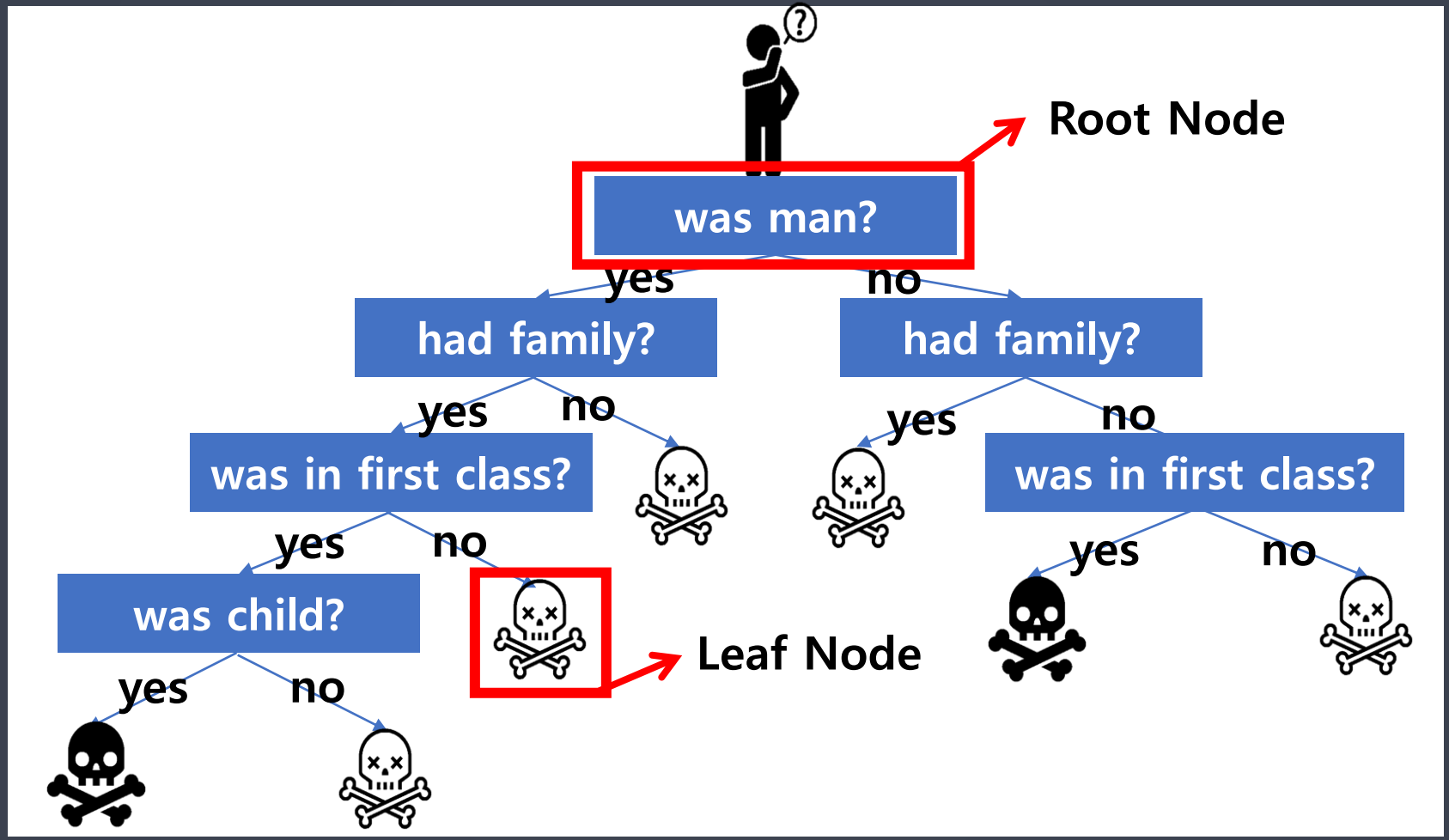
Machine Learning

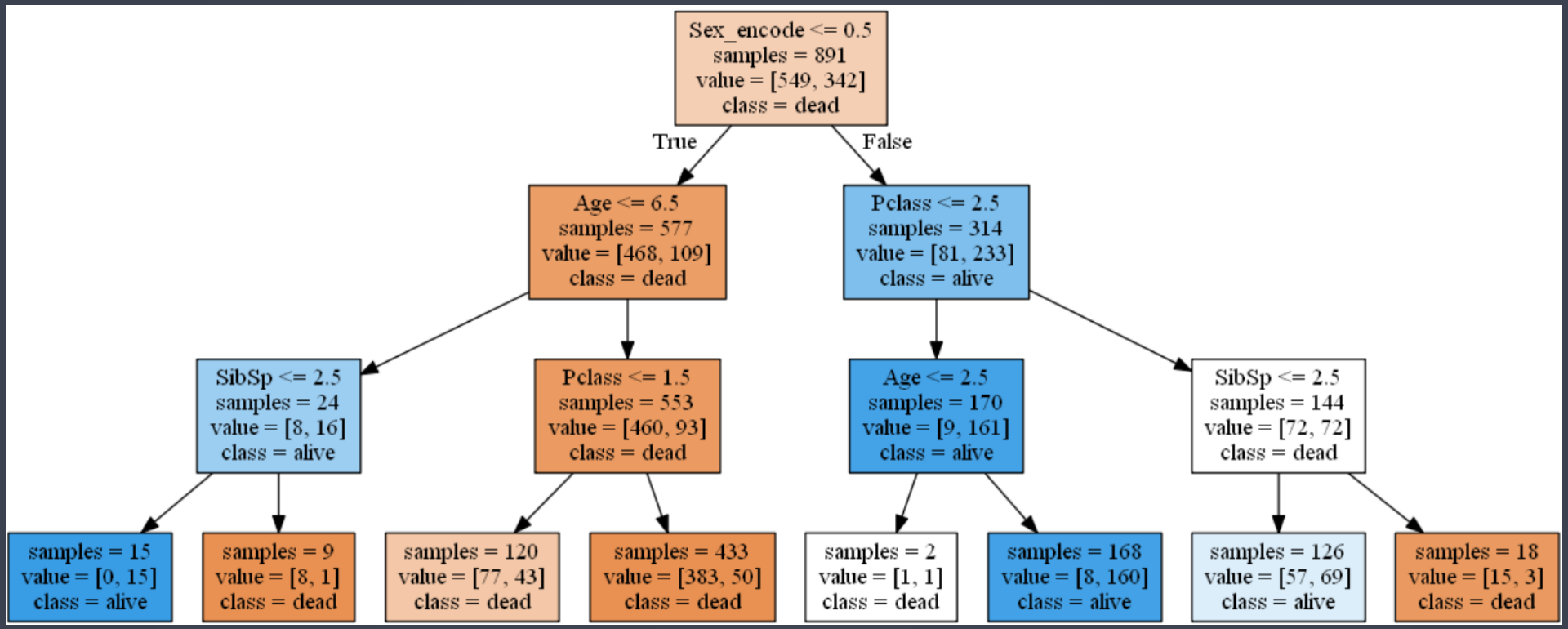
Decision Tree

Decision Tree

의사결정나무



Decision Tree



Decision Tree

의사결정나무

- 타깃 값이 한 개인 리프 노드를 순수 노드라고 한다.
- 모든 노드가 순수 노드가 될 때 까지 학습하면 복잡해지고 과대적합이 된다.
- 새로운 데이터 포인트가 들어오면 해당하는 노드를 찾아 분류는 가장 많은 클래스, 회귀는 평균 값을 갖는다.

Decision Tree

과대적합 제어

- 노드 생성을 미리 중단하는 사전 가지치기(pre-pruning)와 트리를 만든 후에 크기가 작은 노드를 삭제하는 사후 가지치기(pruning)가 있다.
  (sklearn은 사전 가지치기만 지원)
- 트리의 최대 깊이나 리프 노드의 최대 개수를 제어
- 노드가 분할 하기 위한 데이터 포인트의 최소 개수를 지정

Decision Tree

주요 매개변수(Hyperparameter)

- 트리의 최대 깊이 : max_depth
- 리프 노드의 최대 개수 : max_leaf_nodes
- 리프 노드가 되기 위한 최소 샘플의 개수 : min_samples_leaf

Decision Tree

장단점

- 만들어진 모델을 쉽게 시각화할 수 있어 이해하기 쉽다. (white box model)
- 각 특성이 개별 처리되기 때문에 데이터 스케일에 영향을 받지 않아 특성의 정규화나 표준화가 필요 없다.
- 훈련데이터 범위 밖의 포인트는 예측 할 수 없다. (ex : 시계열 데이터)
- 가지치기를 사용함에도 불구하고 과대적합되는 경향이 있어 일반화 성능이 좋지 않다.

머신러닝(Machine Learning) 실습 – ex02

Iris Dataset

머신러닝(Machine Learning) 실습 – ex02

붓꽃의 품종 분류

setosa, versicolor, virginica 종 분류

꽃잎petal, 꽃받침sepal의 폭과 길이

사전에 준비한 데이터를 이용하므로 지도 학습

3개의 붓꽃 품종에서 고르는 분류classification

클래스class: 가능한 출력값. 즉 세개의 붓꽃 품종

레이블label: 데이터 포인트 하나에 대한 출력