기상환경 데이터와 기계학습을 활용한 미세먼지 농도 예측

사조(전다함, 신준화, 장석규, 김병찬, 박종훈)

1. 서론

미세먼지는 우리가 매 순간 마주하고 있는 대기의 한 부분으로, 그 작은 크기에도 불구하고 건강에 상당한 영향을 끼치는 요소이다. 최근 미세먼지 농도의 상승은 그 영향력을 한층 강조하며 미세먼지는 단순히 호흡기 문제만이 아니라 심혈관계 질병 등 다양한 건강 이슈를 야기할 수 있음이 입증되어 왔다. 따라서 이러한 위협을 사전에 인식하고 대응하기 위해 미세먼지 농도를 정확히 예측하는 것은 절실한 과제로 대두되었다.

이러한 배경 속에서 본 연구는 미세먼지 예측 모델의 개발에 집중하였다. 우리는 이 모델을 통해 미세먼지 노출을 최소화하고, 개인의 건강을 유지하는 데 도움을 줄 수 있는 정보를 제공하려 한다. 또한, 미세먼지의 원인과 그에 따른 농도 변화의 패턴을 이해함으로써 보다 체계적이고 효율적인 대응 방안을 마련하는데 기여하려는 목적을 가지고 있다. 본 연구는 새로운 예측 모델의 구축과 더불어 미세먼지의 복잡한 생태계에 대한 이해를 깊게 하려는 시도로써, 이는 결국 우리가 미세먼지로 인한 위협에서 보다 안전하게 보호받을 수 있게끔 도움을 줄 것으로 기대된다. 우리의 연구가 미세먼지 문제 해결의 일환으로서 기여하고, 사람들이 건강한 환경에서 생활할 수 있도록 지원하길 바라는 바이다.

2. 데이터 수집

미세먼지 농도(㎍/㎥)에 영향을 줄 것으로 보이는 다양한 데이터를 수집했다. 14개의 주요 특성들, 평균기온(℃), 일 강수량(mm), 평균 풍속(m/s), 평균 상대습도(%) 등을 포함한 데이터를 모아봤다. 지역은 춘천, 수원, 부산, 대전, 광주 이렇게 5개 지역에 초점을 맞춰 데이터를 수집했다.

(1) 기상 데이터: 평균기온, 일 강수량, 평균 풍속, 최다 풍향, 평균 상대습도, 평균 현지기압, 평균 전운량 등의 데이터를 기상청의 종관기상관측(ASOS)을 통해 2018년 5월 15일부터 2023년 5월 14일까지의 5년치 데이터를 모았다.

133	대전	2018-05-15	23.1	0	0.9	43	1002.6	7.4
133	대전	2018-05-16	25.3	0	2.4	73	998.2	9.8
133	대전	2018-05-17	26.3	0.2	1.6	61	996	9.5
133	대전	2018-05-18	20.1	9.9	1.8	85	997.6	10
133	대전	2018-05-19	17.6		3.5	47	1007.7	3.4
133	대전	2018-05-20	16.8		3.2	33	1010.9	3.6
133	대전	2018-05-21	17.5		3.1	55	1010.1	4.8
133	대전	2018-05-22	18.8	6.2	1	50	1004.2	7.6
133	대전	2018-05-23	18.9	22.4	1.8	25	1000.4	2.8
133	대전	2018-05-24	18.9		1.9	31	1002.7	1.3
133	대전	2018-05-25	19.6		1.1	29	1000.8	5.6
133	대전	2018-05-26	21.8		1.2	31	1001	0.9
133	대전	2018-05-27	20.9		1	39	1002.6	8.3
133	대전	2018-05-28	23		1.3	39	1002.1	4.9
133	대전	2018-05-29	23.6		1.1	50	1003.4	7
133	대전	2018-05-30	22.4	0	1.4	56	1002.3	6.6
133	대전	2018-05-31	20.9		1.3	41	1003.8	1.6
133	대전	2018-06-01	23.4		1.2	25	1007.1	0
133	대전	2018-06-02	23.4		1.3	21	1007.5	2.6
133	대전	2018-06-03	23.6		1.2	24	1005.5	3.5
133	대전	2018-06-04	23.5	0	1.5	35	1005	7.5
133	대전	2018-06-05	23.9	0.1	1.2	44	1003.9	7.5
133	대전	2018-06-06	24.6		1.1	42	1004	2
133	대전	2018-06-07	24.6		1	44	1003.6	4
133	대전	2018-06-08	22.5		1.3	50	1001.1	4.6
133	대전	2018-06-09	23.1	0.7	3	49	999.1	4.4
133	대전	2018-06-10	21.6	0	3.6	54	996.8	8.3

그림 1. 대전의 기상정보 데이터 중 일부

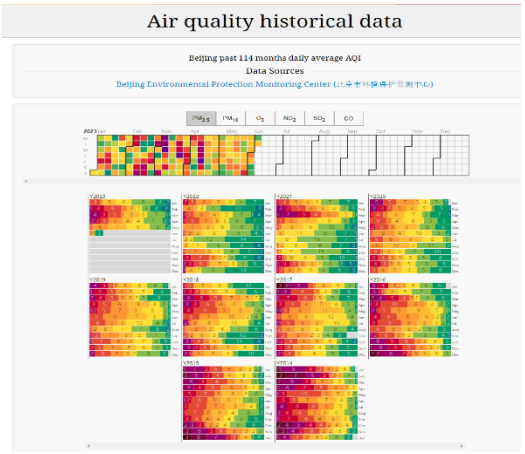


그림 2. 베이징의 일별 미세먼지 정보

(2) 중국 부유먼지, 미세먼지: 베이징의 PM2.5(미세먼지) / PM10 (초미세먼지)의 일별 데이터를 모았다. 이 데이터는 전 세계 도시의 미세먼지 데이터를 기록해 놓는 사이트에서 2018년 5월 15일부터 2023년 5월 14일까지 5년치 데이터를 수집했다.

(3) 교통량 데이터: 고속도로 공공데이터 포탈에서 분기별 톨게이트 교통량을 모은 뒤 특정 지역의 톨게이트만 필터 처리한 후에 통행량을 합쳐서 데이터를 수집했다.

(4) 화력발전량 데이터: 2022년 6월 30일까지의 LNG, 석탄 등을 포함한 화력 발전량 데이터를 수집했다. 수원과 춘천의 경우에는 각각 경기도와 강원도의 데이터로 수집하였다.

(5) 국내 미세먼지 농도 데이터: 기상청 기장자료 개방 포털에서 2018년 5월 15일부터 2023년 5월 14일까지의 5년치 데이터를 모았지만 이 데이터에서는 다른 데이터와 달리 많은 결측치가 발견되었다.

3. 데이터 전처리

(1) 데이터 결측 값 처리

강수량은 비가 오지 않은 날에만 결측 값이 생기므로 0으로 처리했다. 베이징의 미세먼지 수치는 베이징의 전체 먼지 농도 중앙값으로 결측 값을 대체했고, 교통량의 경우에는 요일별로 교통량이 달라서, 같은 요일의 데이터를 이용해 KNN 방법을 통해 결측 값을 처리했다. 모든 결측 값을 처리한 후에 데이터를 정규화했다. 일별 화력 발전량의 결측 값이 있는 행은 제거했고, 일 미세먼지 농도 역시 타겟 라벨이므로 결측 값이 있는 행은 제거했다. 습도 결측 값의 경우는 비가 오지 않은 날의 습도 중앙값으로 대체했다. 마지막으로 평균 전운량과 최대 풍향은 데이터 전체의 평균 값을 이용해서 결측 값을 처리했다.

(2) 정규화

다중 분류에는 최소-최대 정규화 (X – MIN) / (MAX – MIN)를 이용해 모든 특성들의 스케일을 동일하게 맞췄다. 회귀에는 표준화 (X - 평균) / 표준편차를 적용해 모든 특성의 평균을 0, 표준편차를 1로 만들었다. 회귀에서는 일반적으로 정답 데이터에도 표준화 처리를 해서 미세먼지 농도 데이터에도 표준화 과정을 진행했다.

개정 구간		좋 음	보 통	약간 나쁨	나 쁨	매우 나쁨	
예측 농도 (㎍/㎥,일)		0~30	31~80	81~120	121~200	201~300	301~
행동 요령	노약자	-	-	장시간 실외활동 **자제** 가급적	무리한 실외활동 **자제 요청** (특히 호흡기, 심질환자,노약자)	실외활동 **제한**	**실내** 생활
	일반	-	-	-	장시간 무리한 활동 **자제**	실외활동 **자제**	실외활동 자제

(그림 3) 분류용 데이터 프레임에서 사용한 범주화 표

(3) 타겟 라벨 처리

분류용 모델과 회귀용 모델에서 사용하는 데이터가 다르기 때문에, 회귀용 데이터프레임에서는 일 미세먼지 농도를 그대로 사용했다. 반면에 분류용 데이터 프레임에서는 개정 구간에 따라 0~4로 범주화했다.

4. 데이터 시각화 및 분석

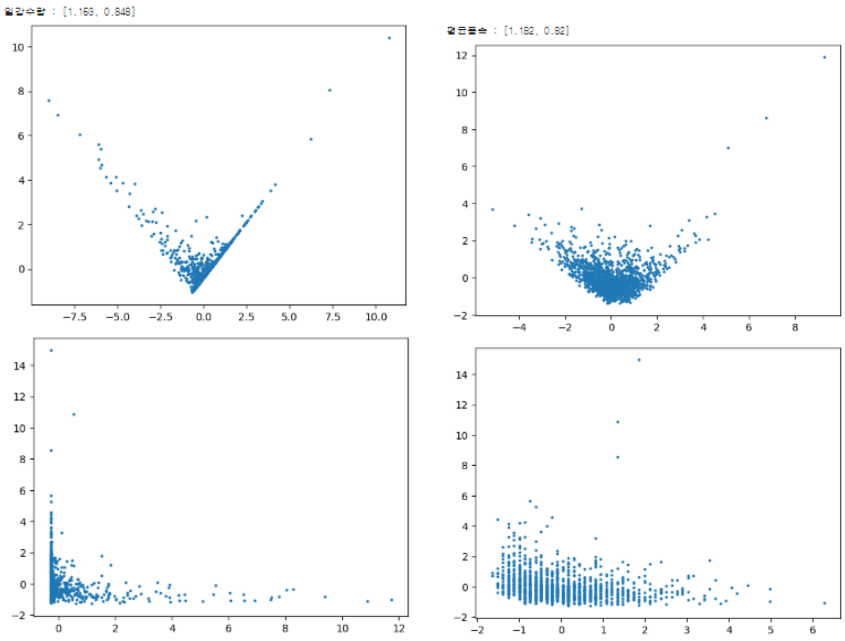
(1). 지역별 Feature, target 간의 선형성 판단

각 지역별로 모두 **regplot**을 이용해 feature, target 간 선형성을 시각화했다.

(그림 4) 데이터를 시각화해 feature들의 관계를 분석하는 과정이다.

(2). 전체 데이터 Feature, target 간 주성분 분석

PCA(Principal Component Analysis)를 이용해 다차원 데이터를 저차원으로 축소하고 데이터의 변동성을 가장 잘 설명하는 주요특성(주성분)을 추출하였다.



(그림 5) 위에는 PCA 그래프, 밑에는 x축이 데이터가 y축이 미세먼지인 그래프

5. 모델 구성 및 정확도 검증(다중 분류)

(1) 나이브 베이즈

특성 간의 독립성을 가정한 확률적 분류 모델으로, 온도, 습도, 바람의 속도와 방향 등이 미세먼지 상태에 독립적으로 영향을 미친다고 가정한다. 수원: 0.537, 부산: 0.373, 대전: 0.600, 광주: 0.681, 춘천: 0.673의 정확도를 보였다.

(2) 다중 로지스틱 회귀:

다항 분류 문제를 다루는 모델로, 각 범주의 확률을 로지스틱 함수를 통해 추정한다. 높은 상관관계가 있는 독립 변수들은 모델의 안정성과 해석력을 저해할 수 있다. 수원: 0.663, 부산: 0.658, 대전: 0.709, 광주: 0.678, 춘천: 0.713의 정확도를 보였다.

(3) LightGBM

Gradient Boosting Tree 방법을 활용해 다중 클래스 문제를 처리하는 모델이다. 범주형 변수를 지원하며 빠른 처리 속도를 보여주며 과적합을 막기 위해 Early stopping 기법을 사용한다. 수원: 0.622, 부산: 0.665, 대전: 0.707, 광주: 0.659, 춘천: 0.708의 정확도를 보였다. 부산을 제외하고는 약간의 성능 저하가 있었는데, 이는 LightGBM의 복잡성이 과적합을 유발하거나, 수원 데이터의 특정 패턴이 이 모델로 잘 학습되지 않았음을 시사한다.

(4) 다중 퍼셉트론

뉴런과 은닉층의 수를 조정하는 인공신경망으로, 복잡하고 유연한 구조를 가진다. 비선형 함수를 이용해 비선형 패턴을 학습하며, 과적합을 막기 위해 Early stopping을 사용한다. 수원: 0.692, 부산: 0.699, 대전: 0.739, 광주: 0.714, 춘천: 0.729의 정확도를 보였고 이는 안정적인 성능을 보여주며, 모델이 데이터의 다양한 부분 집합에서도 잘 일반화될 수 있음을 보여준다. 그러나 정확도가 그다지 높지는 않은데, 모델 구조의 단순성 혹은 특성의 부족 등이 원인일 수 있다.

(5) 균형 랜덤 포레스트

클래스 불균형 문제를 처리하는 데 사용되며 부트스트랩 샘플링과 소수 클래스 오버샘플링을 통해 클래스 간 균형을 맞추고 소수 클래스에 높은 가중치를 부여한다. 수원: 0.337, 부산: 0.380, 대전: 0.354, 광주: 0.336, 춘천: 0.352의 정확도를 보였다. 이 결과는 모델의 성능이 불안정하며, 지역에 따라 미세먼지에 영향을 미치는 요인이 다를 수 있음을 보여준다. 따라서 모델 훈련 시 지역별 특성을 고려하거나, 특성 공학을 이용해 새로운 특성을 생성하거나 하이퍼파라미터 튜닝이 필요할 수 있다.

6. 모델 구성 및 정확도 검증(회귀)

(1) 랜덤 포레스트

여러 개의 결정 트리를 결합한 앙상블 방법으로, 특징 간의 강한 독립성을 가정한다. 온도, 습도, 바람의 속도와 방향 등의 특징들이 독립적으로 미세먼지 상태에 영향을 미친다는 가정을 바탕으로 학습한다. n_estimators, max_depth, max_samples라는 세 가지 하이퍼 파라미터를 조정했고, 결과는 수원 MSE: 0.407, 부산 MSE: 0.400, 춘천 MSE: 0.995, 대전 MSE: 0.427, 광주 MSE: 0.482가 나왔다.

(2) K-최근접 이웃(KNN)

KNN은 인접한 점 k개를 찾아 그들의 평균을 이용하여 결과를 예측하는 알고리즘이다. 결과는 수원 MSE: 0.411, 부산 MSE: 0.357, 춘천 MSE: 1.107, 대전 MSE: 0.457, 광주 MSE: 0.429가 나왔다.

(3) 선형 회귀

선형 회귀는 각 특징들 간의 선형 상관관계를 기반으로 결과를 예측한다. 수원 MSE: 0.481, 부산 MSE: 0.404, 춘천 MSE: 1.149, 대전 MSE: 0.537, 광주 MSE: 0.455가 나왔다.

(4) 의사결정트리

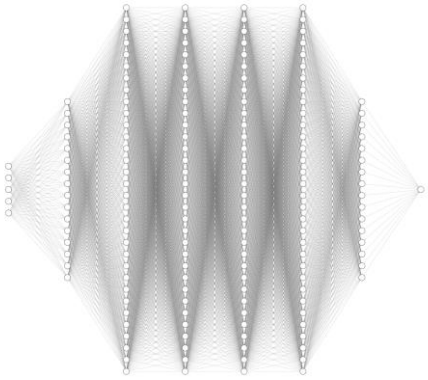
의사결정트리는 특정 기준에 따라 데이터를 구분하는 모델로, max_depth, min_samples_split, min_samples_leaf라는 세 가지 하이퍼 파라미터를 조정하였다. 결과는 수원 MSE: 0.470, 부산 MSE: 0.434, 춘천 MSE: 1.199, 대전 MSE: 0.426, 광주 MSE: 0.488가 나왔다.

(5) 서포트 벡터 머신(SVM)

SVM은 데이터를 가장 잘 나누는 선형 혹은 비선형 경계를 찾는 알고리즘으로, 커널의 종류와 오차 허용 매개변수를 하이퍼파라미터로 조정했다. 결과는 수원 MSE: 0.388, 부산 MSE: 0.313, 춘천 MSE: 1.107, 대전 MSE: 0.427, 광주 MSE: 0.392가 나왔다.

(6) 신경망(NN)

신경망은 인간 뇌의 신경 세포 구조를 모방하여 설계된 기계 학습 방법으로, 모델의 구조와 Dropout, Early stopping 등을 조정하였다. 결과는 수원 MSE: 0.384, 부산 MSE: 0.410, 춘천 MSE: 1.154, 대전 MSE: 0.451, 광주 MSE: 0.427가 나왔다.



(그림 6) 신경망 구조의 모델은 다음과 같다. 원래 사이즈는 위아래로 2배이지만, 가독성을 위해 위처럼 표시하였다.

(7) 장단기 메모리(LSTM)

LSTM은 순환 신경망의 한 종류로, 순차적인 패턴을 학습하는 데 적합하다. 모델의 구조, Dropout 비율, 학습 횟수 등을 조정하였으며 결과는 수원 MSE: 0.439, 부산 MSE: 0.380, 춘천 MSE: 1.041, 대전 MSE: 0.688, 광주 MSE: 0.605가 나왔다.

7. 실제 데이터 테스트

다음은 최종 모델로 설정한 랜덤 포레스트(Random Forest, RF), 신경망(Neural Network, NN), 장단기 메모리(Long Short-Term Memory, LSTM) 모델들을 이용해 2022년 7월 1일 이후 데이터로 진행한 추론 결과이다. 이때 월별, 요일별 평균을 이용하여 정규분포를 따르는 새로운 데이터를 생성했고, 오차율은 다음과 같이 계산했다.

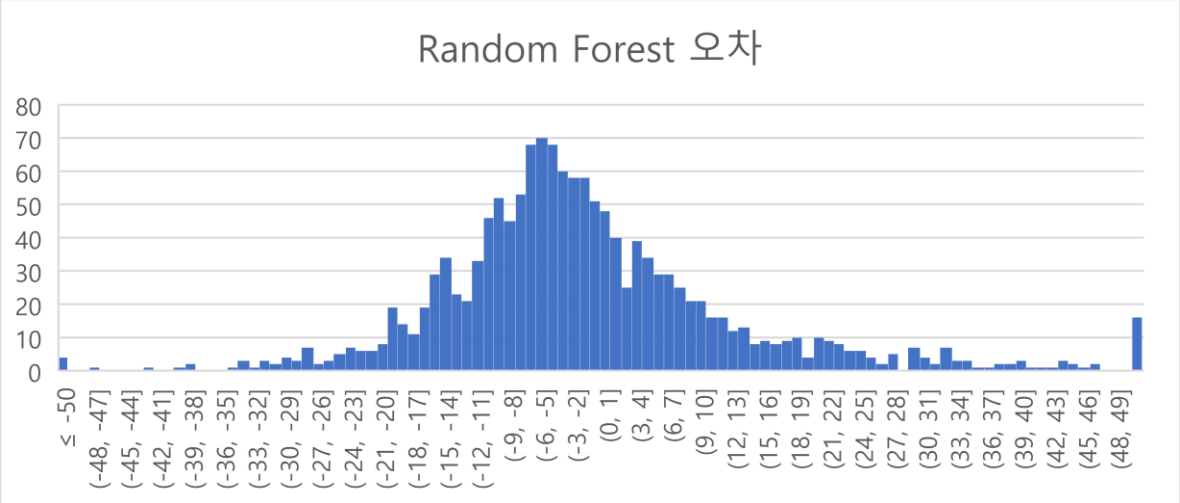
오차율 = |(실제값)-(예측값)|/((실제값))×100%

각 모델의 결과 특성은 다음과 같다.

(1) 랜덤 포레스트 (RF)

대부분의 예측값이 ±15 마이크로그램 범위 내에 있었으며 세 모델 중 가장 안정적인 결과를 보

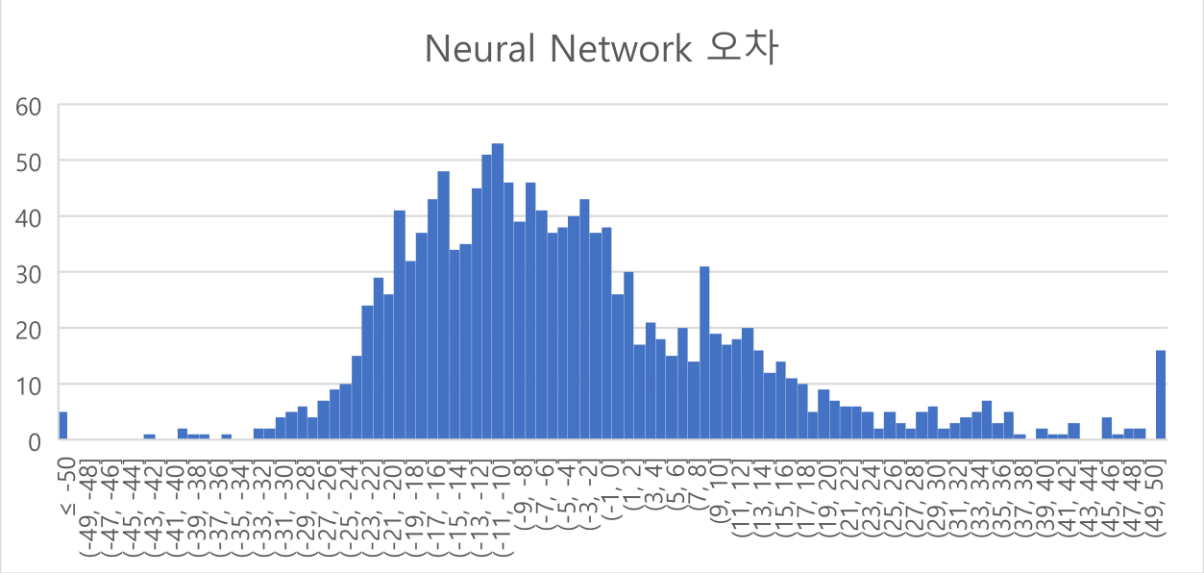
였다. 오차율은 상대적으로 높게 나타났지만, 실제 오차 수치는 수용 가능한 범위 내에 있었다. 오차율은 수원: 27.8%, 부산: 31.0%, 춘천: 19.0%, 대전: 45.0%, 광주: 33.0%가 나왔다.



(그림 7) Random Forest를 테스트했을 때의 오차이다.

(2) 신경망 (NN)

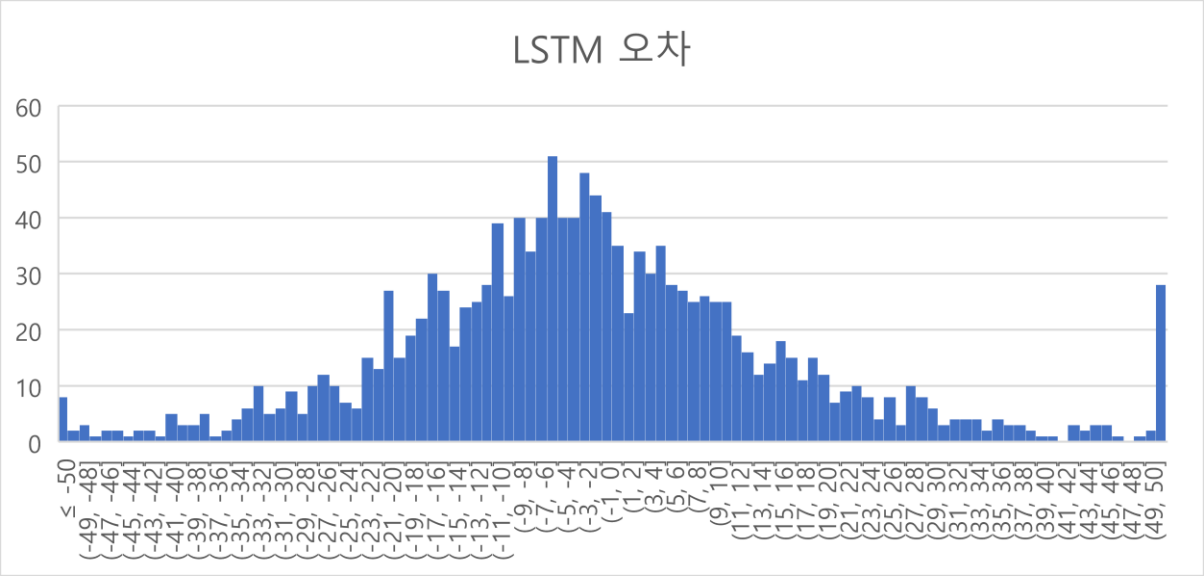
미세먼지 농도가 심각하게 높은 날(150 마이크로그램 이상)에 대한 예측이 부정확했고, 오차율은 수원 : 40.9%, 부산 35.1%, 춘천 31%, 대전 52.8%, 광주 35.9%가 나왔다.



(그림 8) Neural Network를 테스트했을 때의 오차이다.

(3) 장단기 메모리 (LSTM)

랜덤 포레스트와 유사한 결과를 보였지만, 실제로는 그렇지 않은 경우에도 미세먼지 농도가 매우 높다(150 이상)고 예측하는 경우가 있었다. 오차율은 수원 36.2%, 부산 29.2%, 춘천 29.8%, 대전 45.1%, 광주 33.5%가 나왔다.



(그림 9) LSTM을 테스트했을 때의 오차이다.

8. 결론, 한계점 및 느낀 점

(1) 결론

본 연구에서는 다중 분류 모델과 회귀 모델을 활용하여 미세먼지 예측의 최적 모델을 탐색하였고 그 결과, 랜덤 포레스트 모델이 미세먼지 분석에 가장 효과적인 것으로 판단되었다.

(2) 한계점

본 연구에서는 기상 관련 데이터 추출에는 성공하였으나, 교통량이나 화력 발전량 등 추가적인 변수들의 데이터 수집은 쉽지 않았다. 또한, 지역별 특성인 산림 면적이나 해안과의 거리 등을 고려하고 싶었으나, 다양한 지역의 데이터를 수집하고 분석하는 것은 현실적으로 어려운 점이었다.. 하이퍼파라미터 튜닝에 있어서도 제한적인 시도를 하였기 때문에, 모델의 정확도가 일부 아쉬움을 남겼고, 마지막으로 미세먼지 등급의 분포에서 0단계와 1단계의 비율이 2, 3단계에 비해 높아 데이터 불균형 문제를 겪었다.

(3) 느낀 점

본 프로젝트는 혁신적이거나 기술적 능력을 크게 향상시키는 주제는 아니었지만, 일정 계획에 따라 차근차근 목표를 달성하는 과정이 유의미하였다. 모델 설계에서 정답이 존재하지 않기 때문에, 하이퍼파라미터 설정에서 어려움을 겪었고 이로 인해 정확도가 아주 높지 않은 것이 아쉬웠으며 특히 신경망을 이용한 학습과 예측이 가장 좋은 성능을 보일 것으로 기대하였으나, 랜덤 포레스트가 가장 우수한 성능을 보였다는 사실이 놀라웠다.