# Module IV

## Linear Regression & Decision Tree

아이디케이스퀘어드 양승준 / sidney.yang@idk2.co.kr
https://www.heartcount.io

# 데이터 분석 주요기술

과거

Looking Backward

Looking Forward

미래

- **DESCRIBE (기술 분석) - 엑셀, 대쉬보드**
  - 데이터 특성과 모양을 요약
- **EXPLORE (탐험적 분석) - 데이터 시각화 도구**
  - 가설수립·데이터 감 잡기 위해 패턴 탐험
- **PREDICT/INFER (예측·추론 분석) - 통계/ML**
  - 패턴(모형)을 통해 주어진 문제를 예측·설명

# 선형회귀분석 (Linear Regression Analysis)

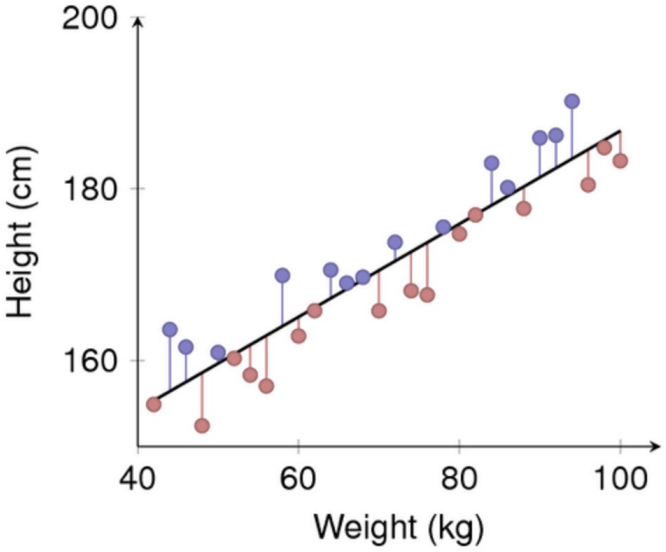
**Supervised Machine-Learning**

**Regression Model:** Y가 숫자형 변수(매출)인 경우

**Classification Model:** Y가 범주형 변수(성별)인 경우

## Linear Regression

- 가장 오래되고, 널리 쓰이고, 결과를 이해하기 쉬운 지도학습 알고리즘
- 독립변수(X)를 가지고 숫자형 종속변수(Y)를 가장 잘 설명·예측**(Best Fit)**하는 선형 관계(Linear Relationship)를 찾는 방법 중 하나
- X가 범주형 변수(성별)인 경우, 집단(남·녀) 간 Y값의 차이를 분석




### 계산방법 (Least Squares)

X와 Y 사이에 선형적 관계가 있다는 가정 하에 실제 Y값과 예측한 Y값의 차이를 최소화하는 방정식을 계산

$$Y = b_0 + b_1X + \text{error}$$

- $b_0$ : Y축 절편(Intercept); 예측변수가 0일 때 기대 점수를 나타냄
- $b_1$ : 기울기로 X가 한 단위 증가했을 때의 Y의 평균적 변화값을 나타냄

*참고: http://students.brown.edu/seeing-theory/regression/index.html

# 선형회귀분석 (Linear Regression Analysis)

## P-Value (Probability-Values)

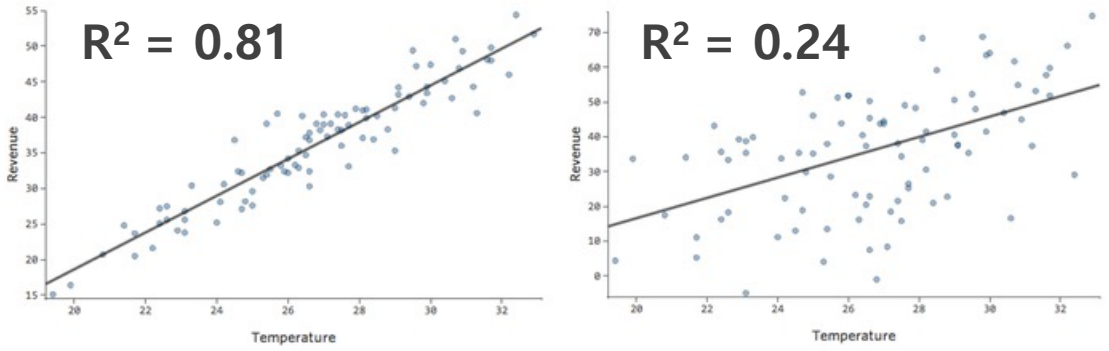
**Q. X와 Y 사이에 통계적으로 유의미한 관계가 있나?**

- Statistical Significance (통계적 유의성)
- 데이터를 통해 확인한 관계가 우연히 나왔을 확률
- P값이 0.03: 데이터에서 발견한 관계가 운일 확률 3%
- 관계의 세기(Size of an Effect)를 나타내는 것은 아님

## $R^2$ (R-SQUARED; 결정계수)

**Q. X가 Y를 얼마나 잘 설명/예측하는가?**

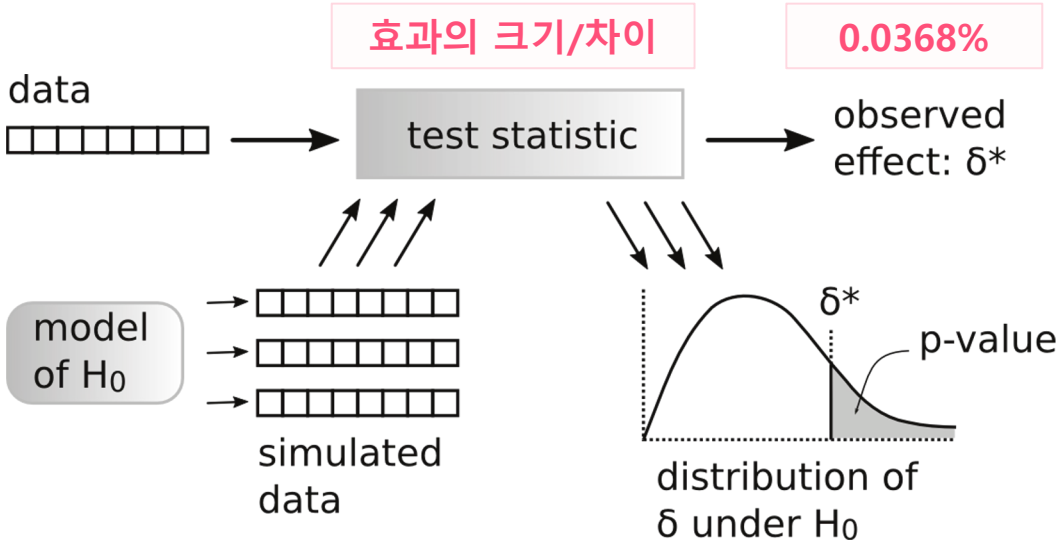
- Goodness of Fit: X로 설명할 수 있는 Y 변화량의 크기

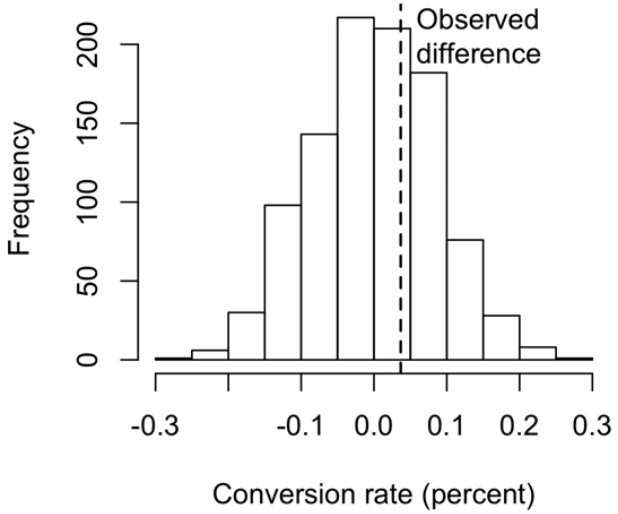


# Statistical Significance and P-Value(Probability-Values)

| Outcome | Campaign A | Campaign B |
|---|---|---|
| Conversion | 200 | 182 |
| No Conversion | 23539 | 22406 |
| Conversion Rate | 0.8425% | 0.8057% |

1. 캠페인 A의 전환률이 0.0368% 높음; 차이가 의미가 있나?
2. 통계적 유의성 검증이 꼭 필요한가? (이 정도면 작은 샘플을 사용하여 일반화하는 일을 걱정할 필요없는 빅데이터 아닌가?)
3. **통계적 유의성: 두 캠페인 사이의 전환율 차이가 우연은 아닌가?**
   - 둘 간 전환율 차이가 없다고 가정($H_0$)하고 두 캠페인 결과를 하나로 섞는다
   - 섞은 데이터에서 23739, 22588개를 Resampling하여 두 캠페인 간 전환율 차이를 기록 (1,000번 반복)
   - 전환율 차이(test statistic; 검정통계량)가 > 0.0368%보다 큰 경우의 확률을 계산: 30.8% (P-Value: 0.308)
   - 두 캠페인 간 실제로 전환률 차이가 없는 경우에도 우연의 작용으로 0.0368% 차이 정도는 흔하게 발생.

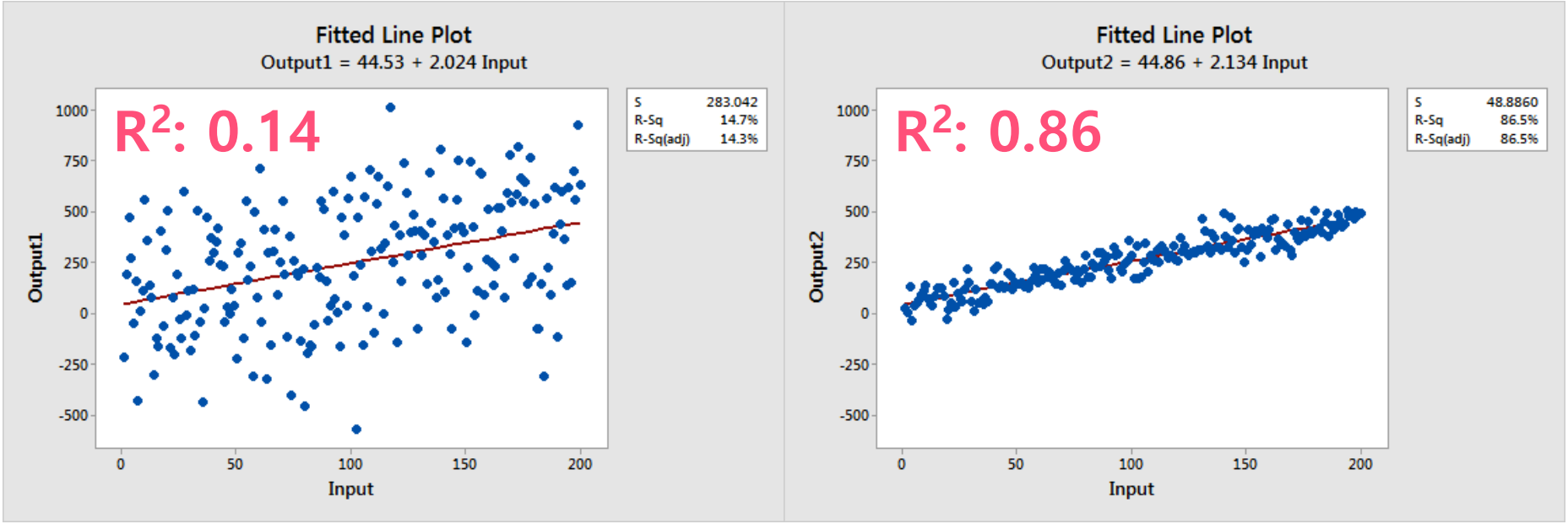


# 선형회귀분석: 결정계수($R^2$: R-SQUARED)

## 낮은 결정계수가 반드시 나쁜 (Inherently Bad) 것은 아님

- 동일한 회귀방정식: Y = 44 + 2*X; P < 0.001
- 우측 모형이 좌측 모형보다 예측 정확도($R^2$)는 매우 높음
- 변수 간 경향성은 동일: X 1단위 증가 → Y 2단위 증가



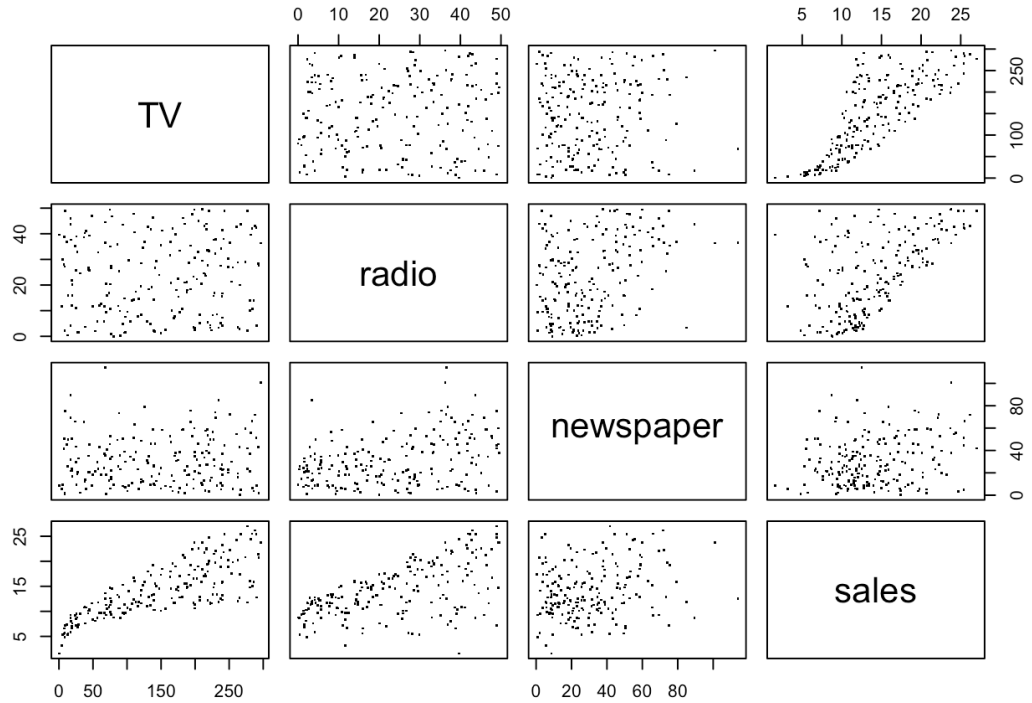

**Y값을 정확히 예측하기 위해선 $R^2$ 값이 중요**
**하지만, 경향성 정보가 중요한 경우 $R^2$ 가 낮다고 꼭 나쁜 모형은 아님**

# Simple Linear Regression Analysis – Advertisement

매출에 미치는 매체 영향에 상호작용이 없다고 가정하였을 때
TV, Radio, 신문 중 Sales 증가에 가장 효과적인 광고매체는?

강의 보조자료 참고(클릭)

Advertising.xlsx



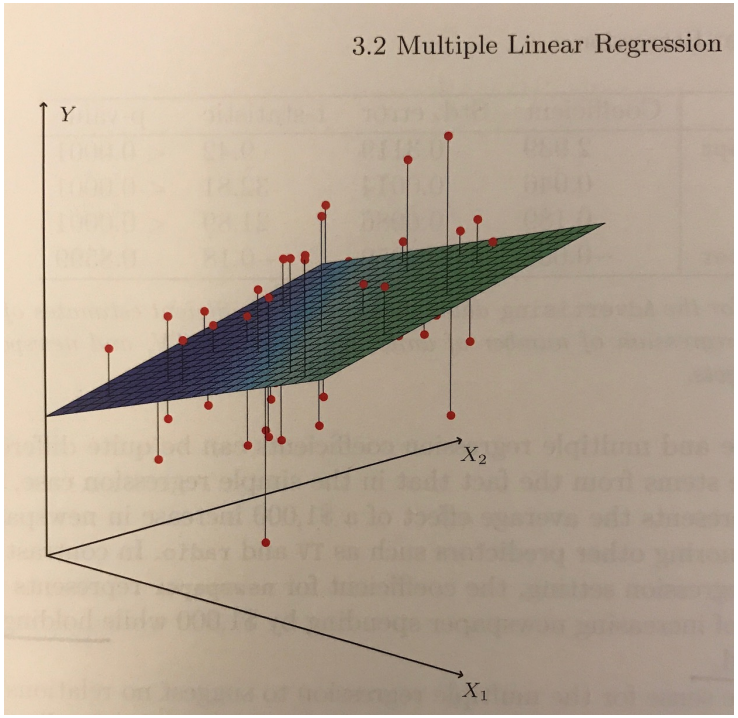
# Multiple Linear Regression Analysis – Advertisement

이번에는 변수 2개[TV, Radio]를 사용하여 Sales와의 관계를
설명·예측하는 회귀모형을 만들어 봅시다.

강의 보조자료 참고(클릭)

$$Y = b_0 + b_1X_1 + b_2X_2$$

$$\text{Sales} = \mathbf{2.9} + \mathbf{0.045} \times \text{TV} + \mathbf{0.187} \times \text{Radio}$$




SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.94720339 |
| R Square | 0.897194261 |
| Adjusted R Square | 0.896150548 |
| Standard Error | 1.681360913 |
| Observations | 200 |

ANOVA

| | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Regression | 2 | 4860.2348 | 2430.1174 | 859.6177183 |
| Residual | 197 | 556.91398 | 2.8269745 | |
| Total | 199 | 5417.1488 | | |

| | *Coefficients* | *tandard Erro* | *t Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | 2.921099912 | 0.2944897 | 9.9191929 | 4.56556E-19 |
| X Variable 1 | 0.045754815 | 0.0013904 | 32.908708 | 5.43698E-82 |
| X Variable 2 | 0.187994227 | 0.00804 | 23.382446 | 9.77697E-59 |

# Decision Tree and Classification

**Supervised Machine-Learning**

**Regression Model:** Y가 숫자형 변수(매출)인 경우

**Classification Model:** Y가 범주형 변수(성별)인 경우

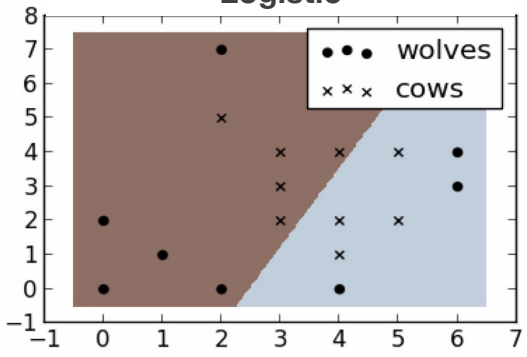
- **Decision Tree:** 의사결정트리; 대표적 Classification Model
- **Classification:** 서로 다른 집단을 구분하는 규칙(경계) 찾기

## Classification: 집단을 구분하는 경계 찾기

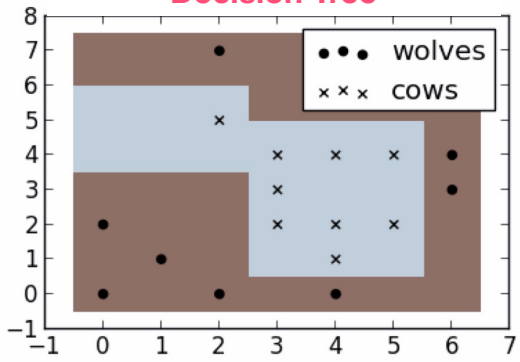
**Linear Classifier**

**Non-Linear Classifier**

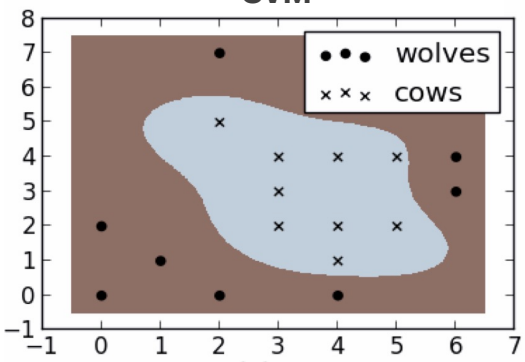
Logistic




Decision Tree




SVM

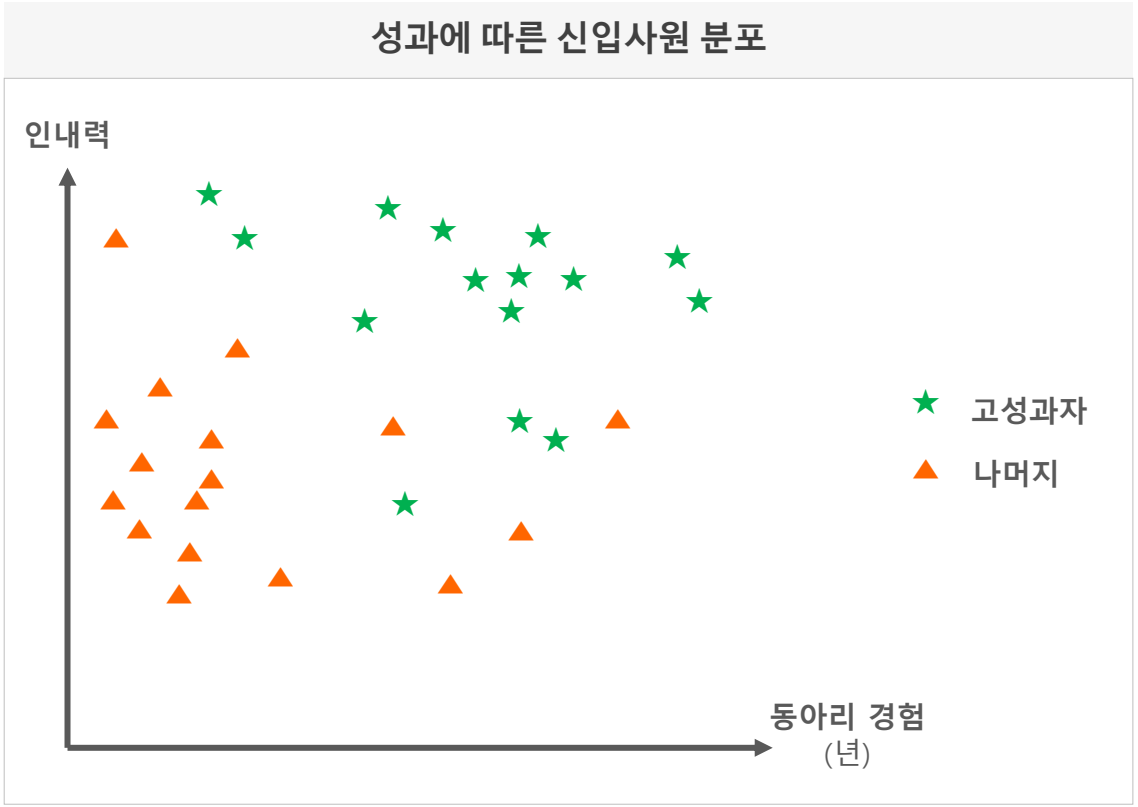


# Decision Tree - Minimizing Entropy

## Decision Tree Algorithm

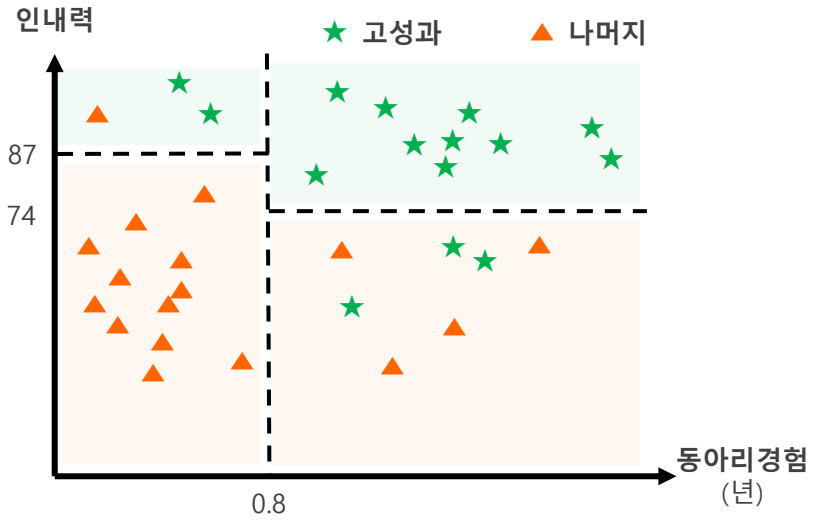
- **Purity:** *엔트로피를 최소화하도록(= 끼리끼리 모이도록) 공간 구획
- **Homogeneity:** 동질적 집단이 밀집한 세그먼트의 논리적 규칙 찾기

***엔트로피(Entropy): Measure of Impurity**

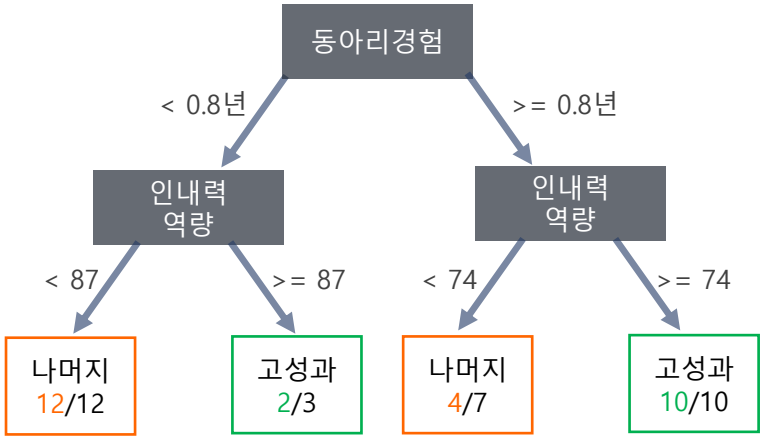


# Decision Tree - Minimizing Entropy

## 성과에 따른 신입사원 분포



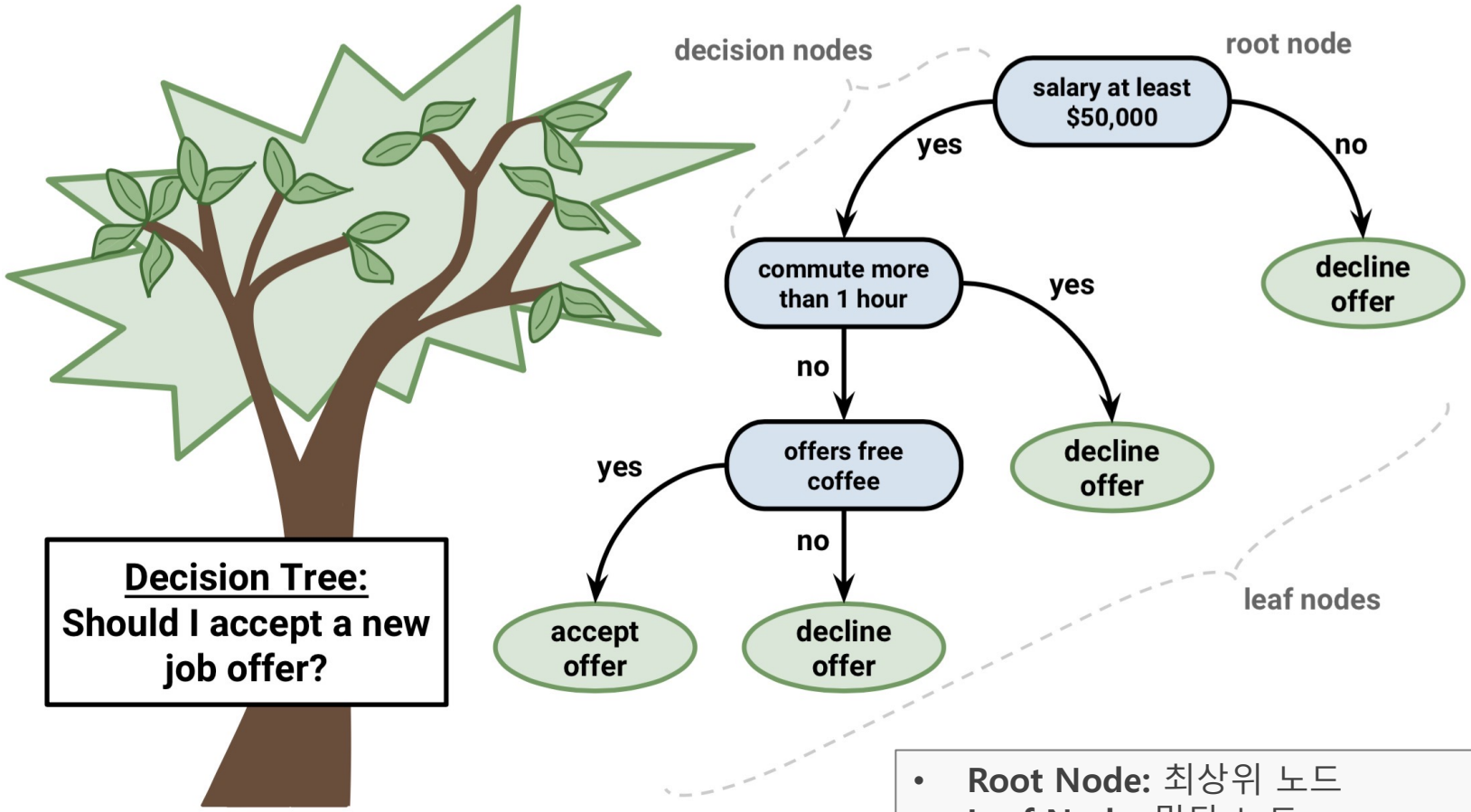

## 고성과 신입사원 분류모형 (의사결정트리)




## 분류규칙

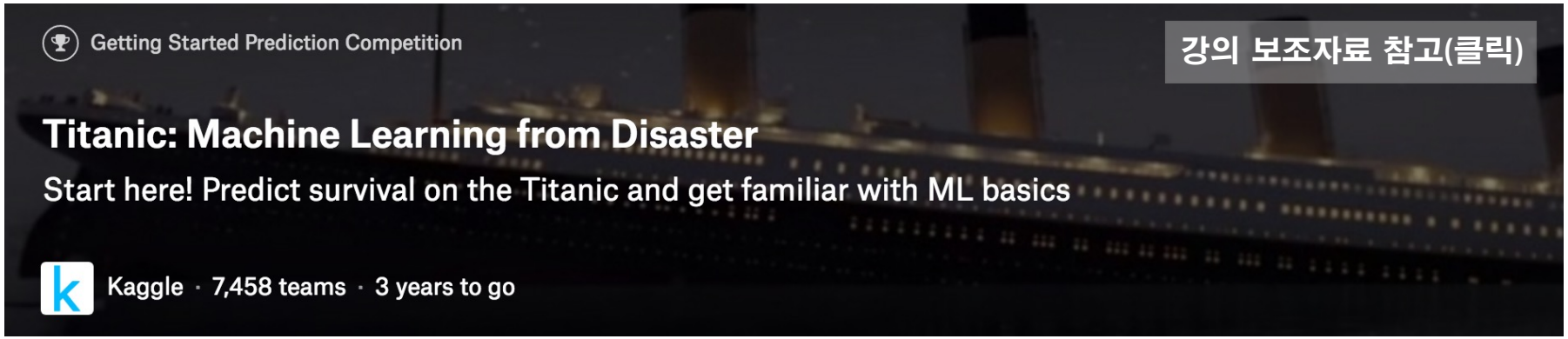
| 규칙 (Rule Set) | 확률 (Probability) |
|---|---|
| IF (동아리경험 < 0.8년) and (인내력역량 < 87) then Class = 저성과 신입사원 | 100% (12/12) |
| IF (동아리경험 < 0.8년) and (인내력역량 >= 87) then Class = 고성과 신입사원 | 67% (2/3) |
| IF (동아리경험 >= 0.8년) and (인내력역량 < 74) then Class = 저성과 신입사원 | 57% (4/7) |
| IF (동아리경험 >= 0.8년) and (인내력역량 >= 74) then Class = 고성과 신입사원 | 100% (10/10) |

# Decision Tree – Terminology



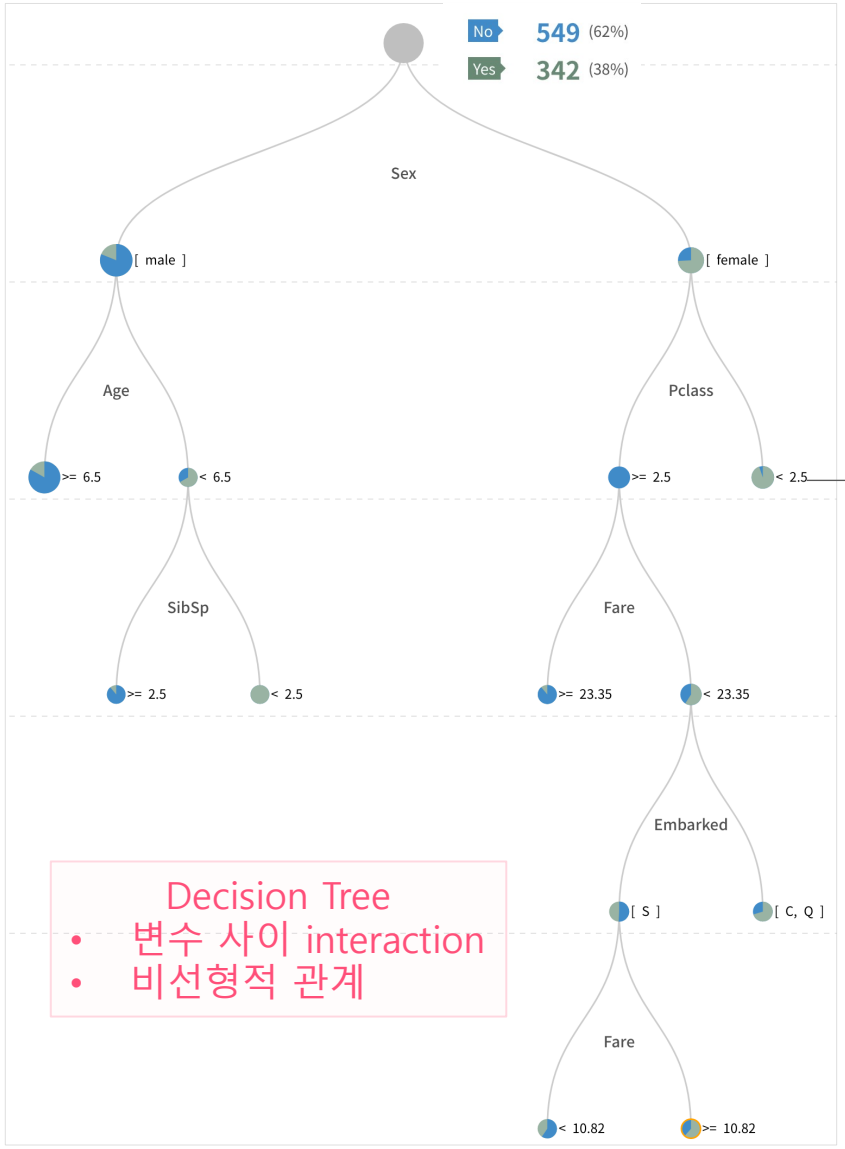

- **Root Node:** 최상위 노드
- **Leaf Node:** 말단 노드
- **Decision Node:** 의사결정 노드
- **Splitting:** 동질적 집단으로 쪼개는 일
- **Pruning:** 트리가 너무 길어지지 않게 하는 일

# Decision Tree – Titanic Dataset




| Variable | Definition | Key |
|---|---|---|
| survival | Survival **(target)** | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex | |
| Age | Age in years | |
| sibsp | # of siblings / spouses aboard the Titanic | Sibling = brother, sister, stepbrother, stepsister<br>Spouse = husband, wife (mistresses and fiancés were ignored) |
| parch | # of parents / children aboard the Titanic | Parent = mother, father<br>Child = daughter, son, stepdaughter, stepson |
| ticket | Ticket number | |
| fare | Passenger fare | |
| cabin | Cabin number | |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

# Decision Tree – Titanic Dataset




Decision Tree
- 변수 사이 interaction
- 비선형적 관계

**세그먼트 특성**

| | |
|---|---|
| 클래스 | Yes |
| 크기 | 170 레코드 |
| 순도 | 94.71% (161/170) |
| 타겟 비율 | 47.08% (161/342) |
| 평균 KPI 차이 | +56.32 ( 94.71 – 38.38 ) |

**세그먼트 분류규칙**

Sex = [female]

Pclass < 2.5

| confusion matrix | Yes (Predicted) | No (Predicted) |
|---|---|---|
| Yes (Actual) | **227** **True Positive** | 115 False Negative |
| No (Actual) | 28 False Positive | 521 True Negative |

$$\text{Precision(Y)} = \frac{227}{227 + 28} \quad \text{89\%}$$

$$\text{Recall(Y)} = \frac{227}{227 + 115} \quad \text{66\%}$$

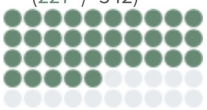
# Decision Tree – Confusion Matrix

## Confusion Matrix: 분류 모형의 성능을 평가하는 방법
이해하기는 쉬운데 용어가 어려움

- **True Positive:** 맞는 걸 맞다고 하는 것
- **True Negative:** 아닌 걸 아니라고 하는 것
- False Positive (I형 오류) : 아닌데 맞다고 하는 것 (거짓을 믿는 것)
- False Negative (II형 오류) : 긴데 아니다고 하는 것 (참을 거부하는 것)

| Confusion Matrix | Yes (Predicted) | No (Predicted) |
|---|---|---|
| Yes (Actual) | **227** **True Positive** | 115 False Negative TYPE II Error |
| No (Actual) | 28 False Positive TYPE I Error | **521** **True Negative** |

**Recall(Y)** 재현률 **66%** = $\frac{227}{227 + 115}$ 66.4% (227 / 342)



**Precision(Y)** 정밀도 **89%** = $\frac{227}{227 + 28}$

**Accuracy** 모형 정확도 **84%** = $\frac{227 + 521}{227 + 115 + 28 + 521}$