

12/04 · Logistic Regression 구현 (Tensorflow
Sklearn
이진분류)

→ 복습도 출점 admission data2 구현 !!

~~설명~~

→ 만들 model이 좋을지 어떻게 판단하나요??
(아님)



Evaluation (평가)를 진행!!

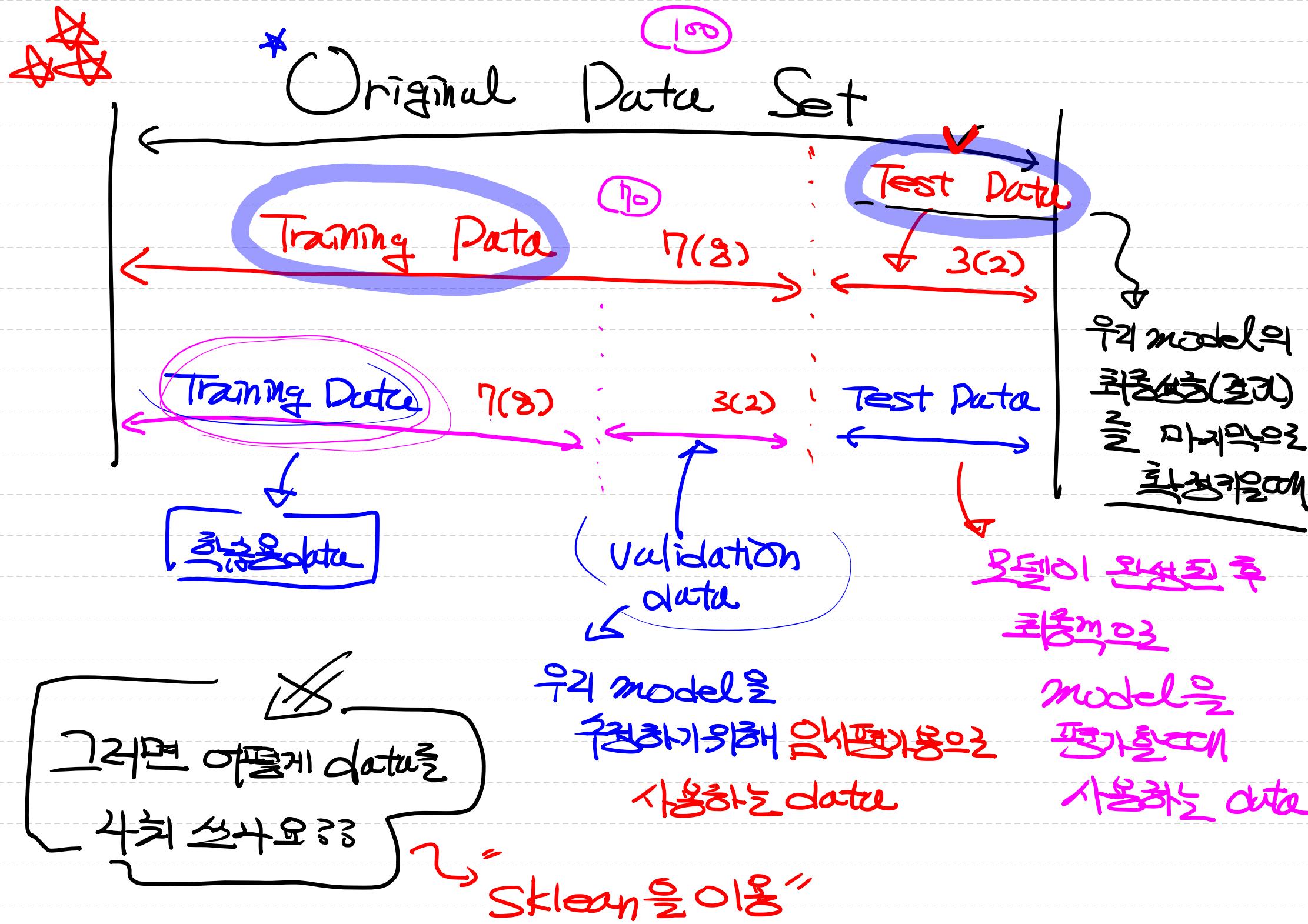
데이터를 분할하기

(
학습하고
평가하기)

~~설명~~

* 학습에 사용된
data를

평가에 사용하지
않아요



○ 우리는 지금 logistic Regression을 이용해
binary classification 작업을 수행!!

→ model 구현 후 평가를 진행

~~※※※~~

↳ 평가기준 (Metrics)

✓ ~~※※※~~

Confusion Matrix

(0, 1)

✓

		* 실제 정답	
		True	False
model 예측값	True	(TP) True Positive	(FP) False Positive
	False	(FN) False Negative	(TN) True Negative

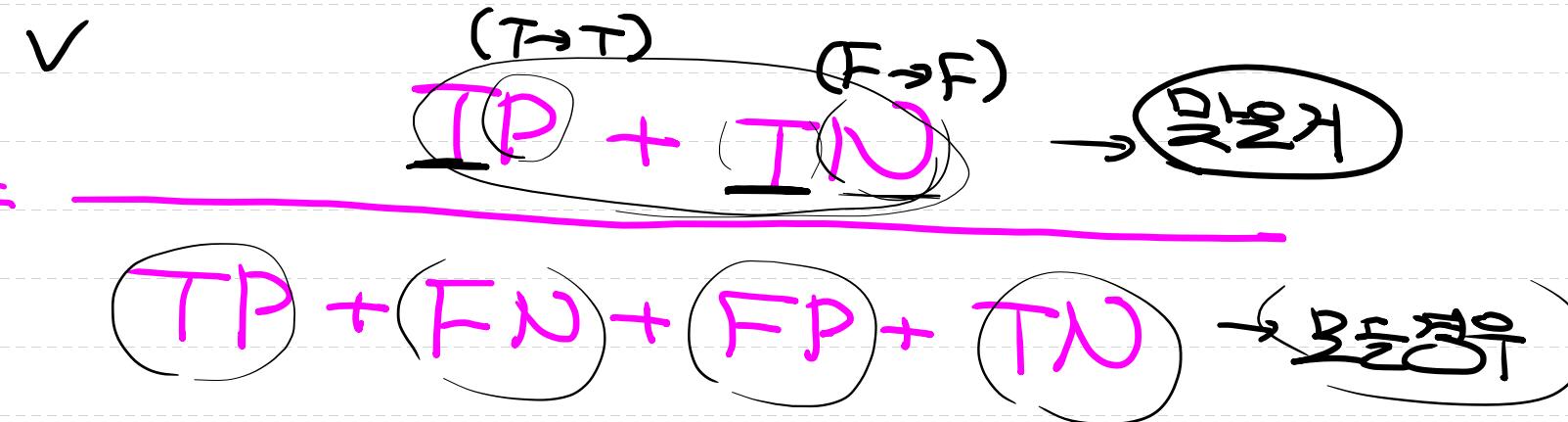
○ 우리 model의 평가기준을 여쭤보자면 뭐니요

Classification의 평가기준 중 가장 대표적인 건.

(분류모델)

① Accuracy (정확도) ~ accuracy이용
이런어떻게 구하나요??

Accuracy =



→ “가장 적고난이한 성능평가 지표”

○ 희귀罕见 Data

...		0
...		0
...	X	0을 헷갈

* model

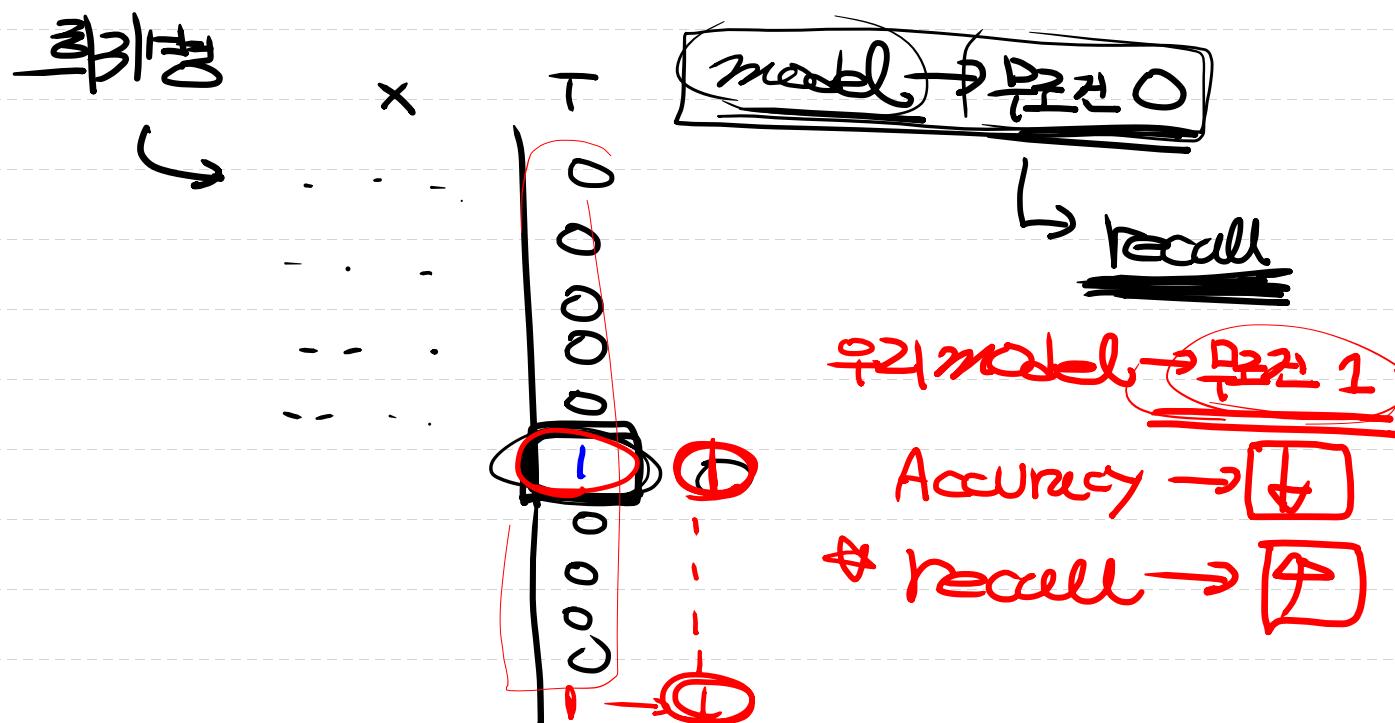
0을 헷갈

↳ 뜨겁다 뜯어요

↳ Data의 풍향이 심한경우!!

② Recall (재현율, hit rate)

$$\text{Recall} \rightarrow \frac{\text{TP}[\text{정답 T}, \text{예측 T}]}{\text{TP} + \text{FN} [\text{정답 T}, \text{예측 F}]} \Rightarrow \text{AI 실제 True 의 것 중에 } \\ \text{우리의 model이 } \text{True 라고 } \text{전하고 } \underline{\text{예측한 비율}}$$



③ Precision (정밀도)

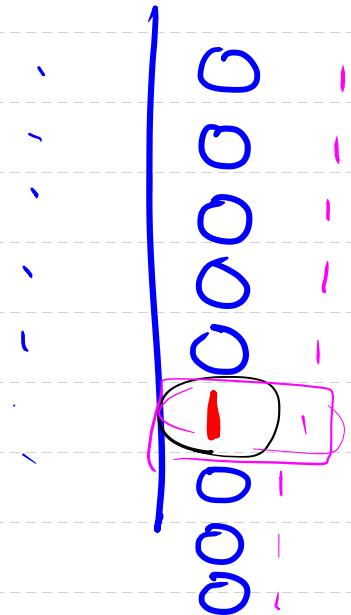
*

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

→ 우리의 Model이
True로 분류한
것 중 정밀로
True인거

[정답 T]
 [예측 T] [정답 F]
 [예측 F]

학습
data



model → 0

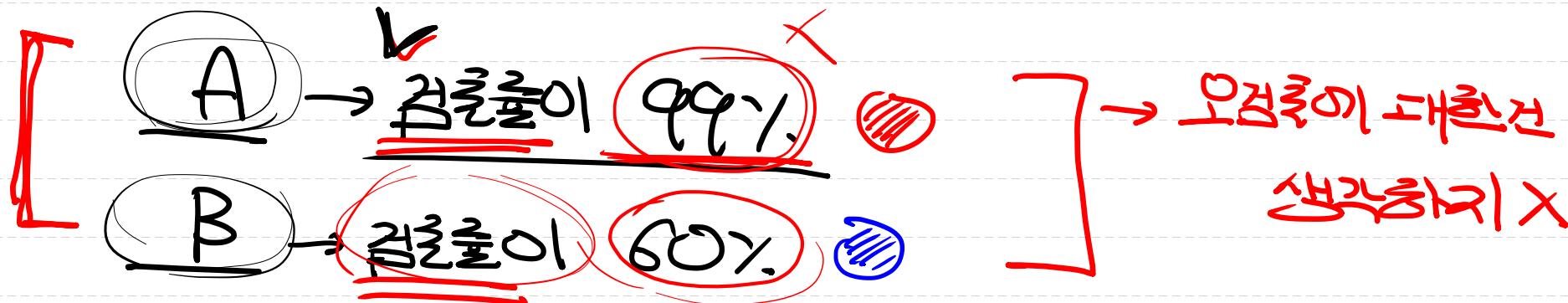
→ Accuracy ↑
recall ↑

model → 1 → Accuracy ↑
recall ↑

Precision ↓

설명해주세요

ex) 사진이 있는 고양이 검출 model을 개발



↑ Recall ↔ Precision [‘好坏识别率’]

④ F1 Score ↑

조화평균

$$F1\text{ score} = \frac{2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}}{\text{Precision} \times \text{Recall}}$$

⑤ Fall-out

ROC curve, AUC,

log-loss

‘검출률’ 같은 조건
많이 사용하는 평가지표

기본적으로 알아야 하는 내용은 많이 설명되었어요!!

추가적으로 알아야 하는 시즌!!

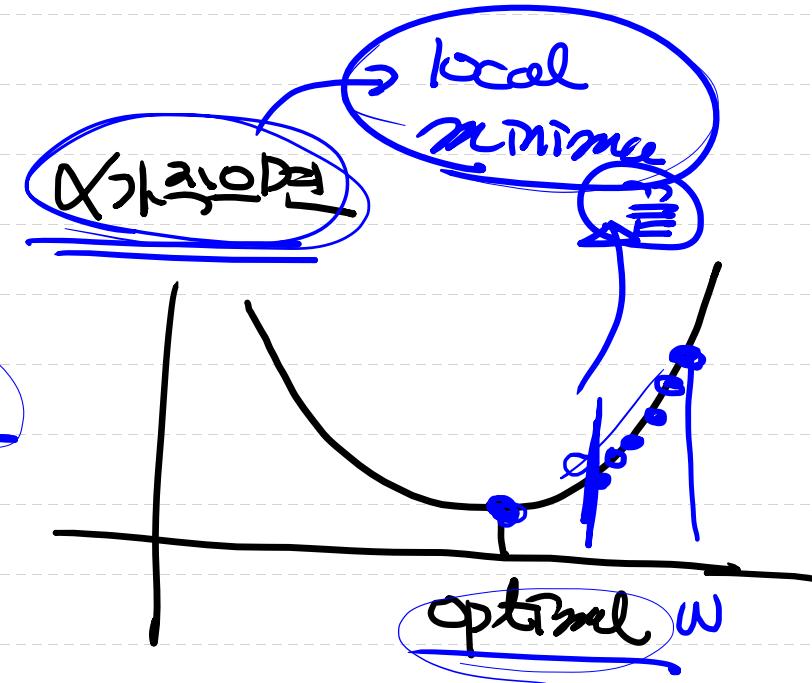
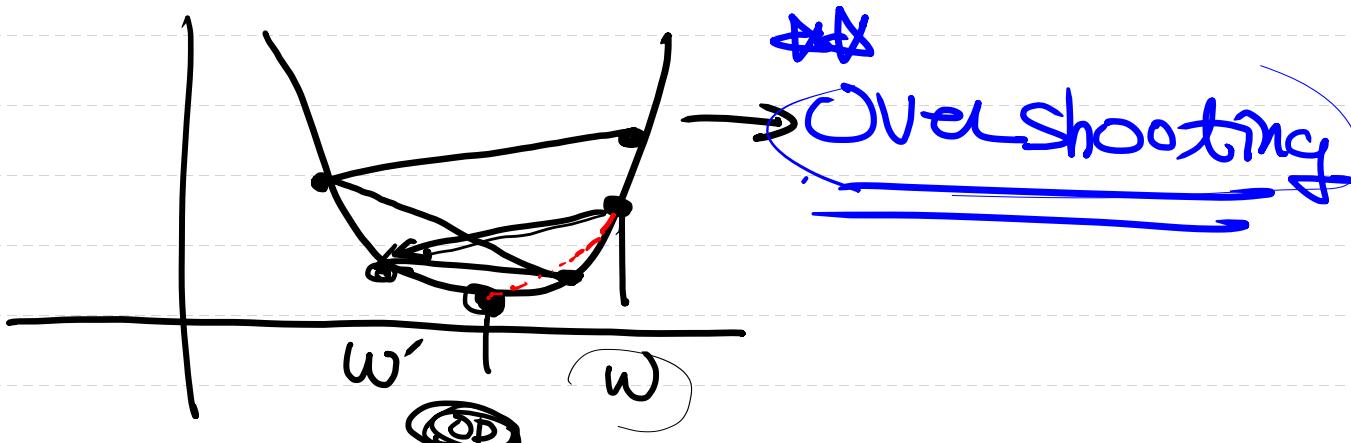
① Learning rate →
(hyperparameter)

→ 가작을 놓아요 ~ $1e^{-4}$

$$(w) = w - \alpha \frac{\partial \text{loss}}{\partial w}$$

learning rate

α 가 크면 최적의 w 를
놓아요



② Normalization (정규화)

머신러닝 할 때, 특성들이 0~1 사이의 값으로 정규화

- | ① 각 feature의 scale을 동일하게 만들어요.
- | ② Overfitting을 피하고 Noise를 감소시켜요.

① Min-Max Normalization (최대, 최소)

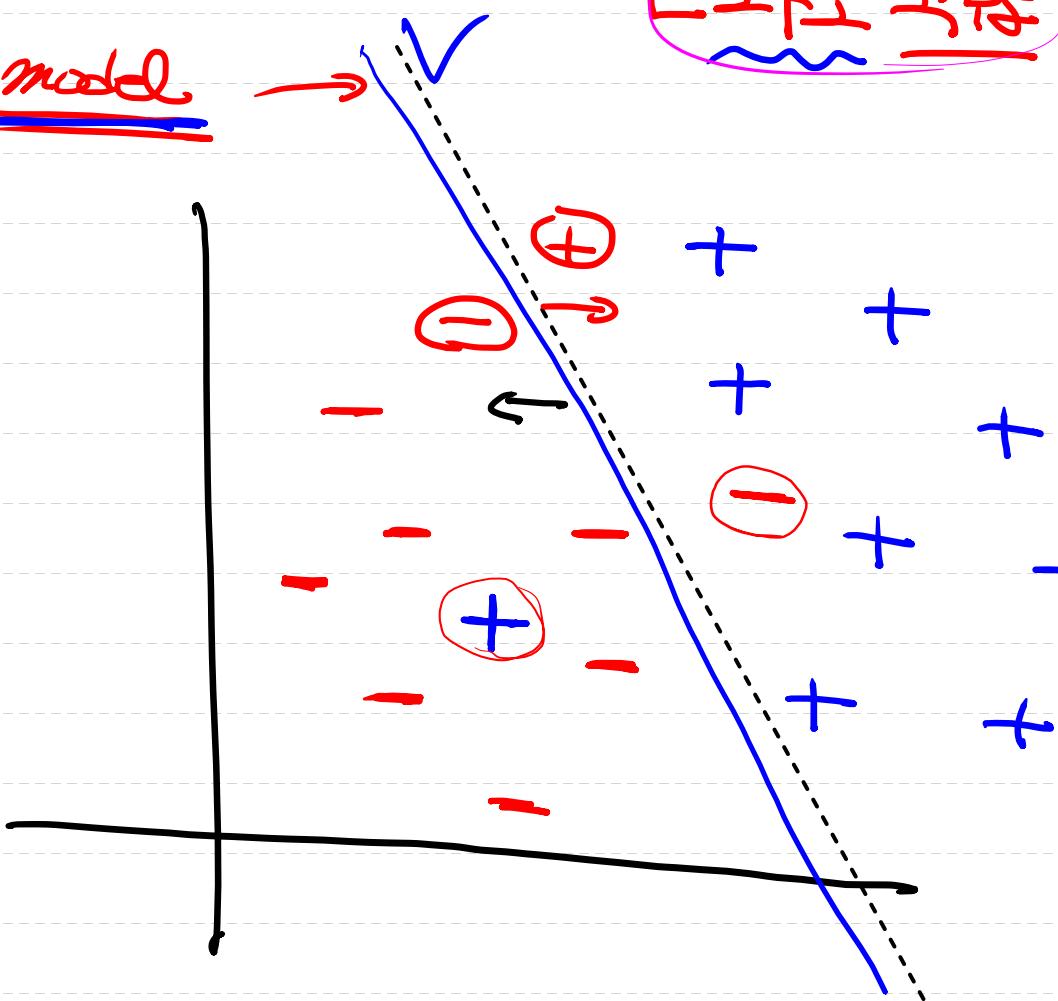
② Standardization (평균, 표준편차)

데이터가 이걸로 뺍해요 ~ 예제

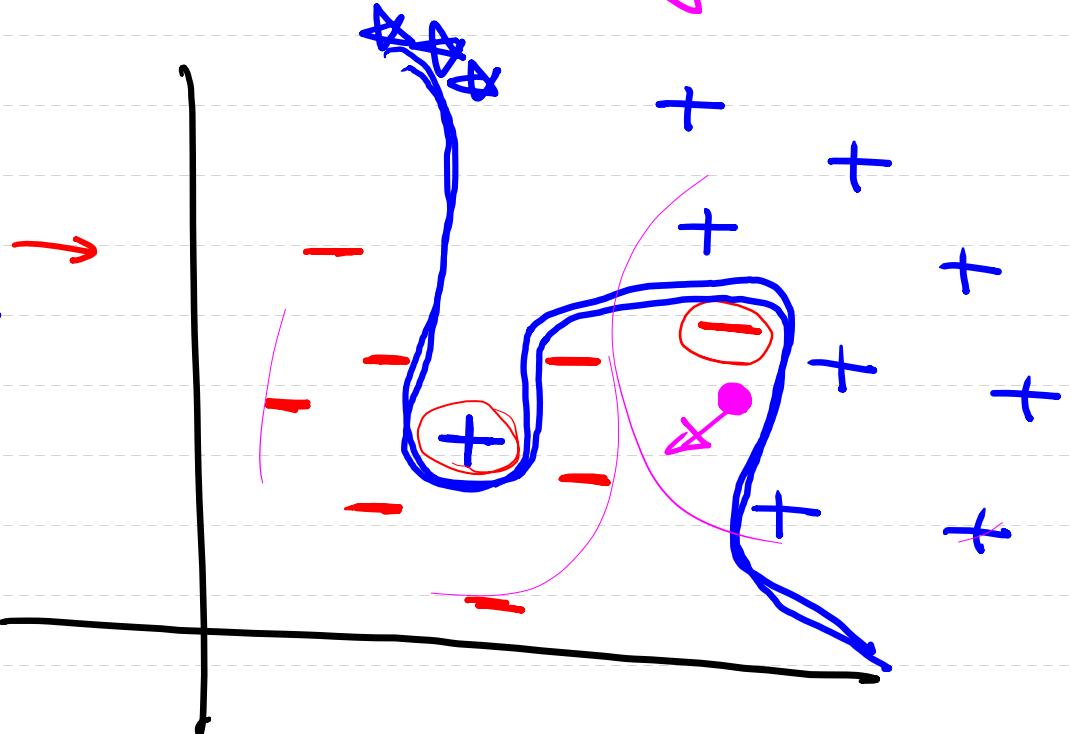
③ Overfitting

※ 과적합 (Overfitting) → "Overfitting"
※ 과대적합 (데이터 모델이 너무 잘 만들었어요)
※ 과소적합 (모델 학습이 안되었어요) Underfitting

model



Underfitting



Overfitting

Training Data Set이 아주 쪽壑을 model에 모은

Test Data에 잘 안되는 경우

→ 거의 흔들 브론해요 (이 정도를 줄이기 고민)

How??

①

많은 Training Data를 이용

Overfitting의 주된 이유는 데이터가 허락하는 때문

②

Feature의 개수를 줄여요

→ 다중공선성
VIF

③ Deep Learning → Dropout 기법 이용

④ Weight 값이 너무 커지면 오宦도로 정의학으로 조절

$$w' = w - \alpha \frac{\partial \text{loss}}{\partial w}$$



→ 제제 (Regularization)

↳ [L1 제제]

L2 제제

→ 구식이 풀라요

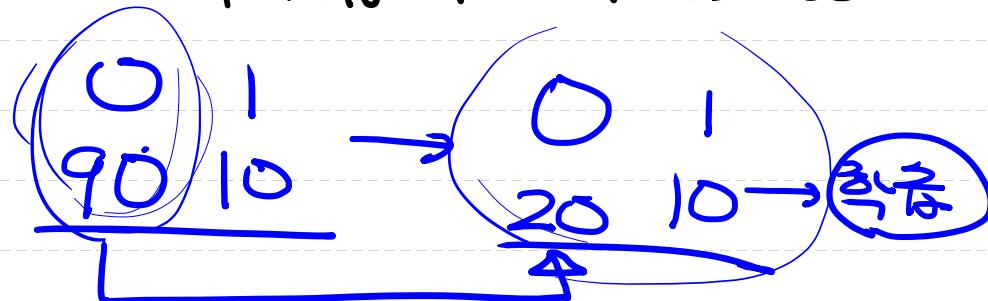
④ Imbalanced Data problem

“데이터 불균형 문제”

→ 희귀병 Data

부도가 많아 예측을 위한 Data

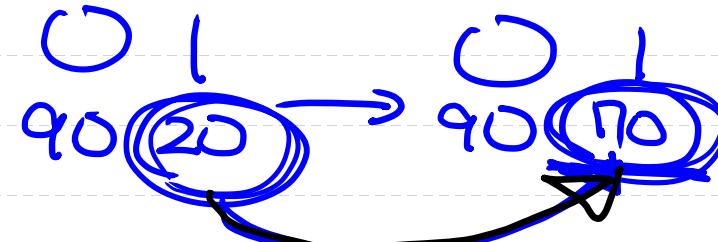
(1) Under Sampling



(2) Over Sampling

데이터를 늘려요

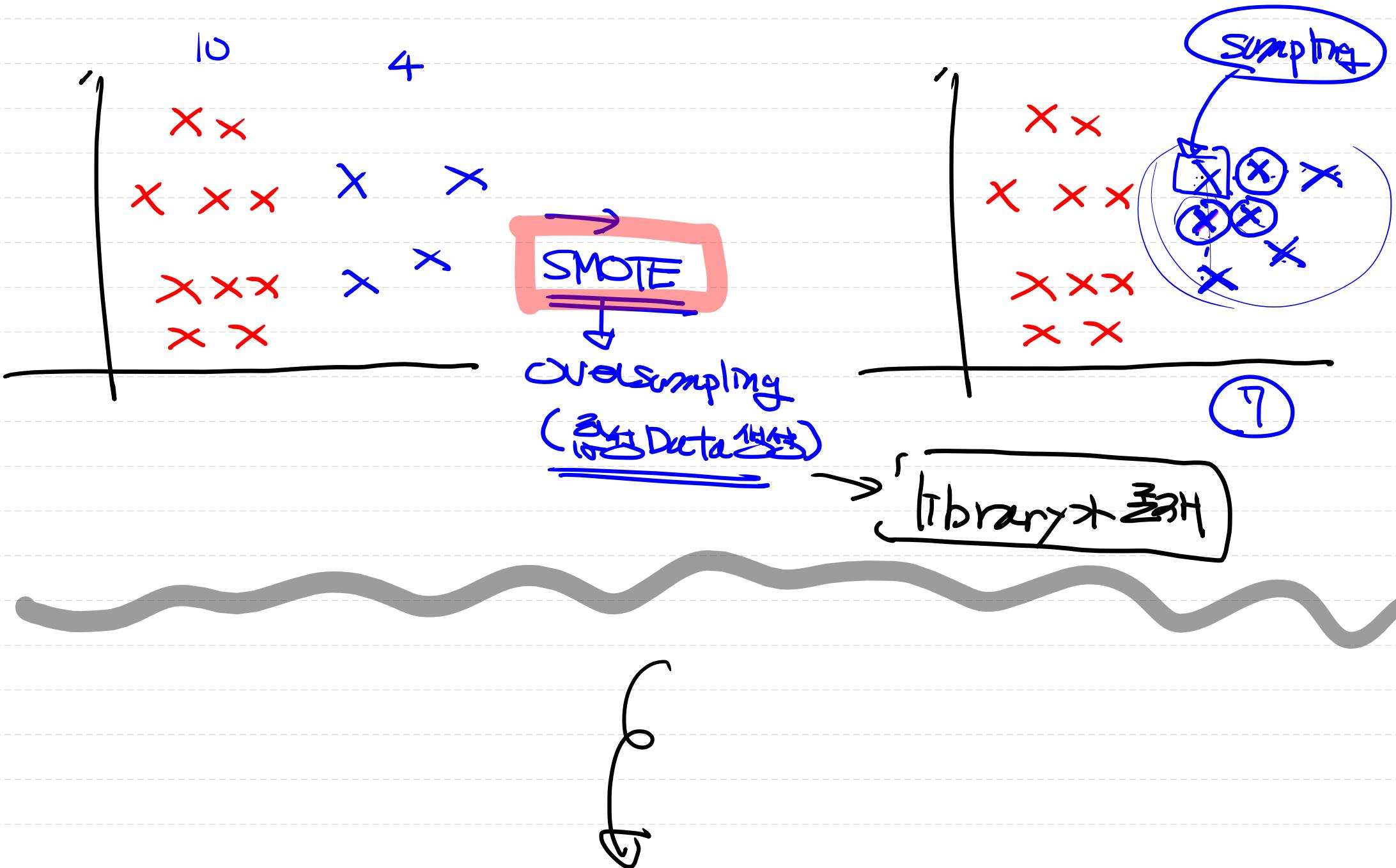
↳ 복제



↳ 같은데이터 ↑ → overfitting

(2) 증명 Data를 생성

→ SMOTE 증명 2/13

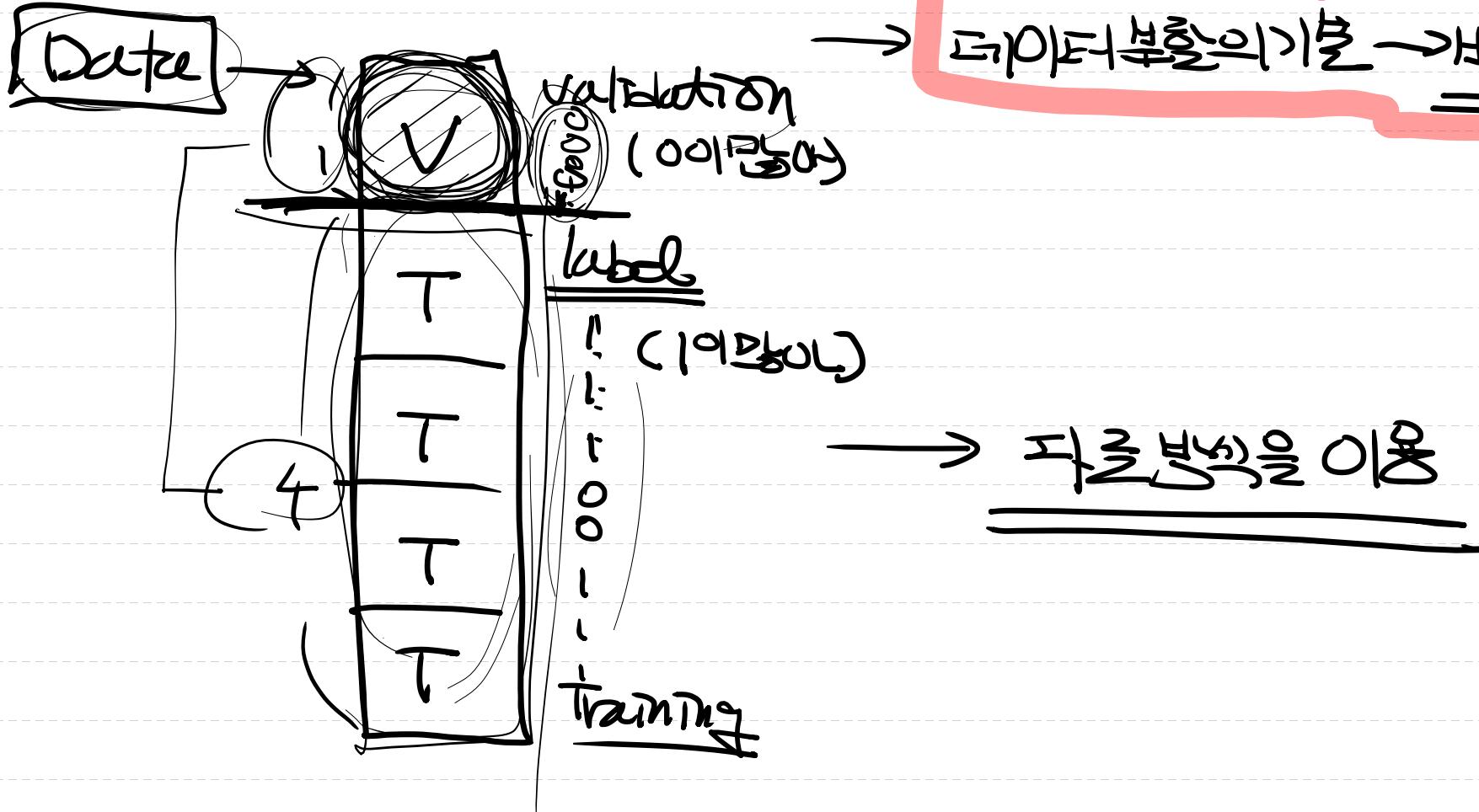


⑤

Validation을 어떻게 하시죠??

Evaluation → Test data 이용

Validation data 이용



평가용

데이터 분할의 기준 → 과대적합이 발생 X

* ~~별도로 validation set을 설정하는 것은 좋지 않다.~~

k-Fold Cross Validation

→ Validation data set이 가지고 있는 bias 문제

↙ 작은 data를 이용하여 Validation이 가능해짐

(k) 폴드 개수 지정

(5)

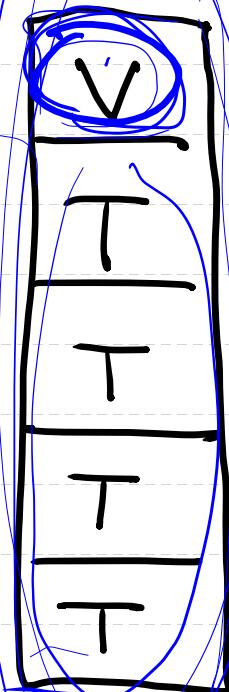
하지만 시각화 오래걸리는 단점도 존재!

작은

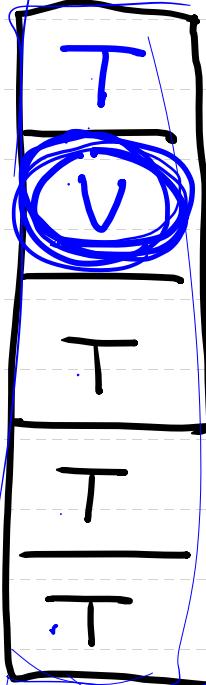
data



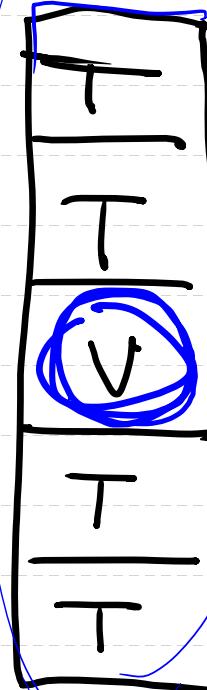
Fold 1



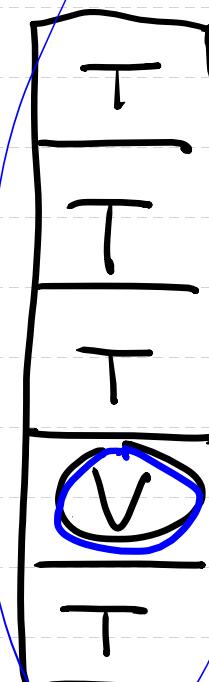
Fold 2



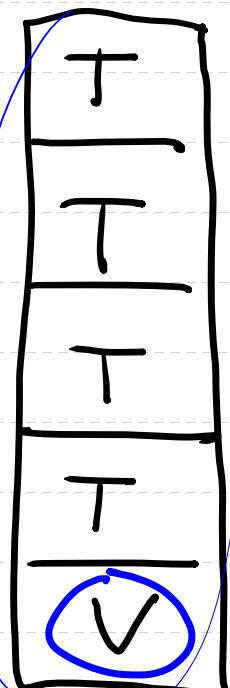
Fold 3



Fold 4

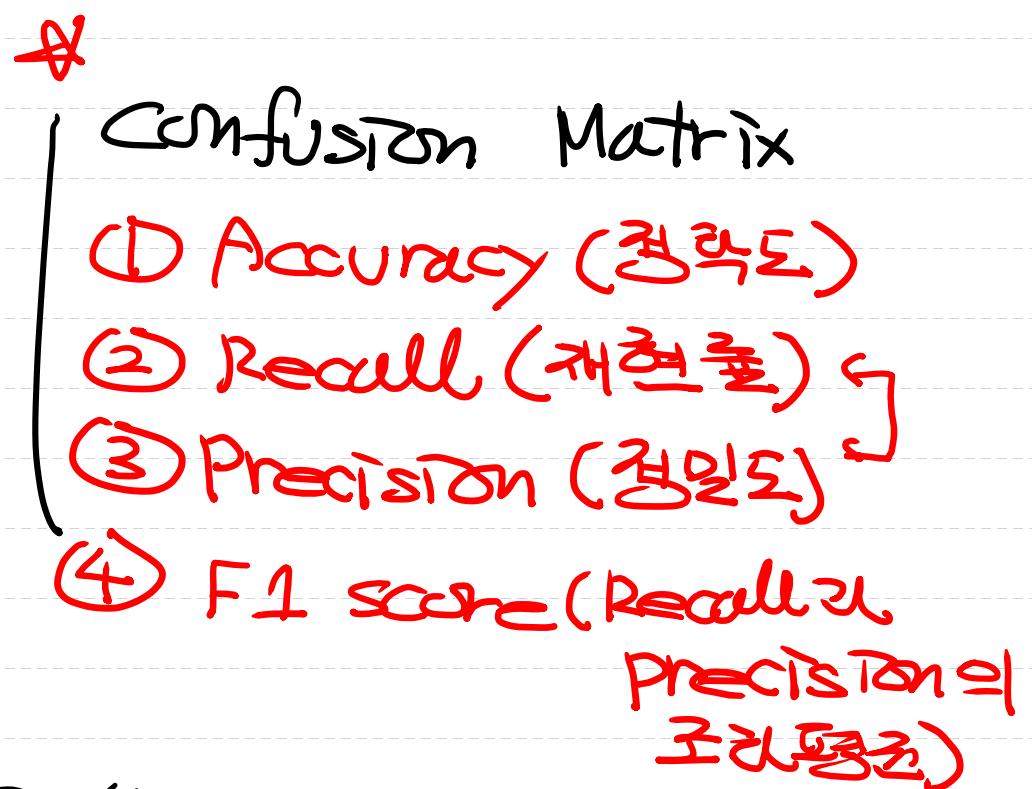


Fold 5



결과
결과
결과
결과
결과

- Metrics (평가기준)



- 추가적으로 알아야 할 개념들 !!

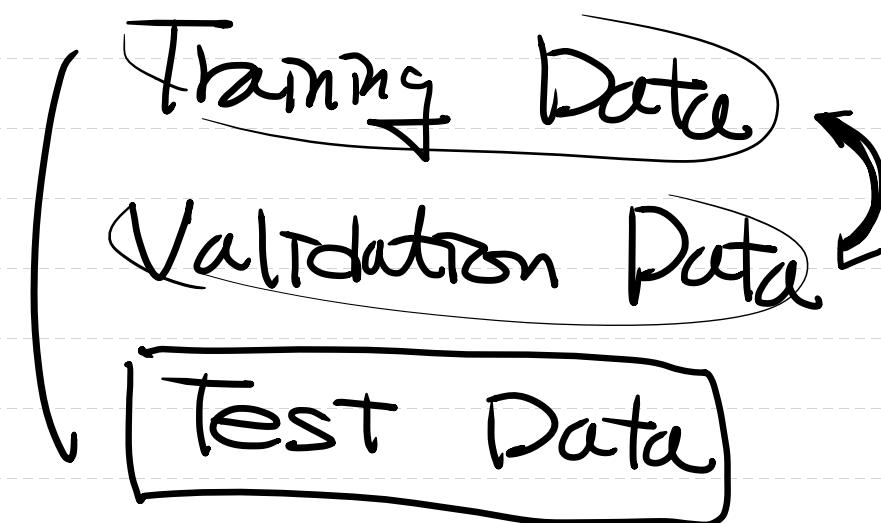
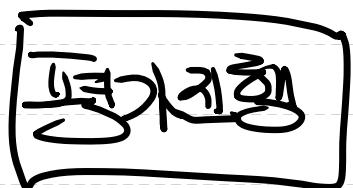
- ① Learning rate [표로 이동으로 보면 Overshooting , 즉으면 Local Minima]
- ② Normalization [Min-Max scaling
Standardization]
- ③ Overfitting → 피하려면 어떻게 해야요 ?

④ Imbalanced data problem

Undersampling

Oversampling → SMOTE

⑤ * k - Fold cross Validation



① Logistic Regression Model을 구현

→ 풋가기 전략!

- 데이터셋 → Wisconsin Breast Cancer Data