[TIL] 24/12/05_PPT 내용

PPT 구성 정리

분석 사항

- 어떠한 주제로 분석을 했는지에 대한 내용 작성

> 💡 주제 연관성
>
> - 항생제 사용을 통한 치료 성공률 분석에서 예상한 수치가 나오지 않음
> - MDR을 통한 감염 치료 효과를 분석하여 개선된 지표가 나왔으니, 다시 위의 주제에 반영한다면
> - 궁극적으로 치료 성공률 향상에 도움이 될 것

사용 데이터

- MIMIC III 데이터셋에 대한 설명
  - 의료 관련 기초 지식에 대한 설명을 덧붙여야 함
  - 데이터 특징 및 요약된 구조(모든 테이블에 대한 소개 지양, 주요 테이블만) 설명

> 주제1에 기반한 내용입니다.

데이터 전처리

- Intro: 사용한 데이터 테이블 소개 + 이유 설명
  - 테이블 간의 연관성 or ERD 이용

- 전처리

  0. 공통
     - ❄️ 결측치 제거
       - 추가 근거 필요 → 미믹 데이터가 가진 특성과 결합하는 것이 좋아 보임
     - 날짜 타입 정리 : Datetime

  1. ICUSTAYS + ADDMISSIONS 테이블 Left join → ICUSTAYS에 merge
     - 환자가 병원에서 사망했는지 여부에 대한 플래그인 `hospital_expire_flag` 추가
     - 필요한 컬럼만 유지

  2. 항생제 리스트업
     - V1: 공식 문서를 사용해 항생제 키워드를 30개로 정리 → 141개로 분류되는 문제가 있었음
     - V2: 재필터링을 통해 V1에서 고려하지 못했던 부분(이유: 동일 성분이나 목적에 따라 다르게 네이밍됨) 반영

  3. ICU 환자의 감염 여부 파악
     - LABEVENTS 이용

  4. 항생제 처방 여부 파악
     - 3번의 내용(환자 감염 관련 검사 결과)에 병합

Feature Engineering

- 항생제 치료 기간 계산
- ❄️ 감염 검사 결과 → 하위 2가지의 대한 당위성 필요
  - mean, max로 요약
  - 누락된 값은 0으로 대체

- 환자를 기준으로 데이터 정리
- 검사 요약 데이터를 환자 기준 데이터에 병합

Modeling

- 학습 데이터 정리
  - Feature(X)
    - `hospital_expire_flag` : Target 이기 때문에 정리
    - `subject_id`, `hadm_id` , `icustay_id_y`: 고유 ID(학습에서는 불필요)이므로 정리
    - `drug` : 약물 이름에 대해서까지 분류를 하진 않았기 때문에 정리
  - Label(y)
    - `hospital_expire_flag` : Target(이진)
      - 0: 생존
      - 1: 환자가 (병원 내에서) 사망

- ❄️ 모델 선정 → MIMIC 데이터의 특성에 맞게 이유를 설정하는 것이 좋아보임
  - 이진 분류 문제
    - **클래스 불균형** 문제로 인해 다수의 생존 데이터에 과적합될 가능성이 있음

      https://bluediary8.tistory.com/132

  - RF
    - 과적합 방지에 유리하다는 측면에서 사용하기 좋음
    - (클래스 불균형 문제를 해결하는 클래스 가중치 설정이 가능함 → 이 내용의 경우 실제 코드에서는 적용하지 않았기 때문에 언급하지 않는 것이 좋을 것 같기도....?)
  - XGB
    - 타당한 이유를 찾아야 함

성능 지표

- 키워드 필터링 과정에 따라 분류
  - 각 과정의 순서
    - 튜닝 전
    - ❄️ 튜닝 후(튜닝 값은 무엇인지, 더 성능이 개선된 이유는 무엇인지에 대한 당위성 필요)

주제2 내용 추가

-

회고

- 정규화 추가
  - 모델 학습 이전에 이진 분류의 특성에 따른 정규화 과정이 추가된다면 더 좋은 성능을 낼 수 있을 것으로 예상

- 리샘플링
  - 이진 분류의 클래스 불균형 문제를 해결하기 위해 오버 샘플링, 언더 샘플링, 하이브리드 샘플링 등의 작업을 진행한다면 → 성능 지표에 더 좋은 영향이 있을 것으로 보임