스토리텔링보다 프로젝트 본 목적에 집중하기

필요한 데이터셋만 갖춰서 분석하기 (급선무)

분석에 사용할 만한 데이터셋: 코로나가 포함되어있는 데이터셋

- 감염질병으로 사망한 경우 (코로나도 포함되어 있는지 확인) -> 사용해 분석

- 격리 병동 위치 정보

- 시도별 코로나 현황 (가장 적합한 데이터셋)

전처리하기

- 다른 질병 빼고 코로나만 데이터셋 정제

동까지 구하는건 불가능, 불필요해보임, 구까지 묶어서 병동으로 이동하는 경우가 있기 때문

목적: 지역별 코로나로 인한 사망, 위증증 분포 분석 -> 적절한 위치에 병상 확보

독립 변수: 코로나 사망자 수, 환자 수

종속 변수: 필요한 병상 수 (예측 모델에 필요한 데이터셋)

-> 맞춰서 전처리하기

전처리하며 사용할 모델 찾아보기

- 예측: decision tree, 회귀 분석

모델 돌릴 때 변수는 몇 개정도가 적절한가?

- 변수 개수가 중요 X, 종속변수와의 관계성을 얼마나 잘 보여주는가가 중요

- 필요없는 변수 포함되면 성능 저하될 수 있기 때문

- 결과 예측에 마이너스

- 개수보다 연관성 높은 변수 셀렉하기

- 2개 ~ 3000개

- 상관관계 보통 0.4 이상, 0.7 이상이면 높다고 판단

- 변수 수 부족하면 0.4, 충분하면 기준 높임

역할 분담

- 머신러닝 맡은 사람 중 후처리에 대한 역할은 필요 없음

- 데이터셋 전처리 / 모델 파라미터 조절하며 성능 체크 / 성능 시각화

- 기간 안겹침 -> 전처리부터 빠르게

- 모델 별로 들어가는 인풋 다름

- 모델 3개 정도 찾아서 1명씩 맡고 성능 비교해 시각화하기

예측으로는 시각화한적이 많이 없음

분류할 때 머신러닝 많이 쓰기 때문

예측 - 성능 그래프로 시각화

논문 많이 봐야할 듯

이번주안으로 무조건 데이터셋 전처리 끝내기 (모델 돌릴 수 있게)

다음주에 모델 성능 확인

하이퍼 파라미터 값 조정 -> 성능 향상

마지막 주에 시각화 + 스토리텔링