ISSN(Print): 1225-7672 / ISSN(Online): 2287-822X
DOI https://doi.org/10.11001/jksww.2021.35.6.417

앙상블 머신러닝 모형을 이용한 하천 녹조발생 예측모형의 입력변수 특성에 따른 성능 영향

Effect of input variable characteristics on the performance of an ensemble machine learning model for algal bloom prediction

강병구·박정수*
Byeong-Koo Kang·Jungsu Park*

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

ABSTRACT

Algal bloom is an ongoing issue in the management of freshwater systems for drinking water supply, and the chlorophyll-*a* concentration is commonly used to represent the status of algal bloom. Thus, the prediction of chlorophyll-*a* concentration is essential for the proper management of water quality. However, the chlorophyll-*a* concentration is affected by various water quality and environmental factors, so the prediction of its concentration is not an easy task. In recent years, many advanced machine learning algorithms have increasingly been used for the development of surrogate models to prediction the chlorophyll-*a* concentration in freshwater systems such as rivers or reservoirs. This study used a light gradient boosting machine(LightGBM), a gradient boosting decision tree algorithm, to develop an ensemble machine learning model to predict chlorophyll-*a* concentration. The field water quality data observed at Daecheong Lake, obtained from the real-time water information system in Korea, were used for the development of the model. The data include temperature, pH, electric conductivity, dissolved oxygen, total organic carbon, total nitrogen, total phosphorus, and chlorophyll-*a*. First, a LightGBM model was developed to predict the chlorophyll-*a* concentration by using the other seven items as independent input variables. Second, the time-lagged values of all the input variables were added as input variables to understand the effect of time lag of input variables on model performance. The time lag (i) ranges from 1 to 50 days. The model performance was evaluated using three indices, root mean squared error-observation standard deviation ration (RSR), Nash-Sutcliffe coefficient of efficiency (NSE) and mean absolute error (MAE). The model showed the best performance by adding a dataset with a one-day time lag (i=1) where RSR, NSE, and MAE were 0.359, 0.871 and 1.510, respectively.

Received 3 November 2021, revised 21 November 2021, accepted 23 November 2021.
*Corresponding author: Jungsu Park (E-mail: parkjs@hanbat.ac.kr)

• 강병구 (석사과정) / Byeong-Koo Kang (Master Student)
대전광역시 유성구 동서대로 125, 34158
125, Dongseo-daero, Yuseong-gu, Daejeon 34158, Republic of Korea

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

The improvement of model performance was observed when a dataset with a time lag up of about 15 days (i=15) was added.

Key words: Ensemble machine learning, Gradient boosting decision tree(GBDT), Machine learning, Water quality management, Water quality prediction

주제어: 앙상블 머신러닝, 그레디언트 부스팅 의사결정 나무, 머신러닝, 수질관리, 수질예측

1. 서 론

하천 유역에서 발생하는 점·비점오염원에 의한 영양염류 등 오염물질의 유입과 기후변화에 따른 수온 상승 등의 다양한 요인에 의해 하천에서 조류(algae) 발생이 지속되고 있다 (Lim and An, 2018). 조류의 발생은 하천 및 취수원 수질의 안정적 관리를 위해 고려해야 할 중요한 인자 중 하나로 지속적인 관리가 필요하다. 조류의 관리를 위해서는 조류 발생 정도를 정량화하여 분석하는 것이 중요하며, chlorophyll-*a*(Chl-*a*) 농도는 조류의 발생 정도를 확인하기 위해 널리 사용되는 대표적인 수질항목이다 (Kwak, 2021; Shin et al., 2017). 효율적 수질관리를 위해서는 수질현황을 확인하고 그 변화를 사전에 예측하는 것이 필요하다. 이를 위해 수질, 환경조건, 수문 기상인자 등에 기반한 다양한 수질예측 모형이 개발되고 활용되어왔으며, 최근에는 고도화된 데이터 분석 기술에 기반한 머신러닝(machine learning) 모형을 이용하여 수질변화를 예측하기 위한 연구도 활발해지고 있다 (Lee et al., 2020; Lim and An, 2018; Park et al., 2015).

Park et al. (2015)은 대표적인 머신러닝 알고리즘(algrorithm)인 artificial neural networks(ANN)과 support vector machine(SVM)을 이용하여 Chl-*a* 농도를 예측하는 모형을 제시하였으며, Kwon et al. (2018)은 ANN 및 SVM 알고리즘과 인공위성의 이미지 자료를 활용하여 Chl-*a* 농도를 예측하였다. Lim and An (2018)은 시계열 자료의 분석에 좋은 성능을 보이는 딥러닝(deep learning) 알고리즘 중 순환신경망 recurrent neural networks(RNN)과 long short-term memory(LSTM) 알고리즘을 이용하여 오염부하량을 예측하였다. Lee et al. (2020)은 대표적인 ensemble 머신러닝 알고리즘인 random forest(RF)와 gradient boosting decision tree(GBDT) 등을 이용하여 낙동강 중류지역의 Chl-*a*를 예측하는 모형을 구축하였다.

머신러닝 모형은 예측대상 항목의 물리, 화학, 생물학적 특성에 기반한 계수를 별도로 구하지 않고 독립변수를 입력받은 후 구축된 모형의 내부 알고리즘에 따라 예측의 대상인 종속변수의 예측을 수행하게 된다. 따라서, 예측대상 항목의 특성을 고려하고 다양한 입력변수가 모형의 성능에 미치는 영향을 분석하여 모형의 성능과 실제 적용성을 높일 수 있다.

입력자료에 기반하여 예측을 수행하게 되므로 시계열 데이터 분석, 언어분석, 이미지분석 등에 범용적으로 적용할 수 있도록 개발된 머신러닝 알고리즘이 다양한 분야에 유연하게 활용될 수 있으며, 수질분야에서도 머신러닝 모형의 활용이 빠르게 늘어나고 있다. 다양한 머신러닝 모형 중 여러 개의 모형의 결과를 결합하여 모형의 성능을 향상시키는 방식을 이용하는 ensemble 모형은 딥러닝 모형에 비해 상대적으로 모형의 구축이 복잡하지 않으면서도 우수한 예측성능을 보여 최근까지도 다양한 분야에 활발히 사용되고 있다 (Belgiu and Drăguţ, 2016; Dietterich, 2000; Zhou, 2021). 본 연구에서는 ensemble 모형의 대표적인 알고리즘인 GBDT중 모형구축에 적용하는 입력변수를 줄이는 내부 알고리즘을 통해 모형의 구현속도를 높이면서도 우수한 성능을 유지하는 장점을 가지고 있어 가장 널리 사용되는 GBDT 알고리즘 중 하나인 light gradient boosting machine(LightGBM)을 이용하여 우리나라 금강유역의 대표적 취수원인 대청호의 Chl-*a* 농도를 예측하는 모형을 구축하고, 모형의 구축에 사용된 입력자료가 모형의 성능에 미치는 영향을 분석하였다.

2. 재료 및 실험방법

2.1 연구대상지역

금강 본류에 위치한 대청호는 1981년 완공된 대청댐에 의해 조성된 인공호수로 유역면적 3,204 km^2, 총 저수량 14.9억m^3으로 소양호와 충주호에 이어 대한민국에서 세 번째로 큰 규모를 가지고 있다 (Fig. 1)(K-water, 2021).

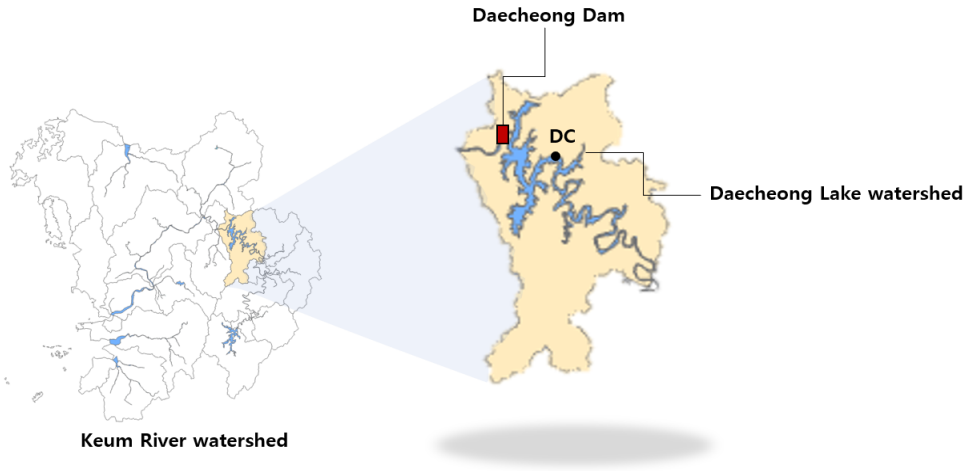


Fig. 1. Research site.

Table 1. Characteristics of input variables

Variables	Average	Standard deviation	Max.	Min.
Temp	17.7	8.4	33.7	2.4
pH	8.1	0.8	10.4	6.6
EC	150.0	19.0	220.0	72.0
DO	10.0	1.8	16.5	4.1
TOC	2.4	0.4	5.0	1.5
TN	1.5	0.4	3.0	0.1
TP	0.014	0.010	0.088	0.001
Chl-*a*	8.7	10.0	157.8	0.5

대청호는 금강 유역의 홍수 관리에 중요한 역할을 하고 있으며, 대전 및 충청지역 등에 용수를 공급하는 주요 상수원으로 주기적으로 발생하는 녹조문제 등 수질관리가 중요한 호소이다.

본 연구에서는 환경부 국립환경과학원 실시간 수질정보시스템에서 제공되는 수질 자동 측정망 자료 중 대청호지점(DC)에서 2012년 7월 1일부터 2021년 6월 30일까지 측정된 일별 측정자료를 활용하였다 (NIER). 연구에 이용된 수질 측정 항목은 수온(Temp), 수소이온농도(pH), 전기전도도(EC), 용존산소(DO), 총유기탄소(TOC), 총질소(TN), 총인(TP), 및 Chl-*a*의 총 8개이다 (Table 1).

2.2 LightGBM 모형

LightGBM은 Microsoft에서 개발한 머신러닝 알고리즘으로 XGBoost와 함께 최근 가장 널리 사용되는 GBDT 알고리즘중 하나이다 (Chen and Guestrin, 2016; Ke et al., 2017; Ma et al., 2018; Oh et al., 2021; Su and Zhao, 2020; Zhang et al., 2018). GBDT는 대표적인 ensemble 머신러닝 알고리즘으로 weak learner로 불리는 여러 개의 단위 모형으로 구성되어 있다. 각 단계의 weak leaner는 이전 단계 weak learner의 예측값과 실측값의 차이인 잔차(residual)를 예측하며, 이러한 단위 모형이 연쇄적으로 구성되어 모형이 실측값에 가까운 값을 예측하게 된다 (Fig. 2)(Chen and Guestrin, 2016; Ke et al., 2017).

LightGBM은 모형구축에 사용되는 입력변수 및 자료수를 줄이기 위해 Gradient-based One-Side Sampling (GOSS)와 Exclusive Feature Bundling(EFB)의 2가지 알고리즘을 적용하여 모형의 구현속도를 높이면서도 충분한 예측성능을 유지하도록 구성되었다 (Ke et al., 2017). GOSS는 모형의 손실함수(loss function)의 미분을 통해 구해지는 Gradient(GR)가 클수록 해당 자료에 의한 정보획득량(information gain)이 커질수 있다는 점을 이용하여 입력자료를 GR의 절대값이 큰 순서대로 배열후 GR이 큰 입력자료의 a*100%를 선택하고, 선택되지 않은 자료 (1-a)*100%중 b*100%를 선택하여 모형

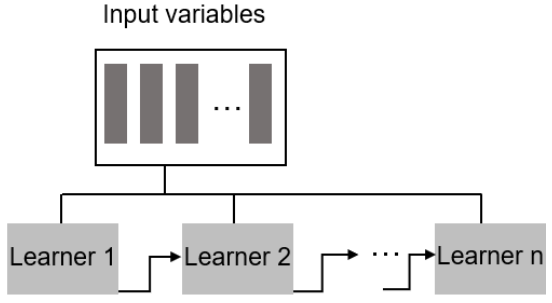


Fig. 2. A schematic of GBDT algorithm.

구축에 사용되는 입력자료의 수를 줄이는 방식으로, a는 모형구축에 사용된 GR의 큰 입력자료의 비율을 b는 모형구축에 사용된 GR이 작은 입력자료의 비율을 의미한다 (Ke et al., 2017). EFB는 고차원의 입력자료(high-dimensional data)의 경우 변수공간이 희박한(sparse) 특성이 있음을 이용하여 서로 다른 독립변수를 단일변수로 합쳐서(bundling) 사용해도 모형의 성능을 유지할 수 있음을 이용하여 입력변수를 줄여 모형의 효율성을 향상시키는 방식이다 (Ke et al., 2017).

2.3 입력 자료

2.3.1 결측치 전처리

측정한 8개의 항목은 Temp 7.7%, pH 7.5%, EC 7.2%, DO 8.5%, TOC 17%, TN 11.8%, TP 14.7%, Chl-*a* 12%의 결측치를 포함하고 있었으나, 결측이 발생하는 기간이 대부분 강우가 발생하지 않은 평상시로 수질의 변동이 크지 않은 구간이었다. 본 연구에서는 결측이 발생한 자료로부터 거리가 가까운 k개의 다른 자료를 이용하여 결측을 보정하는 K-Nearest Neighbor(KNN) 방법을 이용하여 모형의 결측값을 보정하였으며, KNN의 구현은 python open source library인 scikit-learn을 활용하였다 (Pedregosa et al., 2011).

2.3.2 모형 입력 자료 구축

LightGBM을 이용한 Chl-*a* 예측 모형이 구축을 위해 2012년 7월 1일부터 2019년 9월 30일까지의 측정자료를 모형의 학습(training)에 2019년 10월 1일부터 2021년 6월 30일까지의 측정자료를 학습된 모형의 평가(testing)에 활용하여 training과 testing에 사용된 자료의 비율은 8:2로 구성하였다. 종속변수인 Chl-*a*의 예측을 위해 Temp, pH, EC, DO, TOC, TN 및 TP의 7개 항목이 모형의 입력변수로 이용되었다. 모형의 구축은 LightGBM open source library를 이용하였으며, 모형 hyperparameter 최적화는 python open source library인 scikit-learn의 grid search를 이용하여 수행되었다 (LightGBM; Pedregosa et al., 2011).

2.3.3 차분 값의 활용

구축된 LightGBM 모형에 과거에 측정된 값을 입력변수로 추가하여 포함하는 것이 모형의 성능에 미치는 영향을 분석하기 위해 각 수질항목별로 1일(t-1)~50일(t-50)까지 1일씩의 차분(time lag)을 적용한 값을 모형에 적용하여 성능을 비교하였다.

2.4 LightGBM 모형 성능 검정 및 비교

LightGBM의 예측 성능 비교를 위해 root mean squared error-observation standard deviation ration(RSR), nash-sutcliffe coefficient of efficiency(NSE), mean absolute error(MAE) 3개의 평가지수를 이용하여 성능 비교를 하였다 (Eq 1, 2, 3). RSR은 0~1의 범위를 가지며 0에 가까울수록 모형의 성능이 좋은 것을 의미하고, RSR<0.7인 경우 모형이 종속변수를 잘 예측한 것으로 판단한다 (Bennett et al., 2013; Moriasi et al., 2007). NSE는 -∞ ~ 1.0의 범위를 가지며 1에 가까울수록 모형의 성능이 우수함을 의미한다. MAE는 각 실측값에 대해 모형을 통해 구한 예측값과의 차이의 절대값을 구한 후 그 총합을 실측 자료 수로 나누어 산정되며 그 값이 작을수록 모형의 성능이 우수함을 나타낸다.

$$RSR = \frac{\sqrt{\sum_{t=1}^{n}(M_{t,obs} - M_{t,m})^2}}{\sqrt{\sum_{t=1}^{n}(M_{t,obs} - \overline{M_{t,obs}})^2}} \quad (1)$$

$$NSE = 1 - \frac{\sum_{t=1}^{n}(M_{t,m} - M_{t,obs})^2}{\sum_{t=1}^{n}(M_{t,obs} - \overline{M_{obs}})^2} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} | M_{t,m} - M_{t,obs} | \quad (3)$$

$M_{t,obs}$: Observed value at time t, $M_{t,m}$: Predicted value at time t

$\overline{M}_{obs}$: Average of observed values, n: Total number of observed values

3. 결과 및 고찰

3.1 LightGBM 예측결과

본 연구에서는 LightGBM을 이용하여 구축된 Chl-*a* 예측모형의 성능에 입력변수의 차분값이 미치는 영향을 분석하였으며, 이를 위하여 우선 별도의 차분값이 없이 전체 입력변수를 적용하여 구성한 모형의 성능을

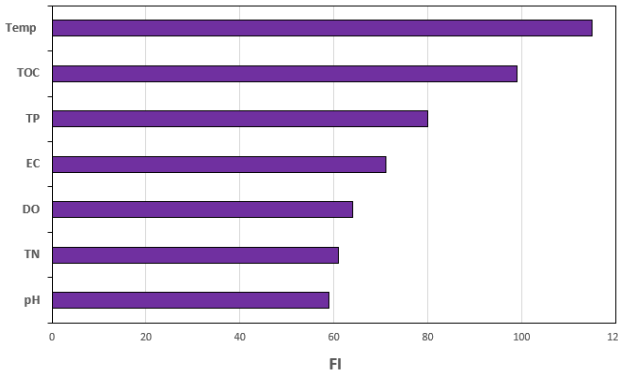


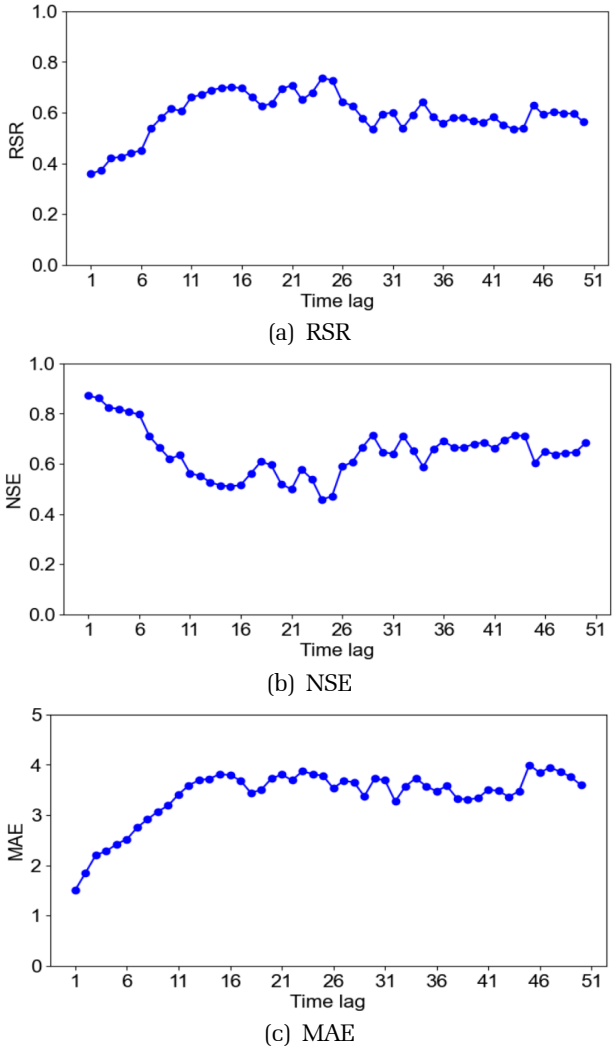
Fig. 3. FI of LightGBM model.

확인하였다. 입력변수에 별도의 차분값 적용 없이 최적화한 모형의 구축결과 RSR, NSE, MAE는 각각 0.741, 0.451, 3.763으로 산출되었다. LightGBM은 모형의 성능에 영향을 미친 각 입력변수의 상대적 중요도(feature importance: FI)를 확인할 수 있으며, 산출된 FI는 Fig. 3과 같다. FI 분석결과 모형 성능에 영향을 미치는 가장 중요한 변수는 Temp였으며, TOC, TP, EC, DO, TN, pH 순으로 높은 중요도를 가지는 것으로 분석되었다.

3.2 차분값에 따른 모형성능 비교

과거의 자료를 활용하여 구축된 입력자료가 현시점의 수질예측 성능 개선에 미치는 영향을 분석하기 위해, Chl-*a*에 대하여 1~50일의 차분을 구하였으며 최적화된 LightGBM 모형에 각 일수별 차분을 독립변수로 추가하여 새로운 LightGBM 모형을 구축하여 차분으로 적용한 일수의 차이가 모형이 성능에 미치는 영향을 분석하였다. 구축된 모형의 성능을 RSR, NSE 및 MAE의 3가지 지수를 이용하여 비교한 결과 3가지 지수 모두 유사한 결과를 보여주었다. RSR과 MAE는 0에 가까울수록 모형이 실측값을 잘 예측하는 것을 나타내고 지수값이 커지는 것은 모형의 정확도가 낮아짐을 의미하며, NSE는 1에 가까울수록 모형이 실측값을 잘 예측하는 것을 나타내고 지수값이 작아지는 것은 모형의 정확도가 낮아짐을 의미한다. 세 가지 평가지수 모두 모형의 성능은 1일의 차분을 적용할 경우 가장 많이 개선되는 경향을 보였으며, 이후 차분이 증가함에 따라 모형의 성능개선 효과가 점점 작아져 약 15일 이후에는 입력변수에 차분을 적용함에 따른 모형성능 효과는 크지 않으며, 차분일수에 따라 약간의 증감이 있으나 차분을 적용하지 않은 경우와 유사한 수준을 보여주는 것으로 분석되었다 (Fig. 4).

모형의 성능이 가장 크게 개선된 1일의 차분을 적용한 경우 RSR, NSE, MAE는 각각 0.359, 0.871, 1.510으로 산출되었다. 차분을 적용하지 않은 경우와 1일의 차분을 적용한 모형의 실측값(observation)과 예측값(prediction)을 (Fig. 5)에 비교하였다. 실측값과 모형의 예측값을 비교한 결과 모형의 성능 평가에 활용된 지수인 RSR, NSE, MAE가 각각 차분을 적용하기 전과 후에 0.741에서 0.359로, 0.451에서 0.871로, 3.763에서 1.510로 개선되었으며, Fig. 5에 제시된 바와 같이 1:1 line에 근접하여 분포하여 차분적용에 따른 모형성능의 향상을 확인할 수 있었다.



(a) RSR

(b) NSE

(c) MAE

Fig. 4. Evaluation of model simulation using different time lags of Chl-*a* concentration.

1일의 차분을 적용한 모형의 입력변수의 FI를 산정한 결과 1일 전의 Chl-*a* 농도가 전체 변수의 중요도 총합의 41%를 차지하여 가장 높은 영향을 미치는 것으로 분석되었으며, 차분을 적용하지 않을 경우와는 다르게 수온이 아닌 TOC가 2번째로 높은 FI값을 가지는 입력변수로 분석되었다 (Fig. 6).

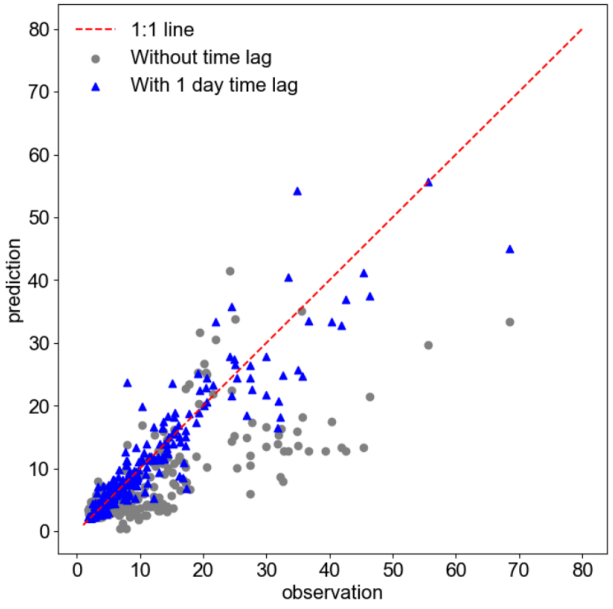


Fig 5. A comparison of model prediction with different time lag.

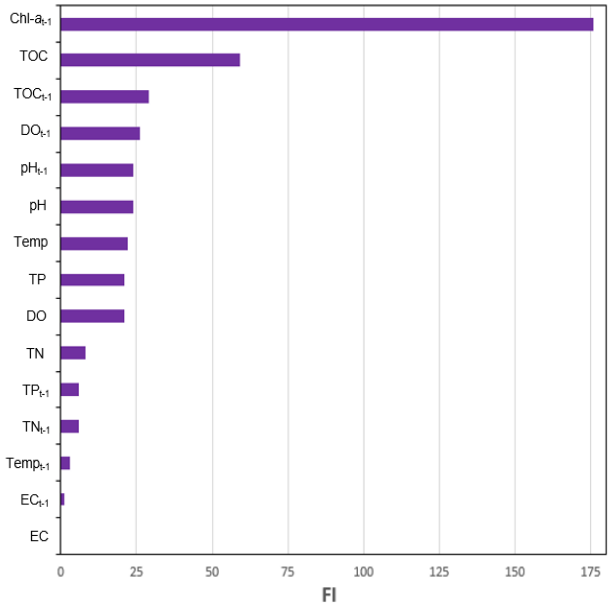


Fig. 6. FI of lightGBM using 1 day time lag of input variables.

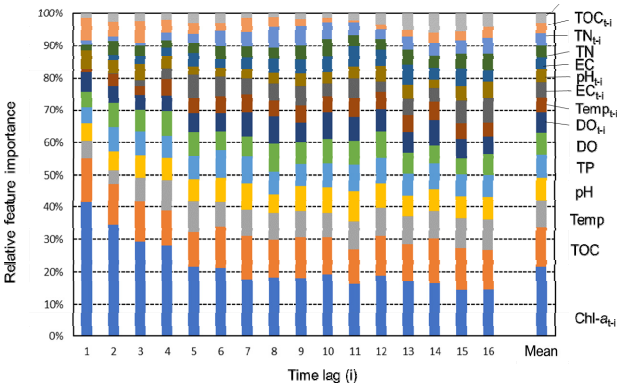


Fig. 7. FI of LGMB with different time lags of input variables.

차분을 다르게 적용한 모형의 성능에 영향을 미치는 입력변수의 상대적 중요도를 분석하기 위해 16일까지의 차분을 적용한 각 모형의 입력변수별 FI를 비교하였다. 각 모형의 FI 총합은 일정하지 않으므로 입력변수별 상대적인 중요도를 확인하기 위해 각 차분별로 모형의 FI의 총합을 구한 후 각 입력변수의 FI를 이 총합으로 나누어 총합이 100%가 되도록 표준화하여 그 결과를 Fig. 7에 제시하였다. 1일의 차분을 적용할 경우에는 전일의 Chl-*a* 농도가 다른 입력변수에 비해 매우 높은 영향을 미치는 것으로 분석되었으나, 차분을 증가시킴에 따라 그 상대적 중요도는 점차 감소하여 약 15일 이후에는 TOC와 유사한 FI를 가지는 것으로 분석되었다. 16일까지의 입력변수별 FI의 평균값은 차분을 적용한 Chl-*a*가 가장 높았으며, 다른 항목의 경우 EC와 TN 이외의 항목은 동일 일자의 입력변수가 차분을 적용한 경우보다 높은 FI를 가지는 것으로 분석되었다.

3.3 머신러닝 모형 성능에 대한 입력변수 영향에 대한 고찰

본 연구를 통해 과거의 측정값을 모형의 입력변수로 활용함에 모형의 성능에 미치는 영향과 차분일수에 따른 FI의 변화를 확인하였다. 차분을 적용하는 것은 이전의 측정정보를 이용하여 미래의 수질을 예측하는 것으로 그 차이가 작을수록 예측성능의 향상에 도움을 줄 수 있을 것을 추정할 수 있다. 모형의 성능 향상을 위해 입력자료를 추가적으로 확보하기 위해서는 자료수집을 위한 인력과 소요시간의 증가 및 이에 따른 비용의 증가가 필요해 현실적으로 제한이 있는 경우가 많다. 따라서, 추가적인 자료의 확보 없이 이

전의 측정정보를 활용하여 모형의 성능을 향상하는 것은 실무적인 측면에서 수질의 변화를 예측하는데 유용하게 활용될 수 있다. 본 연구에서는 차분의 일수가 모형의 성능에 미치는 영향을 분석하였으며 LightGBM 모형에서 약 2주 정도의 차분을 적용한 입력변수를 적용하는 것은 모형이 성능향상에 도움을 줄 수 있는 것을 확인할 수 있었다. LightGBM 모형의 입력변수의 FI는 머신러닝 모형의 내부 알고리즘에 의한 의사결정과정에서 입력변수가 모형성능에 미치는 영향을 제시하는 것으로 예측의 대상이 되는 종속변수와 예측에 활용되는 독립변수간의 물리적 혹은 화학적 인과관계를 설명하지는 않는다. 하지만 모형에 영향을 주는 입력변수의 특성에 대한 정보를 제시하여 모형의 성능향상에 도움을 줄 수 있는 정보를 제시할 수 있다. 머신러닝 모형은 내부 알고리즘의 복잡성으로 모형이 결과에 영향을 미치는 요인을 해석하거나 물리적 화학적 현황과 연관성을 확인하는 것이 쉽지 않으며 이는 black box 모형인 머신러닝 모형의 단점 중 하나로 제시되고 있다. FI는 머신러닝 모형의 성능에 영향을 미치는 요인을 정량적으로 제시하는 대표적인 방법 중 하나이며, 최근에는 머신러닝 모형의 예측결과에 대한 해석을 위한 설명 가능한 인공지능(Explainable machine learning, XAI) 등에 대한 연구도 이루어지고 있다. 향후 머신러닝 모형의 결과를 해석하고 모형성능에 영향을 미치는 영향인자에 대한 지속적인 연구를 통해 머신러닝 모형이 성능을 개선하고 모형의 실용성과 활용성을 높일 수 있을 것이다.

4. 결 론

본 연구에서는 LightGBM을 이용하여 입력변수 특성에 따른 하천 Chl-*a* 농도를 예측하는 모형의 성능을 비교하였다. 입력변수에 차분을 적용한 자료의 활용이 모형성능에 미치는 영향을 분석하기 위하여 1일(t-1)～50일(t-50)의 차분을 적용하였을 때 모형의 성능을 3가지 지표를 이용하여 비교한 결과 유사한 경향을 보여주었다. 1일의 차분을 적용할 경우 모형의 성능이 가장 크게 개선이 되었으며, 1일의 차분을 적용한 모형의 입력변수의 FI 분석결과 1일전의 Chl-*a* 농도의 FI가 전체의 41%로 가장 높은 영향을 미치는 것으로 분석되었고, 당일은 TOC가 2번째로 높은 영향을 미치는 입력변수로 분석이 되었다.

차분의 적용에 따라 입력변수가 모형의 성능에 미치는 영향을 비교하기 위해 16일까지의 차분을 적용한 값을 LightGBM의 입력변수로 이용한 결과 16일간의 평균 FI는 Chl-*a*가 가장 높은 것으로 분석되었으나, 차분이 커질수록 점차 감소하여 15일 이후에는 TOC와 유사한 수준을 보이는 것으로 분석되었다.

본 연구를 통하여 입력변수에 차분을 적용하는 것이 일정기간까지는 LightGBM 모형의 성능향상에 도움을 주는 것을 확인하였다.

사 사

본 연구는 환경부의 재원으로 한국환경산업기술원의 수생태계 건강성 확보 기술개발사업의 지원을 받아 연구되었습니다(과제번호 : 2020003030006).

References

Belgiu, M. and Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions, ISPRS J. Photogramm. Remote Sens., 114, 24-31.

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*. 1-15.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst., 30, 3146-3154.

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/ (May 22, 2021).

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.

Lee, S.M., Park, K.D. and Kim, I.K. (2020). Comparison of machine learning algorithms for Chl-*a* prediction in the middle of Nakdong river (focusing on water quality and quantity factors), J. Korean Soc. Water Wastewater, 34(4), 277-288.

LightGBM. https://lightgbm.readthedocs.io/ en/latest/ (August, 2021).

Lim, H.S. and An, H.U. (2018). "Prediction of pollution loads in Geum River using machine learning", *Proceedings of the Korea Water Resources Association Conference*, Korea Water Resources Association.

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.

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 (May 22, 2021).

Oh, H.R., Son, A.L., and Lee, Z. (2021). Occupational accident prediction modeling and analysis using SHAP, J. Digit. Contents Soc., 22(7), 1115-1123.

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.

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.

Su, Y. and Zhao, Y. (2020). Prediction of downstream BOD based on light gradient boosting machine method, IEEE, 127-130.

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.

Zhou, Z.H. (2021). *Machine Learning.* 181-210.