Regular paper | Journal of the KIECS. pp. 531-538, vol. 19, no. 3, Jun. 30. 2024, t. 125, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2024.19.3.531

머신러닝 학습 알고리즘을 이용한 광주천 수질 분석에 대한 예측 모델 연구

정유정* · 이정재**

A Study on the Prediction Model for Analysis of Water Quality in Gwangju Stream using Machine Learning Algorithm

Yu-Jeong Jeong* · Jung-Jae Lee**

요 약

수질 환경의 중요성이 강조되고 있는 가운데 광주광역시 도시 하천의 수질개선을 위한 수질 지표는 수생태계에 영향을 미치는 중요한 요소로 정확한 예측이 필요하다. 본 연구에서는 XGBoost와 LightGBM 머신러닝 알고리즘을 활용하여 광주천의 중요한 지점인 하류 평촌교(PyeongchonBr)와 상류 방학교(BangHakBr_Gwangjucheon1) 수계의 수질 검사 항목 중 통계적 검증 결과 유의미한 항목인 질소(TN), 질산염(NO3), 암모니아 양(NH3) 세 가지 수질 지표를 예측하는 연구를 수행하였고, 회귀 모델 평가 지표인 RMSE를 이용하여 예측 모델의 성능을 평가하였다. 수계별 개별적인 모델을 구현하여 교차 검증 후 성능을 비교한 결과, XGBoost 모델이 뛰어난 예측 능력을 보였다

ABSTRACT

While the importance of the water quality environment is being emphasized, the water quality index for improving the water quality of urban rivers in Gwangju Metropolitan City is an important factor affecting the aquatic ecosystem and requires accurate prediction. In this paper, the XGBoost and LightGBM machine learning algorithms were used to compare the performance of the water quality inspection items of the downstream Pyeongchon Bridge and upstream BanghakBr_Gwangjucheon1 water systems, which are important points of Gwangju Stream, as a result of statistical verification, three water quality indicators, Nitrogen(TN), Nitrate(NO3), and Ammonia amount(NH3) were predicted, and the performance of the predictive model was evaluated by using RMSE, a regression model evaluation index. As a result of comparing the performance after cross-validation by implementing individual models for each water system, the XGBoost model showed excellent predictive ability.

키워드

Machine Learning, Boost algorithm, LightGBM, Gwangju Stream, Water Pollution
머신러닝, 부스팅 알고리즘, 광주천, 수질 오염, LightGBM

* 호남대학교 AI교양대학 교수(narimono@naver.com)
** 교신저자 : 송원대학교 컴퓨터정보학과
• 접 수 일 : 2024. 04. 18
• 수정완료일 : 2024. 05. 15
• 게재확정일 : 2024. 06. 12
• Received : Apr. 18, 2024, Revised : May. 15, 2024, Accepted : Jun. 12, 2024
• Corresponding Author : Jeong-Jae Lee
Computer Information Division, SongWon University,
Email : jjalee@songwon.ac.kr

Ⅰ. 서 론

수질 환경은 우리의 삶과 생태계에 미치는 영향이 커지고 있는 중요한 요소로 도시 지역에서는 수질 관리가 더욱 중요하며, 하천은 이러한 도시 지역에서 많은 인구와 산업 활동에 노출되어있다. 생태계적인 환경변화로 인한 자연적이고 인위적인 인자로 인하여 오염이 증가하면서 수질 악화 문제가 크게 대두되고 있으며 이로 인한 녹조 현상도 심각하게 발생되고 있는 경우도 있다. 녹조 발생을 미리 경보하기 위한 인공지능 알고리즘 개발에 필요한 데이터 불균형 문제를 해결하는 방안을 주요 연구 대상으로 하는 연구[1]와 실시간 강우량을 입력 데이터로 사용함으로써 수질데이터의 변화량을 측정하기 위해 기존 설계 강우량 사용 시 실시간성을 반영하지 못하는 단점을 대폭 개선하는 연구도 진행되고 있다[2].

특히 광주광역시의 하천인 광주천은 이러한 수질 문제로부터 영향을 받고 있다. 광주광역시는 한국의 중요한 도시 중 하나로, 농업과 산업 등 다양한 활동이 활발하게 이루어지고 있다. 이에 따라 수질 오염 또한 산업 배출물과 생활 폐수 등으로 인해 심각한 수준으로 증가하고 있으며 특히 광주천은 광주광역시의 중심부를 흐르는 주요 하천으로서, 주변 지역의 산업 활동과 인구 밀도가 높아 수질 문제가 심각한 수준에 이르고 있다. 수질 지표 예측은 환경 보전 및 생태계 관리에 있어서 중요한 주제로 여러 연구들이 진행되어왔다. 기존 연구들은 DO(:Dissolved Oxygen), BOD(:Biochemical OxygenDemand), COD(:Chemical Oxygen Demand)을 통계적 모델을 활용하여 수질 지표를 예측하였다.

본 연구에서는 기존의 DO(:Dissolved Oxygen), BOD(:Biochemical OxygenDemand), COD(:Chemical Oxygen Demand)에만 수질분석을 의존하였던 문제점을 보완하기 위해 광주천 상류지점과 하류지점 측정 데이터 중 유의미한 데이터를 찾아 통계적 검증에 의해 분석 후 XGBoost와 LightGBM(:Light Gradient Boosting Machine)[3]이라는 최신 머신러닝 알고리즘을 적용하여 광주천의 수질을 예측하고, 수질 개선에 기여할 수 있는 방안을 모색하고자 한다.

머신러닝(Machine Learning) 알고리즘은 성능 최적화 뿐만 아니라 다양한 산업 분야에서 최적화 문제에 적용하기 위해 응용되고 있다[4].

실험에 사용할 데이터는 공공 포털 데이터 광주광역시 광주보건환경연구원 수질측정망 데이터 2020년에서 2022년 월별 측정 데이터 21개소 취수원 중 광주천을 흐르는 하류, 상류 평촌교(PyeongchonBr)와 방학교(BangHakBr_Gwangjucheon1) 두개의 수계로부터 다양한 데이터를 수집 받아 수질관리에 있어서 중요한 항목인 18개 항목은 데이터들을 일정한 시간 간격으로 측정된 것으로, 각 수질 지표들에 대한 값을 포함하고 있다. 통계적 검증을 통해 유의미한 데이터를 가지고 있는 TN(:Total Nitrogen), NO3(:Nitrate), NH3(:Ammonia) 세 가지 수질 지표를 예측하는 연구를 수행하였다. 이 세 변수들을 모델링 및 예측에 활용하여 수질의 변화를 예측하여 중요한 정보를 제공 가능하도록 하기 위해 데이터 전처리를 수행하여 수질 지표와 해당 수질 데이터를 분리하였고 XGBoost와 LightGBM(:Light Gradient Boosting Machine) 모델의 하이퍼파라미터를 최적화하는 작업을 진행한 후 그리드 서치 기법을 활용하여 최적의 하이퍼파라미터 조합을 찾은 후 모델의 성능 평가를 위해 평촌교(PyeongchonBr)와 방학교(BangHakBr_Gwangjucheon1) 두 개의 수계로부터 수집된 데이터를 각각 분리하여 모델 학습에 활용하였다. 이후 XGBoost와 LightGBM모델을 학습시켜 TN, NO3, NH3 예측 모델을 개별적으로 구현하였다. 회귀 모델 평가 지표인 RMSE(:Root Mean Squared Error)를 사용하여 예측 모델의 성능을 평가 하였다.

Ⅱ. 관련 연구

2.1 수질 환경의 오염 지표 성분

수질 오염 총량관리란 과학적 토대 위에서 수계 구간별 목표 수질을 설정하고, 목표 수질을 달성·유지하기 위한 허용 부하량을 산정하여, 해당 총량관리 단위유역 내에서 배출되는 오염물질의 총량이 목표수질을 달성할 수 있는 허용 부하량 이내로 규제 또는 관리하는 제도를 말한다[5].

목표수질은 총량관리 목표 설정을 위한 기준치로서 하천의 용도(상수원수, 농업용수 등), 오염원 밀도, 지역개발 정도, 환경기초시설 투자 정도, 수량 및 수질,

수중 생태계의 건전성 등을 고려하여 수계 구간별로 설정하며, 수질평가는 산정 시점으로부터 과거 3년간 측정한 결과를 토대로 산식에 따라 평균 수질을 산정하여 당해 목표수질 지점의 수질 변동을 확인한다[5].

국내 수질 오염 총량 관리제와 관련한 연구사례를 살펴보면 김규완은 의암댐 수질자료 중 pH, DO, T-N, T-P를 선택하여 변동 특성을 분석하고 수질개선의 필요성을 파악하기 위해 비모수적 통계방법 중 맨-켄달 검정법을 사용하여 수질변화의 통계적 경향성을 분석하여 정량적인 수질자료를 제시하였다[6].

수질 성분은 다음과 같은 지표로 구분한다. 먼저 수온은 기온의 변화에 의해 결정되지만, 생활하수 등에 영향을 크게 받는 도심하천의 경우, 농촌 하천보다 상대적으로 수온 변화가 크게 나타난다.

TN(:Total Nitrogen)은 물에 함유된 전체 질소의 양을 나타내는 지표로 질소는 생물의 성장에 필요한 영양소이지만, 과도한 질소 배출은 물의 오염을 유발할 수 있다. TN은 물의 영양 상태를 평가하는 데 사용되며 농업 및 산업 활동으로 인해 증가할 수 있다. NO3(:Nitrate)는 물에 함유된 질산염의 양을 나타내며 질산염은 농업 및 산업 활동으로 인해 물에 노출될 수 있으며, 수질 오염의 주요 원인 중 하나이다.

NH3(:Ammonia)는 물에 함유된 암모니아의 양을 나타내며 암모니아는 물의 질을 평가하고 질소 오염의 정도를 추정하는 데 사용된다. 높은 NH3 값은 물의 오염 정도가 높을 수 있음을 나타낸다.

2.2 XGBoost

XGBoost는 트리 기반의 앙상블 학습에서 널리 알려져 있는 알고리즘 중 하나이다. 부스팅 기법을 이용하여 구현한 알고리즘은 AdaBoost(:Adaptive Boosting)와 Gradient Boost가 있다. XGBoost는 GBM(:Gradient Boosting Machine)에 기반하고 있지만, GBM은 학습 속도가 느리고 과적합의 문제를 발생시키는 단점이 있다[7]. 이를 보완하기 위해 더 빠른 속도를 보여주고 자체 과적합 규제 기능으로 과적합에 좀 더 강한 내구성을 가지고 병렬 학습이 지원되는 라이브러리가 XGBoost이다[8].

XGBoost는 먼저 약한 예측 모델을 구성한 뒤 학습 셋과 일치도를 평가한다. 이후 경사 하강법을 이용하여 일치도가 높아지는 방향의 기울기를 설명변수로 하는 새로운 약한 예측 모델을 구성한다.

앞선 과정을 반복하여 최종적으로 여러 예측 모델을 앙상블 하여 이때 활용하는 오차 함수에 따라 다양한 예측 모델이 다르게 나타날 수 있는데 예측 오차의 최소 자승 형태의 오차 함수를 활용한다면 그 기울기는 예측 과정의 잔차가 되어 회귀 함수를 단순화할 수 있다[9].

XGBoost 모델을 수식적으로 표현할 때 약한 예측 모델의 파라미터를 구하는 과정은 식 (1)과 식 (2)와 같다. $m-1$번째에서 m번째의 앙상블 모델로 업데이트 하는 과정이며, F_m는 m번째의 앙상블 모델을, f_m은 m번째의 약한 예측 모델을, γ_m은 m번째 모델을 업데이트하기 위한 승수(Multiplier)를 의미한다. i은 Training set의 관찰 값 수를, j는 m번째 모델의 잎의 숫자이다[10].

$$l=(y_i,F_{m-1}(x_i))=\frac{(y_i-F_{m-1}(x_i))^2}{2} \quad \cdots(1)$$

$$F_m(x)=F_{m-1}(x)+\gamma_m f_m(x) \quad \cdots(2)$$

XGBoost는 결정 트리의 분화 적정성을 평가하는 목적함수에 페널티 항을 추가하는 방식으로 이루어지며 식 (3)과 같다. 페널티에 대한 정리는 식 (4)와 같다. 식 (4)에서 T_m는 m번째 모델의 잎의 숫자, w는 잎의 값을 나타내는 벡터이며 α와 λ는 각각 정규화 상수이다[1].

$$L=\sum_i l(y_i,F_{m-1}(x_i)+f_m(x_i))+ohm(f_m)\cdots(3)$$

$$ohm(f_m)=\alpha T_m+\frac{1}{2}\lambda \parallel w \parallel^2 \quad \cdots(4)$$

2.3 LightGBM

LightGBM은 Microsoft에서 개발한 머신러닝 알고리즘으로 XGBoost와 함께 부스팅 계열 알고리즘에서 많이 사용되고 있는 GBDT(:Gradient Boosting Decision Tree) 기반 앙상블 모델 중 하나이다[11]. LightGBM은 분류와 회귀 등 여러 학습 작업에 사용이 되며 효율적인 병렬

1) https://xgboost.readthedocs.io

학습이 지원된다. LightGBM은 모형을 구축할 때 사용하는 변수와 자료의 수를 줄여 작업에 소요되는 시간을 줄이기 위해 GOSS(:Gradinet based One Side Sampling)와 EFB(:Exclusive Feature Bundling) 2가지 알고리즘을 적용하여 모형 구축에 소요되는 시간을 줄이고, 충분한 예측성능을 유지할 수 있도록 구성되었다[12].

III. 실험 결과

3.1 대상 지역 현황

광주지역의 수계망 발달은 지질 분포와 지질구조 등의 영향을 크게 받고 있으며, 영산강 본류는 양 산맥과 평행하게 NE-SW 방향으로 조립질 화강암류 지역을 사행하면서 충적평야를 관류하고 있으며, 무등산에서 발원하는 광주천은 풍영정천, 황룡강 등과 합류하여 나주평야로 흘러간다. 이러한 수문 지형·지질적 조건은 지하수의 유동계를 지배하는 중요한 요인 중에 하나로(최희철, 2002) 대수층의 토양이나 암석의 종류, 공극의 크기 및 형태 등은 투수계수와 저류계수 등과 같은 수문 지질적 특성을 결정한다[13]. 광주천은 남동-북서 방향으로 흐르기 때문에 금남로, 충장로 등이 광주천에 평행한 방향으로 뻗고, 이 도로에 교차하는 중앙로, 광남로 등은 북동-남서방향으로 달리게 되어 광주천의 흐름 방향이 도시 간선도로의 방향을 결정하는 데 영향을 미친 것이다[14]. 본 연구에서는 상류와 하류 수질 예측 측정값이 유의미한 의미를 가지고 있고 정확한 수질 예측 변화가 있는지 연구하기 위해 무등산에서 내려오는 증심천이 합류하는 상류 방학교 수계와 광주광역시 북구 충효동 평촌교로 흘러 영산강으로 합류하기 전 하류 지점의 수계 데이터 측정 자료를 사용하였다.

3.2 모델링

수질 농도 예측하기 위한 모델의 데이터 셋은 국가통계포털 사이트에서 데이터를 수집하였으며, 데이터 셋은 공공포털데이터 광주광역시 광주 보건환경연구원 수질측정망 데이터 2020년에서 2023년 월별 측정 데이터 21개소 취수원 중 광주천을 흐르는 하류, 상류의 대표적인 평촌교(PyeongchonBr)와 방학교(BangHakBr_Gwangjucheon1) 두 개의 수계로부터 TN, NO3, NH3 지표들의 변화를 포함한 다양한 데이터를 수집하였다. 본 연구에서는 두 개 그룹 18개의 수질 검사 항목을 t-검정을 통해 TN p-value :0.000145, NO3 p-value:0.000431, NH3 p-value: 0.000204 결과로 유의미한 p-value를 가지고 있으므로 XGBoost와 LightGBM 분석에 적합한 파라미터 인자로 선택하여 학습을 진행하였고, XGBoost와 LightGBM 모델을 학습시켜 TN, NO3, NH3 예측 모델을 개별적으로 구현하였다. 두 모델의 예측 성능을 비교하기 위해 각 수계별로 학습된 모델을 교차 검증 수행 후 RMSE로 평가하였다.

3.3 모델링 전처리

데이터 셋을 확인하기 위해 수계별 그룹으로 평균을 이용하여 수질의 오염도 특성을 확인하였다. 모델링 결측치를 잘못된 값으로 채운다면 편향된 추정치와 왜곡된 통계력 및 결론을 얻을 수 있다 .결측치를 처리하기 위해서는 누락된 데이터를 제거하거나 결측치를 특정 값으로 보간하는 방법 등 다양한 방법이 존재한다[15].

본 연구에 사용되는 학습 데이터와 검증, 평가 데이터의 결측치를 처리하기 위해 데이터의 mean, std, min, 사분위수 25% , 50% , 75% 값을 고려하여 실행하였다. 그림 1에서 평균(mean) 과 중간 값을 비교하면 왜곡의 정도를 파악할 수 있다.

PyeongchonBr Group describe

	TN	NO3	NH3
count	43.000000	43.000000	43.000000
mean	3.947860	2.336233	0.571349
std	1.741089	1.241162	0.930581
min	1.208000	0.306000	0.000000
25%	2.706500	1.567500	0.101500
50%	3.734000	2.222000	0.188000
75%	4.841000	3.117500	0.768500
max	9.172000	5.649000	4.533000

BangHakBr_Gwangjucheon1 Group describe

	TN	NO3	NH3
count	43.000000	43.000000	43.000000
mean	7.448070	3.553349	2.034860
std	4.931045	2.181005	2.352046
min	0.879000	0.242000	0.021000
25%	2.751500	1.855000	0.186000
50%	7.498000	3.209000	1.374000
75%	11.231000	4.745500	2.863500
max	18.570000	11.714000	11.465000

그림 1. 평균, 표준편차, 최솟값, 최댓값, 25%,50%, 75% 사분위수 데이터.

Fig. 1 mean, standard deviation, minimum, maximum, 25%, 50%, 75% quartile data

3.4 모델 학습

전처리가 완료된 테스트 데이터 셋을 통해 학습한 모델로 추론이 가능해진다. 학습 데이터에서 유의미한 데이터를 가지고 있는 TN, NO3, NH3 3개의 변수로 각각 학습을 진행하였고, XGBoost와 LightGBM 모델을 학습시켜 TN, NO3, NH3 예측 모델을 개별적으로 구현하였다. 다음은 모델의 적합성을 알아보기 위하여 학습 셋에서 80% 학습과 나머지 20%는 모델이 유의미하는지 검증하는 데 사용하였다.

TN, NO3, NH3 예측 모델에 대한 예측하기 위한 최적의 XGBoost 및 LightGBM 모델을 찾기 위해 각 모델에 대해 그리드 탐색(Grid SearchCV)을 수행하여 최적의 하이퍼파라미터를 찾은 후 해당 하이퍼파라미터를 사용하여 모델을 학습하고 평가하였다. LightGBM모델에 대한 최적 하이퍼파라미터는 learning_rate 0.01, max_depth 3, n_estimators 100, XGBoost 모델에 대한 최적 하이퍼파라미터는 learning_rate 0.1, max_depth: 4, n_estimators 200으로 사용하였다. 결정트리의 개수가 많을수록 모델의 복잡성이 증가하고 과적합의 가능성이 높아 질 수 있다.

3.4 모델 평가

본 연구에서는 XGBoost와 LightGBM 두 가지 머신러닝 알고리즘을 활용하여 평촌교(PyeongchonBr)와 방학교(BangHakBr_Gwangjucheon1) 두 개의 수계로부터 TN, NO3, NH3 수질 지표 학습 결과를 5개의 서로 다른 부분으로 나눈 후 각각의 부분을 테스트 세트로 사용하고 나머지 부분을 훈련 세트로 사용하여 모델을 5번 학습하고 평가하는 방법으로 하는 5-fold cross-validation 교차 검증으로 테스트 세트에 대한 모델의 성능을 평균하여 최종 Cross-validation score를 계산하였고 Cross-validation score 전의 RMSE 평가 지표는 표 1, 표 2의 결과로 나타냈으며 교차 검증 후 평촌교(PyeongchonBr) 수계의 TN, NO3, NH3의 RMSE 지표는 그림 2로 나타냈으며 방학교(BangHakBr_Gwangjucheon1) 수계의 TN, NO3, NH3의 RMSE 지표는 그림 3으로 나타냈다. 그림 4는 평촌교(PyeongchonBr)와 방학교(BangHakBr_Gwangjucheon1) 수계 교차 검증 RMSE(:Root Mean Squared Error) 두 개 지표 결과를 모두 비교하였다.

실험 결과, XGBoost 모델이 평촌교(PyeongchonBr) 및 방학교(BangHakBr_Gwangjucheon1) 수계에서 모든 변수에 대해 상대적으로 낮은 RMSE 값을 보였다. 특히, TN 및 NO3 변수에서 XGBoost 모델의 성능이 높았으며, LightGBM 모델은 상대적으로 높은 RMSE 값을 보였다. 이는 XGBoost가 두 수계의 변수를 예측하는 데 더 효과적인 모델임을 알 수 있었다.

표 1. 평촌교(PyeongchonBr) 수계의 XGBoost와 LightGBM 모델의 RMSE

Table 1. The RMSE of XGBoost and LightGBM models for the PyeongchonBr watersheds

PyeongchonBr	XGBoost	LightGBM
TN	0.468	2.030
NO3	0.254	1.200
NH3	0.432	0.878

표 2. 방학교(BangHakBr_Gwangjucheon1) 수계의 XGBoost와 LightGBM 모델의 RMSE

Table 2. The RMSE of XGBoost and LightGBM models for the BangHakBr_Gwangjucheon1 watersheds

PyeongchonBr	XGBoost	LightGBM
TN	0.992	4.756
NO3	0.369	1.548
NH3	0.468	1.984

Ⅳ. 결 론

수질 지표 예측은 수질 개선을 위해 중요한 역할을 수행한다. 광주광역시 도시 하천의 수질 개선을 위해서는 수질 지표를 정확히 예측하여 수생 생태계에 미치는 영향을 파악하는 것이 필수적이다. 따라서 이 연구에서는 머신 러닝 알고리즘을 활용하여 수질 지표 예측을 수행하였다. 두 모델은 각 수질 지표별로 독립적으로 학습되었으며, 수질 데이터에 대해 XGBoost와 LightGBM 모델을 평가하고자 하였다.

교차 검증을 통해 각 모델의 성능을 평가하였고 평

촌교(PyeongchonBr) 그룹에서는 XGBoost 모델이 LightGBM 모델보다 모든 변수(TN, NO3, NH3)에 대해 더 낮은 RMSE 값을 보였다. 따라서 평촌교(PyeongchonBr) 그룹의 경우 XGBoost 모델이 더 우수한 예측 성능을 보이며, 특히 TN 변수에서의 성능 차이가 두드러졌다.

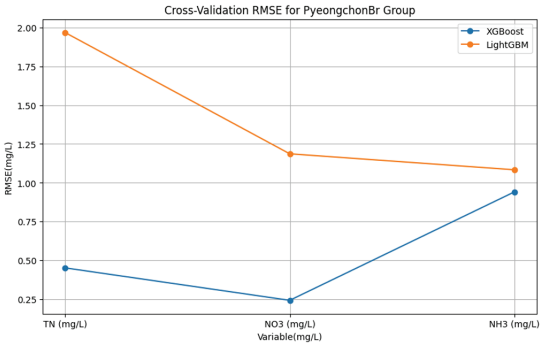

그림 2. 평촌교(PyeongchonBr) 그룹의 TN, NO3, NH3의 RMSE 지표
Fig. 2 Cross-Validation RMSE for PyeongchonBr Group

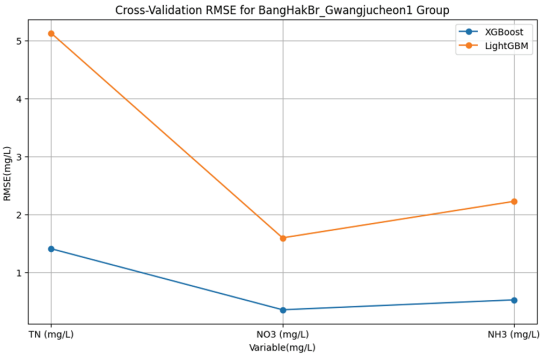

그림 3. 방학교(BangHakBr_Gwangjucheon1) 수계의 TN, NO3, NH3의 RMSE 지표
Fig. 3 Cross-Validation RMSE for BangHakBr_Gwangjucheon1 Group

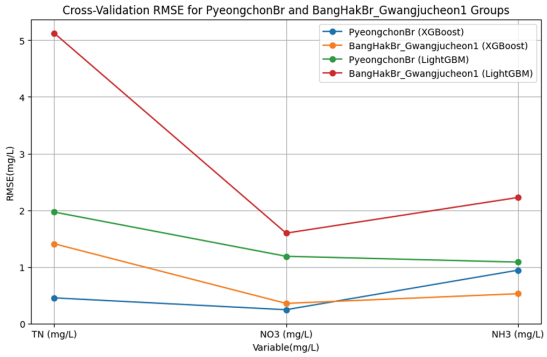

그림 4. 평촌교(PyeongchonBr) and 방학교(BangHakBr_Gwangjucheon1) 수계 교차검증 RMSE
Fig. 4 Cross-Validation RMSE for PyeongchonBr and BangHakBr_Gwangjucheon1 Groups

방학교(BangHakBr_Gwangjucheon1) 그룹에서도 XGBoost 모델이 LightGBM 모델보다 모든 변수에 대해 더 낮은 RMSE 값을 보였다. 방학교(BangHakBr_Gwangjucheon1) 그룹 역시 XGBoost 모델이 더 나은 성능을 보이며, 특히 TN 변수에서의 성능 차이가 크다는 것을 보였다. 이는 TN 변수에 대한 예측 성능에서 XGBoost 모델이 더 뛰어났다. 향후 유사한 수질 데이터에 대한 예측 모델을 개발할 때는 XGBoost 모델 예측 성능이 높을 것으로 사료된다.

향후 연구과제로는 더 나은 광주광역시의 수질예측 모델을 개발하기 위해 방학교(BangHakBr_Gwangjucheon1) 그룹의 수질 예측 결과에서 LightGBM 모델의 RMSE 값이 매우 높게 나타난 것으로 보아 이는 해당 지역의 수질 데이터가 다른 요인들로 인해 XGBoost와 LightGBM 모델 모두에 대해 예측이 어려웠음을 보였다. 하류와 상류지점 두 그룹 간의 수질 데이터 차이는 지역적 특성, 인구밀도와 산업 구조, 수질 관리 정책, 측정 시기 및 방법 등 다양한 요인들의 영향으로 인한 것으로 분석된다. 이러한 다양한 요인들을 고려하여 수질 데이터를 해석하고 이러한 연구들을 통해 더욱 정확하고 효율적인 수질 예측 모델을 개발하여 광주광역시 수질 환경 보전과 관리에 기여하고자 한다.

감사의 글

본 논문은 2024년도 송원대학교의 학술연구비 지원 사업으로 수행되었음.(과제번호:A-2014-49)

References

[1] J. Shin, S. Lee, M. Kim, and H. Park, “Imbalanced data augmentation for algal blooming warning AI model,” *Journal of the Information Technology and Applied Engineering*, vol. 11, no. 1, 2021, pp. 15-23.

[2] J. Kang, J. Park, S. Han, and K. Kim, “Development of Machine Learning based Flood Depth and Location Prediction Model,” *Journal of the Korea Institute of Electronic Communication Sciences*, vol. 18, no. 1, Feb. 2023, pp. 91-98.

[3] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, *in Proc. KDD'16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, Aug. 2016, pp. 785-794.

[4] S. Lee and G. Seok, “A Study Machine Learning Algorithms based on Embedded Processors Using Genetic Algorithm,” *Journal of the Korea Institute of Electronic Communication Sciences*, vol. 19, no. 2, Apr. 2024, pp. 417-426.

[5] The Ministry of Environment, "Operation manual for total water pollution management," *Report,* 2004.

[6] G. Kim, “A Study on the Analysis of Water Quality Trends in Rivers Using Nonparametric Statistical Tests,” Master’s Thesis, *Yonsei University*, 2015.

[7] H. Oh, A. Son, and Z. Lee, "Occupational accident prediction modeling and analysis using SHAP", *Journal of Digital Contents Society*, vol. 22. no. 7, July 2021, pp. 1115-1123.

[8] J. Shin, S. Park, and J. Shon, "Prediction of Semiconductor Exposure Process Measurement Results using XGBoost," *Journal of the Korean Society for Information Processing*, vol. 28, no. 1, May 2021, pp. 505-508.

[9] H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Institute of Mathematical Statistics*, vol. 29, no. 5, Oct. 2001, pp. 1189-1232.

[10] 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,” *Transactions of the Korean Institute of Electrical Engineers*, vol. 68, no. 12, 2019, pp. 1704-1710.

[11] G. Ke, Q. Meng, T. Finely, T. Wnag, W. Chen, W. Ma, Q. Ye, and T. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol 30, 2017, pp. 3146-3154.

[12] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, "Study on a prediction of P2P network loan default based on the machine learning- 34 -LightGBM and XGboost algorithms according to different high dimensional data cleaning," *Electronic Conference Research and Applications*, vol. 31, Sept. 2018, pp. 24-39.

[13] H. Choi, “Development of Investigation and Management Method for Contamination of Groundwater in Gwangju Metropolitan City,” *A Study on the Environmental Technology Development Center in Gwangju 2002*, Gwangju, Korea, 2002, pp. 60-63.

[14] J. Kim and H. Yang, “A Study on the Securing of the River Maintenance Water in Gwangju Stream,” *Technical Report*, Aug. 2002.

[15] J. Kaiser, “Dealing with missing values in data,” *Journal of Systems Integration,* vol. 5, no. 1, Nov. 2014, pp. 42-51.

저자 소개

정유정(Yu-Jeong Jeong)

1992년 조선대학교 전산학과 졸업(이학사)
1997년 조선대학교 대학원 전산통계학과 졸업(이학석사)
2010년 조선대학교 대학원 전산통계학과 졸업(이학박사)
2022년 ~현재 호남대학교 AI교양대학 조교수
※ 관심분야 :빅데이터, 데이터마이닝,인공지능, 영상 처리

이정재(Jung-Jae Lee)

1986년 조선대학교 전산기공학과(공학사)
1986년 조선대학교 대학원 전산기공학과(공학석사)
1997년 조선대학교 대학원 전산통계학과(이학박사)
1997년~현재 송원대학교 컴퓨터정보학과 교수
※ 관심분야 : 인공지능, 빅데이터, 헬스케어