한국정보전자통신기술학회논문지(jkiiect)'21-8, Vol.14 No.4
논문 21-14-04-328 http://dx.doi.org/10.17661/jkiiect.2021.14.4.328

기상 데이터와 대기 환경 데이터 기반 (초)미세먼지 분석과 예측

박홍진*

Analysis and Prediction of (Ultra) Air Pollution based on Meteorological Data and Atmospheric Environment Data

Hong-Jin Park*

요 약 석면, 벤젠과 같이 발암물질 1급인 미세먼지는 각종 질병에 원인이 되고 있다. 초 미세먼지 확산은 코로나 바이러스 확산의 중요한 원인중 하나이다. 본 논문은 2015년부터 2019년까지 서울시 평균 기온, 강수량, 평균 풍속등의 기상 데이터와 SO2, NO2, O3,등의 대기 환경 데이터를 기반으로 미세먼지와 초 미세먼지를 분석하고 예측한다. 계절별과 월별로 미세먼지와 초미세먼지 현황을 파악·분석하며 미세먼지를 예측하기 위해 기계학습 모델 중 선형회귀, SVM, 앙상블 모델을 이용하여 비교 분석하였다. 또한 미세먼지와 초 미세먼지 발생에 영향을 미치는 중요한 피쳐(속성)를 파악한다. 본 논문이 파악한 결과 3월에 가장 (초)미세먼지가 높았고, 8월에서 9월까지 (초)미세먼지가 가장 낮았다. 기상 데이터일 경우 (초)미세먼지에 가장 영향을 미치는 데이터가 평균 기온이며, 기상 데이터와 대기 환경 데이터일 경우 NO2가 (초)미세먼지 발생에 가장 크게 작용하였다.

Abstract Air pollution, which is a class 1 carcinogen, such as asbestos and benzene, is the cause of various diseases. The spread of ultra-air pollution is one of the important causes of the spread of the corona virus. This paper analyzes and predicts fine dust and ultra-air pollution from 2015 to 2019 based on weather data such as average temperature, precipitation, and average wind speed in Seoul and atmospheric environment data such as SO2, NO2, and O3. Linear regression, SVM, and ensemble models among machine learning models were compared and analyzed to predict fine dust by grasping and analyzing the status of air pollution and ultra-air pollution by season and month. In addition, important features(attributes) that affect the generation of fine dust and ultra-air pollution are identified. The highest ultra-air pollution was found in March, and the lowest ultra-air pollution was observed from August to September. In the case of meteorological data, the data that has the most influence on ultra-air pollution is average temperature, and in the case of meteorological data and atmospheric environment data, NO2 has the greatest effect on ultra-air pollution generation.

Key Words : Air pollution, Ultra-air pollution, Meteorological data, Atmospheric environment data, Linear regression, SVM, Ensemble models

1. 서론

대기 중의 대기오염 물질은 미량으로 존재하지만 직접적이고 장기적인 노출로 인해 인체에 각종 질병을 야기시키는 것 뿐만아니라 사망률도 증가시키는 원인이 되고 있다. 미세먼지는 석면, 벤젠와 같이 세계보건기구(WHO)에서 1군 발암물질로 분류하고 있다. 미세먼지는 아황산 가스,

This Paper was supported by research Fund of Sangji University in 2019.
*Dep. of Computer Engineering, Sangji University(hjpark1@sangji.ac.kr)
Received August 10, 2021 Revised August 10, 2021 Accepted August 21, 2021

질소 산화물, 납등을 포함하는 대기오염 물질로 대기 중에 장기간에 떠다니는 입자의 지름이 10㎍/㎥ 이하의 먼지이며, PM10 이라고 한다. 입자의 지름이 2.5㎍/㎥ 이하인 경우에는 PM2.5라고 하며, 초미세먼지 혹은 극미세먼지라고 한다. 미세먼지는 폐렴, 폐암 발생을 증가시키고 폐기능을 저하시키며, 만성호흡기질환자의 증상을 악화시킬 수 있다. 특히 만성폐쇄성폐질환(COPD)의 급성 악화를 유발하기도 한다. 질병관리본부 연구 보고서에 따르면, 미세먼지 농도가 10㎍/㎥ 증가할 때마다 만성폐쇄성폐질환으로 인한 입원률이 2.7%, 사망률은 1.1% 증가하고 초미세먼지 농도가 10㎍/㎥ 증가할 때마다 폐암 발생률이 9% 증가하는 것으로 보고된 바 있다. 또한, 최근 연구에 의하면 PM2.5 초미세먼지 증가는 신종 코로나바이러스(COVID-19) 확산의 중요한 원인으로 평가되고 있다. 초미세먼지 PM 2.5의 1㎍/m^3에 증가에 코로나바이러스의 대략 12% 증가하였으며[1], 초미세먼지 PM 2.5의 1㎍/m^3 에 증가에 따른 사망률을 8% 증가되었다[2]. PM2.5 초미세먼지 증가는 COVID-19의 확산에 큰 원인이 임을 알 수 있다.

본 논문은 2015년부터 2019년까지 서울시의 기상 데이터와 대기 환경 데이터를 이용하여 PM10인 미세먼지와 PM2.5인 초미세먼지를 분석하고 농도를 예측한다. 기상청에서 제공하는 평균 기온, 평균 풍속, 강수량등의 기상 데이터와 에어 코리아에서 제공하는 SO_2, NO_2, O_3등의 대기 환경 데이터를 기반으로 서울시 미세먼지와 초미세먼지를 분석·예측한다. 서울시의 미세먼지와 초미세먼지를 계절별, 월 미세먼지 현황을 파악·분석하며, 미세먼지 발생 영향에 미치는 중요한 피쳐(속성)을 파악한다. 또한, 기계학습에 사용되고 있는 선형회귀, SVM과 앙상블 모델을 이용하여 미세먼지를 예측한다.

2. 선행 연구

2010년 9월부터 2012년 12월까지 지상으로부터 약 21m 높이의 대구대학교 공과대학 6호관 옥상에 1.5m 높이의 PM2.5 채취기를 이용하여 총 260개의 시료를 채취하여 인자분석, 역궤적 군집분석을 위해 통계 분석 모델인 HYSPLIT 4모델과 PSCF 모델을 이용한 잠재적 오염원 위치를 추정하고자 하였다. 결과적으로 PM 2.5의 경우 주로 몽골, 중국의 내몽골, 안희성, 강소성, 하난성 지역이 장거리 이동되어 국내에 영향에 미치는 것으로 파악하였다[3].

선형 예측 방법인 MLR(Multi-linear Regresion), 비선형 예측방법인 SVR(Support Vector Regression)를 사용하여 대구 지역 통합 미세먼지 농도를 예측하고 각 알고리즘 성능을 RMSE(Root Mean Square Error)로 비교하였다. 2013년부터 2015년까지 일 평균 대구 지역 대기질 정보(SO_2, NO_2, O_3, CO, PM10)와 기상정보(기온, 강수량, 풍속)로 구성하였고 대구 지역 통합 미세먼지 예측 결과 MLR보다 SVR이 더 좋은 성능을 나타내었다[4].

에어 코리아의 대기오염 측정망에서 측정한 PM10의 이력 데이터와 기상청에서 제공하는 기상자료 개방포털에서 제공하는 서울의 일별 평균기온, 최저기온, 최고기온, 일강수량, 최대순간풍속, 평균풍속, 최대순간 풍속풍향 데이터를 2015년 1월 1일부터 2018년 9월 30일까지 데이터를 활용하여 LSTM(Long Short Term Memory)모델을 이용하여 미세먼지 농도를 예측하였다. 은닉층을 고정한 후에 에포크(epoch)를 변화시키면서, 에러율을 비교하였고, 에포크를 고정하면서, 은닉층을 변화시키면서 에러율을 비교하여 측정하여 미세먼지 농도 예측 방법을 비교평가하였다[5].

기상 데이터와 대기질 정보, 교통 데이터, 발전량 데이터를 이용하여 사용되는 기법으로 XGBoost, 랜덤포레스트, SVM, 인공신경망 알고리즘을 적용하여 모델간 비교 분석하였다. 비교분석한 결과 인공신경망 기법이 가장 예측 정확도가 높은 것으로 확인하였다[6].

베이징 PM2.5 데이터를 가지고 2010년 1월 1일부터 2014년 12월 31일까지 5년간 PM2.5, 강

우, 풍속등 7개의 환경 측정 항목을 매시간 측정한 자료를 가지고 LSTM 단독모델로 구성된 기존 방법과 CNN과 LSTM을 결합하는 방법의 예측 성능을 비교하는 실험을 수행하였다. 기존 방법보다 제안한 방법이 일관되게 향상된 결과를 보였다[7].

기계학습 방법인 앙상블을 이용한 대기질의 시계열 예측 방법과 지리적 가중회귀모델(GWR)을 이용한 공간 속성 기반 예측 방법을 혼합한 앙상블 모형을 통해 새로운 대기질 예측 방법을 제시하였다. 전라북도 혁신도시를 2019년 6월부터 11월 중순까지 91개 셀 단위로 수집하여 데이터화 하였다. 미세먼지 및 악취유발 물질 농도는 시계열 속성과 더불어 공간적인 속성에 의해 영향을 받는다는 것을 보여주였다[8].

3. 미세먼지 분석과 예측

3.1 미세먼지 분석

본 논문은 대기 환경 데이터와 기상 데이터를 2015년 1월부터 2019년 12월까지 월 별 평균 데이터를 사용하였다. 기상 데이터는 기상청에서 제공하는 월별 평균 기온, 평균 기온차이, 강수량, 강수일수, 상대습도량, 평균운량, 일조시간, 평균풍속, 최대 풍속등을 고려하였다. 강수일수는 일강수량이 0.1mm 이상인 날의 수를 의미한다. 대기 환경 데이터는 에어 코리아(https://www.airkorea.or.kr/)에서 제공하는 SO_2, NO_2, O_3, CO, PM10, PM2.5 데이터를 사용하였다. 본 논문에서 사용된 기계학습 모델은 선형회귀와 SVM, 여러개의 앙상블(ensemble) 모델이다. 앙상블은 하나의 모델만을 학습시켜 사용하지 않고 여러 모델을 학습시켜 결합하는 방식으로 예측하는 방식이다. 본 논문에서 사용된 앙상블 모델은 보팅(voting), 램덤 포레스트(Random Forest), 엑스트라 램덤 트리(Extra Random Tree), GMB(Gradient Boost Machine), XGBoost(eXtreme Gradient Boosting)이다.

보팅은 서로 다른 모델을 가지고 여러개의 분류기를 결합하여 최종 결과 값을 선정하는 형태로 가중치가 동일하면 다음 같은 수식으로 정의 될 수 있다. 여기에서 C_i는 분류기의 예측 값이다.

$$\hat{y} = \mathrm{mode}\{C_1(x), C_2(x), \ldots C_m(x)\}$$

램덤 포레스트는 배깅의 가장 대표적인 방식으로 기본적으로 100개의 의사결정 나무를 사용하고, 부트스트랩을 수행하여 여러개의 서브세트(subset)의 임의로 생성되며 전체 데이터 수는 동일 하지만 부트스트랩을 사용하기 때문에 각 분류기에서 사용된 데이터는 중첩되어 생성될 수 있다.

엑스트라 램덤 트리는 랜덤 포레스 방식과 비슷하나, 랜덤 포레스트 방식은 주어진 샘플에서 모든 변수에 대한 정보 이득을 계산하고 가장 높은 정보 이득을 가지고 트리를 가지를 분리하는데 반에 엑스트라 램덤 트리 방식은 가지를 분리할 때 임의의 속성을 선택하여 그 속상에 대해서만 최적의 정보 이득을 가지로 분리한다.

GBM은 에이다부스트와 비슷하나 가중치 업데이트를 경사 하강법을 사용하는 방식이다. 즉, GBM 방식은 경사 하강법을 이용하여 오류를 최소화 시키는 방향으로 반복적으로 가중치를 수정하는 방식이다.

XGBoost기법은 여러개의 회귀나무를 이용해서 오차값을 줄여나가면서 최적의 회귀나무를 찾아가는 방식이다.

[그림 1]은 2015년에서 2019년까지 서울시 미세먼지와 초미세먼지 현황을 나타내고 있다. 전체적인 추세선을 보면 PM10인 미세먼지는 해마다 줄어들고 있음을 알 수 있으며, PM2.5인 초미세먼지는 줄어들고 있지 않고 해마다 비슷한 농도도 분포되고 있음을 알 수 있다.

[표 1]은 2015년부터 2019년까지 서울시 기상 데이터의 분석을 보여주고 있다. [표 1]의 기온 평균을 보면, 2015년부터 2019년까지 월 전체 평균 기온이 13.3℃이고, 표준편차가 10,3℃, 월 평균 가장 낮은 온도가 영하 4℃이며 월 평균 가장 높은 온도는 28.8℃이다.

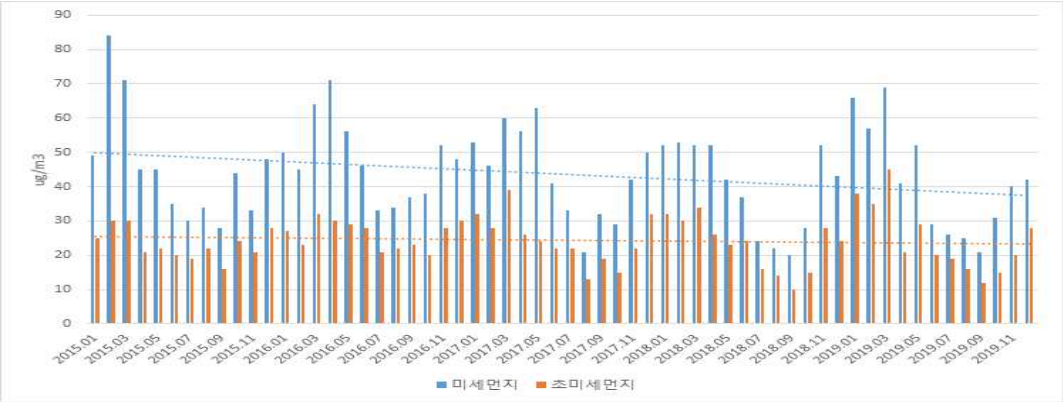


그림 1. 2015년~2019년 서울시의 (초)미세먼지 현황
Fig. 1. (Ultra) air pollution status in Seoul from 2015 to 2019

[표 2]는 2015년부터 2019년까지 서울시 대기 환경 데이터의 분석을 보여주고 있다. 환경부에서 제시한 국가 기준을 보면 SO_2의 경우 평균 0.005ppm로 좋은 수준(0.02ppm이하)이다, NO_2의 경우 국내 기준이 연간 평균치 0.03ppm 이하이나 0.03ppm으로 같음으로써 다른 대기 오염물질에 비해 상대적으로 NO_2 수치가 높음을 알 수 있다.

표 1. 2015년~2019년 서울시 기상 데이터 분석
Table 1. Analysis of weather data in Seoul from 2015 to 2019

구분	기온 평균 (℃)	평균 일교차(℃)	상대 습도 (%)	평균 운량 (10%)	일조 시간 (hr)	평균 풍속 (m/s)	최대 풍속 (m/s)	강수 량 (mm)	강수 일수 (일)
Count	60	60	60	60	60	60	60	60	60
Mean	13.3	9.22	58.2	4.7	213	2.17	7.35	86.8	8.2
std	10.3	1.35	7.66	1.15	41	0.39	1.33	103	3.5
min	-4	6.1	45	2.8	109	1.4	4.2	1	1.3
25%	4.60	8.10	53	3.9	185	1.9	6.62	26.7	5.7
50%	14	9.2	57.5	4.6	212	2.1	7.25	55	8.1
75%	22.7	10.2	63	5.32	236	2.4	8	100.	10
max	28.8	11.9	77	7.9	314	3.1	11.8	621	20

NO_2의 주요 배출원인 자동차와 발전소등이다. O_3,의 경우 평균 0.024ppm로 좋은 수준(0.03ppm이하)이다. CO의 경우 0.52ppm로 좋은 수준(2ppm이하)이다.

표 2. 2015년~2019년 서울시 대기 환경 데이터 분석
Table 2. Analysis of air environment data in Seoul from 2015 to 2019

구분	SO_2 (ppm)	NO_2 (ppm)	O_3 (ppm)	CO (ppm)	PM10 (㎍/m³)	PM2.5 (㎍/m³)
Count	60	60	60	60	60	60
Mean	0.005	0.030	0.024	0.525	43.70	24.31
std	0.001	0.006	0.009	0.114	14.15	6.981
min	0.003	0.017	0.009	0.4	20	10
25%	0.004	0.025	0.016	0.400	33.00	20.00
50%	0.005	0.031	0.025	0.5	43.5	23.5
75%	0.005	0.035	0.03	0.6	52	29
max	0.007	0.04	0.043	0.8	84	45

미세먼지 PM10인 경우 국가 기준이 연평균 50㎍/m³이하인데 서울시는 43.5㎍/m³ 을 나타내고 있고, 초 미세먼지 PM2.5인 경우 국가 기준이 연평균 15㎍/m³이하인데 23.5㎍/m³로 초과함

으로써 초미세먼지는 좋은 수준(15이하)이 아니고 보통(16~35)인 수준에 있음을 알수 있다.

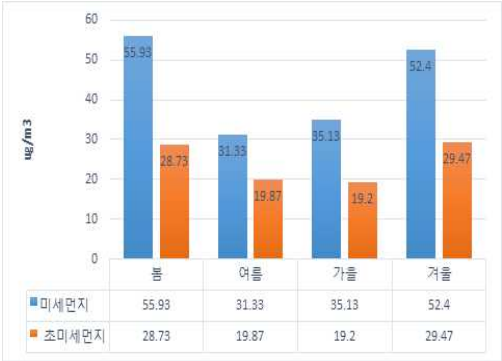


그림 2. 2015년~2019년 서울시의 계절별 평균 (초)미세먼지 현황

Fig. 2. Average (ultra) air pollution status by season in Seoul from 2015 to 2019

[그림 2]는 2015년부터 2019년까지 서울시의 계절별 평균 미세먼지와 초미세먼지 현황이다. 미세먼지 농도는 겨울과 봄에 농도가 높았으며, 여름과 가을에는 상대적으로 미세먼지 농도가 낮았다. 국내 봄에는 이동성 저기압과 중국의 황사를 동반한 미세먼지 발생이 크고, 상대적으로 비가 많이 내리는 여름에는 초미세먼지 농도가 낮았으며, 계절중에 대기의 순환이 원할한 가을도 미세먼지 농도가 낮았으며, 겨울에는 난방 연료 등로 증가로 다시 미세먼지 농도가 높음을 알 수 있다.

[그림 3]은 2015년부터 2019년까지 서울시의 월 평균 미세먼지와 초 미세먼지 현황을 나타내고 있다. 국내 미세먼지와 초 미세먼지는 3월에 가장 높음을 알수 있다. 3월에는 중국 몽골발 황사의 원인으로 미세먼지와 초 미세먼지의 농도가 가장 높다. 미세먼지 농도가 3월 가장 높다가 점점 낮아지는 추세가 가고 늦은 여름인 8월에 미세먼지 농도가 가장 낮았다. 가을 중간인 10월까지 미세먼지 농도가 낮다가 11월 상승하여 계속해서 미세먼지 농도가 상승하였다.

3.1 미세먼지 예측

미세먼지 예측 비교 평가를 하기 위해 2015년부터 2018년까지 데이터를 학습(train) 데이터로 사용하고 2019년 데이터를 테스트(test) 데이터로 사용하여 미세먼지 예측을 비교평가 한다.

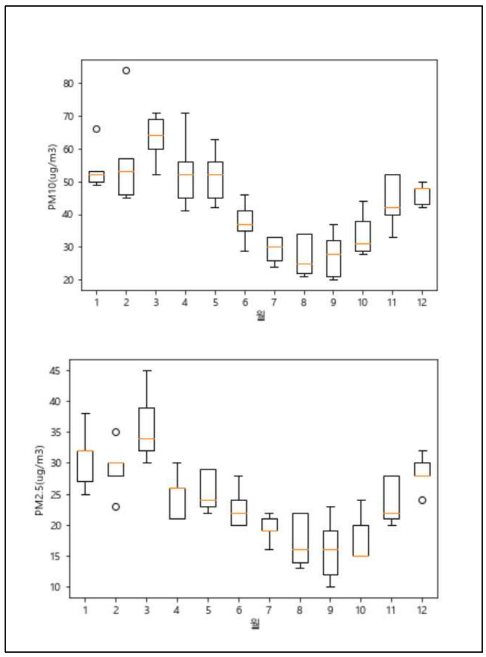


그림 3. 2015년~2019년 서울시 월 평균 (초)미세먼지 현황

Fig. 3. Average (ultra) air pollution status by month in Seoul from 2015 to 2019

미세먼지 농도를 예측하기 전에 서로 다른 입력 변수 값 범위를 일정한 수준으로 맞추는 작업을 피처 스케일링이라고 하는데 본 논문에서는 정규화(normalization)로 피처 스케일링을 수행하였다. 정규화는 데이터 값을 0과 1상의 범위로 변환하는 값이다. 정규화된 값 x_{i_new}은 원래 값에서 피처 x의 최솟값을 뺀 값을 피처 x의 최댓값과 최솟값의 차이로 나눈 값으로 변환되는 값이다.

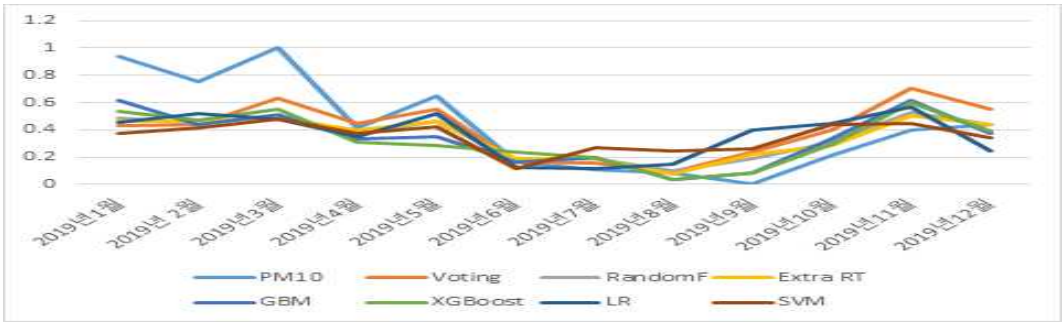


그림 4. 기상 데이터 기반 PM10 예측 비교 평가
Fig. 4. Comparative evaluation of PM10 prediction based on meteorological data

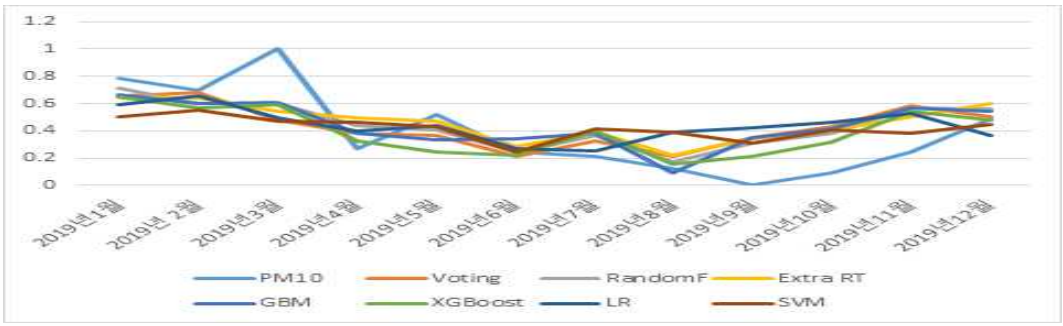


그림 5 기상 데이터 기반 PM2.5 예측 비교 평가
Fig. 5. Comparative evaluation of PM2.5 prediction based on meteorological data

$$x_i\_new = \frac{x_i - \min(x)}{\max(x) - \min a(x)}$$

본 논문에서 사용된 모델은 선형회귀, SVM, 보팅, 랜덤 포레스트, 엑스트라 랜덤 트리, GBM, XGBoost으로 총 7개의 기계학습 모델을 비교 평가하였다. 여러개의 개별 모델을 결합하여 예측하는 보팅 모델에 사용된 개별 모델은 선형회귀와 의사결정회귀, k-최근접 이웃회귀를 사용하였다. 랜덤 포레스트와 익스트림 랜덤 트리는 의사결정나무의 개수를 100개로 하고, 불순도는 지니 계수(gini index)를 사용하였다.

각 모델을 평가하기 위해 사용된 성능지표는 MAE(Mean Absolute Error), MSE(Mean Squared Error), RMSE(Root Mean Squared Error), MSLE(Mean Squared Log Error))을 이용하였다.

표 3. 기상 데이터 기반 PM10 예측 모델 성능 평가
Table 3. Performance evaluation of PM10 prediction model based on meteorological data

Model	MAE	MSE	RMSE	MSLE
Linear Regression	0.21277	0.072138	0.268585	0.021552
SVM	0.22422	0.078119	0.279498	0.033771
Voting	0.185765	0.058869	0.242630	0.024322
Random Forest	0.167638	0.053977	0.232330	0.021100
Extra Random Tree	**0.164796**	0.056047	0.236743	0.022139
GMB	0.178213	**0.051767**	**0.227524**	**0.020153**
XGBoost	0.185822	0.054529	0.233516	0.021552

표 4. 기상 데이터 기반 PM2.5 예측 모델 성능 평가
Table 4. Performance evaluation of PM2.5 predicti on model based on meteorological data

Model	MAE	MSE	RMSE	MSLE
Linear Regression	0.205593	0.066618	0.258105	0.020575
SVM	0.210423	0.063383	0.251761	0.031416
Voting	0.184228	0.057985	0.240800	0.029199
Random Forest	0.170909	0.044497	0.210945	0.023157
Extra Random Tree	0.188841	0.051649	0.227265	0.026674
GMB	8849250	0.049784	0.223125	0.026279
XGBoost	**0.16619**	**0.04195**	**0.20483**	**0.0205**

[그림 4]과 [그림 5], [표 3]과 [표 4]는 기상청에서 제공하는 월별 평균 기온, 평균 기온차이, 강수량, 강수일수, 상대습도량, 평균운량, 일조시간, 평균풍속, 최대 풍속의 서울시 기상 데이터 가지고 2015년부터 2018년까지의 데이터로 학습한 후에 2019년을 예측한 PM10의 미세먼지와 PM2.5의 초미세먼지 농도의 예측 결과이다.

2019년 3월에 PM2.5인 초미세먼지는 2015년에서 2019년까지 비교하여 가장 높았으며, 같은해 9월에 PM10인 미세먼지는 다른 해에 비해 매우 낮아서 실제 값과 예측 값이 약간의 차이를 보이고 있다.

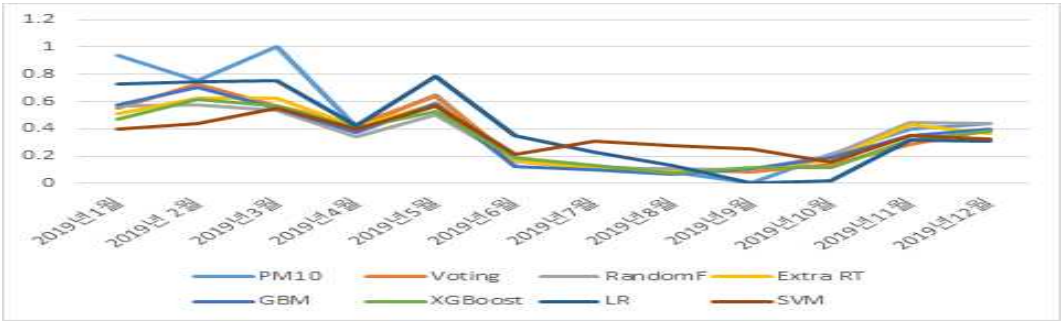


그림 6. 기상 데이터와 대기 환경 데이터 기반 PM10 예측 비교 평가
Fig. 6. Comparative evaluation of PM10 prediction based on weather data and atmospheric environm ent data

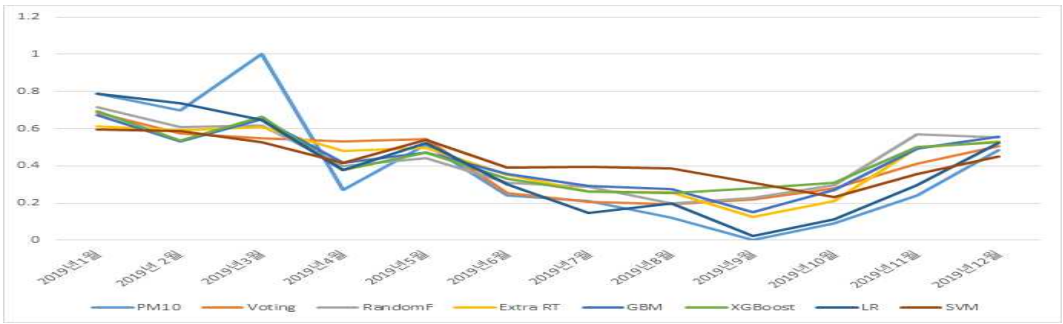


그림 7. 기상 데이터와 대기 환경 데이터 기반 PM2.5 예측 비교 평가
Fig. 7. Comparative evaluation of PM2.5 prediction based on weather data and atmospheric environ ment data

모델의 예측값과 실제값의 차이를 보여주는 성능지표 MAE를 기준으로 미세먼지는 엑스트라 램덤 트리가 초미세먼지는 XGBoost가 가장 잘 예측하였다. 그 다음 순위는 램덤 포레스트가 미세먼지나 초미세먼지에서 좋은 예측을 보여주였다. 모델의 예측값과 실제 값의 차의 면적 합을 나타내는 MSE는 미세먼지인 경우에 GBM, 램덤 포레스트 XGBoost 순으로 잘 예측하였다. 초미세먼지 경우 XGBoost가 가장 잘 예측하였고 램덤 포레스트, GBM 순으로 좋은 예측을 보여주고 있다.

표 5. 기상 데이터와 대기 환경 데이터 기반 PM10 예측 모델 성능 평가

Table 5. Performance evaluation of PM10 prediction model based on weather data and atmospheric environment data

Model	MAE	MSE	RMSE	MSLE
Linear Regression	0.118757	**0.01987**	**0.14096**	0.212779
SVM	0.192449	0.063255	0.251507	0.02754
Voting	**0.10128**	0.03054	0.174770	0.010701
Random Forest	0.118428	0.035334	0.187974	0.012370
Extra Random Tree	0.10400	0.030347	0.174204	0.010630
GMB	0.102808	0.029846	0.172762	**0.01039**
XGBoost	0.130304	0.038735	0.196813	0.014160

표 6. 기상 데이터와 대기 환경 데이터 기반 PM2.5 예측 모델 성능 평가

Table 6. Performance evaluation of PM2.5 prediction model based on weather data and atmospheric environment data

Model	MAE	MSE	RMSE	MSLE
Linear Regression	**0.0704**	**0.0130**	**0.1144**	**0.0048**
SVM	0.17775	0.04545	0.213273	0.02277
Voting	0.13752	0.03459	0.18587	0.01617
Random Forest	0.14924	0.03354	0.18316	0.01671
Extra Random Tree	0.14430	0.03050	0.17466	0.01386
GMB	0.15204	0.02937	0.17137	0.01431
XGBoost	0.15029	0.0319	0.17860	0.01692

[그림 6]과 [그림 7], [표 5]과 [표 6]은 기상청에서 제공하는 기상 데이터와 에어 코리아에서 제공하는 대기 환경 데이터를 가지고 2015년부터 2018년까지의 데이터로 학습한 후에 2019년을 예측한 PM10의 미세먼지와 PM2.5의 초미세먼지 농도의 예측 결과이다.

성능지표 MAE으로 보면 미세먼지는 보팅, 초미세먼지는 선형회귀가 가장 잘 예측하였다. 그 다음으로 미세먼지는 GBM, 엑스트라 램덤 트리 순으로 좋은 예측을 했다. 초미세먼지에서는 보팅, XGBoost 순이였다.

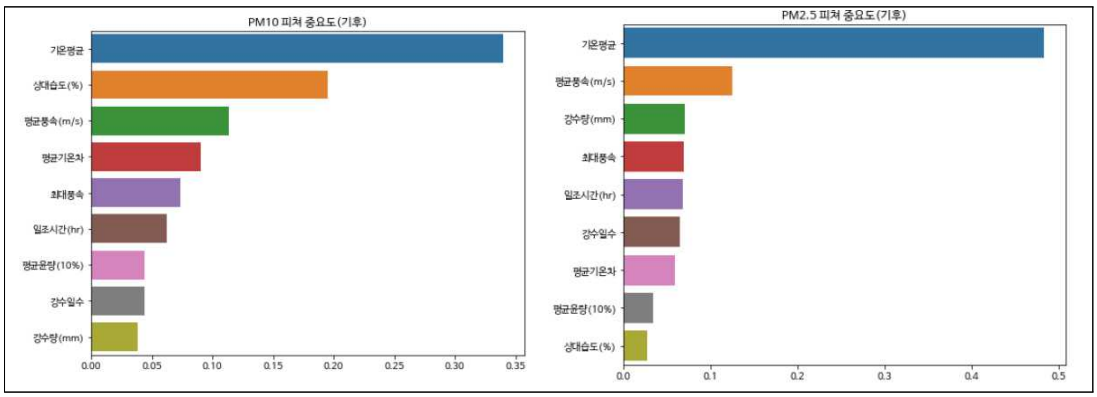


그림 8. 기상 데이터 기반 (초)미세먼지 피쳐 중요도

Fig. 8. (Ultra) air pollution feature importance based on Meteorological data

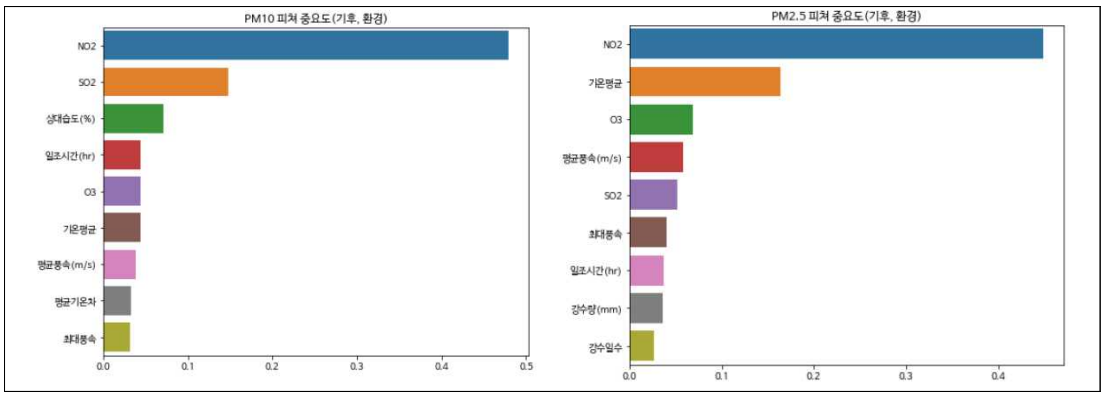


그림 9 기상 데이터와 대기 환경 데이터 기반 (초)미세 먼지 피쳐 중요도

Fig. 9. Feature importance of (ultra) air pollution based on meteorological data and atmospheric environment data

모델의 예측값과 실제 값의 차의 면적 합을 나타내는 MSE는 미세먼지인 경우에 선형회귀, GBM, XGBoost 순으로 잘 예측하였다. 초미세먼지 경우 선형회귀가 가장 잘 예측하였고 GBM, XGBoost 순으로 좋은 예측을 보여주고 있다.

[그림 8]은 기상 데이터를 기반으로 미세먼지와 초미세먼지에 영향을 미치는 중요한 피쳐를 나타내는 그림이다. [그림 8]에서 알수 있듯이 기상 데이터를 기반으로 보면 미세먼지나 초미세먼지에 가장 많이 영향을 주는 피쳐는 평균 기온이다. 미세먼지에서는 상대습도, 평균 풍속, 초미세먼지는 평균 풍속과 강수량이 평균 기온 다음으로 영양을 주는 피쳐임을 알수 있다.

[그림 9]는 기상 데이터와 대기 환경 데이터를 기반으로 미세먼지와 초미세먼지에 영향을 미치는 중요한 피쳐를 나타내는 그림이다. 기상 데이터와 대기 환경 데이터를 보면 미세먼지나 초미세먼지에 가장 많이 영향을 주는 피쳐는 NO2이다. 미세먼지에서는 SO2, 상대습도, 초미세먼지는 평균 기온과 O3 그 다음으로 영양을 주는 피쳐임을 알수 있다.

3. 결론

1급 발암물질로 분류되고 있는 미세먼지 확산은 코로나바이러스 확산의 큰 원인 중 하나이며 초미세먼지 증가에 따른 사망률도 증가하고 있다. 본 논문은 2015년부터 2019년까지 서울시의 대기 환경 데이터와 기상 데이터를 기반으로 미세먼지와 초미세먼지를 분석하고 미세먼지에 중요하게 영향을 미치는 요수를 파악하고, 미세먼지를 예측하였다.

본 논문에서 분석한 결과 2015년분터 2019년까지 서울시 미세먼지는 해마다 조금씩 줄어 들고 있었으며, 초미세먼지는 해마다 비슷한 농도로 분포되고 있었다. 또한, 서울시의 (초)미세먼지는 사계절에서 늦여름과 가을 중간인 8월에서 9월까지가 낮았으며, 11월부터 점점 증가하여 다음해 봄인 3월에는 (초)미세먼지가 가장 높았다.

본 논문에서 사용된 모델은 선형회귀, SVM와 앙상블 모델인 보팅, 램덤 포레스트, 엑스트라 램덤 트리, GBM, XGBoost으로 총 7개의 모델을 이용하여 미세먼지를 예측하였다. 결과적으로 기상 데이터만을 가지고 예측할 경우 PM10인 미세먼지인 경우 엑스트라 램덤트리와 GBM가 우수하게 예측하였고, PM2.5인 초미세먼지일 경우 XGBoost가 우수하게 예측하였다. 또한, 기상

데이터와 대기 환경 데이터를 같이 입력 데이터로 할 경우 미세먼지는 보팅과 선형회귀 모델이 우수하였고, 초미세먼지는 선형회귀가 다른 모델보다 우수하였다. 또한, 기상 데이터에서는 평균기온이 가장 중요하게 (초)미세먼지에 영향을 미치는 피쳐이고, 기상데이터와 대기 환경 데이터를 기반할때는 NO2가 (초)미세먼지에 가장 중요하게 영향을 미치는 피쳐었다.

REFERENCES

[1] M. Travaglio, Y. Yu, R. Popovic, L. Selley, N. Lea and L. M. Martins, "Links between air pollution and COVID-19 in England", Jour. of Environmental Pollution, Vol. 268, Jan. 2021

[2] X. Wu, R. C. Nethery, B.M. Sabath, D. Braun and F. Dominici, "Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study", Jour. of Science Advances, Vol 6, No. 45, Nov. 2020

[3] K. C. Lee and I. J. Hwang, " Characteristics of PM2.5 in Gyeongsan Using Statistical Analysis", Jour. of Korean Society for Atmospheric Environment, Vol. 31, No. 6, pp. 520-529, Dec. 2015

[4] S. W. Joun, J. Y. Choi and J. H. Bae, " Performance Comparison of Algorithms for the Prediction of Fine Dust Concentration" , Conf. of Korea Information Science Society, pp. 775-777, Dec. 2017

[5] Y. M. Seo and J. H. Yom, "Comparison of LSTM-based Fine Dust Concentration Prediction Method using Meteorology Data", Conf. of Korea Society of Surveying, Geodesy, Photogrammetry, and Cartography, pp. 117-120, Mar. 2019

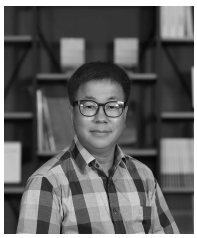
[6] S. H. Sung, S. J. Kim and M. H. Ryu, "A Comparative Study on the Performance of Machine Learning Models for the Prediction of Fine Dust: Focusing on Domestic and Overseas Factors", Jor. of Korea Society of Innovation, Vol. 15, Num. 4, pp. 339-357, Nov. 2020

[7] C. H. Hwang and K. W. Shin, "CNN-LSTM Combination Method for Improving Particular Matter Contamination (PM2.5) Prediction Accuracy", Journal of the Korea Institute of Information and Communication Engineering, Vol. 24, No. 1, pp. 57-64, Jan. 2020

[8] J. Y. Lee, M. J. Choi and J. K. Yang, "Ensemble Method for Predicting Particulate Matter and Odor Intensity", Jour. of the Society of Korea Industrial and Systems Engineering, Vol. 42, No.4, pp. 203-210, Dec. 2019

저자약력

박 홍 진(Hong-Jin Park) **[종신회원]**



- 1995년 8월 : 중앙대학교 일반대학교 컴퓨터공학과(공학석사)
- 2001년 8월 : 중앙대학교 일반대학교 컴퓨터공학과(공학박사)
- 2001년 9월 ~ 현재 : 상지대학교 컴퓨터공학과 교수

〈관심분야〉 시스템, 지능형 서버 관리, 빅데이터 분석