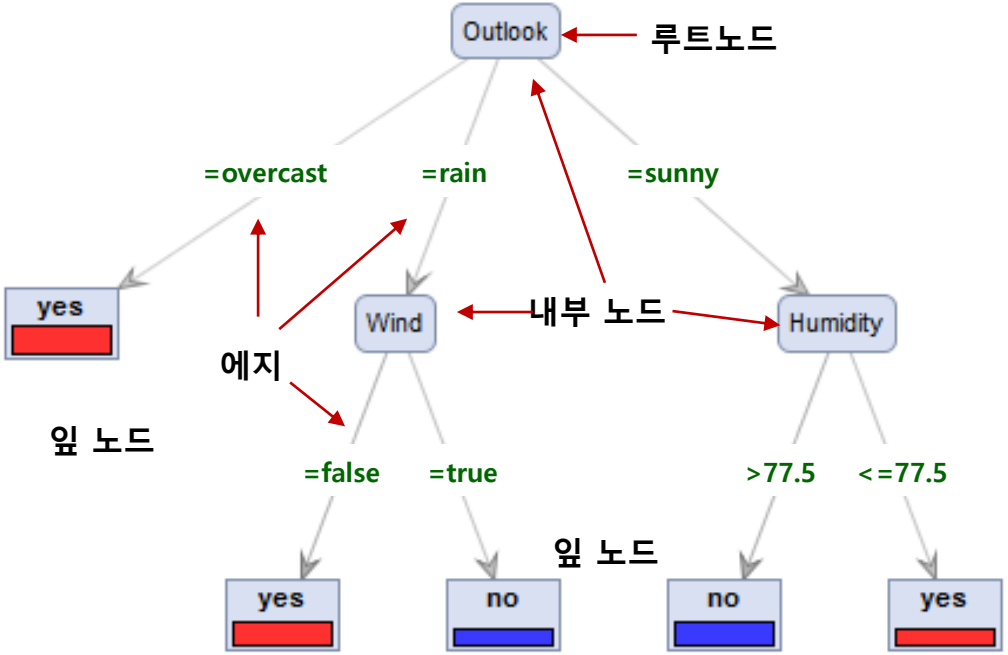
Data Science

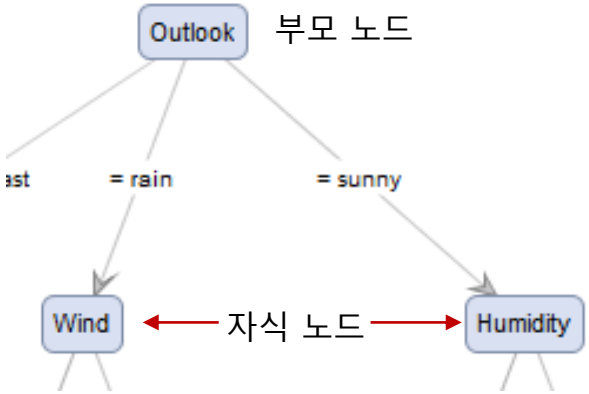
04.01-01 의사결정나무 소개, 적용, 생성 알고리즘 - Decision Tree

4.1 의사결정나무 (Decision Tree)

- 의사결정나무(Decision Tree : DT)
  - 트리(나무)구조: 나무를 거꾸로,
  - 노드(node)와 에지(edge)로 구성, 루트노드(root node)부터 시작
  - 에지(edge): node 와 node 를 연결, 부모노드 → 자식노드
  - 내부노드(internal node), 잎 노드(leaf node)



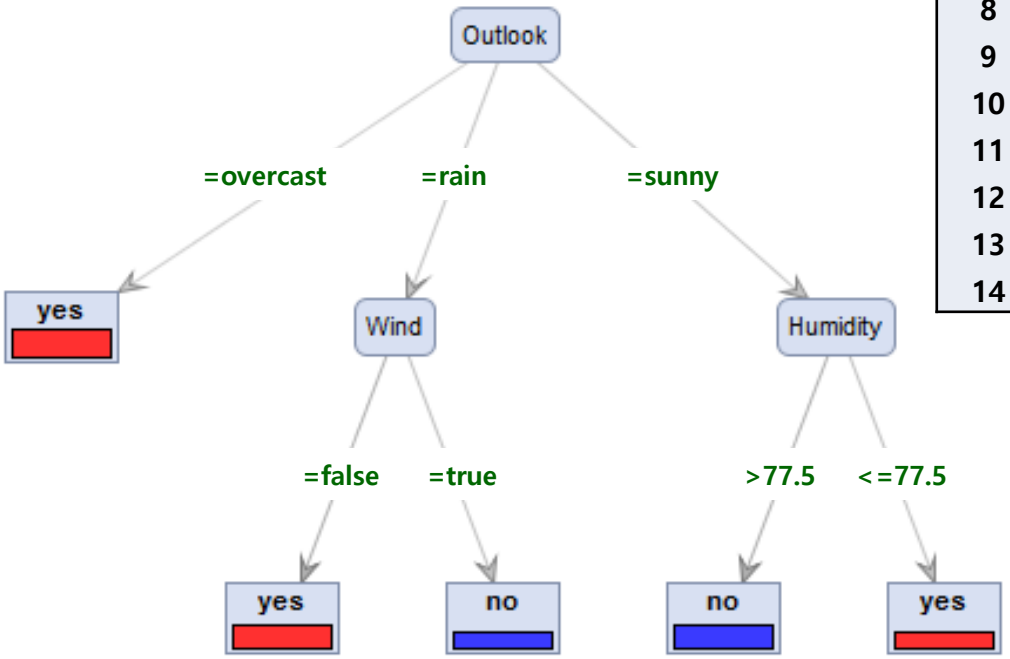
Tree model induced with Golf dataset



부모, 자식 노드

4.1 의사결정나무 (Decision Tree)

- 의사결정나무(Decision Tree : DT)
  - 내부노드(internal node): 속성 테스트
  - 에지(edge): 속성 조건
  - 잎 노드(leaf node): 클래스(라벨)



Tree model induced with Golf dataset

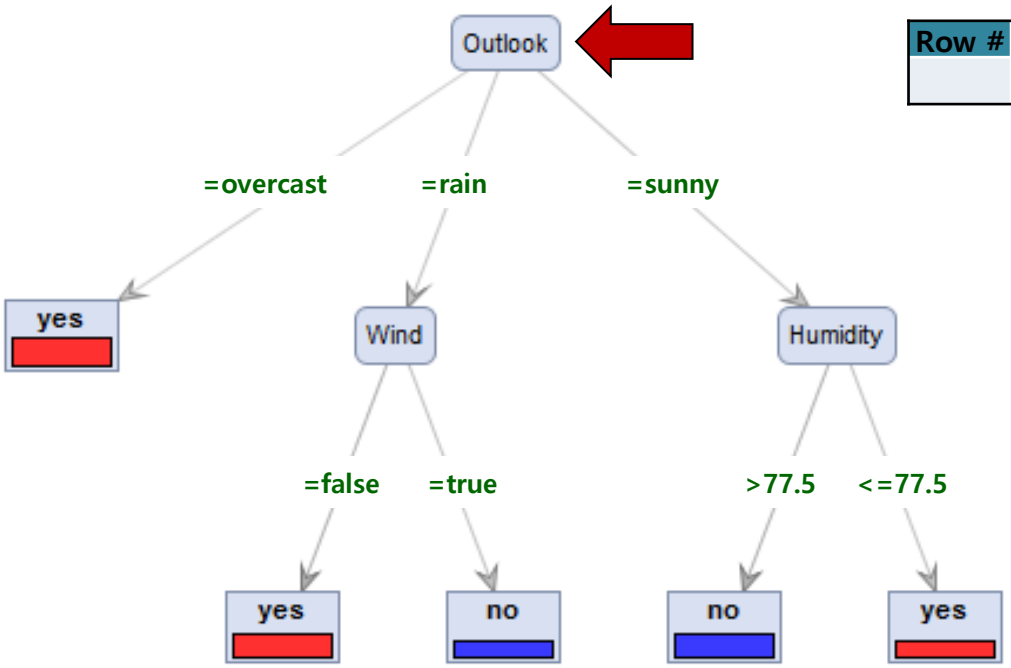
Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	no
2	sunny	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	false	yes
6	rain	65.0	70.0	true	no
7	overcast	64.0	65.0	true	yes
8	sunny	72.0	95.0	false	no
9	sunny	69.0	70.0	false	yes
10	rain	75.0	80.0	false	yes
11	sunny	75.0	70.0	true	yes
12	overcast	72.0	90.0	true	yes
13	overcast	81.0	75.0	false	yes
14	rain	71.0	80.0	true	no

Golf Dataset

목표변수

4.1 의사결정나무 (Decision Tree)

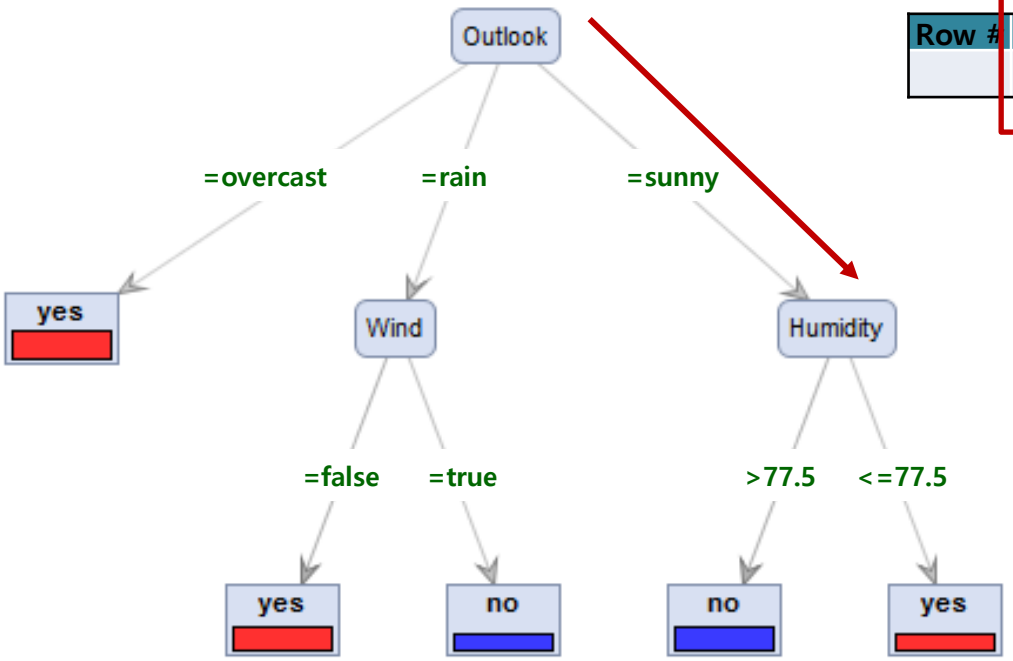
- 의사결정나무(Decision Tree : DT) 모델의 적용
  - 루트부터 시작
  - 각 노드에서 속성값을 테스트하고 테스트 결과에 해당되는 에지를 따라 자식노드 이동
  - 잎(leaf) 노드 도착하면 라벨 클래스로 예측



Row #	Outlook	Temperature	Humidity	Wind	Play
	sunny	85.0	85.0	false	?

4.1 의사결정나무 (Decision Tree)

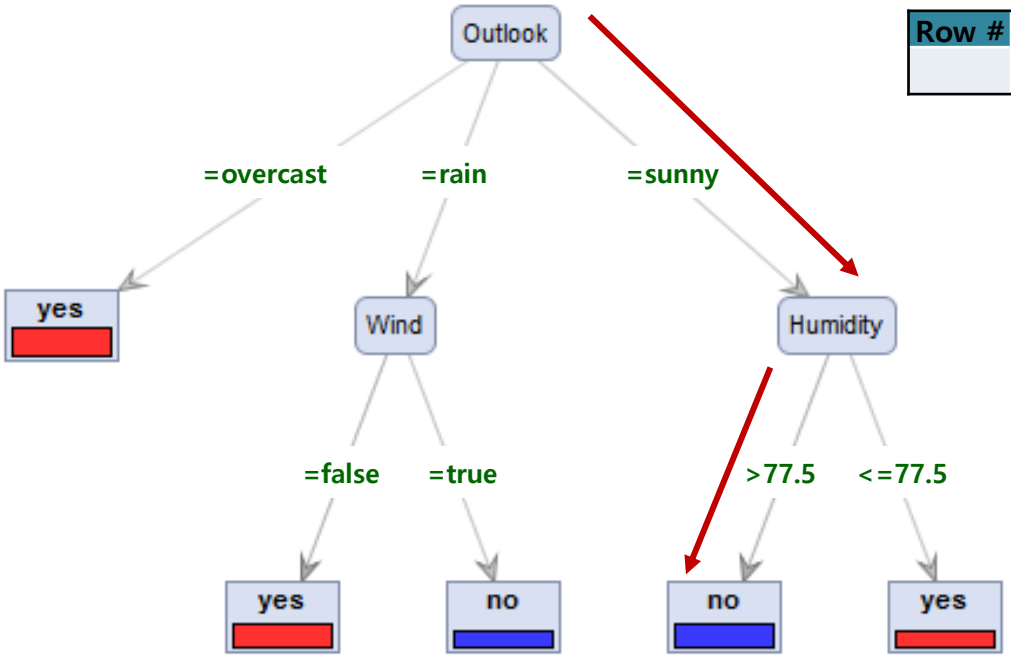
- 의사결정나무(Decision Tree : DT) 모델의 적용
  - 루트부터 시작
  - 각 노드에서 속성값을 테스트하고 테스트 결과에 해당되는 에지를 따라 자식노드 이동
  - 잎(leaf) 노드 도착하면 라벨 클래스로 예측



Row #	Outlook	Temperature	Humidity	Wind	Play
	sunny	85.0	85.0	false	?

4.1 의사결정나무 (Decision Tree)

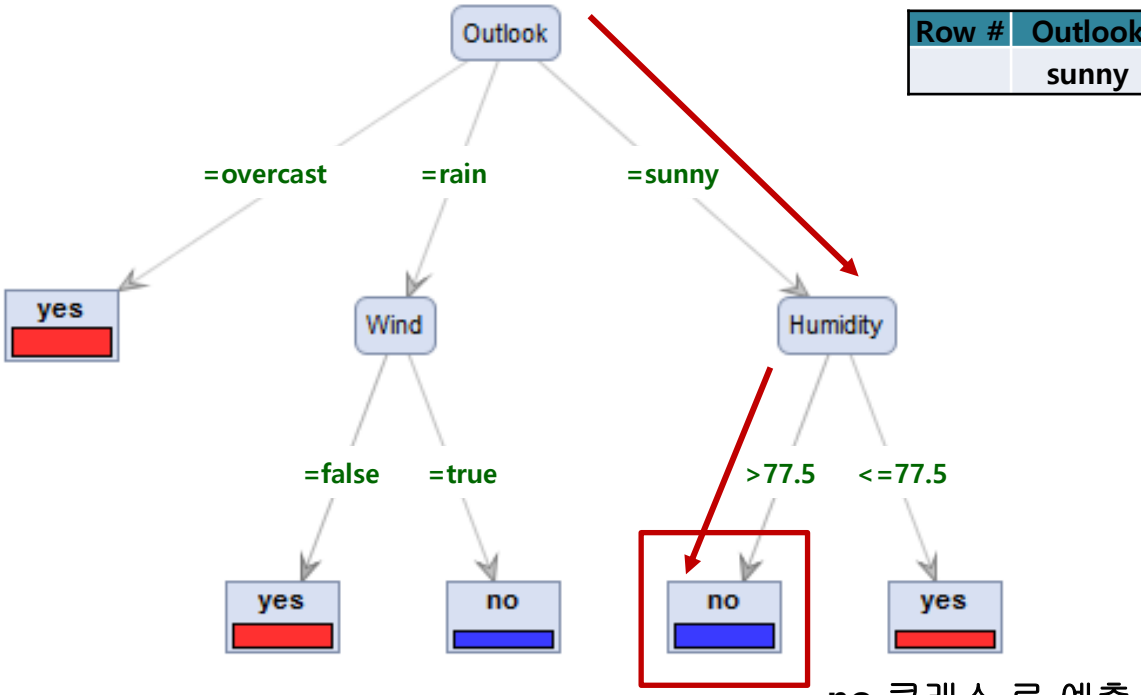
- 의사결정나무(Decision Tree : DT) 모델의 적용
  - 루트부터 시작
  - 각 노드에서 속성값을 테스트하고 테스트 결과에 해당되는 에지를 따라 자식노드 이동
  - 잎(leaf) 노드 도착하면 라벨 클래스로 예측



Row #	Outlook	Temperature	Humidity	Wind	Play
	sunny	85.0	85.0	false	?

4.1 의사결정나무 (Decision Tree)

- 의사결정나무(Decision Tree : DT) 모델의 적용
  - 루트부터 시작
  - 각 노드에서 속성값을 테스트하고 테스트 결과에 해당되는 에지를 따라 자식노드 이동
  - 잎(leaf) 노드 도착하면 라벨 클래스로 예측



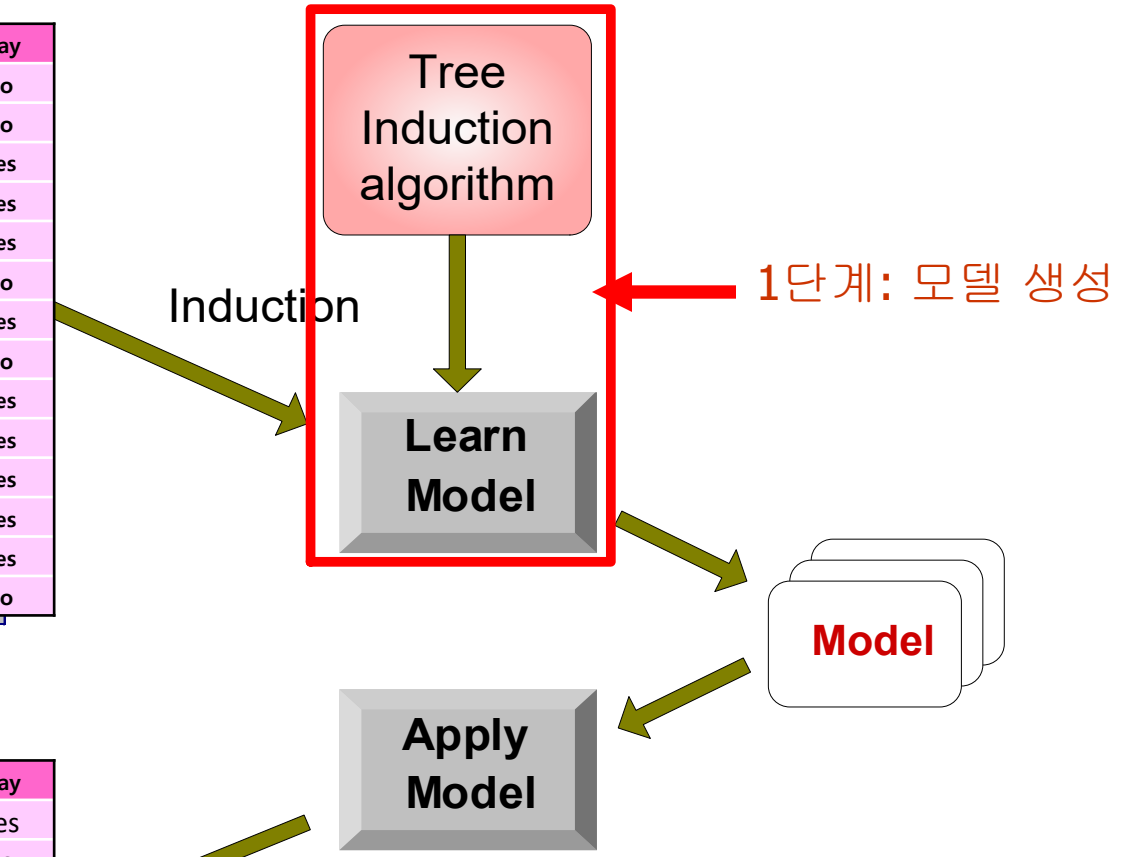
Row #	Outlook	Temperature	Humidity	Wind	Play
	sunny	85.0	85.0	false	?

no 클래스 로 예측

Decision Tree Classification Task

Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	no
2	sunny	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	false	yes
6	rain	65.0	70.0	true	no
7	overcast	64.0	65.0	true	yes
8	sunny	72.0	95.0	false	no
9	sunny	69.0	70.0	false	yes
10	rain	75.0	80.0	false	yes
11	sunny	75.0	70.0	true	yes
12	overcast	72.0	90.0	true	yes
13	overcast	81.0	75.0	false	yes
14	rain	71.0	80.0	true	no

Training Set



Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	yes
2	overcast	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	true	yes
6	rain	65.0	70.0	true	no
15	No	Large	67K	?	

Test Set

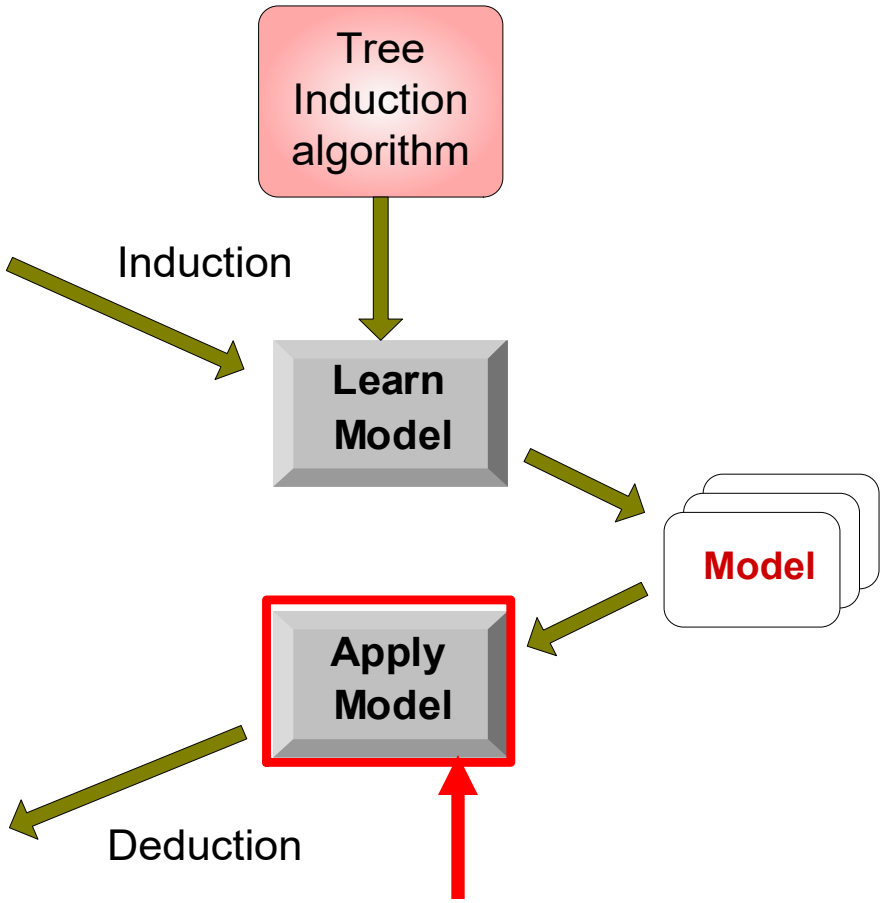
Decision Tree Classification Task

Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	no
2	sunny	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	false	yes
6	rain	65.0	70.0	true	no
7	overcast	64.0	65.0	true	yes
8	sunny	72.0	95.0	false	no
9	sunny	69.0	70.0	false	yes
10	rain	75.0	80.0	false	yes
11	sunny	75.0	70.0	true	yes
12	overcast	72.0	90.0	true	yes
13	overcast	81.0	75.0	false	yes
14	rain	71.0	80.0	true	no

Training Set

Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	yes
2	overcast	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	true	yes
6	rain	65.0	70.0	true	no

Test Set



2단계: 새로운 데이터에 모델을 적용하여 예측

Decision Tree Induction

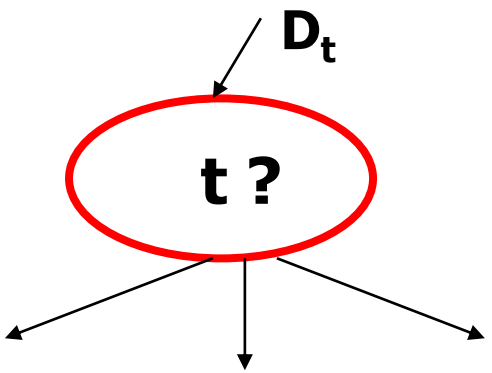
- Many Algorithms:
  - Hunt's Algorithm (기본)
  - CART
  - ID3, C4.5
  - SLIQ, SPRINT

General Structure of Hunt's Algorithm

- D_t: node t 에 주어진 학습용 데이터셋
- Procedure:
  - 경우 1. D_t 가 class y_t 에 속하는 사례들로만 구성됨. 순수한(pure) 경우
    → t 는 는 잎노드(leaf node)이며 class label 은 y_t
  - 경우 2. D_t 가 empty set 인 경우 → t 는 잎노드 이고 default class, y_d 로 label.
  - 경우 3. D_t 가 서로 다른 class 의 사례 들로 혼합 되어 있음, 순수하지 않음(not pure) → 속성을 하나를 선택한 후 속성 테스트를 사용하여 D_t 을 split (분할) 하여 split 된 데이터셋들을 하나씩 자식 노드들에게 넘긴다.
  - 재귀적으로 자식 노드에 대하여 procedure 를 적용한다.

Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	no
2	sunny	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	false	yes
6	rain	65.0	70.0	true	no
7	overcast	64.0	65.0	true	yes
8	sunny	72.0	95.0	false	no
9	sunny	69.0	70.0	false	yes
10	rain	75.0	80.0	false	yes
11	sunny	75.0	70.0	true	yes
12	overcast	72.0	90.0	true	yes
13	overcast	81.0	75.0	false	yes
14	rain	71.0	80.0	true	no

Golf Dataset

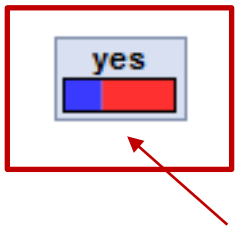


Hunt's Algorithm

* Class label : Play

1. Begin: All the records are assigned to the root

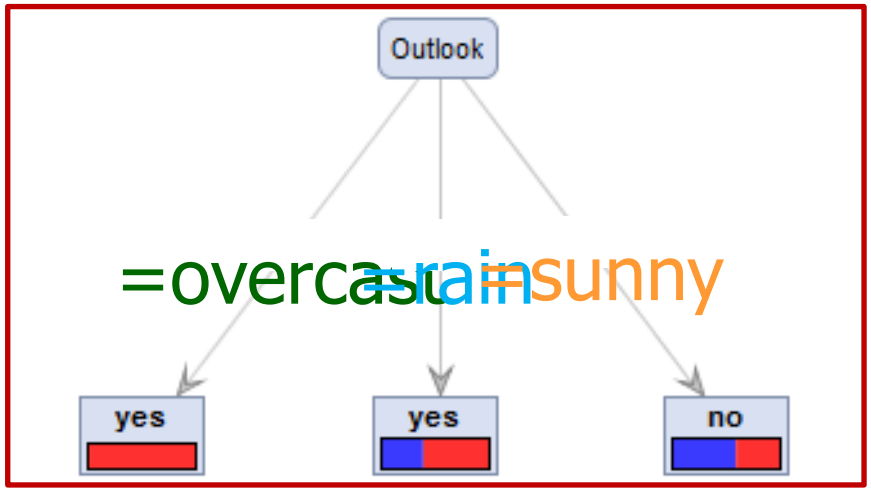
Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	no
2	sunny	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	false	yes
6	rain	65.0	70.0	true	no
7	overcast	64.0	65.0	true	yes
8	sunny	72.0	95.0	false	no
9	sunny	69.0	70.0	false	yes
10	rain	75.0	80.0	false	yes
11	sunny	75.0	70.0	true	yes
12	overcast	72.0	90.0	true	yes
13	overcast	81.0	75.0	false	yes
14	rain	71.0	80.0	true	no



Not pure

Hunt's Algorithm

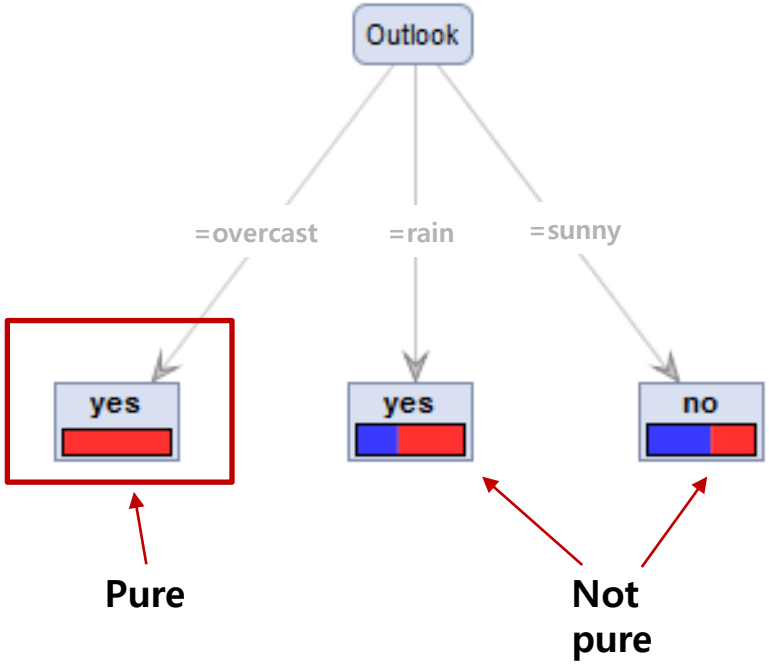
1. Begin: All the records are assigned to the root
2. Split on Outlook



Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	no
2	sunny	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	false	yes
6	rain	65.0	70.0	true	no
7	overcast	64.0	65.0	true	yes
8	sunny	72.0	95.0	false	no
9	sunny	69.0	70.0	false	yes
10	rain	75.0	80.0	false	yes
11	sunny	75.0	70.0	true	yes
12	overcast	72.0	90.0	true	yes
13	overcast	81.0	75.0	false	yes
14	rain	71.0	80.0	true	no

Hunt's Algorithm

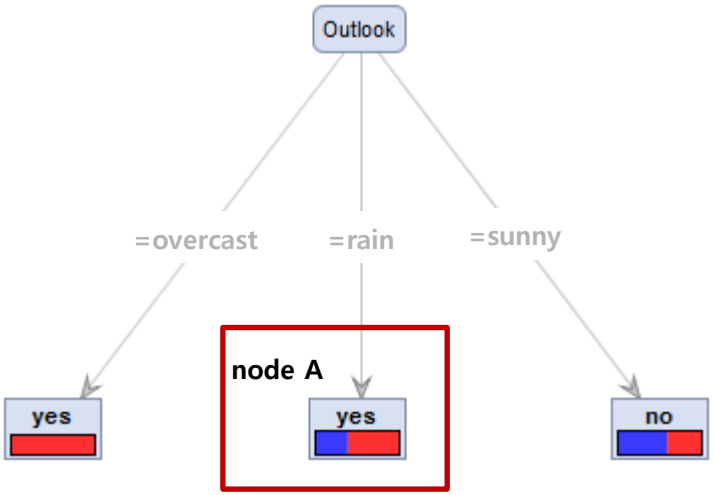
1. Begin: All the records are assigned to the root
2. Split on Outlook



Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	no
2	sunny	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	false	yes
6	rain	65.0	70.0	true	no
7	overcast	64.0	65.0	true	yes
8	sunny	72.0	95.0	false	no
9	sunny	69.0	70.0	false	yes
10	rain	75.0	80.0	false	yes
11	sunny	75.0	70.0	true	yes
12	overcast	72.0	90.0	true	yes
13	overcast	81.0	75.0	false	yes
14	rain	71.0	80.0	true	no

Hunt's Algorithm

1. Begin: All the records are assigned to the root
2. Split on Outlook
3. Data set assigned to node A

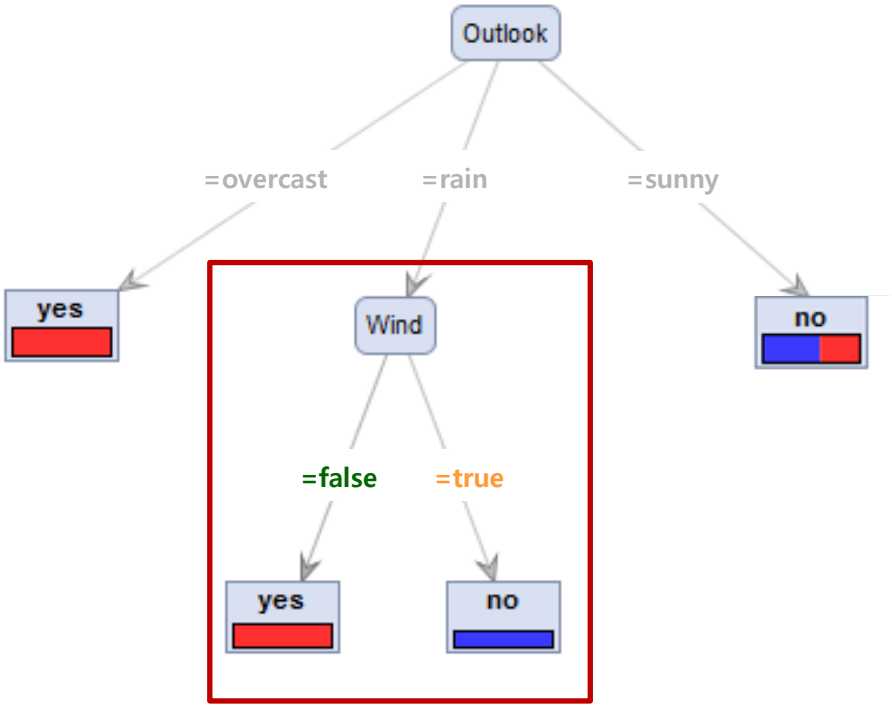


Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	no
2	sunny	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	false	yes
6	rain	65.0	70.0	true	no
7	overcast	64.0	65.0	true	yes
8	sunny	72.0	95.0	false	no
9	sunny	69.0	70.0	false	yes
10	rain	75.0	80.0	false	yes
11	sunny	75.0	70.0	true	yes
12	overcast	72.0	90.0	true	yes
13	overcast	81.0	75.0	false	yes
14	rain	71.0	80.0	true	no

Hunt's Algorithm

1. Begin: All the records are assigned to the root
2. Split on Outlook
3. Data set assigned to node A
4. Split on Wind

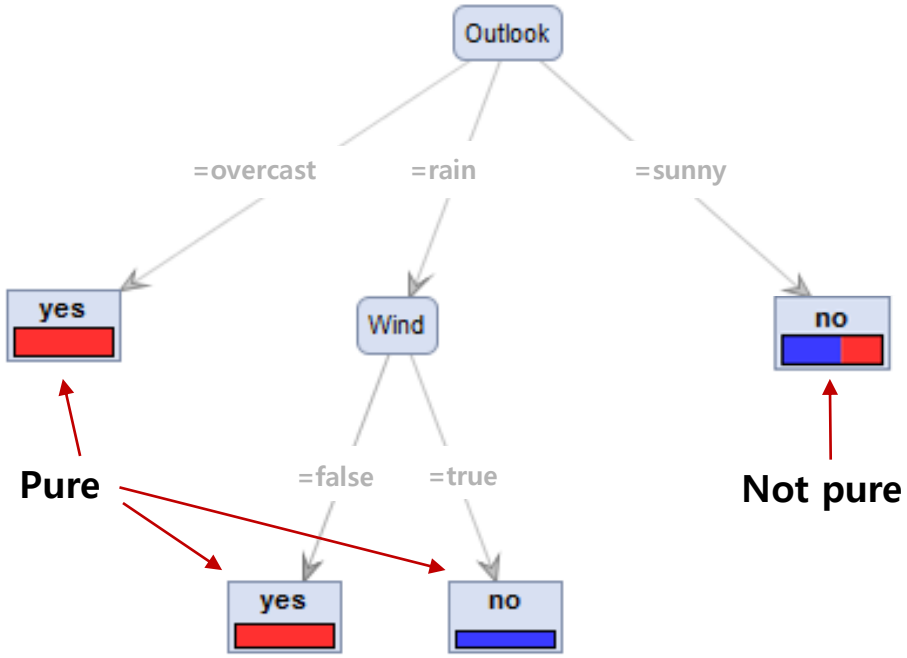
Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	no
2	sunny	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	false	yes
6	rain	65.0	70.0	true	no
7	overcast	64.0	65.0	true	yes
8	sunny	72.0	95.0	false	no
9	sunny	69.0	70.0	false	yes
10	rain	75.0	80.0	false	yes
11	sunny	75.0	70.0	true	yes
12	overcast	72.0	90.0	true	yes
13	overcast	81.0	75.0	false	yes
14	rain	71.0	80.0	true	no



Hunt's Algorithm

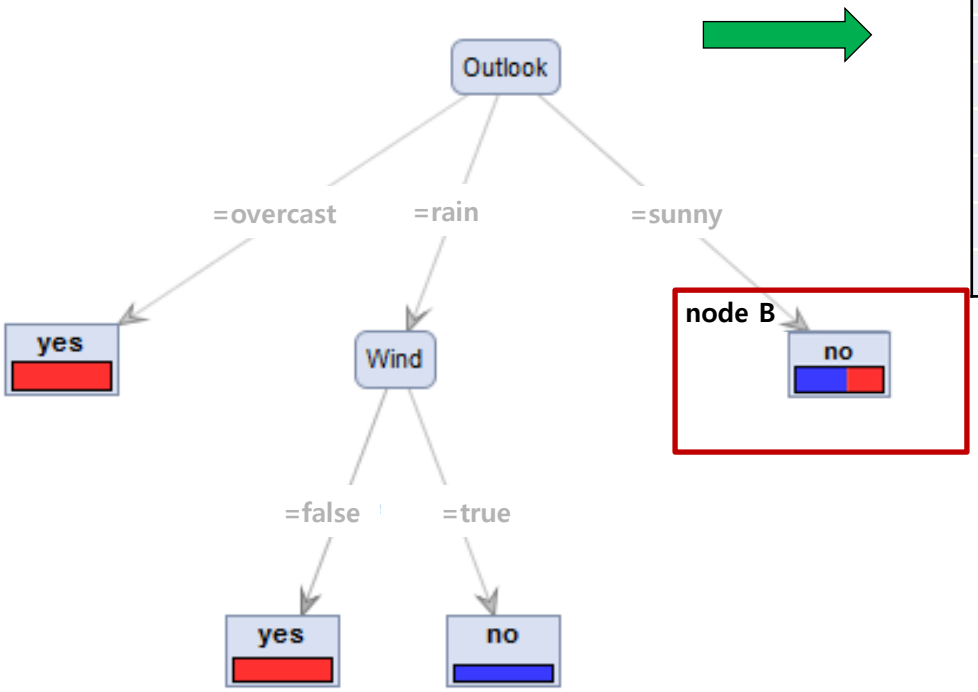
1. Begin: All the records are assigned to the root
2. Split on Outlook
3. Data set assigned to node A
4. Split on Wind

Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	no
2	sunny	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	false	yes
6	rain	65.0	70.0	true	no
7	overcast	64.0	65.0	true	yes
8	sunny	72.0	95.0	false	no
9	sunny	69.0	70.0	false	yes
10	rain	75.0	80.0	false	yes
11	sunny	75.0	70.0	true	yes
12	overcast	72.0	90.0	true	yes
13	overcast	81.0	75.0	false	yes
14	rain	71.0	80.0	true	no



Hunt's Algorithm

1. Begin: All the records are assigned to the root
2. Split on Outlook
3. Data set assigned to node A
4. Split on Wind
5. Data set assigned to node B

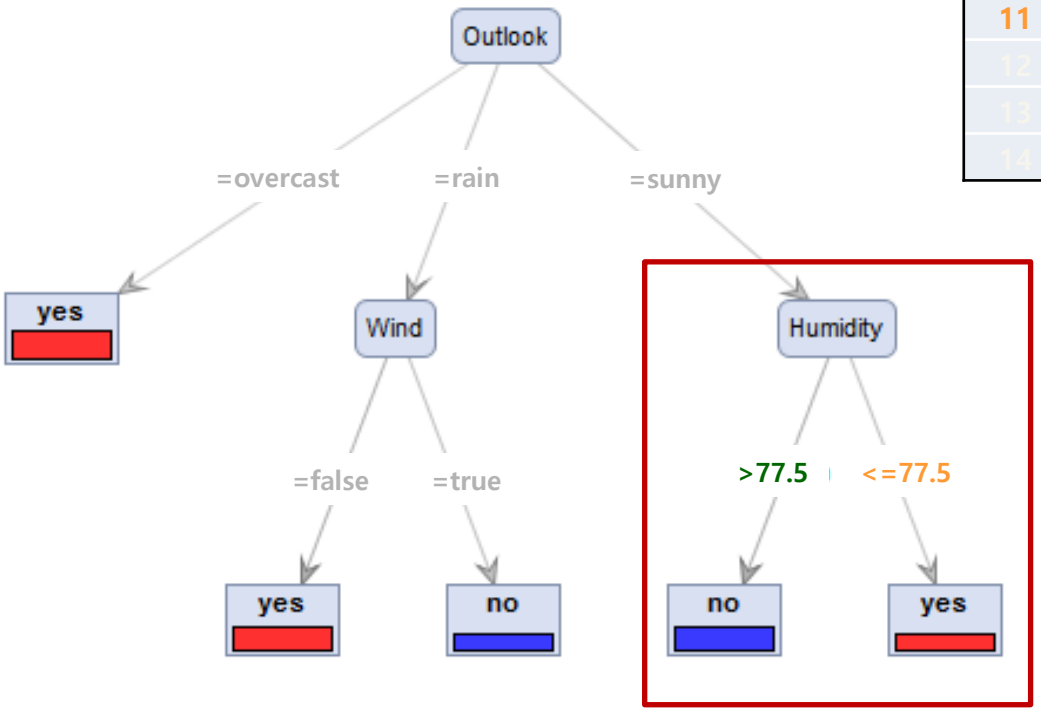


Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	no
2	sunny	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	false	yes
6	rain	65.0	70.0	true	no
7	overcast	64.0	65.0	true	yes
8	sunny	72.0	95.0	false	no
9	sunny	69.0	70.0	false	yes
10	rain	75.0	80.0	false	yes
11	sunny	75.0	70.0	true	yes
12	overcast	72.0	90.0	true	yes
13	overcast	81.0	75.0	false	yes
14	rain	71.0	80.0	true	no

Hunt's Algorithm

1. Begin: All the records are assigned to the root
2. Split on Outlook
3. Data set assigned to node A
4. Split on Wind
5. Data set assigned to node B
6. Split on Humidity

Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	no
2	sunny	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	false	yes
6	rain	65.0	70.0	true	no
7	overcast	64.0	65.0	true	yes
8	sunny	72.0	95.0	false	no
9	sunny	69.0	70.0	false	yes
10	rain	75.0	80.0	false	yes
11	sunny	75.0	70.0	true	yes
12	overcast	72.0	90.0	true	yes
13	overcast	81.0	75.0	false	yes
14	rain	71.0	80.0	true	no

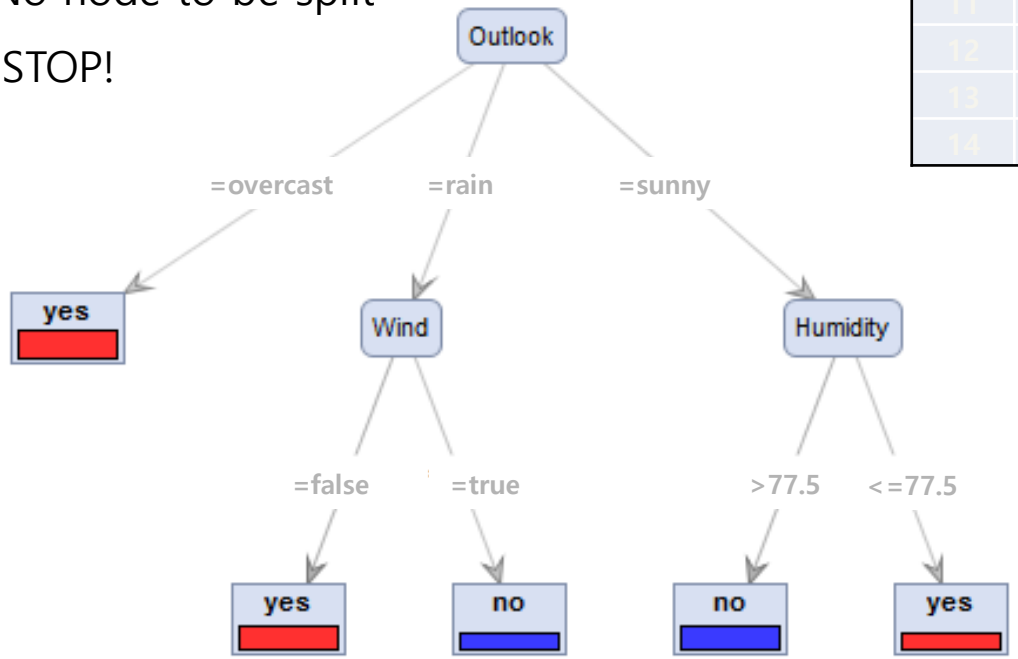


Hunt's Algorithm

1. 시작: 모든 record가 Root 노드 에 주어짐.
2. Split on Outlook
3. Data set assigned to node A
4. Split on Wind
5. Data set assigned to node B
6. Split on Humidity
7. No node to be split

   STOP!

Row #	Outlook	Temperature	Humidity	Wind	Play
1	sunny	85.0	85.0	false	no
2	sunny	80.0	90.0	true	no
3	overcast	83.0	78.0	false	yes
4	rain	70.0	96.0	false	yes
5	rain	68.0	80.0	false	yes
6	rain	65.0	70.0	true	no
7	overcast	64.0	65.0	true	yes
8	sunny	72.0	95.0	false	no
9	sunny	69.0	70.0	false	yes
10	rain	75.0	80.0	false	yes
11	sunny	75.0	70.0	true	yes
12	overcast	72.0	90.0	true	yes
13	overcast	81.0	75.0	false	yes
14	rain	71.0	80.0	true	no



All the leaf nodes are pure!
No node to be split
Stop!

Tree Induction

- 탐욕적 전략(Greedy algorithm)
  - 정하여진 기준을 최적화 시켜주는 속성 테스트에 기반하여 사례집합을 나눈다.
  - 과거, 미래에 영향 받지 않고 현재 노드 만 고려함.
  - 전역적 최적(globally optimal) 이 아닌 지역적 최적(locally optimal) 결과가 나올 수 있음.