**문제해결 빅데이터 활용 프로젝트 기획안**

기획안 작성일자 : 2022. 09. 22

|  |  |
| --- | --- |
| 조 명 | 3조 : 그만피시조 |
| 조 원 | 조장: 성기호 / 조원: 민홍기, 정서연, 박준수, 이중훈 |
| 프로젝트 주제 및 개요 | 생체 지표를 통한 흡연 분류 모델 생성  (https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking) |
| 프로젝트 수행 방향 | ● 수행방향  1. 주제 선정 및 데이터 수집  2. 데이터 전처리  - 사용 데이터 전 처리  A. 데이터 도메인 분석  B. 분류형 데이터와 연속형 데이터의 처리  C. 특성 별 분포 확인을 위해 bar plot 형식으로 시각화  D. 각 수치의 이상치 분포를 확인하기 위해 box plot 으로 시각화  3. 분류모델 분담 및 조사 진행  - 분담모델 : Random Forest / Logistic Regression / XGBoost / LightGBM  / Support Vector Machine / Decision Tree  - 추가모델(미반영) : K-Nearest Neighbor / Naive Bayes,Adaboost / MLP  4. 모델링 학습  - 스케일러를 통해 데이터 스케일을 적용  - 각 모델 별 Hyper parameter tunning  - Feature 중요도 분석 및 이상치 처리  - 각 모델 별 최종 모델 선정  5. 평가 지표에 따른 학습 결과 분석  - 평가 지표 : ROC curve, Accuracy, Precision, Recall, F1 score  - 모델 별 평가 지표 비교를 위한 그래프 시각화  6. 최적 모델 선정  Decision Tree, Random Forest, XGB, LGBM와 같은 Tree 구조 모델, Logistic Regression, Support Vector Machine 모델에 대한 장, 단점을 비교한 결과, 모든 성능 수치에서 높게 나타난 LGBM을 최적 모델로 선정.  ● 수행도구, 데이터소개  - 사용언어 : python  - 수행도구 : pycharm, Jupyter Notebook  - 패키지 : Numpy, Pandas, Sklearn, Matplotlib, Seaborn |
| 프로젝트 조직  (구성원 및 역할) | ● 역할분담  - 조장 (성기호) : Decision Tree 모델 테스트 및 분석, 일정 조율 및 기획서 작성, 데이터 시각화  - 조원1 (정서연) : Random Forest 모델 테스트 및 분석, 데이터 전처리  - 조원2 (박준수) : Logistic Regression 모델 테스트 및 분석, 데이터 전처리, ppt 작성, WBS 관리  - 조원3 (민홍기) : XGBoost, Light GBM 모델 테스트 및 분석, 데이터 전처리, 발표 자료 작성 및 발표  - 조원4 (이중훈) : Support Vector Machine 모델 테스트 및 분석, 데이터 시각화 |
| 프로젝트 추진 일정 | ● 일정  9/20~ 9/22 : 주제 선정 및 주제 별 자료 조사  9/22 ~9/23 : 기획안 초안 작성 및 담당 모델 분배  9/23 ~ 9/27 : WBS 작성 및 일정 수립  9/27 ~ 10/4 : 모델 연구 및 테스트  (9/27~9/29 : 1차 제작 및 피드백, 9/29~10/4 : 2차 제작 및 피드백)  10/4 ~ 10/5 : 최종 모델 점검  10/4 ~ 10/6 : 발표 자료 정리 및 발표 준비 |