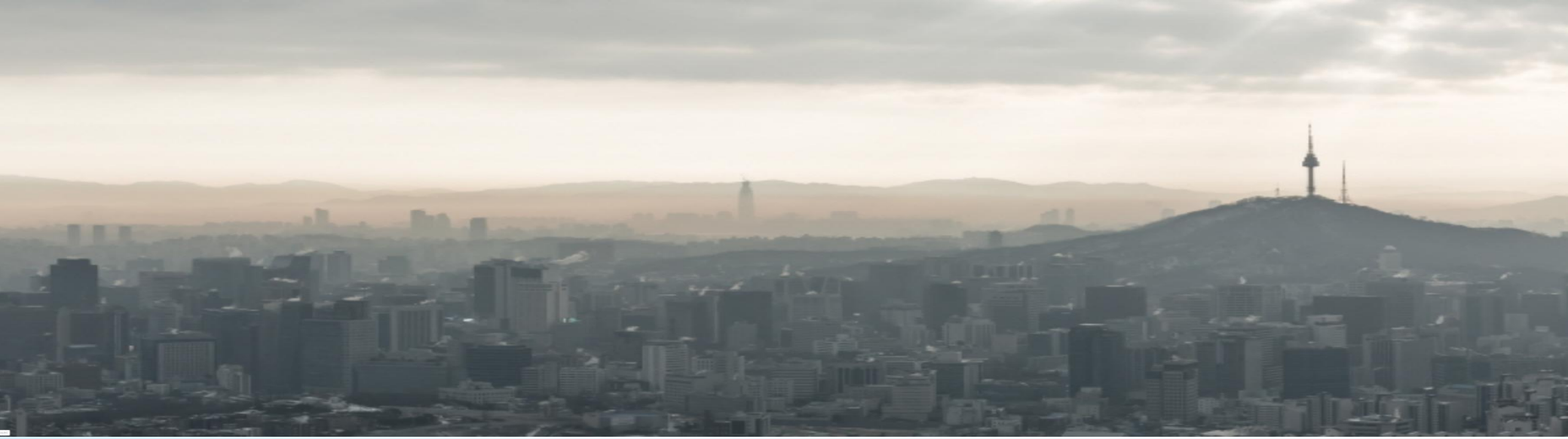


# 미세먼지 유발 영향인자 분석 및 대안제시

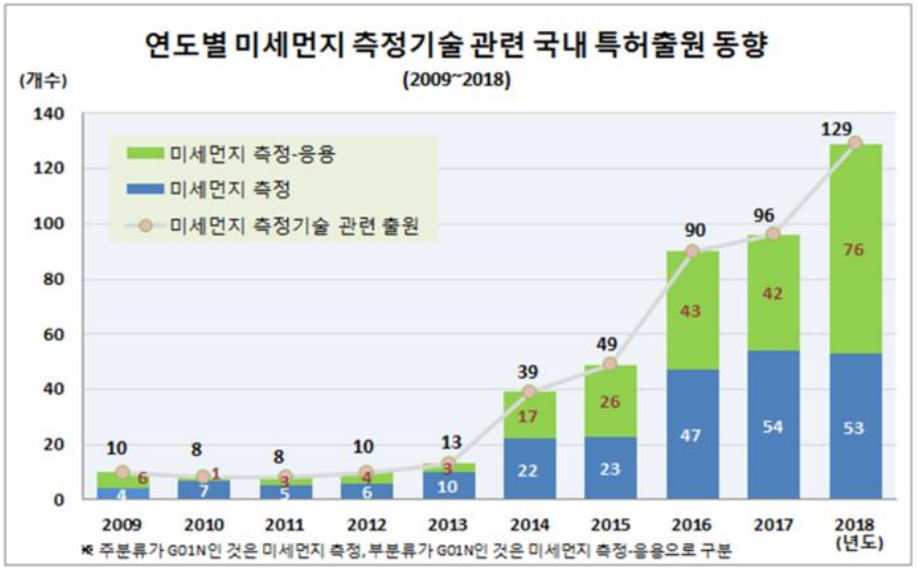
# 과제 정의

미세먼지는 눈에 보이지 않을 만큼 매우 작기 때문에 대기 중에 머물러 있다 호흡기를 거쳐 폐 등에 침투하거나 혈관을 따라 체내로 이동하여 들어감으로써 건강에 나쁜 영향을 미칠 수도 있다.1)

세계보건기구(WHO)는 미세먼지(PM10, PM2.5)에 대한 대기질 가이드라인을 1987년부터 제시해 왔고, 2013년에는 세계보건기구 산하의 국제암연구소(IARC, International Agency for Research on Cancer)에서 미세먼지를 사람에게 발암이 확인된 1군 발암물질(Group 1)로 지정하였다.

**미세먼지에 대한 사회적 관심은 매년 늘어나는 추세**

**미세먼지에 관한 연구의 필요성 증가**




출처: 특허청

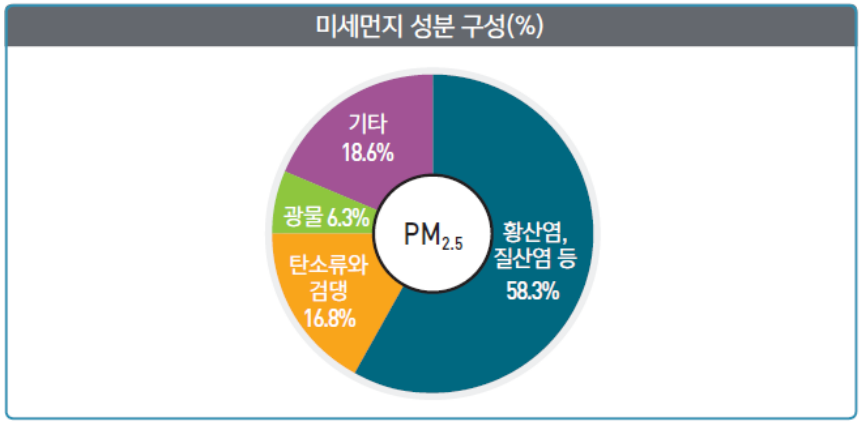
1) Jennifer A(2014) Fine particulate matter air pollution and cognitive function among U.S. older adults. Journals of Gerontology Series B: Psychological Sciences and Social Sciences Vol. 70 No. 2 322p ~ 330p 1079-5014 SCI(E)

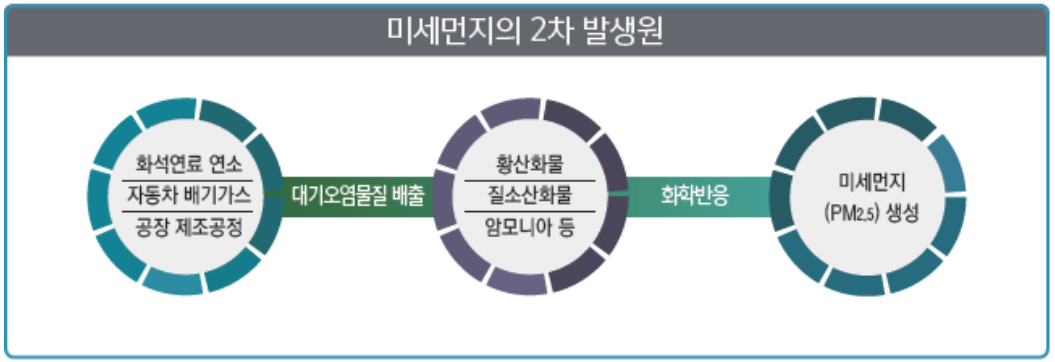
# 과제 정의

미세먼지는 일반적으로는 대기오염물질이 공기 중에서 반응하여 형성된 덩어리(황산염, 질산염 등)와 석탄 · 석유 등 화석연료를 태우는 과정에서 발생하는 탄소류와 검댕, 지표면 흙먼지 등에서 생기는 광물 등으로 구성

미세먼지는 1차와 2차적 발생으로 나눠진다.
- 1차적 발생: 굴뚝 등 발생원에서부터 고체 상태의 미세먼지로 나오는 경우
- 2차적 발생: 가스 상태로 나온 물질이 공기 중의 다른 물질과 화학반응을 일으켜 미세먼지가 되는 경우

## 우리는 이렇게 발생하는 미세먼지에 대해서 알아보도록 하겠다.




출처: 환경부(2016), 미세먼지 도대체 뭘까

# 분석 계획



2차적 발생에 의한 미세먼지는 각종 **유발 요인에 의해 생성**되며, 미세먼지를 이루는 성분은 그 미세먼지가 발생한 **지역이나 계절, 기상조건 등에 따라 달라질 수 있다**. 따라서 기상 관측 자료를 통해 미세먼지의 발생 원인을 파악해 영향 인자를 선정하고, 발생량을 예측해보고자 한다. 그리고 나아가서는 분석 결과를 토대로 미세먼지를 줄일 수 있는 방안에 대해 찾아보고자 한다.

- 2019년 7월 부터 2020년 6월까지의 미세먼지 및 기상 정보를 모아 놓은 “AIR_POLLUTION.csv” 데이터 이용
- 2차 미세먼지 발생원인을 기반으로 가설 수립
- 데이터의 이상치와 결측치 ,분포를 확인해서 데이터 정제
- 탐색적 분석을 통해 변수와의 관계를 확인해서 가설을 점검
- 모델 선정 및 영향 인자 분석으로 모델 설계
- 해당 모델에 적용시켜 예측 모델 개발 및 평가
- 개선안 도출 및 실무 적용 방안 제시

| | | | | |
|---|---|---|---|---|
| 1 | MeasDate | 측정일자 | 제외 | 연속형 |
| 2 | PM10 | 미세먼지10㎍/㎥ | 목표변수 | 연속형 |
| 3 | O3 | 오존 농도 | 설명변수 | 연속형 |
| 4 | NO2 | 이산화질소 농도 | 설명변수 | 연속형 |
| 5 | CO | 일산화탄소 농도 | 설명변수 | 연속형 |
| 6 | SO2 | 아황산가스 농도 | 설명변수 | 연속형 |
| 7 | TEMP | 기온(℃) | 설명변수 | 연속형 |
| 8 | RAIN | 강수량(mm) | 설명변수 | 연속형 |
| 9 | WIND | 풍속(m/s) | 설명변수 | 연속형 |
| 10 | WIND_DIR | 풍향(16방위) | 설명변수 | 연속형 |
| 11 | HUMIDITY | 습도(%) | 설명변수 | 연속형 |
| 12 | ATM_PRES | 현지기압(hPa) | 설명변수 | 연속형 |
| 13 | SNOW | 적설(cm) | 설명변수 | 연속형 |
| 14 | CLOUD | 전운량(10분위) | 설명변수 | 연속형 |

# 데이터 현황

결측치:
- 366개 데이터 중 S02 C0, Pm10, 03, N-2에 null데이터가 존재하는 것으로 보임.
- C0데이터의 경우 null데이터가 많아 보이므로 평균치를 넣어줄 계획
- PM10의 데이터의 결측치를 수정하는 과정에서 CO 칼럼의 null데이터 이외에 칼럼의 결측치가 같이 지워짐

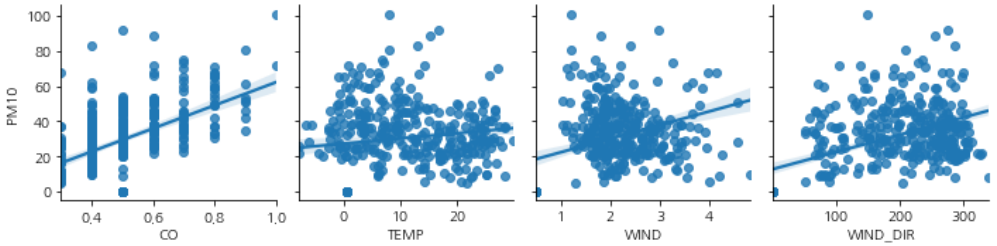
이상치:
- ATMPRESS의 경우 단위가 너무 커서 박스 플롯에서 높게 나타나는 것 같아보인다.
- 데이터별로 단위가 차이가 나서 박스 플롯 비교가 힘들어 보인다. 따라서 이상치 확인을 위해 scale을 변환시켜서 확인해 봄

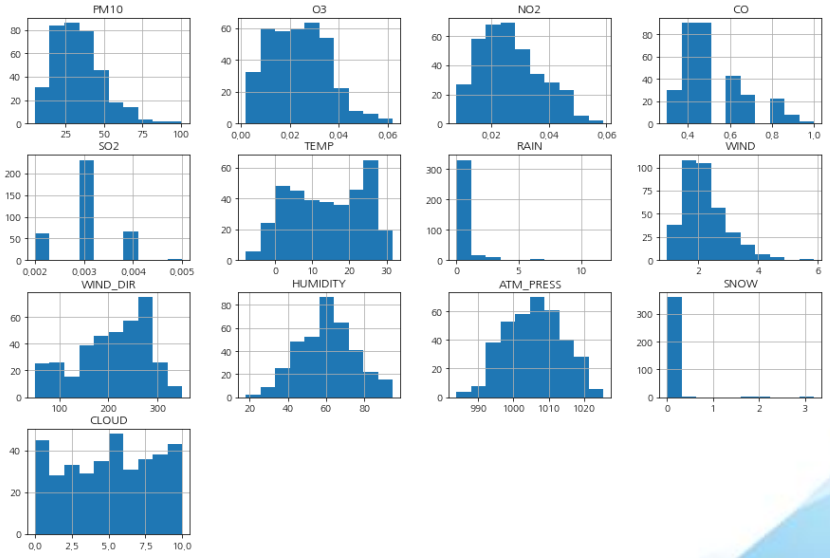
Scaling 이후
- 다른 데이터보다는 PM10과 RAIN SNOW데이터에서 어느정도 이상치가 있는 것으로 보여진다.
- 하지만 PM10의 경우 unique로 데이터 확인해본 결과 미세먼지의 농도가 나쁜 날일 경우 충분히 나올 수 있는 데이터라고 생각이 되어 놔두도록 한다. 또한 눈이나 비같은 경우 최대치의 경우 눈은 3cm이고, 비는 11mm이다. 이런 측청 치는 홍수나, 폭설 등 기존보다 많이 올 수 있는 경우가 존재할 수 있다고 생각해 남겨두었다.

# 탐색적 분석

변환
- 데이터의 변환은 크게 필요가 없어 보인다.
- 일부 데이터의 단위가 다른 경우는 선형회귀 모델 시 스케일을 진행해 변수 비교를 실시하도록 하겠다.

분포
- 데이터의 분포를 보면 PM10, O3, NO2가 비슷한 분포를 보이는 것처럼 보인다.
- 또한 pairplot을 통해 데이터와 선형성을 보이는 일부 변수를 뽑아서 다시 스캐터플롯을 그려본 CO나 TEMP에서 약간의 선형성을 보이는 것을 확인했다.
- HUMIDITY와 ATM_PRESS도 비슷한 분포를 보이고 있다.



# 탐색적 분석

상관관계
- 상관계수를 부여하니 CO, SO2, NO2 순으로 높은 양의 상관관계를 보였고,
- 다음으로 TEMP가 다음으로 음의 상관관계를 보였다. 5개의 변수를 선택해보면 ATM_PRESS가 상대적으로 높은 상관관계를 보였다.

| | PM10 |
|---|---|
| PM10 | 1.000 |
| O3 | -0.052 |
| NO2 | 0.396 |
| CO | 0.588 |
| SO2 | 0.429 |
| TEMP | -0.310 |
| RAIN | -0.121 |
| WIND | -0.100 |
| WIND_DIR | 0.020 |
| HUMIDITY | -0.150 |
| ATM_PRESS | 0.253 |
| SNOW | -0.020 |
| CLOUD | -0.172 |

가설 검정
- 초기 가설로 미세먼지를 발생시킬 수 있는 요인들이 있다는 가설을 세웠었다. 그래프의 분포를 보면 PM10과 CO는 상관관계가 있는 것처럼 보였고, O3, NO2 또한 PM10과 비슷한 분포를 보이고 있다.

- 상관계수 분석에서 미세먼지를 발생시키는 CO와 SO2, NO2에서 PM_10과 상관관계가 높게 나타났다. 또한 날씨와 관련된 TEMP나 ATM_PRESS도 상관계수가 높게 나왔다.
- 또한 그래프 분석에서도 CO, WIND, WIND_DIR에서 관계가 크지는 않으나 있어 보였다.

- 따라서 '미세먼지를 발생시키는 요인' 으로 화석 연료 사용으로 나오는 유해물질 간에 관계가 있을 것이라는 가정에 힘을 실을 수 있을 것이라고 본다.

- 모델링 기법을 진행하면서 해당 칼럼들에 관심을 가지고 수치를 확인해야겠다.

# 모델링 & 요약

※ 모델링
다중회귀분석, 의사결정나무, 랜덤포레스트, 그레디언트 부스팅을 사용할 계획이다.

1) 다중 회귀분석
   - 후진 제거법으로 변수 선정
   - 초기 7개의 변수를 설정 하였을 때, 설명력 68.3%였으나 NO2와 SNOW의 p-value가 높아서 개수를 6개로 줄여 봄
   - 6개로 줄였을 때 설명력이 65.3%로 올랐음. NO2의 경우 화학적 반응을 일으킬 수 있는 변수이며, 이전 탐색적 분석에서 높은 수치를 가지고 있었기에 남겨두었다.

변수 7개의 OLS

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | PM10 | R-squared: | 0.689 |
| Model: | OLS | Adj. R-squared: | 0.683 |
| Method: | Least Squares | F-statistic: | 113.2 |
| Date: | Thu, 04 Mar 2021 | Prob (F-statistic): | 1.25e-86 |
| Time: | 00:00:18 | Log-Likelihood: | -1385.8 |
| No. Observations: | 365 | AIC: | 2788. |
| Df Residuals: | 357 | BIC: | 2819. |
| Df Model: | 7 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | -33.0019 | 5.661 | -5.830 | 0.000 | -44.135 | -21.869 |
| O3 | 519.0347 | 71.406 | 7.269 | 0.000 | 378.604 | 659.465 |
| NO2 | 90.4226 | 126.053 | 0.717 | 0.474 | -157.478 | 338.323 |
| CO | 90.2500 | 6.856 | 13.163 | 0.000 | 76.766 | 103.734 |
| SO2 | -634.4147 | 160.673 | -3.948 | 0.000 | -950.400 | -318.430 |
| RAIN | -1.8409 | 0.634 | -2.904 | 0.004 | -3.088 | -0.594 |
| WIND | 4.3349 | 1.207 | 3.593 | 0.000 | 1.962 | 6.708 |
| SNOW | -0.7749 | 2.615 | -0.296 | 0.767 | -5.917 | 4.367 |

| | | | |
|---|---|---|---|
| Omnibus: | 114.787 | Durbin-Watson: | 1.229 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 433.362 |
| Skew: | 1.346 | Prob(JB): | 7.88e-95 |
| Kurtosis: | 7.609 | Cond. No. | 894. |

변수 6개의 OLS

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | PM10 | R-squared: | 0.663 |
| Model: | OLS | Adj. R-squared: | 0.657 |
| Method: | Least Squares | F-statistic: | 117.1 |
| Date: | Thu, 04 Mar 2021 | Prob (F-statistic): | 3.17e-81 |
| Time: | 00:01:17 | Log-Likelihood: | -1396.5 |
| No. Observations: | 364 | AIC: | 2807. |
| Df Residuals: | 357 | BIC: | 2834. |
| Df Model: | 6 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 30.4235 | 0.594 | 51.242 | 0.000 | 29.256 | 31.591 |
| O3 | 78.6408 | 12.669 | 6.207 | 0.000 | 53.726 | 103.556 |
| NO2 | 6.3900 | 21.963 | 0.291 | 0.771 | -36.804 | 49.584 |
| CO | 12.3341 | 1.002 | 12.307 | 0.000 | 10.363 | 14.305 |
| SO2 | -94.0355 | 29.416 | -3.197 | 0.002 | -151.886 | -36.185 |
| RAIN | -1.2677 | 0.627 | -2.023 | 0.044 | -2.500 | -0.035 |
| WIND | 3.8015 | 1.129 | 3.368 | 0.001 | 1.582 | 6.021 |

| | | | |
|---|---|---|---|
| Omnibus: | 87.702 | Durbin-Watson: | 1.297 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 289.945 |
| Skew: | 1.056 | Prob(JB): | 1.09e-63 |
| Kurtosis: | 6.829 | Cond. No. | 118. |

# 모델링 & 요약

## 2) 의사결정나무

- 트레이닝셋과 데이터 셋 7:3비율로 나눠서 테스트 진행
- 하이퍼 파라미터는 GridSearch 활용하려고 했으나, 그리드 서치는 성능이 잘 안나오므로 하이퍼 파라미터 튜닝을 진행하기로 계획

- 직접 하이퍼 파라미터 튜닝을 한 모델과 그리드 서치를 진행한 모델의 test score의 차이는 0.133이 나는 것을 확인함.

- 해당 모델로 변수 중요도에 대해 확인한 결과 ATM-PRESS, CO에서 강한 중요도를 보여주고, WIND_DIR, O3, TEMP는 순으로 다른 변수보다는 낮은 중요도를 보여주는 것을 확인함.

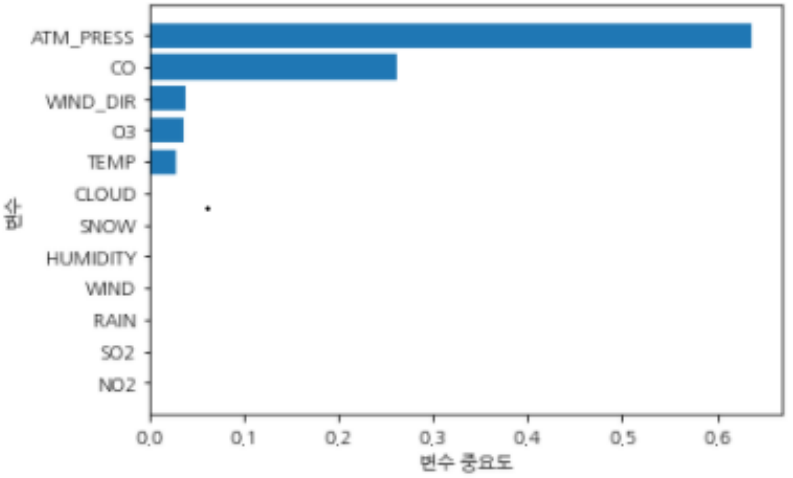
```
### GridSearchScore
dt_model = DecisionTreeRegressor(min_samples_leaf= 3, min_samples_split=8,
                                 max_depth=7,random_state=1234)
dt_result = dt_model.fit(df_train_x, df_train_y)
# Train 데이터 설명력
print("GS Score on training set: {:.3f}".format(dt_model.score(df_train_x, df_train_y)))
#test데이터 설명력
print("GS Score on test set: {:.3f}".format(dt_model.score(df_test_x, df_test_y)))
```

```
GS Score on training set: 0.875
GS Score on test set: 0.440
```

```
## 해당 모델 사용

### 직접 구한 매개변수
dt_model = DecisionTreeRegressor(min_samples_leaf= 18, min_samples_split=8,
                                 max_depth=5, random_state=1234)
dt_result = dt_model.fit(df_train_x, df_train_y)
# Train 데이터 설명력
print("Score on training set: {:.3f}".format(dt_model.score(df_train_x, df_train_y)))
#test데이터 설명력
print("Score on test set: {:.3f}".format(dt_model.score(df_test_x, df_test_y)))
```

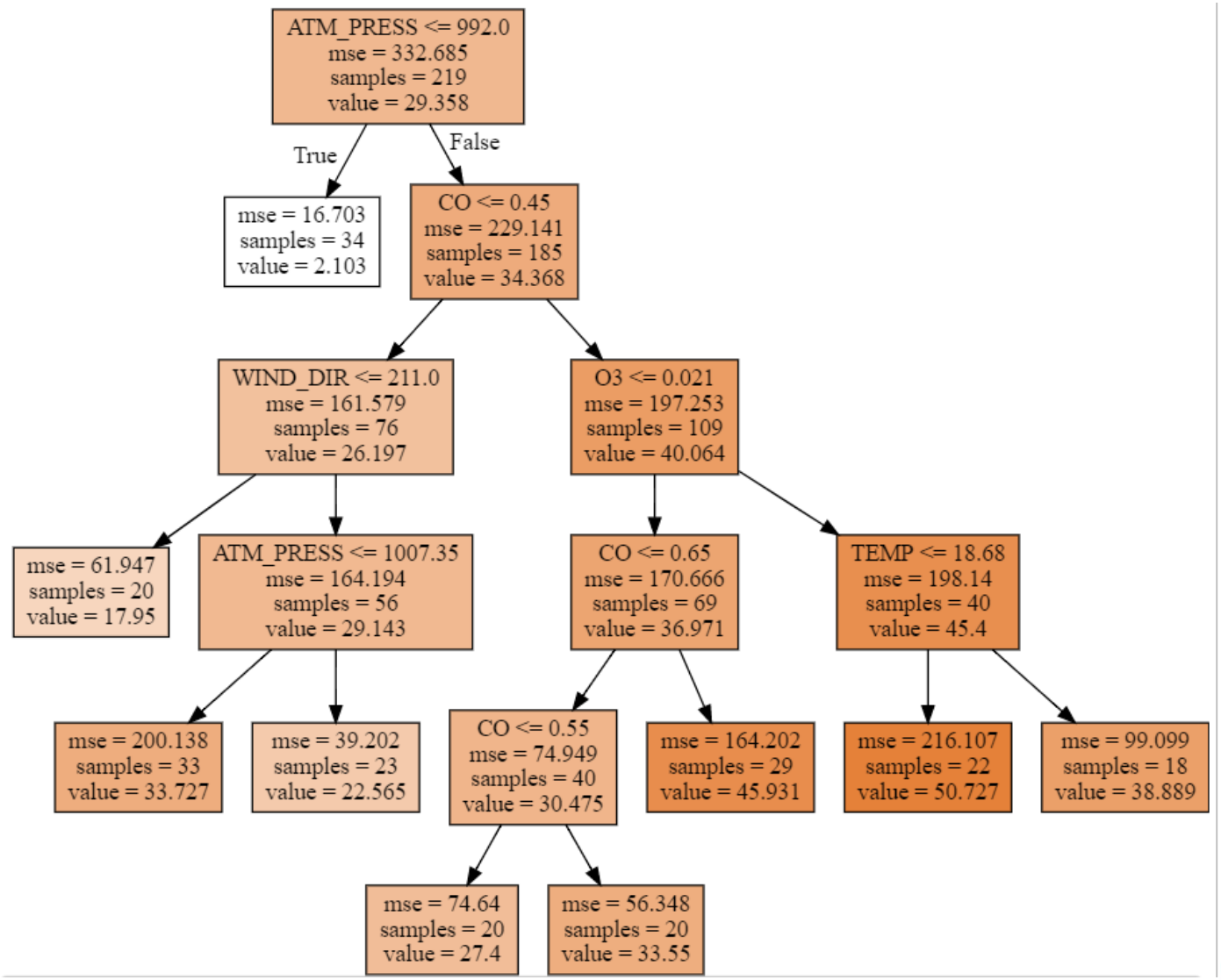
```
Score on training set: 0.681
Score on test set: 0.573
```



# 모델링 & 요약

## 2) 의사결정나무

- 트리 시각화



# 모델링 & 요약

## 3) 랜덤포레스트

- 트레이닝셋과 데이터 셋 6:4비율로 나눠서 테스트 진행
- 하이퍼 파라미터 튜닝을 진행 ( n_estimators = 40, min_samples_leaf = 2, min_samples_split=6, max_depth=5 선정)

- 하이퍼 파라미터 튜닝 전에는 트레이닝 set에 과적합이 되어 있었으나, 튜닝 이후 테스트 데이터의 성능이 0.013만큼 줄긴했으나 그보다 트레이닝 데이터의 과적합이 많이 줄게 되었다.

- 해당 모델의 중요도를 구한 결과 변수의 중요도는 ATM_PRESS, CO, O3, WIND_DIR, WIND 순으로 나타났다.

```
rf_uncustomized = RFR(random_state=1234)
rf_uncustomized.fit(df_train_x, df_train_y)

# Train 데이터 설명력
print("Score on training set: {:.3f}".format(rf_uncustomized.score(df_train_x, df_train_y)))

#test데이터 설명력
print("Score on test set: {:.3f}".format(rf_uncustomized.score(df_test_x, df_test_y)))
```

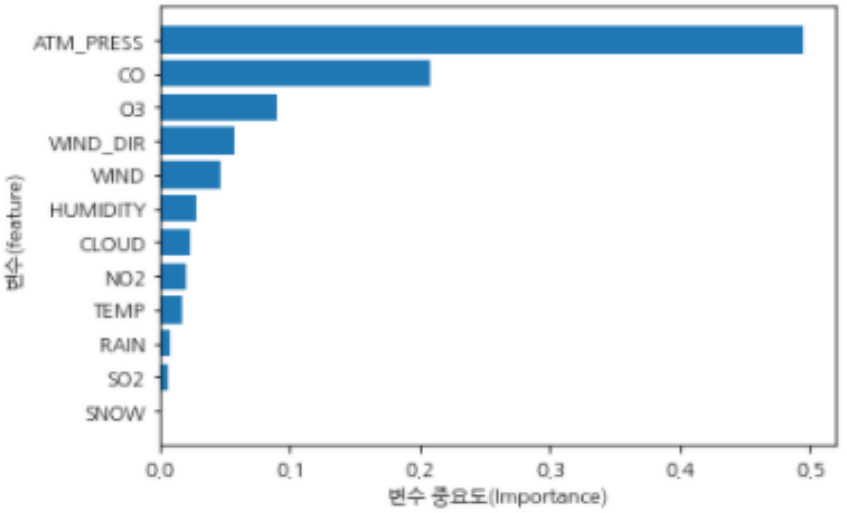
```
Score on training set: 0.944
Score on test set: 0.640
```

```
rf_final = RFR(random_state=1234, n_estimators = 40, min_samples_leaf = 2,
               min_samples_split=6, max_depth=5)
rf_final.fit(df_train_x, df_train_y)

# Train 데이터 설명력
print("Score on training set: {:.3f}".format(rf_final.score(df_train_x, df_train_y)))

#test데이터 설명력
print("Score on test set: {:.3f}".format(rf_final.score(df_test_x, df_test_y)))
```

```
Score on training set: 0.796
Score on test set: 0.627
```



## 4) 그레디언트 부스팅

- 트레이닝셋과 데이터 셋 6:4비율로 나눠서 테스트 진행
- 하이퍼 파라미터 튜닝을 진행 (n_estimators = 60, min_samples_leaf = 2,
- min_samples_split=16, max_depth=3, learning_rate=0.1선정)

- 하이퍼 파라미터 튜닝 이후 초기값보다 많이 나아진 모습을 보임. test성능이 0.02줄어든 반면, training data에 대한 과적합이 크게 줄었음을 확인할 수 있다.

- 해당 모델의 중요도를 구한 결과 변수의 중요도는 ATM_PRESS, CO, O3, WIND_DIR, TEMP 순으로 나타났다.

```
gb_uncustomized = GBR(random_state=1234)
gb_uncustomized.fit(df_train_x, df_train_y)

# Train 데이터 설명력
print("Score on training set: {:.3f}".format(gb_uncustomized.score(df_train_x, df_train_y)))

#test데이터 설명력
print("Score on test set: {:.3f}".format(gb_uncustomized.score(df_test_x, df_test_y)))
```

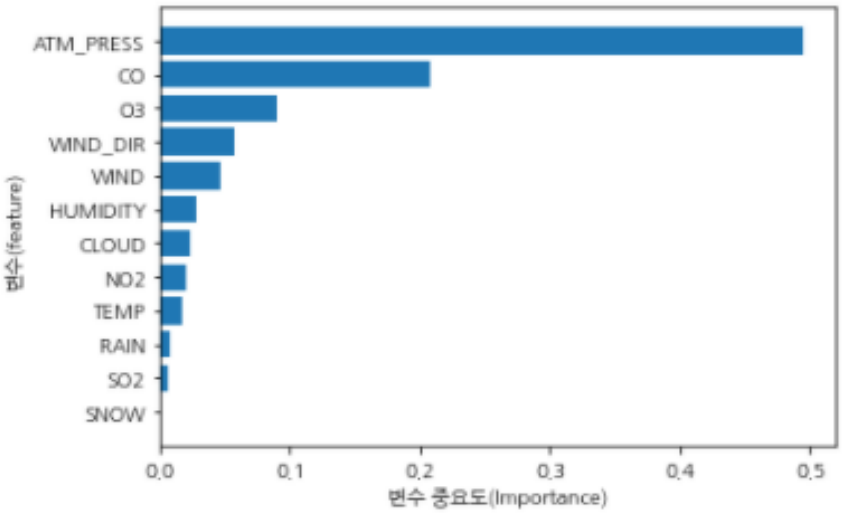
```
Score on training set: 0.967
Score on test set: 0.695
```

```
gb_final = GBR(random_state=1234, n_estimators = 60, min_samples_leaf = 2,
               min_samples_split=16, max_depth=3, learning_rate=0.1)
gb_final.fit(df_train_x, df_train_y)

# Train 데이터 설명력
print("Score on training set: {:.3f}".format(gb_final.score(df_train_x, df_train_y)))

#test데이터 설명력
print("Score on test set: {:.3f}".format(gb_final.score(df_test_x, df_test_y)))
```

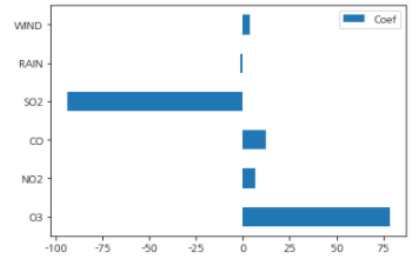
```
Score on training set: 0.894
Score on test set: 0.693
```



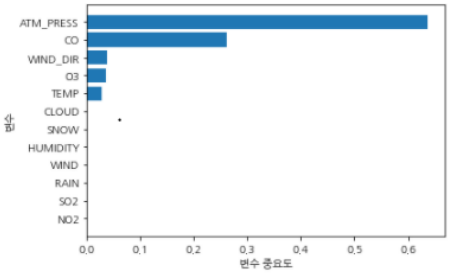
# 모델링 & 요약

※ 모델별 주요 변수 정리

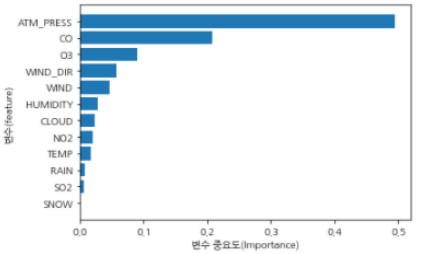
- 선형 회귀에서 scaling 후 구한 변수의 중요도는 SO2, O3, CO, NO2, WIND, RAIN 순서대로 나타났다.



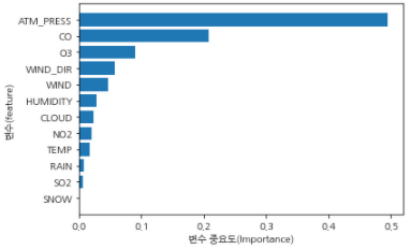

- DT 모델에서는 ATM-PRESS, CO에서 강한 중요도가 나타났고, WIND_DIR, O3, TEMP는 순으로 낮은 중요도가 나타났다.




- RF 모델의 중요도를 구한 결과 변수의 중요도는 ATM_PRESS, CO, O3, WIND_DIR, WIND 순으로 나타났다. HUMIDITY 변수가 나타났다.




- GB 에서는 변수의 중요도는 ATM_PRESS, CO, O3, WIND_DIR, TEMP 순으로 나타났다. ATM_PRESS, CO, O3는 특히 강한 중요도를 보여준다. RF와 마찬가지로 6번째 순위에 HUMIDITY(습도)가 등장했다.

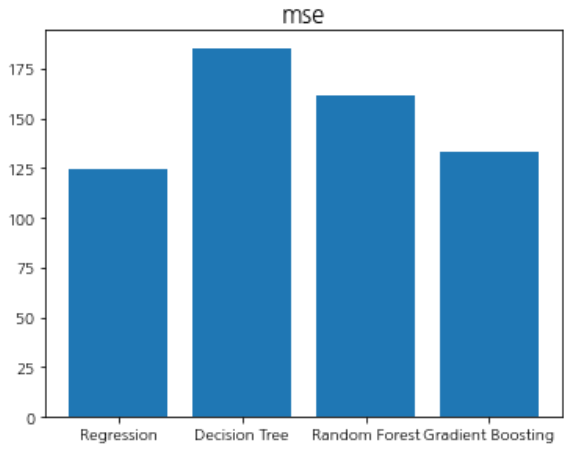


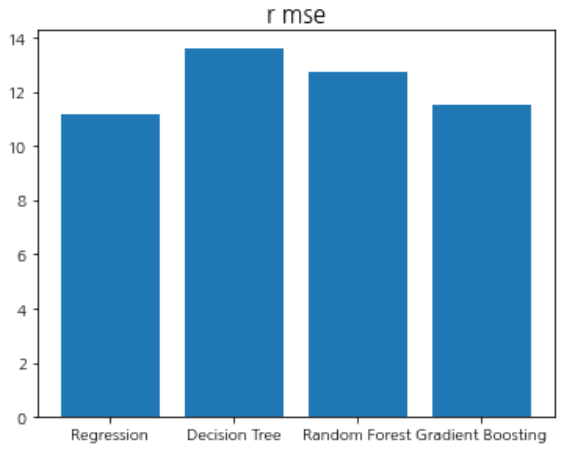
# 모델링 & 요약

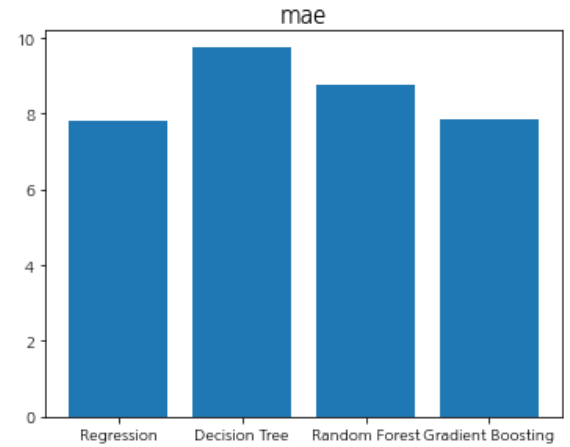
※ 모델별 평가 시각화 그래프

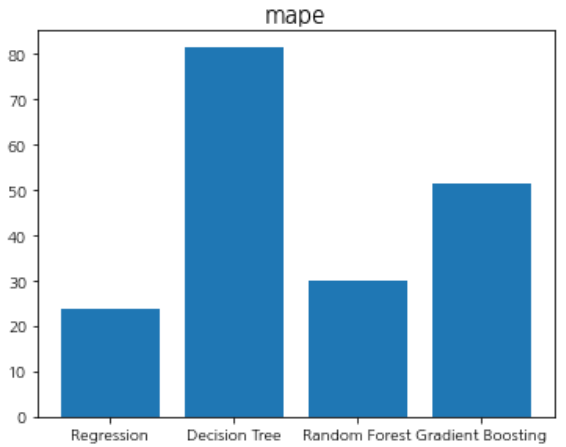
- Regression 모델이 모든 평가 지표에서 가장 점수가 낮게 나왔기에 성능이 가장 좋다고 할 수 있는 반면에 Decision Tree모델의 성능이 가장 좋지 않다고 할 수 있다.

- 평가지표 별 모델의 정확도 "추세"는 거의 동일하게 나타난다. 다만 mape에서 GB모델이 RF보다 안 좋은 점수를 보여준다.



# 결론 및 대안제시

결론

- 초기 가설로 미세먼지를 발생시키는 요인으로 화석연료 사용으로 인한 미세먼지가 증가했을 것(1차적 발생)이라고 보았고, 미세먼지를 야기시키는 요인으로 기상현상이 영향을 끼칠 것(2차적 발생)으로 보았다.

- 탐색적 분석과 모델링 기법을 통해 분석한 결과 실제로 화석연료의 사용으로 인해 증가하는 변수들인 SO2(아황산가스), CO(이산화질소), NO2(일산화탄소) 변수가 나타났었다. 특히 이산화질소는 다른 여러 모델에서도 등장하며, 미세먼지 농도와 연관이 깊어 보였다.

- 모델링 기법을 진해하면서 주요한 변수로 ATM_PRESS(기압)와 O3(오존 농도)가 꾸준히  미세먼지에 영향을 미치는 변수로 나왔었다. 또한 WIND_DIR(풍향) 변수도 자주 나오며 PM10변수에 영향을 주는 것을 확인했다.

- 해당 변수들은 기상조건에 해당하는 변수들로서 오존 농도와 기압 그리고 풍향 등에 의해서 미세먼지가 영향을 받는다고 할 수 있다.

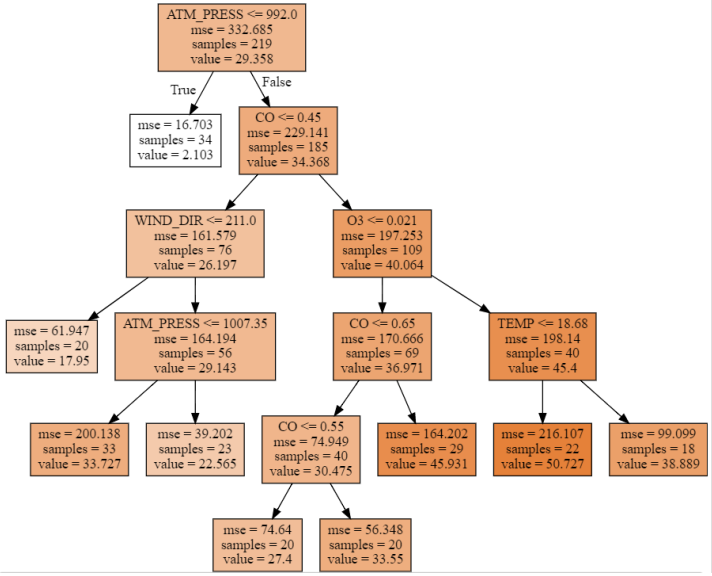
# 결론 및 대안제시

## 임계치 제시

- 4가지 모델링 기법을 대상으로 진행한 정확도 검증에서 의사결정나무가 가장 좋은 성능을 보였다. 따라서 의사결정나무를 통해 임계치를 생각해보도록 하겠다.

- ATMPRESS <=992.0 에서는 mse값이 16으로 떨어졌으며, value(PM_10) 또한 2점대를 기록했다.
- ATMPRESS >=992.0 에서는 CO로 구분한다.
- CO가 0.45보다 클 경우 O3가 0.021보다 큰지 작은지에 따라 구분하며, CO가 0.45보다 작을 때는 WIND_DIR이 211보다 큰지 작은지에 따라 구분되었다.

- 따라서 ATMPRESS<=992와 CO <=0.45, O3 <=0.021, WIND_DIR <=211.0 수치 등을 통해 미세먼지 현황을 예측해볼 수 있을 것으로 기대된다.



# 결론 및 대안제시

## 대안제시

- 1차적 발생 요인을 대상으로 미세먼지를 줄이기 위해서는 우선 PM_10과 가장 관계가 깊은 이산화질소를 낮출 수 있는 방법을 생각해봐야 한다. 특히 이산화질소 등의 물질은 화학 연료를 사용할 때 발생하기에 이를 줄일 수 있는 방법에 대해 꾸준한 발전과 고찰이 필요해 보인다.

- 2차적 발생 요인을 대상으로 지표면 부근에서 배출된 대기오염 물질은 '기온 역전층'을 벗어나지 못하면서 축척이 되어 고농도 오염 사례를 유발[2] 하게 된다는 사례처럼 기압을 풍향과 연관시켜 기상변화에 따라 미세먼지 현황을 미리 예측할 수 있을 것으로 생각된다.

- 기상 현상과 임계치를 통한 현황을 분석해서 미세먼지를 저감시킬 수 있을만한 방안을 미리 생각하고 '기온 역전층'이 발생하게 될 시점을 예측할 수 있다면, 해당 기간에는 미세먼지가 머물지 않고 날려보낼 수 있을 만한 방법을 생각해볼 수 있다.

2) 박세환(2016), 미세먼지의 유해성과 기후변화의 상관관계 분석보고서

통찰, 아이디어, 애로사항

- 실제 논문 등을 통해 가설을 설정하고 자료를 봤을 때는 막연한 느낌이 있었다. 하지만 직접 데이터를 모델에 넣고 돌려보면서 결과가 나오는 것을 보면서 일부 가설을 조금 더 검증할 수 있었던 것 같다. 따라서 초기에 가설을 세우기 어려울 지라도 어느 정도의 가설을 세워놓고 점차 보충하고 수정해가는 방법을 사용해도 괜찮을 거 같다는 생각을 했다.

- 분석을 조금 더 다양하게 하고, 여러 시도를 해보려고 해도 시간이 너무 부족해서 조금 더 설명력이 있는 결과를 도출하지 못한 것 같아서 아쉬웠다. 또한 변수들의 설명이 별도로 있었지만, 경험이 부족해서 일부 변수들의 경우 해당 데이터가 현실 세계의 어떤 데이터를 나타내는지 이해하기 어려웠던 부분이 있었던 것 같다.

# 핵심인자 정리 템플릿(참조용_파일첨부)



| 변수 번호 | 변수 | 변수 설명 | 변수 역할 | 변수 형태 | 분석 제외 사유 | 탐색적 기법 | | | 모델링 기법 | | | | | 총점 | 선정 (사유) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 그래프 | 검정 | 상관분석 | 회귀분석 | DT | RF | GB | 사례연구 | | |
| 1 | MeasDate | 측정일자 | 제외 | 연속형 | 회귀분석을 진행하는데 오류가 생김 | | | | | | | | | | |
| 2 | PM10 | 미세먼지10㎍/㎥ | 목표변수 | 연속형 | | | | | | | | | | | |
| 3 | O3 | 오존 농도 | 설명변수 | 연속형 | | | | | 2 | 4 | 3 | 3 | 논문 참조 | 3 | 화학적 반응을 일으키는 매개체로서 오존에 의해서 미세먼지를 만드는 다양한 물질이 생성될 가능성이 있다. |
| 4 | NO2 | 이산화질소 농도 | 설명변수 | 연속형 | | | 4 | 3 | 4 | | | | 논문 참조 | | |
| 5 | CO | 일산화탄소 농도 | 설명변수 | 연속형 | | | 3 | 1 | 3 | 2 | 2 | 2 | 논문 참조 | 2 | 탐색적 기법은 물론 모델링 기법에서도 자주 등장하는 변수였으며, 높은 영향을 끼쳤다. |
| 6 | SO2 | 아황산가스 농도 | 설명변수 | 연속형 | | 1 | 1 | 2 | 1 | | | | 논문 참조 | 4 | 아황산 가스는 화석연료의 사용으로 발생하는 물질로 미세먼지를 발생시켰다고 할 수 있으며, 탐색적 기법과 회귀분석에서 자주 나타났다. |
| 7 | TEMP | 기온(°C) | 설명변수 | 연속형 | | | | 4 | | 5 | | 5 | | | |
| 8 | RAIN | 강수량(mm) | 설명변수 | 연속형 | | | | | 6 | | | | 논문 참조 | | |
| 9 | WIND | 풍속(m/s) | 설명변수 | 연속형 | | 2 | 3 | | 5 | | 5 | | | | |
| 10 | WIND_DIR | 풍향(16방위) | 설명변수 | 연속형 | | 3 | 6 | | | 3 | 4 | 4 | 논문 참조 | 5 | 풍향은 미세먼지를 일으키는데 대기이동 측면에서 영향을 끼칠 수 있다고 봤다. |
| 11 | HUMIDITY | 습도(%) | 설명변수 | 연속형 | | | | | | | 6 | 6 | | | |
| 12 | ATM_PRESS | 현지기압(hPa) | 설명변수 | 연속형 | | | 5 | 5 | | 1 | 1 | 1 | 논문 참조 | 1 | 분석 중에 특히 모델링 기법 중 트리 계열의 모델에서 매우 높은 점수를 보였다. 또한 기압은 풍향과 같은 맥락에서 영향을 끼칠 것으로 봤다. |
| 13 | SNOW | 적설(cm) | 설명변수 | 연속형 | 선형회귀모델에서 유의미해 보였으나 p-value가 너무 높아 제외하였음 | | | | | | | | | | |
| 14 | CLOUD | 전운량(10분위) | 설명변수 | 연속형 | 여러 기법을 사용해봤지만 주요 변수로 등장하지 않음 | | | | | | | | | | |
| | | | | | | | | | | | | | 논문 참조 | | |

사례 연구 및 점수 계산만 진행