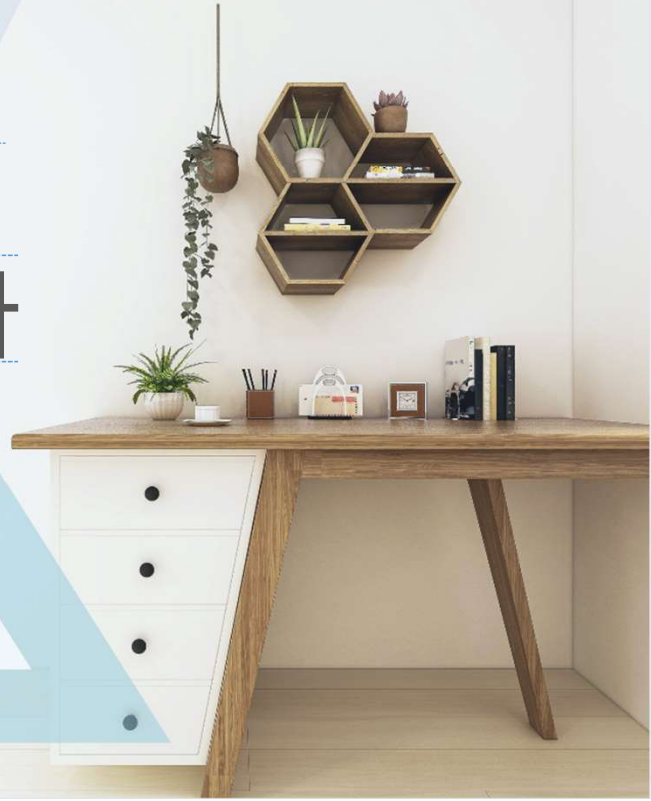
# 10강 모형 비교평가 I

통계·데이터과학과 장영재 교수

한국방송통신대학교

# 목차

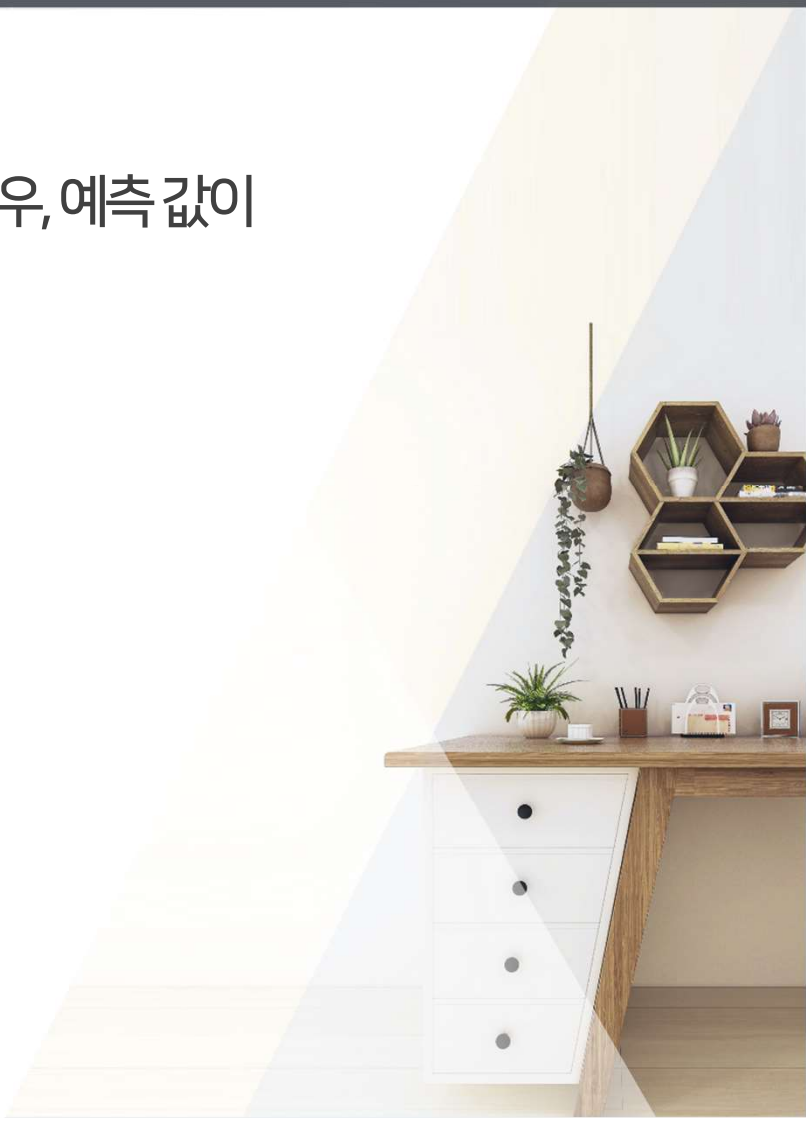




# 1. 평가모형

# 01 평가 모형

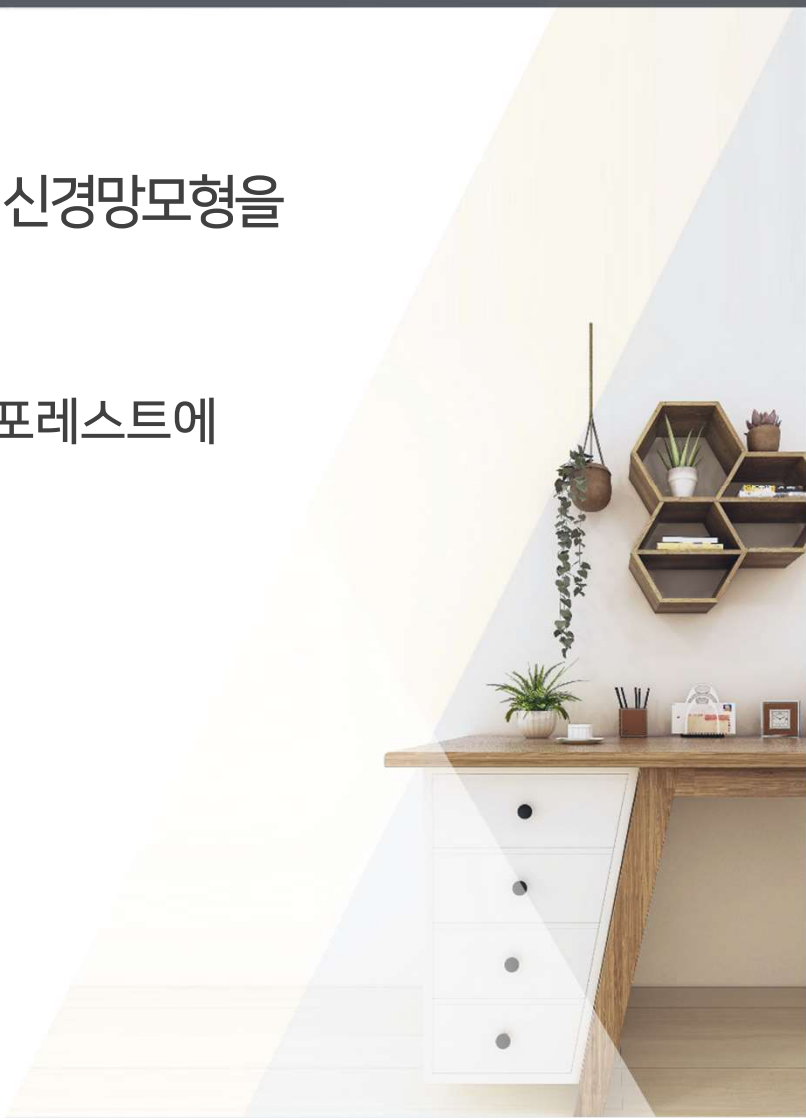
- 목표변수가 존재할 때 다양한 방법으로 모형을 구축하는 경우, 예측 값이 실제 값과 동일 또는 유사하다면 예측이 잘되었다고 평가
  - 데이터마다 예측력을 평가하여 최적의 모형을 선택



# 01 평가 모형

## 1 연속형 목표 변수

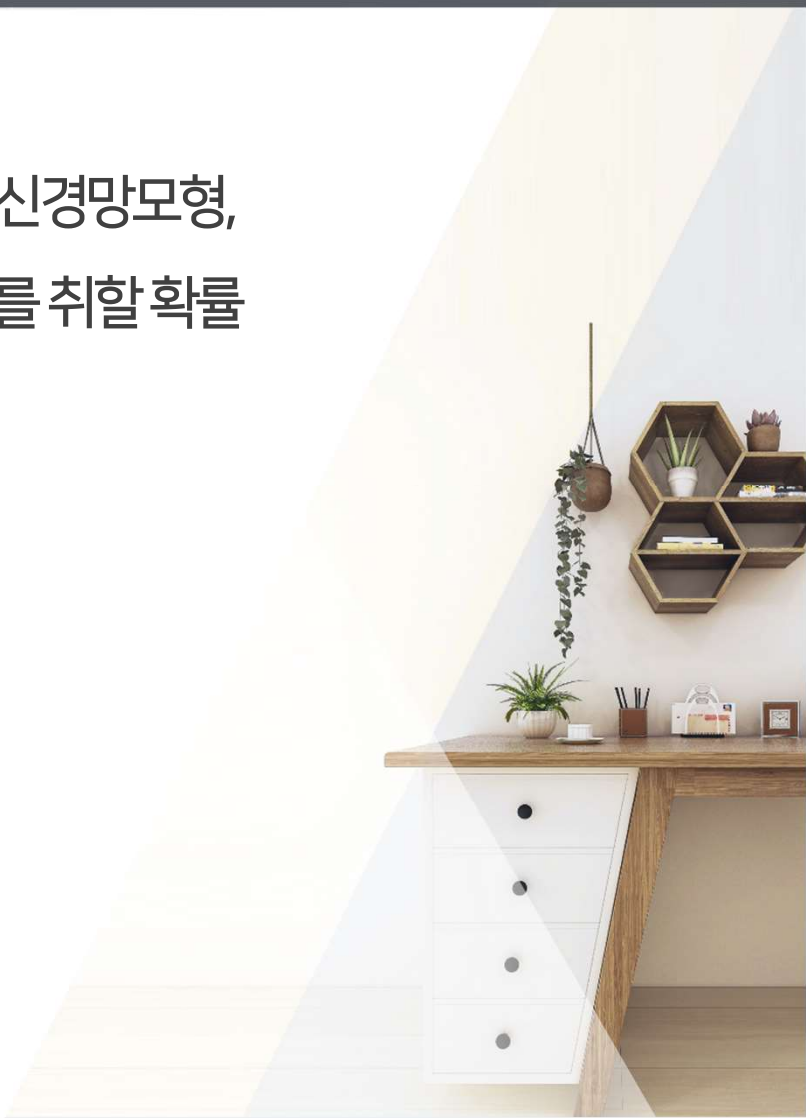
- 목표변수가 연속형일 때 선형회귀모형, 회귀나무모형, 또는 신경망모형을 구축하여 각 객체의 목표변수의 예측 값 산출
  - 선형회귀모형, 회귀나무모형, 신경망모형, 그리고 랜덤포레스트에 의해 생성된 모형을 비교 평가



# 01 평가 모형

## 2 이항형 목표 변수

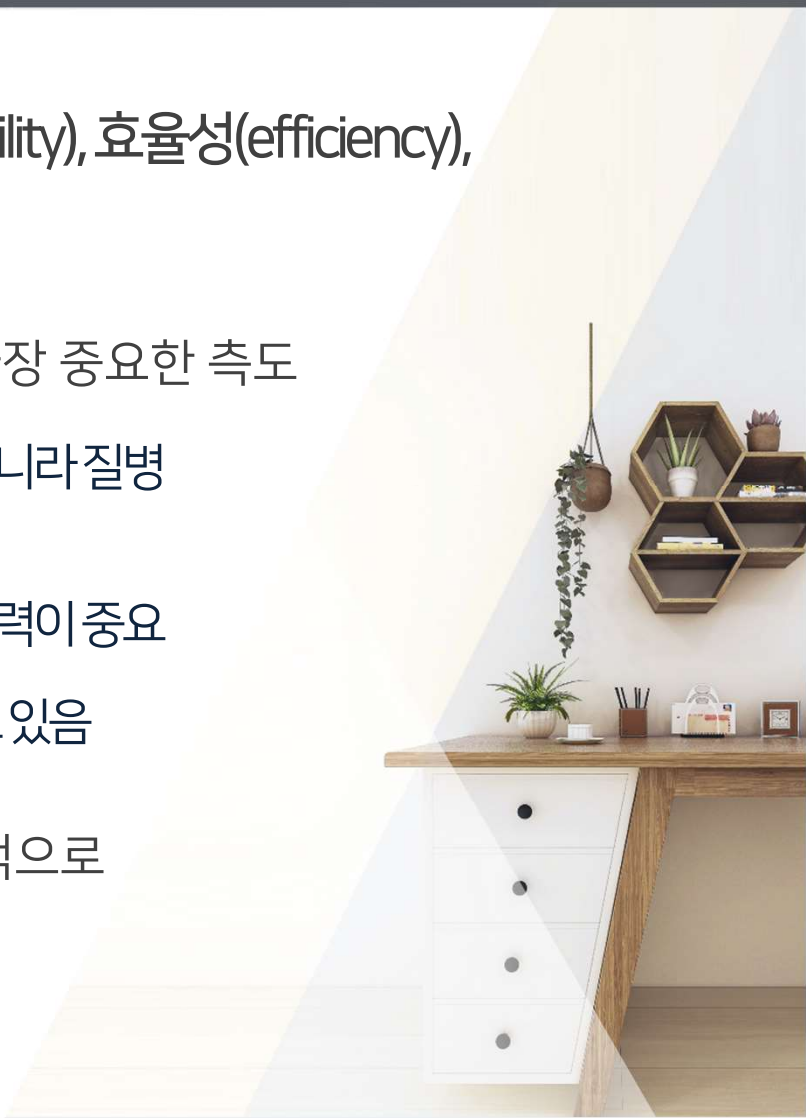
- 목표변수가 이항형일 때 로지스틱회귀모형, 분류나무모형, 신경망모형, 앙상블 (배깅, 부스팅, 랜덤포레스트) 등을 사용해서 각 범주를 취할 확률을 계산
  - 목표변수의 예측 값을 구하여 모형을 비교 평가



# 2. 평가측도

# 02 평가 측도

- 모형 선택시 예측력(prediction power), 해석력(interpretability), 효율성(efficiency), 안정성(stability) 등 다양한 측면을 고려
  - 데이터마이닝의 주목적이 예측이기 때문에 예측력이 가장 중요한 측도
    - 의학연구의 경우 또는 신용평가 분야에서는 예측 뿐 만아니라 질병 예방을 위한 입력변수의 해석 또한 중요한 요소
    - 신용평가 분야에서 평가의 객관성을 설명하는 경우 해석력이 중요
    - 반면, 해석은 중요하지 않으나 예측만 잘하면 되는 경우도 있음
  - 응용분야에 따라 어떤 요소가 중요한지 고려하여 종합적으로 모형을 평가하여 선택하여야 함



# 02 평가 측도

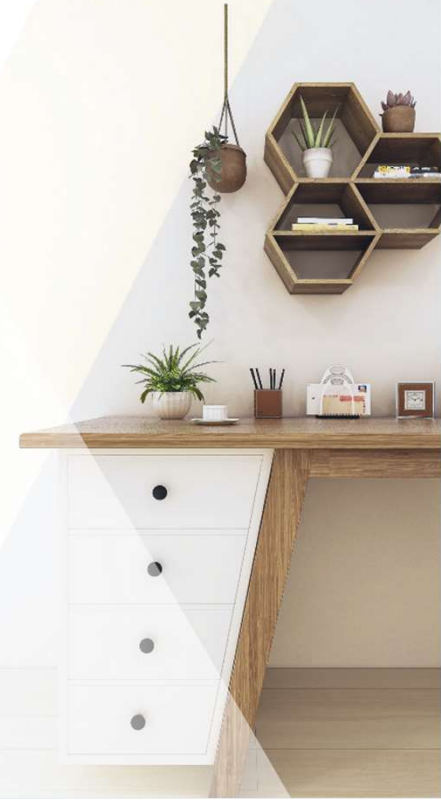
## 1 연속형 목표 변수

- 목표변수가 연속형인 경우에 모형의 예측력 측도로서 MSE( Mean Squared Error) 또는 MAE(Mean Absolute Error)를 주로 사용

$$MSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / n$$

$$MAE = \sum_{i=1}^{n} |y_i - \hat{y}_i| / n$$

- 관측값($y_i$)과 예측값($\hat{y}_i$) 사이에 차이가 적을수록 예측을 잘한 것이므로 MSE 또는 MAE가 작을수록 그 모형의 예측력은 높다고 할 수 있음
  - 관측값과 예측값을 가로 및 세로축으로 놓고 그린 산점도(scatter plot)가 45도 대각선을 중심으로 모여 있으면 예측력이 좋다고 할 수 있음

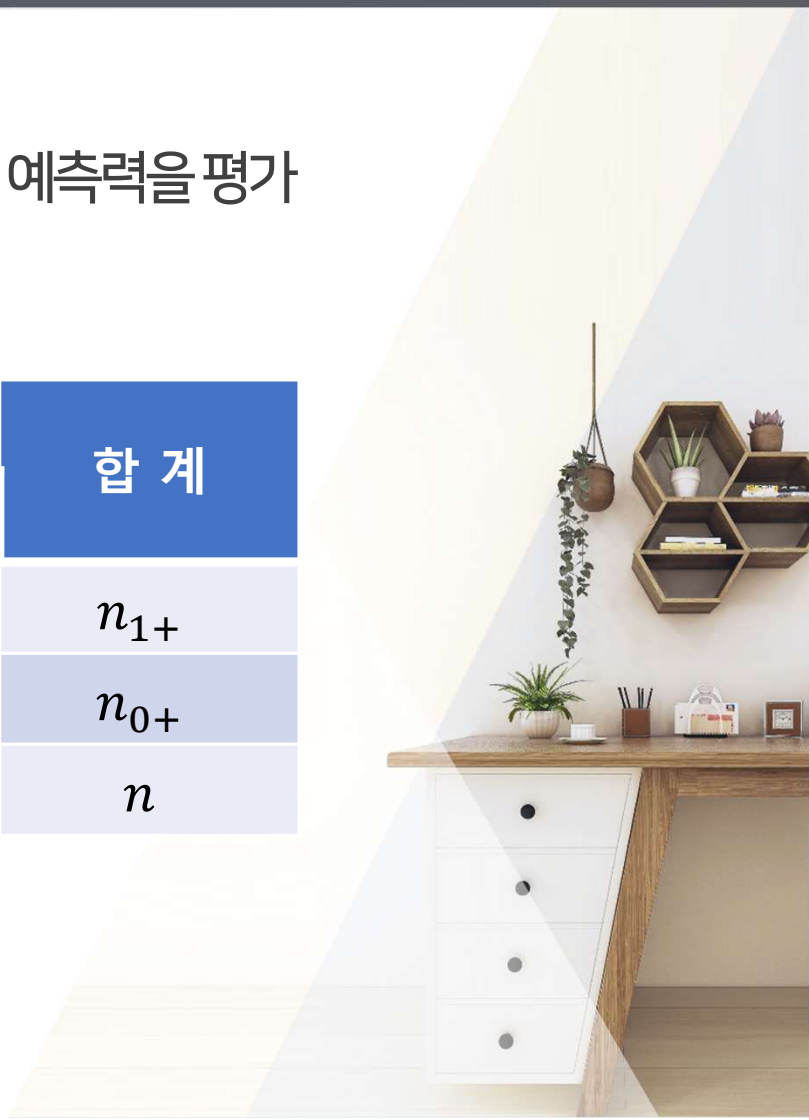


## 2 이항형 목표 변수

### 실제 범주와 예측 범주를 분류표로 만든 정오분류표를 통해 예측력을 평가

- 목표변수의 예측 값을 구하여 모형을 비교 평가

| | | 예측 범주 | | 합 계 |
|---|---|---|---|---|
| | | 1 | 0 | |
| 실제 범주 | 1 | $n_{11}$ | $n_{10}$ | $n_{1+}$ |
| | 0 | $n_{01}$ | $n_{00}$ | $n_{0+}$ |
| 합 계 | | $n_{+1}$ | $n_{+0}$ | $n$ |



# 02 평가 측도

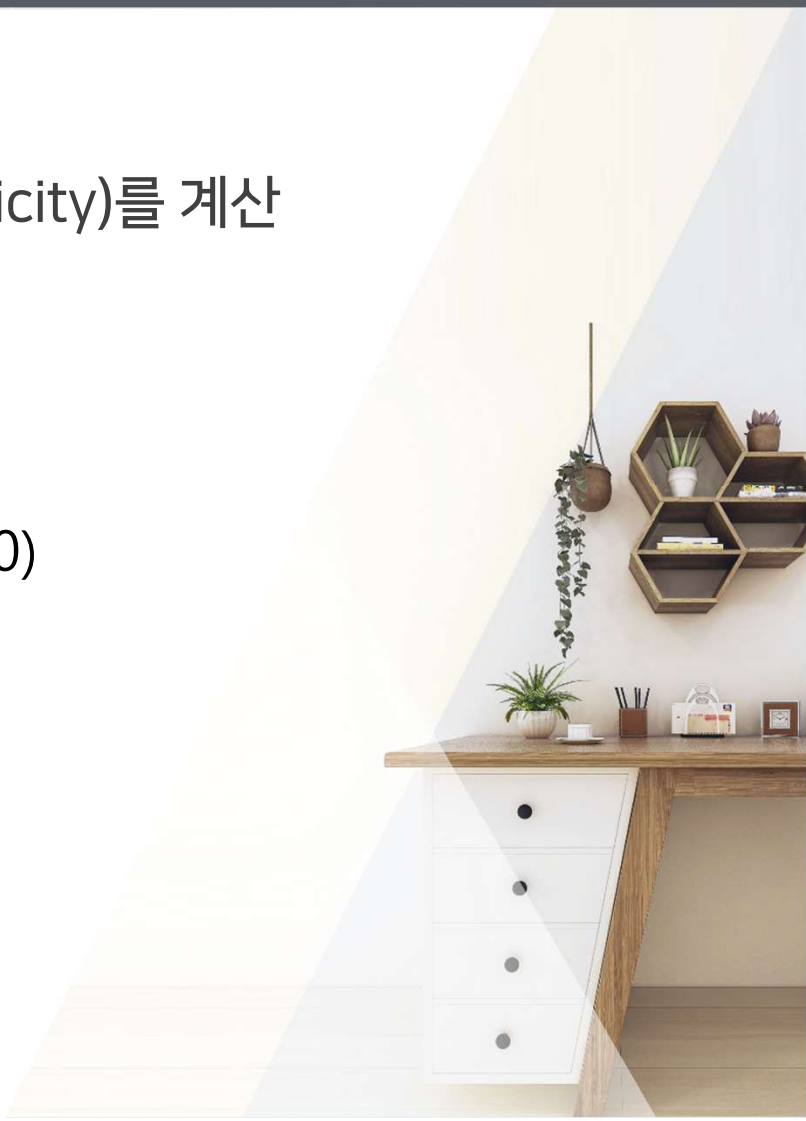
## 2 이항형 목표 변수

- 예측력의 측도로 민감도(sensitivity)와 특이도(specificity)를 계산

  - 민감도 = $\Pr(\hat{Y} = 1 \mid Y = 1) = n_{11} / n_{1+}$
  - 특이도 = $\Pr(\hat{Y} = 0 \mid Y = 0) = n_{00} / n_{0+}$
  - 예측정확도 = $\Pr(\hat{Y} = 1, Y = 1) + \Pr(\hat{Y} = 0, Y = 0)$

    $= (n_{11} + n_{00}) / n$
  - 오분류율 = $\Pr(\hat{Y} \neq 1, Y = 1) + \Pr(\hat{Y} \neq 0, Y = 0)$

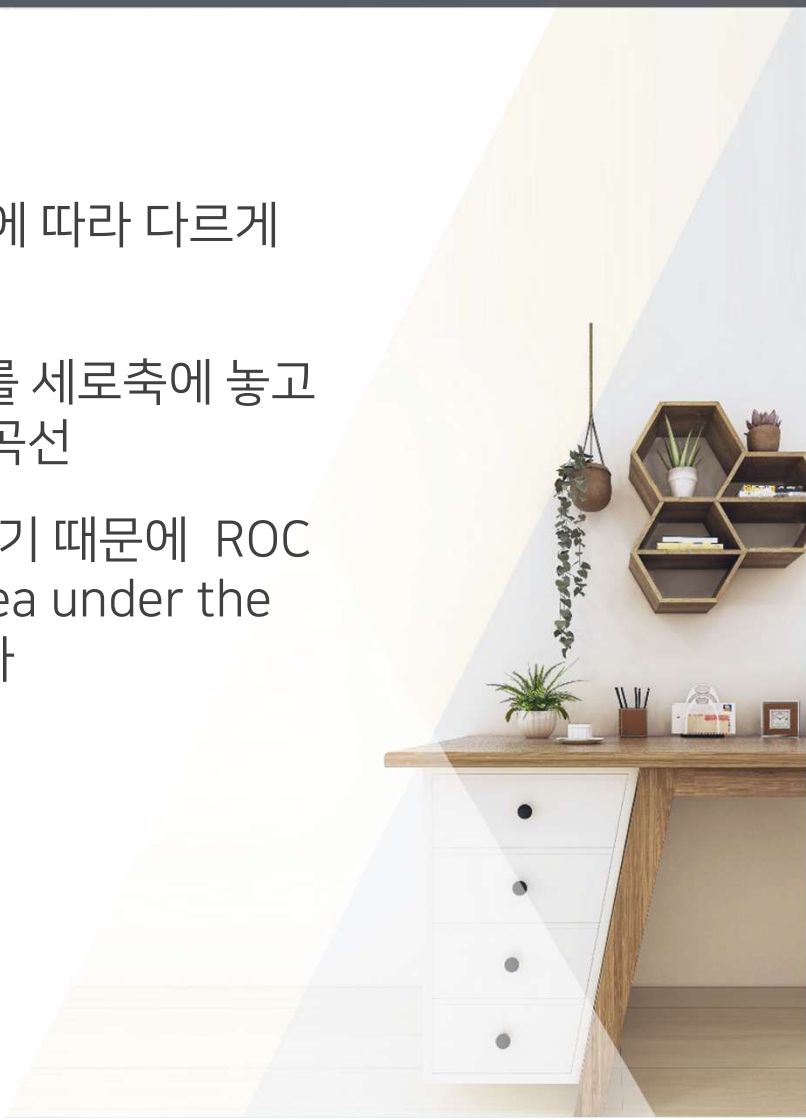
    $= (n_{10} + n_{01}) / n$



# 02 평가 측도

## 2 이항형 목표 변수

- 민감도와 특이도는 임계치에 따라 달라지고, 임계치는 상황에 따라 다르게 결정
- 여러 가능한 임계치에 대해 (1-특이도)를 가로축에, 민감도를 세로축에 놓고 그린 그래프를 ROC(receiver operating characteristic) 곡선
- 민감도와 특이도가 모두 높을수록 예측력이 좋다고 할 수 있기 때문에 ROC 곡선이 좌상단에 가까울수록 ROC 곡선 아래 면적(AUC; area under the ROC curve)이 커지고, AUC가 클수록 예측력이 좋다고 평가



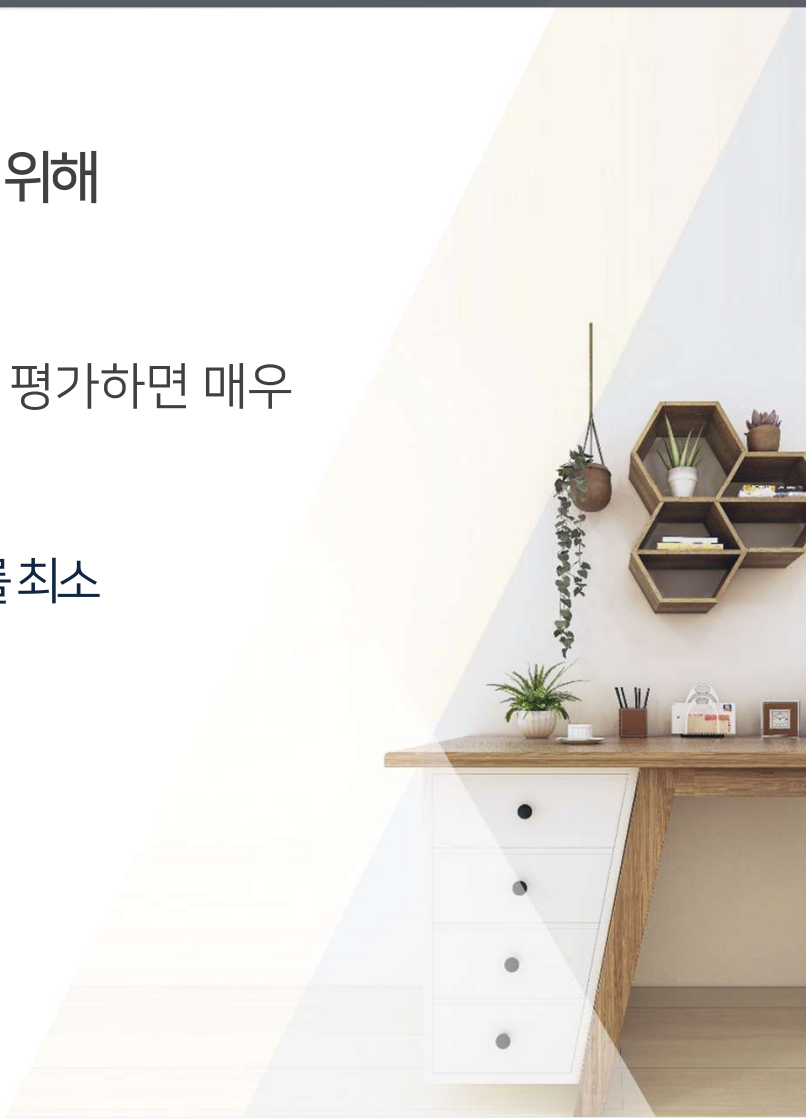
# 3.데이터 분할에 의한 타당도 평가

# 03 데이터 분할에 의한 타당도 평가

- 예측 및 분류모형에 있어서 성능평가는 모형의 실제 활용을 위해 매우 중요한 과정

  - 대체로 모형을 적합한 데이터를 가지고 그대로 예측력을 평가하면 매우 높은 성능을 나타내는 경우가 많음

    - 모형을 적합(훈련)시키면서 이미 이 데이터를 바탕으로 오류를 최소화하는 규칙이 적용되었기 때문

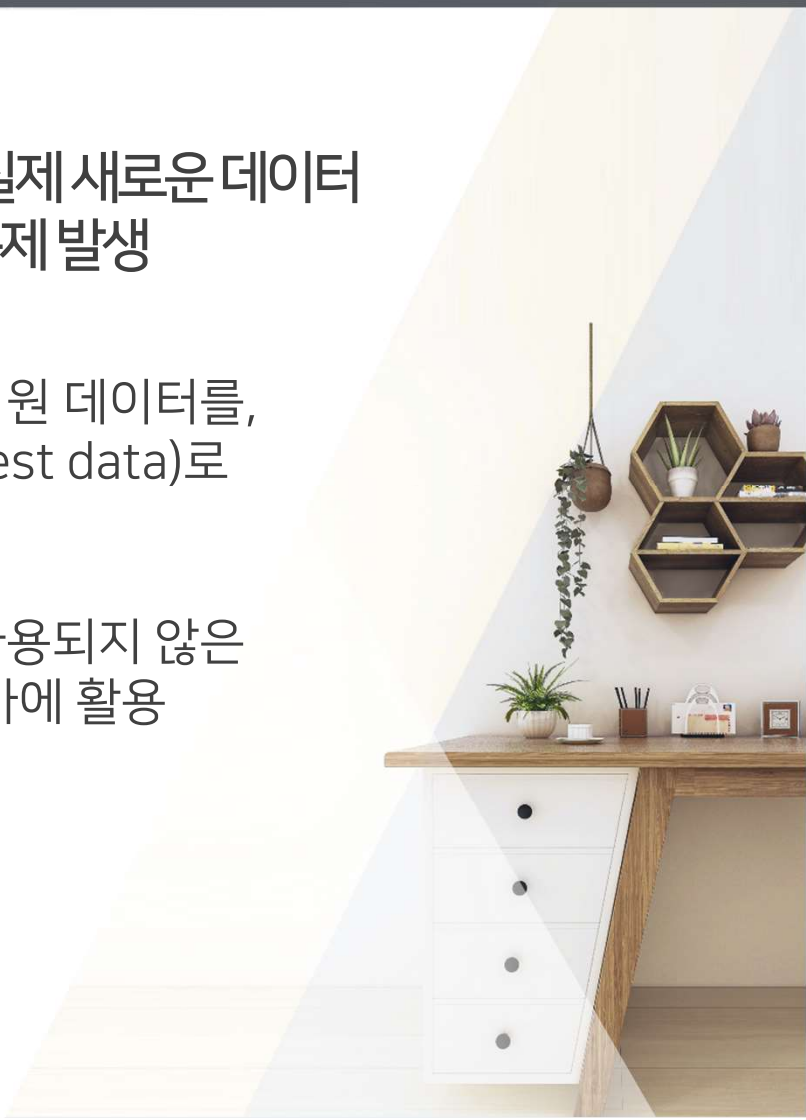


# 03 데이터 분할에 의한 타당도 평가

- 경우에 따라서는 모형 적합 시 높은 성능을 나타내는 모형이 실제 새로운 데이터에 대해서는 매우 낮은 예측력을 보이는 과적합(overfitting) 문제 발생

  - 적절하고 객관적으로 모형의 예측력을 평가하기 위해서는 원 데이터를, 모형을 적합하는 훈련데이터(train data)와 검증데이터(test data)로 분할할 필요

  - 모형훈련에만 사용되는 훈련 데이터와 이 훈련과정에서 사용되지 않은 성능평가용 검증 데이터로 나누어 모형적합 및 예측력 평가에 활용



다음시간 안내

# 11강.
# 모형비교평가 II