미세먼지 농도 예측을 위한 머신러닝 모델 성능 비교

윤정은, 이영재*
국립안동대학교, *한국전자통신연구원
yje3058@naver.com, *lyj4295@etri.re.kr

Performance Comparison of Machine Learning Models for Prediction of Fine Dust Concentration

Jungeun Yoon, *Youngjae Lee
Dept. of Multimedia Engineering, *Regional Industry IT Conversion Team
Andong National Univ., *ETRI

요 약

미세먼지는 심혈관 및 호흡기 질환 등 건강에 악영향을 미치므로 국민들은 미세먼지에 대해 예민하게 반응하고 있으며 미세먼지의 예측 연구의 중요성이 높아지고 있다. 미세먼지를 예측하면 사전에 대응책을 마련할 수 있어 생활과 경제에 도움이 될 수 있다. 따라서 본 논문에서는 미세먼지를 예측하기 위해 기상 데이터와 대기질 데이터를 활용한다. 이를 기반으로 미세먼지를 예측하기 위해 기계학습 모델인 SVM, 앙상블 모델들을 이용하여 비교분석하고 가장 좋은 성능을 나타낸 모델은 예측에 영향을 주는 중요 특성을 분석하였다.

Ⅰ. 서 론

미세먼지에 노출되면 시력 상실 위험이 높아지며 심근경색과 만성신장질환 발생 위험까지 증가할 정도로 심혈관 및 호흡기 질환 등 건강에 악영향을 미친다. 미세먼지 노출은 사망률 증가와 연관됐으며 폐질환, 심혈관 질환 등 다양한 질환과 연관되어 있어 심각성이 더욱 크다[1]. 또한 미세먼지에 장기간 노출될 경우 면역력이 급격히 저하됨에 따라 감기, 천식과 같은 기관지염 등의 호흡기 질환에 노출되기 쉽다. 호흡기질환뿐만 아니라 심혈관 질환, 안구질환 등 각종 질병에 노출될 수 있다. 특히 초미세먼지는 직경 2.5㎛ 이하의 입자를 가지고 있어 체 내 기관지 및 폐 깊숙한 곳까지 침투하여 각종 질환을 유발할 수 있으므로 미세먼지의 위험성 및 심각성이 더욱 크다[2].

미세먼지를 동반한 대기오염은 국내의 주요 문제 중 하나이며 미세먼지의 악영향으로 인해 건강에 미치는 우려가 커지고 있다. 이에 따라 대기오염 측정에 대한 중요성이 더욱 강조되고 있으며 더 높은 정확도의 예측 모델을 요구하고 있다. 따라서 본 논문에서는 더 높은 정확도의 예측 모델을 개발하기 위해 머신러닝을 활용할 것이며 미세먼지 농도에 영향을 끼치는 특성들을 크게 2가지 특성으로 분류하여 해당 데이터를 사용한 머신러닝 모델을 실험하고 예측에 가장 영향을 많이 주는 특성에 대해 분석 연구를 진행하였다.

Ⅱ. 본론

표 1. 데이터 수집 내용

데이터 제공	데이터
에어코리아	SO2, NO2, O3, CO, PM10, PM2.5
기상자료개방포털	기온, 풍속, 풍속, 풍향, 습도, 증기압, 이슬점 온도, 현지 기압, 해면 기압, 전운량, 중하층운량, 시정, 지면 온도

본 논문은 안동지역의 미세먼지를 예측하기 위해서 대기 환경 데이터와 기상 데이터를 2019년 12월부터 2022년 6월까지 1시간 간격으로 측정된 데이터를 사용하였다. 표 1과 같이 대기 환경 데이터는 에어코리아에서 제공하는 데이터를 사용하였으며, 기상 데이터는 기상자료개방 포털의 데이터를 수집하였다.

본 논문에서 사용된 머신러닝 모델은 SVM(Support Vector Machine), 앙상블 모델인 RF(Random Forest), GBM(Gradient Boost Machine), XGBoostt(eXtreme Gradient Boosting)이며 연구 방법은 그림 1과 같다.

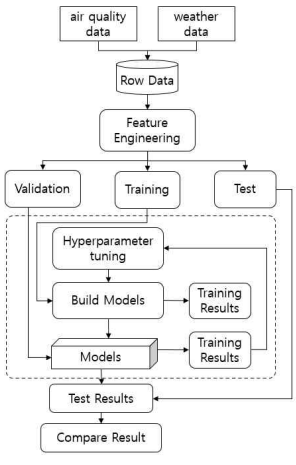

그림 1. 미세먼지 예측 모델 구조

데이터는 90:10 비율로 2019년 12월부터 2021년 12월 데이터를 학습(train)데이터로 사용하였으며 테스트(Test)데이터는 2022년 6월 데이터로 사용하였다. 모델별로 미세먼지 예측을 비교하기 위해 각 모델은 그림 1과 같이 파이프라인을 구성하여 전처리 및 최적의 매개변수를 찾는다. 교차검증을 통해 최적의 매개변수를 찾고 매개변수를 활용하여 테스트 데이터로 검증한다. 각 모델을 평가하고 비교·분석하기 위해 사용된 성능지

표는 MSE, RMSE, MAE, R^2를 이용하였다. 표 2와 그림 2는 4가지 머신러닝 모델을 사용한 미세먼지 농도 예측 결과이다. 표 2와 같이 MAE를 기준으로 XGBoost 모델이 가장 잘 예측한 것을 볼 수 있다. MSE, RMSE, R^2는 XGBoost, GBM 모델이 동일한 값으로 가장 잘 예측한 것을 볼 수 있다. 표 3은 기존 미세먼지 예측 연구의 결과로 표2와 비교하여 기존 연구 결과보다 향상된 예측 성능을 보여준다.

표 2 . PM2.5 예측 모델 성능 평가

Model	MAE	MSE	RMSE	R^2
SVM	4.23	32.37	5.69	0.75
RF	4.24	33.74	5.81	0.74
GBM	4.07	*31.25	*5.59	*0.76
XGBoost	*4.02	*31.25	*5.59	*0.76

표 3. 기존 미세먼지 예측 연구 결과

Model	Outcome	Literature
SVM	RMSE = 12.29	Kim et al.,[3]
RF	MAE = 4.48, RMSE = 6.12	Shim., S[4]
GBM	R^2 = 0.654	Suh., K[5]
XGBoost	RMSE = 6.41	Won et al.,[6]

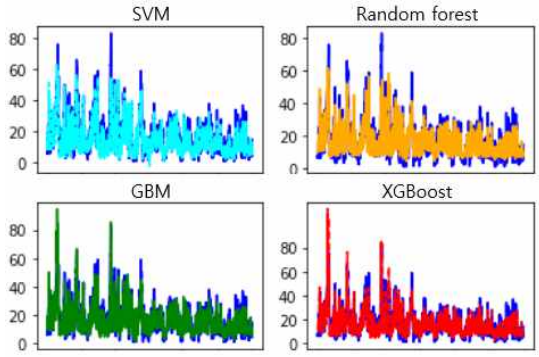

그림 2 . PM2.5 예측 모델 성능 평가 그래프

그림 2의 x축은 timeline으로 2022년 1월부터 2022년 6월이며 y축은 미세먼지의 예측값이다. 그림 2의 그래프를 보면 GBM과 XGBoost는 SVM, RF 모델보다 2022년 3월에서 5월까지의 데이터를 더욱 잘 예측하는 것을 볼 수 있으며 성능지표도 다른 모델이 비해 좋은 지표를 보여주었다.

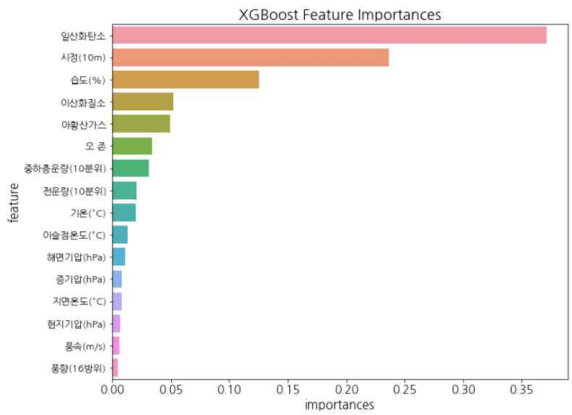

그림 3 . XGBoost 모델의 특성 중요도

그림 3은 가장 좋은 성능을 보였던 XGBoost 모델의 특성 중요도 그래프로 미세먼지 예측에 영향을 미치는 중요한 특성을 분석하기 위한 그림이다. 그림 3에서 알 수 있듯이 미세먼지에 가장 많이 영향을 주는 특성은 일산화탄소, 대기의 혼탁 정도를 나타내는 시정, 습도가 예측에 많은 영향을 주는 특성임을 알 수 있다. 또한 이산화질소, 아황산가스, 오존 특성과 같은 대기질 데이터가 대체로 영향이 있는 것을 알 수 있다.

Ⅲ. 결론

미세먼지의 위험성이 대두됨에 따라 예측 연구에 관한 중요성이 더욱 강조되고 있으며 더 높은 정확도의 예측 모델을 요구하고 있다. 따라서 본 논문에서는 크게 기상, 대기질 2가지 특성으로 분류한 데이터를 사용하여 4종류의 머신러닝 모델을 실험하였으며 예측에 가장 영향을 많이 주는 특성에 대해 분석 연구를 진행하였다.

4가지 머신러닝 모델의 예측 성능을 비교 분석한 결과 미세먼지를 예측하기 위해 사용된 모델 중 XGBoost는 MAE 4.02, MSE 31.25, RMSE 5.59, R^2 0.76으로 가장 좋은 성능을 보였다. 또한 기존 미세먼지 예측 연구 대비 향상된 예측 성능을 보여주었다. 미세먼지 예측에 영향을 주는 특성을 분석한 결과 일산화탄소, 시정, 습도가 많은 영향을 주는 편이었고 이외의 이산화질소, 아황산가스, 오존 등 대기질 데이터가 미세먼지를 예측하기 위한 특성으로 영향이 있음을 확인할 수 있었다.

향 후 본 논문에서 활용한 데이터는 시계열 기반 데이터이므로 시계열 데이터에 특화된 딥러닝 모델인 CNN, RNN, LSTM 모델을 활용하여 예측 연구를 진행하고자 한다.

ACKNOWLEDGMENT

본 논문은 안동지역 '중소기업 ICT 융합기술 경쟁력 강화 사업 (23AD1100)'의 연구결과로 수행되었음

참 고 문 헌

[1] Yeo, M., Kim, Y., "Trends of the PM10 concentrations and high PM10 concentration cases in Korea," Journal of Korean Society for Atmospheric Environment, pp. 249-264, Oct. 2022.

[2] Lee, H., Yoon, T., Lee, Y, . "Empirical Analysis of Fine Particular Matter (PM2.5) Concentrations and Domestic Tourism Demand," Korean Energy Economic Review, pp. 161-185, Seb. 2022.

[3] Kim, M., Jeong, H,. "Development of machine learning based prediction of particlulate matter concentration in Seoul.," Journal of the Korean Data And Information Science Society, pp. 1095-1111, Nov. 2022.

[4] Shim, S., "Performance Comparison of Machine Learning-Based Particulate Matter Prediction Models Using Meteorological Factors and Air Pollution Material Factors Data", master's thesis, Sejong University, Feb. 2020.

[5] Suh, K., "Artificial Intelligence-Based Sensor Data Prediction Technique for Improving the Accuracy of Low-cost Ultra-fine Dust Sensor." Proceedings of KIIT Conference, pp. 179-183. Oct. 2022.

[6] Won, D., Kim, S., Kim, Y., Song, G., "Prediction of Fine Dust in Gyeonggi-do Industrial Complex using Machine Learning" Methods. Journal of KIISE, pp. 764-773, Jul. 2021.