-bieu-do-hop, Box Plot.
5. Wikipedia. Accessed: 6 July 2022 [Online]. Available https://vi.wikipedia.org/wiki/C%C3%A2y_quy%E1%BA%BFt_%C4%91%E1%BB%8Bnh, Cây quyết định.
6. Wikipedia. Accessed: 6 July 2022 [Online]. Available https://vi.wikipedia.org/wiki/M%C3%A1y_vect%C6%A1_h%E1%BB%97_tr%E1%BB%A3, Máy vectơ hỗ trợ.
7. Wikipedia. Accessed: 6 July 2022 [Online]. Available https://vi.wikipedia.org/wiki/M%E1%BA%A1ng_th%E1%BA%A7n_kinh_nh%C3%A2n_t%E1%BA%A1o, Mạng thần kinh nhân tạo.
8. Wikipedia. Accessed: 6 July 2022 [Online]. Available https://en.wikipedia.org/wiki/Coefficient_of_determination, Coefficient of determination.
9. Wikipedia. Accessed: 6 July 2022 [Online]. Available https://vi.wikipedia.org/wiki/Sai_s%E1%BB%91_to%C3%A0n_ph%C6%B0%C6%A1ng_trung_b%C3%ACnh, Sai số toàn phương trung bình.
10. Wikipedia. Accessed: 6 July 2022 [Online]. Available https://en.wikipedia.org/wiki/Root-mean-square_deviation, Root-mean-square deviation.
11. Wikipedia. Accessed: 6 July 2022 [Online]. Available https://en.wikipedia.org/wiki/Mean_absolute_error, Mean absolute error.