# 데이터 과학

## L14: Decision Trees

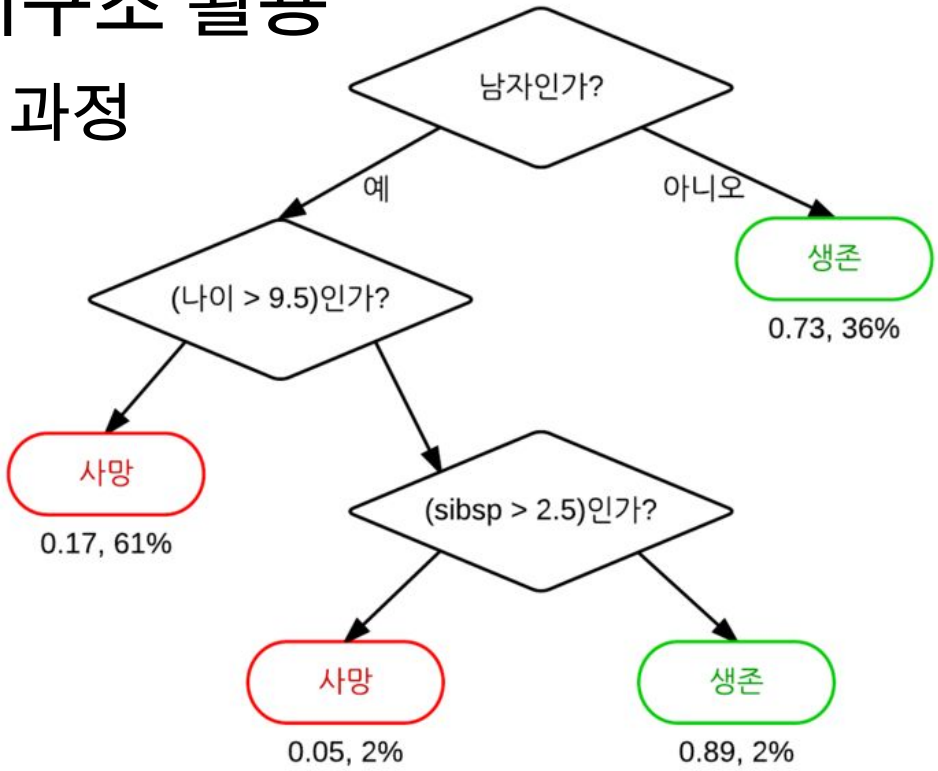
Kookmin University

# 목차

- **Decision Trees**
- Decision Tree 만들기
- Types of Questions
- Pros/Cons of Decision Trees

# Decision Trees

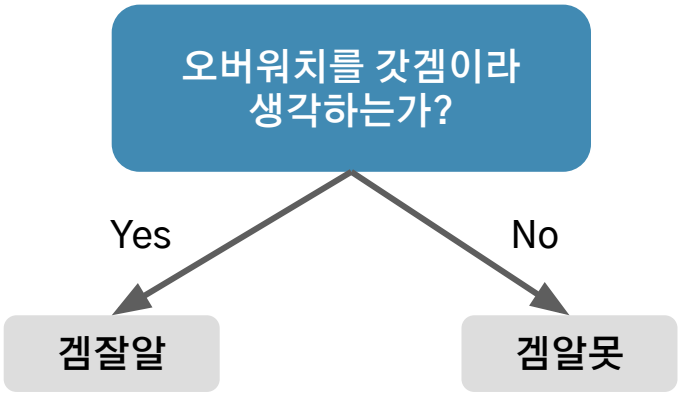
- 비모수적 지도학습 방법
- 결과를 예측하는데 트리구조 활용
  - 스무고개놀이와 유사한 과정
- 분류 트리, 회귀 트리

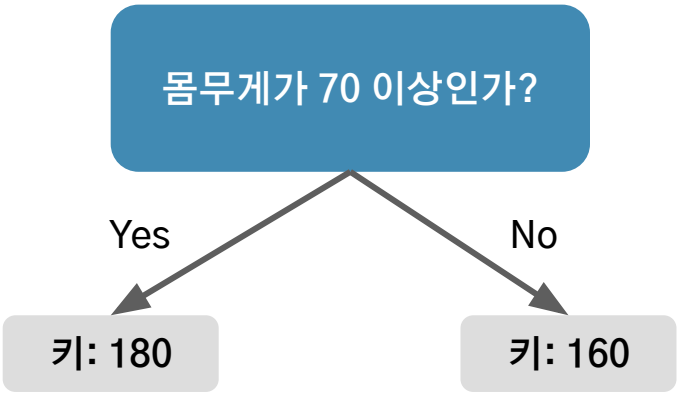



타이타닉호 탑승객의 생존 여부를 나타내는 결정 트리
출처: 위키백과 – 결정 트리 학습법

# Types of Decision Trees

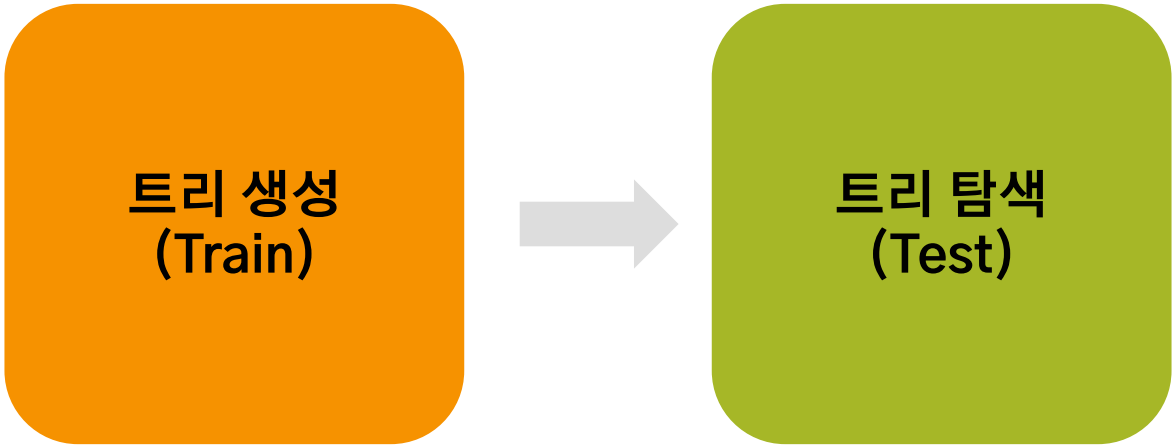
- Classification Trees
  - 예측 결과 = 분류
- Regression Trees
  - 예측 결과 = 실수 값



# Decision Tree Learning

- 좋은 트리를 생성하기위해 어떤 질문을 해야할까?




Decision Tree Learning Process

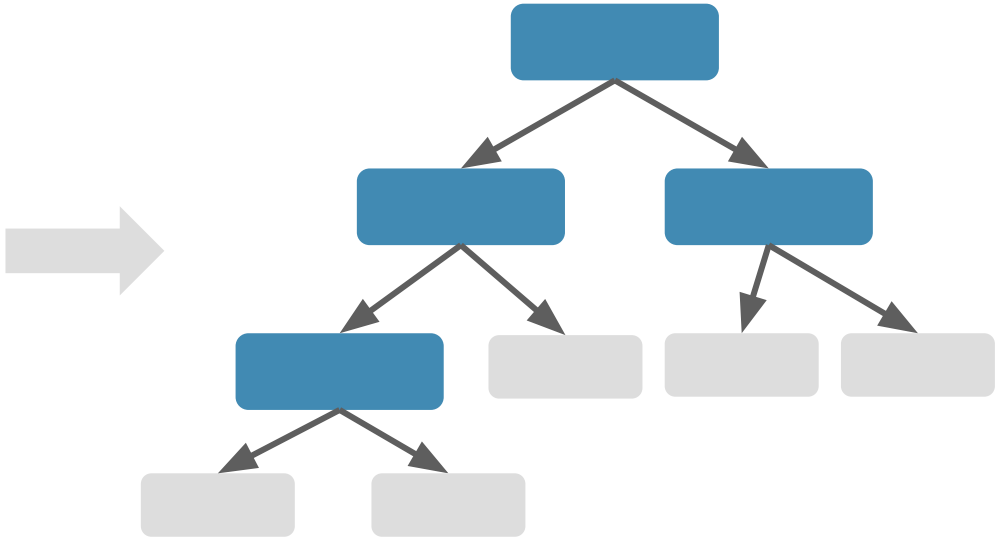
# 목차

- Decision Trees
- **Decision Tree 만들기**
- Types of Questions
- Pros/Cons of Decision Trees

# Decision Tree 만들기

- 데이터로부터 Decision Tree를 만드는 방법?
  - 컬럼 중 하나를 골라 질문하기
  - 남은 데이터에서 다른 컬럼 골라 질문하기
  - …

| 수업? | 과제? | 비? | 커피? |
|---|---|---|---|
| X | X | X | X |
| O | O | X | X |
| O | O | O | O |
| O | X | ? | O |
| … | … | … | … |



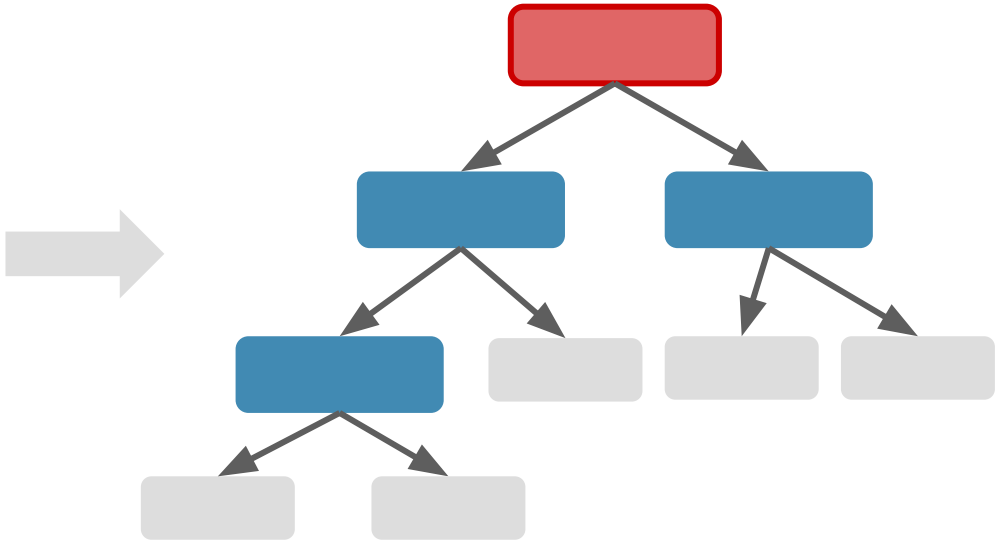
데이터: 국민대 학생은 어떤 경우에 커피를 마실까?

# Decision Tree 만들기

- 수업, 과제, 비 중에서 어떤 것을 가장 처음 질문해야 좋을까? → 가장 분류를 잘하는 것!

| 수업? | 과제? | 비? | 커피? |
|---|---|---|---|
| X | X | X | X |
| O | O | X | X |
| O | O | O | O |
| O | X | ? | O |
| … | … | … | … |

데이터: 국민대 학생은 어떤 경우에 커피를 마실까?

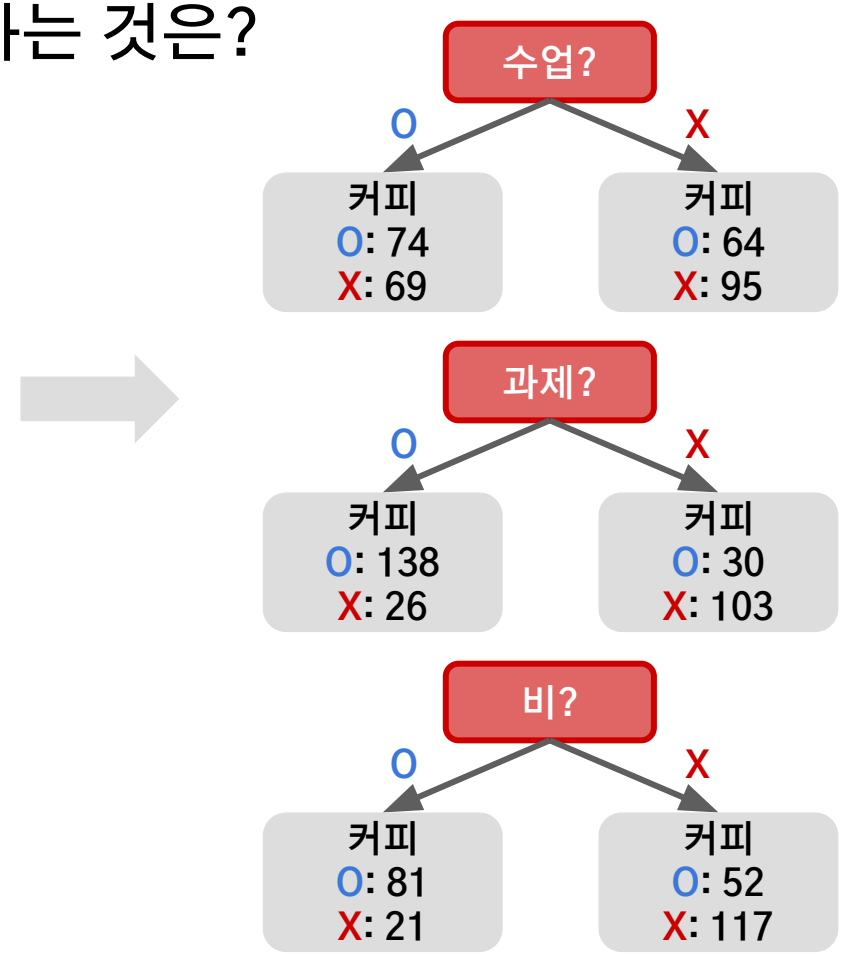


# Decision Tree 만들기

- 수업, 과제, 비 질문에 따른 커피 섭취 여부 분포 확인
  - 셋 중 가장 분류를 잘하는 것은?

| 수업? | 과제? | 비? | 커피? |
|---|---|---|---|
| X | X | X | X |
| O | O | X | X |
| O | O | O | O |
| O | X | ? | O |
| … | … | … | … |

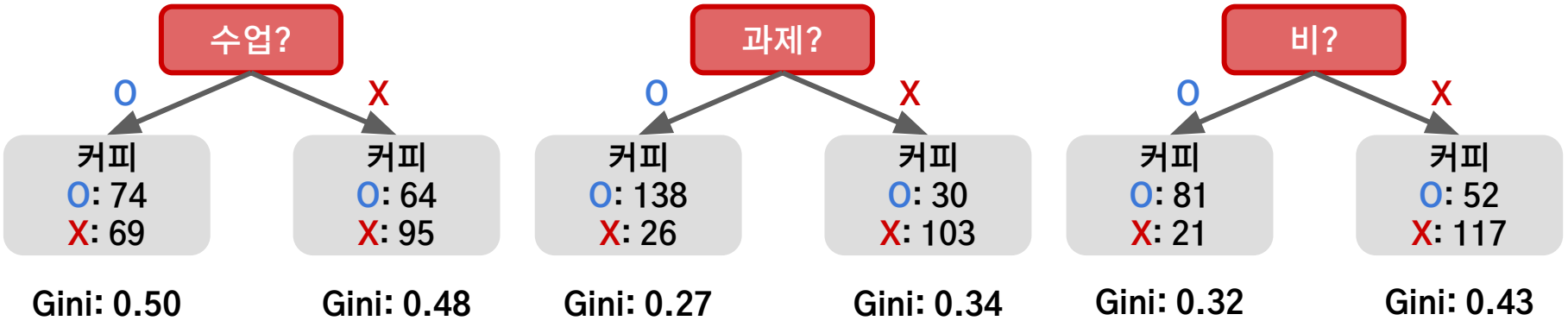
데이터: 국민대 학생은 어떤 경우에 커피를 마실까?



# Impurity

- 가장 분류를 잘한다 → 분류 결과의 불순도가 낮다
- Gini impurity: 불순함의 정도를 측정

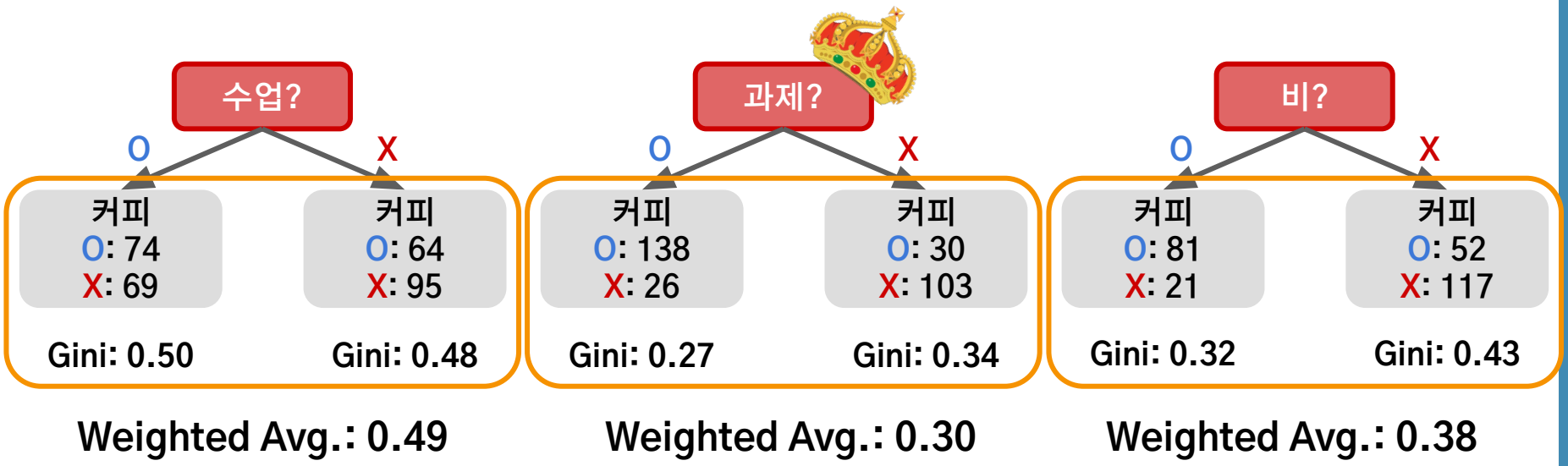
$$I_G(p) = \sum_{i \in C} p_i(1 - p_i) = 1 - \sum_{i \in C} {p_i}^2$$




참고: impurity의 다른 척도로는 entropy도 있음

# Impurity

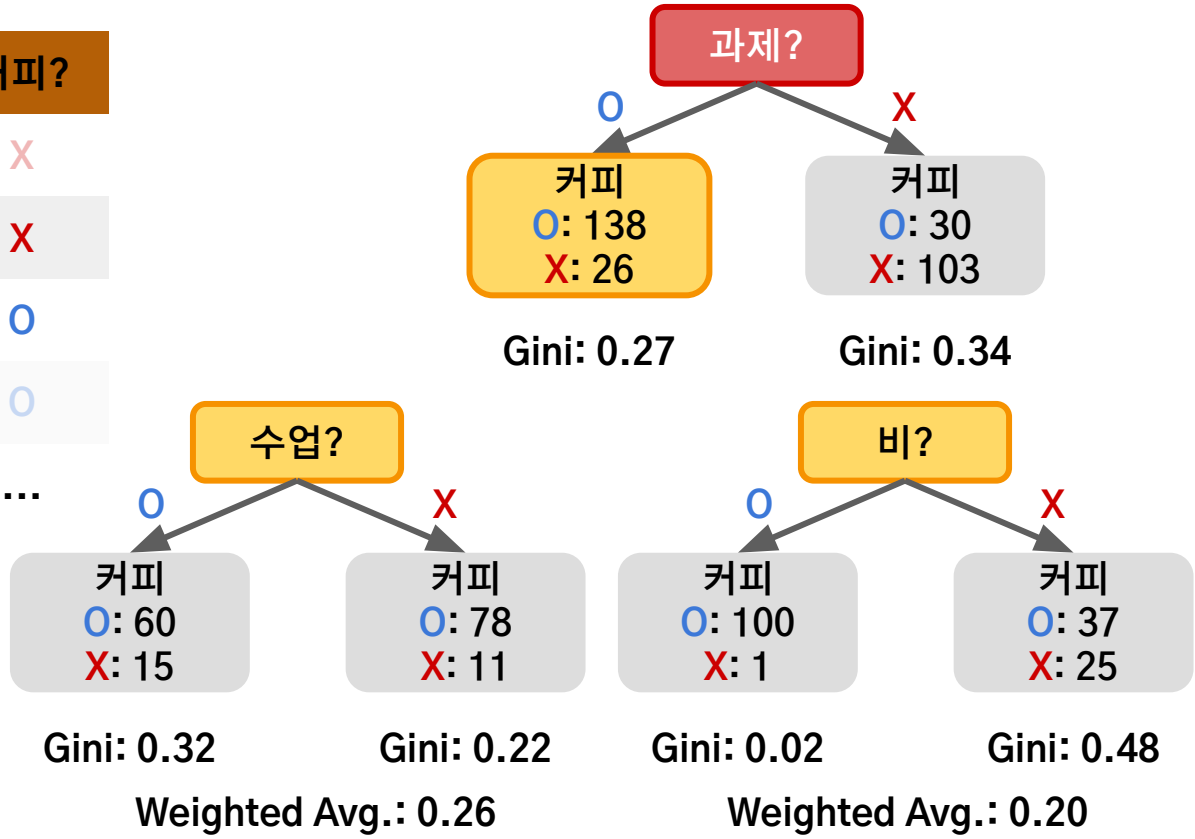
- 수업, 과제, 비 중에서 가장 나은 것은?
  - Gini Impurity의 **가중 평균**으로 비교



# Decision Tree 만들기

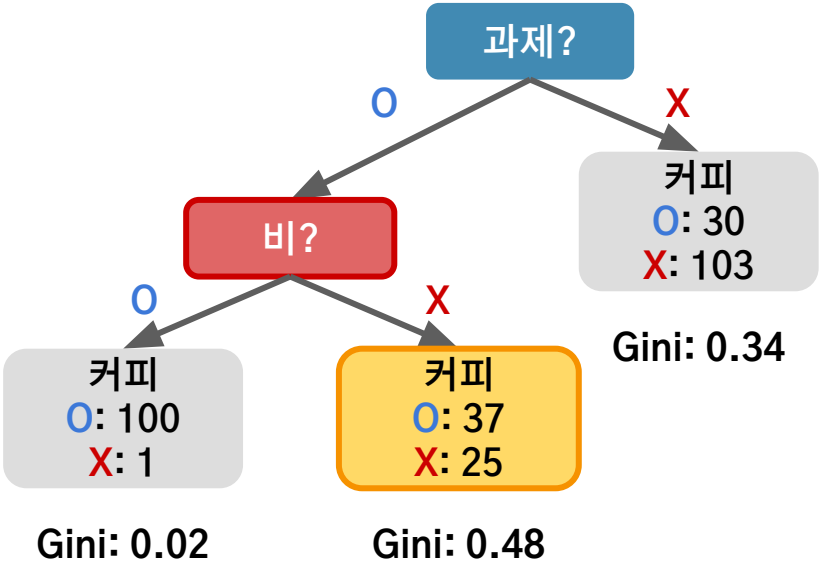
- Leaf 노드에서 같은 과정 반복
  - 예) 수업, 비 중에서 어떤걸로 질문할까?

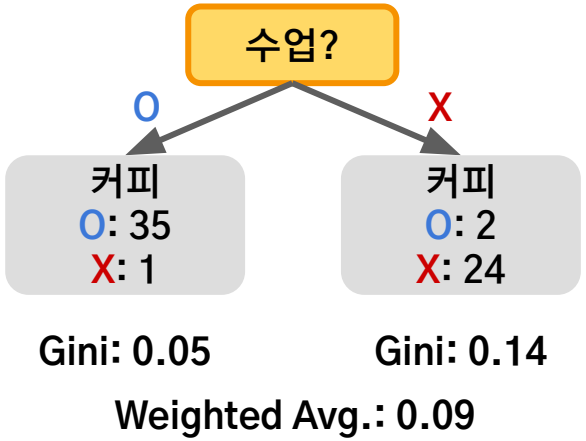
| 수업? | 과제? | 비? | 커피? |
|---|---|---|---|
| X | X | X | X |
| O | O | X | X |
| O | O | O | O |
| O | X | ? | O |
| … | … | … | … |



# Decision Tree 만들기

- Leaf 노드에서 같은 과정 반복



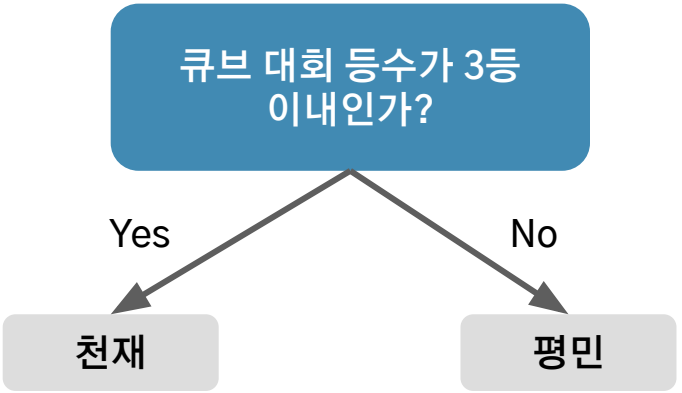



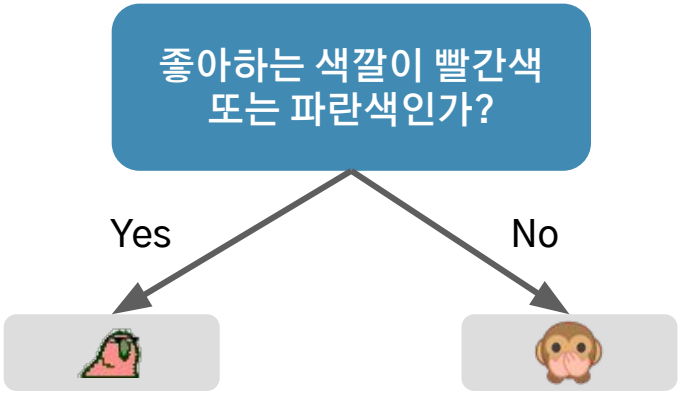
# 목차

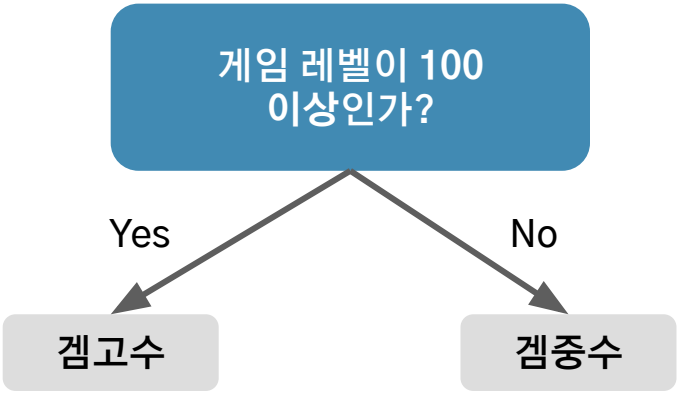
- Decision Trees
- Decision Tree 만들기
- **Types of Questions**
- Pros/Cons of Decision Trees

# Types of Questions

- Yes/No
- Numeric data
- Rank data
- Categorical data



# Numeric Data

- 수치 데이터에서 나누는 지점을 선택하는 방법?
  - 정렬 후 사이 값들에 대해 Gini Impurity 계산 및 비교

| 수면시간? | 커피? |
|---|---|
| 7.5 | O |
| 5.9 | X |
| 6.5 | X |
| 4.1 | O |
| 2.3 | O |
| 9.1 | X |

| 수면시간? | 커피? |
|---|---|
| 2.3 | O |
| 3.2 → Gini: 0.40 | |
| 4.1 | O |
| 5.0 → Gini: 0.25 | |
| 5.9 | X |
| 6.2 → Gini: 0.44 | |
| 6.5 | X |
| 7.0 → Gini: 0.50 | |
| 7.5 | O |
| 8.3 → Gini: 0.40 | |
| 9.1 | X |

# Rank Data

- 랭킹 데이터 에서 나누는 지점을 선택하는 방법?
  - 1등 부터 차례대로 Gini Impurity 계산 및 비교

| 석차 | 커피? |
|---|---|
| 1 | O |
| 3 | X |
| 5 | X |
| 2 | O |
| 4 | O |
| 6 | O |

**1등 이상? Gini: 0.40**

**2등 이상? Gini: 0.33**

**3등 이상? Gini: 0.44**

**4등 이상? Gini: 0.42**

**5등 이상? Gini: 0.40**

# Categorical Data

- 여러 선택지가 있는 데이터에서 나누는 지점을 선택하는 방법?
  - 모든 category 조합에 대해 비교

| 전공 | 커피? |
|---|---|
| 소융 | O |
| 소융 | X |
| 경영 | X |
| 자동차 | O |
| 소융 | O |
| 경영 | X |

**소융? Gini: 0.44**

**경영? Gini: 0.25**

**자동차? Gini: 0.40**

**소융 or 경영? Gini: 0.4**

**소융 or 자동차? Gini: 0.25**

**경영 or 자동차? Gini: 0.44**

# 목차

- Decision Trees
- Decision Tree 만들기
- Types of Questions
- **Pros/Cons of Decision Trees**

# Pros/Cons of Decision Trees

- 장점
  - 빠른 모델 구축 및 분류 속도
  - 데이터 가공의 필요성이 적음
  - 직관적인 모델 구축 원리
  - 예측 결과의 해석 및 이해가 쉬움
    - Tree의 Path를 통해 분류된 이유 설명 가능
- 단점/한계
  - 오버피팅
  - 약간의 데이터 차이로 트리 구조가 바뀔 수 있음
  - 계층적 구조로 인한 에러 전파

# Questions?