Research Paper

A Comparative Study on the Application of Boosting Algorithm for Chl-a Estimation in the Downstream of Nakdong River

Sang-Min Lee¹ · Il-Kyu Kim^{2,†}

¹Division of Environmental System Sciences, Pukyong National University

²Department of Environmental Engineering, Pukyong National University

(Received October 19, 2020; Revised December 11, 2020; Accepted January 5, 2021)

Objectives : To estimate algae of Chlorophyll-a (Chl-a) with machine learning algorithms, water quality and quantity data of the downstream region of Nakdong River area were used.

Methods : At first, the correlation analysis was studied about Chl-a, water quality and quantity data. We have extracted ten important factors for water quality and quantity data about HC (Hapcheon Changnyeong), CH (Changnyeong Haman). Algorithms estimated how ten factors affected Chl-a occurrence each sites. We used algorithms about decision tree, random forest, elastic net, gradient boosting with Python.

Results and Discussion : The MSE (Mean of Square Error), RMSE (Root Mean Square Error), R^2 (Coefficient of determination) values were used to evaluate excellent algorithms. The gradient boosting showed 56.47 of MSE, 7.51 of RMSE, 0.78 of R^2 values for the HC site and 63.82 of MSE, 7.99 of RMSE, 0.76 of R^2 values for the CH site. Estimation value for the four algorithms was also verified through the ROC (Receiver Operation Characteristic) curve and AUC (Area Under Curve). As a result of the verification, the AUC value was 0.961 at HC site and the AUC value was 0.885 at CH site. So the gradient boosting algorithm's ability to interpret seemed to be excellent.

Conclusions : The gradient boosting algorithm showed excellent results for HC and CH sites.

Keywords : Chlorophyll-a, Machine Learning, Gradient Boosting, R^2 , ROC Curve

The Korean text of this paper can be translated into multiple languages on the website of <http://jksee.or.kr> through Google Translator.

[†] Corresponding author E-mail: kimilky523@gmail.com Tel: 051-629-6528 Fax: 051-629-6523

연구논문

낙동강 하류지역 Chl-a 추정을 위한 boosting 알고리즘 적용에 관한 비교 연구

이상민¹ · 김일규^{2,*}

¹부경대학교 지구환경시스템과학부

²부경대학교 환경공학과

목적: 조류의 Chl-a를 추정하기 위해 낙동강 하류 지역의 수질 및 수량 데이터를 활용해 기계 학습 알고리즘을 사용하였다.

방법: 처음에는 Chl-a와 수질 및 수량 데이터 간의 상관관계 분석을 실시하였으며, HC, CH의 수질 및 수량 데이터에 대해 중요도가 높은 10가지 요소를 추출하였다. 알고리즘을 통해 10가지 요인이 각 지점의 Chl-a 발생에 어떤 영향을 미치는지 추정하였다. 4가지 알고리즘인 의사 결정 트리, 랜덤 포레스트, 엘라스틱 넷, 그레디언트 부스팅 알고리즘을 Python으로 수행하였다. MSE, RMSE, R²값을 사용하여 우수한 알고리즘을 평가하였다.

결과 및 토의: 그레디언트 부스팅은 HC지점에서 MSE 56.47, RMSE 7.51, R² 0.78을 나타내었으며, CH지점에서 MSE 63.82, RMSE 7.99, R² 0.76을 나타내었다. ROC곡선 및 AUC를 활용해 4가지 알고리즘에 대한 추정값도 평가되었다. 평가 결과 HC지점의 AUC값은 0.961, CH지점의 AUC값은 0.885으로 그레디언트 부스팅 알고리즘이 두 지점에서 다른 알고리즘에 비해 비교적 우수한 성능을 보여주었다.

결론: 그레디언트 부스팅 알고리즘이 HC와 CH지점에 대해 우수한 결과를 보여주었다.

주제어: 클로로필-a, 머신러닝, 경사하강증폭법, 결정계수, 수신자판단속성곡선

1. 서론

낙동강은 유역면적이 23,727 km²이며, 중·하류 지역에는 환경오염 유발원이 산재해 있으며 특히 하류인 부산 및 경남 지역은 취수원이 밀집되어 있어 오염에 취약한 지역이다.¹⁾ 기후변화로 인해 낙동강은 수자원확보 등 여러 취지의 목표로 4대강 사업이 진행되어 왔고 현재는 8개의 보가 건설되었다. 보의 건설로 인해 낙동강 수계는 체류시간이 변하게 되었으며 폐쇄성 수역의 특성을 나타내고 있다.²⁾ 최근 들어 낙동강 하류는 매년 여름철 폭염과 가뭄으로 인해 조류가 대량 발생하는 조류경보가 발령되고 있으며, 특히 2018년 8월 22일에는 합천 창녕보의 유해성 남조류 세포수가 1,264,052 cell/mL³⁾를 기록하기도 하였다. 이 사례는 4대강 보 건설 이후 16개의 모든 보에서 찾아 볼 수 없는 가장 높은 수치를 기록하였으며 낙동강 하류는 상류와 중류의 지속적인 영양염류 유입 및 축적으로 인해 조류가 빈번히 발생하고 있다.⁴⁾

Chl-a 농도는 현장 샘플링 시 조류 증식 측정 모니터링에서 광합성색소의 하나인 조류 생물량 지표로 활용하고 있다.⁵⁾

본 연구에서는 낙동강 하류의 수질인자와 수량인자를 활용해 Chl-a 농도로 조류 발생량을 추정하고자 하였다. 앞선 연구로 낙동강의 수질 측정 자료를 활용한 통계분석인 상관관계분석을 통해 조류 증식에 영향을 미치는 수질인자에 관한 연구, 최근에는 위성 영상자료의 다중회귀분석 방법, 머신러닝을 활용한 Chl-a 농도를 추정한 연구 결과가 있다. 위성영상을 활용한 머신러닝 연구는 위성영상의 분광특성을 이용하여 농도값을 추정하는 방식으로 하천 조류발생의 모니터링 추정식은 존재한다. 하지만 모델을 사용하는 하천의 환경 변화에 따라 농도 추정식의 정확도가 떨어진다는 문제점이 있다고 알려져 있다.⁶⁾ 국내에서도 수질자료를 통한 연관된 수질인자를 찾아내어 머신러닝 알고리즘인 신경망 또는 PNN (Probabilistic Neural Network)을 활용해 낙동강 Chl-a 농도를 추정한 사례가 있으나^{7,8)} 신경망 또는 PNN 알고리즘을 통해 수질인자만 이용하여 Chl-a를 예측하였다. 조류 발생 연구는 수질, 기상, 수리 영향인자를 포함하여 복합적으로 그 영향에 대한 평가가 이루어져야 한다. 조류의 성장 인자로는 수질인자인 영양물질, 일사량, 수온, 체류시간 등이 있으며 그 외

Fig. 1. Sampling sites HC, CH in the Nakdong River.

하천의 물리적 조건인 보의 위치, 길이와 폭, 수심, 저수용량에도 영향을 받는다.⁹⁾ 이와 같이 조류성장에 영향을 주는 인자는 매우 다양하다. 이번 연구에서는 여러 머신러닝 알고리즘 중 최근 많이 활용되고 있는 앙상블 방식의 bagging과 boosting의 대표적인 알고리즘을 모두 이용하여 조류가 수시로 자주 발생하고 있는 낙동강 하류지역의 Chl-a 농도를 추정하고자 하였다. 아울러 수질측정자료 및 보 운영 자료인 수량자료를 같이 활용하여 Chl-a 농도 추정능력을 향상시키고자 하였다.

최근 국내에서 머신러닝을 활용하여 낙동강 조류 성장에 관한 연구로는 낙동강 주류 구간에서 발생하는 남조류의 특성을 파악하기 위해 조류경보제의 발령 기준을 범주형 목표변수로 하여 주요 영향인자에 대한 남조류 발생 조건을 decision tree를 통해 연구하였다.⁴⁾ 그리고 random forest를 통해 낙동강의 Chl-a 농도를 예측한 연구 사례도 있으며¹⁰⁾ 스페인의 Trasona 인공 연못에서는 gradient boosting을 활용해 다양한 수질인자를 변수로 하여 남조류 개체수와 Chl-a 농도를 비교적 정확하게 예측한 사례도 있다.¹¹⁾ 마지막으로 분석기법 중 대표적인 예측기법으로 회귀분석법이 있다. Elastic net은 ridge regression 특징과 lasso regression의 특징을 모두 가지고 있고 비교적 데이터가 큰 set에서 용이하게 작동하는 대표적인 회귀분석 알고리즘으로 알려져 있어¹²⁾ 이번 연구에서는 이러한 gradient boosting, decision tree (CART), elastic net, random forest 등 4가지 알고리즘을 비교 분석하였다.

2. 실험방법

본 연구는 조류를 발생시키는 다양한 인자 중 영양염류가 집중적으로 유입되는 낙동강 하류 지역의 2개보인 HC(합천창녕보)와 CH(창녕함안보)에 걸쳐서 보 건설 이후인 2012년 6월부터 2019년 11월까지 주기적으로 측정되고 있는 16개의 수질인자, 5개의 수량 인자와 연관된 상관관계를 도출하였다. 도출된 인자 중 머신러닝을 이용하여 상관도를 평가하였으며 머신러닝은 Python을 활용하여 낙동강 하류 2개 보에 적합한 Chl-a 농도 추정 알고리즘을 구현하고자 하였다.

2.1. 조사지점 및 시기

2012년에 건설된 낙동강 보 지점으로 조류가 수시로 자주 발생하고 있는 하류지점인 Fig. 1의 HC, CH를 대상으로 하였다. HC지점은 연장 322.5 m, 관리수위 10.5 m, 저수용량 70.0 백만 m³이며 CH는 연장 549.3 m, 관리수위 5 m, 저수용량 100.9백만 m³로 낙동강 8개보 중 하류지역을 대표하는 보이다. 수질측정 자료는 수질측정 자료를 획득할 수 있는 HC 지점의 덕곡, CH 지점의 함안지점을 대상으로 하였다. 연구 기간으로는 조류발생 측정시점인 2012년 6월부터 2019년 11월까지로 하였으며 수량자료 또한 WEIS (Water Environment Information System, <http://water.nier.go.kr>)에서 제공되는 두 보의 자료를 이용하였다.

2.2. 자료수집

수질자료는 WEIS에서 매주 1회 측정되고 있는 Chl-a, pH, DO (dissolved oxygen), BOD, COD, SS, 총질소(T-N), 총인(T-P), TOC, 수온, 전기전도도(E·C), 용존총질소(DTN), 암모니아성질소(NH₃-N), 질산성질소(NO₃-N), 용존총인(DTP), 인산염인(PO₄-P)를 이용하였으며 수량자료는 보하류수위(DWL, Downstream Water Level), 저수량(Flux), 공용량(WV, Water Volume), 유입량(Inflow), 총 방류량(TFR, Total Flow Rate) 자료를 활용하였다. 매일 측정되는 수량자료는 주간단위로 측정되는 수질측정 자료를 중심으로 21종의 변수로 데이터 set을 구성하였다. 데이터 set 구성기간으로는 연구기간인 2012년 6월부터 2019년 11월까지 384주간으로 구성하였으며 training data는 80%, testing data는 20%로 설정하였다.

2.3. 분석 알고리즘

2.3.1. Decision Tree

Decision tree는 보통은 독립변수(dependent variable)가 범주형 변수일 때 수립된 전략을 수립하여 규칙을 정하고 독립변수(dependent variable)를 분류한다. 나무구조를 아래로 도표화하여 결과를 예측하여 하나의 tree 형태를 구성한다. Tree의 각 마디를 node라고 하며 가장 위에 위치한 node를 뿌리마디(root node), branch는 중간마디(internal node)를 이어주는 선을 말한다. 중간에 위치한 마디는 중간 마디(internal node), 끝에 위치한 마디를 끝마디(terminal node)라 부르며 깊이(depth)는 가지를 형성하고 있는 마디의 개수를 말한다.¹³⁾ 이번 연구에 활용한 decision tree의 알고리즘은 Classification And Regression Trees (CART) 알고리즘으로 자료를 크기순으로 배열하여 순위를 매긴 다음 순위의 합을 통해 차이를 비교하는 순위합검정을 적용하는 비모수적 방법 및 회귀 기법으로 결과를 도식화하여 분석하기 때문에 이해하기 쉬워 널리 활용되고 있다. 전처리 과정(scaling)이 대체적으로 필요하지 않으며 비선형(nonlinear) 관계는 결과에 크게 영향을 미치지 않는 이점이 있다. 반면 비교적 복잡한 tree를 생성하는 training set은 설명이 가능하지만 검증 세트는 잘 설명하지 못하는 과적합이 발생할 수 있으며 작은 변화의 데이터가 크게 변할 수 있다. 각 단계에서 최적화를 지역적으로 실행하여 전체를 근사적으로 최적화하는 탐욕적 알고리즘(greedy algorithm)으로 전체적으로 최적의 decision tree를 보장하지는 못한다. Decision tree의 CART 알고리즘은 손실함수를 최소화 하면서 구해지며 다음 식 (1)은 CART 알고리즘의 손실함수를 보여준다.

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

m 은 표본의 수를 나타내며, m_l 는 왼쪽 마디의 표본수를 표시한다. G_i (i =left, right)는 마디 i 의 불순도(impurity)를 표시

하며 일반적으로 지니(Gini), 엔트로피(entropy), 분류오류(misallocation) 지표를 대부분 사용한다. 지니(Gini)의 경우 아래 식 (2)와 같으며 $p_{i,k}$ 는 노드 i 에서 분류집단 k 에 속하는 확률이다.

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

아래 식 (3)과 (4)는 각각 엔트로피와 분류오류를 나타낸다.¹⁴⁾

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

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

2.3.2. Random Forest

Random forest 알고리즘은 부모가 없는 뿌리 마디(root node), 리프 노드가 아닌 중간 마디(internal node), 단말 마디(terminal node or leaf) 등의 마디(node)와 에지(edge)의 집합으로 이루어져 decision tree에서 파생한 알고리즘이다.¹⁵⁾ 단일 decision tree는 한정된 시간 내에 수행하기 위해 최적의 해 대신 현실적으로 만족할 만한 수준의 해를 구하는 기법으로 한 머신러닝 방법으로 최적의 decision tree를 학습한다는 보장이 어렵다. 반면 random forest 기반의 머신러닝은 주어진 데이터를 활용해 무작위 방식으로 형성된 다수의 decision tree를 구성하여 최적의 학습 모델을 찾는 방법이다.¹⁶⁾ 이런 고유의 특성은 각 tree들이 예측한 결과가 역 상관(decorrelation)되어서 결과를 일반화 하도록 유도하며 random forest의 강인성으로 인해 노이즈가 포함된 데이터에 등 다양한 분야에서 활용되고 있다.¹⁷⁾ Random forest는 정밀을 요하는 분류와 회귀, 클러스터링 등 집단 학습을 기반으로 한 알고리즘으로 다수의 결정 tree를 임의로 학습하는 앙상블 방법이다. 여러 결정 tree를 형성하는 학습 단계는 입력 벡터가 입력되었을 때 분류를 하거나 예측하는 테스트 단계로 넘어가게 되어 검출, 분류, 회귀 등 다양하게 활용된다.¹³⁾

2.3.3. Elastic Net

Ridge regression은 추가 제약 조건으로 계수 제공의 합을 최소화하며 λ 는 기존의 RSS (residual sum of squares)의 비중을 조절하기 위한 추가 제약 조건의 hyper parameter이다. λ 가 커져 정규화 정도가 커지게 되면 가중치 값들은 작아지게 되고 반대로 λ 가 작아지면 정규화 정도가 작아져 λ 가 0이 되면 일반 선형 회귀모형이 된다. Ridge regression 식은 (5)와 같이 나타낼 수 있다.¹⁸⁾

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

Table 1. Hyperparameter about the algorithm.

Algorithm	Hyperparameter tuning
Decision tree	Criterion (mse), Max depth (5.0), Min samples leaf (1.0)
Random forest	Max depth (5.0), N estimators (10), Random state (none)
Elastic net	Alpha (1.0), L1 ratio (0.5), Random state (none)
Gradient boosting	Learning rate (0.1), Max depth (3.0), Subsample (1.0)

Lasso regression은 추가적인 조건으로 가중치 절대값 합을 최소화하기 위해 실행되며 계수의 절대값은 상수로 계수를 0으로 만드는 경향이 있으며 식은 (6)과 같이 나타낼 수 있다.¹⁹⁾

$$\beta^* = \underset{\beta}{\operatorname{argmin}} (\|y - X\beta\|^2 + \lambda \|\beta\|_1) \quad (6)$$

Elastic net은 L_1 과 L_2 규제를 적용한 선형 조합으로 ridge regression과 lasso regression 정규화가 결합된 선형 알고리즘이다. 식은 (7)과 같이 나타낼 수 있으며 가중치 절대값 합과 제곱합을 동시에 제약 조건으로 가지는 모형이다.¹²⁾

$$\beta^* = \underset{\beta}{\operatorname{argmin}} (\|y - X\beta\|^2 + \lambda^2 \|\beta\|^2 + \lambda_1 \|\beta\|_1), \quad a = \frac{\lambda_2}{(\lambda_2 + \lambda_1)}$$

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

Elastic net은 RSS을 최소로 만드는 회귀계수를 최소화하는 가장 단순한 회귀모형 예측법이다. 하지만 최소자승법으로 여러 설명변수를 모형화할 경우 다중공선성이 발생하는 문제가 있을 수 있으며 예측변수가 많아질수록 예측 능력은 증가할 수 있으나 새로운 자료에 대한 과적합 문제가 발생 가능한 단점이 있다. 그러나 elastic net은 다중공선성 및 과적합을 다루는데 적합하고 변수가 다양한 대용량 자료를 분석할 때 효율적으로 적용 가능하다.²⁰⁾

2.3.4. Gradient boosting

Boosting 알고리즘은 좀 더 정확한 strong learner을 만들기 위해 부정확한 weak learner을 혼합하는 알고리즘이다. 첫 번째로 정확도가 다소 낮은 tree를 형성한 이후, 향후 발견된 오류는 두 번째 tree에서 보완된다. 이런 방식으로 약점을 보완하여 다음 tree를 만들게 되고 나중에는 strong learner를 구축하게 된다. Loss Function (손실함수, J)을 통해 예측 알고리즘의 오류를 수치화 및 정량화하며, 알고리즘은 gradient descent를 사용하여 loss Function 값을 최소화하고 알고리즘 내에서 parameter를 찾는다. Gradient boosting은 이런 parameter loss Function (손실함수, J)을 최소화 하는 과정을 모델 함수(f_i) 공간에서 실행한다, loss Function (손실함수, J)를 알고리즘 parameter가 아니라 아래와 같은 (8)번 수식에 의해 학습된

tree 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 함수의 미분값은 그전까지 학습된 알고리즘의 약점을 나타내는 역할을 한다. 그 다음 tree 알고리즘의 피팅을 수행할 때 해당 미분값을 사용하여 약점을 보완하는 방식으로 성능을 향상시킨다.²¹⁾ Gradient boosting은 decision tree sequence를 여러번 학습시키며 학습된 알고리즘은 트레이닝 인스턴스 레이블과 실제 레이블을 예측하고 비교한다. 부정확한 예측으로 인해 훈련 인스턴스에 더 중점을 두기 위해서는 데이터 set의 레이블이 다시 지정된다. Gradient boosting은 반복적으로 decision tree에 의한 실수를 정정하는 것으로 알려져 있으며 최신 버전은 정규화를 통해 병렬처리, tree 정리, 결측값을 처리하는 최적화된 gradient boosting으로 과적합 및 바이어스를 방지하는 장점이 있다.²²⁾

학습과정에서 과적합이 발생하지 않고 좋은 성능을 나타내기 위하여 이번 연구에서 앞서 살펴본 4가지 알고리즘의 하이퍼파라미터 튜닝을 Table 1과 같이 수행하였다.

2.3.5. RMSE (Root Mean Square Error), R^2 (Coefficient of determination)을 통한 알고리즘 성능 평가

RMSE는 추정값과 실제값과 차이를 나타내는 척도로 알고리즘의 추정능력과 정확도를 평가할 수 있다. RMSE는 잔차(계산값과 이론값의 차)의 제곱합을 평균한 값의 제곱근으로 실제값들의 상호간 편차를 의미한다. 표준편차를 일반화시킨 척도로 실제값과 추정값과의 편차를 알려 주는데 많이 활용되고 있다. RMSE와 표준편차는 추정값과 실제 환경에서 관찰되는 값의 차이를 나타낸다. 실제값에 대한 추정값의 RMSE 값은 다음식 (9)과 같이 구할 수 있다.²³⁾

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

RMSE값은 작아질수록 추정값과 실제값의 차이가 없다는 것을 의미하며 0에 가까워질수록 좋은 성능의 알고리즘으로 판단할 수 있다. R^2 은 실제값과 추정값 사이에서 알고리즘의 적합한 정도를 측정하여 그 값이 1이 된다고 하면 실제값과 추정값의 차이는 없다는 의미이다. 실제값을 t_i , 추정값을 y_i 그리고 관찰된 샘플 n 의 평균 $\bar{t}$ 는 식 (10)과 같이 나타낼 수

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

있다.

$$\bar{t} = \frac{1}{n} \sum_{i=1}^n t_i \quad (10)$$

이 식에 의해서 R^2 은 식 (11)과 같이 구할 수 있다.¹¹⁾

$$R^2 = 1 - \frac{\sum_{i=1}^n (t_i - y_i)^2}{\sum_{i=1}^n (t_i - \bar{t})^2} \quad (11)$$

최근 환경분야 및 다양한 분야에서 훈련데이터로 훈련한 알고리즘의 성능 평가를 위해 RMSE, MSE, R^2 등이 많이 사용되고 있다.²⁴⁻²⁷⁾

2.3.6. ROC (Receiver Operation Characteristic) Curve를 활용한 추정값 정확성 평가

ROC분석은 의학계에서 진단 여부를 판단하고자 하는 경우 많이 사용하고 있으며 최근에는 질환의 유무를 예측하고 진단에 활용하고자 하는 테스트의 효율성을 평가하기 위해 활용되고 있다.²⁸⁾ 최근에는 ROC가 환경 분야에서 수질 예측지표 및 위해도 평가 기준을 설정하고자 연구에 적용되고 있다.^{29,30)} ROC는 알고리즘의 추정능력을 판단하는데 활용되고 있으며 ROC 아래면적인 AUC (Area Under Curve)는 값이 0 ~ 1까지의 값을 가지며 1의 값에 가까울수록 좋은 성능의 알고리즘으로 평가하고 있다.³¹⁾ AUC는 보통 5등급의 평가 기준으로 분류하여 **Table 2**와 같이 분류할 수 있다.

ROC는 TPR (True Positive Rate)과 위양성율 FPR (False Positive Rate)을 통해 시각화 할 수 있다. TPR은 추정의 정확성에 대한 값을 나타내고 FPR은 추정의 부정확성에 대한 값을 나타내어 이러한 값들로 ROC를 시각화 할 수 있다.³²⁾ 낙동강 하류의 조류관리단계를 판별하기 위하여 HC, CH 두 지점의 Chl-a 추정값이 조류발생 지표로 효과적인지 확인할 필요가 있다. 추정값이 평균이상으로 나타날 경우 민감도(sensitivity)와 특이도(specificity)가 높게 나타나 효과적인 지표로 볼 수 있다. 민감도는 실제 Chl-a가 두 지점의 조류발생 측정기간 동안 평균이상으로 HC지점은 21.99, CH지점은 23.85 이상 발생하는 Chl-a 평균이상의 부영양화된 수질로 분류되는 TPR을 말하며, 특이도는 Chl-a가 Chl-a 평균미만의 양호한 수질로

분류되는 FPR로 나타났다. TPR은 다시 민감도로 나타내어지고 TP (True Positive), FN (False Negative)으로 식 (12)와 같이 계산된다. 반면 FPR은 1-민감도로 표현되고 TN, FP로 식 (13)로 계산된다.

$$\text{양성율 (TPR)} = \frac{TP}{FN + TP} \quad (12)$$

$$\text{위양성율 (FPR)} = \frac{TN}{TN + FP} \quad (13)$$

3. 결과 및 고찰

3.1. 수질 및 수량항목 상관관계분석 결과

조류는 수질 및 수량인자 등 다양한 영향인자에 의해 발생하며 이는 또 서로 영향을 미치며 조류가 발생하므로 조류 성장지표인 Chl-a와 영향인자에 대한 상관관계를 해석할 필요가 있다. Chl-a에 영향을 미치는 두 보의 주요 수질, 수량인자와 Chl-a와의 상관관계 분석 결과는 **Fig. 2**와 같다.

상관도가 떨어지는 요인을 제거하기 위해 중요인자를 추출하였으며 변수 선택 결과 10개의 변수가 선택되고 선택된 변수의 상관계수는 절대값이 대부분 0.1 이상일 때 머신러닝 알고리즘이 정확하게 구현되었다. Chl-a와 상관관계를 Pearson 상관계수를 통해 분석한 결과 절대값이 높은 기준으로 HC지점은 BOD (0.6), pH (0.5), DO (0.4), PO₄-P (-0.4), COD (0.4), DTP (-0.3), SS (0.2), NH₃-N (-0.2), 수온(-0.1), T-P (-0.003) 순으로 나타났으며, CH지점은 BOD (0.5), COD (0.5), pH (0.5), TOC (0.4), PO₄-P (-0.3), NH₃-N (-0.3), DO (0.2), DTP (-0.2), SS (0.1), E·C (-0.1) 순으로 나타났다. 특히 BOD, COD, pH, PO₄-P는 상관계수 절대값이 0.3 이상으로 유의한 상관성을 나타내었다. 대표적인 유기물 수질 지표인 BOD, COD, TOC는 호기성 미생물의 합성을 위한 산소량과 수중 유기물의 양, 유기물 산화에 필요한 산소의 양 등으로 Chl-a와 높은 상관성을 보이고 있으며 조류성장에 따른 자생 유기물질과 관련이 있는 것으로 추정해 볼 수 있다.³³⁾ 수온은 기온과 밀접한 관계를 가진 수질인자로 조류는 이산화탄소를 탄소원으로 이용하여 광합성을 할 때 수온은 이산화탄소를 분리하는데 중요한 역할을 하며 조류의 성장 수온은 25~35℃가 일반적이며 최적 성장 수온은 약 30℃로 알려져 있다.³⁴⁾ DO는 두 지점 모두 상관계수 0.2 이상으로 적지만 유의한 상관성을 나타내었다. 자연수계에서 가장 중요한 성분 중 하나로 주로 수온 및 기온의 변화에 따라 변화하며 일반적으로는 조류의 광합성 작용 등에 의해 증가한다고 보고되어 있으며 수심별 Chl-a와 유사한 경향을 보인다.³⁵⁾ 이외의 영양염류인 NH₃-N, PO₄-P, DTP에서도 절대값 0.2 이상의 유의한 상관성을 나타내었다. NH₃-N은 질소의 형태에 따라 조류 성장에 영향을 주는 요인으로 볼 수 있으며 조류의 성장에 필요한

Fig. 2. Scatter plot between target output and input variables about HC, CH.

질소원으로 우선 $\text{NH}_3\text{-N}$ 을 이용하고 $\text{NH}_3\text{-N}$ 이 소모되었을 때 $\text{NO}_3\text{-N}$ 을 이용하기 때문이다.³⁶⁾ $\text{PO}_4\text{-P}$ 는 두 지점에서 뚜렷하게 음의 상관성이 증가하였다. $\text{PO}_4\text{-P}$ 는 조류 성장 과정에서 우선적으로 소비되므로 조류가 많이 발생하는 시기에는 농도가 낮게 되어 그 결과 음의 상관성을 보이는 것으로 볼 수 있다.³⁷⁾ 그리고 두 지점 모두 유의한 상관관계를 나타낸 pH는 식물 플랑크톤이 증식하면서 광합성과 밀접한 관계가 있고 수중의 CO_3^{2-} 및 HCO_3^- 가 흡수되어 pH가 증가하게 되므로 pH는 조류성장의 간접지표로도 활용이 가능하다. 이번 연구에서 E·C는 수계를 구분하거나 연속적인 수질 변화를 비교적 용이하게 파악하는 지표로 특히 조류가 성장하였을 때 수계 상층에서 높게 나타난다. 수계 내 SS농도는 집중호우 등 농경지, 도로 등과 같은 외부 비점오염원의 영향을 받는 동시에 일부분은 조류와 같은 미생물 농도인 내부원의 영향을 받는다.

이전 낙동강 Chl-a 연구사례로 낙동강 하류지역에서 수질, 수리 및 기상인자와 Chl-a 사이의 상관관계분석 결과가 있었으며 Chl-a에 대한 영향인자 중 수질인자는 수온, pH, DO, BOD, COD, T-N, $\text{NO}_3\text{-N}$, $\text{PO}_4\text{-P}$ 8가지 요인이 유의한 상관관계를 나타내었다.³⁸⁾ 이번 연구 결과와 유사한 지역인 낙동강 하류 합천창녕 지점에서 2012년 1월부터 2016년 10월까지

의 14개 수질인자 및 Chl-a와의 상관관계분석 결과 pH, DO, BOD, COD, SS, T-N, T-P, 수온, $\text{NH}_3\text{-N}$, $\text{NO}_3\text{-N}$, 등이 유의한 상관관계($p < 0.01$)를 나타내고 창녕합안보는 pH, DO, BOD, COD, T-N, T-P, TOC, $\text{NH}_3\text{-N}$, $\text{NO}_3\text{-N}$, $\text{PO}_4\text{-P}$ 등이 유의한 상관관계($p < 0.01$)를 나타낸 연구 결과가 있었다.³⁹⁾ 이는 이번 연구에서 두 지점인 HC지점은 pH, DO, BOD, COD, SS, T-P, 수온, $\text{NH}_3\text{-N}$ 등 8개의 인자가 일치하였고 CH지점은 pH, DO, BOD, COD, TOC, $\text{NH}_3\text{-N}$, $\text{PO}_4\text{-P}$ 등 7개 인자가 일치하였다. 지점별로 상관관계를 나타내는 인자가 다른 이유는 조류 발생과 직접적인 연관이 있는 Chl-a 농도는 보편 수리환경적 인자와 수질 특성 그리고 2015년 여름의 경우 대구, 의령, 합천 지점의 기상청 강수량 자료는 135.8 mm로 다른 해에 비해 낮은 강수량을 나타내는 등 이와 같은 기상인자들이 복합적으로 영향을 미치는 지역 특성에 기인한 것으로 보인다.³⁹⁾

3.2. 알고리즘 성능 지표를 활용한 알고리즘 성능 평가 결과

본 연구에서 4가지 알고리즘의 성능을 평가하기 위해 MSE, RMSE, R^2 을 활용하였으며 알고리즘 성능 지표 비교는 Table 3과 같다.

HC지점에서의 MSE는 decision tree가 171.74, random forest는 61.07, elastic net 75.57, gradient boosting 56.47로 나타났으

Table 3. Algorithms Performance Comparison with MSE, RMSE, R^2 about HC, CH.

	HC				CH			
	Decision tree	Random forest	Elastic net	Gradient boosting	Decision tree	Random forest	Elastic net	Gradient boosting
MSE	171.74	61.07	75.57	56.47	108.21	71.52	111.01	63.82
RMSE	13.11	7.81	8.69	7.51	10.40	8.46	10.54	7.99
R^2	0.50	0.76	0.72	0.78	0.65	0.74	0.61	0.76

Fig. 3. Comparisons of R^2 results using four estimative algorithms about HC.

Fig. 4. Comparisons of R^2 results using four estimative algorithms about CH.

며 CH지점에서의 MSE가 decision tree 108.21, random forest 71.52, elastic net 111.01, gradient boosting 63.82로 나타났다.

그리고 HC지점에서의 RMSE는 decision tree가 13.11, random forest는 7.81, elastic net 8.69, gradient boosting 7.51로 나타났다

Fig. 5. Measured and estimated values of decision tree models about HC, CH.

Fig. 6. Measured and estimated values of random forest about HC, CH.

으며 CH지점에서는 decision tree 10.40, random forest 8.46, elastic net 10.54, gradient boosting 7.99로 나타났다. HC지점의 R^2 은 decision tree가 0.50, random forest는 0.76, elastic net 0.72, gradient boosting 0.78로 나타났으며 CH지점에서는 decision tree 0.65, random forest 0.74, elastic net 0.61, gradient boosting 0.76로 나타났다. 두 지점의 R^2 값을 비교한 결과는 Fig. 3, 4에서 볼 수 있다.

HC지점의 gradient boosting의 MSE가 56.47, RMSE는 7.51, R^2 은 0.78로 알고리즘의 성능이 높게 나타났으며 다음 성능을 나타낸 알고리즘은 random forest, elastic net, decision tree순으로 나타났다. CH지점은 gradient boosting의 MSE가 63.82, RMSE는 7.99, R^2 은 0.76으로 알고리즘 성능이 높게 나타났으며 다음 성능을 나타낸 알고리즘은 random forest, decision tree, elastic net으로 나타났다. R^2 의 범위는 0~1 사이의 값을 가지며 1에 가까울수록 종속변수와 독립변수와의 상관성이 크다는 것을 나타낸다. 이번 연구에서 HC, CH 두 지점 모두 gradient boosting이 종속변수가 독립변수 변동에 따라 비교적 정확하게 추정할 수 있다고 설명할 수 있으나 추가적인 검증 을 위해서 분류기 특성 분석 평가지표인 ROC 커브를 이용하여 재검증하였다. 실제값과 추정값을 비교한 결과는 Fig. 5

~8에서 볼 수 있다.

Gradient boosting 알고리즘은 풍력발전을 위한 확률적 바람 예측방법⁴⁰⁾, 태양광 발전을 위한 다중사이트 예측⁴¹⁾, 뉴욕지역에서의 짧은 기간 동안의 폐기물 발생 예측⁴²⁾과 같이 환경 분야에 많이 적용되어 환경과 연관된 문제를 해결하기 위해 활용되고 있다. 뿐만 아니라 조류 발생의 생물학적 매개변수를 효과적으로 추정하는 알고리즘으로 이용되고 있으며¹¹⁾ 혐기성 조건인 유기성폐기물을 효율적으로 처리하기 위한 적절한 소화시설 설치⁴³⁾와 효율을 극대화하기 위한 저비용으로 도심의 에어로졸 모니터링²⁵⁾ 등 여러 환경 분야인 수질, 대기, 폐기물 분야에서 다양하고 활발하게 적용되고 있다. 특히 gradient boosting 알고리즘은 regression 분석과 classification 분석을 동시에 수행할 수 있다. 훈련데이터로부터 하나의 함수를 추정하기 위한 머신러닝의 한 방법인 지도학습의 한 종류로 변수가 많은 함수에 대해 활용도가 높은 알고리즘이다.⁴⁴⁻⁴⁶⁾ 최근 스페인의 Trasona 저수지를 대상으로 하여 gradient boosting 알고리즘을 활용하여 남조류 개체수와 Chl-a 농도 예측 연구에 활용된 연구결과가 있어¹¹⁾ 환경 분야에서 다양한 목표변수 예측연구를 위한 머신러닝 알고리즘으로 적합한 것으로 판단된다.

Fig. 7. Measured and estimated values of elastic net about HC, CH.

Fig. 8. Measured and estimated values of gradient boosting about HC, CH.

3.3. ROC Curve를 활용한 추정값 정확성 평가 결과

두 지점의 민감도와 특이도에 대한 그래프를 Fig. 9, 10에서 확인할 수 있다. 그리고 HC, CH지점의 4가지 알고리즘에 대해 ROC를 시각화하고 AUC를 산출하였다. 산출결과 HC지점에서 elastic net의 AUC는 0.886 decision tree의 AUC는 0.905, random forest의 AUC는 0.955, gradient boosting의 AUC가 0.961로 나타났다. CH지점에서 elastic net의 AUC는 0.811, decision tree의 AUC는 0.848, random forest의 AUC는 0.881, gradient boosting의 AUC가 0.885로 나타났다. Gradient boosting은 AUC값이 HC지점은 0.961, CH지점은 0.885로 두 지점에서 gradient boosting 알고리즘의 추정값 평가가 우수하게 나타났다. Gradient boosting의 AUC값은 HC지점의 Chl-a 농도 21.99 초과여부에 대하여 0.961 (95% 신뢰구간, 0.923~0.999), CH지점의 경우 Chl-a 농도 23.85 초과여부에 대하여 0.885 (95% 신뢰구간, 0.812~0.958)로 나타났다. AUC값이 1에 근접할 경우 알고리즘 추정값의 평가가 우수하며 두 지점 모두 AUC값이 다른 알고리즘에 비해 1에 근접하여 gradient boosting 알고리즘이 우수한 것으로 판단된다.

Fig. 9. Receiver operating characteristic curve about HC.

4. 결론

본 연구에서 대표적인 낙동강 하류 보인 HC, CH 두 지점의 수질과 수량항목에 대해 Chl-a와의 상관관계 분석을 통해 중요한 인자를 추출하였다. 그리고 Chl-a를 종속변수로 하여 추출된 10가지 인자를 활용하여 4가지 알고리즘을 비교 분석하였으며 결과는 다음과 같다.

1) HC, CH 두 지점 모두 BOD, SS, pH, DO, COD 등이

Fig. 10. Receiver operating characteristic curve about CH.

공통적으로 중요인자로 평가되었다. 알고리즘의 추정능력을 평가하기 위해 MSE, RMSE, R^2 을 통해 평가한 결과 낙동강 하류의 두 지점에서 gradient boosting 알고리즘이 우수하였다. 그리고 두 지점의 Chl-a 농도 실제값에 대한 추정값의 test score는 random forest의 HC지점이 76%, CH지점이 71%, gradient boosting의 HC지점은 78%, CH지점이 74%로 나타났다.

2) 낙동강 하류의 경우 알고리즘 성능은 두 보간 편차가 거의 발생하지 않았다. 그 원인으로 중요인자인 변수 간 중요도 차이가 크지 않았으며 추출된 10가지 인자 중 두 보의 공통적인 여러 중요인자가 영향을 미친 것으로 판단된다. 공통의 상위 5가지 인자는 향후 Chl-a 농도 추정에 중요한 자료로 활용될 것으로 판단된다.

3) 추정값의 정확성을 평가하기 위해 ROC 분석을 두 지점에 실시하였다. 4가지 알고리즘 중 두 지점에서 gradient boosting 알고리즘의 AUC값이 0.8 이상으로 추정값의 정확성이 높게 나타났다. 알고리즘별로 일부 성능차이는 있었으며 gradient boosting 알고리즘은 이번 연구와 같이 수질, 수량 데이터를 기반으로 하여 조류가 수시로 자주 발생하고 있는 낙동강 하류지역의 Chl-a 농도 추정에 적용 가능한 알고리즘으로 평가된다.

4) 이번 연구를 통해 향후 조류가 빈번히 발생하고 있고 오염물질이 지속적으로 유입되어 조류가 수시로 자주 발생하는 낙동강 전 지역으로 활용도 가능하며 다른 지역으로도 알고리즘이 적용 가능하다는 많은 공공데이터가 축적되어 있는 4대강 보 중심으로 연구가 가능하다고 판단된다. 그리고 좀 더 정확한 추정을 위해서는 기상자료 등 다양한 공공데이터를 활용한다면 이번 알고리즘의 수질 추정 분석 성능은 더욱 높아질 것이다.

Acknowledgement

본 연구는 2019년도 부경대학교 연구년 교수 지원사업에 의하여 연구되었습니다.

References

1. J. J. Yu, K. L. Lee, H. J. Lee, J. W. Hwang, H. S. Lyu, L. Y. Shin, A. R. Park, S. U. Chen, Relations of nutrient concentrations on the seasonality of algal community in the Nakdong River, Korea, *J. Korean Soc. Water Environ.*, 31(2), 110-119(2015).
2. S. H. Lee, B. R. Kim, H. W. Lee, 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(2014).
3. Ministry of Environment, Water Environment Information System (WEIS), <http://water.nier.go.kr/algae/algaePreMeasure.do>(2012).
4. W. S. Jung, B. G. Jo, Y. D. Kim, S. E. Kim, A study on the characteristics of cyanobacteria in the mainstream of Nakdong river using decision trees, *J. Wet. Res.*, 21(4), 312-320(2019).
5. J. Gregor, B. Marsalek, Freshwater phytoplankton quantification by chlorophyll-a: a comparative study of in vitro, in vivo and in situ methods, *Water Res.*, 38(3), 517-522(2004).
6. D. J. Chun, J. Eun, Application method of remote sensing method for monitoring the water quality of big river, *KEI Environmental Forum*, 214, 21(2017).
7. B. G. Park, A study for estimation of chlorophyll-a in a mid-lower reach of the Nakdong River using a neural network, Master's Thesis, Pukyong National University, Busan, Korea, 29(2015).
8. H. G. Kim, Prediction of chlorophyll-a in the middle reach of the Nakdong River at Maegok using artificial neural networks, Master's Thesis, Busan National University, Busan, Korea, 6-13(2017).
9. D. Caissie, M. G. Satish, N. El-Jabi, Predicting water temperatures using a deterministic model: application on Miramichi River catchment (New Brunswick, Canada), *J. Hydrol.*, 336(3-4), 303-315(2007).
10. D. H. Kim, J. H. Yom, Machine learning based estimation of Chlorophyll-a concentrations in the Nakdong River using satellite imagery, in proceedings of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, KSGPC, Yongin, Korea, pp. 231-236(2018).
11. P. J. G. Nieto, E. García-Gonzalo, F. S. Lasheras, J. R. A. Fernandez, C. D. Muniz, F. J. Cos Jues, Cyanotoxin level prediction in a reservoir using gradient boosted regression trees: a case study, *Environ. Sci. Pollut. Res.*, 25, 22658-22671(2018).
12. H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. R. Statist. Soc. B*, 67(2), 301-320(2005).
13. J. Y. Cho, Odor compounds forecasting in Daecheong water intake station using machine learning models, Doctor's Thesis, Chungnam National University, Daejeon, Korea(2019).
14. K. Y. Park, J. W. Ko, A short guide to machine learning for economists, *The Korean Journal of Economics*, 26(2), 367-408 (2019).
15. L. Rokach, O. Maimon, Decision trees, *Data Mining and Knowledge Discovery Handbook*, Springer, Boston, MA,

- USA, pp. 165-192(2005).
16. A. C. Müller, S. Guido, Introduction to Machine Learning with Python, O'Reilly Media, Inc., Sebastopol, CA, USA(2016).
17. C. W. Kim, Y. G. Seo, Design and performance prediction of ultra-low flow hydrocyclone using the random forest method, J. Korean Soc. Manuf. Technol. Eng., 29(2), 83-88(2020).
18. A. E. Hoerl, R. W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics, 12(1), 55-67(1970).
19. R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Statist. Soc. B, 58(1), 267-288(1996).
20. J. A. Lee, J. E. Yoo, Exploration of predictors to teacher efficacy via elastic net, Asian J. Education, 20(1), 149-172 (2019).
21. J. S. Heo, D. H. Kwon, J. B. Kim, Y. H. Han, C. H. An, Prediction of cryptocurrency price trend using gradient boosting, KIPS Trans. Softw. and Data Eng., 7(10), 387-396 (2018).
22. T. H. Krishna, C. Rajabhushanam, G. Michael, R. Kavitha, Liver disorder prognosis with Apache spark random forest and gradient booster algorithms, IJITEE, 8(953), 615-620(2019).
23. R. J. Hyndman, A. B. Koehler, Another look at measure of forecast accuracy, Int. J. Forecast., 22(4), 679-688(2006).
24. R. Lawrence, A. Bunn, S. Powell, M. Zambon, Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis, Remote Sens. Environ., 90(3), 331-336(2004).
25. N. E. Johnson, B. Bonczak, C. E. Kontokosta, 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(2018).
26. J. Fan, X. Ma, L. Wu, F. Zang, X. Yu, W. Zeng, Light gradient boosting machine: an efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data, Agric. Water Manag., 225, 105758(2019).
27. L. Wei, C. Huang, Z. Wang, Z. Wang, X. Zhou, L. Cao, 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(2019).
28. C. E. Metz, Basic principles of ROC analysis, Semin. Nucl. Med., 8(4), 283-298(1978).
29. A. M. Morrison, K. Coughlin, J. P. Shin, B. A. Coull, A. C. Rex, Receiver operating characteristic curve analysis of beach water quality indicator variables, Appl. Environ. Microbiol., 69(11), 6405-6411(2003).
30. D. B. McLaughlin, Assessing the predictive performance of risk-based water quality criteria using decision error estimate from receiver operating characteristics (ROC) analysis, Integr. Environ. Assess. Manage., 8(4), 674-684(2012).
31. D. T. Do, N. Q. K. Le, Using extreme gradient boosting to identify origin of replication in *Saccharomyces cerevisiae* via hybrid features, Genomics, 112(3), 2445-2451(2020).
32. S. S. Dhaliwal, A. A. Nahid, R. Abbas, Effective intrusion detection system using XGboost, Information, 9(7), 149(2018).
33. G. H. Kim, K. Y. Jung, J. S. Yoon, S. U. Cheon, 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(2013).
34. X. Zeng, M. K. Danquah, X. D. Chen, Y. Lu, Microalgae bioengineering: from CO₂ fixation to biofuel production, Renewable Sustainable Energy Rev., 15(6), 3252-3260(2011).
35. S. Noh, Y. Shin, H. Choi, J. Lee, J. Lee, D. Rhew, Characteristics of algae occurrence on environmental changes, J. Environ. Impact Assess., 24(3), 278-286(2015).
36. B. H. Zheng, C. J. Cao, Y. W. Qin, M. S. Huang, Analysis of nitrogen distribution characters and their sources of the major input rivers of three gorges reservoir, Huan Jing Ke Xue., 29(1), 1-6(2008).
37. F. I. Kappers, The cyanobacterium *Microcystis aeruginosa* Kg. and the nitrogen cycle of the hypertrophic lake Brielle (The Netherlands), Developments in Hydrobiology, J. Barica, L. R. Mur (Eds.), Springer, Dordrecht, Netherlands, pp. 37-43(1980).
38. J. S. Lim, Y. W. Kim, J. H. Lee, T. J. Park, I. G. Byun, Evaluation of correlation between chlorophyll-a and multiple parameters by multiple linear regression analysis, J. Korean Soc. Environ. Eng., 37(5), 253-261(2015).
39. S. Y. Jung, I. K. Kim, 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(2017).
40. M. Landry, T. P. Erlinger, D. Patschke, C. Varrichio, Probabilistic gradient boosting machines for GEFCom2014 wind forecasting, Int. J. Forecast., 32(3), 1061-1066(2016).
41. C. Persson, P. Bacher, T. Shiga, H. Madsen, Multi-site solar power forecasting using gradient boosted regression trees, Sol. Energy, 150, 423-436(2017).
42. N. E. Johnson, O. Ianiuk, D. Cazap, L. Liu, D. Starobin, G. Dobler, M. Ghandehari, Patterns of waste generation: a gradient boosting model for short-term waste prediction in New York City, Waste Manage., 62, 3-11(2017).
43. D. D. Clercq, Z. Wen, F. Fei, Determinants of efficiency in anaerobic bio-waste co-digestion facilities: a data envelopment analysis and gradient boosting approach, Appl. Energy, 253, 113570(2019).
44. V. N. Vapnik, Statistical Learning Theory, Wiley-Interscience, New York, USA, pp. 1-768(1998).
45. J. H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal., 38(4), 367-378(2002).
46. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, New York, USA, pp. 1-745 (2009).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors and Contribution Statement

Sang-Min Lee

Division of Environmental System Sciences, Pukyong National University, Ph.D. Candidate, ORCID® 0000-0002-1738-9759: Conceptualization, Data analysis, Visualization, Writing-original draft.

Il-Kyu Kim

Department of Environmental Engineering, Pukyong National University, Professor, ORCID® 0000-0002-9859-2475: Methodology, Supervision, Validation, Writing-review editing.