

머신러닝

머신러닝 이해와 파이프라인

머신러닝 이해

인공지능

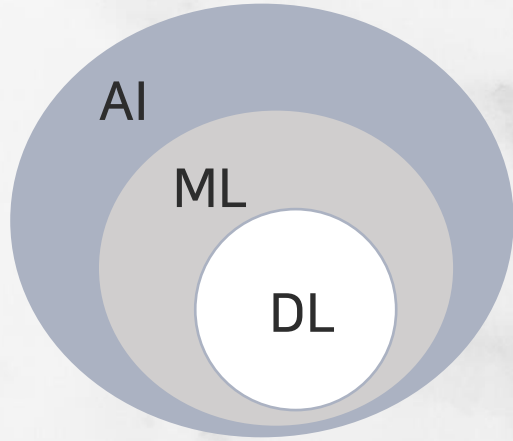
인공지능(Artificial Intelligence)

인간의 학습능력, 추론능력, 지각능력을
인공적으로 구현하려는 컴퓨터 과학의 분야

사람처럼 학습하고 추론할 수 있는
시스템을 만드는 기술

머신러닝과 딥러닝을 포괄하는 종합적인 분야

인공지능



인공지능(Artificial Intelligence)

Artificial Intelligence



사람이 해야 할 일을
기계가 대신할 수 있는
자동화

Machine Learning



데이터로부터
의사결정을 위한
패턴을 기계가
스로 학습

Deep Learning



인공신경망
기반의 모델로 비정형
데이터로부터 특징
추출 및 판단까지
기계가 한번에 수행

머신러닝



머신러닝(Machine Learning)



인공지능의 한 분야로 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야임



전통적인 방식은 주어진 문제를 해결하기 위해 사람이 규칙을 프로그래밍하여 정답을 예측함



머신러닝은 데이터를 학습하여 스스로 규칙을 찾아내서 정답을 예측함

머신러닝

💡 머신러닝(Machine Learning)

⚙️ 전통적인 프로그램 예측 방식



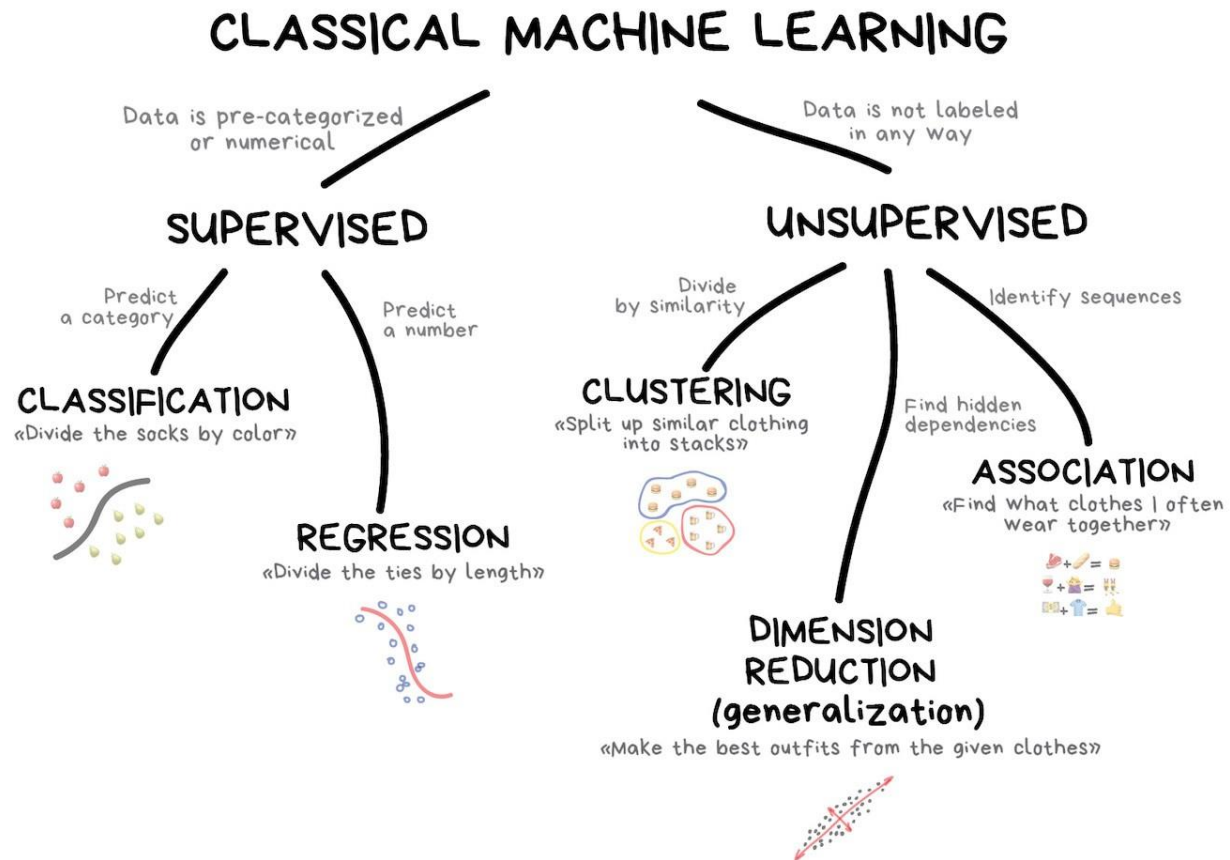
⚙️ 머신러닝



머신러닝의 학습방식

💡 머신러닝의 분류

머신러닝은 학습 방식에 따라 지도 학습, 비지도 학습으로 분류



머신러닝의 학습방식



지도 학습(Supervised Learning)

모델(알고리즘)에 입력 값(특성)과 출력 값(정답)을 같이 넣어 학습시키는 방식

분류
(Classification)

- 예측하고자 하는 정답 값이 범주형 데이터인 경우
- Binary, Multi-class, Multi-label

회귀
(Regression)

- 예측하고자 하는 정답 값이 수치형 데이터인 경우

머신러닝의 학습방식

지도 학습(Supervised Learning)

체질량 지수	가족력	콜레스테롤	...	정상여부
18.3	없음	100.7	...	정상
22.1	없음	98.7	...	정상
25.5	있음	190.9	...	비정상
17.1	없음	150.7	...	정상
28.5	있음	140.1	...	비정상
16.5	있음	111.2	...	정상
18.3	없음	190.7	...	정상
20.5	있음	190.9	...	비정상
17.3	없음	170.71	...	정상
18.3	없음	160.7	...	정상

머신러닝의 학습방식

비지도 학습(Unsupervised Learning)

정답이 없는 데이터를 학습하는 방식

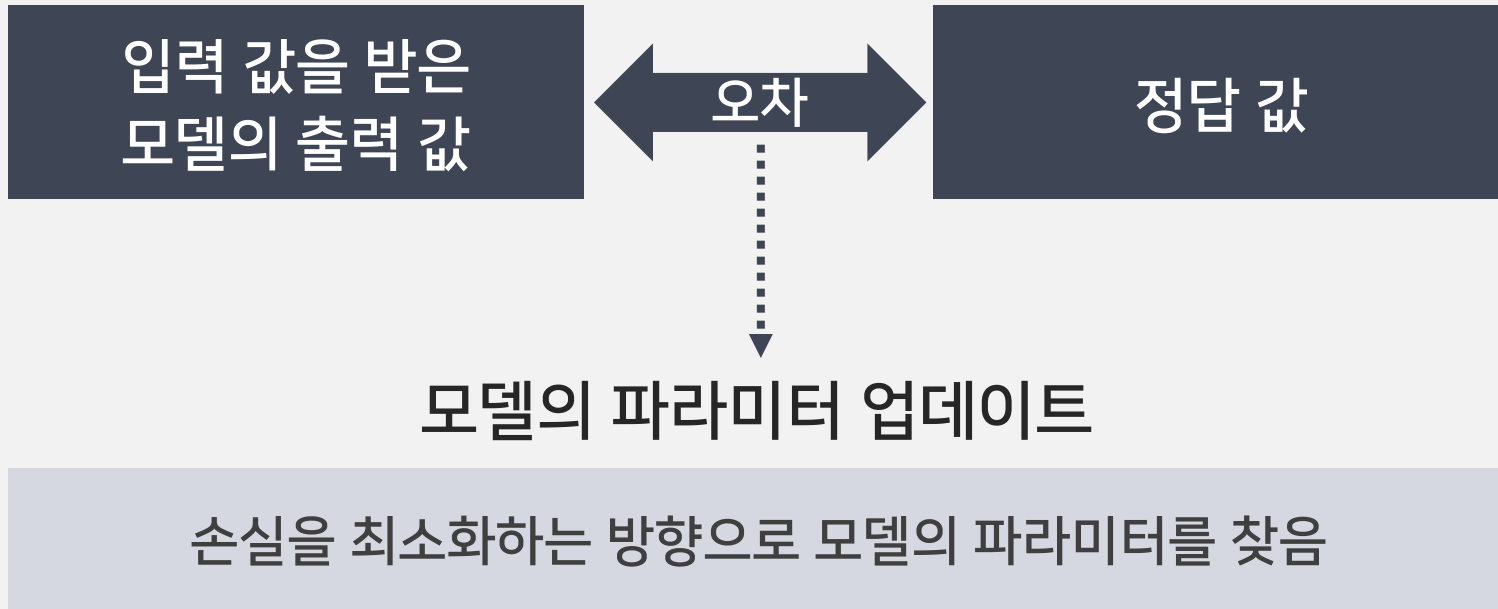
군집화(Clustering)

- 주어진 데이터를 유사한 데이터들의 그룹으로 나누는 것

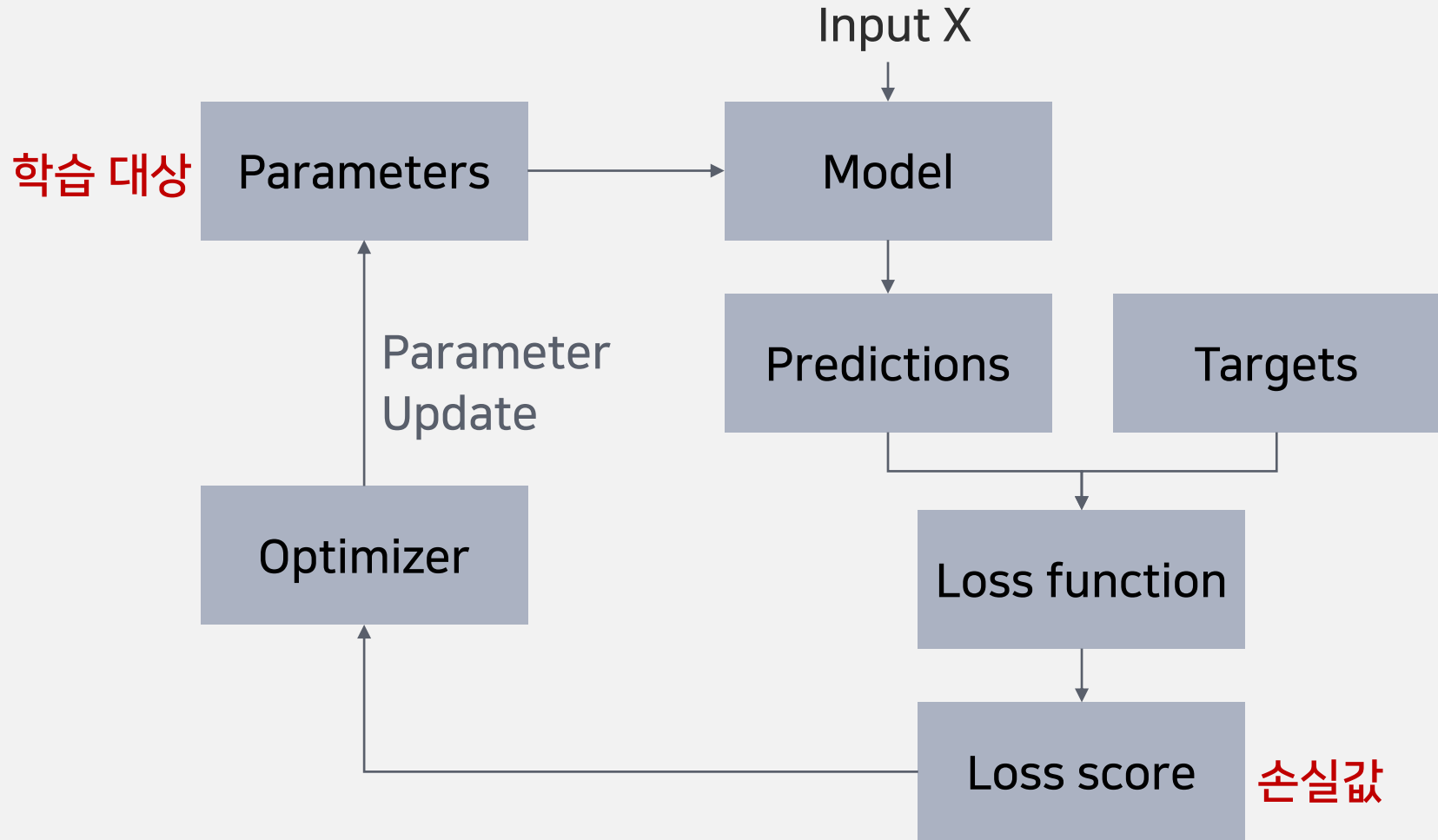
차원 축소
(Dimensionality
reduction)

- 고차원의 데이터를 저차원의 데이터로 변환하는 방법

머신러닝 학습 원리

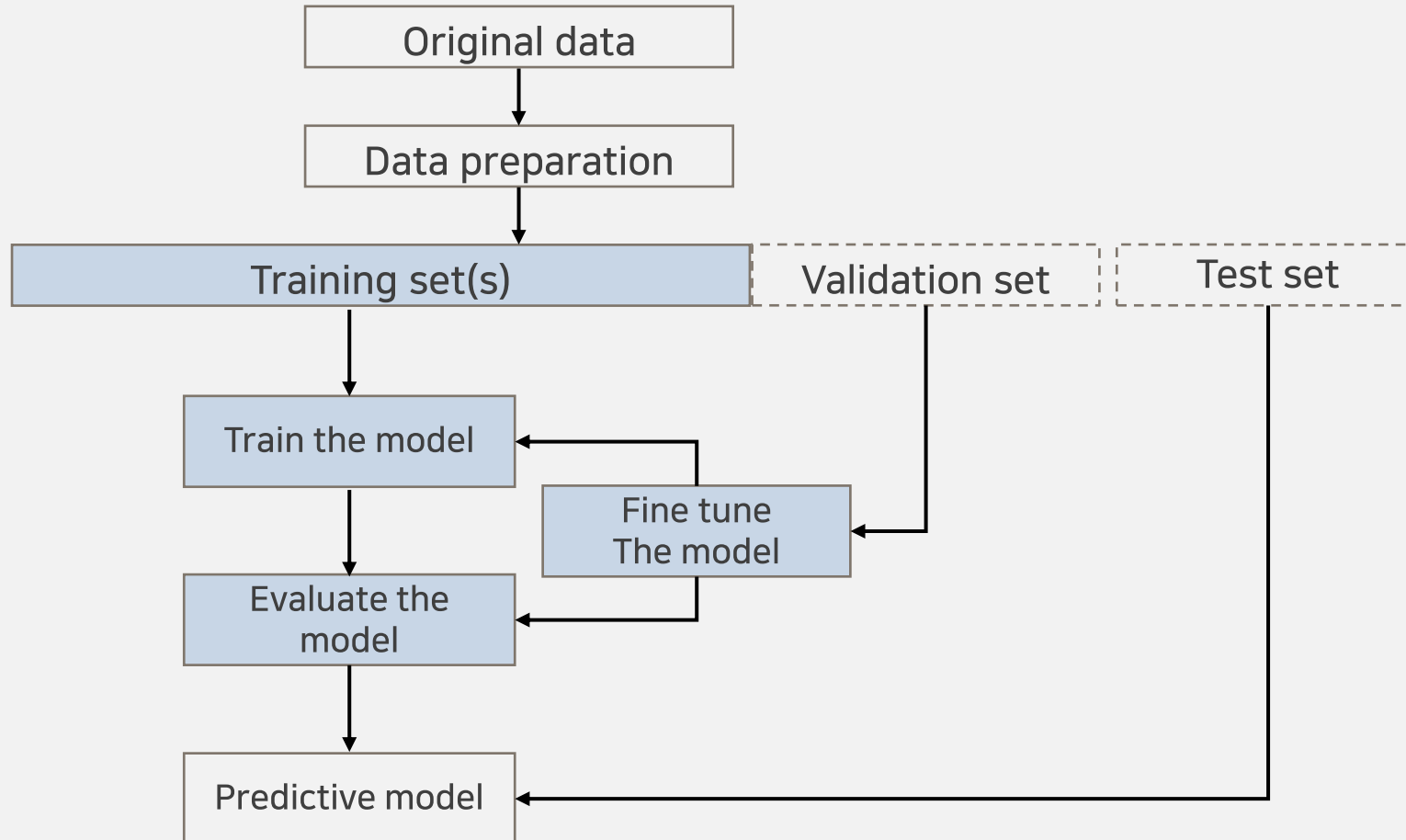


머신러닝 학습 원리



머신러닝 워크플로우(Workflow)

데이터를 전처리하고 모델 학습 및 성능을 평가하는 과정



머신러닝 기초용어

1 Feature, 독립변수, 설명변수

▹ 학습데이터의 특성

2 class, label, target, 종속변수

▹ 정답 데이터

체질량 지수	가족력	콜레스테롤	...	정상여부
18.3	없음	190.7	...	정상
...
20.5	있음	101.9	...	비정상

Feature

↓
Label

머신러닝 기초용어

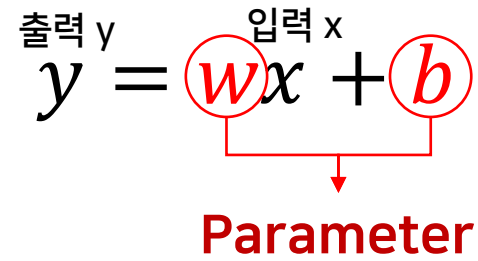
3 Parameter

◀ 모델이 학습과정에서 업데이트하는 파라미터

선형회귀 모델 예시

$$\text{출력 } y = \text{입력 } x \cdot w + b$$

Parameter



4 Hyper parameter

◀ 사용자가 직접 설정하는 파라미터

머신러닝 기초용어

5 Loss, 손실

◀ 정답 값과 예측 값의 오차를 표현하는 지표

6 Metrics, 평가지표

◀ 모델의 성능을 평가할 때 사용하는 지표

머신러닝 파이프라인 이해

머신러닝 파이프라인



머신러닝 파이프라인(ML Pipeline)



머신러닝 기술을 활용함에 있어서 초기 기획부터 데이터 수집·가공, 분석과 사후관리까지 일련의 전체 과정



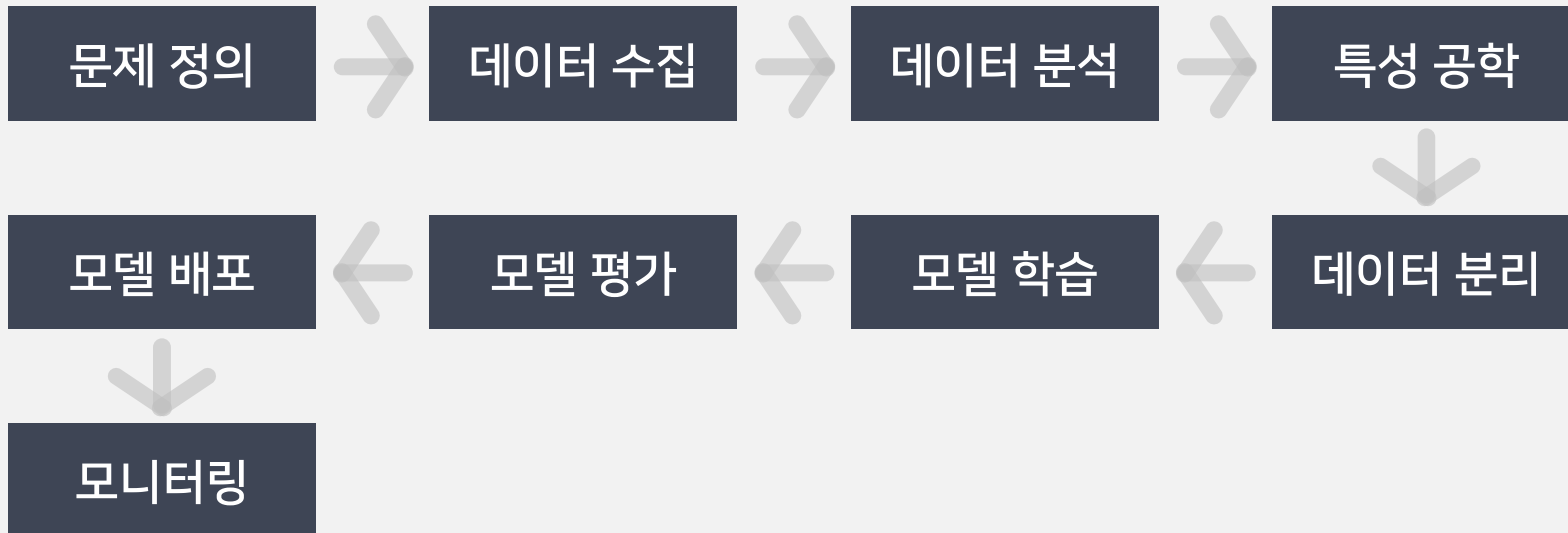
문제 정의부터 데이터 수집, 전처리, 학습, 모델 배포, 모니터링까지 전 과정을 순차적으로 처리하도록 설계된 머신러닝 아키텍처

머신러닝 파이프라인

💡 머신러닝 파이프라인(ML Pipeline)

⚙️ 파이프라인이란?

한 데이터 처리 단계의 출력이 다음 단계의 입력으로 이어지는 형태로 연결된 구조



문제 정의

비즈니스 목적에 맞게 문제를 구체화하는 단계

머신러닝
타당성 확인

- 머신러닝 모델을 이용하여 어떠한 이익을 얻을 수 있는지 등을 파악

데이터 수집 방안
정의

- 내부데이터가 있는지에 대한 여부와 외부데이터 수집 가능 여부 등을 파악

지도학습 or
비지도 학습

- 레이블 되어 있는 데이터가 있는지에 대한 여부 등을 파악하여 학습 유형을 정의

문제 정의

비즈니스 목적에 맞게 문제를 구체화하는 단계

회귀 or 분류

- 예측하고자 하는 타겟값에 따라 회귀인지 분류인지 정의

성능 측정 지표
선택

- 적절한 평가지표를 선택



데이터 수집

데이터 수집(Data Collection)

주어진 문제를 해결하기 위한 데이터 수집 및 처리하는 단계

데이터 구조 확인



테스트 세트를 생성하기 위해 데이터 구조를 파악



학습 및 평가에 사용할 수 없는 데이터를 제거

데이터 수집

데이터 수집(Data Collection)

테스트 세트 생성



모델 평가를 위해 테스트 세트를 생성



일반적으로 20~30% 비율의 데이터를 테스트 세트로 분리



테스트 세트에 대해서는 절대 EDA를 진행하면 안됨

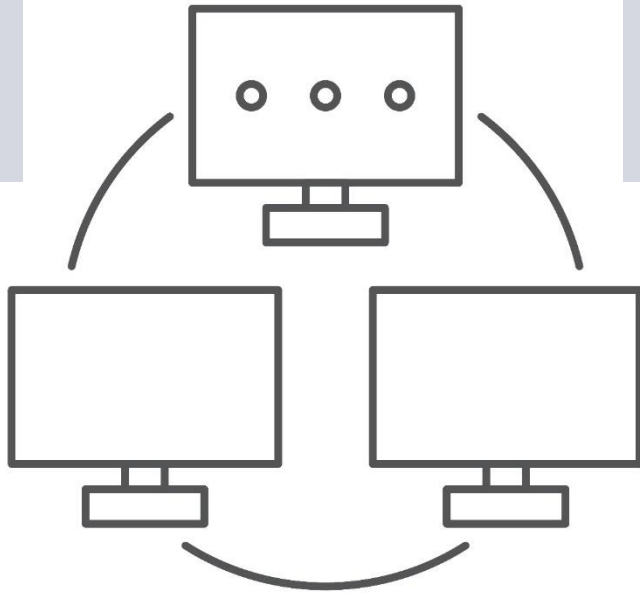


데이터 분석, 특성 공학

💡 데이터 분석(Data analysis)

데이터 이해를 위한
탐색과 시각화

학습 데이터에
대해서만 탐색을
진행



데이터 분석, 특성 공학

💡 특성 공학(Feature Engineering)



데이터 정제

이상치 제거,
결측치 처리 등

범주형 특성 인코딩 (Feature Encoding)

범주형 특성을
숫자형태로 변환

특성 스케일링 (Feature Scaling)

모든 특성의 범위를
같도록 만들어 주는
방법



데이터 분리, 모델 학습

💡 데이터 분리(Data Split)

모델 평가 전에 모델의 성능을 검증하는 검증 세트 생성

대표적인 데이터 분리 방식



Hold-out

A circular diagram representing a dataset. A thick, dark gray arc on the right side of the circle indicates the portion of the data that is held out for testing. The rest of the circle is a thin, light gray line.



K-fold

A circular diagram representing a dataset. A thick, dark gray arc on the right side of the circle indicates one of the K folds used for testing. The rest of the circle is a thin, light gray line.

데이터 분리, 모델 학습

모델 학습(Train Model)

모델 선택과 훈련



다양한 머신러닝 모델을 활용하여 학습을 진행

검증 세트를 이용하여 모델 평가



학습된 모델이 좋은 성능을 보일 수 있는지를 검증



모델 평가, 모델 배포, 모니터링

모델 평가 (Model Evaluation)

- 테스트 세트를 이용하여 모델 평가
- 테스트 세트에 대한 평가가 좋지 못하면 이전 단계들로 돌아감

모델 배포 (Model Serving)

- 학습된 모델을 시스템에 적용
- REST API를 통해 질의할 수 있는 전용 웹 서비스 등에 학습된 모델을 배포

모니터링 (Monitoring)

- 모델 실전 성능 모니터링
- 모델 성능이 감소하는 상황이 감지되면 이전 단계들로 돌아감

