JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 1, pp. 1-4, Feb. 2024 (작성 금지)

공간 빅데이터를 활용한 경남지역 미세먼지 예측모델 연구

A Study on Fine Dust Prediction Model for Gyeongnam Region Using Spatial Big Data

[요 약]

본 연구는 경남의 미세먼지 농도에 대기유해물질과 기상환경 요인들이 미치는 영향을 분석하여 '경남형 미세먼지 예측 모델'을 개발하는 것을 목표로 한다. 이를 위해 6종의 대기유해물질 데이터와 7종의 기상환경 데이터를 포함한 30만 건 이상의 데이터셋을 구축, Linear Regression·MLPRegressor·CNN 알고리즘 기반의 인공지능 모델을 통해 분석을 진행하였다. 연구 결과, 미세먼지와 초미세먼지는 강한 양의 상관관계를 보였으며, 이산화질소는 전구물질로 작용해 자외선과 반응하며 오존이 급증하는 것을 확인하였다. 또한, 미세먼지와 초미세먼지가 일산화탄소, 이산화질소와 강한 상관관계를 보이며, 농도 변화에 큰 영향을 미치는 것을 확인할 수 있었다. 다양한 환경 데이터를 활용하여 개발한 '경남형 미세먼지 예측 모델'은 경남의 대기환경 정책 수립의 근거자료로 활용될 수 있을 것이다.

[Abstract]

This study aims to develop a "Gyeongnam Fine Dust Prediction Model" by analyzing the impact of air pollutants and meteorological factors on fine dust concentration in Gyeongnam. A dataset comprising over 300,000 entries, including six types of air pollutants and seven types of meteorological data, was constructed. The analysis was conducted using AI models based on Linear Regression, MLPRegressor, and CNN algorithms. The results indicated a strong positive correlation between fine dust and ultrafine dust, with nitrogen dioxide acting as a precursor that reacts with ultraviolet light, leading to a surge in ozone levels. Additionally, fine and ultrafine dust showed significant correlations with carbon monoxide and nitrogen dioxide, affecting concentration changes. The developed Gyeongnam Fine Dust Prediction Model can serve as a foundational resource for establishing air quality policies in the region.

색인어 : 공간빅데이터, 인공지능, 예측모델, 미세먼지, 대기유해물질

Keyword : Spatial Big Data, Artificial Intelligence, Predictive Model, Fine Dust, Air Pollutants

http://dx.doi.org/10.9728/dcs.2024.25.1.1 (작성 금지)

Received 23 September 2023; **Revised** 22 October 2023
Accepted 05 November 2023 (작성 금지)

***Corresponding Author;**

Tel:

E-mail:

Ⅰ. 서 론

미세먼지는 전 세계적으로 가장 중요한 대기환경 및 호흡기 건강 문제의 원인으로 대두되고 있으며, 그 위험성은 날로 증가하고 있다. 환경과 인체에 유해한 미세먼지는 공기 중에 떠다니는 입자상 물질로, 크기에 따라 PM2.5와 PM10으로 나뉜다. 이 중 PM2.5, 즉 초미세먼지는 크기가 매우 작아 인간의 폐 깊숙이 침투할 수 있어 호흡기 질환, 심혈관 질환, 그리고 암을 유발하는 주요 원인으로 지목되고 있다. 세계보건기구(WHO)도 초미세먼지를 1급 발암물질로 지정하며 그 위험성을 경고한 바 있다[1]. 따라서 미세먼지와 초미세먼지 농도를 예측하고 이를 관리하는 것은 대기환경의 개선과 국민 호흡기 건강 보호를 위해 필수적이라 할 것이다.

한국의 미세먼지 문제는 대기오염물질의 국내·외 발생 원인이 복합적으로 작용하면서 더욱 심화되고 있다. 중국 등에서 유입되는 외부 요인뿐만 아니라 국내의 교통, 산업 활동, 난방 등 다양한 내적 요인도 미세먼지 농도를 높이는 주요 요인이 된다[2]. 특히, 경상남도는 서부의 대규모 산업단지와 선박 운항이 잦은 남해안의 지리적 특성으로 인해 미세먼지 농도의 변동성이 매우 크다[3]. 이에 따라, 경남 지역에 특화된 미세먼지 예측 모델을 개발하는 것은 지역 맞춤형 대기 질 관리에 있어 매우 중요한 선결과제로 인식되고 있다.

미세먼지 예측에서 가장 중요한 변수 중 하나는 기상데이터이다[4]. 기온, 풍속, 습도, 기압 등의 기상 요소는 미세먼지의 확산과 축적에 영향을 미치며, 이를 바탕으로 미세먼지 농도의 시공간적 변화를 예측할 수 있다. 예를 들어, 풍속이 강할 경우 대기 중에 정체되어 있던 미세먼지가 다른 지역으로 이동하거나 확산될 수 있으며, 반대로 대기 정체 시에는 미세먼지 농도가 급격히 상승할 수 있다. 기온과 습도 또한 미세먼지의 화학적 변형과 응축에 영향을 미쳐 농도에 변화를 일으킨다. 이처럼 기상 조건은 대기 중 미세먼지 농도 변화에 중대한 역할을 하기 때문에, 이를 정확하게 분석하는 것은 미세먼지 예측 모델의 정확성을 높이는 데 필수적이다.

또한, 미세먼지 농도 예측을 위해서는 측정소별 대기오염물질 농도 데이터가 중요한 역할을 한다. 경남 지역에 설치된 여러 대기오염 측정소에서 수집된 데이터는 특정 지역(구간)에서의 미세먼지 농도를 시간별·공간별로 파악할 수 있게 한다. 측정소별 데이터는 특정 지역의 오염원 특성, 기상 조건, 지리적 환경 등에 따라 미세먼지 농도가 어떻게 변화하는지 실시간으로 제공하기 때문에 지역 맞춤형 예측 모델 구축에 필수적인 자료다. 전술하였듯, 경남 지역은 산업 활동, 교통량, 해안과 인접한 특성 등으로 인해 대기오염물질 농도가 변화무쌍하며, 산업단지와 해안이 복합적으로 존재하는 지리적 특성으로 보다 정교한 지역 맞춤형 예측 모델이 요구된다.

이에 본 연구에서는 다양한 변수의 영향을 받는 경남의 기상 데이터와 측정소별 대기오염물질 농도 데이터를 기반으로 경남 지역에 특화된 미세먼지 예측 모델을 개발하고자 한다. 이를 위해 경남 지역의 다양한 환경적 특성과 오염물질 발생 원인을 종합적으로 고려한 공간 빅데이터 분석을 수행하고 있다. 기상데이터는 미세먼지 농도 변화에 영향을 미치는 주요 요인을 파악하는 데 도움을 주며, 측정소별 대기오염물질 농도 데이터는 특정 지역의 오염 패턴을 명확히 파악할 수 있는 실질적인 자료를 제공한다. 이 두 가지 카테고리의 데이터를 결합한 예측 모델은 경남 지역의 시공간적 미세먼지 농도를 보다 정밀하게 예측할 수 있을 것이며, 이를 통해 지역 맞춤형 대기질 관리 방안을 마련하는 데 기여할 수 있을 것이다.

궁극적으로 본 연구는 경남 지역의 특수성을 고려한 맞춤형 미세먼지 예측 모델을 개발함으로써 지역 주민의 호흡기 건강관리에 활용되고, 경남의 실효성 있는 미세먼지 저감 대책 수립에 중요한 인사이트를 제공할 것이다.

Ⅱ. 이론적 배경

2-1 공간 빅데이터 정의 및 특성

「국가공간정보 기본법」 제2조에 따르면, 공간정보란 '지상·지하·수상·수중 등 공간상에 존재하는 자연적 또는 인공적인 객체에 대한 위치정보 및 이와 관련된 공간적 인지 및 의사결정에 필요한 정보'를 뜻한다. 즉, 공간정보는 지도 위에 표현이 가능한 모든 정보를 의미하며, 공간 빅데이터는 이러한 공간정보를 포함한 대규모 빅데이터를 일컫는다.

공간 빅데이터는 교통관리, 도시계획, 환경 모니터링 등 다양한 분야에서 광범위하게 사용되고 있으며, 공적 영역에서 활용도가 점차 확대되고 있다[5]. 본 연구에서는 경상남도 전 지역(측정소)의 대기오염물질 데이터와 기상환경 데이터라는 공간적 개념에서 누적된 많은 양의 데이터를 활용해 유의미한 패턴을 연구하고, 정책적 의사결정에 도움을 줄 수 있는 근거를 제공하고 있다.

2-2 선행연구 검토

미세먼지 등 기수행된 대기유해물질 예측모델 개발 연구와 관련된 다양한 문헌 자료들을 참조하여 연구의 방향을 설정하였다. 이성구(2019)는 국내 환경을 고려해 중국발 미세먼지 데이터와 국내의 기상데이터 및 미세먼지 데이터를 활용한 딥러닝 모델을 개발했으며[6], 이종영 등(2020)은 시계열 데이터와 공간정보를 동시에 고려하는 새로운 대기질 예측 앙상블(ensemble) 모델을 개발하였다[7]. 또한 임준묵(2019)은 미세먼지 농도와 상관관계가 높을 것으로 밝혀진 기상자료와 대기질 관련 환경자료를 활용한 머신러닝 기법으로 예측의 정확도를 높였다[4]. 김혜림(2021) 등은 지방도시에서 노후산업단지가 있는 지역을 선정하여 미세먼지를 생성하는 요인을 분석하고, 미세먼지 발생을 예측할 수 있는 모형을 개발하였다. 개발의 결과물은 스마트산단의 발전을 촉진하

는 데 기여하였다[8]. 또한 성상하(2020) 등은 XGBoost, Random Forest, Support Vector Machine, Artificial Neural Network의 알고리즘을 활용해 미세먼지 수치를 예측하였고, 여러 모델 간 비교·분석을 통해 변수의 중요도를 분석하였다[9].

선행된 연구에서는 다양한 머신러닝 및 딥러닝 기법을 활용한 미세먼지 예측 모델이 개발되었고, 시계열 데이터와 공간정보를 결합한 앙상블 모델과 중국발 미세먼지와 국내 데이터를 연계한 연구들이 주를 이루었다. 그러나 각 지역별 대기환경 정책을 수립해야 함에도, 일률적인 모델로 실제 정책의 근거로 활용하기에 정확도가 떨어지는 측면이 있어 왔다. 따라서 경남의 시공간적 빅데이터 분석을 통해 예측의 정확도를 높인 '경남형 미세먼지 예측 모델'의 효용은 더욱 커질 수 있을 것이다.

2-3 대기오염물질 측정

환경부 산하 한국환경공단에서는 2004년 4월부터 전국의 대기측정망에서 측정되는 미세먼지를 포함한 대기오염 데이터를 수집 및 관리하는 국가대기오염정보관리시스템(NAMIS)을 구축하여, 대기오염물질 농도 정보를 제공하고 있다. 또한 2005년 12월 28일에는 '에어코리아(Air Korea)'라는 전국 실시간 대기오염도 공개 웹 플랫폼을 구축·공개하여 누구나 쉽게 대기오염물질 데이터에 접근할 수 있도록 하고 있다[10].

각 지자체에서도 별도의 대기오염물질 농도의 정보제공 플랫폼을 구축해 주민의 호흡기 건강관리를 위한 알권리 확보에 노력하고 있다. 경상남도에서는 'Air경남' 웹 플랫폼을 구축해 경남 전역의 47개 측정소에서 측정하는 미세먼지(PM10), 초미세먼지(PM2.5), 오존(O3), 일산화탄소(CO), 아황산가스(SO2), 이산화질소(NO2), 일산화질소(NO), 질소산화물(NOx) 등의 농도를 1시간 단위로 측정·제공하고 있다[11].

Ⅲ. 데이터 분석 방법

3-1 분석 대상

분석 대상은 경남의 47개 대기측정소에서 측정되며, Air경남에서 제공하는 미가공 데이터를 활용하였다. 분석 범위는 2021년 1월 1일 00시부터 2023년 12월 31일 23시까지 3개년의 미세먼지, 초미세먼지, 오존 등을 포함한 약 110만 개 대기오염물질 데이터를 수집하였다. 수집한 데이터 중 1차 상관분석을 통해 상관성이 거의 없는 일산화질소, 질소산화물, 이동차데이터 등을 소거하였다. 또한, 한국전력거래소에서 수집한 화력 및 전력 발전 자료는 결측치 양이 너무 많아 보간이 불가하므로 대상 변수에서 제외하였다.

이 밖에도 기상청에서 수집한 기상 데이터(기온, 강수량, 풍속, 습도, 일조, 적설, 전운량)를 미세먼지 농도에 영향을 미칠 것으로 가정하고 변수로 추가하였다. 결측치를 보완하기 위해 선형보간을 수행하였고, 각 변수에 대한 정규화를 진행하였다. 완성된 데이터셋은 시계열 데이터로서 최대한 그 특성을 살릴 수 있도록 하여 전처리 후 최종 데이터셋을 구축하였다.

최종 데이터셋은 총 301,356개의 데이터이며, '기온(°C)', '강수량(mm)', '풍속(m/s)', '습도(%)', '일조(hr)', '적설(cm)', '전운량(10분위)', '초미세먼지', '미세먼지', '오존', '일산화탄소', '아황산가스', '이산화질소'의 13개 피처(feature)로 구성된 데이터셋을 구축하였고, 최종적으로 Fig 1과 같은 각 데이터에 대한 기술통계량을 분석 대상으로 확정하였다.

	Temperature(°C)	Precipitation (mm)	Wind speed (m/s)	humidity(%)
count	301386.000000	301386.000000	301386.000000	301386.000000
mean	0.575235	0.009427	0.105032	0.663602
std	0.179112	0.027932	0.089993	0.227700
min	0.000000	0.000000	0.000000	0.000000
25%	0.437613	0.000000	0.034247	0.489796
50%	0.594937	0.001283	0.082192	0.693878
75%	0.725136	0.006576	0.150685	0.857143
max	1.000000	1.000000	1.000000	1.000000

	sunlight (hr)	Snow cover (cm)	Total cloud cover (decile)	PM25
count	301386.000000	301386.000000	301386.000000	301386.000000
mean	0.292575	0.052050	0.502398	0.072163
std	0.408332	0.157114	0.389320	0.047162
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.004663	0.000000	0.040359
50%	0.000000	0.006930	0.600000	0.062780
75%	0.700000	0.024555	0.900000	0.094170
max	1.000000	1.000000	1.000000	1.000000

PM10	O3	CO	SO2	NO2
301386.000000	301386.000000	301386.000000	301386.000000	301386.000000
0.027259	0.190043	0.116348	0.115892	0.137127
0.027539	0.103858	0.046594	0.039565	0.102102
0.000000	0.000000	0.000000	0.000000	0.000000
0.014648	0.112994	0.085714	0.083333	0.066667
0.022461	0.186441	0.114286	0.125000	0.100000
0.032227	0.254237	0.142857	0.125000	0.166667
1.000000	1.000000	1.000000	1.000000	1.000000

그림 1. 최종 데이터셋에 대한 기술통계량
Fig. 1. Descriptive statistics for the final data-set

3-2 데이터 분석 및 시각화 결과

각 데이터에서 유의미한 패턴과 변수 간의 상관관계를 규명하기 위해 Matplotlib, Altair 등의 Python 라이브러리를 이용, 빅데이터 분석 시각화를 진행하였다. Fig 2는 최근 3년(2021~2023) 동안 경남 전체 지역에 대한 미세먼지·초미세먼지 농도의 시계열 분석을 시각화한 자료다.

초미세먼지는 3년 동안 두드러진 큰 변화가 없지만, 미세먼지가 눈에 띄게 증가하는 시점에 함께 증가하는 패턴을 미루어 초미세먼지와 미세먼지는 함께 증가하는 패턴이 있는 것을 유추할 수 있다. 연구의 예측 모델의 타깃이 되는 미세먼지와 초미세먼지는 80:20 비율로 분리하여 각각 훈련(Train)과 테스트(Test) 데이터셋을 설정하였다.

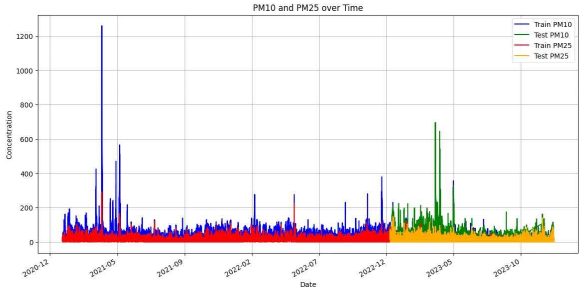

그림 2. 최근 3년간 미세먼지와 초미세먼지 농도 추이
Fig. 2. Trends in fine dust and ultrafine dust concentrations over the past three years

Fig 3은 시간대별 대기오염물질 6종에 대한 농도 추이 분석 결과다. 자동차, 화학공장, 정유공장 등에서 배출되는 전구물질인 휘발성유기화합물(VOCs)이 자외선과 광화학 반응을 일으켜 오존을 생성하고, 12시부터 오후 3시까지 낮 시간대에 급격하게 높아지고 있는 것으로 판단된다.

이와 반대로 이산화질소는 오전 6시부터 11시까지 높은 수치를 유지하다가, 오존 농도가 급증하는 낮 시간대에 낮아지는 것을 확인할 수 있다. 이산화질소는 대기 중 부유하며 자동차, 발전소, 공장에서 나오는 화학물질과 반응하여 오존을 생성하는 전구물질(precursor) 역할을 한다. 이는 이른 오전부터 생산활동을 시작하며 자동차, 공장, 발전소, 일터에서 생긴 이산화질소가 오전 시간대에 오존을 만들어 오후 시간대에 높은 농도로 변화하는 데 기인하는 것을 확인할 수 있으며, 음의 상관관계를 통해 전술한 내용을 유추할 수 있다.

그 외 대기오염물질은 크게 유의미한 패턴을 띄지는 않지만, 전체적으로 일조량과 오존 농도가 높은 오후 시간대에 미세먼지, 초미세먼지도 완만한 감소 내지 유사한 수준을 보이고 있다.

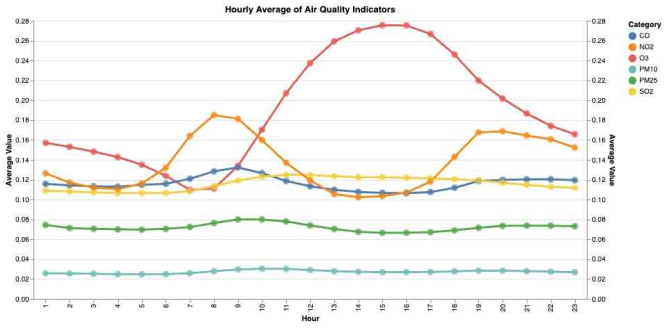

그림 3. 시간대별 대기오염물질 6종 농도 추이
Fig. 3. Concentration trends of six types of air pollutants by time zone

Fig 4는 6종의 대기오염물질을 월별 누적 분석한 그래프로, 대기 중 부유하는 오염물질의 총량을 확인하기 위함이다. 전체적인 대기오염물질은 1~3월에 가장 높고, 7~9월에 가장 낮은 것으로 분석된다. 미세먼지는 3~4월에 가장 높고 초미세먼지는 여름을 제외하면 유사하게 많은 양이 배출되는 것을 알 수 있다.

최근 3년(2021~2023) 간 대기오염물질 6종의 누적 농도를 비교한 Fig 5에서도 가장 높은 비율을 차지하는 대기오염물질은 오존과 이산화질소다. 이를 통해 여름철 높은 일조량, 오존 농도가 높은 낮 시간대에 대기오염물질 농도가 전체적으로 낮아질 수 있을 것이라 유추할 수 있다. 이러한 현상은 다양한 요소가 복합적으로 작용할 수 있지만, 예상되는 주요 원인은 강한 태양광, 기온의 상승, 태풍 등으로 인한 대기 순환, 장마로 인한 높은 습도와 비, 식물의 광합성 등이 될 수 있다. 그러나 오존과 같은 특정 오염물질은 여름철 급격히 증가할 수 있으므로, 대기질 변화를 종합적으로 고려하여 관리할 수 있는 정책 설정이 필요하다.

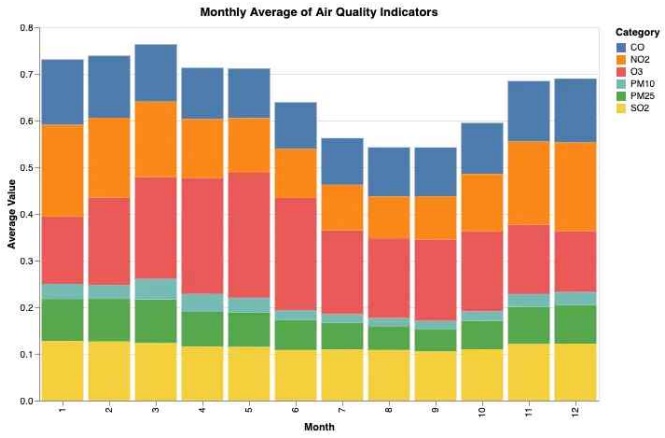

그림 4. 월별 대기오염물질 6종 농도 추이
Fig. 4. Concentration trends of six types of air pollutants by Month

대기오염물질 배출량은 누적 및 단일로도 큰 차이가 없음

을 확인할 수 있다(Fig 5). 대기오염물질을 구성하는 세부 오염물질의 구성이나 양의 변화에서도 유의미한 패턴이 발견되지 않았다.

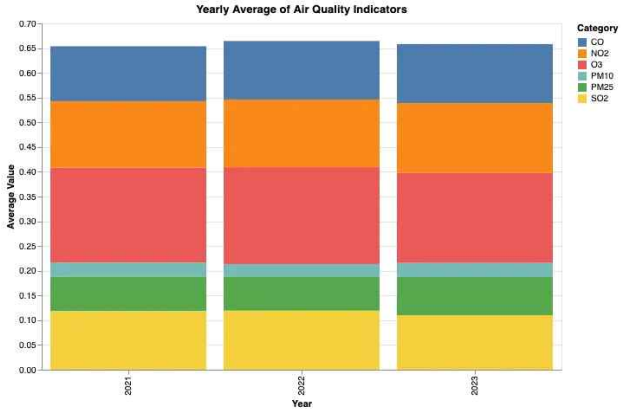

그림 5. 연도별 대기오염물질 6종 농도 추이

Fig. 5. Concentration trends of six types of air pollutants by Year

Fig 6은 변수 간의 상관관계와 유의미한 패턴을 찾기 위해 진행한 상관분석 결과로, 전체적으로 대기오염물질 변수 간 양의 상관관계를 띄고, 기상데이터 변수 간 얕은 음의 상관관계를 보인다. 오존은 풍속(0.40), 일조(0.38), 기온(0.35)과 양의 상관관계를 보이고, 습도(-0.46)와는 음의 상관관계를 보인다.

미세먼지와 초미세먼지는 상호 강한 양의 상관관계를 보이며, 초미세먼지에는 미세먼지(0.69), 일산화탄소(0.42), 이산화질소(0.41)가 강한 양의 상관성을 띤다. 미세먼지에는 초미세먼지(0.69)가 유일하며 강력한 양의 상관성을 띠고 있다.

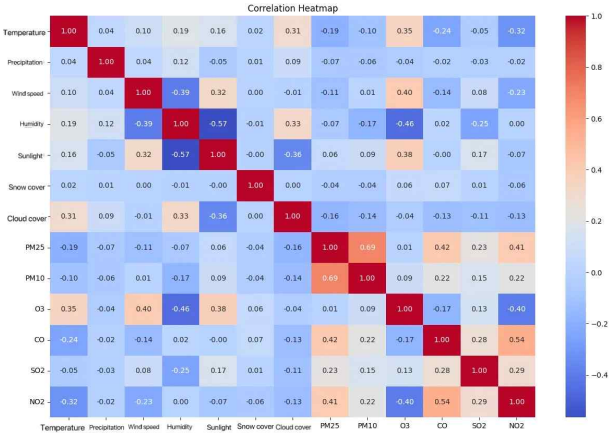

그림 6. 대기유해물질 6종 및 대기환경데이터 7종의 상관분석

Fig. 6. Correlation analysis of 6 types of air pollutants and 7 types of atmospheric environmental data

Ⅳ. 공간빅데이터 기반 인공지능 미세먼지 예측 모델 개발

4-1 Linear Regression 모델

선형 회귀는 종속변수 y와 한 개 이상의 독립변수(또는 설명변수) x와의 선형 상관관계를 모델링하는 회귀분석 기법으로, 80:20의 비율로 훈련과 테스트셋을 설정하였다. 성능 검증 평가에 대하여 초미세먼지에 대한 MSE는 1.918683027603784e-32, R^2 Score는 0.8이며, 미세먼지에 대한 MSE는 1.6494840385609694e-33, R^2 Score는 0.8로 모델의 성능이 매우 우수한 수준(0.8)으로 나왔다. 그러나 오버피팅(과적합)과 데이터의 일반화 능력 등을 고려하여 추가적인 교차검증(cross-validation)을 통해 모델이 잘 작동하는지 검토할 필요가 있다.

Fig 7과 같이 미세먼지와 초미세먼지는 서로 증가하는, 즉 미세먼지 농도가 높을수록 초미세먼지 농도도 높아지는 밀접한 관계가 있는 것으로 분석된다. 이에 두 변수의 관계를 직관적으로 보여주는 교차분석을 진행하였다.

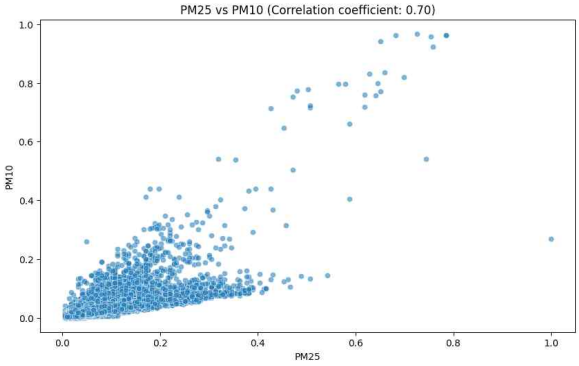

그림 7. PM2.5와 PM10의 상관성 회귀분석

Fig. 7. Correlation regression analysis between PM2.5 and PM10

초미세먼지와 미세먼지가 함께 증가하는 대각의 선형적 경향이 보이며, 0.0과 0.4의 사이 구간에서 클러스터를 형성하고 있다. 이는 특정 시간대나 지역에서 초미세먼지와 미세먼지 값이 비슷한 패턴을 보일 수 있음을 뜻한다.

다만 완벽한 직선 형태로 분포되지 않고, 곡선의 형태를 보이므로 초미세먼지와 미세먼지의 관계가 단순히 선형적인 비례 관계가 아닌, 다소 복잡한 상호 작용이 존재할 수 있음을 시사하고 있다. 또한, 데이터 포인트들이 대체로 직선 주변에 밀집되어 있는데, 이는 모델이 데이터의 일반적인 패턴을 잘 학습했음을 의미한다. 훈련 데이터와 테스트 데이터에 대한 정확도가 비슷하다는 점을 보아 오버피팅이 발생하지 않았음을 유추할 수 있다.

4-2 MLPRegressor 모델

MLPRegressor(Multi-layer Perceptron regressor)는 인공신경망(Artificial Neural Network, ANN)을 기반으로 한 다층 퍼셉트론 회귀 모델로, 비선형 관계를 학습하여 연속형 값을 예측하는 데 사용된다. 이번 예측 분석에서도 80:20의 비율로 훈련(Train)과 테스트(Test) 데이터셋을 설정하였으며, 2023년 1월부터 2024년 1월까지의 미세먼지 예측값과 실제값을 비교하는 그래프를 시각화하면 Fig 8과 같다.

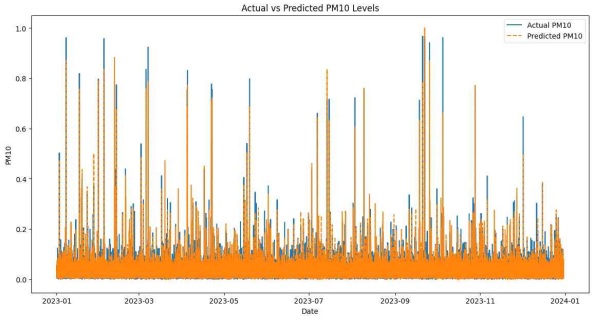

그림 8. 미세먼지에 대한 실제값과 예측값 비교
Fig. 8. Comparison of actual and predicted values for fine dust

Fig 8에서 알 수 있듯, 전체적인 경향에 대해 예측을 잘하는 것을 알 수 있다. Train MSE(Mean Squared Error, 평균제곱오차)는 0.29, Train MAE(Mean Absolute Error, 평균절대오차)는 0.09, Test MSE와 Test MAE도 각각 0.37, 0.12로 데이터에 대해 모델이 예측값과 실제값 사이의 오차가 극히 적어 성능이 매우 뛰어난 모델임을 알 수 있다.

특히, 훈련 데이터와 테스트 데이터에서의 성능이 유사하다는 것은 모델이 과적합(overfitting) 되지 않고 일반화(generalization) 능력이 뛰어나다는 것으로, 모델의 안정성이 확보된 것으로 판단된다.

잔차 플롯의 목적은 모델이 예측에 있어 체계적인 오류를 가지고 있는지를 확인하는 것이다. Fig 9는 잔차가 0을 중심으로 고르게 분포되어 있는 경우이므로 모델이 데이터를 잘 예측하고 있다는 것을 의미한다.

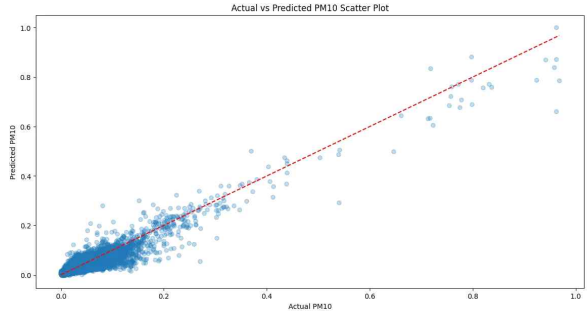

그림 9. 잔차 시각화 그래프
Fig. 9. Residual visualization graph

MLPRegressor는 일반적으로 피처의 중요도를 직접적으로 제공하지 않지만, 모델의 예측 성능이 피처의 순서를 무작위로 섞었을 때 얼마나 감소하는지를 측정하는 순열 중요도(permutation importance)를 사용해 피처의 중요도를 평가할 수 있다. Fig 10에서 보듯이, 미세먼지에는 초미세먼지가 압도적인 영향력을 행사하며, 그 뒤로 습도, 기온, 아황산가스, 이산화질소, 오존, 일조, 일산화탄소, 풍속, 강수량 등의 순으로 영향을 미치고 있다.

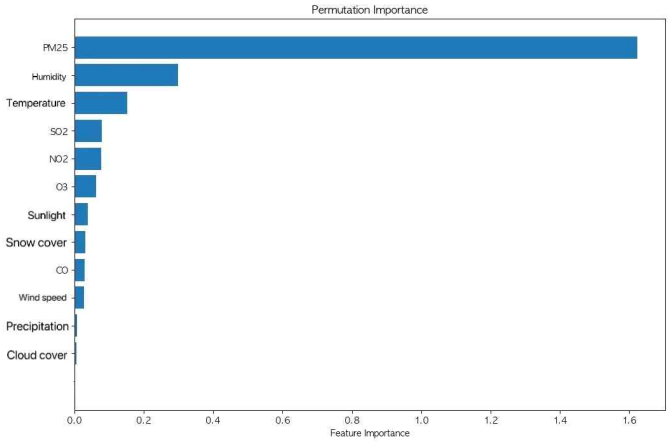

그림 10. 머신러닝 모델의 변수 중요도
Fig. 10. Variable importance in machine learning models

4-3 CNN 모델

CNN(Convolutional Neural Network)은 다양한 분야에서 널리 쓰이는 인공지능 신경망 아키텍처로, 시계열 예측 분야에서도 뛰어난 성능을 발휘한다. 확보한 대기오염물질 데이터셋과 기상 데이터셋이 각각 미세먼지 예측에 어떤 영향을 미치는지와 모델 성능을 비교하기 위해 각각의 데이터셋을 별도로 사용하거나 결합하여 비교·분석하는 실험을 진행하였다.

통일성을 위한 변수 통제를 위해 80:20의 비율로 훈련(Train)과 테스트(Test) 셋을 설정하였고, 에포크(epoch)는 50으로 통일하였다. Table 1과 같이, MSE 값은 [Weather only→All data→Pollution only] 순으로 높게 나타났다. MSE 값은 수치가 낮을수록 좋은 모델이므로 [Pollution only→All data→Weather only] 순으로 정확도가 높은 것이다. 따라서 기상데이터는 미세먼지와 초미세먼지 분석에 중요도가 낮고 대기오염물질은 중요하다고 할 수 있으며, 미세먼지에 대기오염물질의 상관성이 높은 것을 알 수 있다.

표 1. 데이터 종류별 MSE 값 비교
Table 1. Comparison of MSE values by data type

Data type	MSE value
Pollution only	1.0914277
Weather only	6.8926544
All data	3.4465205

Fig 11은 미세먼지 예측 모델을 검증하기 위한 그래프로, 미세먼지 예측값(주황색)이 실제값(파란색)과 크게 벗어나지 않고 유사한 패턴을 따르고 있음을 알 수 있다. 피크값에 있어서도 어느 정도는 예측을 하고 있으나, 크게 튀는 값들에 대해서는 제대로 예측하지 못하고 상대적으로 낮은 값을 예측하고 있다.

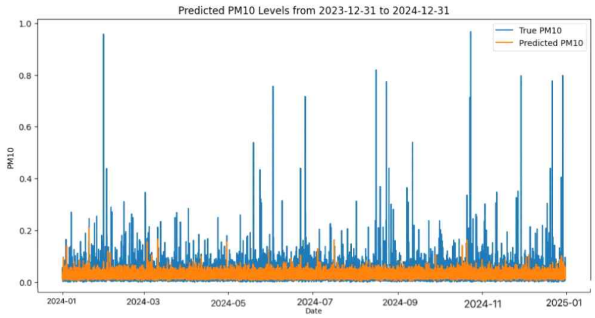

그림 11. 2024.1.1.~2024.12.31.까지 미세먼지(PM10) 예측
Fig. 11. Fine dust (PM10) forecast from 2024.1.1. to 2024.12.31.

Fig 12는 초미세먼지 예측 모델의 검증 그래프로, 초미세먼지 예측값(주황색)이 실제값(파란색)과 크게 벗어나지 않고 유사한 패턴을 따르고 있음을 알 수 있다. 하지만, 낮음~중간 수준의 피크값은 정밀하게 예측하던 미세먼지 예측 결과와 달리, 피크값을 제대로 예측하지 못하고 있다. 급격하게 미세먼지나 초미세먼지가 높아지는 것은 현재의 변인들만으로 패턴이나 원인을 명확하게 파악하기 어렵고, 또 다른 변인들에 의해 통제받는 것을 유추할 수 있다.

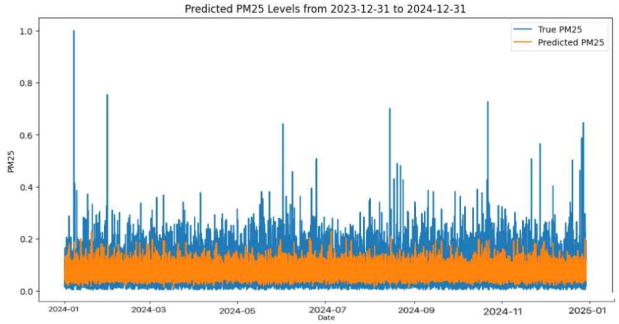

그림 12. 2024.1.1.~2024.12.31.까지 초미세먼지(PM2.5) 예측
Fig. 12. Ultra Fine dust (PM2.5) forecast from 2024.1.1. to 2024.12.31.

V. 결론

본 연구는 대기유해물질 6종과 기상환경 데이터 7종이 미세먼지와 초미세먼지에 미치는 영향을 분석하고 있다. 분석 결과, 미세먼지 발생에 영향을 미칠 것으로 가정한 환경요인에 대한 변수들이 경남의 환경 조건에서 모두 양의 상관관계를 보이는 것은 아님을 확인할 수 있었다.

상관분석을 통해 변수 중요도를 측정한 결과, 미세먼지와 초미세먼지는 아주 강한 양의 상관관계를 보이고 있다. 즉 미세먼지나 초미세먼지는 하나가 증가하면 다른 하나의 수치도 증가하며, 동시에 상호 가장 주요한 발생 요인으로 밝혀졌다. 또한, 미세먼지와 초미세먼지는 1~4월 가장 높은 패턴을 보이는 것으로 분석되었으며, 이는 봄철 황사, 꽃가루 등이 주요한 원인임을 유추할 수 있다.

오전 6시부터 11시까지는 자동차, 발전소, 공장에서 발생하는 화학물질에 반응하는 이산화질소가 전구물질 역할을 하여 일조량이 급격하게 느는 12시부터 15시까지 자외선과 광화학 반응을 일으켜 오존이 급증하는 것을 확인하였다. 따라서 해당 시간대에 대기오염물질 배출사업장과 공장 밀집지역 등의 배출을 분산시키거나, 해당 구역의 배출 상황을 집중 관리하는 등의 조치로 대기유해물질 배출의 절대량을 줄이는 방안이 필요할 것으로 판단된다. 아울러 이산화질소의 경우 휘발유·가스차 등에서 많이 배출됨에 따라 저공해사업을 노후 경유차에서 노후 휘발유·가스차로 확대하고, 일산화탄소는 가스열펌프(GHP) 저감장치 부착을 지원, 오존은 발생원인물질인 질소산화물을 배출하는 배출사업장에 대한 관리를 강화하는 등의 대책을 마련해 나가야 할 것이다.

이 밖에도 미세먼지·초미세먼지가 일산화탄소, 이산화질소와 강한 상관도를 보이며, 농도 변화에 직접적인 영향을 미치는 것으로 확인하였다. 공학적 관점에서 1차 생성 요인을 줄이면 2차 생성물을 자연스레 줄일 수 있음을 이용하여 일산화탄소와 이산화질소의 저감을 통해 미세먼지와 초미세먼지를 예방·관리하는 조치가 필요하다.

또한, 현재 경남도의 '미세먼지 계절관리제'는 12월 1일부터 이듬해 3월 31일까지로 설정하여 대기오염물질 배출일 줄이고 집중 관리하는 조치를 시행하고 있다. 하지만 분석 결과와 같이 경남도는 12월보다 4~5월 대기유해물질 농도가 높은 것으로 분석되어, 미세먼지 계절관리제 추진 기간에 대한 탄력적 운용이 필요해 보인다.

본 연구는 미세먼지 농도 예측을 위해 다양한 대기환경 변수를 활용하여 상관관계와 중요도를 분석했음에도 불구하고, 더욱 다양한 기상 데이터나 화력 발전량, 전력 발전량, 항만 발생물질, 중국발 대기오염물질, 지형 데이터 등의 변수들은 고려하지 못하였다. 이러한 내재적 한계에도 불구하고, 경남의 환경 특성만을 고려하여 대기유해물질과 기상환경 데이터에 대한 공간 빅데이터 분석을 수행하여 '경남형 미세먼지 예측 모델'을 개발한 데서 연구의 의의가 있을 것이다. 향후 더

욱 다양한 변인들을 추가하여 모델의 정밀도와 강건성을 높이는 연구가 필요할 것이다.

참고문헌

[1] J. B. Kim, “Assessment and Estimation of Particulate Matter Formation Potential and Respiratory Effects from Air Emission Matters in Industrial Sectors and Cities/Regions”, *Journal of Korean Society of Environmental Engineers,* Vol. 39, No. 4, pp. 220-228, April. 2017.
http://dx.doi.org/10.4491/KSEE.2017.39.4.220

[2] M. W. Choi, “An Overview on China's Recent Air Pollution Regulation and Management Policy”, *Environmental and Resource Economics Review,* Vol. 27, No. 3, pp. 399-420, September. 2018.
http://dx.doi.org/10.15266/KEREA.2018.27.3.569

[3] C. S. Sim, “Study on Integrated Management of Particulate Matter Pollution”, Korea Environment Institute, Sejong, December. 2021.

[4] J. M. Lim, “An Estimation Model of Fine Dust Concentration Using Meteorological Environment Data and Machine Learning”, *Journal of Information Technology Services,* Vol. 18, No. 1, pp.173-186, March. 2019.
https://doi.org/10.9716/KITS.2019.18.1.173

[5] D. H. Kim, “The Use of Spatial Big Data for Planning Policy Support”, Korea Research Institute for Human Settlements, Sejong, December. 2014.

[6] S. K. Lee, “Research of Particulate Matter Prediction Modeling Based on Deep Learning”, Master´s degree, Sungkyunkwan University, 2019.

[7] J. Y. Lee, M. J. Choi, Y. I. Joo and J. K. Yang, “Ensemble Method for Predicting Particulate Matter and Odor Intensity”, *Journal of Korean Society of Industrial and Systems Engineering*, Vol. 42, No. 4, pp.203-210, December. 2019.
https://doi.org/10.11627/jkise.2019.42.4.203

[8] H. L. Kim and T. H. Moon, “Machine learning-based Fine Dust Prediction Model using Meteorological data and Fine Dust data”, *Journal of the Korean Association of Geographic Information Studies,* Vol. 24, No. 1, pp.92-111, March. 2021.
http://dx.doi.org/10.11108/kagis.2021.24.1.092

[9] 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”, *Innovation studies,* Vol. 15, No. 4, pp.339-357, November. 2020.
https://doi.org/10.46251/INNOS.2020.11.15.4.339

[10] Air Korea, Available: https://www.airkorea.or.kr/

[11] Air Gyeongnam, Available: https://golib.gyeongnam.go.kr/