



# 의사결정나무 모델: 분류 나무 (Classification Tree)





# Key words

#CART #C4.5 #C5.0 #CHAID  
#불순도 #지니 불순도 #엔트로피

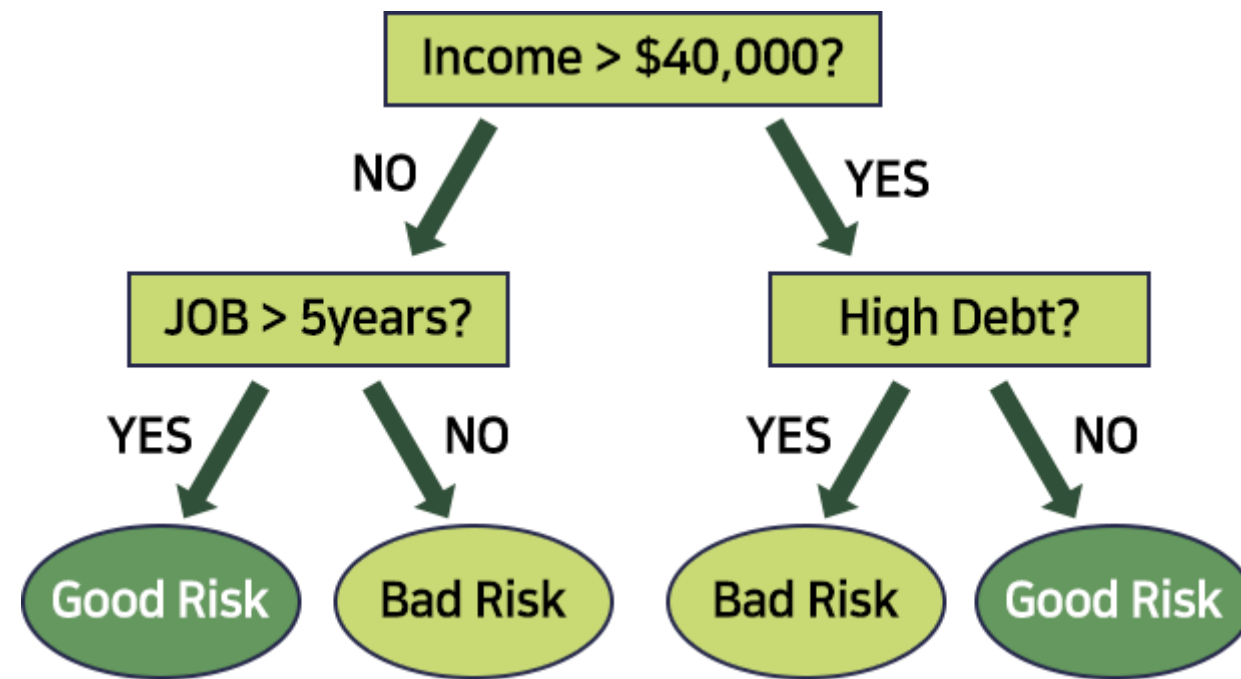
# 의사결정나무 개요

## 의사결정 나무 모형

- 의사결정규칙(decision rule)을 나무 구조로 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분할하는 방식으로 분류 및 예측하는 분석 방법.
  - 목표변수가 범주형인 경우 : 분류 나무(classification tree)
  - 목표변수가 연속형인 경우 : 회귀 나무(regression tree)

# 의사결정나무 개요

## 의사결정 나무의 구조



- 뿌리마디 (root node) : 시작되는 마디로 전체 자료로 구성됨.
- 자식마디 (child node) : 하나의 마디로부터 분리되어 나간 2개 이상의 마디들.
- 부모마디 (parent node) : 주어진 마디의 상위마디.

# 의사결정나무 개요

## 의사결정 나무의 구조



- 끝마디 (terminal node) : 자식마디가 없는 마디.
- 중간마디 (internal node) : 부모마디와 자식마디가 모두 있는 마디.
- 깊이 (depth) : 뿌리마디부터 끝마디까지 중간마디의 수.



# 의사결정나무 개요

## 의사결정 나무의 종류

- 의사결정나무 알고리즘은 CHAID, CART, C5.0, QUEST 등과 이들의 장점을 결합한 다양한 알고리즘이 있음.
- 주요 알고리즘의 특징

	분리기준	특징
CART	<ul style="list-style-type: none"><li>분류나무(범주형 목표변수) : 지니불순도</li><li>회귀나무(연속형 목표변수) : 분산감소량</li></ul>	<ul style="list-style-type: none"><li>항상 이진분리</li><li>개별 특성변수 및 특성변수의 선형결합 형태의 분리기준도 가능</li></ul>
C4.5, C5.0	<ul style="list-style-type: none"><li>엔트로피 불순도로 구한 정보이득</li></ul>	<ul style="list-style-type: none"><li>범주형 특성변수는 다진분리</li><li>연속형 특성변수는 이진분리</li></ul>
CHAID	<ul style="list-style-type: none"><li>분류나무(범주형 목표변수) : 카이제곱 통계량</li><li>회귀나무(연속형 목표변수) : ANOVA F 통계량</li></ul>	<ul style="list-style-type: none"><li>다진분리</li><li>변수간 통계적 관계에 기반</li></ul>

# 의사결정나무 개요

## 의사결정나무 분석절차

- 나무의 성장 (growing) : 각 마디에서 적절한 최적의 분리규칙 (splitting rule)을 찾아 나무를 성장시킴.  
정지 규칙(stopping rule)을 만족하는 경우는 성장을 중단.
- 가지치기 (pruning) : 오류율(error rate)을 크게 할 위험이 높거나 부적절한 추론 규칙을 가지고 있는 가지를 제거.
- 타당성 평가 : 평가자료(test data)를 이용하여 의사결정나무를 평가.
- 해석 및 예측 : 구축된 나무 모형을 해석하고 분류 및 예측 모형을 설정.

# 의사결정나무 개요

## 나무의 성장(growing)

- 상위노드로부터 하위노드로 나무 구조를 형성하는  
매 단계마다 분리규칙(어느 특성변수로 어떻게 분할할 것인가)을 선택함.
- 분리규칙의 형태(이진분할의 경우)
  - 연속형 특성변수 : 분리에 사용될 특성변수  $X$ 와 분리점  $c$ 를 이용하여  
 $X < c$  면 왼쪽 자식마디, 그렇지 않으면  
오른쪽 자식마디로 자료를 분리.
  - 범주형 특성변수 : 분리에 사용될 특성변수  $X$ 가 가지는 전체범주 중  
부분집합인  $A$ 를 이용하여  $X \in A$  면 왼쪽 자식마디,  
그렇지 않으면 오른쪽 자식마디로 자료를 분리.



# 의사결정나무 개요

## ■ 나무의 성장(growing)

- 분류기준은 해당 노드에서 그 기준으로 하위노드를 분기하였을 때, 하위 노드 내에서는 동질성이 하위 노드 간에는 이질성이 가장 커지도록 선택됨.

# 분류나무

## 분류나무의 분리규칙 탐색

- 불순도

$Y$ 의 범주가  $j = 1, \dots, C$ 로 구성될 때  $t$ 노드에서의 불순도  $\text{imp}(t)$ 는 다음과 같이 정의됨.

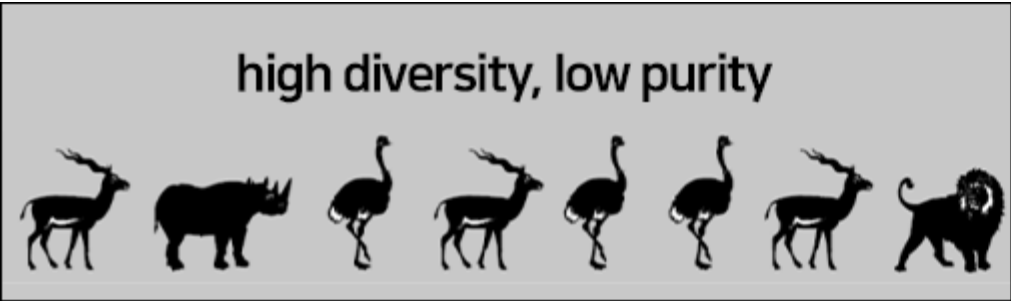
$p(j|t)$  :  $t$  마디로 분류했는데 범주  $j$ 에 속할 확률이라고 할 때,

- 지니 불순도 (gini impurity) :  $\text{imp}(t) = 1 - \sum_{j=1}^C p(j|t)^2$
- 엔트로피 불순도 (entropy impurity) :  $\text{imp}(t) = -\sum_{j=1}^C p(j|t) \log p(j|t)$
- 불순도가 클수록 자식노드 내 이질성이 큼을 의미함. 불순도가 가장 작아지는 방향으로 가지분할을 수행함.

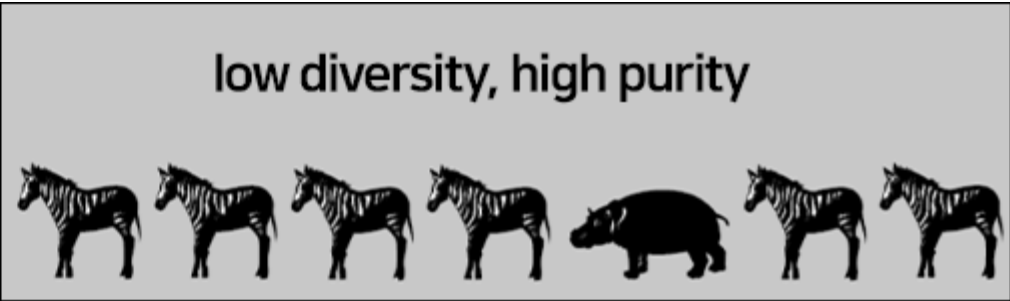
# 분류나무

## 분류나무의 분리규칙 탐색

- 지니불순도 산출 예시



$\text{Pr}(\text{interspecific encounter}) = 1 - 2(3/8)^2 - 2(1/8)^2 = .69$



$\text{Pr}(\text{interspecific encounter}) = 1 - (6/7)^2 - 2(1/7)^2 = .24$

## 분류나무

### 분류나무의 분리규칙 탐색

- 불순도의 향상된 정도 (Goodness of split)  $G(s, t)$

어떤 부모마디  $t$ 에서 분리기준  $s$ 로 분리한 뒤 생성된 두 자식마디를  $t_L, t_R$ 이라고 할 때,

$$G(s, t) = \text{imp}(t) - \frac{N(t_L)}{N(t)} \cdot \text{imp}(t_L) - \frac{N(t_R)}{N(t)} \cdot \text{imp}(t_R)$$

- $N(t)$  :  $t$  마디에서 자료의 수
- $\text{imp}(t)$  :  $t$  마디에서의 불순도



# 분류나무

## I 분류나무의 분리규칙 탐색

- 모든 특성변수와 그 특성변수의 모든 가능한 분리점에 대하여  $G(s, t)$ 를 구한 뒤  $G(s, t)$ 가 가장 큰 특성변수 및 분리점을 해당 마디에서의 분리기준으로 정함.

# 분류나무

## 불순도 감소량을 이용한 분리기준 예시

- 부모마디
  - $\text{imp}(t) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$
- 자식마디 (X1으로 분리 :  $s1$  ( M vs F ))
  - $\text{imp}(t_L) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$
  - $\text{imp}(t_R) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.4444$
- 자식마디 (X2으로 분리 :  $s2$  (고졸 vs 대졸))
  - $\text{imp}(t_L) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$
  - $\text{imp}(t_R) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.4444$

X1 (성별)	X2 (학력)	Y(구매여부)
M	대졸	1
M	대졸	0
F	고졸	1
F	대졸	0
F	고졸	1

# 분류나무

## 불순도 감소량을 이용한 분리기준 예시

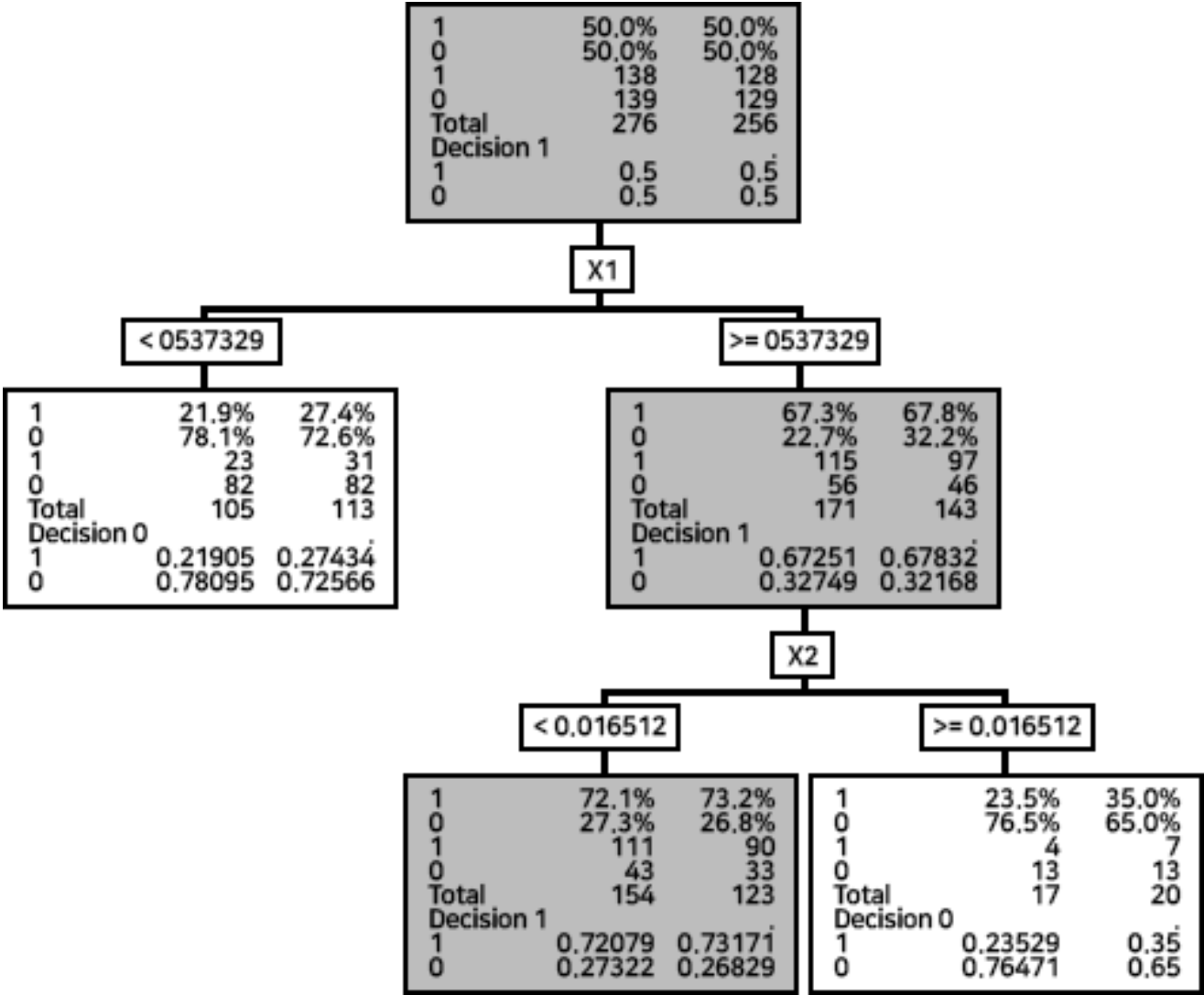
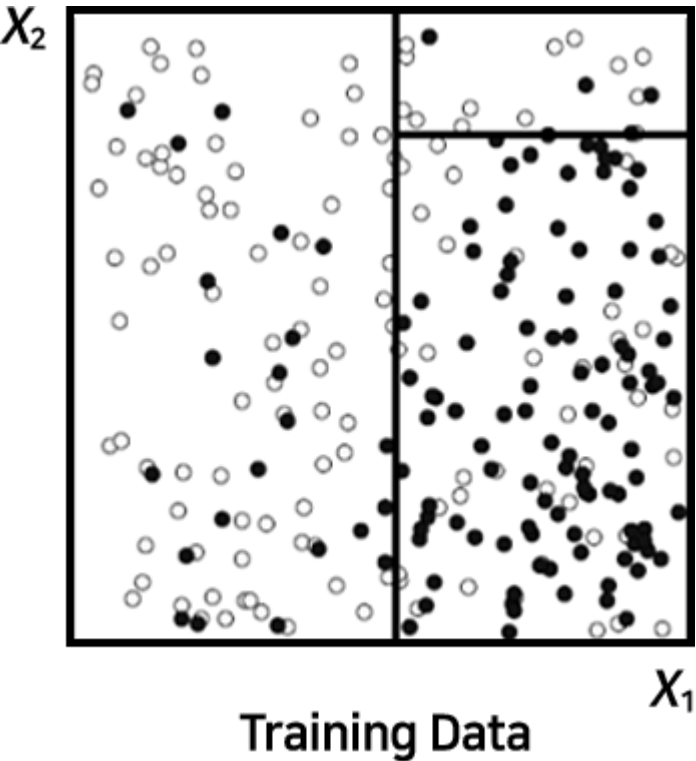
- 불순도가 향상된 정도
  - $G(s1,t) = 0.48 - \frac{2}{5} \times 0.5 - \frac{3}{5} \times 0.4444 = 0.01336$
  - $G(s2,t) = 0.48 - \frac{2}{5} \times 0 - \frac{3}{5} \times 0.4444 = 0.21336$   
→ s2가 더 좋은 분리기준임.

X1 (성별)	X2 (학력)	Y(구매여부)
M	대졸	1
M	고졸	0
F	대졸	1
F	대졸	0
F	고졸	1


# 분류나무

## 분류나무 적합 결과

- 분류경계면이 축에 수직임.







# 의사결정나무 모델: 회귀 나무 (Regression Tree)



The background features several thick, wavy, ribbon-like lines in shades of green, yellow, and blue. A small, glossy purple sphere is positioned near the top center, resting on one of the upper wavy lines.

# Key words

#분산감소량 #정지규칙 #가지치기

# 회귀나무

## 회귀나무(Regression Tree)의 분리규칙 탐색

- 분산의 감소량
  - 각 그룹(자식노드) 내에서의 목표변수의 분산이 작을수록, 그룹 내 이질성이 작은 것으로 볼 수 있음.
  - 자식노드로 분리했을 때 분산의 감소량이 가장 커지도록 하는 분리규칙을 탐색.

# 회귀나무

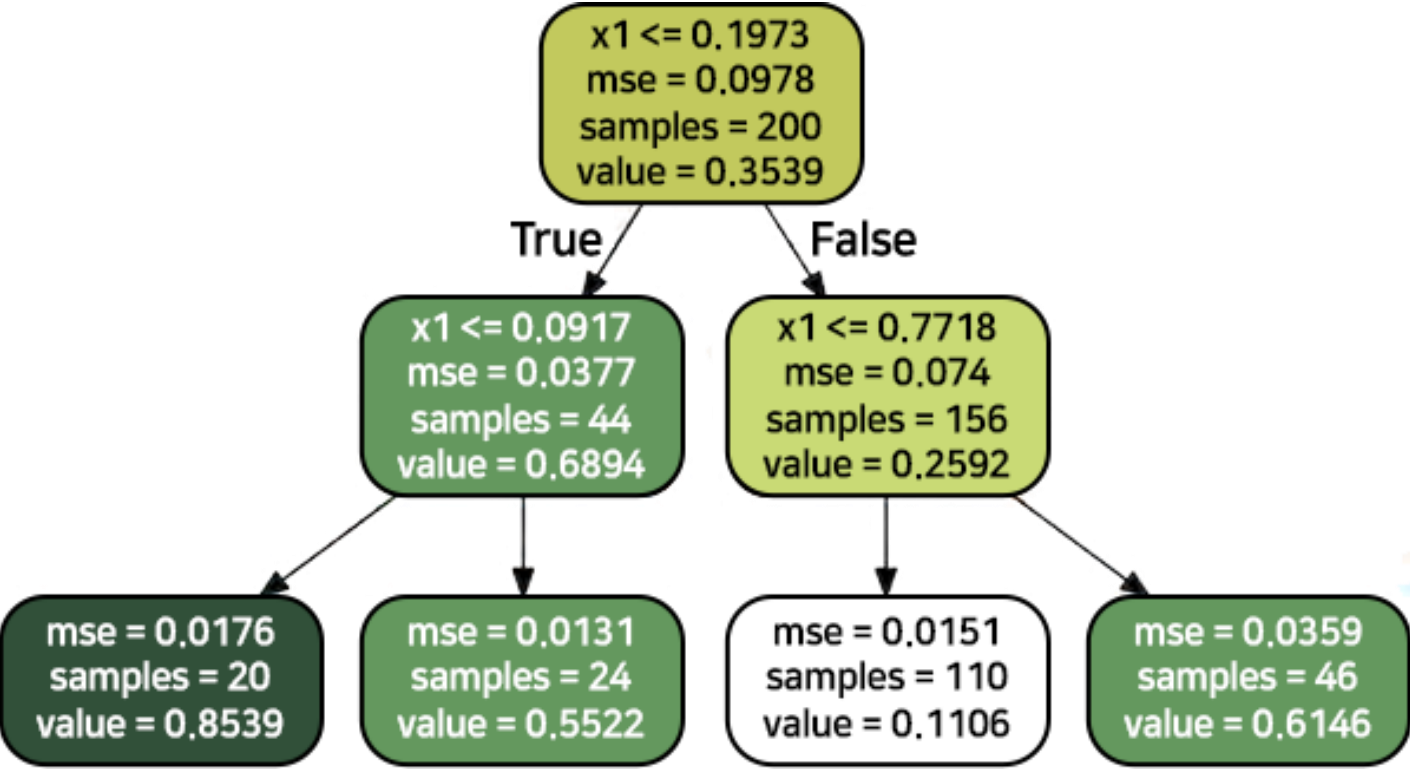
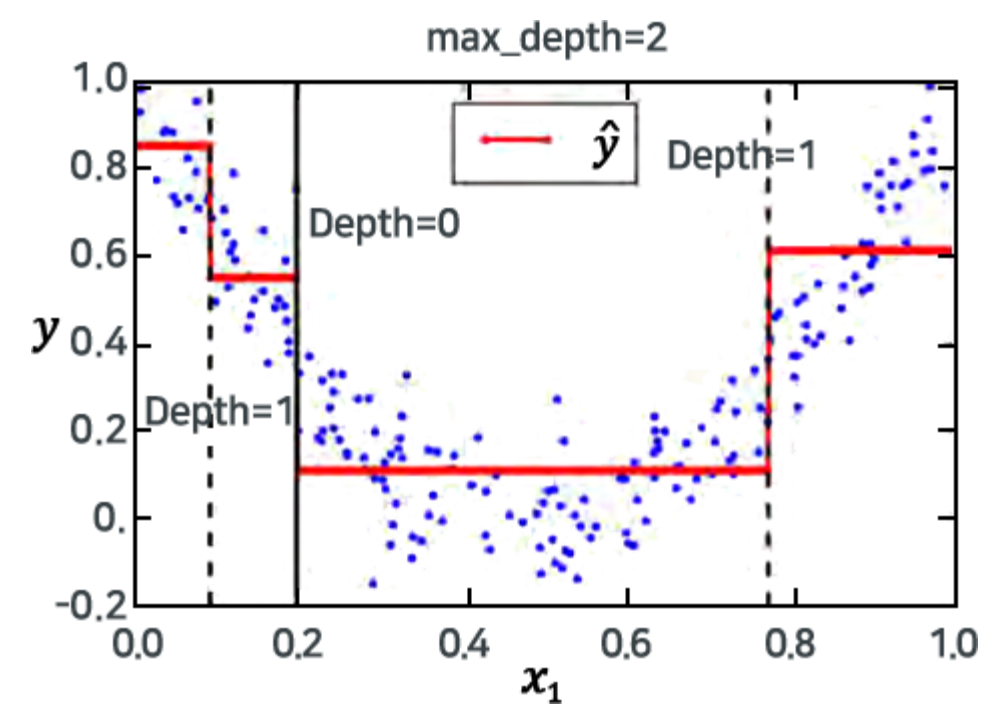
## I 회귀나무(Regression Tree)의 분리규칙 탐색

- ANOVA의 F 통계량
  - F값이 클수록 그룹(자식노드) 간에 평균차이가 있다는 것이므로, 그룹 간 이질성이 큰 것으로 볼 수 있음.
  - F값이 가장 커지게 되는 분리규칙을 탐색.



# 회귀나무

## 회귀나무 적합 결과



# 정지규칙과 가지치기

## 의사결정나무의 과적합 방지 방법

- 지나치게 많은 마디를 가지는 의사결정나무는 새로운 자료에 적용할 때 예측오차가 매우 커지는 과적합(overfitting) 상태가 됨.
- 이를 방지하기 위한 방법으로 정지규칙 또는 가지치기 방법을 사용함.

# 정지규칙과 가지치기

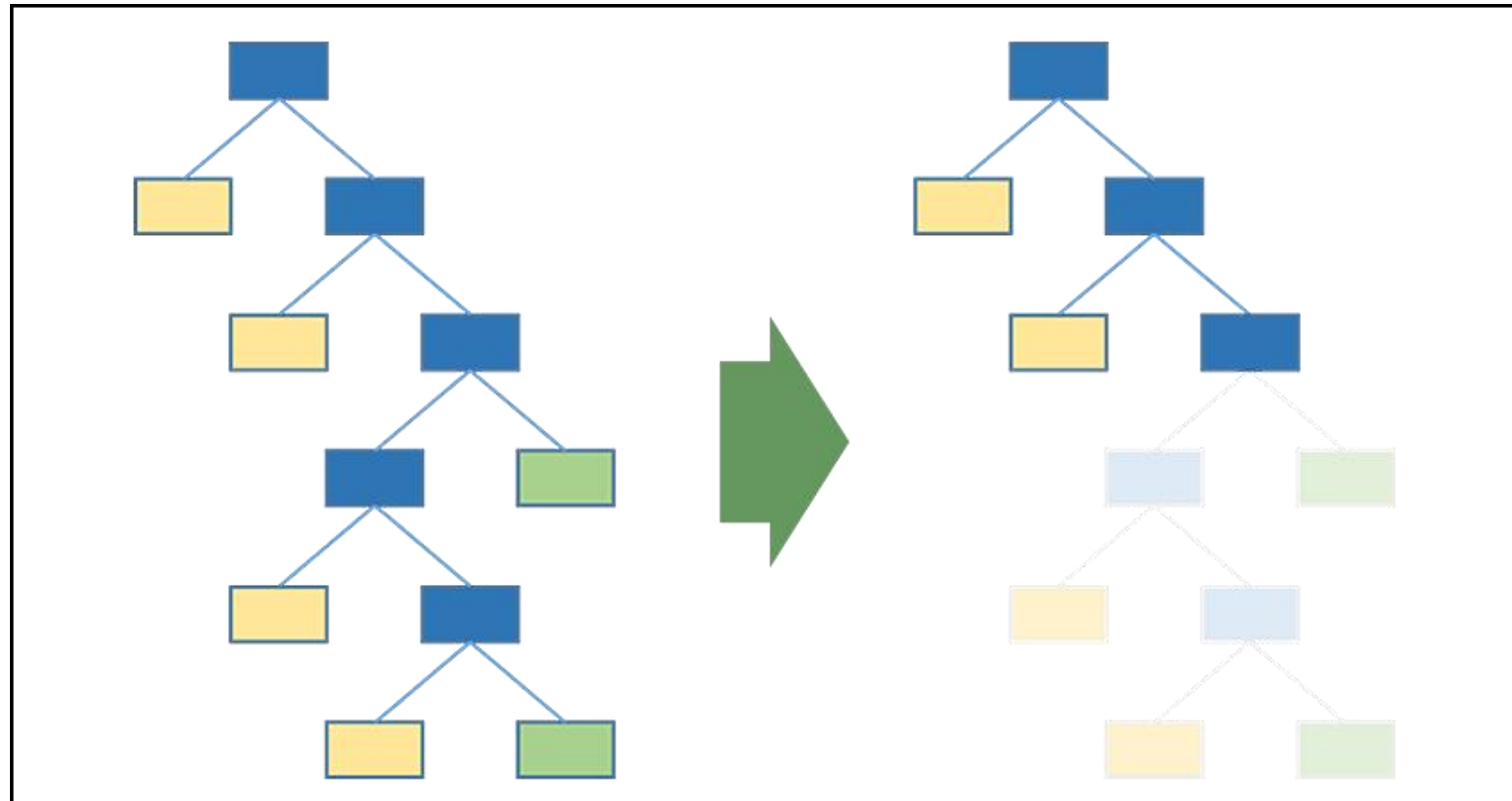
## 의사결정나무의 과적합 방지 방법

- 정지규칙 (stopping rule)
  - 다음의 경우에 더 이상 분리하지 않고 나무가 성장을 멈추도록 함.
    - 모든 자료의 목표변수 값이 동일할 때.
    - 마디에 속하는 자료의 개수가 일정 수준보다 적을 때.
    - 뿌리마디로부터의 깊이가 일정 수준 이상일 때.
    - 불순도의 감소량이 지정된 값보다 적을 때.

# 정지규칙과 가지치기

## 의사결정나무의 과적합 방지 방법

- 가지치기 (pruning)
  - 성장이 끝난 나무의 가지를 제거하여 적당한 크기를 가지도록 함.
  - 적당한 크기를 결정하는 방법은 검증용 자료(validation data)에 대한 예측 오류가 가장 작은 나무 모형을 찾는 것이 일반적이며, 이 과정은 의사결정나무 모형 알고리즘 내에 자동화되어 있는 경우가 많음.





# 의사결정나무모형 특징

장점	단점
<ul style="list-style-type: none"><li>• 이해하기 쉬운 규칙을 생성함 : if-then-else 방식</li><li>• 특성변수 및 목표변수 둘 다 연속형, 범주형 자료 모두 취급함.</li><li>• 데이터의 전처리가 거의 필요하지 않음.</li><li>• 이상치에 덜 민감.</li><li>• 모형에 가정이 필요 없는 비모수적 모형.</li></ul>	<ul style="list-style-type: none"><li>• 훈련결과가 불안정함.</li><li>• 모든 분할은 축에 수직임.</li><li>• 나무가 깊어질수록 과적합으로 예측력이 저하되며, 해석이 어려워짐.</li></ul>