2025년도 공공기관 용역과제

AI개발 수행내역서

과제명	AI 기반 정수장 수질(pH) 예측모델 개발 및 음용 판단 여부 시스템 구축
담당자	류건우

2025년 12 월 28 일

AI개발 수행내용

1. 사업과제 : AI 기반 정수장 수질(pH) 예측모델 개발 및 시각화 시스템 구축

2. 개요 및 현황

2.1 추진배경 및 목적

○ 최근 기후 변화와 원수 수질 변동성 증가로 인해 정수장 운영의 불확실성이 확대되고 있음

○ 기존 정수장 운영은 기준치 초과 여부에 대한 사후 대응 중심으로 이루어지고 있어, 약품 사용의 비효율이 발생

○ 공공 수질 데이터(경도, 탁도, 잔류염소 등)를 활용한 AI 기반 예측 모델을 통해 pH 변화를 사전에 예측하고 약품 운전 부담 및 음용 안전성을 선제적으로 판단할 필요가 있음

2.2 과제 범위

과제구분		내용
AI	AI기반 수질예측모델 구현	공공 정수장 수질 데이터 수집 및 DB 구축
		수질 항목 전처리 및 결측치 처리
		상관관계 분석 및 주요 변수 선정
		RandomForest 기반 PH 예측 모델 학습
		RMSE, MAE, R^2 기반 모델 성능 평가
		예측모델 웹기반 시스템 구축
		테스트
시각화	수질 입력 기반 pH 시각화	수질 입력 기반 pH 예측 시연 화면 구성
		연·월·지역별 pH 예측 결과 시각화
		예측 결과에 대한 오차 및 잔차 분석
		테스트
		통합테스트 및 시운전

2.3 과제 추진 방법

1) 구축 대상 선정 기준

○ 데이터 접근성 및 활용성

■ 공공데이터 포털을 통해 제공되는 정수장 수질 데이터를 활용하여 데이터 수집 및 관리 용이

■ 이미 구축된 api를 데이터베이스로 활용

■ pH, 탁도, 잔류염소 등 정수 처리 공정에 직접적인 영향을 미치는 핵심 수질 변수가 포함되어 있어 예측 모델 학습에 적합함

○ 예측모델 개발 효율성

■ 수질 데이터는 연·월 단위로 정리되어 있어 시계열적 패턴 학습 및 예측이 가능

■ 수질 항목 간 상관관계 분석을 통해 모델 입력 변수 축소 및 해석 가능성 확보

○ 수질 관리 기여도 및 유지 비용 절감

■ pH 예측 결과를 활용하여 약품 투입 여부 및 운전 조건 판단 가능

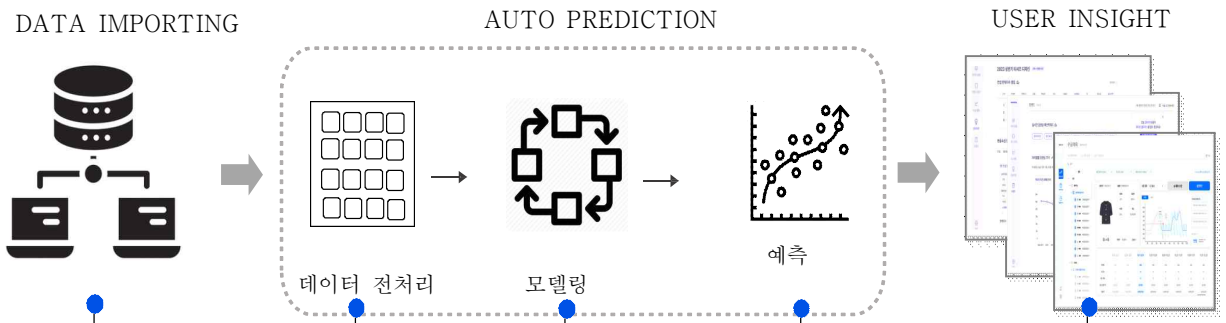
■ 불필요한 약품 사용 감소를 통해 운영 비용 절감 효과 기대

■ 수질 이상 징후를 사전에 파악함으로써 음용 안전성 확보 및 사회적 비용 감소에 기여

2) AI 예측 분석모델 적용 대상

	수집 데이터	예측모델인자(독립변수)	AI예측 분석 대상
정수장 수질 관리	- 정수장 월별 수질 측정 데이터 - 지역별 정수장 운영 데이터	- 경도(HR) - 증발잔류물(RE) - 질산성질소(MON) - 브롬산염(BRO) - 알루미늄(AL) - 클로로포름(CF) - 황산이온(SO), 탁도(TU), 잔류염소(RC)	- pH 예측 - 약품 투입 필요 여부 판단 - 음용 가능 여부 판단 - 지역·월별 수질 상태 비교 분석

3) AI 분석모델 구축 프로세스



정수장 월별 데이터 수집, 지역별 정수장 수질 항목 데이터 확보, 데이터 구조로 정리하여 시계열 분석 기반 마련

수질 항목의 결측지 처리 및 이상치 제거, 문자형 수질 값을 수치형 데이터 변환, 모델 학습을 위한 학습 데이터 / 테스트 데이터 분리

RandomForest Regressor 기반 pH 예측 모델 설계, 하이퍼파라미터 튜닝을 통한 모델 성능 개선

지역·월별 pH 자동 예측, 예측된 pH 값과 탁도, 잔류염소를 결합해 음용 여부, 약품 투입 필요 여부 판단

지역별 수질 상태 비교 및 위험 지역 식별, 약품 운전 의사결정 지원, 음요 가능 여부 확인, 미래 시점에 대한 수질 예측

연구개발 주요 결과물

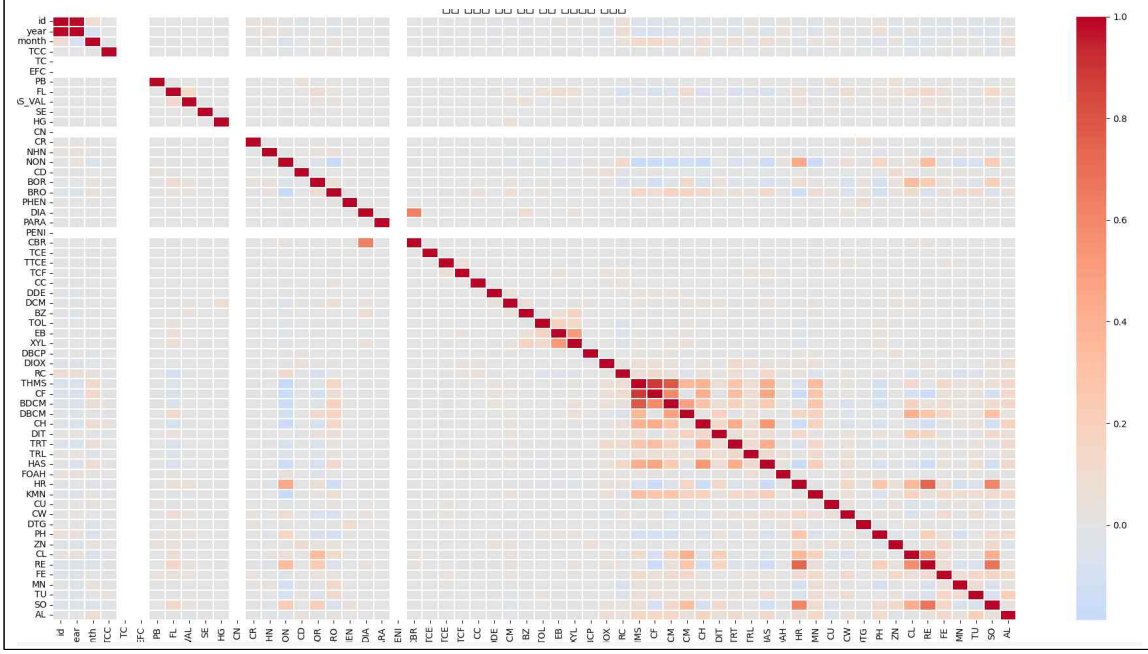
1. 데이터 수집

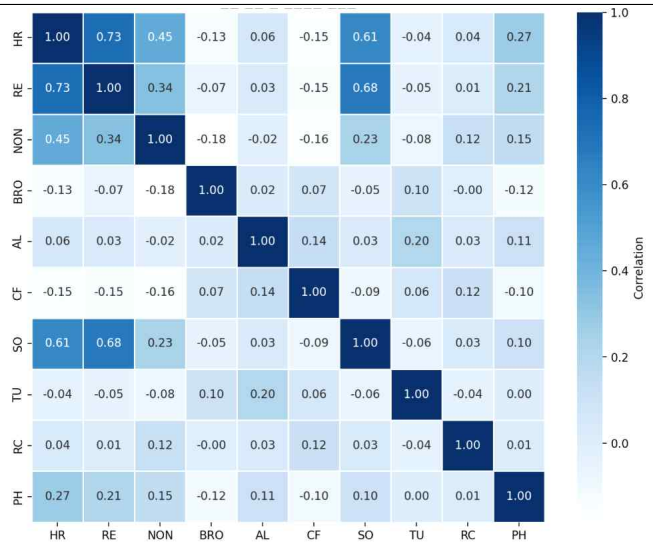
- 한국수자원공사 5년간 상수도법정수질정보(api) : 2020년 ~ 2025년

이름	타입	스키마
water_quality		CREATE TABLE water_quality (id INTEGER PRIMARY KEY AUTOINCREMENT, year REAL,month REAL,region TEXT,facility TEXT,TCC TEXT,TC TEXT,EFC TEXT,PB REAL
id	INTEGER	"id" INTEGER
year	REAL	"year" REAL
month	REAL	"month" REAL
region	TEXT	"region" TEXT
facility	TEXT	"facility" TEXT
TCC	TEXT	"TCC" TEXT
TC	TEXT	"TC" TEXT
EFC	TEXT	"EFC" TEXT
PB	REAL	"PB" REAL
FL	REAL	"FL" REAL
AS_VAL	REAL	"AS_VAL" REAL
SE	REAL	"SE" REAL
HG	REAL	"HG" REAL
CN	REAL	"CN" REAL
CR	REAL	"CR" REAL
NHN	REAL	"NHN" REAL
NON	REAL	"NON" REAL
CD	REAL	"CD" REAL
BOR	REAL	"BOR" REAL
BRO	REAL	"BRO" REAL
PHEN	REAL	"PHEN" REAL
DIA	REAL	"DIA" REAL
PARA	REAL	"PARA" REAL
PENI	REAL	"PENI" REAL
CBR	REAL	"CBR" REAL
TCE	REAL	"TCE" REAL
TTCE	REAL	"TTCE" REAL
TCF	REAL	"TCF" REAL
CC	REAL	"CC" REAL
DDE	REAL	"DDE" REAL
DCM	REAL	"DCM" REAL
BZ	REAL	"BZ" REAL
TOL	REAL	"TOL" REAL
EB	REAL	"EB" REAL
XYL	REAL	"XYL" REAL
DBCP	REAL	"DBCP" REAL
DIOX	REAL	"DIOX" REAL
RC	REAL	"RC" REAL
THMS	REAL	"THMS" REAL
CF	REAL	"CF" REAL
BDCM	REAL	"BDCM" REAL
DBCM	REAL	"DBCM" REAL
CH	REAL	"CH" REAL
DIT	REAL	"DIT" REAL
TRT	REAL	"TRT" REAL
TRL	REAL	"TRL" REAL
HAS	REAL	"HAS" REAL
FOAH	REAL	"FOAH" REAL

1. 데이터 분석

2.1 수질데이터 상관관계(Heatmap)





○ EDA 히스토그램 및 히트맵 분석을 통해 수질 변수별 분포 특성, 정규성 여부를 확인하고 데이터 변환 및 변수 간 상관관계를 분석함

○ pH와 각 입력 변수 간의 직접적인 상관관계를 분석한 결과, 대부분의 변수는 약한 양의 상관관계를 보임

○ 특히 경도(HR), 증발잔류물(RE), 질산성질소(NON)은 상대적으로 pH와 높은 상관성을 보임

○ 반면, 탁도(TU), 잔류염소(RC)는 pH와 상관성은 낮지만 정수 공정 및 약품 운전 판단에 중요한 운영 변수이므로 모델 입력 변수로 유지.

○ 수질 예측 모델링을 위한 대상 설정 : 정수장 pH 예측 및 약품 운전 판단

○ 전력량, 약품사용량에 영향을 미치는 요인 분석

■ 수질 변수 : 경도(HR), 증발잔류물(RE), 질산성질소(NON), 황산이온(SO), 탁도(TU), 잔류염소(RC) 등

■ 환경 변수 : 계절별 수질 변화 특성 (월별·연도별 패턴)

id	year	month	region	facility	TCC	TC	EFC	PB	FL	AS_VAL	SE	HG	CN	CR	NHN	NON	CD	BOR	BRO	PHEN	DIA	PARA	PENI	CBR	TCE	TTCE	TCF	CC
필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터
1	2020.0	1.0	서울특별시	강북	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	2020.0	1.0	서울특별시	광암	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	2020.0	1.0	서울특별시	구의	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	2020.0	1.0	서울특별시	뚝도	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	2020.0	1.0	서울특별시	암사	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	2020.0	1.0	서울특별시	영등포	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	2020.0	1.0	부산광역시	덕산	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	0.07	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	2020.0	1.0	부산광역시	명장	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	2020.0	1.0	부산광역시	범어사	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	2020.0	1.0	부산광역시	화명	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	0.06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	2020.0	1.0	대구광역시	가창	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	2020.0	1.0	대구광역시	고산	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	2020.0	1.0	대구광역시	공산	0	불검출	불검출	0.0	0.16	0.0	0.0	0.0	0.0	0.0	0.0	2.3	0.0	0.02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	2020.0	1.0	대구광역시	군위	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	2020.0	1.0	대구광역시	동부	0	불검출	불검출	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.7	0.0	0.02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

2. 데이터 학습 및 모델정의

3.1 모델정의 및 컴파일

○ 모델 정의 : RandomForest 모델

```
rf = RandomForestRegressor(
    n_estimators=300,
    min_samples_leaf=5,
    random_state=RANDOM_STATE,
    n_jobs=-1
)

rf.fit(X_train, y_train)
```

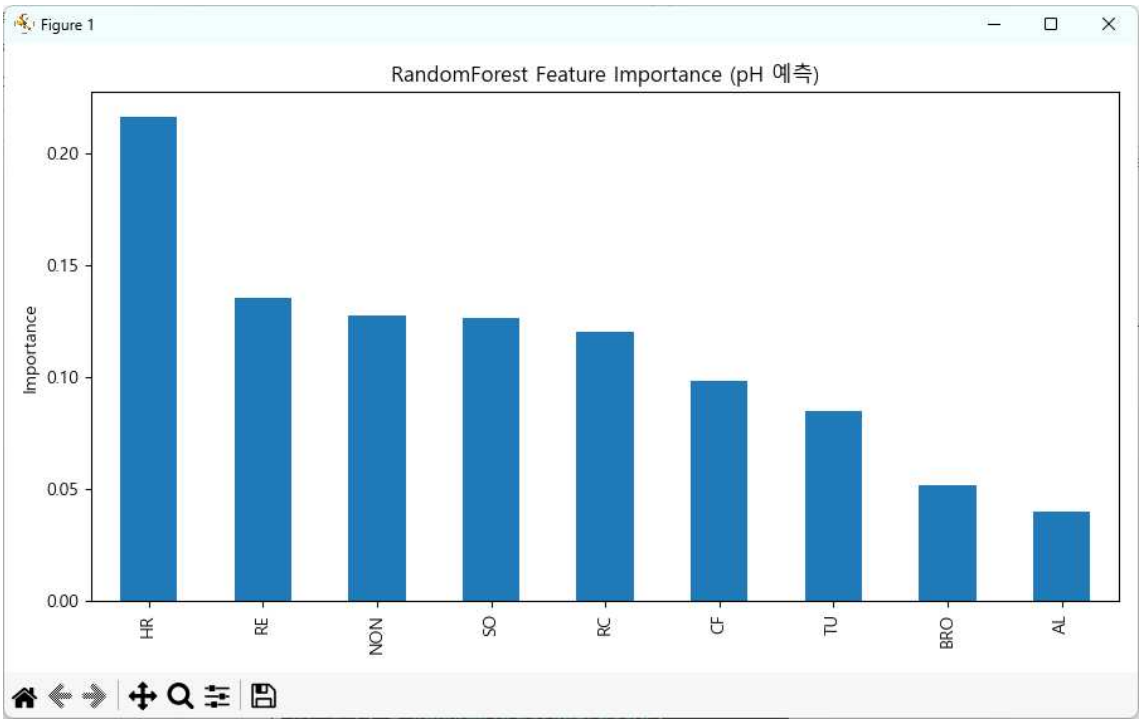
3.2 모델학습 및 학습 시각화

○ 모델 학습 및 성능 지표

```
(.venv) PS C:\Users\kwr51\Desktop\test> & C:/Users/kwr51/Desktop/test/.venv/Scripts/python.exe c:/Users/kwr51/Desktop/test/forestsave.py
모델 학습 데이터 수: 29798

RandomForest 회귀 성능
RMSE : 0.3112
MAE  : 0.2348
R²   : 0.3723
```

○ 공공 수질 데이터에 있는 pH는 미세한 변화만으로도 의미가 있고, 실제 데이터 특성상 대부분의 정수장의 데이터 값은 정상 범주에 속해있기 때문에 R2 값이 낮게 나옴.

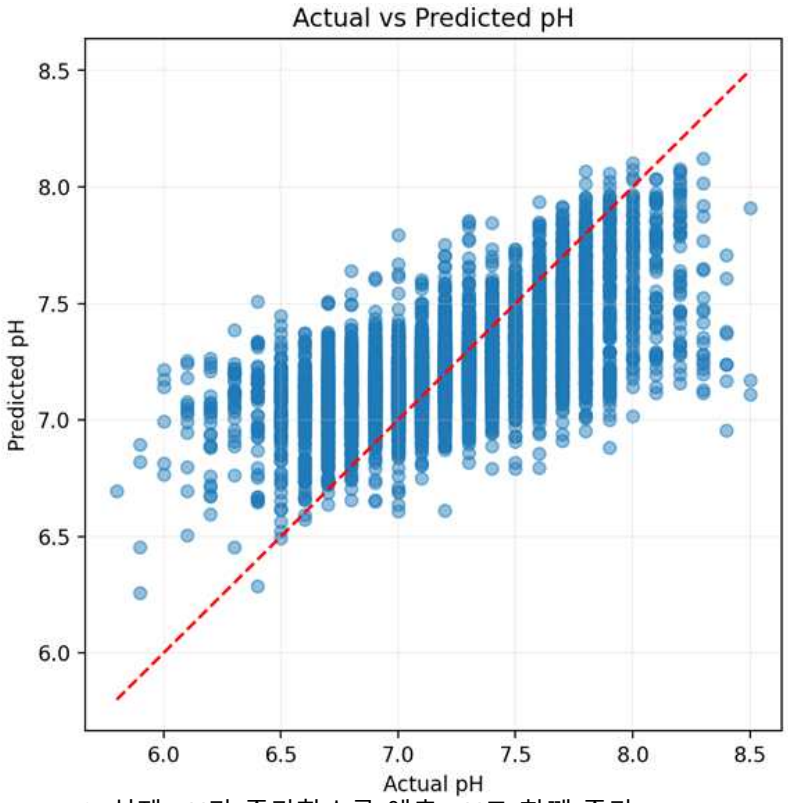


3.3 모델 예측

○ 예측값 vs 실제값 비교

```
# ========================
# 9. 예측 vs 실제 시각화
# ========================
plt.figure(figsize=(6, 6))
plt.scatter(y_test, y_pred, alpha=0.5)
plt.plot([y_test.min(), y_test.max()],
         [y_test.min(), y_test.max()],
         'r--')

plt.xlabel("Actual PH")
plt.ylabel("Predicted PH")
plt.title("Actual vs Predicted PH")
plt.tight_layout()
plt.show()
```

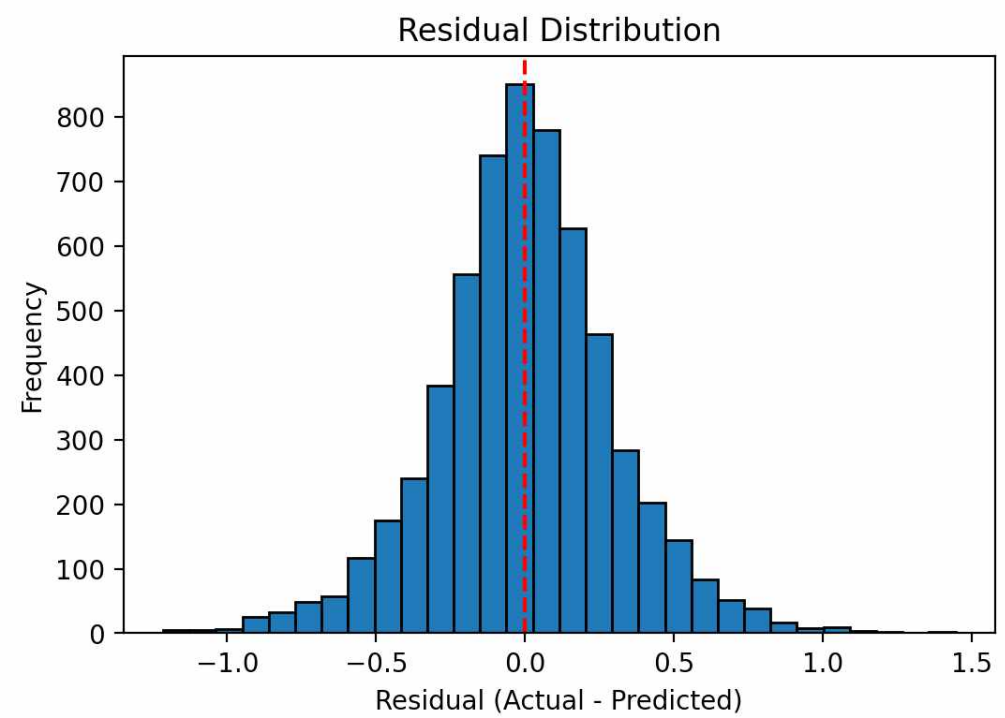


○ 실제 pH가 증가할수록 예측 pH도 함께 증가

pH 값이 7.0 ~ 7.8 구간에 집중.
실제 수질 데이터 대부분이 정상 범위에 분포하기 때문.
그래서 극단값(6.0 이하, 8.5 이상)은 상대적으로 적음.

오차패턴, 일부 점들이 기준선 위 아래로 벗어나 있는 이유는,
자연 환경 요인, 측정 오차, 약품 투입량과 수온, 유량 등의 환경적인 요인

○ Residual Distribution(잔차 분포)



○ 평균 잔차가 0에 가까움, 모델이 특정 방향으로 과대/과소 예측하지 않은 것.

○ 대부분의 잔차가 -0.3 ~ +0.3 구간에 집중
극단값인 1.0 이상은 소수

○ 이는 일시적인 수질 이상, 측정 오차 등 외부요인에 의한 것.

3. 프로토타이핑(화면)

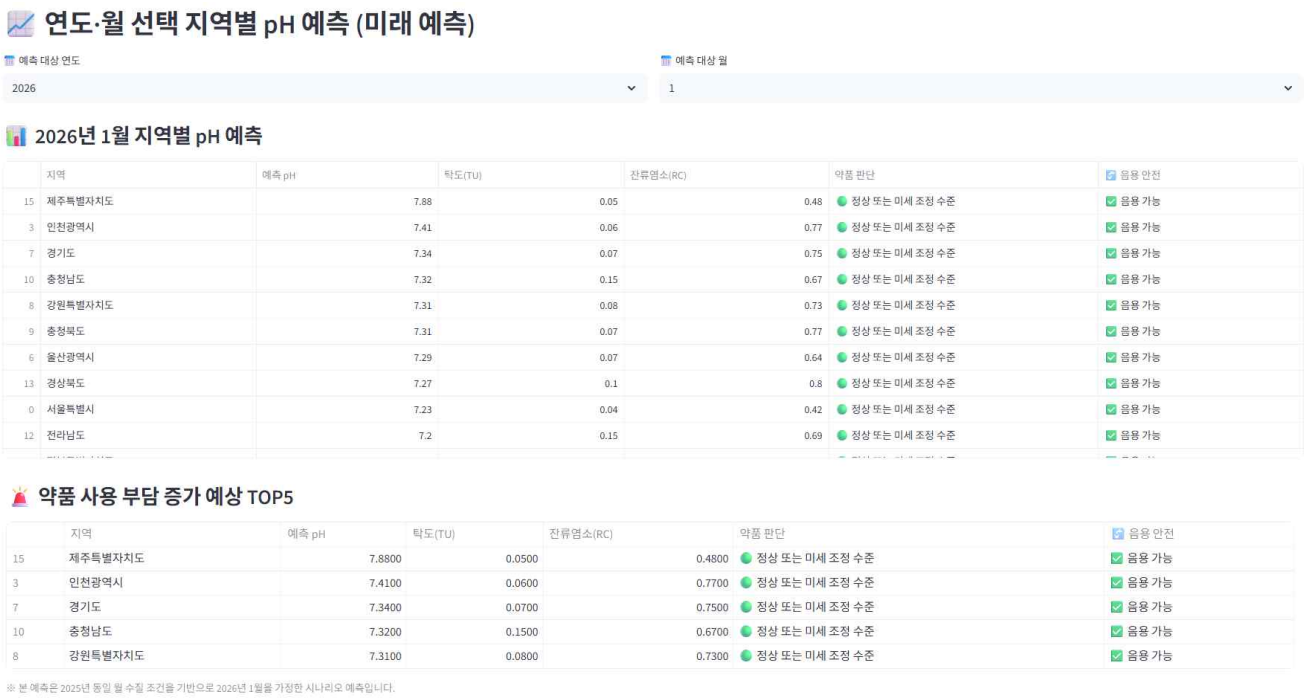
4.1 수질 입력 후 pH 예측

○ 각 요소에 영향을 받은 pH 예측

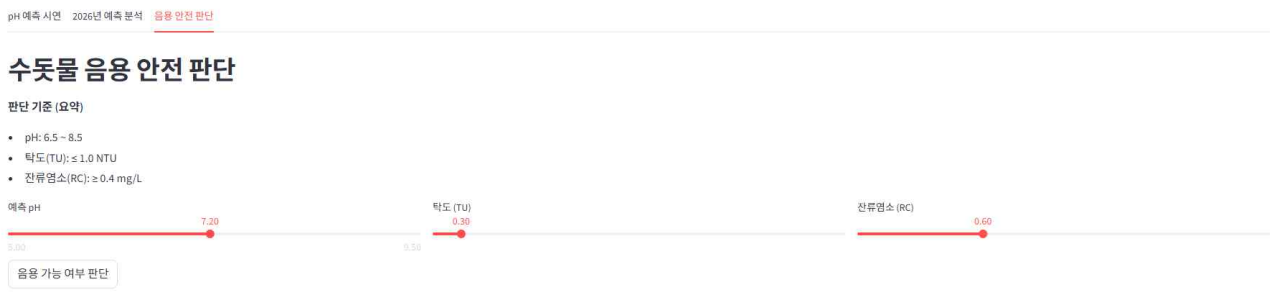
- 경도, 증발잔류물 등 입력

4.2 미래 연도 예측 분석

- 예측 대상 연도, 예측 대상 월 입력



4.3 음용 안전 판단 분석



4. 결론 및 제안

○ pH 예측 성능(R^2 약 0.31, RMSE 약 0.3 수준)을 확보하였으나, 일부 구간에서 오차가 발생하여 추가적인 보완이 필요함

→ 운영 변수(약품 투입량), 환경 변수(계절·강수량 등) 추가 필요

○ 지역·월별 예측 결과에서 pH 값의 변동 폭이 존재함

→ 원수 수질 특성 및 정수장 운영 조건 차이에 따른 영향으로 판단됨

○ Feature Importance 분석 결과, 탁도(TU), 잔류염소(RC) 등 공정 핵심 지표가 pH 예측에 주요한 영향을 미침

○ 본 분석을 통해 약품 사용 부담이 예상되는 지역 및 시점을 사전에 식별 가능

→ 지역·계절별 특성을 반영한 세분화된 데이터셋 구축 필요

→ 향후 약품 사용량 예측 및 정수장 운영 최적화 전략 수립에 활용 가능