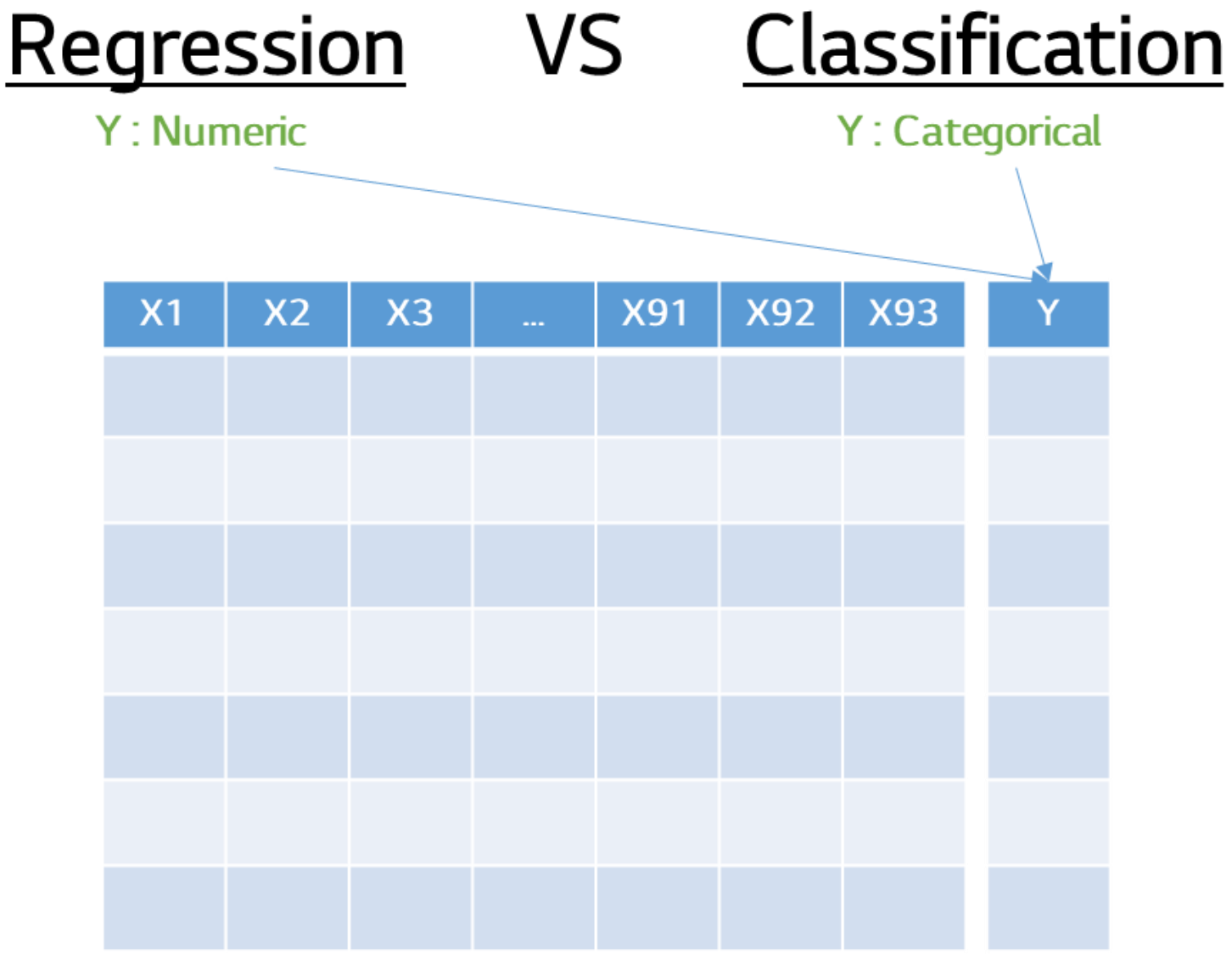
Classification Problem

Data Scientist
안건이

목차

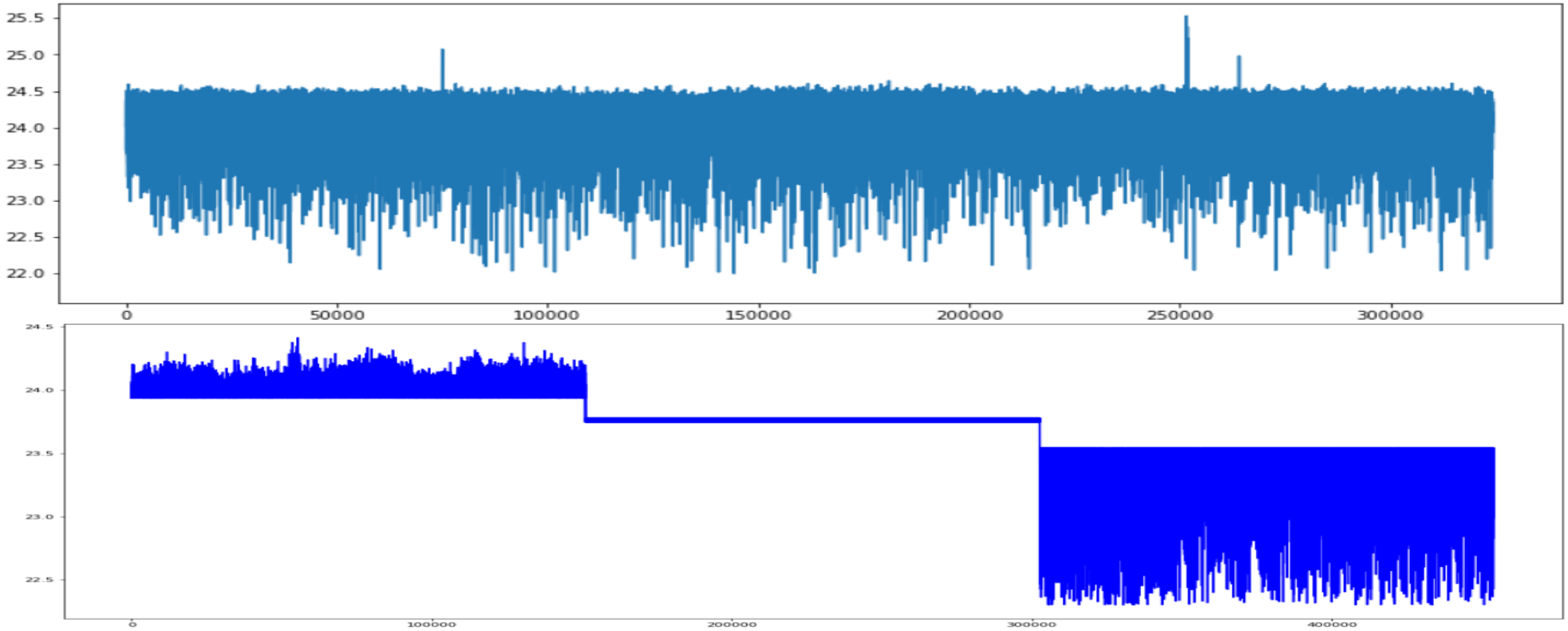
Regression vs Classification

- 모든 Model(Algorithm)은 Loss Function을 가지고 있음
  - 크게는 Regression, Classification



Why Classification ?

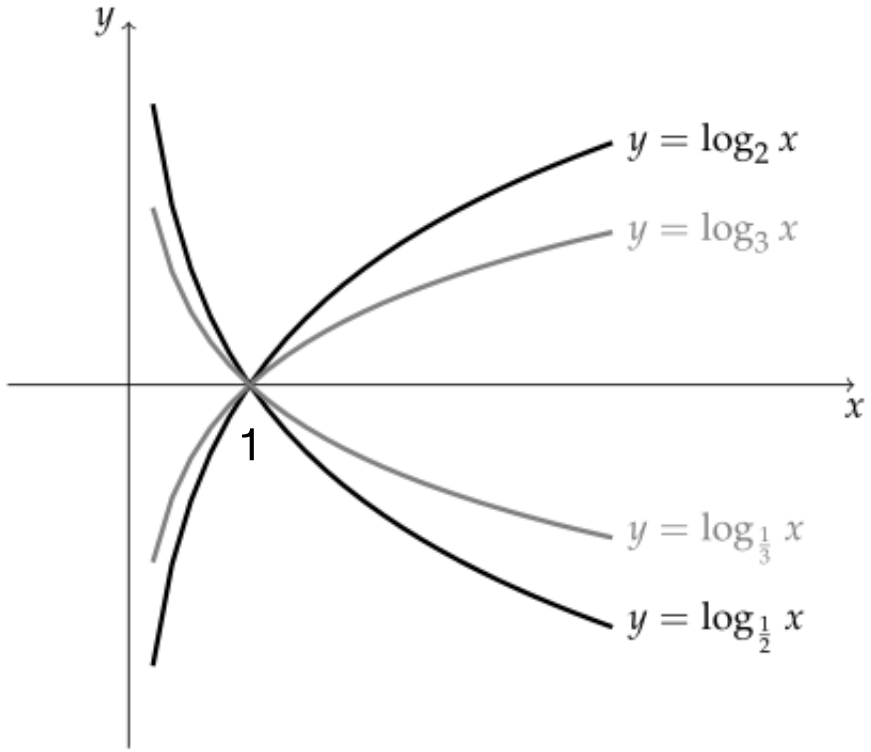
- 결국 우리의 목적은 무엇인가?
  - 고효율(상위 10%) vs 저효율(하위 10%)를 구분 짓는 변수는 무엇인가?
  - 고효율(상위 10%) vs 중효율(중위 10%)를 구분 짓는 변수는 무엇인가?
- 굳이 불필요한 데이터까지 학습해야 하는가?
  - Complexity만 증가함, 타켓한 데이터만 집중하자
  - Regression의 경우 복잡도가 매우 상승함
    - “0~100” vs “1 또는 0” → 어떤게 맞추기 쉬울까?
- Regression에도 데이터 Imbalance가 존재한다고 생각함
  - 개인적인 생각 (토론)

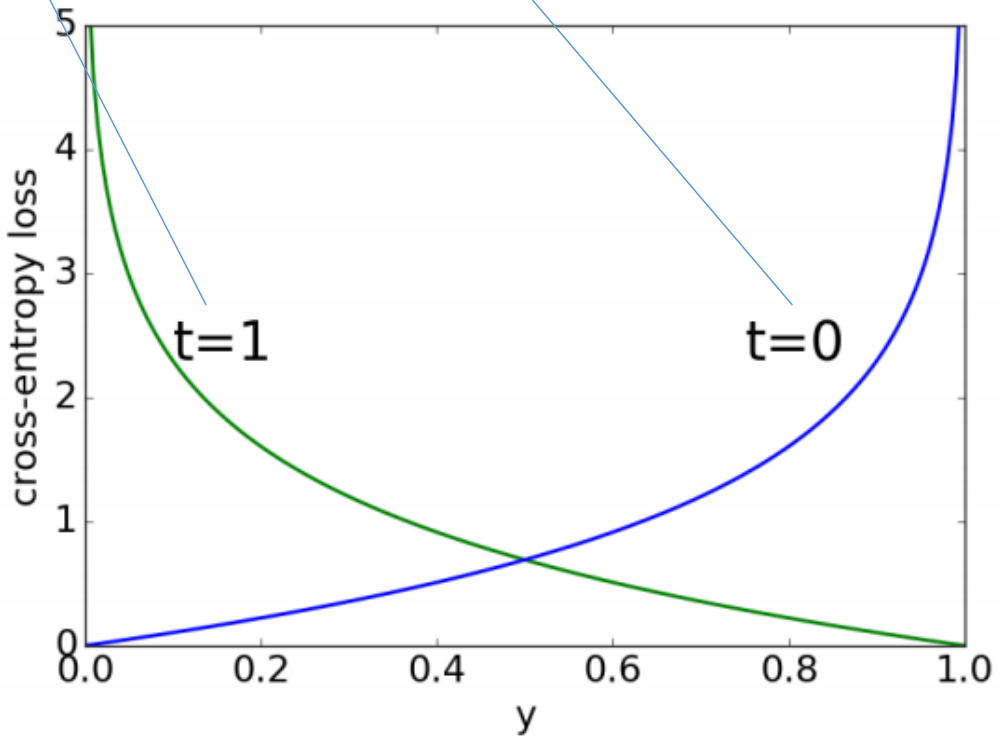


Classification Loss Function Remind

- Classification Loss Function
  - 예측 값을 확률로 뱉음

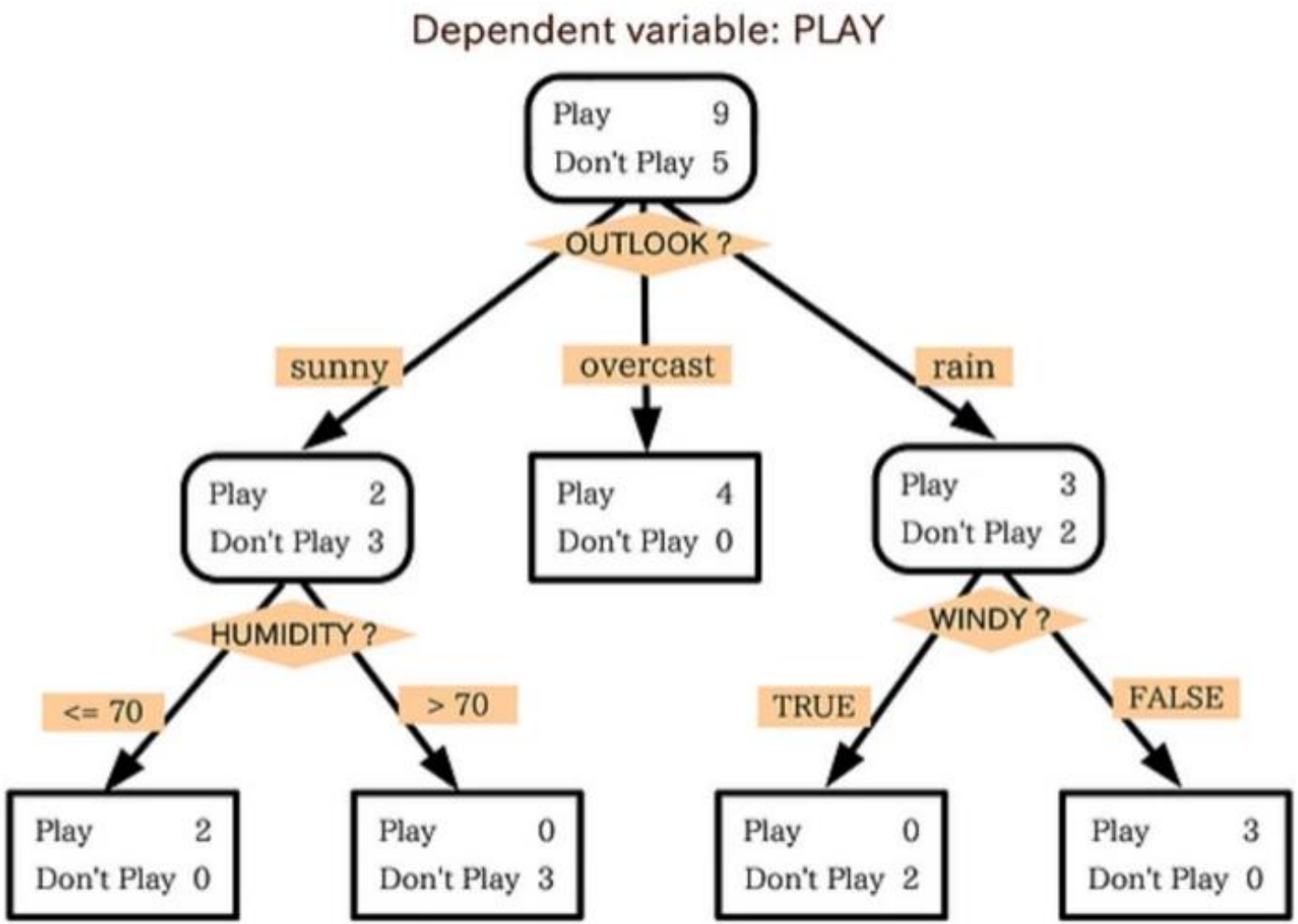
$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot \log(\hat{y_i}) + (1-y_i) \cdot \log(1-\hat{y_i})$$





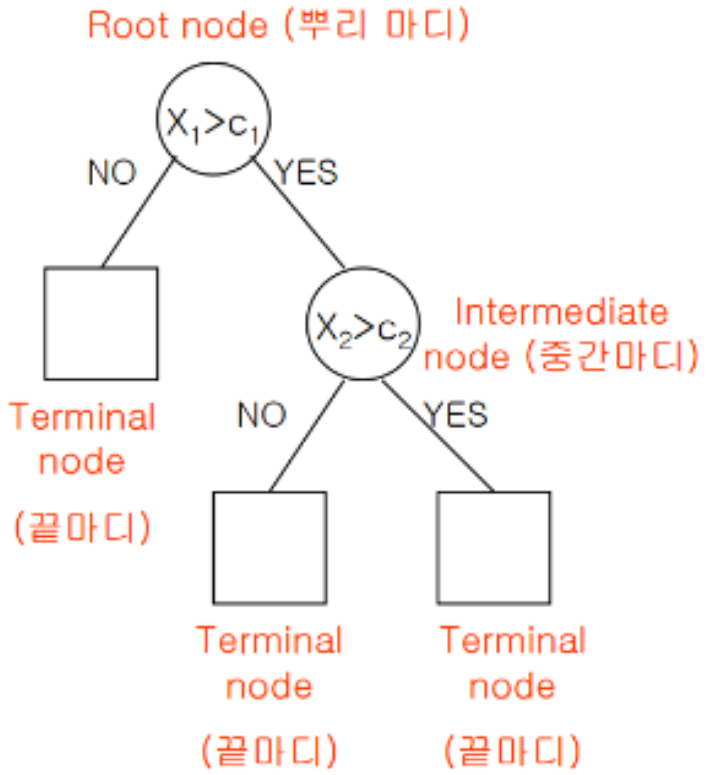
Decision Tree

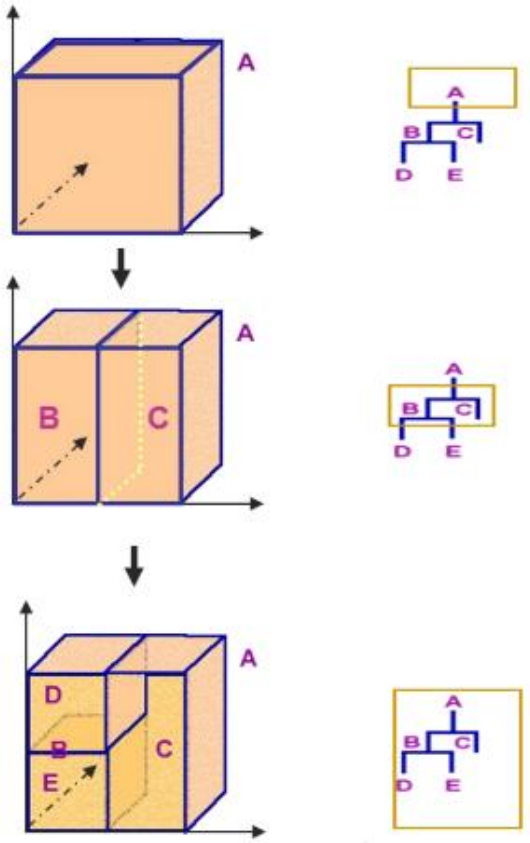
- Decision Tree → 의사결정나무
  - 데이터를 분석하여 이들 사이에 존재하는 패턴을 예측 가능한 규칙들의(Rules) 조합으로 나타냄
  - 모양이 '나무'와 같다고 해서 의사결정나무라고 불림
  - 질문을 던져서 대상을 좁혀나가는 '스무고개 ' 놀이와 비슷한 개념



Decision Tree

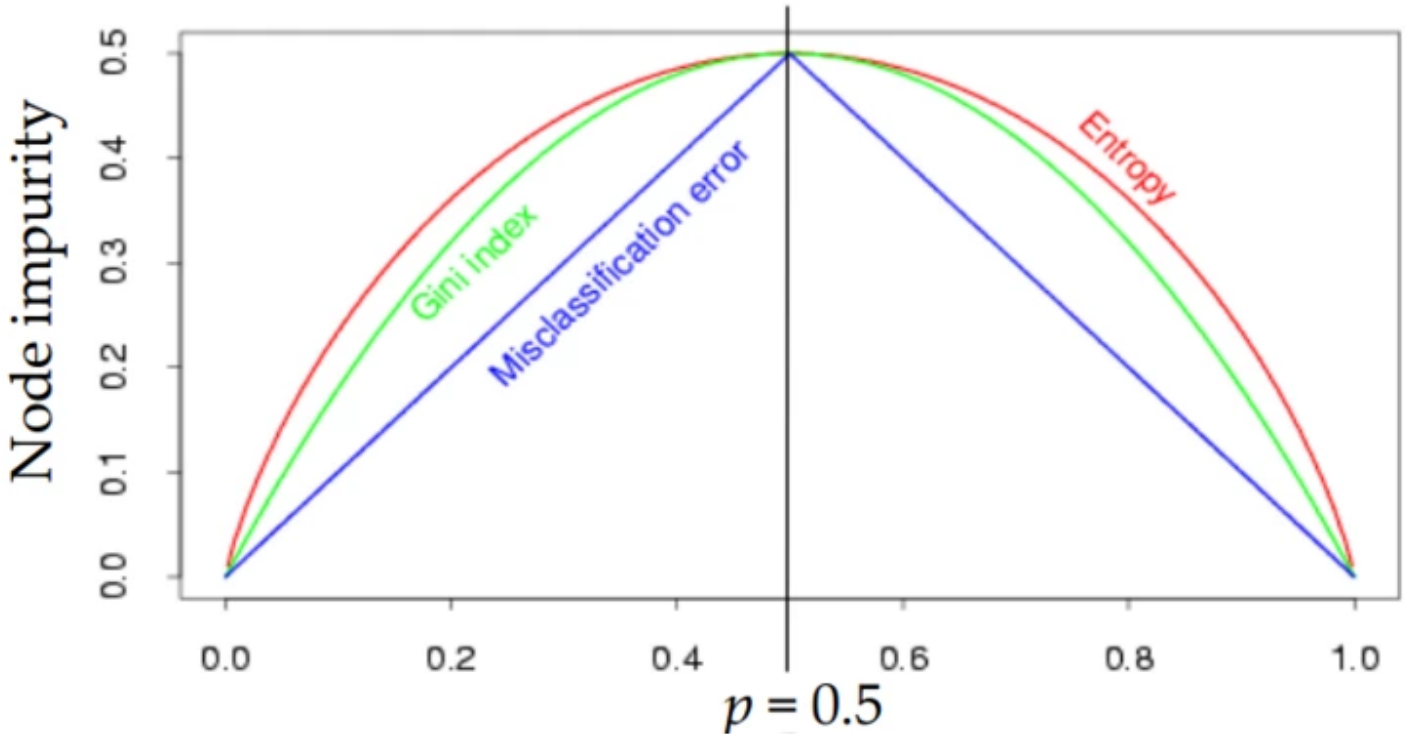
- Decision Tree → 의사결정나무
  - Decision Tree는 분류(Classification)과 회귀(Regression)이 모두 가능함
  - Linear Regression과 다르게 Model의 Complexity를 극한으로 높일 수 있음
    - Deep를 늘림
    - Leaf Node를 늘림
  - 만약 Terminal node 수가 3개뿐이라면 새로운 데이터가 100개, 1000개가 주어진다고 해도 의사결정나무는 딱 3종류의 답(Rule)만을 출력하게 됨





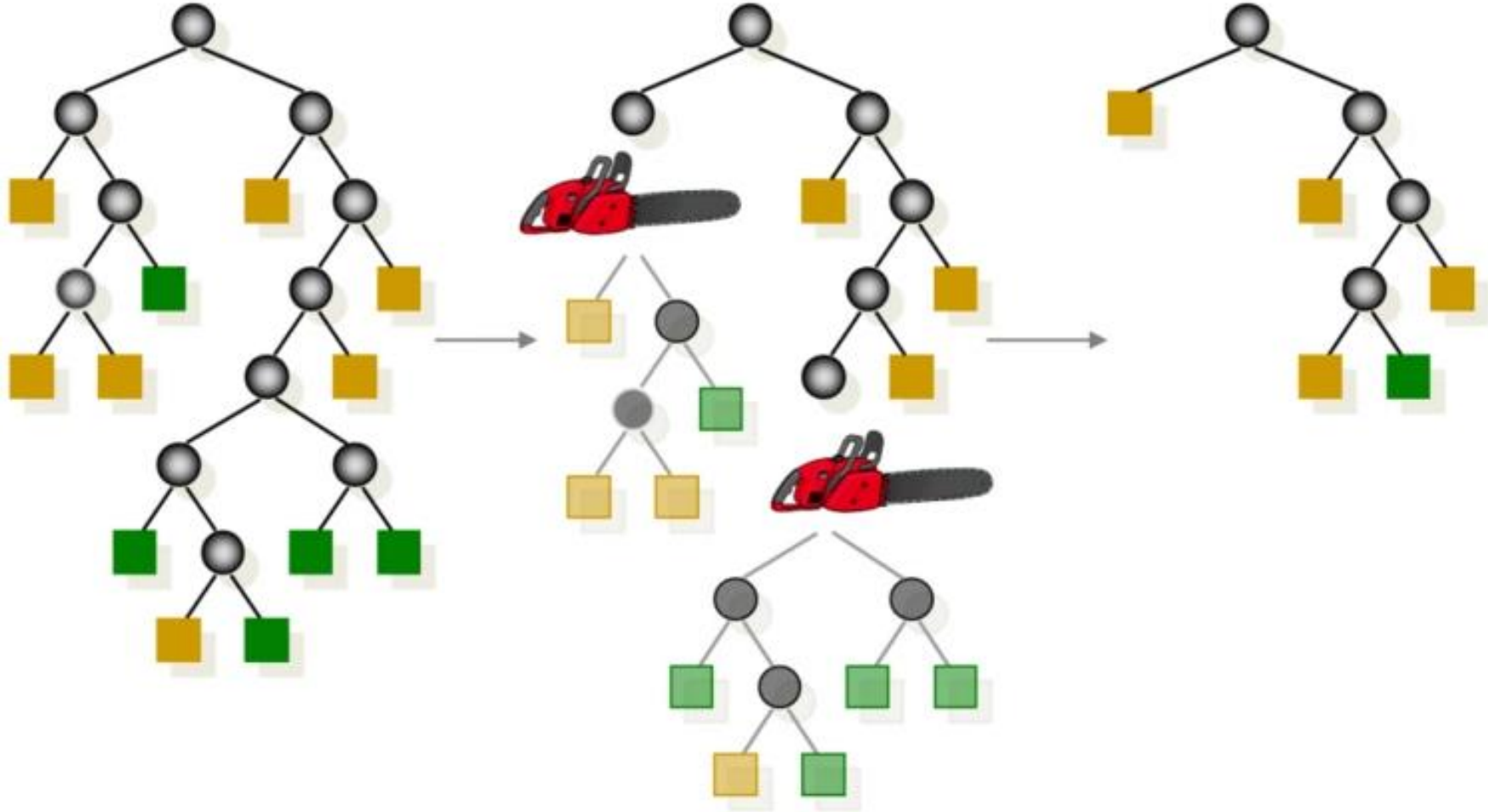
Decision Tree

- Decision Tree → 의사결정나무
  - 학습 방법
    - 순도(Homogeneity)를 최대로 증가시키는 방향
    - 불순도(Impurity) 혹은 불확실성(Uncertainty)을 최소로 감소시키는 방향
    - 순도 증가/불확실성 감소를 정보이론에서는 정보획득(Information gain)이라고 함
  - 순도 계산 방식
    - 엔트로피(Entropy)
    - 지니 계수(Gini Index)
    - 오분류 오차(Misclassification Error)
      - 미분이 불가능하기 때문에 잘 사용하지 않음



Decision Tree

- Decision Tree → 의사결정나무
  - 가지치기 (Pruning)
    - 모든 Terminal node의 순도가 100%인 상태를 Full tree라고 함
    - Full tree를 생성한 뒤 적절한 수준에서 terminal node를 잘라 줘야함
    - 분기가 너무 많으면 학습데이터에 대한 Overfitting될 염려가 있기 때문



Decision Tree

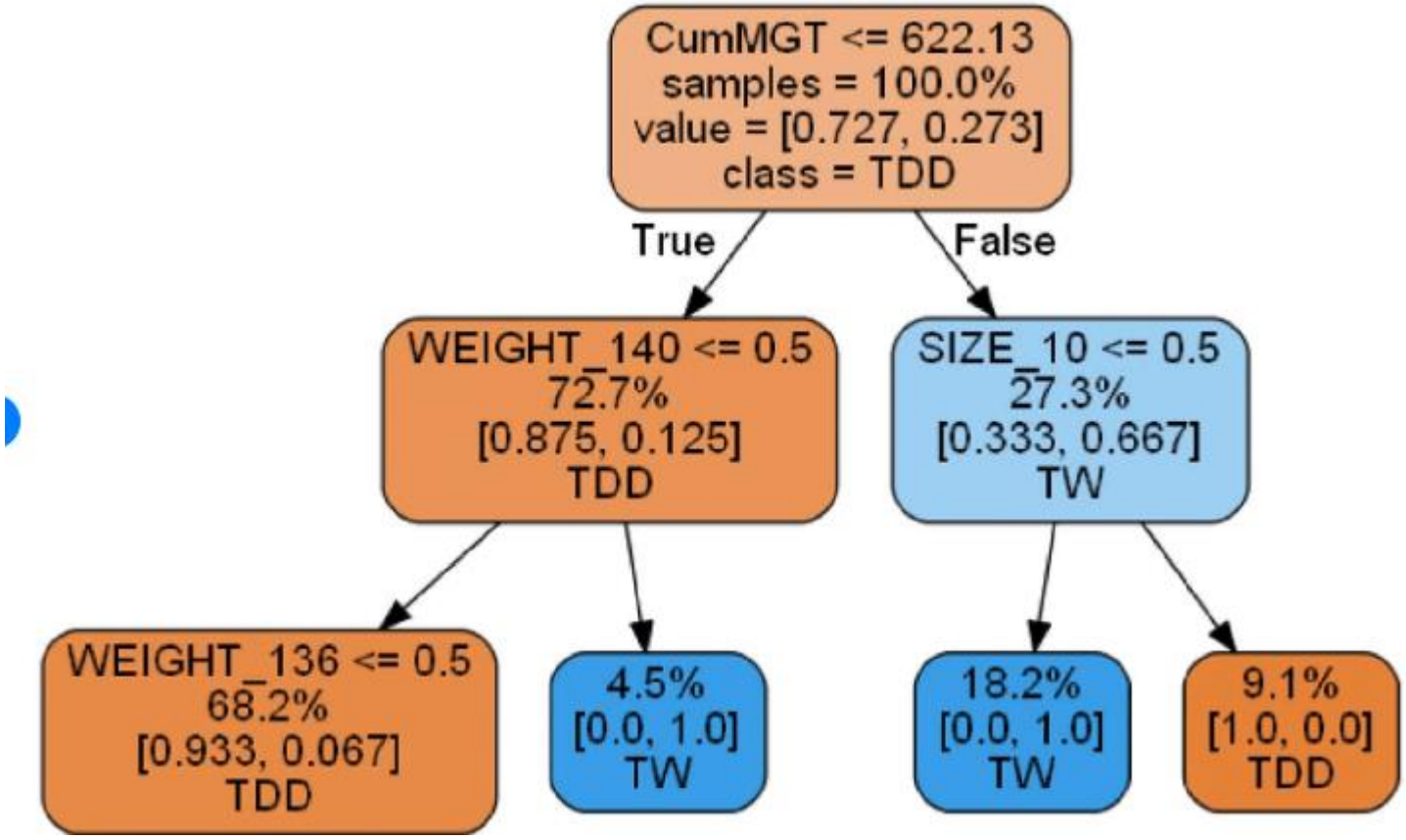
- Decision Tree → 의사결정나무
  - Confusion Matrix

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

ROC Curve, AUC ...

Decision Tree

- Decision Tree → 의사결정나무
  - Rule Extraction
    - 가장 중요하다고 생각함
    - Simple 하지만 직관력이 있음
      - Simple is the Best



Q & A