K-NET ML/DL STUDY GROUP MEETING #4

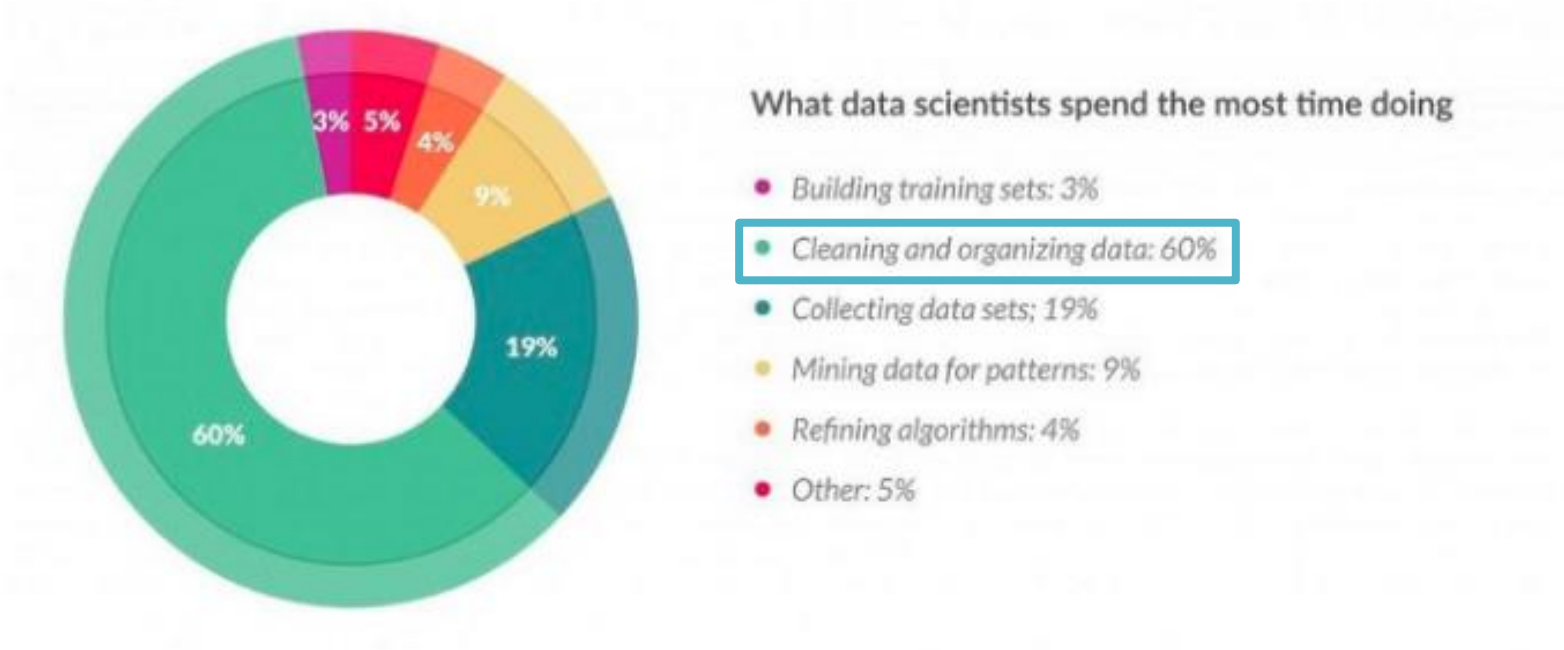
Hyungyu Kim
K-NET
2023-04-14

이번 시간에 다룰 주제

- 머신러닝 프로젝트 처음부터 끝까지

데이터 과학자를 가장 괴롭히는 것

- 데이터 전처리(데이터 품질을 높이는 것)에 가장 시간을 많이 사용



Source: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

데이터 전처리는 왜 중요할까?

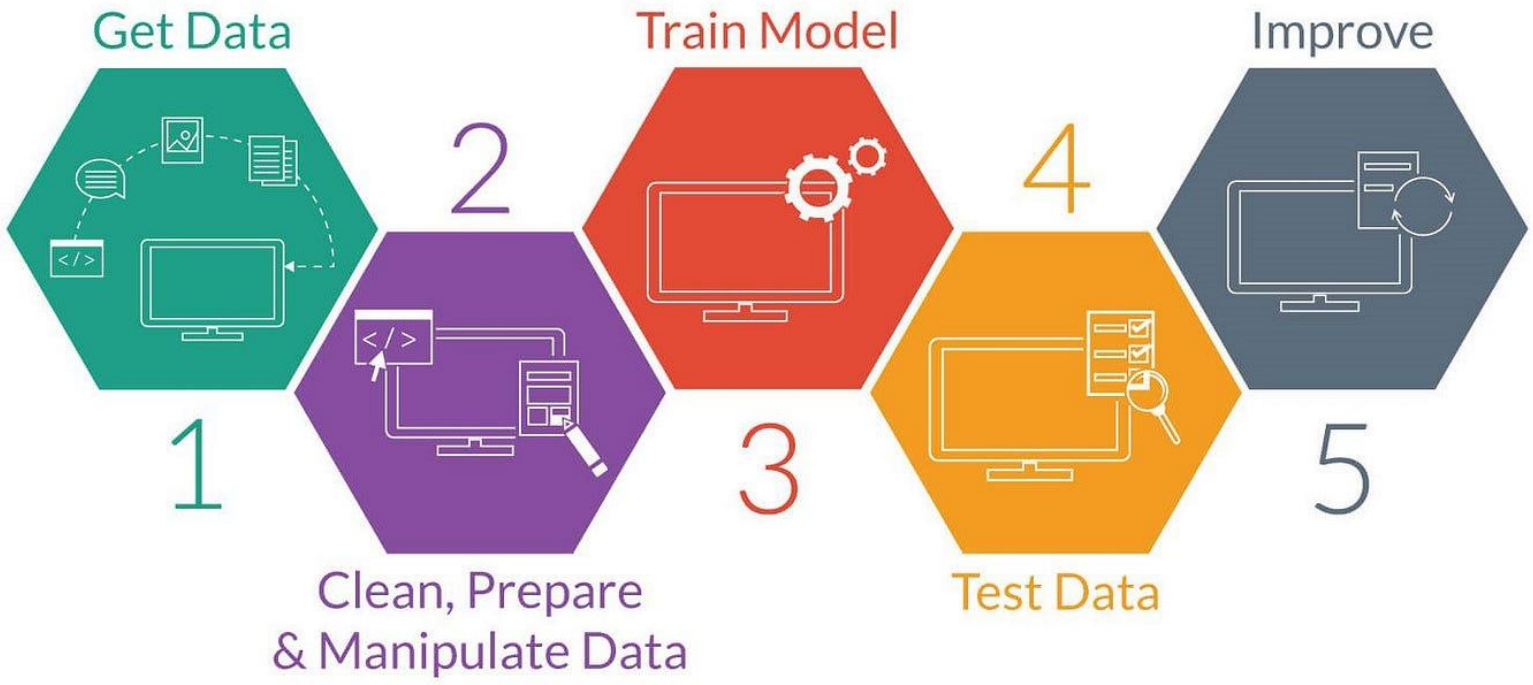
- 데이터 전처리를 하지 않으면 모델의 성능이 떨어지거나 잘못된 판단을 내릴 수 있다.
  - 결측치, 이상치, 잡음 처리 : 삭제/대체값 사용/가만히 놔두기
  - 피처(입력) 스케일링 : 피처가 갖는 값들의 범위가 서로 비슷해야 함
  - 범주형 변수 처리 : 원-핫-인코딩 등 숫자로 변환
  - 변수 선택(feature selection) : 예측력이 있는 변수 사용하기
  - 기타 도메인 지식 활용 : 새로운 특성(변수) 추출
- 따라서 탐색적 데이터 분석(EDA)을 통해 데이터를 자세히 들여다보는 과정이 필수

EDA를 할 때 중점적으로 볼 사항

- 각 변수의 유형과 속성 파악하기(범주형 or 연속형, 명목형 or 순서형)
- 결측치(누락된 값)가 있는지 확인
- 이상치(극단적으로 크거나 작은 값)가 있는지 확인 : boxplot 시각화로 확인 가능
- 피처(입력)와 타겟(출력) 간 상관관계 분석 : 예측력이 있는 변수를 선택
- 피처와 타겟의 분포 시각화하기 : 필요하다면 스케일링, 로그 변환 등을 실시
- 데이터 시각화: 히스토그램, 박스플롯, 산점도 등으로 데이터를 눈에 보이게 만들기
- 논외) 피처 간의 상관관계: 다중공선성 (ML보단 통계학에서 중요)

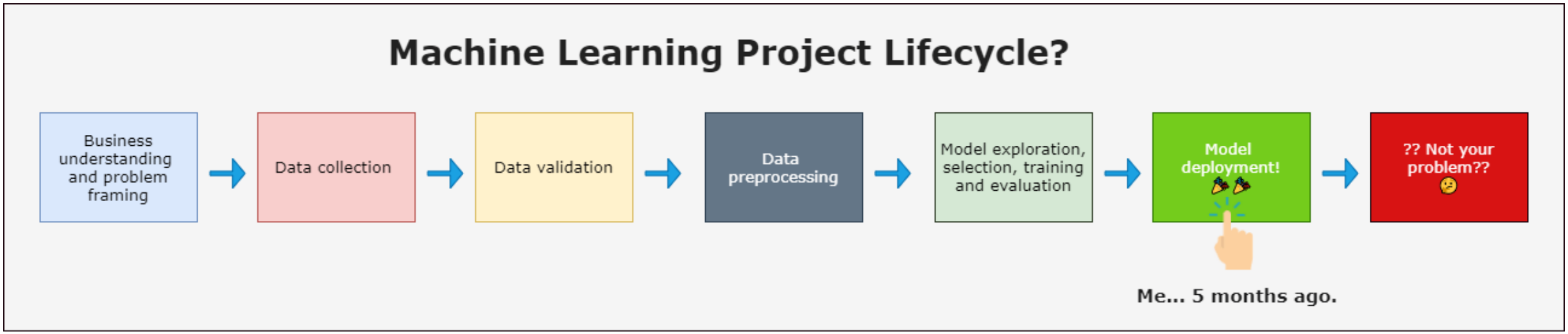
머신러닝 프로젝트의 순서

- 데이터 수집 – 데이터 전처리 – 모델 훈련 – 모델 평가 – 모델 성능 향상(모델 튜닝)
- 여기에는 중요한 과정 한 개가 빠져 있는데, 뭘까요?



머신러닝 프로젝트

- 바로 “문제 정의”입니다. 데이터 전처리와 함께 문제 정의도 아주 중요합니다.



프로젝트 예시 코드

- 이제 여러분이 기다리던 코드를 보겠습니다. 오늘 코드는 좀 어지러울 거에요
- 주의 사항 1: ML이 처음이신 분들은 오늘 코드를 너무 진지하게는 보지 마시고 차분하게 감상(?)하듯이 보세요.
- 주의 사항 2: 이제부터 나올 numpy, pandas, 기타 등등 라이브러리의 메서드를 처음부터 다 외우려 하지 마세요. 아시죠? ☺

요약

- 데이터 전처리가 왜 중요한지를 살펴봤습니다.
- EDA 시에 중점적으로 봐야 할 사항을 살펴봤습니다.
- 머신러닝 프로젝트의 일반적인 순서를 살펴봤습니다.
- 머신러닝 프로젝트의 예시 코드를 살펴봤습니다.

요청 사항

- 다음 시간부터 여러분들이 발표할 차례네요!
- 저희 스터디에서는 따로 교재/레퍼런스를 지정하지 않았으므로, ChatGPT의 답변이나 인터넷에 있는 모든 자료를 참고하셔도 됩니다.
- 발표 준비하실 때 너무 방황하지 마시라고 제가 참고용 가이드라인을 매 회차마다 드릴 예정입니다.
- 이것은 어디까지나 참고용이므로, 본인이 원하는 발표 방향이 있으면 그 방향으로 진행하셔도 됩니다.
- 발표가 끝나면 다른 분들은 Padlet에 발표자에게 적절한 피드백을 남겨주세요!

다음 시간에 다룰 주제

- 선형 회귀와 경사 하강법