ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 48, No. 7, pp. 764-773, 2021. 7
https://doi.org/10.5626/JOK.2021.48.7.764

기계학습을 활용한 경기도 산업단지 미세먼지 예측

(Prediction of Fine Dust in Gyeonggi-do Industrial Complex using Machine Learning Methods)

원 동 준 † 김 선 겸 †† 김 영 훈 ††† 송 규 원 ††††
(Dong-Jun Won) (Sun-Kyum Kim) (Yeonghun Kim) (Gyuwon Song)

요 약 최근 미세먼지의 다양한 예측 모델들을 통한 연구가 이루어지고 있지만 현재 PM_{10} 농도 예측에 치중되어 있어 $PM_{2.5}$ 농도를 예측할 수 있는 모델 개발이 필요한 상황이다. 본 논문은 최근 약 2년간의 반월시화국가산업단지의 대기질, 기상, 교통 데이터를 수집하여 미세먼지($PM_{2.5}$)와 미세먼지(PM_{10}), 이산화황(SO_2), 이산화질소(NO_2), 일산화탄소(CO), 오존(O_3), 온도, 습도, 풍향, 풍속, 강수량, 도로 구간별 차량속도 변수간의 상관관계 분석 및 회귀분석을 통해 변수의 유의성을 파악하고, 산업단지의 시간대별 $PM_{2.5}$를 예측하는 데 활용하였다. 인공지능 기반의 Random Forest, XGBoost, LightGBM, Deep neural network과 Voting 모델을 통해 산업단지의 시간별 PM2.5 농도를 예측하고, RMSE를 기준으로 비교분석을 진행하였다. 예측 결과 RMSE는 각각 6.27, 6.41, 6.22, 6.64, 6.12로 각 모델 모두 에어코리아에서 예측하는 모델의 10.77에 비해 매우 높은 성능을 보여주었다.

키워드: 미세먼지, $PM_{2.5}$, 반월시화산업단지, 예측, 인공지능, 딥러닝

Abstract Recently, research on fine dust has been conducted through various prediction techniques. However, currently the research focused on PM_{10} concentration prediction, and thus it is necessary to develop a model capable of predicting $PM_{2.5}$ concentration. In this paper, we have collected air quality, weather, and traffic of the Banwol Shihwa National Industrial Complex in the recent two years. The significance of the variable been identified through correlation analysis and regression analysis among $PM_{2.5}$ and PM_{10}, SO_2, NO_2, CO, O_3, temperature, humidity, wind direction, wind speed, precipitation, road section vehicle speed for each vehicle. Next, the data has been used to predict $PM_{2.5}$ concentration based on time in the industrial complex. Through the artificial intelligence techniques, Random Forest, XGBoost, LightGBM, Deep neural network and Voting models, $PM_{2.5}$ concentration industrial complexes been predicted on an hourly basis, and comparative analysis been conducted based on RMSE. As a result of prediction, RMSE was 6.27, 6.41, 6.22, 6.64, and 6.12, respectively, and each technique showed very high performance compared to 10.77 of the technique predicted by Air Korea.

Keywords: fine dust, $PM_{2.5}$, Banwol Shihwa National Industrial Complex, prediction, artificial intelligence, deep learning

· 본 연구는 차세대융합기술연구원(AICT-2021-0001)의 지원을 받아 수행되었습니다.

† 비 회 원 : 현대트랜시스 매니저
whon1090@hyundai-transys.com
†† 정 회 원 : 한국건설기술연구원 미래스마트건설연구본부 수석연구원
(KICT)
sunkyumkim@kict.re.kr
(Corresponding author임)
††† 비 회 원 : 차세대융합기술연구원 기술기획팀 선임연구원
toya84@snu.ac.kr
†††† 정 회 원 : 차세대융합기술연구원 기술기획팀 선임연구원(AICT)
gyuwon.song@snu.ac.kr
(Corresponding author임)

논문접수 : 2020년 6월 18일
(Received 18 June 2020)
논문수정 : 2021년 4월 11일
(Revised 11 April 2021)
심사완료 : 2021년 4월 14일
(Accepted 14 April 2021)

정보과학회논문지 제48권 제7호(2021. 7)

1. 서 론

최근 대기오염물질에 대한 국민적 관심과 우려가 증가하고 있다. 특히 초미세먼지($PM_{2.5}$)는 입자의 크기가 10㎛ 이하인 미세먼지(PM_{10})보다 훨씬 작은 2.5㎛ 이하인 먼지를 말하며, 폐·호흡기·심혈관계 질환 등 호흡을 통한 인체 급성 피해와 조기 사망을 증가시키는 만성 피해를 주는 것으로 보고되고 있다[1]. 이에 대한 대응으로 세계보건기구(WHO)는 $PM_{2.5}$를 1급 발암물질로 지정하였으며[2], 국가적으로도 미세먼지와 초미세먼지를 새로운 환경정책의 전환요인으로 인식하고, 이에 대한 위해성 논의를 확산하고 있다. 미세먼지 발생 저감을 위하여 국가 및 지자체 단위에서 적극적인 대책을 수립하고 있으며, 미세먼지의 예측 또한 그 필요성이 높아짐에 따라 다양한 예측 모델들을 통한 연구가 이루어지고 있다.

하지만 현재 활용되고 있는 미세먼지 예측 모델들은 예측력이 높지 않은 수준이며, PM_{10} 농도 예측에 치중되기 때문에, 보다 신속하고 정확하게 $PM_{2.5}$ 또한 예측할 수 있는 모델 개발이 필요한 상황이다. 최근에는 일별 미세먼지 발생원과 미세먼지 영향에 대한 연구가 중요하게 인식되고 있으나, 시간 기반의 미세먼지 예측 연구는 미흡하다.

경기도는 2018년 기준 $PM_{2.5}$ 연평균 농도 25$\mu g/m^3$로 대기환경기준(15$\mu g/m^3$)을 초과하고, WHO의 권고 기준인 10$\mu g/m^3$보다 두 배 이상 높은 수치를 기록하였다[1]. 미세먼지 배출의 경우 사업장이 전국 기준 38%로 배출량 1위이며, 경기도 기준으로도 18%로 2위를 차지할 정도로 사업장이 미세먼지 생성 원인의 매우 큰 비중을 차지한다[1].

경기도의 대표적인 사업장으로서 반월시화국가산업단지는 수도권 인구 분산 정책의 일환으로 서울과 경기도 각지에 산재한 중소기업, 공해업체의 공장들을 안산시 성곡동, 원시동(반월) 및 시흥시 정왕동(시화) 일대에 이전, 계열화하여 육성할 목적으로 조성된 산업단지이다. 반월시화국가산업단지는 서쪽으로는 서해와 동쪽으로는 시흥시와 안산시를 비롯하여 의왕시, 군포시, 수원시가 위치하고 있다. 이러한 반월시화산업단지는 사업장의 미세먼지를 분석하는 데 매우 적합한 지역적·환경적인 특성을 가지고 있다.

본 논문은 최근 약 2년간의 국가측정망(에어코리아), 기상관측센터, 경기도교통정보센터로부터 반월시화산업단지의 대기질, 기상, 교통 데이터를 수집하여 이를 통해 반월시화국가산업단지의 $PM_{2.5}$와 PM_{10}, 이산화황(SO_2), 이산화질소(NO_2), 일산화탄소(CO), 오존(O_3), 온도, 습도, 풍향, 풍속, 강수량, 도로 구간별 차량 속도 변수간 유의성을 파악하여, 해당 결과를 토대로 시간대별 $PM_{2.5}$를 예측한다. 이를 위해 산업단지의 계절별 평균 $PM_{2.5}$ 농도를 통해 계절 변수가 미세먼지에 주는 영향을 확인하고, $PM_{2.5}$와 대기질 및 기상, 교통 데이터와의 회귀분석을 통해 영향도를 분석하며, 피어슨 상관계수를 통해 해당 변수 간의 상관관계를 파악한다. 또한 예측력을 높이기 위해 해당 변수들을 활용하여, 인공지능 알고리즘인 머신러닝 기반 Random Forest, LightGBM, XGBoost와 딥러닝 기반 Deep Neural Network(DNN) 모델을 통해 반월시화국가산업단지의 시간별 $PM_{2.5}$ 농도를 예측하고, 에어코리아에서 제공하는 24시간예측이동평균과 비교분석한다.

본 논문의 2장에서는 관련 연구와 반월시화산업단지의 실제 $PM_{2.5}$ 농도를 확인한다. 3장에서는 데이터 전처리 및 지역별 미세먼지 분석 및 상관관계 분석, 회귀분석을 통해 미세먼지 원인 변수 데이터를 분석하고, 4장에서는 다양한 인공지능 기반 모델들을 통해 반월시화국가산업단지의 시간에 따른 $PM_{2.5}$ 농도를 예측한다. 마지막으로 5장의 결론을 통해 본 논문을 마무리한다.

2. 관련 연구 및 배경 지식

본 논문은 개별 도시 및 어린이집 주변이나 교통흐름에 따른 도로 주변 등의 특정지역의 미세먼지 농도 예측 관련 연구를 설명하고, 본 논문의 분석 및 예측을 위해 반월시화산업단지의 계절별 실제 미세 농도를 알아본다.

2.1 도시 미세먼지 연구

구윤서 등[3]은 2006년 1월 1일부터 2009년 5월 31일까지 신경망 모형, 회귀모형, 의사결정모형을 각각 수행하여 빈도수가 높은 지수를 이용해 서울남동지역을 분석하였다. 당일예보모형과, 내일예보모형 두 가지로 나누어 분석하였고, 각각의 모형의 예보 지수중에서 최빈값을 최종 예보 지수로 정하였다. 분석 결과 당일예보모형 80.8%, 내일예보모형 72.4%의 예측정확도를 기록하였다.

차진욱 등[4]은 에어코리아에서 제공하는 대기질 측정자료, 기상청에서 제공하는 기상 데이터를 바탕으로 회귀분석, 상관관계 분석 및 ANN, KNN을 알고리즘을 응용하여 미세먼지 농도 수치를 예측하였다. ANN 알고리즘을 통한 예측 값을 KNN 알고리즘으로 분류하였고 ANN과 KNN을 사용한 단일 예측 모델의 정확도 62.27%, 58.41%보다 좋아진 83.4%를 기록하였다.

권준현 등[5]은 2011년 1월부터 2014년 6월 서울 종로구 대기질, 기상 데이터를 바탕으로 τ값의 변화에 따른 분위수 부스팅 예측 분석을 진행하였다. 기존 의사결정나무 모형과 비교하여 분위수 부스팅의 예측 정확도는 2.6% 증가하였다.

성민기[6]는 2014년부터 2017년까지 대기질, 기상, 중국 7개 도시의 미세먼지 농도를 기계 학습을 이용하여

예측을 실시하였다. Ridge, Lasso, ElasticNet을 이용한 정규화 선형 회귀 모델에 CNN을 이용한 모델을 생성하여, 단순 정규화 선형회귀모델과 비교하였다. Ridge모형을 통해 도출된 예측 값과 실제 값의 차이를 Average Pooling, Flatten Layer를 사용한 CNN 모델로 학습하여 오차를 보정하였고, 단순 정규화 선형회귀모델에 비해 3%의 성능 개선을 이루었다.

2.2 특정지역 미세먼지 연구

신익희 등[7]은 지능형 도시에서 미세먼지 예측을 위한 심층 학습 기법 평가를 하였다. 어린이집처럼 데이터의 특성이 많은 경우 합성곱 신경망으로 입력 데이터에서 특징을 추출한 후, 특징을 LSTM 신경망으로 학습하는 CNN-LSTM 신경망이 적합하며, 스마트 가로등 같이 특성이 적을 경우 CNN-LSTM 신경망은 입력 데이터에서 특징을 추출하기 어렵고, LSTM신경망은 예측 정확도가 떨어져 다층 신경망이 미세먼지 예측에 적합하다는 결과를 얻었다.

이홍석 등[8]은 도심지 교통흐름 및 미세먼지 예측을 위한 딥러닝 LSTM 프레임워크를 연구하였고, 미세먼지 미세먼지와 교통흐름 사이의 상관관계는 멀티-채널 혹은 3차원 ConvLSTM을 적용할 수 있는 것을 확인하였다.

2.3 산업단지 미세먼지 농도

그림 1은 반월시화산업단지의 계절별 미세먼지 평균 농도를 보여준다.

반월시화산업단지의 미세먼지 농도는 봄과 겨울에 높으며 여름과 가을에는 낮은 수치를 나타낸다. 봄에는 이동성 저기압과 건조한 지표면의 영향으로 고농도의 미세먼지가 발생할 가능성이 크기 때문에 비교적 높은 수치를 보인다. 여름철에는 비로 인해 대기오염물질이 빗물에 씻겨 내려감으로써 대기가 깨끗해져, 평균 미세먼지 농도는 낮아진다. 가을에는 원활한 대기의 순환이 이루어지기 때문에 상대적으로 미세먼지 농도가 낮아진다. 겨울에는 일반적으로 미세먼지 농도가 높다고 생각하지만, 난방을 위한 연료사용 등으로 인해 평균 미세먼지 농도는 높은 수치를 보인다. 결론적으로 해당 결과를 반영하여 미세먼지 예측 모델에 계절변수를 활용한다.

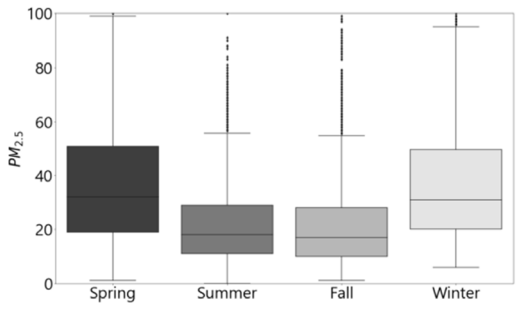


그림 1 계절별 미세먼지 농도

Fig. 1 Average Concentration of Fine Dust Based on the Season

3. 데이터 분석

3.1 수집 및 활용 데이터

본 논문에서는 데이터 분석 및 예측을 위해 2017년 12월부터 2019년 10월까지 약 23개월간의 국가측정망(에어코리아)로부터 얻은 반월시화국가산업단지의 시간대별 대기질 확정 데이터를 활용하였다[9]. 동기간의 기상 데이터는 대기질 국가측정망과 인접한 위치에 있는 데이터를 수집하였고[10], 구간별 차량속도 데이터는 경기도교통정보센터에 요청한 반월시화국가산업단지의 시간별 통행속도 내부 자료를 활용하였다[11].

수집한 데이터의 각 측정망은 시의 중심이 아닌 건물 옥상이나 산 등에 인적이 드문 곳에 위치하고 있으며, 측정되는 장소 기준으로 활용 데이터를 시를 대표하는 데이터로 활용한다. 표 1은 활용 데이터를 보여준다.

활용 데이터 중 $PM_{2.5}$와 PM_{10}, SO_2, NO_2, CO, O_3, 온도, 습도, 풍속, 풍향은 기록된 값으로 활용하였다. 강수량의 경우 비가 오지 않은 날과 미미한 날도 많았기 때문에 강수의 유무로 값을 이용하였다. 교통 데이터의 경우, 반월시화국가산업단지의 도로 데이터를 활용하였으며, 도로교통법 기준에 따라 원활, 보통, 정체 3단계에 따라 가변수 처리하여 각각 0, 1, 2 값으로 범주화하였다. 또한 계절의 특성을 적용하기 위해 봄, 여름, 가을, 겨울을 가변수(Dummy Variable) 처리하여 각각 0, 1, 2, 3으로 범주화하여 활용하였다

3.2 데이터 전처리

에어코리아에서 제공하는 데이터와 기상, 통행속도의 데이터 형식이 모두 달라 해당 데이터 형식을 모두 통

표 1 활용 데이터

Table 1 Used Data

Classification	Feature	Measurement period
Air Quality	$PM_{2.5}$	1 hour (23 months)
	PM_{10}	
	SO_2	
	NO_2	
	CO	
	O_3	
Weather	Temperature	
	Humidity	
	Wind Speed	
	Wind Direction	
	Precipitation	
Traffic Speed	Vehicle speed by section	

일한 후 데이터 정규화를 진행하였다.

활용 데이터 중 대기질, 기상, 교통 데이터는 수일의 결측치가 있었으며, 이를 보완하기 위해 결측치가 존재하는 시간의 이전 시간 값으로 대체하여 사용하였다. 또한, 연속형 변수인 PM_{10}, SO_2, NO_2, CO, O_3, 온도, 습도, 풍속, 풍향 변수는 정확한 예측을 위해 식 (1)의 min-max scale 방식을 이용해 0부터 1까지 정규화 하였다.

$$X = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

3.3 변수간의 관계 분석

본 논문은 반월시화국가산업단지의 $PM_{2.5}$와 대기질, 기상 교통 요인의 인과관계를 분석하기 위해 선형회귀분석(Linear Regression Analysis)을 이용한다. 선형회귀분석은 종속 변수 y와 한 개 이상의 독립 변수 x와의 선형 관계를 나타내는 기법이다. 이를 통해 R-Squared 값과 P-value, VIF (Variance inflation factor)를 도출할 수 있다. R-Squared 값은 $x = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$ 로 나타내며, $\bar{y}$는 모든 y값의 평균, y_i는 측정된 값 $\hat{y}_i$는 예측값을 의미한다. R-squared는 0과 1 사이의 값으로 나타나는데, 0에 가까울수록 설명력이 낮고, 1에 가까울수록 높다고 해석할 수 있다. P-value는 유의확률이며, 가설이 참이라고 가정할 때 얻은 결과보다 극단적인 결과가 실제로 관측될 확률이며, P-value가 0.05 미만일 때 해당 변수는 유의하다고 할 수 있다. VIF(Variance Inflation Factor)는 분산팽창요인이라 불리며, 회귀 왜곡을 피하기 위해 사용한다. R-Squared값이 높아 회귀 식의 설명력은 높지만, 개별 독립 변수의 P-value 값이 커서 개별 인자들이 유의하지 않는 경우가 있다. 이런 경우 독립변수들 간에 높은 상관관계가 있다고 의심되며 이때 VIF 값을 이용한다. VIF는 식 (2)로 계산하며, 이 값이 10을 넘긴다면 다중공선성이 있다고 판단한다.

$$VIF_i = \frac{1}{1 - R_i^2} \quad (2)$$

본 논문은 산업단지의 $PM_{2.5}$와 나머지 대기질, 기상, 교통 데이터와의 관계를 분석하기 위해 독립 변수 x에는 대기질, 기상, 교통 변수와 종속 변수 y값에는 $PM_{2.5}$ 값을 적용하여 다중회귀분석을 통해 R-squared, P-value 및 VIF 값을 도출하였다. 회귀분석 결과의 R-squared 값은 0.804로 종속 변수 $PM_{2.5}$와의 다른 독립 변수와 밀접한 관계가 있음을 보여주었다.

표 2는 각각의 독립변수간 P-value와 VIF 값을 보여준다. P-value의 경우 모든 변수에 대하여 0.05 미만이기 때문에 모두 유의하다. 다중공선성의 경우 VIF 값이 모두 4 미만이므로 변수 간 관계없이 독립적이라 할 수 있다. 그러므로 모든 변수는 독립 변수 간 상관성을 유지하며 유의미하다. 이것은 반월시화국가산업단지의 $PM_{2.5}$와 각각의 변수는 관련이 있다는 것을 나타내며, 해당 변수 모두 예측 모델에 활용하였다.

표 2 P-value 및 VIF
Table 2 P-value and VIF

Features	P-value	VIF
PM_{10}	0.0	2.209040
SO_2	0.0	1.059211
NO_2	0.0	3.275532
CO	0.0	3.098946
O_3	0.0	2.214173
Temperature	0.007	1.655853
Humidity	0.0	1.680792
Wind Speed	0.0	1.757550
Wind Direction	0.0	1.324629
Precipitation	0.001	1.052670
Weather	0.0	1.123635
Traffic	0.01	1.166763

표 3 PM2.5 및 PM10 Pearson 상관계수
Table 3 PM2.5 & PM10 Pearson Correlation Coefficient

Features	$PM_{2.5}$	PM_{10}
$PM_{2.5}$	1	0.87
PM_{10}	0.87	1
SO_2	0.19	0.15
NO_2	0.56	0.5
CO	0.69	0.64
O_3	-0.05	0.0
Temperature	-0.24	-0.25
Humidity	0.11	-0.01
Wind Speed	0.0	-0.14
Wind Direction	-0.23	0.03

표 3은 변수간의 상관관계를 피어슨 상관계수(Pearson's coefficient)[12]를 통해 산업단지의 $PM_{2.5}$, PM_{10}과 각 변수들 간 관련성을 나타낸다. $PM_{2.5}$는 PM_{10}과 0.87로 매우 강한 양의 상관관계, CO와 0.69로 강한 양의 상관관계, NO_2와 0.56으로 양의 상관관계, SO_2와 0.19로 양의 약한 양의 상관관계를 가지며, 온도, 풍향과 각각 -0.24, -023으로 약한 음의 상관관계를 가진다. PM_{10}의 경우 CO와 0.64로 강한 양의 상관관계, NO_2와 0.5로 양의 상관관계, SO_2와 0.15로 약한 양의 상관관계를 가지며 온도, 풍속과 각각 -0.25, -0.14로 약한 상관관계를 가진다.

4. 예측

4.1 예측 모델

본 논문에서 실험을 위해 PM_{10} 예측 논문에서 자주 사용되는 머신러닝 Random Forest, XGBoost를 활용

하였다. 또한 Goss, leaf-wise 방식을 통해 속도와 성능 보완을 이룬 LightGBM과 입력 변수들 간 비선형 조합이 가능하고 다변량 변수를 dropout방식을 통해 변수를 추출하여 우수한 성능을 보이는 딥러닝 기반 DNN 총 네 가지 모델을 사용하였다. 마지막으로 네 알고리즘 중 성능이 우수한 두 모델을 조합한 Voting 모델을 통해 반월시화국가산업단지의 시간별 $PM_{2.5}$ 농도 예측을 진행하였다.

4.1.1 Random Forest

Random Forest [13]는 다수의 의사결정 나무(Decision Tree)를 결합하여 반응변수를 예측함으로써 의사결정 나무의 불안정성과 낮은 예측력을 보완한다. 하나의 데이터의 bootstrap을 이용하여 다양한 데이터 모델링을 하고 이를 집계(Aggregating)한 후에 최종모형을 만드는 bagging 방식이다.

4.1.2 XGBoost(eXtreme Gradient Boosting)

XGBoost [14]는 여러 개의 의사결정 나무를 조합하는 알고리즘이다. XGBoost는 과적합, 속도 문제를 해소하고 병렬 연산이 가능하도록 구조를 변형하여 기존 GBM (Gradient Boosting Model)의 문제점을 해결했다. XGBoost는 의사결정 나무의 분화 적정성을 평가하는 목적함수에 페널티 항을 추가하는 방식으로 이루어지며, 손실함수를 최소화하면서 과적합을 방지한다.

4.1.3 LightGBM(Light Gradient Boosting Model)

LightGBM [15]은 Gradient boosting 모델의 특징인 모든 노드를 분리하는 방법을 leaf-wise 방식을 통해 비대칭적인 트리를 생성하여 보완하였다. XGBoost와 같은 GBDT(Gradient Boosting Decision Tree) 기법을 기반으로 GOSS(Gradient-Based One-Side Sampling) 등을 통해 속도를 보완하였고, GBDT의 각 노드에서 분기점이 나뉠 때, 잘 맞는 노드를 기준으로 분리하며, 잘 맞지 않는 노드는 분기점으로 선택되지 않는 장점을 지닌다.

4.1.4 DNN(Deep Neural Network)

DNN [16]은 입력 층과 출력 층 사이에 여러 개의 은닉층들로 이뤄진 ANN(Artificial Neural Network)이다. DNN은 역전파(Backpropagation) 알고리즘을 사용하여 하위층에서 상위층으로 전파되면서 오차를 계산한다. 이를 바탕으로 목표 값에서 입력 값으로 역전파하여 각 층에 연결된 가중치를 수정함으로써 출력 값과 목표 값의 오차를 줄여 모델의 성능을 향상시킨다. 본 논문에서는 은닉층의 활성함수로는 ReLU(Rectified Linear Unit)를 사용하여 예측을 진행하였다.

4.1.5 Voting 모델

Voting 모델[17]은 정확도가 높은 강한 모델을 하나 사용하는 것보다, 정확도가 낮은 약한 모델을 여러 개 조합 하는 방식이 정확도가 높다는 방법에 기반한 모델이다. 학습된 여러 기계학습 알고리즘 모델을 투표를 통해 최종 예측 결과를 결정하여 새로운 모델을 만드는 방식으로 진행한다. 이를 통해 단일 모형을 사용할 때보다, 성능 분산이 감소하여 성능이 향상되고, 모형의 과적합을 방지한다. 본 논문에서는 Root-mean-square deviation (RMSE)와 정확도 값을 기반으로 두 모델을 선정하여 Python Scikit-learn [18]의 Default 값인 Hard Voting 방식을 통해 모델을 결합하였다.

4.1.6 에어코리아 모델

앞서 설명한 다섯 가지의 예측모델과 비교를 위해 국가측정망인 에어코리아에서 사용하는 24시간예측이동평균 방법을 활용하였다. 에어코리아의 예측 식은 식 (3), (4), (5)와 같다[19].

$$C_{24E} = [C_{12} \times 12 + C_4 \times 12] \div 24 \quad (3)$$

$$C_4 = (C_{ai} + C_{a(i-1)} + C_{a(i-2)} + C_{a(i-3)}) \div 4 \quad (4)$$

C_{12}는 과거 12시간의 평균이며, C_4는 과거 4시간의 평균이다. C_i는 기준시간 측정농도, i는 기준시간이다.

$$C_i < M \rightarrow C_{ai} = C_i \quad (5)$$

여기서 M은 에어코리아에서 제공하는 기준인, PM_{10}의 경우 70$\mu g/m^3$, $PM_{2.5}$의 경우 30$\mu g/m^3$을 적용한다.

4.2 예측 방법

본 논문에서는 미세먼지 예측을 위해 표 2의 $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3, 온도, 습도, 풍향, 풍속, 강수량, 계절, 도로 구간별 차량속도 13가지의 변수를 모두 활용하여 2017년 12월 1일부터 2019년 10월 22일까지 시간별 데이터 총 15,917개를 사용하였다.

그림 2는 실험에서 사용한 데이터 예측 방법을 나타낸다. 본 논문의 예측 모델은 시간별로 수집한 대기질, 기상, 교통 데이터를 8:2로 분할하여 12,731개를 훈련용 데이터로 모델 학습에 사용하고 3,183개를 테스트 데이터로 사용하여 예측을 진행한다. 평가 지표로는 인공지능 모델 예측에서 가장 많이 사용되는 RMSE를 활용하

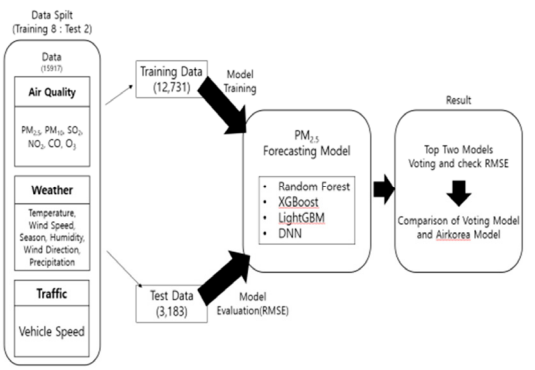


그림 2 데이터 예측 방법

Fig. 2 Forecasting Data Method

였다. 훈련용 데이터 학습을 위해 인공지능 모델 Random Forest, XGBoost, LightGBM, DNN 네 모델의 하이퍼파라미터 최적화를 진행하고, 학습한 모델을 바탕으로 테스트 데이터에 $PM_{2.5}$ 농도 예측을 진행하여 RMSE 값을 측정하였다. 이후 RMSE 수치가 좋은 두 모델을 결합한 Voting 모델을 생성하여 예측을 진행하였고, 에어코리아 예측 모델의 RMSE와 비교분석을 진행하였다.

예측을 위해 사용한 평가지표 RMSE는 평균제곱근편차로, 예측 값과 실제 값의 차이에서 제곱한 값들을 더하고 전체 개수로 나누어 루트를 씌운 값이며, 크기에 의존하는 에러(Scale-dependent Errors)이다. 예측하려는 값의 크기가 크면 더불어 같이 증가하고, 예측하고자 하는 값의 크기가 작으면 같이 감소한다. 식 (6)은 RMSE의 수식이며, y_i는 실제 값, $\hat{y_i}$,는 예측 값을 나타낸다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \quad (6)$$

4.3 실험 최적화

실험을 위해 머신러닝 각 모델의 하이퍼파라미터 최적화는 모두 Random Search를 이용하여 진행하였다. Random Search는 파라미터(parameter)의 범위를 설정한 후, 파라미터의 설정지점부터 샘플링을 통해 최적의 파라미터 값을 선정하는 방식이다. 각각의 매개변수 설정 시 균일분포로 샘플링하여 파라미터 결과가 개선되는 경우에만 이전 파라미터 값을 갱신하는 모델이다. 이러한 Random search를 통해 Random Forest, XGBoost, LightGBM, DNN의 파라미터를 최적화하였으며, 각 모델에 대한 최적화 값은 표 4-7로 나타내었다.

Random Forest는 각각의 의사결정 나무의 최대 깊이를 설정하는 max_depth, 총 의사결정 나무 개수를 설정하는 n_estimators, 나무의 내부 노드 분할에 필요한 최소 샘플 수를 지정하는 min_samples_split, 노드에 개별 잎의 최소 샘플 수를 지정하는 min_samples_leaf, 전체 특성의 수를 설정하는 max_features를 이용하였다. XGBoost는 max_depth, n_estimators, L2 정규화 reg_lambda, L1 정규화 reg_alpha 네 가지 파라미터를 사용하였다. LightGBM은 max depth, n_estimators, reg_lambda, reg_alpha와 학습률을 설정하는 learning rate, 훈련 데이터의 사용 비율을 설정하는 bagging_fraction을 활용하였다. DNN은 전체 딥러닝 횟수인 epoch, 한번에 가지고 오는 크기를 나타내는 batch size, 레이어값인 layer_dense, 학습률을 설정하는 learning rate를 이용하여 모델의 파라미터 최적화를 진행하였다.

표 4 Random Forest 하이퍼파라미터
Table 4 Random Forest HyperParameter

Parameter	Optimization value
max_depth	19
n_estimators	3000
min_samples_split	3
min_samples_leaf	1
max_feature	4

표 5 XGBoost 하이퍼파라미터
Table 5 XGBoost HyperParameter

Parameter	Optimization value
max_depth	5
n_estimators	670
reg_lambda	0.21482758620689657
reg_alpha	0.41965517241379313

표 6 LightGBM 하이퍼파라미터
Table 6 LightGBM HyperParameter

Parameter	Optimization value
max_depth	13
n_estimators	2467
reg_lambda	0.01
reg_alpha	0.8350000000000001
learning_rate	0.5421250000000001
bagging_fraction	0.41725

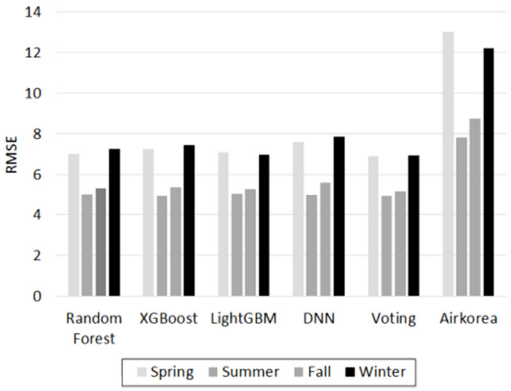
표 7 DNN 하이퍼파라미터
Table 7 DNN HyperParameter

Parameter	Optimization value
batch_size	256
epoch	50
layers_dense	64
optimizer	adam(default value)
learning_rate	0.001

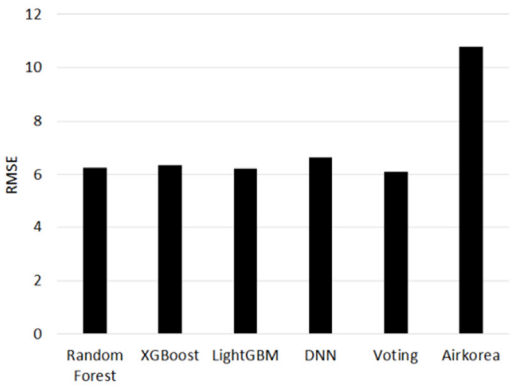
4.4 실험 결과

반월시화국가산업단지의 $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3, 온도, 습도, 평균 풍속, 풍향, 강수량, 계절변수, 교통변수를 활용하여 Random Forest, XGBoost, LightGBM, DNN 네 가지 모델을 통해 $PM_{2.5}$ 농도를 예측하고 RMSE를 통해 정확도를 평가하였다. 위의 네 모델 중 RMSE 값을 기반으로 성능이 가장 좋은 머신러닝 Bagging 기법인 Random Forest와 Boosting 기법 LightGBM을 VotingRegressor를 통해 결합하여 새로운 Voting 모델을 만들었다. Random Forest, XGBoost, LightGBM, DNN, Voting, 에어코리아에서 제공하는 24시간예측이동평균모델을 RMSE를 통해 비교하였고, 모두 소수점 둘째 자리까지 반올림하여 그림 3에서 (a)는 계절별로, (b)는 전체 평균으로 나타내었다.

(a)와 (b) 모두 Voting 모델이 성능이 가장 우수하였다. (a)는 봄과 겨울 모두 미세먼지가 여름과 가을에 비해 상대적으로 높기 때문에 RMSE값도 다소 높게 측정



(a) Forecasting Models for RMSE by Seasons



(b) Forecasting Models for RMSE by Average

그림 3 예측 모델의 RMSE

Fig. 3 Forecasting Models for RMSE

되었다. (b)의 경우 Random Forest, XGBoost, LightGBM, DNN, Voting, 에어코리아 모델의 RMSE는 각각 6.27, 6.41, 6.22, 6.64, 6.12, 10.77로 Voting 모델이 가장 높은 성능을, 다음으로 LightGBM이 높은 성능을 나타냈으며 DNN이 에어코리아 모델을 제외하고 상대적으로 가장 낮은 성능을 보여주었다. DNN의 경우, 학습하는 데이터의 수가 많지 않아 상대적으로 낮은 RMSE 값을 보였다. LightGBM의 경우는 XGBoost의 level-wise 방식을 통해 트리를 분할하는 방식보다, LightGBM의 leaf-wise 방식을 통해 각 잎의 노드에서 분기점이 나뉠 때, 잘 맞는 노드를 분리하여 나눴기 때문에 더 좋은 예측 값을 가졌다고 볼 수 있다. Voting 모델은 RMSE를 기준으로 가장 성능이 우수한 Random Forest와 LightGBM 두 모델을 활용하였고, 결과적으로 Voting과 에어코리아 예측 모델 대비 약 56.82% 향상된 결과를 기록하였다. Voting 모델은 앞선 단일 기법기반의 모델들보다 다소 높은 성능을 보여주었고, 이는 서로 다른 방식의 두 앙상블 모델인 Random Forest의 Bagging과 LightGBM의 Boosting을 Voting하여, 각 모델의 예측 값 평균을 통해 성능 분산을 감소시켜 예측 값과 실제 값의 오차를 줄인 결과이다.

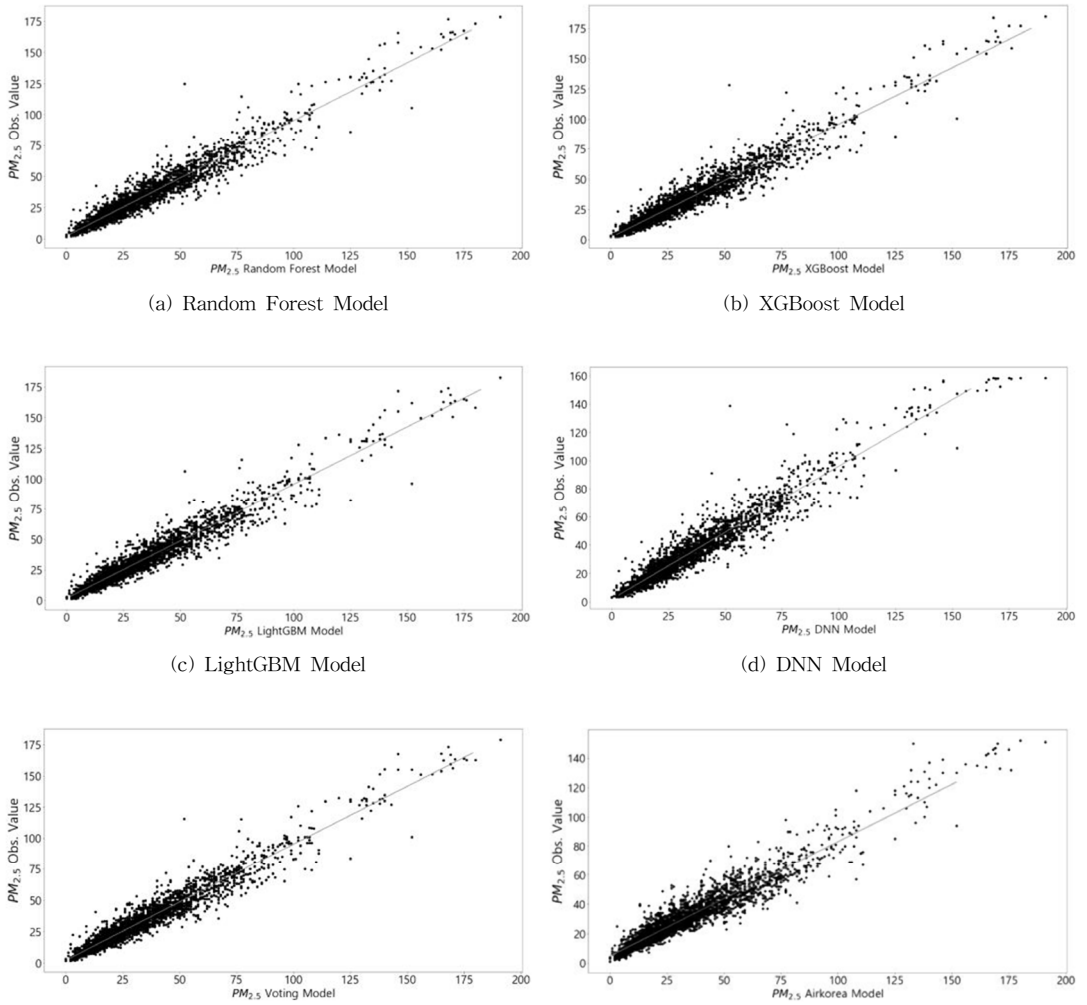
보다 세부적으로 확인하기 위해 그림 4에서는 $PM_{2.5}$ 농도, 그림 5에서는 시간에 따른 실제 값과의 비교 결과를 나타내었다. 그림 4는 가로측은 예측 모델이며 세로는 실제 값이고 중앙 선에 값이 집중될수록 정확하다는 것을 의미한다. 전반적으로 Voting 모델이 저·고농도 모두에서 잘 모사되고 있으며, 분산정도가 크지 않아 이를 통해 가장 성능 높은 것을 알 수 있었다.

그림 5는 가로는 시간이며 진한 색의 그래프는 예측 모델, 연한 색의 그래프는 실제 값이고, 이 그래프의 경우 두 개의 그래프가 많이 겹칠수록 정확한 것을 의미한다. 앞선 결과와 마찬가지로 전반적으로 Voting 모델이 예측 모델의 값과 실제값의 겹침이 가장 많았으며, 가장 $PM_{2.5}$ 농도가 높았던 2018년의 2월과 4월, 2019년의 2월, 2019년 4월경에서 명확히 보여주었다. 결과적으로 Voting 모델이 가장 우수한 성능을 보여준다고 할 수 있다.

5. 결 론

본 논문에서는 최근 약 2년간의 대기질, 기상, 교통 데이터를 통해 반월시화국가산업단지의 $PM_{2.5}$ 생성에 유의미한 영향을 미치는 변수가 무엇인지 상관관계 분석과 회귀분석을 통해 확인하였다. 먼저 계절별 미세먼지 평균 농도를 통해 계절에 따른 $PM_{2.5}$의 변화를 통해 계절 변수가 주는 영향을 확인하였다. 이후 회귀분석을 진행하여 R-Squared, P-value, VIF를 통해 변수 간 관계를 파악하였다. 결정계수인 R-Squared 값은 0.804로 해당 변수들이 $PM_{2.5}$ 농도와 유의미한 관계를 나타내었고, 모든 변수의 P-value가 0.05 미만으로 유의하고, VIF 값이 10 미만으로 변수간 관계없이 독립적임을 보였다. 결과적으로 PM_{10}, SO_2, NO_2, CO, O_3의 대기질과 온도, 습도, 풍향, 풍속, 계절변수, 강수량의 기상, 반월시화국가산업단지의 도로 구간별 차량속도 교통 데이터는 틀림없이 $PM_{2.5}$에 영향을 주는 변수라고 할 수 있다.

이후 $PM_{2.5}$에 영향을 미치는 대기질, 기상, 교통 데이터를 활용하여 2017년 12월 1일부터 2019년 10월 22일까지 시간별 반월시화국가산업단지의 $PM_{2.5}$ 농도를 예측하였다. 인공지능 기반의 Random Forest, XGBoost, LightGBM, DNN을 통해 반월시화국가산업단지의 시간별 $PM_{2.5}$ 농도를 RMSE 기준으로 예측하고, 비교분석을 진행하였다. 예측 결과 RMSE값이 6.27, 6.42, 6.22,



(a) Random Forest Model (b) XGBoost Model

(c) LightGBM Model (d) DNN Model

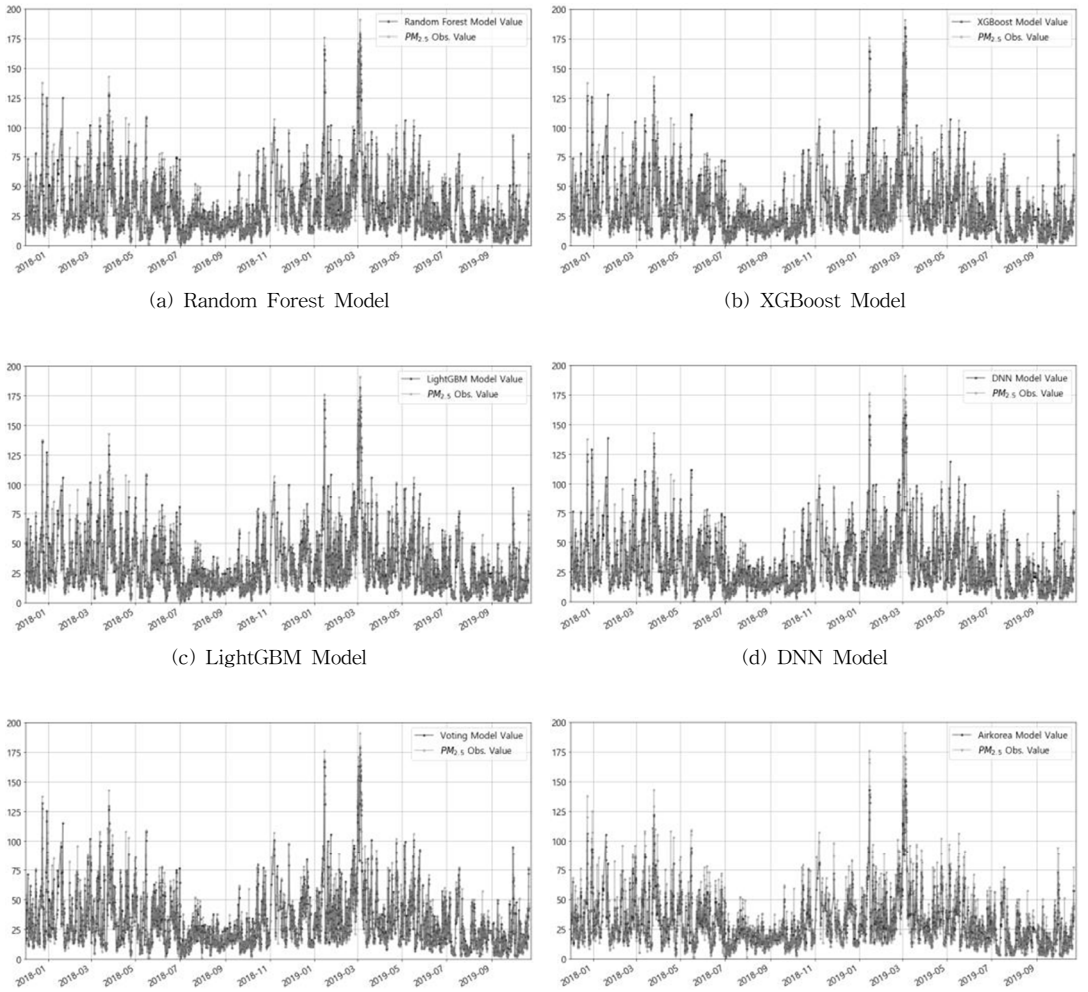
(e) Voting Model (f) Airkorea Model

그림 4 $PM_{2.5}$ 농도에 따른 실제 값과의 비교

Fig. 4 Comparison of Forecasting Models and Real(Observed) data based on $PM_{2.5}$

6.64로 개별 모델 모두 에어코리아에서 예측 모델인 10.77에 비해 매우 높은 성능을 보여준다. 이것은 단순히 이전 시간대의 $PM_{2.5}$만으로 예측한 에어코리아 예측 모델보다 대기질, 기상, 교통 등 실제로 영향을 주는 변수를 포함하여 예측했을 때 정확한 예측 결과를 얻을 수 있는 것을 보인다. 또한, 예측 결과가 가장 좋은 두 모델 Random Forest와 LightGBM을 Ensemble Voting을 통해 결합하여, 단일 모델보다 향상된 결과인 RMSE 6.12를 기록하였고 결과적으로 Voting Model과 에어코리아 예측모델을 비교하였을 때 약 56.82% 향상된 결과를 기록하였다. 이는 Random Forest의 Bagging과 LightGBM의 Boosting 서로 다른 방식의 두 앙상블 모델을 Voting하여, 각 모델의 예측 값 평균을 통해 성능 분산을 감소시켜 예측 값과 실제 값의 오차를 최적화한 결과이다.

현재 $PM_{2.5}$의 인체 유해성에 대해 국가 및 지자체 단위에서 저감을 위한 대책이 수립되고 있으며, 최근 환경부에서는 미세먼지 어린이집·노인시설 밀집지역 대상으로 집중관리구역을 지정하고, 생활밀착형 지원사업을 추진하고 있다. 그러나 아직 소수의 시·도를 중심으로 대책이 마련되고 있거나 명확하지 않은 기준으로 공통적으로 저감 대책이 시행되고 있다. 이에 본 논문은 경기도의 대표적인 사업장으로서 반월시화국가산업단지를 상관관계분석과 회귀분석을 통하여 계량적인 방법으로



(a) Random Forest Model (b) XGBoost Model

(c) LightGBM Model (d) DNN Model

(e) Voting Model (f) Airkorea Model

그림 5 $PM_{2.5}$ 농도에 따른 실제 값과의 비교

Fig. 5 Comparison of Forecasting Models and Real(Observed) data based on $PM_{2.5}$

기존 미세먼지 PM_{10}뿐만 아니라 초미세먼지 $PM_{2.5}$ 농도를 예측 및 분석을 하여 기존 연구화 차별화 되며, 실효성있는 미세먼지 저감 대책 및 효과적인 대기관리 방안 마련에 도움이 될 것으로 기대한다.

향후에는 중국 대기질 데이터를 활용하여 국외의 대기질을 포함한 미세먼지 관련 요인이 경기도 지역에 어떠한 영향을 주는지 분석하고, 예측모델을 연구할 예정이다.

References

[1] GRI, *The key task of improving fine dust is to strengthen the management capability of emission facilities in the workplace*, 2019.

[2] J. Jun, J. Lee, I. Kong, G. Sung, and G. Jung, "Development of fine dust and harmful substances to remove High-efficiency Cabin filter," *Proceedings of the Korean Society Of Automotive Engineers*, pp. 4-7, May. 2018.

[3] Y. Koo, H. Yun, H. Kwon, and S. Yu, "A Development of PM10 Forecating System," *Jouranl of Korean Society for Atmospheric Environment*, Vol. 26, No. 6, pp. 666-682, Nov. 2010.

[4] J. Cha and J. Kim, "Development of Data Mining Algorithm for Implementation of Fine Dust Nemerical Prediction Model," *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 22, No. 4, pp. 595-601, Apr. 2018.

[5] J. Kwon, Y. Lim, and H. Oh, "Particulate Matter Prediction using Quantile Boosting," *Journal of*

Applied Statistics, Vol. 28, No. 1, pp. 88-92, Feb. 2015.
[6] M. Seong, "Study on improvement of predicting fine dust concentration (master's thesis)," Sungkyunkwan University, pp. 1-42, Apr. 2019.
[7] I. Shin, Y. Moon, and Y. Lee, "Deep Learning Models Conformity Assessment for Particulate Matter Prediction in Smart Cities," *Journal of the Korean Institute of Information Scientists and Engineers*, Vol. 25, No. 12, pp. 610-615, Dec. 2019.
[8] H. Yi, K. N. Bui, and C. Seon, "A Deep Learning LSTM Framework for Urban Traffic Flow and Fine Dust Prediction," *Journal of the Korean Institute of Information Scientists and Engineers*, Vol. 47, No. 3, pp. 292-297, Jan. 2020.
[9] Airkorea. Available "https://www.airkorea.or.kr/web"
[10] KMA. Available "https://data.kma.go.kr/cmmn/main.do"
[11] The province of Gyeonggi. Available "https://gits.gg.go.kr"
[12] J. Cha and J. Kim, "Analysis of fine dust correlation between air quality and meteorological factors using SPSS," *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 22, No. 5, pp. 722-727, May. 2018.
[13] Breiman. L, "Random Forests," *Machine Learning*, Vol. 45, pp. 5-32, Oct. 2001.
[14] Y. Lee, H. Kim, D. Lee, C. Lee, and D. Lee, "Validation of Forecasting Performance of Two=Stage Probabilistic Solar Irradiation and Solar Power Forecasting Algorithm using XGBoost," *The Transactions of the Korean Institute of Electrical Engineers*, Vol. 68, No. 12, pp. 1704-1710, Dec. 2019.
[15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W, Ma, Q, Ye, and T,-Y, Liu (2017), "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Journal of Advances in Neural Information Processing Systems*, pp. 3149-3157, Morgan Kaufmann, San Mateo, 2017.
[16] J. Choi, D. Lee, J. Kim, and G. Jung, "Air pollution prediction using deep learning based model," *Proceedings of Korea Computer Congress 2019*, pp. 859-861, Jun. 2019. (in Korean)
[17] S. Symeonidis, D. Effrosynidis, J. Kordonis, and A. Arampatzis, "A Voting Classification Approach for Twitter Sentiment Analysis," *Proc. of the 11th international workshop on semantic evaluation 2017*, pp. 704-408, Aug. 2017. (in Canada)
[18] Scikit-learn. Available, "https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingRegressor.html"
[19] Airkorea. Available, "https://www.airkorea.or.kr/web/khaiInfo?pMENU_NO=129"

원 동 준

2018년 국민대학교 빅데이터경영통계학과 졸업(학사). 2020년~현재 현대트랜시스 매니저. 관심분야는 데이터 분석, 컴퓨터 비전, 자연어 처리

김 선 겸

2010년 세종대학교 컴퓨터공학과 졸업(학사). 2016년 연세대학교 컴퓨터과학과 졸업(박사). 2017년~2019년 한국과학기술정보연구원 박사후 연구원. 2019년~2020년 차세대융합기술연구원 선임연구원. 2020년~현재 한국건설기술연구원 수석연구원. 관심분야는 모바일 네트워크, 데이터 분석, 블록체인

김 영 훈

2008년 서울대학교 기계항공공학부 졸업(학사). 2014년 서울대학교 기계항공공학부 졸업(박사, 석박사통합). 2014년~2016년 KIST 바이오닉스연구단 박사후 연구원 2018년~현재 차세대융합기술연구원 선임연구원 겸 기술기획팀장. 관심분야는 인간-로봇 상호작용

송 규 원

2006년 아주대학교 정보및컴퓨터공학 졸업(학사). 2016년 과학기술연합대학원대학교(KIST) 졸업(박사). 2016년~2019년 KIST 영상미디어연구센터 박사후 연구원. 2019년~현재 차세대융합기술연구원 데이터과학연구실 실장. 관심분야는 데이터 과학 및 공학, 엣지컴퓨팅