Machine Learning

Decision Tree

Dept. SW and Communication Engineering

Prof. Giseop Noh (kafa46@hongik.ac.kr)

Contents

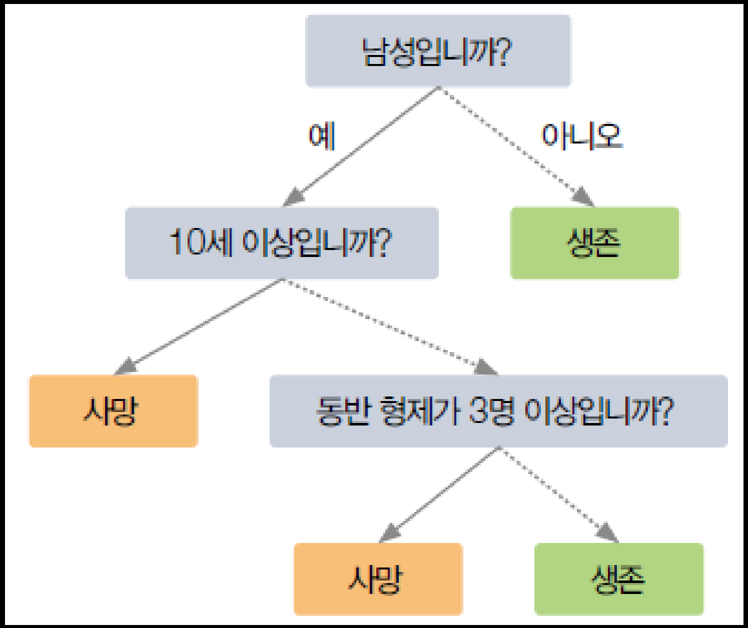
Decision Trees in the Context of Supervised Learning

Decision Trees in the Context of Supervised Learning

Titanic Survivor Prediction

- Supervised learning data used to predict Titanic survivors
- Classification can be performed using Decision Trees

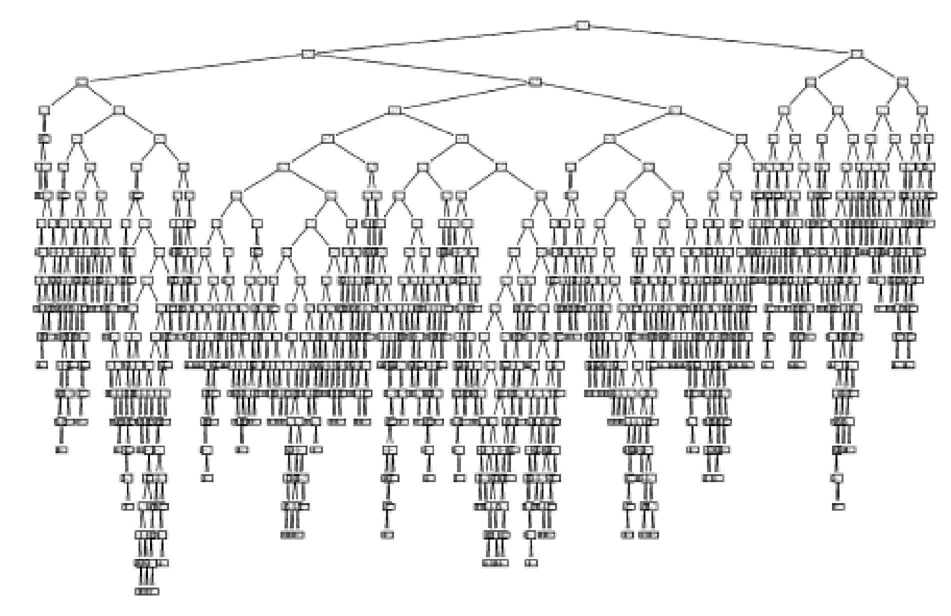
Name	Age (x_1)	Siblings (x_2)	Gender (x_3)	Survival (y_i)
Angela	3	1	Female	Survived
James	7	2	Male	Survived
Robert	4	3	Male	Died
Kelly	33	0	Female	Survived
Austin	35	0	Male	Died
Annie	19	4	Male	Died
Anis	29	0	Male	Died
DiCaprio	8	3	Male	?



Basic Concepts of Decision Trees

Ideas on Decision Tree

- **A method for expressing a series of rules (questions) that split data attributes into a tree structure**
- **The rules are like 'if-else' statements in programming**
- **A tree represents a process of classification or decision making**
- **The final nodes of the tree (leaf nodes) show**
  - The predicted class (for classification)
  - The predicted value (for regression)



Example of Decision Tree

■ Example of a Decision Tree

- Akinator Game: Similar to the “Twenty Questions (스무고개)” guessing game

https://kr.akinator.com/

Structure of Decision Tree Models

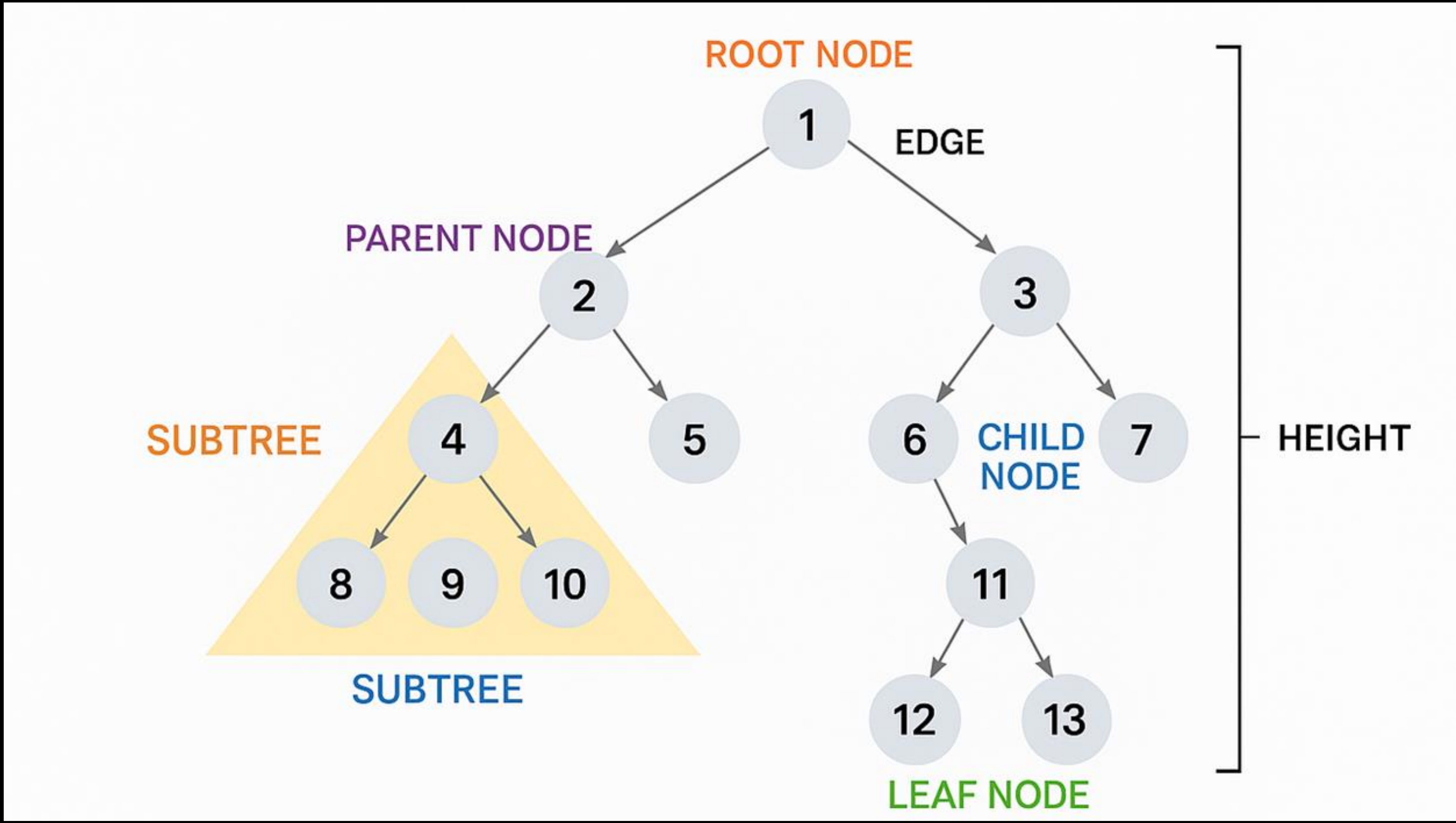
Tree Component

- **Node**
  - Root Node: The starting point of the tree
  - Parent Node: A node that has one or more child nodes
  - Child Node: A node that descends from a parent node
  - Sibling Node: Nodes that share the same parent
  - Leaf Node: A node that has no children

- **Edge**
  - A line that connects one node to another

- **Height**
  - The depth from the root node to the deepest leaf node

Visual Representation of Tree



Working Principles of Decision Tree

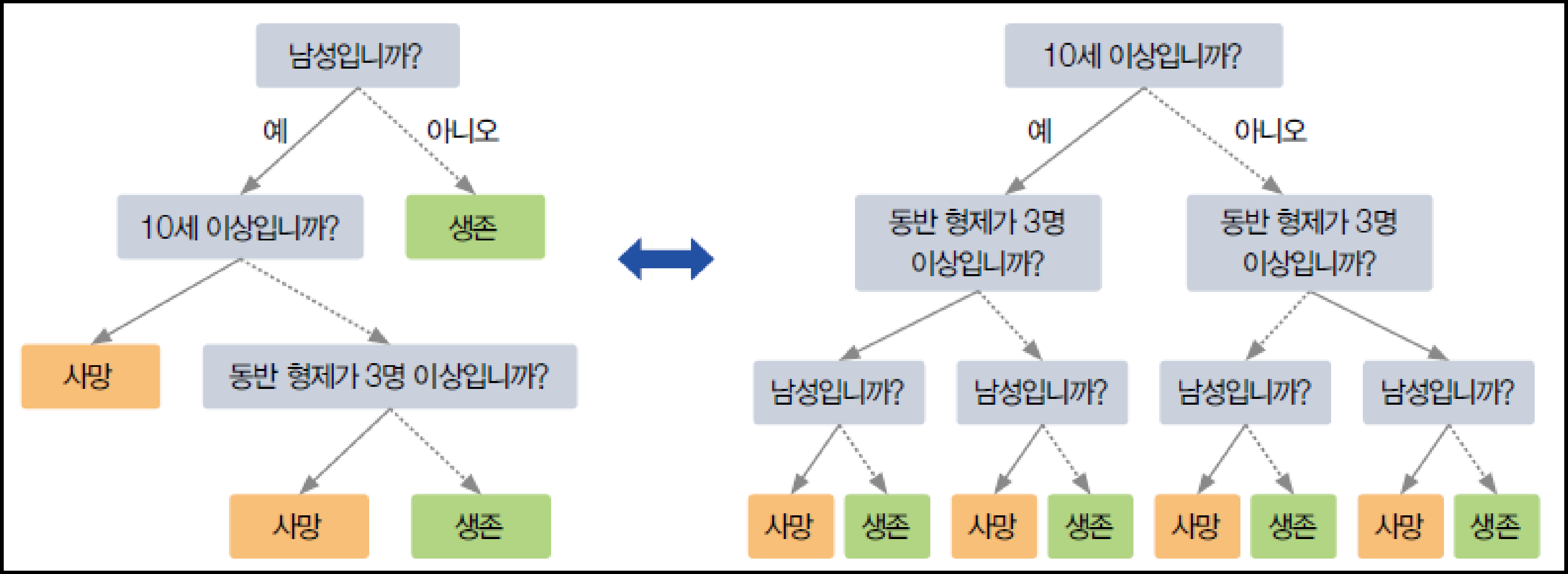
Decision Tree Splitting Rules

- **Determining the split rules that define each node in the decision tree**
  - The leaf nodes contain the final class or predicted value
  - Internal (parent) nodes encode if-else conditions

- **Splitting attributes**
  - These are the if-else conditions at parent nodes

- **Choosing the right splitting attributes is essential.**
  - Entropy
  - Gini Index

Example of Node Splitting



Recap: Information Entropy

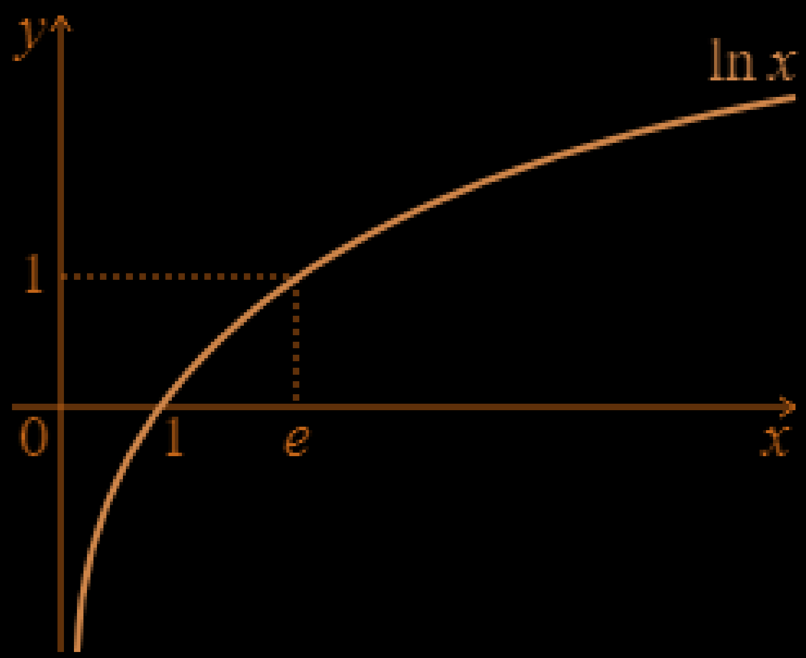
- **Information, *I***
  - 도대체 정보를 어떻게 표현할까?
  - 어떤 정보가 가치 있을까?
  - 내일은 해가 동쪽에서 뜬다.
  - 내일은 해가 서쪽에서 뜬다.
  - 교수님은 강의가 있는 날 출근하신다.
  - 교수님은 내일 퇴직하신다.

:

$$Information\ (I) \propto \frac{1}{p(x)} = p(x)^{-1}$$

$$x: random\ variable$$

Recap: Which Function we choose in Information



$$I(x) \propto \frac{1}{p(x)} = p(x)^{-1}$$

x: *random variable*

$$I(x) = \log_a \frac{1}{P(x)} = \log_a P(x)^{-1} = -\log_a P(x) \propto -\ln P(x)$$

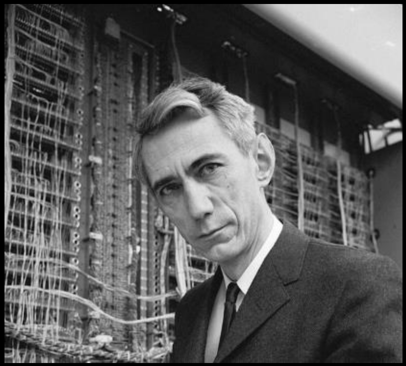
Recap: Information Entropy

Information Entropy

- Expected Information of Individual Events
- It becomes easier if you think in terms of the formula for average.
- Average

  → Expected Value: Multiply each outcome by its probability

  → Then sum up everything.



Claude Shannon (1916~2001)
섀넌에 의해 제안되어
'섀넌 엔트로피'라고 불리기도 함

$$H(P) = H(x) = E_{x\sim P}[I(x)] = E_{x\sim P}[-\log P(x)]$$

$$= -\sum_{x} P(x) \cdot \log P(x) = \sum_{i=1}^{n} p_i \cdot \log p_i$$

Information Entropy

Entropy

- A Measure of Information (Uncertainty) Based on Probability
- Entropy quantifies the amount of uncertainty using the probability of events
- When selecting a splitting attribute at a node, the information gain is calculated.
- The attribute that minimizes entropy is chosen to build the decision tree

$$H(D) = -\sum_{i=1}^{n} p_i \log p_i$$

- D: Data set
- n: Number of target classes to be classified
- p_i: Probability of the i-th class in the data set (i.e., proportion of data with class i)

Comparison of Entropy Level

- Low entropy (확률이 높다, 뻔하다, 확실하다. etc.)
  - One class dominates → Low uncertainty → Less information
- High entropy (확률이 낮다. 어찌될지 모른다. 불확실하다. etc.)
  - Classes are evenly mixed → High uncertainty → More information

Information Gain & ID3

Toy Example: Entropy Computation

No.	Age (x_1)	Income (x_2)	Student (x_3)	Credit Rating (x_4)	Purchase (y)
1	Teen	High	No	Fair	No
2	Teen	High	No	Excellent	No
3	Young Adult	High	No	Fair	Yes
4	Middle Aged	Medium	No	Fair	Yes
5	Middle Aged	Low	No	Fair	Yes
6	Middle Aged	Low	Yes	Excellent	No
7	Young Adult	Low	Yes	Excellent	Yes
8	Teen	Medium	No	Fair	No
9	Teen	Low	Yes	Fair	Yes
10	Middle Aged	Medium	Yes	Fair	Yes
11	Teen	Medium	Yes	Excellent	Yes
12	Young Adult	Medium	No	Excellent	Yes
13	Teen	High	Yes	Fair	Yes
14	Middle Aged	Medium	No	Excellent	No

- Probability of purchasing a computer: 9/14
- Probability of not purchasing a computer: 5/14

$$H(D) = \left(-\frac{9}{14}\ln\frac{9}{14}\right) + \left(-\frac{5}{14}\ln\frac{5}{14}\right)$$

$$= 0.95$$

Information Gain

■ Information Gain

- A metric used to evaluate the effectiveness of each attribute when selecting a splitting criterion for nodes in the training data
- If data is split based on the attribute that maximizes information gain at each node,

  ➔ resulting tree structure will be the most efficient in terms of classification

■ The greater the Information Gain, the more informative the attribute

$$Gain(D, A) = H(D) - H_A(D)$$

- $H(D)$: 전체 데이터 집합 D 에 대한 엔트로피
  - 데이터를 분할하기 전, 현재 전체 데이터가 얼마나 불확실하고 섞여 있는지를 수치로 표현
- $H_A(D)$: 속성 A를 기준으로 데이터를 분할했을 때의 전체 엔트로피
  - 속성 A로 데이터를 나눠본 뒤, 나눠진 각 그룹의 엔트로피를 계산하고,
  - 그 그룹들의 크기를 고려해 전체 평균 엔트로피를 구한 것!

Information Gain

Information Gain

$$Gain(D, A) = H(D) - H_A(D)$$

- If splitting on attribute A results in lower entropy, then A is a good splitting attribute
- To calculate $H_A(D)$, we compute the entropy for each subset of data for each value of A, weighted by the size of each subset

$$H_A(D) = \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \cdot H(D_v)$$

- $Values(A)$: Set of possible values for attribute A
- D_v : Subset of D where attribute A has value v
- $|D|$: Total number of instances in dataset D

Tree Growth

■ Model Construction Based on Information Gain

- ID3 Algorithm is used for building a decision tree
- The process of building a decision tree is commonly referred to as tree growth
- At each node, a splitting attribute is selected such that entropy is minimized as much as possible

반복적으로 둘로 나누는 (영어, 다이코터마이즈)

버전 3 알고리즘

■ ID3 (Iterative Dichotomiser 3)

- An algorithm that iteratively divides the dataset
  - Top-down greedy optimization algorithm that selects the attribute with the highest information gain
  - At each node, the attribute that yields the maximum information gain is selected to split the data

Information Gain Calculation Dataset

No.	Age (x_1)	Income (x_2)	Student (x_3)	Credit Rating (x_4)	Purchase (y)
1	Teen	High	No	Fair	No
2	Teen	High	No	Excellent	No
3	Young Adult	High	No	Fair	Yes
4	Middle Aged	Medium	No	Fair	Yes
5	Middle Aged	Low	Yes	Fair	Yes
6	Middle Aged	Low	Yes	Excellent	No
7	Young Adult	Low	Yes	Excellent	Yes
8	Teen	Medium	No	Fair	No
9	Teen	Low	Yes	Fair	Yes
10	Middle Aged	Medium	Yes	Fair	Yes
11	Teen	Medium	Yes	Excellent	Yes
12	Young Adult	Medium	No	Excellent	Yes
13	Young Adult	High	Yes	Fair	Yes
14	Middle Aged	Medium	No	Excellent	No

Information Gain Based on Age Attribute

No.	Age (x_1)	Income (x_2)	Student (x_3)	Credit Rating (x_4)	Purchase (y)
1	Teen	High	No	Fair	No
2	Teen	High	No	Excellent	No
3	Young Adult	High	No	Fair	Yes
4	Middle Aged	Medium	No	Fair	Yes
5	Middle Aged	Low	Yes	Fair	Yes
6	Middle Aged	Low	Yes	Excellent	No
7	Young Adult	Low	Yes	Excellent	Yes
8	Teen	Medium	No	Fair	No
9	Teen	Low	Yes	Fair	Yes
10	Middle Aged	Medium	Yes	Fair	Yes
11	Teen	Medium	Yes	Excellent	Yes
12	Young Adult	Medium	No	Excellent	Yes
13	Young Adult	High	Yes	Fair	Yes
14	Middle Aged	Medium	No	Excellent	No

Information Gain Based on Age Attribute

$$Gain(D, age) = H(D) - H_{age}(D) \qquad H_A(D) = \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \cdot H(D_v)$$

$$\left.\begin{aligned} P(Yes) &= \frac{9}{14} \\ P(No) &= \frac{5}{14} \end{aligned}\right\} \text{구매 확률}$$

$$H(D) = \left(-\frac{9}{14}\ln\frac{9}{14}\right) + \left(-\frac{5}{14}\ln\frac{5}{14}\right) \approx 0.94$$

$$H_{age}(D) = \frac{5}{14}\left(\left(-\frac{2}{5}\ln\frac{2}{5}\right) + \left(-\frac{3}{5}\ln\frac{3}{5}\right)\right)$$

Teen entropy

$$+ \frac{4}{14}\left(-\frac{4}{4}\ln\frac{4}{4}\right)$$

Young Adult entropy

$$+ \frac{5}{14}\left(\left(-\frac{3}{5}\ln\frac{3}{5}\right) + \left(-\frac{2}{5}\ln\frac{2}{5}\right)\right)$$

Middle Aged entropy

$$= 0.693$$

$$Gain(D, age) = H(D) - H_{age}(D)$$

$$= 0.94 - 0.693 = 0.247$$

Information Gain Based on Income Attribute

No.	Age (x_1)	Income (x_2)	Student (x_3)	Credit Rating (x_4)	Purchase (y)
1	Teen	High	No	Fair	No
2	Teen	High	No	Excellent	No
3	Young Adult	High	No	Fair	Yes
4	Middle Aged	Medium	No	Fair	Yes
5	Middle Aged	Low	Yes	Fair	Yes
6	Middle Aged	Low	Yes	Excellent	No
7	Young Adult	Low	Yes	Excellent	Yes
8	Teen	Medium	No	Fair	No
9	Teen	Low	Yes	Fair	Yes
10	Middle Aged	Medium	Yes	Fair	Yes
11	Teen	Medium	Yes	Excellent	Yes
12	Young Adult	Medium	No	Excellent	Yes
13	Young Adult	High	Yes	Fair	Yes
14	Middle Aged	Medium	No	Excellent	No

Information Gain Based on Income Attribute

$$Gain(D, income) = H(D) - H_{income}(D) \qquad H_A(D) = \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \cdot H(D_v)$$

$$\left.\begin{array}{l} P(Yes) = \frac{9}{14} \\ \\ P(No) = \frac{5}{14} \end{array}\right\} \text{구매 확률}$$

$$H(D) = \left(-\frac{9}{14}\ln\frac{9}{14}\right) + \left(-\frac{5}{14}\ln\frac{5}{14}\right)$$

$$\approx 0.94$$

$$H_{income}(D) = \frac{4}{14}\left(\left(-\frac{2}{4}\ln\frac{2}{4}\right) + \left(-\frac{2}{4}\ln\frac{2}{4}\right)\right)$$

High entropy

$$+ \frac{6}{14}\left(\left(-\frac{4}{6}\ln\frac{4}{6}\right) + \left(-\frac{2}{6}\ln\frac{2}{6}\right)\right)$$

Medium entropy

$$+ \frac{4}{14}\left(\left(-\frac{3}{4}\ln\frac{3}{4}\right) + \left(-\frac{1}{4}\ln\frac{1}{4}\right)\right)$$

Low entropy

$$= 0.911$$

$$Gain(D, income) = H(D) - H_{income}(D)$$

$$= 0.94 - 0.911 = 0.029$$

Information Gain Based on Student Attribute

No.	Age (x_1)	Income (x_2)	Student (x_3)	Credit Rating (x_4)	Purchase (y)
1	Teen	High	No	Fair	No
2	Teen	High	No	Excellent	No
3	Young Adult	High	No	Fair	Yes
4	Middle Aged	Medium	No	Fair	Yes
5	Middle Aged	Low	No	Fair	Yes
6	Middle Aged	Low	Yes	Excellent	No
7	Young Adult	Low	Yes	Excellent	Yes
8	Teen	Medium	No	Fair	No
9	Teen	Low	Yes	Fair	Yes
10	Middle Aged	Medium	Yes	Fair	Yes
11	Teen	Medium	Yes	Excellent	Yes
12	Young Adult	Medium	No	Excellent	Yes
13	Young Adult	High	Yes	Fair	Yes
14	Middle Aged	Medium	No	Excellent	No

Information Gain Based on Credit Attribute

$$Gain(D, student) = H(D) - H_{student}(D) \qquad H_A(D) = \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \cdot H(D_v)$$

$$\left.\begin{aligned} P(Yes) &= \frac{9}{14} \\ P(No) &= \frac{5}{14} \end{aligned}\right\} \text{구매 확률}$$

$$H(D) = \left(-\frac{9}{14}\ln\frac{9}{14}\right) + \left(-\frac{5}{14}\ln\frac{5}{14}\right)$$

$$\approx 0.94$$

$$H_{student}(D) = \frac{7}{14}\left(\left(-\frac{6}{7}\ln\frac{6}{7}\right) + \left(-\frac{1}{7}\ln\frac{1}{7}\right)\right)$$

Excellent entropy

$$+ \frac{7}{14}\left(\left(-\frac{3}{7}\ln\frac{3}{7}\right) + \left(-\frac{4}{7}\ln\frac{4}{7}\right)\right)$$

Fair entropy

$$= 0.789$$

$$Gain(D, student) = H(D) - H_{student}(D)$$

$$= 0.94 - 0.789 = 0.151$$

Information Gain Based on Credit Attribute

No.	Age (x_1)	Income (x_2)	Student (x_3)	Credit Rating (x_4)	Purchase (y)
1	Teen	High	No	Fair	No
2	Teen	High	No	Excellent	No
3	Young Adult	High	No	Fair	Yes
4	Middle Aged	Medium	No	Fair	Yes
5	Middle Aged	Low	Yes	Fair	Yes
6	Middle Aged	Low	Yes	Excellent	No
7	Young Adult	Low	Yes	Excellent	Yes
8	Teen	Medium	No	Fair	No
9	Teen	Low	Yes	Fair	Yes
10	Middle Aged	Medium	Yes	Fair	Yes
11	Teen	Medium	Yes	Excellent	Yes
12	Young Adult	Medium	No	Excellent	Yes
13	Young Adult	High	Yes	Fair	Yes
14	Middle Aged	Medium	No	Excellent	No

Information Gain Based on Credit Attribute

$$Gain(D, credit) = H(D) - H_{credit}(D) \qquad H_A(D) = \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \cdot H(D_v)$$

$$\left.\begin{aligned} P(Yes) &= \frac{9}{14} \\ P(No) &= \frac{5}{14} \end{aligned}\right\} \text{구매 확률}$$

$$H(D) = \left(-\frac{9}{14}\ln\frac{9}{14}\right) + \left(-\frac{5}{14}\ln\frac{5}{14}\right)$$

$$\approx 0.94$$

$$H_{credit}(D) = \frac{8}{14}\left(\left(-\frac{6}{8}\ln\frac{6}{8}\right) + \left(-\frac{2}{8}\ln\frac{2}{8}\right)\right)$$

Excellent entropy

$$+ \frac{6}{14}\left(\left(-\frac{3}{6}\ln\frac{3}{6}\right) + \left(-\frac{3}{6}\ln\frac{3}{6}\right)\right)$$

Fair entropy

$$= 0.892$$

$$Gain(D, credit) = H(D) - H_{credit}(D)$$

$$= 0.94 - 0.892 = 0.048$$

Select the Best Information Gain

Select Max

Information Gain

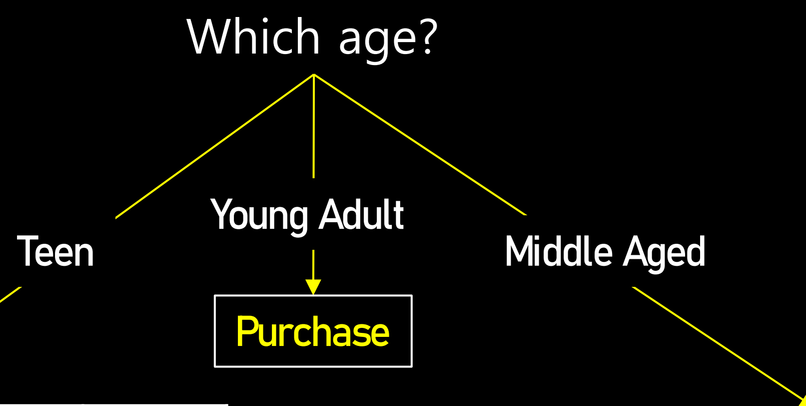
$$\boxed{Gain(D, age) = H(D) - H_{age}(D) = 0.94 - 0.693 = 0.247}$$

$$Gain(D, income) = H(D) - H_{income}(D) = 0.94 - 0.911 = 0.029$$

$$Gain(D, student) = H(D) - H_{student}(D) = 0.94 - 0.789 = 0.151$$

$$Gain(D, credit) = H(D) - H_{credit}(D) = 0.94 - 0.892 = 0.048$$

Construct Top-level



No.	Age (x_1)	Income (x_2)	Student (x_3)	Credit (x_4)	Purchase (y)
1	Teen	High	No	Fair	No
2	Teen	High	No	Excellent	No
8	Teen	Medium	No	Fair	No
9	Teen	Low	Yes	Fair	Yes
11	Teen	Medium	Yes	Excellent	Yes

No.	Age (x_1)	Income (x_2)	Student (x_3)	Credit (x_4)	Purchase (y)
4	Middle Aged	Medium	No	Fair	Yes
5	Middle Aged	Low	No	Fair	Yes
6	Middle Aged	Low	Yes	Excellent	No
10	Middle Aged	Medium	Yes	Fair	Yes
14	Middle Aged	Medium	No	Excellent	No

Compute Teen Age Information Gain

Teen

No.	Age (x_1)	Income (x_2)	Student (x_3)	Credit (x_4)	Purchase (y)
1	Teen	High	No	Fair	No
2	Teen	High	No	Excellent	No
8	Teen	Medium	No	Fair	No
9	Teen	Low	Yes	Fair	Yes
11	Teen	Medium	Yes	Excellent	Yes

Select Max

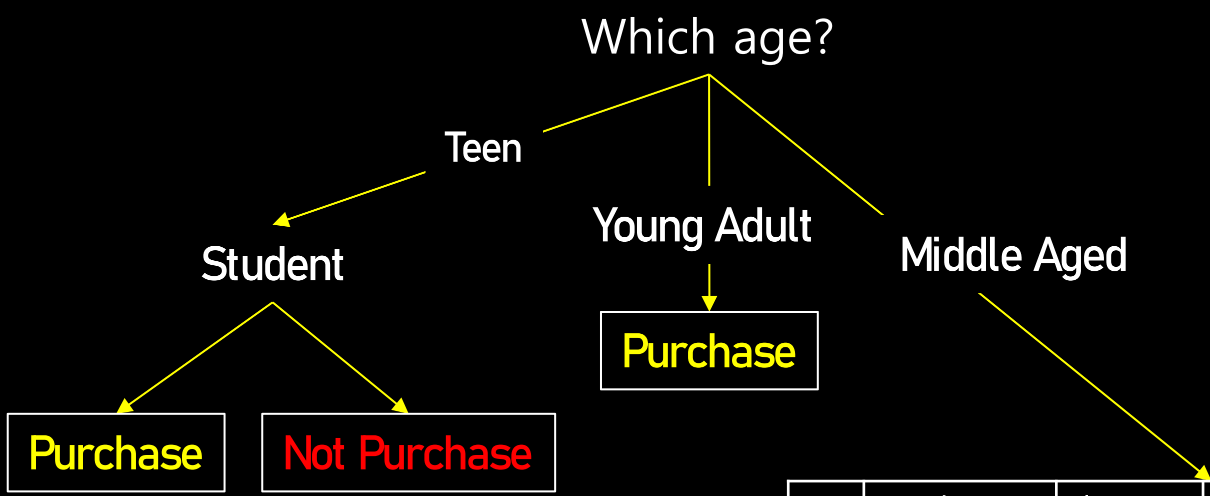
Information Gain

$$Gain(D_{teen}, income) = H(D_{teen}) - H_{income}(D_{teen}) = 0.971 - 0.4 = 0.571$$

$$\boxed{Gain(D_{teen}, student) = H(D_{teen}) - H_{student}(D_{teen}) = 0.971 - 0.0 = 0.971}$$

$$Gain(D_{teen}, credit) = H(D_{teen}) - H_{credit}(D_{teen}) = 0.971 - 0.951 = 0.02$$

Construct Second Level



No.	Age (x_1)	Income (x_2)	Student (x_3)	Credit (x_4)	Purchase (y)
1	Teen	High	No	Fair	No
2	Teen	High	No	Excellent	No
8	Teen	Medium	No	Fair	No
9	Teen	Low	Yes	Fair	Yes
11	Teen	Medium	Yes	Excellent	Yes

No.	Age (x_1)	Income (x_2)	Student (x_3)	Credit (x_4)	Purchase (y)
4	Middle Aged	Medium	No	Fair	Yes
5	Middle Aged	Low	No	Fair	Yes
6	Middle Aged	Low	Yes	Excellent	No
10	Middle Aged	Medium	Yes	Fair	Yes
14	Middle Aged	Medium	No	Excellent	No

Compute Middle Aged Information Gain

Middle Aged

No.	Age (x_1)	Income (x_2)	Student (x_3)	Credit (x_4)	Purchase (y)
4	Middle Aged	Medium	No	Fair	Yes
5	Middle Aged	Low	No	Fair	Yes
6	Middle Aged	Low	Yes	Excellent	No
10	Middle Aged	Medium	Yes	Fair	Yes
14	Middle Aged	Medium	No	Excellent	No

Select Max

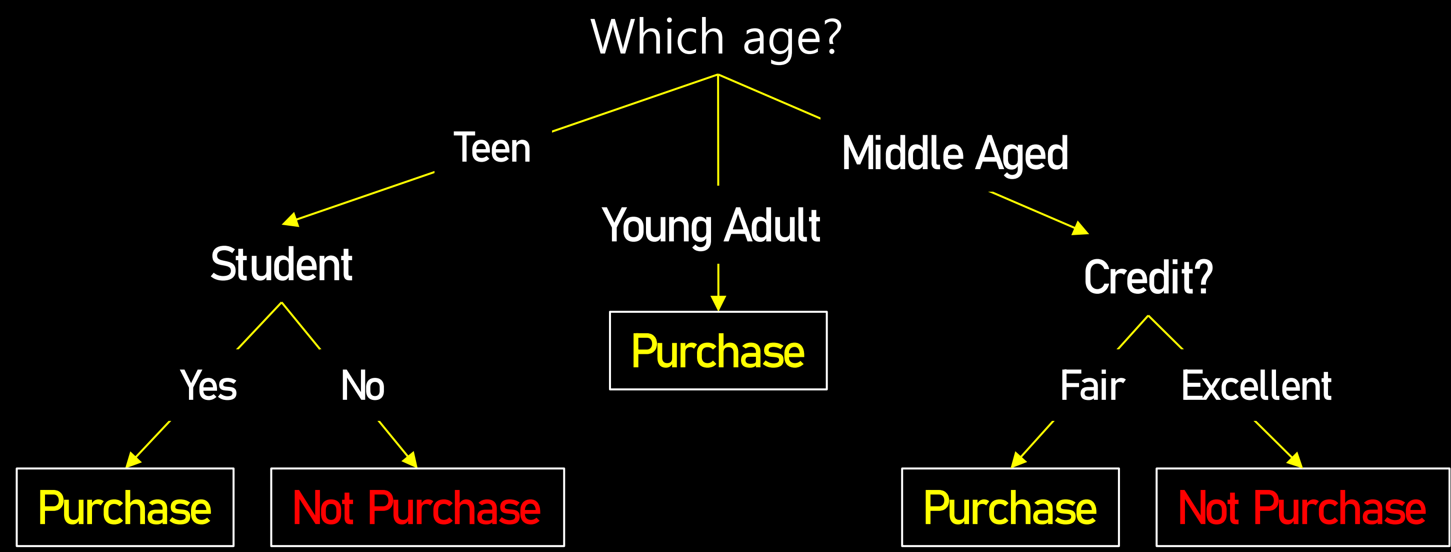
Information Gain

$$Gain(D_{middle\ aged}, income) = H(D_{middle\ aged}) - H_{income}(D_{middle\ aged}) = 0.971 - 0.951 = 0.02$$

$$Gain(D_{middle\ aged}, student) = H(D_{middle\ aged}) - H_{student}(D_{middle\ aged}) = 0.971 - 0.951 = 0.02$$

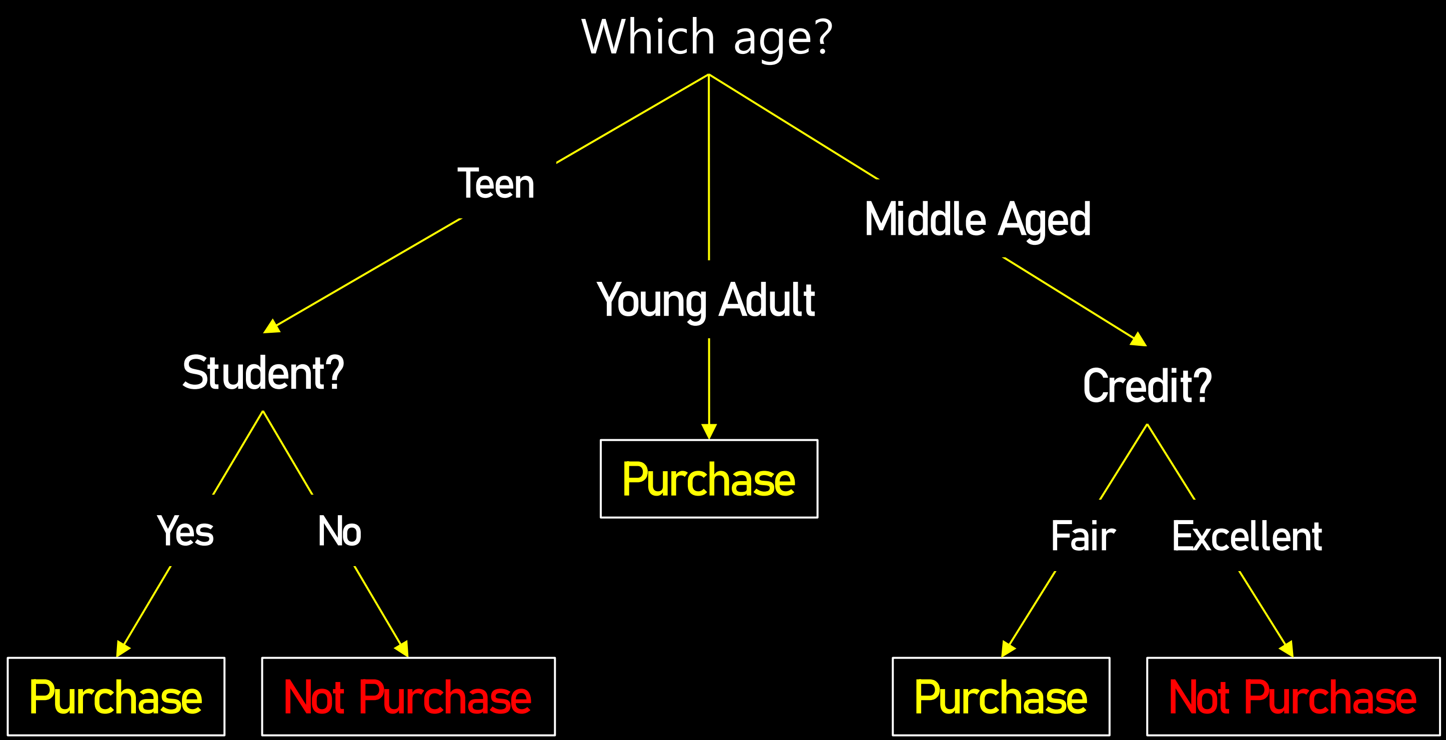
$$\boxed{Gain(D_{middle\ aged}, credit) = H(D_{middle\ aged}) - H_{credit}(D_{middle\ aged}) = 0.971 - 0.0 = 0.971}$$

Construct Second Level



No.	Age (x_1)	Income (x_2)	Student (x_3)	Credit (x_4)	Purchase (y)
4	Middle Aged	Medium	No	Fair	Yes
5	Middle Aged	Low	No	Fair	Yes
6	Middle Aged	Low	Yes	Excellent	No
10	Middle Aged	Medium	Yes	Fair	Yes
14	Middle Aged	Medium	No	Excellent	No

Construct Second Level



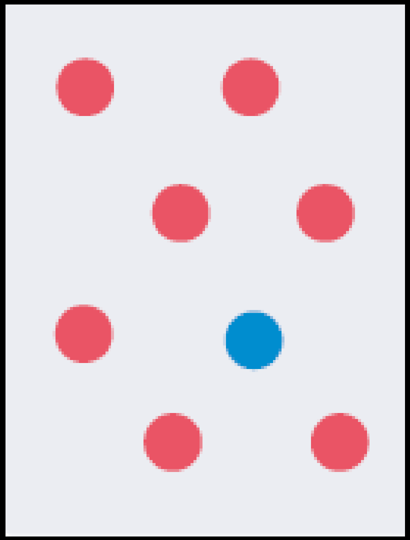
GINI & CART Algorithm

Impurity

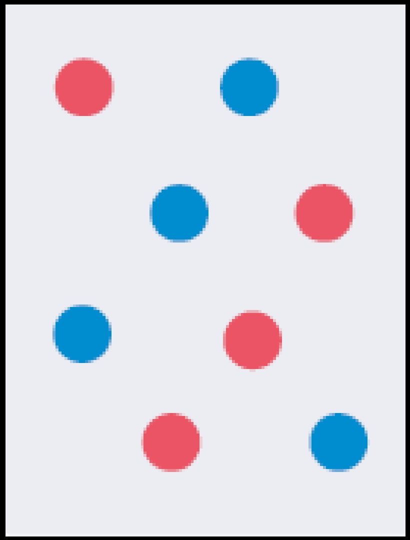
■ Impurity (불순도)

- If the data contains only one color (i.e., one class), the impurity is low
- If the data contains a mix of different colors (i.e., multiple classes evenly mixed)

  ➔ Impurity is high



Low impurity
(mostly one class)



High impurity
(balanced class mix)

Gini Index

Gini Index

- A metric that measures the impurity of a dataset
- Gini index ranges from 0 to 1
  - Gini = 0 → Pure node (perfectly classified)
  - Gini = 1 → Maximum impurity (completely mixed classes)
- The lower the Gini index, the purer the node

$$GINI(D) = 1 - \sum_{j} p(j)^2$$, where $p(j)$: $Propotion\ \ of\ class\ \ j\ \ in\ \ dataset\ D$

CART Algorithm

- Uses the Gini index to split nodes
- Grows the tree by selecting the split that minimizes the Gini index

Dataset

번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
1	청소년	고소득층	아니오	좋음	미구매
2	청소년	고소득층	아니오	아주 좋음	미구매
3	청년	고소득층	아니오	좋음	구매
4	중년	중소득층	아니오	좋음	구매
5	중년	저소득층	예	좋음	구매
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매

Gini Index - 1st split

First Attribute Split – Gini Index of the Root Node

$$GINI(D_{root}) = 1 - \left(\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right) = 0.49$$

- 4 out of 7 data points → Yes (Purchase)
- 3 out of 7 data points → No (No Purchase)

Gini Index Based on Age Attribute

- **Split the node based on the Age attribute**
  - The age values are categorized as:
    - Teen = 0
    - Young Adult = 1
    - Middle Aged = 2

- **Applying CART algorithm, we typically perform binary splits**
  - We can try two types of binary splits
    - Teen vs. Young Adult + Middle Aged
    - Teen + Young Adult vs. Middle Aged

Age Attribute 1. Teen vs. Young Adult + Middle Aged

- 나이 속성 기준의 지니계수

$$GINI(D_{청소년}) = 1 - \left\{ \left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2 \right\} = 0$$

$$GINI(D_{청년,중년}) = 1 - \left\{ \left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2 \right\} = 0.32$$

$$불순도\ 감소량 = GINI(D_{루트}) - \left(\frac{N_{청소년}}{N_{루트}} \times GINI(D_{청소년}) + \frac{N_{청년,중년}}{N_{루트}} GINI(D_{청년,중년}) \right)$$

$$= 0.49 - \left\{ \left(\frac{2}{7} \times 0\right) + \left(\frac{5}{7} \times 0.32\right) \right\} = 0.26$$

번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
1	청소년	고소득층	아니오	좋음	미구매
2	청소년	고소득층	아니오	아주 좋음	미구매
3	청년	고소득층	아니오	좋음	구매
4	중년	중소득층	아니오	좋음	구매
5	중년	저소득층	예	좋음	구매
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매

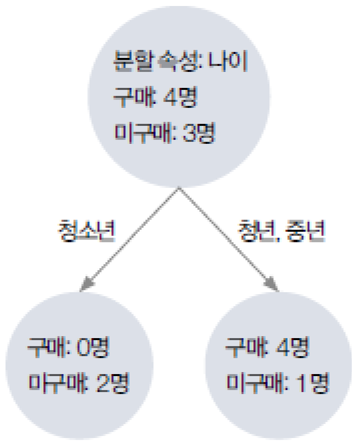


그림 8-10 나이 속성에 따른 컴퓨터 구매 여부 데이터: 청소년 / 청년 및 중년

Age Attribute 2. Teen + Young vs. Middle Aged

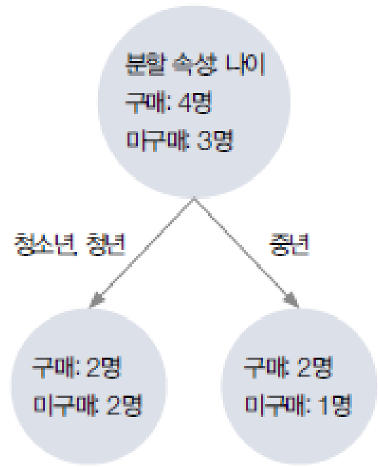
$$GINI(D_{\text{청소년, 청년}}) = 1 - \left\{ \left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2 \right\} = 0.5$$

$$GINI(D_{\text{중년}}) = 1 - \left\{ \left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right\} = 0.44$$

$$\text{불순도 감소량} = GINI(D_{\text{뿌리}}) - \left(\frac{N_{\text{청소년, 청년}}}{N_{\text{뿌리}}} \times GINI(D_{\text{청소년, 청년}}) + \frac{N_{\text{중년}}}{N_{\text{뿌리}}} GINI(D_{\text{중년}}) \right)$$

$$= 0.49 - \left\{ \left(\frac{4}{7} \times 0.5 \right) + \left(\frac{3}{7} \times 0.44 \right) \right\} = 0.02$$

번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
1	청소년	고소득층	아니오	좋음	미구매
2	청소년	고소득층	아니오	아주 좋음	미구매
3	청년	고소득층	아니오	좋음	구매
4	중년	중소득층	아니오	좋음	구매
5	중년	저소득층	예	좋음	구매
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매

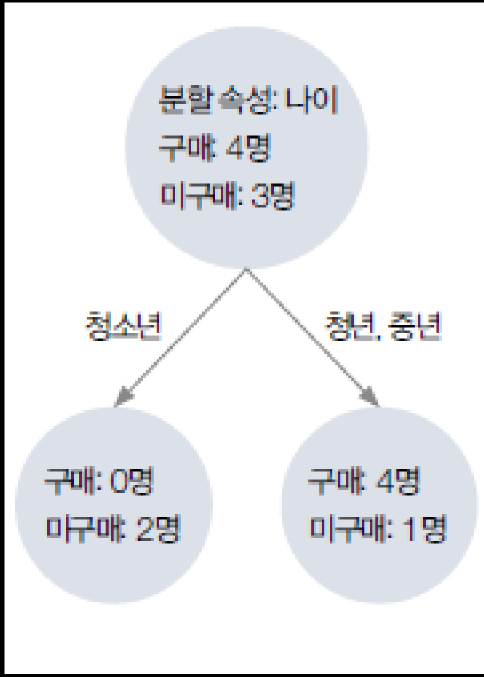


Gini Index Based on Age Attribute

- **Teen vs. Young Adult + Middle Aged**

$$= 0.49 - \left(\frac{2}{7} \times 0 + \frac{5}{7} \times 0.32\right) = 0.26$$

Select a Case with Higher Value



- **Teen + Young Adult vs. Middle Aged**

$$= 0.49 - \left(\frac{4}{7} \times 0.5 + \frac{3}{7} \times 0.44\right) = 0.02$$

Gini Index Based on Income Attribute

- **Split the node based on the Income attribute**
  - The age values are categorized as:
    - High = 0
    - Medium = 1
    - Low = 2

- **Applying CART algorithm, we typically perform binary splits**
  - We can try two types of binary splits
    - High vs. Medium + Low
    - High + Medium vs. Low

Income Attribute 1. High vs. Medium + Low

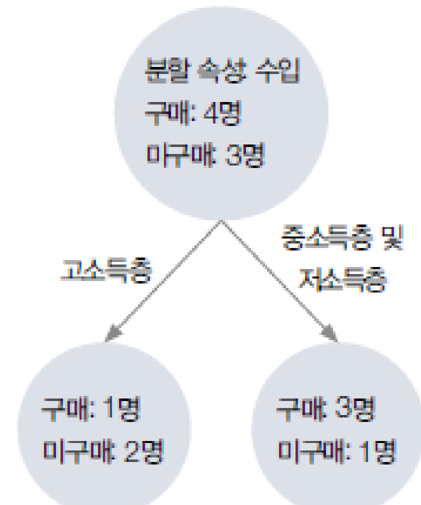
$$GINI(D_{고소득층}) = 1 - \left\{ \left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right\} = 0.44$$

$$GINI(D_{중,저소득층}) = 1 - \left\{ \left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2 \right\} = 0.375$$

$$\text{불순도 감소량} = GINI(D_{루트}) - \left(\frac{N_{고소득층}}{N_{루트}} \times GINI(D_{고소득층}) + \frac{N_{중,저소득층}}{N_{루트}} GINI(D_{중,저소득층}) \right)$$

$$= 0.49 - \left\{ \left(\frac{3}{7} \times 0.44\right) + \left(\frac{4}{7} \times 0.375\right) \right\} = 0.087$$

번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
1	청소년	고소득층	아니오	좋음	미구매
2	청소년	고소득층	아니오	아주 좋음	미구매
3	청년	고소득층	아니오	좋음	구매
4	중년	중소득층	아니오	좋음	구매
5	중년	저소득층	예	좋음	구매
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매



Income Attribute 2. High + Medium vs Low

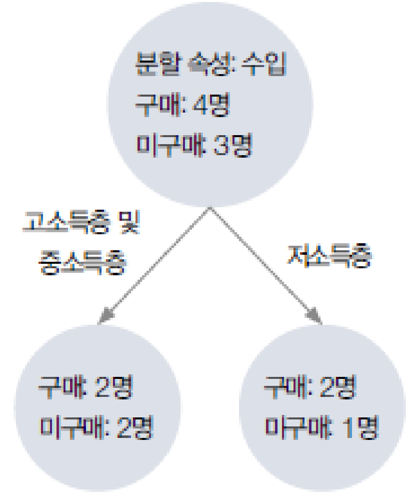
$$GINI(D_{\text{고,중소득층}}) = 1 - \left\{ \left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2 \right\} = 0.5$$

$$GINI(D_{\text{저소득층}}) = 1 - \left\{ \left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right\} = 0.44$$

$$\text{불순도 감소량} = GINI(D_{\text{뿌리}}) - \left(\frac{N_{\text{고,중소득층}}}{N_{\text{뿌리}}} \times GINI(D_{\text{고,중소득층}}) + \frac{N_{\text{저소득층}}}{N_{\text{뿌리}}} GINI(D_{\text{저소득층}}) \right)$$

$$= 0.49 - \left\{ \left(\frac{4}{7} \times 0.5 \right) + \left(\frac{3}{7} \times 0.44 \right) \right\} = 0.02$$

번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
1	청소년	고소득층	아니오	좋음	미구매
2	청소년	고소득층	아니오	아주 좋음	미구매
3	청년	고소득층	아니오	좋음	구매
4	중년	중소득층	아니오	좋음	구매
5	중년	저소득층	예	좋음	구매
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매



Gini Index Based on Income Attribute

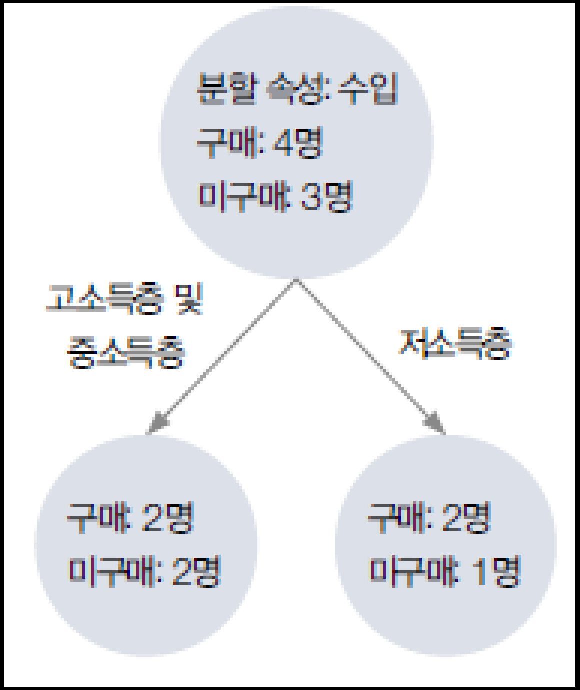
- **High vs. Medium + Low**

$$= 0.49 - \left(\frac{3}{7} \times 0.44 + \frac{4}{7} \times 0.375\right) = 0.087$$

Select a Case with Higher Value

- **High + Medium vs. Low**

$$= 0.49 - \left(\frac{4}{7} \times 0.5 + \frac{3}{7} \times 0.44\right) = 0.02$$



Gini Index Based on Student Attribute

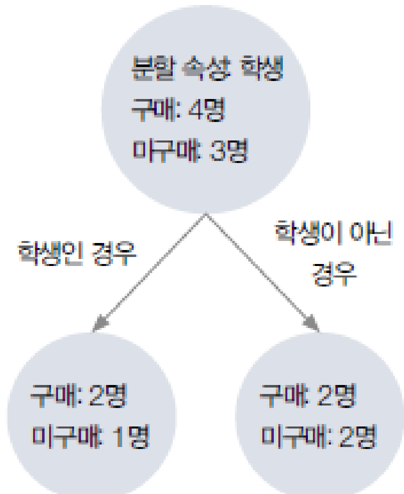
$$GINI(D_{\text{학생}}) = 1 - \left\{ \left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2 \right\} = 0.5$$

$$GINI(D_{\text{학생 아님}}) = 1 - \left\{ \left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right\} = 0.44$$

$$\text{불순도 감소량} = GINI(D_{\text{뿌리}}) - \left(\frac{N_{\text{학생}}}{N_{\text{뿌리}}} \times GINI(D_{\text{학생}}) + \frac{N_{\text{학생 아님}}}{N_{\text{뿌리}}} GINI(D_{\text{학생 아님}}) \right)$$

$$= 0.49 - \left\{ \left(\frac{4}{7} \times 0.5\right) + \left(\frac{3}{7} \times 0.44\right) \right\} = \boxed{0.02}$$

번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
1	청소년	고소득층	아니오	좋음	미구매
2	청소년	고소득층	아니오	아주 좋음	미구매
3	청년	고소득층	아니오	좋음	구매
4	중년	중소득층	아니오	좋음	구매
5	중년	저소득층	예	좋음	구매
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매



GINI Index Based on Credit Attribute

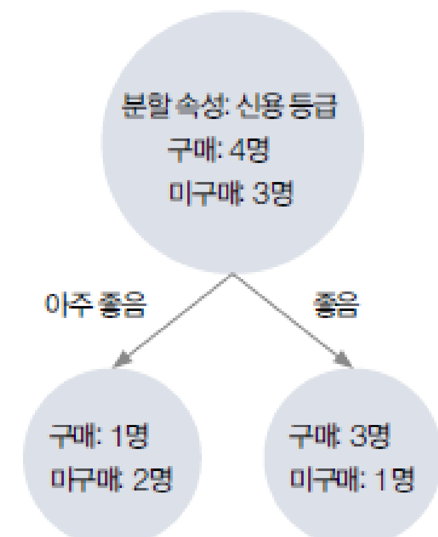
$$GINI(D_{\text{아주 좋음}}) = 1 - \left\{ \left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 \right\} = 0.44$$

$$GINI(D_{\text{좋음}}) = 1 - \left\{ \left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2 \right\} = 0.375$$

$$\text{불순도 감소량} = GINI(D_{\text{뿌리}}) - \left(\frac{N_{\text{아주 좋음}}}{N_{\text{뿌리}}} \times GINI(D_{\text{아주 좋음}}) + \frac{N_{\text{좋음}}}{N_{\text{뿌리}}} GINI(D_{\text{좋음}}) \right)$$

$$= 0.49 - \left\{ \left(\frac{3}{7} \times 0.44\right) + \left(\frac{4}{7} \times 0.375\right) \right\} = \boxed{0.087}$$

번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
1	청소년	고소득층	아니오	좋음	미구매
2	청소년	고소득층	아니오	아주 좋음	미구매
3	청년	고소득층	아니오	좋음	구매
4	중년	중소득층	아니오	좋음	구매
5	중년	저소득층	예	좋음	구매
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매



Split Root Node

■ **Choose the Attribute with the Greatest Impurity Reduction (Gini Gain)**

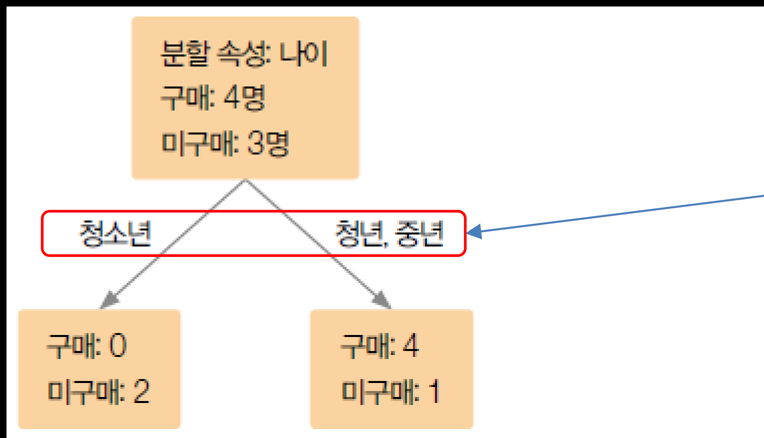


그림 8-16 나이 속성(청소년/청년 및 중년)을 통한 노드 분할

Attribute	Gini Gain
Age	0.26
Income	0.02
Student	0.02
Credit	0.087

- Second Split Decision (on Right Node): Young + Middle Group

$$GINI(D_{\text{청년, 중년}}) = 1 - \left\{ \left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2 \right\} = 0.32$$

Gini Index - 2nd split

Second node

- **Second Split Decision (on Right Node): Young + Middle Group**

번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
3	청년	고소득층	아니오	좋음	구매
4	중년	중소득층	아니오	좋음	구매
5	중년	저소득층	예	좋음	구매
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매

$$GINI(D_{\text{청년, 중년}}) = 1 - \left\{ \left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2 \right\} = 0.32$$

Age Attribute: Yong vs. Middle Aged

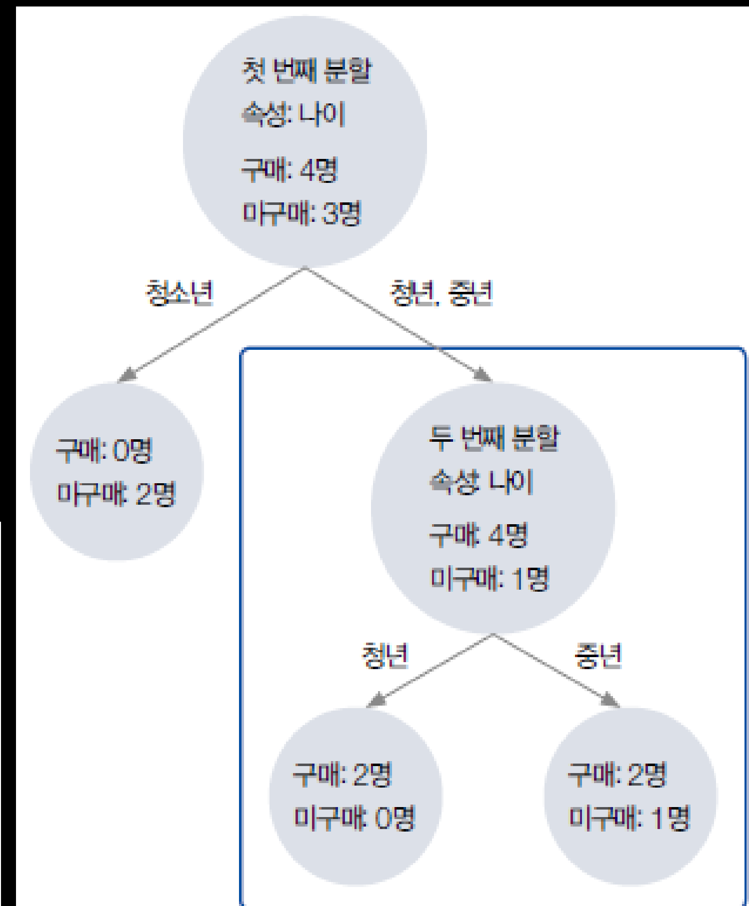
번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
3	청년	고소득층	아니오	좋음	구매
4	중년	중소득층	아니오	좋음	구매
5	중년	저소득층	예	좋음	구매
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매

$$GINI(D_{청년}) = 1 - \left\{ \left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2 \right\} = 0$$

$$GINI(D_{중년}) = 1 - \left\{ \left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 \right\} = 0.44$$

$$\text{불순도 감소량} = GINI(D_{부모}) - \left(\frac{N_{청년}}{N_{부모}} \times GINI(D_{청년}) + \frac{N_{중년}}{N_{부모}} GINI(D_{중년}) \right)$$

$$= 0.32 - \left\{ \left(\frac{2}{5} \times 0\right) + \left(\frac{3}{5} \times 0.44\right) \right\} = 0.056$$



Income Attribute 1. High vs. Medium + Low

번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
3	청년	고소득층	아니오	좋음	구매
4	중년	중소득층	아니오	좋음	구매
5	중년	저소득층	예	좋음	구매
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매

$$GINI(D_{\text{고소득층}}) = 1 - \left\{ \left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2 \right\} = 0$$

$$GINI(D_{\text{중, 저소득층}}) = 1 - \left\{ \left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2 \right\} = 0.375$$

$$\text{불순도 감소량} = GINI(D_{\text{부모}}) - \left(\frac{N_{\text{고소득층}}}{N_{\text{부모}}} \times GINI(D_{\text{고소득층}}) + \frac{N_{\text{중,저소득층}}}{N_{\text{부모}}} GINI(D_{\text{중,저소득층}}) \right)$$

$$= 0.32 - \left\{ \left(\frac{1}{5} \times 0 \right) + \left(\frac{4}{5} \times 0.375 \right) \right\} = 0.02$$

Income Attribute 2. High + Medium vs Low

번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
3	청년	고소득층	아니오	좋음	구매
4	중년	중소득층	아니오	좋음	구매
5	중년	저소득층	예	좋음	구매
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매

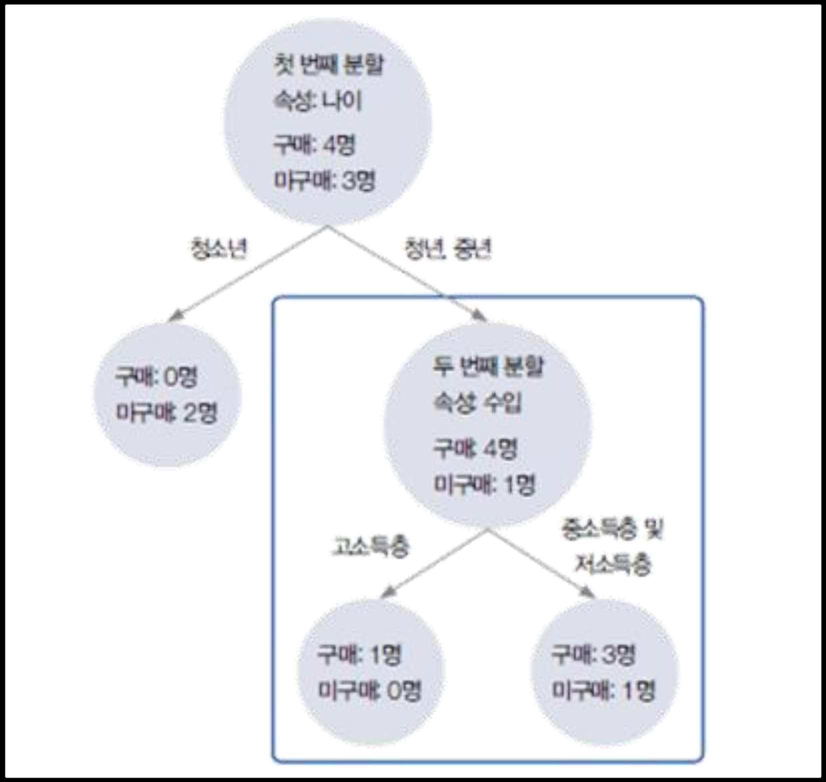
$$GINI(D_{\text{고,중소득층}}) = 1 - \left\{ \left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2 \right\} = 0$$

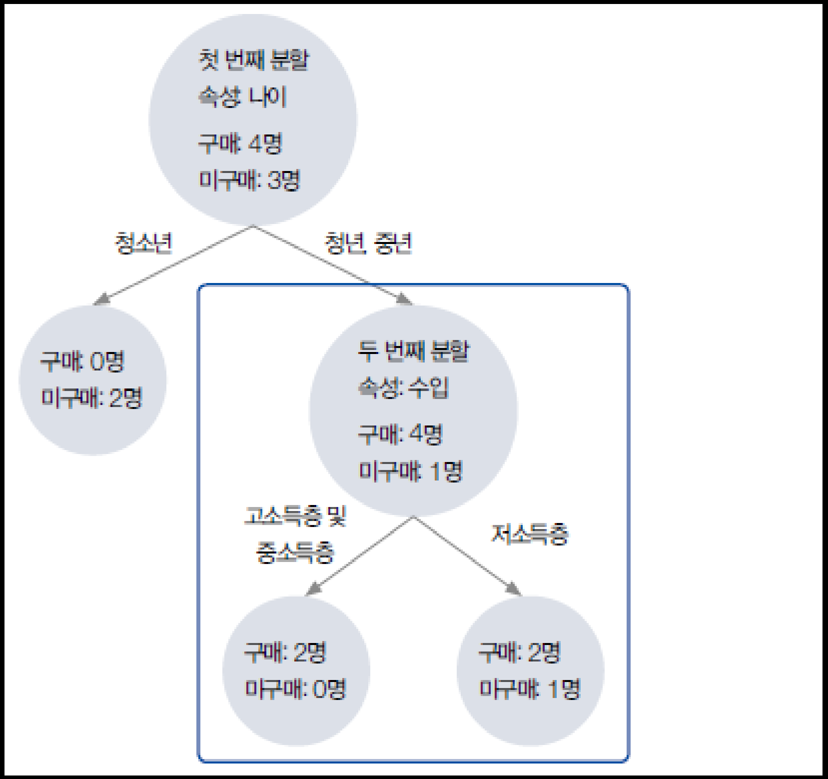
$$GINI(D_{\text{저소득층}}) = 1 - \left\{ \left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right\} = 0.44$$

$$\text{불순도 감소량} = GINI(D_{\text{부모}}) - \left(\frac{N_{\text{고,중소득층}}}{N_{\text{부모}}} \times GINI(D_{\text{고, 중소득층}}) + \frac{N_{\text{저소득층}}}{N_{\text{부모}}} GINI(D_{\text{저소득층}}) \right)$$

$$= 0.32 - \left\{ \left(\frac{2}{5} \times 0\right) + \left(\frac{3}{5} \times 0.44\right) \right\} = 0.056$$

Select Max GINI Gain in Income Attribute





GINI Gain: 0.02

GINI Gain: 0.056

GINI Index Based on Student Attribute

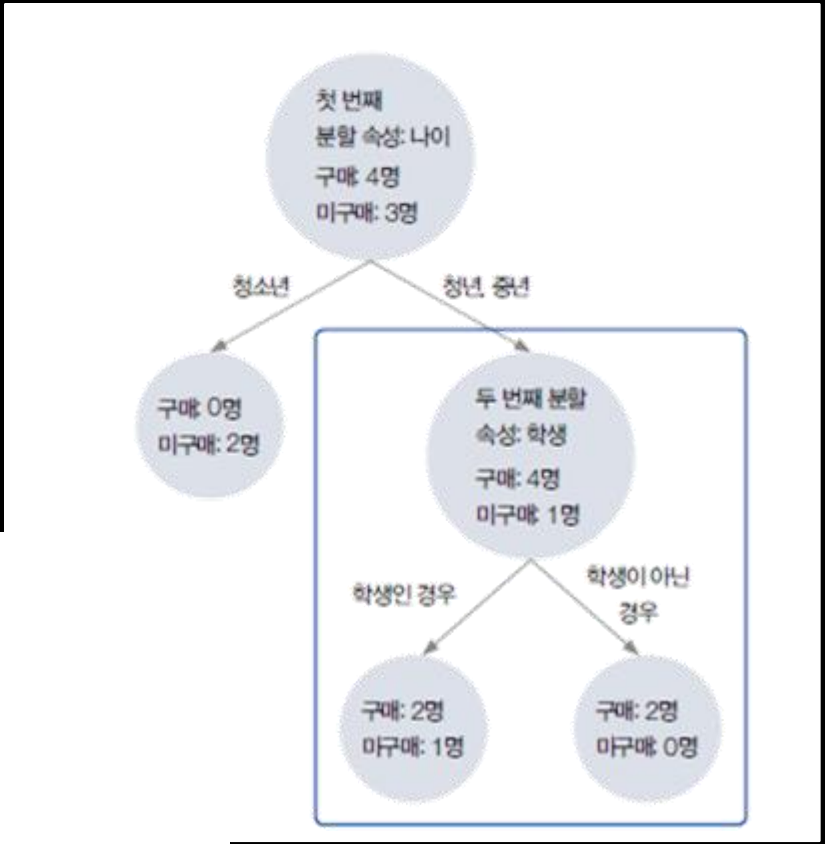
번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
3	청년	고소득층	아니오	좋음	구매
4	중년	중소득층	아니오	좋음	구매
5	중년	저소득층	예	좋음	구매
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매

$$GINI(D_{\text{학생}}) = 1 - \left\{ \left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right\} = 0.44$$

$$GINI(D_{\text{학생 아님}}) = 1 - \left\{ \left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2 \right\} = 0$$

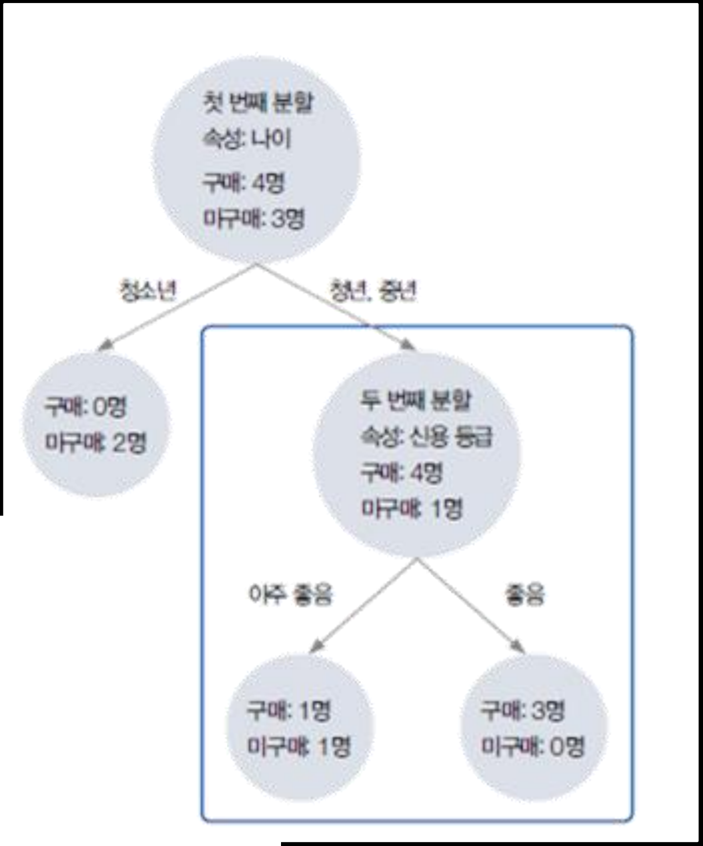
$$\text{불순도 감소량} = GINI(D_{\text{부모}}) - \left(\frac{N_{\text{학생}}}{N_{\text{부모}}} \times GINI(D_{\text{학생}}) + \frac{N_{\text{학생 아님}}}{N_{\text{부모}}} GINI(D_{\text{학생 아님}}) \right)$$

$$= 0.32 - \left\{ \left(\frac{3}{5} \times 0.44\right) + \left(\frac{2}{5} \times 0\right) \right\} = 0.056$$



GINI Index Based on Credit Attribute

번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
3	청년	고소득층	아니오	좋음	구매
4	중년	중소득층	아니오	좋음	구매
5	중년	저소득층	예	좋음	구매
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매



$$GINI(D_{\text{아주 좋음}}) = 1 - \left\{ \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right\} = 0.5$$

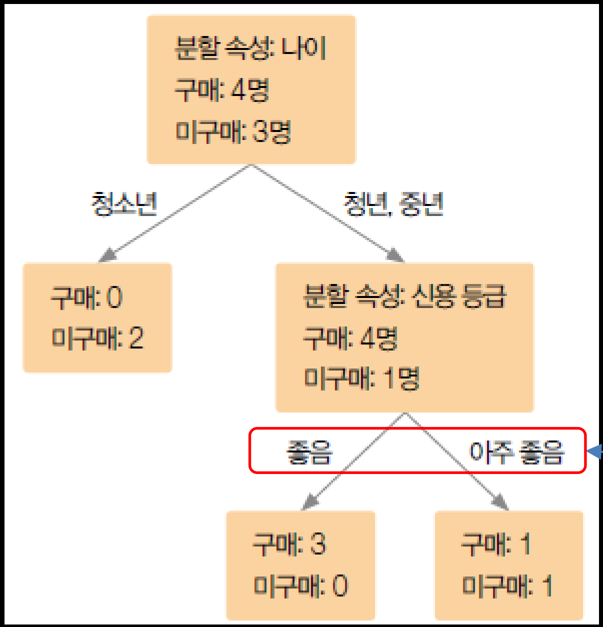
$$GINI(D_{\text{좋음}}) = 1 - \left\{ \left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2 \right\} = 0$$

$$\text{불순도 감소량} = GINI(D_{\text{부모}}) - \left(\frac{N_{\text{아주 좋음}}}{N_{\text{부모}}} \times GINI(D_{\text{아주 좋음}}) + \frac{N_{\text{좋음}}}{N_{\text{부모}}} GINI(D_{\text{좋음}}) \right)$$

$$= 0.32 - \left\{ \left(\frac{2}{5} \times 0.5 \right) + \left(\frac{3}{5} \times 0 \right) \right\} = 0.12$$

Split Node in Next Level

- **Choose the Attribute with the Greatest Impurity Reduction (Gini Gain)**



Attribute	Gini Gain
Age	0.056
Income	0.056
Student	0.056
Credit	0.12

Gini Index - 3rd split

GINI Index Based on Age Attribute

번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매

Third Split Attribute Decision:

Data where the individual is a Young Adult or Middle Aged and has an Excellent credit rating

$$GINI(D_{\text{아주 좋음}}) = 1 - \left\{ \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right\} = 0.5$$

GINI Index Based on Age Attribute

번호	나이(x_{i1})	수입(x_{i2})	학생 여부(x_{i3})	신용 등급(x_{i4})	구매 여부(y_i)
6	중년	저소득층	예	아주 좋음	미구매
7	청년	저소득층	예	아주 좋음	구매

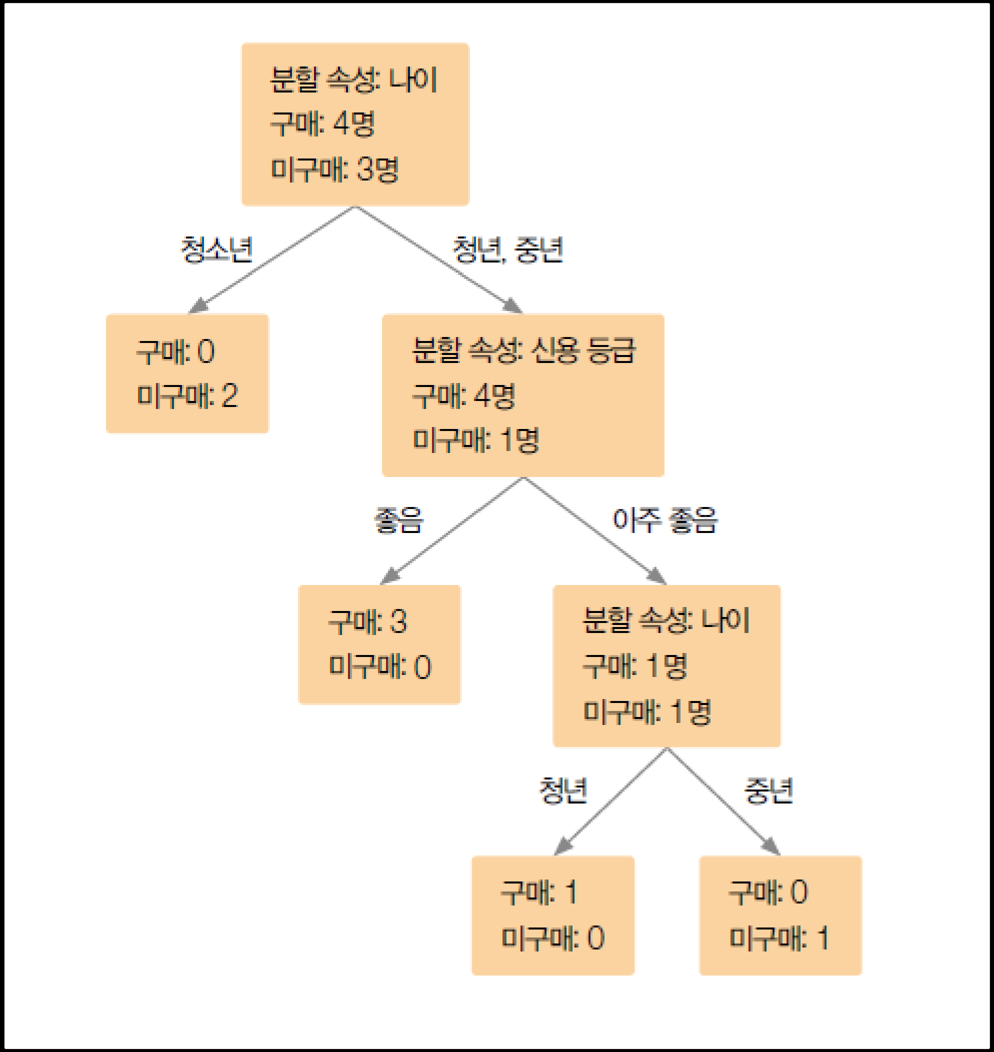
$$GINI(D_{중년}) = 1 - \left\{ \left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2 \right\} = 0$$

$$GINI(D_{청년}) = 1 - \left\{ \left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2 \right\} = 0$$

$$\text{불순도 감소량} = GINI(D_{부모}) - \left(\frac{N_{중년}}{N_{부모}} \times GINI(D_{중년}) + \frac{N_{청년}}{N_{부모}} GINI(D_{청년}) \right)$$

$$= 0.5 - \left\{ \left(\frac{1}{2} \times 0\right) + \left(\frac{1}{2} \times 0\right) \right\} = 0.5$$

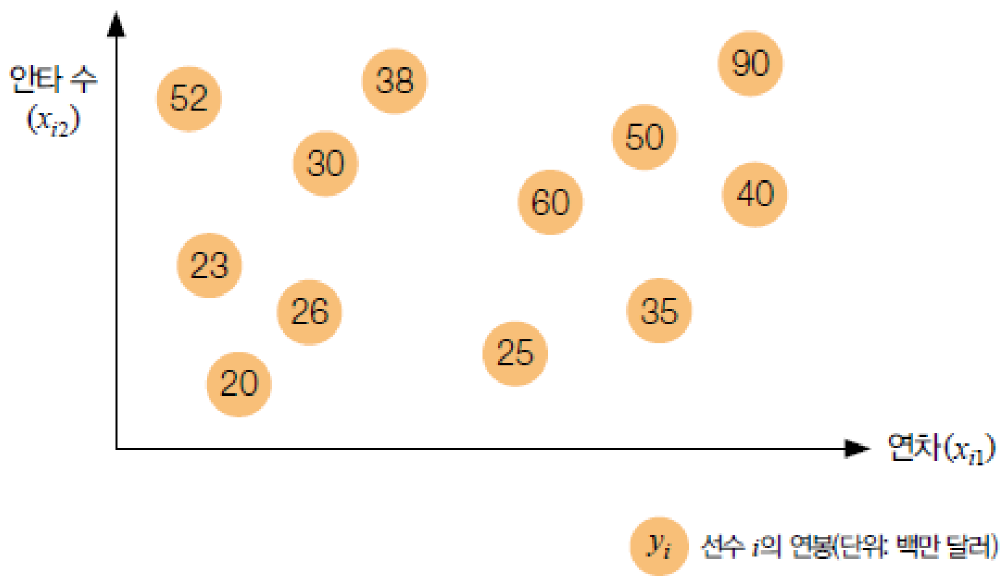
Final Decision Tree based on GINI Gain



Application of Decision Trees to Regression

Decision Trees to Regression

- **In regression problems, the leaf nodes of a decision tree can contain continuous predicted values**
- **Let's examine how to build a regression tree using the dataset below, following Step 1 to Step 3**



선수	연차	안타 수	연봉
1	2	80	52
2	3	40	23
3	4	20	20
4	6	65	30
5	5	30	26
6	8	85	38
7	11	25	25
8	12	60	60
9	13	75	50
10	13	30	35
11	15	90	60
12	15	60	40

Variance Reduction from Splitting (Split Score)

Variance of Dataset

$$VAR_{total} = \frac{1}{N_{total}} \sum_{i=1}^{N_{total}} (y_i - \bar{y})^2$$

- VAR_{total} : Variance of the entire dataset
- N_{total}: Total number of data points
- y_i: Target value of each data point
- $\bar{y}$: Mean of the target values

Variance Reduction from Splitting (Split Score)

$$Var\ Reduction = VAR_{total} - \left(\frac{N_{\leq}}{N} \times VAR_{\leq} + \frac{N_{>}}{N} \times VAR_{>}\right)$$

- N: Total number of Dataset
- $N_{\leq}$: Number of the data less than or equal to criteria
- $VAR_{\leq}$: Variance of the data less than or equal to criteria
- $N_{>}$: Number of the data greater than criteria
- $VAR_{>}$: Variance of the data greater than criteria

Calculating Variance Reduction Based on Age (Years)

Variation Reduction Example (Split at 2 years)

$$= VAR_{total} - \left(\frac{N_{\le}}{N} \times VAR_{\le} + \frac{N_{>}}{N} \times VAR_{>}\right) = 187.19 - \left(\frac{1}{12} \times 0 + \frac{11}{12} \times 185.4\right)$$

$$= 17.2$$

선수	연차	안타 수	연봉
1	2	80	52
2	3	40	23
3	4	20	20
4	6	65	30
5	5	30	26
6	8	85	38
7	11	25	25
8	12	60	60
9	13	75	50
10	13	30	35
11	15	90	60
12	15	60	40

기준		분산(기준 이하)	분산(기준 초과)	분산 감소량
연차	2	0.0	185.4	17.2
	3	210.2	182.4	0.1
	4	208.2	160.9	14.4
	5	162.2	151.7	32.0
	6	129.7	148.8	46.9
	8	116.6	166.7	45.6
	11	105.1	104.0	82.5
	12	186.7	92.21	32.0
	13	171.5	100.0	27.6

Max Var. Reduction

Calculating Variance Reduction Based on Age (Years)

Variation Reduction Example (Split at 40 battings)

$$= VAR_{total} - \left(\frac{N_{\le}}{N} \times VAR_{\le} + \frac{N_{>}}{N} \times VAR_{>}\right) = 187.19 - \left(\frac{5}{12} \times 29.25 + \frac{7}{12} \times 162.60\right)$$

$$= 69.03$$

선수	연차	안타 수	연봉
1	2	80	52
2	3	40	23
3	4	20	20
4	6	65	30
5	5	30	26
6	8	85	38
7	11	25	25
8	12	60	60
9	13	75	50
10	13	30	35
11	15	90	60
12	15	60	40

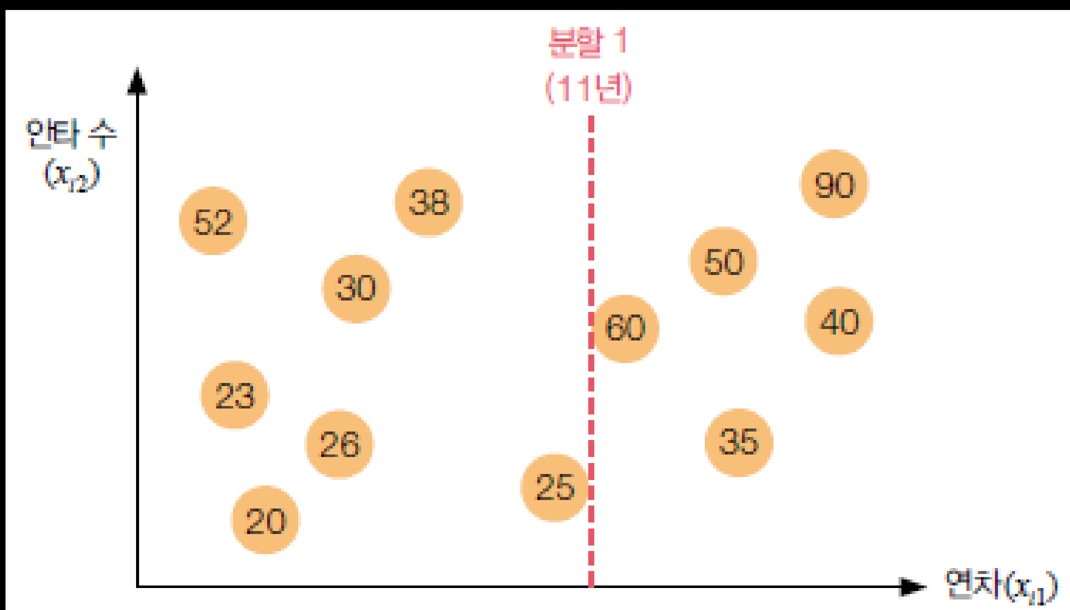
기준		분산(기준 이하)	분산(기준 초과)	분산 감소량
안타 수	20	0,0	171,1	30,28
	25	6,25	163,84	49,61
	30	29,25	162,84	49,61
	40	29,25	162,60	69,03
	60	166,20	113,6	42,09
	65	146,2	62,0	69,03
	75	160,66	82,66	46,02
	80	172,69	121,0	23,11
	85	157,28	0,0	43,0

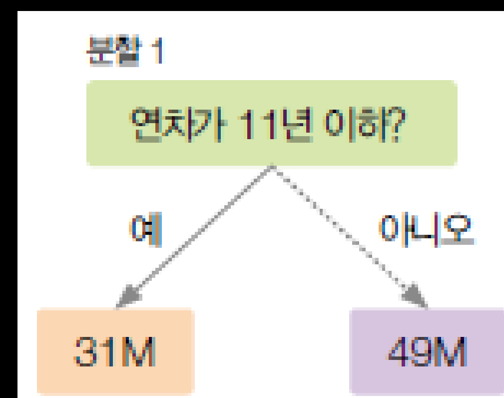
Max Var. Reduction

Splitting Tree to Regression (1/3)

■ First Split Based on Maximum Variance Reduction

- Splitting by Years (x_1 = 11 years)
  - Left Node: ≤ 11 years → Mean: 31M
  - Right Node: >11 years → Mean: 40M





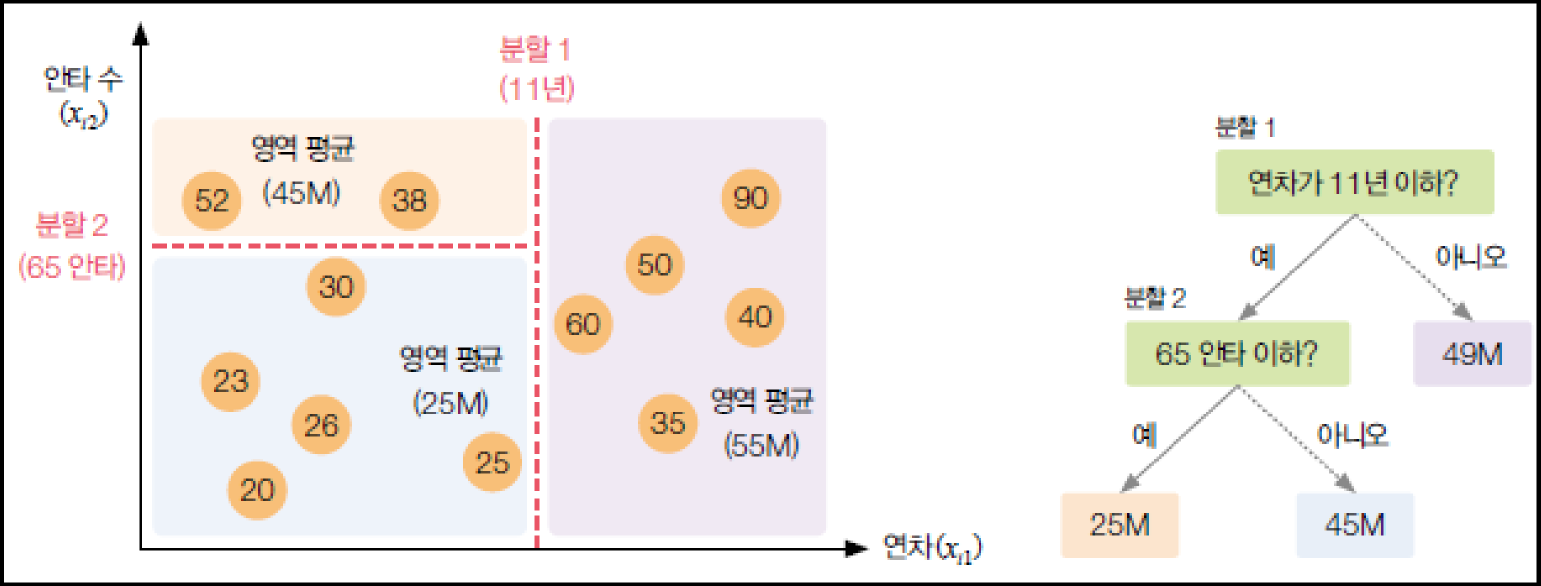
Splitting Tree to Regression (2/3)

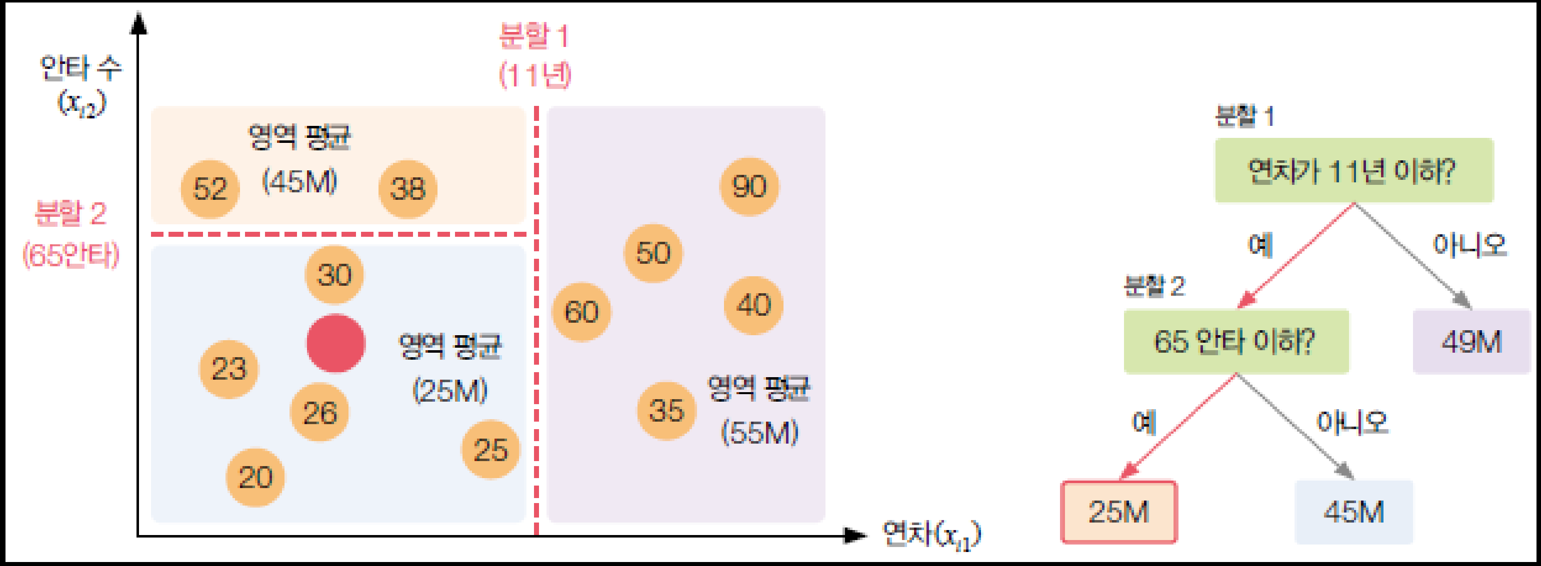
- **Second Split Point Based on Remaining Subset (≤ 11 years only)**

기준		분산(기준 이하)	분산(기준 초과)	분산 감소량
안타 수 (연차 11년 이하)	20	0	100.9	18.62
	25	6.25	108.16	26.05
	30	6.88	116.18	35.75
	40	2.25	82.66	66.67
	65	10.95	49.0	83.27
	80	111.88	0.0	9.19

Max Var. Reduction

Splitting Tree to Regression (3/3)





Implementing Decision Trees in Code

Implementing Decision Trees in Code

- **Use codes from Prof.**
- **Alternatively, practice codes from Textbook (GitHub repository)**
  - https://github.com/KMA-AIData/ML
  - Notebook File:
    - Import necessary libraries and packages
    - Load the dataset
    - Preprocess the data
    - Build a decision tree model
    - Train the model
    - Evaluate model performance on the test set
    - Visualize model performance using scatter
    - Visualize tree structure using graphviz

18_CH08_실습_의사결정나무.ipynb

수고하셨습니다 ..^^..
Thank you!