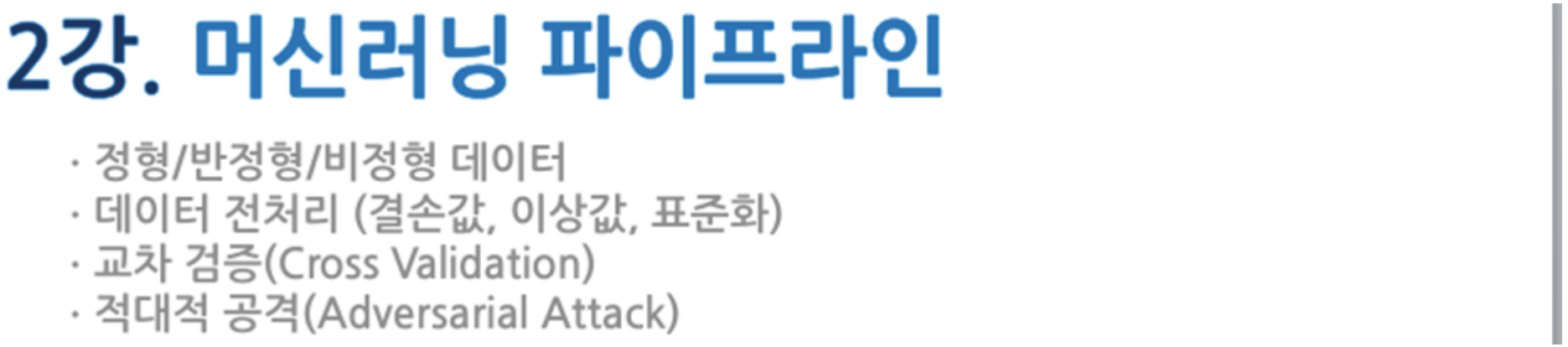
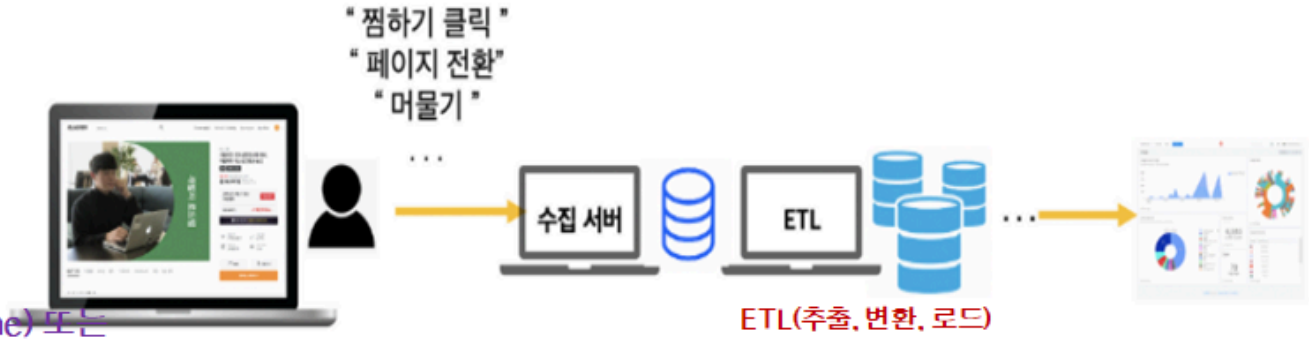
머신러닝 파이프라인-결측치/이상치/표준화/교차검증/적대적공격/AutoML등



머신러닝 파이프라인

머신러닝 전체 과정을 순차적으로 처리하는 일련의 프로세스



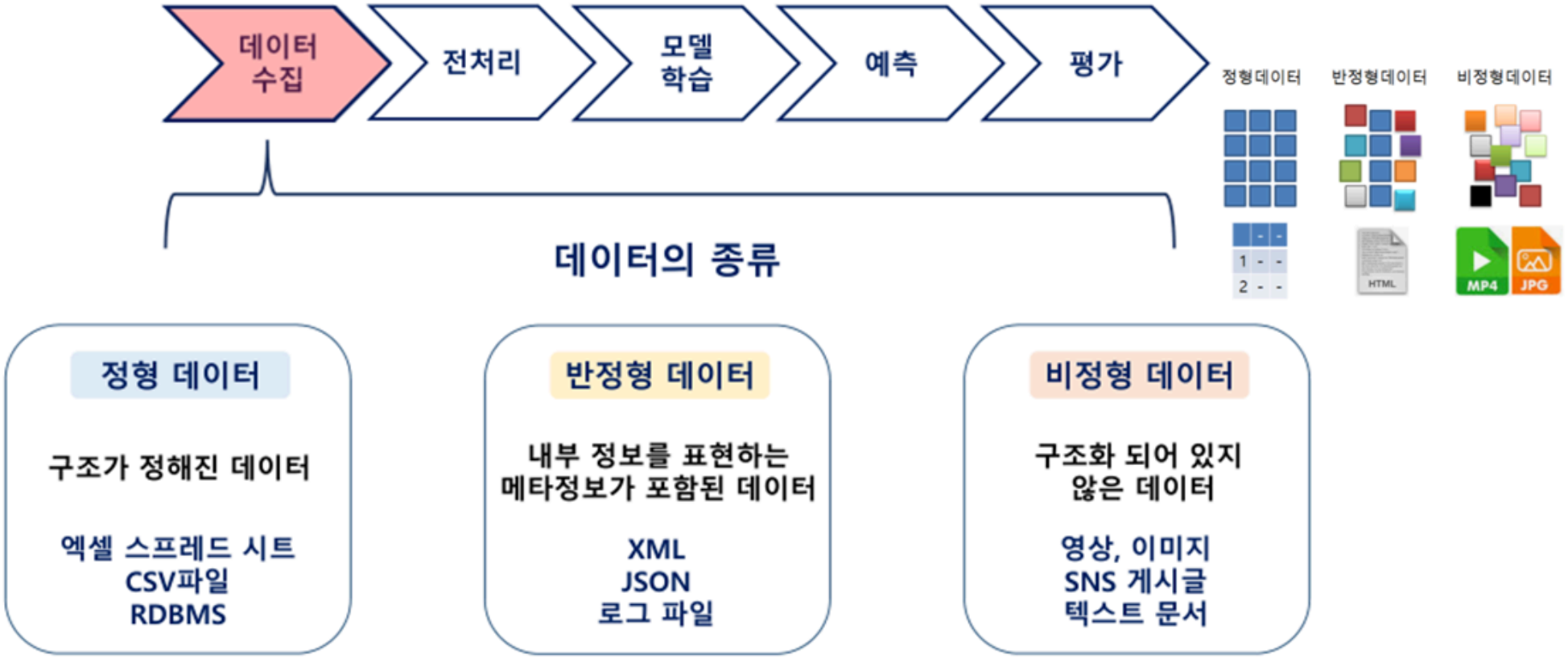
ETL(추출, 변환, 로드)

[데이터가 시각화되기까지]

컴퓨터 과학에서 파이프라인(pipeline) 또는
데이터 파이프라인(data pipeline)

- ✓ 한 데이터 처리 단계의 출력이 다음 단계의 입력으로 이어지는 형태로 연결된 구조를 가리킴

머신러닝 파이프라인


데이터 수집
전처리
모델 학습
예측
평가
정형데이터
반정형데이터
비정형데이터
데이터의 종류
정형 데이터
구조가 정해진 데이터
엑셀 스프레드 시트
CSV파일
RDBMS
반정형 데이터
내부 정보를 표현하는
메타정보가 포함된 데이터
XML
JSON
로그 파일
비정형 데이터
구조화 되어 있지
않은 데이터
영상, 이미지
SNS 게시글
텍스트 문서

머신러닝 파이프라인

- 데이터를 분석할 수 있도록 데이터를 가공하는 작업
- Raw Data를 Clean Data로 만드는 작업
- 데이터 없이는 분석이 불가하기에 대단히 중요한 단계
- Garbage In – Garbage Out

결측값 처리 이상값 처리 표준화

머신러닝 파이프라인

이름	나이	키	몸무게
김OO	25	171	70
이OO	30	175	결측값
박OO	27	180	90

제거

대치 → 80

결측값 처리 이상값 처리 표준화

머신러닝 파이프라인

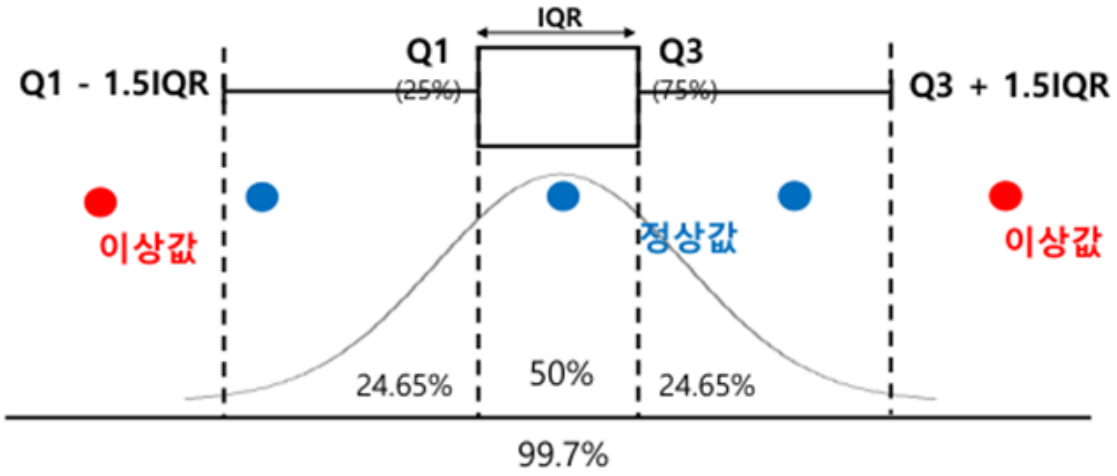
OO마을 주택 가격

A	1000	E	1200
B	1250	F	900
C	950	G	1080
D	1100	H	5000 이상값

결측값 처리 이상값 처리 표준화

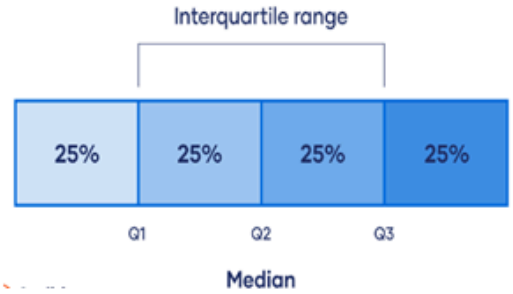
머신러닝 파이프라인

BOX PLOT

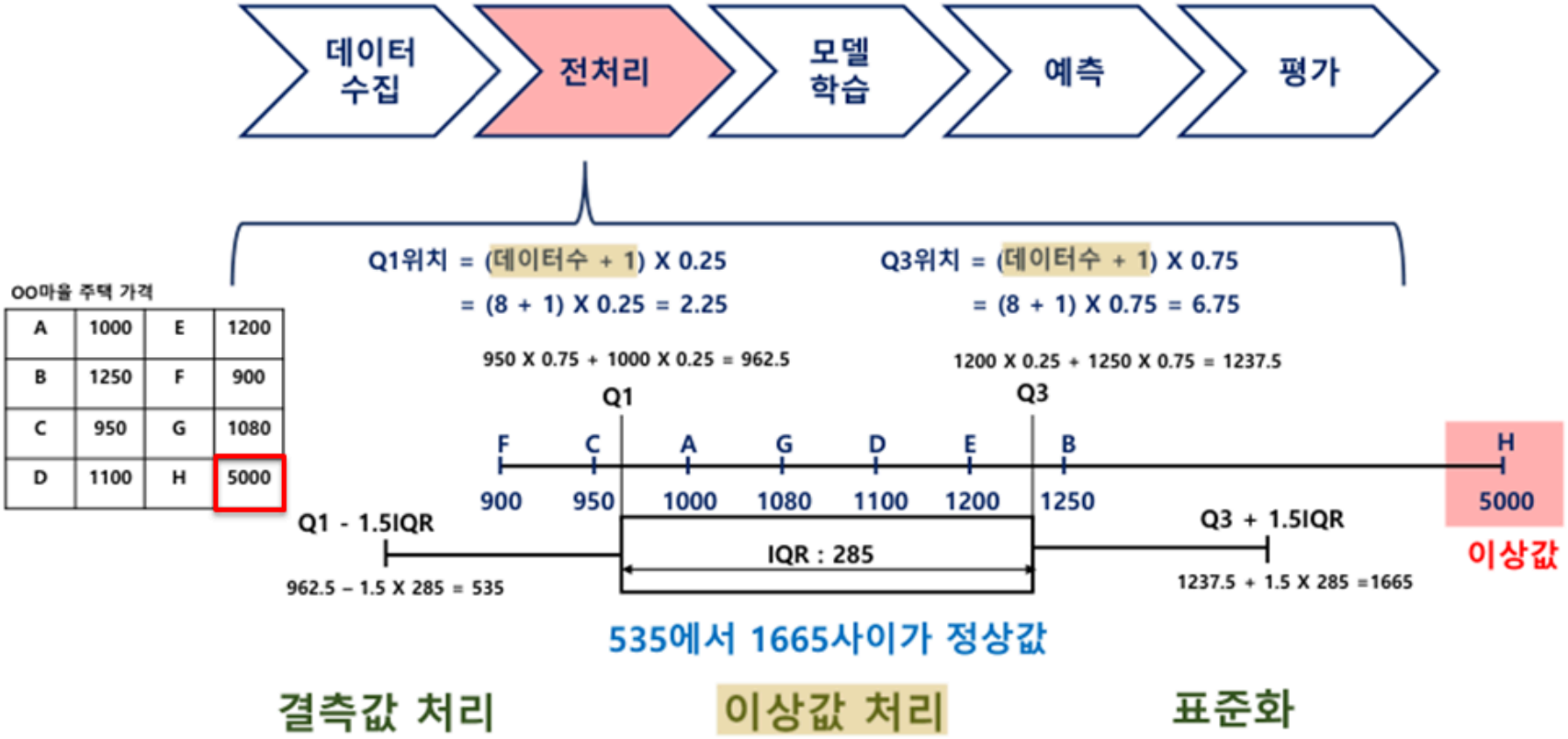


결측값 처리 이상값 처리 표준화

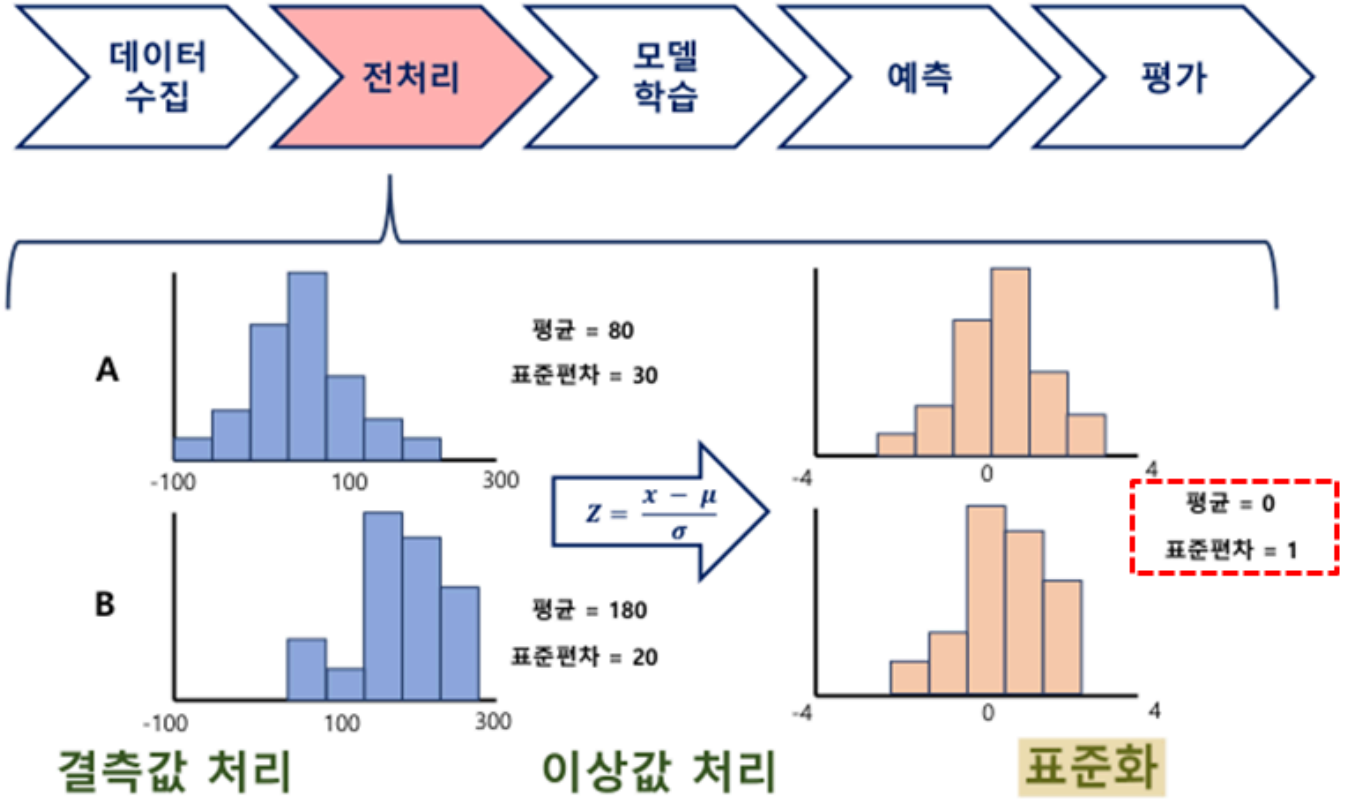
IQR(사분위수 범위) - 데이터 세트를 사분위수로 나누어 표시 여부를 측정한 것



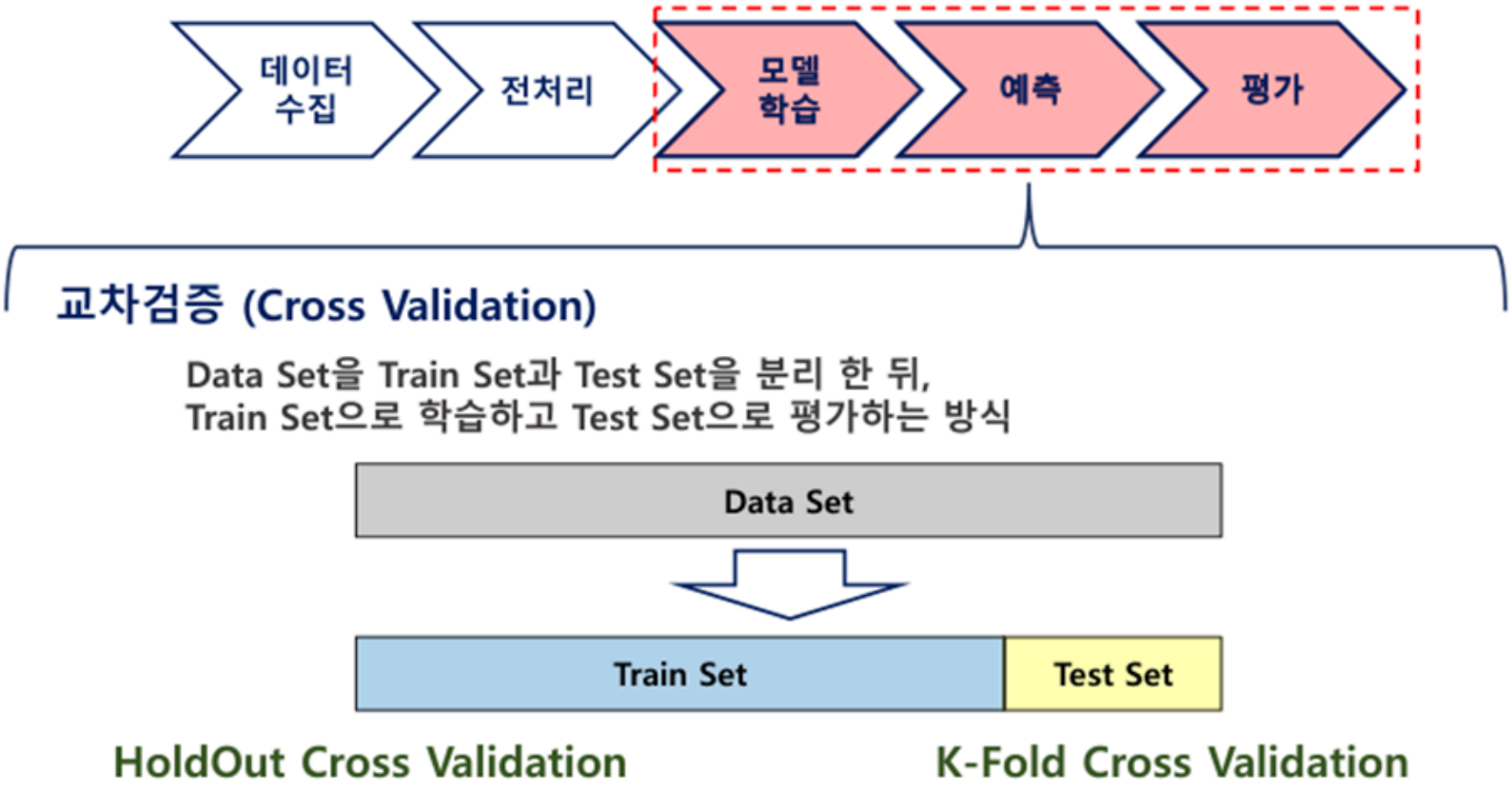
머신러닝 파이프라인


데이터 수집
전처리
모델 학습
예측
평가
Q1위치 = (데이터수 + 1) X 0.25
= (8 + 1) X 0.25 = 2.25
Q3위치 = (데이터수 + 1) X 0.75
= (8 + 1) X 0.75 = 6.75
OO마을 주택 가격
A 1000 E 1200
B 1250 F 900
C 950 G 1080
D 1100 H 5000
950 X 0.75 + 1000 X 0.25 = 962.5
1200 X 0.25 + 1250 X 0.75 = 1237.5
Q1
Q3
F C A G D E B H
900 950 1000 1080 1100 1200 1250 5000
Q1 - 1.5IQR
962.5 – 1.5 X 285 = 535
IQR : 285
Q3 + 1.5IQR
1237.5 + 1.5 X 285 =1665
이상값
535에서 1665사이가 정상값
결측값 처리
이상값 처리
표준화

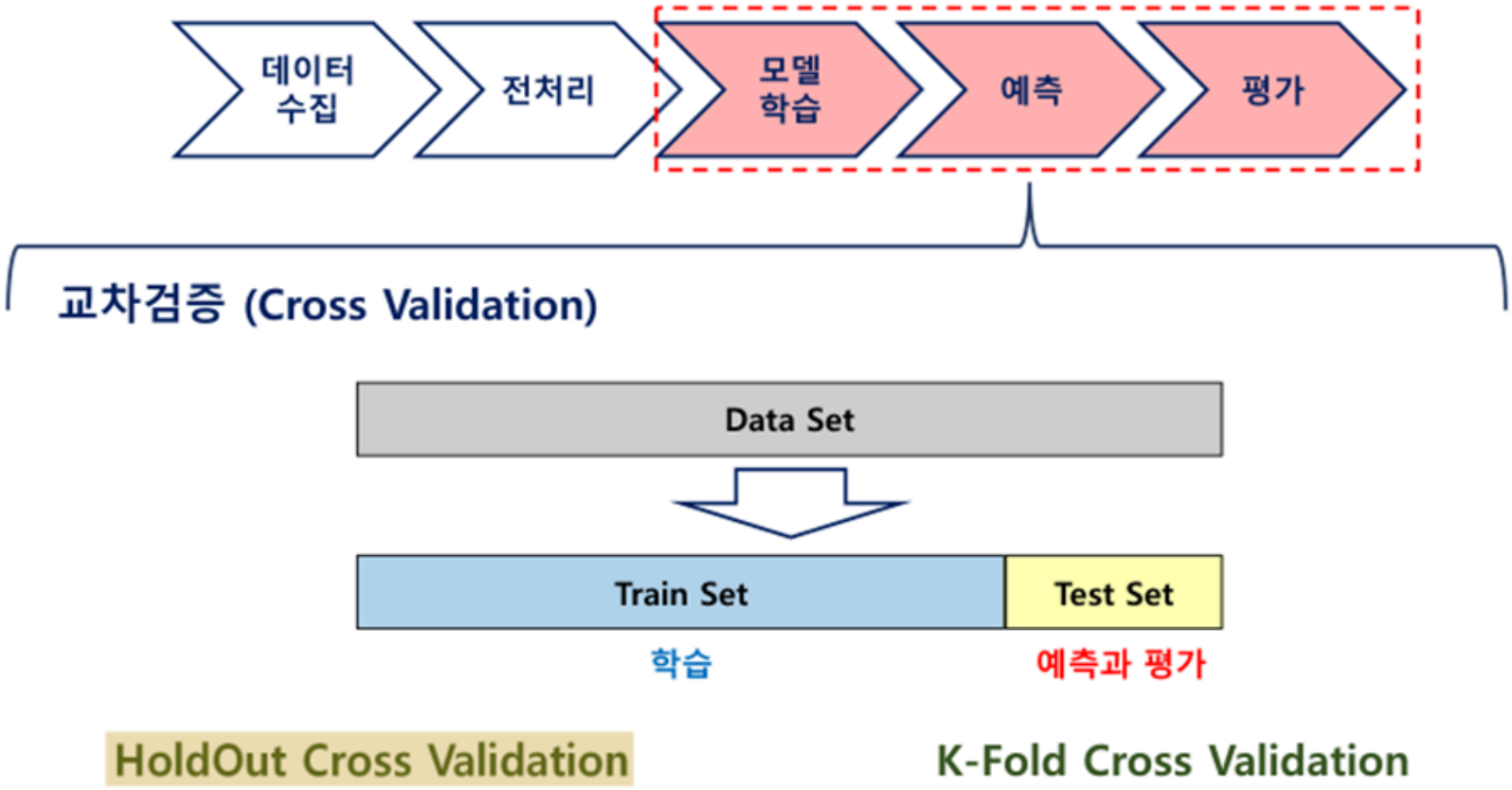
머신러닝 파이프라인


데이터 수집
전처리
모델 학습
예측
평가
A
평균 = 80
표준편차 = 30
-100
100
300
B
평균 = 180
표준편차 = 20
-100
100
300
$Z = \frac{x - \mu}{\sigma}$
-4
0
4
-4
0
4
평균 = 0
표준편차 = 1
결측값 처리
이상값 처리
표준화

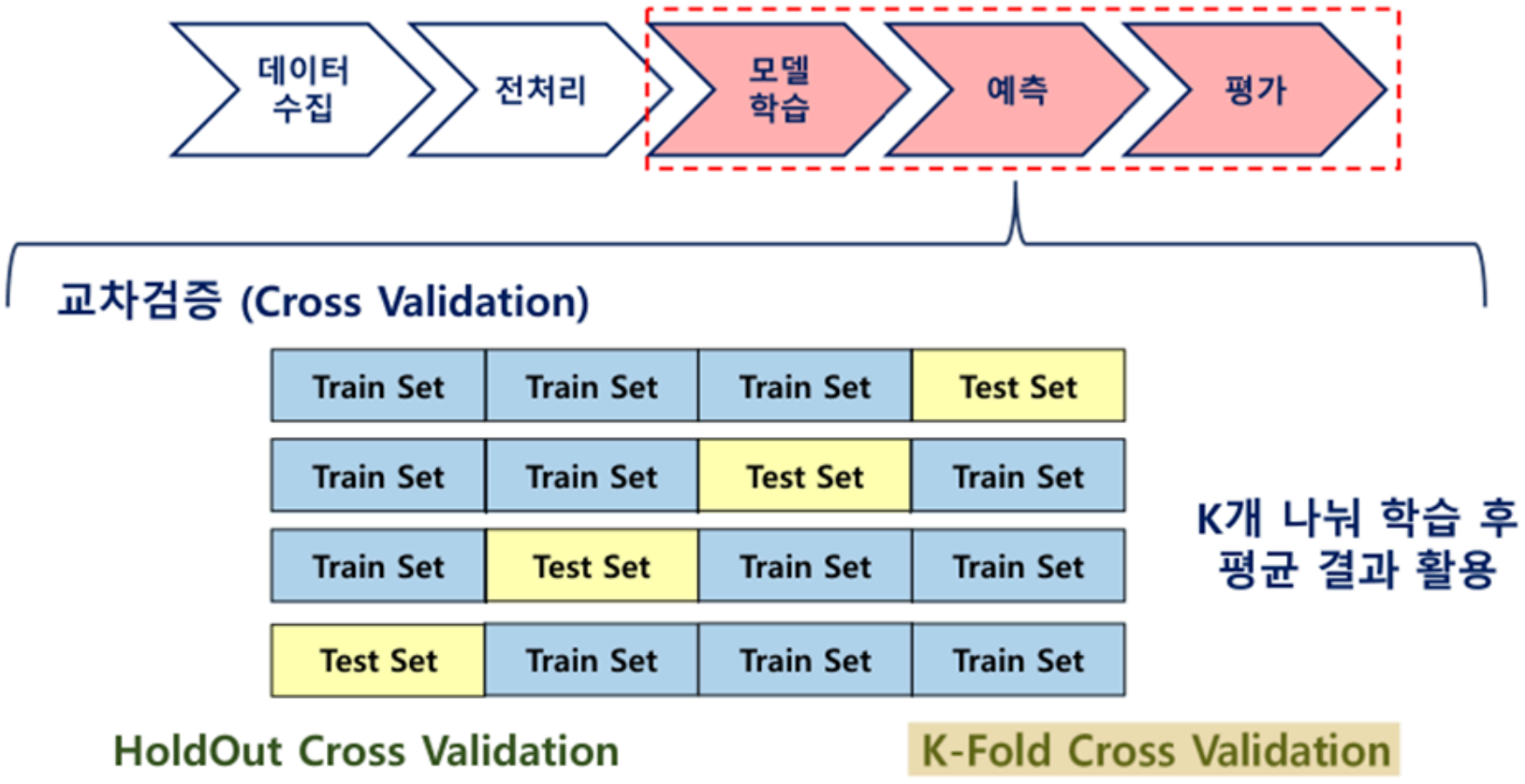
머신러닝 파이프라인


데이터 수집
전처리
모델 학습
예측
평가
교차검증 (Cross Validation)
Data Set을 Train Set과 Test Set을 분리 한 뒤,
Train Set으로 학습하고 Test Set으로 평가하는 방식
Data Set
Train Set
Test Set
HoldOut Cross Validation
K-Fold Cross Validation

머신러닝 파이프라인


데이터 수집
전처리
모델 학습
예측
평가
교차검증 (Cross Validation)
Data Set
Train Set
Test Set
학습
예측과 평가
HoldOut Cross Validation
K-Fold Cross Validation

머신러닝 파이프라인


데이터 수집
전처리
모델 학습
예측
평가
교차검증 (Cross Validation)
Train Set
Test Set
K개 나눠 학습 후 평균 결과 활용
HoldOut Cross Validation
K-Fold Cross Validation

머신러닝 파이프라인

교차검증 (Cross Validation)

1학년 100명 → 25명
2학년 100명 → 25명
3학년 100명 → 25명
4학년 100명 → 25명

Train Set	Train Set	Train Set	Test Set
1학년 100명	2학년 100명	3학년 100명	4학년 100명

학습이 엉망!

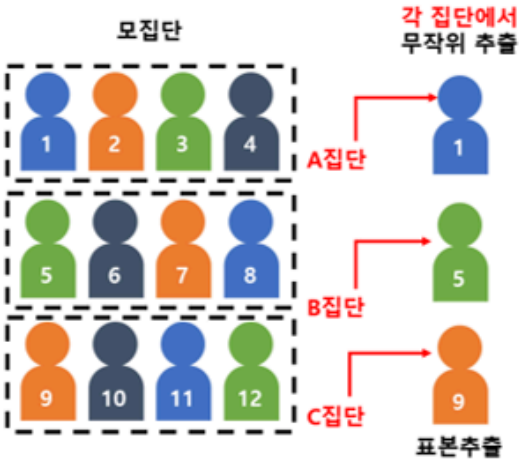
데이터를 어떻게 하면 골고루 분리가 가능할까?

층화추출법 (Stratified random sampling)

층내는 동질적, 층간은 이질적 특성을 가지도록 구성하여 추출

→ 적은 비용으로 더 정확한 모델 학습이 가능

머신러닝 파이프라인



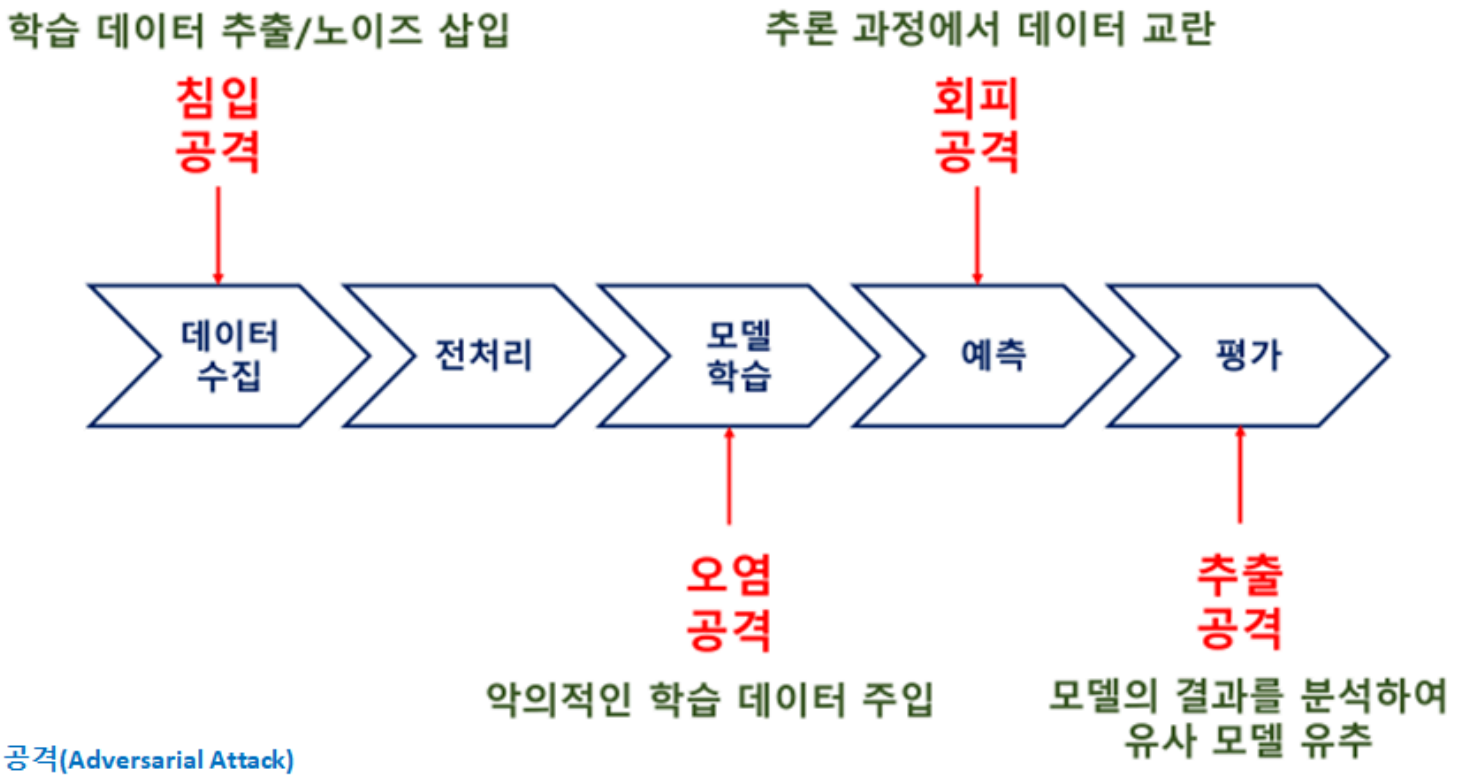
계통추출법

- ✓ 체계적 표집, 체계적 추출법(systematic sampling)
- ✓ 첫 번째 요소는 무작위로 선정한 후 목록의 매번 k번째 요소를 표본으로 선정하는 표집방법
- ✓ 모집단의 크기를 원하는 표본의 크기로 나누어 k를 계산(여기서 k는 표집간격)

층화 추출법

- ✓ 모집단을 먼저 서로 겹치지 않는 여러 개의 층으로 분할한 후, 각 층에서 단순 임의추출법에 따라 배정된 표본을 추출하는 방법
- ✓ 모집단의 분할이 되는 부모집단을 층(stratum)이라고 하고, 각 층에서 임의추출을 하는 표본추출방법
- ✓ *예) 전국 가구를 모집단으로 하는 "생활실태조사"*

적대적 공격(Adversarial Attack)



적대적 공격(Adversarial Attack)

- ✓ 인공지능(AI) 시스템의 취약점을 이용해 의도적으로 오작동이나 잘못된 판단을 유도하는 공격 기법
- ✓ 눈에 띄지 않는 작은 변화를 가한 입력(adversarial examples)을 넣어 모델의 정상 작동을 방해하는 것을 목표로 함

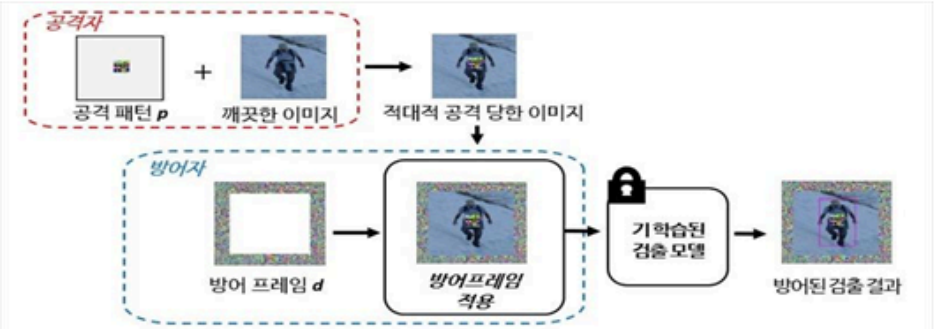
적대적 공격(Adversarial Attack)

적대적 공격 대응방안

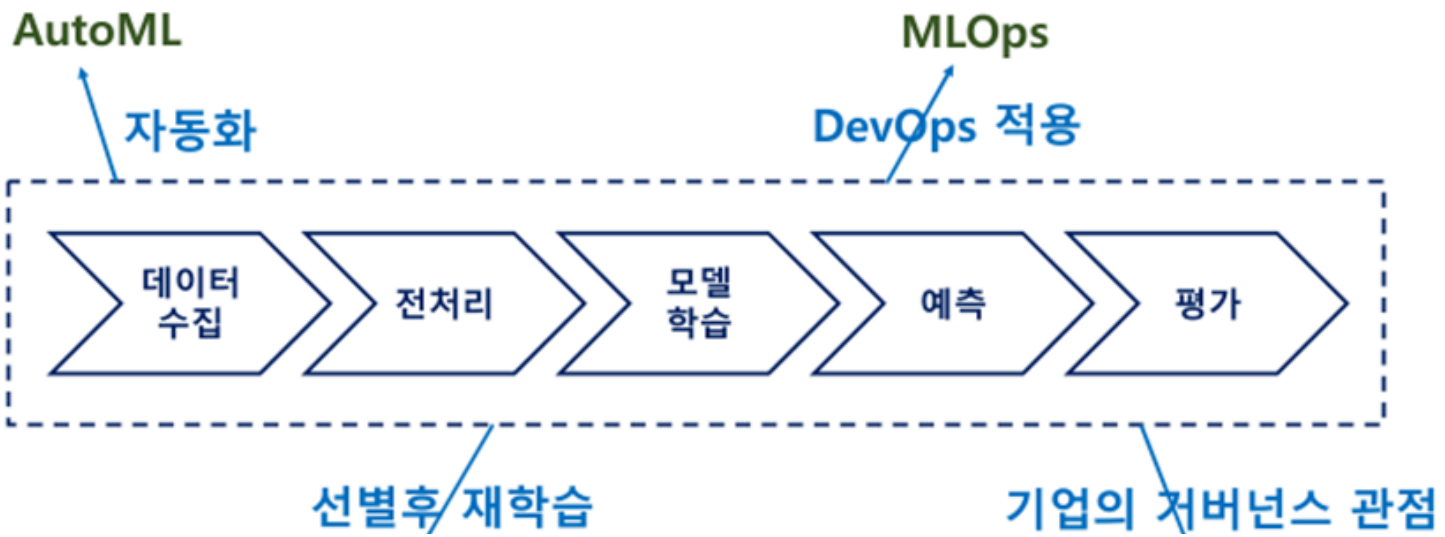
적대적 훈련 가능한 모든 적대적 사례를 포함하여 학습

결과값 차단 결과 값이 노출되지 않도록 하거나, 분석 하지 못하게 변환

탐지 차단 적대적 공격 모델을 추가하여 차이점 비교



머신러닝 파이프라인의 확장



1. What Is AutoML?

1.1 Overview

AutoML (automated machine learning) refers to the automated end-to-end process of applying machine learning in real and practical scenarios.

A typical machine learning model includes the four following steps: