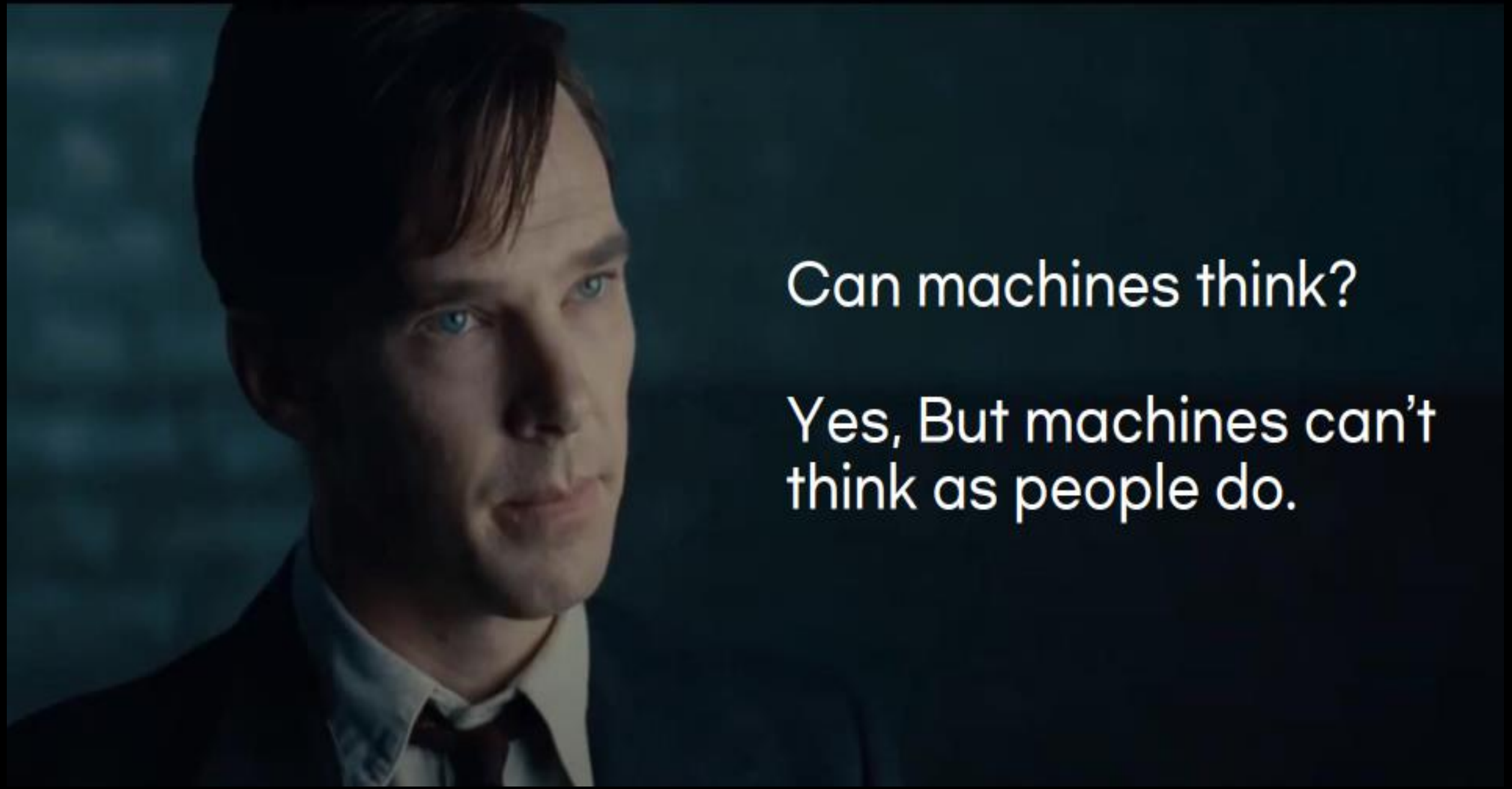
#5. 6장 국민청원 데이터 시각화와 분류

https://bit.ly/2023어쩌다텍스트분석



영화 : 이미테이션 게임

책의 6,7,8장의 목차를 살펴볼게요!

6장 국민청원 데이터 시각화와 분류

- 학습 세트와 시험 세트 만들기 / LightGBM

7장 '120다산콜재단' 토픽 모델링과 RNN, LSTM

- 학습-시험 데이터 세트 분리하기 / Bidirectional LSTM

지도학습

8장 인프런 이벤트 댓글 분석

- 군집화 하기 / KMeans

비지도학습

[6장]국민청원 데이터 시각화와 분류 : 머신러닝을 통한 텍스트 분류

모델 LightGBM

https://github.com/akngs/petitions

README.md

청와대 국민청원 사이트의 만료된 청원 데이터 모음.

데이터

petition.csv

- 전체 데이터

petition_corrupted.csv

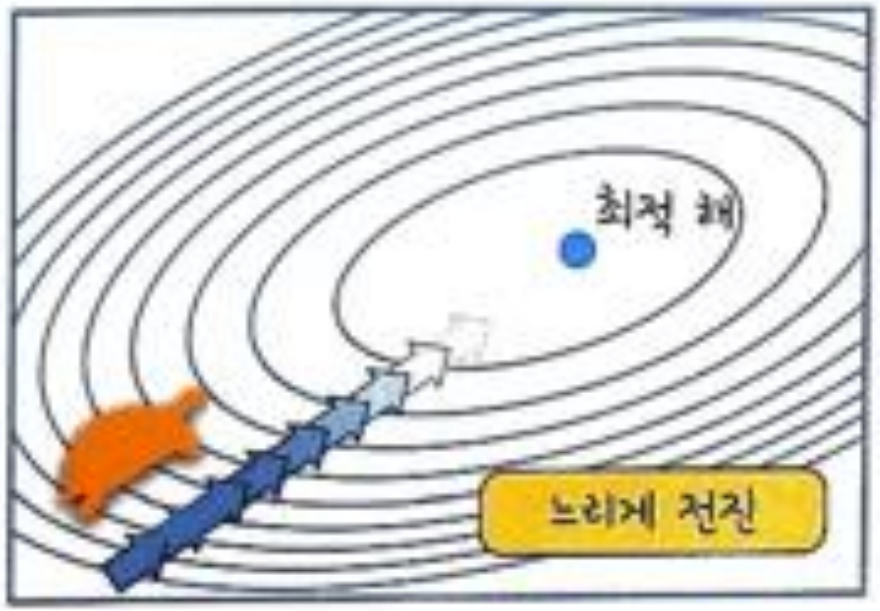
- 전체 행 중에서 5%는 임의 필드 1개에 결측치 삽입
- 범주(category)가 '육아/교육'이고 투표수(votes)가 50건 초과이면 20% 확률로 투표수에 결측치 넣기
- 나머지는 전체 데이터와 동일

petition_sampled.csv

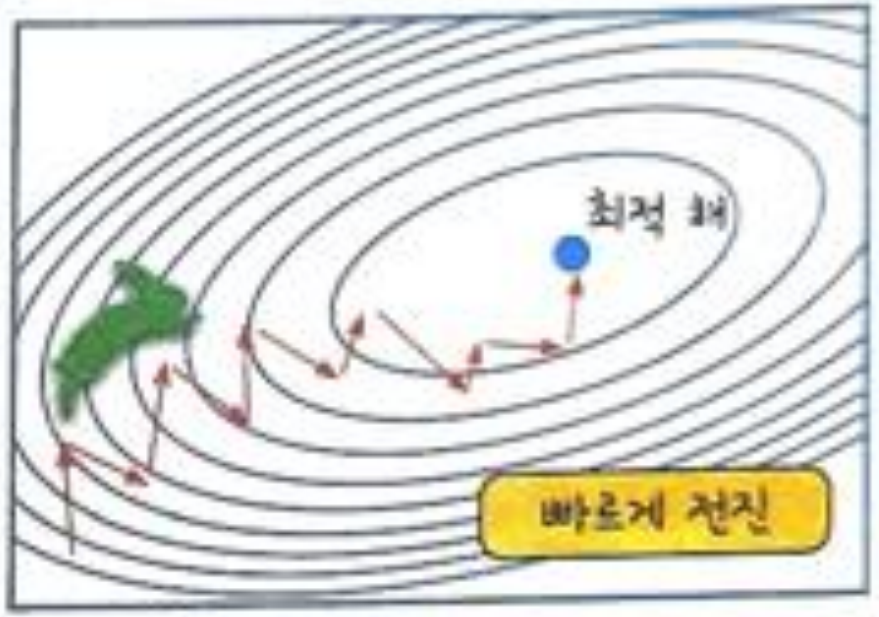
- 전체 데이터 중 5%만 임의추출한 데이터

오늘의 모델 : LightGBM

확률적 경사하강법(Gradient Decent)



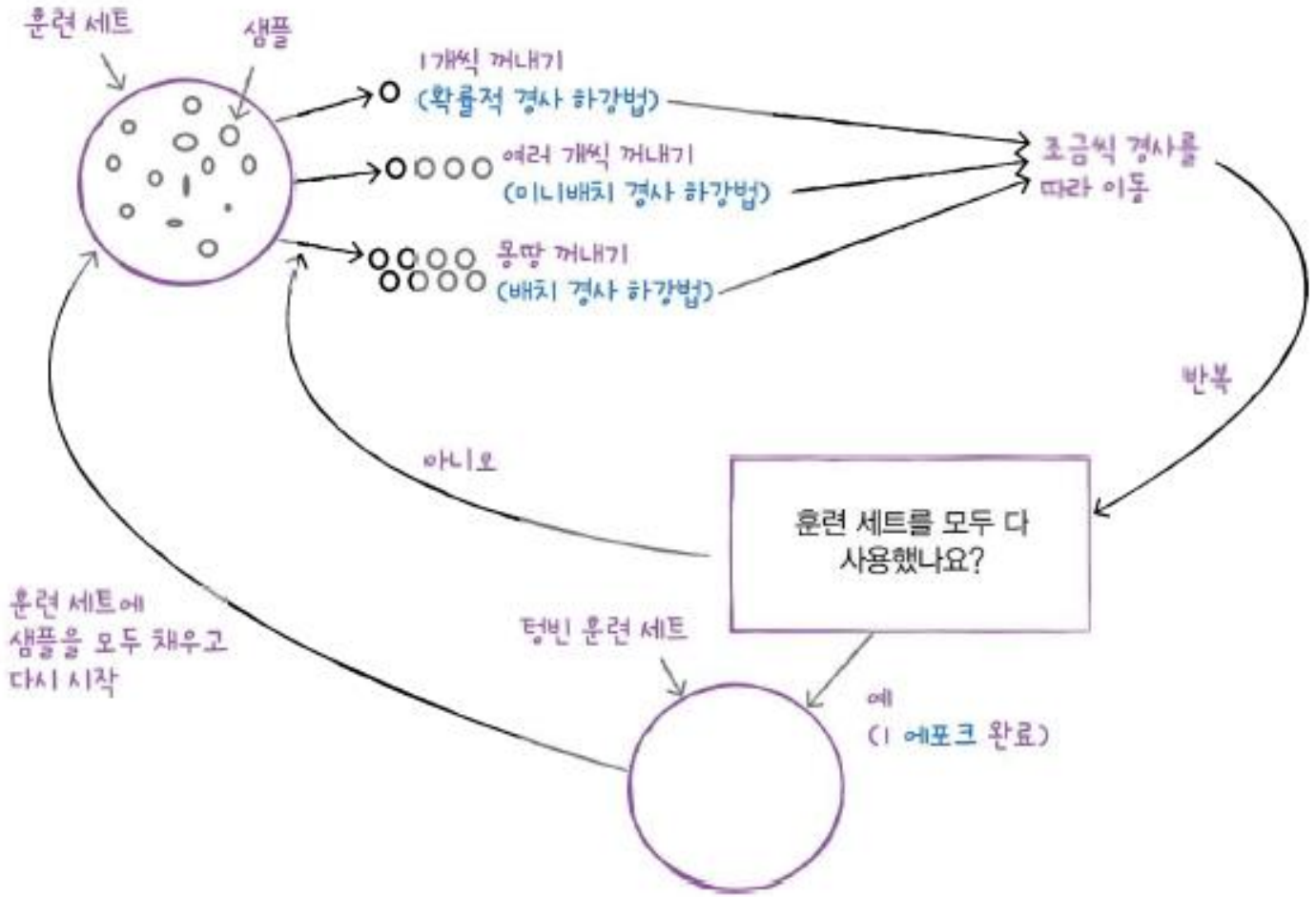
경사 하강법



확률적 경사 하강법

오늘의 모델 : LightGBM

확률적 경사하강법(Stochastic gradient decent)



오늘의 모델 : LightGBM

- 그레이디언트 부스팅(Gradient Boosting)

경사하강법을 사용하여(Gradient) 결정 트리를 계속 추가하면서 가장 낮은 곳을 찾아 이동하는 방법. 천천히 조금씩 이동해야 하므로 깊이가 얕은 트리를 이용함

- 히스토그램 기반 그레이디언트 부스팅

입력 특성을 256개 구간으로 나누고 256개 구간 중 하나를 떼어놓고 누락된 값을 위해서 사용

- LightGBM

MS에서 만든 히스토그램 기반 그레이디언트 부스팅 라이브러리

[6장]국민청원 데이터 시각화와 분류 : 머신러닝을 통한 텍스트 분류

모델 LightGBM

https://github.com/akngs/petitions

README.md

청와대 국민청원 사이트의 만료된 청원 데이터 모음.

데이터

petition.csv

- 전체 데이터

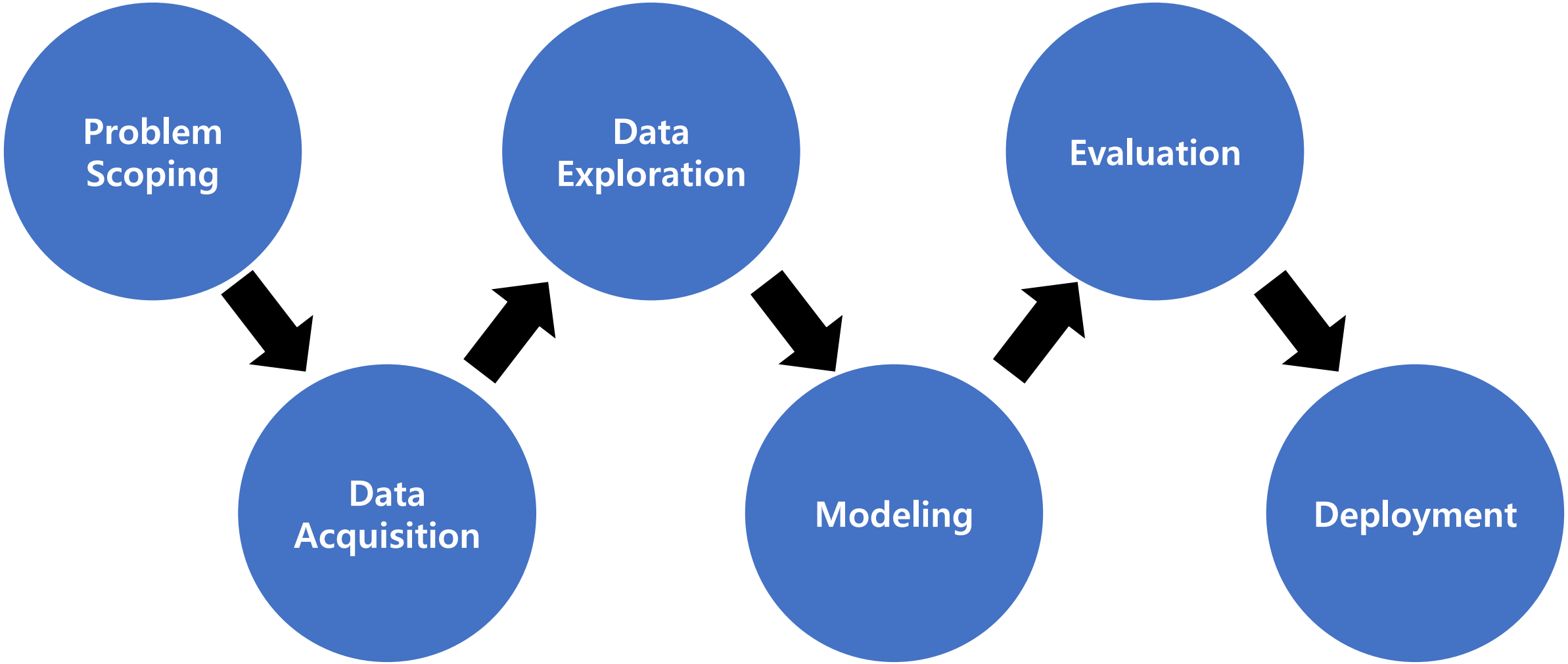
petition_corrupted.csv

- 전체 행 중에서 5%는 임의 필드 1개에 결측치 삽입
- 범주(category)가 '육아/교육'이고 투표수(votes)가 50건 초과이면 20% 확률로 투표수에 결측치 넣기
- 나머지는 전체 데이터와 동일

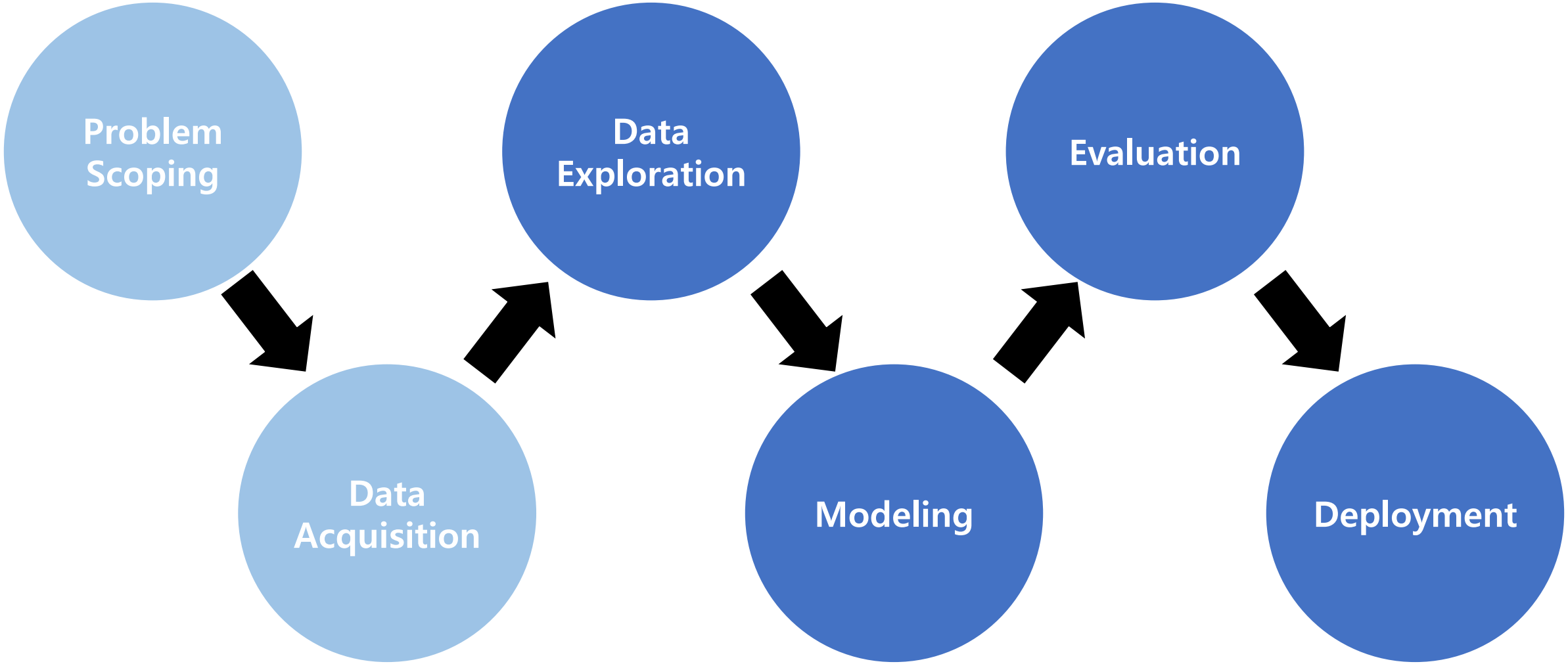
petition_sampled.csv

- 전체 데이터 중 5%만 임의추출한 데이터

AI 프로젝트 작업 흐름도

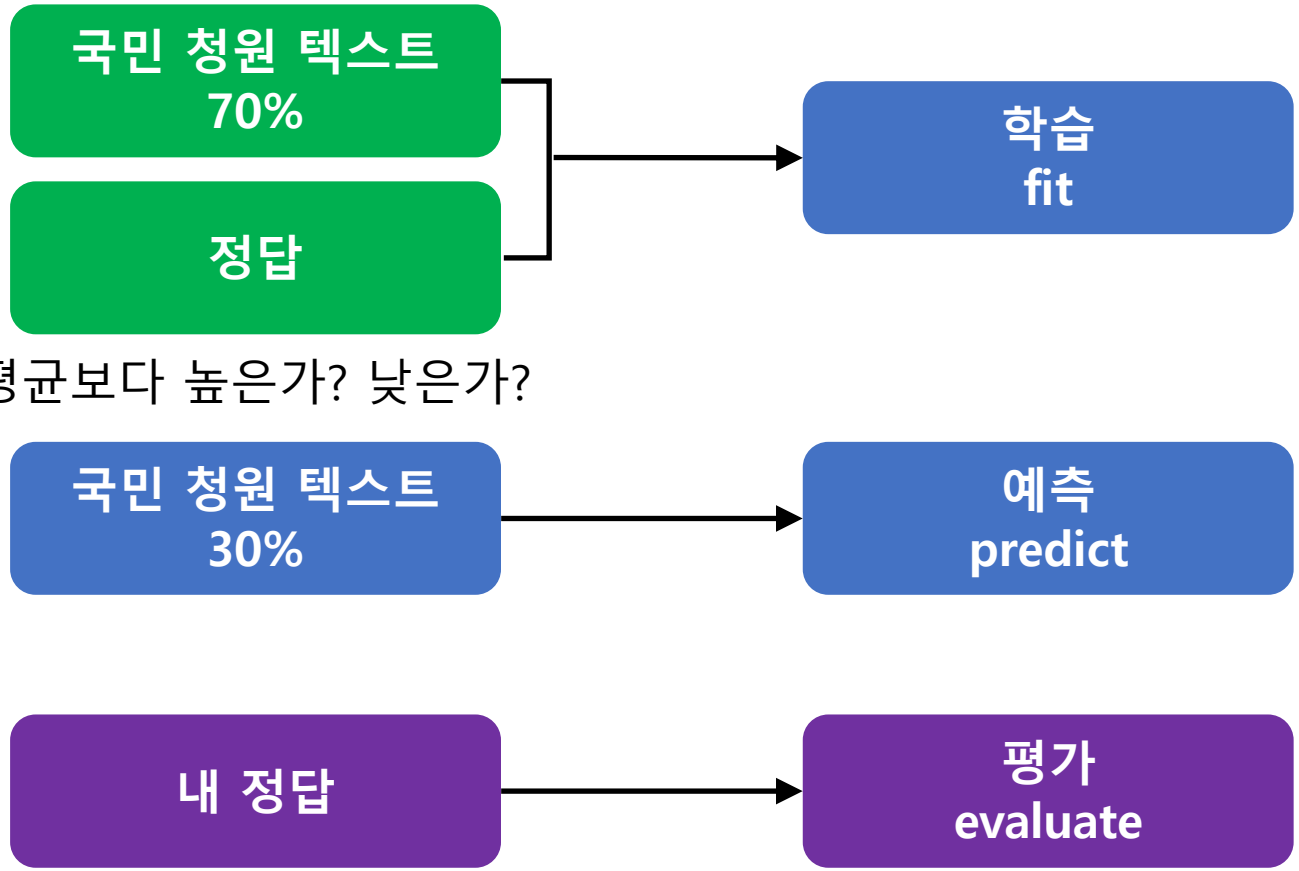


AI 프로젝트 작업 흐름도



국민청원 데이터 시각화 분류 : 머신러닝으로 데이터 이진 분류하기

국민청원 텍스트 지도학습 : 모델 - LightGBM



이번 6장도 정말...고생했다...