Decision Tree

: 의사 결정 규칙을 나무구조로 나타내어 전체 데이터를 소집단으로 분류하거나 예측하는 방법

1. ID3

- 불순도 지표 : Entropy
- Entropy 란?
  - 무질서도를 정량화 해서 표현한 값 = 데이터의 불확실성
  - 어떤 집합의 Entropy 가 높을 수록 그 집단의 특징을 찾는 것은 어려움 (Entropy 가 낮을 수록 좋음)

$$Entropy(A) = -\sum_{k=1}^{m} p_k \log_2(p_k)$$ ※ k는 클래스를 의미

- ID3 알고리즘 이란?
  - Entropy 지수를 이용한 알고리즘
  - Entropy 지수를 통해 Information Gain 도출 ※ Information gain : 상위 노드의 Entropy - 하위 노드 Entropy
  - Information Gain이 크게 나오는 변수 A를 기준으로 선택
  - Information Gain이 클수록 엔트로피를 많이 줄였다는 의미

$$Gain(S, A) = E(S) - \underbrace{I(S, A)}_{\sum \frac{|S_i|}{|S|} \cdot E(S_i)}$$

2. CART

- 지니 지수
  - 데이터의 통계적 분산 정도를 정량화해서 표현한 값
  - 지니 지수가 높을수록 그 집단의 데이터가 분산되어 있다. ⇒ 지니 지수가 낮은 방향으로 분류하기

$$Gini(A) = \sum_{j=1}^{2} \frac{|D_j|}{|D|} \times Gini(D_i)$$

$$Gini(D_i) = 1 - \sum_{j=1}^{x} P_j^2$$

- CART 알고리즘이란?
  - Gini index를 이용한 알고리즘
  - Binary Split 을 전제로 분석
  - 데이터의 대상 속성을 얼마나 잘못 분류할지 계산

3. 가지치기

- Pre pruning : 트리의 최대 depth 나 분기점의 최소 개수를 미리 지정
- Post pruning : 트리를 만든 후 데이터 포인트가 적은 노드를 삭제/병합