ISSN(Print): 1225-7672 / ISSN(Online): 2287-822X
DOI https://doi.org/10.11001/jksww.2022.36.4.239

수질자료의 특성을 고려한 앙상블 머신러닝 모형 구축 및 설명가능한 인공지능을 이용한 모형결과 해석에 대한 연구

Development of ensemble machine learning model considering the characteristics of input variables and the interpretation of model performance using explainable artificial intelligence

박정수*
Jungsu Park*

국립한밭대학교 건설환경공학과
Department of Civil and Environmental Engineering, Hanbat National University

ABSTRACT

The prediction of algal bloom is an important field of study in algal bloom management, and chlorophyll-*a* concentration(Chl-*a*) is commonly used to represent the status of algal bloom. In, recent years advanced machine learning algorithms are increasingly used for the prediction of algal bloom. In this study, XGBoost(XGB), an ensemble machine learning algorithm, was used to develop a model to predict Chl-*a* in a reservoir. The daily observation of water quality data and climate data was used for the training and testing of the model. In the first step of the study, the input variables were clustered into two groups(low and high value groups) based on the observed value of water temperature(TEMP), total organic carbon concentration(TOC), total nitrogen concentration(TN) and total phosphorus concentration(TP). For each of the four water quality items, two XGB models were developed using only the data in each clustered group(Model 1). The results were compared to the prediction of an XGB model developed by using the entire data before clustering(Model 2). The model performance was evaluated using three indices including root mean squared error-observation standard deviation ratio(RSR). The model performance was improved using Model 1 for TEMP, TN, TP as the RSR of each model was 0.503, 0.477 and 0.493, respectively, while the RSR of Model 2 was 0.521. On the other hand, Model 2 shows better performance than Model 1 for TOC, where the RSR was 0.532. Explainable artificial intelligence(XAI) is an ongoing

pp. 209-218
pp. 219-228
pp. 229-237
pp. 239-248

Received 12 August 2022, revised 15 August 2022, accepted 15 August 2022.
*Corresponding author: Jungsu Park(E-mail: parkjs@hanbat.ac.kr)

• 박정수 (조교수) / Jungsu Park (Assistant professor)
대전광역시 유성구 동서대로 125, 34158
125, Dongseo-daero, Yuseong-gu, Daejeon 34158, Republic of Korea

field of research in machine learning study. Shapley value analysis, a novel XAI algorithm, was also used for the quantitative interpretation of the XGB model performance developed in this study.

Key words: Ensemble machine learning, Explainable artificial intelligence, Machine learning, Water quality management, Water quality prediction

주제어: 앙상블 머신러닝, 설명가능한 인공지능, 머신러닝, 수질관리, 수질예측

1. 서 론

하천, 호소 등에서 발생하는 조류는 점·비점오염원 및 기온, 강우 등 다양한 자연적·인위적 요인에 영향을 받게 되며 이러한 조류의 발생 정도는 취수원 및 정수장 수질관리에 중요한 지표중 하나로 지속적인 관리가 필요하다. 효율적인 수질관리를 위해서는 하천 및 호소내 조류발생 변화에 대한 예측이 중요하며 다양한 관련 연구가 지속되고 있다. Chlorophyll-a(Chl-*a*) 농도는 수중에 발생하는 조류의 양을 정량적으로 나타내는 대표적 수질항목중 하나이며, Chl-*a*는 수질, 기상 등 다양한 조건에 복합적인 영향을 받게 되어 정확한 예측이 쉽지 않다. 최근 수년간 예측의 대상이 되는 Chl-*a*와 다양한 영향 인자간 복잡성과 비선형관계를 반영할 수 있는 다양한 머신러닝 모형을 Chl-*a* 예측에 적용하기 위한 연구가 지속되고 있다 (Kwak, 2021; Park et al., 2015; Park et al., 2022; Shin et al., 2017).

머신러닝을 수질예측에 활용하기 위해서는 독립변수인 입력자료와 예측의 대상이 되는 종속변수를 이용하여 모형을 구축하고, 구축된 모형을 통해 예측대상이 되는 미래의 종속변수값을 예측하는 지도학습 모형이 주로 활용되고 있으며, 가장 대표적인 머신러닝 알고리즘중 하나인 artificial neural networks, support vector machine과 함께 random forest(RF), gradient boosting decision tree(GBDT) 등 다양한 ensemble 머신러닝 모형 등이 활용되고 있다 (Kwak, 2021; Kwon et al., 2018; Liu and Lu, 2014; Shin et al., 2017; Singh et al., 2011). Ensemble 머신러닝은 bagging, boosting 등 다양한 방법을 통해 단일 모형의 결과를 종합하여 단일 모형을 사용하는 것 보다 안정적이고 정확한 예측이 가능한 장점이 있어 최근까지도 수질예측을 포함한 다양한 분야에서 널리 활용되는 대표적인 머신러닝 알고리즘이다 (Dietterich, 2000; Hollister et al., 2016; Ma et al., 2018; Park et al., 2020; Park, 2021).

머신러닝 모형은 별도의 물리·화학·생물학적 실험을 통한 계수의 산정 등이 필요하지 않아 상대적으로 빠르게 모형을 구축할 수 있는 장점이 있으나, 모형의 성능이 입력자료의 특성에 많은 영향을 받게 되어 모형의 학습(training)과 성능의 평가(testing)를 위해 충분한 자료의 확보가 필수적이다. 더욱이 우리사회의 다양한 분야의 데이터에 기반하여 구축된 모형을 수질예측에 활용하기 위해서는 수질자료의 특성을 반영할 수 있는 적정한 입력자료의 확보와 전처리가 필요하다 (Park, 2021). 또한 블랙박스 모형의 특성상 모형 결과에 대한 이해와 모형에 영향을 주는 다양한 영향요인 등에 대한 정량적인 분석이 쉽지 않은 것은 머신러닝 모형의 단점중 하나로 제시되고 있다. 설명가능한 인공지능(XAI: explainable artificial intelligence)은 머신러닝 모형의 결과에 대한 정량적 해석을 제시하여 머신러닝 모형의 한계를 극복하고 모형의 활용성을 높일 수 있어 최근 관심이 높아지고 있는 연구분야이다 (Gunning et al., 2019; Park et al., 2022; Ribeiro et al., 2016; Shrikumar et al., 2016).

본 연구에서는 대표적인 ensemble 머신러닝 모형중 하나인 GBDT 알고리즘을 이용하여 호소내 Chl-*a* 농도를 예측하는 모형을 구축하였으며, 비지도 학습모형인 k-means clustering(KMC) 알고리즘을 이용하여 입력자료를 군집화하여 입력자료의 특성을 반영한 모형의 구축이 모형성능에 미치는 영향을 분석하였다. 또한 XAI 알고리즘을 이용하여 모형구축 결과에 대한 정량적인 해석을 수행하였다.

2. 재료 및 실험방법

2.1 연구대상지역 및 분석자료

본 연구의 대상인 대청호는 금강 상류 지역에 위치하며 유역면적 4,134 km^2, 총저수량 14.9억m^3으로 연간 약 16억m^3의 용수를 공급하는 금강 유역 주요 상수원 중 하나로 지속적인 조류 관리가 중요한 지역이다 (Fig. 1)(K-water, 2022).

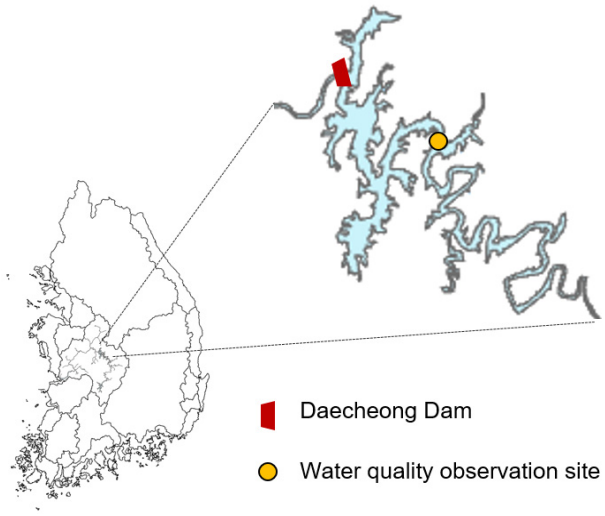


Fig. 1. Daecheong Dam and water quality monitoring site locations in this study.

본 연구에서는 환경부 국립환경과학원 물환경정보시스템 자동측정망 대청호지점에서 2013년 4월 2일부터 2021년 9월 29일까지 측정된 일별 측정자료와 기상청 기상자료 개방포털의 일별 기상자료를 모형의 구축 및 분석에 활용하였다 (KMA, 2022; NIER, 2022).

모형의 구축 및 분석에는 자동측정망에서 측정된 수온(TEMP), 수소이온농도(pH), 전기전도도(EC), 용존산소(DO), 총유기탄소(TOC), 총질소(TN), 총인(TP), 및 Chl-*a*(CHLA)의 8개 수질항목과 강수 계속시간(RAIN_DUR), 일강수량(RAINFALL), 합계 일조시간(SUN_DUR)의 3개 기상 항목이 활용되었다 (Table 1).

수질측정자료중 2014년 9월 25일부터 2014년 12월 31일까지 및 2015년 10월 7일부터 2016년 4월 1일까지의 자료는 장기적인 결측치 등을 포함하고 있어 분석에서 제외하였다. 실제 분석에 활용된 자료도 결측치를 포함하고 있었으며, TOC, TN 및 TP가 각각 15.1%, 8.7%, 8.0%의 결측치를 포함하고 있으며 그 외 항목은 5% 이내의 결측치를 포함하는 것으로 확인되었다. 결측이 발생한 구간은 대부분 수질변동이 크지 않은 구간으로 모형구축을 위해 K-Nearest Neighbor(KNN) 방법을 이용하여 결측치에 대한 보정을 수행하였다. KNN은 결측된 자료로부터 가까운 k개의 자료를 이용하여 결측값을 보정하는 방법으로 본 연구에서는 python(version 3.7.13) open source library인 scikit-learn (version 1.0.2)을 이용하여 KNN(k=3)을 이용한 보정을 수행하였다 (Pedregosa et al., 2011).

2.2 GBDT 모형구축

Ensemble 머신러닝 모형은 모형이 학습데이터에 과적합(overfitting) 되는 등의 문제를 해결하기 위해 단일 모형인 weak learner를 다수 생성하고 각 weak learner의 예측결과를 종합하여 최종적인 결론을 도출하는 방식으로 RF와 GBDT 등이 대표적인 알고리즘이다. RF는 각각의 weak learner의 생성 결과를 독립적으로 활용하는 반면, GBDT는 전단계 weak learner의 결과가 다음 단계 weak learner에 활용되어 점진적으로 모형의 성능이 향상되는 구조로 구성된 boosting 기반 알고리즘으로(Fig. 2), 다양한 분야에서 우수한 성능을 보여 최근까지도 널리 활용되고 있다 (Chen and Guestrin, 2016; Shin et al., 2020; Zhang et al., 2018). GBDT는 실측값(y_i)과 모형의 예측값($\hat{y_i}$)간의 차이로 계산되는 손실함수(M)와 각 weak learner(w_k)의 regularization 함수(Ω)로 구성된 목적함수(J)의(Eq. 1)

Table 1. Characteristics of input variables

Variables	Average	Standard deviation	Max	Min
TEMP(℃)	18.4	8.3	3.2	33.6
PH	8.1	0.8	6.6	10.4
EC(μS/cm)	152.4	17.3	72	220
DO(mg/L)	9.9	1.7	4.1	15.1
TOC(mg/L)	2.4	0.4	1.5	5
TN(mg/L)	1.5	0.4	0.1	3
TP(mg/L)	0.012	0.008	0	0.078
CHLA(mg/m^3)	8.6	9.5	0.5	157.8
RAIN_DUR(h)	2.4	5.1	0	133
RAINFALL(mm)	3.5	11.4	0	179.1
SUN_DUR(h)	6.9	4	0	13.9

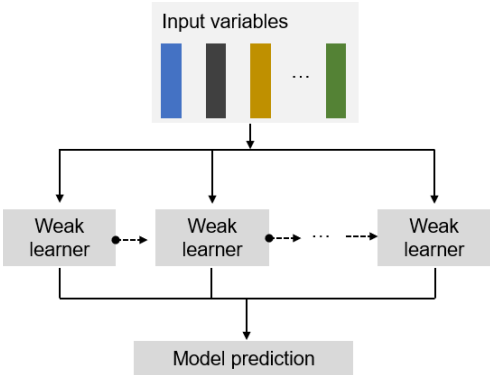


Fig. 2. A schematic of GBDT algorithm.

값을 최소화하도록 모형의 최적화가 수행된다 (Chen and Guestrin, 2016; Friedman, 2001; Zhang et al., 2018).

$$J=\sum_{i=1}^{n} M(y_i,\hat{y_i})+\sum_{i=1}^{K}\Omega(w_k) \quad (1)$$

본 연구에서는 GBDT 모형의 대표적인 알고리즘중 하나인 XGBoost(XGB)(version 0.90)를 활용하여 모형을 구축하였으며, 모형의 구성은 python open source library인 scikit-learn을 활용하고 grid search 방법을 이용하여 모형의 최적화를 수행하였다 (Pedregosa et al., 2011). 모형에 입력자료는 예측대상인 CHLA를 종속변수로 CHLA를 제외한 7개의 수질항목과 3개 기상항목을 포함한 총 11개 항목을 독립변수로 구성하였으며 11개 독립변수의 1일전 측정값을 함께 모형의 입력자료로 활용하였다. 계절적 변화가 뚜렷한 우리나라의 특성을 고려하여 전체 입력자료 중 2013년 4월 2일부터 2019년 12월 31일까지의 자료를 training에 2020년 1월 1일부터 2021년 9월 29일까지의 자료를 testing에 사용하였으며 장기간의 결측으로 제외된 구간을 제외하면 모형의 training과 testing에 활용된 자료의 비율은 각각 0.77과 0.23으로 구성되었다.

2.3 입력자료의 군집화를 통한 모형구축

본 연구에서는 KMC를 이용하여 입력자료의 군집화를 수행하였다. KMC는 사전에 정의된 군집의 수에 맞추어 입력자료를 임의의 군집으로 분류한 후 분류된 군집에 대하여 각 군집의 평균값과 각 입력자료값 간의 유클리디언 거리(euclidean distance)를 계산하고 대상 자료와 가장 가까운 군집으로 자료를 분류하는 과정을 반복하여 최적의 군집을 구성하게 된다 (Ahmad and Dey, 2007; Ayub et al., 2016; Song, 2017).

수질자료의 특성을 반영한 입력자료의 구축이 모형의 성능에 미치는 영향을 분석하기 위해 비지도학습 모형인 KMC를 이용하여 입력자료중 수질오염 현황을 나타내는 대표적인 항목인 TOC, TN, TP 3개 항목과 수온(TEMP)에 대하여 모형의 training에 사용된 기간의 측정값을 기준으로 군집화를 수행하였다. 선정된 4개 항목에 대하여 대상 항목의 값이 높은 구간에 속하는 군집과 낮은 구간에 속하는 군집의 2개 군집을 구성하고 각각의 군집에 최적화된 모형을 구축하였다 (Model 1). 모형의 성능 평가를 위해 testing에 활용되는 자료의 대상 항목 측정값이 높은 구간에 속하는 경우 높은 구간에서 최적화된 모형을 적용하고 낮은 구간에 속하는 경우 낮은 구간에서 최적화된 모형을 적용하여 예측값을 산출하여 모형의 실측값과 비교하여 성능을 평가하였다. 입력자료의 특성을 반영한 군집화된 입력자료로 구축된 Model 1과의 성능 비교를 위해 군집화를 수행하지 않은 전체 입력자료를 모형의 training에 사용하여 모형(Model 2)을 구축하여 그 결과를 Model 1의 결과와 비교하였다.

2.4 설명가능한 인공지능

복잡한 내부 알고리즘을 가지는 머신러닝 모형은 다양한 분야에서 좋은 성능을 보이고 있으나, 블랙박스 모형의 한계로 모형 결과에 대한 정확한 해석이 쉽지 않은 한계를 가지고 있다. XAI는 머신러닝 모형의 결과에 대한 정량적 해석을 가능하게 하여 기존 머신러닝 모형의 한계를 해소할 수 있는 알고리즘으로 제시되고 있다.

본 연구에서는 대표적인 XAI중 하나인 shapley value(SHAP) 분석을 통해 모형의 성능에 입력변수의 특성이 미치는 영향을 분석하였다. SHAP 분석은 특정 입력변수를 제외했을 경우 모형의 결과에 미치는 변화를 확인하여 특정 변수의 기여도를 분석하고 이를 통해 각 입력변수에 대한 SHAP를 산정하여, 모형 구축에 미치는 변수의 중요도를 정량적으로 제시하는 방법이다 (Lundberg et al., 2018; Lundberg and Lee, 2017). 본 연구에서는 SHAP 분석을 이용하여 본 연구에서 구축된 XGB 모형의 예측결과에 입력자료가 미치는 영향에 대한 해석 결과를 제시하였다.

2.5 모형 성능 평가

구축된 XGB 모형의 성능을 평가하기 위해 mean absolute error(MAE), root mean squared error(RMSE) 및 root mean squared error-observation standard deviation ration(RSR)의 3개 지수를 이용하였다 (Eq. 2, 3, 4). 산출식에서 y_t와 $\hat{y}_t$은 각각 시간 t에서의 실측값과 모형의 예측값을 나타내며, $\overline{y_t}$는 실측값의 평균을 n은 자료수를 나타낸다. MAE와 RMSE는 계산된 지수 값이 0에 가까울수록 모형의 예측성능이 우수함을 나타낸다. RSR은 0~1의 범위의 값을 가지며 RSR<0.7인 경우 모형을 통해 구한 결과가 실측값을 잘 예측하는 것으로 판단한다 (Bennett et al., 2013; Moriasi et al., 2007).

$$MAE = \frac{1}{n}\sum_{t=1}^{n} | y_t - \hat{y}_t | \quad (2)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}} \quad (3)$$

$$RSR = \frac{\sqrt{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}}{\sqrt{\sum_{t=1}^{n}(y_t - \overline{y_t})^2}} \quad (4)$$

3. 결과 및 고찰

3.1 입력자료 군집화 결과

입력자료의 특성을 모형구축에 반영하기 위해 수행된 4개 수질항목(TEMP, TOC, TN, TP) 각각에 대한 군집화를 통해 구성된 군집의 각 수질항목과 예측대상 종속변수인 CHLA의 관계를 Fig. 3에 제시하였다. 각 수질항목은 2개의 군집(Low range, high range)으로 분류되었으며, 각 항목별로 low range 및 high range에서의 평균값과 각 군집에서의 평균 Chl-a 농도를 비교하여 Table 2에 제시하였다. 군집화 수행결과 TEMP, TOC, TP의 경우 평균이 높은 high range의 군집에서 더 높은 Chl-*a* 농도를 가지는 반면 TN의 경우 TN 측정

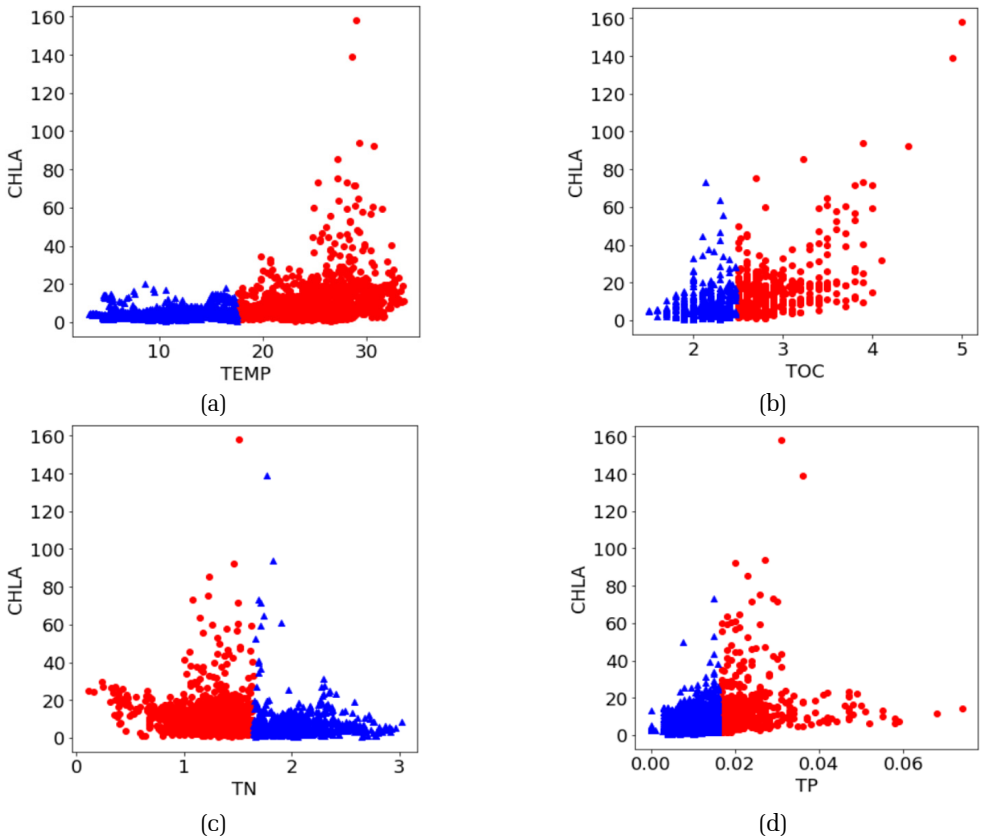


Fig. 3. Distribution of clustered input water quality data used for the model training.

Table 2. Characteristics of clustered input variables used for the model training

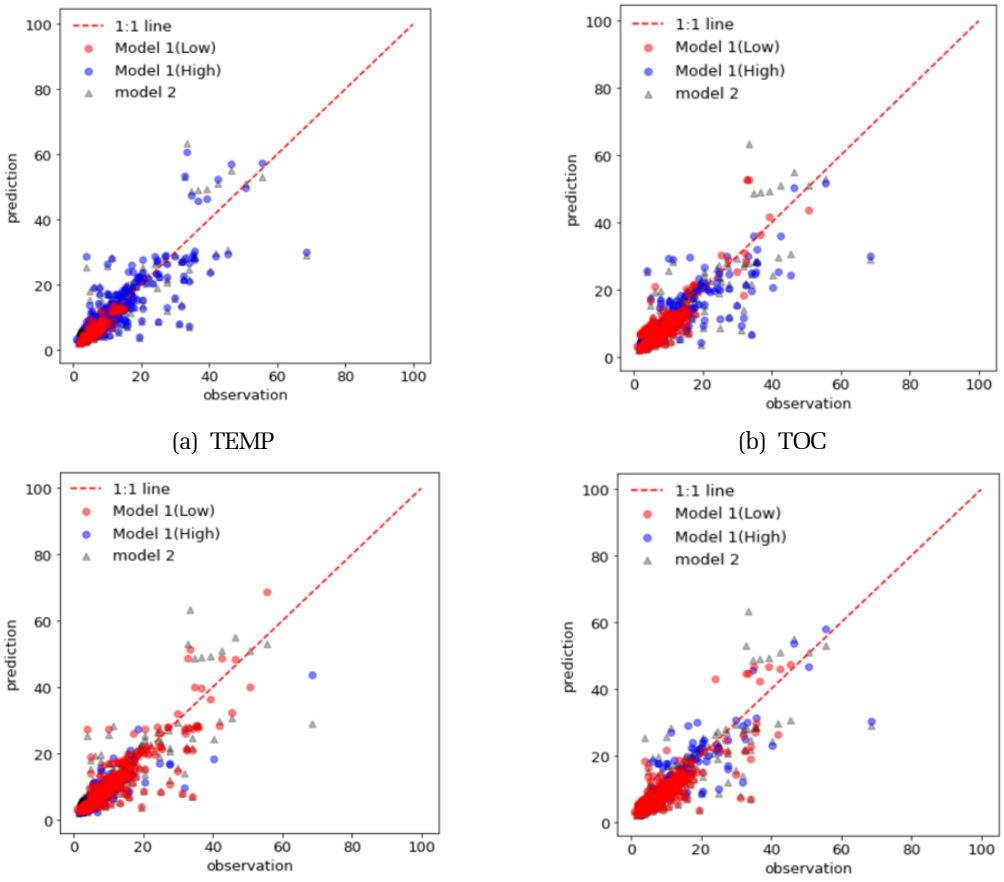
Variables	Range	Average	Average of Chl-a concentration(mg/㎥)
TEMP(℃)	Low	10.1	4.7
	High	25.0	11.4
TOC(mg/L)	Low	2.2	5.8
	High	2.8	12.9
TN(mg/L)	Low	1.3	9.4
	High	2.0	6.8
TP(mg/L)	Low	0.01	6.4
	High	0.024	16.0

값의 평균이 낮은 low range 군집에서 더 높은 Chl-*a* 농도를 가지는 것으로 분석되어 다른 수질항목과 차이를 보였으며, 이러한 경향은 Fig. 3에서도 시각적으로 확인할 수 있다.

3.2 모형예측결과

본 연구에서는 XGB를 이용하여 Chl-*a* 농도의 예측모형을 구축하였으며, 입력자료의 특성을 반영하여 4가지 수질항목(TEMP, TOC, TN, TP)에 대하여 각각 구축된 Model 1과 전체자료를 활용하여 구축된 Model 2의 예측 결과를 비교 평가하였다 (Table 3).

모형 성능의 비교 결과 TEMP, TN, TP 3가지 항목에 대해서는 낮은 측정값과 높은 측정값을 기준으로 각각의 군집에 대한 별도의 모형을 구축한 Model 1이 전체자료를 모두 활용하여 구축된 Model 2 보다 개선된 예측성능을 보이는 것으로 분석되었다. 특히 TN의 경우 Model 1과 Model 2의 MAE가 각각 1.990과 2.201로, RMSE가 각각 3.954와 4.315로 RSR이 각각 0.477과 0.521로 개선되어 입력자료의 특성을 고려한 군집화를 통한 모형의 성능 개선 효과가 가장 높은 것으로



(a) TEMP (b) TOC

(c) TN (d) TP

Fig. 4. Comparison of model predictions. Model 1 was developed using clustered data while Model 2 was developed using entire data before clustering.

Table 3. Model evaluation results

Model	Variables	MAE	RMSE	RSR
Model 1	TEMP	1.960	4.168	0.503
	TOC	2.328	4.407	0.532
	TN	1.990	3.954	0.477
	TP	2.123	4.084	0.493
Model 2	-	2.201	4.315	0.521

분석되었다. 반면 TOC의 경우 Model 1의 MAE, RMSE 및 RSR이 각각 2.328, 4.407 및 0.532로 각각의 군집에 최적화된 Model 1이 전체자료를 사용한 Model 2에 예측성능이 저하되는 경향을 보였다.

보다 세부적인 비교를 위해 Model 1의 대상 항목의 값이 낮은 군집에서 구축된 모형의 적용결과(Model 1(High)), 높은 군집에서 구축된 모형의 적용결과 (Model 1(Low)) 및 전체자료를 적용하여 구축된 모형의 결과(Model 2)를 각각 비교하여 Fig. 4에 제시하였다.

TN이 경우 군집화 분석을 통해 확인한 바와 같이 TN이 낮은 군집에서 CHLA가 더 높은 특성을 가지고 있으며, 모형의 구축결과 Model 1(Low)의 예측결과가 높은 CHLA 구간에서 전체입력자료를 활용한 Model 2보다 실측값을 더 잘 예측하여 1:1 line에 근접하여 분포하고 있는 것을 확인할 수 있었다. 이러한 예측 특성 등의 영향으로 군집별로 별도의 모형을 구축한 Model 1이 좀더 높은 예측성능을 보이는 것으로 판단된다. 반면 TOC의 경우 Model 1(high)의 예측값이 Model 1에 비해 상대적인 오차가 크게 나타나는 경향을 시각적으로 확인할 수 있다.

3.3 설명가능한 인공지능을 이용한 모형 결과분석

SHAP 분석은 머신러닝 모형구축에 사용된 입력변수의 상대적 중요도를 정량적으로 제시하며 개별 측정값에 대한 해석이 가능한 장점이 있어 다양한 분야에서 관련 연구가 지속되고 있다 (Ekmekcioğlu et al., 2022; Mangalathu et al., 2020; Park et al., 2022). Park et al. (2022)는 SHAP 분석과 통계적분석방법 등을 통해 Chl-*a* 농도를 예측하는 모형에 입력변수가 미치는 영향을 정량적으로 분석하였으며 이를 기반으로 입력변수의 우선순위에 따라 모형을 구축하고 변수의 선정이 모형의 성능에 미치는 영향을 분석하여 제시하였다.

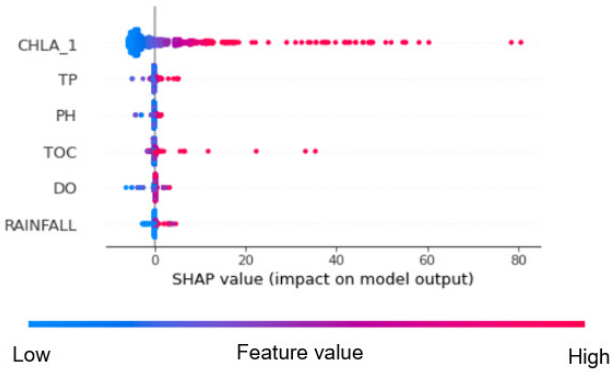


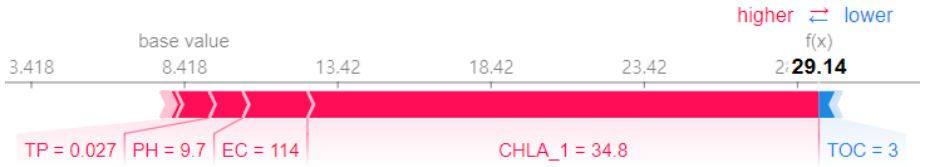
Fig. 5. SHAP summary plot of Model 2.

본 연구에서는 XGB를 이용하여 구축된 Model 2의 SHAP 분석을 통해 모형 예측결과에 영향이 큰 5가지 독립변수의 SHAP 값을 Fig. 5에 제시하였으며, 1일 전의 Chl-*a* 농도가 모형구축에 가장 높은 영향을 미치는 것을 확인할 수 있다.

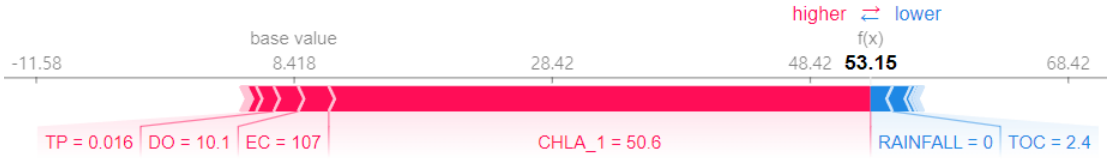
보다 세부적인 분석을 위해 모형의 예측값과 실측값의 차이가 상대적으로 큰 2020년 9월 16일 및 2020년 9월 26일의 개별 측정결과에 대한 SHAP 분석을 수행하였다. 2020년 9월 16일의 경우 CHLA 실측값과 Model 2 모형의 예측값은 각각 68.5 mg/㎥ 및 29.14 mg/㎥로 모형이 실측값보다 훨씬 적은 값을 예측한 것을 확인하였다 (Fig. 6 and Fig. 7). Fig. 6(a)는 2020년 9월 16일의 개별 예측값에 대한 SHAP 분석결과로 모형의 예측값에 모형구축에 사용된 각 독립변수가 미치는 영향을 보여준다.

Fig. 6(a)에서 보여주는 바와 같이 CHLA_1>EC>PH>TP 순으로 해당 시점의 측정값이 base value보다 CHLA를 높게 예측하는데 영향을 주고 있으며, TOC는 반대로 CHLA의 예측값을 낮게 예측하는 방향으로 영향을 주게 됨을 확인할 수 있다. 해당일의 예측결과는 전일의 Chl-*a* 측정값인 CHLA_1가 가장 많은 영향을 미치는 것으로 분석되어 기상 및 수질오염 등 다른 조건 보다 전일의 조류농도 자체가 당일의 조류농도에 영향을 미치는 주요 인자임을 정량적으로 보여주었다. 모형에서는 전일의 Chl-*a* 실측값인 34.8 mg/㎥를 이용하여 CHLA를 예측하게 되는데 2020년 9월 16일의 경우 전일에 비해 68.5 mg/㎥까지 급격히 Chl-*a*가 증가한 것이 모형이 CHLA를 실측값보다 낮게 예측하는 한 원인이 된 것으로 판단된다.

반면 2020년 9월 26일의 경우 CHLA 실측값과



(a) Observation on September 16, 2020



(b) Observation on September 26, 2020

Fig. 6. SHAP force plot of Model 2.

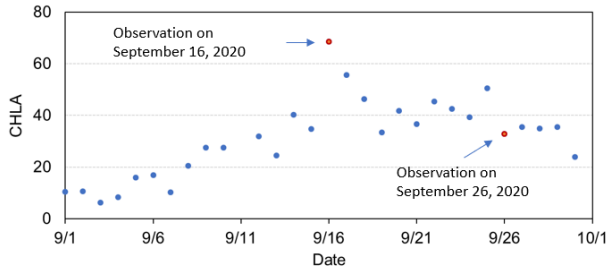


Fig. 7. Observed CHLA between September 1, 2020 and September 30, 2020.

Model 2의 예측값은 각각 32.8 mg/m³ 및 53.15 mg/m³로 모형이 예측값이 실측값보다 훨씬 높은 것을 확인하였다. Fig. 6(b)에서 확인할 수 있는 바와 같이 전일의 Chl-*a* 측정값(CHLA_1)이 50.6 mg/m³로 비교적 높아서 모형이 CHLA를 실측값 보다 높게 예측하는 한 원인이 된 것으로 판단된다.

4. 결 론

본 연구에서는 입력변수의 특성에 따라 변수를 군집화하고 각 군집에 최적화된 XGB 모형을 구축하여 변수의 군집화가 모형의 성능에 미치는 영향을 분석하였다. TEMP, TOC, TN, TP 4개 변수에 대해 측정값이 낮은 군집과 높은 군집의 2가지로 입력자료를 군집화하고 각각의 군집에 최적화된 모형을 구축하였다(Model 1). Model 1의 성능을 평가한 결과 TEMP, TN, TP의 3가지 항목을 기준으로 군집화를 수행한 경우 전체자료를 모두 사용한 모형(Model 2)에 비해 성능이 향상된 것으로 분석되었다. 전체자료를 이용하여 모형을 구축한 경우 RSR은 0.521이었으며, TN을 기준으로 군집화를 수행한 모형의 경우 RSR이 0.477로 가장 높은 성능의 향상을 보여주었다. 반면 TOC를 기준으로 군집화를 수행한 모형의 경우 RSR이 0.532로 전체자료를 사용한 모형에 비해 성능이 낮아지는 것으로 분석되었다.

본 연구에서는 SHAP 분석을 통해 머신러닝 모형의 예측성능에 미치는 입력변수의 영향을 확인하고 개별 측정값에 대한 세부적인 분석을 통해 XAI를 이용하여 머신러닝 모형의 결과에 대한 가능한 해석을 제시하였다.

머신러닝 모형은 입력자료를 기반으로 모형의 최적화를 수행하게 되며 입력자료의 특성이 모형의 성능에 많은 영향을 미치게 된다. 이러한 머신러닝 모형을 좀더 효율적으로 적용하기 위해서는 물환경분야 측정자료의 특성에 맞는 모형구축을 위한 다양한 연구가 필요하다. 본 연구에서는 입력자료의 특성을 고려한 별도의 최적화된 모형의 구축이 모형의 성능에 미치는 영향을 분석하였으며, 향후 입력자료의 특성을 고려한 최적화된 모형의 구축을 위한 지속적인 연구가 필요할 것으로 생각된다.

References

Ahmad, A., and Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data, Data Knowl. Eng., 63, 503-527.

Bennett, N.D., Croke, B.F., Guariso, G., Guillaume, J.H., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S.,

Newham, L.T., Norton, J.P. and Perrin, C. (2013). Characterising performance of environmental models, Environ. Modell. Softw., 40, 1-20.

Chen, T. and Guestrin, C. (2016). "Xgboost: A scalable tree boosting system", *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17 August, San Francisco, CA, USA. Association for computing Machinery.

Dietterich, T.G. (2000). Ensemble methods in machine learning, *In international workshop on multiple classifier systems, June, Berlin, Heidelberg*. 1-15.

Ekmekcioğlu, Ö., Koc, K., Özger, M., and Işık, Z. (2022). Exploring the additional value of class imbalance distributions on interpretable flash flood susceptibility prediction in the Black Warrior River basin, Alabama, United States, J. Hydrol., 610, 127877.

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine, Ann. Stat., 1189-1232.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.-Z., 2019. XAI—explainable artificial intelligence, Sci. Robot. 4(37).

Hollister, J.W., Milstead, W.B. and Kreakie, B.J. (2016). Modeling lake trophic state: A random forest approach, Ecosphere, 7, e01321.

KMA Korea Meteorological Administration, open met data portal, https://www.data.kma.go.kr/ (April 1, 2022).

Kwak, J. (2021). A study on the 3-month prior prediction of Chl-*a* concentraion in the Daechong lake using hydrometeorological forecasting data, J. Wetl. Res., 23(2), 144-153.

K-water Mywater https://www.water.or.kr/ (June 1, 2022).

Kwon, Y.S., Baek, S.H., Lim, Y.K., Pyo, J., Ligaray, M., Park, Y. and Cho, K.H. (2018). Monitoring coastal chlorophyll-a concentrations in coastal areas using machine learning models, Water 10(8), 1020.

Liu, M., and Lu, J. (2014). Support vector machine—an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river?, Environ. Sci. Pollut. R., 21, 11036-11053.

Lundberg, S.M., Erion, G.G., and Lee, S.I. (2018). Consistent individualized feature attribution for tree ensembles, https://arxiv.org/abs/1802.03888

Lundberg, S.M. and Lee, S.I. (2017). "A unified approach to interpreting model predictions", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q. and Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning, Electron Commer. Res. Appl., 31, 24-39.

Mangalathu, S., Hwang, S.H., and Jeon, J.S. (2020). Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach, Eng. Struct., 219, 110927.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D. and Veith, T.L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Am. Soc. Agric. Biol. Eng., 50, 885-900.

NIER National Institute of Environmental Research, realtime water information system http://www.koreawqi.go.kr/index_web.jsp (April 1, 2022).

Park, J. (2021). The effect of input variables clustering on the characteristics of ensemble machine learning model for water quality prediction, J Korean Soc. Wat. Environ., 37(5), 335-343.

Park, J., Lee, W.H., Kim, K.T., Park, C.Y., Lee, S. and Heo, T.Y. (2022). Interpretation of ensemble learning to predict water quality using explainable artificial intelligence, Sci. Total Environ., 832, 155070.

Park, J., Park, J.H., Choi, J.S., Joo, J.C., Park, K., Yoon, H.C., Park, C.Y., Lee, W.H., and Heo, T.Y. (2020). Ensemble Model Development for the Prediction of a Disaster Index in Water Treatment Systems, Water, 12, 3195.

Park, Y., Cho, K.H., Park, J., Cha, S.M. and Kim, J.H. (2015). Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Sci. Total Environ., 502, 31-41.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011). Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., 12, 2825-2830.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should I trust you?" explaining the predictions of any classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery And Data Mining*, 1135–1144.

Shin, C.M., Min, J.H., Park, S.Y., Choi, J., Park, J.H., Song, Y.S. and Kim, K. (2017). Operational water quality forecast for the Yeongsan river using EFDC model, J. Korean Soc. Water Environ., 33(2), 219-229.

Shin, Y., Kim, T., Hong, S., Lee, S., Lee, E., Hong, S., Lee,

C., Kim, T., Park, M.S., and Park, J. (2020). Prediction of chlorophyll-a concentrations in the Nakdong River using machine learning methods, Water, 12, 1822.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A., 2016. Not just a black box: learning important features through propagating activation differences arXiv preprint arXiv: 1605.01713.

Singh, K.P., Basant, N., and Gupta, S. (2011). Support vector machines in water quality management, Anal. Chim. Acta., 703, 152-162.

Song, J. (2017). K-Means cluster analysis for missing data, J. Korean Data Anal. Soc., 19, 689-697.

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B. and Si, Y. (2018). A data-driven design for fault detection of wind turbines using random forests and XGboost. IEEE Access, 6, 21020-21031.