# 데이터 품질/대표성 없는 훈련 데이터

---

## 데이터

- Instances 및 해당 Instances의 Attributes에 대한 모음이라고 표현할 수 있음
- Attributes는 인스턴스의 속성 또는 특성으로 Instances와 Attributes는 아래와 같이 다른 이름으로 표현 가능

**Attributes (variable/feature/characteristics/field)**

**Instances (sample/ record/ point/ case/ object)**

| Outlook | Temperature | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | Hot | High | False | No |
| Sunny | Hot | High | True | No |
| Overcast | Hot | High | False | Yes |
| Rainy | Mild | Normal | False | Yes |
| ... | ... | ... | ... | ... |

*=> Attributes: 인스턴스의 특성 측정 (variable/feature/characteristics/field)*

*=> Instances: 개념의 개별적이고 독립적 인 예 (sample/record/point/case/object)*

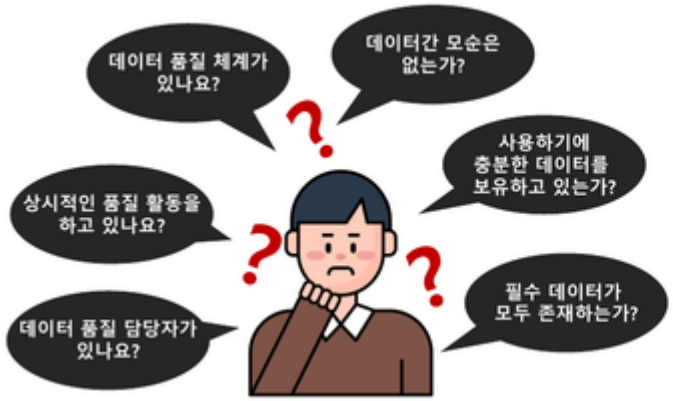
---

## 데이터 품질이란?

- "데이터의 최신성, 정확성, 상호연계성 등을 확보하여 이를 사용자에게 유용한 가치를 줄 수 있는 수준"으로 정의 - 데이터가 특정 목적에 적합한 정도를 가리키는 개념

  - 데이터 품질 표준은 기업이 비즈니스 목표를 달성하기 위해 데이터 기반 의사 결정을 내릴 수 있도록 보장함
  - 중복 데이터, 누락값, 이상값과 같은 데이터 문제를 제대로 해결하지 않으면 비즈니스 결과에 부정적인 영향이 발생할 위험이 높아짐
  - Gartner 보고서에 따르면 데이터 품질 저하로 인해 조직은 매년 평균 USD 1,290만에 달하는 비용을 지출하고 있음
  - 그 결과, 데이터 품질 저하와 관련된 부정적인 영향을 완화하기 위한 데이터 품질 도구가 등장했음

---




급변하는 대·내외 환경에서
사용자에게 신속한 의사결정을 지원할
수 있는 고품질의 데이터 제공(활용)이
비즈니스 성패와 직결

↓

**"데이터 품질관리 체계 필요"**

---

## 데이터 품질 문제

- Missing values(결측 데이터), Noise and outliers(노이즈 및 이상값), Nomalization(정규화) 문제로 발생

  - Missing values(결측 데이터)

        - 결측 데이터 발생 및 처리
        - 결측 데이터 탐지 방법
        - 결측값 보정(imputation)
    - Noise and outliers(노이즈 및 이상값)
        - 이상값 원인 및 처리
    - Feature scaling
        - Normalization(정규화)
        - Standardization(표준화)

---

## 머신러닝의 주요 도전 과제

- 학습 알고리즘을 선택해서 데이터에 훈련시키는 과정에서 문제가 될 수 있는 두 가지
    - 나쁜 알고리즘
    - 나쁜 데이터
- 1) 충분하지 않은 양의 훈련 데이터
    - 대부분의 머신러닝 알고리즘이 잘 작동하려면 데이터가 많아야 함
    - 아주 간단한 문제에서조차도 수천개의 데이터가 필요하고, 이미지 음성 인식 같은 복잡한 문제라면 수백만개가 필요할지도 모름
- 2) 대표성 없는 훈련 데이터

  - 일반화 하려는 사례들을 대표하는 훈련세트를 사용하는 것이 중요하지만, 어려울 때가 많음
  - 샘플이 작으면 샘플링 잡음(sampling noise) 발생 - 샘플링 잡음은 우연에 의한 대표성 없는 데이터를 뜻함
  - 매우 큰 샘플도 표본 추출 방법이 잘못되면 대표성을 띠지 못할 수도 있음 - 샘플링 편향(smpling bias)
- 3) 낮은 품질의 데이터
  - 훈련 데이터가 에러, 이상치 (outlier), 잡음으로 가득하다면 머신러닝 시스템이 내재된 패턴을 찾기 어려워 잘 작동하지 않을 것임
- 4) 관련 없는 특성
  - 성공적인 머신러닝 프로젝트의 핵심 요소는 훈련에 사용할 좋은 특성들을 찾는 것 -> 특성 공학(feature engineering)
    - 특성 선택(feature selection) - 가지고 있는 특성 중에서 훈련에 가장 유용한 특성을 선택
    - 특성 추출(feature extraction) - 특성을 결합하여 더 유용한 특성을 만듦
- 5) 훈련 데이터 과대/과소적합

---

## <<<참조자료 사이트>>>

1.AI 훈련용 빅데이터 2026년 고갈…문제점과 대책은
2."학습 데이터 부족" AI개발에 닥친 난관… 차세대 모델 개발 지연
3.AI 데이터 고갈 위기-"2년 후 AI 성장 멈출 수도"…데이터 절벽 '경고'
4.Undersampling과 Oversampling이란?
5.불균형 데이터(imbalanced data) 처리를 위한 샘플링 기법
6.샘플링 편향이라는 문제
7.데이터 품질(Data Quality)이란?
8.데이터에 제값 매기는 데이터 품질 관리

---