Machine Learning

들어가며.

2024.07.01

머신러닝 ?

- 기계가 일일이 코드로 명시하지 않은 동작을 데이터로부터 학습하여 실행할 수 있도록 하는 알고리즘을 개발하는 연구 분야

- Ex. 현재 내가 살고 있는 집을 팔려고 한다. 얼마가 적절한 가격일까?
- 집의 가격을 결정하기 위해 주변의 다른 집들의 가격들을 확인해 보았다. 그리고 집의 가격을 결정하는 몇 가지 요인을 찾을 수 있었다.

    역에서 가까운지?

    집이 오래 되었는지?

    주차장이 넓은지?

머신러닝 ?

머신러닝의 3요소

1. 알고리즘 : 문제를 풀기 위한 의사 결정 과정
   집이 낡았음 -> 주차장이 좁음 -> 역에서 가까움 -> 5억원!

2. 데이터 : 주관적인 기준이 아닌 수치에 근거한 객관적인 기준
   가깝다 -> 10km 이내
   오래 되었다 -> 10년이 넘었다
   주차장이 크다 -> 주차장이 100m2보다 크다

3. 학습 : 데이터로부터 최적의 값을 찾는 과정
   수치에 들어갈 값을 컴퓨터가 스스로 찾는다! -> 10km, 10년, 100m2

머신러닝이란 풀고자 하는 문제의 정답과 데이터를 주고 기계를 학습시켜서 정답을 맞추게 하는 것이다.

머신러닝 알고리즘의 종류

1. 지도 학습 (Supervised Learning)

- 정답이 있는 문제를 풀 경우

2. 비지도 학습 (Unsupervised Learning)

- 정답이 없는 문제를 풀 경우

정답의 종류에 따른 지도 학습

1. 회귀 (Regression)

- 정답이 연속형 변수일 때

2. 분류(Classification)

- 정답이 비연속형(범주형) 변수일 때

회귀(Regression)

- 회귀(regress)의 원래 의미는 옛날 상태로 돌아가는 것을 의미

- 정답이 연속형 변수일 때

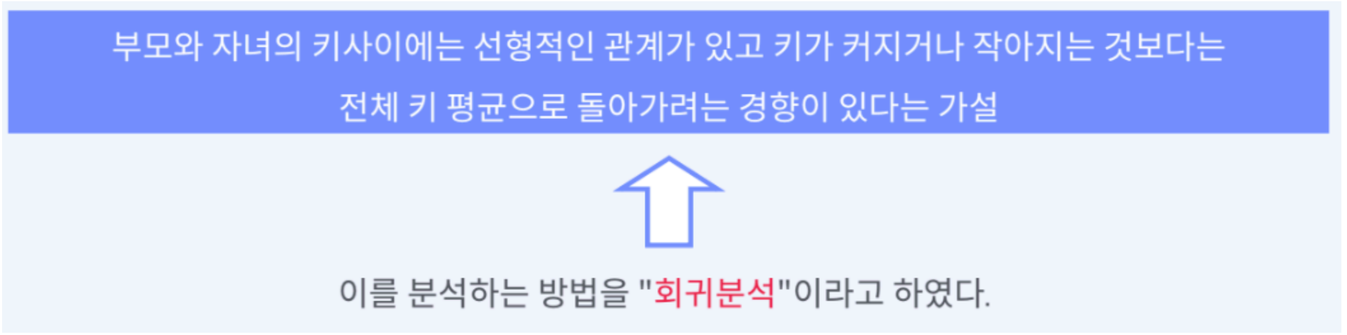
- 연속형이란?

  - 값이 정수처럼 명확하지 않음

eg) 키, 몸무게, ...

- 회귀 분석 문제

  - 키를 이용해 몸무게를 예측하기



분류(Classification)

정답이 비연속형(범주형) 변수일 때

- 범주형 변수란?
    - 값이 정수처럼 명확함

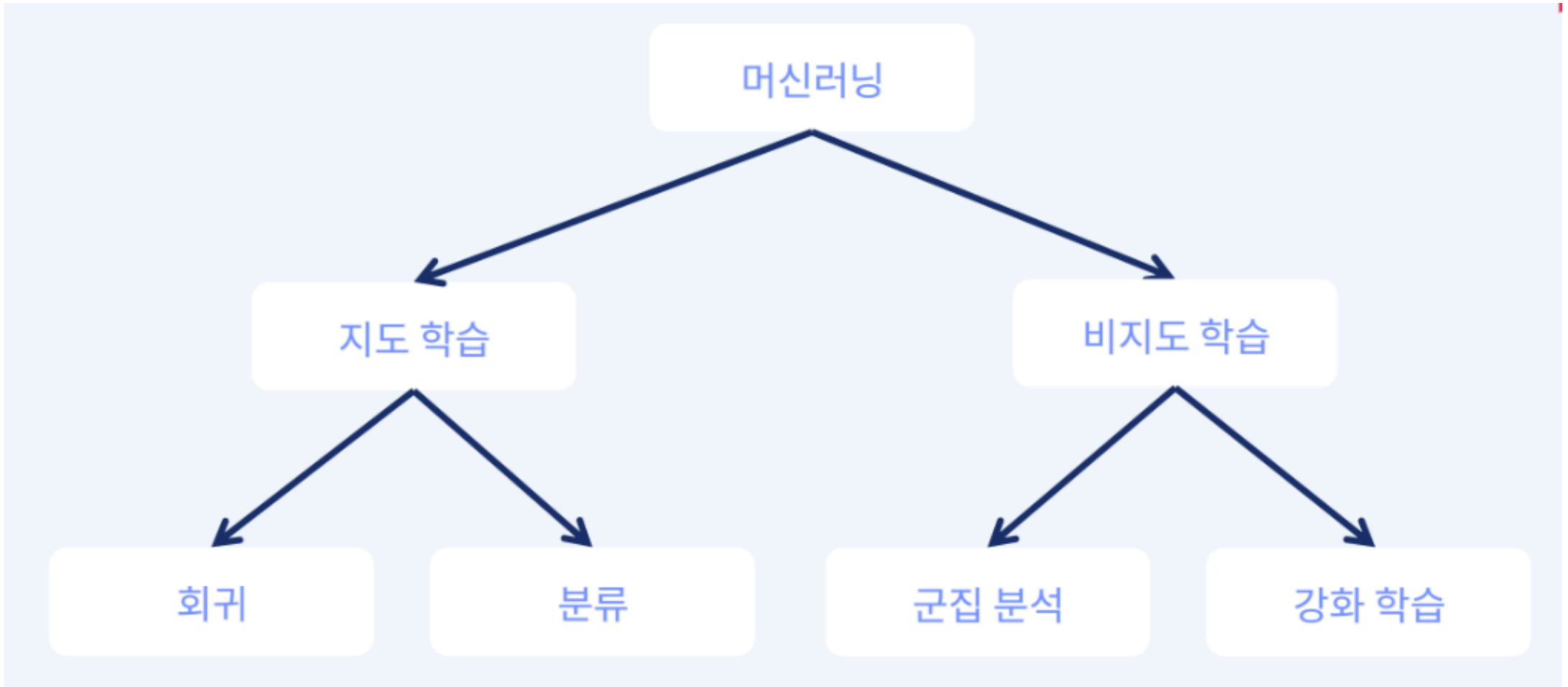
eg) 성별, 도시, ...

- 분류 문제
    - 키와 몸무게를 이용해 성별을 맞추기

비지도 학습의 종류

1. 군집분석 (Clustering) : 주어진 데이터가 어떻게 구성 되었는지 알아내려는 분석 방법

2. 강화학습 (Reinforcement Learning) : 행동에 따른 보상을 최대화 시키는 학습 방법


머신러닝
지도 학습
비지도 학습
회귀
분류
군집 분석
강화 학습

Model Selection

1. 모델

2. Train data, Test data

3. 과대적합, 과소적합

4. Cross Validation

모델

[모델의 정의]

- 어떤 X가 주어 졌을 때 f 라는 함수를 통해 y라는 값을 도출하는 과정
- 이 때의 함수 f 를 모델 또는 알고리즘이라고 부른다.

[모델의 수식]

- y = f(X)
    - X : 데이터
    - y : 예측값

[모델의 목적]

- 데이터를 이용해 값을 예측

[모델의 평가]

- 모델이 값을 잘 예측하는지 평가

데이터의 종류

Train Data

- 학습에 사용되는 데이터

Test Data

- 학습에 사용되지 않은 데이터

- 모델이 실제로 잘 예측하는지 알기 위해서는 학습에 사용되지 않은 데이터를 이용해 평가해야 한다.

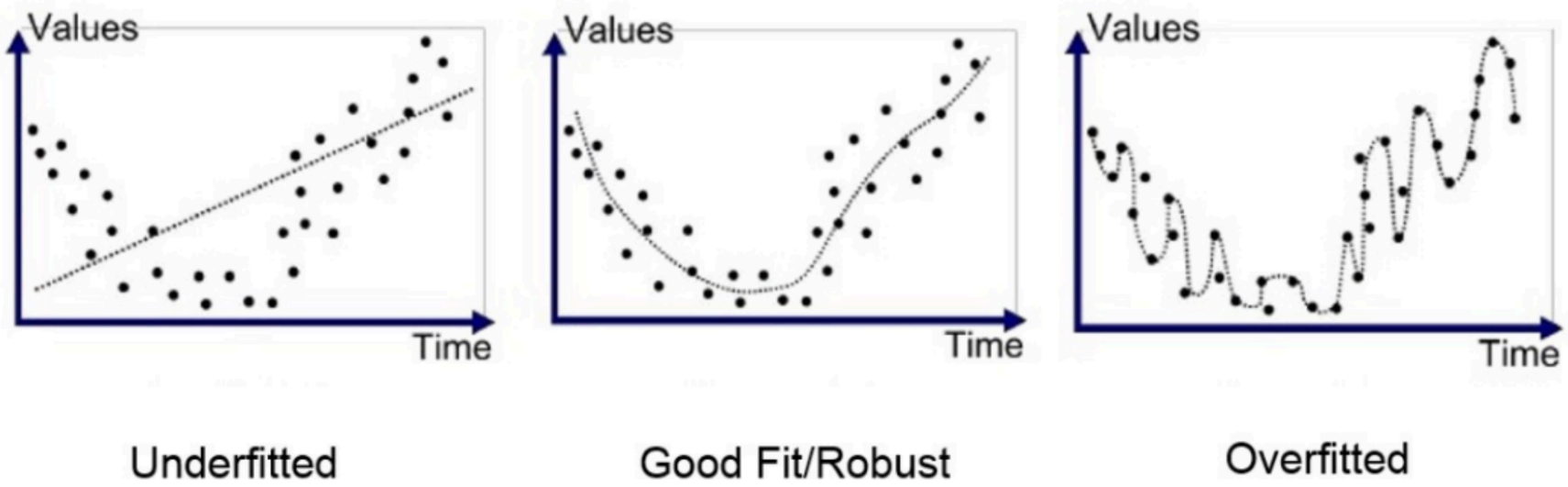
모델 평가와 데이터의 관계

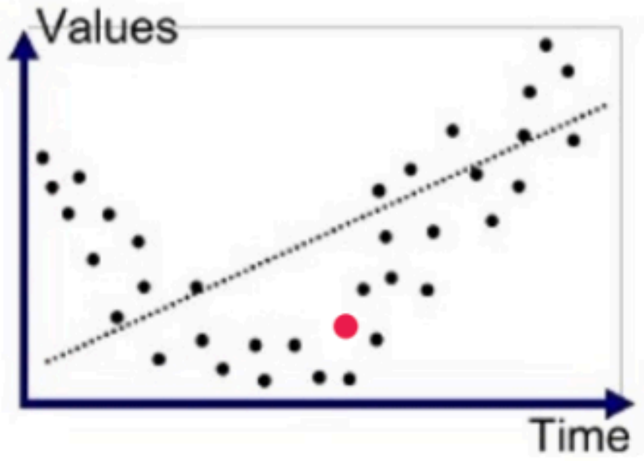
Underfitting (과소 적합)

학습 데이터를(Train data) 잘 맞추지 못하는 현상

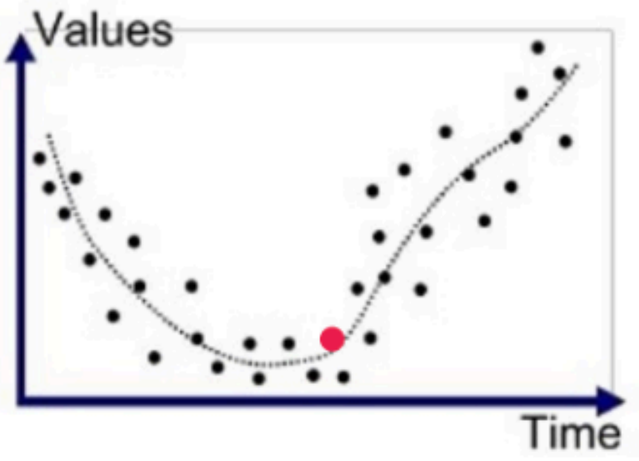
Overfitting (과대 적합)

학습 데이터(Train data)는 잘 맞추지만 학습 데이터 외에는 잘 맞추지 못하는 현상

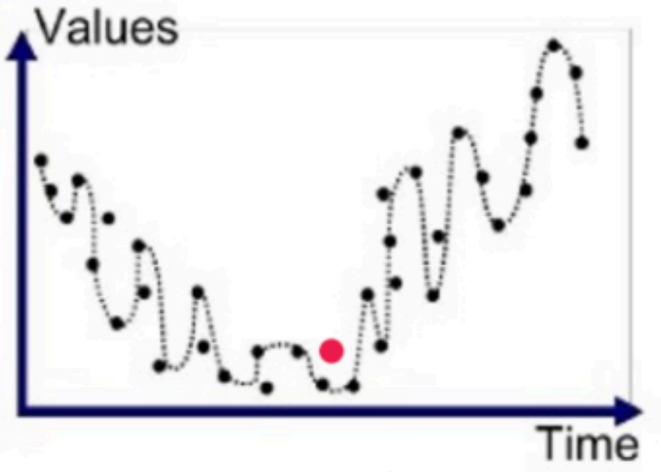




Underfitted



Good Fit/Robust



Overfitted

Underfitting을 확인하는 방법

- Train data로 학습된 모델을 Train data로 평가한다.
- Train data를 잘 맞추지 못 한다면 Underfitting 상태

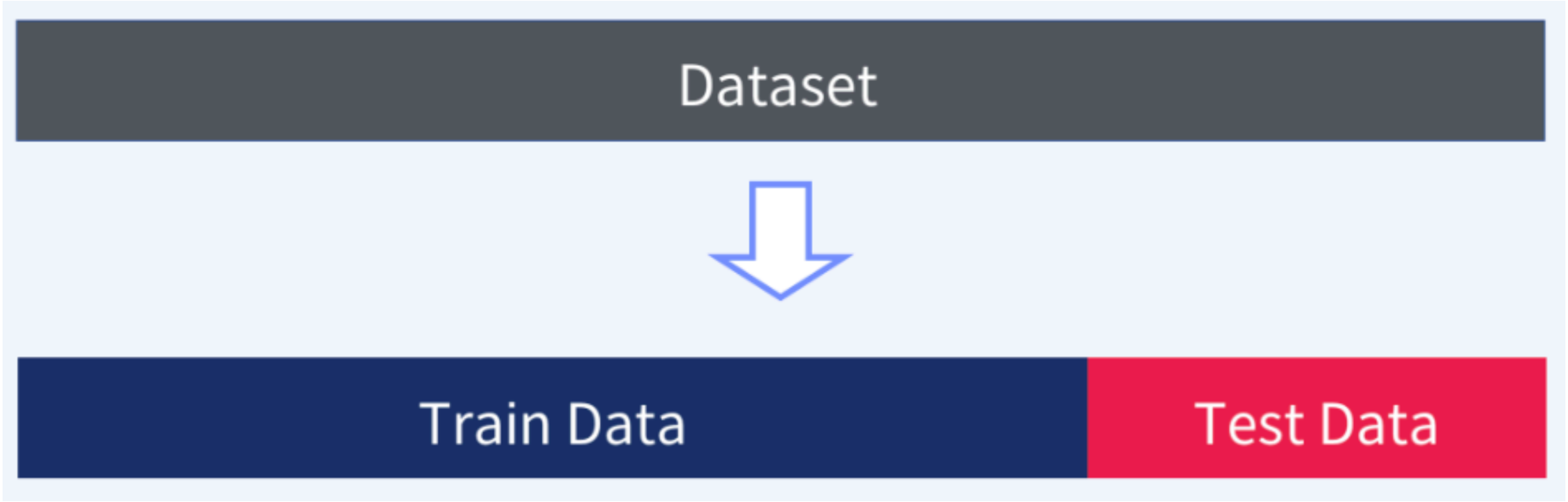
Overfitting을 확인하는 방법

- Train data을 잘 학습한 모델을 Test data로 평가한다.
- Train data는 잘 맞추지만 Test data를 잘 맞추지 못한다면 Overfitting 상태

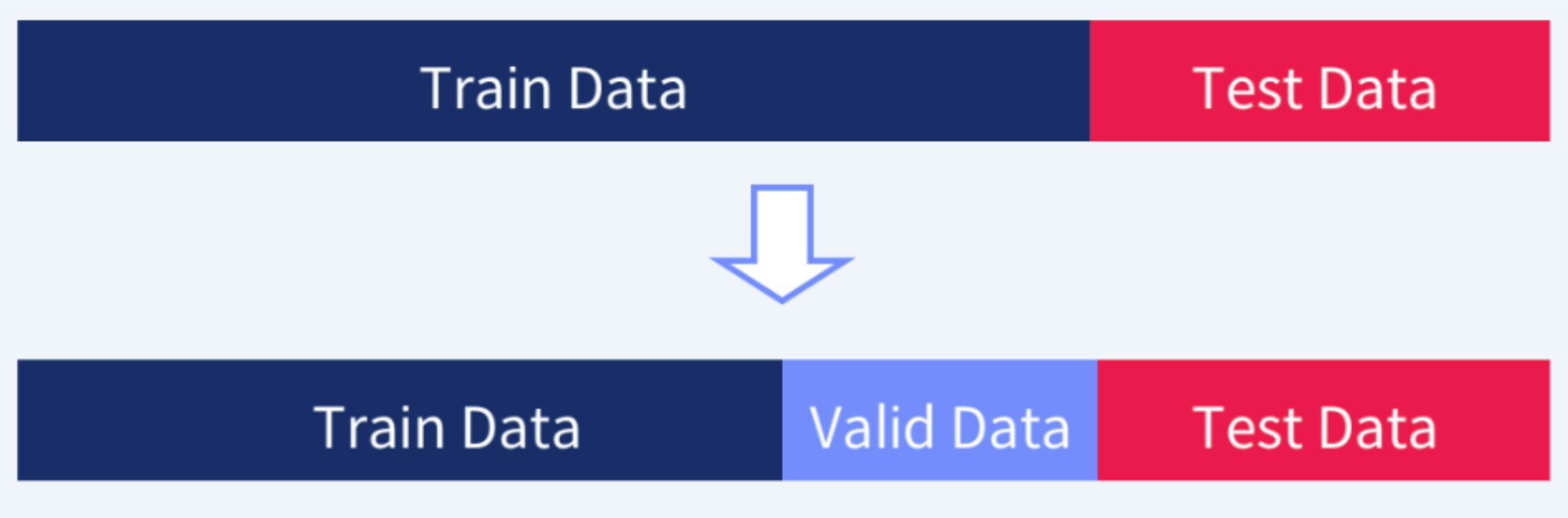
Data Split ?

데이터를 Train data와 Test data로 나누는 것

Train / Test 로 나누기



Train / Valid / Test 로 나누기



각 데이터의 용도

Train Data

- 학습에 사용되는 데이터

Valid Data

- 학습이 완료된 모델을 검증하기 위한 데이터

- 학습에 사용되지는 않지만 관여하는 데이터

Test Data

- 최종 모델의 성능을 검증하기 위한 데이터

- 학습에도 사용되지 않으며 관여하지도 않는 데이터

Train / Valid / Test 로 나누기

Valid Data

- > 학습에 사용되지는 않지만 관여하는 데이터

Overfitting

Valid 데이터에 Overfitting 될 수 도 있음

Cross Validation

Valid 데이터를 고정하지 않고 계속해서 변경함으로써 Overfitting 되는 것을 막기 위한 방법

Cross Validation 종류

1. LOOCV (Leave One Out Cross Validation)
2. K-Fold

LOOCV (Leave One Out Cross Validation)

하나의 데이터를 제외하고 모델을 학습한 후 평가

LOOCV (Leave One Out Cross Validation)

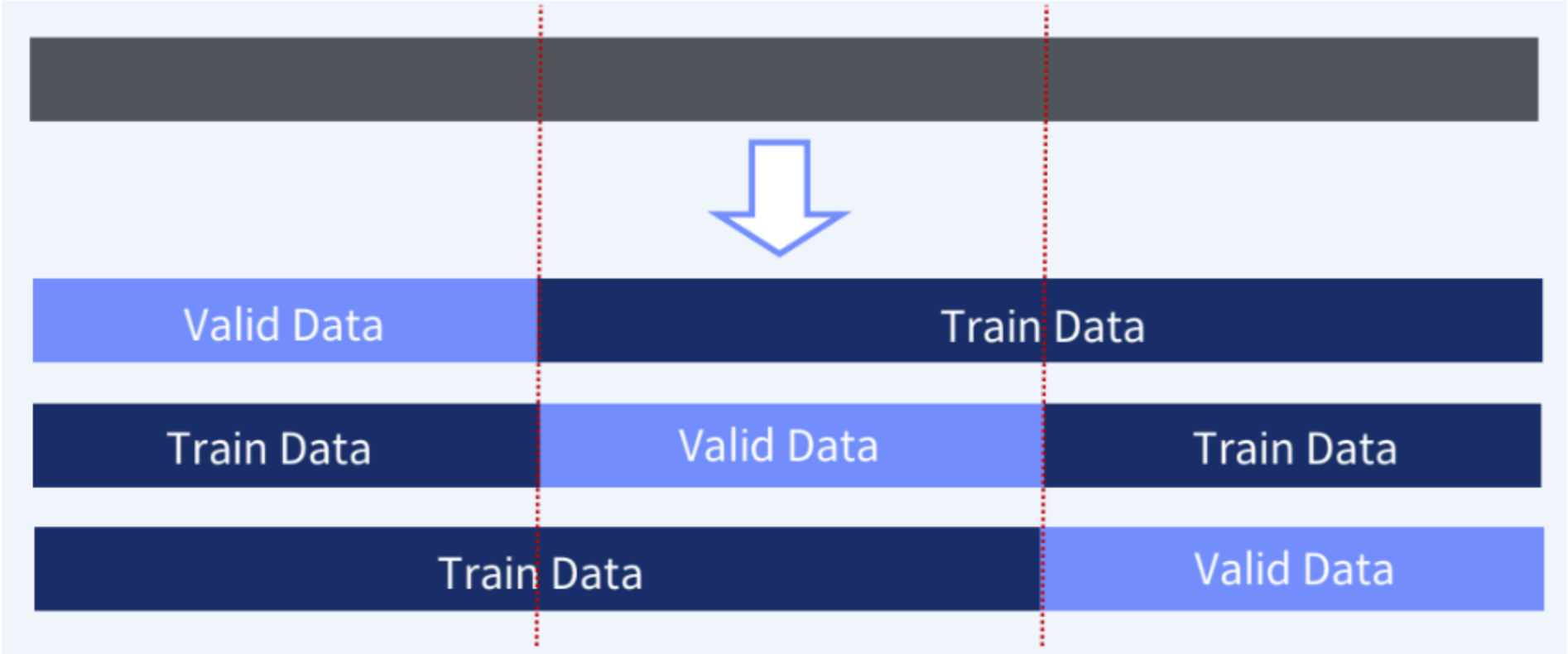
하나의 데이터를 제외하고 모델을 학습한 후 평가

- 데이터 개수 만큼의 모델을 학습해야 한다.
- 데이터가 많을 경우 시간이 오래 걸린다.

K-Fold

- 데이터를 K 개로 분할한 후 한 개의 분할 데이터를 제외한 후 학습에 사용

- 제외된 데이터는 학습이 완료된 후 평가에 사용



Cross Validation 평가

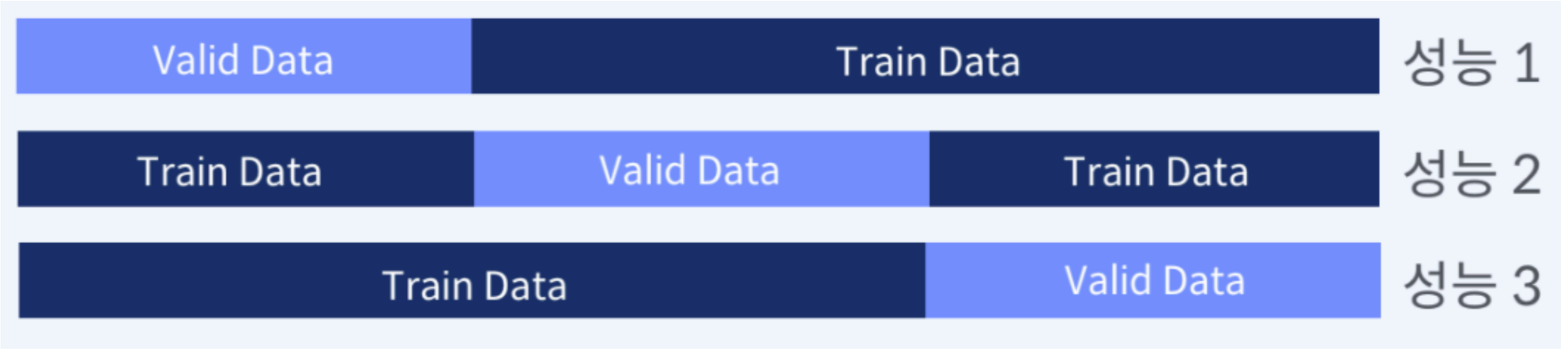
1. Cross Validation을 이용하면 방법에 따라 K개의 평가지표가 생성

2. 생성된 평가 지표의 평균을 이용해 모델의 성능을 평가

3. 전체 Train 데이터를 이용해 모델 학습

Cross Validation 평가

step 1) Cross Validation을 이용하면 방법에 따라 K개의 평가지표가 생성

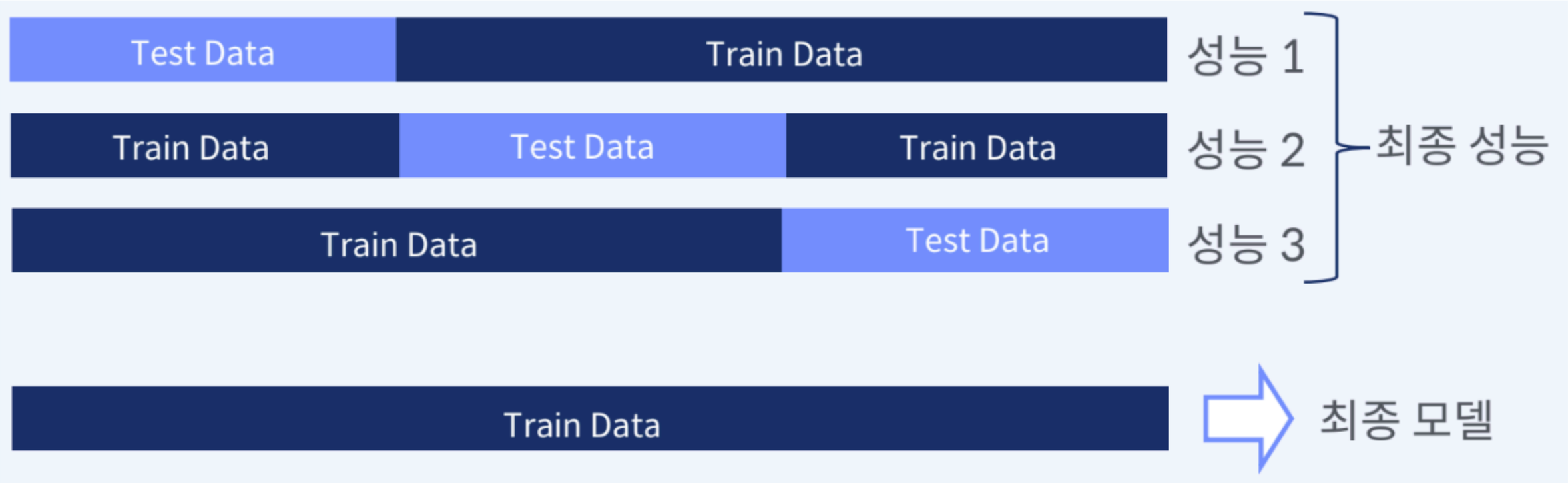


Cross Validation 평가

Step 2) 생성된 평가 지표의 평균을 이용해 모델의 성능을 평가

Cross Validation 평가

Step 3) 전체 Train 데이터를 이용해 모델 학습



데이터 분석과
Machine Learning

들어가며.

2024.07.01

1. 일반적인 데이터 분석

일상에서의 데이터 분석

우리의 뇌는 끊임없이 데이터 분석을 하고 있다.

수강생	방법	사용 유무
A	R	X
A	Python	X
A	Excel	O
B	R	X
B	Python	X
B	Excel	X

문제 정의 → “A와 B는 데이터 분석 경험이 있는 것일까?”

A, B 의 특징 정리

수강생	R	Python	Excel
A	X	X	O
B	X	X	X

사람의 경험 기반 분류 모델

“A는 데이터 분석 경험이 있다.
B는 데이터 분석 경험이 없다.”

1. 일반적인 데이터 분석

통계적인 데이터 분석

수집하고 다룰 수 있는 데이터의 크기가 작고 연산 능력이 부족한 시기에는

통계 분석 및 시각화를 활용한 방법을 주로 사용.

예측의 정확도를 높이는 것 보다는 분석의 실패 가능성을 줄이는 것을 목표로 함

1. 일반적인 데이터 분석

통계적인 데이터 분석

[데이터 수집]

[데이터 정제]

방법 X→제거

결측값→제거

수강생	방법	사용 유무
A	R	X
A	Python	X
A	Excel	O
A	가위	O
B	R	X
B		O
B	Python	O
B	Excel	X

문제 정의

내용 이해

“A와 B는 데이터 분석 경험이 있는 것일까?”

데이터 수집 비용 ↑ 서버 연산 능력 ↓
→ 보유 데이터로 분석 실패 가능성 ↓
→ 개별 특성값의 설명 중요도 ↑

통계 분석 및 시각화

“모형을 단순화하여
모형의 신뢰도와
개별 인자들의 설명 가능성에
초점을 맞춰서 분석함”

머신러닝 데이터 분석

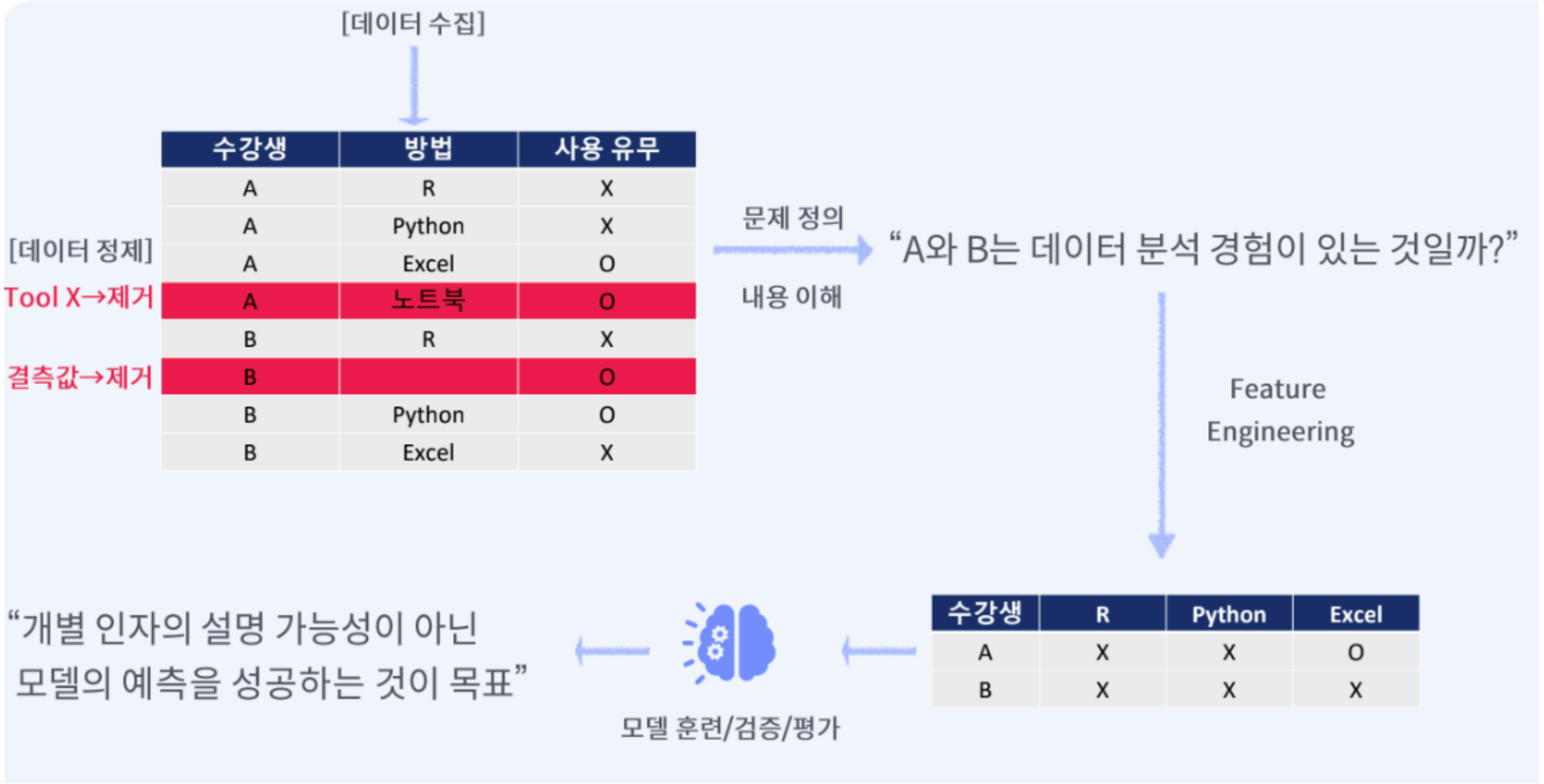
빅데이터와 복잡한 연산이 가능해진 이후에는

분석의 실패 가능성이나 개별 인자의 설명 가능성을 높이는 것이 아닌,

머신러닝 기법을 활용하여 Feature Engeneering 등의

최소한의 데이터 정제를 통해 최대한 많은 인자로 예측을 성공하는데 집중함

머신러닝 데이터 분석


[데이터 수집]
수강생 방법 사용 유무
A R X
A Python X
A Excel O
A 노트북 O
B R X
B O
B Python O
B Excel X
[데이터 정제]
Tool X→제거
결측값→제거
문제 정의
내용 이해
“A와 B는 데이터 분석 경험이 있는 것일까?”
Feature Engineering
수강생 R Python Excel
A X X O
B X X X
모델 훈련/검증/평가
“개별 인자의 설명 가능성이 아닌 모델의 예측을 성공하는 것이 목표”

통계적 데이터 분석 vs 머신러닝 데이터 분석

예제	통계적 데이터 분석	머신러닝 데이터 분석
데이터 분석가 여부 판단	어떤 방법이 데이터 분석가 여부와 가장 상관성이 높은지 분석	데이터 분석가인지 아닌지 예측 정확도에 집중
보험료 예측	고객의 어떤 특징이 보험료 결정에 가장 큰 영향을 주는지 분석	보험료를 정확하게 예측하는 것에 집중
직원 이탈 요인 분석	어떤 이유로 직원들이 이탈하는지 분석	직원들의 특징에 따라 이탈하게 될 직원 수에 집중
	“ 데이터 수집 비용이 높고, 서버 연산 능력이 낮아서, 분석 실패 확률이 낮아야 할 때 유리 ”	“ 데이터 수집 비용이 낮고, 서버 연산 능력이 높아서, 예측 성공 확률이 높아야 할 때 유리 ”