2023년 K-water 대국민 물 빅데이터 공모전 수행 결과보고서

제 목	다목적 댐의 상태 평가와 예측을 위한 머신러닝 기반 모델 개발		
부 문	데이터 융합		
성 명	팀 장	차수빈	010-7570-4629
		이화여자대학교 통계학과	chasubin02@gmail.com
	팀 원	배주원	이화여자대학교 컴퓨터공학전공
		박신영	이화여자대학교 휴먼기계바이오공학부

Ⅰ. 과제 목표

국내에는 총 21개의 다목적 댐이 설치되어 있다. 이들은 안정적인 수자원 공급을 위한 핵심 시설이며, 따라서 현재 다목적 댐들의 상태 점검과 기능 확인이 중요한 과제이다. 이를 위해 머신러닝을 활용하여 각 댐의 상태를 평가하는 회귀 모델을 개발하는 것을 목표로 설정하였다. 각 댐의 수문 정보와 기상 정보를 입력하면 해당 댐의 저수량을 예측하는 회귀 모델을 개발하였으며, 이후 댐의 활용능력과 수질 상태를 종합적으로 고려하여 각 댐의 현황을 파악하였다. 이러한 모델이 실제 현장에서 적용되면, 다목적 댐의 상태를 신속하게 평가하고 이상 상황에 대비하는 데 도움이 되리라 기대한다.

Ⅱ. 활용 데이터

자료들을 성격과 목적에 맞게 4가지로 분류하고 가공한 후, 이를 하나의 CSV 파일로 정리했다. 관측 기간은 2019년부터 2022년으로 설정하였으며, 수질 데이터를 제외한 나머지 자료는 모두 일 단위로 구성되었다.

1) 수문자료

- 환경빅데이터플랫폼, 한국수자원공사, 『다목적댐 운영 정보(일자료)』
- 물정보포털(MyWater), 운영현황_댐/보 일별 수문자료

강우량, 유입・방류량, (현재) 저수량, 저수율(%)을 수집하였다. 강우량 결측치는 해당 연도・월의 최빈값으로 보완하였다.

2) 제원 정보

- 한국수자원공사, 사전정보공표_다목적댐 및 용수댐 현황
- 한국수자원공사, 댐관리규정(2015)_제2장

다목적댐 21곳의 총저수량, 유효저수량, 홍수조절용량, 비활용용량을 수집하였으며, 저수지 용량 배분에 따라 이수용량을 계산하였다.

3) 기상자료

- 환경빅데이터플랫폼, 한국수자원공사, 『 관측소별 기상관측정보』
- 국가수자원관리종합정보시스템, 수문기상_실시간 기상자료
- 농촌진흥청, 기상모니터링_기상통계
- 국립농업과학원, 농업 기상정보_주산지 기상분석
- 한국농촌경제연구원, 관측 기상_지역별 기상정보

다목적댐이 위치한 지역 혹은 인근 지역의 평균습도, 평균기온, 평균풍속, 합계일사량을 수집했다. 결측치는 습도와 일사량은 해당 연도/월의 최빈값으로, 온도는 평균값으로 대체하였다. 풍속은 봄과 가을에는 평균값으로, 여름과 겨울에는 최빈값으로 처리했다.

4) 수질 자료

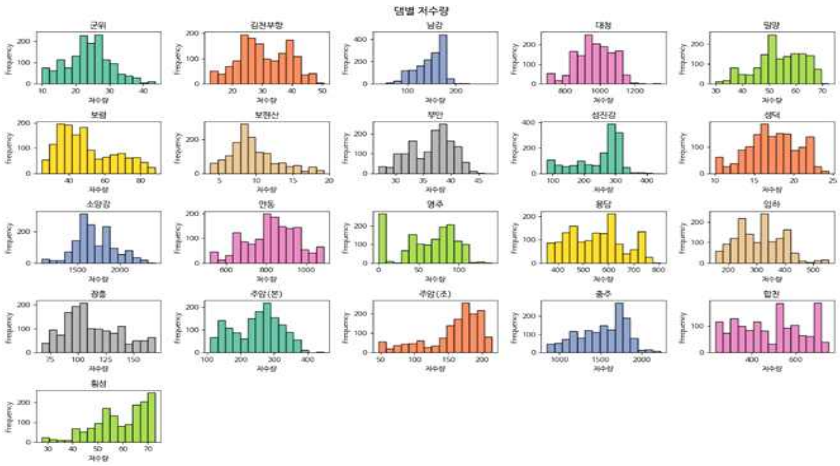
- 물환경정보시스템, 수질측정망 자료조회(호소)
- 물정보포털(MyWater), 운영현황_다목적댐 수질자료
- 환경부, 「수질오염 실태 보고」, 호소의 수질현황

월별로 10가지 검사항목을 수집하여 호소의 생활 환경기준에 따라 7개 등급으로 분류하였다. 등급 산정이 어려운 경우 총유기탄소량(TOC)만을 사용하여 등급을 결정하였다. 만약 특정 월의 데이터가 모두 결측치라면 해당 댐의 등급 최빈값으로 대체하였다.

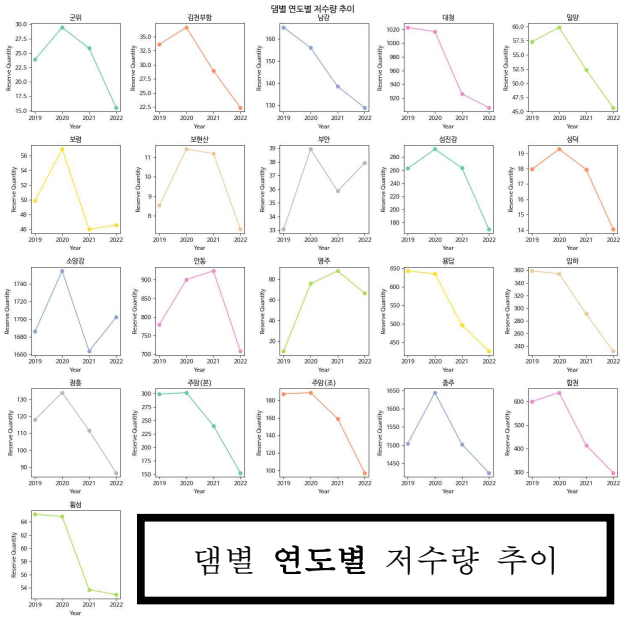
Ⅲ. 주요 내용

1) EDA(데이터 탐색)

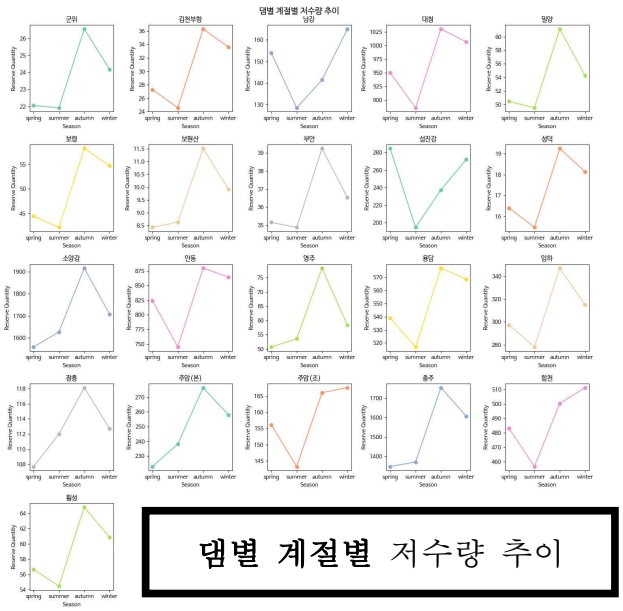
댐별 특징을 파악하기 위해 EDA를 수행하였다. 하지만 모델을 테스트할 때는 어떤 데이터가 들어올지 예측할 수 없기에, 학습용 데이터에 대해서만 진행하였다.



먼저, 저수량을 분석한 결과 댐 대부분에서 저수량이 거의 일정하게 유지되는 것을 확인했다. 또한, 데이터의 기술 통계량과 왜도 분석 결과 댐 대부분에서 비대칭성이 낮음을 확인했다. 하지만, 댐별 평균 저수량 값에 큰 차이가 있다는 점을 파악했다. 한편 연도별 또는 계절별로 저수량 추이를 분석했을 때, 댐에 따라 다양한 양상이 나타났다. 특히 여름에 비가 집중적으로 내리는 한국의 특성 때문에, 계절별 저수량 추이도 다르게 나타났다. 이를 바탕으로 연도와 계절을 새로운 변수로 추가하였다.



댐별 **연도별** 저수량 추이



댐별 계절별 저수량 추이

Feature 변수들에 대해서도 각 변수의 분포, 계절별 분포, 그리고 연도별 분포를 분석하고 시각화하였다. 이 과정에서 왜도를 파악한 후 기술통계량을 확인하여 변수들을 직접 비교하였다. 특히 강우량, 유입량, 그리고 방류량은 대다수 댐에서 왜도가 약 10 정도로 나타났으며, 이를 통해 해당 변수들이 왜곡된 분포를 가지고 있음을 파악할 수 있었다. 유입량과 방류량은 댐 간에 차이가 크다는 사실도 확인 가능했다. 한편, 평균습도, 평균기온, 평균풍속, 그리고 합계일사량은 댐별 차이가 크지 않았으며, 댐 내에서도 비교적 안정적인 수준을 보이는 것으로 확인되었다.

더 자세한 시각화 결과는 EDA.ipynb 파일과 태블로 자료를 통해 확인할 수 있다.

- K-Water DashBoard(태블로 시각화)

2) 군집 분석

댐마다 다양한 특징을 확인한 후, 유사한 특성을 가진 댐을 그룹화하기 위해 군집 분석을 수행하였다. 이에 KMeans 알고리즘을 사용하였는데, KMeans는 이상치에 민감하게 반응하므로 이상치 제거를 우선적으로 수행하였다. 평균습도와 평균기온은 데이터 특성상 이상치 발생 가능성이 상대적으로 적은 변수이기에, 이 두 변수에 대해서만 이상치 제거를 하였다. 다음으로, 2~5개 사이에서 군집 개수를 최적화하기 위해 각 경우마다 평균 실루엣 계수를 비교하였고, 그 결과 4개의 군집이 가장 적절한 것으로 확인되어 최종적으로 댐들을 4개의 군집으로 분류하였다.

• 그룹 1 (군위, 김천부항, 남강, 밀양 등 12개소)

그룹1의 경우 전체 21개의 다목적댐 중 12개가 속해있는 것을 보아 절반 이상의 댐들을 포함한 집단이다. 해당 그룹에 속한 댐들의 저수량 데이터를 살펴보면 다른 그룹들에 비해 저수량(reserve_qy) 평균값이 확연히 낮음을 알 수 있으며, 유입량, 방류량의 평균값 또한 4개 그룹 중 가장 낮은 것으로 나타난다. 또한 그룹1에 속한 댐들의 저수율 평균을 살펴보면 약 55%인 것으로 드러나는데, 이 또한 4개 그룹들 중 가장 낮은 수치이며 이에 따라 그룹1에 속하는 댐은 저수율이 더 이상 낮아지지 않도록 지속적인 관리가 필요하다고 볼 수 있다. 기상 데이터의 경우 다른 그룹들에 비해 눈에 띄는 특징은 없으나, 평균 풍속과 평균 습도가 비교적 일정한 것으로 나타난다.

• 그룹 2 (소양강, 충주)

그룹 2는 약 1600(백만m3) 정도의 평균 저수량을 가지며 전체 저수량은 2500(백만m3) 이상으로, 댐의 크기가 가장 큰 그룹이다. 그러나 저수율은 약 57%로, 저수량 관리에 지속적인 주의가 필요한 상황이다. 특히 2020년에 최고치의 저수량을 기록한 이후, 2021년부터는 점차적으로 감소하고 있다. 이러한 결과는 댐 관리 및 운영에 있어 주의가 필요하며, 상황에 맞는 대응이 이루어져야 함을 강조하고 있다.
유입량과 방류량 역시 다른 그룹에 비해 약 4배 정도 크다. 따라서 유입・방류량 관리에 더욱 관심을 기울여야 하는 지역이기도 하다.

기후 조건은 상대적으로 낮은 기온과 작은 강우량·일사량이 특징이다. 또한 습도와 풍속 변화가 큰 지역이다.

• 그룹 3 (섬진강, 용담, 임하, 주암(본댐), 합천)

그룹 3의 평균 저수량은 약 250에서 550 정도이며, 총저수량은 대략 450에서 820 정도 범위이다. 또한 그룹 내 저수량 차이가 가장 큰 편이다. 하지만 저수율은 주로 60% 미만이다. 특히 2020년에는 다른 그룹에 비해 높은 저수량을 기록했으나, 그 후 급격한 감소 추세를 보이고 있어 저수량 관리에 주의가 필요한 상황이다.

유입량·방류량의 경우 다른 댐들에 비해서는 작은 편이다. 강우량, 풍속도 마찬가지로 다른 그룹에 비해 상대적으로 작은 편이며, 그에 반해 습도와 일사량은 큰 범위를 보인다.

• 그룹 4 (대청, 안동)

저수량과 총저수량이 두 번째로 큰 댐들로서, 저수량이 안정적으로 관리되며 변동성이 크지 않다. 관측 기간 동안 댐의 저수율은 대부분 60% 이상으로 유지되었고, 이는 댐의 높은 저장 능력을 시사한다. 특히 그룹 4는 대규모 댐인 그룹 2와 비교하여 더 높은 저수율을 보이며, 이는 군집 분석 결과에서 중요한 차이점으로 나타나고 있다.

유입량과 방류량은 다른 댐들에 비해 상당히 큰 것으로 나타나며, 대부분이 10 이상의 값을 가진다. 더불어 유입량이 방류량보다 조금 더 큰 경향을 보인다. 습도는 일반적으로 다른 댐들보다 낮으며, 기온과 풍속은 조금 더 큰 값을 나타낸다. 강우량은 평균 수준을 보이며, 일사량은 큰 변동성을 가진 지역으로 나타난다.

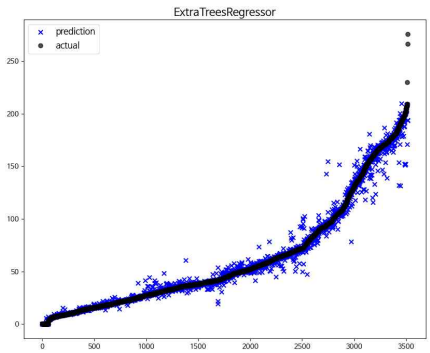
군집 분석 결과, 군집마다 고유한 데이터 특징을 확인하였다. 이에 따라 각 그룹의 차이를 최대한 유지하기 위해 군집 별로 전처리를 진행했다. 먼저, 변수들의 왜도를 평가하여 1 이상이면 로그 변환을 수행하였다. 이후 RobustScaler를 적용하여 데이터의 범위를 조정했다. 더 나아가, 범주형 변수에 대해서는 원-핫 인코딩을 하여 모델링에 더욱 적합한 형태로 변환하였다. 이러한 다양한 전처리 과정을 통해 모델 학습에 사용되는 데이터의 품질을 향상했다.

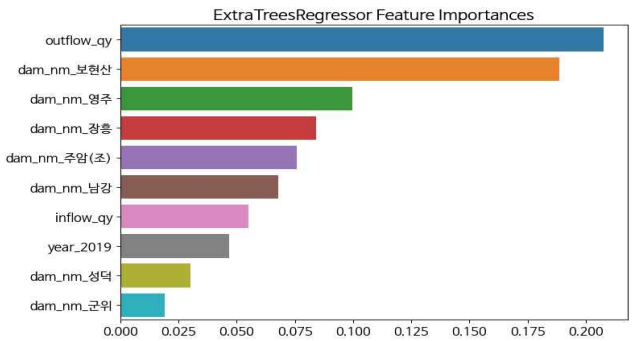
3) 회귀분석(모델링)

21개 다목적댐의 현재 저수량 예측을 위해 강우량, 유입량·방류량, 평균습도, 평균온도, 평균풍속, 합계일사량, 댐 이름, 연도, 계절 변수로 회귀 모델링을 수행하였다. Pycaret을 사용해 다양한 회귀 알고리즘의 성능을 평가하고, 평균 제곱근 오차(RMSE)를 기준으로 RandomForest, CatBoost, XGBoost, ExtraTrees, LightGBM 알고리즘을 선택하였다. 그룹별로 해당 알고리즘으로 모델을 구축하고, RMSE와 Adjusted R^2로 예측 성능을 평가했다. 또한, 최적의 모델로 테스트 데이터의 예측을 수행하였다. 아래는 그룹별 모델링 결과이다.

- 그룹 1 (군위, 김천부항, 남강, 밀양 등 12개소)

ExtraTreesRegressor의 기본 모델의 성능이 가장 좋았다.
(RMSE: 5.072, Adjusted R^2: 0.9904)

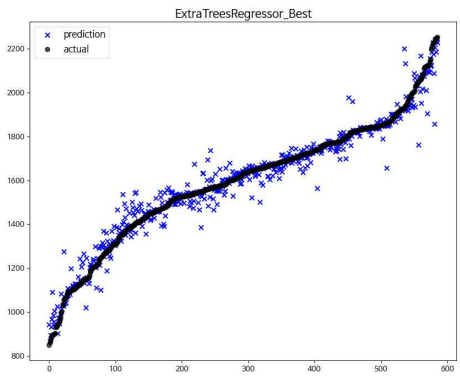


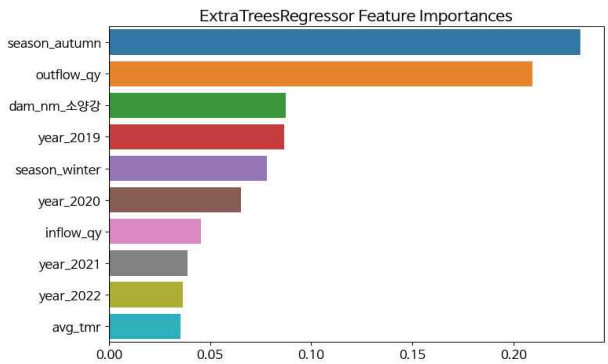


피처 중요도 분석을 통해 방류량이 저수량 예측에 큰 영향을 미쳤으며, 전체 21개 댐들 가운데 절반 이상이 속해있기 때문에 댐 이름 무엇인지 또한 저수량 예측에 상당한 영향을 주었음을 확인할 수 있다.

- 그룹 2 (소양강, 충주)

파라미터 튜닝된 ExtraTreesRegressor의 성능이 가장 좋았다.
(RMSE: 52.185, Adjusted R^2: 0.9651)

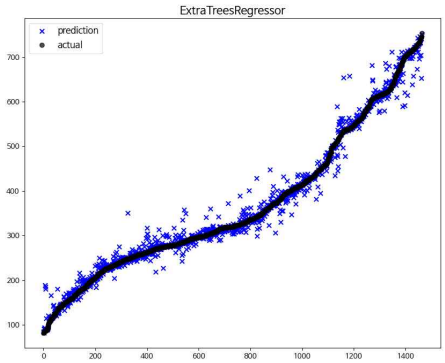


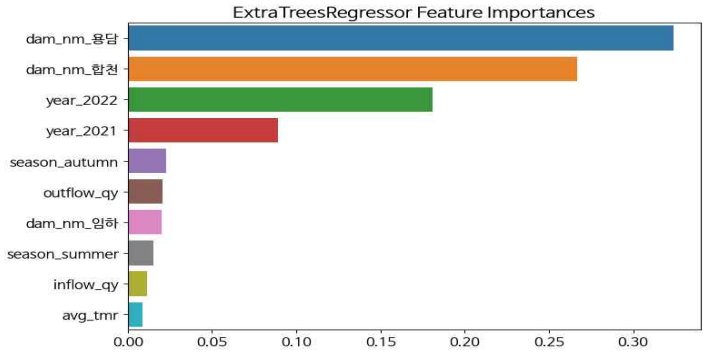


피처 중요도 분석을 통해 연도와 계절이 저수량 예측에 큰 영향을 미친다는 점을 확인할 수 있다. 또한, 유입량과 방류량이 댐의 저수량에 미치는 영향이 상당하다는 것을 확인할 수 있으며, 특히 방류량이 유입량보다 저수량에 미치는 영향이 더 크다는 사실도 확인할 수 있다. 마지막으로 기온 역시 댐의 저수량에 상당한 영향을 미치는 것을 확인할 수 있다.

• 그룹 3 (섬진강, 용담, 임하, 주암(본댐), 합천)

가장 성능이 좋았던 것은 ExtraTreesRegressor의 기본 모델이었다. (RMSE: 16.460, Adjusted R2: 0.9895)

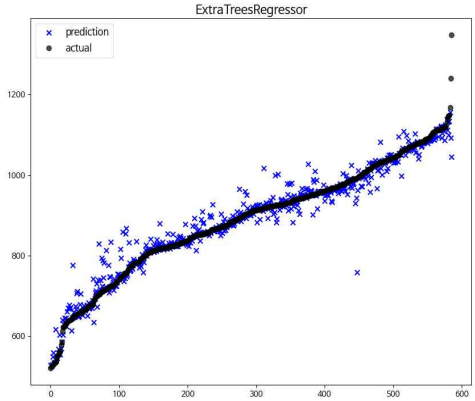


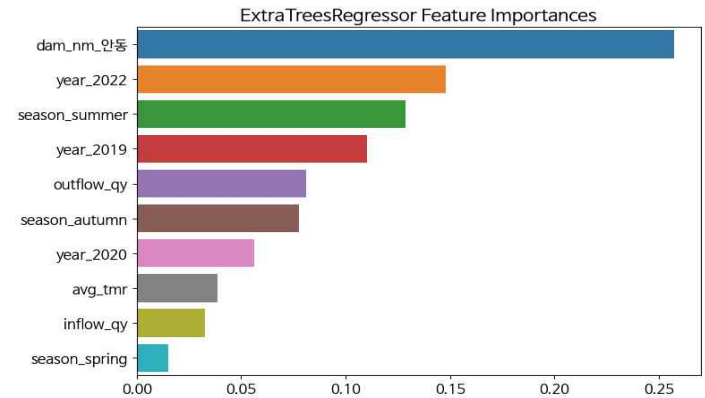


피처 중요도 분석을 통해 댐의 이름과 연도가 저수량 예측에 가장 큰 영향을 미침을 파악할 수 있다. 또한 다른 그룹들과 마찬가지로 유입량·방류량이 댐의 저수량에 미치는 영향이 크며, 평균 기온 또한 상대적으로 저수량 예측에 큰 영향을 미치는 것을 확인할 수 있다.

• 그룹 4 (대청, 안동)

파라미터 튜닝을 거치지 않은 기본 ExtraTreesRegressor의 성능이 가장 좋았다. (RMSE: 29.469, Adjusted R^2: 0.9556)



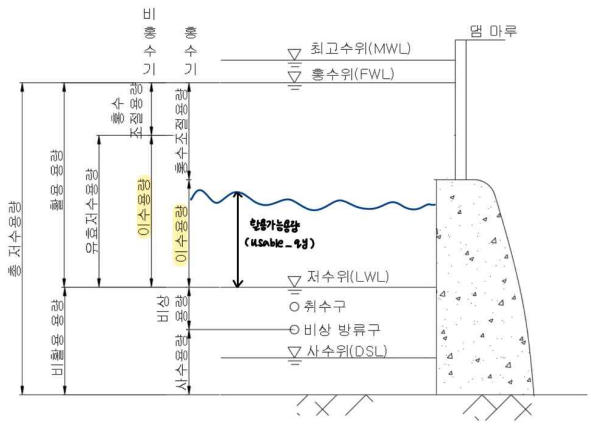


그룹 4의 경우, 댐마다 저수량 차이가 뚜렷함을 확인할 수 있다.

연도와 계절에 따라서도 저수량에 차이가 나타난다. 또한, 방류량이 댐의 저수량에 미치는 영향이 큰 것을 확인할 수 있으며, 기온 역시 댐의 저수량에 상당한 영향을 미치는 것을 확인할 수 있다.

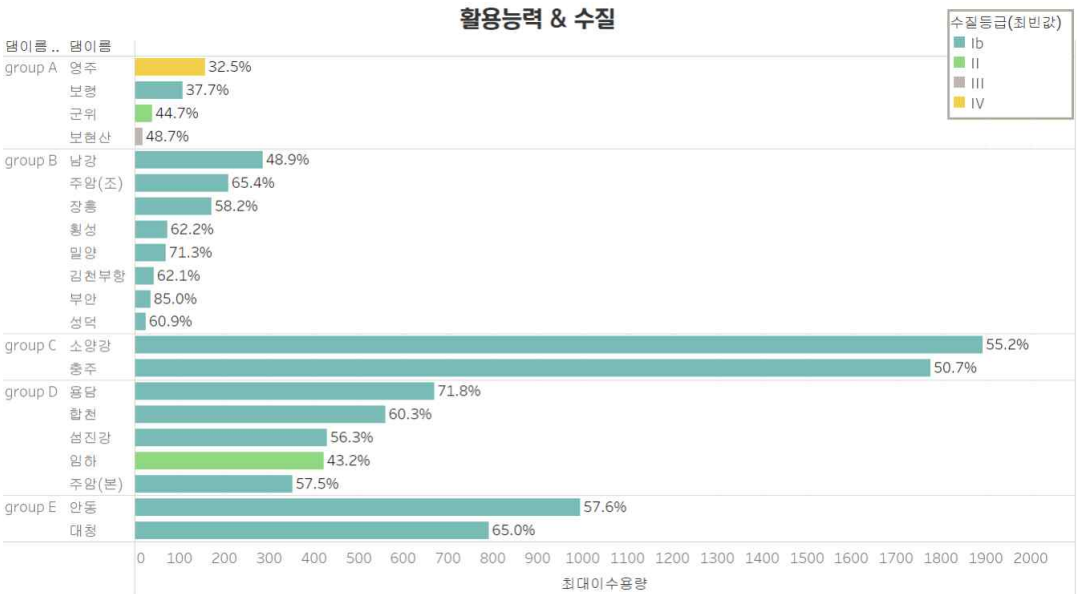
IV. 결과 및 기대 효과

댐의 가장 중요한 역할은 물을 저장 및 조절하여 필요할 때 활용할 수 있도록 하는 것인데, 전체 21개 댐별로 역할을 잘 수행하고 있는지 그 상태를 파악하기 위해 '활용 능력'이라는 지표를 산정하였다. 활용 능력은 (최대)이수용량에 대한 활용 가능 용량의 비율로, 댐마다 이수용량에 비해 실제로 물을 얼마나 활용할 수 있을지를 나타낸 지표이다.



[변수 설명]
- tot_qy: 총저수량
- valid_qy: 유효저수용량
- flood_qy: 홍수조절용량
- unused_qy: 비활용용량
- usable_qy: 활용 가능 용량
- maximum_use_qy: 이수용량
- efficiency: 활용 능력

모델링을 통해 저수량을 예측한 후 저수량, 수질 등급, 그리고 활용 능력에 따라 전체 21개 댐을 5개의 그룹으로 분류하였다.



• 그룹 A (군위, 영주, 보령, 보현산)

저수량이 다른 그룹에 비해 현저히 낮으며 수질 등급 또한 좋지 않다. 최대 이수용량 자체도 크지 않지만, 활용 능력 역시 약 30~40% 수준으로 활용 가능 용량의 비율이 작게 나타난다. 따라서 해당 그룹의 댐들은 저수량, 수질, 활용 능력 등 전체적인 부분에서 개선이 필요하며, 지속적인 관리가 이루어져야 한다. 특히 영주댐의 경우 수질 등급이 미흡하며 상당량의 오염 물질이 있기에, 수질 관리에 집중해야 한다.

• 그룹 B (김천부항, 주암(조), 장흥, 횡성, 밀양, 부안, 성덕)

그룹 B에 속한 댐들은 다른 그룹에 비해 저수량이 크게 낮거나 높지 않은 특징을 갖고 있다. 이 그룹의 수질 등급은 모두 Ib로 우수하며, 활용 능력 역시 50~70% 정도로 높은 수준을 보인다. 특히 부안댐은 평균 활용 능력이 85%로 매우 뛰어나다. 최대 이수용량은 다른 댐들에 비해 작지만, 활용 능력이 크기 때문에 이 댐들의 저수량을 효과적으로 관리하고 일정 수준을 유지하기 위한 노력이 필요하다.

• 그룹 C (소양강, 충주)

최대 이수용량이 가장 높은 댐들이며, 수질 등급 또한 모두 Ib로 우수하다. 다만, 댐들의 활용 능력은 약 50% 정도로 다른 댐들과 비교했을 때 우수하지 않은 편이다. 따라서, 양질의 물을 효과적으로 활용할 수 있도록 수문 관리 최적화를 위한 노력이 필요하다.

• 그룹 D (섬진강, 용담, 주암(본), 합천, 임하)

저수량이 보통 수준이며 수질 등급도 양호하다. 또한 그룹C와 달리 최대 이수용량이 큰 편은 아니지만, 활용 능력은 좋은 편이기 때문에 현재의 상태를 유지하는 데 집중하는 것이 중요하다. 다만 임하댐의 경우 수질 등급과 활용 능력이 다른 댐들에 비해 낮은 편이기에 지속적인 관리 및 개선이 필요하다.

• 그룹 E (대청, 안동)

그룹 E에 속한 댐들의 경우, 저수량도 그룹C와 함께 많은 축에 속한다. 수질 등급 또한 양호하며 활용 능력이 약 60% 정도로 5개의 그룹 중 가장 우수한 상태이다. 따라서, 안정적으로 양질의 이수 공급이 가능할 것으로 전망되는 지역이다.

특정 부분에 대한 개선이 이루어져야 한다거나, 현 상태를 유지하는 데 집중해야 하는 등 그룹 별로 처한 상황은 각각 다르지만, 모델링 결과를 바탕으로 추후 댐의 관리 방향을 설정하는 데 중요한 지침을 얻을 수 있었다는 점에서 유의미한 결과라고 해석할 수 있다. 이를 통해 지속적인 발전과 효율적인 자원 활용을 통한 더 나은 수자원의 미래를 향해 전진하는 데 기여할 것으로 기대한다.