2024년도 공공기관 용역과제

AI개발 수행내역서

과제명	AI기반 뇌졸증 예측모델 개발 및 시각화
담당자	우성민

2025년 8 월 1 일

AI개발 수행내용

1. 사업과제 : OO병원 AI기반 뇌졸증 예측모델 개발 및 시각화

2. 개요 및 현황

2.1 추진배경 및 목적

○ 장기간 축적된 데이터베이스를 기반으로하여 인공지능 기반의 예측모델에 대한 수요가 점진적 증대 예상

○ 기상이변으로 인한 환경 변화와 재해 발생가능성이 높아짐에 따라 기후와 환경에 대한 정확한 예측이 중요해지고 있는 상황

○ 환경 데이터를 분석하고 복잡한 패턴을 학습함으로써, 환경오염에 대한 영향을 보다 정밀하게 예측하고자 함

○ 환경오염 대상 중에서 하수를 대상으로 하여 시범적 모델을 구축하여 향후 대기오염, 소각 등의 다른 환경업무로 확대하고자 함

2.2 과제 범위

과제구분		내용
AI	AI기반 수질예측모델 구현	원시 데이터 수집 및 데이터셋 구축
		데이터 전처리, 표준화, 상관관계 분석 (EDA도구 활용)
		예측모델 선정 및 학습
		RMSE, MAE 등 평가지표를 활용한 모델 성능 평가
		웹 API 및 프로토타입 구축
		예측모델 웹기반 시스템 구축
		테스트
시각화	실시간 환경계측정보 연계 및 시각화	환경계측정보 실시간 연계 모형 구현
		상황관리 대시보드 등 시각화 구현 (BI 시각화 도구 활용)
		예측모델 시각화
		테스트
		통합테스트 및 시운전

2.3 과제 추진 방법

1) 구축 대상 선정 기준

○ 데이터 접근성 및 활용성

■ 데이터 수집 및 관리의 용이성

■ 정부 및 공공기관에서 이미 구축된 데이터베이스 활용 여부

■ 종속변수에 영향을 미치는 다양한 독립변수에 대한 정보 포함여부를 통한 모델학습의 유용성

○ 예측모델 개발 효율성

■ 모델 학습 및 평가 과정 간소화를 위한 다른 환경 기초데이터에 비해 변수가 상대적으로 단순한 구조 여부

■ 개발된 모델을 통해 다른 환경기초 데이터에 적용가능 여부

○ 환경문제 해결 기여도 및 경제성

■ 예측모델을 통해 환경관리에 상대적 기여도가 높은 지 여부(ex. 오염도 저감, 에너지 절감 등)

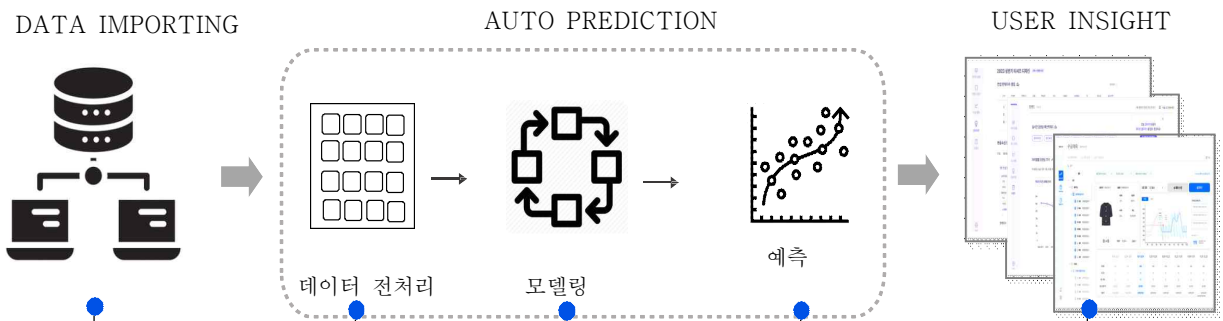
■ 운영 효율성을 높여 비용절감 효과 여부

■ 환경문제 해결을 통한 사회적 비용감소 효과 여부

2) AI 예측 분석모델 적용 대상

환경관리 기능	수집 데이터	예측모델인자(독립변수)	AI예측 분석 대상
하수	- 일별 하수처리 운영 데이터 - 일자, 날씨, 온도 외에 수질 핵심 변수가 포함된 데이터셋 - 생물반응조 전·중·후 수질 데이터	- 수질변수 : pH, BOD,TOC, TN, TP, SS 등 - 운영변수 : 슬러지처리량, 처리시간, 약품사용량 등 - 환경변수 : 강수량, 기온, 계절 등 - 장비변수(하수처리시설) : 송풍기, 슬러지 탈수기	- 전력량(에너지 절감) - 방류수질 예측 - 유입량, 수질기준, 에너지 사용량 등 상관관계 분석 - 법적기준적합여부

3) AI 분석모델 구축 프로세스



일별 하수처리 운영데이터 수집 및 임포팅
일자, 날씨, 온도, PH, BOD, TOC 등 핵심변수 포함

변수별 상관관계 분석
데이터 표준화 및 전처리 (이상치, 결측치 처리)
시계열 데이터를 고려하여 학습, 테스트 데이터 분리

모델 비교 및 최적 모델 도출(ex. LSTM)
시간적 추이/변화에 따른 추론이 가능한 시계열 모델
활용
모델 정의 및 학습

예측값 계산
슬러지 발생량, 유입 및 방류수질 초과 횟수 분석
전력량 예측

기상, 슬러지 발생량 을 통한 공정시 약품처리, 처리시간에 의한 법적적합도 분석
전력량 등 에너지 수요 예측을 통한 에너지 효율성 향상

연구개발 주요 결과물

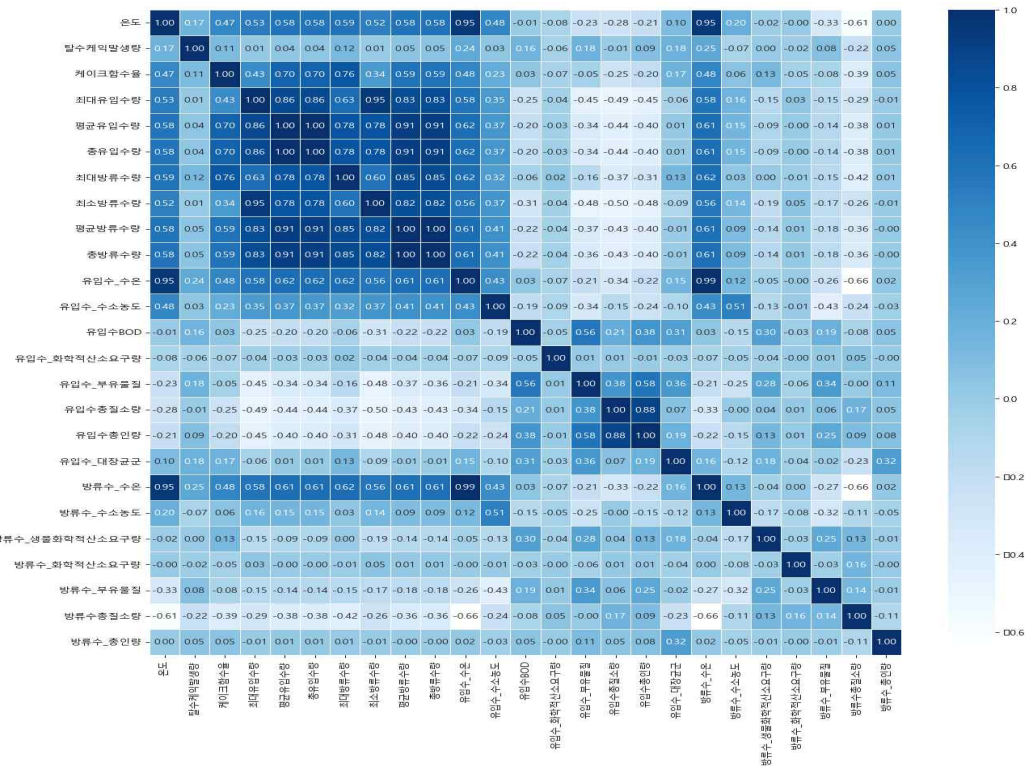
1. 데이터 수집

- OO환경공단 5년간 하수 데이터(엑셀) : 2019년 ~2023년
- OO환경공단 5년간 전력량 데이터(엑셀) : 2019년 ~2023년

일자	온도	탈수케익발생량	케이크함수율	최대유입수량	평균유입수량	총유입수량	최대방류수량	최소방류수량	평균방류수량	총방류수량	산화구A_용존산소량농도A	산화구A_용존산소량농도B	산화구B_용존산소량농도C	산화구B_용존산소량농도
2022-01-01	3.4	0	399	114	285	6845	377	100	277	6652	1.51	7.07	1.09	
2022-01-02	-1.2	0	382	117	299	7171	382	107	285	6842	1.5	8.2	1.07	
2022-01-03	-1.9	0	346	111	272	6523	327	104	263	6303	1.35	6.34	0.95	
2022-01-04	-2.5	0	341	101	275	6600	446	88	250	6008	1.34	6.56	1.04	
2022-01-05	-2.8	0	342	106	267	6409	401	86	255	6131	1.34	6.67	1.05	
2022-01-06	-2.2	0	353	106	295	7078	388	94	268	6432	1.34	6.45	1.04	
2022-01-07	-1.6	0	354	107	273	6543	321	90	269	6466	1.36	6.24	1	
2022-01-08	0.3	0	356	108	266	6393	329	87	263	6303	1.37	5.88	0.92	
2022-01-09	1.3	0	378	87	266	6393	348	72	264	6341	1.43	5.74	0.91	
2022-01-10	-0.1	80.6	378	87	256	6138	348	72	253	6073	1.44	5.87	0.91	
2022-01-11	-7.5	0	336	117	256	6139	304	94	253	6075	1.5	5.64	0.99	
2022-01-12	-6.9	81.2	349	115	268	6430	312	99	265	6354	1.43	6.45	1.1	
2022-01-13	-5.6	0	349	118	270	6487	318	98	265	6366	1.42	7.41	1.23	
2022-01-14	-4.7	0	365	104	265	6369	0	0	0	0	1.4	6.94	1.24	
2022-01-15	-0.1	0	354	107	263	6318	0	0	0	0	1.4	6.56	1.13	
2022-01-16	-2.7	0	353	107	262	6276	340	98	248	5962	1.45	6.11	1.15	
2022-01-17	-5	81.2	338	110	263	6311	315	97	249	5995	1.46	6.49	1.24	
2022-01-18	-5	80.6	358	106	270	6488	325	92	256	6164	1.64	7.25	1.39	
2022-01-19	-4.7	0	339	107	252	6036	315	102	249	5991	1.87	7.6	1.34	
2022-01-20	-5.4	80.2	361	112	262	6287	349	102	248	5973	1.87	6.03	1.3	
2022-01-21	-3.1	79.8	361	112	262	6287	339	100	255	6137	1.87	6.03	1.3	
2022-01-22	-0.7	0	400	100	262	6296	319	89	249	5981	1.82	6.38	1.35	
2022-01-23	1.1	0	385	97	264	6333	352	82	250	6016	1.04	6.06	1.07	
2022-01-24	4.5	79.9	361	101	263	6315	349	89	250	5999	0.98	5.59	1.11	

2. 데이터 분석

2.1 수질데이터 상관관계(Heatmap)



○ EDA 히스토그램, 히트맵 변수별 분포를 통해 정규분포 여부, 데이터 변환(ex. 로그변환) 필요성 및 변수 간 관계를 유추

○ 수질 예측 모델링을 위한 대상 설정 : 전력량 약품사용량

○ 전력량, 약품사용량에 영향을 미치는 요인 분석

- ■ 수질변수 : BOD, TOC, TN, TP, SS
- ■ 운영변수 : 처리공정, 슬러지처리방법 등
- ■ 환경변수 : 강수량, 기온, 계절 등

2.2 탐색적 데이터 분석

○ 결측치 및 중복값 통계

- 결측치 및 중복값 통계 분석 내용 기입

Overview | Alerts 23 | Reproduction

Dataset statistics

Number of variables	25
Number of observations	365
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	1
Duplicate rows (%)	0.3%
Total size in memory	71.4 KiB
Average record size in memory	200.4 B

Variable types

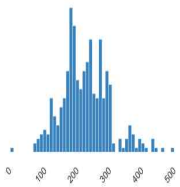
Numeric	23
Categorical	2

○ 주요 변수별 데이터 분포(Histogram)

- 주요 별수별 데이터 분포 분석 결과 기입

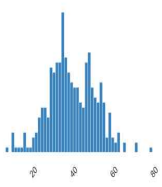
유입수BOD
Real number (ℝ)
HIGH CORRELATION

Distinct	271	Minimum	4.4
Distinct (%)	74.2%	Maximum	511
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	235.92548	Memory size	3.0 KiB



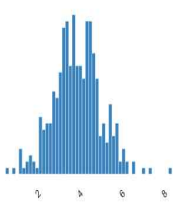
유입수총질소량
Real number (ℝ)
HIGH CORRELATION

Distinct	355	Minimum	5.923
Distinct (%)	97.3%	Maximum	78.871
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	39.013205	Memory size	3.0 KiB



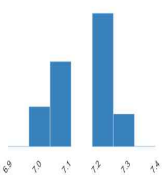
유입수총인량
Real number (ℝ)
HIGH CORRELATION

Distinct	342	Minimum	0.44
Distinct (%)	93.7%	Maximum	8.331
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	3.8489041	Memory size	3.0 KiB



유입수_수소농도
Real number (ℝ)
HIGH CORRELATION

Distinct	7	Minimum	6.9
Distinct (%)	1.9%	Maximum	7.4
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	7.1545479	Memory size	3.0 KiB



○ 데이터 전처리

First rows | Last rows

	온도	탈수케익발생량	케이크함수율	최대유입수량	평균유입수량	총유입수량	최대방류수량	최소방류수량	평균방류수량	총방류수량	유입수_수온
0	3.40	0.00	399	114	285	6845	377	100	277	6652	9.20
1	-1.20	0.00	382	117	299	7171	382	107	285	6842	9.80
2	-1.90	0.00	346	111	272	6523	327	104	263	6303	9.80
3	-2.50	0.00	341	101	275	6600	446	88	250	6008	9.60
4	-2.80	0.00	342	106	267	6409	401	86	255	6131	9.20
5	-2.20	0.00	353	106	295	7078	388	94	268	6432	9.50
6	-1.60	0.00	354	107	273	6543	321	90	269	6466	9.70
7	0.30	0.00	356	108	266	6393	329	87	263	6303	9.40
8	1.30	0.00	378	87	266	6393	348	72	264	6341	9.50
9	-0.10	80.60	378	87	256	6138	348	72	253	6073	9.70

3. 데이터 학습 및 모델정의

3.1 모델정의 및 컴파일

○ 시계열 모델 정의 : LSTM

```
# LSTM 모델 정의
model = Sequential()
model.add(LSTM(64, activation='tanh', return_sequences=True, input_shape=(seq_length, X_train.shape[2])))
model.add(Dropout(0.2))
model.add(LSTM(32, activation='tanh'))
model.add(Dropout(0.2))
model.add(Dense(1))
```

```
/home/was/.local/lib/python3.9/site-packages/keras/src/layers/rnn/rnn.py:204: UserWarning: Do not pass an `input_shape`/`input_dim` argument to a layer.
When using Sequential models, prefer using an `Input(shape)` object as the first layer in the model instead.
  super().__init__(**kwargs)
```

○ 모델 컴파일

```
# 모델 컴파일
model.compile(optimizer='adam', loss='mean_squared_error')
```

3.2 모델학습 및 학습 시각화

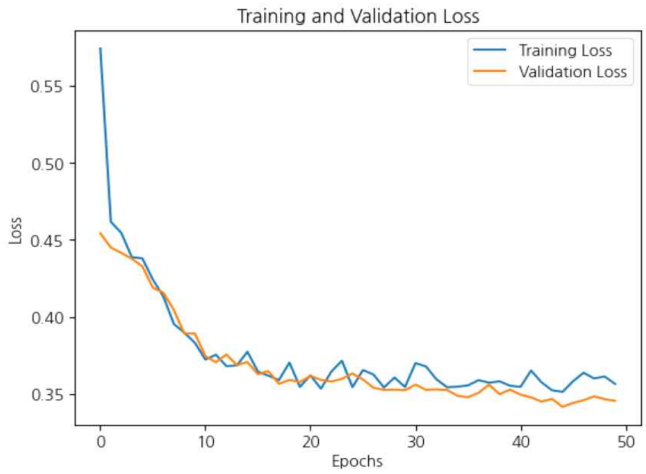
○ 모델 학습

```
# 모델 학습 및 history 저장
history = model.fit(X_train, y_train, epochs=50, batch_size=16, validation_data=(X_test, y_test), verbose=2, shuffle=False)
```

```
Epoch 1/50
53/53 - 2s - 45ms/step - loss: 0.5745 - val_loss: 0.4542
Epoch 2/50
53/53 - 0s - 4ms/step - loss: 0.4618 - val_loss: 0.4451
Epoch 3/50
53/53 - 0s - 4ms/step - loss: 0.4545 - val_loss: 0.4416
Epoch 4/50
53/53 - 0s - 4ms/step - loss: 0.4388 - val_loss: 0.4378
Epoch 5/50
53/53 - 0s - 4ms/step - loss: 0.4380 - val_loss: 0.4328
Epoch 6/50
53/53 - 0s - 4ms/step - loss: 0.4243 - val_loss: 0.4190
Epoch 7/50
53/53 - 0s - 4ms/step - loss: 0.4132 - val_loss: 0.4157
Epoch 8/50
53/53 - 0s - 4ms/step - loss: 0.3954 - val_loss: 0.4044
Epoch 9/50
53/53 - 0s - 4ms/step - loss: 0.3897 - val_loss: 0.3892
Epoch 10/50
53/53 - 0s - 4ms/step - loss: 0.3831 - val_loss: 0.3893
Epoch 11/50
53/53 - 0s - 4ms/step - loss: 0.3723 - val_loss: 0.3745
Epoch 12/50
53/53 - 0s - 4ms/step - loss: 0.3754 - val_loss: 0.3705
```

○ 학습과정 시각화

```
# 학습 과정 시각화
plt.plot(history.history['loss'], label='Training Loss')
plt.plot(history.history['val_loss'], label='Validation Loss')
plt.title('Training and Validation Loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
plt.show()
```



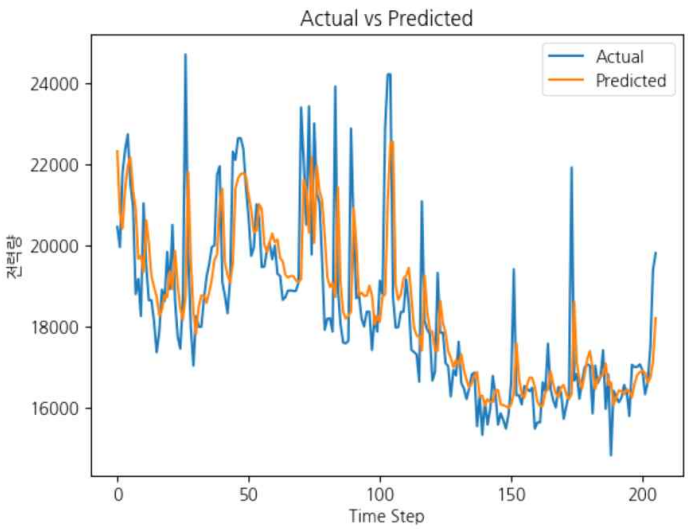
3.3 모델 예측

○ 예측값 vs 실제값 비교

```
# 예측값 역변환
y_pred_inverse = scaler.inverse_transform(np.concatenate((test_scaled[seq_length:, :-1], y_pred), axis=1))[:, -1]
```

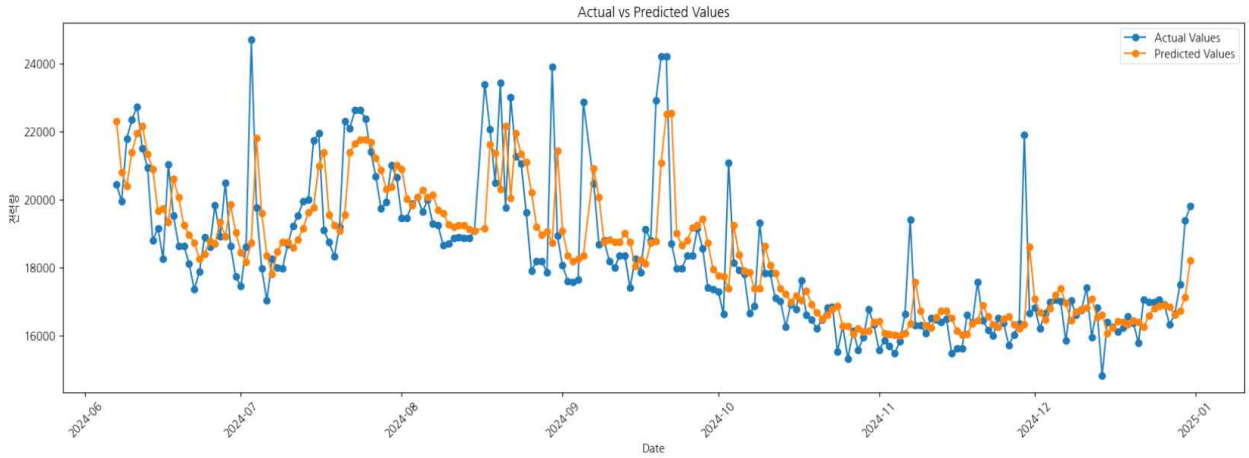
```
# 실제값 역변환
y_test_inverse = scaler.inverse_transform(np.concatenate((test_scaled[seq_length:, :-1], y_test), axis=1))[:, -1]
```

```
# 시각적 비교 그래프
plt.plot(y_test_inverse, label='Actual')
plt.plot(y_pred_inverse, label='Predicted')
plt.title('Actual vs Predicted')
plt.xlabel('Time Step')
plt.ylabel('전력량')
plt.legend()
plt.show()
```



○ 일자별 예측값과 실제값 비교

```
# 일자로 비교한 예측값과 실제값 비교
plt.figure(figsize=(20, 6))
plt.plot(df_concat['일자'][split_index + seq_length:], y_test_inverse, label='Actual Values', marker='o')
plt.plot(df_concat['일자'][split_index + seq_length:], y_pred_inverse, label='Predicted Values', marker='o')
plt.title('Actual vs Predicted Values')
plt.xlabel('Date')
plt.ylabel('전력량')
plt.xticks(rotation=45)
plt.legend()
plt.show()
```



4. 프로토타이핑(화면)

4.1 모델 예측

○ 사업소별/분기별 수질 예측

- 예측 결과 내용 기입 AA
- 예측 결과 내용 기입 BB

○ 전력량 및 유입량 예측결과

- 예측 결과 내용 기입 AA
- 예측 결과 내용 기입 BB

전력량 및 유입량 예측 결과

예측 기간: 7일 예측하기

7일 예측 결과

날짜	예측 유입량	예측 전력량 (kWh)
2025-01-01	9,264	23,177.08
2025-01-02	9,010	22,477.26
2025-01-03	8,913	22,236.11
2025-01-04	8,847	22,082.51
2025-01-05	8,834	22,052.58
2025-01-06	8,787	21,943.05
2025-01-07	8,781	21,925.91