Trees

2022.01.11

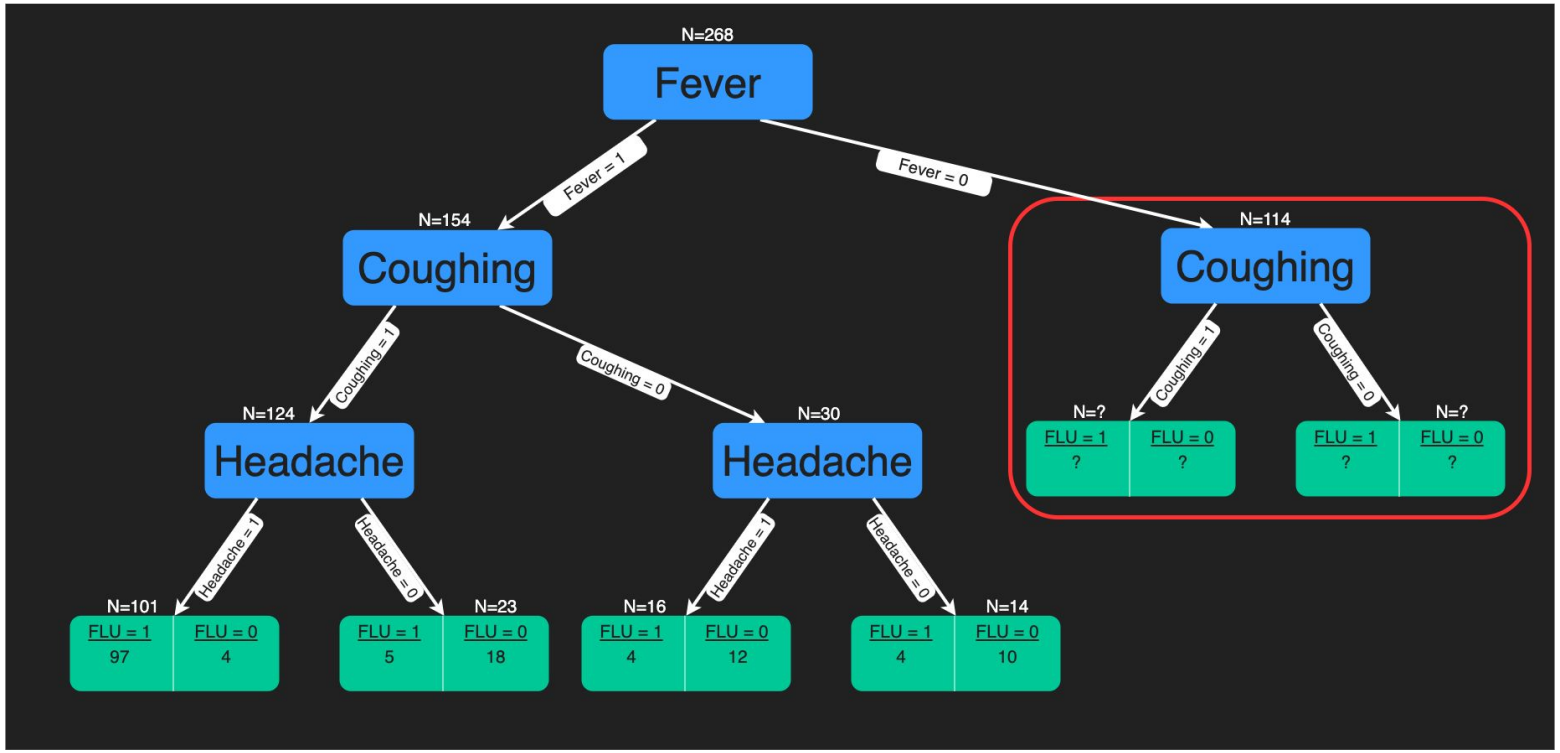
Dongmin Kim (tommy.dm.kim@kaist.ac.kr)

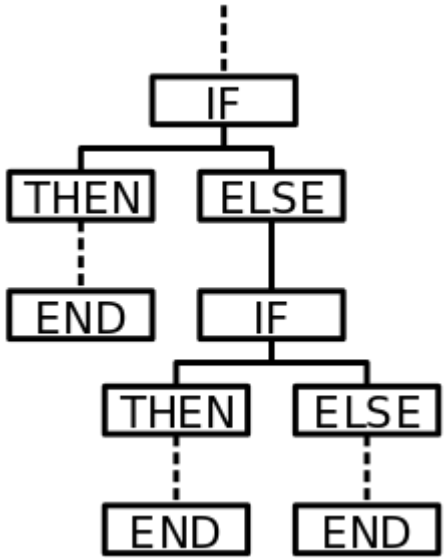
TODO

- Decision Tree Model의 동작 과정
- 좋은 Tree에 대하여
- Kaggle Titanic에 Decision Tree 적용하기

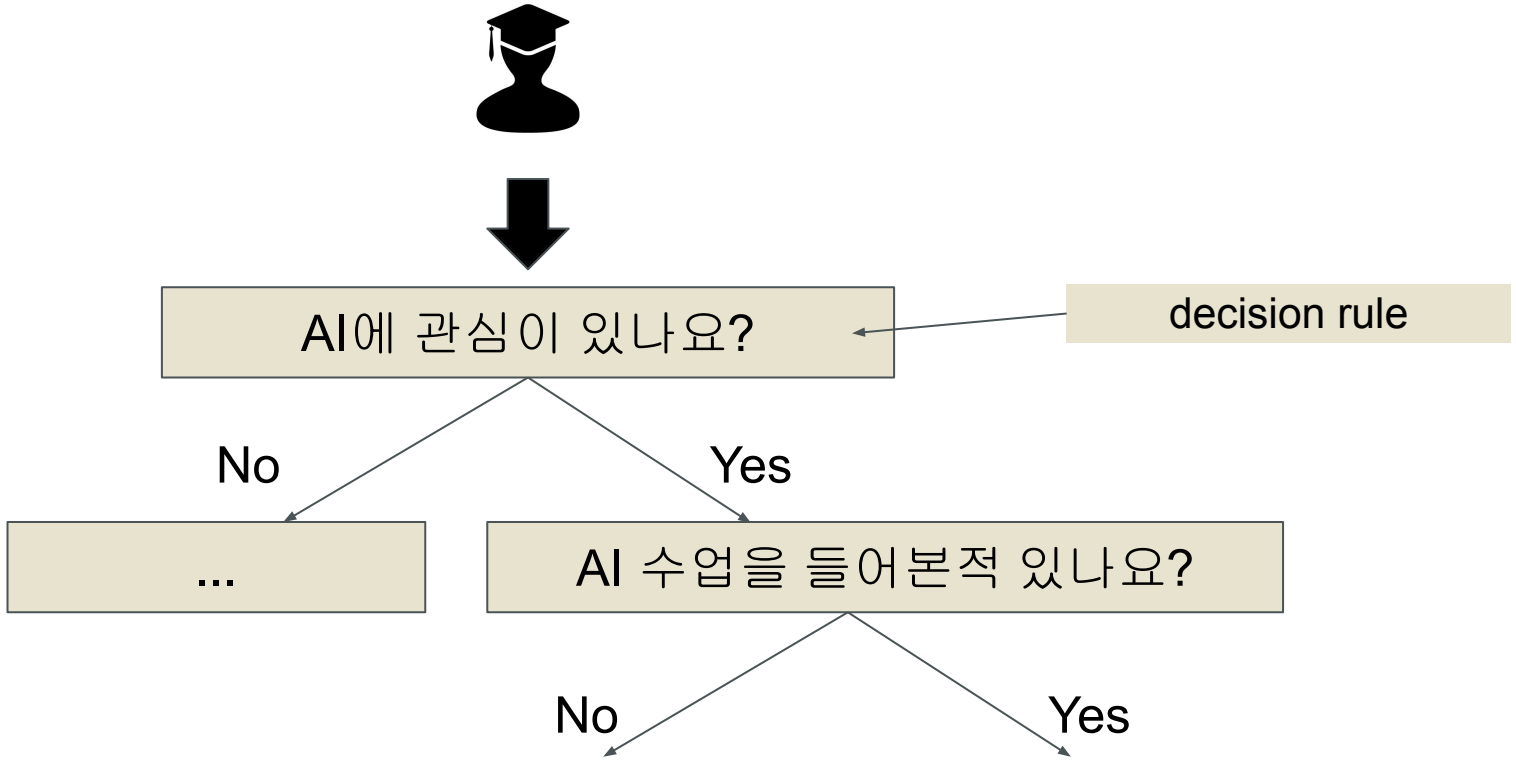
1. Tree의 동작 과정

Decision Tree Model의 동작 과정



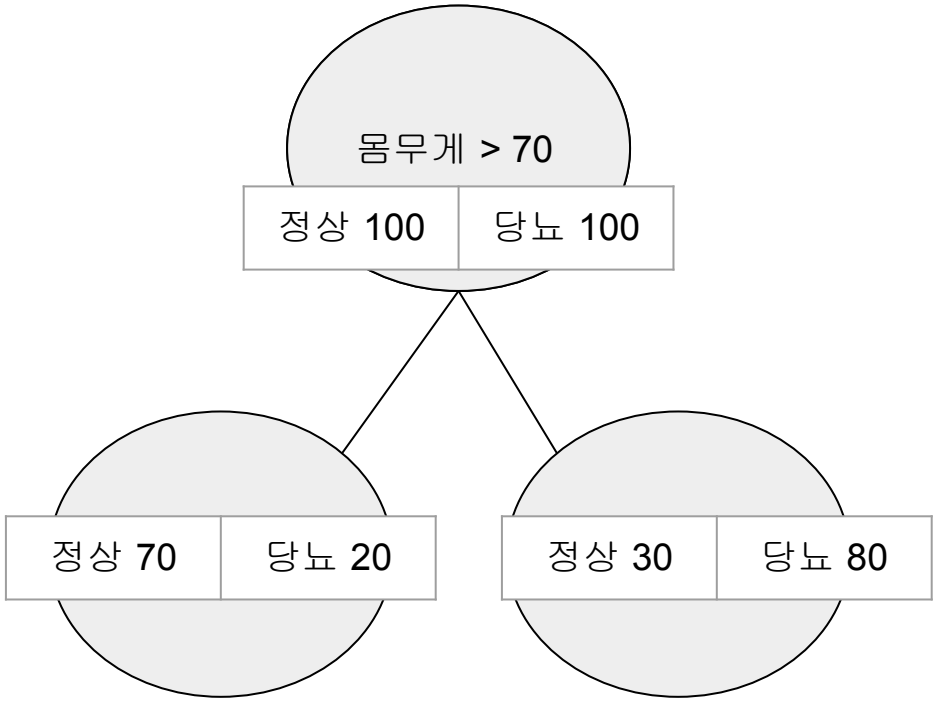


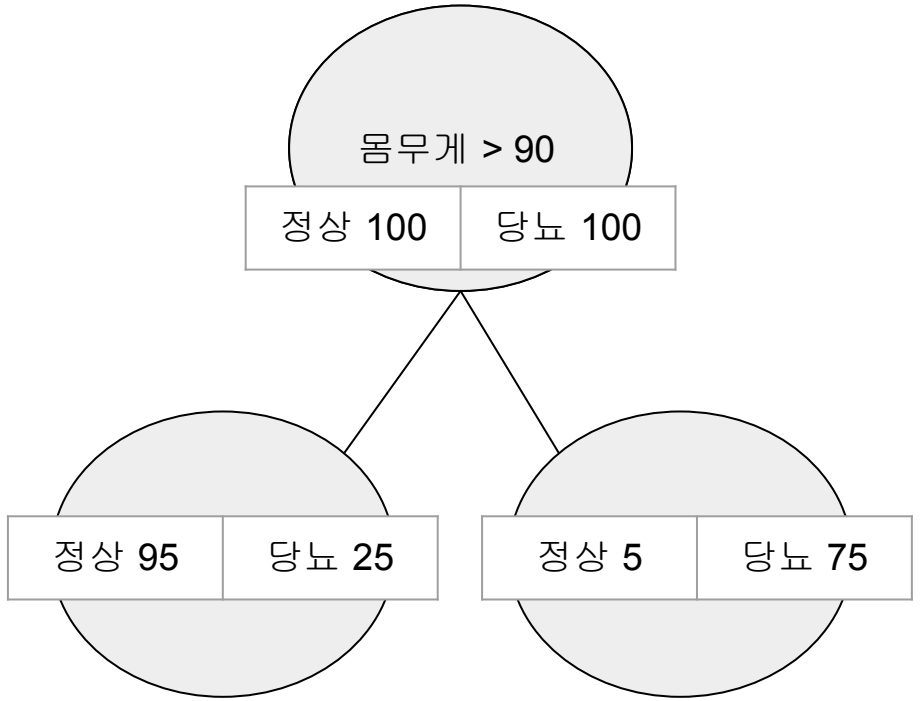
Decision Tree Model의 동작 과정



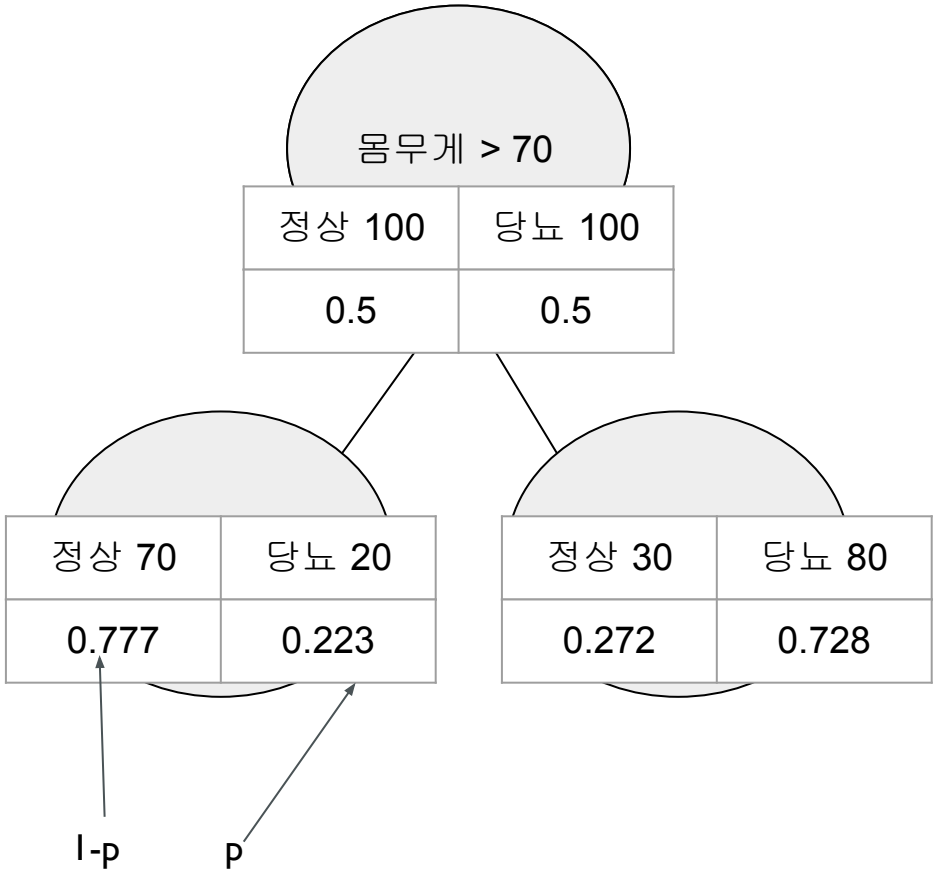
2. 좋은 Tree

조건 1. 그룹을 잘 나누어 주어야 한다



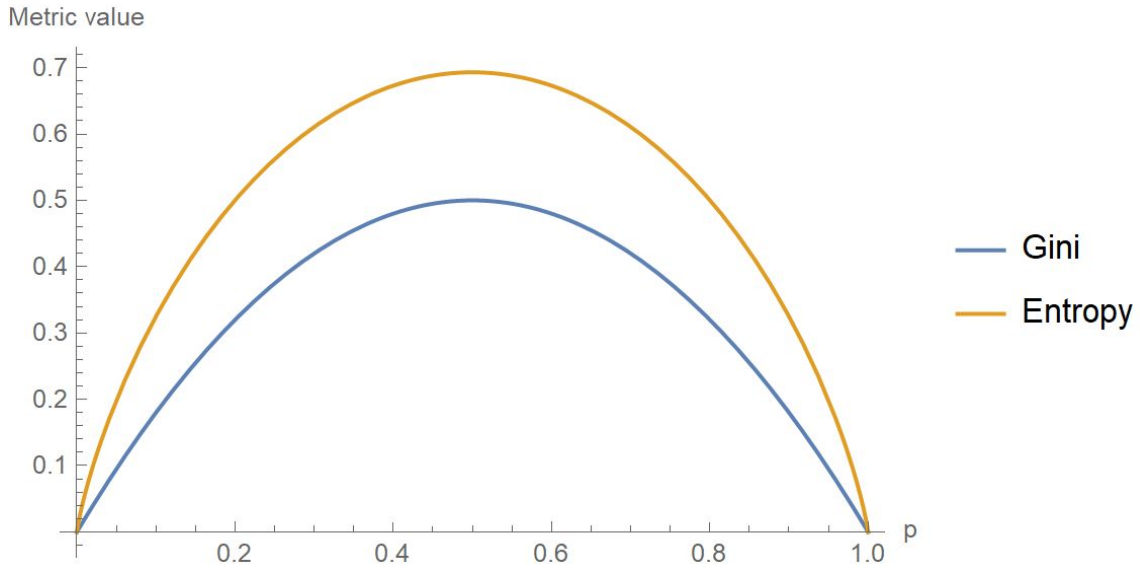


조건 1. 그룹을 잘 나누어 주어야 한다



1. p, 1-p는 각각 당뇨일 확률, 정상일 확률을 의미
2. p가 극단적인 값일수록 (0이나 1에 가까울수록)
   그룹은 잘 나누어졌다고 볼 수 있음
3. p가 얼마나 극단적인 값들을 갖는가? 에 대한 지표가 필요함

조건 1. 그룹을 잘 나누어 주어야 한다



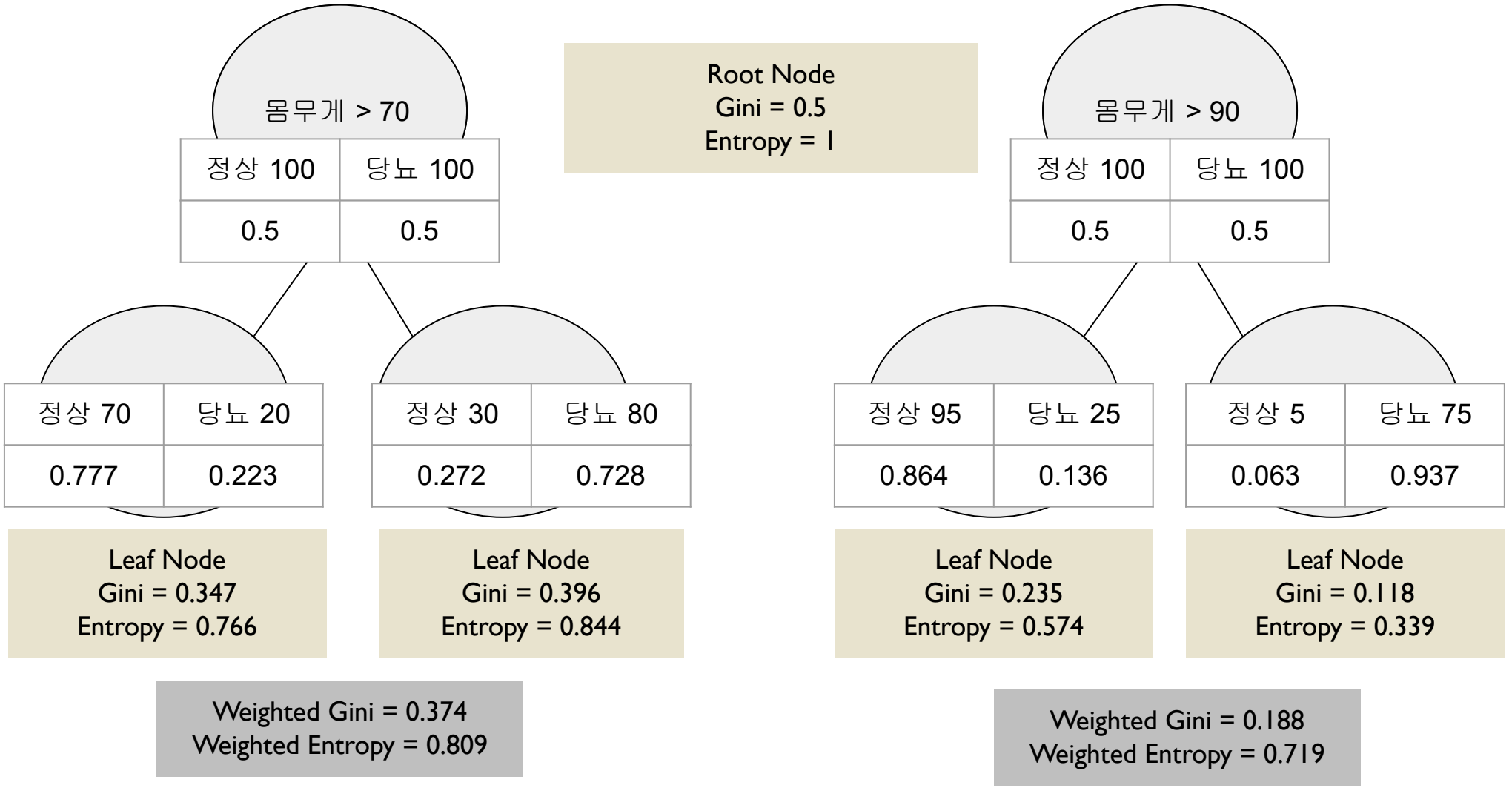
Information Entropy:

$$H = -\sum p(x) \log p(x)$$

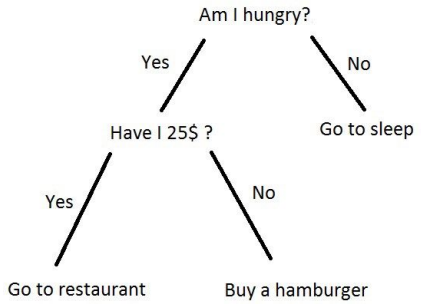
Gini Impurity:

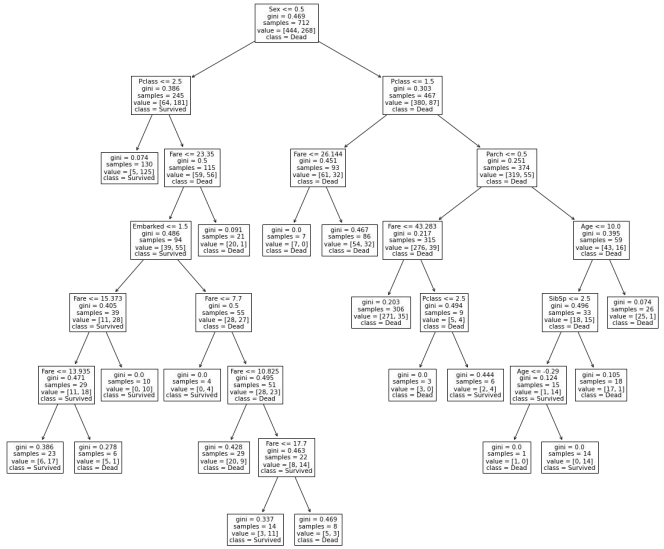
$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

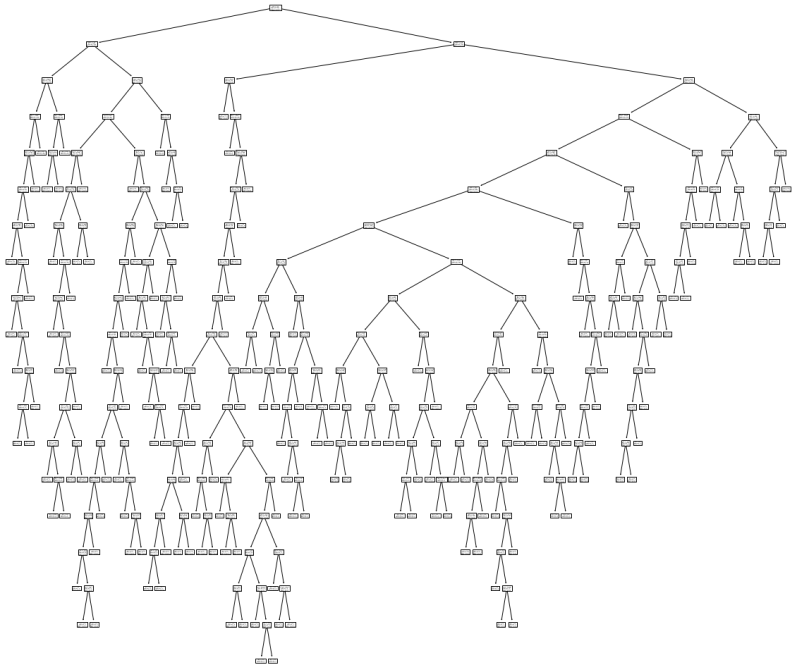
조건 1. 그룹을 잘 나누어 주어야 한다



조건 2. 가능하면 작은 트리가 좋다







Underfitting

Just right

Overfitting

조건 2. 가능하면 작은 트리가 좋다

- Hyperparameters
  - max_depth : tree의 깊이를 제한
  - max_leaf_nodes: leaf node의 개수를 제한
  - min_impurity_decrease: 그룹을 나누어 줄 때, impurity (ex. gini, entropy) 가 일정 이상 감소해야 가지치기를 함

- Feature Importance
  - 당뇨 예측에 있어, 몸무게와 일일 당 섭취량이라는 두가지 feature가 있다고 가정
  - 가장 간단하게는, 이 feature들이 decision tree에 몇번 등장했는지를 기준으로 feature importance를 계산해 볼 수 있음
  - 더 나아가 feature들이 decision tree에 등장할 때 impurity가 얼마나 개선되었는지 합하여 feature importance를 계산해 볼 수 있음 (sklearn decision tree default: gini)

3. Kaggle

Kaggle Titanic에 Decision Tree 적용하기

0. Kaggle 노트북 준비하고, train/val/test split 부분까지 실행시키기

4. 일단 써보기

   4.1. sklearn의 DecisionTreeClassifier를 사용해 모델 트레이닝 시키기

   4.2. sklearn을 활용하여 accuracy 출력, confusion matrix 그리기, classification report 생성해보기, feature importance 뽑아보기

   4.3. sklearn.tree 에 내장된 plot_tree를 이용하여, 생성한 decision tree 모델 plot하기

   4.4. 어떻게 개선시켜 볼 수 있을까요?

5. Hyperparameter Tuning

   5.1. sklearn의 gridsearchCV를 사용하여, 튜닝을 해봅시다. max_leaf_nodes 라는 파라미터 하나에 대해서 여러 시도를 해보세요.
   추가적으로, 이 파라미터를 조절하는 것이 어떤 의미가 있는지, Overfitting의 관점에서 고민해보세요.

   5.2. 튜닝 된 모델들에 대하여, 4.1, 4.2 에서 했던 모든 과정들을 반복해주세요.

[Submission 2]

- 우리의 두번째 모델에 대한 결과를 제출해봅시다.
- test.csv에서 불러온 데이터 셋들에 대해서도, train.csv에서 했던 전처리 과정들을 똑같이 적용해주어야 모델을 쓸 수 있겠죠?