|  |
| --- |
| **제9회「2021 빅콘테스트」데이터 분석 계획서** |

|  |  |  |  |
| --- | --- | --- | --- |
|  | \* 해당란에 체크 표시 | | |
| **참가분야** | □ 이노베이션분야 R 데이터분석분야 | | |
| **세부분야** | □ 루키리그 R 퓨처스리그 □ 챔피언리그 \*데이터분석분야에 한함 | | |
| **세부부문**  \*해당시 체크 | □ 지역활성화 □ 중소기업지원 □ ECO제주  R 홍수ZERO □ 스포츠테크 □ 수산Biz | | |
| **개인/팀여부** | □ 개인 R 팀(구성원 4 명) | **개인/팀명** | 홍수빅타 |
| **지도교사명** | - \*루키리그에 한함(선택) | | |
| **대표ID** | https://github.com/statco19 | | |

※ **5장 내외로** 목차는 준수하여 자유롭게 작성

|  |  |
| --- | --- |
| **분석 주제명** | 댐 유입 수량 예측을 통한 최적의 수량 예측 모형 도출 |
| **분석 배경** | 대한민국의 경우 여름철에 집중되는 강우로 인해 댐의 수위가 한 시기에 집중적으로 높아지는 현상이 매년 반복되고 있다. 원활한 댐 수위 관리와 댐 범람으로 인한 경제적, 사회적 그리고 환경적 피해를 예방하기 위해 댐의 수위를 예측하는 분석은 필수적이다.  이번 분석에서는 K-댐의 유입량에 영향을 주는 여러 요인을 종합적으로 분석하여, 최고의 유입량 예측도를 가지는 모델을 수립하고자 한다. |
| **분석 내용**  **요약** | 환경 빅데이터 플랫폼에서 제공하는 댐 수위 및 강우량 데이터를 사용하여K-댐의 홍수사상번호 26번에 해당하는 유입량을 예측하는 프로젝트이다.  홍수사상번호별로 구분 지어 시계열 데이터로 취급하여 분석하였으며, 수위(E지역)과 유입량의 상관관계가 높다는 점을 주의 깊게 살펴보았다. |
| **분석방법 및**  **계획** | * 데이터 탐색적 자료 분석(EDA) 및 시각화 방법 * 데이터 종류 및 결측치 탐색   1 ~ 26번까지의 홍수사상번호에 따른 6개의 데이터 집단이 주어짐.  모든 데이터는 정수형/실수형 데이터로 주어졌으며, 결측치는 존재하지 않았음.  Table  Description automatically generated   * 데이터 집단 간 데이터 비교   홍수사상번호 별로 데이터를 시각화 해봤을 때, 데이터 집단 간의 큰 데이터 차이를 발견할 수 없었음.  Chart  Description automatically generated  - 이상치 탐색  - 시계열 데이터 특성을 감안한 지수가중이동평균을 활용한 이상치 탐색  지수가중이동평균을 활용한 탐색에서는 이상치를 발견할 수 없었음.  Chart  Description automatically generated with low confidence    - 홍수사상번호별로 각 변수와의 상관관계 분석을 통한 이상치 탐색  Heatmap을 통한 홍수사상번호별 유입량과 변수와의 상관관계 분석을 통해 이상치를 발견할 수 있었음.  5번 홍수에서 수위(D지역)의 값이 모두 동일하여 상관계수가 계산되지 않았음.  21번과 23번 홍수에서 수위(E지역)의 상관계수가 타 집단과는 다르게 현저히 낮음을 발견함.    이상치로 간주되는 홍수의 경우 제외 및 포함하여 결과 비교를 통해 최종 제외 여부를 결정함.  Chart  Description automatically generated with low confidence   * 분석에 적용·활용한 통계·분석 기법 및 방법론   - 분석에 활용한 분석 모델  - Linear Regression  - Decision Tree Model  - XGB Regressor  - SGD Regressor  - Kernel Ridge Regression  - DNN  - 평가 지표 비교  총 4가지의 평가지표를 기준으로 각 모델의 예측력을 비교해볼 계획임.  - RMSE  - RMSLE  - MAPE  -  - 시계열 클러스터링  Silhouette score를 이용하여 군집의 개수를 설정한 이후, 수위(E지역) 그래프의 모양이 비슷한 것끼리 군집화를 진행함.  26번 홍수의 수위(E지역) 그래프의 모양이 다음 중 하나와 비슷하다면, 그 군집에 속한 데이터로 모델 학습을 시도함.  Chart, line chart  Description automatically generated  - Lagging을 활용한 데이터 shift  홍수사상번호마다 앞과 뒤의 데이터를 시프트하여 학습 모델이 주어진 데이터를 시퀀스 데이터로 인식할 수 있게 만들어 줌.  - 로그 변환  홍수 데이터 대부분의 경우 왼쪽으로 몰려있는 형태를 지니고 있음. 하지만 머신러닝은 데이터가 정규분포를 따를 때 학습이 더 잘 되기 때문에 데이터 로그 변환을 통해 최대한 정규 분포와 유사하게 만들어 줌.  - 제곱 값 추가  홍수사상번호마다 유입량과 수위(E지역)의 관계가 상대적으로 convex한 선형 관계로 나타남. 변환을 통해 좀 더 선형적인 관계를 뚜렷하게 갖게 함.   * 분석 결과에 대한 시각화 방법 * 파이썬의 Matplotlib, Pandas, Seaborn 등의 라이브러리를 활용하여 scatter plot, heatmap, line plot, bar plot 등으로 상관관계를 시각화하는 방법을 활용할 계획임. |
| **분석결과 활용 및 시사점** | RMSLE를 평가지표 중 하나로 사용함으로써 예측값이 실제값보다 더 작게 될 경우 큰 페널티를 부여하였다. 이는 유입량을 적게 예측해서 생길 수 있는 홍수 범람의 가능성을 낮추기 위해서이다.  홍수 유입량을 장마철에 적용하여 홍수 범람의 가능성이 있을 시 미리 예고함으로써 주변 지역의 농작물 및 인명 피해 최소화를 기대할 수 있다. |

※ 제출자료는 최종 출품작 평가시 활용될 수 있음