ISSN(Print): 1225-7672 / ISSN(Online): 2287-822X
DOI https://doi.org/10.11001/jksww.2020.34.4.277

머신러닝 기법을 활용한 낙동강 중류 지역의 Chl-a 예측 알고리즘 비교 연구(수질인자 및 수량 중심으로)

Comparison of machine learning algorithms for Chl-a prediction in the middle of Nakdong River (focusing on water quality and quantity factors)

이상민[1]·박경덕[2]·김일규[1,*]
Sang-Min Lee[1]·Kyeong-Deok Park[2]·Il-Kyu Kim[3,*]

[1]부경대학교 환경공학과
[2]부경대학교 마린융합디자인공학과
[1]*Department of Environmental Engineering, Pukyong National University*
[2]*Department of Marine Design Convergence engineering, Pukyong National University*

ABSTRACT

In this study, we performed algorithms to predict algae of Chlorophyll-a (Chl-a). Water quality and quantity data of the middle Nakdong River area were used. At first, the correlation analysis between Chl-a and water quality and quantity data was studied. We extracted ten factors of high importance for water quality and quantity data about the two weirs. Algorithms predicted how ten factors affected Chl-a occurrence. We performed algorithms about decision tree, random forest, elastic net, gradient boosting with Python. The root mean square error (RMSE) value was used to evaluate excellent algorithms. The gradient boosting showed 10.55 of RMSE value for the Gangjeonggoryeong (GG) site and 11.43 of

Received 8 July 2020, revised 11 August 2020, accepted 14 August 2020.
*Corresponding author: Il-Kyu Kim (E-mail: ikkim@pknu.ac.kr)

• 이상민 (박사) / Sang-Min Lee (Ph.D. candidate)
부산광역시 남구 용소로 45, 48513
45 Yongso-ro, Nam-gu, Busan 48513, Republic of Korea

• 박경덕 (박사) / Kyeong-Deok Park (Ph.D. candidate)
부산광역시 남구 용소로 45, 48513
45 Yongso-ro, Nam-gu, Busan 48513, Republic of Korea

• 김일규 (교수) / Il-Kyu Kim (Professor)
부산광역시 남구 용소로 45, 48513
45 Yongso-ro, Nam-gu, Busan 48513, Republic of Korea

RMSE value for the Dalsung (DS) site. The gradient boosting algorithm showed excellent results for GG and DS sites. Prediction value for the four algorithms was also evaluated through the Receiver operating characteristic (ROC) curve and Area under curve (AUC). As a result of the evaluation, the AUC value was 0.877 at GG site and the AUC value was 0.951 at DS site. So the algorithm's ability to interpret seemed to be excellent.

Key words: Chlorophyll-a (Chl-a), Machine learning, Gradient boosting, Nakdong River, RMSE, ROC curve

주제어: 클로로필-a, 머신러닝, 경사하강증폭법, 낙동강, 평균제곱근편차, 수신자판단속성곡선

1. 서 론

낙동강은 강원도 태백시에서 발원하여 영남지역을 관통하고 남해에 이르기까지 총 유로연장 525.8 km로 한반도에서 압록강 다음으로 두 번째로 긴 강이며, 유역면적은 23,860 km² 이다. 이번 연구에서는 낙동강에서 조류가 빈번히 발생하고 있는 지점인 낙동강 중류지역으로 상류지역에서의 축산폐수, 생활하수와 농경지 유출수 등 점오염원의 영향을 받을 것으로 예상되는 최대상수원 지역인 강정고령보 수계와 경산, 영천, 대구를 가로질러 흐르는 금호강의 유입으로 인한 산업폐수, 생활폐수의 영향이 있을 것으로 보이는 달성보 수계를 대상으로 하였다 (Lee, 2013). 특히 낙동강은 비점오염원에 노출되어 있으며 중・하류구간은 인구밀집 지역 및 공업지역으로 인해 그 수질을 관리하기 어려운 수계이다 (Hwang, 2012). 전 지구적 기후변화로 인한 재해예방, 수자원확보, 수질개선 등을 목표로 4대강 사업이 진행되어 8개의 보가 완공되었다. 보가 완공됨으로 인하여 낙동강의 수계는 체류시간의 변화로 폐쇄성 수역의 특성을 나타내고 있으며(Lee et al., 2014) 여름철에는 기온상승에 의한 *Microsystis sp.* 등의 남조류 증식으로 생태계와 인체에 악영향을 끼칠 수 있으며 이는 신경독성 및 급성간질환과 연관이 있다고 알려져 있다 (Falconer and Humpage, 2005).

과거 낙동강의 Chl-a 연구 방향은 하천의 수질자료를 이용하여 Chl-a 성장에 영향을 미치는 수질인자에 대한 통계분석인 상관분석, 다중회귀분석 등을 진행해 왔으며 최근에는 머신러닝을 통해 위성 영상자료 등을 이용하여 Chl-a를 예측한 사례가 있다. 하지만 조류발생 모니터링 방법으로 Chl-a의 위성영상 분광특성을 이용하여 농도 값을 예측하는 방식은 위성영상 밴드를 이용한 모델식이 다양하게 존재 하지만 예측 모델을 사용하기 위한 수계 환경변화를 반영하는 농도 계산식의 정확도가 떨어지는 문제점이 존재하는 것으로 알려져 있다 (Chun and Eun, 2017). 그리고 국내에서도 수질자료를 활용하여 머신러닝 알고리즘인 인공신경망을 통해 낙동강 Chl-a를 예측한 사례가 있으나(Park, 2015; Kim, 2017) 인공신경망 알고리즘과 수질자료만 활용하였으며 Chl-a를 정확하게 예측하기 위해선 조류 발생에 영향을 미치는 다양한 데이터를 통해 적합한 알고리즘을 찾을 필요가 있다.

이번 연구는 Python 프로그램을 기반으로 최근 낙동강 Chl-a 농도 예측으로 조류와 연관된 국내에서 많이 사용되고 있는 알고리즘을 비교하고자 하였으며 특히 조류가 빈번히 발생하고 있는 낙동강 중류지역의 Chl-a 농도를 우수하게 예측하는 알고리즘을 수행하고자 하였다. 최근 국내에서는 낙동강 조류에 관한 연구로 decision tree를 활용한 사례가 있으며 낙동강 본류 구간의 남조류 발생특성을 파악하고자 조류경보제 기반의 발령기준을 범주형 목표변수로 하여 주요 영향인자에 따른 남조류 발생 조건을 연구한 결과가 있다 (Jung et al., 2019). 머신러닝 기반으로 random forest를 활용하여 낙동강의 Chl-a 농도를 예측한 연구사례도 있다 (Kim et al., 2018). 그리고 스페인의 트라소나 저수지에 활용된 gradient boosting은 다양한 수질인자를 통하여 남조류 개체수와 Chl-a 농도 예측을 정확하게 구현 하였다 (Nieto et al., 2018). 마지막으로 분석기법 중 대표적인 예측기법인 회귀분석에서 ridge regression과 lasso regression의 장점을 모두 가지고 있고 큰 데이터셋에서 잘 작동하는 대표적인 회귀분석 알고리즘인 elastic net 또한 활용할 필요가 있으며 이번 연구에서는 이러한 4가지 알고리즘을 활용하고자 하였다.

조류의 발생에 영향을 미치는 인자는 매우 다양하며, 수질 인자뿐만 아니라 기상 및 수리 영향인자를 포함한 복합적인 평가가 이루어져야 한다. 조류의 성장에 영향을 미치는 여러 인자는 영양물질(N, P), 일사량, 수온, 물순환정체(체류시간) 등이 있으며 수질인

자 뿐만 아니라 수체의 물리적 조건인 수체 위치, 길이와 폭, 수심, 저수용량에도 영향을 받는다 (Caissie et al., 2007). 따라서 이번 연구는 조류 발생 여러 인자 중 오염물질이 집중적으로 유입되는 낙동강 중류 지역의 2개보에 걸쳐서 보 설치 이후에 지속적으로 측정·관찰되고 있는 Chl-a와 수질측정자료, 보 운영 자료인 수량자료를 같이 활용하여 특징적인 상관관계를 우선 도출하여 낙동강 중류지역의 Chl-a를 예측하고자 하였다. 이번 연구는 향후 낙동강 중류 보 구간의 수질관리정책 수립을 위한 기초 자료를 제공하는 데 목적이 있다.

2. 연구방법

2.1 조사지점 및 시기

본 연구에서는 물환경측정망 및 수질측정망의 자료 중 낙동강권역에서 2012년에 건설된 낙동강의 보건설 지점으로 조류가 빈번히 발생하고 있는 중류지점인 Fig. 1의 GG(강정고령보), DS(달성보)을 대상으로 하였다. GG지점은 연장 953.5 m, 관리수위 19.5 m, 저수용량 107.7백만 m^3이며 DS지점은 연장 580 m, 관리수위 14 m, 저수용량 56백만 m^3로 낙동강 8개보 중 중류지역을 대표하는 보이다. 수질자료는 수질자료를 획득할 수 있는 GG의 다사, DS의 논공지점을 대상으로

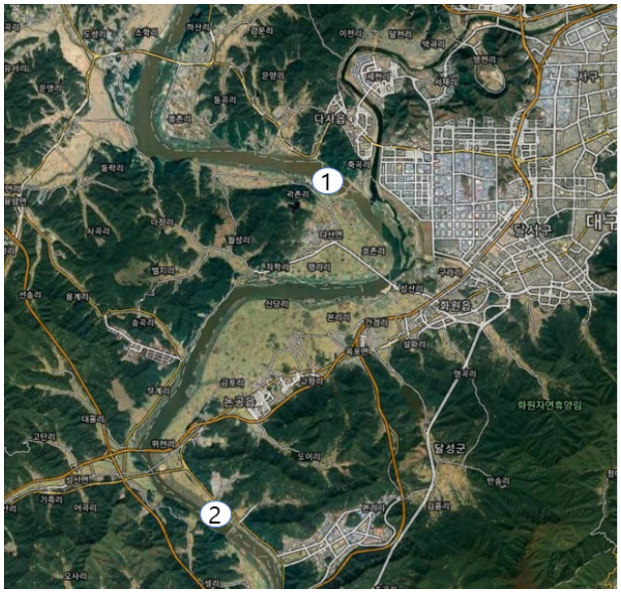


Fig. 1. Sampling sites (1) GG and (2) DS in the Nakdong River.

하여 두 지점의 조류발생 측정시점인 2012년 6월부터 2019년 11월까지로 하였으며 수량자료는 물환경정보시스템에서 제공하는 자료를 이용하였다.

2.2 자료수집 및 기계학습

수질자료는 수질측정망 보 지점(하천)의 자료는 물환경정보시스템(http://water.nier.go.kr)의 매주 1회 측정되고 있는 Chl-a, pH, DO (dissolved oxygen), BOD, COD, SS, 총질소(T-N), 총인(T-P), TOC, 수온(W·T), 전기전도도(E·C), 용존총질소(DTN), 암모니아성질소(NH_3-N), 질산성질소(NO_3-N), 용존총인(DTP), 인산염인(PO_4-P)를 이용하였고, 수량자료는 보 하류 수위(DWL, Downstream water level), 저수량(Flux), 공용량(WV, Water volume), 유입량(Inflow), 총 방류량(TFR, Total flow rate) 자료를 추출하여 이용하였다.

데이터 set은 주로 수질자료를 중심으로 주간단위로 구성되어 매일 측정되는 수량자료를 주단 단위로 조정하였으며 2012년 6월부터 2019년 11월까지 총 384주간의 데이터를 통합하여 21종의 변수로 데이터 set을 구성하였다. Training data는 80%, testing data는 20%로 설정하여 두 지점의 Chl-a를 예측하는 4가지 알고리즘을 수행하였다.

2.3 분석방법

2.3.1 상관관계 분석

머신러닝 기반 Chl-a 예측 알고리즘을 수행하기 위하여 본 연구에서 이용하는 수질자료와 수량자료에 대해 상관관계 분석을 우선 실시하였다. 수질과 수량 요소간의 상관성이 있는 항목을 우선 찾아내고 중요도가 높은 항목에 대해 decision tree, random forest, elastic net, gradient boosting의 알고리즘을 실행하였다.

2.3.2 Decision tree

Decision tree는 일반적으로 종속변수가 이산형 변수일 때 의사 결정 규칙을 정하고 독립변수를 분류한다. 하향식 나무구조로 도표화하여 결과를 예측하는 방법으로 하나의 나무 형태로 구성되어 있고 나무의 각 마디를 노드라고 하며 맨 위에 위치한 노드를 뿌리 노드(root node), 중간에 이어지는 선들은 가지(branch)라 한다. 중간에 위치한 노드는 중간 노드(internal

node), 가장 아래에 위치한 노드를 잎사귀 노드(leaf node)라 부르며 가지를 이루고 있는 마디의 개수를 깊이(depth)라고 한다 (Cho, 2019). Classification and Regression Trees (CART)라는 decision tree는 비모수적인 분류 및 회귀 기법으로 결과를 시각적으로 이해하고 분석하기 쉽기 때문에 널리 사용되고 있으며 스케일링(scaling) 같은 전처리 과정이 상대적으로 필요하지 않으며 비선형 관계가 결과에 큰 영향을 미치지 않는 장점이 있다. 반면 아주 복잡한 tree를 생성하는 훈련 세트는 매우 잘 설명할 수 있지만 검증 세트는 잘 설명하지 못하는 over fitting 문제가 발생할 수 있으며 데이터의 작은 변화에 결과가 크게 변할 수 있다. 각각의 단계에서 지역적 최적화를 구해 전체 최적화를 근사화하는 greedy 알고리즘이므로 전체적으로 최적인 decision tree를 보장하지는 못한다. decision tree는 CART 손실함수를 최소화해서 구해진다. 아래 식 (1)은 CART 손실함수를 보여준다.

$$J(k,t_k)=\frac{m_l}{m}+\frac{m_r}{m}G_r \tag{1}$$

m은 표본 수, m_l는 왼쪽 노드의 표본수를 표시한다. Gi(i=left, right)는 노드 i 의 불순도(impurity)를 표시하는데 통상 지니(Gini), 엔트로피(entrophy), 분류오류(misallocation) 지표를 많이 사용한다. 지니(Gini)의 경우 아래 식 (2)와 같으며 $p_{i,k}$는 노드 i에서 분류집단 k에 속하는 확률이다.

$$G_i=1-\sum_{k=1}^{n}p_{i,k}^2 \tag{2}$$

아래 식 (3)과 (4)는 각각 엔트로피와 분류오류를 나타낸다(park and Ko., 2019).

$$G_i=-\sum_{k=1}^{n}p_{i,k}\log(p_{i,k}) \tag{3}$$

$$G_i=1-\max(p_{i,k}) \tag{4}$$

2.3.3 Random forest

Random forest 알고리즘은 루트 노드(root node), 내부 노드(internal node), 터미널 노드(terminal node or leaf) 등의 노드(node)들과 에지(edge)들의 집합으로 이루어지는 decision tree에서 파생한 방법이다 (Rokach and Maimon, 2005). 단일의 decision tree는 휴리스틱 기법을 기반으로 한 머신러닝 방법으로 최적의 decision tree를 학습한다는 보장이 어려운 반면, random forest 기반의 기계학습은 주어진 데이터를 통해 무작위 방식으로 구성된 다수의 decision tree를 만들어 최적의 학습 모델을 찾는 방법으로 알려져 있다 (Müller and Guido, 2016). 이러한 특성은 각 tree들이 예측한 결과가 상관화 되지(decorrelation) 않아서 결과적으로 일반화 된 결과를 유도하는 것뿐만 아니라 random forest는 노이즈가 포함된 데이터에 대해서도 강인성을 갖기 때문에 다양한 분야에서 활용되고 있다 (Kim and Seo, 2020). Random forest는 집단 학습을 기반으로 고정밀 분류, 회귀, 크러스트링을 구현하는 알고리즘으로 여러 개의 결정 tree들을 임의로 학습하는 앙상블 방법으로 크게 다수의 결정 tree를 구성하는 학습 단계는 입력 벡터가 들어 왔을 때 분류하거나 예측하는 테스트 단계로 구성되어 검출, 분류, 회귀 등에도 활용된다 (Cho, 2019).

2.3.4 Elastic net

Ridge regression 식 (5)는 계수의 제곱의 합을 최소화하는 것을 추가적인 제약으로 하며 λ는 기존의 잔차 제곱합과 추가 제약 조건의 비중을 조절하기 위한 hyper parameter이다. λ가 커지면 정규화 정도가 커지므로 가중치의 값들은 작아지며 λ가 작아지면 정규화 정도가 작아지고 λ가 0이 되면 일반선형 회귀모형이 된다 (Hoerl and Kennard, 1970).

$$\hat{\beta}=\underset{\beta}{\operatorname{argmin}}\left(\|y-X\beta\|^2+\lambda\|\beta\|^2\right) \tag{5}$$

Lasso regression 식 (6)은 가중치의 절대값의 합을 최소화하는 것을 추가적인 조건으로 실행되며 계수의 절대값은 상수로 계수를 0으로 만드는 경향이 있다 (Tibshirani, 1996).

$$\hat{\beta}=\underset{\beta}{\operatorname{argmin}}\left(\|y-X\beta\|^2+\lambda\|\beta\|_1\right) \tag{6}$$

Elastic net은 L_1과 L_2페널티의 조합으로 ridge regression와 lasso regression 정규화가 결합된 알고리즘으로 식은 (7)과 같이 나타낼 수 있으며 가중치의

절대값의 합과 제곱합을 동시에 제약 조건으로 가지는 모형이다 (Zou and Hastie., 2005).

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\| y - X\beta \|^2 + \lambda^2 \| \beta \|^2 + \lambda_1 \| \beta \|_1),$$

$$a = \frac{\lambda_2}{(\lambda_2 + \lambda_1)}$$

$$\hat{\beta} = \underset{\beta \text{ subject to } (1-\alpha)\| \beta \|_1 + \alpha \| \beta \|^2}{\operatorname{argmin}} \| y - X\beta \|^2, \quad (7)$$

Elastic net은 잔차 제곱합을 최소로 만드는 회귀계수를 구하는 최소제곱이 가장 단순한 회귀모형 예측법이고 최소제곱법으로 다수의 설명변수를 모형화 하는 경우 다중공선성이 발생하는 문제가 있을 수 있으며 예측변수가 많아질수록 예측력은 증가하나 새로운 자료에 대한 over fitting 문제가 발생 가능하다. 그러나 elastic net은 다중공선성 및 over fitting을 다루는데 적합하고 변수가 많은 대용량 자료를 분석할 때 유용하게 적용된다 (Lee and Yoo, 2019).

2.3.5 Gradient boosting

머신러닝에서 boosting은 부정확한 weak learner을 혼합하여 좀 더 정확하고 strong learner을 만드는 방식이다. 우선 정확도가 낮은 첫 번째 tree 모델을 만들고 난 이후, 예측에서 발견된 오류는 두 번째 tree 모델에 보완된다. 이런 방법으로 다음 tree 모델에서 약점을 계속 보완하여 향후에는 강한 학습기를 구축한다. Loss function(손실함수, J)는 예측 모델의 오류를 정량화하며, loss function 값을 최소화하여 알고리즘 내 파라미터를 찾기 위하여 머신러닝 알고리즘은 gradient descent를 사용한다. gradient boosting은 이러한 파라미터 손실함수 최소화 과정을 모델 함수(f_i) 공간에서 수행하며, 손실함수를 알고리즘 파라미터가 아니라 다음과 같은 (8)번 수식에 의해 현재까지 학습된 tree 모델 함수로 미분하며 수식에서 ρ 는 학습률을 나타낸다.

$$f_{i+1} = f_i - \rho \frac{\delta J}{\delta f_i} \quad (8)$$

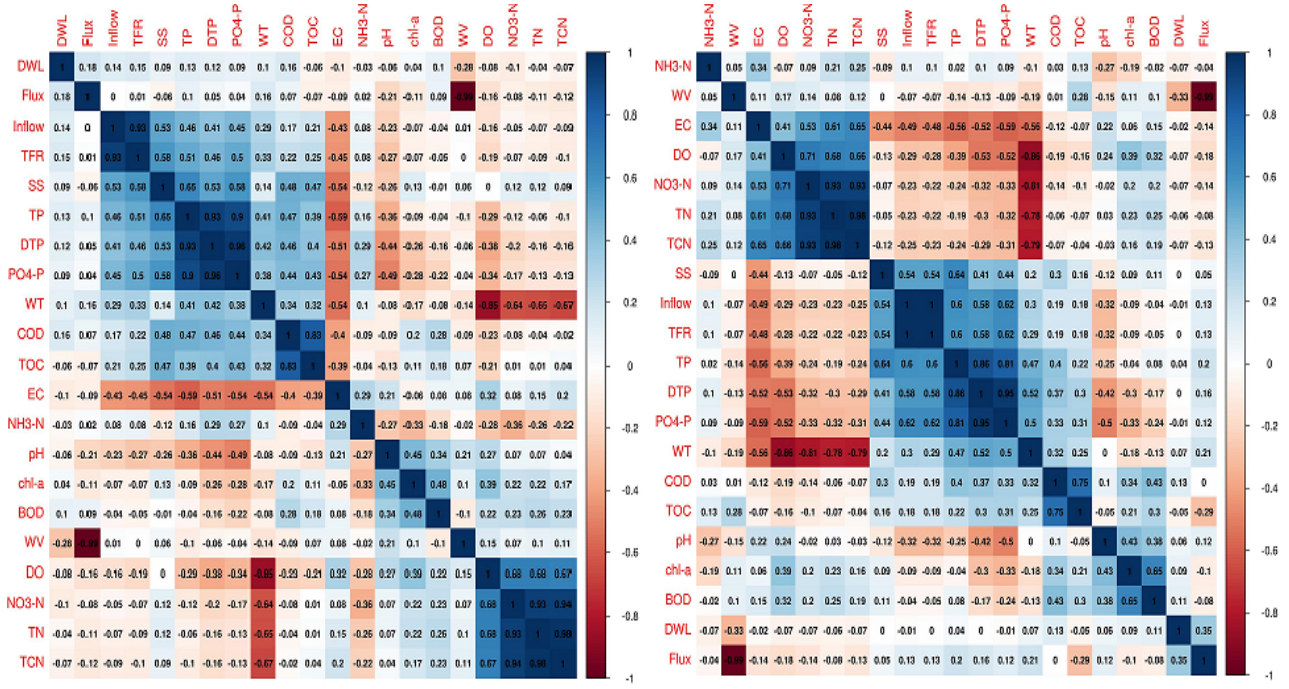
Gradient boosting 알고리즘에서 tree 알고리즘 함수 미분값은 현재까지 학습된 알고리즘의 약점을 나타내는 역할을 하며, 다음 tree 알고리즘의 fitting을 수행할 때 그 미분값을 사용하여 약점을 보완하여 성능을 향상시킨다 (Heo et al., 2018). Gradient boosting은 decision tree 시퀀스를 반복적으로 학습시키며 각 반복에서 알고리즘은 현재 그룹을 사용하여 각 트레이닝 인스턴스 레이블을 예측하고 실제 레이블을 예측하고 비교한다. 빈약한 예측으로 훈련 인스턴스에 더 중점을 두기 위해 데이터 세트의 레이블이 다시 지정되며 gradient boosting은 반복에서 decision tree에 의한 실수를 정정한다. 최신 버전은 병렬처리, 트리정리, 결측값 처리 및 정규화를 통해 최적화 된 gradient boosting으로 over fitting이나 바이어스를 방지하는 장점이 있다 (Krishna et al., 2019).

3. 결과 및 고찰

3.1. 수질항목별 상관분석 결과

조류는 수질, 수량, 기상인자 등 다양한 영향인자에 영향을 받고 발생한다. 그리고 이러한 영향인자는 조류발생에 서로 영향을 미치므로 Chl-a와의 상관관계를 통해 해석하여 적용할 필요가 있다. Chl-a에 영향을 미치는 주요 영향인자를 파악하기 위해 각 보별 Chl-a와 수질, 수량과의 상관관계를 분석하였으며 그 결과는 Fig. 2와 같다.

상관도가 떨어지는 요인을 제거하기 위한 전단계로 상관관계 분석을 수행하였고 중요인자를 추출하였으며 두 지점의 상관계수 절대값이 0.1 이상과 인자의 수가 10가지일 경우 머신러닝 알고리즘이 보다 정확하게 구현되었다. Chl-a와 상관관계를 pearson 상관계수를 통해 분석한 결과 GG지점은 BOD, DO, pH, SS, WV, TOC, NH_3-N, E·C, Flux, T-P, DS지점은 BOD, DO, COD, TOC, DTP, SS, E·C, T-P, PO_4-P, NH_3-N으로 나타났다. 특히 BOD, COD, DO, pH는 상관계수가 0.3이상으로 유의한 상관관계를 나타내었다. pH는 조류 및 박테리아 성장에 영향을 미치는 영향인자며, 대부분의 조류는 중성영역인 pH 7~9 사이를 선호하며 질산염(NO_3^-)을 질소원으로 사용할 때는 pH는 상승하고, 암모니아(NH_4^-)를 질소원으로 사용할 때는 pH가 낮아져 DO와도 연관이 있다. 조류는 다른 환경 요인이 만족된다 해도 어떤 pH의 범위 내에서만 잘 자라고 pH가 맞지 않으면 사멸하거나 활성이 떨어지므로(Twisti et al., 1988) 조류와 pH는 상



(A) (B)

Fig. 2. Scatter plot between target output and input variables about (A) GG and (B) DS.

관성이 높다. 그리고 유기물 수질 지표인 BOD, COD, TOC 등은 Chl-a와 높은 상관성을 보이고 있으며 조류성장에 따른 자생 유기물질과 관련이 있는 것으로 추정해 볼 수 있다 (Kim et al., 2013). 이외의 영양염류인 질소와 인에서도 상관성이 나타났으며 NH_3-N, DTP는 상관계수 절대값 0.3이상의 유의성을 나타내었으며 T-P, PO_4-P, DTP 등도 절대값 0.1 이상의 유의한 상관성을 나타내었다. 이번 연구의 유의한 수질인자 중 E·C는 수계의 구분이나 수질의 연속적 변화를 쉽게 파악하는 지표이며 조류가 증식하였을 때 수계의 상층 E·C는 높게 나타나며 SS는 집중호우의 경우 농경지, 도로 등과 같은 비점오염원에서 유입되어 높아지는 경향이 있으며 일정 부분은 조류와 같은 미생물 농도의 영향을 받는다. 수량자료인 WV, Flux 또한 작지만 0.1 이상의 상관성을 나타내었는데 조류의 성장에 영향을 미치는 여러 인자 중 물순환정체 등 수체의 물리적 조건인 저수용량 등에도 영향을 받는다 (Caissie et al., 2007). 낙동강 지역의 수질, 수리 및 기상인자와 Chl-a 농도 사이의 상관관계 분석결과 Chl-a에 대한 주요 영향인자 중 수질인자는 W·T, pH, DO, BOD, COD, T-N, NO_3-N, PO_4-P 8가지가 유의한 상관관계를 나타내었으며 유량, 유속 및 수심 등 수량인자도 영향인자로 나타난 연구 사례가 있다 (Lim et al., 2015). 특히 이번 연구결과와 지역이 일치하는 낙동강 중류 지역 강정고령보 지점으로 2012년 1월부터 2016년 10월까지 14개 수질인자와 Chl-a와의 상관관계 분석 결과 pH, DO, BOD, COD, SS, T-N, W·T, NH_3-N, NO_3-N, PO_4-P 등이 유의한 상관관계($p < 0.01$)를 나타내고 달성보는 pH, DO, BOD, COD, TN, T-P, W·T, NH_3-N, NO_3-N 등이 유의한 상관관계를 나타낸 연구결과가 있었으며 (Jung and Kim, 2017) 이번 연구 결과의 중요 수질인자와 대부분 일치하였다.

3.2. RMSE를 통한 알고리즘 성능평가

Root mean square error (RMSE)는 예측값과 측정값과 차이를 나타내는 척도로 예측력과 정확도를 확인할 수 있다. RMSE는 잔차(관측에서 나타나는 오차)의 제곱합을 산출 평균한 값의 제곱근으로서 측정값들 간의 상호간 편차를 의미한다. 표준편차를 일반화시킨 척도이며 예측값과 측정값과의 차이를 알려주는데 많이 사용하고 있는 척도로 RMSE와 표준편차는 개별 측정값이 중심으로부터 얼마나 멀리 떨어져 있는 정도를 나타낸다. 측정값(estimated parameter) θ에 대한

Table 1. Algorithms Performance Comparison between the estimated and measured values by analysis method

	GG				DS			
	Decision Tree	Random Forest	Elastic Net	Gradient Boosting	Decision Tree	Random Forest	Elastic Net	Gradient Boosting
MSE	259.83	136.57	198.35	**111.35**	161.63	152.98	246.52	**130.61**
RMSE	16.12	11.69	14.08	**10.55**	12.71	12.37	15.70	**11.43**

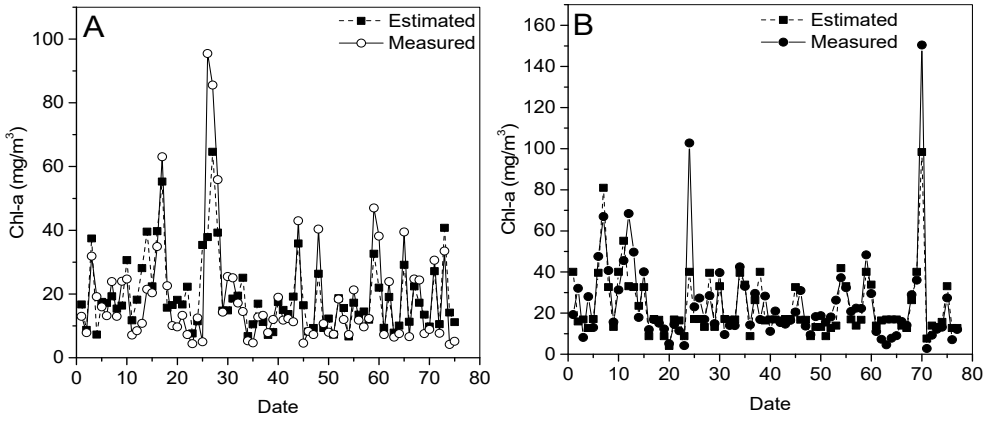


Fig. 3. Measured and estimated values of decision tree models about (A) GG and (B) DS.

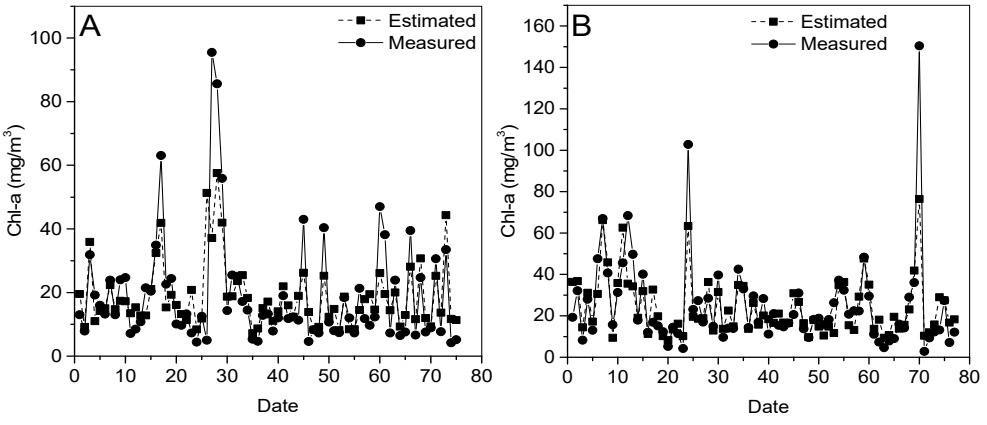


Fig. 4. Measured and estimated values of random forest about (A) GG and (B) DS.

예측값(estimator) $\hat{\theta}$의 RMSE 값은 식 (9)과 같이 구할 수 있다.

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{(\theta-\hat{\theta})^2} \quad (9)$$

RMSE값은 작을수록 예측값과 측정값의 차이가 작다는 것을 의미하며 0에 가까울수록 좋은 성능을 보인다고 할 수 있다. RMSE와 Mean squared error (MSE)는 통계모델링에서 이론적으로 관련되어 현재까지도 널리 사용되고 있다 (Hyndman and Koehler., 2006). 환경 관련분야에서도 훈련데이터로 훈련한 알고리즘의 성능 평가를 위해 RMSE가 많이 사용되고 있다 (Lawrence et al., 2004; Johnson et al., 2018; Krishna et al., 2019; Fan et al., 2019; Wei et al., 2019).

이번 연구에서 알고리즘 성능을 평가하기 위해 잔차오차기반 지표인 MSE, RMSE를 활용해 평가하였다. 데이터 set을 기반으로 하는 알고리즘 지표 비교는 Table 1과 같다.

GG지점에서의 MSE는 decision tree가 259.83, random forest는 136.57, elastic net 198.35, gradient boosting

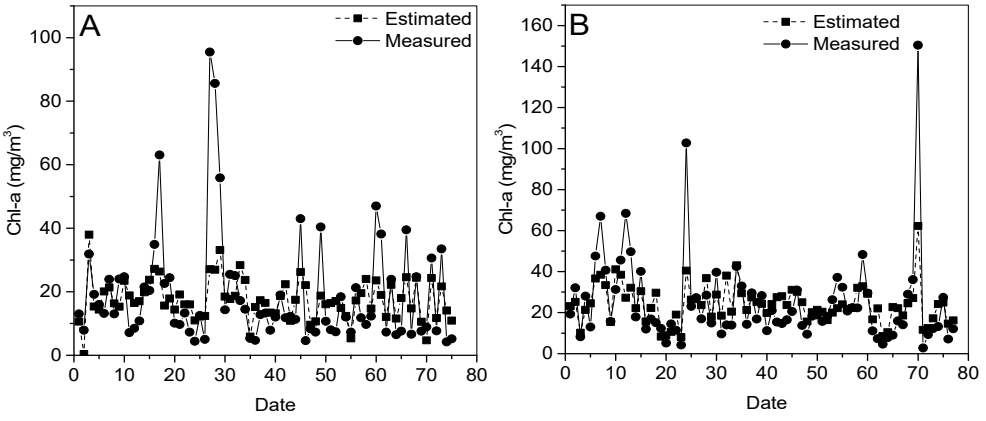


Fig. 5. Measured and estimated values of elastic net about (A) GG and (B) DS.

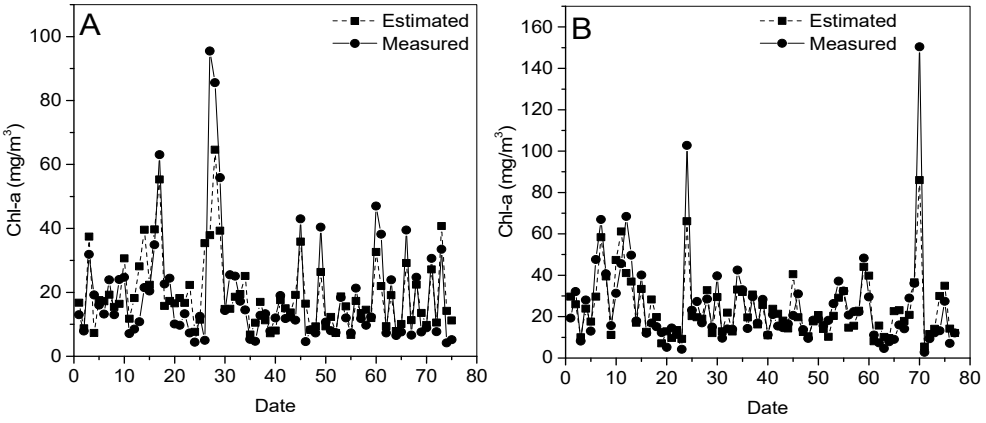


Fig. 6. Measured and estimated values of gradient boosting about (A) GG and (B) DS.

111.35로 나타났으며 DS지점에서의 MSE가 decision tree 161.63, random forest 152.98, elastic net 246.52, gradient boosting 130.61로 나타났다. 그리고 GG지점에서의 RMSE는 decision tree가 16.12, random forest는 11.69, elastic net 14.08, gradient boosting 10.55로 나타났으며 DS지점에서는 decision tree 12.71, random forest 12.37, elastic net 15.70, gradient boosting 11.43으로 나타났다. GG지점의 gradient boosting의 MSE가 111.35, RMSE은 10.55로 알고리즘의 성능이 높게 나타났으며 다음으로는 random forest, elastic net, decision tree 순으로 알고리즘 성능을 나타냈다. DS지점은 gradient boosting의 MSE가 130.61, RMSE은 11.43으로 알고리즘의 성능이 높게 나타났다. DS지점은 그 다음으로 좋은 성능을 나타낸 알고리즘은 random forest, decision tree, elastic net으로 나타났다. 이번 연구에서 GG, DS 두 지점 모두 gradient boosting이 좋은 성능의 알고리즘으로 평가되었다. 예측값과 측정값을 비교한 결과는 Fig. 3～6에서 볼 수 있다.

3.2 Receiver operating characteristic (ROC) 커브를 통한 예측값 정확성 평가

ROC분석은 의학 분야에서 많이 사용되며 최근에도 영상장비와 같이 질환의 유무와 예측을 진단하는 테스트의 효율성 평가에 주로 활용되어 왔다 (Metz, 1978). 최근 수질에 대한 예측지표나 위해도 기준을 설정하는 등 환경 분야에도 연구가 적용되고 있다 (Morrison et al., 2003; McLaughlin, 2012; Song et al., 2017). ROC커브는 알고리즘의 예측능력을 검증하는데 사용되며, ROC커브에서 커브 아래영역의 면적인 Area under curve (AUC)는 그 값이 0에서부터 1까지의 값을 가진다. 그리고 AUC는 1의 값에 가까워 질 때 좋은 알고리즘으로 평가받는다 (Do and Le, 2020). AUC값에 대한 평가 기준은 보통 5등급으로 분류하고 Table 2와 같이 해석할 수 있다.

Table 2. AUC performance metric about the algorithm

AUC	grade	interpretation
0.9 - 1.0	A	excellent
0.8 - 0.9	B	good
0.7 - 0.8	C	fair
0.6 - 0.7	D	poor
0.5 - 0.6	E	fail

ROC커브의 양성율(TPR, True positive rate)은 예측의 정확도에 대한 값을 나타내고 위양성율(FPR, False positive rate)은 예측의 부정확성에 대한 값을 나타내어 그 값들로 표현할 수 있다 (Dhaliwal et al., 2018; Krishna et al., 2019). Chl-a 예측값을 두 지점의 조류발생 지표로 사용하기 위해 이 지표가 조류관리단계를 판별하는데 효과적인지 확인할 필요가 있으며 지표가 효과적일 경우 예측값은 두 지점의 Chl-a 농도 평균이상 높은 민감도(sensitivity)와 특이도(specificity)로 나타낸다. 민감도는 실제 Chl-a가 두 지점의 조류발생 측정기간 동안 평균이상으로 GG지점은 17.26, DS지점은 23.79 이상 발생하는 부영양화된 수질로 분류되는 양성율(TPR)을 말하며, 특이도는 Chl-a가 평균미만으로 발생하는 양호한 수질로 분류되는 위양성율(FPR)로 나타낸다. 양성율(TPR)은 다시 민감도로 나타내어지고 True Positive (TP), False Negative (FN)으로 (10)식과 같이 계산된다. 반면 위양성율(FPR)은 1-민감도로 표현되고 True Negative (TN), False Positive (FP)로 식 (11)로 계산된다.

$$양성율(TPR) = \frac{TP}{FN + TP} \tag{10}$$

$$위양성율(FPR) = \frac{TN}{TN + FP} \tag{11}$$

이번 연구에서 4개 알고리즘의 두 지점 Chl-a 농도 평균 초과여부의 정확성은 민감도와 특이도를 통해 확인 할 수 있으며 민감도와 특이도에 대한 시각화는 Fig. 7~8에서 확인할 수 있다. Chl-a 농도를 통해 정확한 조류발생을 예측하기 위해 Chl-a 농도 예측은 가급적 양성율이 크고 위양성율이 작아야 의미 있는 지표로 활용 가능하다. 따라서 두 지점의 4개 알고리즘에 대한 ROC커브를 표현하고 AUC값을 도출하였다. 정확성 평가결과 GG지점에서 decision tree의 AUC는 0.790, elastic net의 AUC는 0.867, random forest의 AUC는 0.869, gradient boosting의 AUC가 0.877로 나타났다. DS지점에서 decision tree의 AUC는 0.878, elastic net의 AUC는 0.880, random forest의 AUC는 0.931, gradient boosting의 AUC가 0.951로 나타났다. Gradient boosting의 AUC값은 GG지점의 Chl-a 농도 17.26 초과여부에 대하여 0.877(95% 신뢰구간, 0.789~0.965), DS지점의 경우 Chl-a 농도 23.79 초과여부에 대하여 0.951(95% 신뢰구간, 0.923~0.999)으로 나타났다. AUC값이 1에 가까울 때 알고리즘의 예측값 평가가 우수하며(Do and Le, 2020), 두 지점 모두 AUC값이 1에 가까운 gradient boosting 알고리즘이 우수하게 나타났다.

Gradient boosting 알고리즘(Breiman et al., 1984; Vapnik, 1998; Friedman, 2002; Hastie et al., 2009)은 바람과 관련된 여러 수치들을 예측하고(Landry et al., 2016), 다양한 지역에서의 태양광 예측(Persson et al., 2017), 단기 폐기물 발생 예측(Johnson et al., 2017)과 같이 많은 환경문제를 해결하기 위해 생물학적 파라미터를 예측하는 효과적인 알고리즘으로 이용되고 있다 (Nieto et al., 2018). 최근에는 boosting 기법을 활용한 gradient boosting 알고리즘(Breiman et al., 1984; Vapnik, 1998; Friedman, 2002; Hastie et al., 2009)은 혐기성 조건에서 유기성폐기물 소화시설 설치를 위한 효율적인 결정(Clercq et al., 2019)과 도시환경에서 저비용으로 에어로졸 모니터링(Johnson et al., 2018) 등 폐기물, 대기, 수질 분야에서 활발하고 다양하게 적용되고 있다.

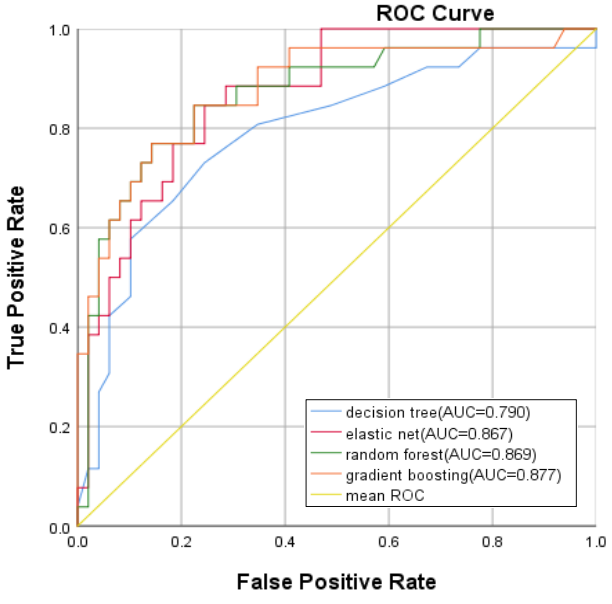


Fig. 7. Receiver operating characteristic curve about GG.

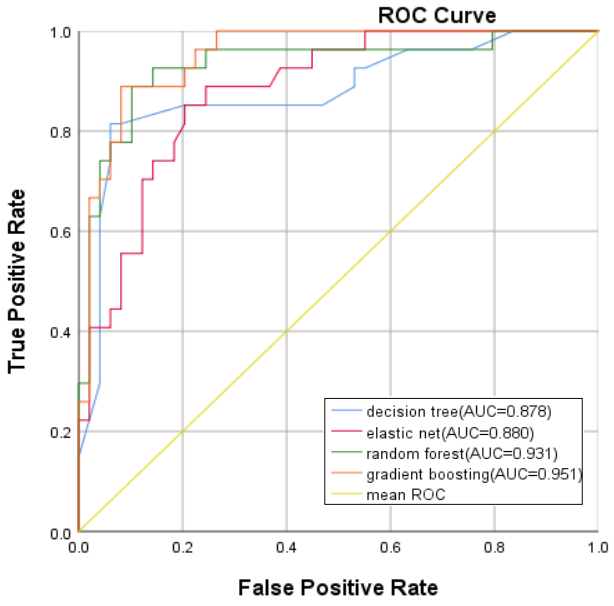


Fig. 8. Receiver operating characteristic curve about DS.

특히 gradient boosting 알고리즘은 회귀와 분류를 동시에 수행할 수 있는 지도학습의 한 종류이며 다변량 함수에 대해 범용근사자로 활용이 되고 있어(Vapnik, 1998; Friedman, 2002; Hastie et al., 2009) 환경관련 분야에서도 다양하고 우수하게 활용되고 있다.

4. 결 론

본 연구에서 낙동강 중류의 2개보 지점에서의 수질과 수량항목에 대한 Chl-a와의 상관성을 통해 10가지 중요한 인자를 추출하였다. 그리고 Chl-a를 목표변수로 하여 두 지점에서 중요인자를 통해 4가지 알고리즘을 비교 분석하였으며 그 결과는 다음과 같다.

두 지점은 상관성 중요인자와 인자별 상관계수 값이 서로 상이하였고 그 결과 두 지점별로 예측 성능이 다르게 나타난 것으로 평가된다. 특히 지점별로 변수 중요도는 차이가 났으며 두 지점의 공통된 수질인자로는 BOD, DO, SS, TOC, NH_3-N, E·C, T-P 등으로 평가되었다. 향후 낙동강 중류지역의 조류 발생을 억제하기 위하여 공통된 수질인자에 대한 추가적인 연구가 필요한 것으로 판단된다.

알고리즘의 예측력을 평가하기 위해 RMSE, MSE를 통해 평가한 결과 낙동강 중류의 두 지점 모두 gradient boosting 알고리즘이 우수하게 나타났다. 기존의 SPSS를 활용한 다중회귀분석 방법으로 낙동강 수계 Chl-a를 예측한 RMSE값 보다 우수하게 나타나 다른 통계방법 보다 머신러닝 알고리즘을 이용한 Chl-a 예측이 정확하였다.

예측값에 대한 정확성을 평가하기 위해 ROC커브 분석을 두 지점에 실시하였다. 두 지점 모두 gradient boosting 알고리즘이 예측값의 정확성이 높게 나타났으며 두 지점의 Chl-a 농도 예측값과 측정값에 대한 test score는 GG지점 62%, DS지점 72%로 나타났다. 알고리즘별로 일부 성능차이는 있었으나 크지는 않았으며 gradient boosting 알고리즘이 본 연구와 같이 수계환경 데이터를 기반으로 한 알고리즘으로 적용 가능하였다.

향후 조류가 빈번히 발생하고 있고 오염물질이 지속적으로 유입되어 조류가 빈번히 발생하는 낙동강 하류에 대한 연구도 필요하며 하류지역에도 알고리즘이 적용 가능하다면 보가 설치되어 있는 4대강을 중심으로 연구가 필요하다고 판단된다. 그리고 보다 정확한 예측을 위해서는 조류가 많이 발생하는 수계에 자동측정망 등을 구축하여 시계열적으로 지속적인 측정·분석된 자료와 강수량, 일사량, 운량 등 다양한 기상자료가 추가로 확보된다면 본 알고리즘을 활용한 수질 예측 성능은 더욱 향상될 수 있을 것이다.

사 사

이 논문은 2019년도 부경대학교 연구년 교수 지원사업에 의하여 연구되었음.

References

Breiman. L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and regression trees, Wadsworth Statistics/Probability Series, Wadsworth Advanced Books and Software.

Caissie, D., Satish, M.G., and El-Jabi, N. (2007). Predicting water temperatures using a deterministic model: Application on Miramichi River catchment(New Brunswick, Canada), J. Hydrol., 336, 303-315.

Chun, D.J. and Eun, J. (2017). Application method of remote sensing method for monitoring the water quality of big River, KEI Environmental Forum, 214, 21.

Cho, J. Y. (2019). Odor compounds forecasting in Daecheong water intake station using machine learning models, Doctor's Thesis, Chungnam National University, Daejeon, Korea.

Clercq, D.D., Wen, Z., and Fei, F. (2019). Determinants of efficiency in anaerobic bio-waste co-digestion facilities: A data envelopment analysis and gradient boosting approach, Appl. Energy, 253, 113570.

Dhaliwal, S.S., Nahid, A.A., and Abbas, R. (2018). Effective intrusion detection system using XGboost, Information, 9(7), 149.

Do, D.T. and Le, N.Q.K. (2020). Using extreme gradient boosting to identify origin of replication in Saccharomyces cerevisiae via hybrid features, Genomics. 112(3), 2445-2451.

Falconer, I.R. and Humpage, A.R. (2005). Health risk assessment of cyanobacterial (blue-green algal) toxins in drinking water, Int. J. Environ. Res. Public Health, 2(1), 43-50.

Fan, J., Ma, X., Wu, L., Zang, F., Yu, X., and Zeng, W. (2019). Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological date, Agric. Water Manag., 225, 105758.

Friedman, J.H. (2002). Stochastic gradient boosting, Comput. Stat. Data Anal., 38(4), 367-378.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: date mining, inference and prediction, Springer Series in Statistics, New York, 745.

Heo, J.S., Kwon, D,h., Kim, J,B., Han, Y.H., and An, C.H. (2018). Prediction of cryptocurrency price trend using gradient boosting, KIPS Trans, Softw. Data Eng., 7(10), 387-396.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12(1), 55-67.

Hwang, S.J. (2012). Forecasting system for water quality using artificial neural Networks: The Kangjung-Koryung weir on the Nakdong River, Doctor's Thesis, Keimyung University.

Hyndman, R.J. and Koehler, A.B. (2006). Another look at measure of forecast accuracy, Int. J. Forecast., 22(4), 679-688.

Johnson, N.E., Bonczak, B., and Kontokosta, C.E. (2018). Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment, Atmos. Environ., 184, 9-16.

Johnson, N.E., Ianiuk, O., Cazap, D., Liu, L., Starobin, D., Dobler, G., and Ghandehari, M. (2017). Patterns of waste generation: A gradient boosting model for short-term waste prediction in New York City, J. Waste Manag., 62, 3-11.

Jung, S.Y. and Kim, I.G. (2017). Analysis of water quality factor and correlation between water quality and Chl-a in middle and downstream weir section of Nakdong River, J. Korean Soc. Environ. Eng., 39(2), 89-96.

Jung, W.S., Kim, B,G., Kim, Y.D., and Kim, S.E. (2019). A study on the characteristics of cyanobacteria in the mainstream of Nakdong river using decision trees, J. Wetl. Res., 21(4), 312-320.

Kim, C.W. and Seo, Y.G. (2020). Design and performance prediction of ultra-low flow hydrocyclone using the random forest method, J. Korean Soc. Manuf. Technol. Eng., 29(2), 83-88.

Kim, D.H. and Yom, J.H. (2018). Machine Learning Based Estimation of Chlorophyll-a Concentrations in the Nakdong River Using Satellite Imagery, J. Korean Soc, Geom. atics., 4, 231-236.

Kim, G.H., Jung, K.Y., Yoon, J.S., and Cheon, S.U. (2013). Temporal and spatial analysis of water quality data observed in lower watershed of Nam River Dam, J. Korean Soc. Hazard Mitig., 13(6), 429-437.

Kim, H.G. (2017). Prediction of chlorophyll-a in the middle reach of the Nakdong River at Maegok using artificial neural networks, Department of Integrated Biological Science, Master's Thesis, The Graduate School Busan National University, Busan, Korea.

Krishna, T.H., Rajabhushanam, C., Michael, G., and Kavitha, R. (2019). Liver disorderprognosis with Apache spark random forest and gradient booster Algorithms, IJITEE, 8, 2278-3075.

Landry, M., Erlinger, T.P., Patschke, D., and Varrichio, O. (2016). Probabilistic gradient boosting machines for Gefcom 2014 wind forecasting, Int. J. Forecast, 32(3), 1061-1066.

Lawrence, R., Bunn, A., Powell, S., and Zambon, M. (2004). Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis, Remote Sens. Environ., 90(3), 331-336.

Lee, H.W. (2013). A study on nutrient mass balance of the weir sections in the middle of Nakdong River basin, Master's Thesis, Department of Environment Engineering Graduate School Yeungnam University, Gyeongsan, Gyeongbuk, Korea.

Lee, J.A. and Yoo, J.E. (2019). Exploration of predictors to teacher efficacy via elastic net, Asian J. Education, 20(1), 149-172.

Lee, S.H., Kim, B.R., and Lee, H.W. (2014). A study on water

quality after construction of the weirs in the middle area in Nakdong River, J. Korean Soc. Environ. Eng., 36(4), 258-264.

Lim, J.S., Kim, Y.W., Lee, J.H., Park, T.J., and Byun, I.G. (2015). Evaluation of Correlation between Chlorophyll-a and Multiple Parameters by Multiple Linear Regression Analysis, J. Korean Soc. Environ. Eng., 37(5), 253-261.

McLaughlin, D.B. (2012). Assessing the predictive performance of risk-based water quality criteria using decision error estimate from receiver operating characteristics(ROC) analysis, Integr. Environ. Asses., 8(4), 674-684.

Metz, C.E. (1978). Basic principles of ROC analysis, Seminars in the Nuclear Medicine, 8(4), 283-298.

Morrison, A.M., Coughlin, K., Shin, J.P., Coull, B.A., and Rex, A.C. (2003). Receiver operating characteristic curve analysis of beach water quality indicator variables, Appl. Environ. Microb., 69(11), 6405-6411.

Nieto, P.J.G., Gonzalo, E.G., Lasheras, F.S., Fernandez, J.J.R., Muniz, C.D., and Cos Jues, F.J. (2018). Cyanotoxin level prediction in a resevoir using gradient boosted regression trees: A case study, Environ. Sci. Pollut. R., 25, 22658-22671.

Müller, A.C., and Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists, O'Reilly Media, Inc.

Park, B.G. (2015). A study for estimation of chlorophyll-a in a mid-lower reach of the Nakdong River using a neural network, Master's Thesis, Department of Civil Engineering, The Graduate School Pukyong Natioal University, Busan, Korea.

Park, K.Y., and Ko. J.W. (2019). A short guide to machine learning for economists, Korean J. Econ., 26(2), 367-408.

Persson, C., Bacher, P., Shiga, T., and Madsen, H. (2017). Multi-site solar power forecasting using gradient boosted regression trees, J. Sol. Energy, 150, 423-436.

Rokach, L., and Maimon, O. (2005). Decision Trees In Data Mining and Knowledge Discovery Handbook, Springer, Boston, MA.

Song, S.S., Park, J.J., Kang, T.T., Kim, Y.S., Kim, J.Y., and Kang, T.K. (2017). Accuracy evaluation and alert level setting for real-time cyanobacteria measurement using receiver operating characteristic curve analysis, J. Korean Soc. Water Environ., 33(2), 130-139.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, Series B (Methodological), 58(1), 267–288.

Twisti, H., Edeards. A.C., and Codd, G.A. (1988). Algae growth respones to waters of contrasting tributaries of the river Dee, North-East Scotland, Water Res., 32(8), 2471-2479.

Vapnik, V. (1998). Statistical learning theory, Wiley-Interscience, New York.

Wei, L., Huang, C., Wang, Z., Wang, Z., Zhou, X., and Cao, L. (2019). Monitoring of urban black-odor water based on Nemerow index and gradient boosting decision tree regression using UAV-borne hyperspectral imagery, Remote Sens., 11(20), 2402.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.