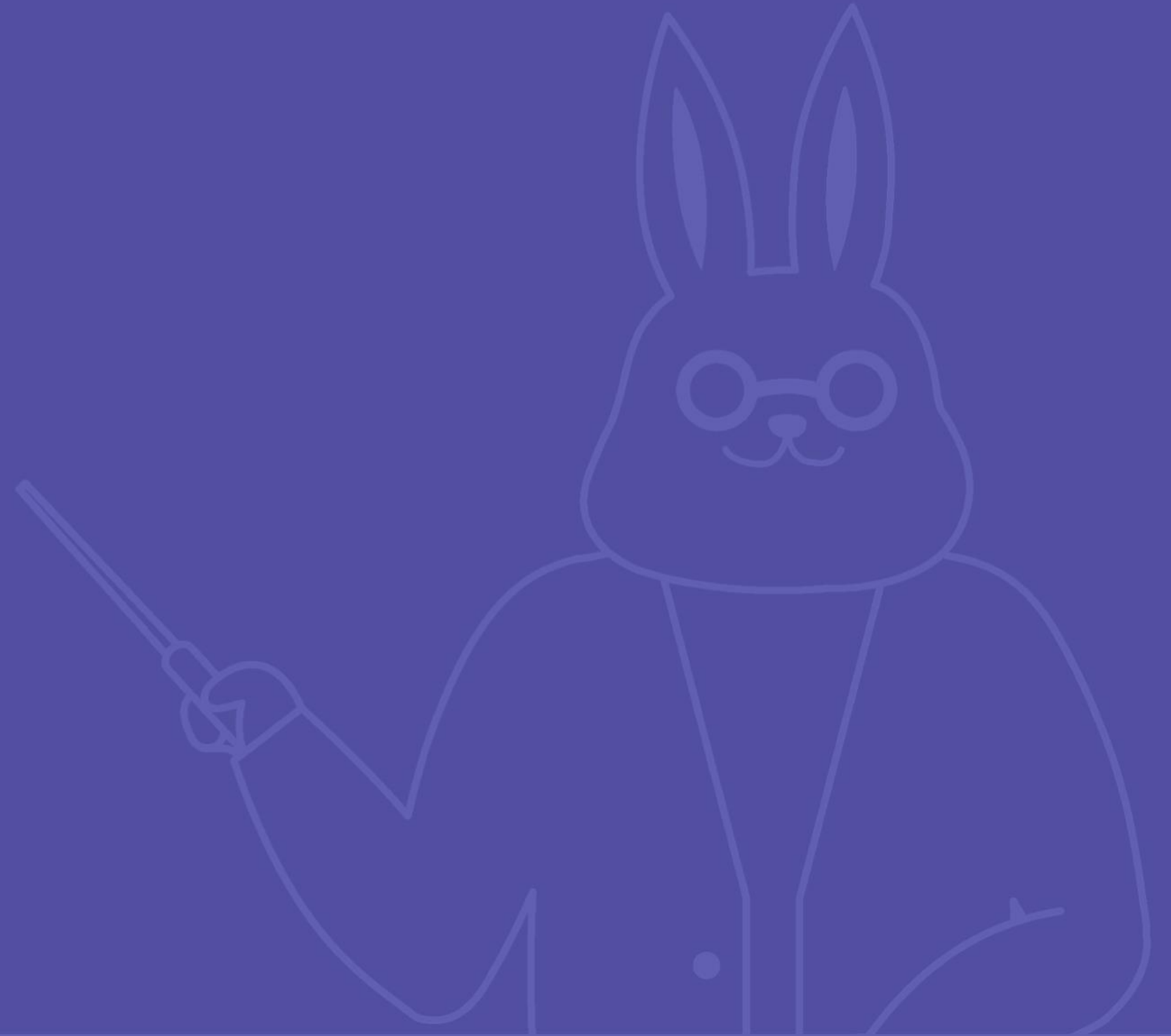




# 머신러닝 시작하기

04 지도학습 - 분류



# 목차

- 01. 분류 개념 알아보기
- 02. 의사결정나무 - 모델 구조
- 03. 의사결정나무 - 불순도
- 04. 분류 평가 지표

01

# 분류 개념 알아보기



## ✓ 가정해보기

해외 여행을 준비하고 있다고 가정하기

완벽한 여행을 위해 항공 지연을 피하고자 함

기상 정보(구름 양, 풍속)를 활용하여  
해당 항공의 지연 여부를 예측할 수 있다면?



✔ 문제 정의와 해결 방안

문제 정의

$X$

$Y$

- 데이터: 과거 기상 정보(풍속)과 그에 따른 항공 지연 여부
- 목표: 현재 풍속에 따른 항공 지연 여부 예측하기

해결 방안

분류 알고리즘

$X$

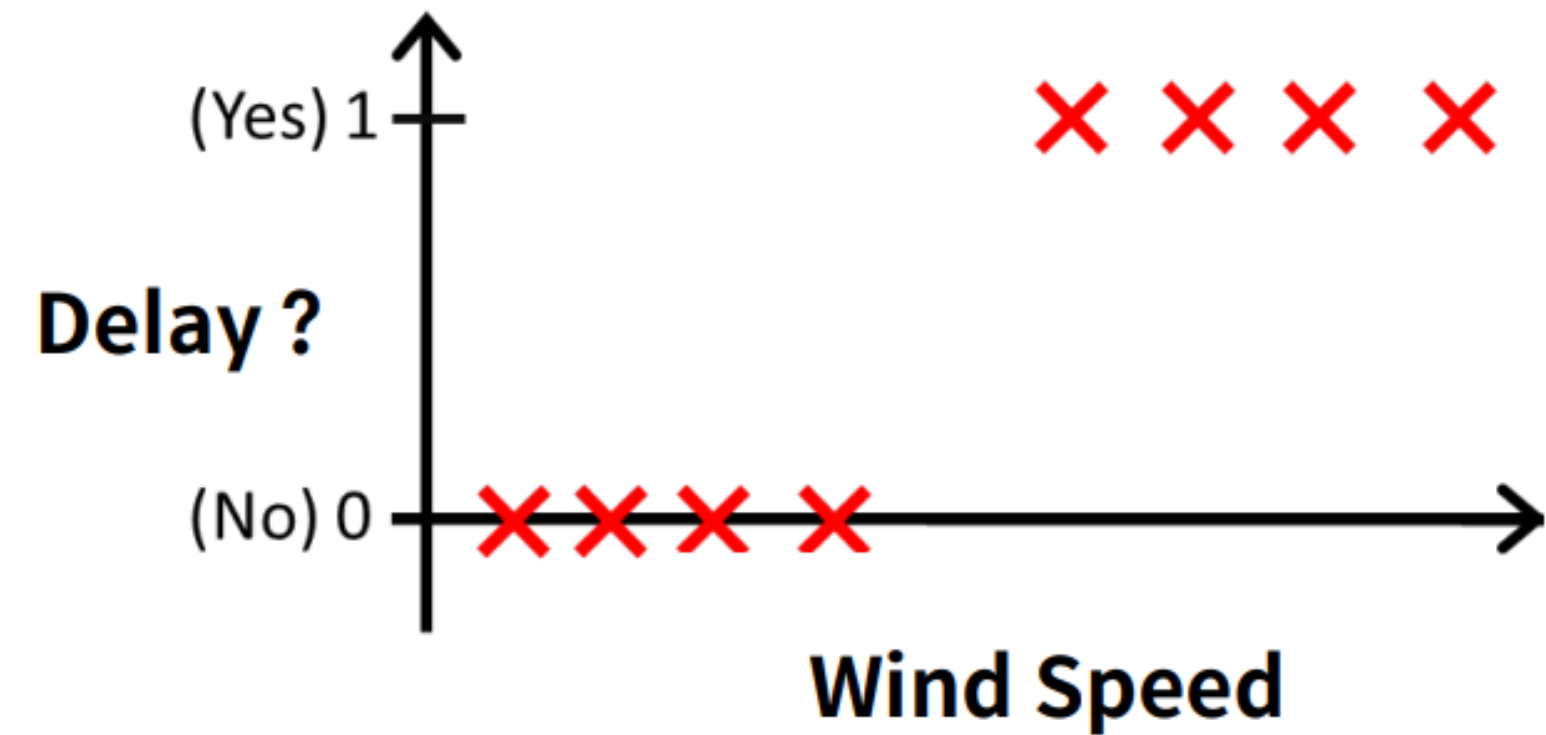
$Y$

풍속(m/s)	지연 여부
2	No
4	Yes
3	No
1	No

## ✓ 분류란?

주어진 입력 값이 **어떤 클래스에 속할지**에 대한 결과 값을 도출하는 알고리즘

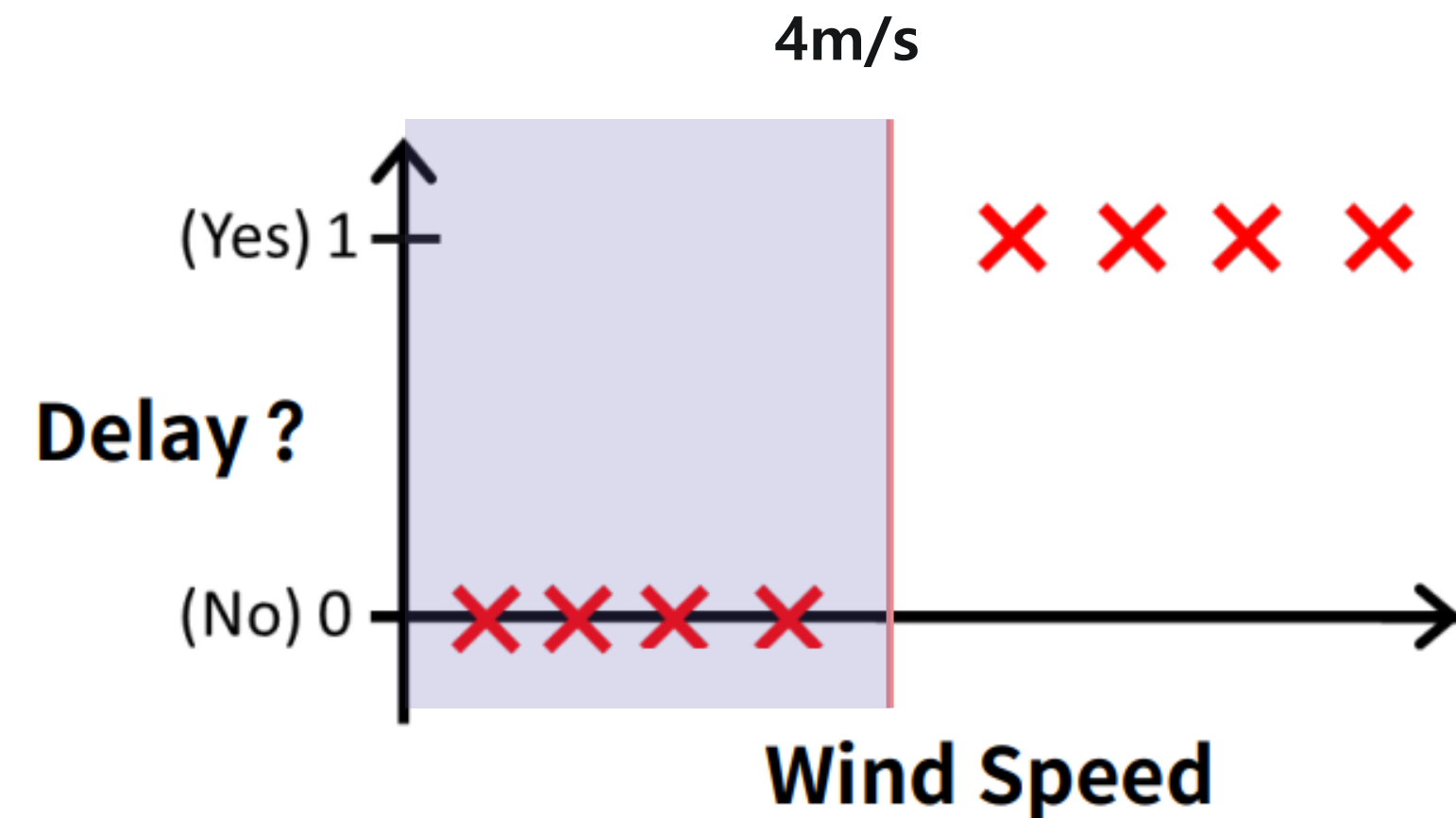
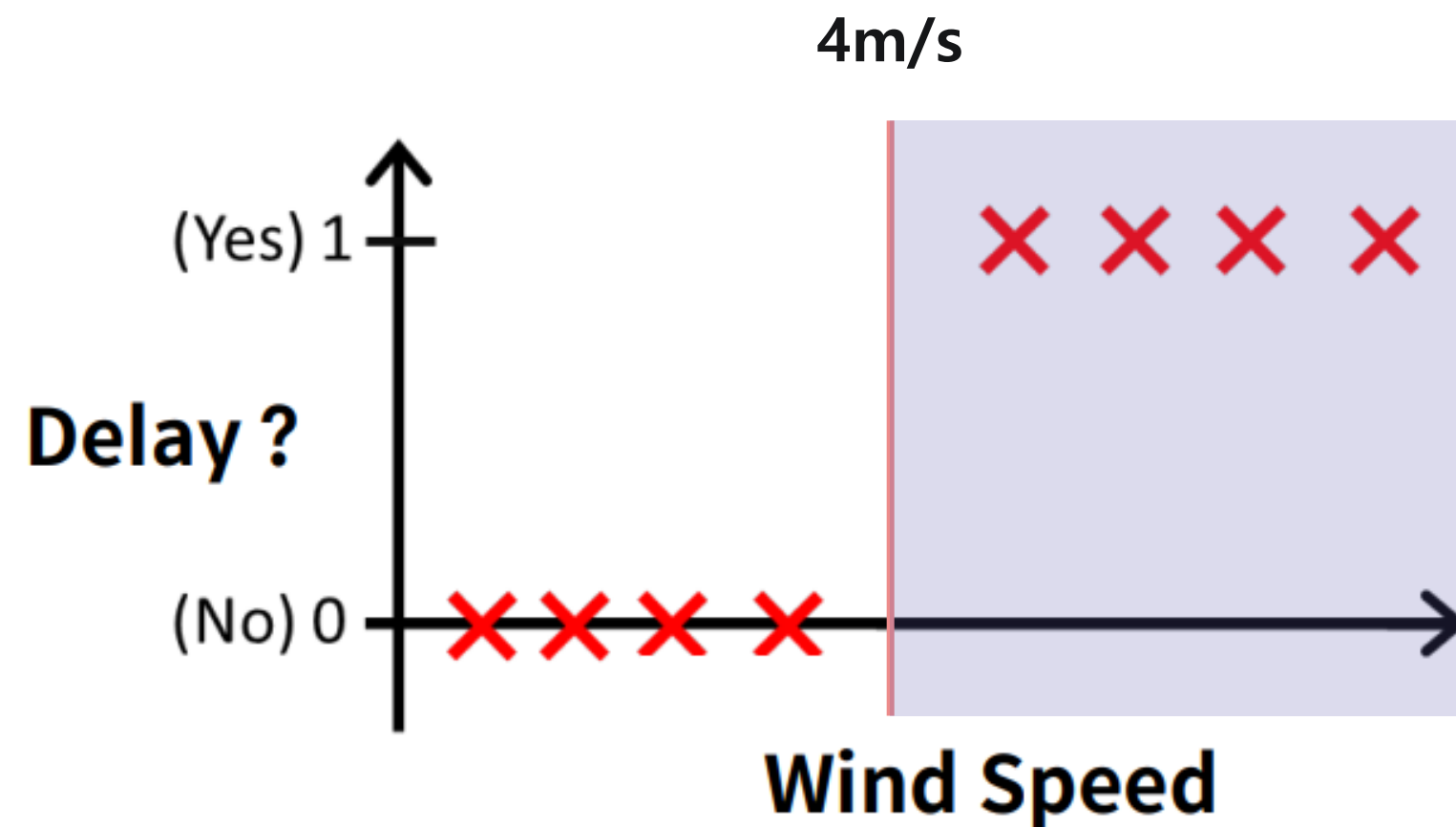
다양한 분류 알고리즘이 존재하며,  
예측 목표와 데이터 유형에 따라 적용



## ✓ 항공 지연 문제 해결하기

풍속  $4\text{m/s}$  를 기준으로 지연 여부를 나눠보자

- 풍속  $4\text{m/s}$  보다 크면 지연
- 풍속  $4\text{m/s}$  보다 작으면 지연 없음



✔ 어떠한 분류 알고리즘이 있을까?

분류 문제에 다양한 머신러닝 **모델**을 사용하여 해결

트리 구조 기반	의사결정나무, 랜덤포레스트, ...
확률 모델 기반	나이브 베이즈 분류기, ...
결정 경계 기반	선형 분류기, 로지스틱 회귀 분류기, SVM, ...
신경망	퍼셉트론, 딥러닝 모델, ...
...	...



02

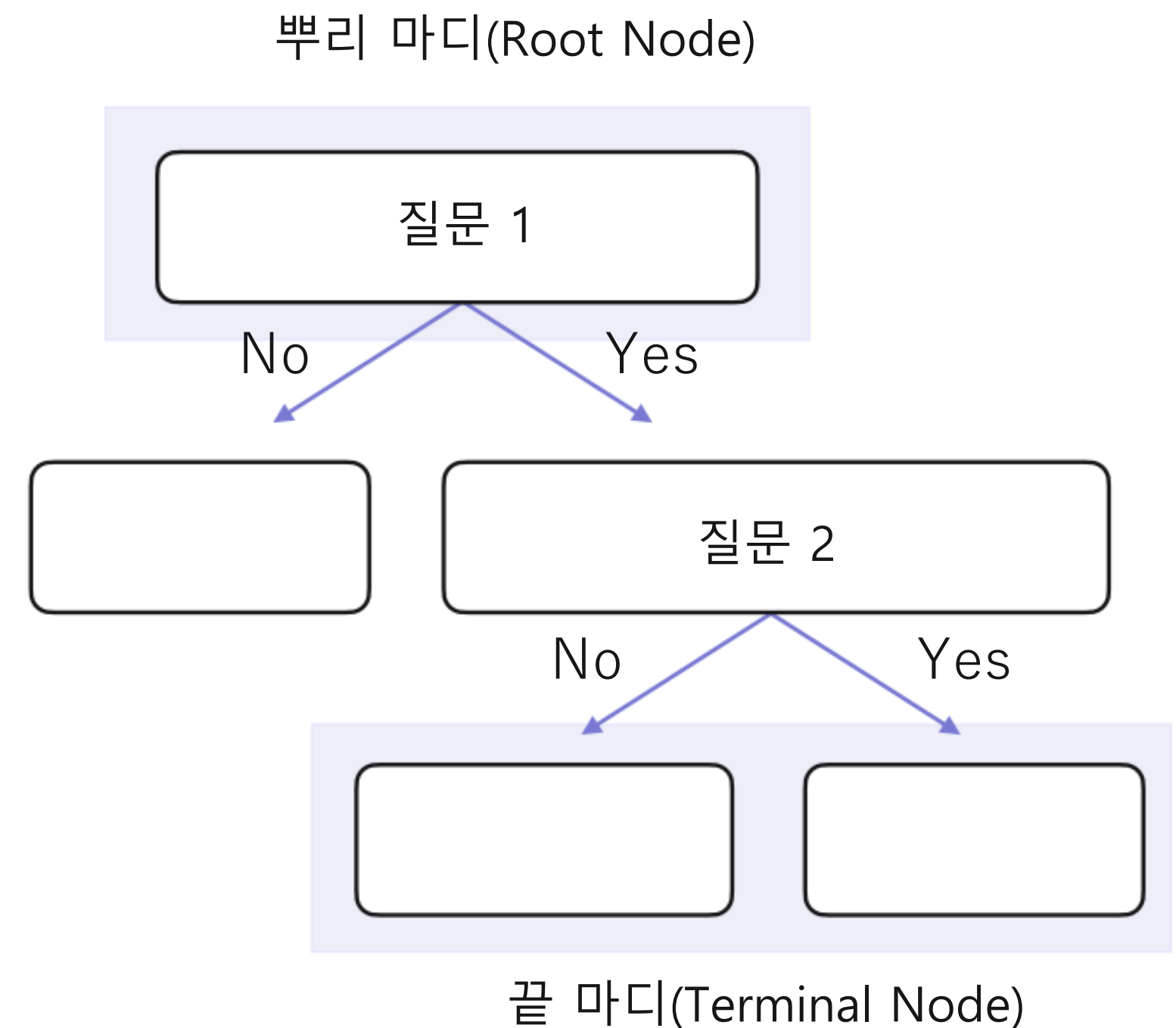
# 의사결정나무 - 모델 구조



## ✔ 의사결정나무(Decision Tree)란

스무고개와 같이 특정 질문들을 통해  
정답을 찾아가는 모델

최상단의 **뿌리 마디**에서  
마지막 **끝 마디**까지 아래 방향으로 진행



✔ 의사결정나무로 이해하기

항공 지연 데이터

풍속(m/s)	지연여부
1	No
1.5	No
2.5	No
5	Yes
5.5	Yes
6.5	Yes



뿌리 마디(Root Node)



풍속(m/s)	지연여부
1	No
1.5	No
2.5	No

끝 마디(Terminal Node)

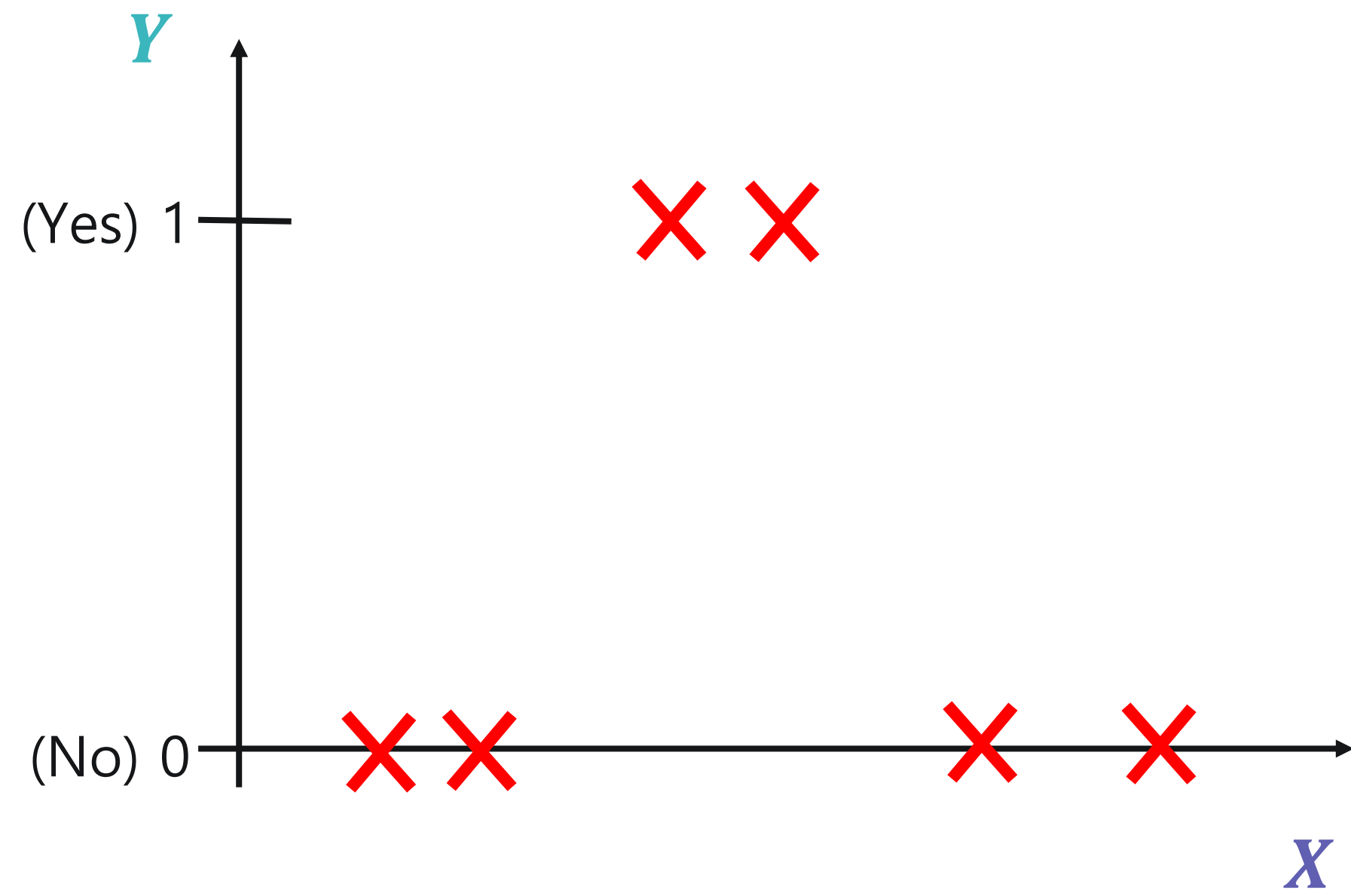
풍속(m/s)	지연여부
5	Yes
5.5	Yes
6.5	Yes

끝 마디(Terminal Node)

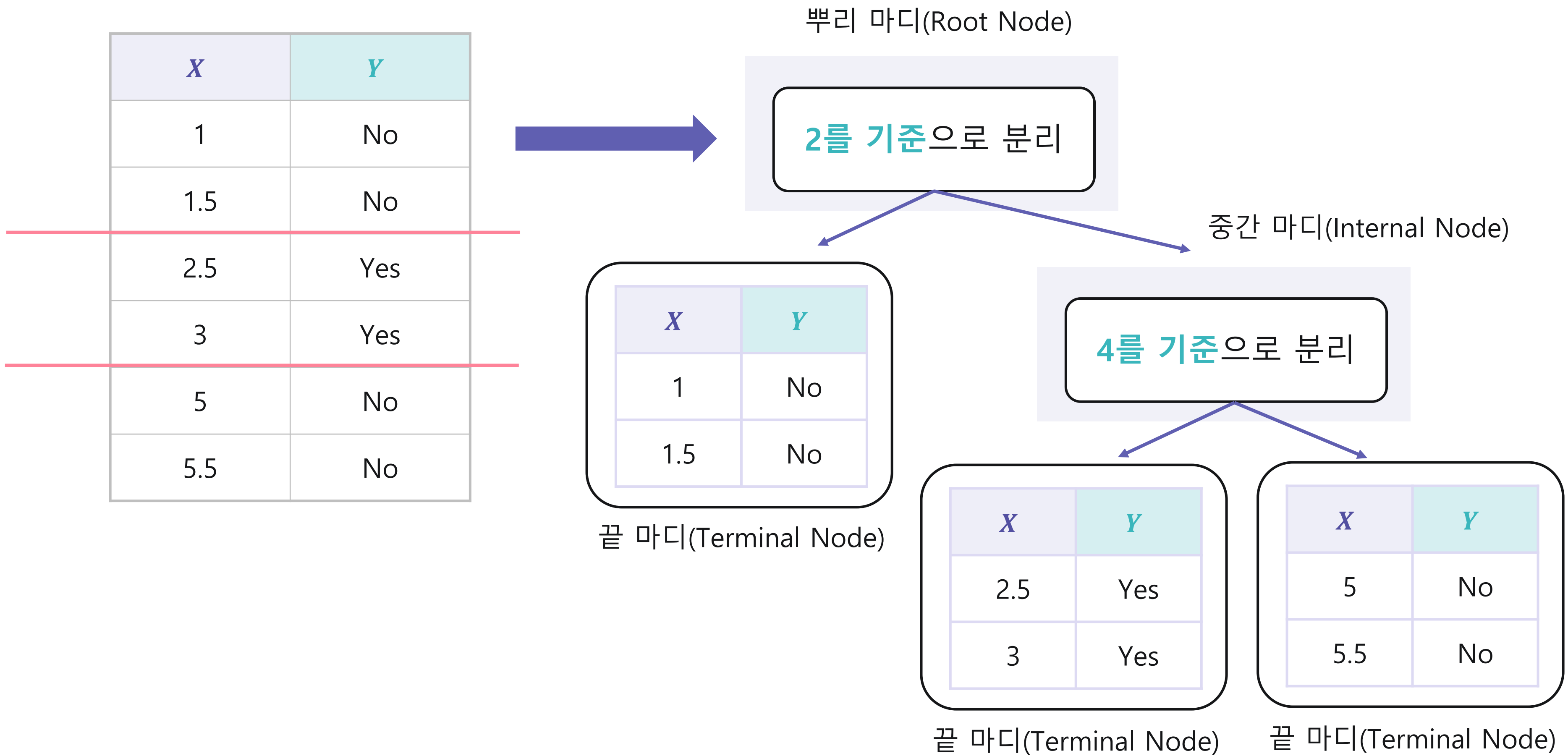
✓ 중간 마디 추가하기

아래와 같은 데이터는 어떻게 나뉘야 할까?

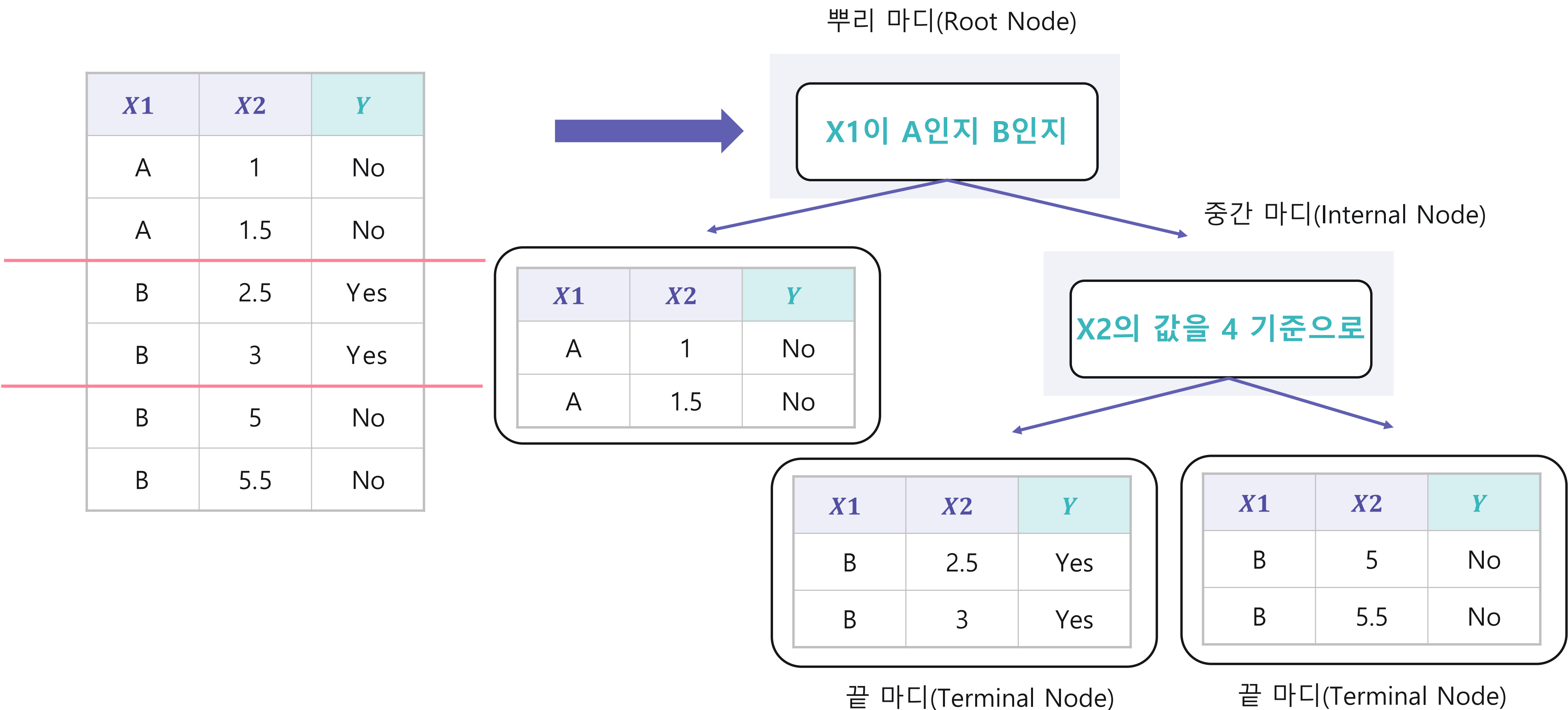
<i>X</i>	<i>Y</i>
1	No
1.5	No
2.5	Yes
3	Yes
5	No
5.5	No



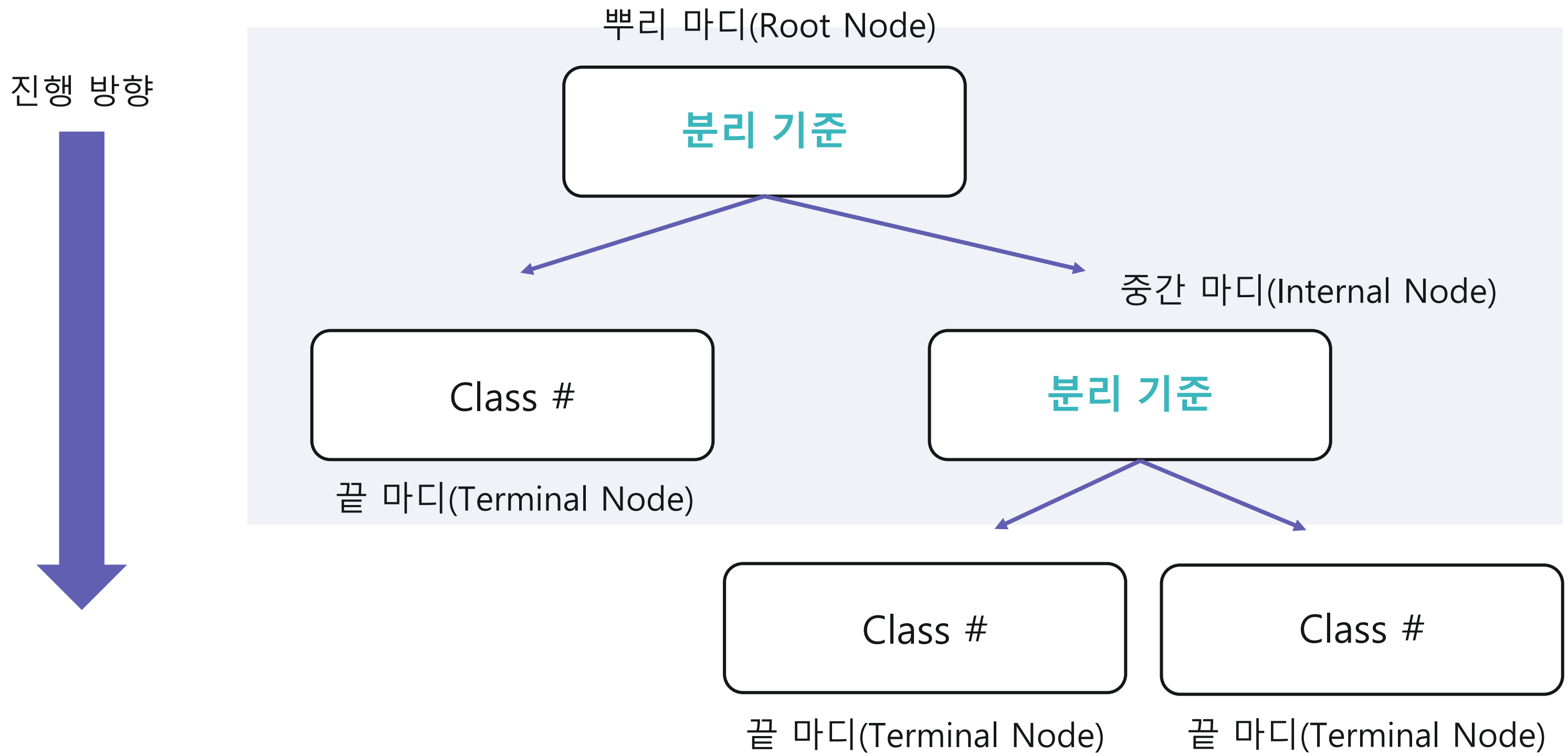
중간 마디 추가하기



✔ 2개 이상의 feature 데이터의 경우

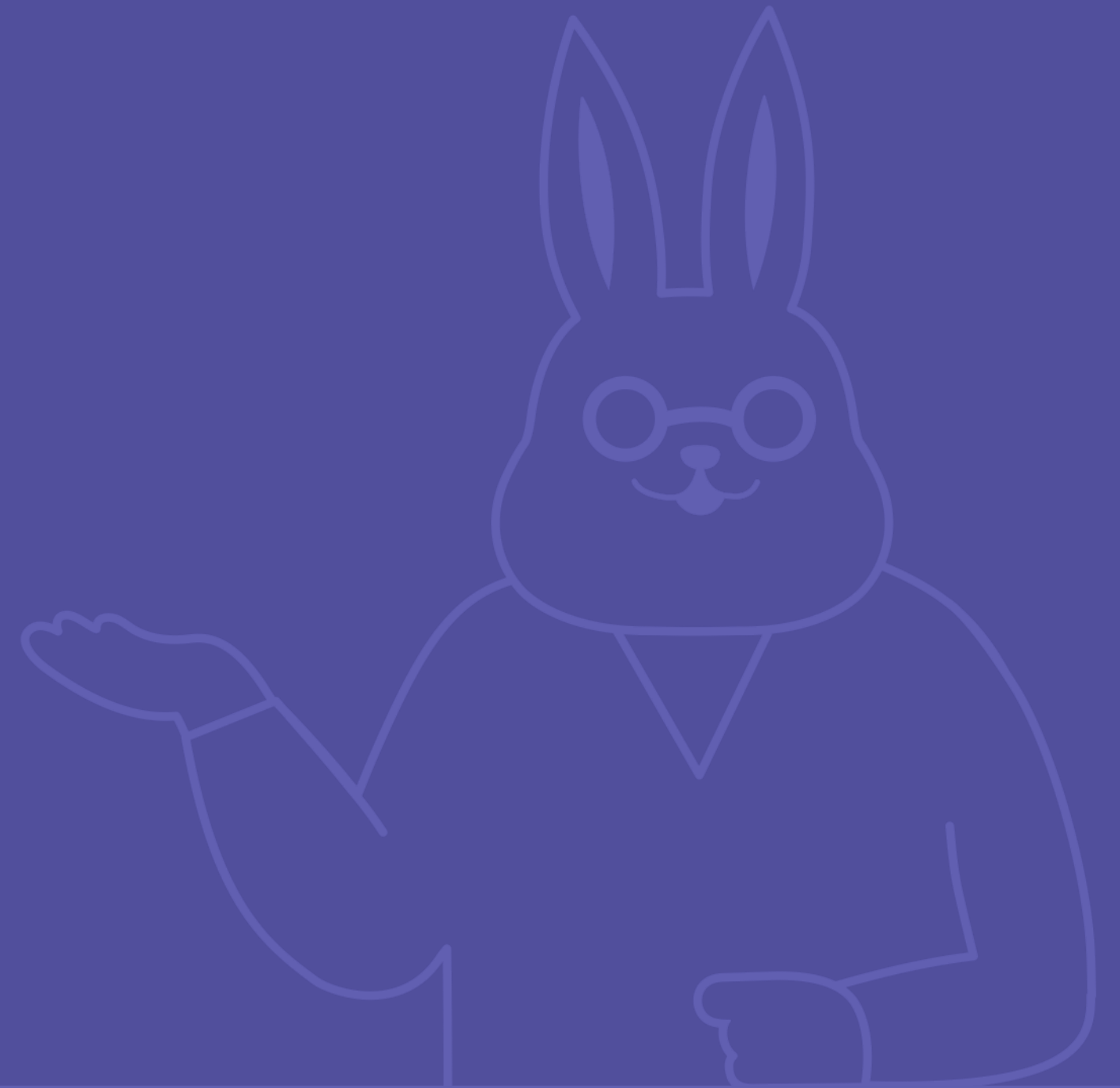


✔ 의사결정나무 구조 살펴보기



03

# 의사결정나무 - 불순도

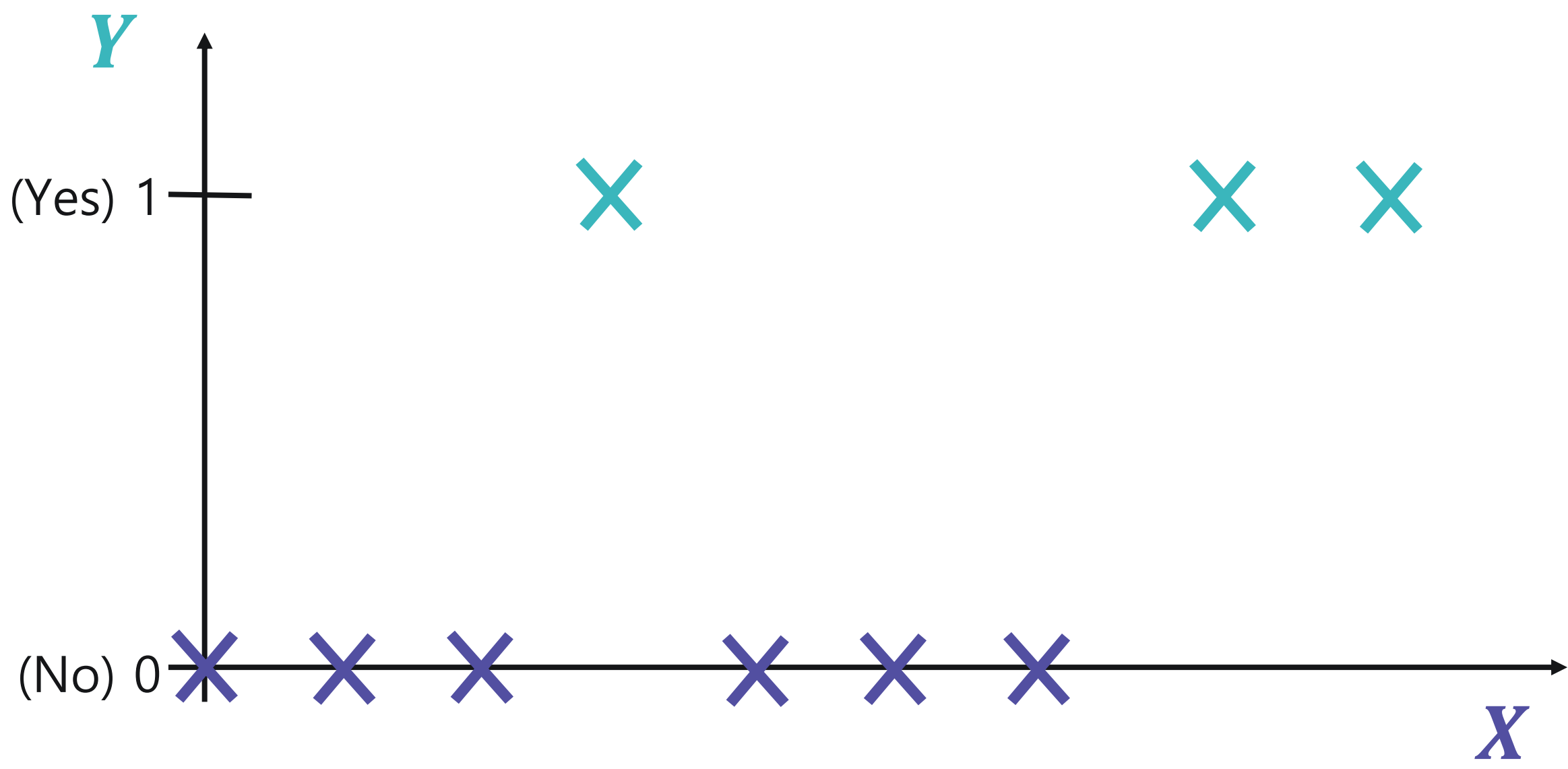




✔ 의사결정나무 분리 기준 알아보기

아래와 같은 데이터는 어떻게 나눠야 할까?

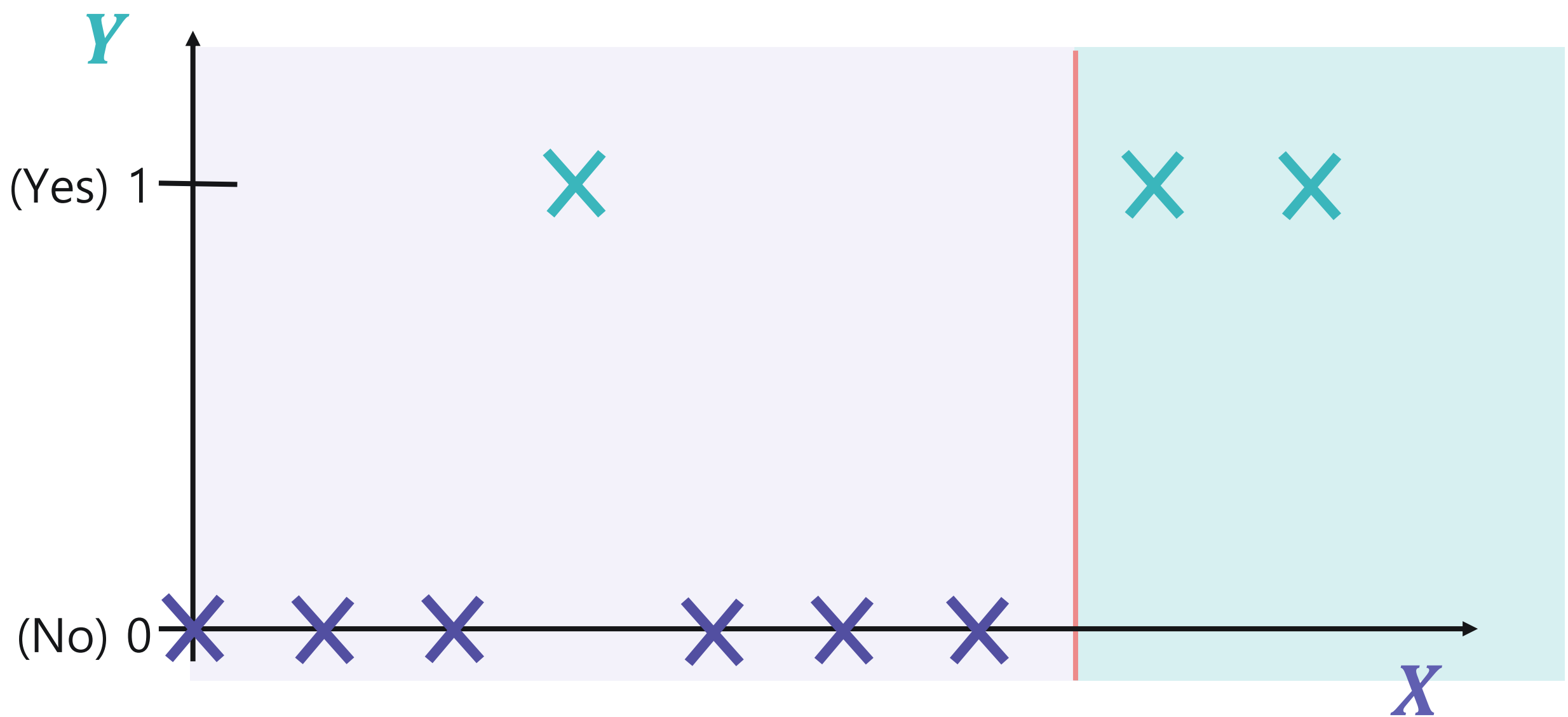
<i>X</i>	<i>Y</i>
1	No
2	No
3	Yes
4	No
5	No
6	No
7	Yes
8	Yes



✔ 의사결정나무 분리 기준 알아보기

데이터의 불순도(Impurity)를 최소화하는 구역으로 나누자!

<i>X</i>	<i>Y</i>
1	No
2	No
3	Yes
4	No
5	No
6	No
7	Yes
8	Yes



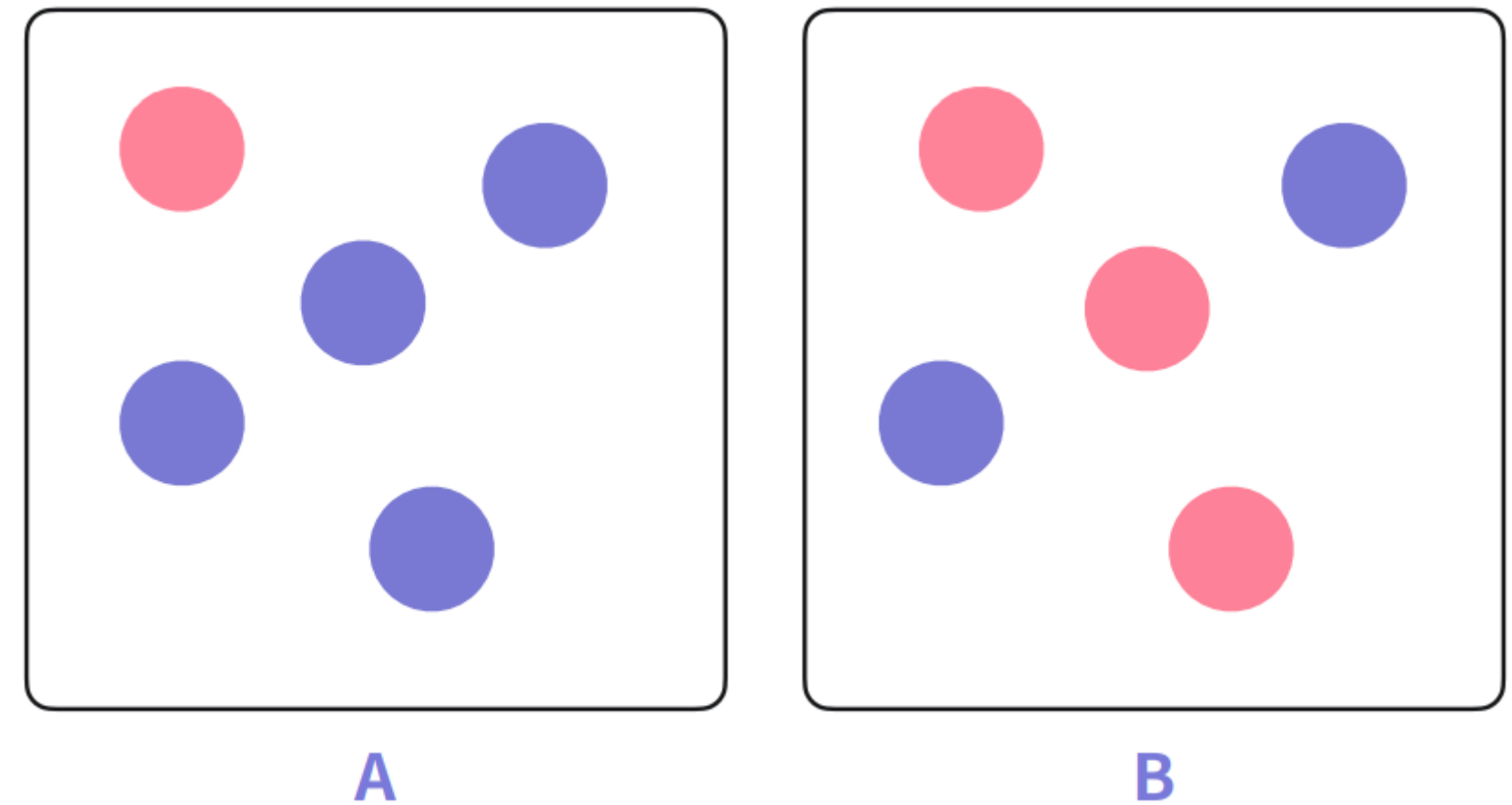
## ✓ 불순도(Impurity)

- 불순도

다른 데이터가 섞여 있는 정도

- 데이터 셋 A와 B 중 불순도가 더 낮은 것은?  
정답은 A!

데이터의 개수가 적기 때문에 눈으로 확인  
그렇다면 수많은 데이터가 존재할 때  
불순도는 어떻게 측정할 수 있을까?



## ✓ 불순도 측정 방법, 지니 불순도(Gini Impurity)

- 지니 계수(Gini Index)

해당 구역 안에서 특정 클래스에 속하는 데이터의 비율을 모두 제외한 값 즉, **다양성을 계산**하는 방법

- 지니 불순도(Gini Impurity)

$$\text{Gini Index} = 1 - (\text{yes의 확률})^2 - (\text{no의 확률})^2$$

$$\text{Gini Impurity} = \frac{n_1}{N} \text{Gini}_1 + \frac{n_2}{N} \text{Gini}_2$$

$n_i$ : i번째 자식 마디의 데이터 개수

N: 부모 마디의 데이터 개수

✔ 지니 불순도(Gini Impurity) 계산하기

**Gini Index** =  $1 - (yes\text{의 확률})^2 - (no\text{의 확률})^2$

**Gini Impurity** =  $\frac{n_1}{N} Gini_1 + \frac{n_2}{N} Gini_2$

X	Y
1	No
2	No
3	Yes
4	No
5	No
6	No
7	Yes
8	Yes

Yes	No	Gini index
0	2	$1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$
3	3	$1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = \frac{1}{2}$

**Gini Impurity** =  $\frac{2}{8} \times 0 + \frac{6}{8} \times \frac{1}{2} = \frac{3}{8}$

✓ 지니 불순도(Gini Impurity) 계산하기

$$\text{Gini Index} = 1 - (yes\text{의 확률})^2 - (no\text{의 확률})^2$$

$$\text{Gini Impurity} = \frac{n_1}{N} Gini_1 + \frac{n_2}{N} Gini_2$$

X	Y
1	No
2	No
3	Yes
4	No
5	No
6	No
7	Yes
8	Yes

Gini Impurity = 0.375

Gini Impurity = 0.467

Gini Impurity = 0.438

Gini Impurity = 0.367

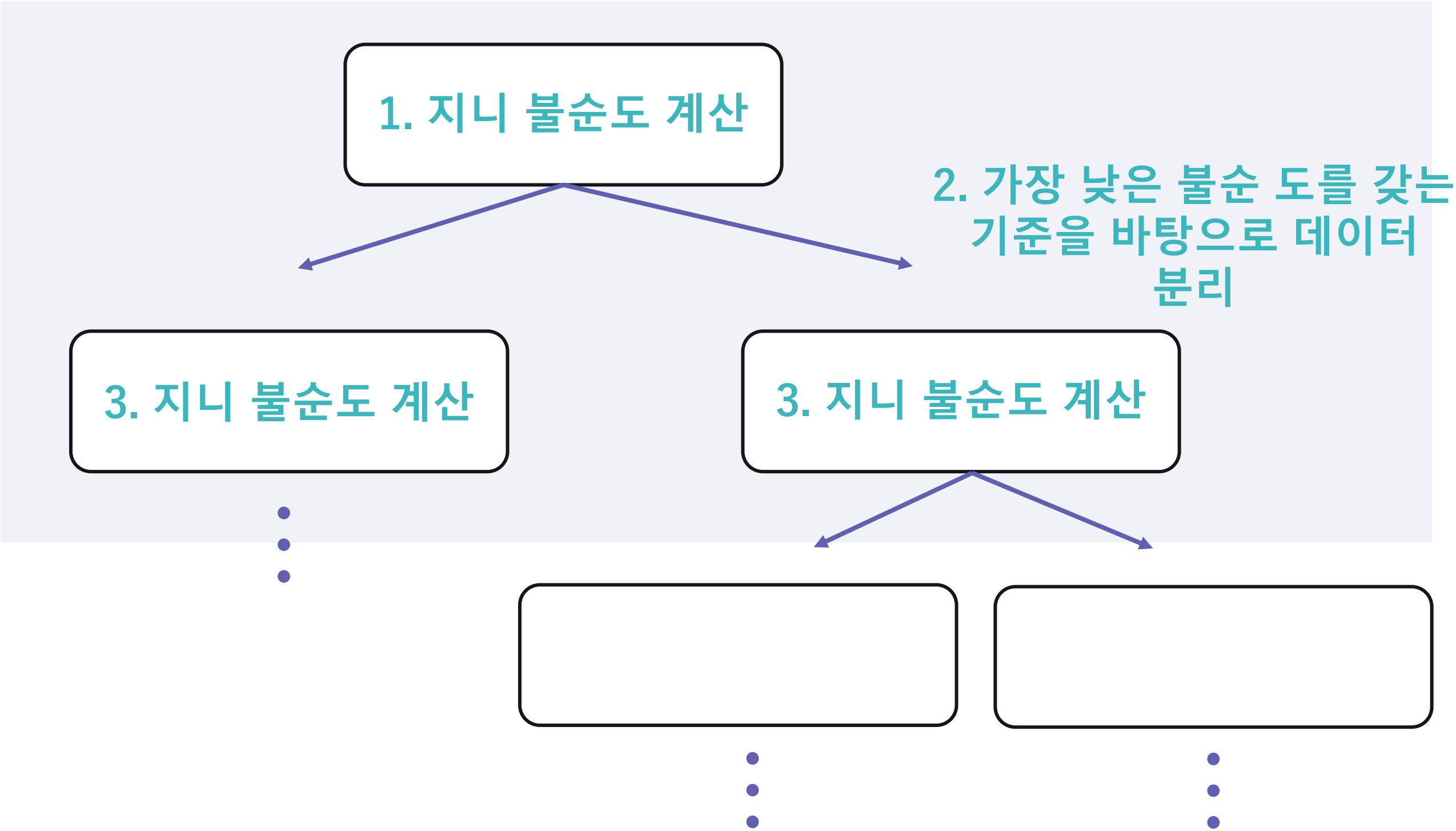
Gini Impurity = 0.208



가장 낮은 Gini 불순도를  
갖는 기준을 선택

✔ 지니 불순도(Gini Impurity) 적용하기

진행 방향

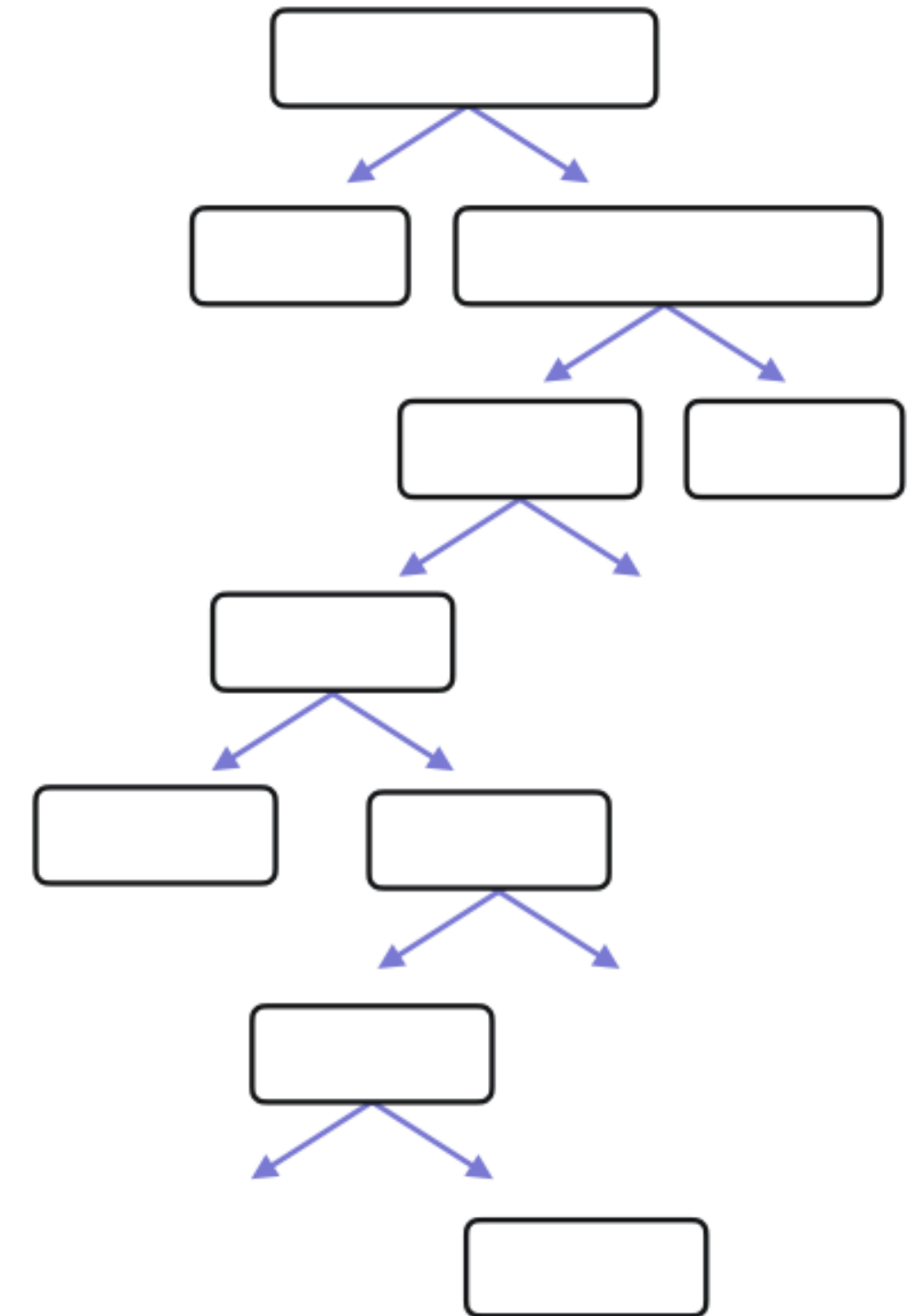


## ☑ 의사결정나무의 깊이의 trade-off

의사결정나무의 깊이가 깊어질 수록 세분화해서 나눌 수 있음

하지만 너무 깊은 모델은 과적합을 야기할 수 있음

⇒ 데이터에 따라 다를 수 있지만 **너무 깊은 모델은 지양**





### ✓ 의사결정나무의 특징

- 결과가 직관적이며, 해석하기 쉬움
- 나무 깊이가 깊어질수록 과적합(Overfitting) 문제 발생 가능성이 매우 높음
- 학습이 끝난 트리의 작업 속도가 매우 빠르다

04

# 분류 평가 지표



✔ 혼동 행렬(Confusion Matrix)

분류 모델의 성능을 평가하기 위함

		예측	
		Positive	Negative
실제	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

## ✓ 혼동 행렬(Confusion Matrix)

**True Positive:** 실제 **Positive** 인 값을 **Positive** 라고 예측(정답)

**True Negative:** 실제 **Negative** 인 값을 **Negative** 라고 예측(정답)

**False Positive:** 실제 **Negative** 인 값을 **Positive** 라고 예측(오답) – 1형 오류

**False Negative:** 실제 **Positive** 인 값을 **Negative** 라고 예측(오답) – 2형 오류

## ✓ 정확도(Accuracy)

전체 데이터 중에서 제대로 분류된 데이터의 비율로,  
**모델이 얼마나 정확하게 분류**하는지를 나타냄

일반적으로 분류 모델의 주요 평가 방법으로 사용됨

그러나, 클래스 비율이 **불균형** 할 경우  
평가 지표의 신뢰성을 잃을 가능성이 있음

$$Accuracy = \frac{TP+TN}{P+N}$$

$$P: TP + FN,$$

$$N: TN + FP$$

## ✓ 정밀도(Precision)

모델이 Positive라고 분류한 데이터 중에서 실제로 Positive인 데이터의 비율

### Negative가 중요한 경우

즉, 실제로 Negative인 데이터를 Positive라고 판단하면 안되는 경우 사용되는 지표

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

## ✓ Negative가 중요한 경우

스팸 메일 판결을 위한 분류 문제

해당 메일이 스팸일 경우 **Positive**,  
스팸이 아닐 경우 즉, 일반 메일일 경우 **Negative**

일반 메일을 **스팸 메일(Positive)**로 잘못 예측했을 경우  
중요한 메일을 전달받지 못하는 상황이 발생할 수 있음

## ✓ 재현율(Recall, TPR)

실제로 Positive인 데이터 중에서  
모델이 Positive로 분류한 데이터의 비율

### Positive가 중요한 경우

즉, 실제로 Positive인 데이터를  
Negative라고 판단하면 안되는 경우 사용되는  
지표

$$Recall = \frac{TP}{TP+FN} = \frac{TP}{P}$$



## ✓ Positive가 중요한 경우

악성 종양 여부 판결을 위한 검사

악성 종양일 경우 **Positive**,

악성 종양이 아닐 경우 즉, 양성 종양일 경우 **Negative**

악성 종양(Positive)을 **양성 종양(Negative)으로 잘못 예측**했을 경우  
제 때 치료를 받지 못하게 되어 생명이 위급해질 수 있음

## ✓ 다양한 분류 지표의 활용

분류 목적에 따라 다양한 지표를 계산하여 평가

- 분류 결과를 전체적으로 보고 싶다면 → **혼동 행렬(Confusion Matrix)**
- 정답을 얼마나 잘 맞췄는지 → **정확도(Accuracy)**
- FP 또는 FN의 중요도가 높다면 → **정밀도(Precision), 재현율(Recall)**

# 크레딧

/\* elice \*/

코스 매니저

이해솔

콘텐츠 제작자

이해솔

강사

이해솔

감수자

-

디자이너

강혜정

# 연락처

TEL

070-4633-2015

WEB

<https://elice.io>

E-MAIL

[contact@elice.io](mailto:contact@elice.io)

