Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2024.47.3.068

Vol. 47, No. 3 : 68-74, September 2024
ISSN : 2005-0461 / EISSN : 2287-7975

Machine Learning Based Coagulant Rate Decision Model for Industrial Water Treatment Plant

Kyungsu Park* · Yu-jin Lee* · Haneul Noh* · Jun Heo** · Seung Hwan Jung***†

*Department of Business Administration, Pusan National University
**Korea Water Resources Corporation
***School of Business, Yonsei University

머신러닝 기반의 공업용수 정수장 응집제 주입률 결정

박경수* · 이유진* · 노하늘* · 허 준** · 정승환***†

*부산대학교 경영학과
**한국수자원공사
***연세대학교 경영학과

This study develops a model to determine the input rate of the chemical for coagulation and flocculation process (i.e. coagulant) at industrial water treatment plant, based on real-world data. To detect outliers among the collected data, a two-phase algorithm with standardization transformation and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is applied. In addition, both of the missing data and outliers are revised with linear interpolation. To determine the coagulant rate, various kinds of machine learning models are tested as well as linear regression. Among them, the random forest model with min-max scaled data provides the best performance, whose MSE, MAPE, R^2 and CVRMSE are 1.136, 0.111, 0.912, and 18.704, respectively. This study demonstrates the practical applicability of machine learning based chemical input decision model, which can lead to a smart management and response systems for clean and safe water treatment plant.

Keywords : Industrial Water Treatment Plant, Coagulant Rate, Random Forest Model

1. 서 론

현대 사회에서 물은 산업, 생활 및 다양한 활동을 위한 핵심 자원 중 하나로, 깨끗하고 안전한 물의 공급은 국가 및 지역 사회의 핵심 이슈이며, 산업단지에서도 공업용수의 처리 및 정수장 운영은 필수적이다. 또한 기상 및 환경적 요인 변화, 인구 증가, 산업화 등으로 인해 이러한 수질 관리의 중요성은 나날이 증대되고 있다. 장마철 유속의 변화 및 난류 증가로 인해 원수의 수소 이온 농도(pH)가 낮아지거나 탁도가 증가하기도 하고, 다른 여러 가지 이유로 미생물의 함량, 전기전도도, 알칼리도가 변화하기도 한다. 이렇듯 변화하는 원수의 상태에 대응하여 수자원을 처리하고 정수장을 운영하여 양질의 용수를 제공해야 한다. 이를 위해 본 연구에서는 울산에 소재하는 온산국가산업단지에 공업용수를 제공하는 온산공단 정수장 데이터를 기반으로 정수를 위한 약품투입량 결정 문제를 다룬다.

온산공단 정수장에서의 약품 투입률은 Jar-test에 의해 만들어진 약품투입용 조견표를 참고하여 근무자가 실시

Received 6 August 2024; Finally Revised 30 August 2024; Accepted 2 September 2024
† Corresponding Author : seunghwan.jung@yonsei.ac.kr

간적으로 수질변화에 대응하며 약품을 투입한다. 이러한 조견표는 통제된 환경의 실험실에서 특정 지표만을 참고하여 도출된 것이므로 실제 정수장 환경 및 상황과는 다소 차이가 있을 수 있다. 또한 근무자의 주관적 판단에 따라 실제 약품 투입률이 달라지므로 근무자의 경험 및 숙련도에 따라 약품투입률 및 정수 결과가 달라지며, 신규 근무자의 경우에는 약품투입률 결정에 큰 어려움이 있을 수 있다. 이에 따라 현장에서는 깨끗하고 안전한 물에 대한 스마트 관리 및 대응체계 구축에 대한 수요가 높아지고 있는 실정이다.

따라서 본 연구에서는 수소 이온 농도(pH), 탁도(Turbidity), 알칼리도(Alkalinity) 등 온산 정수처리장에서 수집하고 있는 원수 데이터 및 과거 근무자들의 실제 약품투입률을 활용하여 머신러닝 기반의 응집제 투입률을 결정하는 알고리즘을 제시한다. 또한 원활한 학습을 위해 수집된 데이터의 결측치 및 이상치 처리 프로세스에 대해서도 다룬다.

본 논문의 구성은 다음과 같다. 제2장에서는 선행연구를 수행한다. 제3장에서는 정수처리 프로세스 및 데이터에 관해 설명하고, 제4장에서는 이러한 데이터의 결측치 및 이상치 처리 프로세스에 대해 다룬다. 제5장에서는 응집제 투입률 예측 모형 및 결과를 제시하며, 제6장에서는 결론을 토의한다.

2. 선행 연구

머신러닝을 비롯한 여러 인공지능 기법은 다양한 분야에서 활용된다. 수자원 처리 과정 및 수질 분석에도 다양한 인공지능 기법이 사용되고 있으며 이와 관련된 선행 연구는 다음과 같다.

XGBoost는 복수의 모델을 학습하며 일반 앙상블 모델과 달리 부스팅 방식을 사용하여 효율성이 높은 Gradient boosting 기반 머신러닝 기법이다[8]. Sim et al.[9]에서는 제주도 지하수의 주요 오염물질 중 하나인 염소이온의 농도를 예측하고자 XGBoost regression를 활용하였다. 특히 제주도 지하수에 주요이온분석을 실시하여 주된 양이온, 음이온을 구분하고 그 외에 수온, 수소이온농도지수(pH), 전기전도도(EC)를 측정하여 학습 변인으로 사용한다.

랜덤 포레스트(Random Forest, RF)는 XGBoost와 같은 앙상블 기법이지만 부스팅이 아닌 배깅 방식을 적용하는 방식으로 예측에 자주 활용되며, 무작위 복원 추출로 이상치 및 잡음(Outliers)의 영향을 줄여 높은 성능을 보인다. 응집은 물의 절반 이상에 있는 표층수의 탁도와 색을 처리하기 위한 필수 단위 공정으로 적절한 양의 응집제 투입이 중요하며, Achite et al.[1]에서는 수처리시설(Water treatment plant)에서 수집한 원수 생산량, 탁도수, 전기전도도, 부유물질 등을 학습 변인으로 하여 응집제 투입률을 예측하였다. Park et al.[5] 등에서 연구되었듯이 RF는 주어진 입력에 대해 예측값을 출력하는 모델이기 때문에 출력값을 최적화하는 입력값을 찾기 위해 최적화 알고리즘과 결합하기도 하며, Achite[1]에서는 RF에 유전 알고리즘(Genetic Algorithm, GA) 최적화를 적용한 형태의 학습모델을 활용하였다.

인공신경망(Artificial Neural Network, ANN), 합성곱신경망(Convolutional Neural Network, CNN) 등의 딥러닝 기법도 정수 처리 공정과 관련하여 활용되기도 한다. Arismendy et al.[2]는 산업의 기하급수적인 성장으로 인한 폐수 배출의 환경 영향 완화를 위해 ANN, SVR, RF와 선형회귀분석(Linear regression) 등 다양한 모델을 활용하여 화학적 산소 요구량(Chemical Oxygen Demand, COD)을 예측하였다. 학습 변인으로 유량, 부유물질, 질소, pH 등을 사용하였으며 예측값을 바탕으로 인터페이스를 구축하여 폐수처리장 운영과 관련된 의사결정을 지원한다. Egbueri[3]에서는 최근 개발도상국의 산업화 속도가 지속적으로 증가함에 따라 지속적인 산업 수질 평가의 필요성을 바탕으로 수자원의 부식 가능성을 예측하였다. 염화물-황산염 질량 비율(CSMR), 라슨-스콜드 지수(LSI), 랑겔리에 지수(LI), 리즈나 안정성 지수(RSI) 등을 학습 변인으로, ANN, 다중회귀분석(Multiple Regression, MR)을 학습 모델로 삼아 예측을 진행하였으며 높은 예측 성능을 보였다.

Lee et al.[6]에서는 낙동강 중류 지역의 조류 발생을 모니터링하기 위해 엽록소 수를 예측하였다. 수질자료인 pH, 용존 산소, 생화학적 산소요구량, 화학적 산소요구량, 부유물질, 총질소, 총인, 총유기탄소, 수온, 전기전도도, 용존 총질소, 암모니아성질소, 질산성질소, 용존 총인, 인산염인과 수량자료인 보 하류 수위, 저수량, 공용량, 유입량, 총 방류량 등을 학습 변인으로 활용하였다. 낙동강 중부 내 2개의 보 지점에서 학습 변인과 예측값인 클로포필-a(Chl-a)와의 상관성을 통해 10가지 중요 인자를 추출하고, 의사결정나무, RF, Elastic-Net, Gradient Boosting모델을 활용 및 예측하여 수질 모니터링에 일조하였다.

Park et al.[7]에서는 응집제 투입률 결정을 위해서 Jar-test 데이터를 활용하여 원수 탁도를 바탕으로 목표 침전치 탁도를 위한 최소 응집제 투입률을 결정하는 강화학습을 만들었다. Kim et al.[4]에서는 정수장 운영 근무자의 패턴을 학습하여 K-means와 Gradient Boosting Regression(GBR)을 바탕으로 군집 및 분류 모델을 만들어 약품투입률을 결정하려 하였다.

3. 정수처리 프로세스 및 데이터 수집

3.1 정수처리 프로세스

본 연구에서는 울산광역시에 위치하고 있는 한국수자원공사 온산정수장에 대해 다룬다. 온산정수장은 낙동강 하구의 원동 취수장으로부터 직선거리 24km의 관망을 통하여 원수(raw water)를 공급받아 온산공업단지의 140여 고객(화학 및 정유 공장 등)의 공업용수를 위해 사용된다. 이때 원동 취수장에서 온산 정수장까지 도달하는 데에는 주변 관망의 상황과 여건에 따라 달라질 수 있지만 대략 4.5시간 정도의 시간이 소요된다. 정수장에 도착한 원수는 소독 공정, 화학작용제 혼화, 응집침전, 배수 및 공급 등의 과정을 거치게 된다. 이러한 정수처리 과정에서는 급속혼화기, 완속응집기 등을 이용하여 원수의 각종 유기물, 미생물, 현탁 물질 등이 응집되어 생성되는 플록(Floc)의 형성을 돕고 이러한 플록이 크고 무거워지면 침전지(Sedimentation basin)에서 제거되며, 플록이 제거된 용수는 배수지(Distribution reservoir)를 통해 고객에게 제공된다. 용수들의 침전지에서의 체류시간은 약 5시간이며, 배수지에서는 약 2.5시간 정도 체류 후 고객들에게 제공된다.

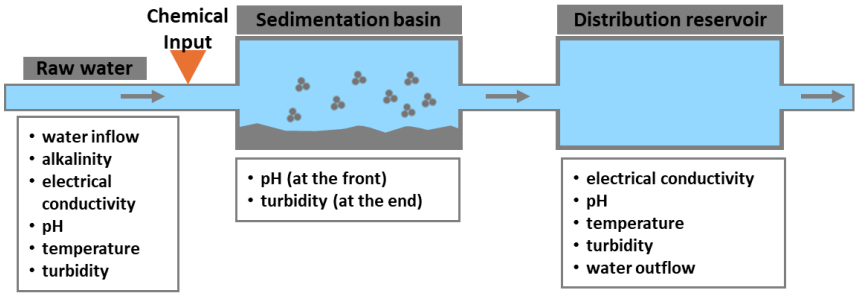
온산정수장은 이러한 응집 및 플록 형성을 위해 $PACS_2$라는 응집제를 사용하고 있으며, 근무자의 판단에 의해 그 투입률이 결정된다. 이때, 근무자는 현재 유입되는 원수의 상태, 침전지 및 배수지의 상황 등에 실시간적으로 대응하기 위해, 미리 통제된 환경에서 수행된 Jar-test에 의해 도출된 조견표를 참고하며 그 투입률을 결정하게 된다. 하지만 이러한 조견표는 실제 정수처리 장과는 다른 실험실의 통제된 환경에서 도출된 것이므로 정수처리장에 완전히 맞지 않을 수 있으며, 이를 보정하기 위해 근무자의 주관적 판단에 의해 실제 약품투입률이 결정되고 있는 실정이다. 본 연구에서는 기존의 상황에 따른 응집제 투입률을 학습하여 머신러닝 기반의 응집제 투입률 결정 모델을 개발하고자 한다.

3.2 데이터 구성 및 특징

온산정수장에서는 <Figure 1>과 같이 원수, 침전, 배수지 등에서 다양한 데이터를 수집하여 활용하고 있다. 온산정수장에 유입되는 원수에서는 원수유입량(Inflow), 알칼리도(Alkalinity), 전기전도도(Electrical conductivity), pH, 수온, 탁도(Turbidity)를 측정하게 되며, 이러한 원수는 투입된 응집제와 혼화되어 침전지에 유입된다. 원수가 침전지에 유입된 직후에 다시 한 번 pH를 측정하며, 침전지를 통과하면서 플록이 형성 및 침전되며 정수처리가 진행된다. 침전지의 말단에는 부유물질, 바이러스, 세균, 콜로이드성 물질 등이 대다수 제거된 후 수질을 체크하기 위해 다시 한 번 탁도를 측정한다. 이후 정수처리된 용수들이 배수지에 모이게 되며, 유출량, 전기전도도, pH, 수온, 탁도를 모니터링한다.

이러한 데이터들은 대부분 수질에 직간접적인 영향을 미치거나 응집제의 성능에 영향을 미치므로 응집제 투입률의 결정에 중요한 요인이 된다. 주요 지표들에 대한 설명은 아래와 같다.

- 알칼리도: 물 속의 염기성 물질 농도를 측정하는 지표로써 응집에 영향을 미친다. 높은 알칼리도는 응집 및 침전제의 성능을 감소시킬 수 있으며, 낮은 알칼리도는 일부 물질의 침전을 유발할 수도 있다.
- 전기전도도: 물이 전기를 얼마나 잘 전도하는지를 나타내는 측정 항목으로, 물 속의 용해 물질의 양과 종류에 영향을 받는다. 따라서 일반적으로 높은 전기전도도는 물 속에 용해된 물질의 양이 많음을 나타내며, 이는 낮은 수질을 의미할 수도 있다.
- pH: 물의 산성 또는 염기성 정도를 측정하는 지표로, 응집 및 침전 프로세스, 물질의 용해도, 미생물 활동 등에 영향을 미친다.
- 수온: 응집 및 침전 프로세스에 직접적인 영향을 미친다. 일반적으로 낮은 온도는 응집 및 침전의 속도를 저하시킬 수 있다.
- 탁도: 물의 탁함 정도를 나타낸다.



<Figure 1> Flow of the Water and Data Collection at the Water Treatment Plant

<Table 1> Feature Information for the Model

Location	Data	Number of data	Number of missing data	2022.09 ~ 2023.08		
				range	average	standard deviation
Raw water	water inflow	17516	4	6487.90~415299328.33	34356.09	3137495.07
	alkalinity	17518	2	0.00~72.32	50.43	12.39
	electrical conductivity	17517	3	101.00~586.00	329.79	125.83
	pH	17517	3	6.74~9.00	7.79	0.36
	temperature	17517	3	0.01~29.37	14.91	8.05
	turbidity	17517	3	0.84~181.84	6.37	12.88
coagulant rate		17514	6	0.00~15.94	4.76	2.91
Sedimentation basin	pH (at the front)	17516	4	6.62~8.19	7.53	0.37
	turbidity (at the end)	17516	4	0.20~2.67	0.66	0.21
Distribution reservoir	electrical conductivity	17516	4	113.53~571.81	345.44	125.15
	pH	17516	4	6.78~8.15	7.64	0.34
	temperature	17516	4	4.08~29.99	17.36	7.88
	turbidity	17516	4	0.00~2.42	0.59	0.22
	water outflow	17516	4	3761.09~15000	11154.57	670

본 연구에서는 2022년 9월 1일부터 2023년 8월 31일까지 1년간 수집된 자료를 사용하였으며, 30분 간격으로 데이터를 추출하였다. <Table 1>에서는 수집된 데이터 및 특성을 요약하고 있다.

<Table 1>에서 제공하는 범위, 평균, 표준편차는 제4장에서 수행한 결측치 및 이상치 처리 후의 데이터를 분석한 결과이다. 대부분의 지표들이 큰 폭으로 변화하고 있는 것을 알 수 있다. 또한 배수지에서 수집한 데이터들은 이미 응집제 투입 이후에 수집된 데다 침전지의 데이터와 유사하였으므로 배수지의 전기전도도, pH, 수온, 탁도, 유출량은 분석에서 제외하였다.

4. 결측치 및 이상치 처리 프로세스

4.1 결측치 처리

<Table 1>에서 알 수 있듯이, 일부 데이터 컬럼에서는 측정 기기 오류 등으로 인한 결측치가 발생하였다. 본 데이터는 30분 간격의 시계열 데이터이므로 선형 보간법(Linear interpolation)을 통해 결측치를 보완하였다. 예를 들어 10시 정각의 값이 100이고 11시 30분의 값이 250이라면, 10시 30분의 값은 150, 11시 정각의 값은 200으로 처리하였다.

4.2 이상치 처리

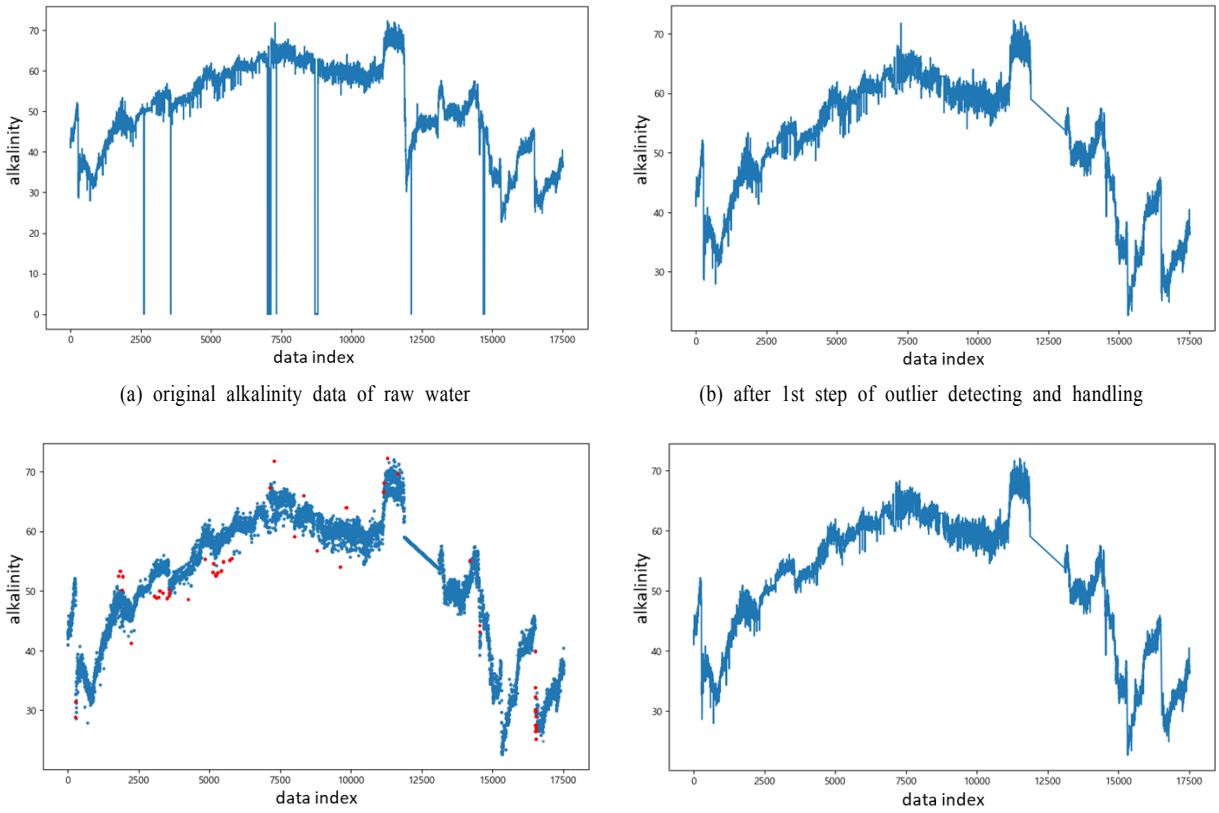
전문가와의 토의를 통해 수집된 데이터는 측정 기기의 오류 등으로 인해 많은 이상치를 포함하고 있음을 확인하였으며, 응집제 투입률 결정 모델 생성에 앞서 이상치 처리를 위한 프로세스를 정립하였다. 다양한 이상치 제거 프로세스를 적용하고 결과를 비교한 결과, 본 연구에서는 이상치 처리를 위해 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)을 기반으로 크게 두 단계로 구성하였다. 이러한 이상치 처리 프로세스는 각 데이터 컬럼별로 수행하였다.

(단계 1) 일별 데이터를 기반으로 한 월별 이상치 처리:

일차적으로 월별 데이터에 대해 표준화(Standardzation)만을 적용한 이후에 DBSCAN을 통해 이상치를 식별하였다. 이때, 코어를 형성하기 위한 최소 데이터 수(minPt)는 2, 6, 24, 48로 변화시켜보며 이상치 식별 결과를 확인하였으며, 이 중 보수적으로 이상치를 식별하기 위해 minPt = 48로 설정하여 이상치를 식별하였다. 참고로 48개의 데이터는 하루 동안 수집된 데이터의 양과 동일하다. 사용된 코어중심반경은 $\varepsilon = 0.5$ 로 설정하였다. 식별된 이상치는 앞서 결측치 처리와 동일하게 선형 보간법을 이용하여 값을 대체하였다.

(단계 2) 추세 및 잔차를 기반으로 한 이상치 처리:

본 연구의 데이터들은 시계열 데이터이므로 이상치 처리를 위해 데이터 값의 추세 또한 고려하였다. 동일한 값도 근처 데이터 값 및 추세에 따라 정상치일 수도, 이상치일 수도 있기 때문이다. 이를 위해 이동평균값을 활용하여 추세를 파악하고 이동평균값과의 차이를 잔차로 정의하였다. 이러한 잔차가 큰 값은 추세를 따르지 않는 이상치 데이터라고 판단하였으며 이를 구분하기 위해



(a) original alkalinity data of raw water

(b) after 1st step of outlier detecting and handling

(c) after 2nd step of outlier detecting and handling (red dots: detected outliers in this step)

(d) final alkalinity data

<Figure 2> Detecting and treating outliers

DBSCAN(minPt = 2, ε = 0.5)을 적용하였다. 또한, 직전 6시간, 24시간, 72시간의 데이터를 활용하여 이러한 잔차를 계산하고 각각의 경우에 대해 DBSCAN을 적용하여 이상치를 파악하였다. 즉, 동일한 데이터 컬럼에 대해 DBSCAN을 세 번 적용하고 이 중 한 번이라도 이상치로 판단되는 경우에는 최종적으로 이상치로 판단하고 선형 보간법을 적용하였다.

<Figure 2>는 이러한 이상치 처리 프로세스를 원수의 알칼리도 데이터에 적용한 예시를 보여준다.

5. 응집제 투입률 예측

5.1 상관분석

먼저 각 데이터 간의 상관계수는 아래 <Table 2>와 같다. 하지만 일부 데이터 컬럼 간의 상관분석 결과는 앞선 3.2절의 설명과 다르다. 예를 들어, 알칼리도가 높거나 전기전도도가 높으면 약품투입률은 높아져야 한다. 하지만 원수의 알칼리도와 전기전도도의 약품 투입률의 상관계수는 각각 -0.724와 -0.802로 오히려 설명과 반대의 결과이며, 이는 영향력이 다른 여러 변수들이 복합적으로 약품투입률에 영향을 미치기 때문으로 판단된다. 즉, 다수의 데이터에서 전기전도도는 높지만 탁도가 낮아 약품투입률이 적었다면 전기전도도와 약품투입률의 상관계수는 앞의 설명과는 반대로 음수가 나올 수도 있다.

5.2 데이터 학습 모델

원수의 유입량, 알칼리도, 전기전도도, pH, 수온, 탁도, 침전지에서의 pH 및 탁도를 바탕으로 실제 약품투입률을 학습하였다. 2022년 9월 1일부터 2023년 8월 31일까지 1년간의 데이터 중 10달치의 데이터를 학습에 사용하였으며 2달치의 데이터는 테스트를 위해 사용하였다. 또한 각 계절마다 데이터의 특성에 다소 차이가 있어 테스트 데이터는 각 계절별로 2주치 데이터를 사용하였다. 즉, 테스트 데이터는 2022년 10월 8일~2022년 10월 21일, 2023년 1월 8일~2023년 1월 21일, 2023년 4월 8일~2023년 4월 21일, 2023년 7월 8일~2023년 7월 21일이다.

학습 모형으로는 다중선형회귀모델과 대표적인 머신러닝 기반 모델들인 랜덤포레스트(Random forest), XGBoost, Stochastic Gradient Descent(SGD), Support Vector Machine

<Table 2> Correlation Analysis Results

	raw water						Sedimentation basin		Distribution reservoir
	water inflow	alkalinity	electrical conductivity	pH	temperature	turbidity	pH	turbidity	water outflow
coagulant rate	-0.012	-0.724	-0.802	-0.641	0.726	0.637	-0.845	0.442	0.046

<Table 3> Validation Results

		Linear Regression	Random Forest	XGBoost	SGD	SVM	DNN
Data without scaling	MSE	0.865	0.074	0.099	3.32E+23	13.411	0.086
	MAPE	0.376	0.259	0.26	3.84E+21	1.697	0.288
	R^2	0.885	0.99	0.987	-4.35E+43	-0.789	0.961
	CVRMSE	20.274	5.94	6.861	2.51E+23	79.838	11.842
Data with log transformation	MSE	0.881	0.074	0.169	4.903	4.326	1.634
	MAPE	0.243	0.034	0.065	0.558	0.396	13.689
	R^2	0.882	0.99	0.977	0.344	0.423	-75.337
	CVRMSE	26.161	7.553	11.472	61.257	57.852	229.34
Data with min-max scaling	MSE	0.865	0.072	0.181	0.875	1.131	14.943
	MAPE	0.376	0.246	0.301	0.36	0.467	24.364
	R^2	0.885	0.99	0.976	0.883	0.849	-830.724
	CVRMSE	20.274	5.833	9.265	20.394	23.011	1724.851
Data with standardization scaling	MSE	0.865	0.074	0.114	0.909	1.04	18.563
	MAPE	0.376	0.254	0.269	0.399	0.448	23.599
	R^2	0.885	0.99	0.985	0.879	0.861	-23.08
	CVRMSE	20.274	5.909	7.362	20.788	22.161	-24663.807

(SVM), Deep Neural Network(DNN) 모델들을 적용하였으며, 이상치 처리 직후 데이터 외에도 로그 변환, min-max 정규화, 표준화(Standardization)를 통한 전처리도 활용하여 가장 적합한 모델을 찾으려 하였다.

모델의 성능 평가를 위해서 과적합 문제를 줄이기 위해 일반적으로 사용되는 5-fold 교차검증을 실시하였다. 즉, 랜덤하게 5개의 fold를 구성하였으며, 5번 수행한 검증 결과의 평균값은 <Table 3>에 정리되어 있다. 약품투입률은 최소 0.0008에서 최대 15.9434로 큰 변동성을 가지므로 일반적으로 사용되는 MSE, MAPE, R^2 외에도 CVRMSE (Coefficient of Variation of the Root Mean Square Error)를 참고하여 모델의 성능을 평가하였으며, 최종적으로 min-max 정규화를 적용한 랜덤포레스트 모델이 가장 좋은 모델로 평가되었다. 참고로 로그 변환을 제외한 대부분의 모델에서 MAPE는 MSE 값에 비해 매우 높으며, 이는 절대오차가 상대적으로 작음에도 불구하고 실제 응집제 투입률이 매우 작은 경우들이 자주 있어 발생하는 결과이다.

5.3 예측 결과

최종적으로 이상치 처리 프로세스, min-max 정규화 및 랜덤포레스트를 적용하여 각 계절별로 2주치 데이터에 대해 테스트한 결과는 <Table 4>와 같다. 또한 이때 사용된 결정트리의 개수는 10개이며 과적합을 방지하기 위해 노드를 분할하기 위한 최소한의 데이터 수는 5로 지정하였다. 검증 결과와 비교하여 성능이 다소 하락하였지만, 1.135의 MSE, 0.912의 R^2, 18.704의 CVRMSE에서 알 수 있듯이 여전히 우수한 성능을 보여주었다.

<Table 4> Test Results

	Random Forest with min-max scaling
MSE	1.136
MAPE	0.111
R^2	0.912
CVRMSE	18.704

6. 결 론

본 연구에서는 온산 정수처리장에서 2022년 9월 1일부터 2023년 8월 31일까지 1년간 실제로 수집된 원수 유입량, 알칼리도, 전기전도도, pH, 수온, 탁도, 침전지에서

의 pH 및 탁도, 근무자가 실제로 투입한 응집제 투입률을 바탕으로, 정수처리를 위해 투입하는 응집제의 양을 결정하기 위한 머신러닝 기반의 모델을 개발하였다. 다양한 종류의 대표적인 머신러닝 기반의 모델 및 선형회귀 모델을 테스트한 결과, min-max 정규화를 적용한 랜덤포레스트 모델이 가장 뛰어난 성능을 보여주었으며, 테스트 결과 MSE는 1.136, R^2는 0.912, CVRMSE는 18.704로 우수한 성능을 보여주었다.

또한 머신러닝 모델 개발에 앞서, 정규화 변환 및 DBSCAN을 활용하여 두 단계로 이루어진 이상치 처리 프로세스를 개발하였다. 즉, 이상치 탐지를 위해 단순히 월별로 DBSCAN을 적용하는 것 외에도 추세 및 잔차 계산을 통한 이상치 탐지를 적용하였다. 이상치 및 결측치는 선형 보간법을 통해 보정되었다.

본 연구의 한계점으로는 응집제의 투입률을 결정하기 위해 기존 근무자의 패턴을 학습했다는 점을 꼽을 수 있다. 기존 근무자의 투입률이 최적의 투입률을 보장하지는 않기 때문에 본 연구에서 제시하는 모델 또한 최적의 응집제 투입률을 결정하는 모델은 아니다. 그럼에도 불구하고 기존의 근무자에 의한 운영과 유사한 수준의 정수장 관리가 가능하며 Jar-test 외의 참고자료를 제시할 수 있다는 점에서 의미가 있다. 이를 통해 깨끗하고 안전한 정수 처리장의 스마트한 관리 및 대응 시스템에도 도움이 될 수 있다. 더 장기간에 거친 연구 및 테스트, 더 다양한 모델 적용, 파라미터 개선을 통한 모델 개선 등은 향후 연구가 될 수 있다.

Acknowledgement

This work was supported by Yonsei Business Research Institute and National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2022R1C1C101173111).

References

[1] Achite, M., Samadianfard, S., Elshaboury, N., and Sharafi, M., Modeling and optimization of coagulant dosage in water treatment plants using hybridized random forest model with genetic algorithm optimization, *Environment, Development and Sustainability*, Vol. 25, No. 10, 2023, pp. 11189-11207.

[2] Arismendy, L., Cárdenas, C., Gómez, D., Maturana, A., Mejía, R., and Quintero, M.C.G., Intelligent system for the predictive analysis of an industrial wastewater treatment process, *Sustainability*, Vol.12, No.16, 2020, p. 6348.

[3] Egbueri, J.C., Predicting and analysing the quality of water resources for industrial purposes using integrated data-intelligent algorithms, *Groundwater for Sustainable Development*, Vol. 18, 2022, p. 100794.

[4] Kim, J., Kang, B., and Jung, H., Determination of coagulant input rate in water purification plant using K-means algorithm and GBR algorithm, *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 25, No. 6, 2021, pp. 792-798.

[6] Lee, S., Park, K., and Kim, I., Comparison of machine learning algorithms for Chl-a prediction in the middle of Nakdong River (focusing on water quality and quantity factors), *Journal of the Korean Society of Water and Wastewater*, Vol. 34, No. 4, 2020, pp. 277-288.

[5] Park, J.J., Kim, J.W., Jang, S.Y., and Lee, S.Y., Machine Learning and Genetic Algorithm Integration to Optimize Paraffin Coating Uniformity, *Transactions of the Korean Society of Mechanical Engineers - A*, Vol. 48, No. 6, 2024, pp. 397-403.

[7] Park, J., Heo, H., Seo, J., Kim, T., Sim, M.K., and Kang, M., Optimization of Coagulant Dosage Rate using Reinforcement Learning in Water Purification Plants, *Spring Annual Conference of IEIE*, June, Republic of Korea, 2022.

[8] Seong, J.M. and Yoon, B.J., Analysis of the Impact Factors of Peak and Non-peak Time Accident Severity Using XGBoost, *Journal of the Society of Disaster Information*, Vol. 20, No. 2, 2024, pp. 440-447.

[9] Sim, D., Lee, J., Jang, J., and Lee, M., Prediction of chloride concentration in groundwater on Jeju Island using XGBoost regression machine learning, *Journal of the Geological Society of Korea*, Vol. 58, No. 2, 2022, pp. 243-255.

ORCID

Kyungsu Park | http://orcid.org/0000-0002-5386-5222
Yu-jin Lee | http://orcid.org/0000-0002-7584-9397
Haneul Noh | http://orcid.org/0009-0001-3164-4867
Jun Heo | http://orcid.org/0009-0003-5031-4040
Seung Hwan Jung | http://orcid.org/0000-0002-3044-3879