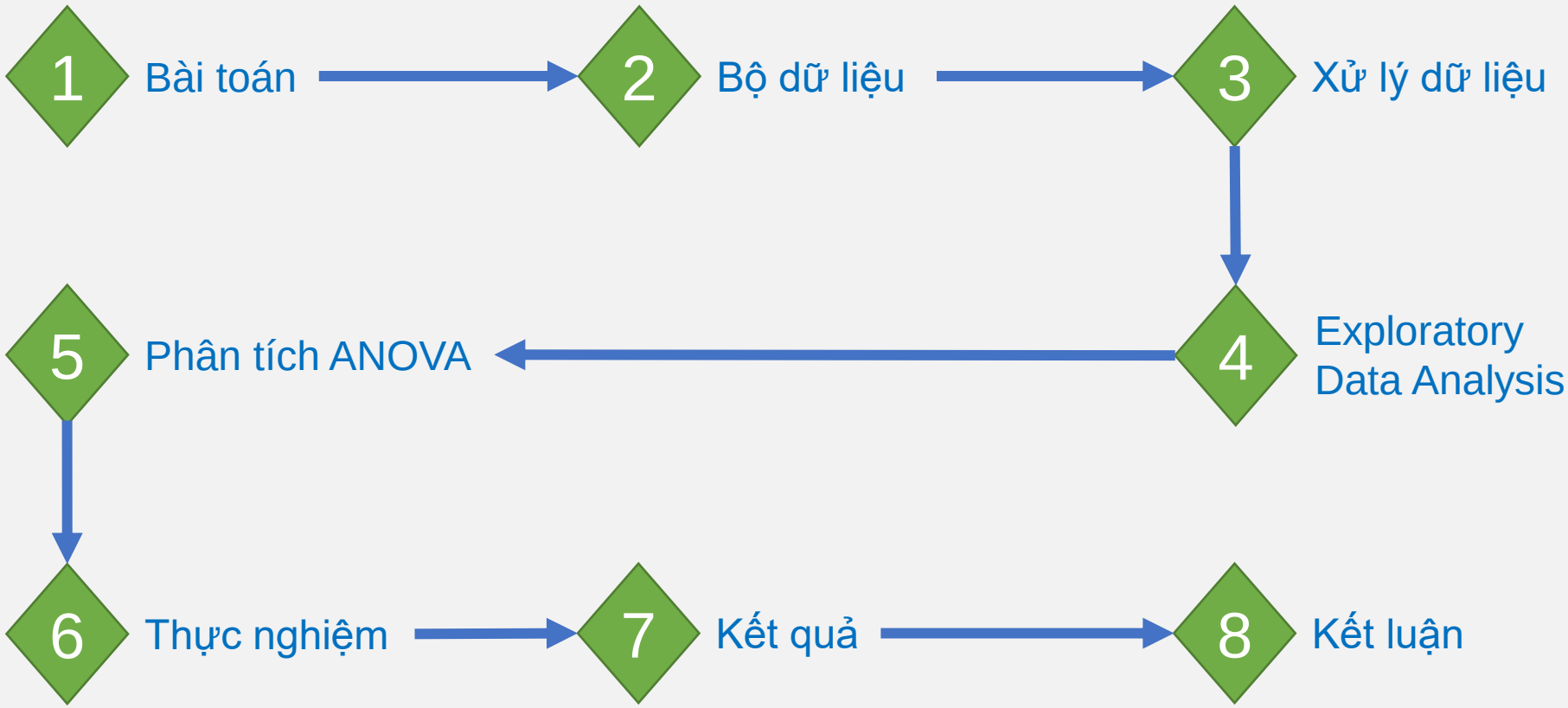
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC & KỸ THUẬT THÔNG TIN

PHÂN TÍCH & XÂY DỰNG MÔ HÌNH DỰ ĐOÁN NỒNG ĐỘ CO TRONG KHÔNG KHÍ

GVHD: Đỗ Trọng Hợp

SVTH: Team 21 - Phạm Đức Thể

Nội Dung


1
Bài toán
2
Bộ dữ liệu
3
Xử lý dữ liệu
4
Exploratory
Data Analysis
5
Phân tích ANOVA
6
Thực nghiệm
7
Kết quả
8
Kết luận

Bài Toán

- 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.

Bộ Dữ Liệu

Thông tin	Nội dung
Tên bộ dữ liệu	Air Quality Data Set (UCI)
Nguồn thu thập và cách thức thu thập	Dữ liệu 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. Thiết bị được đặt trên cánh đồng ở một khu vực ô nhiễm nghiêm trọng trong một thành phố của Ý. Dữ liệu được ghi lại từ tháng 3 năm 2004 đến tháng 2 năm 2005 (một năm).
Số thuộc tính	15
Số dòng dữ liệu	9357
Missing values	Các giá trị bị thiếu được gắn thẻ với giá trị -200.
Thông tin tác giả	Saverio De Vito (saverio.devito@enea.it), ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development

Bộ Dữ Liệu

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 theo 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)

Bộ Dữ Liệu

Index	Thuộc tính	Ý nghĩa
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	Nhiệt độ (°C).
13	RH	Relative Humidity - Độ ẩm tương đối (%).
14	AH	Absolute Humidity - Độ ẩm tuyệt đối.

Bộ Dữ Liệu

	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

Xử lý dữ liệu

Datatype

```
<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
```

Kiểu dữ liệu gốc

```
<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
```

Kiểu dữ liệu đã được chuẩn hóa

Xử lý dữ liệu

Missing values

	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

- Bỏ thuộc tính NMHC_GT vì tỉ lệ missing values > 50%.
- Vì thuộc tính CO_GT là thuộc tính mục tiêu của bài toán nhưng lại có tỉ lệ missing values khá cao (~18%), nên chúng tôi đề xuất 2 chiến lược xử lý là:
  - REMOVE: Xóa tất cả các dòng dữ liệu bị missing values của thuộc tính CO_GT.
  - MEAN: Điền các missing values bằng giá trị trung bình của thuộc tính CO_GT.
- Các thuộc tính missing values khác chúng ta sẽ xử lý bằng cách điền bằng giá trị trung bình (mean) của từng thuộc tính.

Exploratory Data Analysis

Thống kê mô tả

	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

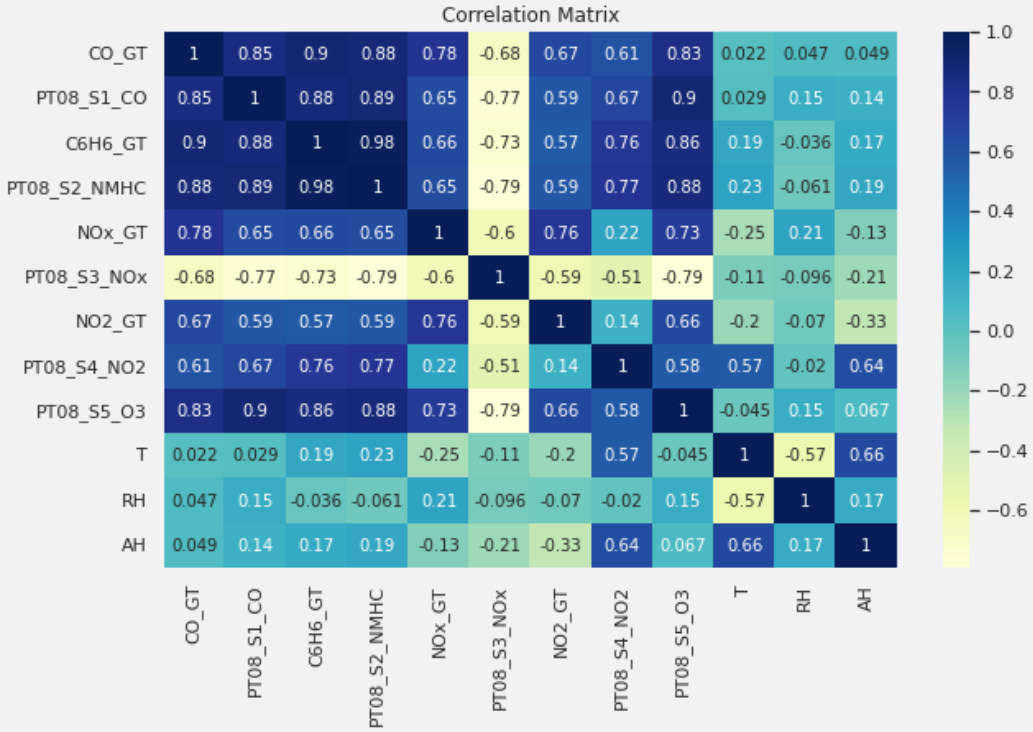
Air Quality–REMOVE

	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	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

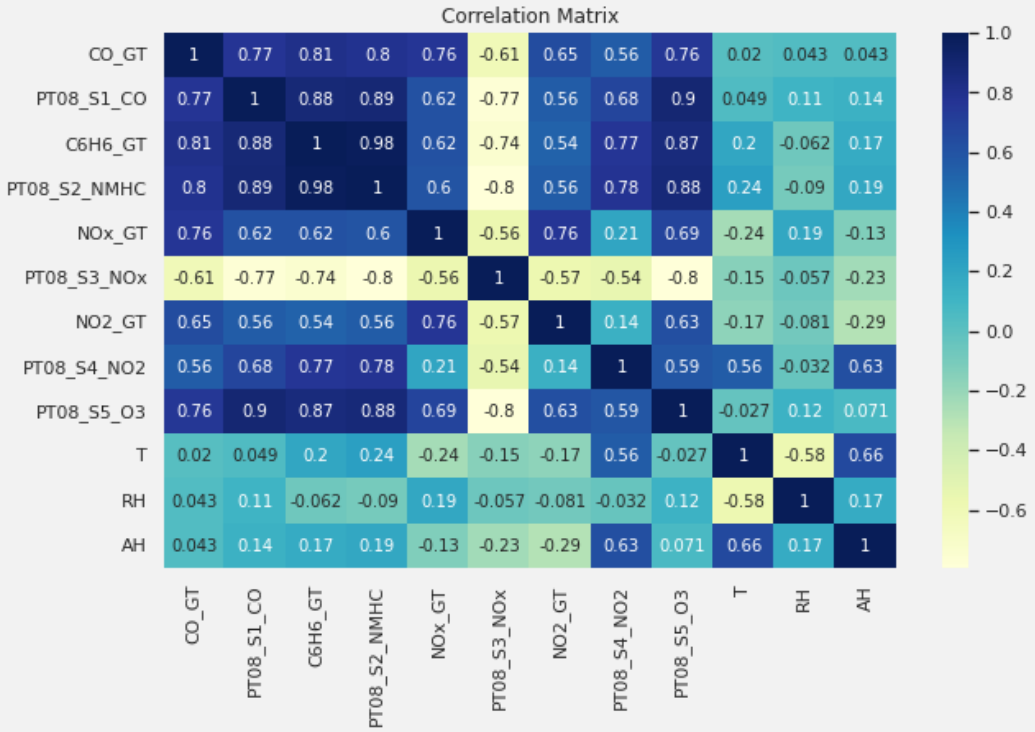
Air Quality–MEAN

Exploratory Data Analysis

Ma trận tương quan



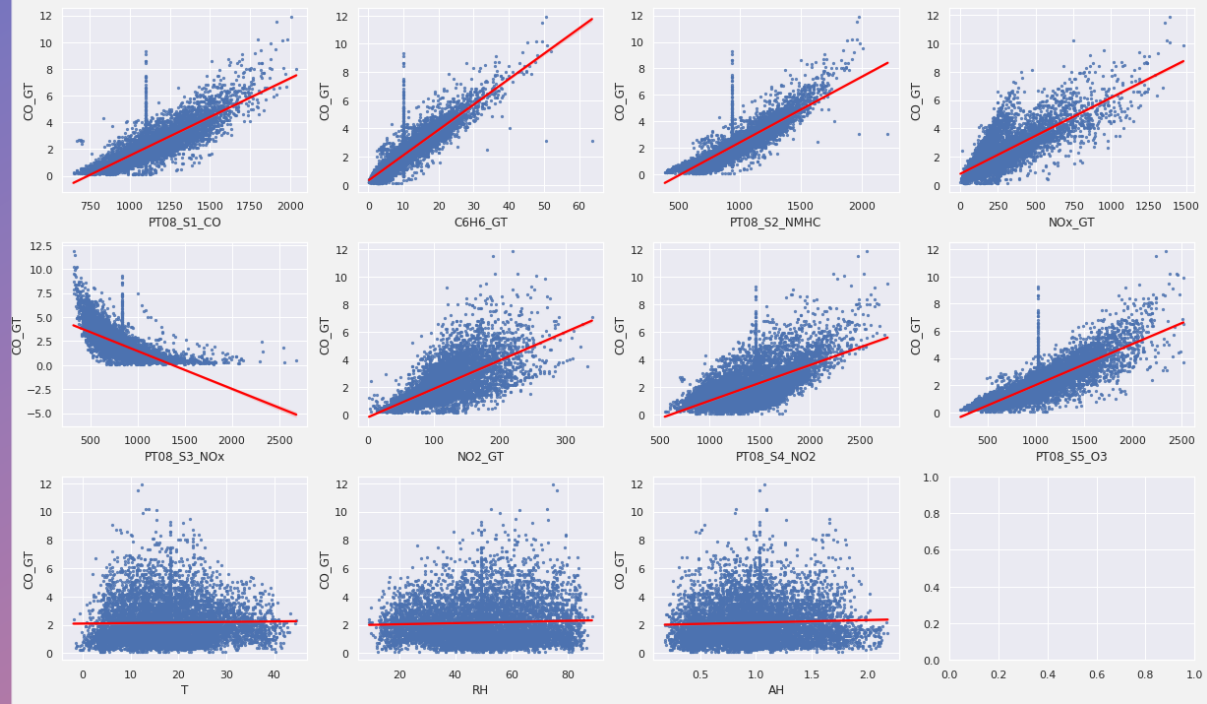
Air Quality–REMOVE



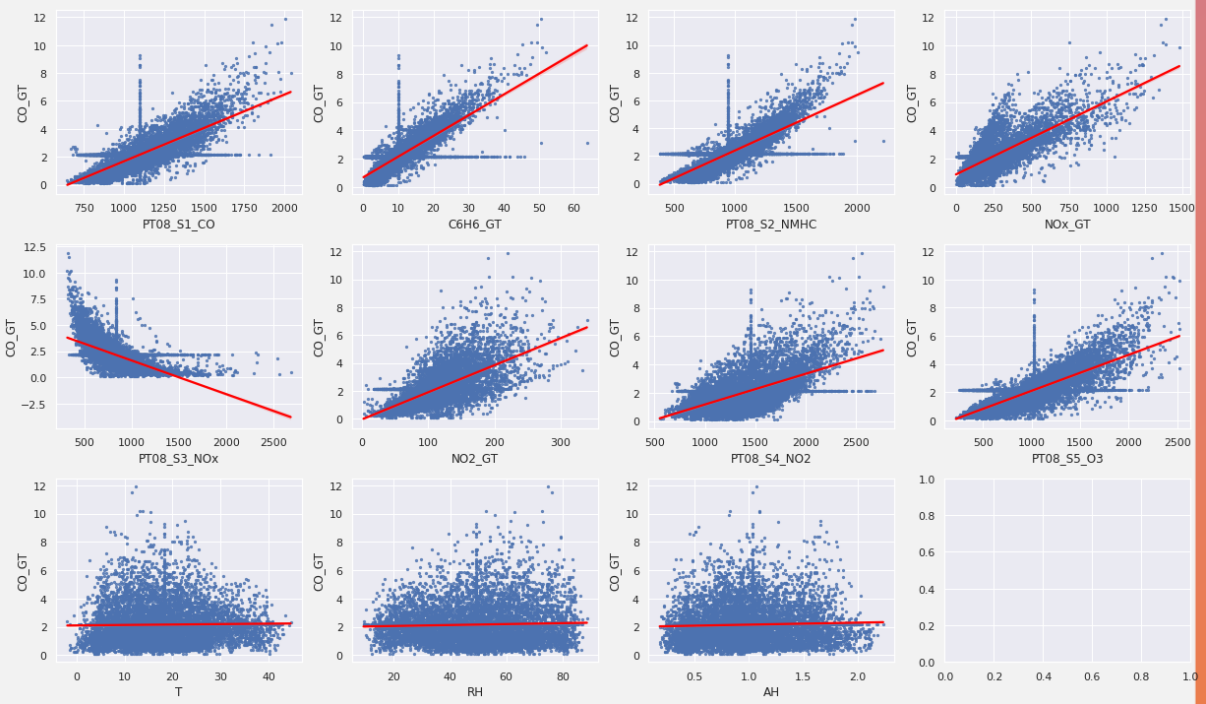
Air Quality–MEAN

Exploratory Data Analysis

Regression plot



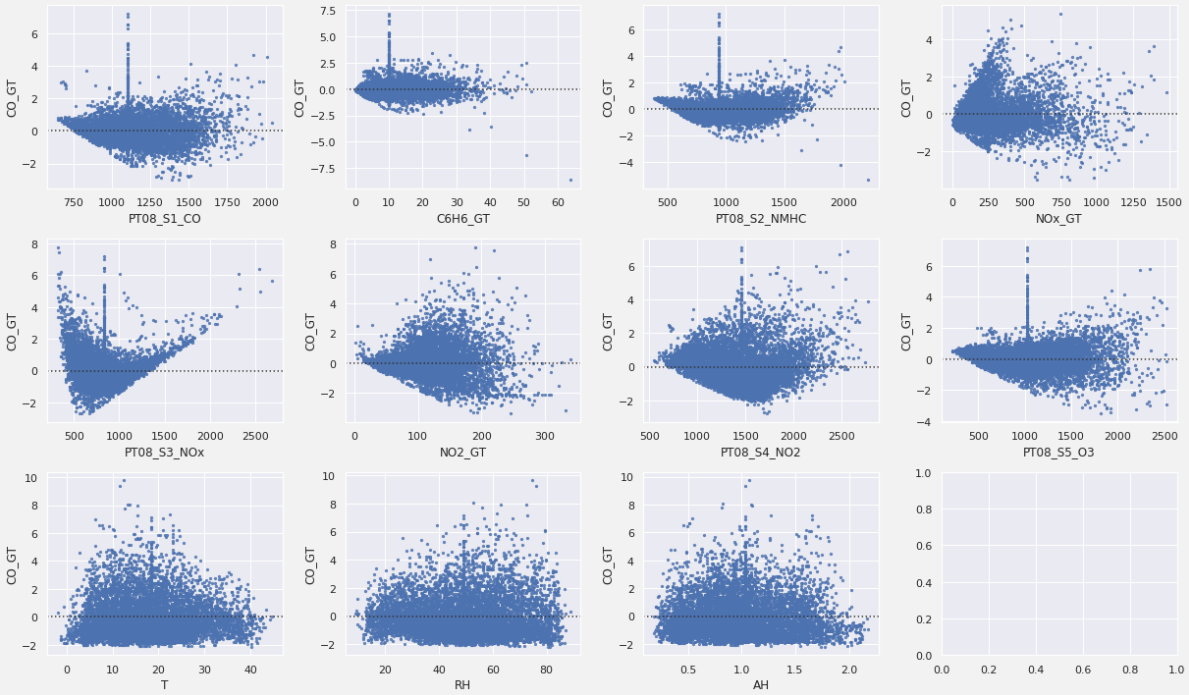
Air Quality–REMOVE



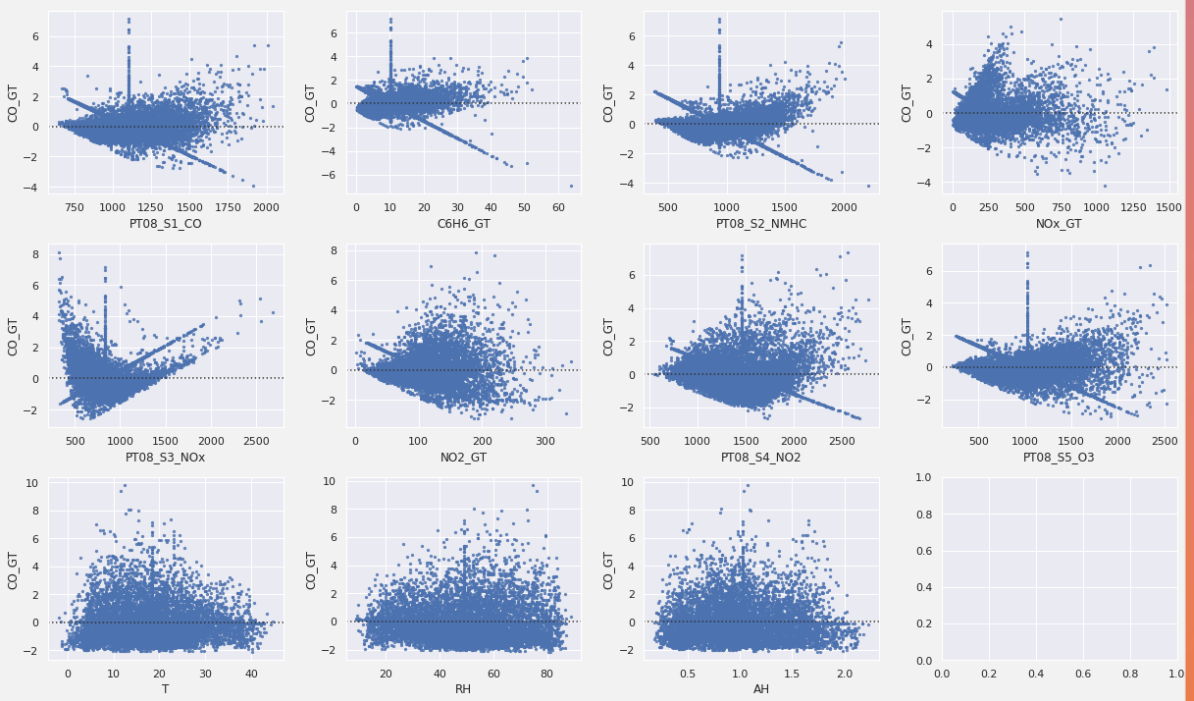
Air Quality–MEAN

Exploratory Data Analysis

Residual plot



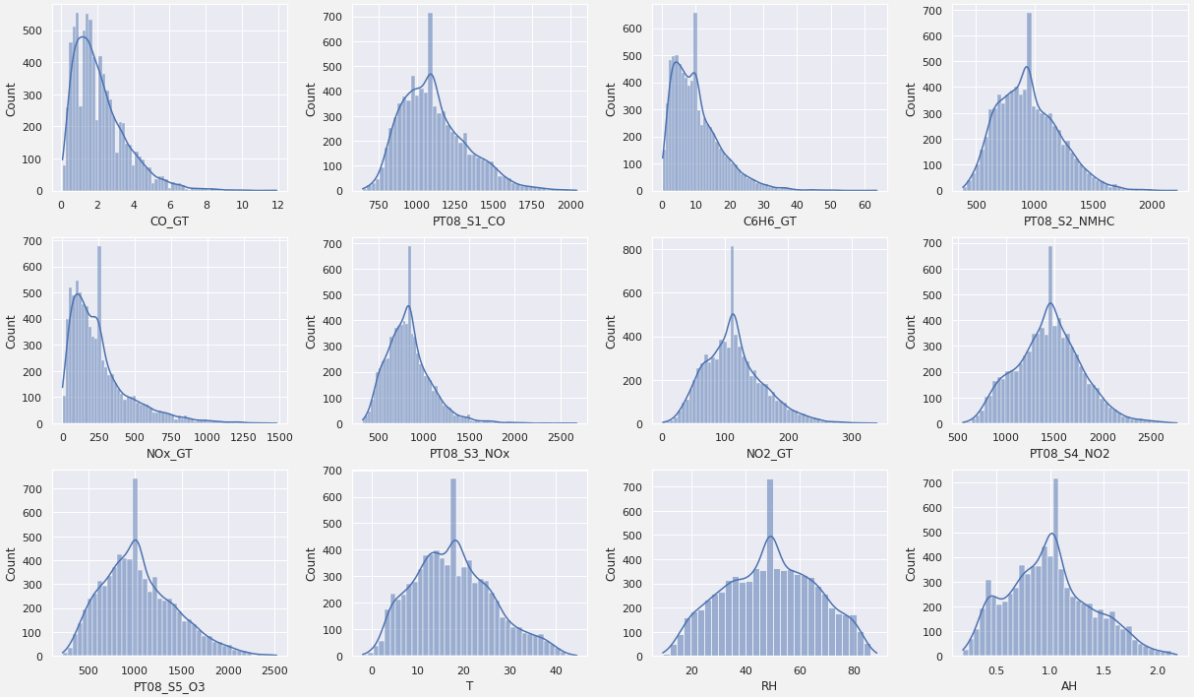
Air Quality–REMOVE



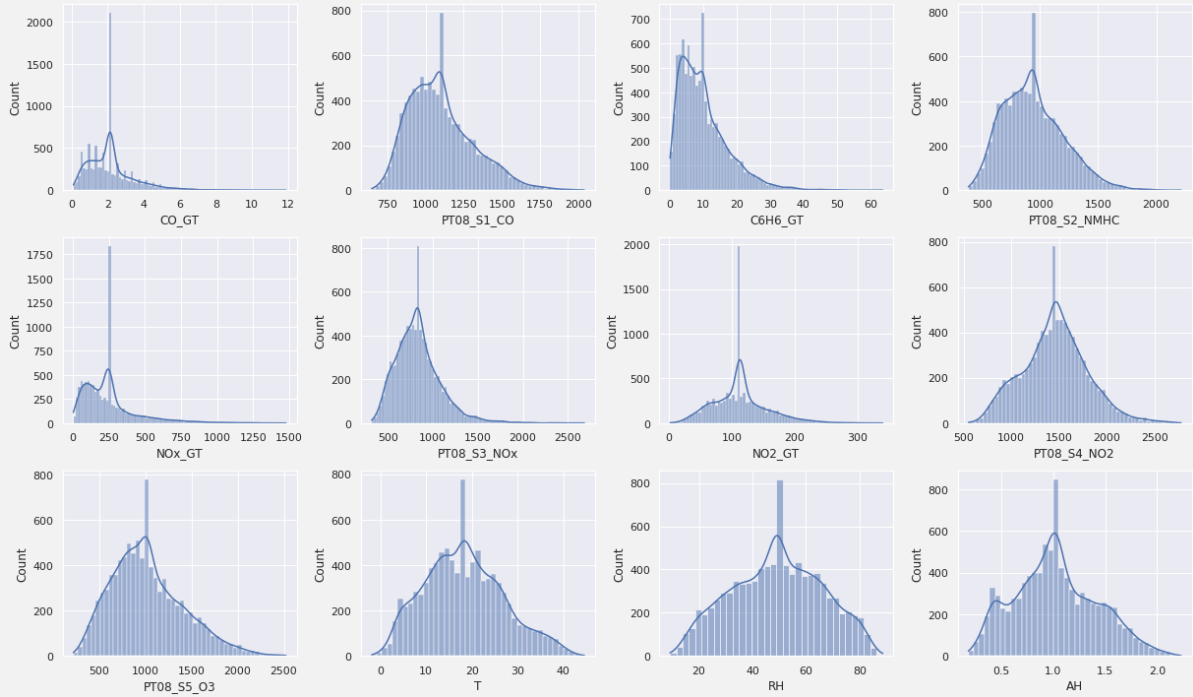
Air Quality–MEAN

Exploratory Data Analysis

Histogram plot



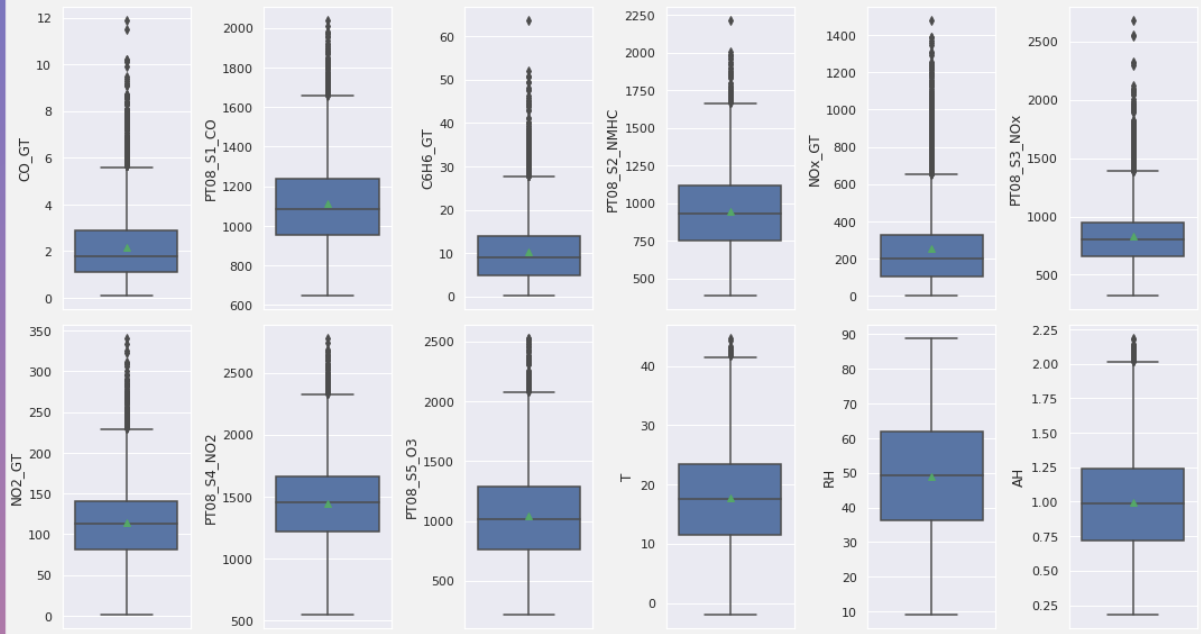
Air Quality–REMOVE



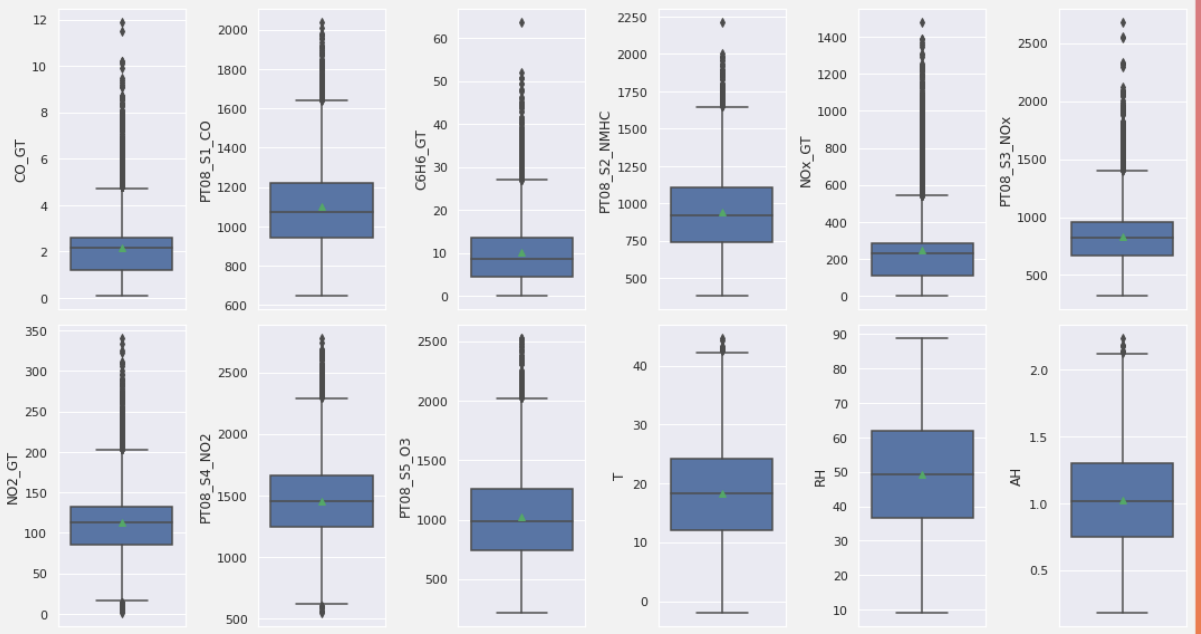
Air Quality–MEAN

Exploratory Data Analysis

Box plot



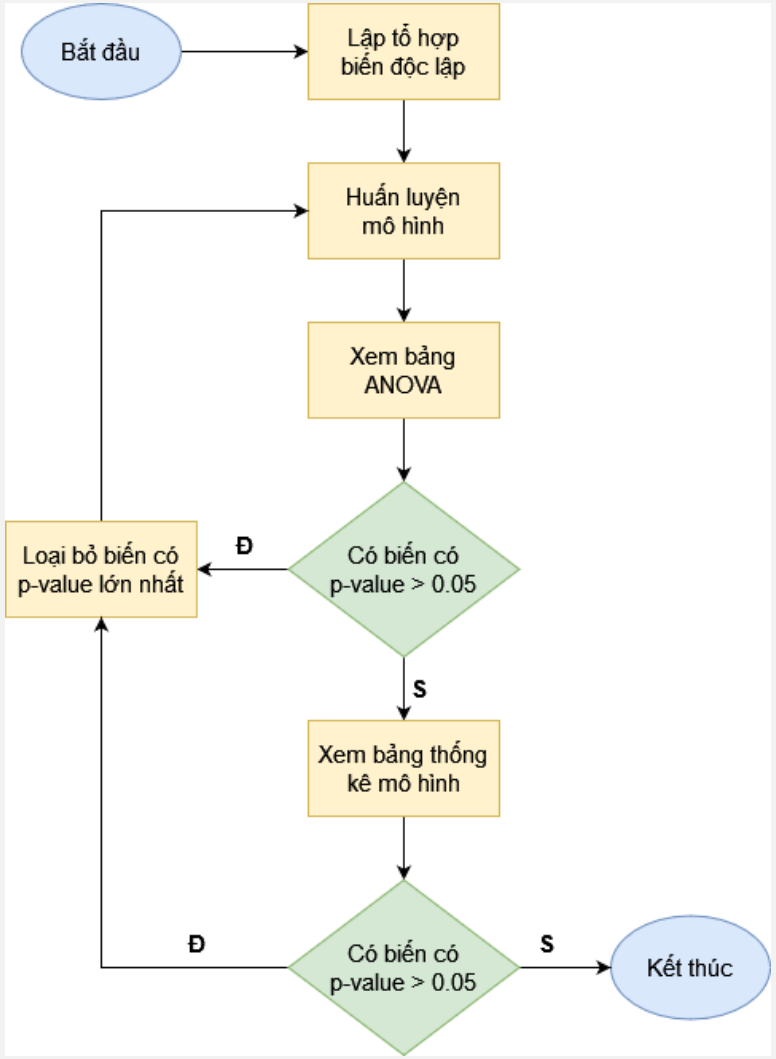
Air Quality–REMOVE



Air Quality–MEAN

Phân Tích ANOVA

Quy trình phân tích ANOVA



Phân Tích ANOVA

Air Quality–REMOVE

ANOVA đơn thuộc tính

	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

ANOVA lần 1

	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

ANOVA lần 2

Phân Tích ANOVA

Air Quality–REMOVE

```
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
```

Summary mô hình hồi quy lần 1

ANOVA đơn thuộc tính

```
             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
```

ANOVA lần 3

Phân Tích ANOVA

Air Quality–REMOVE

ANOVA đơn thuộc tính

```
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
```

Summary mô hình hồi quy lần 2

Phân Tích ANOVA

Air Quality–REMOVE

ANOVA đơn thuộc tính

Mô hình hồi quy có dạng:

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

Bỏ được 2 thuộc tính: PT08_S2_NMHC và AH

Phân Tích ANOVA

Air Quality–REMOVE

	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	***

ANOVA tương tác 2 thuộc tính

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

Phân Tích ANOVA

Air Quality–REMOVE

	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

ANOVA tương tác 2 thuộc tính

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

Phân Tích ANOVA

Air Quality–REMOVE

ANOVA tương tác 2 thuộc tính

Mô hình hồi quy có dạng:

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

Phân Tích ANOVA

Air Quality–MEAN

ANOVA đơn thuộc tính

Mô hình hồi quy có dạng:

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

Không bỏ được thuộc tính nào (giống với bộ dữ liệu gốc).

Phân Tích ANOVA

Air Quality–MEAN

ANOVA tương tác
2 thuộc tính

Mô hình hồi quy có dạng:

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

Thực Nghiệm

Chia dữ liệu train/test

Với từng bộ dữ liệu, chúng tôi chia 80% cho tập training và 20% cho tập testing.

		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.

Thực Nghiệm

Áp dụng các thuật toán ML và DL:

- **Thuật toán ML**
  - Linear Regression
  - Decision Tree Regression
  - Random Forest Regression
  - Support Vector Regression
- **Thuật toán DL**
  - Neural Network

Thực Nghiệm

Độ đo đánh giá:

❖ **Sử dụng các độ đo sau:**

- R squared (R^2)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)

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

Kết quả train/test

			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.

Kết quả

Nhận xét kết quả trên tập test

- 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.
- 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.
- 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).

Kết luận

Đã làm được:

- Phân tích và xây dựng mô hình dự đoán nồng độ CO trong không khí dựa trên bộ dữ liệu Air Quality.
- 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.

Kết luận

Đã làm được:

- Áp 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.

Kết luận

Hướng phát triển:

- Bộ dữ liệu: Xử lý các missing values tốt hơn nữa, ta có thể 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 đoan tốt hơn nữa.

THANKS FOR WATCHING!