**2021 1학기 인공지능 기말 프로젝트**

1. **과제 공지**
   1. 마감 기한: 2021년 6월 18일 금요일11:59 PM 블랙보드에 제출
   2. 기말고사 대체 과제
   3. 이전 김현철 교수님의 기계학습 수업을 수강하여 수행한 프로젝트 재사용 금지 (채점시 이전 제출물과 대조하여 표절 검사 수행함)
2. **개요**
   1. 두개의 전처리가 완료된 데이터 셋 제공
   2. 제공된 데이터 중 하나를 선택하여 세개의 분류 모델 설계, 데이터에 대한 모델 성능 비교
   3. 데이터 가공 및 수정 불가
   4. 과제에서 시도하여야 하는 분류 모델:
      1. Decision Tree
      2. Logistic Regression
      3. Multilayer Perceptron(MLP)
3. **사용 데이터 셋 (하나 선택하여 진행)**
   1. Titanic data
      1. y: Survived
      2. 원출처: https://www.kaggle.com/c/titanic
      3. 전처리된 데이터: https://raw.githubusercontent.com/inikoreaackr/datasets/main/titanic.csv
   2. Wine data
      1. y: best\_quality
      2. 원출처: https://archive.ics.uci.edu/ml/datasets/wine+quality
      3. 전처리된 데이터: https://raw.githubusercontent.com/inikoreaackr/datasets/main/wine\_quality.csv
4. **과제 도구**
   1. 권장: python 기반의 Jupyter Notebook 환경(Google Colab), scikit-learn 라이브러리
   2. 그 외 python에서 사용 가능한 기계학습 라이브러리 사용 가능, 본인이 직접 구현한 기계학습 코드 사용 가능
   3. GUI기반의 기계학습 도구(Weka, Orange, RapidMiner 등등) 사용 금지
5. **실험 진행 및 보고서 내용**
   1. 실험 내용에 대한 전체 요약
   2. 데이터 원 출처에서 데이터 설명을 읽고 숙지한 후, 선택한 데이터에 대한 explanatory data analysis (feature 및 instance 개수, 각 feature의 의미, 분류 목적, 데이터 분포 등)
   3. 실험 설계 및 방법(진행 내용은 최대한 구체적으로 작성)
   4. 비교를 위해 시도한 trial 및 error에 대한 설명 (Default 모델만 사용하는 것이 아니라, 모델의 parameter 설정에 다양한 변화를 주는 과정이 포함되어 있어야 함. 그리고 validation을 통해 **각각의** 모델에 대해 최적의 parameter을 선정하는 과정이 포함되어 있어야 함.)
   5. 실험 결과 비교
   6. 데이터 셋을 train set, validation set, test set으로 구분하여 사용해야 함. Validation 과정은5-fold 또는10-fold cross-validation을 사용해 검증해야 함. **최종적인 모델 평가는 반드시 test set을 통해 평가.**
   7. 결론: 선택한 데이터에 대하여 모든 모델의 최종 성능을 자유롭게 비교
6. **제출물**
   1. 보고서 및 프로그램 코드 제출
   2. 사용한 모델 및 코드는 모두 채점할 때 실행 가능해야 함.
   3. Jupyter Notebook 환경에서 작업 시 .ipynb 파일에 보고서 작성(markdown cell로 작성, 보고서 코멘트(#)로 작성 금지) 및 프로그램 코드를 동시에 제출 가능. 이 때, .ipynb 파일에 자신이 코드를 실행해서 나온 결과물이 같이 출력되어 있어야만 함. (결과물 출력이 없으면 해당 항목에 대해 채점X)
   4. 기타 환경에서 작업 시, 프로그램 코드 및 보고서를 압축하여 제출(보고서 양식은 따로 없음). 이 때, 자신이 실행한 실험의 결과물을 반드시 캡쳐하여 보고서 내에 포함시켜야만 함.
7. **과제 수행 시 순서**
   1. 데이터 셋 선정
   2. 데이터 셋 파악 및 분석
   3. 데이터 셋 전처리
   4. 데이터 셋 분리(train-test split)
   5. 각 분류 모델에 대하여
      1. 모델 생성 및 훈련
      2. 모델 성능 평가 (cross validation)
      3. i, ii단계를 반복하며 최적의 모델 파라미터 선정, 최종 모델 선택
      4. test셋에 대하여 최종 모델의 성능 평가
   6. 모든 분류 모델의 test 성능 평가 비교 및 분석
8. **Help Session**
   1. 데이터 셋 선정에 어려움을 겪고 있거나, 과제 진행에 궁금한 점이 있거나, 진행 상황에 대해 상담이나 논의가 필요한 경우 등등, 도움이 필요한 경우 조교 이메일([koreaAIassistant@gmail.com](mailto:koreaAIassistant@gmail.com))에게 메일을 보내서 해결할 수 있도록 하기 바람.