미세먼지 원인 분석 및 제안

B반 1조 장선영

배경

- **미세먼지란?**

  이황산가스, 질소산화물, 납, 오존, 일산화탄소 등을 포함하는 대기오염 물질로 자동차, 공장, 조리 과정 등에서 발생하며 대기 중 장기간 떠다니는 입경 10㎛ 이하의 미세한 먼지

- **미세먼지가 우리 삶에 미치는 영향**

  - **건강**
  미세먼지에 장기간 노출되게 되면 사람의 인체 건강에 상당히 안 좋은 영향 유발

  - **환경**
  토양 황폐화로 비닐하우스에 쌓이면 일조량 감소 및 농작물의 광합성 방해

  - **경제**
  미세먼지로 인한 불황 및 미세먼지에 민감한 반도체나 디스플레이 등 산업에 미치는 악영향으로 경제적 손실 발생 (현대경제연구원 출처 경제적 손실을 연간 4조230억 원 규모로 추정)

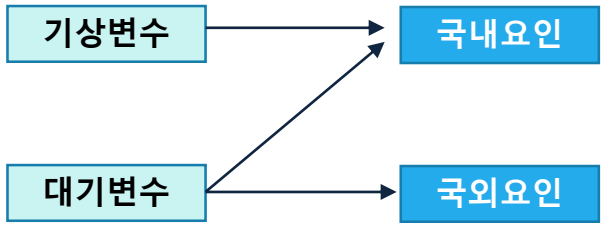
사전 자료 조사

- **대기 오염**

환경부의 국내 환경동향 보고에서 미세먼지의 발생원인은 다음과 같이 분석하고 있다

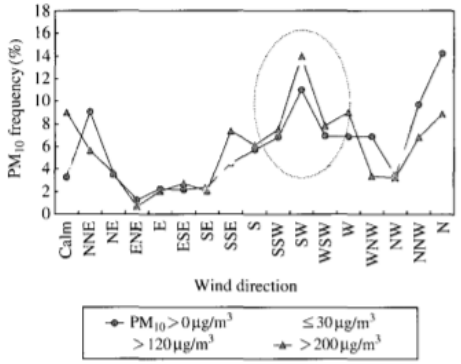
1) 국내 요인
국내에서 생성된 미세먼지의 비중 50~70%, 서울의 미세먼지 가운데 51%는 국내 생성
국내 미세먼지 배출 원인 중 68.2%는 화석연료를 연소시키면서 배출하는 입자들

2) 국외 요인
국내 미세먼지의 43% 정도는 중국 발로 공장의 매연이나 자동차의 배기가스, 사막의 황사 등이 원인



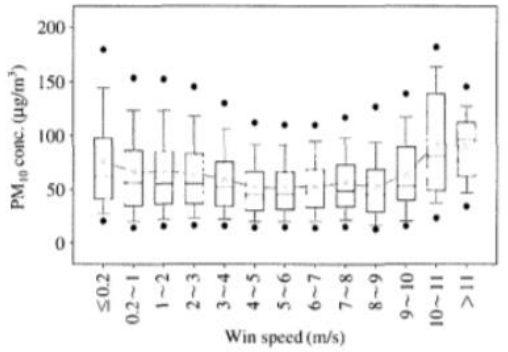
- **기상정보**

1) 풍향



서풍계열 바람이 불 때 미세먼지 농도 高

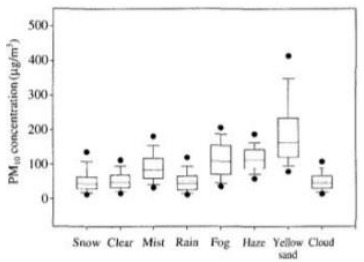
2) 풍속



0~6 m/s 구간은 무풍 시 가장 높은 농도, 풍속이 증가함에 따라 농도 감소,
6 m/s 초과 구간에서는 증가함에 따라 농도도 상승

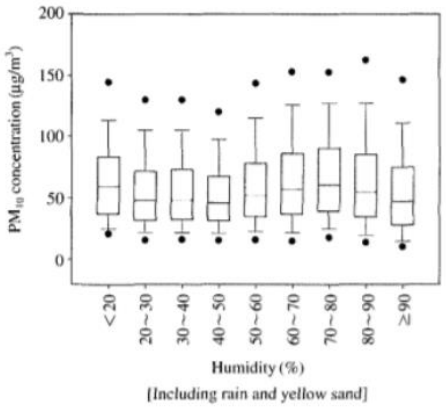
3) 일기 유형

Weather condition	N	Mean (μg/m³)	S.D. (μg/m³)
Clear	18,666	54.0	31.0
Cloud	19,591	51.2	30.4
Rain	6,557	50.6	34.7
Snow	791	53.1	38.2
Mist	13,583	92.3	47.2
Fog	955	114.1	54.5
Haze	140	117.4	40.2
Yellow sand	875	196.0	113.4



황사와 안개가 낀 날씨일 때 미세먼지 농도 高

4) 습도



60~90%일 때 미세먼지 농도 高
높은 습도가 대기 중의 황산염과 질산염 등의 2차 먼지 생성을 촉진하기 때문

분석 목표

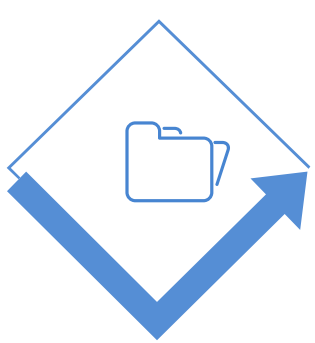
배경 및 사전 조사 결과 여러 요인들이 미세먼지 농도에 영향을 미치고 있음을 알 수 있었다.
이에 주어진 미세먼지 관련 데이터도 이러한 배경과 조사 결과와 동일한지 확인할 필요성이 있으며 이에 근거하여 대안 또한 제시할 필요성이 있다.

따라서 분석 목표를 요약하면 다음과 같다.

미세먼지의 경우 다방면으로 악영향을 끼치기 때문에 어떤 인자가 미세먼지 농도에 영향을 주는지 파악하여 미세먼지 농도를 감소시킬 수 있는 방안을 찾고자 함

분석 계획

Process



1. 데이터 구성

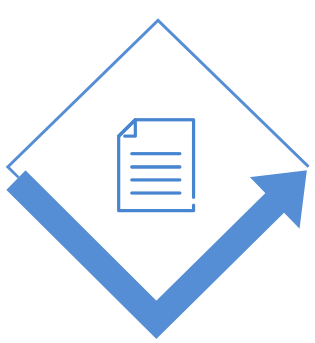
- 데이터 가져오기
- 데이터 확인하기

2. 데이터 품질

- 결측치 확인 및 처리
- 이상치 확인 및 처리
- 파생변수

3. 시각화

- 목표변수
- 설명변수
- 파생변수



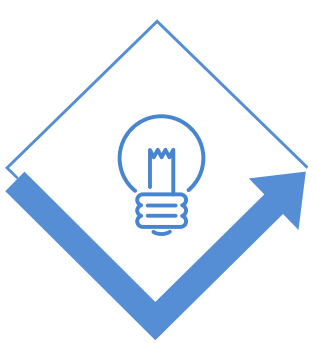
4. 검정

- 상관분석
- 2-Sample t-test

5. 모델링

- 다중회귀분석
- 의사결정나무
- 랜덤포레스트
- 그래디언트 부스팅

6. 모델 평가



7. 결론

데이터 구성 및 품질 확인

1. 데이터 구성

- 기간: 2019년 7월 1일 ~ 202-년 6월 30일
- 목표변수: PM10
- 설명변수: O3, NO2, CO, SO2, TEMP, RAIN, WIND, WIND_DIR, HUMIDITY, ATM_PRESS, SNOW, CLOUD

<참고>
국내환경부 기준

미세먼지농도 (㎍/㎥, 일평균)	좋음	보통	나쁨	매우나쁨
PM10	0~30	31~80	81~150	151이상

WHO 기준

미세먼지농도 (㎍/㎥, 일평균)	좋음	보통	나쁨	매우나쁨
PM10	0~20	21~45	46~75	76이상

2. 데이터 품질 확인

- 결측치

변수	개수
PM10	1
O3	1
NO2	55
CO	1
SO2	1

- 2020-05-42 에서 다수의 결측치 발견
  → 전체 제거
- NO2의 경우 계절 별 NO2 평균으로 입력

- 이상치

사용자 정의 함수로 각 변수별 IQR방식으로 이상치를 탐지

But,

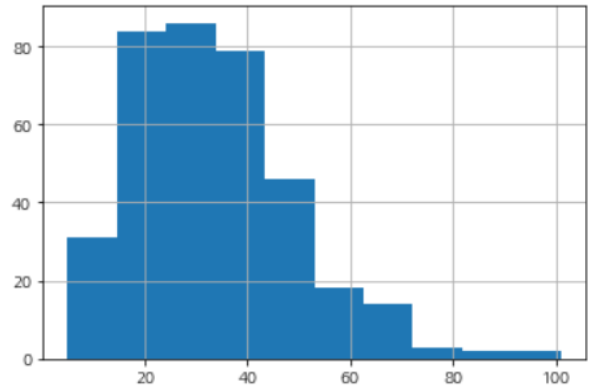
따로 이상치에 대한 처리는 하지 않음

- 파생변수 설정

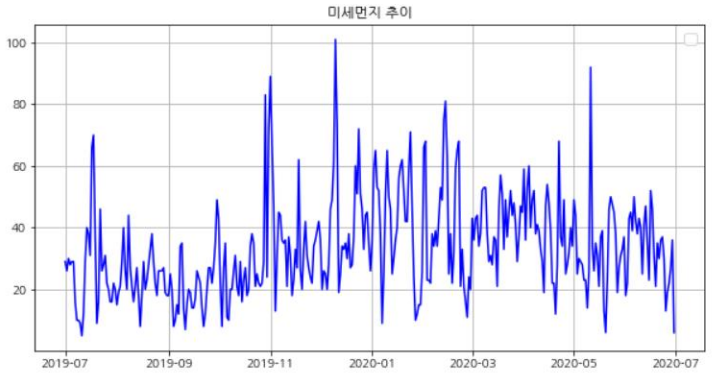
- PM10
: WHO 기준으로
좋음/보통/나쁨/매우 나쁨 범주화
- WIND_DIR
: 풍향을 8방위로 세부 범주화
- WIND
: 풍속에 따른 세부 범주화
- 계절에 따른 구분
- 비/눈 유무에 따른 구분

시각화

1. 목표변수 (PM10)



히스토그램: 분포 확인



선 그래프: 시간에 따른 미세먼지 추이

히스토그램과 선 그래프를 통해
미세먼지 추이 및 분포 확인

2. 설명변수 (1)

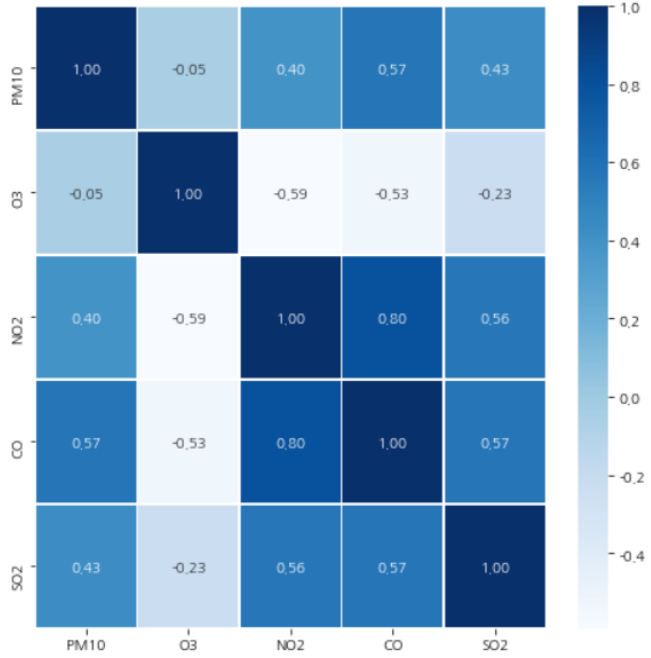
주어진 x변수들을 크게 두 가지 기준으로 분류

1) 기상변수

: TEMP, RAIN, WIND, WIND_DIR, HUMIDITY, ATM_PRESS, SNOW, CLOUD

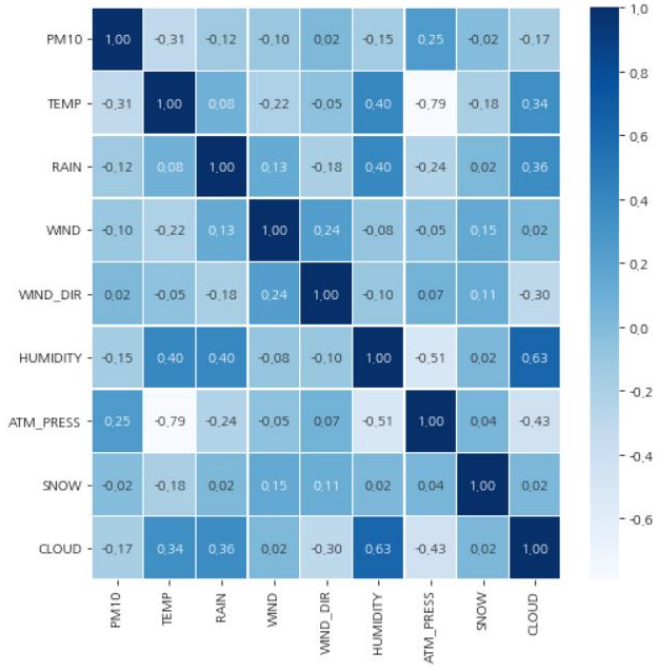
2) 대기변수

: O3, NO2, CO, SO2



→ 기상변수는 국지성 특징으로 국내요인

→ PM10과 NO2, CO1, SO2 상관성 有



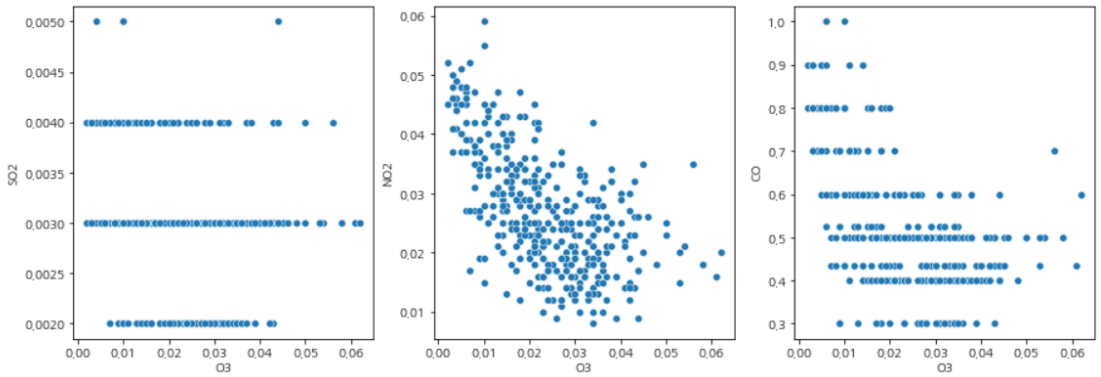
→ 대기변수는 국내/국외요인

→ PM10과 별다른 상관성 無

시각화

2. 설명변수 (2)

1) 기상변수(상세)



→ O3: PM10에 직접적으로 영향 X

BUT, SO2, NO2, CO와는 산포도를 통해 음의 상관성 有

즉, O3는 대기오염에 직접적인 영향을 주진 않지만 간접적으로 대기오염과 연관이 있다는 것을 알 수 있음

<상관성 분석 실시>
PM10과 O3

```
Correlation Analysis
corr:-0.052
p-value:0.324
```

PM10과 O3은 상관성이 있다고 할 수 없다. (H0 채택)

<상관성 분석 실시>
PM10과 NO2/CO/SO2

```
Correlation Analysis
corr:0.396
p-value:0.000
```

PM10과 NO2는 약한 상관성이 있다고 할 수 있다. (H0 기각)

```
Correlation Analysis
corr:0.573
p-value:0.000
```

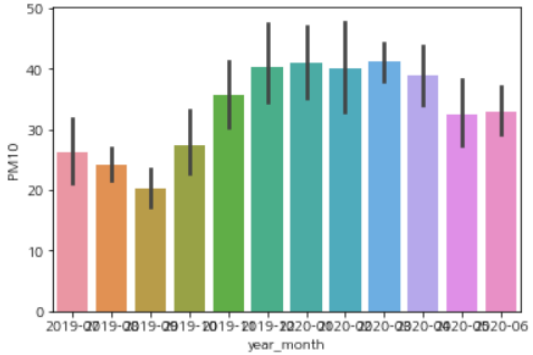
PM10과 CO는 상관성이 있다고 할 수 있다. (H0 기각)

```
Correlation Analysis
corr:0.429
p-value:0.000
```

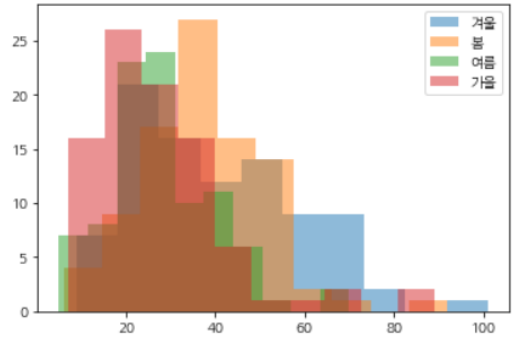
PM10과 SO2는 상관성이 있다고 할 수 있다. (H0 기각)

3. 파생변수

1) 계절



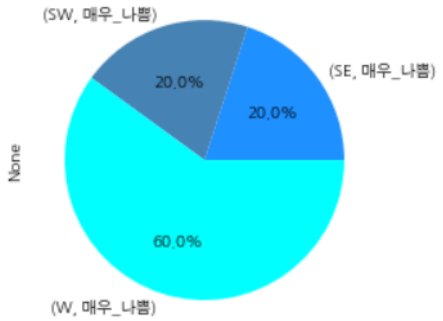
→ 계절별 상이한 PM10

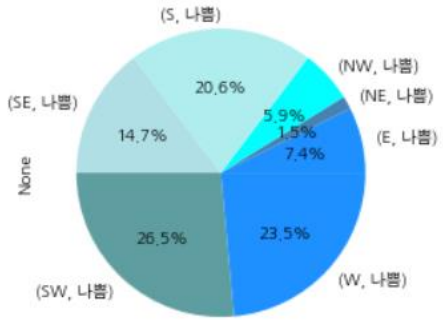


→ 겨울에 특히 PM10 수치가 높음

2) 풍향

- N/S/W/E/SW/SE/NE/NW로 설정
- PM10: WHO기준으로 좋음/보통/나쁨/매우 나쁨으로 설정



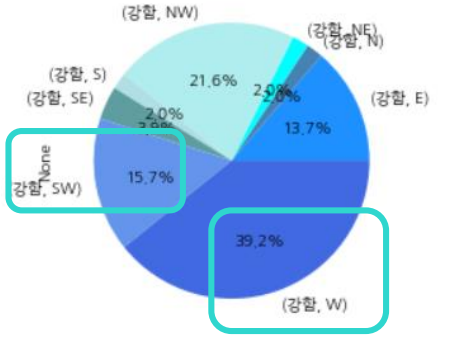


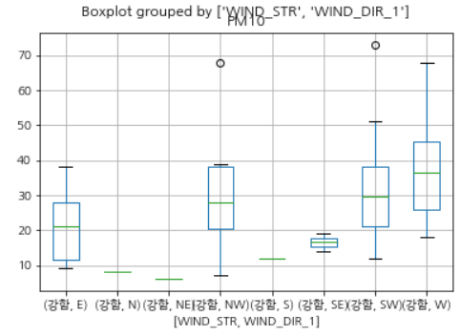
PM10이 '매우 나쁨' & '나쁨'의 경우 W, SW가 50% 이상을 차지
→ 중국의 영향 가능성 大 (추가 근거 확보 필요)

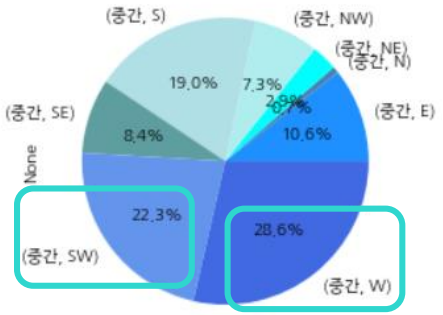
시각화

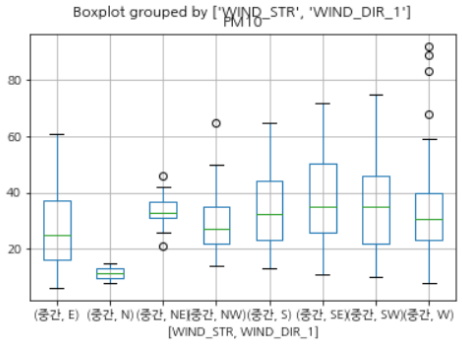
3. 파생변수

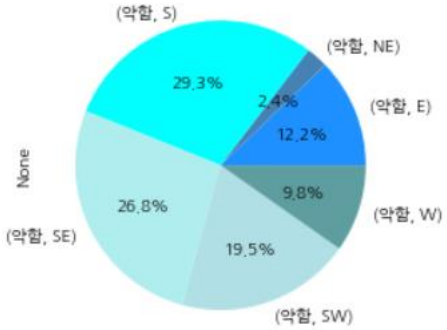
3) 풍속

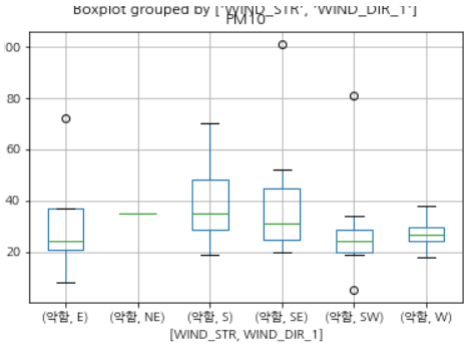













```
corr_ana(df_raw_h4['WIND'], df_raw_h4["PM10"])
```

```
Correlation Analysis
corr:-0.033
p-value:0.773
```

(예상) 국외 요인 중 중국발 미세먼지 영향이 클 것이다.

→ [가설] 풍속이 강하고, 서풍이면 PM10이 높을 것 이다.

상관분석 결과, 풍속과 PM10은 상관성 無

BUT

Boxplot을 참고하면, 풍속이 강함/중간일 때, W, SW에서의 PM10 수치가 높거나 위로의 이상치가 존재하는 것을 알 수 있음

<모델링에 활용할 유발 인자 선택>

NO2 / CO / SO2 / **기압** / **기온** / HUMIDITY / **계절**(파생변수) / **풍향** (파생변수) / **풍속** (파생변수)

모델링

다중회귀분석

OLS Regression Results

Dep. Variable:	PM10	R-squared:	0.489
Model:	OLS	Adj. R-squared:	0.480
Method:	Least Squares	F-statistic:	57.09
Date:	Mon, 08 Nov 2021	Prob (F-statistic):	2.57e-49
Time:	01:30:52	Log-Likelihood:	-1408.6
No. Observations:	365	AIC:	2825.
Df Residuals:	358	BIC:	2852.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	324.1706	140.082	2.314	0.021	48.683	599.658
O3	631.5493	72.518	8.709	0.000	488.935	774.164
NO2	405.6936	120.883	3.356	0.001	167.964	643.423
CO	64.4865	7.629	8.453	0.000	49.484	79.489
TEMP	-0.6233	0.115	-5.406	0.000	-0.850	-0.397
WIND_DIR	0.0356	0.010	3.595	0.000	0.016	0.055
ATM_PRESS	-0.3461	0.139	-2.489	0.013	-0.620	-0.073

Omnibus:	96.804	Durbin-Watson:	1.165
Prob(Omnibus):	0.000	Jarque-Bera (JB):	265.903
Skew:	1.242	Prob(JB):	1.82e-58
Kurtosis:	6.364	Cond. No.	2.66e+05

```
MAPE(df_test_y, reg_y_pred)
```
31.146956791735402

- R-squared: 48.6%의 설명력
- Prob(F-statistics): 2.57e-49 (통계적으로 유의)

최종 모델 회귀식

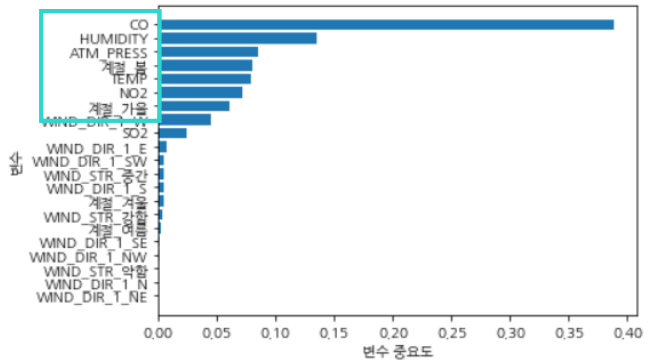
Y_hat = 324.1706 + 631.5493 O3 + 405.6936 NO2 + 64.4865 CO – 0.6233 TEMP – 0.3451 ATM_PRESS

MAPE 값: 31.13

50이하로 매우 합리적인 예측이라 판단 가능 (Tofallis, 2016)

랜덤포레스트

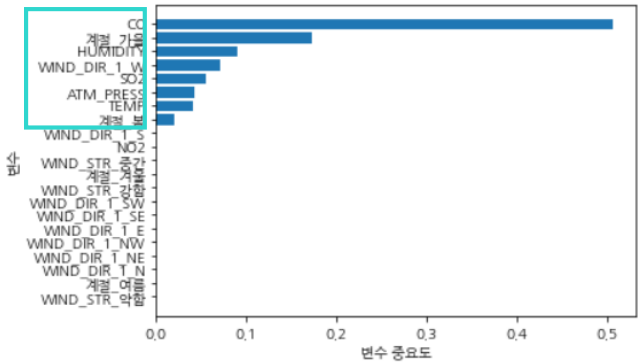
Score on training set:0.698
Score on test set:0.447



- N_estimator: 50
- Leaf: 5
- Split : 10
- Max depth: 8

의사결정나무

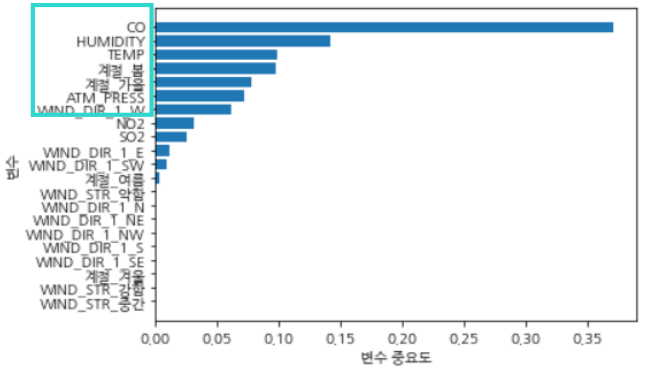
Score on training set:0.454
Score on test set:0.222



- Leaf : 10
- Split : 22
- Max depth: 4

그래디언트 부스팅

Score on training set:0.721
Score on test set:0.450

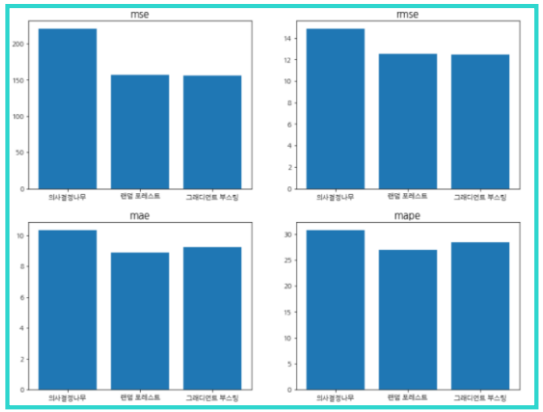


- N_estimator: 70
- Leaf: 22
- Max depth: 3
- Learning rate: 0.1

결론

결론

- **분석 결과**
  - PM10 수치는 기상요인은 상관성이 약하고 대기요인과 강한 상관성을 가짐
  - 계절은 PM10 관련해서 상당히 중요한 요인으로 작용
  - 8방위 풍향 기준, 서풍/남서풍의 경우 미세먼지가 '나쁨' / '매우 나쁨'이 많음
    → 중국발 가능성 高
    (추가 근거 필요)



- **모델링**
  - 다중회귀분석/의사결정나무/랜덤포레스트/그래디언트 부스팅
  - 랜덤포레스트와 그래디언트 부스팅이 높은 정확도를 보임
    BUT, 50% 미만의 모델 정확도로 실무에서 쓰이기엔 부적합 판단
  - 공통적인 주요 변수
    : CO, HUMIDITY, ATM_PRESS, 계절

제안

- **국외요인**
  : 분석 결과, 주된 요인인 중국발 미세먼지 절감을 위해 국가간 협의 必
- 예측 모델링을 통해 향후 미세먼지 예측을 통해 실생활 대비

한계점

- 데이터 양 부족으로 인한 낮은 test 정확도
- '(예상 가설) 국외 요인 중 중국발 미세먼지 영향이 큼' 을 증명하기 위해 추가 자료 필요
  <예시>
  - 우리나라 서쪽(백령도) / 동쪽(울릉도) / 남쪽(제주도)을 대표하는 지역의 상세 미세먼지 비교
  - 실제로 중국발 미세먼지인지 확인을 위한 중국내 기간 내 미세먼지 자료
- COVID19 영향 고려 X
  - COVID19로 인한 제조생산활동 감소 등의 이유로 국내/국외 미세먼지 발생 원인 감소에 대한 고려는 일절 하지 않음

(참고)

측정 일자	변수	변수 설명	변수 역할	변수 형태	분석 제외 사유	탐색적 기법			모델링기법			
						그래프	검정	상관분석	회귀분석	DT	RF	GB
1	MeasDate	측정일자	제외	연속형	제외							
2	PM10	미세먼지10㎍/㎥	설명변수	연속형		line, histogram						
3	O3	오존 농도	설명변수	연속형	간접적으로 영향	heatmap, box plot, scatterplot		상관성 분석	O			
4	NO2	이산화질소 농도	설명변수	연속형		heatmap, box plot, scatterplot		상관성 분석	O	O	O	O
5	CO	일산화탄소 농도	설명변수	연속형		heatmap, box plot, scatterplot		상관성 분석	O	O	O	O
6	SO2	아황산가스 농도	설명변수	연속형		heatmap, box plot, scatterplot		상관성 분석		O	O	O
7	TEMP	기온(°C)	설명변수	연속형		heatmap		상관성 분석	O	O	O	O
8	RAIN	강수량(mm)	설명변수	연속형	상관성 낮음	heatmap	2 sample t-test					
9	WIND	풍속(m/s)	설명변수	연속형	*범주형으로 변경후 사용	heatmap, box plot, pie chart		상관성 분석		O	O	O
10	WIND_DIR	풍향(16방위)	설명변수	연속형	*범주형으로 변경후 사용	heatmap, box plot, pie chart		상관성 분석	O	O	O	O
11	HUMIDITY	습도(%)	설명변수	연속형		heatmap				O	O	O
12	ATM_PRESS	현지기압(hPa)	설명변수	연속형	*범주형으로 변경후 사용	heatmap			O	O	O	O
13	SNOW	적설(cm)	설명변수	연속형	상관성 낮음	heatmap	2 sample t-test					
14	CLOUD	전운량(10분위)	설명변수	연속형	상관성 낮음	heatmap						
기타	계절	사계절	파생변수	범주형		heatmap, box plot, histogram, bar plot	2 sample t-test	상관성 분석		O	O	O


End of Document