[직접분석 보고서]

팀 명		Q.E.D.
과제명		미션 : □미래가치 □경제활력 ■민생·안전
		AcuaForecast 4대강 녹조 예측을 위한 머신러닝 기반 시스템
활용 데이터	공공	물환경정보시스템 – 녹조(조류) 정보(과거 수질 자료)
	민간	

과제 개요(150자)

저희 팀은 4대강에 녹조의 발생을 예측하는 머신러닝 기반 예측 모델을 개발했습니다. 이 모델을 활용하여 요소들의 변화 추세를 기반으로 값을 입력하면 녹조 발생을 예측할 수 있고, 이를 토대로 사전에 녹조 대비 조치를 취할 수 있습니다.

활용 데이터 및 분석도구

- **활용 데이터**

: 환경부 산하기관인 국립환경과학원의 물환경정보시스템에서 얻은 녹조 정보 데이터를 활용했습니다. 해당 데이터는 CSV 파일로 2016년부터 2022년까지의 일 단위 시계열 데이터를 사용하였으며, 데이터가 주어진 것은 2000년부터 있었지만, 녹조에 영향을 주는 요인들의 값 중에 결측치가 많아서 유의미하지 않은 데이터로 판단하여 2016년부터 2022년 데이터로 분석을 진행하였습니다.
저희가 수집한 녹조 데이터의 feature로는 강의 유형별 분류, 지점 명, 채수 위치, 조사일, 수온, pH, DO, 투명도, 탁도, Chl-a(엽록소-a의 농도), 유해 남조류 세포 수, Microcystis, Anabaena, Oscillatoria, Aphanizomenon(녹조에 관련된 미세조류의 종류 4가지), 지오스민, 2MIB(일부 녹조 종류가 생성하는 냄새 성분), Microcystin-LR(Microcystis가 생성하는 독성 물질) 으로 총 18개로 구성되어있습니다.

- **분석 도구**

: Jupyter Notebook 환경에서 Python을 활용한 pandas, seaborn, matplolib, numpy 라이브러리를 통해 데이터 전처리와 시각화를 진행하였고, sklearn, tensorflow, keras 라이브러리를 활용하여 머신러닝을 진행하였습니다.

- **라이브러리 버전**

Python version: 3.9.12
pandas version: 1.4.2
numpy version: 1.21.5
matplotlib version: 3.5.1
seaborn version: 0.11.2
tensorflow version: 2.11.0
sklearn version: 1.0.2
keras version: 2.10.0

창의성

현재 물환경정보시스템에서 운영 중인 조류 예측정보에서는 유해 남조류와 수온의 예측을 진행하고 있습니다. 저희가 개발한 예측 모델은 기존의 조류 예측정보 시스템과는 다르게, 유해 남조류 수를 제외한 수온, pH, 투명도, 탁도와 같이 쉽게 관측할 수 있는 요소들을 활용하여 녹조를 예측할 수 있는 장점이 있습니다. 이를 통해 물환경정보시스템에 추가하여 더 효과적인 녹조 예측과 대비를 진행할 수 있을 것입니다.

적합성

- **문제 정의**

현재 녹조 문제는 해마다 여름이 다가올수록 대두되는 환경 이슈입니다. 녹조 현상이 발생함에 따라 수생 생태계 파괴, 수질오염, 독성 미세조류에 의해 오염된 농작물 섭취 등의 다방면의 문제를 유발합니다. 그래서 저희 팀은 이러한 문제를 수집한 데이터를 활용해 분석하고 대응해야 할 필요성을 느꼈습니다.

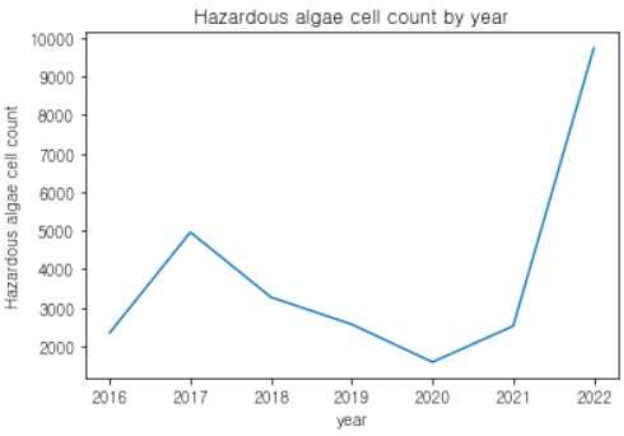
[녹조 관련 뉴스 기사]

광주 서구, 풍암호수 녹조 현상 해소 집중
(https://www.gjdream.com/news/articleView.html?idxno=616190)

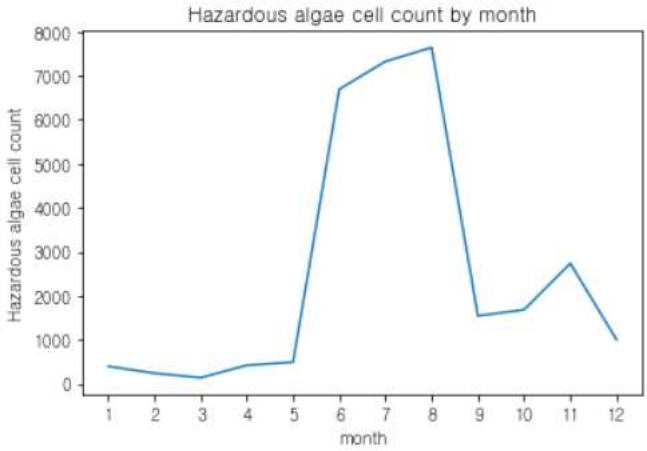
부산에 이어 거제 앞바다에 녹조 현상 관찰…원인 조사
(https://www.yna.co.kr/view/AKR20220816106800052)

- **데이터 확인 및 시각화**

저희 팀은 2016년부터 2022년까지의 7년간, 4대강의 여러 채수 위치 지점을 기준으로 한강 5,249개, 낙동강 6,424개, 금강 2,197개, 영산강 2,548개로 총 16,418개의 데이터를 확보했습니다.



낙동강 연 단위 남조류 세포 수



낙동강 월 단위 남조류 세포 수

위의 그래프는 모든 데이터를 병합한 것을 순서대로 연 단위, 월 단위의 남조류 세포 수 그래프를 시각화한 결과입니다. 보는 것과 같이, 녹조 수치가 2020년을 기준으로 다시 상승하는 것을 볼 수 있으며, 기온이 높은 5월~9월 사이의 녹조 발생률이 높다는 것을 알 수 있습니다.

	분류	지점명	채수위치	조사일	수온(℃)	pH	DO(㎎/L)	투명도	탁도	Chl-a (㎎/㎥)	유해남조류 세포수 (cells/㎖)	Microcystis	Anabaena	Oscillatoria	Aphanizomenon	지오스민	2MIB	Microcystin-LR (㎍/L)
0	호소	주암호	신평교	2016.01.05	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	0.0	0.0	0.0	NaN	NaN	NaN
1	호소	주암호	신평교	2016.01.12	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	0.0	0.0	0.0	NaN	NaN	NaN
2	호소	주암호	신평교	2016.01.20	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	0.0	0.0	0.0	NaN	NaN	NaN
3	호소	주암호	신평교	2016.01.27	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	0.0	0.0	0.0	NaN	NaN	NaN
4	호소	주암호	신평교	2016.02.02	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	0.0	0.0	0.0	NaN	NaN	NaN

위 그림의 raw 데이터를 확인한 결과 데이터양은 많지만, 결측값 또한 많은 것을 볼 수 있었고, 이를 처리하고 필요한 feature 들만 남길 수 있도록 전처리 과정이 필요했습니다.

- 데이터 전처리

저희가 확보한 데이터 안에서 Microcystis, Anabaena, Oscillatoria, Aphanizomenon 칼럼은 남조류의 세포 수를 결정하는 변수들로 제거했으며, 냄새와 독성 성분 또한 녹조 발생 이후의 파생 변수들이므로 지오스민, 2MIB, Microcystin-LR 칼럼을 제거했습니다.

분석에 사용할 주요 변수로는 남조류 세포 수를 결정변수로 설정하고, 지점 명, 채수 위치, 조사일을 제외한 변수들을 녹조 발생의 주요 요인들로 설정하였습니다. 각각의 설명변수들은 녹조 발생에 큰 영향을 주기 때문에, 행 데이터 안에 결측값이 하나라도 있는 경우 해당 행을 삭제했습니다. 남조류 세포 수가 0인 경우 즉, 녹조 발생이 일어나지 않은 데이터는 학습 과정에 혼동을 줄 수 있으므로 제외했습니다. 국가의 녹조 발령기준인 조류경보제를 활용하여 남조류 세포 수가 10,000 이상일 때 발령되는 경계 발령을 나타내는 파생 변수인 "경계 발령" 변수를 추가했으며. 이진 분류를 위해 수치가 10,000 이상이면 1로, 그렇지 않으면 0으로 설정하여 모델에 넣을 수 있도록 데이터 전처리를 완료했습니다.

	분류	채수위치	조사일	수온(℃)	pH	DO(㎎/L)	투명도	탁도	Chl-a (㎎/㎥)	유해남조류 세포수 (cells/㎖)	경계발령
0	호소	낙동강(해평)	2016-08-22	24.4	8.0	8.1	1.9	3.8	8.0	240.0	0
1	호소	낙동강(해평)	2016-08-29	21.1	7.5	7.5	1.5	5.3	6.3	405.0	0
2	호소	낙동강(해평)	2016-09-05	20.9	8.1	8.6	1.3	4.3	7.2	677.0	0
3	호소	낙동강(해평)	2016-09-12	19.7	8.2	8.6	1.8	3.9	8.6	149.0	0
4	호소	낙동강(해평)	2016-09-19	22.8	8.0	7.7	1.4	2.0	11.3	168.0	0

- 전처리 된 데이터 스케일링

: 데이터 특성 간의 크기 차이를 보정하고 모델의 학습 및 예측 성능을 향상 시키기 위해서 MinMaxScaler를 사용하였습니다.

```
scaler = MinMaxScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

- **ANN 모델 설계**

: 스케일링까지 마친 데이터를 기반으로 모델을 만들었습니다. 이 모델은 데이터의 복잡한 패턴을 학습하기 위해 비선형 관계를 모델링할 수 있는 다층 구조를 가지고 있습니다. 각 층마다 뉴런의 개수를 64, 32, 32로 설정하여 모델이 데이터의 다양한 특징과 패턴을 학습할 수 있도록 하였습니다. 또한, ReLU 활성화 함수를 사용하여 비선형성을 도입하였으며, 이를 통해 모델이 복잡한 패턴을 학습할 수 있게 되었습니다. 마지막 출력층은 녹조 발생 여부를 이진 분류로 나타내기 위해 Sigmoid 활성화 함수를 사용하였습니다. 이를 통해 모델은 입력된 데이터를 기반으로 녹조 발생 여부를 예측할 수 있게 되었습니다.

```
model = Sequential()

model.add(Dense(64, input_dim=X_train_scaled.shape[1], activation='relu'))
model.add(Dense(32, activation='relu'))
model.add(Dense(32, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
```

- **ANN 모델 학습**

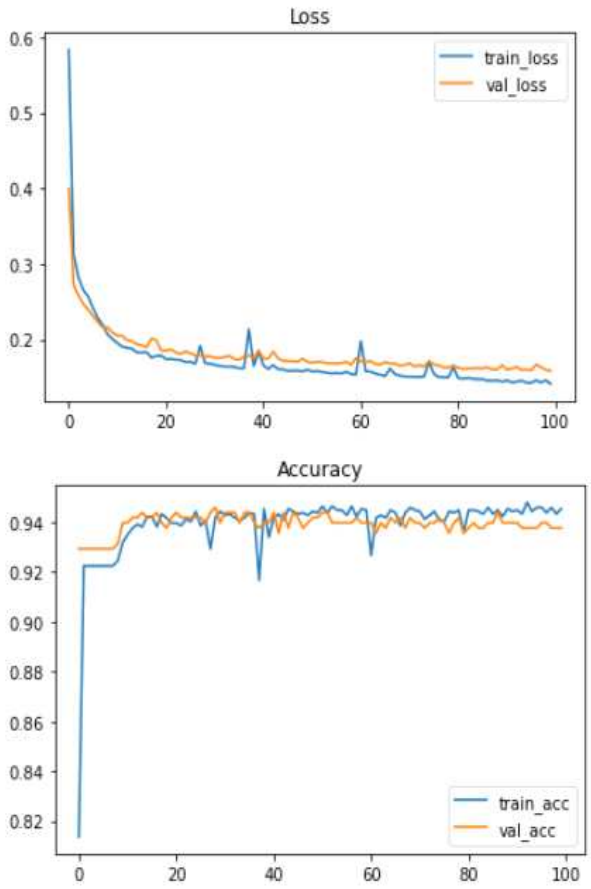
: 모델 학습 시 과적합을 방지하고 모델의 일반화 성능을 높일 수 있게 조기 종료(Early Stopping)를 이용하여 학습을 진행하였습니다.

```
# EarlyStopping 콜백 정의
early_stopping = EarlyStopping(monitor='val_loss', patience=10, restore_best_weights=True)

# 모델 컴파일
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

# 모델 학습
history = model.fit(X_train_scaled, y_train, validation_data=(X_test_scaled, y_test),
                    epochs=1000, batch_size=64, callbacks=[early_stopping])
```

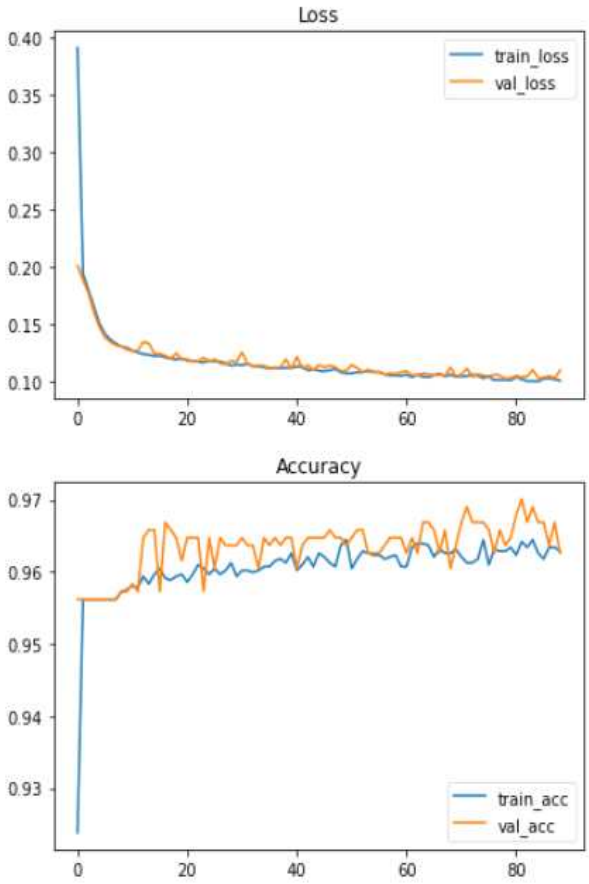
- **Accuracy & Loss**

: Accuracy와 Loss는 머신 러닝 모델의 성능을 평가하고 측정하기 위해 사용되는 대표적인 지표입니다. Accuracy(정확도)는 모델이 올바르게 분류한 샘플의 비율을 나타냅니다. 이는 전체 샘플 중에서 올바르게 예측한 샘플의 비율로 계산됩니다. Accuracy가 높을수록 모델의 분류 성능이 좋다고 평가할 수 있습니다. Loss(손실값)는 모델의 예측과 실제 값 사이의 차이를 측정하는 지표입니다. 손실값은 주로 손실 함수(loss function)를 통해 계산되며, 모델이 예측의 오류를 최소화하기 위해 학습 중에 최적화되는 값입니다.

아래의 그래프를 보면 모델이 학습을 진행함에 따라 점차 개선되어가며 최종적으로는 Accuracy가 높고 Loss가 낮아지는 성능이 좋은 모델이 되는 것을 바로 볼 수 있습니다.



낙동강 녹조 예측 모델의 Loss, Accuracy



4대강 통합 녹조 예측 모델의 Loss, Accuracy

- **4대강 통합을 한 이유**

: 처음에 학습을 진행할 때는 한강, 낙동강, 금강, 영산강 따로 진행하여서 훌륭한 결과물이 나왔었지만, 이렇게 학습한다면 지역별로 과적합 문제에 도달할 수 있다고 판단하여 모든 데이터를 모아서 학습을 진행하였고 결과물도 훌륭하게 나오게 되었습니다.

- **활용**

위의 그림과 같이 각 요소의 값을 입력하면 녹조 발생 여부를 예측하는 모델로 활용할 수 있습니다.

활용성

강에 설치된 댐, 보 등에 수온, DO, pH, 투명도 등 각 요소를 실시간으로 측정하고 기록할 수 있는 장비들을 설치합니다. 이러한 장비들을 통해 수집된 실시간 데이터를 활용하여 시계열 데이터 분석을 수행하면 각 요소의 값들을 예측할 수 있습니다. 이후 예측된 값들은 저희 팀이 개발한 모델에 적용하여 녹조 발생 여부를 예측하는 데 사용될 수 있습니다.

녹조가 발생할 것으로 예측되면, 현재의 요솟값과 예측된 값을 비교한 후, 녹조 발생을 예방하기 위한 대응책을 시행할 수 있습니다. 예를 들어, pH가 낮아져 산성으로 기울어진 경우, 천연 염기성 물질인 베이킹소다 등을 강물에 첨가하여 pH를 높여 정상 수치로 조절하여 녹조 발생을 방지할 수 있습니다. 또한, 투명도가 낮아져서 녹조 발생 가능성이 예상되는 경우, 강물을 여과하거나 침전제를 사용하여 불순물을 응집시키고 제거함으로써 투명도를 개선할 수 있습니다.

이러한 방법들은 녹조 예방과 대응을 위한 효과적인 전략으로 활용될 수 있습니다. 상황에 따라 적절한 방법을 선택하여 녹조 발생을 최소화하고 물 환경을 관리하는 데 도움이 될 것입니다.

정책 활용

위와 같은 활용을 통해 측정 위치별로 지속적인 관리가 필요한 요소들을 알아낼 수 있습니다.

예를 들어, pH가 계속해서 높아지거나 낮아지는 경향을 보인다면, 주변에 있는 농가에서 사용하는 비료의 산성 또는 염기성 특성을 조사하여 적절한 pH 조절 정책을 수립할 수 있습니다. 이를 통해 농가에서 사용되는 비료의 pH 영향을 고려하여 토양 pH를 적절히 조절할 수 있습니다. 또한, 주변에 공장이 위치하고 공업폐수가 배출되는 상황이라면 폐수 처리 시스템을 활용하여 pH를 조절하고 적절하게 처리하는 정책을 시행할 수 있습니다. 이를 통해 주변 환경에 부정적인 영향을 줄이고 pH 관리를 효과적으로 수행할 수 있습니다.

따라서, pH 관리 정책을 통해 지속적인 모니터링과 조절을 수행함으로써 녹조 발생 가능성을 감소시키고 물 환경을 보호하는 데 기여 할 수 있습니다. 이는 지속 가능한 자원 관리와 생태계 보전을 위한 중요한 정책 수립의 일환으로 이어질 수 있습니다.

기대효과

작년 여름 자전거로 인천부터 부산까지 국토 종주를 하던 도중 이상한 냄새를 맡았습니다. 강 쪽을 봤더니 녹조가 발생해있는 것을 직접 목격했습니다. 그 순간, 뉴스에서 녹조 사진을 보는 것만으로는 이해하지 못했던 녹조 문제의 심각성을 몸소 인지하게 되었습니다.

저희 팀이 개발한 모델을 활용하여 녹조 발생을 억제한다면, 미관적인 측면이 개선될 것이고, 좋지 않은 냄새 또한 크게 감소할 것입니다. 이러한 변화가 일어난다면, 자전거를 타고 지나가는 사람들과 함께 산책하는 동네 주민들, 그리고 휴식을 취하러 온 외부인들까지 모두에게 쾌적한 환경을 제공할 수 있을 것입니다.

강물 환경의 개선은 지역 경제와 관광산업에 긍정적인 영향을 미칩니다. 이는 관광객들뿐만 아니라 지역 주민들에게도 삶의 질을 향상시킬 수 있는 기회를 제공합니다. 지역 경제의 발전으로 일자리가 창출되며, 관광산업의 성장은 지역 문화와 예술에 관한 관심을 높입니다. 이는 지역 주민들에게 안정적인 경제적 기회와 다양한 문화적 활동을 제공하여 삶의 질을 향상시키는데 도움이 됩니다.