서울의 미세먼지

성균관 대학교
김한비 2021712203

목차

01 연구 소개

02 데이터 수집 및 전처리

03 데이터 모델링

04 결론 및 토의

연구 소개

- 미세먼지 소개.
- 서울의 미세먼지 현황소개.
- 연구 개요 및 목적.

미세먼지 소개

미세먼지 소개:

먼지: 대기 중에 떠다니거나 흩날려 내려오는 입자상 물질.

미세먼지(PM10): 입자 지름이 10 μm 보다 작은 먼지.

초미세먼지(PM2.5): 입자 지름이 2.5 μm 보다 작은 먼지.

미세먼지의 위험성:

국제암연구소(IARC)에 의해 2013년 1군 발암물질 지정.

서울의 미세먼지 현황 소개

상대적으로 높았던 도시: 중국 인도 중동국가 소속.

가장 높았던 도시: 중국 청두 405 μg/ m^3

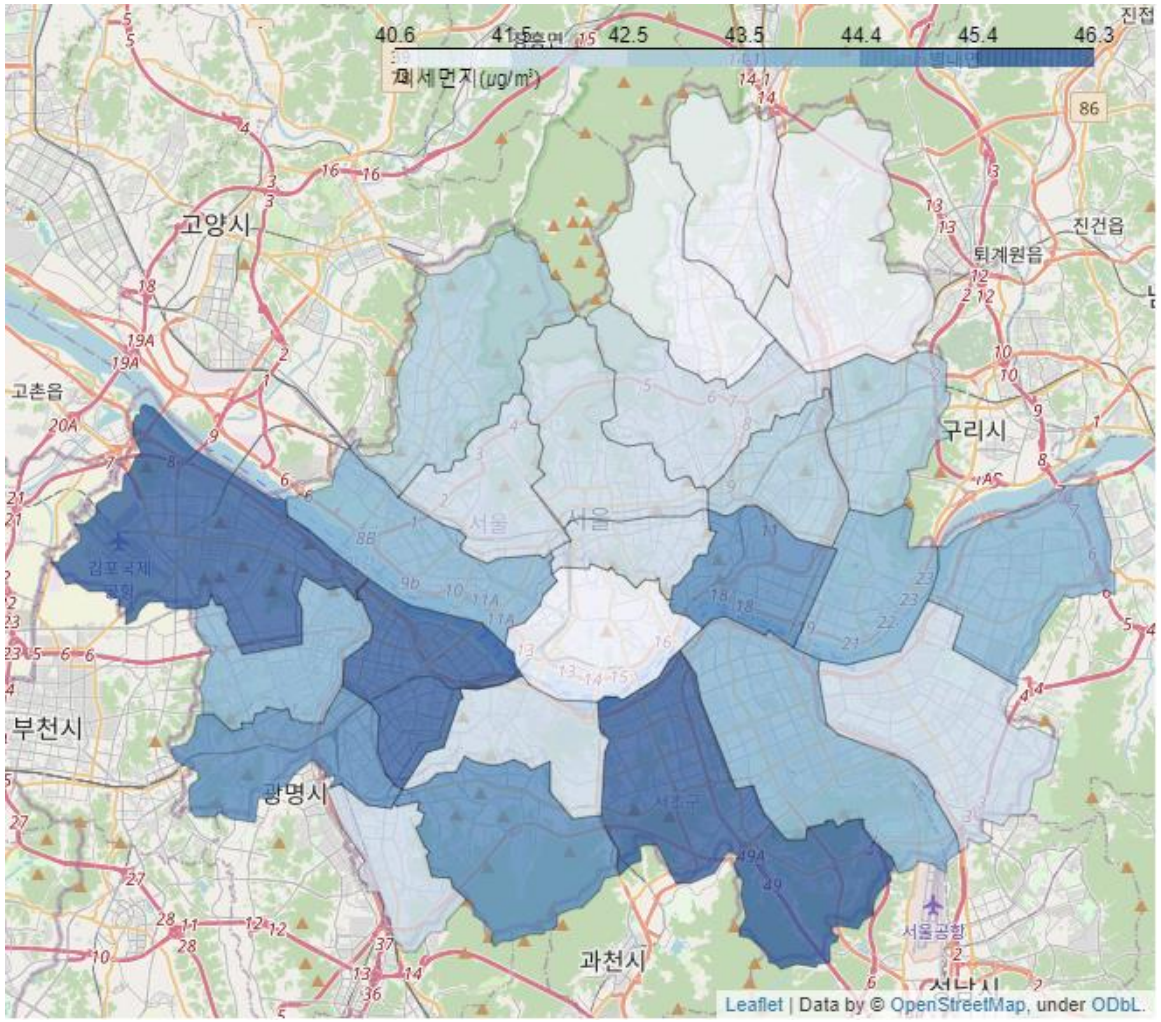
상대적으로 낮았던 도시: 캐나다 호주 뉴질랜드 소속.

가장 낮았던 도시: 캐나다 벤쿠버 12 μg/ m^3

서울(45 μg/ m^3)은 세계 주요도시 169개중 중상위권(69위).

가장 높았던 구: 영등포구 48 μg/ m^3

가장 낮았던 구: 용산구 40 μg/ m^3



연구 개요 및 목적

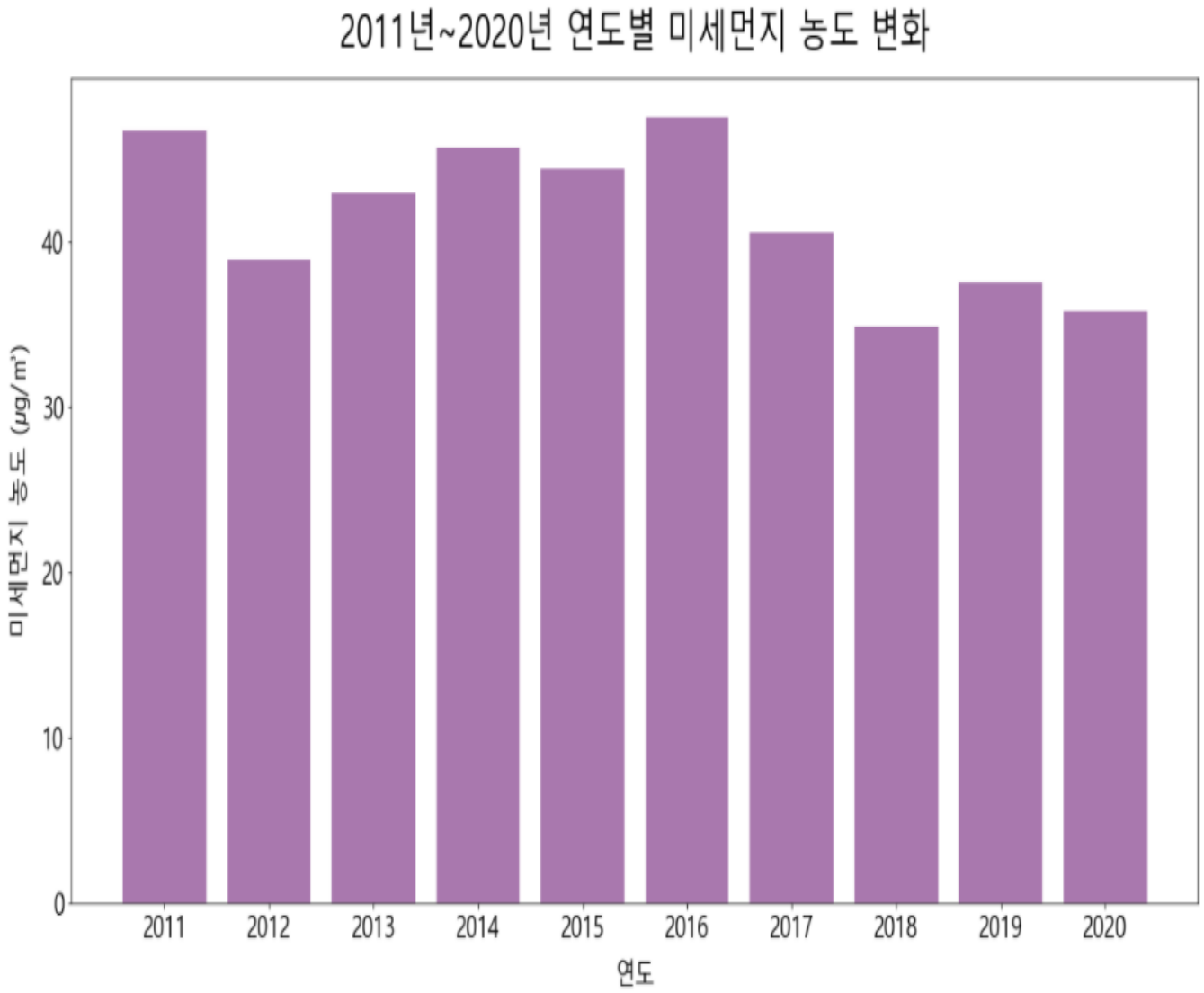
연구 시작 배경:

40 μg/ m^3 후반을 넘기던 2011년 부터 어느정도 감소세를 보이며 현재는 대략 30 μg/ m^3 후반 수준으로 감소하였다. 서울의 미세먼지 농도는 비교적 위험한 수준은 아니지만, 미세먼지의 위험성을 인지하고 이에 따른 대비책이 필요.

연구 목적:

미세먼지에 영향을 끼치는 요인에 대한 분석

미세먼지 농도를 예측하여 실생활 실내외 여가 계획에 사용



데이터 수집 및 전처리 과정

- 데이터 수집 및 전처리.
- 이상치 제거.
- 상관관계분석.
- 최종 데이터 소개.

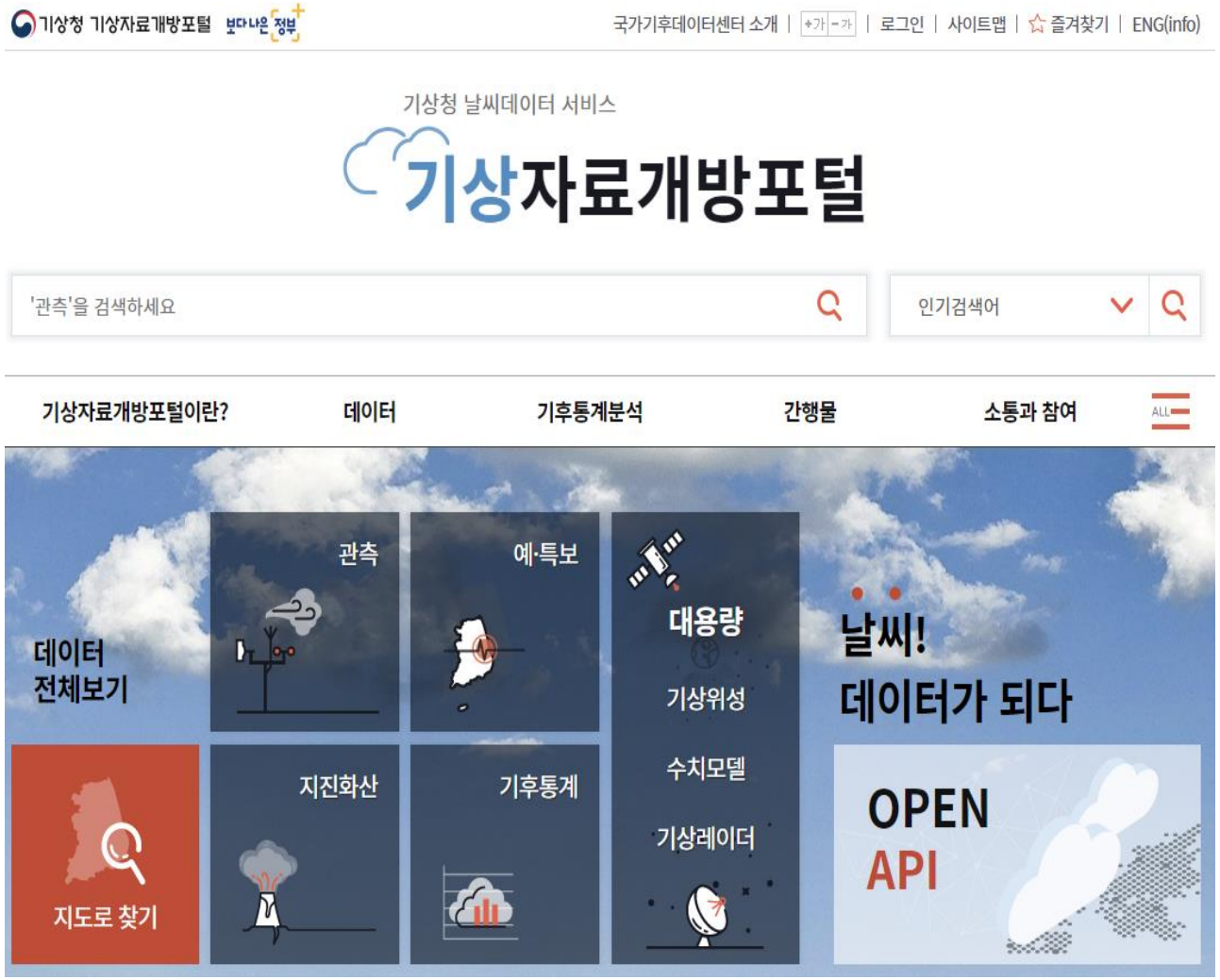
데이터 수집 및 전처리

1	측정일시	측정소명	이산화질소	오존	이산화탄소	아황산가스	미세먼지	초미세먼지
2	20110101	강남구	0.052	0.004	0.7	0.007	66	
3	20110101	강남대로	0.075	0.003	1.4	0.008	70	
4	20110101	강동구	0.036	0.007	0.9	0.007	56	

- 대기오염 정보 데이터 (기상청)
- 측정일시 변경 'YYYY-MM-DD'
- 측정소명에서 '중구' 인 곳만 추출
- 초미세먼지 속성은 NULL 값이 많음으로 제거

1	format: day	hour	value location:60_127 Start : 20110101
2	1	0	-9.6
3	1	100	-7.1
4	1	200	-5.5

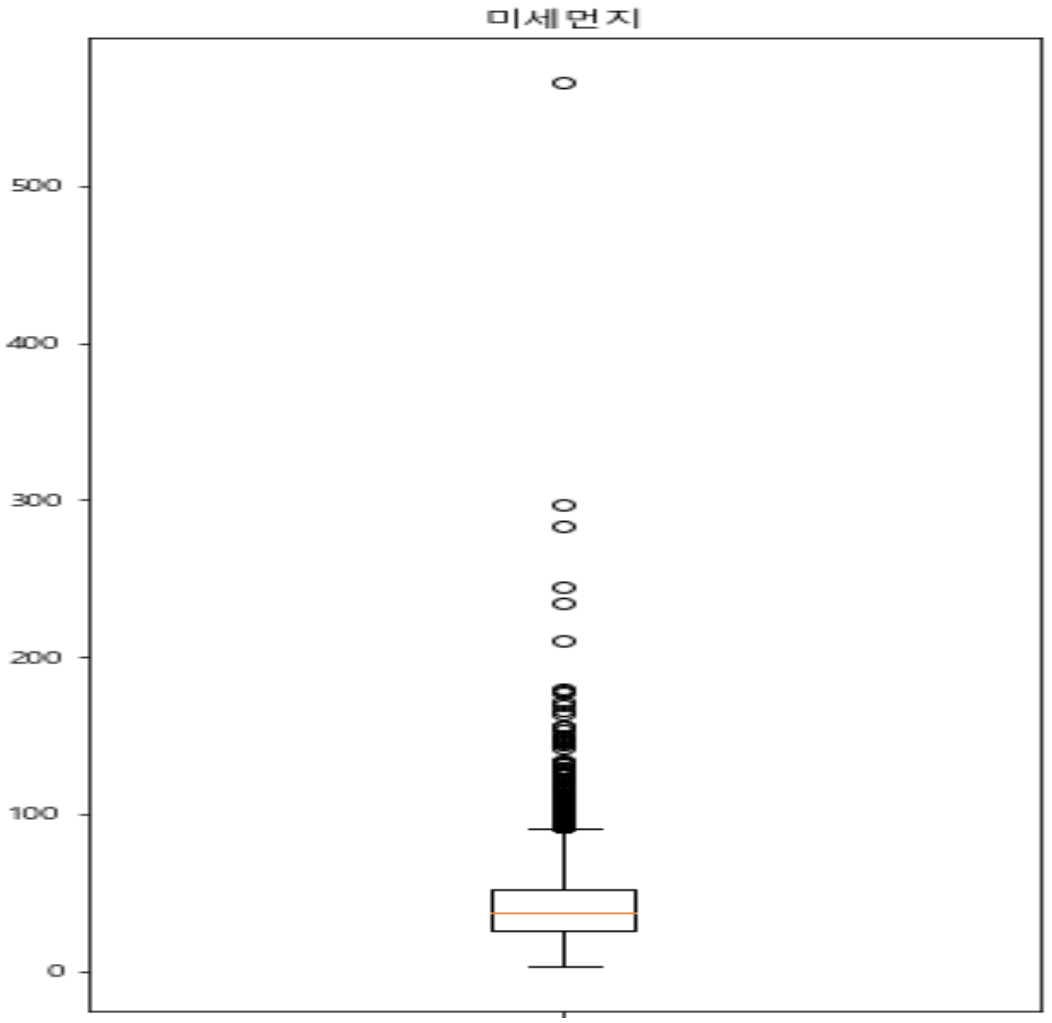
- 교남동 강수 / 강수형태 / 기온 / 습도 / 풍속 / 풍향 / 하늘상태 데이터 (기상청)
- 측정일시 변경 'YYYY-MM-DD'
- 시간 별 데이터를 하루 평균으로 묶어서 계산



출처: https://data.kma.go.kr/cmmn/main.do

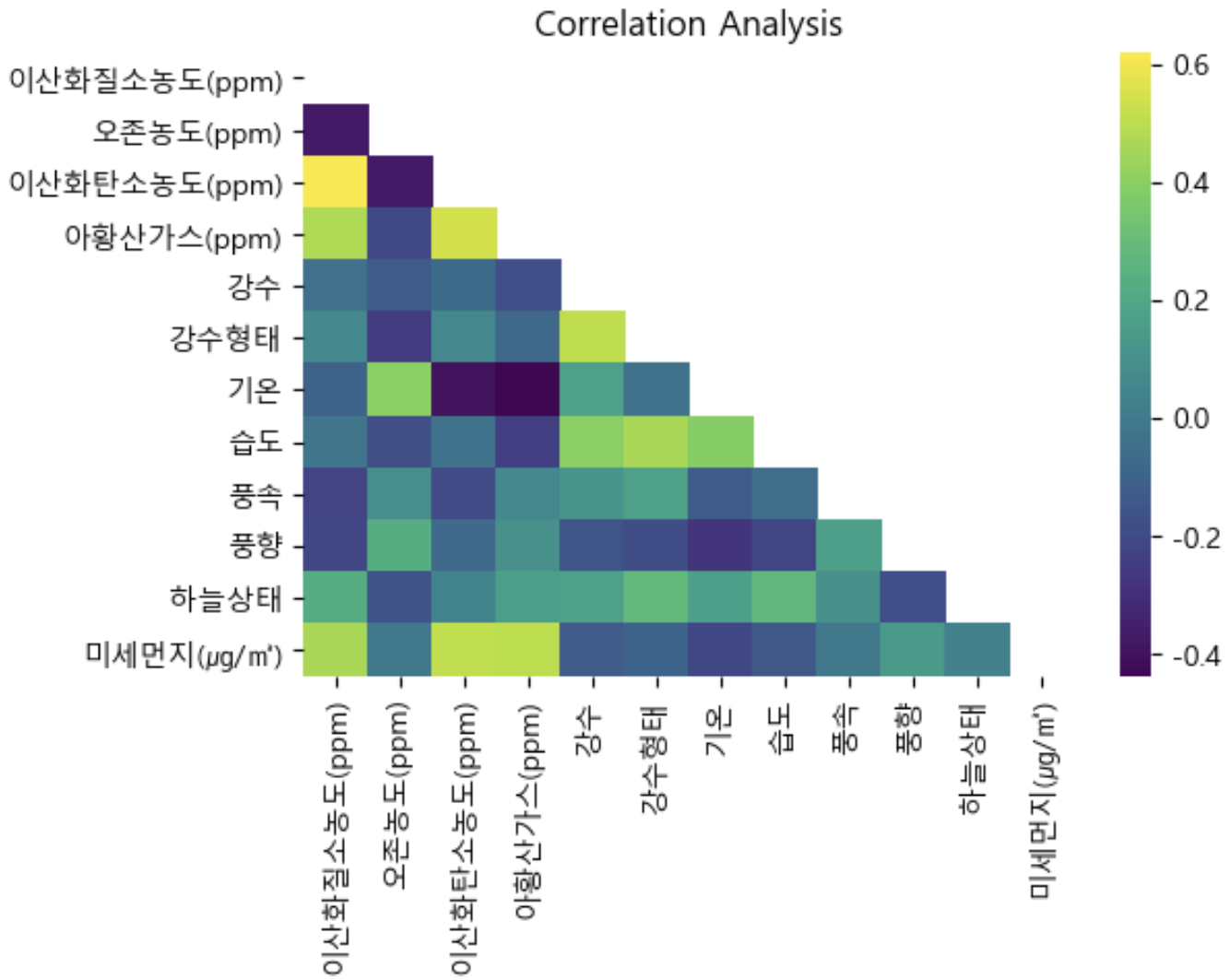
이상치 제거

- 미세먼지 농도 평균 = 45 μg/ m^3
- 미세먼지 농도 표준편차 = 25 μg/ m^3
- 이상치 규정 = 평균 + (표준편차 X 4) = 150 μg/ m^3
- 150 μg/ m^3 이상인 미세먼지 농도 데이터는 총 24개 존재.
- 이는 전체 데이터의 0.007% 해당.



상관 관계 분석

- 미세먼지와 다른 변수 사이에 상관 관계 분석.
- 상관 관계의 절대 값이 1에 가까울수록 상관 관계가 큼
- 양수는 양의 상관관계, 비례관계
- 음수는 음의 상관관계, 반비례관계



상관 관계 분석 결론

- 이산화질소, 이산화탄소, 아황산가스는 미세먼지와의 상관관계계수가 높고, P-value도 0.05보다 작음으로 통계적으로 유의미하다 판단.
- 오존농도, 풍속, 하늘상태의 경우 상관관계계수가 낮음으로 미세먼지와 상관관계에 있다고 말하기 힘들다.
- 기상데이터의 경우, 상대적으로 대기오염 데이터보다 낮은 상관관계계수를 지님.

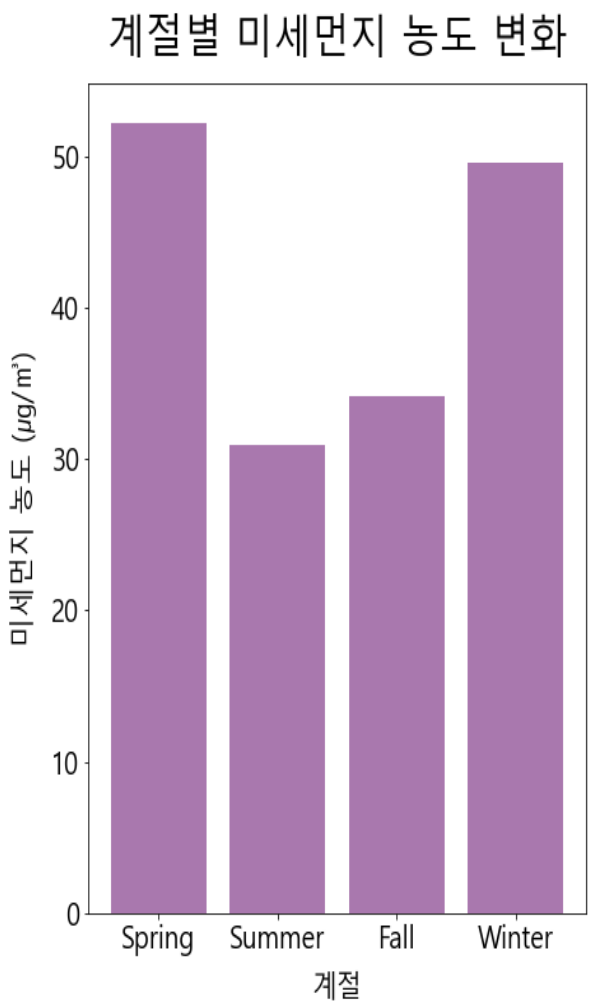
	Factor	Corr Value	P-Value
0	이산화질소농도(ppm)	0.46	0.00
1	오존농도(ppm)	-0.00	0.81
2	이산화탄소농도(ppm)	0.51	0.00
3	아황산가스(ppm)	0.50	0.00
4	강수	-0.13	0.00
5	강수형태	-0.10	0.00
6	기온	-0.21	0.00
7	습도	-0.15	0.00
8	풍속	-0.01	0.75
9	풍향	0.14	0.00
10	하늘상태	0.03	0.04

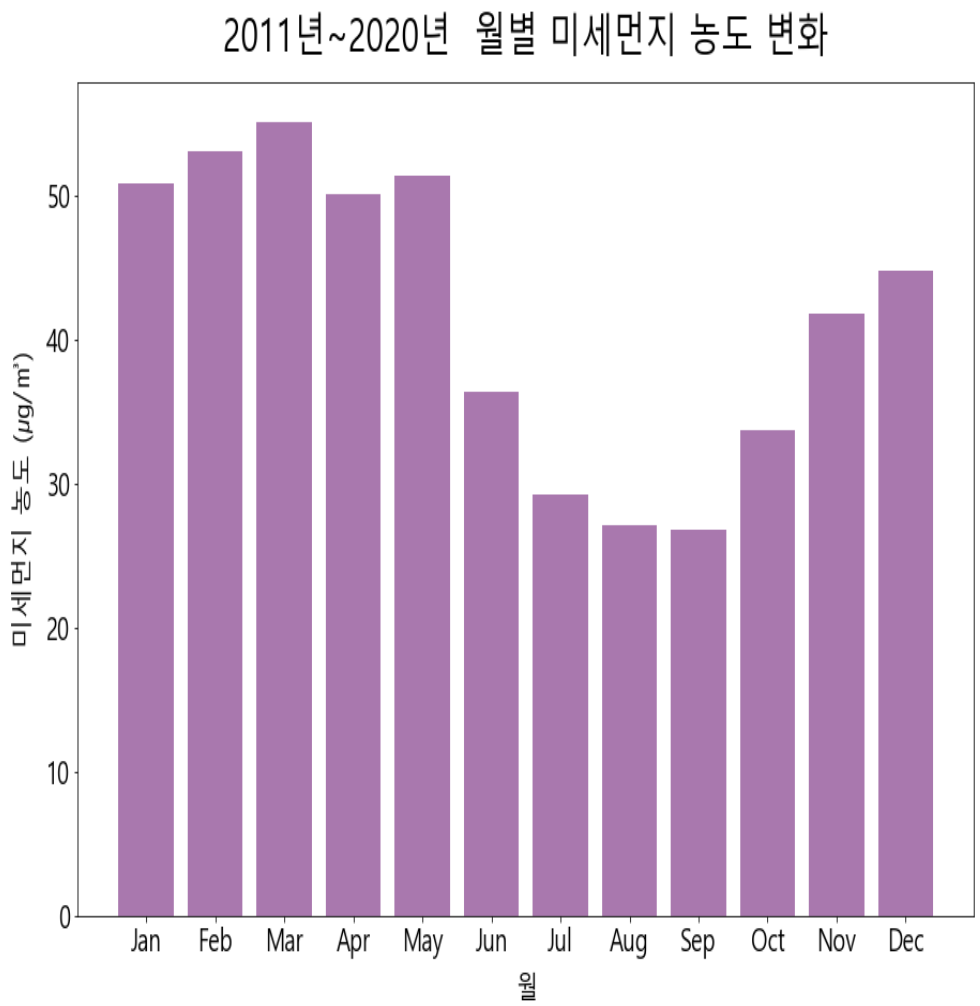
기상 데이터의 효용성 판단

- 계절별 미세먼지 농도 변화를 보면 상대적으로 여름과 가을이 봄과 겨울에
비해 미세먼지 농도가 낮음.

- 계절별 뚜렷한 차이를 보이는 것으로 보아, 계절에 영향을 미치는요인 또한
미세먼지에 영향을 줄 수 있음.

- 당일 기상데이터가 아닌 전날 기상 데이터와의 관계는?





전날의 기상 / 대기오염 상관관계

- 이산화질소, 이산화탄소, 아황산가스 대기오염 데이터의 경우,
오히려 상관관계 계수가 감소.

- 기상데이터의 경우, 습도, 하늘상태를 제외한 나머지 데이터는
조금이나마 상관관계계수가 증가.

전날 미세먼지 농도는 다음날 미세먼지 농도와의
큰 상관관계 계수를 지님.

	Factor	Correlation Value	P-Value
0	전날이산화질소농도(ppm)	0.39	0.00
1	전날오존농도(ppm)	-0.03	0.05
2	전날이산화탄소농도(ppm)	0.40	0.00
3	전날아황산가스(ppm)	0.41	0.00
4	전날강수	-0.18	0.00
5	전날강수형태	-0.17	0.00
6	전날기온	-0.23	0.00
7	전날습도	-0.12	0.00
8	전날풍속	-0.07	0.00
9	전날풍향	0.15	0.00
10	전날하늘상태	0.00	0.85
11	전날 미세먼지(㎍/㎥)	0.62	0.00

최종 선정

당일: 이산화질소 / 이산화탄소 / 아황산가스 / 습도

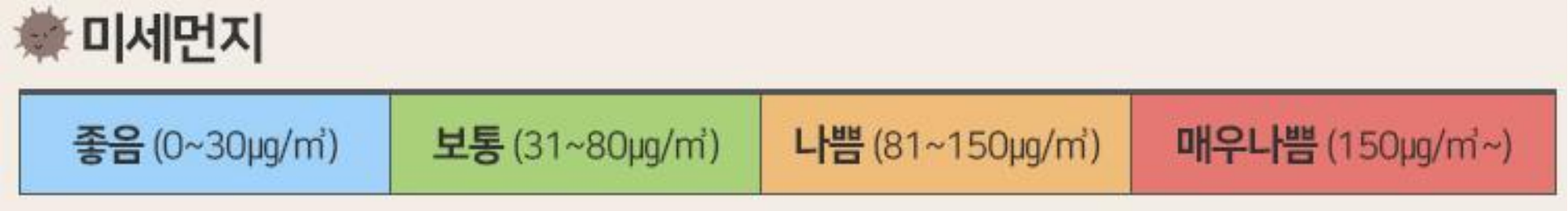
이전날: 강수 / 강수형태 / 기온 / 풍속 / 풍향 / 미세먼지농도

제거: 오존농도 / 하늘상태

최종 데이터 소개

	측정일시	이산화질소농도(ppm)	이산화탄소농도(ppm)	아황산가스(ppm)	습도	전날 미세먼지(㎍/㎥)	전날강수	전날강수형태	전날기온	전날풍향	전날풍속	하늘상태	미세먼지(㎍/㎥)	등급
1513	2015-02-23	0.024	0.6	0.005	63.000000	245.0	0.000000	0.000000	1.158333	267.333333	4.116667	1.625000	566.0	아주나쁨
121	2011-05-02	0.039	0.7	0.004	50.333333	180.0	0.000000	0.000000	11.787500	233.958333	3.550000	2.250000	297.0	아주나쁨
77	2011-03-19	0.027	0.2	0.006	44.625000	41.0	0.000000	0.000000	7.754167	214.875000	4.162500	3.500000	283.0	아주나쁨

- 중구 교남동에서 측정된 대기 오염 정보 및 기상 정보 데이터.
- 2011-01-01 ~ 2020-12-31 사이의 (3585, 14)개의 데이터 존재.
- '등급'의 경우, 정부에서 제공하는 미세먼지 농도 가이드라인에 따라 계산. (아래 등급 표 참조)



데이터 모델링 과정

- 데이터 모델 선정 과정.
- 그리드 서치를 이용한 파인 튜닝.
- 주요 요인 분석.

모델 선정 과정

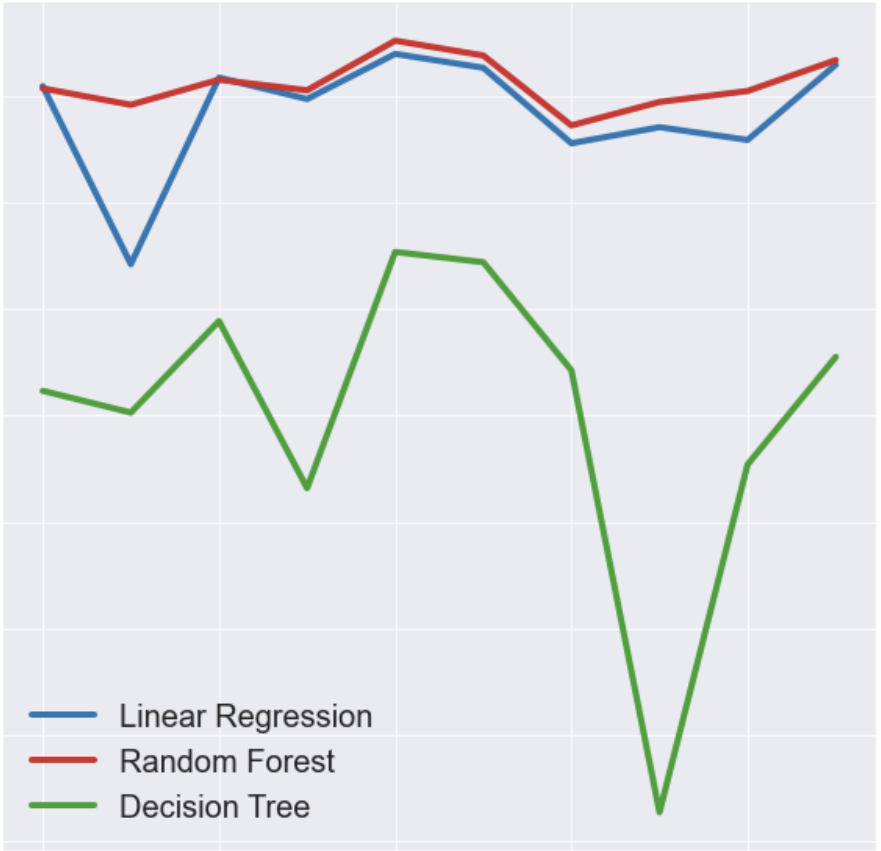
사용 모델 소개:

1. **Linear Regression**
2. **Random Forest**
3. **Decision Tree**

선정 과정:

1. Parameter 변경 없이 디폴트 비교.
2. Cross validation 10번을 진행하여 RMSE, R2 비교 및 평균 측정.
3. TEST SET을 이용하여 RMSE, R2 Accuracy 측정.

RMSE



R2 Score

최종선정 모델:
Random Forest

MODEL	RMSE 평균	RMSE 편차	R2 평균	R2 표준편차	RMSE	R2	Accuracy
Linear Regression	13.37	5.44	0.62	0.083	13.36	0.60	73%
Random Forest	12.67	4.14	0.66	0.037	12.99	0.62	72%
Decision Tree	18.52	7.80	0.27	0.114	18.65	0.22	64%

그리드 서치 튜닝과정

변수 명	변수 설명	설정에 사용한 값
n_estimators	트리의 개수	[10, 50]
max_features	노드를 나눌 때 사용할 Feature 개수	[5, 10]
max_depth	각 트리 별 최대 레벨	[10, 50, None]
bootstrap	샘플링 과정에서의 replacement 허용 유/무	[True, False]

```
12.926201874226892 {'bootstrap': True, 'max_depth': 10, 'max_features': 5, 'n_estimators': 10}
12.500262623798736 {'bootstrap': True, 'max_depth': 10, 'max_features': 5, 'n_estimators': 50}
13.040352047791476 {'bootstrap': True, 'max_depth': 10, 'max_features': 10, 'n_estimators': 10}
12.799619687452992 {'bootstrap': True, 'max_depth': 10, 'max_features': 10, 'n_estimators': 50}
13.1086569838215 {'bootstrap': True, 'max_depth': 50, 'max_features': 5, 'n_estimators': 10}
12.607030229405009 {'bootstrap': True, 'max_depth': 50, 'max_features': 5, 'n_estimators': 50}
13.280153539094387 {'bootstrap': True, 'max_depth': 50, 'max_features': 10, 'n_estimators': 10}
12.713519435466015 {'bootstrap': True, 'max_depth': 50, 'max_features': 10, 'n_estimators': 50}
13.082415512700218 {'bootstrap': True, 'max_depth': None, 'max_features': 5, 'n_estimators': 10}
12.550560364609375 {'bootstrap': True, 'max_depth': None, 'max_features': 5, 'n_estimators': 50}
13.045141145716759 {'bootstrap': True, 'max_depth': None, 'max_features': 10, 'n_estimators': 10}
12.744953861970483 {'bootstrap': True, 'max_depth': None, 'max_features': 10, 'n_estimators': 50}
13.077023775009465 {'bootstrap': False, 'max_depth': 10, 'max_features': 5, 'n_estimators': 10}
12.713221456509421 {'bootstrap': False, 'max_depth': 10, 'max_features': 5, 'n_estimators': 50}
16.20607821056895 {'bootstrap': False, 'max_depth': 10, 'max_features': 10, 'n_estimators': 10}
16.22028980844443 {'bootstrap': False, 'max_depth': 10, 'max_features': 10, 'n_estimators': 50}
13.304443318567962 {'bootstrap': False, 'max_depth': 50, 'max_features': 5, 'n_estimators': 10}
12.82788605313027 {'bootstrap': False, 'max_depth': 50, 'max_features': 5, 'n_estimators': 50}
17.75730388922617 {'bootstrap': False, 'max_depth': 50, 'max_features': 10, 'n_estimators': 10}
17.725286016870005 {'bootstrap': False, 'max_depth': 50, 'max_features': 10, 'n_estimators': 50}
13.372486610503817 {'bootstrap': False, 'max_depth': None, 'max_features': 5, 'n_estimators': 10}
12.817903352055694 {'bootstrap': False, 'max_depth': None, 'max_features': 5, 'n_estimators': 50}
17.778768980563953 {'bootstrap': False, 'max_depth': None, 'max_features': 10, 'n_estimators': 10}
17.70395389015848 {'bootstrap': False, 'max_depth': None, 'max_features': 10, 'n_estimators': 50}
```

n_estimaotrs = 50
max_features = 5
max_depth = 10
boostrap = True

RMSE = 12.50 (0.44 감소)
R2 = 0.63 (0.01 증가)
Accuracy = 73% (1% 증가)

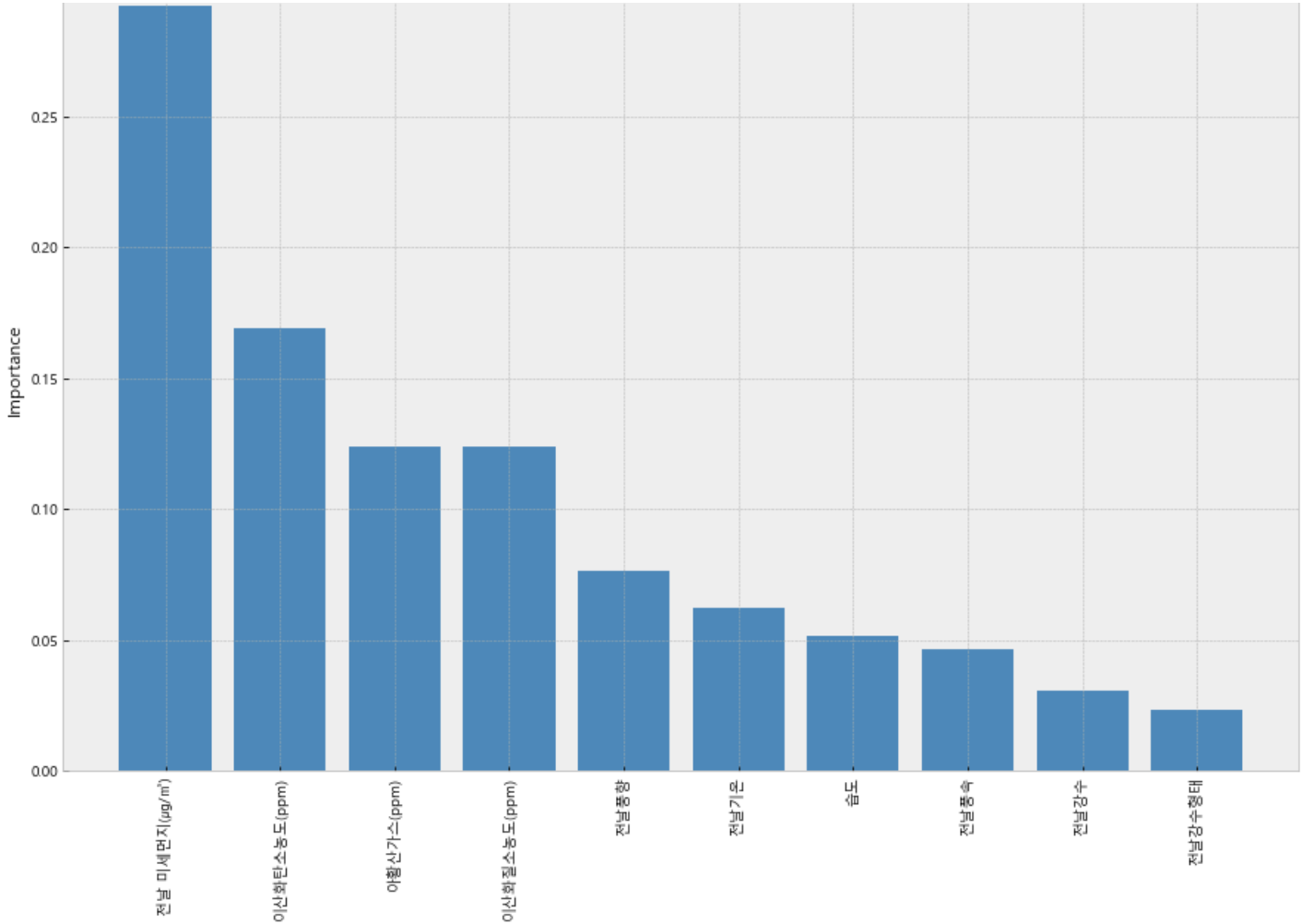
결론:
차이가 크지는 않지만, 튜닝 과정을 통해 더 좋은 예측 결과를 보여준다.

주요 요인 분석

- 각각 랜덤포레스트의 변수로써 결과를 도출하는데 어느 정도의 중요성을 가지는지 측정 (Feature Importance)

- 전날 미세먼지의 경우 미세먼지를 예측하는데, 전체 변수 10가지 중 총 30%의 중요도를 보이며 가장 높았다.

- 상관 관계 분석과 마찬가지로, 기상 데이터에 비해, 대기오염 정보 데이터가 미세먼지 농도를 예측 하는데, 중요도가 높았다.



최종 결론 및 향후 과제

- 결론 및 한계점 해결 방안 제시
- 향후 과제

결론 및 한계점 해결 방안

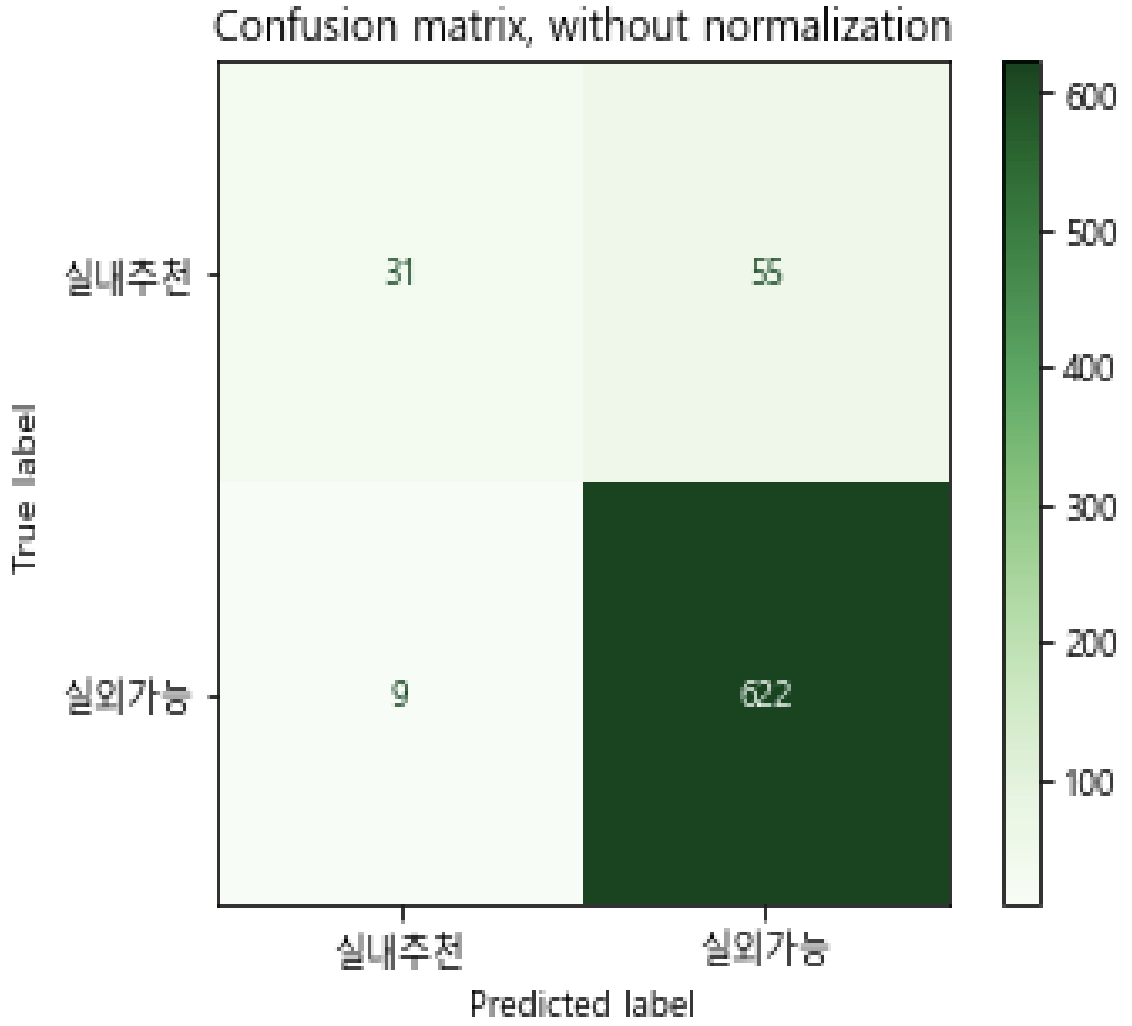
최종 선정된 모델은 Random Forest이며 Fine Tuning을 통해 최종적으로 RMSE가 12.50, R2 = 0.63, Accuracy가 73% 인 미세먼지 예측 모델을 만들 수 있다.

그러나 미세먼지 농도를 정확하게 예측한다고 말하기에는 힘들다.

따라서, 미세먼지를 등급 별로 분류하여 새로운 모델을 만들거나, 일정 미세먼지 농도를 '실내 추천', '실외 추천' 으로 구분하여 분류하면 좀 더 정확한 결과를 만들 수 있다.

미세먼지 농도를 45 $\mu g/m^3$ 미만일 경우를 '실외가능' 이상인 경우를 '실내추천 ' 으로 분류하여 Fine Tuning을 거치지 않은 Random Forest Classification 모델을 통해, 대략 80% 정확도를 가지고 분류가 가능하였다.

해당 모델을 발전시키거나, 다른 분류 모델의 정확성을 높이면, 앞선 대기오염 데이터와 기상 데이터를 통해 연구 목적인 실내외 여가 계획을 수립하는데 도움이 될 것이라 판단된다.



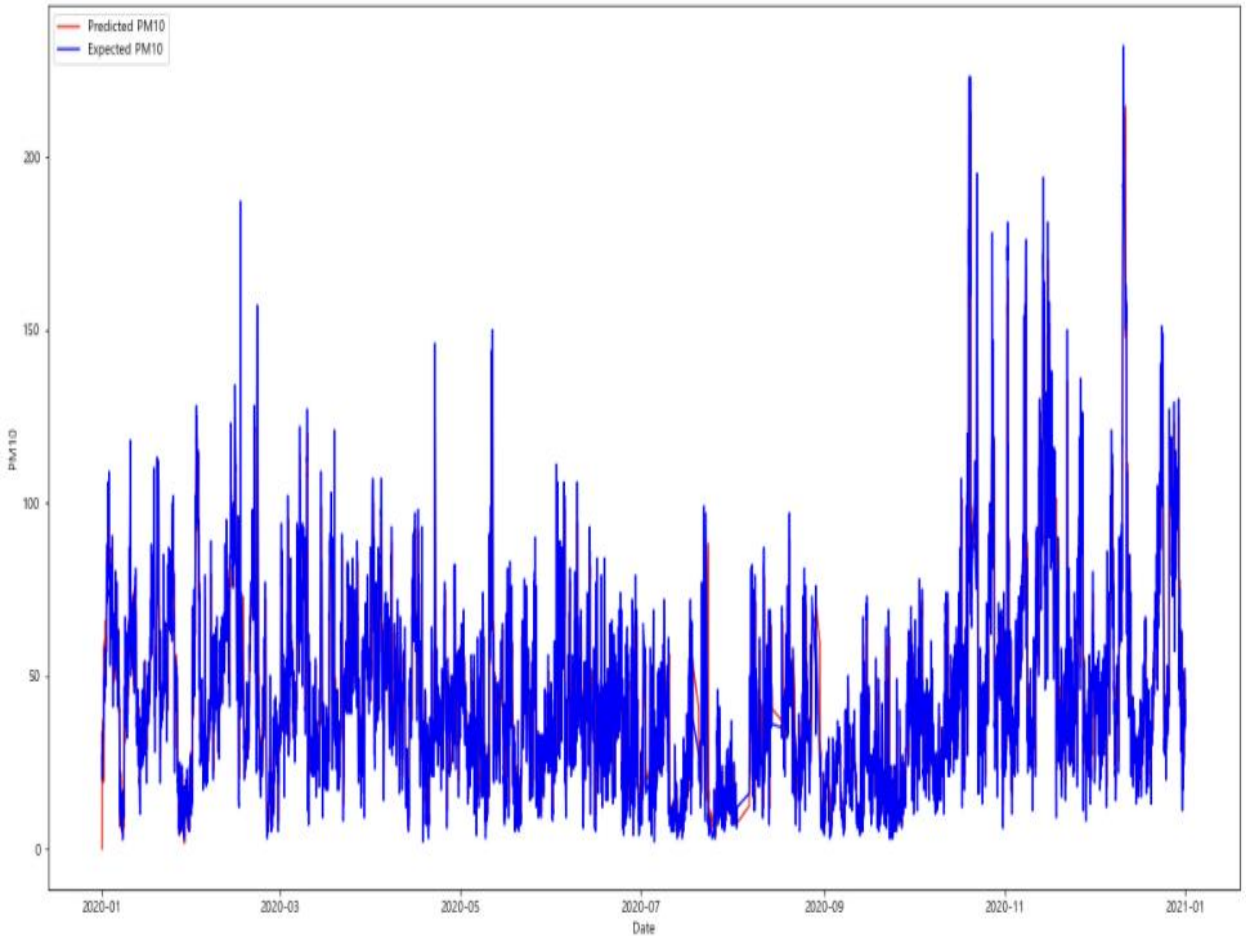
향후 과제

모델링 과정에서 새로운 변수 추가시키기.

- 한국의 계절 특성상 겨울에는 북서풍, 여름에는 남서풍이 분다.
- 중국, 몽골에서의 대기오염,기상데이터 추가.

미세먼지 농도 예측하기.

- 시계열 분석을 통해 미세먼지 농도 예측하기.



2020년 미세먼지 농도 시계열 분석 그래프

감사합니다.