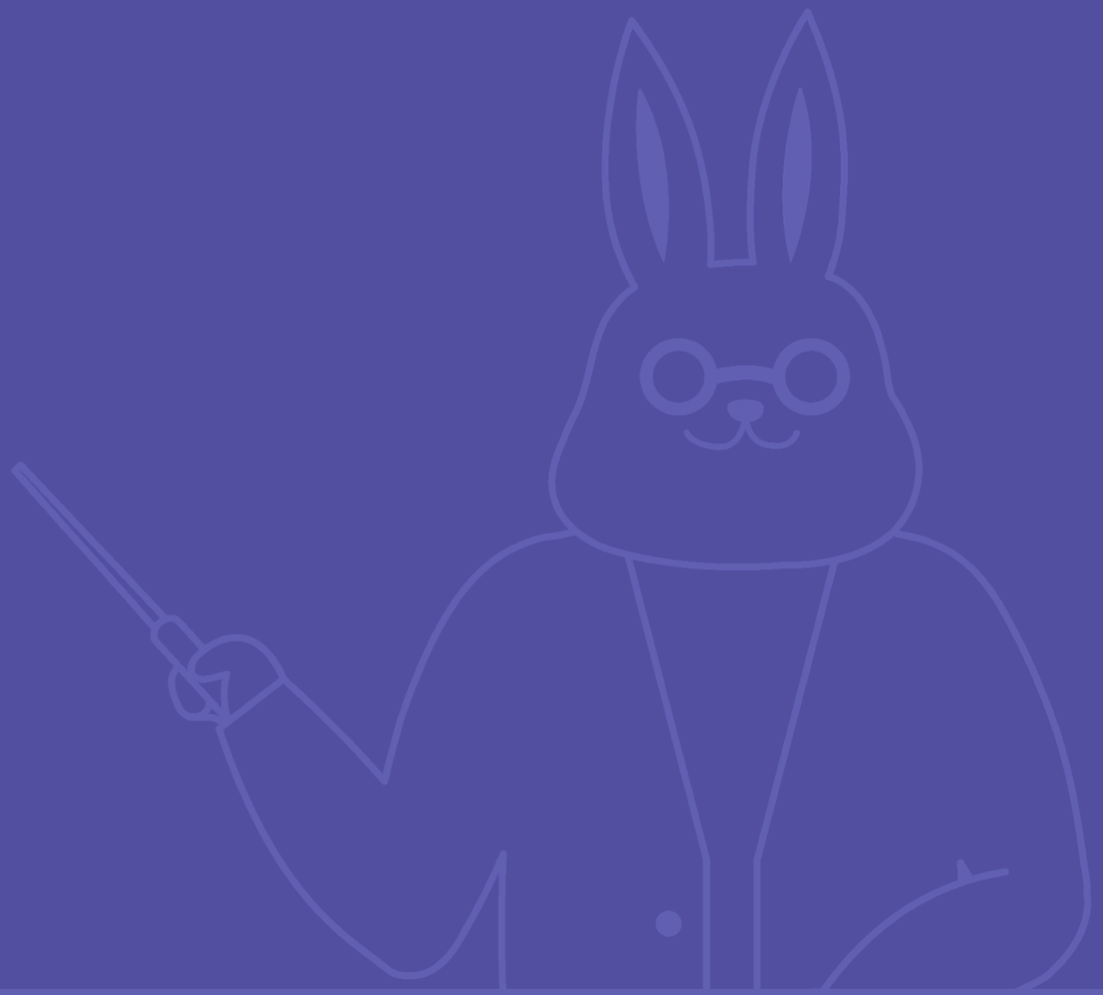
머신러닝 시작하기

04 지도학습 - 분류



목차

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 여부**를 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 분석 알고리즘

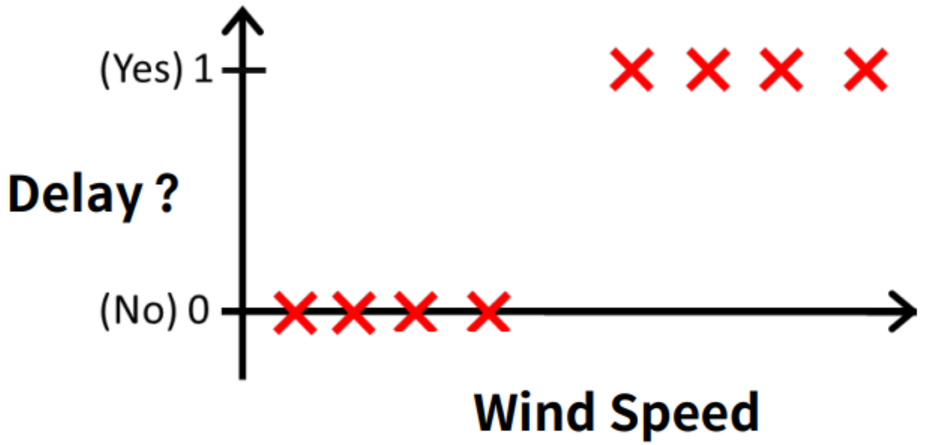
X Y

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

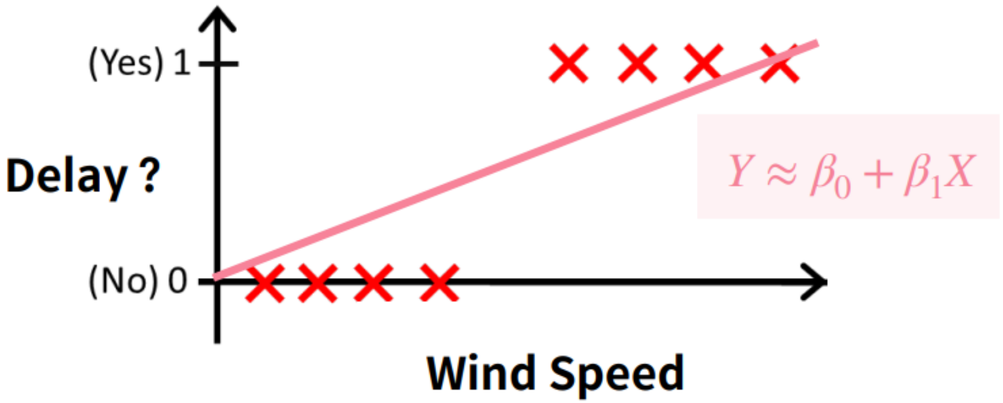


분류 문제에 회귀 알고리즘 적용하기

일반적인 회귀 알고리즘은 분류 문제에
그대로 사용할 수 없다!

Why?
회귀 결과 값은 **-∞ ~ +∞**의 값을 가질 수 있음

Q. 우리의 목표는
지연 여부 판별인데 결과값이 1000이라면?



그렇다면 어떻게 해야 할까?

분류 문제에 다양한 머신러닝 모델을 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, …
…	…

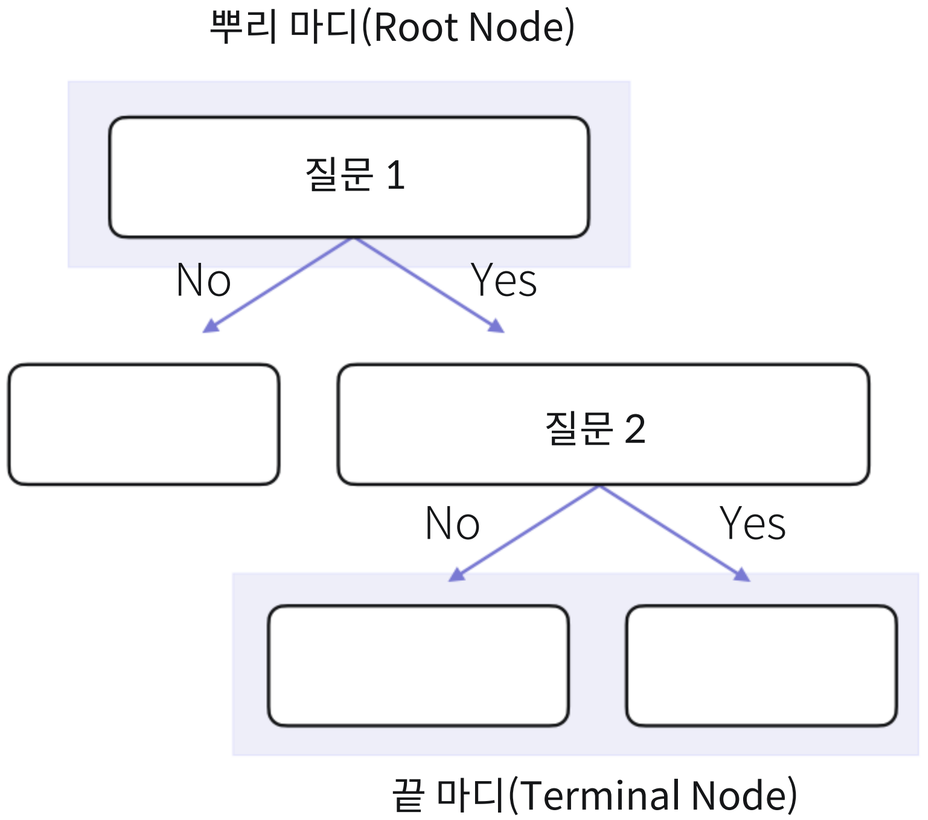
02

의사결정나무 - 모델 구조

의사결정나무(Decision Tree)란

스무고개와 같이 특정 질문들을 통해
정답을 찾아가는 모델

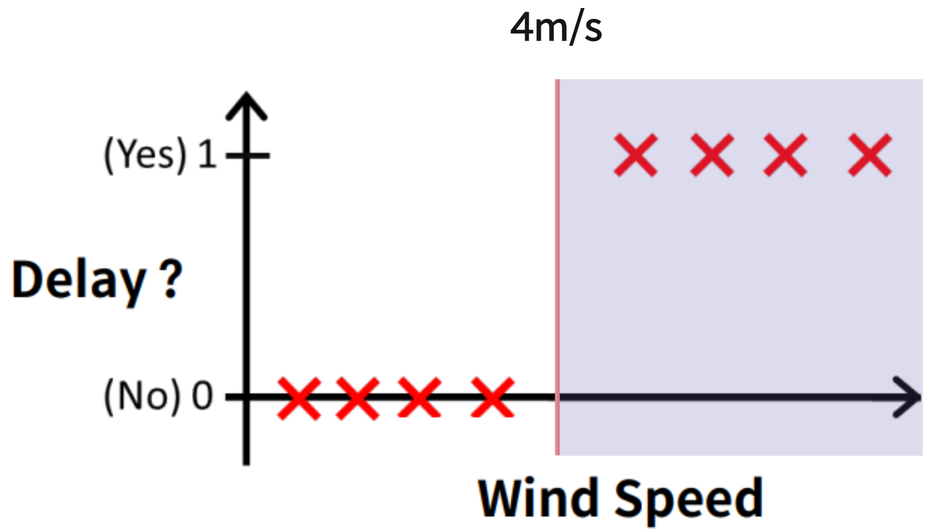
최상단의 **뿌리 마디**에서
마지막 **끝 마디**까지 아래 방향으로 진행

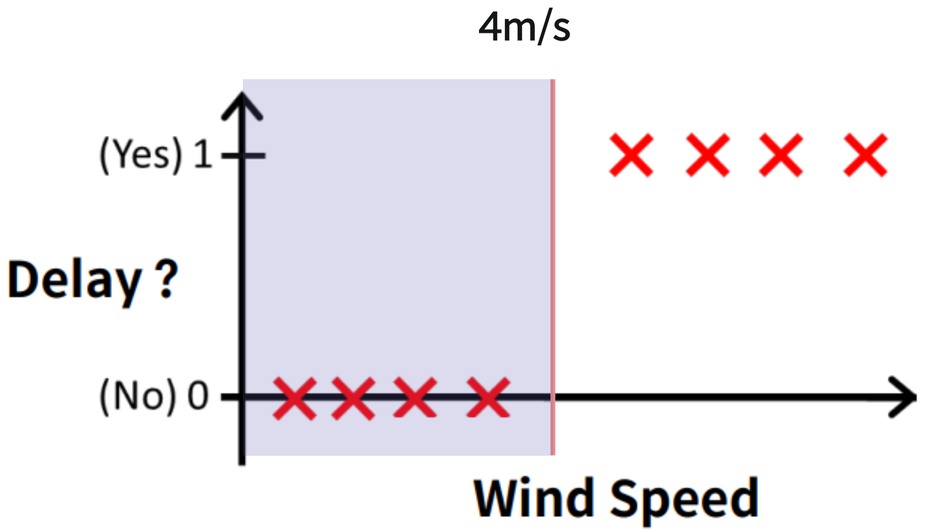


항공 지연 문제 해결하기

풍속 4m/s 를 기준으로 지연 여부를 나눠보자

- 풍속 4m/s 보다 크면 지연
- 풍속 4m/s 보다 작으면 지연 없음





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 4m/s 를 기준으로 분리

풍속(m/s)	지연여부
1	No
1.5	No
2.5	No

끝 마디(Terminal Node)

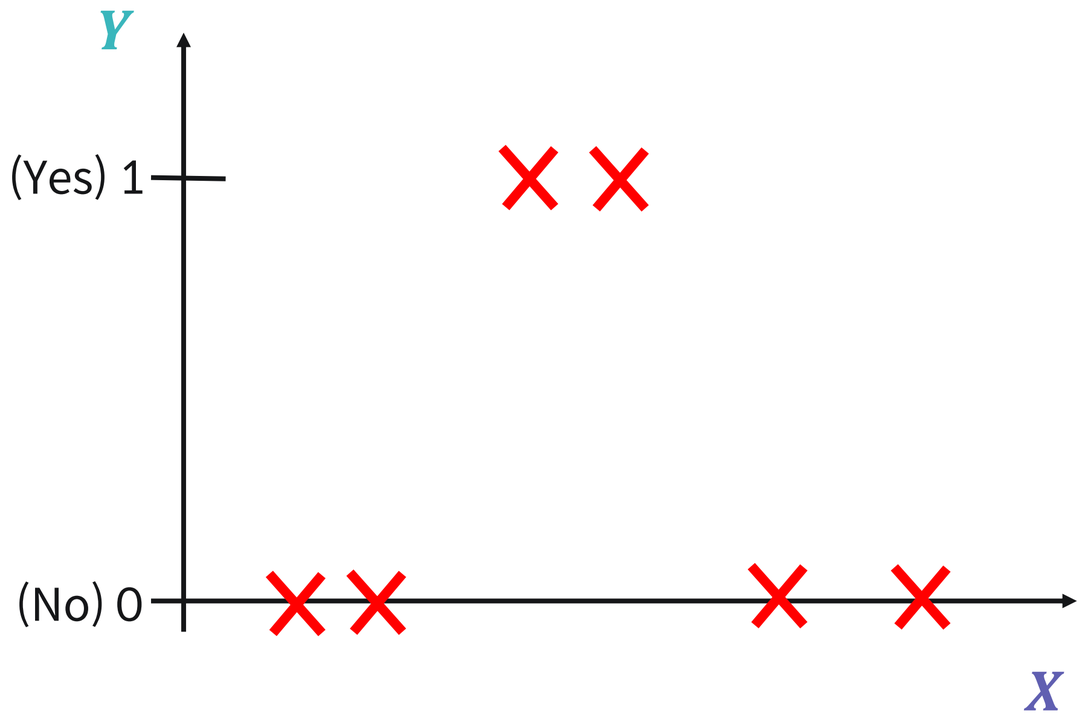
풍속(m/s)	지연여부
5	Yes
5.5	Yes
6.5	Yes

끝 마디(Terminal Node)

중간 마디 추가하기

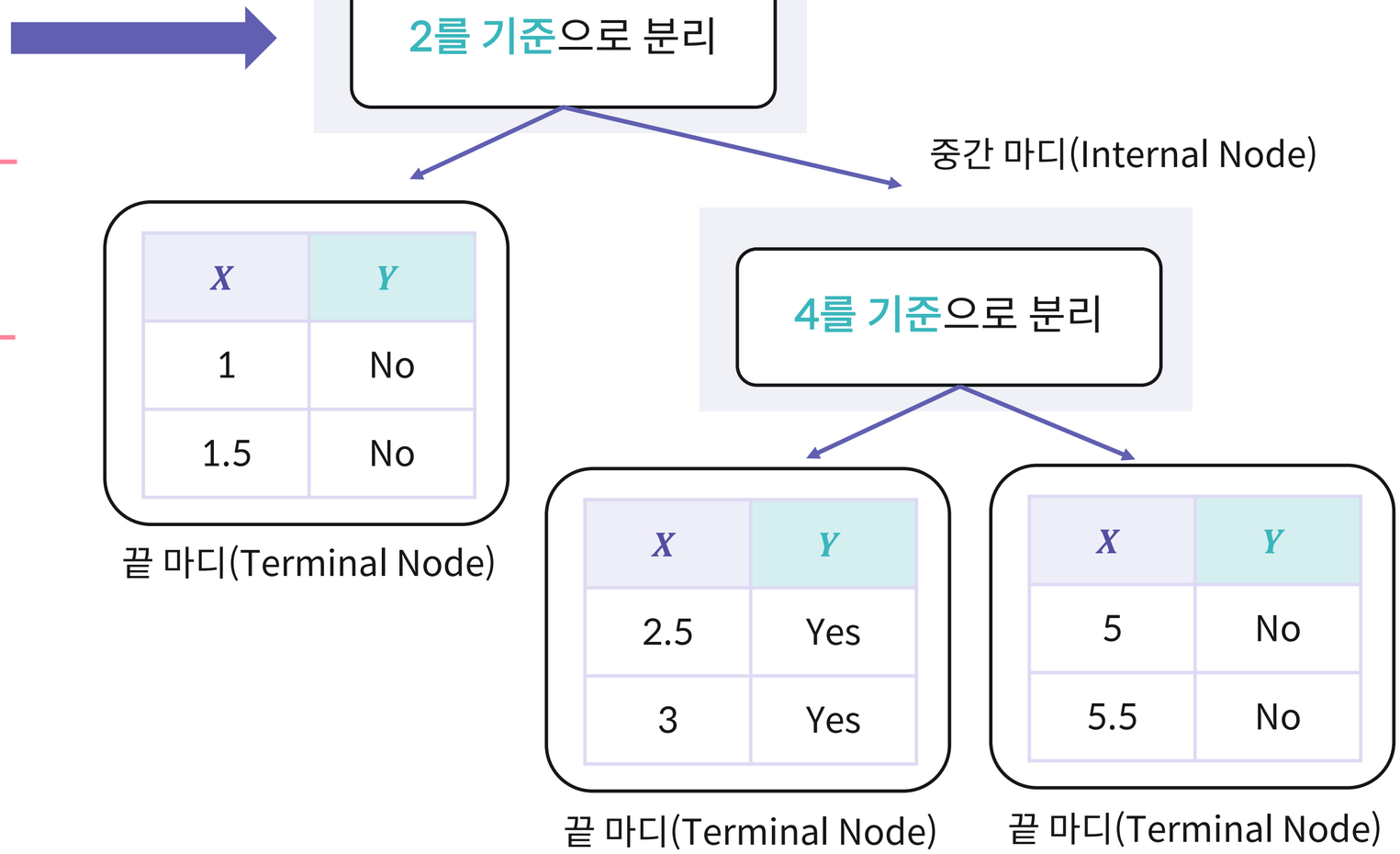
아래와 같은 데이터는 어떻게 나눠야 할까?

X	Y
1	No
1.5	No
2.5	Yes
3	Yes
5	No
5.5	No



중간 마디 추가하기

X	Y
1	No
1.5	No
2.5	Yes
3	Yes
5	No
5.5	No



X	Y
1	No
1.5	No

X	Y
2.5	Yes
3	Yes

X	Y
5	No
5.5	No

2개 이상의 feature 데이터의 경우

$X1$	$X2$	Y
A	1	No
A	1.5	No
B	2.5	Yes
B	3	Yes
B	5	No
B	5.5	No

뿌리 마디(Root Node)

X1이 A인지 B인지

중간 마디(Internal Node)

$X1$	$X2$	Y
A	1	No
A	1.5	No

X2의 값을 4 기준으로

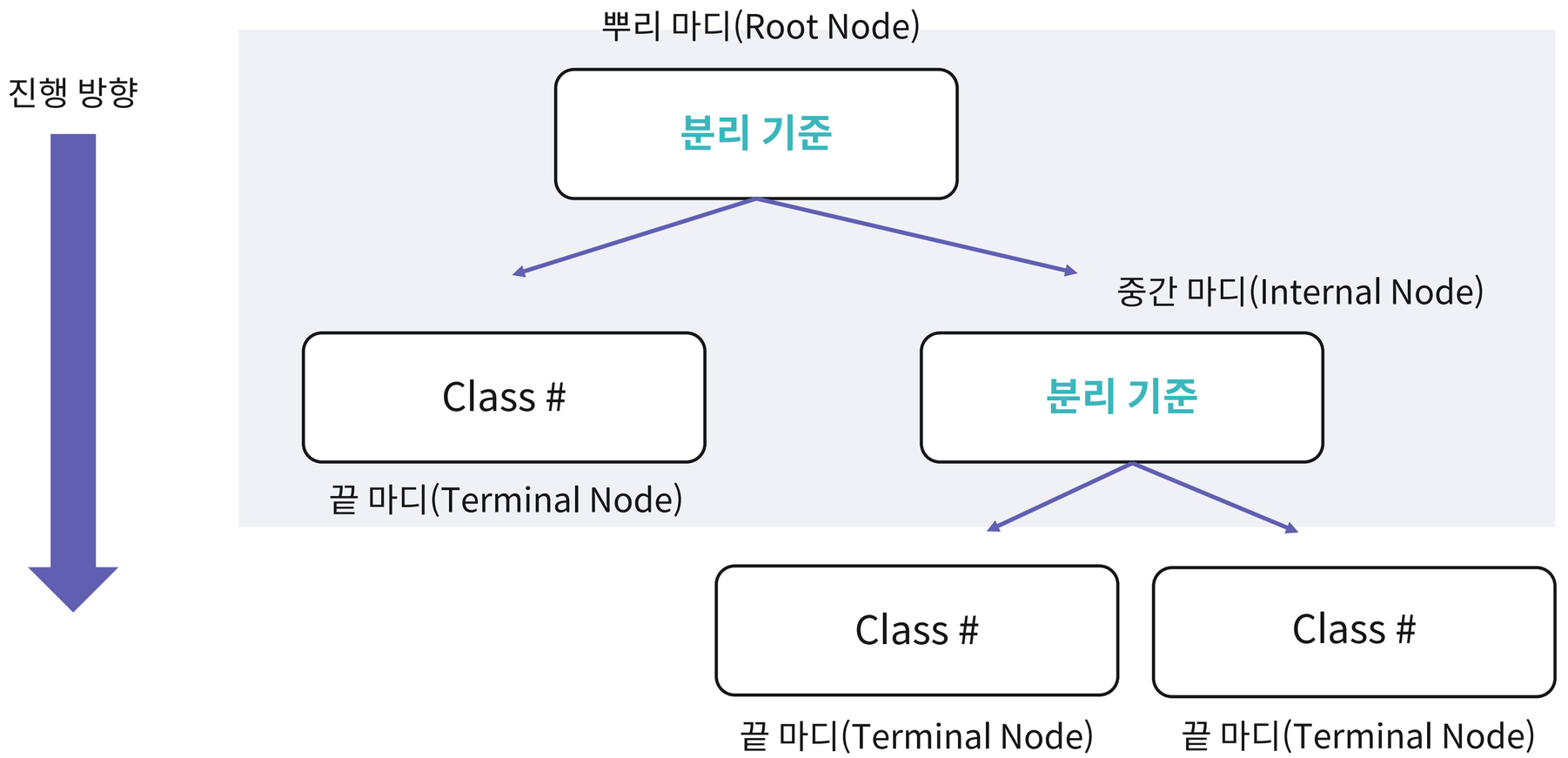
$X1$	$X2$	Y
B	2.5	Yes
B	3	Yes

끝 마디(Terminal Node)

$X1$	$X2$	Y
B	5	No
B	5.5	No

끝 마디(Terminal Node)

의사결정나무 구조 살펴보기



03

분류 평가 지표 (1)

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**

예시

전체 100개 항공기 관련 정보를 활용하여 지연 여부 예측을 실시했을 때 결과

실제 결과	예측 결과
지연 O : 20개 지연 X : 80개	지연 O : 60개 지연 X : 40개

예시

		예측	
		Positive	**Negative**
실제	**Positive**	True Positive : 20개	False Negative : 0개
	Negative	False Positive : 40개	True Negative : 40개

정확도(Accuracy)

전체 데이터 중에서 제대로 분류된 데이터의 비율로,
모델이 얼마나 정확하게 분류하는지를 나타냄

일반적으로 분류 모델의 주요 평가 방법으로 사용됨

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

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

$$P: TP + FN,$$

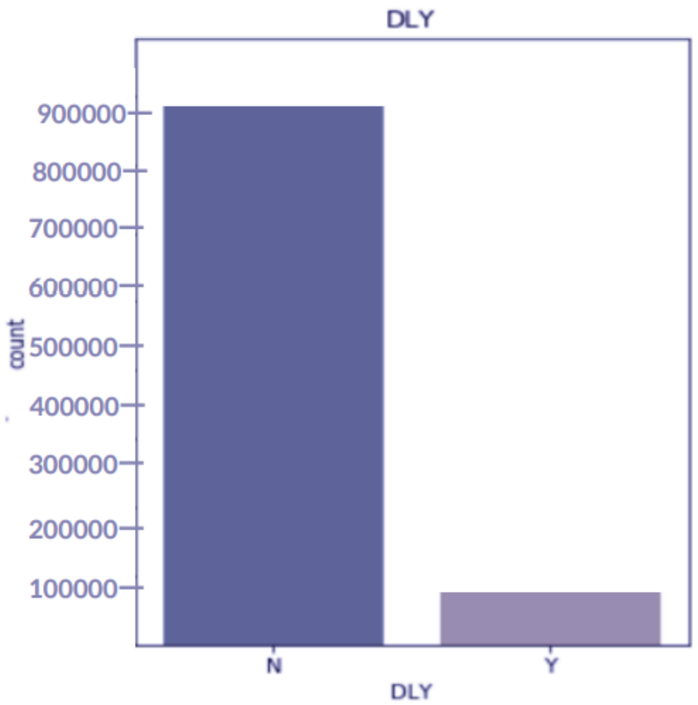
$$N: TN + FP$$

불균형한 클래스에서의 정확도

만약 전체 100만개 항공 데이터 중
90만개가 정상 운행,
10만개만이 지연 운행인 데이터를 예측하고자 할 때,

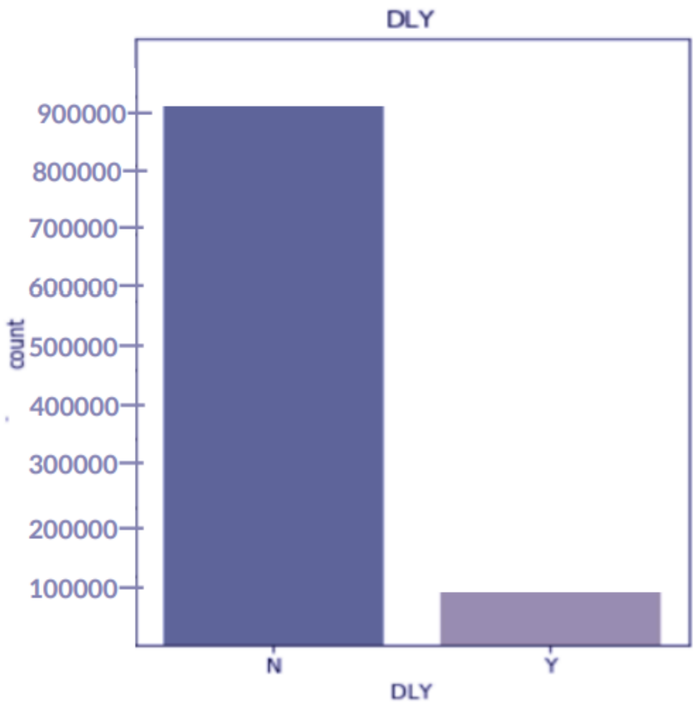
분류 모델 A가 전체 결과가 모두 지연되지 않았다고 예측할 경우

$$Accuracy = \frac{0개+90만개}{100만개} = 90\%$$



불균형한 클래스에서의 정확도

$$Accuracy = \frac{0개 + 90만개}{100만개} = 90\%$$



수치 상으로는 맞으나,
해당 모델에 대한 정확도가 90% 라고 단정하기에는 위험함
다양한 평가 지표 고려 필요성

04

분류 평가 지표 (2)

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 예측**했을 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

FPR(False Positive Rate)

실제로 Negative인 데이터 중에서
모델이 positive로 **잘못** 분류한 데이터의 비율

$$FPR = \frac{FP}{FP+TN} = \frac{FP}{N}$$

FPR 지표와 비정상 사용자 검출 예시

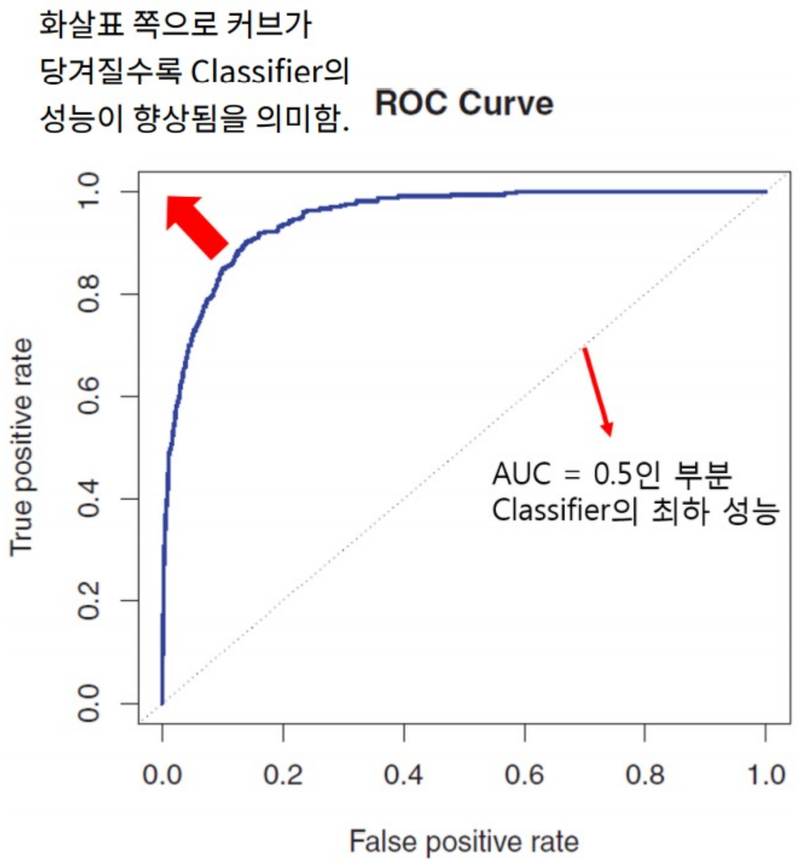
- 게임에서 비정상 사용자 검출 시 **FPR이 높다.**
  =정상 사용자를 **비정상 사용자**로 검출하는 경우가 많다.
- 이 때 비정상 사용자에 대해서 계정정지 등 페널티를 부여할 경우 선의의 사용자가 피해를 입게 될 가능성이 높음

→ 이는 곧 게임에 대한 충성도를 떨어트리는 계기가 될 수 있음

ROC Curve와 AUC

X축을 FPR(False Positive Rate)
Y축을 TPR(True Positive Rate)
로 두고 시각화한 그래프

ROC Curve 아래 면적인
AUC(Area Under Curve)를
이용해 모델의 성능을 평가



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)**
- 실제 Positive 데이터에 중요도가 높다면 → **정밀도(Precision), 재현율(Recall)**
- 실제 Negative 데이터에 중요도가 높다면 → **FPR**
- FPR과 TPR의 변화에 따른 모델의 전체적인 성능 → **ROC 그래프 및 AUC**

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