ML SESSION

#3
Cross Validation

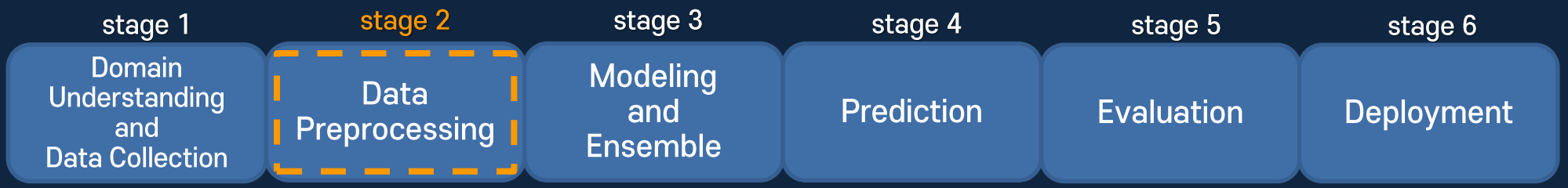
DATA&ANALYTICS

INDEX

0 우수과제팀

팀명	모델 성능
과적합의노예조_2번째 제출	1.399
교수님저희싫어하시조_1번째 제출	1.421
머린이탈출하조_1번째 제출	1.423
머신러닝정복해조_2번째 제출	1.484
A+만들어조_1번째 제출	1.486

1 Data Split



Data Preprocessing (데이터 전처리)

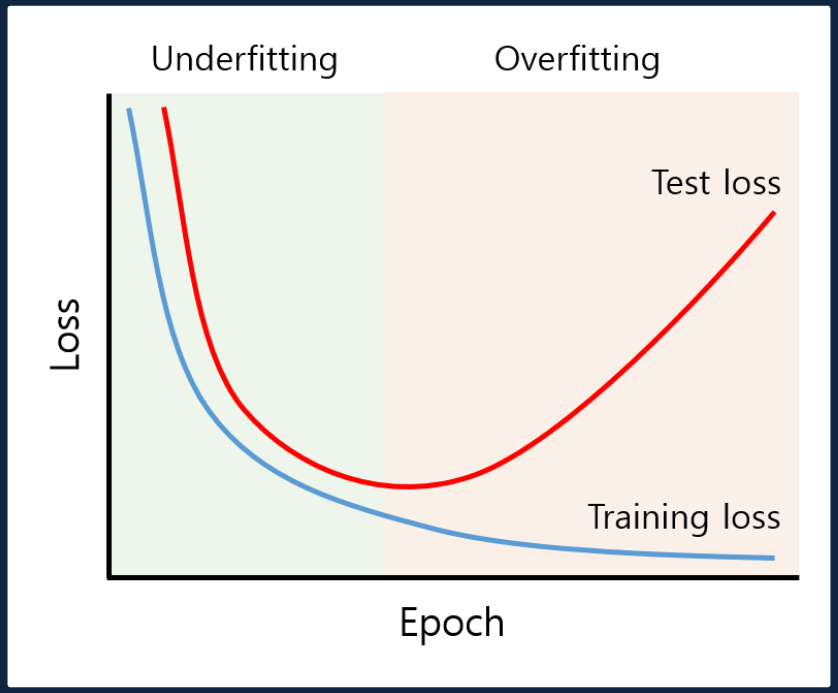
- Feature 만들기
- 결측값, 이상치 처리

- train, test 데이터 분리

- 기존에 배웠던 방법 사용시 과적합의 문제가 발생하기 때문에 train, test, validation 데이터까지 나눠줘야 함

1 Data Split

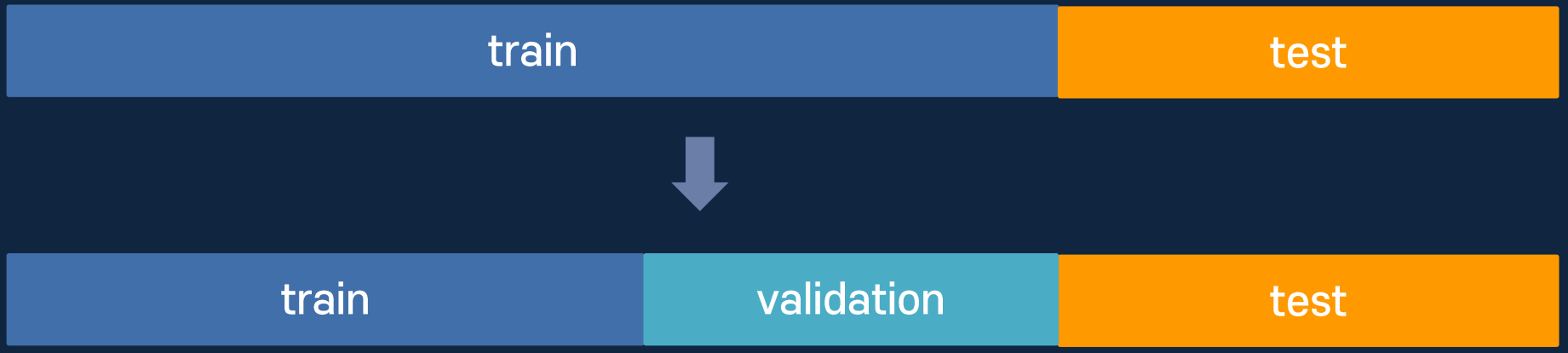
기본적인 데이터 분리 방법



우리의 목적은 학습을 통해
모델의 underfitting된 부분을 제거해 나가면서
overfitting이 발생하기 직전에 학습을 멈추는 것!

1 Data Split

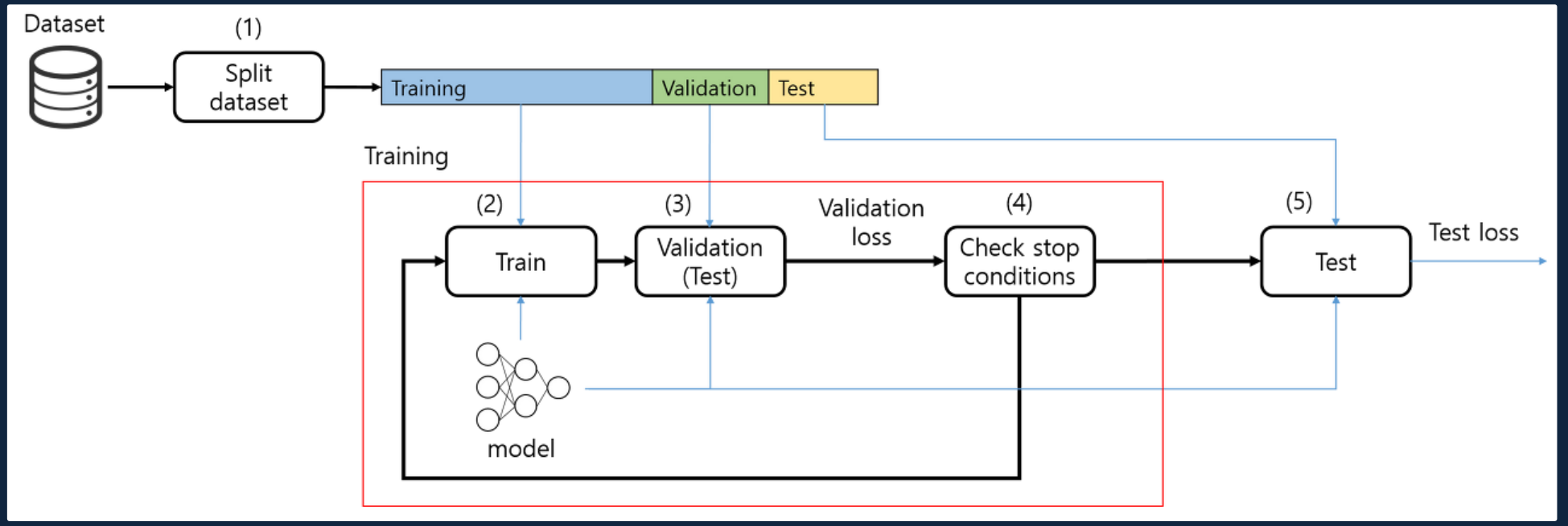
validation 데이터 분리 방법



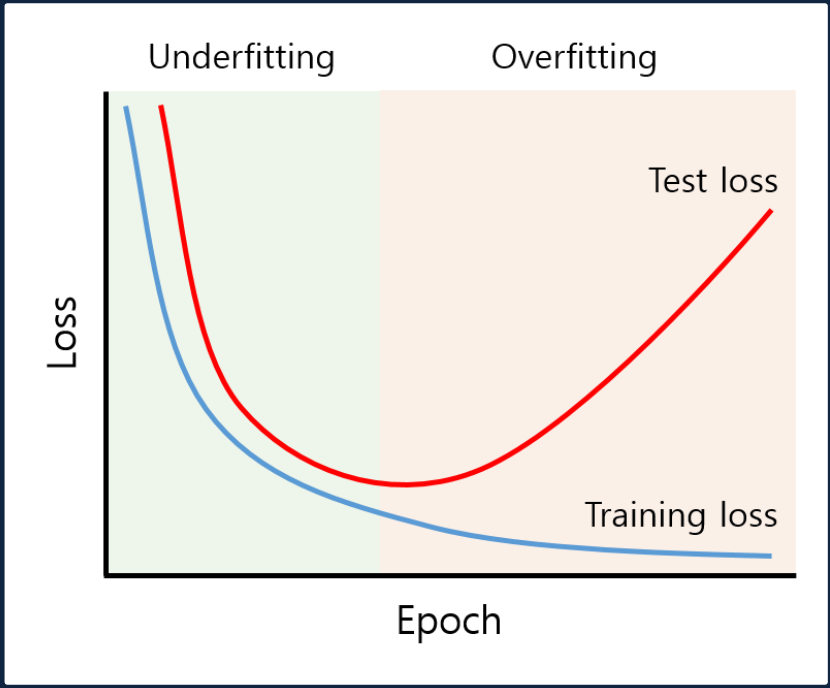
- Validation 데이터는 별도로 만들어진 dataset이 아니라 train에서 추출된 dataset임
- 모델의 파라미터 추정에는 train을 사용, 하이퍼파라미터 설정에는 validation을 사용, 테스트 셋에 모형을 적용시켜 정확도를 측정
- 파라미터 (모델 내부에서 결정되는 변수), 하이퍼파라미터 (모델 세부 조정 값)

1 Data Split

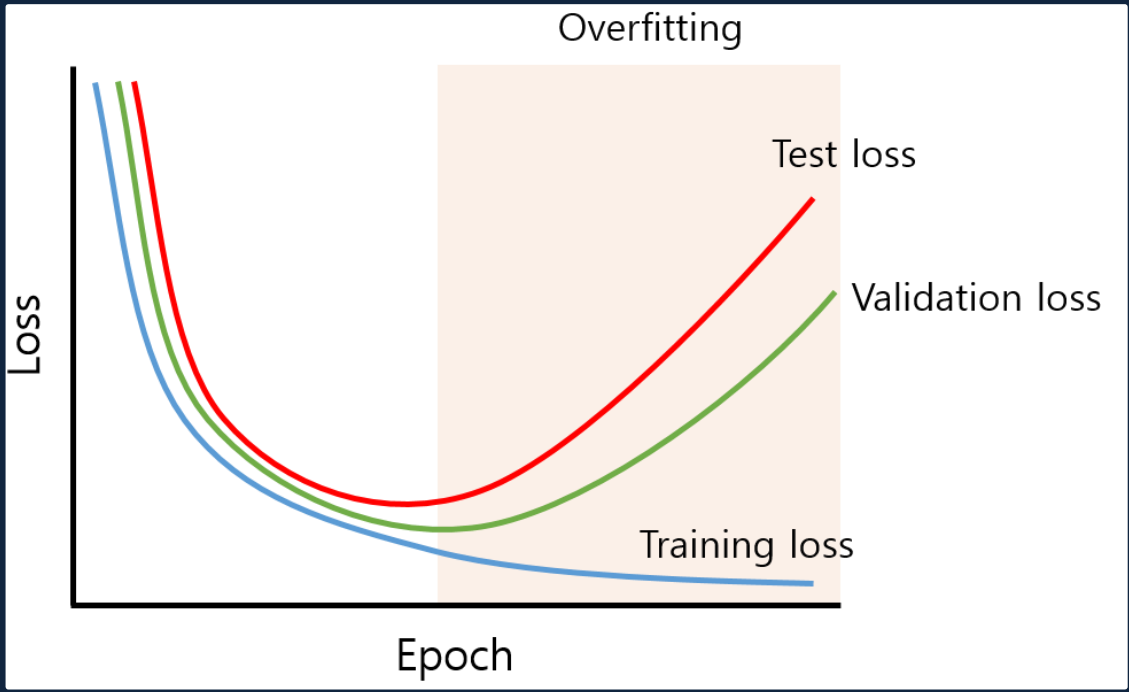
Validation dataset을 이용한 학습 과정



1 Data Split



Train, test로 split



Train, test, validation으로 split

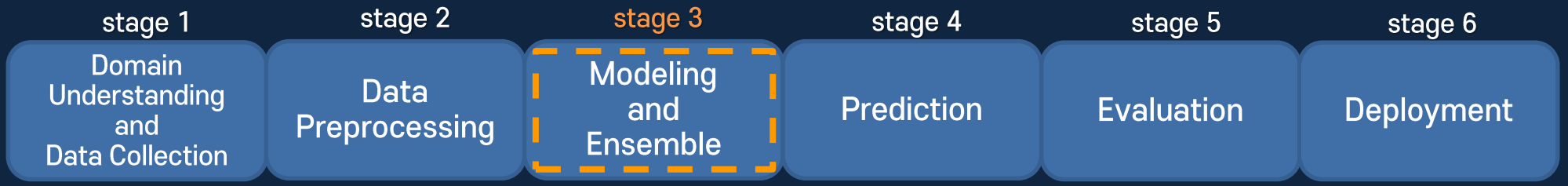
1 Data Split

분리한 데이터들의 차이점

	Training dataset	Validation dataset	Test dataset
학습 과정에서 참조할 수 있는가?	O	O	X
모델의 인자값 (가중치) 설정에 이용되는가?	O	X	X
모델의 성능 평가에 이용되는가?	X	O	O

- Test dataset은 모델의 성능을 평가하는 데이터이기 때문에 중간 과정에 영향X
- 모델의 인자값 (가중치)는 모델을 '학습'시키는 과정에서 업데이트 되는 것이고 validation dataset은 모델의 학습이 끝나고 하이퍼파라미터를 최적화하는 데 사용되는 것!
- Validation dataset은 최종 모델을 평가하는 것이 아닌 학습 과정에서 성능을 평가함

2 Cross Validation



Cross Validation (교차 검증)

- 모델에서 사용되는 하이퍼파라미터를 조정하고 과적합을 막기 위해 사용하는 검증 방식

- 모델의 학습 과정에서 train, validation데이터를 나눌 때 단순히 1번 나누는 게 아니라 K번 나누고 각각의 학습 모델의 성능을 비교하여 평균 값을 모델의 성능으로 판단

2 Cross Validation

교차검증 사용 이유

- 데이터셋이 부족할 때 적용하는 방법

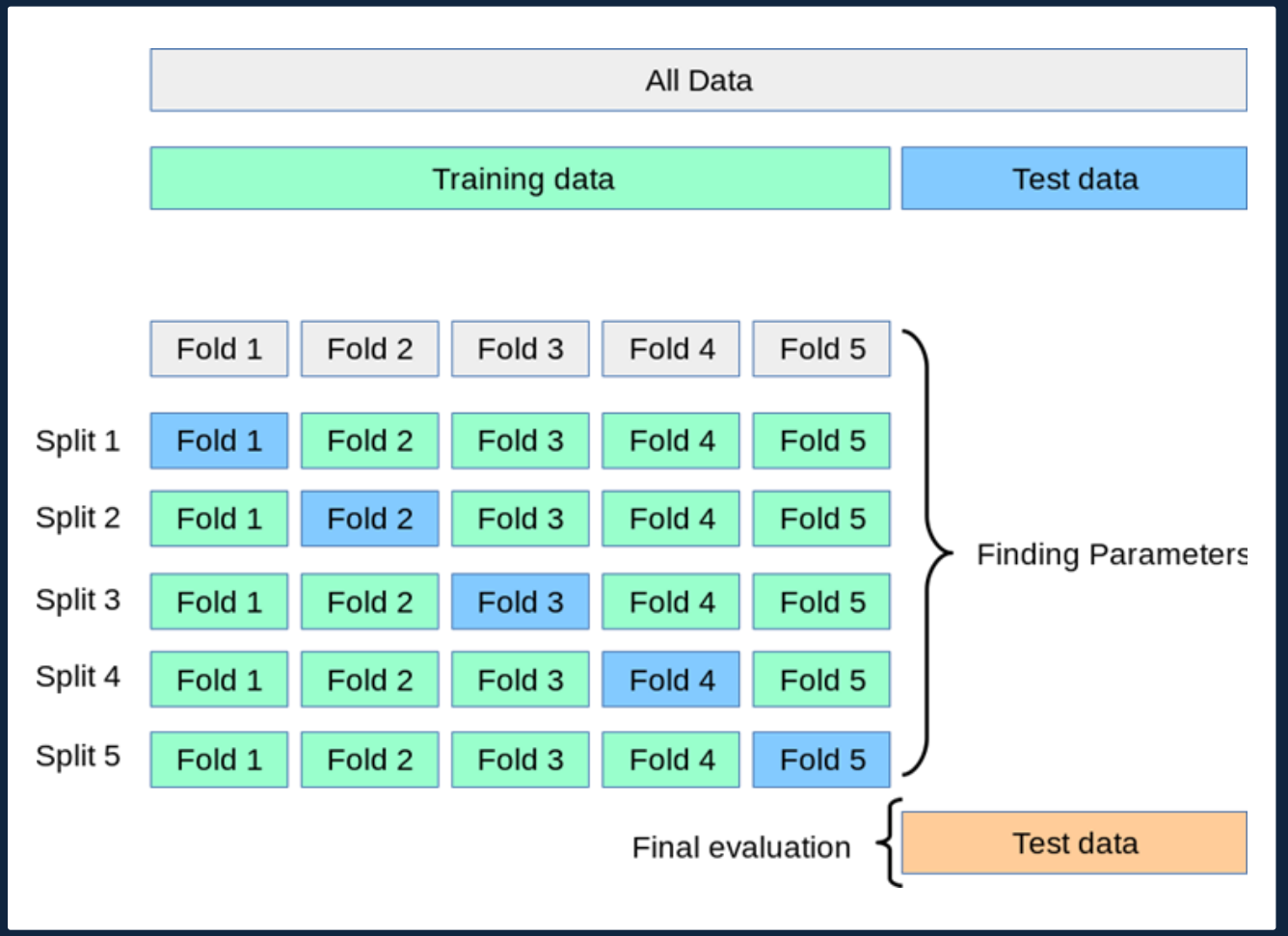
전체 데이터가 학습/검증으로 한번에 나누기 작은 경우 여러번 데이터를 나누고
각 교차검증마다의 모델 성능을 비교하는 방식으로 학습을 진행하면
데이터가 부족한 문제를 보완할 수 있음

- 보다 일반화된 모델 성능 평가 가능

여러차례 나누는 교차검증 방식을 통해 전체 데이터 전 범위를 학습하고
검증 데이터로 성능을 평가 함으로서 보다 일반화된 모델을 생성할 수 있음

2 Cross Validation

k-fold cross validation (k겹 교차검증)



1. Train 데이터를 K 등분 (이미지의 경우 K=5)

2. 1/5을 validation으로, 4/5 를 train으로

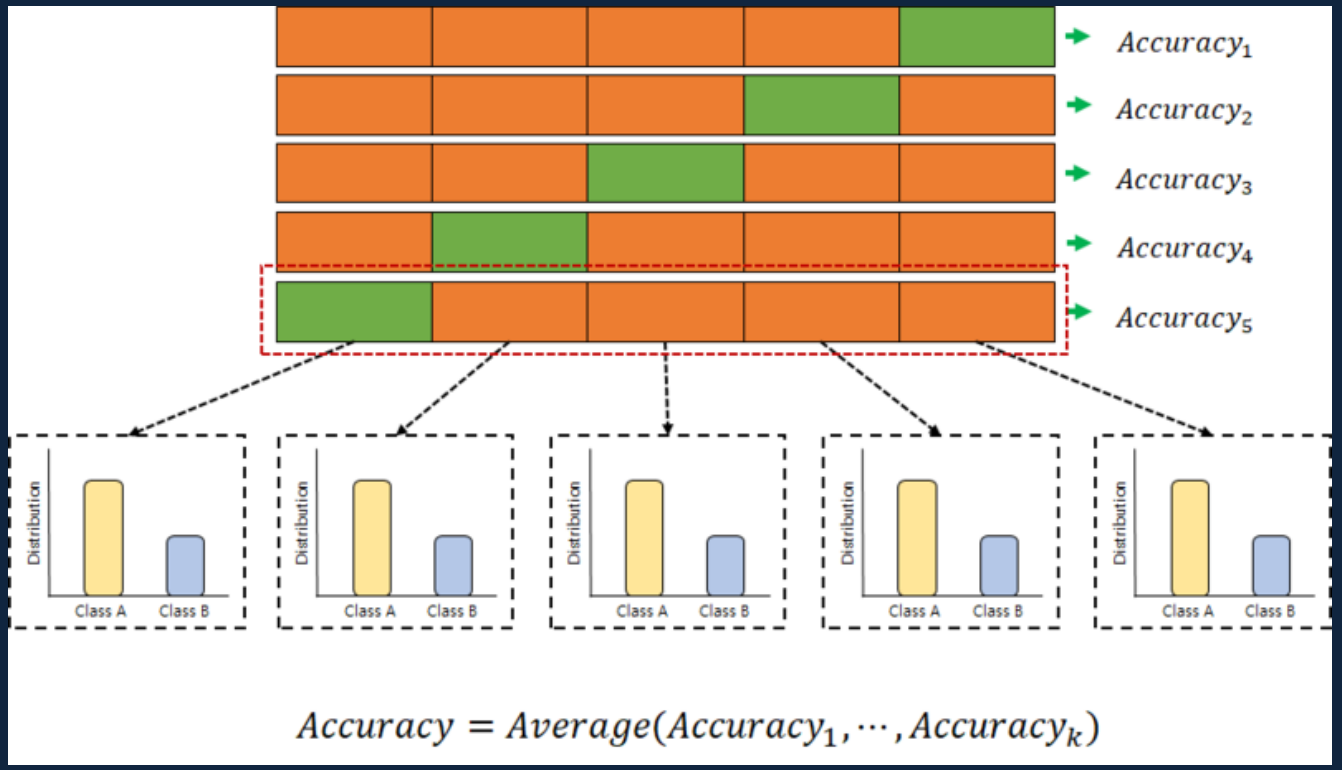
3. Validation을 바꿔가며 성능 평가

→ 총 5개의 성능 결과가 나올 것이고 5개의 평균이 해당 학습 모델의 성능임

2 Cross Validation

Stratified k-fold cross validation (계층별 k겹 교차검증)

- 데이터 클래스 별 분포가 불균형한 상황에서 사용하는 방법
- 데이터 클래스 별 분포를 고려해서 데이터 폴드 세트를 만드는 방법이 계층별 k-겹 교차 검증



THANK YOU