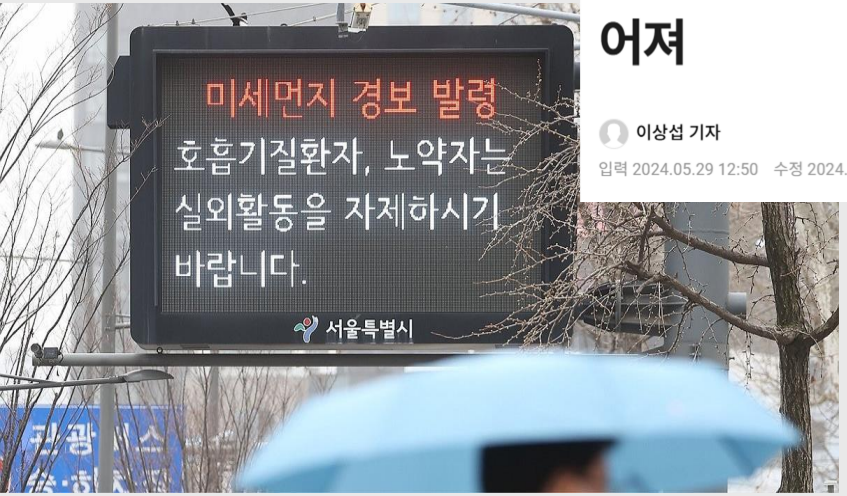
대기오염물질과 호흡기 질환 예측

MZ오피스

2021204097 김나연 (Jenny)
2020204023 신해리 (Samuel)
2019510035 백지헌 (Amanda)
2020204066 송승규 (Scarlett)
2021204045 이성민 (Teemo)

목차

1.1. 프로젝트 배경



의료와사회 : 톡톡뉘우스

“쉽게 나아지지 않는 기침”...환절기 지나도 호흡기 질환 유행 길어져

이상섭 기자

입력 2024.05.29 12:50 수정 2024.05.29 13:21

특화섹션 esc섹션

천식 환자 더 힘든 미세먼지·꽃가루의 계절...악화 피하려면 [ESC]

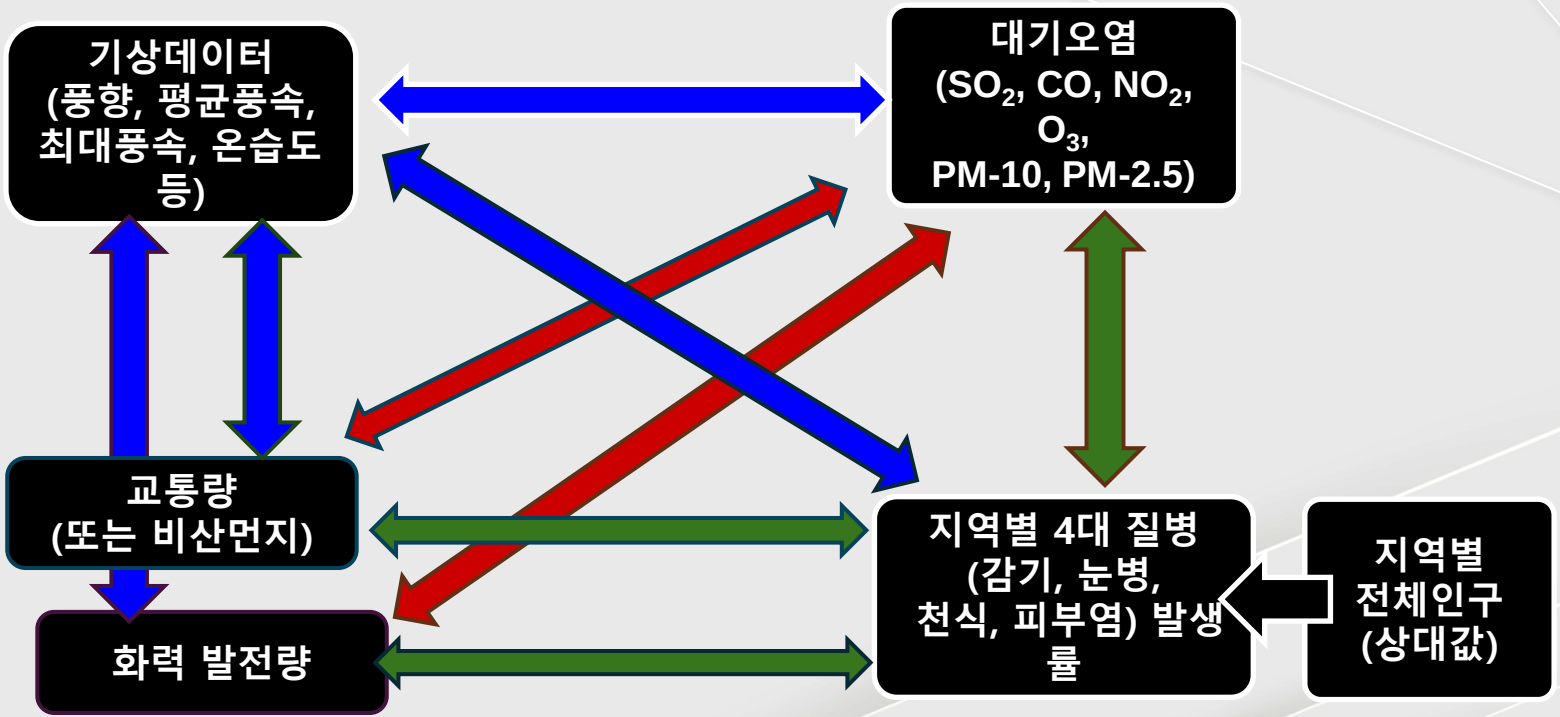
건강 : 천식

원인물질 노출되면 호흡 곤란

흡입용 스테로이드 가장 효과적

노년층, 폐렴 등 예방접종 필요

1.2. 프로젝트 목표



질병과 바람을 타고 확산하는 여러 오염원 간의 상관관계를 바탕으로 환자 발생률 예측

2.1. 데이터 수집

2017년~2022년(5년간), 각 시군구 별, 1일 단위

- 호흡기 질환 발생량 (481643, 7)
- 대기오염 물질 발생량 (385738, 9)
- 화력발전소 발전량 (188102, 8)
- 도로비산먼지 발생량 (53391, 12)
- 풍향 및 풍속 등 기상 데이터 (212546, 16)

 # 별첨1

질병 발생률(인구수 대비)

DATE	시군구	인구수	감기 발생률(%)	눈병 발생률(%)	천식 발생률(%)	피부염 발생률(%)
2020-12-23	서울 영등포구	372301.0	0.3	0.06	0.03	0.08

대기오염물질

Unnamed: 0	DATE	시군구	SO2	CO	O3	NO2	PM10	PM25
281919	2021-03-22	전남 고흥군	0.00119	0.5	0.038286	0.002905	38.380952	9.47619

화력발전량

DATE	지역	위도	경도	발전소 이름	연료원	발전량(MWh)	용량(MW)
2019-09-15	경기 평택복합	37.005715	126.798259	평택화력발전소	LNG	51595.0	869

도로재비산먼지

DATE	시군구	도로명	시작점	종점	측정거리(km)	기온(℃)	습도(%)	재비산먼지 평균농도(㎍/㎥)	상태	lat	long
2020-08-19	광주 광산구	임방울대로	광주광역시 광산구 우산동 535-4 도	광주광역시 광산구 쌍암동 696도	4.05	35	55	10.0	매우좋음	35.160911	126.819395

풍향,풍속

Unnamed: 0	관리관서	지점번호	시군구	지점주소	위도	경도	일시	평균풍속(m/s)	최대풍속(m/s)	최대풍속 풍향(deg)	최대풍속 시각
184343	대구(구 143)	276	경북 청송군	경상북도 청송군청송읍 길안청송로1591-9(덕리 519-1) 청송군공동협력 기상관측소	36.4351	129.0401	2018-05-29	0.8	3.4	110.0	17:32

2.2. 데이터 전처리 및 통합

- **기상 관측소가 없는 지역 결측치 (지리적 군집화)**
  - 기상관측소가 있는 지역을 중심으로 위경도 기준 지리적 군집화하여 최근접 기상 관측소의 같은 날 데이터로 보간
- **대기오염물질의 정규화 및 결측치 처리**
  - 오염물질별 상이한 스케일 정규화
  - 동일한 시군구의 결측치 전,후 3일간의 평균으로 보간

- **질병발생률의 결측치 처리**
  - 주변 날에서 값을 분배하여 보간 (예: 7,n,5 -> 5,4,3)

- **질병 발생 수 피쳐 엔지니어링**
  - 행정구역별 인구수로 나누어 발생률로 환산하였음
  - 천식 발생률 주 데이터로 사용

2.2. 데이터 전처리 및 통합

- **풍향 + 화력발전+ 도로먼지 데이터 통합**
  - 행정구역별 위경도 데이터를 기준으로, 모든 데이터가 고유 날짜 포함하도록 확장
  - 행정구역 데이터와 화력발전, 풍향풍속, 도로비산먼지 데이터 프레임을 각각 결합
- **결합된 데이터 군집화**
  - 각 데이터 셋의 위도와 경도 정보에 대하여 K-Means 클러스터링 수행
  - 위도와 경도 기반으로 병합된 데이터에 클러스터 할당
- **공간적 결측치 보간**
  - 클러스터 근접도를 기반으로 가중 평균을 사용하여 각 데이터프레임의 위도와 경도에 대한 결측값을 채우기
  - 데이터 구조화 후, 역거리 가중치 방법을 통한 결측값 채우기

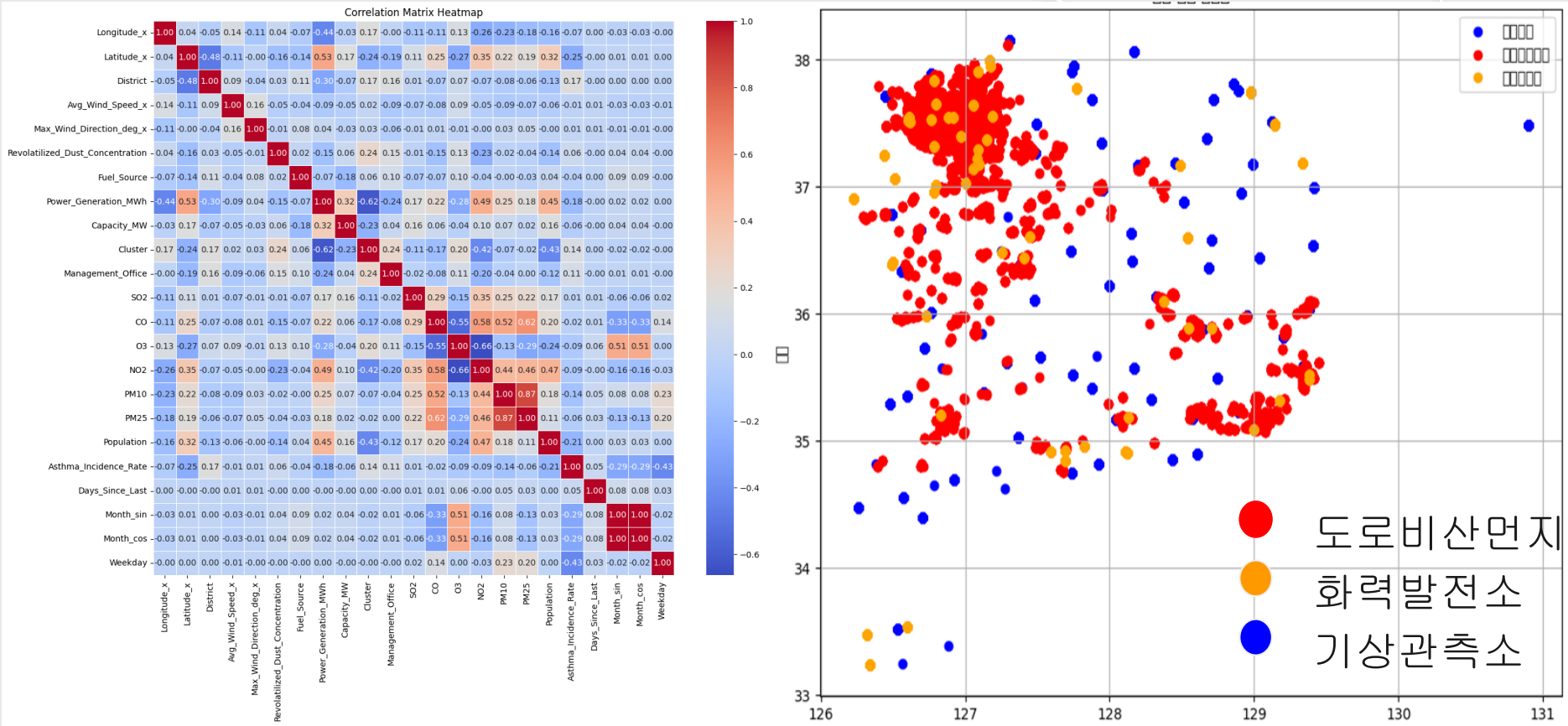
2.2. 데이터 전처리 및 통합

- **시간 단위 통일**
  - 모든 데이터셋의 시계열을 DATETIME 형태로 통합
- **공간 단위 통일**
  - 모든 데이터셋의 공간적 차원을 좌표가 존재하는 모든 위치로 통합

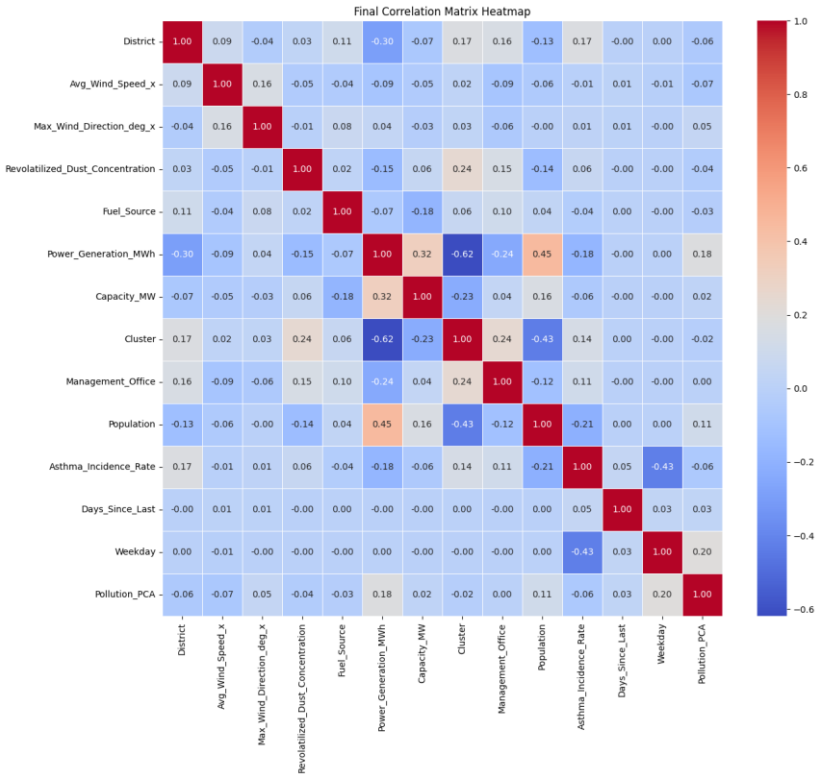
- **풍향+ 질병+ 대기오염 데이터 통합**
  - 화력발전+도로먼지와 DATE 및 시군구 기준 통합
- **피쳐 엔지니어링**
  - DATE 중 계절성 정보인 '월' 값을 코사인 변환
  - 범주형 변수 제거 또는 원 핫 인코딩

총 603,805 건의 데이터셋 구축

2.3. EDA


Correlation Matrix Heatmap
Longitude_x
Latitude_x
District
Avg_Wind_Speed_x
Max_Wind_Direction_deg_x
Revolatilized_Dust_Concentration
Fuel_Source
Power_Generation_MWh
Capacity_MW
Cluster
Management_Office
SO2
CO
O3
NO2
PM10
PM25
Population
Asthma_Incidence_Rate
Days_Since_Last
Month_sin
Month_cos
Weekday
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
38
37
36
35
34
33
126
127
128
129
130
131
도로비산먼지
화력발전소
기상관측소

2.4. 다중공선성 진단 및 PCA



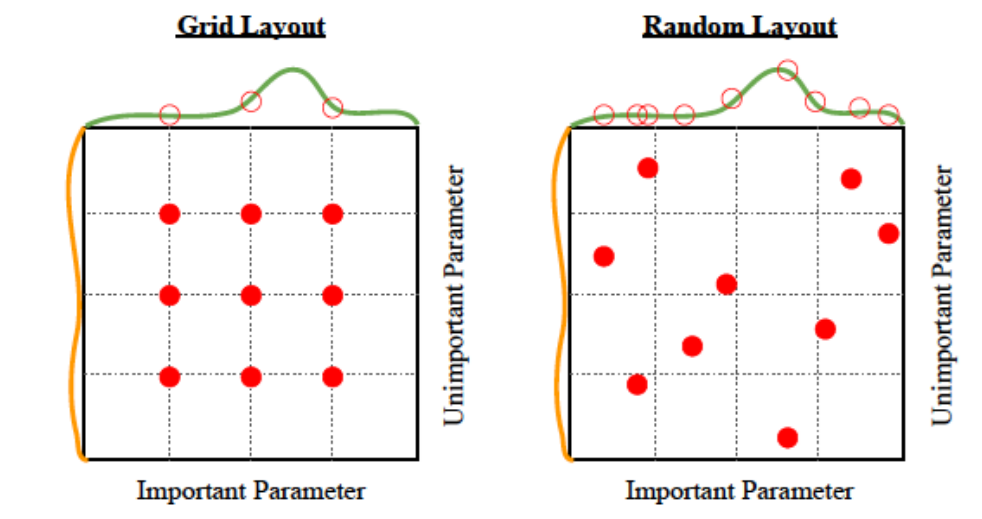
2	District	5.130162
3	Avg_Wind_Speed_x	9.637511
4	Max_Wind_Direction_deg_x	25.955418
5	Revolatilized_Dust_Concentration	13.096702
6	Fuel_Source	1.114279
7	Power_Generation_MWh	5.044930
8	Capacity_MW	10.183735
9	Cluster	4.965898
10	Management_Office	4.579210
11	SO2	10.073390

지면상의 한계로 일부만 출력

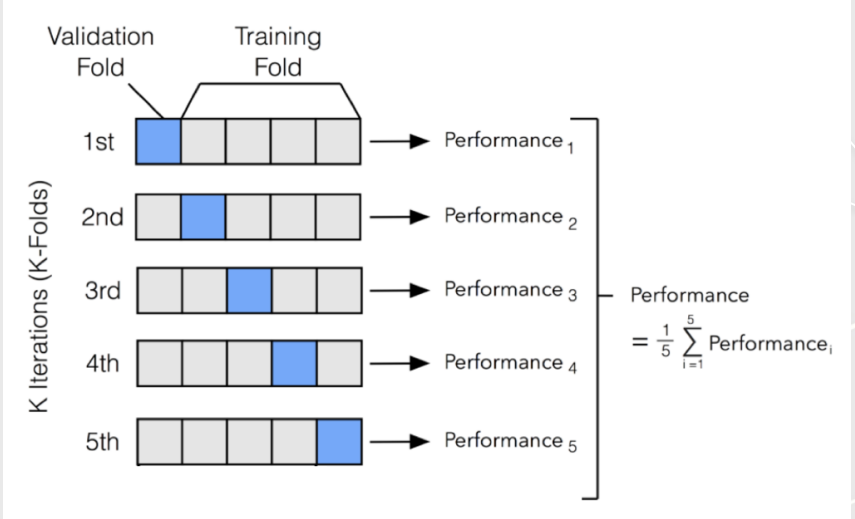
3.1. 모델 선정

- **XGBoost** : 높은 정확도와 일반화 성능, 빠른 학습 속도
- **Random Forest** : 과적합 방지, 높은 예측 성능
- **Support Vector Machine** : 과적합이 되는 경우가 적음.
- **Linear Regression** : 빠른 계산 속도
- **LightGBM** : 높은 정확도, 빠른 학습 속도
- **CatBoost** : 과적합 방지, 빠른 학습 속도
- **GBM** : 과적합 방지, 높은 정확도와 일반화 성능

3.2. 모델 학습



최적의 하이퍼파라미터를 랜덤 서치 방식으로 탐색

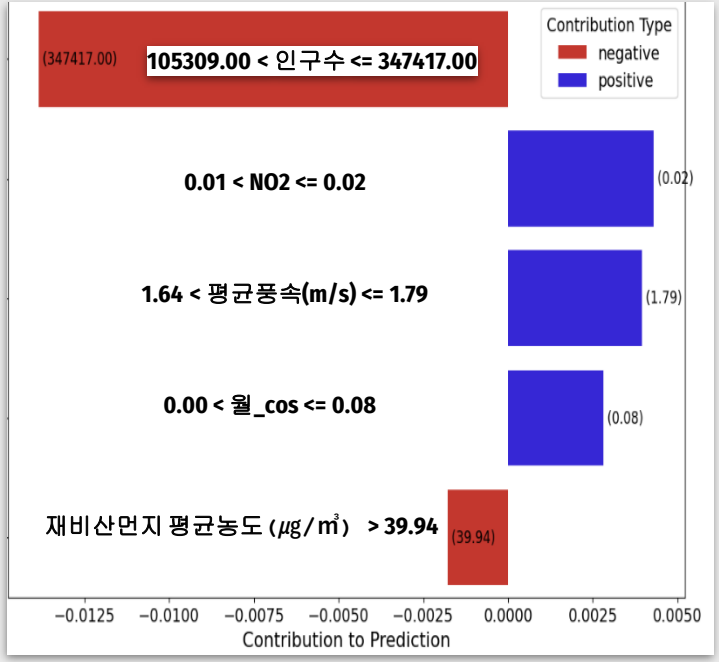


전체 데이터의 특성을 반영하기 위한 **k-fold** 기법

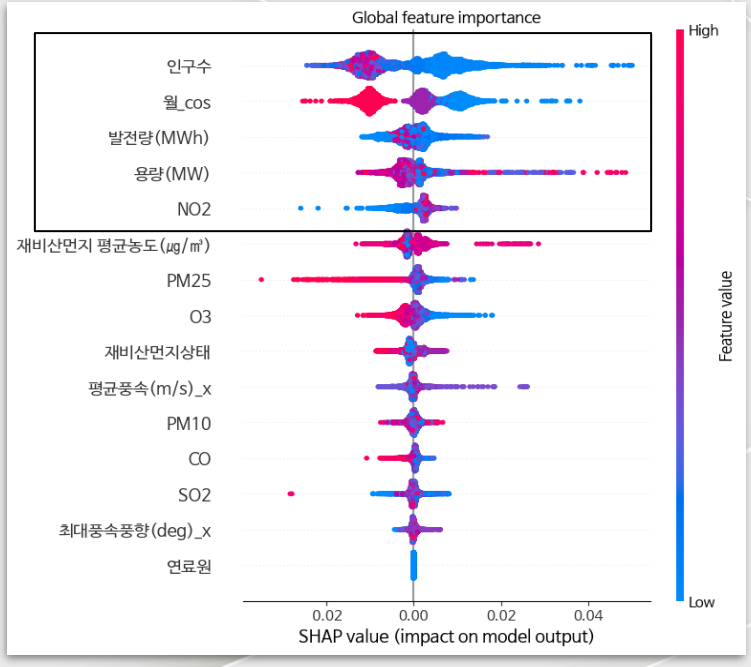
3.3. 모델 학습 결과 및 앙상블

	RMSE (Fold 별 평균, K=5)	R2	MAE	MSE	MedAE	EVS	Adjusted R^2
XGBoost	0.01841 (0.038)	0.778092	0.011533	0.000339	0.006851	0.778352	0.763259
RandomForest	0.020474 (0.0161)	0.725534	0.012682	0.000419	0.008094	0.725646	0.707188
SVM	0.053062 (0.0634)	-0.843479	0.04537	0.002816	0.040283	0.029403	-0.966707
LinearRegression	0.024541 (0.0384)	0.605666	0.016594	0.000602	0.011904	0.605937	0.579307
LightGBM	0.017876 (0.266)	0.790782	0.011145	0.00032	0.006699	0.791674	0.776797
CatBoost	0.018961 (0.349)	0.764606	0.011323	0.00036	0.006499	0.76537	0.748871
GBM	0.019336 (0.0067)	0.762267	0.011912	0.000357	0.007027	0.800382	0.755346
Stacking Ensemble	0.017266	0.804797	0.010687	0.000298	0.006376	0.805667	0.791749

4-2. 결과 해석 - XGBOOST

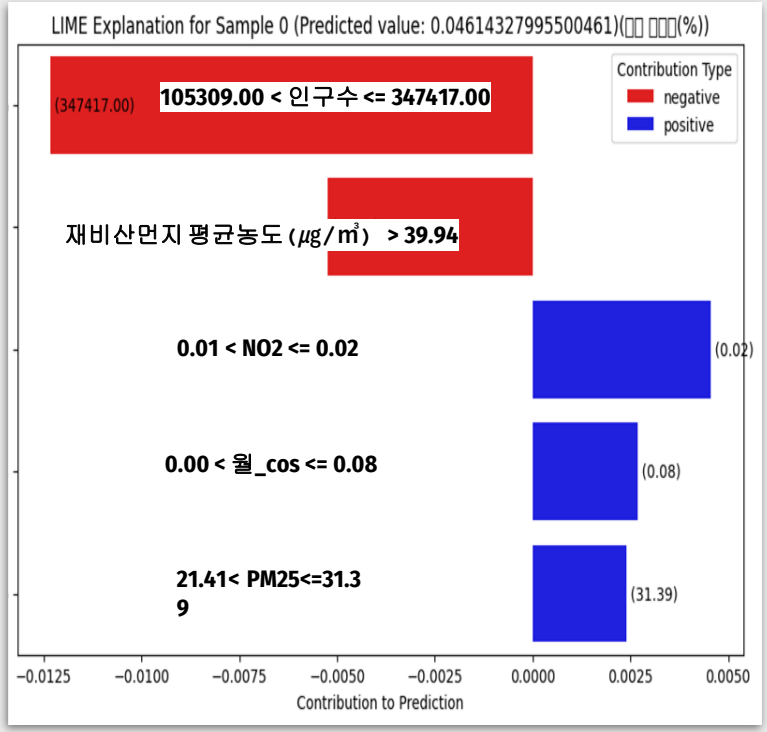


LIME

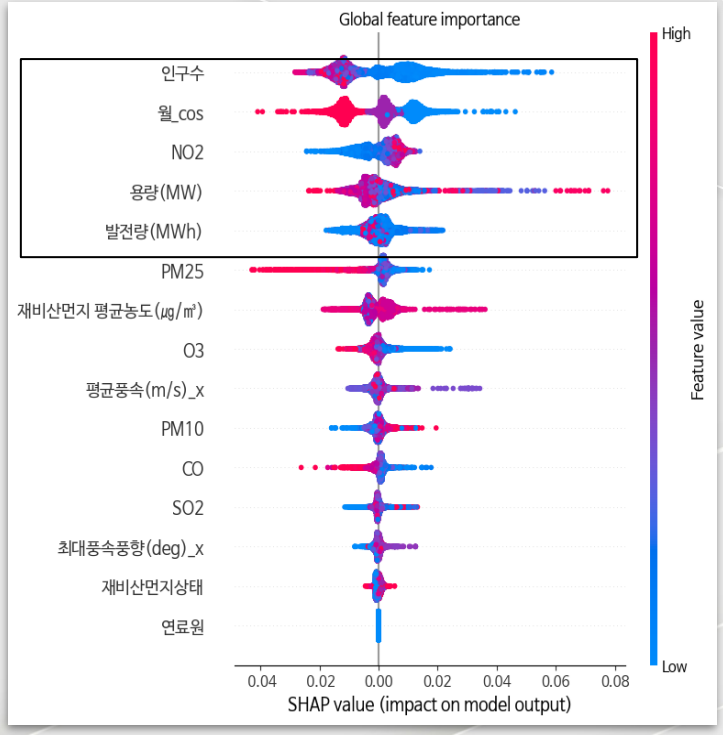


SHAP

4-2. 결과 해석 - LightGBM

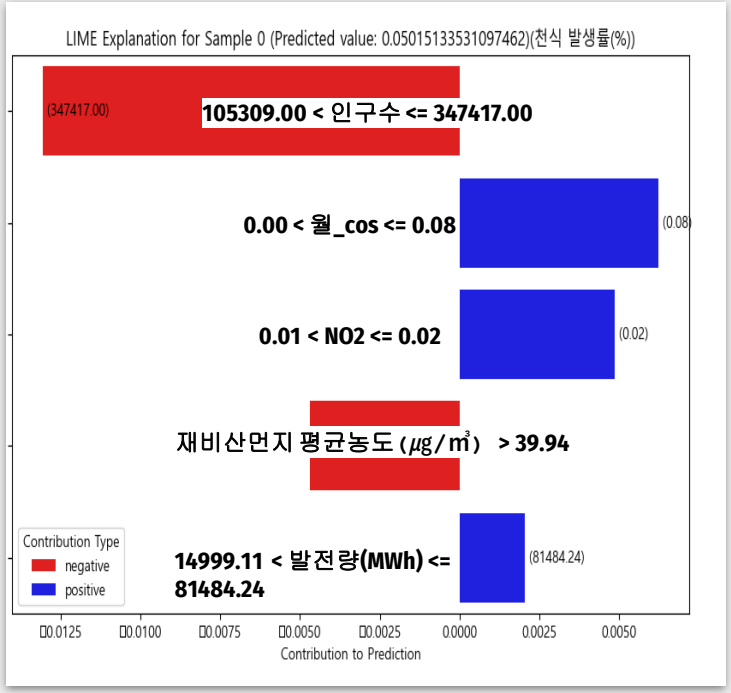


LIME

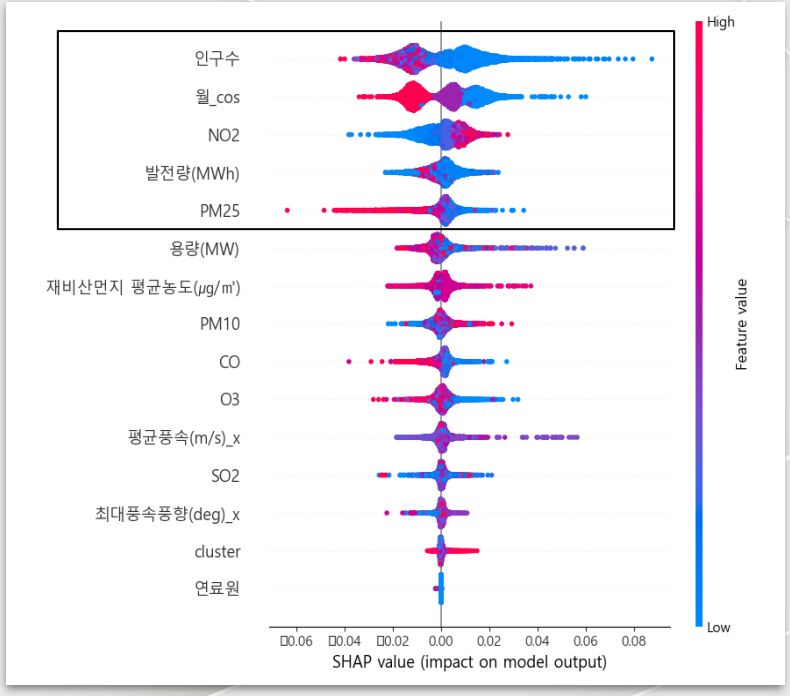


SHAP

4-2. 결과 해석 - GBM



LIME



SHAP

5. 프로젝트의 시사점, 활용 방안, 한계점

- **시사점**

화력발전량, 오존 및 미세먼지 농도, 도로비산먼지 수집량 등이 질병 예측에 유의한 영향력을 제공함

- **활용방안**

질병 환자수 예측

질병 예방 및 대비

질병 연구 데이터 분석

- **한계점**

제한된 데이터와 모델링 환경

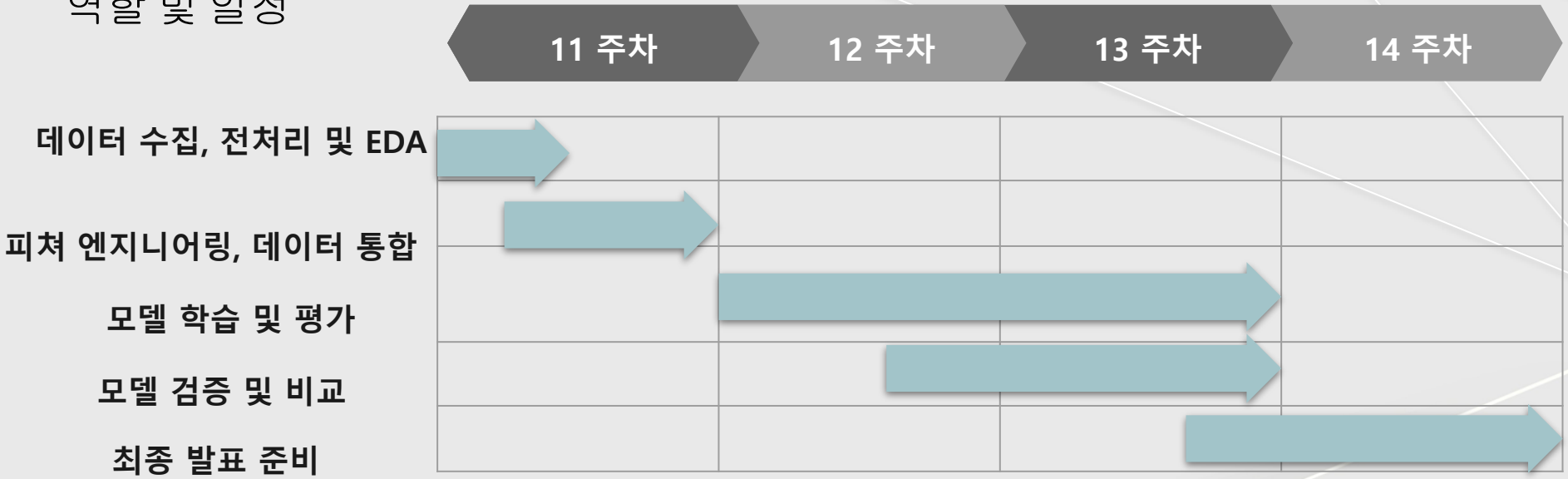
하이퍼파라미터 튜닝 시 랜덤 서치의 한계점

다중공선성 해소를 위한 PCA가 해석하기 어려움

모델의 해석력의 한계

시계열 특성을 충분히 반영하지 못함

역할 및 일정



- **데이터 전처리 및 통합:** 백지헌, 김나연
- **모델링:** 송승규, 이성민, 신해리
- **발표 자료 제작 및 통합:** 백지헌, 송승규
- **발표 (영상):** 이성민
- **프로젝트 매니지먼트 :** 백지헌

발표를 들어주셔서 감사합니다.

#별첨 1. 참고문헌

- https://www.hani.co.kr/arti/specialsection/esc_section/1134521.html
- https://www.rapportian.com/news/articleView.html?idxno=207486

별첨 2. 로우데이터 출처

- 호흡기 질환 발생량 데이터
  - 출처 : https://kosis.kr/statHtml/statHtml.do?orgId=584&tblId=DT_58401_B001040
- 대기오염 및 미세먼지 데이터
  - 출처 https://www.bigdata-environment.kr/user/data_market/detail.do?id=e09c4940-38bb-11ea-be28-4fa0eb812a46
- 화력 발전량 데이터
  - 출처 : https://www.data.go.kr/data/15069345/fileData.do?recommendDataYn=Y
- 기상 데이터
  - 출처 : https://data.kma.go.kr/climate/RankState/selectRankStatisticsDivisionList.do
- 도로비산먼지 데이터
  - 출처 : https://www.cleanroad.or.kr/index.do