

경상남도의회 대학생 인턴십 제1기

공간빅데이터를 활용한 경남지역 미세먼지 예측모델 연구

2024. 7. 31.

인턴명 : 하이든(인제대학교)

경상남도의회

— 목 차 —

I . 서론	01
1. 연구 배경	01
2. 연구 목적	02
3. 연구 내용 및 방법	03
II . 이론적 배경	04
1. 공간 빅데이터 정의 및 특성	04
2. 선행연구 검토	04
3. 대기오염물질 측정	05
III. 공간빅데이터 기반 미세먼지 예측 모델 개발	07
1. 분석 대상	07
2. 분석 방법	08
3. 예측 모델 및 정확도 평가	13
IV. 결론	20

공간빅데이터를 활용한 경남지역 미세먼지 예측모델 연구

인제대학교 하이든

I. 서론

1. 연구 배경

- 미세먼지의 발생원인(표1)은 국내와 국외로 구분되며, 국내 원인은 인위적 배출원으로부터 직접적으로 배출되거나, 대기 중 오존과 산소, 라디칼(Hydroxyl) 등 전구물질의 형태로 화학반응을 통해 2차 생성되기도 하며, 일부는 자연적으로 발생되기도 함
- 국외 주요 원인으로는 인접국인 중국에서 발생한 고농도 미세먼지가 대기 흐름에 의해 강한 서풍을 타고 이동하며 국내로 유입되는 경우가 이에 해당함

표 1 | 미세먼지 발생원인

구분		발생원인
국내원인	직접배출	• 자동차 연료 연소, 사업장 연소, 생물성 연소과정 등
	2차생성	• 질소산화물(NOx), 암모니아(NH3), 황산화물(SOx), 휘발성유기화합물(VOCs) 등이 대기 중 수증기 등과 반응하여 생성
	자연발생	• 소금입자(해염 등), 광물입자(황사 등), 생물성입자(꽃가루 등)
국외원인		• 중국 등에서 발생한 고농도 미세먼지가 강한 바람 등의 기상 영향으로 국내로 유입

자료) 고농도 미세먼지 대응실무매뉴얼(2021)

- 대기오염물질 중 입자상 존재하는 미세먼지(Particulate Matter; PM)는 크기에 따라 공기역학적 직경 $10\mu\text{m}$ 이하인 경우 미세먼지(PM10)로, $2.5\mu\text{m}$ 이하는 초미세먼지(PM2.5)로 분류됨
- 특히 초미세먼지는 머리카락 직경의 20분의 1보다 작은 미세한 입자이며, 기관지 점막을 통해 걸러지지 않고 폐포까지 직접 침투하여 인체에 유해한

영향(호흡기질환, 암 발병, 어린이들의 폐 성장 악화, 기관지염, 천식 증상 악화 등)을 끼치며, 조기사망률을 증가시키는 것으로도 밝혀짐

- 미세먼지·초미세먼지는 산성비와 사막화 진행을 가속화하며 다양한 인간생활 및 환경에 악영향을 끼침
 - 나노 단위의 공정을 거쳐 만드는 반도체 및 디스플레이 산업에서는 아주 작은 이물질로도 불량 제품이 생길 수 있는데, 이러한 이유로 산업 전반에서 불량률을 높이는 등 악영향을 끼치고 있으며, 항공기와 선박의 운영제한 등 사회적·경제적으로 다양한 형태의 부정적 영향을 미침
- 각 지역에 설치되는 도시대기측정망 또는 도로변대기측정망은 「대기오염측정망 설치운영 지침」에 근거하여 해당 지자체에서 신설·이전·폐쇄를 주관하고 있으나, 여전히 측정 누락지역이 발생하고 있고, 실시간 측정이 불가능한 한계 존재
 - 각 측정소는 근교 건물 옥상 등에 설치되는 경우가 많아 중심가 및 주거단지 등 실제 생활권의 농도 측정이 불가하여 DB로서의 신뢰성 부족, 활용도 저조의 문제가 발생함
 - 또한 각 측정소에서는 미세먼지 농도를 1시간 간격으로 제공하고 있어 미세먼지 예·경보시스템에 대한 낮은 적시성 문제가 상존함

2. 연구 목적

- 경상남도에서는 ‘Air경남 대기환경정보’라는 웹기반 플랫폼을 구축하여 경남 지역의 실시간 대기질 현황을 직관적으로 제공하고 있으나, 실시간 농도의 현황을 보여주는 데 그치고 있어 예방·대비의 목적으로 활용되지 못하고 사후 처리만 가능함
 - 또한 다양한 대기환경 연구가 발달한 수도권 지역과 달리 경남 지역의 대기오염물질 분석 및 예측 모델 연구는 전무한 수준임
 - 이러한 이유로 경남 지역의 특성을 고려하여 대기오염물질에 대한 공간 정보를 수집해 AI·빅데이터 분석하고, 미세먼지 발생을 예측할 수 있는 모델의 개발이 필요함
 - 이에 본 연구에서는 AI·빅데이터 등의 ICT 기술을 활용하여 Air경남에서 제공하는 대기유해물질 6종의 데이터와 기상청에서 제공하는 기상 데이터

7종을 원 데이터(raw data)로 삼아, 다양한 AI 알고리즘을 활용하여 분석하였고, 정밀하고 고도화된 예측력을 지닌 미세먼지 예측 모델을 개발하고자 함

3. 연구 내용 및 방법

- 본 연구는 대기환경 측정소별 대기오염물질 데이터를 기준으로, 기상청에서 수집한 기상 데이터를 전처리 후 AI·빅데이터 분석함
 - 최종 데이터셋은 301,356개의 데이터이며, 2021년 1월 1일부터 2023년 12월 31일까지(3년), 1,095일을 00시부터 23시까지 1시간 간격으로 구분하여 측정한 데이터임
 - 맷플롯립(Matplotlib)¹⁾, 알테어(Altair)²⁾ 등의 라이브러리(Library)³⁾를 이용해 빅데이터 분석 및 시각화를 진행하였고, 선형회귀(LinearRegression)⁴⁾, MLPRegressor, CNN(Convolutional Neural Network)⁵⁾ 등의 AI 모델을 활용하여 미래 미세먼지 데이터를 예측함
 - 13개의 피쳐(feature)⁶⁾와 미세먼지, 초미세먼지의 레이블(label)로 구성된 데이터셋을 지역별(측정소별)로 다양하게 채택하여 사용함
 - 80:20, 90:10 두 가지 데이터 분할 비율로 랜덤하게 2회씩 분류하여 훈련용 데이터와 검증용 데이터를 구축하였으며, 평균제곱오차(Mean Squared Error, MSE)와 결정계수(R^2 Score), 순열 중요도(Permutation Importance)⁷⁾를 기준으로 모델 간 성능을 비교·분석함

1) Python 프로그래밍 언어 및 수학적 확장 NumPy 라이브러리를 활용한 플로팅 라이브러리
2) Python에서 쉽게 차트나 그래프를 그릴 수 있도록 해주는 라이브러리로, 그래프를 그리는 코드를 작성할 필요가 없는 대신 적절한 옵션을 설정하고 원하는 그래프의 형태를 정의할 수 있는 직관적인 인터페이스를 제공함
3) API를 기반으로 대상 환경(플랫폼: Linux, macOS, Windows)에서 바로 실행될 수 있도록 모듈화된 프로그램 모음. 라이브러리는 혼자서 동작하는 완전한 프로그램이 아닌, 특정한 부분 기능만을 수행하도록 제작된, 컴파일되어 기계어의 형태로(또는 대상 플랫폼에 따라서는 바이트코드로) 존재하는 프로그램임
4) 선형회귀(Linear Regression)는 $y=f(x)+\varepsilon$ 의 형태로 입력 변수 x 와 출력 변수 y 사이의 선형 상관관계를 모델링하는 분석 기법임
5) 필터링 기법을 인공지능망에 적용하여 이미지를 효과적으로 처리할 수 있는 심층 신경망 기법으로 행렬로 표현된 필터의 각 요소가 데이터 처리에 적합하도록 자동으로 학습되는 과정을 통해 이미지를 분류하는 기법임. 시계열 예측에도 효과적임.
6) 데이터 특성을 나타내는 것으로 데이터 표에서 열(Column)을 지칭함
7) 피쳐가 모델의 전체 예측에 얼마나 중요한지를 측정한 것으로, 피쳐의 순서를 무작위로 섞었을 때 모델의 예측 성능이 얼마나 감소하는지를 측정함

II. 이론적 배경

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 뜻함
 - 즉, 공간정보는 지도 위에 표현이 가능한 모든 정보를 의미하며, 공간빅데이터는 이러한 공간정보를 포함한 대규모 빅데이터를 일컫음. 공간빅데이터는 교통관리, 도시계획, 환경 모니터링 등 다양한 분야에서 광범위하게 사용되고 있으며, 공적 영역에서 활용도가 점차 확대되고 있음
 - 본 연구에서도 경남도 전 지역(측정소)의 대기오염물질 데이터와 기상환경 데이터라는 공간빅데이터를 활용해 유의미한 패턴을 연구하고 정책적 의사결정에 도움을 줄 수 있는 기초자료를 제공하고 있음

2. 선행연구 검토

- 미세먼지 등 대기유해물질 예측모델 개발 연구는 기 수행된 연구들이 많아 다양한 문헌자료들을 참조하여 연구의 방향을 설정할 수 있었음
 - 이성구(2019)는 국내 환경을 고려해 중국발 미세먼지 데이터와 국내의 기상데이터 및 미세먼지 데이터를 활용한 딥러닝 모델을 개발했으며, 양재경 등(2020)은 시계열 데이터와 공간정보를 동시에 고려하는 새로운 대기질 예측 앙상블(Ensemble) 모델을 개발함. 또한 임준목 등(2019)은 미세먼지 농도와 상관관계가 높을 것으로 밝혀진 기상자료와 대기질 관련 환경자료를 활용한 머신러닝기법으로 예측의 정확도를 높였음
 - 김혜림 등(2021)은 지방도시에서 노후산업단지가 있는 지역을 선정해, 미세먼지를 생성하는 요인을 분석하고, 미세먼지 발생을 예측할 수 있는 모델을 개발, 스마트산단의 발전을 촉진하는 데 기여함
 - 성상하 등(2020)은 XGBoost, Random Forest, Support Vector Machine, Artificial Neural Network의 알고리즘을 활용해 미세먼지 수치를 예측했고, 여러 모델 간 비교·분석을 통해 변수의 중요도를 파악함

3. 대기오염물질 측정

- 환경부 산하 한국환경공단에서는 2004년 4월부터 전국의 대기오염측정망에서 측정되는 미세먼지를 포함한 대기오염 데이터를 수집 및 관리하는 국가대기오염정보 관리시스템(NAMIS)을 구축하고 농도 정보를 제공하고 있음
- 2005년 12월 28일에는 ‘에어코리아(Air Korea)’라는 전국 실시간 대기오염도 공개 웹 플랫폼을 구축·공개해 누구나 쉽게 대기오염물질 데이터에 접근할 수 있도록 하고 있음
- 경상남도에서도 ‘Air 경남’ 웹 플랫폼(그림1)을 구축해 경남 전역의 47개 측정소(표2)에서 측정하는 미세먼지, 초미세먼지, 오존, 일산화탄소, 아황산가스, 이산화질소, 일산화질소, 질소산화물 등의 농도를 1시간 단위로 측정·제공하고 있음

그림 1 Air 경남 대기환경정보 웹페이지 발췌

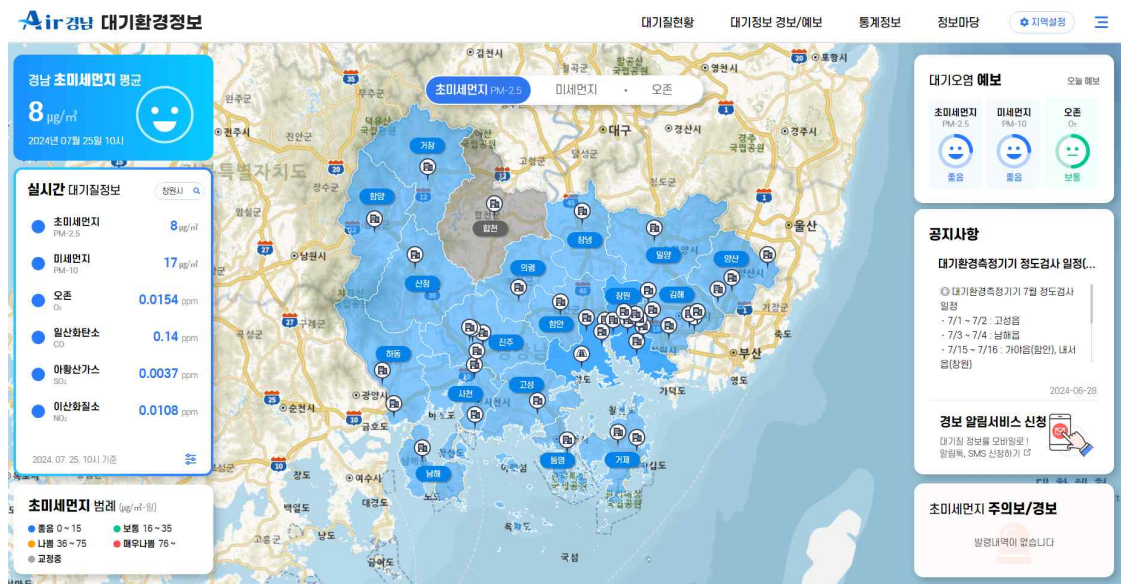


표 2 경남 대기환경측정망 측정지점(47개소)

시·군명	측정소	주소	측정망
창원시	명서동	창원시 의창구 우곡로 101번길 85(명서2동 민원센터)	도시대기
	용지동	창원시 의창구 용지로 239번길 19-4(용지동 행정복지센터)	도시대기
	사파동	창원시 성산구 창이대로 706번길 16-23(사파민원센터)	도시대기

시·군명	측정소	주소	측정망
	성주동	창원시 성산구 외리로 14번길 18(성주민원센터)	도시대기
	웅남동	창원시 성산구 공단로 303(효성굿스프링스)	도시대기
	월영동	창원시 마산합포구 월영동 16길 22(마산합포도서관)	도시대기
	봉암동	창원시 마산회원구 봉암로 148(봉암동 주민센터)	도시대기
	회원동	창원시 마산회원구 회원동 11번길 7(회원1동 주민센터)	도시대기
	내서읍	창원시 마산회원구 내서읍 광령로 8(삼계근린공원)	도시대기
	경화동	창원시 진해구 경화로 16번길 31(병암동주민센터)	도시대기
	반송로	창원시 성산구 중앙동 516-4(시설관리공단 실내 수영장 앞)	도로변대기
	삼진로	창원시 마산합포구 진동면 진동리 635(진동시외버스정류장 앞)	도로변대기
진주시	상봉동	진주시 북장대로 64번길 14(중앙119안전센터)	도시대기
	대안동	진주시 진주대로 1052(IBK기업은행)	도시대기
	상대동	진주시 동진로 279(한국전력공사 진주지점)	도시대기
	정촌면	진주시 정촌면 예하리 1340(예하초등학교 앞, 공원)	도시대기
통영시	무전동	통영시 안개4길 53(무전동주민센터)	도시대기
사천시	사천읍	사천시 사천읍 읍내로 52(사천시니어클럽)	도시대기
	향촌동	사천시 향촌 5길 28 (향촌동주민센터)	도시대기
김해시	삼방동	김해시 활천로 303(신어초등학교상)	도시대기
	동상동	김해시 호계로 517번길 8(동상동 행정복지센터)	도시대기
	장유동	김해시 장유동 능동로 149(서부건강지원센터)	도시대기
	진례면	김해시 진례면 진례로 241(진례면 행정복지센터)	도시대기
	진영읍	김해시 진영읍 김해대로 365번길 6-24(진영읍보건지소)	도시대기
	김해대로	김해시 삼정동 360(인제대역 환승주차장)	도로변대기
밀양시	내일동	밀양시 밀양시 중앙로 346(내일동주민센터)	도시대기
거제시	고현동	거제시 계룡로 125(거제시청)	도시대기
	아주동	거제시 아주동 산 164-1(아주공설운동장)	도시대기
양산시	물금읍	양산시 물금읍 황산로 384(물금읍행정복지센터)	도시대기
	북부동	양산시 북안남5길 21(중앙동주민센터)	도시대기
	삼호동	양산시 삼호9길 11(웅상노인복지회관)	도시대기
의령군	의령읍	의령군 의령읍 서동리 564-2(서동생활공원)	도시대기
함안군	가야읍	함안군 가야읍 함안대로 505(가야읍행정복지센터)	도시대기
창녕군	창녕읍	창녕군 창녕읍 우포2로 1189-35(창녕군보건소 정신건강복지센터)	도시대기
고성군	고성읍	고성군 고성읍 중앙로 35(고성읍보건지소상)	도시대기
남해군	남해읍	남해군 남해읍 남해대로 2745(남해유배문학관)	도시대기

시·군명	측정소	주소	측정망
하동군	하동읍	하동군 하동읍 군청로 23(하동군청)	도시대기
	금성면	하동군 금성면 금성중앙길 14(금성꿈나무어린이집)	도시대기
산청군	산청읍	산청군 산청읍 옥산리 276(산청군청 주차장)	도시대기
함양군	함양읍	함양군 함양읍 고운로 35(함양군청)	도시대기
거창군	거창읍	거창군 거창읍 대평리 1298-1(거창자전거교통안전교육장)	도시대기
합천군	합천읍	합천군 합천읍 대야로 888-21(합천보훈회관)	도시대기
국가 측정망	대산면	창원시 의창구 대산면 갈전로99번길 100	교외대기
	마산항	창원시 성산구 적현로 424(귀곡동)	항만
	부산항	경남 창원시 진해구 신항로 434(안골동)	항만
	저구리	거제시 남부면 저구리 산 116번지	교외대기
	남상면	거창군 남상면 인평길 36 남상면사무소	교외대기

* 지면으로부터 시료채취구 Funnel 까지의 수직 높이

Ⅲ. 공간 빅데이터 기반 미세먼지 예측 모델 개발

1. 분석 대상

○ 데이터 전처리

- Air경남에서 제공하는 47개 측정소 데이터를 기준으로 2021년 1월 1일 00시부터 2023년 12월 31일 23시까지의 3개년 미세먼지, 초미세먼지, 오존 등을 포함한 약 110만 개의 대기오염물질 데이터를 수집함
- 1차 상관분석을 통해 상관성이 거의 없는 일산화질소(NO), 질소산화물(NOX), 이동차데이터 등을 소거함. 한국전력거래소에서 수집한 화력 및 전력 발전 자료는 결측치⁸⁾의 양이 너무 많아 보간⁹⁾이 불가하므로 대상 변수에서 제외함
- 기상청에서 수집한 기상 데이터(기온, 강수량, 풍속, 습도, 일조, 적설, 전운량)를 미세먼지 농도에 영향을 미칠 것으로 가정하고 변수로 추가함. 결측치를 보완하기 위해 선형보간¹⁰⁾을 수행하였고, 각 변수에 대한 정규화¹¹⁾를 진행함

8) 수집된 데이터 셋 중 관측되지 않은 특정 확률변수의 값

9) 결측치를 채우기 위해 인접한 데이터를 이용해 함수(일반적으로 다항식) 조각을 맞추어 데이터의 값을 채우는 프로세스

10) 끝점의 값이 주어졌을 때 그 사이에 위치한 값을 추정하기 위하여 직선 거리에 따라 선형적으로 계산하는 방법

11) 이상현상이 있는 릴레이션을 분해하여 이상현상을 없애는 과정

- 완성된 데이터셋은 시계열 데이터로서, 최대한 그 특성을 살릴 수 있도록 하여 전처리 후 최종 데이터셋을 구축함
- 최종 데이터셋은 총 301,356개의 데이터이며, '기온(°C)', '강수량(mm)', '풍속(m/s)', '습도(%)', '일조(hr)', '적설(cm)', '전운량(10분위)', 'PM2.5', 'PM10', 'O3', 'CO', 'SO2', 'NO2'의 13개 피쳐(feature)로 구성된 데이터셋을 구축함. 최종적으로 그림2와 같이 각 데이터에 대한 기술통계량을 분석 대상으로 확정함

그림 2 각 데이터별(기상데이터 7종+대기유해물질 6종) 기술통계량

	MSR_DT	기온(* C)	강수량(mm)	풍속(m/s)	습도(%)	일조(hr)	적설(cm)	전운량(10분위)
count	3 013860e+05	301386.000000	301386.000000	301386.000000	301386.000000	301386.000000	301386.000000	301386.000000
mean	2.022067e+09	0.575235	0.009427	0.105032	0.663602	0.292575	0.052050	0.502398
std	8.172313e+05	0.179112	0.027932	0.089993	0.227700	0.408332	0.157114	0.389320
min	2.021010e+09	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2.021093e+09	0.437613	0.000000	0.034247	0.489796	0.000000	0.004663	0.000000
50%	2.022070e+09	0.594937	0.001283	0.082192	0.693878	0.000000	0.006930	0.600000
75%	2.023040e+09	0.725136	0.006576	0.150685	0.857143	0.700000	0.024555	0.900000
max	2.023123e+09	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
		PM25	PM10	O3	CO	SO2	NO2	
	301386.000000	301386.000000	301386.000000	301386.000000	301386.000000	301386.000000	301386.000000	
		0.072163	0.027259	0.190043	0.116348	0.115892	0.137127	
		0.047162	0.027539	0.103858	0.046594	0.039565	0.102102	
		0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
		0.040359	0.014648	0.112994	0.085714	0.083333	0.066667	
		0.062780	0.022461	0.186441	0.114286	0.125000	0.100000	
		0.094170	0.032227	0.254237	0.142857	0.125000	0.166667	
		1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

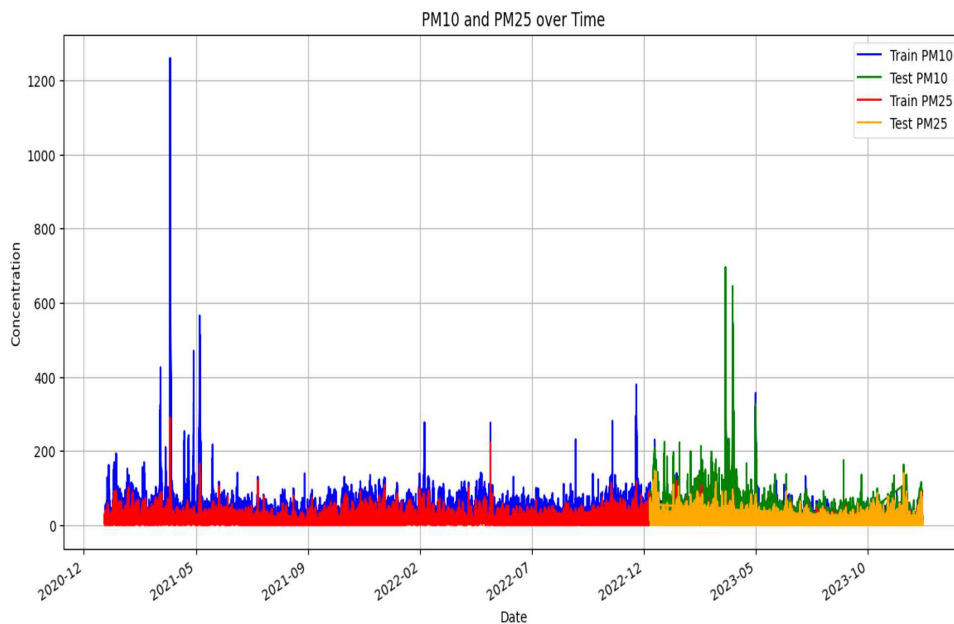
2. 분석 방법

○ 데이터 분석 및 시각화

- 각 데이터에서 유의미한 패턴과 변수 간의 상관관계를 규명하기 위해 Matplotlib, Altair 등의 Python 라이브러리를 이용, 빅데이터 분석 시각화를 진행함. 그림3은 최근 3년(2021~2023) 동안 경남 전체 지역에 대한 미세먼지, 초미세먼지의 농도의 시계열 분석을 시각화한 자료임

- 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 있음
- 본 연구에서 미래 예측 모델의 타깃이 되는 미세먼지와 초미세먼지는 80:20 비율로 분리하여 각각 훈련(Train)과 테스트(Test) 데이터셋을 설정함

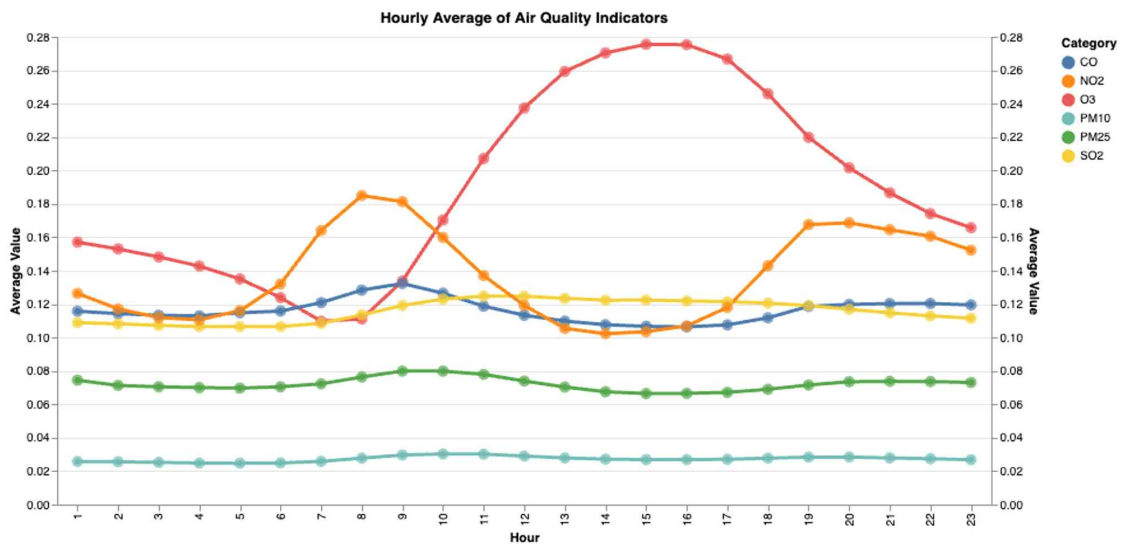
[그림 3] 최근 3년(2021-2023) 미세먼지와 초미세먼지 농도 추이



- 그림4는 시간대별 대기오염물질 6종에 대한 농도 추이에 대한 분석결과임. 자동차, 화학공장, 정유공장 등에서 배출되는 전구물질인 휘발성유기화합물(VOCs)이 자외선과 광화학 반응을 일으켜 생성되는 오존(O3)은 12시부터 시작해 오후 3시까지 낮 시간대에 급격하게 높아짐
- 이와 반대로 이산화질소(NO2)는 오전 6시부터 11시까지 높은 수치를 유지하다가, 오존 농도가 급증하는 낮 시간대에 낮아지는 것을 확인할 수 있음
- 이산화질소(NO2)는 대기 중 부유하며 자동차, 발전소, 공장에서 나오는 화학물질과 반응하여 오존(O3)을 생성하는 전구물질(precursor) 역할을 함. 이는 이른 오전부터 생산활동을 시작하며 자동차, 공장, 발전소, 일터에서 생긴 이산화질소(NO2)가 오전 시간대에 오존(O3)을 만들어 오후 시간대에 높은 농도로 변화하는 데 기인하는 것을 확인할 수 있으며, 음의 상관관계를 통해 전술한 내용을 유추할 수 있음

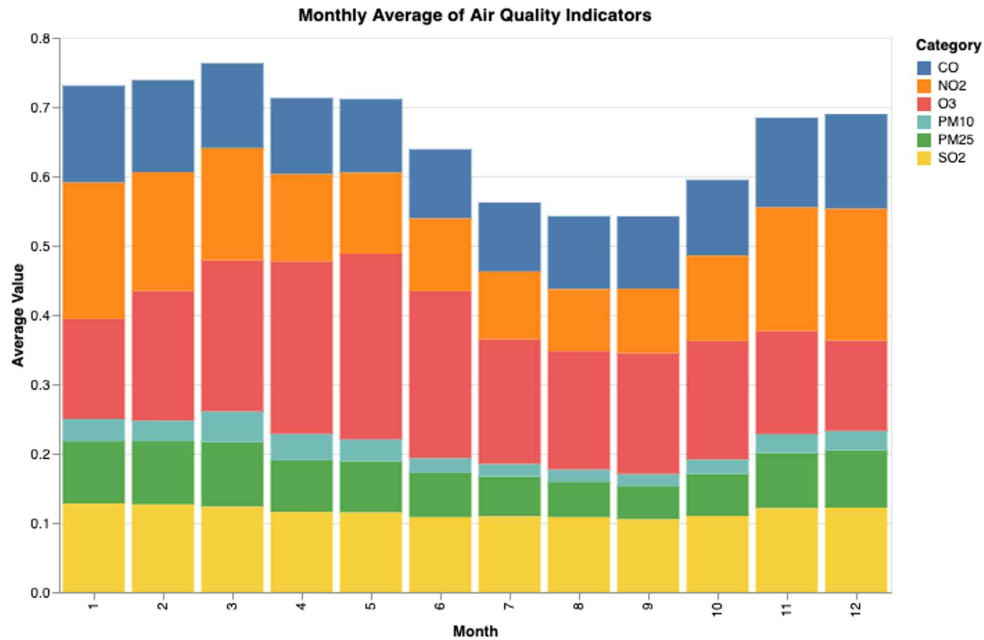
- 그 외 대기오염물질은 크게 유의미한 패턴을 띄지는 않지만, 전체적으로 일조량과 오존 농도가 높은 오후 시간대에 미세먼지, 초미세먼지도 완만한 감소 내지 유사한 수준을 보임

|그림 4| 시간대별 대기오염물질 6종 농도 추이



- 그림5는 6종의 대기오염물질을 월별로 누적 분석한 그래프로, 대기 중 부유하는 오염물질의 총량을 확인하기 위함임. 전체적인 대기오염물질은 1~3월에 가장 높고, 7~9월에 가장 낮아지는 것으로 분석됨
- 미세먼지는 3~4월에 가장 높고 초미세먼지는 여름을 제외하면 유사하게 많은 양이 배출되는 것을 알 수 있음
- 그림4와 마찬가지로, 가장 높은 비율을 차지하는 대기오염물질은 오존(O3)과 이산화질소(NO2)임. 그림4·그림5의 결과를 통해 유추할 수 있는 것은 여름철 높은 일조량, 오존 농도가 높은 낮 시간대에 대기오염물질 농도가 전체적으로 낮아지는 경향을 확인할 수 있음

[그림 5] 월별 대기오염물질 6종 누적 농도 추이



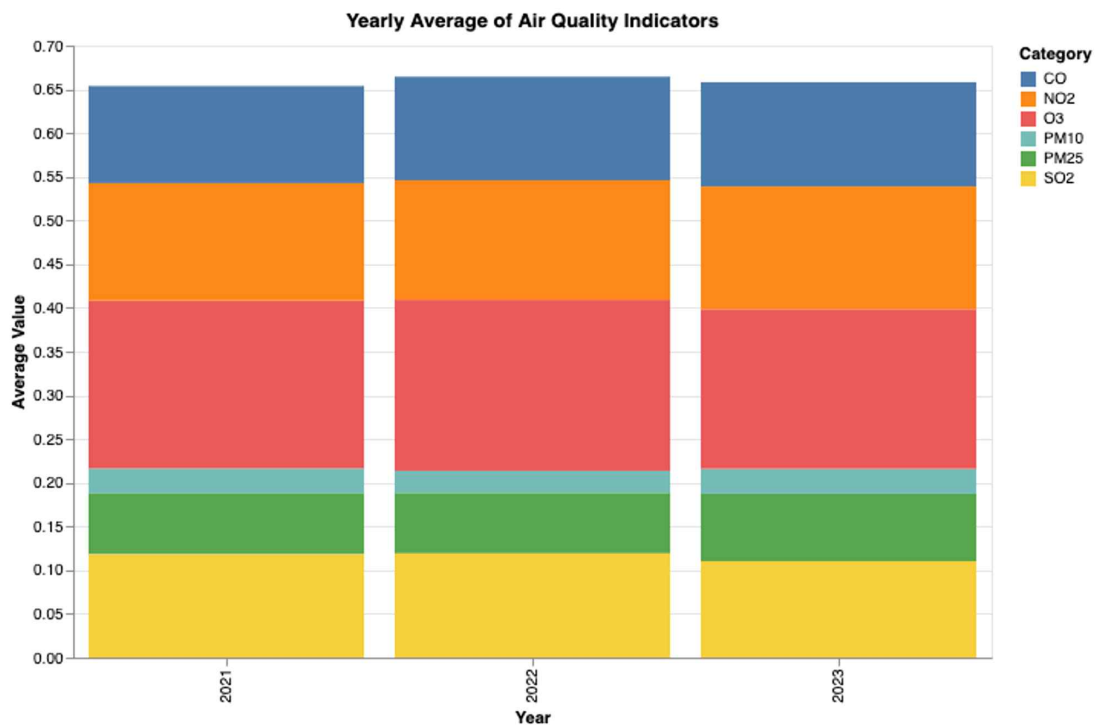
- 이러한 현상의 요인은 복합적으로 작용하는 여러 요소가 있지만 예상되는 주요 원인은 표3과 같음
- 이러한 요인들이 복합적으로 작용하여 여름철, 햇빛이 많은 오후 시간대에 대기오염물질의 농도가 낮아지는 경향을 보이는 것으로 해석할 수 있음
- 그러나 오존과 같은 특정 오염물질은 여름철 급격히 증가할 수 있으므로, 대기질 변화를 종합적으로 고려하여 관리하는 정책 방향설정이 필요함

[표 3] 대기오염물질 농도와 대기환경 영향 요인 분석

영향요인	내용
강한 태양광과 온도 상승	여름에는 태양광이 강하게 내리쬐고 온도가 상승해 대기 중의 오염물질이 화학적으로 변화할 수 있는 환경이 조성됨. 따라서 높은 온도와 햇빛으로 인해 오염물질이 대기 중에서 분해되거나 반응하여 줄어들 수 있음. 다만 오존은 태양빛에 의해 생성되는 2차 오염물질로, 그 양이 증가할 수 있음
대기 순환	여름철에는 대기 순환이 활발해지는 경향이 있어 오염물질이 다른 지역으로 분산되거나 대기 중에서 희석되는 데 도움을 줄 수 있음. 강한 바람이 불면 오염물질이 빠르게 확산되어 농도가 낮아질 수 있음
여름철 장마철로 인한 비와 높은 습도	여름철에는 장마로 인해 비가 자주 오고, 습도가 높아지면서 대기 중의 미세먼지와 오염물질이 세척되거나 침전되어 전체적인 농도가 낮아지는 효과가 있음
식물의 광합성	여름철에는 식물의 생장과 광합성이 활발해지는데, 이 과정에서 이산화탄소를 흡수하고 산소를 방출하면서 대기 질에 긍정적인 영향을 미칠 수 있음

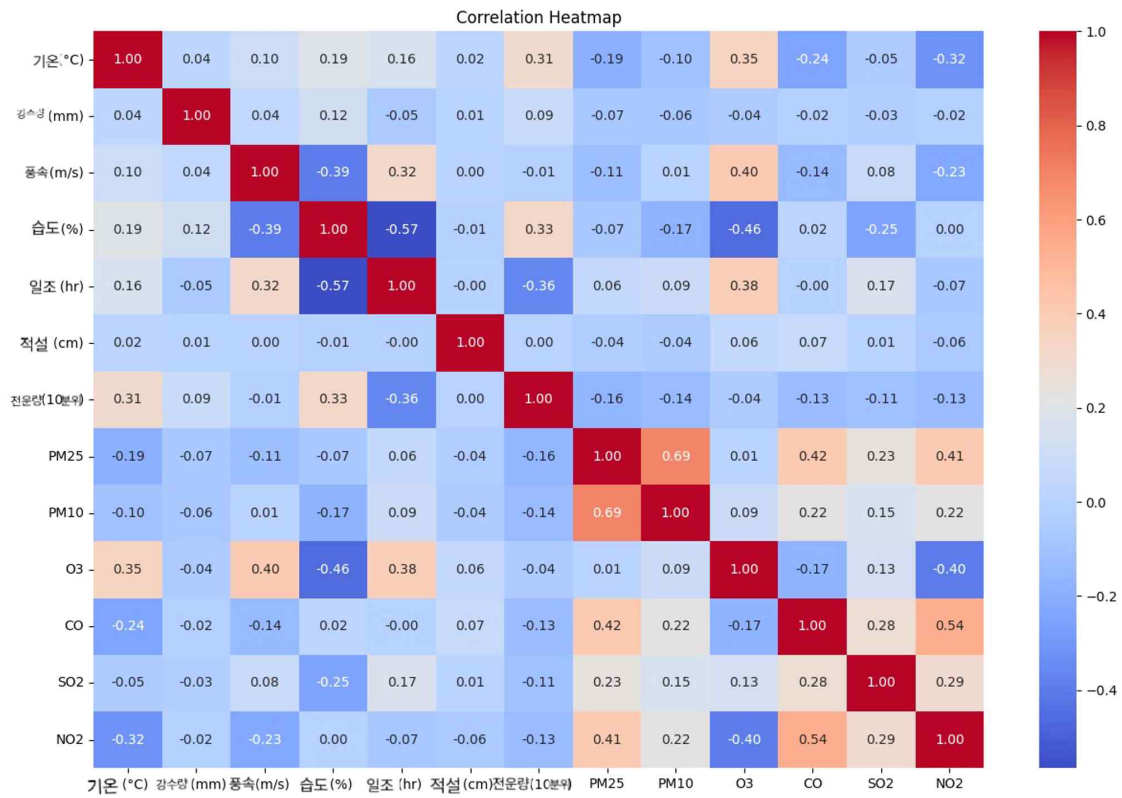
- 최근 3년(2021~2023) 대기오염물질 배출량은 누적 및 단일로도 큰 차이가 없음을 확인할 수 있음(그림6). 대기오염물질을 구성하는 세부 오염물질의 구성이나 양의 변화에서도 유의미한 패턴이 발견되지 않았음

그림 6 연도별 대기오염물질 6종 농도 누적 농도 추이



- 그림7은 변수 간의 상관관계와 유의미한 패턴을 찾기 위해 진행한 상관분석 결과임. 전체적으로 대기오염물질 변수 간 양의 상관관계를 띄고, 기상 데이터 변수 간 얇은 음의 상관관계를 보임
- 오존은 풍속(0.40), 일조(0.38), 기온(0.35)과 양의 상관관계를 보이고, 습도(-0.46)와는 음의 상관관계를 보임. 미세먼지와 초미세먼지는 상호 강한 양의 상관관계를 보임 ※ 괄호의 숫자는 -1~1 사이의 상관계수를 명기한 것임
- 초미세먼지에는 미세먼지(0.69), 일산화탄소(0.42), 이산화질소(0.41)가 강한 양의 상관성을 보임. 미세먼지에는 초미세먼지(0.69)가 유일하며 강력한 양의 상관성을 보임
- 앞선 분석에서도 오존이 미세먼지와 유의미한 패턴을 보이는 것으로 보아 오존의 패턴 변화와 미세먼지·초미세먼지의 관계에 보다 주목할 필요가 있는 것으로 사료됨

그림 7 대기유해물질 6종 및 대기환경데이터 7종의 상관분석



3. 예측 모델 및 정확도 평가

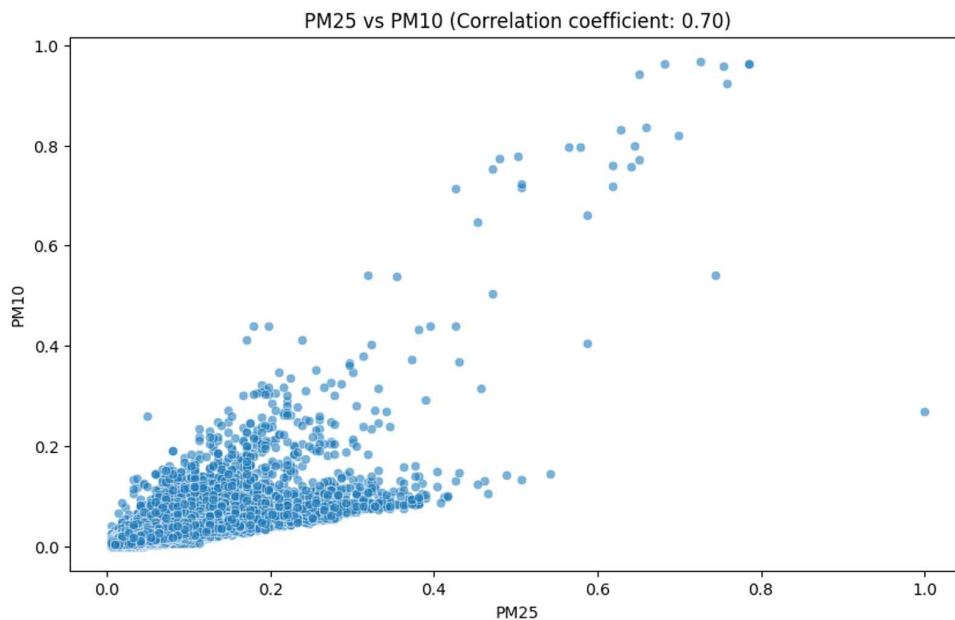
○ Linear Regression(선형 회귀)

- 선형 회귀는 종속변수 y 와 한 개 이상의 독립변수(또는 설명변수) x 와의 선형 상관관계를 모델링하는 회귀분석 기법으로, 80:20의 비율로 훈련과 테스트셋을 설정하였음
- 성능 검증 평가에 대하여 PM2.5에 대한 MSE는 $1.918683027603784e-32$, R^2 Score는 0.8이며, PM10에 대해서는 MSE는 $1.6494840385609694e-33$, R^2 Score는 0.8임
- 평균제곱오차(Mean Squared Error, MSE)는 예측값과 실제값 간의 차이를 제공하여 평균한 값으로, 값이 적을수록 모델의 성능이 우수하다는 것을 의미함. R^2 Score 결정계수는 모델이 종속변수의 분산을 얼마나 잘 해석하는지를 나타내는 0과 1 사이 값이며, 1에 가까울수록 모델이 데이터를 잘 해석한다는 것을 뜻함
- 모델의 성능이 매우 우수한 수준(0.8)으로 나왔지만, 오버피팅(과적합)¹²⁾과

데이터의 일반화¹³⁾ 능력 등을 고려하여 추가적인 검토 및 교차검증 (Cross-validation)을 통해 모델이 잘 작동하는지 추가 검토할 필요가 있음

- 앞선 분석에서 미세먼지와 초미세먼지는 서로 증가하는, 즉 미세먼지 농도가 높을수록 초미세먼지 농도도 높아지는 밀접한 관계가 있는 것으로 분석됨. 이에 두 변수의 관계를 직관적으로 보여주는 교차분석을 진행함(그림8)

그림 8 PM2.5-PM10 상관성 선형회귀분석



- 초미세먼지와 미세먼지가 함께 증가하는 대각의 선형적 경향이 보임. 점들이 0.0과 0.4의 사이 구간에서 클러스터를 형성하는데, 이는 특정 시간대나 지역에서 초미세먼지와 미세먼지 값이 비슷한 패턴을 보일 수 있음을 뜻함
- 다만 완벽한 직선 형태로 분포되지 않고, 곡선의 형태를 보이므로 초미세먼지와 미세먼지의 관계가 단순히 선형적인 비례 관계가 아닌, 다소 복잡한 상호 작용이 존재할 수 있음을 시사함
- 또한 데이터 포인트들이 대체로 직선 주변에 밀집되어 있는데, 이는 모델이 데이터의 일반적인 패턴을 잘 학습했음을 의미함. 훈련 데이터와 테스트 데이터에 대한 정확도가 비슷하다는 점을 보아 오버피팅이 발생하지 않았다는 것을 알 수 있음

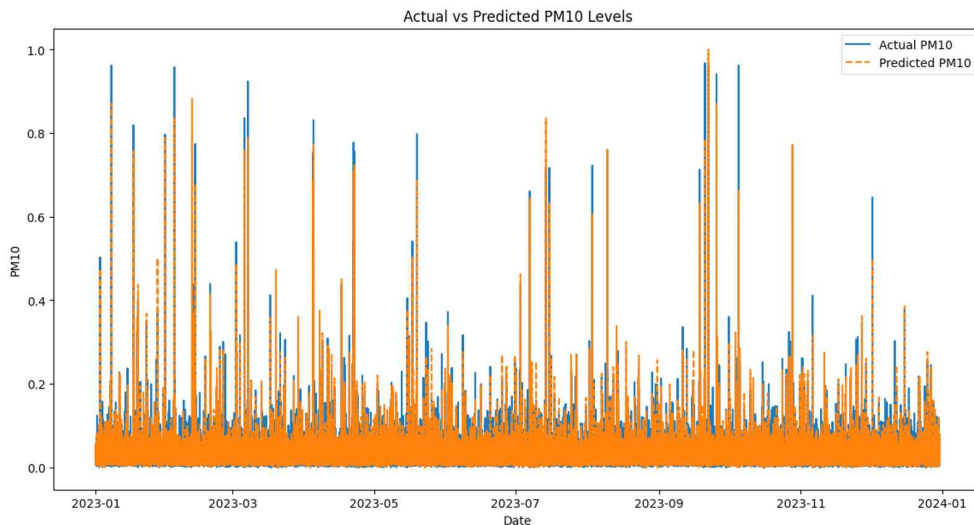
12) 머신러닝 모델이 학습 데이터에 지나치게 최적화되어 새로운 데이터(테스트 데이터)에 대해서는 일반화 능력이 떨어지는 현상

13) 모델이 다양한 실제 상황에서 잘 수행되게 만드는 과정

○ MLPRegressor

- 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 사용됨
- 이번 예측 분석에서도 80:20의 비율로 훈련(Train)과 테스트(Test) 셋을 설정하였으며, 2023년 1월부터 2024년 1월까지의 미세먼지 예측값과 실제값을 비교하는 그래프를 시각화하면 그림9와 같음

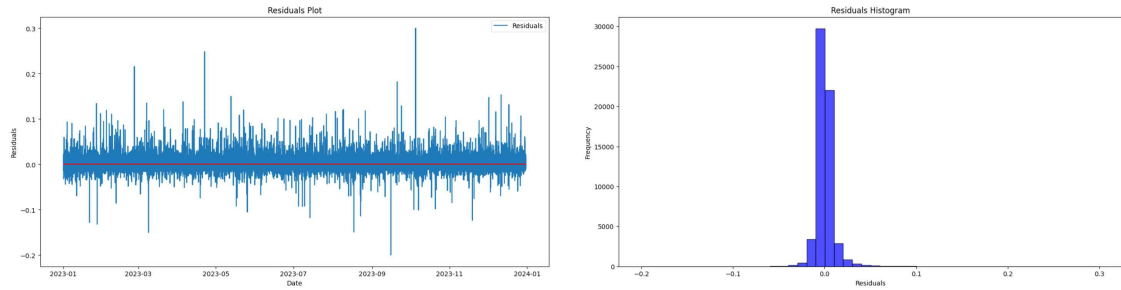
그림 9 미세먼지에 대한 실제값과 예측값 비교



- 그림 9에서 알 수 있듯, 전체적인 경향에 대해 예측을 잘하고 있다는 것을 알 수 있음. 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 있음
- 특히, 훈련 데이터와 테스트 데이터에서의 성능이 유사하다는 것은 모델이 과적합(Overfitting)되지 않고 일반화(Generalization) 능력이 뛰어나다는 것을 시사함
- 잔차¹⁴⁾ 플롯의 목적은 모델이 예측에 있어 체계적인 오류를 가지고 있는지 여부를 확인하는 것임. 그림 10은 잔차가 0을 중심으로 고르게 분포되어 있는 경우이므로 모델이 데이터를 잘 예측하고 있다는 것을 의미함

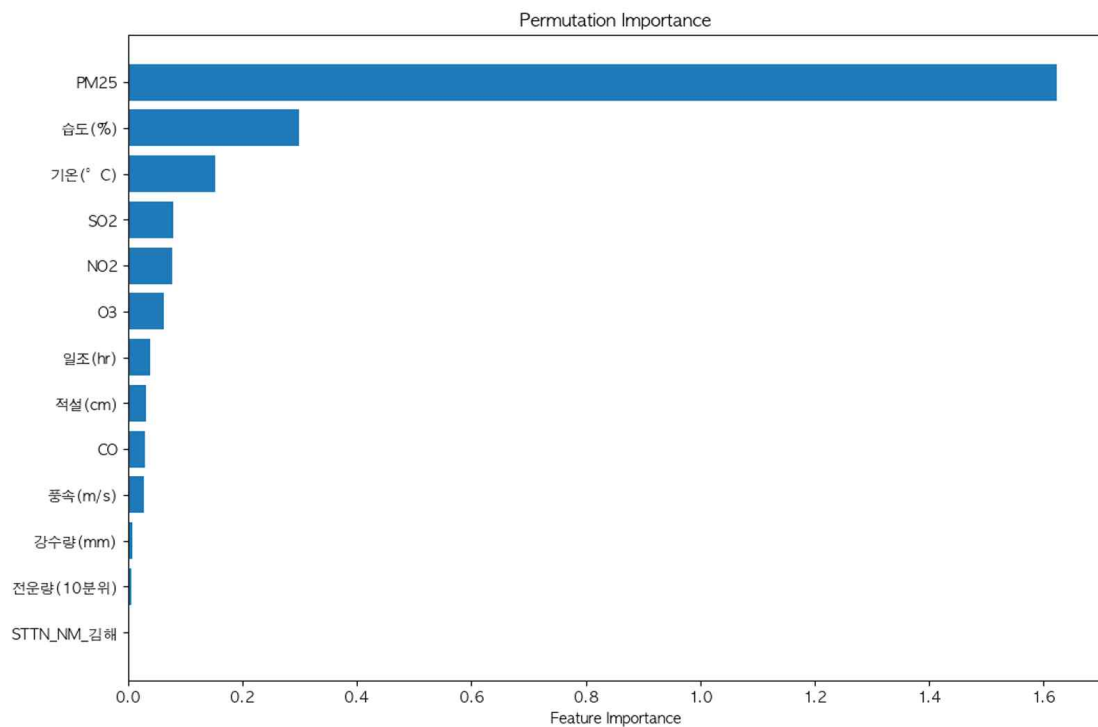
14) 표본(Sample)으로 추정한 회귀식과 실제 관측값의 차이

|그림 10| 잔차 시각화 그래프



- 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 있음
- 그림 11에서 보듯이, 미세먼지에는 초미세먼지가 압도적인 영향력을 행사하며, 그 뒤로 습도, 기온, SO2, NO2, O3, 일조, CO, 풍속, 강수량 등의 순으로 영향을 미치는 것을 확인할 수 있음

|그림 11| 머신러닝 모델의 변수 중요도



○ CNN 모델

- CNN(Convolutional Neural Network)은 다양한 분야에서 널리 쓰이는 인공지능 신경망 아키텍처¹⁵⁾로 시계열 예측 분야에서도 뛰어난 성능을 발휘함
- 확보한 대기오염물질 데이터셋과 기상 데이터셋이 각각 미세먼지 예측에 어떤 영향을 미치는지와 모델 성능을 비교하기 위해 각각의 데이터셋을 별도로 사용하거나 결합하여 비교·분석하는 실험을 진행함. 통일성을 위한 변수 통제를 위해 80:20의 비율로 훈련(Train)과 테스트(Test) 셋을 설정하였고 에포크(epoch)¹⁶⁾는 50으로 통일함
- 표4와 같이, MSE 값은 [Weather only→All data→Pollution only] 순으로 높게 나타남. MSE 값은 수치가 낮을수록 좋은 모델이므로 [Pollution only→All data→Weather only] 순으로 정확도가 높은 것임. 따라서 기상데이터는 미세먼지와 초미세먼지 분석에 중요도가 낮고 대기오염물질은 중요하다고 할 수 있으며, 미세먼지에 대기오염물질의 상관성이 높은 것을 의미함

표 4 데이터 종류별 MSE 값 비교

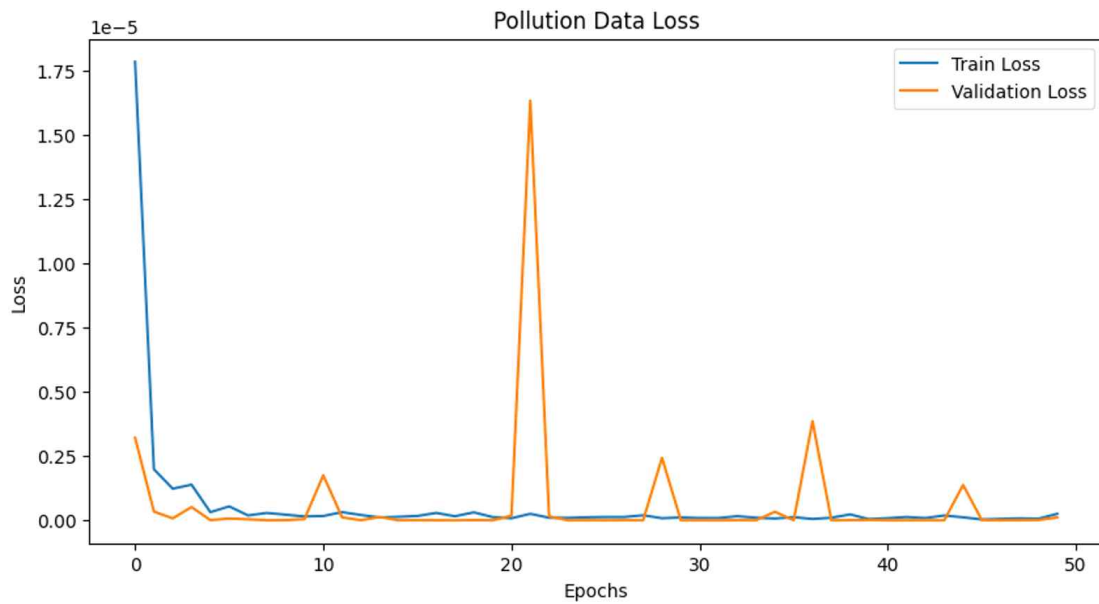
데이터 종류	MSE 값
대기오염물질(Pollution only)	1.0914277481788304e-07
기상데이터(Weather only)	6.892654482726357e-07
기상데이터-대기오염물질(All data)	3.446520580041579e-08

- 그림 12에서 알 수 있듯이, 대기오염물질 데이터의 경우 초기에는 손실 값이 높지만, 학습이 진행됨에 따라 급격히 감소하는 것을 볼 수 있음. 이는 모델이 데이터를 학습하며 예측 성능이 향상되고 있음을 의미함. 다만 특정 에포크(epoch)에서 피크값(튀는값)이 발생하고 있음을 확인할 수 있음

15) 건축학에서 건물의 구조를 건축학적으로 설계하듯 컴퓨터 공학에서는 소프트웨어의 구조(아키텍처)를 단계별로 설계하고 구축하는 것으로, 소프트웨어의 골격이 되는 기본구조임

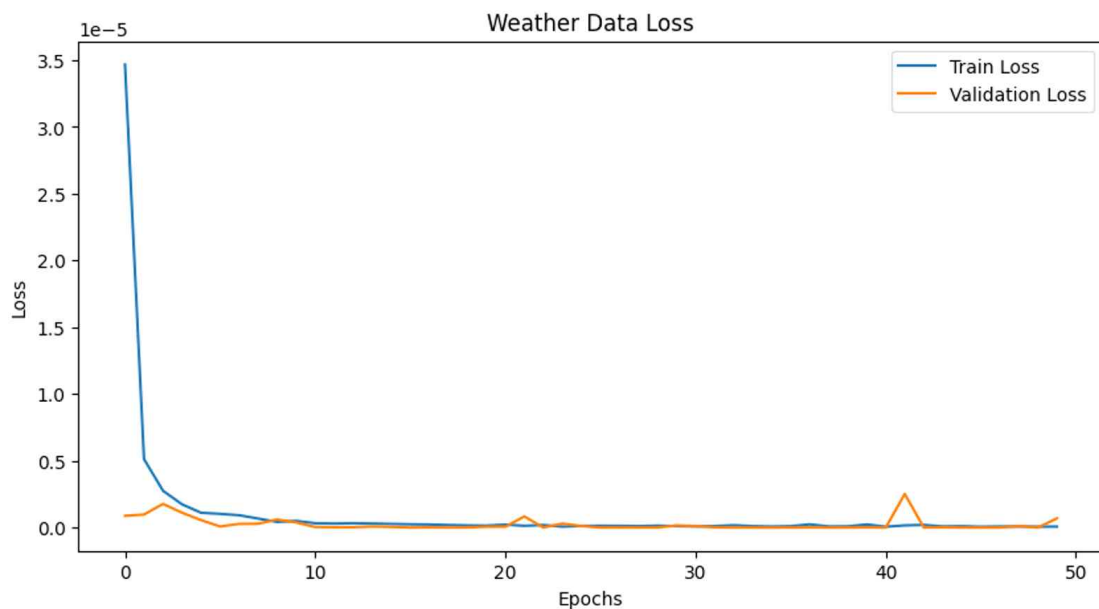
16) 훈련 데이터셋에 포함된 모든 데이터들이 한 번씩 모델을 통과한 횟수로, 모든 학습 데이터셋을 학습하는 횟수를 의미함

|그림 12| 대기오염물질 데이터 손실값 시각화 그래프



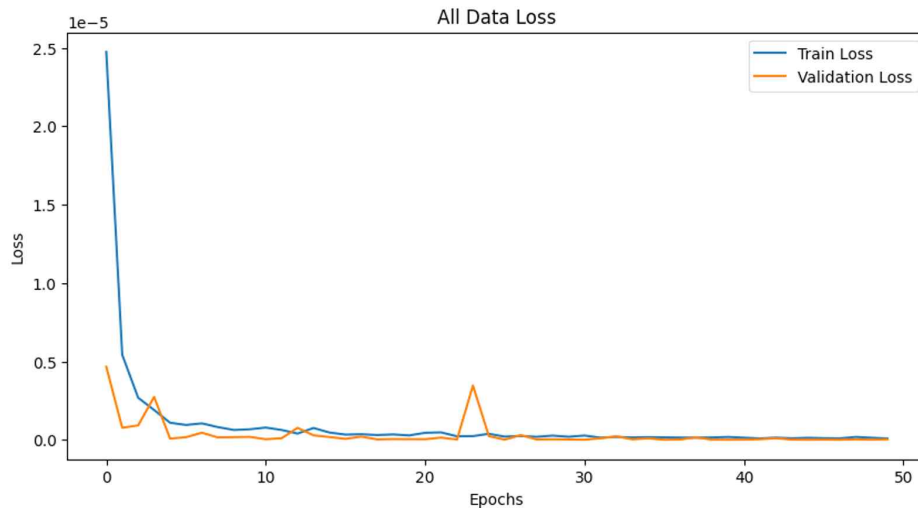
- 그림 13에서도 기상 데이터의 초기 손실 값이 높지만, 학습이 진행됨에 따라 점차 급감하는 것을 볼 수 있음. 피크값(튀는값)이 매우 약하다는 점을 통해해 전체적으로 우수한 성능의 모델임을 확인할 수 있음

|그림 13| 기상 데이터 손실값 시각화 그래프



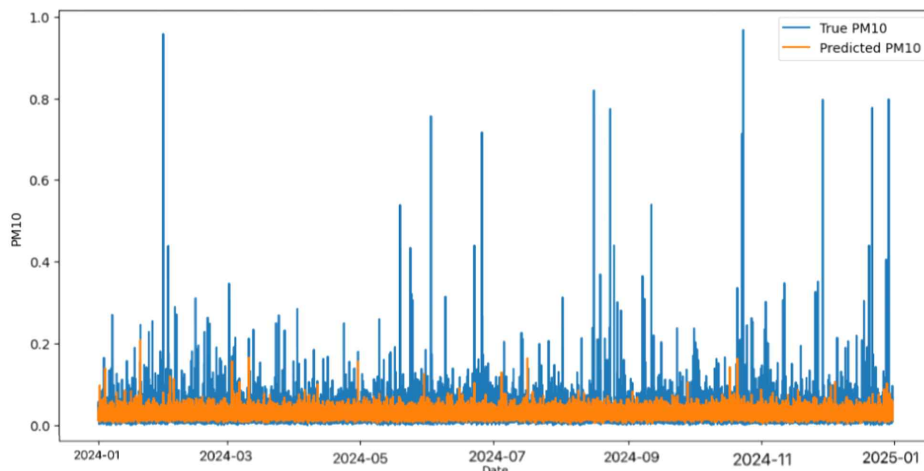
- 그림 14는 분석 결과들을 조합한 전체 데이터(기상+대기오염물질)를 사용하였고, 다양한 매개변수(파라미터)를 조정하여 2023.12.31.~2024.12.31.까지 미래 데이터를 예측하는 CNN 모델을 구축한 결과임. 체적으로 미미한 수준의 피크를 포함해 매우 양호한 수준의 모델임을 확인할 수 있음

그림 14 전체 데이터(기상+대기오염)의 손실값 시각화 그래프



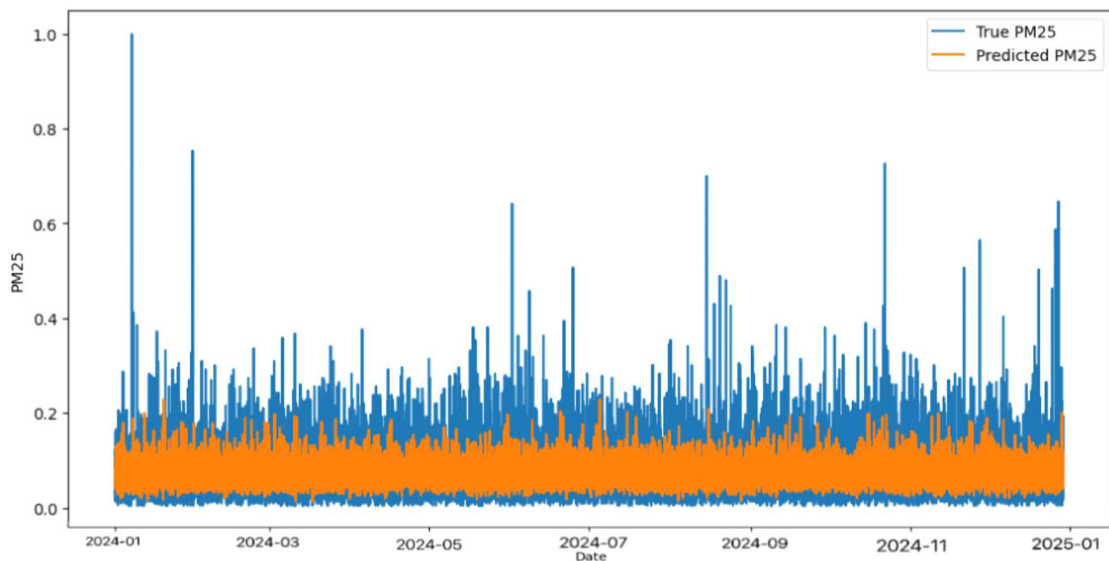
- 그림15는 미세먼지 예측 모델을 검증하기 위한 그래프임. 미세먼지 예측값(주황색)이 실제값(파란색)과 크게 벗어나지 않고 유사한 패턴을 따르고 있음을 알 수 있음. 피크값에 있어서도 어느 정도는 예측을 하고 있으나, 크게 튀는 값들에 대해서는 제대로 예측하지 못하고 상대적으로 낮은 값을 예측하고 있음

그림 15 2024.1.1.~2024.12.31.까지 미세먼지(PM10) 예측



- 그림16은 초미세먼지 예측 모델의 검증 그래프임. 이 경우에도 초미세먼지 예측값(주황색)이 실제값(파란색)과 크게 벗어나지 않고 유사한 패턴을 따르고 있음을 알 수 있음
- 하지만, 낮음~중간 수준의 피크값은 예측을 하던 PM10 예측 결과와 달리, 모델이 피크값을 제대로 예측하지 못하고 있음. 급격하게 미세먼지나 초미세먼지가 높아지는 것은 현재의 변인들만으로 패턴이나 원인을 명확하게 파악하기 어렵고 또 다른 변인들에 의해 통제받는 것을 유추할 수 있음

[그림 16] 2024.1.1.~2024.12.31.까지 초미세먼지(PM2.5) 예측



IV. 결론

1. 미세먼지와 대기유해물질 패턴 및 예측 모델에 대한 고찰

- 본 연구는 대기유해물질 6종과 기상환경 데이터 7종이 미세먼지·초미세먼지에 미치는 영향을 분석하고 있음. 분석 결과, 미세먼지 발생에 영향을 미칠 것으로 가정한 환경요인(대기유해물질 및 기상 데이터)에 대한 변수들이 경남의 환경 조건에서 모두 양의 상관관계를 보이는 것은 아님을 확인할 수 있었음
- 상관분석을 통해 변수 중요도를 측정한 결과, 미세먼지와 초미세먼지는 아주 강한 양의 상관관계를 보임. 즉 미세먼지나 초미세먼지는 하나가 증가하면 다른 하나의 수치도 증가하며, 동시에 상호 가장 주요

한 발생요인으로 밝혀짐. 또한 미세먼지와 초미세먼지는 1~4월 가장 높은 패턴을 보이는 것으로 분석되었으며, 이는 봄철 황사, 꽃가루 등이 주요한 원인임을 유추할 수 있음

- 오전 6시부터 11시까지는 자동차, 발전소, 공장에서 나오는 화학물질에 반응하는 이산화질소가 전구물질(precursor)역할을 해 일조량이 급격하게 느는 12시부터 15시까지 자외선과 광화학 반응을 일으켜 오존이 급증하는 것을 확인함. 이로 인해 미세먼지와 초미세먼지도 급증하는 패턴을 보이고 있음
- 따라서 해당 시간대에 대기오염물질 배출사업장과 공장밀집지역 등의 배출을 분산시키거나 해당 구역의 배출 상황을 집중 관리하는 등의 조치로 배출의 절대량을 줄이는 방안에 대한 논의가 필요함
- 구체적으로 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 것임

2. 대기측정망 생활권 이전을 통한 미세먼지 정보의 신뢰도·정확도 제고 필요

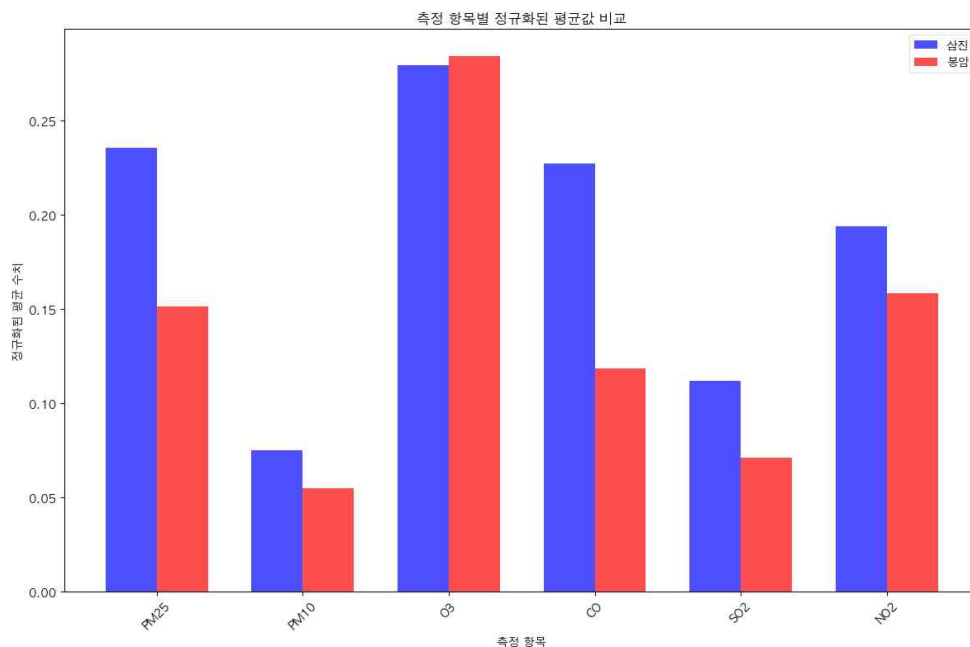
- 현행 대기환경측정망 설치·운영지침(2024)에 따르면, 시료채취구 높이는 원칙적으로 사람이 생활하고 호흡하는 높이인 지상 1.5m 이상~10m 이하로 하되, 불가피할 경우 외부조건에 최대한 영향이 적은 곳을 택하여 높이를 조정할 수 있도록 함. 이때에도 30m를 초과해서는 안 됨
- 경남도 내 대기환경측정망 47개소 중 국가측정망 5개소를 제외한 도관리·운영 측정소는 42개소임. 이중 3개 측정소만이 생활권(도로변)에서 농도를 측정하고 있고, 39개 측정소는 건물 옥상 등 생활권을 벗어난 곳에서 측정하고 있음. 현재 생활권을 이탈한 39개 측정소 중 7개 측정소만 권고 기준을 준수하고 있고, 32개 측정소는 10m를 초과하는 실정임

○ 도로변대기(삼진로)와 도시대기(봉암동)의 대표적 지역을 선정하여 대기유해물질 6종의 농도의 평균값을 비교한 결과(그림17) 오존의 경우만 유사한 수준 보이고, 초미세먼지·미세먼지·일산화탄소·아황산가스·이산화질소 모두 도로변에 설치된 측정소 농도가 압도적으로 높은 것을 확인할 수 있었음

－ 삼진로와 봉암동을 대표 측정소로 선정한 이유는 두 곳의 인구밀도가 각각 367.70명/km², 519명/km² 수준으로 유사한 데 있으며, 실제 도로변의 삼진로 밀도가 더욱 낮음에도 측정된 농도는 더 높은 수준을 보이고 있음

※ 두 측정소의 농도값을 비교한 측정기간은 도로변대기 측정을 시작한 2022.7.5. 16시부터 2023.12.31. 24시까지임

도로변대기(삼진로)와 도시대기(봉암동) 측정소 실제 농도값 비교



○ 본문의 분석 결과에서도 알 수 있듯, 차량의 통행량이 많은 시간대 배기가스 등으로 인한 대기오염이 심각해지는 만큼, 일상생활권을 벗어난 측정소의 설치에 대기유해물질의 정밀한 농도를 측정하기 어려움

○ 경남도에서 관리·운영하는 측정소의 경우, 도민들이 생활하는 높이와 근접한 낮은 곳으로 이전해 미세먼지 등 대기유해물질 농도 측정의 정확도를 높이는 조치가 필요한 것으로 사료됨

3. 대기유해물질 패턴 및 예측 모델을 활용한 미세먼지 관리 전략 수립 필요

- 본 연구의 분석결과, 미세먼지·초미세먼지가 이산화질소, 일산화탄소와 강한 상관도를 보이며 농도 변화에 직접적인 영향을 미치는 것으로 확인됨. 공학적 관점에서 1차 생성 요인을 줄이면 2차 생성물을 자연스레 줄일 수 있음을 이용하여 이산화질소와 일산화탄소의 저감을 통해 미세먼지와 초미세먼지를 예방·관리하는 조치가 필요함
- 또한 현재 경남도의 ‘미세먼지 계절관리제’는 12월 1일부터 이듬해 3월 31일까지로 설정하여 대기오염물질 배출을 줄이고 집중 관리하는 조치를 시행하고 있음
 - 하지만 분석 결과와 같이 경남도는 12월보다 4~5월 대기유해물질 농도가 높은 것으로 분석되어, 미세먼지 계절관리제 추진 기간에 대한 재논의가 필요해 보임

4. 다각적이고 복합적인 환경요인을 추가한 예측모델 개발 필요

- 미세먼지 농도 예측을 위해 다양한 대기환경 변수를 활용하여 상관관계와 중요도를 연구했음에도 불구하고, 더욱 다양한 기상 데이터나 화력 발전, 전력 발전, 항만, 중국발 대기오염물질, 지형 데이터 등 미세먼지에 영향을 줄 수 있는 요인들을 추가하지 못함. 향후 다양한 변인들을 추가하여 모델의 정밀도를 높이는 연구가 필요함
- 본 연구는 이러한 내재적 한계에도 불구하고 그동안 연구되지 않았던 경남도 환경 특성만을 고려하여 대기유해물질 및 기상환경 데이터에 대한 상관도 분석과 미세먼지 예측 모델을 개발한 데 의의가 있음

※ 참고문헌

- 보건복지부·환경부, 고농도 미세먼지 대응실무매뉴얼, 2021.
- 국립재난안전연구원, 미세먼지 재난사태 선포기준 및 대응 방안 연구, 2019.
- 양재경, 공간분석과 머신러닝 융합방법론 기반의 미세먼지 예측, 전북대학교 산학협력단, 2020.

- 안종욱 외 2인, 공간빅데이터 개념 및 체계 구축방안 연구, 한국공간정보학회, 2013.
- 이성구, 딥러닝 기반 국내 미세먼지 예측 모델링 연구, 성균관대학교 삼성학술정보관, 2019.
- 김영희 외 1인, 딥러닝 알고리즘 기반의 초미세먼지(PM2.5) 예측 성능 비교 분석, 융합정보논문지, 2021.
- 임준묵, 기상환경데이터와 머신러닝을 활용한 미세먼지농도 예측 모델, 한국 IT서비스학회지, 2019.
- 김혜림 외 1인, 기상 데이터와 미세먼지 데이터를 활용한 머신러닝 기반 미세먼지 예측 모형, 한국지리정보학회지, 2021.
- Air경남 대기환경정보, www.air.gyeongnam.go.kr.
- 에어코리아, www.airkorea.or.kr.

[부록] 경남 도시대기측정소 현황 및 시료채취구 높이(2024.7.26. 기준)

구분	시군	측정소	시료채취구 높이(m)	채취구 높이 권고 준수	위 치
도시 대기 측정 소	창원	성주동	13.5	X	창원시 성산구 외리로 14번길 18(성주민원센터)
		웅남동	12.2	X	창원시 성산구 공단로 303(효성굿스프링스)
		명서동	12	X	창원시 의창구 우곡로 101번길 28(명서2민원센터)
		용지동	18.5	X	창원시 의창구 용지로 239번길 19-4(용지동 행정복지센터)
		사파동	12.5	X	창원시 성산구 창이대로 706번길 16-23(사파민원센터)
		회원동	15	X	창원시 마산회원구 회원동 11번길 7(회원1동 행정복지센터)
		봉암동	12	X	창원시 마산회원구 봉암로 148(봉암동 행정복지센터)
		월영동	13.6	X	창원시 마산합포구 월영동 16길 22(마산합포도서관)
		내서읍	3.4	O	창원시 마산회원구 내서읍 광령로 8(삼계근린공원)
		경화동	13	X	창원시 진해구 경화로 16번길 31(병암동 행정복지센터)
	진주	상봉동	12	X	진주시 북장대로 64번길 14(중앙119안전센터)
		대안동	18	X	진주시 진주대로 1052(중소기업은행)
		상대동	16	X	진주시 동진로 279(한국전력공사 진주지점)
		정촌면	3.4	O	진주시 정촌면 예하리 1340(예하초등학교 앞, 공원)
	통영	무전동	16.5	X	통영시 안개4길 53(무전동주민센터)
	사천	사천읍	17	X	사천시 사천읍 읍내로 52(사천시니어클럽)
		향촌동	12.8	X	사천시 향촌 5길 28(향촌동행정복지센터)
	김해	동상동	12	X	김해시 호계로 517번길 8(동상동 행정복지센터)
		삼방동	18	X	김해시 활천로 303(신어초등학교)
		장유동	17.5	X	김해시 장유동 능동로 149(김해시 서부보건소)
		진영읍	13	X	김해시 진영읍 김해대로 365번길 6-24(진영읍보건지소)
		진례면	11.5	X	김해시 진례면 진례로 241(진례면 행정복지센터)
	밀양	내일동	15.5	X	밀양시 밀양시 중앙로 346(내일동행정복지센터)
	거제	아주동	4.2	O	거제시 아주동 산164-1(아주공설운동장)
		고현동	18.5	X	거제시 계룡로 125(거제시청)
	양산	북부동	14	X	양산시 북안남5길 21(중앙동 행정복지센터)
		삼호동	12	X	양산시 삼호9길 11(웅상노인복지관)
		물금읍	14.9	X	양산시 물금읍 황산로 384 (물금읍행정복지센터)
	의령	의령읍	3.4	O	의령군 의령읍 서동리 564-2(서동생활공원 지상)
	함안	가야읍	14	X	함안군 가야읍 함안대로 505(가야읍행정복지센터)
	창녕	창녕읍	13.1	X	창녕군 창녕읍 우포2로 1189-35(창녕군보건소 치매안심센터)
	고성	고성읍	12	X	고성군 고성읍 중앙로 35(고성읍보건지소)
	남해	남해읍	3.4	O	남해군 남해읍 남해대로 2745(남해유배문학관)
	하동	하동읍	17.5	X	하동군 하동읍 군청로 23(하동군청)
		금성면	9.5	O	하동군 금성면 금성중앙길 14(금성꿈나무어린이집)
	산청	산청읍	10.5	X	산청군 산청읍 옥산리 276(산청군청, 카페난나 맞은편)
	함양	함양읍	15.5	X	함양군 함양읍 고운로 35(함양군청민원봉사과)
	거창	거창읍	8	O	거창군 거창읍 대평리 1298-1(자전거교통안전교육장)
	합천	합천읍	12	X	합천군 합천읍 대야로 888-21(합천보훈회관)