2024 기계학습 텀프로젝트 안내

주요 일정:

일정	일시
텀프로젝트 phase 1 마감	11 월 25 일(화)
텀프로젝트 phase 2 마감	12 월 5 일(목)
기말고사	12 월 17 일(화) 오후 3 시-4 시 15 분
텀프로젝트 보고서 제출	12 월 23 일(월) 오후 11 시 59 분

- 위 일정은 원활한 시간 분배를 위한 가이드이다. 페이즈 1 와 2 의 결과물은 제출하지 않으며, 12 월 23 일까지 전체 보고서를 블랙보드에 제출해야 한다.

텀프로젝트 개요:

1. Phase 1:
   1.1. 데이터 선정
   1.2. 문제 설정 및 전체적인 과제 설계
2. Phase 2:
   2.1. 데이터 탐색 및 분석 (EDA)
   2.2. 데이터 전처리
   2.3. 여러가지 모델 실험
   2.4. 모델 파인튜닝 및 최종모델 선정
3. 보고서 작성

상세 진행내용:

- 해당 단계의 프로젝트 진행 과정에서 다음과 같은 질문에 대한 답을 고려해야 하며, 해당 답과 분석을 보고서에 포함해야 한다.

Phase 1:

1. 데이터 선정:
   1.1. **[보고서 포함]** 기계학습으로 해결하고자 하는 문제(과제/목표)을 설정한다.
   1.2. **[보고서 포함]** 관심있는 도메인의 아래 제한조건을 만족하는 데이터를 탐색하고, 설명과 출처를 작성한다.
   - 공개 데이터셋 참고 사이트:
     - UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/)
     - KD Nugget (https://www.kdnuggets.com/datasets/index.html)
     - Kaggle (https://www.kaggle.com/)
     - Google Dataset Search (https://toolbox.google.com/datasetsearch) 등등
   - 데이터셋 제한 조건:
     - 최소 10 개 이상의 feature
       - 전처리 과정에서 사용하기 어려운 feature 을 제거한 후, 최소 10 개 이상의 feature(attribute)을 가지고 있어야 함. (categorical dataset 을 one-hot encoding 로 처리한 경우에도 여전히 1 개의 feature 로 count 함)

- 최소 5000 개 이상의 instance
- 금지 데이터셋: KDD Cup 1999, Mushroom dataset

- 참고용 sample dataset:
  - Adult Dataset (https://archive.ics.uci.edu/dataset/2/adult) : 14 개의 feature, 약 5 만개의 instance
  - League of Legends Diamond Ranked Games (https://www.kaggle.com/bobbyscience/league-of-legends-diamond-ranked-games-10-min): 40 개의 feature, 약 9 천 8 백개의 instance
  - Video Game Sales (https://www.kaggle.com/datasets/gregorut/videogamesales) : 11 개의 feature, 약 1 만 6 천개의 instance
  - Mobile App Store (7200 apps) (https://www.kaggle.com/datasets/ramamet4/app-store-apple-data-set-10k-apps) : 17 개의 feature, 약 7 천개의 instance

2. 문제 설정 및 전체적인 과제 설계:
   2.1. **[보고서 포함]** 해당 문제에서 목표를 만족하도록 기계학습 모델을 어떻게 사용할 수 있는가?
   2.2. 해당 문제에서 현재 사용되고 있는 방법이나 해결책이 존재하는가?
   2.3. **[보고서 포함]** 이 문제를 어떻게 정의해야 하는가? (지도/비지도 등, 회귀/분류/군집 등)
   - 한 데이터셋으로 지도/비지도 등, 회귀/분류/군집 등 다양한 모델을 구현할 수 있음.

   2.4. **[보고서 포함]** 모델의 성능은 어떠한 지표로 측정할 것인지 3 개의 성능 지표를 선택한다. 데이터셋의 도메인에서는 왜 선택한 지표들로 성능을 측정해야 하는가? 각 지표가 의미하는 바는 무엇이며,
   2.5. **[보고서 포함]** 성과 측정 방법이 문제의 목표와 일치하는가? 목표를 달성하기 위해 필요한 최소한의 성능은 어느정도 인가?

Phase 2:

3. 데이터 탐색 및 분석 (EDA) (권장: 랜덤시드 설정, 데이터 복사하여 사용):
   3.1. **[보고서 포함]** 선택한 데이터셋의 각 feature 의 설명 및 유형(categorical (ordinal/nominal) / numerical (continuous/discrete)), 그리고 instance 개수는 무엇인가?
   3.2. 데이터셋의 각 속성을 조사:
   - **[보고서 포함]** 데이터 결측치, 이상치 확인
   - 모델에 포함 여부 (유용성)
   - **[보고서 포함]** 데이터 분포 시각화 및 분석
   - **[보고서 포함]** 특징 간 상관관계
   - **[보고서 포함]** 특징 별 적용할 전처리 결정
   - 특징들을 조합하여 새로운 특징 생성 가능성 탐색

4. 데이터 전처리 (권장: 랜덤시드 설정, 데이터 복사하여 사용, 진행한 모든 전처리 내용 **[보고서 포함]**):
   4.1. 이상치 수정하거나 제거
   4.2. 특징 스케일링 (표준화, 정규화 등)
   4.3. 연속형 특징 이산화(discretization)
   4.4. 범주형 특징 변환 (one-hot encoding)
   4.5. 학습에 도움이 되지 않는 특징 제외 (feature selection 방법 1 개 이상 적용)
   4.6. 특징들을 조합하여 새로운 특징 생성
   4.7. 전처리를 순차적으로 수행하는 파이프라인 생성

5. 여러가지 모델 실험 (권장: 랜덤시드 설정):

5.1. **[보고서 포함]** Baseline 모델을 정하고 그 모델보다 CV 성능이 개선된 모델을 찾는 과정을 실험
- 1 개 이상의 앙상블 모델을 포함한 총 5 개의 모델을 선택하여 실험 진행
- baseline 모델은 scikit-learn 의 기본 파라미터로 설정된 모델을 사용하지 않고, 본인이 모델 파라미터를 이해하고 이를 임의로 조정한 모델을 사용

5.2. **[보고서 포함]** 성능을 측정하고 비교 및 분석: 각 모델에 대해 k-fold CV 을 사용 (5 or 10)
- 아래 "보고서 작성 내용"의 "4. 실험 결과 분석"을 참고
- (Optional) 다양한 성능 지표에 따라 달라지는 성능을 분석하고 그 이유를 분석
- 5 가지 모델 훈련, 각 모델의 성능 비교, 이 어플리케이션에 가장 적당하다는 성능 지표를 3 개 정하고, 그것을 비교하교 작성

6. 모델 튜닝 및 최종모델 선정 (권장: 랜덤시드 설정):

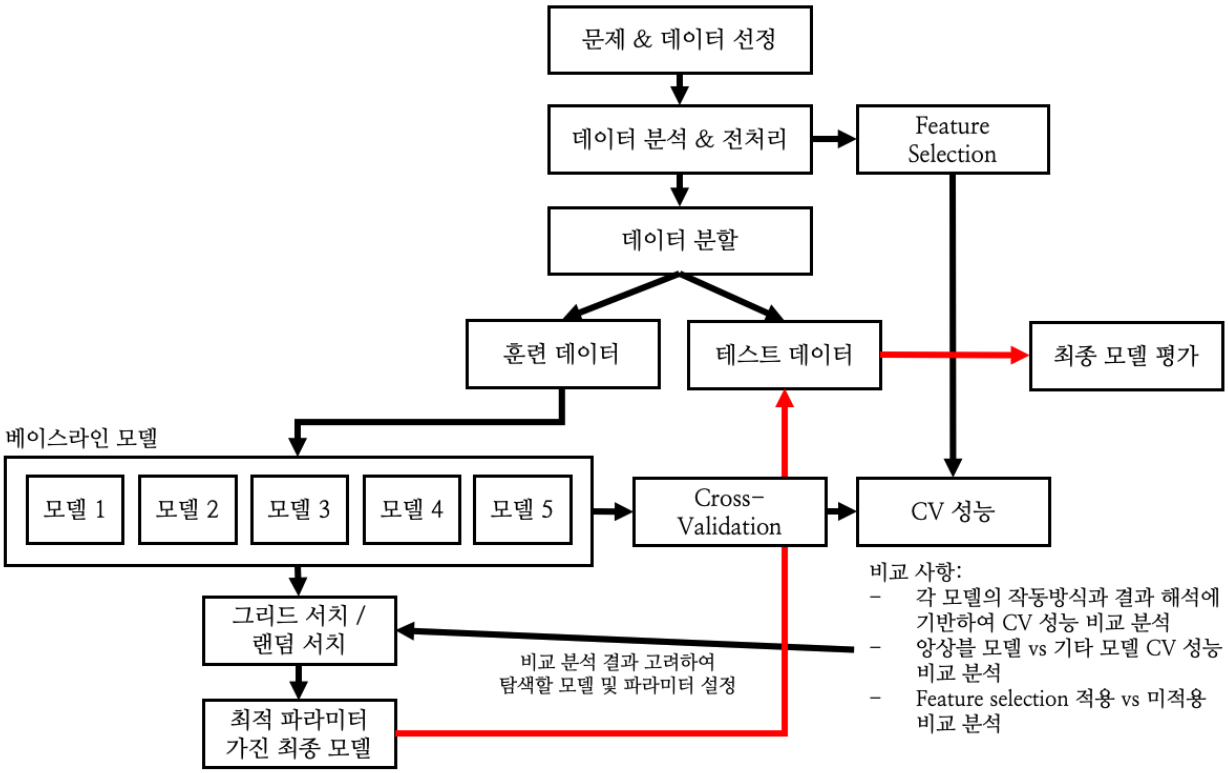
6.1. **[보고서 포함]** 그리드 서치, 랜덤서치를 사용하여 하이퍼파라미터를 조정 (CV 성능으로 평가)
- 단계 "5. 여러가지 모델 실험"의 비교와 분석을 고려하고, 비교 분석 결과를 반영하여 탐색할 1) 모델과 2) 하이퍼파라미터 유형과 범위를 설정
- 이 단계까지 test-set 사용 절대 금지

6.2. **[보고서 포함]** 최종 모델을 선정하여 test-set 성능을 측정
- 최종 모델은 그리드 서치/랜덤 서치(하이퍼파라미터 튜닝)을 진행하여 얻어진 모델이어야 함

보고서 작성 내용:

1. 실험 내용에 대한 전체 요약
2. Phase 1 에서 **[보고서 포함]**으로 표시된 내용 작성
3. 실험 설계 및 방법 (진행 내용은 최대한 구제척으로 작성)
    a. Phase 2 에서 **[보고서 포함]**으로 표시된 내용 작성
    b. 자신이 선정한 모델 및 모델 선정에 대한 이유
        i. 1 개 이상의 앙상블 모델을 포함한 총 5 개의 모델 선택
    c. 분석에 시도한 trial and error 에 대한 설명
        i. Default 모델만 사용하는 것이 아니라, 모델의 parameter 설정에 다양한 변화를 주는 과정이 포함되어 있어야 함. 그리고 validation 을 통해 각각의 모델에 대해 최적의 parameter 를 선정하는 과정이 포함되어 있어야 함
4. 실험 결과 분석
    a. 데이터를 train-set, test-set 으로 구분하고 validation 과정은 cross-validation 을 사용해 검증 (**최종 모델 평가는 반드시 test-set 을 통해 평가.)**
    b. 비교에 사용된 모델의 CV 성능을 비교하고(2.4 에서 선택한 성능 지표들 사용하여 비교), 성능에 차이가 있는 경우, 모델의 결과를 분석할 수 있는 경우 모델을 분석하고, 그렇지 않은 모델인 경우 데이터셋의 특성으로 유추해서 작성해야 함
    c. 앙상블 방법과 CV 성능 비교 및 분석 작성
    d. 사용한 모든 모델에 대하여, feature selection 기법을 적용했을 때와 적용하지 않았을 때의 CV 성능 차이를 비교하고, 차이에 대한 원인을 분석 또는 유추하여 작성하여야 함
    e. 위의 비교와 분석을 고려하고, 비교 분석 결과를 반영하여 탐색할 1) 모델과 2) 하이퍼파라미터의 유형과 범위에 대한 설명을 작성해야 함
5. 결론

(참고용) 실험 개요:



실험 도구:

- Python 기반의 Jupyter Notebook 환경(권장) : scikit-learn 라이브러리(권장), 그 외 python 에서 사용 가능한 기계학습 라이브러리 사용 가능, 본인이 직접 구현한 기계학습 코드 사용 가능(코드 도큐멘테이션 필수)
- GUI 기반의 기계학습 도구(Weka, Orange, RapidMiner 등) 사용 금지

제출물 형태:

- Jupyter Notebook 환경에서 작업 시 .ipynb 파일에 보고서 작성(markdown 으로 작성) 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 채점 X)
- 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.
- 제출물 제목은 **'학번_이름_프로젝트'** 로 한다.
    - Jupyter Notebook(or Google Colab) 환경에서 작업 시 '학번_이름_프로젝트.ipynb'로 저장하여 제출
    - 기타 환경에서는 '학번_이름_프로젝트.zip'으로 제출
        - (zip 내의 모든 파일 이름도 '학번_이름_프로젝트'로 동일해야 함)

성적 비중:

- 중간고사: 30%
- 기말고사: 30%
- 실습 및 과제: 20%

- 텀프로젝트: 20%

보고서 작성에 생성형 AI 을 사용할 경우 어떤 용도로 어떻게 사용하였는지 사사에 작성합니다.

본 보고서는 Microsoft Copilot (버전 GPT-4, Microsoft, https://copilot.microsoft.com/)을 사용하여 초기 메모를 요약하고 최종 초안을 교정했음을 밝힙니다.