2021년 K-water 대국민 빅데이터 공모전 수행 결과보고서

<table>
<tr><td>제 목</td><td colspan="5">하천수질데이터 기반 하천활동등급 개발 및 활용방안</td></tr>
<tr><td>공모전형</td><td colspan="2">대학생</td><td>O</td><td>일반·기업</td><td></td></tr>
<tr><td rowspan="6">성 명</td><td rowspan="2">팀 장</td><td colspan="2">서채영</td><td colspan="2">010-8407-2079</td></tr>
<tr><td colspan="4">숭실대학교 전자공학전공</td></tr>
<tr><td rowspan="4">팀 원</td><td colspan="2">서석주</td><td colspan="2" rowspan="4">숭실대학교
정보통계보험수리학과</td></tr>
<tr><td colspan="2">석지연</td></tr>
<tr><td colspan="2">이동현</td></tr>
<tr><td colspan="2">지혜리</td></tr>
</table>

I. 과제 목표

2020년, 갑작스러운 코로나바이러스의 유행으로 우리의 라이프스타일이 완전히 바뀌었다. 사회적 거리두기 격상에 따라 물이 가지고 있는 관광의 힘 역시 현저히 줄어들었다. 더위를 식히려 발을 담그던 사람들도, 수상 레포츠를 즐기던 사람들도, 러닝을 하는 등 자신만의 방식으로 하천을 즐기고 있던 사람들 모두 집으로 돌아가게 되었다. 잃어버린 하천의 미를 되돌아보며 우리는 만남의 공간, 레저의 공간, 산책의 공간 등 다양한 의미를 지녔던 '하천'을 들여다보았다. 포스트 코로나 시대에 우리가 좀 더 쾌적하고 안전하게 하천 활동을 즐길 수 있도록 현 상황에 맞는 새로운 수질평가지표를 만들어보자는 취지에서 본 프로젝트가 시작되었다.
현재 우리나라는 수질을 생활환경기준에 따라 7단계로 구분하여 평가하고 있다. 해당 기준은 하천의 용도를 구분하는 기준으로 하천 활동에 관한 등급을 나타내기에 악취 또는 녹조 등의 요인을 포함하고 있지 않다는 한계가 있다. 본 프로젝트의 목적은 기존등급에 포함되어 있지 않았던 T-N, 클로로필-a 등의 변수를 추가해 하천활동등급을 개발하는 것이다.

본 프로젝트는 지난 5년(2017년~2021년 5월)의 수질 데이터를 이용하여 하천활동등급 모델을 구축하고, 주별 하천활동등급을 예측하고자 한다. 또한, 많은 사람이 하천을 안전하게 즐길 수 있도록 새로운 평

가지표를 상용화하는 다양한 방안을 제시하고자 한다.

II. 주요 내용

기존 생활환경기준의 요소들을 채택하되, 좋음과 약간 좋음을 다소좋음으로, 약간 나쁨과 나쁨을 다소 나쁨으로 통합하여 5등급제로 산정하였다. 추가적으로, 악취에 영향을 주는 암모니아성 질소, 총질소, 클로로필-a를 추가했다. 자연수의 pH는 대부분 6.5 ~ 8.5 범위이며 범위에서 벗어날 경우 수질이 매우 나쁜 것으로 간주한다. 용존 산소량은 물 속에 포함되어 있는 산소량으로 수질이 좋을수록 높은 값을 가진다. 생화학적산소요구량(BOD)은 호기성 미생물이 일정 기간 동안 물속에 있는 유기물을 분해할 때 필요한 산소의 양이다. 화학적산소요구량(COD)는 유기성 화학물질을 산화시키기 위한 산화제의 양이다. BOD와 COD는 수치가 높을수록 수질이 많이 오염되었음을 시사한다. 총인은 탁도에 영향을 준다. 온도가 낮으면 밀도가 증가하여 탁도가 높아져 총인 농도도 높아져 탁도와 총인 농도가 직접적으로 관련이 있음을 알 수 있었다. 총질소(T-N)는 수중에 포함된 질소화합물의 총량이다. 현재 우리나라의 하천수질기준에는 총질소에 관한 기준이 없어 이와 관련된 논문[1]을 인용하여 기준을 설정하였다. TOC는 총유기탄소량으로 물속에 함유된 유기물 물질의 농도를 의미한다. BOD와 COD보다 빠르고 정확하게 유기물을 측정할 수 있어 최근 수질오염 정도를 나타내는 새로운 지표로 떠오르고 있다. SS는 부유물질량으로 물에 용해되지 않고 부유하는 물질을 의미한다. 분원성 대장균군은 온혈동물의 배설물에서 발견되는 간균으로서 소독되지 않은 조건에서는 상시 존재 가능한 탓에 일부 병원성을 나타내는 균에 의해 인체에 해로운 영향을 끼칠 수 있다. 암모니아성 질소는 암모늄염을 질소량으로 나타낸 것으로, 오염지표 뿐 아니라 수역 부영양화의 요인이 된다. 클로로필-a 역시 부영양화 관련 지표로서 남조류세포수와 함께 환경부 조류경보의 기준이 되지만, 타 변수들과 다른 측정망을 통해 측정되기 때문에 클로로필-a만을 평가요소로 포함하였다.
이 모든 요소들을 종합하여 아래와 같은 등급기준을 만들 수 있었다.

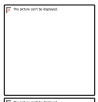
[1] 김학관, 정한석, 배승종. "하천에서의 영양물질 관리를 위한 총질소 환경기준 설정에 관한 연구." 57. 3 (2015): 121-127.

	매우 좋음	다소 좋음	보통	다소 나쁨	매우 나쁨
pH	6.5~8.5	6.5~8.5	6.5~8.5	6.0~8.5	-
BOD	1 이하	3 이하	5 이하	10 이하	10 초과
COD	2 이하	5 이하	7 이하	11 이하	11 초과
DO	7.5 이상	5.0 이상	5.0 이상	2.0 이상	2.0 미만
T-N	1.5 이하	3.0 이하	4.0 이하	8.0 이하	8.0 초과
T-P	0.02 이하	0.1 이하	0.2 이하	0.5 이하	0.5 초과
TOC	2 이하	4 이하	5 이하	8 이하	8 초과
SS	25 이하	25 이하	25 이하	100 이하	100 초과
분원성대장균군	200 이하	200 이하	1000 이하	1000 초과	1000 초과
암모니아성질소	0.5 이하	0.5 이하	0.5 초과	0.5 초과	0.5 초과
클로로필-a	15 미만	15 이상	25 이상	100 이상	100 이상

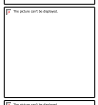
표 1 하천활동등급표

매우 좋음 : 피부에 닿아도 무방하고, 악취가 나지 않아 직접 접촉이 필요한 하천활동이 가능하다.


다소 좋음 : 오염물질이 포함되어 있으나, 피부에 닿아도 무방하고, 악취가 나지 않아 직접 접촉하는 하천활동이 가능하다.



보 통 : 약간의 악취가 있을 수 있고, 피부 접촉에 주의가 필요하지만, 산책 등의 일상적 하천활동은 가능하다.



다소 나쁨 : 피부에 접촉할 시 피부병을 유발할 수 있으며, 악취가 다소 심하여 일상적 하천활동에 지장을 줄 수 있다.

매우 나쁨 : 피부 접촉 시 피부병을 유발할 수 있으며, 악취가 심하고 녹조가 있는 경우가 많아 하천활동을 자제해야 한다.

III. 활용데이터 및 수행내용

1. [2]물환경정보시스템(www.water.nier.go.kr)의 수질측정망의 일별 데이터를 수집하여 분석하였다.0

<A. 데이터 전처리>

2. Basic Pipeline 함수로 기초적인 전처리를 시행하여 다음과 같이 데이터를 가공하였다.

2.1. 데이터가 54개 미만인 측정소 드랍

2.2. str 형식이었던 측정일시를 DateTime형식으로 변환

2.3. 필요하지 않은 컬럼 드랍

[2] 물환경정보시스템(www.water.nier.go.kr)의 수질측정망 자료 중 하천의 2017.1.~2021.5. 데이터

2.4. '정량한계미만'관측치 0으로 대체

2.5. 수치형 변수들 모두 float으로 변환

2.6. 음수인 유량 절댓값 취함

3. 결측값은 다음과 같이 대체하였다.

3.1. 데이터별로 14를 초과하는 pH를 14로, pH 0을 $10^{\wedge}\llbracket -6 \rrbracket$ 으로 대체하고 BoxCox변환 후 정규화

3.2. IterativeImputer에서 ExtraTreeRegressor를 적용하여 결측값을 채움

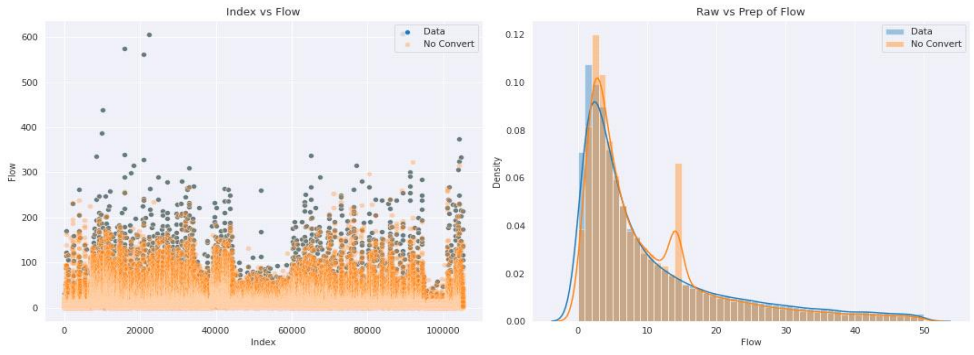
3.2.1. BoxCox변환 후 Imputation을 하여 모델 적용 시의 수치적 under-flow를 피함

3.2.2. ExtraTreeRegressor 모델에 변환을 적용하지 않을 시 밀도가 높은 값에 과적합 되는 경향이 발견

3.2.3. BoxCox, ExtraTreeRegressor, Inverse의 과정을 적용하면 과적합 되는 경향이 사라짐

3.3. Impute된 데이터의 pH의 14를 넘어가는 값을 14로 대체

그림 2 BoxCox 변환 후 ExtraTree Imputation : 원 데이터의 분포와 유사함.



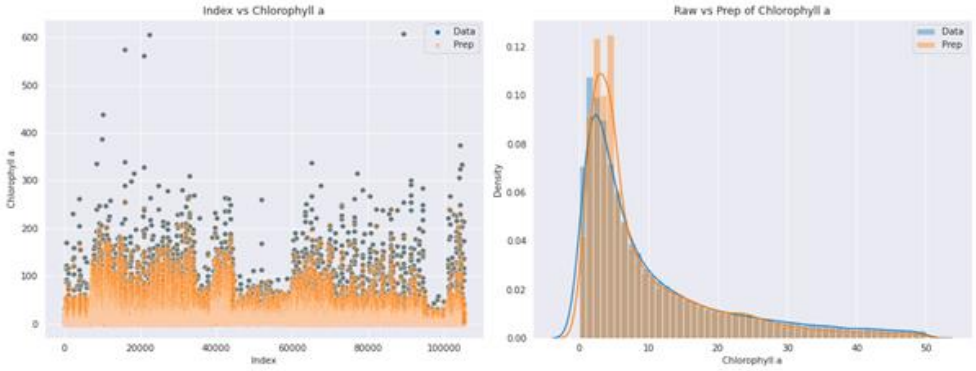


그림 1 BoxCox 변환을 하지 않을 시 ExtraTree Imputation : 원래 데이터의 분포에서 벗어난 값이 보임.

4. 라벨링 : 하천 생활등급표를 참조하여 target 값을 라벨링 하였다.

5. 변환 : 정규성을 띠지 않는 컬럼들에 한하

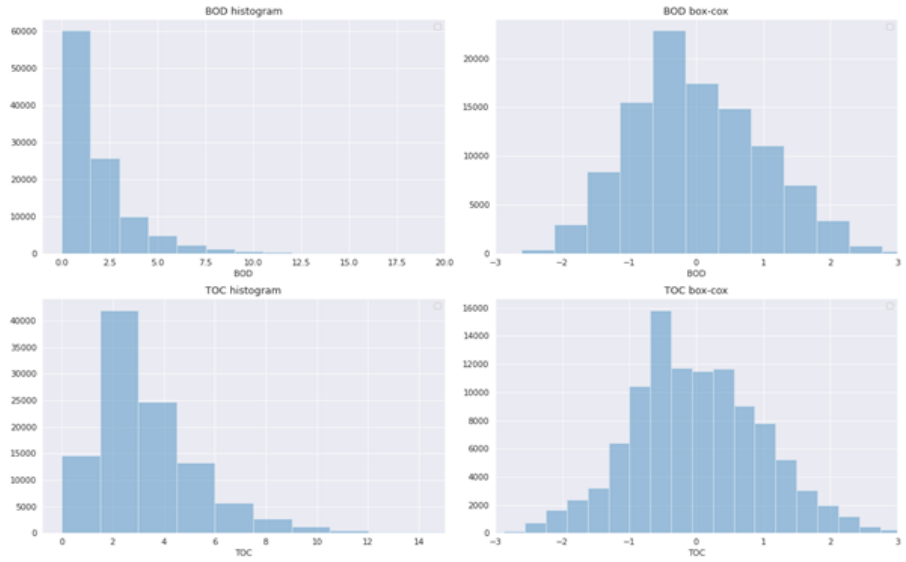


그림 3 BoxCox 변환으로 꼬리가 긴 분포에서 정규성을 띠는 분포로 변환.

여 BoxCo x변환을 적용하여 정규성을 확보하였다.

6. 데이터 준비

하천 생활 등급을 예측하기 위해 수질 데이터 각 요소에 ARIMA모델을 적용하였다. 모델 예측에 사용한 데이터는 "ExtraTree_Data_raw_targeted.csv"이다. 예시로 사용한 데이터는 측정소명이 '서울-가양'인 'pH' 데이터이다. 그림 4는 데이터의 2017년부터 2021년 6월 첫째 주까지의 주별 관측값을 나타낸 것이다.

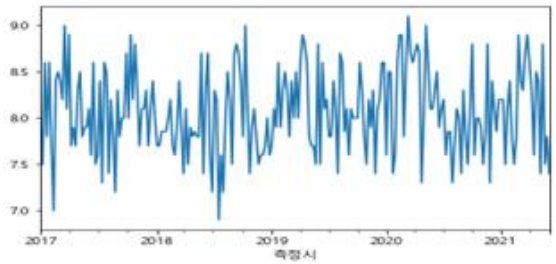


그림 4

7. 정상성 검증

데이터의 정상성(Stationary)을 확인하기 위해 대표적인 테스트인 ADF(Augmented Dickey-Fuller) 테스트를 진행했다. 그림 5는 ADT TEST를 진행하기 위해 만든 함수를 보여준다. '서울-가양'인 'pH' 데이터의 ADF 테스트의 P-Value 값은 0.000003으로 유의수준 0.05이내이다. '서울-가양'인 'pH' 데이터는 정상성을 가진다.

```
In [4]: def adf_test(df):
            result = adfuller(df.values)
            print("ADF Statistics : %f" % result[0])
            print('p-value : %f' % result[1])

In [5]: adf_test(가양.ph)
        ADF Statistics : -5.428888
        p-value : 0.000003
```

그림 5

8. 자기상관계수와 편자기상관계수

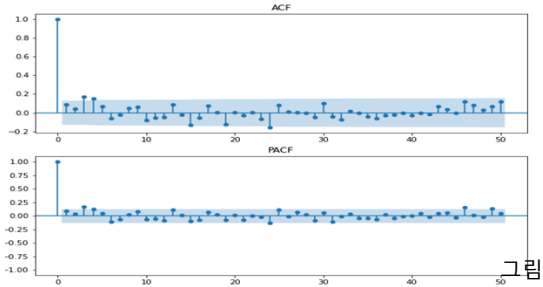


그림 6

확인

ARIMA 모델의 파라미터 값(p, d, q)을 추측하기 위해 자기상관계수(ACF)와 편자기상관계수(PACF) 그래프를 그렸다(그림 6). 하지만 이렇게 추측한 파라미터 값은 최적이 아니다.

9. ARIMA모델 내 최적의 파라미터 값 도출

ARIMA 모델에서 AIC값을 최소로 만드는 파라미터 값을 출력해주는 pmdarima.arima' 모듈 내의 auto_arima 함수를 통해 최적의 파라미터 값을 구해주었다(그림 7). 최적의 파라미터 값은 (4, 1, 0)이다.

그림 7

10. 실제 데이터와 예측 데이터 비교

그림 8는 2017년부터 2021년 6월 첫째 주까지의 실제 데이터 값과 그 데이터를 예측한 값을 비교한 그래프이다.

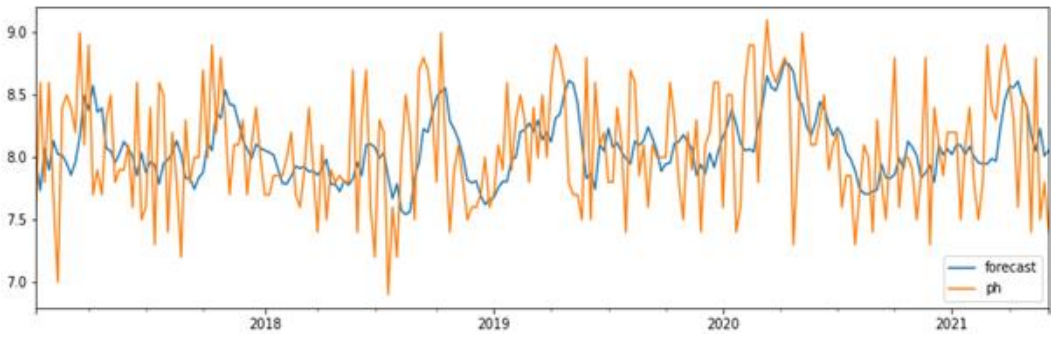


그림 8

11. 모델링 및 3주치 관측 값 예측

ARIMA (4,1,0) 모델을 통해 향후 3주치 관측 값을 예측해보았다(그림 9).

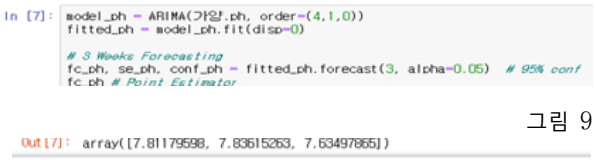


그림 9

'서울-가양'의 pH 향후 3주치 예측값은 7.812, 7.836, 7.635이다.

12. 실제 데이터와 예측값이 포함된 그래프

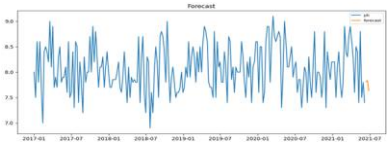


그림 10

그림 10은 3주치 예측값을 포함한 데이터의 그래프이다. 점차적으로 감소하는 추세를 보인다.

13. 요인별 적용

그림 11

모든 수질 요인에 과정 반복
지금까지의 과정을 'pH', 'DO', 'BOD', 'COD', '부유물질', 'TN', 'TP', 'TOC', '수온', '총대장균군수', '암모니아성 질소', '클로로필a','분원성대장균군수', '유량' 데이터에도 진행해 주었다. 각 요소별 예측값들을 데이터프레임으로 만들어주었다(그림 11).

14. 판별 함수 준비
하천활동등급표에 나온 기준을 토대로 각 요소별 등급을 판별해 줄 함수들을 만들어주었다. '매우 좋음' 등급은 1로, '다소 좋음' 등급은 2로, '보통' 등급은 3으로, '다소 나쁨' 등급은 4로, '매우 나쁨' 등급은 5로 라벨링 해주었다.

15. 하천생활등급 예측

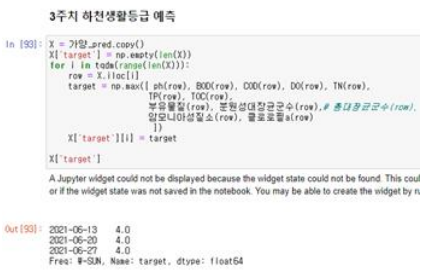


그림 12

각 요소별 등급을 고려하여 3주치 하천생활등급을 예측했다. 각 요소별 등급의 최대값을 그 하천의 하천생활등급으로 지정해주었다(그림 12). 측정소명이 '서울-가양'인 하천의 향후 3주간 하천 생활등급은 모두 '다소 나쁨'에 해당하는 것으로 예측된다.

Ⅳ. 결과 및 기대효과

기존의 수질 데이터를 이용하여 하천생활등급을 도출하는 모델을 만들 수 있었다. 모델을 만드는 과정에서 데이터가 주별로 수집되었다는 특성으로 인해 모델에 계절성을 반영하는 기법인 SARIMA를 적용하지 못하는 문제점이 생겼다. 수질 데이터를 일별로 수집하여 이러한 문제점을 해결한다면, 모델의 예측력을 더 높일 수 있을 것이다.
또, 하천생활등급을 홈페이지와 애플리케이션, 그리고 각 하천 전광판에 게시한다면, 이용자의 실생활에 수질을 적용해보는 기회가 될 것이라 생각한다. 앞의 과정을 거쳐 만든 하천별 생활등급을 여러 가지 플랫폼에 활용한 프로토타입을 제시한다.

그림 13

그림 13에서는 My Water 누리집 첫 화면에 하천생활등급을 그래프와 수치로 보여주고자 한다.

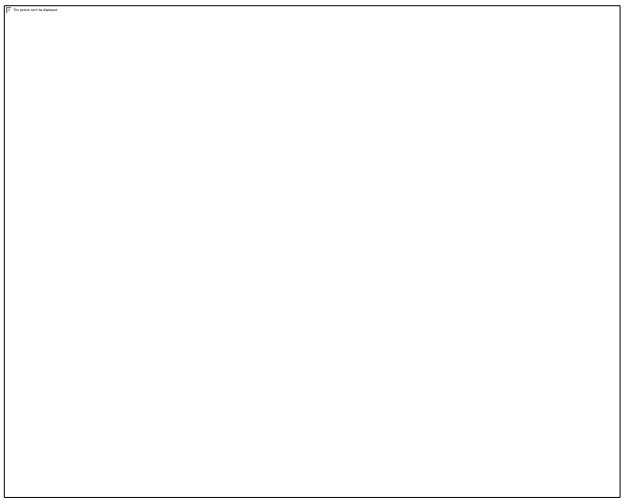


그림 14처럼 쉬운 물통계 페이지에서 하천의 pH, DO, BOD, COD와 같은 수치가 우리에게 어떤 영향을 주는지 충분히 설명한다면 수질에 대한 이용자들의 이해도가 향상할 것으로 사료된다.

그림 14

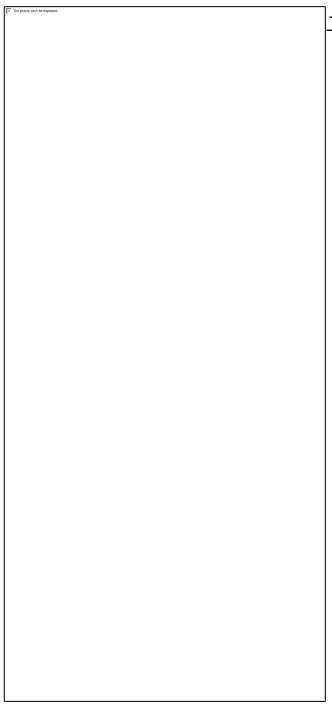
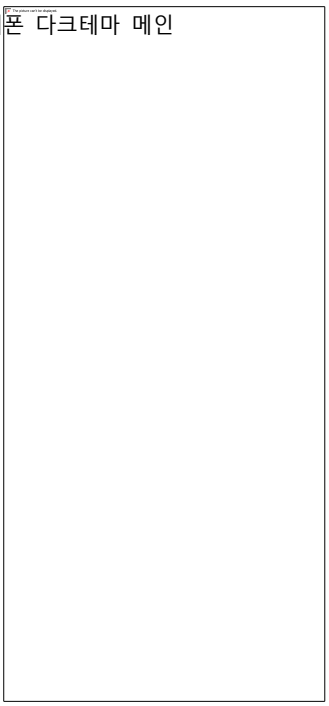


그림 15 아이폰 다크테마 메인



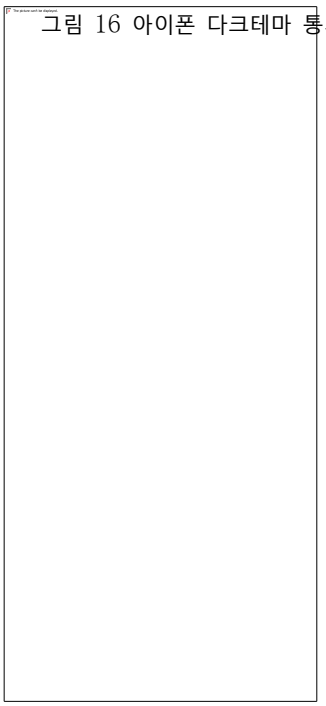


그림 16 아이폰 다크테마 통계

그림 17 삼성 화이트테마 메인

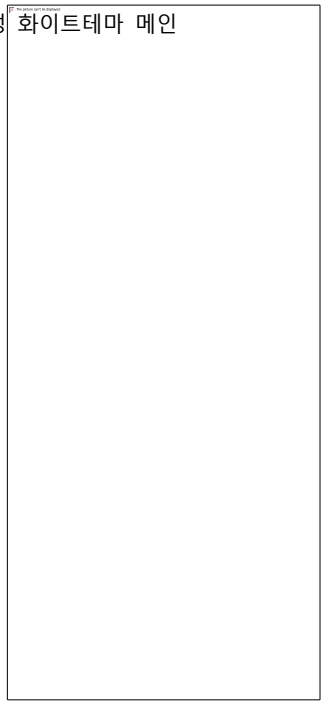


그림 18 삼성 화이트테마 통계

그림 15 ~ 18과 같이 애플리케이션에서는 하천생활등급을 그래프와 캐릭터로 시각화하여 하천 수질에 대한 접근성을 제고할 수 있을 것으로 판단된다.

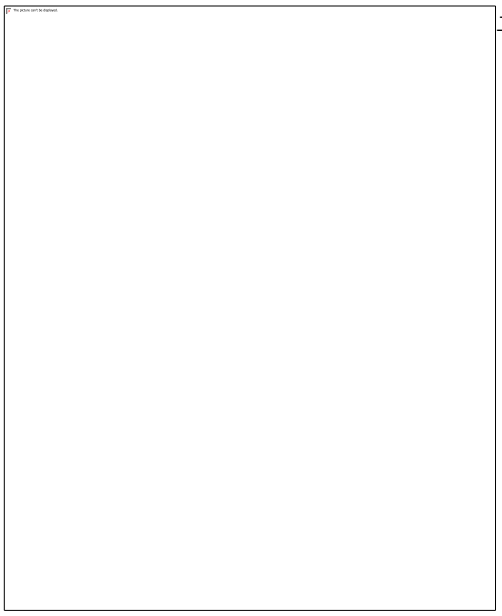


그림 19 전광판 세로

그림 20 전광판 가로

각 하천의 전광판에 그림 19 ~ 20과 같이 하천생활등급을 게시한다면, 국민들의 하천 생활에 유용한 정보를 제공될 것이 기대된다. 그리고, 오수 음용, 부유물질로 인한 부상과 같은 안전사고에 대비할 수 있다는 장점이 있다.