지질학회지 제 58권 제 2호, p. 243-256, (2022년 6월)
J. Geol. Soc. Korea, v. 58, no. 2, p. 243-256, (June 2022)
DOI https://doi.org/10.14770/jgsk.2022.58.2.243

ISSN 0435-4036 (Print)
ISSN 2288-7377 (Online)

<Technical Report>
XGBoost Regression 기계학습을 이용한 제주도 지하수 염소이온 농도예측 연구 사례

심대천 · 이진용‡ · 장지욱 · 이민욱
강원대학교 지질학과

Prediction of chloride concentration in groundwater on Jeju Island using XGBoost regression machine learning

Daecheon Sim · Jin-Yong Lee‡ · Jiwook Jang · Minwook Lee
Department of Geology, Kangwon National University, Chuncheon 24341, Republic of Korea

요 약

기계학습은 현재 다양한 분야에서 우수한 성능과 결과를 보여주는 인공지능 기술이다. 물 연구에서 2017년을 기점으로 기계학습을 적용한 연구사례가 증가하고 있고 XGBoost, Support Vector Machine, Random Forest, Artificial Neural Network와 같은 알고리즘들이 주로 사용된다. 제주도는 수자원의 대부분을 지하수에 의존하고 있으며 자연적인 인자와 인위적인 인자로 인하여 오염부하가 증가하면서 지하수의 수질 악화 문제가 크게 대두되고 있다. 이에 본 연구에서는 제주도 지하수의 주요 오염물질 중 하나인 염소이온(Cl^-)의 농도예측을 위하여 Gradient boosting 알고리즘을 기반으로 한 XGBoost를 이용하여 11개의 지하수질 항목을 입력 인자로 염소이온(Cl^-)을 예측하고자 하였다. 이를 위해 GridsearchCV를 이용하여 모델을 세부조정하였으며 회귀모델 평가 지표인 평균절대오차(Mean Absolute Error, MAE), 평균제곱오차(Mean Squared Error, MSE), 결정계수(R^2), 평균제곱근오차(Root Mean Squared Error, RMSE), 평균절대백분율오차(Mean Absolute Percentage Error, MAPE)를 이용하여 모델을 평가하였다. 본 논문은 기계학습 방법의 하나인 XGBoost를 이용한 지하수 염소이온 농도의 예측 사례를 제공한다.

주요어: XGBoost, 기계학습, 지하수, 염소이온, 제주도

ABSTRACT: Machine learning is an artificial intelligence technology that is currently showing excellent performance and outcome in various fields. In water research, the number of research cases applying machine learning has increased since 2017, and algorithms, such as XGBoost, Support Vector Machine, Random Forest, and Artificial Neural Network are mainly used. Jeju Island have depended on groundwater for most of its water resources, and the problem of the deterioration of groundwater quality has come to the fore. As the pollution load increases due to natural and anthropogenic factors. Therefore, in this study, 11 groundwater quality items were used as input factors to predict the concentration of chloride ion (Cl^-), one of the major pollutants in groundwater on Jeju-do, using XGBoost based on gradient boosting algorithm was intended to predict. For this, the model was fine-tuned by using GridsearchCV, and the model was evaluated by Mean Absolute Error (MAE), Mean Squared Error (MSE), R^2, Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), which are regression model evaluation indicators. This paper provides an example of prediction of groundwater chloride ion concentration using XGBoost, one of the machine learning methods.

Key words: XGBoost, machine learning, groundwater, chloride ion, Jeju Island

‡Corresponding author: +82-33-250-8551, E-mail: hydrolee@kangwon.ac.kr

1. 서 론

기계학습(machine learning)은 인공지능(artificial intelligence)의 일종으로 컴퓨터에 명시적인 프로그램 없이 배울 수 있는 능력을 제공하는 분야로 정의되며 인간이 학습하듯 컴퓨터에도 정보들을 제공하여 스스로 경험하고 학습하게 함으로써 새로운 지식을 창조하는 것을 말한다(Mitchell, 1996; Jordan and Mitchell, 2015). 기계학습의 알고리즘의 유형은 크게, 문제와 답을 갖는 지도학습, 문제만 존재하는 비지도학습, 문제와 답을 찾아야하는 강화학습으로 나눌 수 있다(Bi *et al*., 2019).

한편 홍수, 물 사용량, 수질, 수처리 공정 예측과 같은 물 연구에 대한 기계학습의 이용은 2021년 11월 20일 기준 176건으로 2017년 이후 급격히 증가하는 추세를 보인다(O *et al*., 2022). Cheng *et al*. (2018)는 어류의 3차원 좌표 정보를 바탕으로 어류의 이동을 계산할 수 있게 설계한 Tracking-Learning-Detection (TLD)와 알려지지 않은 어류 이동 매개변수를 분석 및 평가하는 XGBoost를 이용하여 수질 관측 방법을 제시하였다. Singha *et al*. (2021)는 Random Forest (RF), Extreme Gradient Boosting (XGBoost), Artificial Neural Network (ANN), Deep Learning (DL)을 이용하여 Entropy-weighted Water Quality Index (EWQI)를 예측하였으며 XGBoost는 Deep Learning 다음으로 좋은 성능을 나타냈다. Lu and Ma (2020)는 단기간의 수질 자료를 가지고 보다 정확한 예측 결과를 얻기 위하여 XGBoost와 RF에 Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)을 결합하여 모델을 만들었고 CEEMDAN-XGBoost는 수온, 용존산소(DO), 수소이온농도(pH), 전기전도도(EC), 탁도, 형광 용존 유기물 예측에서 우수한 성능을 보였다.

Barzegar *et al*. (2018)은 신뢰할 수 있는 지하수 위해성 지도를 구축하기 위하여 Extreme Learning Machines (ELM), Multivariate Regression Splines (MARS), M5 tree model, Support Vector Regression (SVR)을 통해 얻은 값을 이용하여 ANN을 설계하였다. Naghibi *et al*. (2015)는 지하수의 측정, 보호 및 관리에 대한 발전을 위하여 Boosted Regression Tree (BRT), Classification and Regression Tree (CART), RF을 성능을 비교하였으며 빠르고 정확하며 알맞은 비용으로 효율적인 결과를 도출하였으며 비슷한 유역에서도 같은 모델을 적용할 수 있음을 주장하였다. Knoll *et al*. (2019)는 The European Water Framework Directive의 질산염 감소에 대한 요구사항을 충족하기 위하여 질산염 농도의 공간적 분포는 Geographic Information System (GIS) 기반의 통계적 방법을 이용하였고, 다양한 통계 방법의 비교를 위해 Multiple Linear Regression (MLR), CART, RF, BRT을 사용하였다.

국내에서 기계학습을 지하수에 적용한 사례로 Yoon *et al*. (2016)는 ANN과 SVR을 이용하여 지하수위 시계열 예측모델을 개발하였으며 각 모델 파라미터 값의 범위를 제시하였다. Chung *et al*. (2017)은 ANN을 이용하여 지하수위를 월 단위로 예측하는 모델을 개발하였고 이를 실제 관측공에 적용하였으며 수개월 동안 강수가 없는 제주도 지역에 대한 지하수위 저하 양상을 분석하였다. Kim *et al*. (2019)는 지하수 함양량 산정을 위하여 표고, 경사, 토양과 같은 15개의 항목을 입력 인자로 설정하고 지하수 함양율을 출력 인자로 설정하여 ANN 모델을 개발하였다.

기계학습을 이용한 지하수 염소이온(Cl^-)과 관련된 연구 사례는 Tran *et al*. (2021)에서 RF, XGBoost, CatBoost (CB), Light Gradient Boosting (LGB)을 이용하여 메콩강의 삼각주에서 염분 농도 예측을 하였다. Mosavi *et al*. (2020)는 염도 증가로 인한 심각한 환경 문제에 따른 관리 및 완화를 위하여 지하수 염도 지도화에 기계학습 적용을 제시하였다. Kassem *et al*. (2021)는 Cl^- 농도에 영향을 주는 주요 변수를 선별하기 위하여 ANN 모델을 제시하였다.

국외에 비해 국내에서 기계학습을 이용한 지하수 적용 사례가 적으며 Cl^- 예측 사례는 더욱 적다. 따라서 국내 기계학습을 통한 지하수 연구에 대한 더욱 많은 적용사례가 필요하며 본 연구에서 이러한 간극을 채우기 위한 연구를 수행하였다.

본 연구에서는 제주도의 주요 오염원 중 하나인 염소이온(Cl^-)을 예측하기 위해 제주특별자치도에서 운영하는 지하수위 관측망, 지하수 염분 관측망, 수질 전용 관측망의 2020년 자료를 탐색적 데이터 분석(Exploratory Data Analysis, EDA) 및 전처리(preprocessing)를 진행한 뒤, 구조적 데이터를 사용 및 예측할 때 뛰어난 성능과 변수 중요도를 이용하

여 필요한 변수에 대한 분석이 용이한 XGBoost를 통하여 표고(elevation), 수소이온농도(pH), 전기전도도(EC), 질산성질소(NO_3^--N), 중탄산염(HCO_3^-), 황산염(SO_4^{2-}), 나트륨이온(Na^+), 칼륨이온(K^+), 칼슘이온(Ca^{2+}), 마그네슘이온(Mg^{2+}), 규소이온(Si^{4+})을 입력 인자로 기계학습을 진행하였다(Kwon and Chang, 2021). 또한 회귀모델 평가 지표인 평균절대오차(Mean Absolute Error, MAE), 평균제곱오차(Mean Squared Error, MSE), 결정계수(R^2), 평균제곱근오차(Root Mean Squared Error, RMSE), 평균절대백분율오차(Mean Absolute Percentage Error, MAPE)를 이용하여 XGBoost Regression의 성능을 평가하였다.

2. 연구지역 및 연구방법

2.1 연구지역

제주도는 한반도 남쪽으로부터 90 km 떨어진 곳에 위치하며, 한라산을 중심으로 74 km의 장축과 약 31 km의 폭을 지닌 면적 1,833.2 km^2 의 타원형의 섬이다(Lee *et al*., 2007). 주봉인 한라산을 중심으로 동서 사면은 완만한 경사를 가지고, 남북 사면은 급한 사면을 갖는다. 해발고도를 기준으로 200 m 이하 지역은 제주도 전체면적의 55.3%, 200~500 m 지대 27.9%, 500~1,000 m 지대 12.3%, 1,000 m 이상의 고산지대는 4.5%를 차지하고 있다.

한편 제주도의 지질층서는 하부에서부터 선캄브리아기-팔레오세의 미문상화강암(불국사화강암), 흑운모화강암(대보화강암), 화성쇄설암으로 구성된 기반암, 플라이오세 후기-플라이스토세 전기의 함력 사암, 세립 사암, 사질 이암, 이암과 상부에 유리질 쇄설암, 수성화산쇄설층이 협재된 서귀포층, 파호에호에 용암으로 구성된 표선리현무암군, 역암, 함력 사암, 사암, 함력 사질 이암, 사질 이암, 이암과 같은 쇄설성 퇴적층으로 구성된 탐라층, 탐라층을 관입하는 현무암질 암맥복합체, 파호에호에 용암과 아아 용암으로 구성되는 한라산현무암군, 용암돔과 관입암으로 구성된 백록담조면암군, 분석구의 산체붕괴로 인하여 형성된 화산성 암설사태층, 홀로세의 역암층과 사암층이 호층을 이루는 신양리층, 고해빈-내만사층으로 구성된다(Yoon *et al*., 2014).

제주도는 거의 대부분 수자원을 지하수에 의존하기 때문에 지하수 지속적인 개발 및 관리는 필수적이다. 2017년 기준, 제주도 수자원 개발이용 시설(총 5,981개) 중 지하수는 4,818개 시설과 시설용량 568×10^6 m^3/year으로 전체 시설용량 90.53%를 차지한다. 수자원 이용 또한 243×10^6 m^3/year으로 전체의 81.378%를 차지한다. 지속이용가능량은 652×10^6 m^3/year으로 남부, 동부, 북부, 서부 순으로 가능량이 높으며 지하수 분포면적은 동부, 남부, 북부, 서부 순으로 각 494.7 km^2, 492.2 km^2, 466.1 km^2, 375.3 km^2으로 분포한다. 2022년 기준, 지하수 허가현황은 총 4,806공, 581×10^6 m^3/year로 조사되었다. 지하수 관측망은 지하수위관측, 인공함양관측, 농업용수모니터링이 각 135, 8, 12개 시설을 운영되고 있다(Jeju Special Self-Governing Province, 2022).

한편 제주도 지하수에 대한 연구는 1990년대부터 본격적으로 수행되고 있다(Jung, 2012). 연구결과 제주도 지하수의 부존형태는 담수와 해수의 관계, 지하 지질 분포 특성, 지하수위 분포 특성 등에 따라 상위지하수, 준기저지하수, 기저지하수, 기반암 지하수로 분류된다(Won *et al*., 2006). 수질에 있어서는 NO_3^--N과 Cl^-이 지하수에서 가장 흔히 나타나는 오염물질이며 자연적인 인자와 인위적인 인자로 인해 그 농도가 변동한다. 농업활동과 정화조가 없는 지역에서는 이들 오염물질이 가중되어 지하수 수질이 악화되고 있는 것으로 보고되고 있다(Kim *et al*., 2007).

2.2 자료수집

연구에 사용된 지하수 관측자료는 제주특별자치도에서 1992년부터 운영하고 있는 지하수위 관측망, 지하수 염분 관측정, 수질 전용 관측망의 2020년 관측자료를 이용하였다(Jeju Special Self-Governing Province and Jeju Groundwater Research Center, 2021). 이들 자료는 제주특별자치도 지하수정보시스템(water.jeju.or.kr)에서 취득하였다. 제주특별자치도는 총 104개소의 관측소(지하수 심도가 75 m 이하인 지하수위 관측시설 91개소, 수질전용 관측시설 9개소, 중산간 지역 관측소 4개소)를 대상으로 지하수를 연 2회 채수하여 실내 실험실에서 이온크로마토그래피를 이용해 주요이온 분석을 실시한다. 현장에서 수온, pH 및 EC를 측정하고 채수한 지하수 시료를 이용하여 실내에서 양이온(Na^+, K^+, Mg^{2+}, Ca^{2+}), 음이온(HCO_3^-, NO_3^--N, Cl^-, SO_4^{2-})을 분석한다.

Table 1. Number of data, mean, deviation, minimum, 25%, 50%, 75%, and maximum values of groundwater quality.

Parameter	Elevation	pH	EC	Cl^-	NO_3^--N	HCO_3^-	SO_4^{2-}	Na^+	K^+
Count	208	188	198	203	204	195	204	203	204
Mean	60.052212	7.871277	1034.397475	205.078325	5.811275	61.978974	32.562255	135.251724	48.682843
STD	74.114449	1.248668	3816.320036	844.275769	6.701292	49.468842	102.005712	541.466056	234.514343
Min	1.69	6.2	7.2	4.6	0	0	1.6	0	1.4
25%	16.98	7.2	155.475	10.85	1.5	39.3	3.875	10.1	3
50%	42.96	7.5	259.55	18.3	3.65	51	8.45	15.9	4.3
75%	77.5	8.1	449.75	32.05	7.675	68.8	20	24.95	6.525
Max	546	12.7	30770	6987.7	43.7	609.5	1014.6	4478.7	1927.1

Parameter	Ca^{2+}	Mg^{2+}	NH_4^+-N	NO_2^--N	F^-	Cu^{2+}	Si^{4+}	Sr^{2+}	Zn^{2+}
Count	204	203	188	189	189	204	189	189	163
Mean	18.522549	13.979803	0.070048	0.001757	0.187561	0.00166	17.043915	0.213096	0.117585
STD	37.898146	42.838318	0.600337	0.024149	0.260794	0.009896	4.727312	1.392602	1.409491
Min	0.9	0	0	0	0	0	0	0	0
25%	6.675	3.85	0	0	0.088	0	15.4	0.0414	0
50%	10.25	6.1	0	0	0.114	0	17.3	0.07	0
75%	16.5	12.2	0	0	0.164	0	19.5	0.1226	0.006
Max	412.6	477.8	7	0.332	1.876	0.09	30.1	19.1	18

2.3 전처리

2.3.1 결측치 확인 및 처리

개별 행의 결측치를 직접 엑셀과 같은 표로 확인하는 것은 매우 어려운 일이므로 이에 결측치를 확인하고 패턴을 감지하기 위하여 Python package 중 하나인 Missingno matrix를 사용하였다(Bilogur, 2018).

Missingno matrix와 Python 패키지를 통해 각 수질항목의 결측치 개수는 pH: 20, EC: 10, Cl^-: 5, NO_3^--N: 4, HCO_3^-: 13, SO_4^{2-}: 4, Na^+: 5, K^+: 4, Ca^{2+}: 4, Mg^{2+}: 5, NH_4^+-N: 20, NO_2^--N: 19, F^-: 19, Cu^{2+}: 4, Si^{4+}: 19, Sr^{2+}: 19, Zn^{2+}: 45로 확인되었다. 결측치를 잘못된 값으로 채운다면 편향된 추정치와 왜곡된 통계력 및 결론을 얻을 수 있다(Acock, 2005). 결측치를 처리하기 위해서는 누락된 데이터를 제거하거나 결측치를 특정 값으로 보간하는 방법 등 다양한 방법이 존재한다(Kaiser, 2014).

제주특별자치도 지하수위 관측망, 지하수 염분 관측정, 수질 전용 관측망의 2020년 관측자료의 결측치를 처리하기 위해 18개의 수질항목의 평균, 편차, 최솟값, 최댓값, 4분위 값을 고려하여 Missing 1, Missing 2, Missing 3으로 그룹을 나누었다(표 1). Missing 1에 해당하는 Zn^{2+}, NO_2^--N, NH_4^+-N, Cu^{2+}은 대부분 값이 0을 가지며 이는 불검출을 의미한다. Missing 1의 결측치는 최빈값인 0으로 보간하였다. Missing 2에 해당하는 pH, NO_3^--N, HCO_3^-, F^-, Si^{4+}, Sr^{2+}은 비교적 값들이 큰 차이를 갖지 않는다. Missing 2의 결측치는 평균으로 보간하였다. Missing 3에 해당하는 EC, Cl^-, SO_4^{2-}, Na^+, K^+, Ca^{2+}, Mg^{2+}은 이상값을 가지며 값들의 차이가 매우 크다. Missing 1, Missing 2와 같이 0이나 평균으로 보간할 경우 왜곡된 결과와 초래할 수 있어 결측치를 갖는 해당 열을 삭제하는 방법을 사용하였다(Kaiser, 2014).

2.3.2 이상치 제거

이상치란 관측된 데이터의 범위에서 벗어난 아주 작거나 큰 값을 말한다. 이상치 데이터는 모델의 예측 성능을 하락시키는 불필요한 요소로서 제거가 필수적이다. 표 1를 통하여 각 수질항목의 최솟값, 25%, 중앙값, 75% 최댓값을 분석하여 이상치를 확인하였다. Missing 3의 이상치 제거를 위하여 우선적으로 사분위법(Inter Quantile Range, IQR)을 사용하였다. 사분위법은 데이터를 오름차순으로 정렬하였을 때 중간 50%의 해당되는 데이터들을 말하며 Q_1, Q_2, Q_3 각 0.25, 중앙값, 0.75에 위치하는 데이터

를 기준으로 네 구간으로 나눈다. 사분위법(IQR)을 이용한 이상치은 Q_1, Q_3에 위치한 데이터 값에 1.5* 중앙값을 빼거나 더해 구한 값의 이하, 이상 값들을 말한다(Jeon *et al*., 2020)(식 1). 그러나 높은 값을 가지는 데이터가 상당수 제거되어 사분위법을 이용한 이상치 제거는 사용하지 못하였다(그림 1).

$$IQR = Q_3 - Q_1$$
$$Q_1 - 1.5 \times IQR \leq X \leq Q_3 + 1.5 \times IQR \quad (1)$$

데이터 손실 최소화와 모델 예측 성능 향상을 목적으로 이상치 제거 작업을 진행하였으며 사용한 방법은 다음과 같다. 수질항목들의 값을 오름차순으로

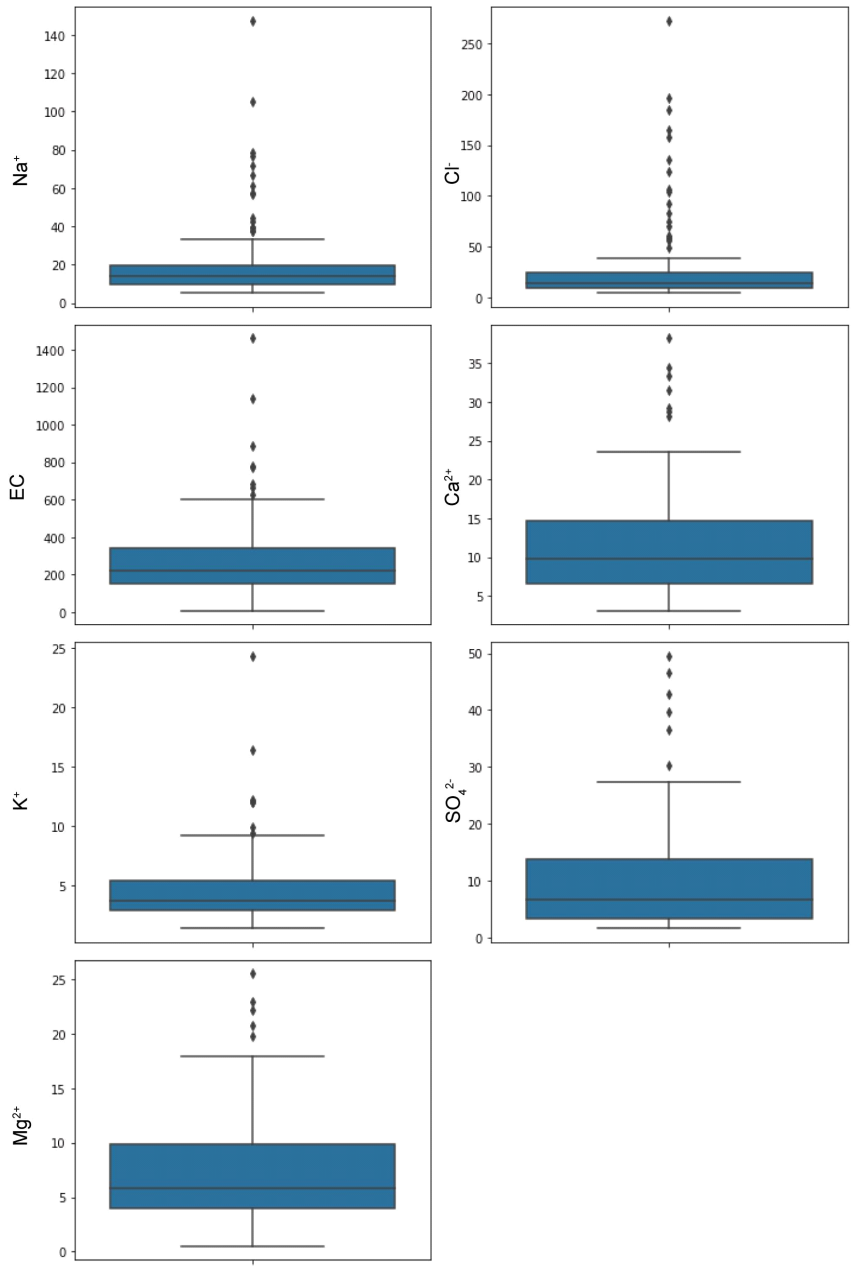

Fig. 1. Box plots of missing 3 after removing missing values and outliers.

Table 2. Percentile distribution of parameters (94%, 95%, 96%, 97%, 98%, 99%) with outliers sorted in ascending order.

Parameter	0.94	0.95	0.96	0.97	0.98	0.99
Elevation	119.351	132.000	170.044	224.1435	274.931	435.724
pH	9.830	10.000	10.700	11.4450	11.940	12.505
EC	1370.000	1795.500	2927.040	6782.6500	17445.000	20755.000
Cl^-	257.700	291.600	589.700	1265.8750	2017.170	4004.490
NO_3^--N	16.230	19.000	20.520	21.8350	29.900	31.210
HCO_3^-	120.800	132.250	134.360	143.9300	148.950	158.245
SO_4^{2-}	47.400	53.450	97.180	135.1400	266.710	405.945
Na^+	148.180	208.525	317.520	941.4200	1703.520	2933.330
K^+	33.010	42.650	67.860	106.8150	410.540	1133.420
Ca^{2+}	35.960	39.300	41.940	53.3750	71.020	86.520
Mg^{2+}	23.020	25.575	28.180	30.3050	60.030	139.765
Si^{4+}	23.030	23.425	23.800	24.0750	25.030	25.940

정렬 후, 구간별 값들의 변화 증가 추세를 파악하였다. 표 2와 같이 0.94, 0.95, 0.96, 0.97, 0.98, 0.99 구간의 값을 분석한 결과 0.96 구간 수질항목들의 값은 0.95 구간보다 EC, Cl^-, SO_4^{2-}, Na^+가 각 1,131.5 μS/cm, 298.1 mg/L, 108.9 mg/L, 43.7 mg/L로 급격한 상승을 보인다. 0.96 이후 구간부터는 더욱 급격한 상승을 나타내기 때문에 0.96 이전 구간의 데이터를 사용하고 해당 구간부터 이상치로 판단하고 제거하였다. 0.95 구간에 해당하는 값들까지 사용하였기에 데이터 손실을 최소화와 모델 예측 성능을 향상시킬 수 있었다.

2.3.3 로그 변환

표고, pH, EC, Cl^-, NO_3^--N, HCO_3^-, SO_4^{2-}, Na^+, K^+, Ca^{2+}, Mg^{2+}의 히스토그램을 살펴보았을 때, 왜도가 양의 값을 가지며(positive skewness) 수질항목별 크기(scale) 또한 다르다. 그러므로 기계학습의 성능을 올리기 위해서는 최대한 정규분포로 만들어야 하며 수질항목들을 크기를 맞추어야(scaling) 한다. 로그 변환은 연속형 자료를 정규화할 때 매우 효과적이며 (O'Hara and Kotze, 2010) 비교적 큰 값을 작게 만들고 왜도와 첨도를 줄여 정규분포를 나타내게 할 수 있다. 위 수질항목들은 log-normal 분포를 나타내며 Gaussian 분포로 변화시키기 위하여 $\log(1+X)$를 취하여 변환하였다(그림 2). $\log(X)$가 아닌 $\log(1+X)$를 사용한 이유는 데이터에 0 값이 존재하여 y가 -무한대(-infinite)의 값을 가지는 것을 방지하기 위함이다. 또한 로그 변환된 수질항목들은 0-7.28의 값의 비교적 유사한 크기로 변환되었다.

2.4 XGBoost

XGBoost는 EXtreme Gradient Boosting의 약자로 XGBoost는 Gradient boosting 알고리즘을 기반으로 한다(Lee and Sun, 2020). Gradient Boosting이란 여러 개의 약한 학습기(weak learner)를 조합하여 사용하는 앙상블 학습(Ensemble Learning) 중 하나인 Boosting 기법에 오류식을 최소화하는 방향성을 가지고 반복적으로 가중치 값을 갱신하는 방법인 경사 하강법을 접목한 것이다. 앙상블 학습이란 여러 개의 분류기를 생성하고 예측을 결합함으로써 보다 정확도가 높은 예측을 도출하는 기법이다. XGBoost는 나무 구조(Tree) 형태로 데이터를 분할하는 의사결정나무(decision tree)를 약한 학습기로 이용한다(Chen and Guestrin, 2016).

본 연구에 사용된 XGBoost는 분류와 회귀에서 우수한 성능을 발휘한다. 기술적인 측면에서 일반적인 Gradient Boosting Machine (GBM)은 순차적으로 가중치를 증가시켜 학습하지만 XGBoost는 병렬 수행을 통하여 비교적 빠른 속도를 나타내며 과적합 규제(regularization) 기능이 내장되어 과적합에 대한 내구성을 가진다(Chen and Guestrin, 2016). 식 (2)는 앙상블 모델을 표현하는 식이다.

$$\hat{y}=\sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (2)$$

$\hat{y}$는 예측값, K는 Tree의 개수, f_k는 F 공간에 k 번째 의사결정나무, F는 모든 의사결정나무의 집합이다. Chen and Guestrin (2016)에서 소개된 XGBoost는 손실함수를 정규화하여 Regularized Boosting 기술을 만들었다(Abou Omar, 2018)(식 3).

$$\iota(\phi)=\sum_{i}^{n} \ell(y_i, \hat{y}_i)+\sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

여기서 좌측항인 $\ell(y_i, \hat{y}_i)$는 실제값과 예측값의 차이인 손실함수(loss function)이며 우측항인 $\Omega(f_k)$ 모델의 복잡도가 높아지지 않게 과적합을 방지하는 정규화항이다(Chen and Guestrin, 2016).

2.5 모델 설정 및 평가

2.5.1 GridsearchCV

기계학습 알고리즘을 조정하는 하이퍼파라미터(hyperparameter)는 모델을 구축할 때, 사용자가 직접 설정해줘야 하는 파라미터이다. XGBoost의 하이퍼파라미터인 Colsample_bytree는 의사결정나무

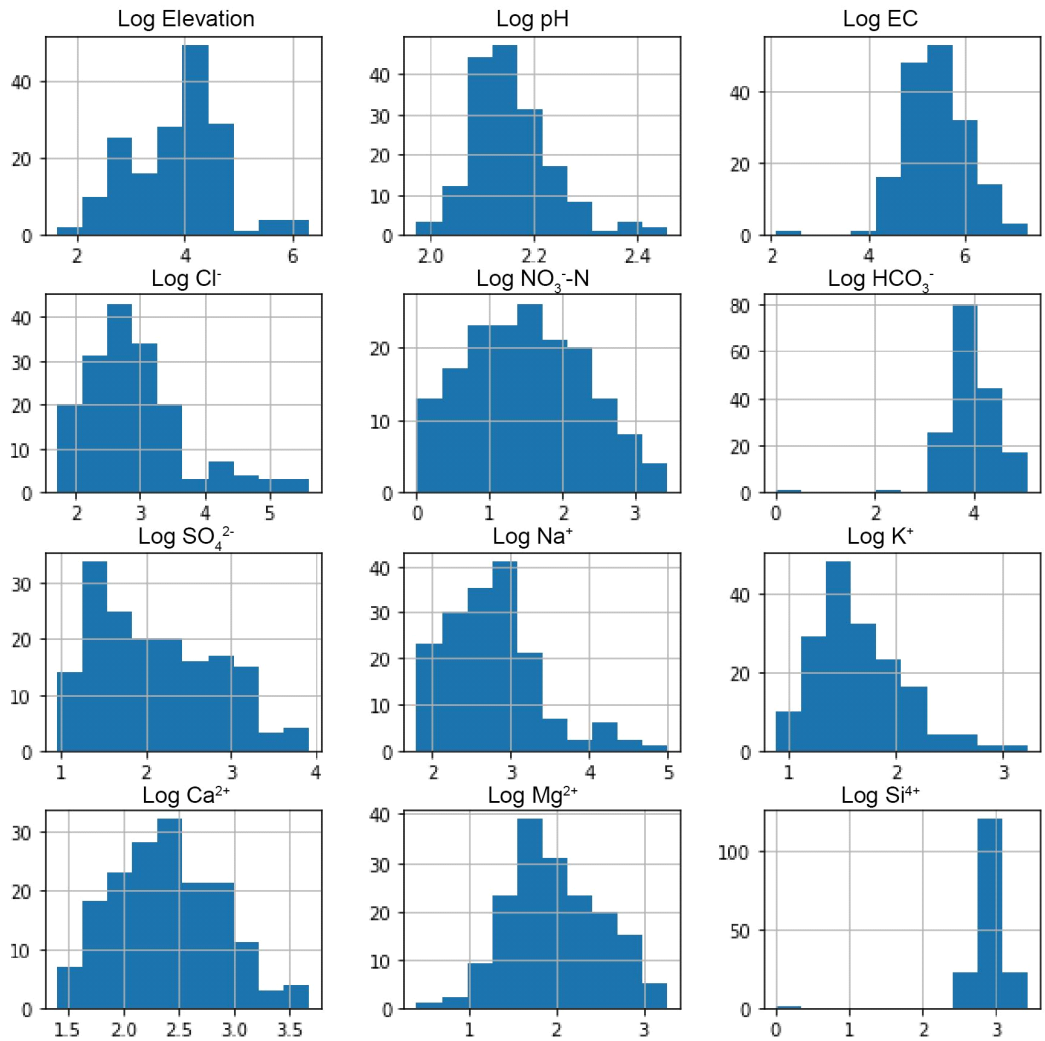

Fig. 2. Normalized and feature scaled variables using log transformation for Cl^- regression.

생성에 필요한 변수(feature) 추출에 사용하며 과적합을 조정한다. Max_depth는 의사결정나무의 최대 깊이를 조정하며 Min_child_weight는 Leaf node에 할당되는 최소 데이터의 수를 의미한다. N_estimators는 Learning_rate (η)는 의사결정나무의 개수를 조정하는 역할을 한다. GridsearchCV는 하이퍼파라미터의 값들을 순차적으로 입력하여 최적의 하이퍼파라미터를 탐색하는 격자 탐색(grid search)과 학습 데이터와 검증 데이터를 K번 나누어 K번의 성능을 평균 값으로 사용하는 교차 검증(cross validation)을 합친 API이다(Lee and Sun, 2020). 본 연구에서는 Learning_rate를 제외한 하이퍼파라미터들의 최솟, 최댓값들을 입력 후 점차적으로 그 범위를 줄여 입력하는 방식으로 XGBoost 알고리즘을 조정하였다. Learning_rate는 GridsearchCV를 사용하지 않고 0.01-0.1 사이에 값 중 최적의 값을 선택하였다.

2.5.2 모델 평가

XGBoost Regression 모델 평가에 사용될 지표는 실제 값과 예측값의 차이를 절대값으로 변환 뒤 합산하여 평균으로 구하는 MAE, 실제값과 예측값의 차이를 제곱하여 평균으로 구하는 MSE, 평균제곱오차에 제곱근을 씌워 구하는 RMSE, 실제값과 예측값의 차이를 실제값으로 나눈 뒤 절대값을 평균으로 구해 백분률로 환산하는 MAPE, 실제값과 예측값의 차이를 제곱한 값의 합(Sum of Squared Residual, SSR)을 실제값과 평균의 차이를 제곱한 것의 합(Sum of Squared Total, SST)으로 나눈 값을 말하며 통산적으로 0에 근접할수록 쓸모없는 회귀식으로 1에 가까울수록 모델이 우수한 성능을 보인다고 평가하는 R^2을 사용하였다. 식에서 y, $\hat{y}$, $\bar{y}$는 각 실제값과 예측값, 평균값을 의미하며 n은 자료의 수를 의미한다(식 4, 5, 6, 7, 8).

$$MAE = \frac{\sum_{i=1}^{n} |y - \hat{y}|}{n} \tag{4}$$

$$MSE = \frac{\sum_{i=1}^{n} (y - \hat{y})^2}{n} \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y - \hat{y})^2}{n}} \tag{6}$$

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y - \hat{y}}{y} \right| \tag{7}$$

$$R^2 = 1 - \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{8}$$

XGBoost와 같은 의사결정나무 기반 모델은 예측하는데 어떠한 변수가 중요하게 작용했는지 보여주는 변수 중요도(feature importance)를 제공한다. 본 연구에서는 변수별 데이터를 분리하는데 쓰인 횟수(weight)로 변수 중요도의 순위를 매겨 Cl^- 예측을 위한 변수를 선별하였다.

3. 결과 및 토론

XGBoost Regression의 예측 성능을 높이기 위하여 불필요한 매개변수 제거 전 수질항목에 대해 피어슨 상관분석과 XGBoost의 변수 중요도(feature importance)를 분석하였고, Cl^-과의 상관성이 낮은 수질항목과 XGBoost Regression에서 중요도가 낮은 매개변수를 비교 및 제거하는 과정을 반복 수행하여 모델 성능을 향상시켰다.

그림 3는 피어슨 상관분석을 이용하여 수질항목과 Cl^-의 관계성을 분석한 결과를 보여준다. SO_4^{2-}, Na^+, EC, F^-, K^+, Mg^{2+}, Ca^{2+}, pH, HCO_3^-, NO_2^--N, NO_3^--N, Cu^{2+}, NH^+-N, Zn^{2+}, Sr^{2+}, 표고, Si^{4+}은 Cl^-과의 상관계수가 각 0.99, 0.98, 0.89, 0.89, 0.78, 0.77, 0.69, 0.58, 0.18, -0.10, -0.15, -0.16, -0.19, -0.22, -0.24, -0.47, -0.56의 값을 가진다.

1차적으로 모든 파라미터를 XGBoost에 입력하여 도출하였을 때 Sr^{2+}, Zn^{2+}, Cu^{2+}, F^-, NO_2^--N, NH_4^+-N의 변수 중요도가 비교적 낮게 나왔으며 Sr^{2+}, Zn^{2+}, Cu^{2+}, NO_2-N, NH_4^+-N은 비교적 낮은 상관관계를 보였다. 변수 중요도와 상관분석 결과, 공통으로 낮은 파라미터를 제외하는 경우보다 낮은 변수 중요도를 갖는 파라미터를 제외했을 때, Sr^{2+}, Zn^{2+}, Cu^{2+},

F^-, NO_2^--N, NH_4^+-N을 제외한 XGBoost의 성능이 향상되었다. 예외적으로 F^-은 비교적 높은 상관계수를 가지지만 변수 중요도는 낮게 나타났다. 결과적으로 Na^+, HCO_3^-, Si^{4+}, K^+, SO_4^{2-}, NO_3^--N, Mg^{2+}, EC, 표고, Ca^{2+}, pH 순으로 XGBoost regression에서 Cl^-를 예측하는데 중요한 변수로 작용하였다(그림 4).

XGBoost를 최적화시키기 위하여 GridsearchCV를 수행하였으며, 이를 위해서는 사용자가 직접 하이퍼파라미터의 값을 설정해야 한다. 학습 단계별 가중치를 얼마나 적용할 지에 대한 하이퍼파라미터인 learning rate와 Tree의 깊이를 결정하는 Max_depth, 과적합을 조정하는데 사용하는 Colsample_bytree, Min_child_weight를 고려하였다(Singha *et al*., 2021). Min_child_weight, Colsamlpe_bytree는 0.5-4, Max_depth는 3-6, N_estimators는 100-500, Learning_rate는 0.04 값을 사용하였다. GridsearchCV을 이용하여 300개의 경우의 수와 5번의 교차검증, 총 1500가

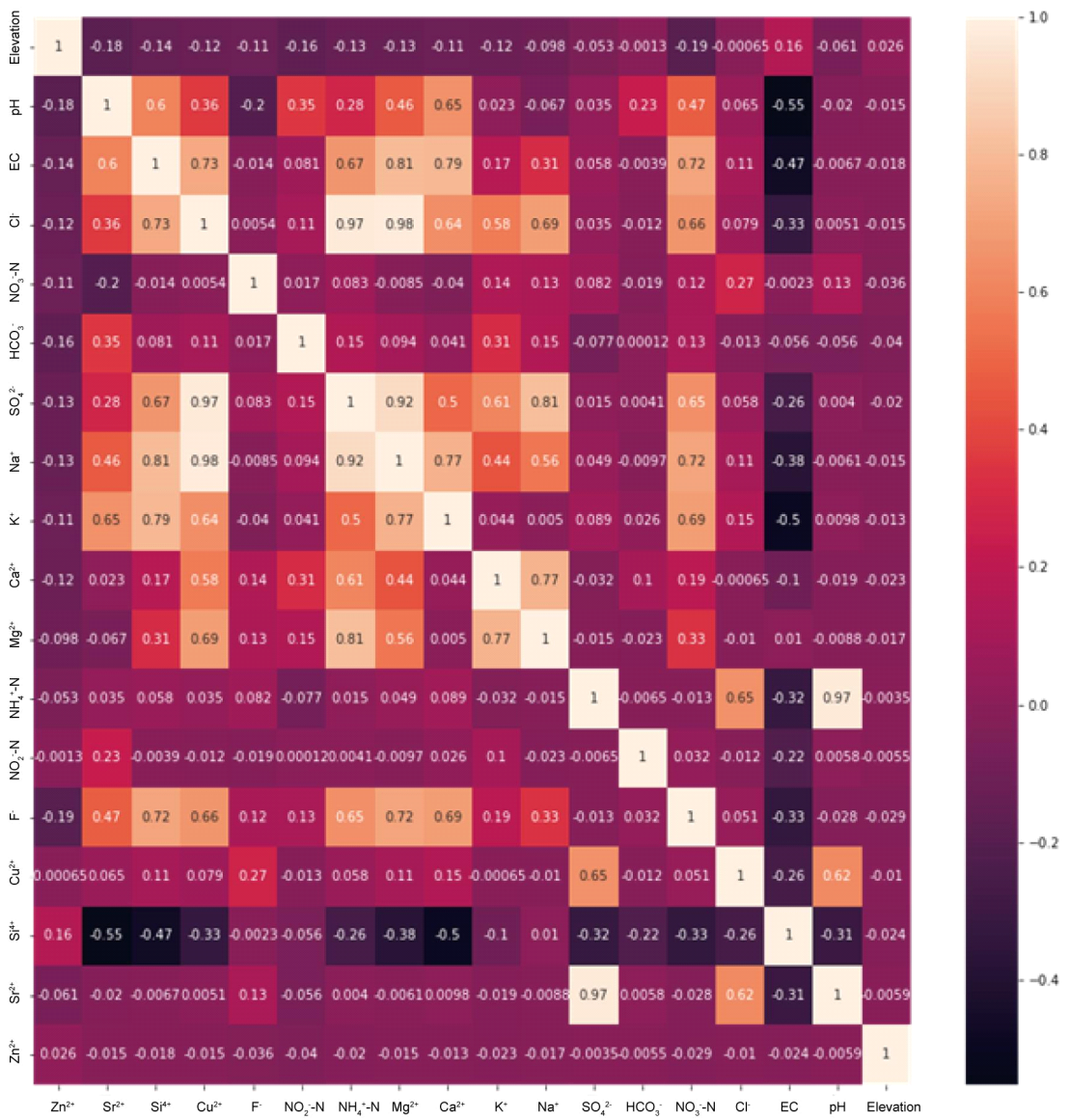

Fig. 3. Pearson correlation coefficients for all variables.

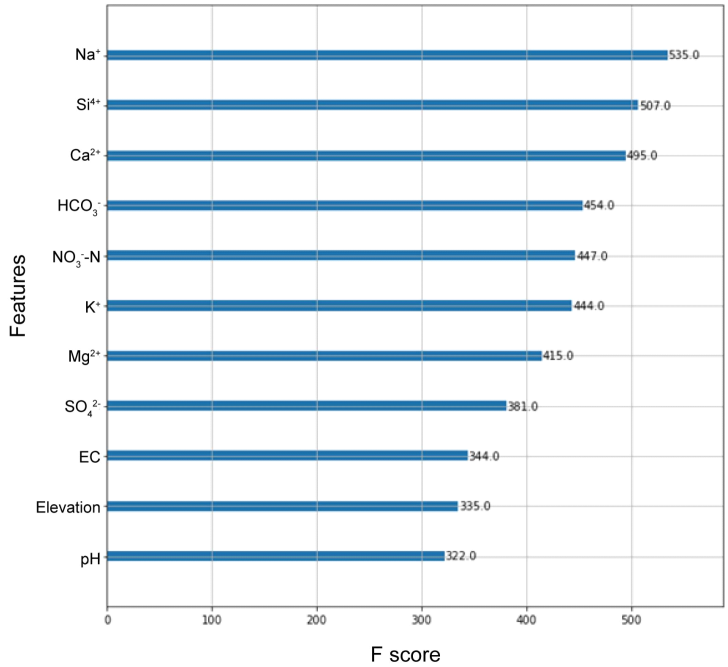

Fig. 4. Feature importance graph when XGBoost regression regressed Cl^-.

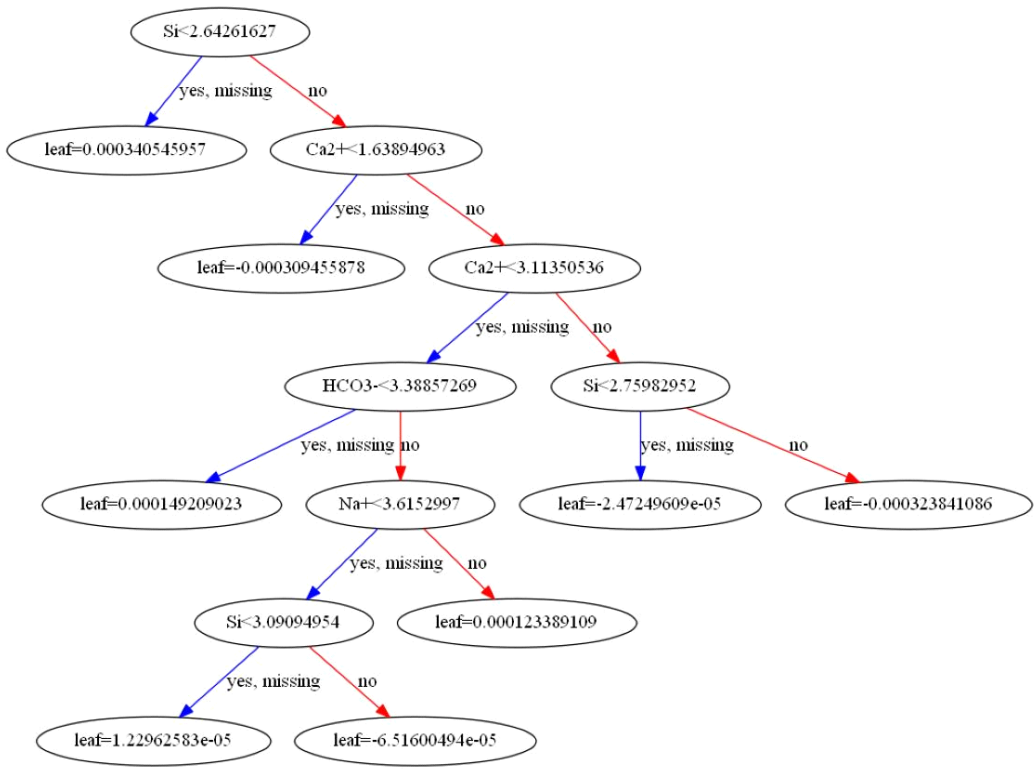

Fig. 5. Visualization of the last tree out of 500.

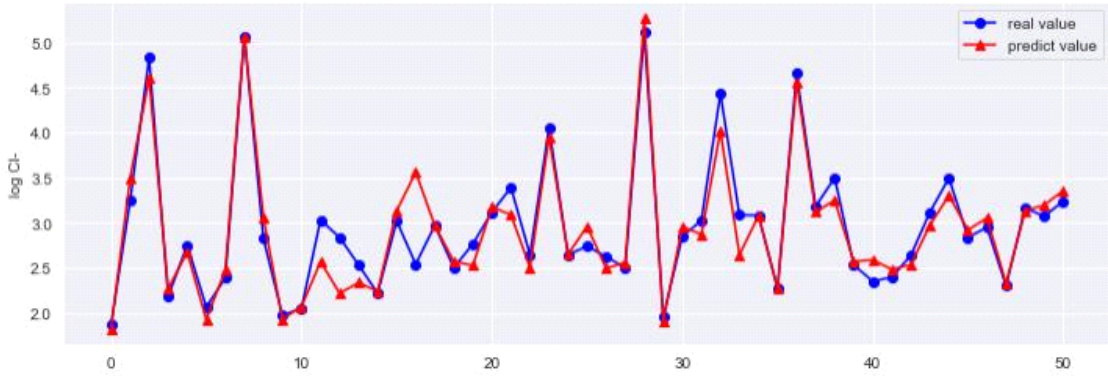

Fig. 6. Comparison graph of real and predict values.

지 경우의 수를 순차적으로 학습시켰으며 최적의 하이퍼파라미터 값은 Colsample_bytree 0.5, Max_depth 6, Min_child_weight 4.0, N_estimators 500, Learning_rate 0.04으로 측정되었으며 이를 사용하였다. N_estimators는 500으로 설정하였으며 그 중 마지막 의사결정나무는 그림 5와 같다.

MAE, MSE, R^2, RMSE, MAPE을 이용하여 XGBoost 모델을 평가하였다(Doreswamy *et al*., 2020). MAE 0.156, MSE 0.055, R^2 0.900, RMSE 0.235, MAPE 0.053의 값으로 평가되었으며 실제값과 예측값의 비교를 위한 시각화는 그림 6과 같다.

4. 결 론

본 연구에서는 XGBoost regression을 이용하여 제주특별자치도 지하수위 관측망, 지하수 염분 관측정, 수질 전용 관측망의 2020년 관측자료에서 추출한 18개의 수질항목 중 11개의 입력값으로 Cl^-을 예측하였다. 상관분석과 변수 중요도 분석을 통하여 Cl^-을 예측하기 위해서는 Na^+, HCO_3^-, Si^{4+}, K^+, SO_4^{2-}, NO_3^--N, Mg^{2+}, EC, 표고, Ca^{2+}, pH의 매개변수를 이용하여 회귀하였을 때 가장 좋은 예측값(R^2=0.904)을 얻었다. 그러나 본 연구에서 어려운 점은 수질측정 자료의 결측치와 이상치 제거에 대한 최적의 방법을 찾는 것이었다. 이상치 처리에 대한 고민과 다양한 요인으로 인한 일부 수질측정자료 값이 비정상적으로 높았으며 Cl^- 농도를 예측하는데 어려움을 겪었다. 그러므로 수질 측정자료의 이상치 처리, 정규화 및 표준화에 대한 추가적인 연구가 필요하다고 판단된다.

제주도 주요 오염물질 중 하나인 Cl^-을 더욱 정확히 예측하려면 주기적인 시계열적인 관측자료와 좌표, 강수량, 해안거리 등 추가적인 매개변수들이 필요하다. 위와 같은 문제점들이 개선된다면 XGBoost 이외의 다른 기계학습 또한 Cl^-을 포함한 다른 오염물질들에 대한 예측 성능이 더욱 향상될 것으로 사료된다.

감사의 글

본 논문은 2022년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2019R1A6A1A03033167). 또한 환경부의 재원으로 한국환경산업기술원의 미세플라스틱 측정 및 위해성평가 기술개발사업의 지원을 받아 연구되었음(2020003110010).

REFERENCES

Abou Omar, K.B., 2018, XGBoost and LGBM for Porto Seguro's Kaggle Challenge: A Comparison. ETH Zurich, Zurich, 16 p.

Acock, A.C., 2005, Working with missing values. Journal of Marriage and Family, 67, 1012-1028.

Barzegar, R., Moghaddam, A.A., Deo, R., Fijani, E. and Tziritis, E., 2018, Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms. Science of the Total Environment, 621, 697-712.

Bi, Q., Goodman, KE., Kaminsky, J. and Lessler, J., 2019, What is machine learning? A primer for the epidemiologist. American Journal of Epidemiology, 188, 2222-2239.

Bilogur, A., 2018, Missingno: a missing data visualization suite. Journal of Open Source Software, 3, 1-4.

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, 785-794.

Cheng, S., Zhang, S., Li, L. and Zhang, D., 2018, Water quality monitoring method based on TLD 3D fish tracking and XGBoost. Mathematical Problems in Engineering, 2018, 5604740.

Chung, I., Lee, J. and Chang, S.W., 2017, Long-term prediction of groundwater level in Jeju Island using artificial neural network model. Journal of the Korean Society of Civil Engineers, 37, 981-987.

Doreswamy, Harishkumar, K.S., Yogesh, K.M. and Gad, I., 2020, Forecasting air pollution particulate matter ($PM_{2.5}$) using machine learning regression models. Procedia Computer Science, 171, 2057-2066.

Jeju Special Self-Governing Province, 2022, https://water.jeju.go.kr/main.cs (April 25, 2022).

Jeju Special Self-Governing Province and Jeju Groundwater Research Center, 2021, Annual Groundwater Monitoring Report. Jeju Special Self-Governing Province, 213 p (in Korean).

Jeon, T.Y., Yu, S.H. and Kwon, H.Y., 2020, Improvement of PM forecasting performance by outlier data removing. Journal of Korea Multimedia Society, 23, 747-755 (in Korean with English abstract).

Jordan, M.I. and Mitchell, M.T., 2015, Machine learning: Trends, perspectives, and prospects. Science, 349, 255-260.

Jung, H.J., 2012, Underground water discourse and policy at JeJu in 1990s. Tamla Culture, 40, 171-224 (in Korean with English abstract).

Kaiser, J., 2014, Dealing with missing values in data. Journal of Systems Integration, 5, 1804-2724.

Kassem, Y., Gökçekuş, H. and Maliha, M.R., 2021, Identifying most influencing input parameters for predicting chloride concentration in groundwater using an ANN approach. Environmental Earth Sciences, 80, 1-16.

Kim, G., Hwang, C.I., Shin, H.J. and Choi, M.R., 2019, Applicability of ground-water recharge rate estimation method based on artificial neural networks in unmeasured areas. Journal of the Geological Society of Korea, 55, 693-701 (in Korean with English abstract).

Kim, G., Kim, J., Won, J. and Koh, G., 2007, Regional trend analysis for groundwater quality in Jeju Island- focusing on chloride and nitrate concentrations. Journal of Korea Water Resources Association, 40, 469-483 (in Korean with English abstract).

Knoll, L., Breuer, L. and Bach, M., 2019, Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Science of the Total Environment, 668, 1317-1327.

Kwon, C.W. and Chang, H.H., 2021, Comparative analysis of traffic accident severity of two-wheeled vehicles using XGBoost. The Journal of The Korea Institute of Intelligent Transport Systems, 20, 1-12 (in Korean with English abstract).

Lee, C.S., Cho, T.C., Lee, S.B. and Won, K.S., 2007, A study of weathering characteristic of Baeknokdam trachyte in Jeju Island. The Journal of Engineering Geology, 17, 235-251 (in Korean with English abstract).

Lee, Y. and Sun, J., 2020, Predicting highway concrete pavement damage using XGBoost. Korean Journal of Construction Engineering and Managment, 21, 46-55 (in Korean with English abstract).

Lu, H. and Ma, X., 2020, Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere, 249, 126169.

Mitchell, T.M., 1996, Machine Learning. McGraw Hill, New York, 432 p.

Mosavi, A., Hosseini, F.S., Choubin, B., Taromideh, F, Ghodsi, M., Nazari, B. and Dieva, A., 2020, Susceptibility mapping of groundwater salinity using machine learning models. Environmental Science and Pollution Research, 28, 10804-10817.

Naghibi, S.A. and Pourghasemi, H.R., 2015, A comparative assessment between three machine learning models and their performance comparison by bivariate and multivariate statistical methods in groundwater potential mapping. Water Resources Management, 29, 5217-5236.

O, W.K., Jang, H.N. and Shin, S.G., 2022, Application of machine learning in water industry: A review. Journal of the Korean Society of Water and Wastewater, 36, 9-21 (in Korean with English abstract).

O'Hara, R. and Kotze, J., 2010, Do not log-transform count data. Methods in Ecology and Evolution, 1, 118-122.

Singha, S., Pasupuleti, S., Singha, S.S., Singh, R. and Kumar, S., 2021, Prediction of groundwater quality using effi-

cient machine learning technique. Chemosphere, 276, 1-11.

Tran, D., Tsujimura, M., Ha, N.T., Nguyen, V.T., Binh, D.V., Dang, T.D., Doan, Q., Bui, D.T., Ngoc, T.A., Phu, L.V., Thuc, P. and Pham, T.D., 2021, Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong delta, Vietnam. Ecological Indicators, 127, 107790.

Won, J.H., Lee, J.Y., Kim, J.W. and Koh, G.W., 2006, Groundwater occurrence on Jeju Island, Korea. Hydrogeology, 14, 532-547.

Yoon, H., Yoon, P., Lee, E., Kim, G.B. and Moon, S.H., 2016, Application of machine learning technique-based time series models for prediction of groundwater level fluctuation to national groundwater monitoring network data. Journal of the Geological Society of Korea, 52, 187-199 (in Korean with English abstract).

Yoon, S., Jung, C.Y., Hyun, W.H. and Song, S.T., 2014, Tectonic history of Jeju Island. Journal of the Geological Society of Korea, 50, 457-474 (in Korean with English abstract).

Received : April 26, 2022
Revised : May 20, 2022
Accepted : May 24, 2022