라인웍스 머신러닝 엔지니어 과제

제공된 데이터를 이용해 머신러닝 모델을 만듭니다.

과제의 목적은 머신러닝을 이용하여 문제를 해결하기 위한 실험 설계와 데이터 분석 능력을 보는데 있습니다. 최종 모델의 성능의 절대 수치는 중요하지 않습니다.

데이터

Synthea를 이용하여 가상으로 생성한 EMR 데이터를 Common Data Model로 변환한 데이터가 제공됩니다.

제공되는 테이블은 다음과 같습니다.

- person
- visit_occurrence
- condition_occurrence
- drug_exposure
- death
- concept

별도의 데이터 설명 문서를 확인하시기 바랍니다.

헬스케어 데이터에 익숙하지 않다면 다음 링크의 자료를 참조 바랍니다.
https://speakerdeck.com/hongwonjun/helseukeeo-deiteo-silseub

문제

주어진 데이터를 이용해 다음 4가지의 문제 중 하나를 선택해 모델을 만듭니다.

1. 재입원 예측 모델
    - 입원했던 환자가 퇴원 후, 30일 이내 재입원할 확률
    - 모델 사용 시점
        - 환자 퇴원 시점
2. 원내 사망 예측 모델
    - 입원 환자가 병원내에서 사망할 확률 예측
    - 모델 사용 시점
        - 환자 입원 시점
        - 환자 입원 후 24시간 시점
3. 입원일 수 / 장기(10일 이상) 입원 예측 모델
    - 입원 환자의 입원일 수 혹은 장기 입원 예측
    - 모델 사용 시점
        - 환자 입원 시점

        - 환자 입원 후 24시간 시점
4. 응급실 방문 예측 모델
    - 환자가 (외래, 입원, 응급을 포함한 모든) 방문을 종료 후, 30일 이내 응급실에 방문할 확률
    - 모델 사용 시점
        - 방문 종료 시점

결과물

- 실험 설계, 실험 내용, 전처리 방법, 모델 설명 등을 문서(ppt, doc, txt 등)로 정리합니다. 형식과 내용은 자유이지만, 문서에 반드시 포함되어야할 내용은 다음과 같습니다.

1. 사용할 데이터 및 Target Label 추출
    - 데이터 수
    - 환자 수
    - 라벨 비율
2. 검증
    - 선택한 검증 방법
    - 선택한 Metric
3. 모델 및 전처리
    - 선택한 모델
    - 전처리 방법
4. 실험 내용
    - 하이퍼 패러미터 변경 실험
    - 입력 데이터 변경 실험
    - 자율적으로 실험

- 주어진 데이터 파일을 입력해 최종 선택된 패러미터로 모델을 만드는 것까지 하나의 코드 파일(py, ipynb)로 정리합니다.