충북대학교 전자정보대학 컴퓨터공학과
[5110129] 오픈소스개발프로젝트
2022년 2학기 기말 프로젝트

<오픈소스 API를 활용한 데이터 수집 및 분석>

본 프로젝트는 데이터 및 주제를 자유롭게 선정하고 아래 단계를 거쳐 해당 데이터를 분석하는 것을 목표로 한다.

1. 데이터 수집

분석 데이터는 본인이 선택한 주제에 맞게 자유롭게 선택하되, 아래 두 가지 방법 중 하나를 이용하여 수집한다.

1) 케글(https://www.kaggle.com/datasets) 또는 K-ICT 센터 (https://kbig.kr/portal/kbig/datacube/dataset/info)에서 원하는 데이터셋을 다운로드

2) 주제에 맞는 웹사이트로부터 웹 크롤링을 통하여 데이터를 수집하여 자신만의 데이터셋을 만들어 활용 (컬럼 3개 이상의 데이터셋)

2. 데이터 분석

Pandas 라이브러리를 사용하여 데이터를 처리, 분석 및 도식화한다. 분석 내용 및 그래프 난이도 등에 따라 가산점 부여 가능.

필수 수행 요소 :

1) groupby를 사용한 데이터 그룹화 후 간단한 통계 분석 (메소드 3가지 이상 사용)

2) matplotlib을 활용한 그래프 그리기 (그래프 종류는 자유, 그래프 2가지 이상)

3) 머신러닝 기법 1개 이상을 사용한 모델 학습 및 모델 평가 수치 계산 (모델의 성능이 좋을 필요는 없음)

3. Github 저장소 활용

2의 과제를 수행하고, 코드를 자신의 개인 repository를 만들어 업로드한다.

4. 결과 해석 및 응용 방향 설계

2번으로부터 나온 데이터셋 분석 결과를 해석해보고, 데이터를 활용하여 어떤 응용 어플리케이션을 만들어볼 수 있을지 설계한다. *어플리케이션이나 시스템을 실제 제작할 필요는 없음. 구체적이고 명확한 설계일수록 높은 점수 부여.

5. 보고서 제출

2, 3, 4번 수행 결과를 정리한 보고서 제출 (자유 형식)

(2번: 표 및 그래프, 3번: 업로드 후 github 캡쳐화면, 4번: 글)

3번에서 자신의 과제를 저장한 Github 저장소 URL를 보고서에 반드시 기입할 것!

<프로젝트 예시(동일하게 하지 말 것)>

1. 심장질환 발생 예측 데이터셋
   (https://www.kaggle.com/fedesoriano/heart-failure-prediction)
2. (1) 각 환자의 심장질환 발생 여부(HeartDisease)를 기준으로 한 환자의 나이(Age), 최대 심박수(maxHR), 콜레스테롤 수치(Cholesterol) 통계량 계산 및 그래프
   (2) 나이 등 환자의 특성에 따른 심장질환 발생 여부를 예측하는 모델 학습 및 모델의 accuracy, F1-score 등 계산
4. 환자의 검진 정보 입력 시 심장질환 등의 질병 발생 예측이 가능한 웹서비스 구상 (시스템 아키텍처, 현재 사용한 데이터에 추가될 수 있는 feature들, 사용 가능한 머신러닝 모델 종류 등)