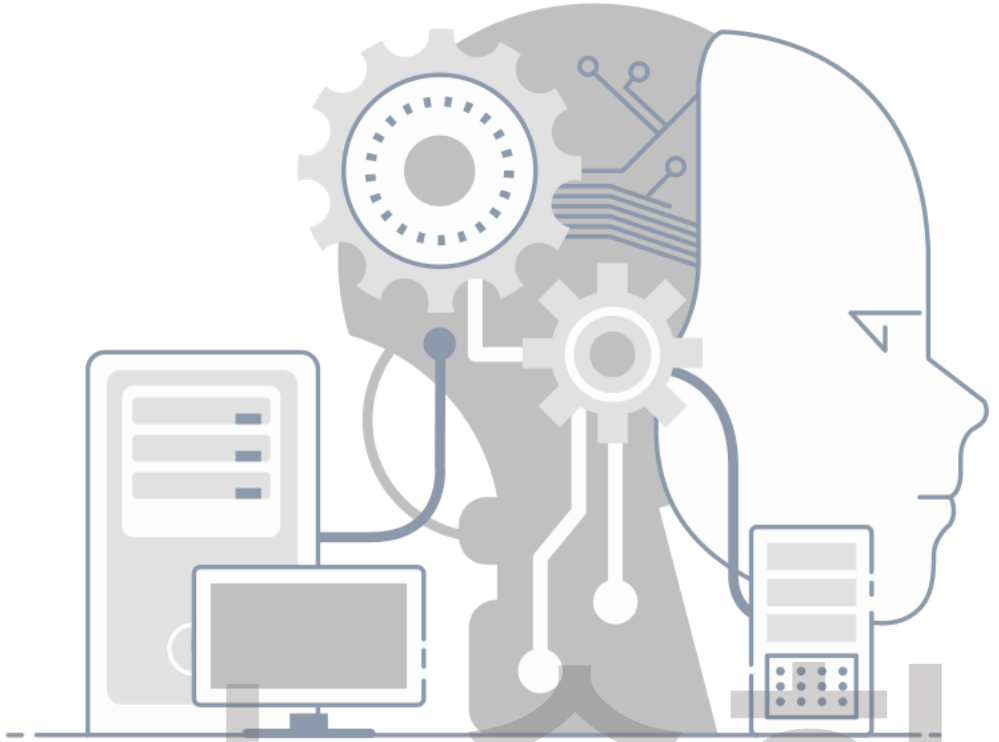
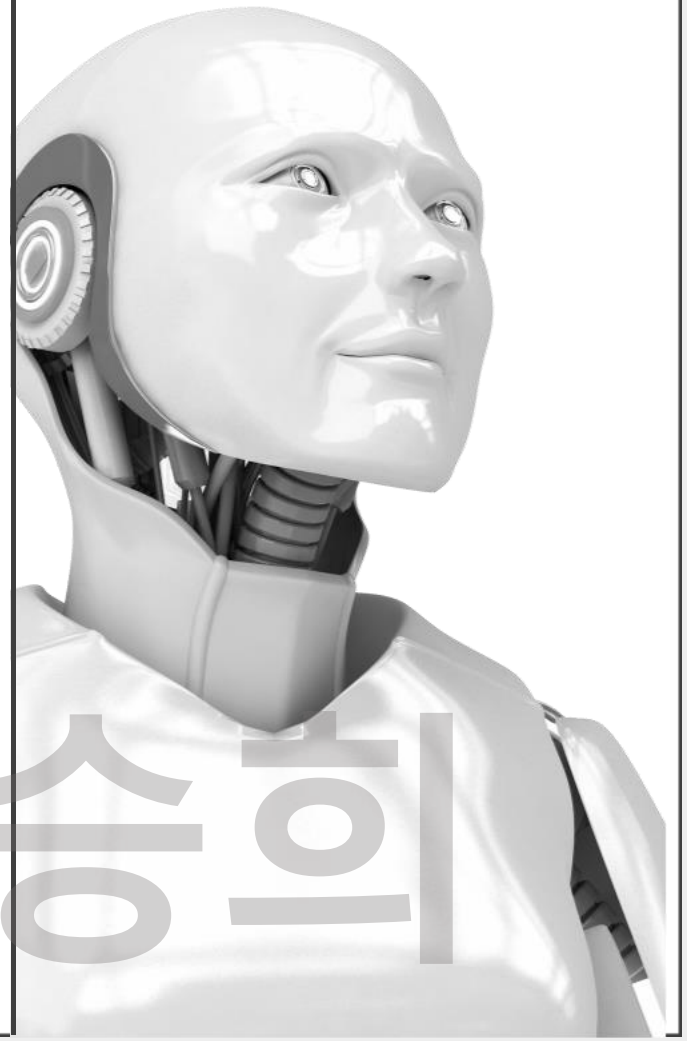
2022 DX Camp

머신러닝 강의 소개



01 강의 개요

교육명	**개발자를 위한 머신러닝 실습**
교육 목표	1. 머신러닝이 익숙하지 않은 개발자를 위한 과정. 2. 머신러닝 워크플로우를 이해하고 프로젝트 설계 및 이해가 가능한 실무자 양성.
교육 대상	DX Camp 참가자 중 수강 희망자
일 정	3시간 x 주 2회 x 2.5주 = 15시간



02 강의 커리큘럼

1강 머신러닝 개요

2강 머신러닝 프로세스

3강 회귀

4강 분류

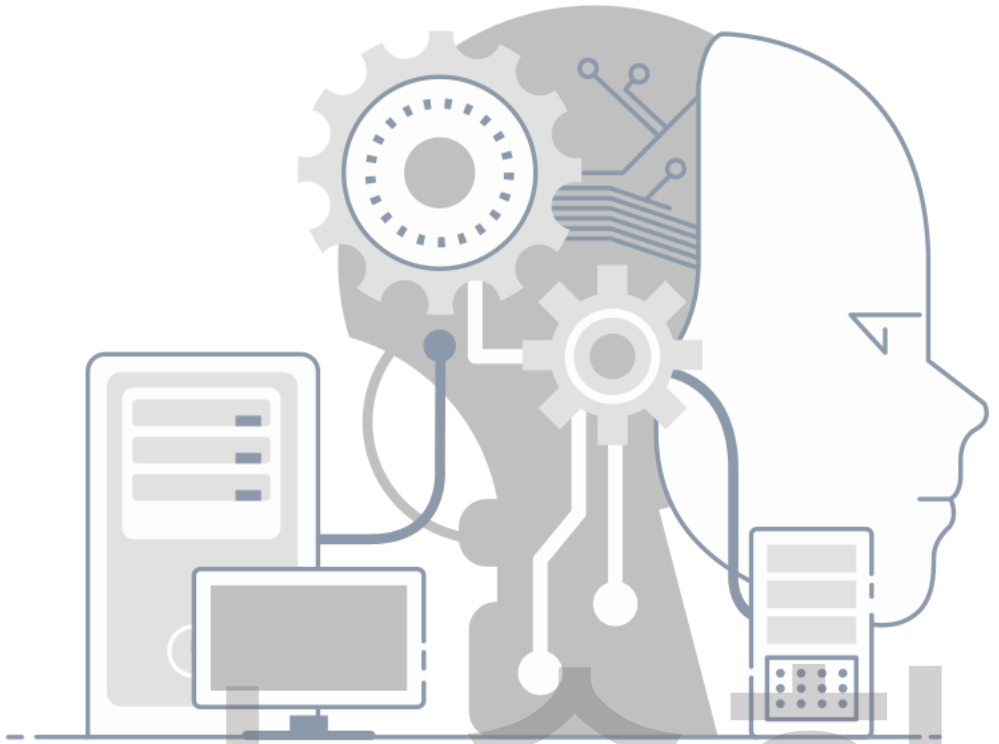
5강 앙상블

6강 추천 시스템

※ 각 강의 당 3시간씩 총 5회 교육이 진행 됩니다.

2022 DX Camp

1강 머신러닝 개요



개발자 머신러닝 시작하기

- 많은 머신러닝 자료들은 연구관점에서 시작한다.
  → 퍼셉트론부터 역전파, 최적화 이론들을 전부 학습한 후에 모델과 코드를 학습한다.

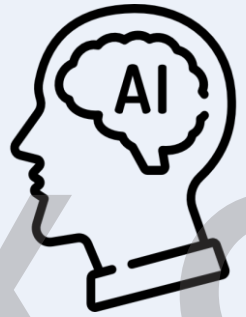
- 수학을 잘 아시는 분에게는 적합할지 몰라도, 비전공자와 수학을 모르는 개발자 입장에서는 진입장벽은 높아만 보일 것이다.

- 게임을 개발하기 위해 우리는 3D 그래픽 지식을 먼저 습득하지 않으며, 시스템에 보안을 도입하기 위해 암호학을 공부하지 않는다.

인공지능, 머신러닝, 딥러닝의 관계

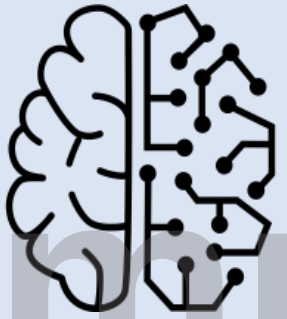
인공지능

사고나 학습 등 인간이 가진 지적 능력을 컴퓨터를 통해 구현하는 기술



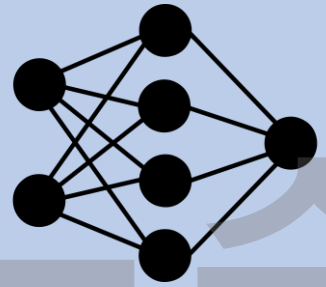
머신러닝

컴퓨터가 스스로 학습하여 인공지능의 성능을 향상 시키는 기술 방법



딥러닝

인간의 뉴런과 비슷한 인공 신경망 방식으로 정보 처리



머신러닝, 딥러닝의 차이

통계 기반 머신러닝

SVM

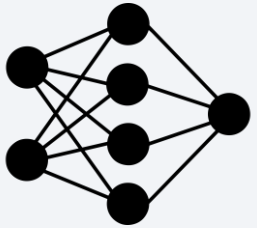
Naive Bayes

kNN

Linear Regression

Decision Tree

VS



뉴럴 네트워크

CNN

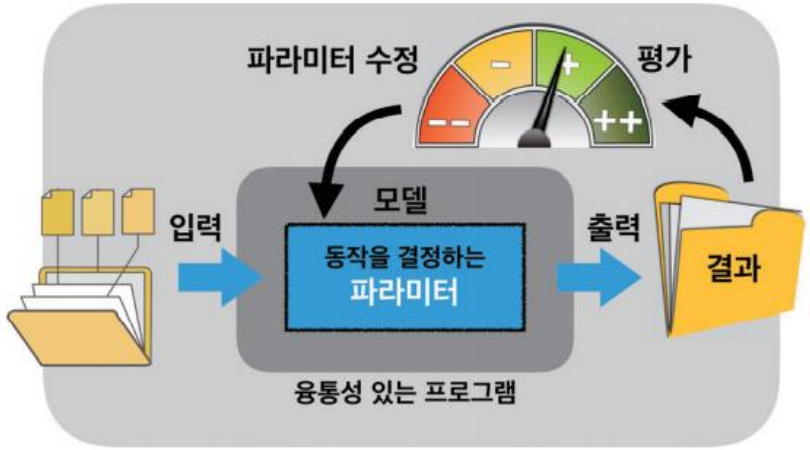
RNN

GAN

LSTM

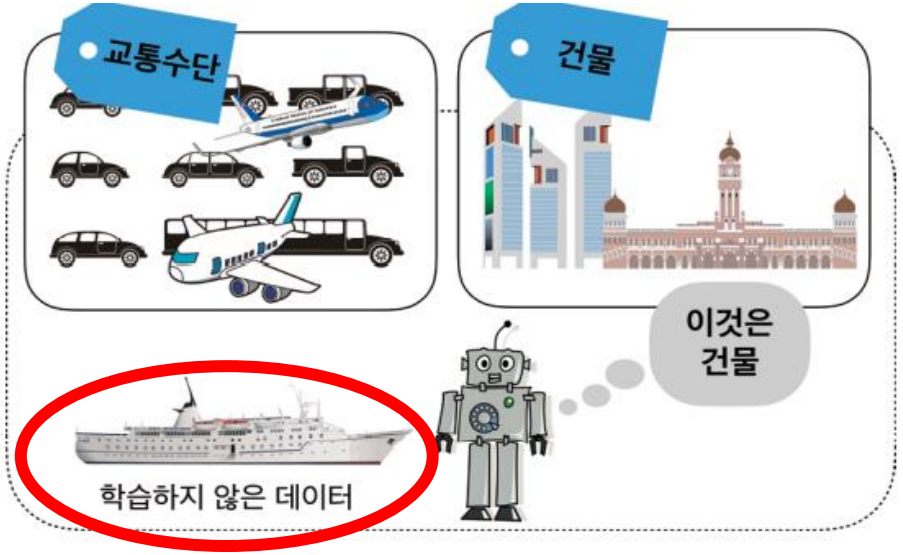
맹목적 딥러닝 사용은 지양하자!

머신러닝은 무엇을 하려는 것인가?



- 머신러닝은 **기계**machine 가 **학습**learning을 하는 것이다.
- 머신러닝을 위해 기계에 수학적 알고리즘을 구축해야 하는데 이것을 **모델**model이라고 부른다.
- 머신러닝 모델은 내부적으로 변경 가능한 **파라미터**parameter에 의해 동작이 결정된다.
- 좋은 동작이 나오도록 파라미터를 변경하는 일을 하는데, 이 과정을 **학습**learning이라고 부른다.

머신러닝의 핵심



학습 데이터가 충분하지 않을 경우
머신러닝은 좋은 성능을 낼 수 없다.

- 머신러닝의 핵심적인 문제는 **알고리즘**과 **데이터**라고 할 수 있다.
- **알고리즘**은 생각보다 중요하지 않다. 하지만, **데이터**에 대한 이해는 반드시 필요하다.
- 데이터 분석을 위해서는 해당 프로젝트에 대한 도메인 지식을 습득하고 주제에 알맞은 데이터를 사용해야한다.
- 참고 사이트: https://www.youtube.com/watch?v=20PlFERKCyo&t=11s

- **데이터 편향**data bias : 확보된 데이터가 대표하는 모집단의 분포를 제대로 반영하지 못하고 일부의 특성만을 가지고 있는 경우

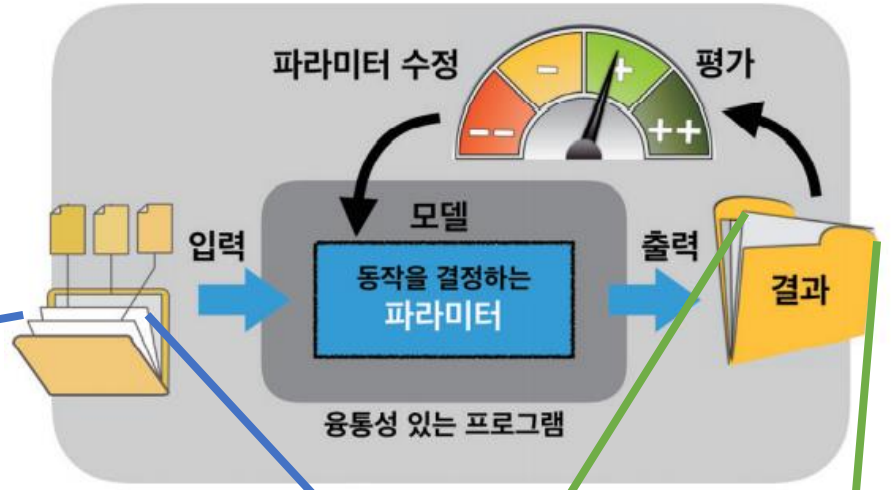
  * 편향의 원인
    - 너무 적은 수의 표본을 추출한 경우
    - 표집 방법이 잘못되어 모집단에 속한 대상을 골고루 추출하지 못 하는 경우.

- **부정확성**inaccuracy : 데이터의 품질이 낮아 많은 오류와 이상치, 잡음을 포함하고 있는 경우

- **무관함**irrelavance : 데이터는 많이 확보했지만, 이 데이터가 담고 있는 특성들이 학습하려고 하는 문제와는 무관한 데이터

머신러닝에서 데이터의 중요성은 아무리 강조해도 지나치지 않다!

머신러닝 학습 방법



X1	X2	X3	X4	X5	X6	Y
19	10.2	182	160	1000.2	0.1	30.5
30	10.8	174	162	2108.0	0.1	12.2
…	…	…	…	…	…	…
21	9.8	168	159	2100.0	0.1	40.3
12	11.2	188	163	1000.0	0.2	40.5

X(입력값) : 독립변수, 입력변수

Y(결과값) : 종속변수, 출력변수

머신러닝 학습 방법

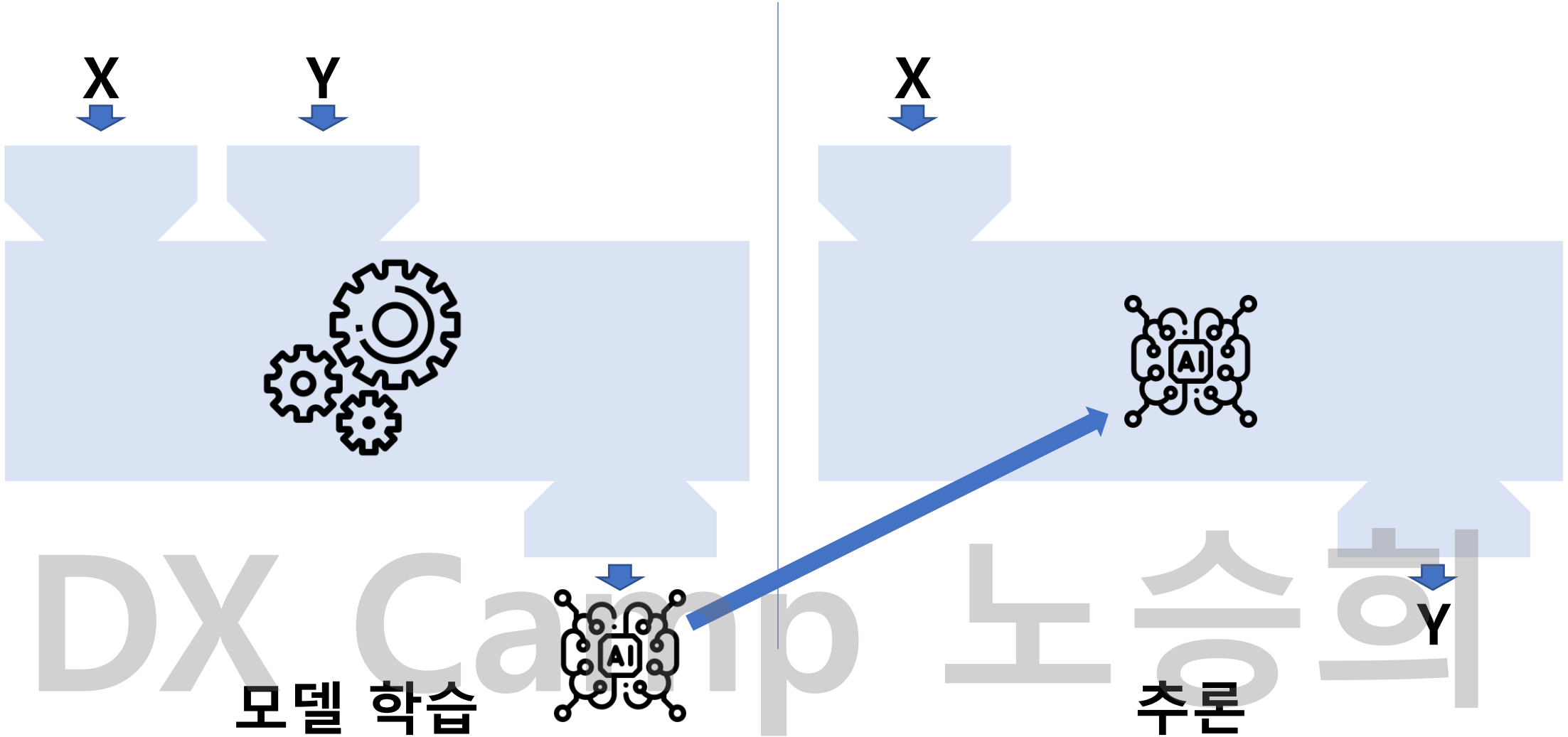
아버지 키	어머니 키
182	160
174	162
…	…
168	159
188	163

자식 키
174.5
172.2
…
168.3
177.5

X(입력값) : 독립변수, 입력변수

Y(결과값) : 종속변수, 출력변수

머신러닝 학습 방법



머신러닝 학습 방법

X1	X2	Y
0	3	6
1	2	9
2	4	18

학습

Y = ?X1 + ?X2

(Y = 5X1 + 2X2)

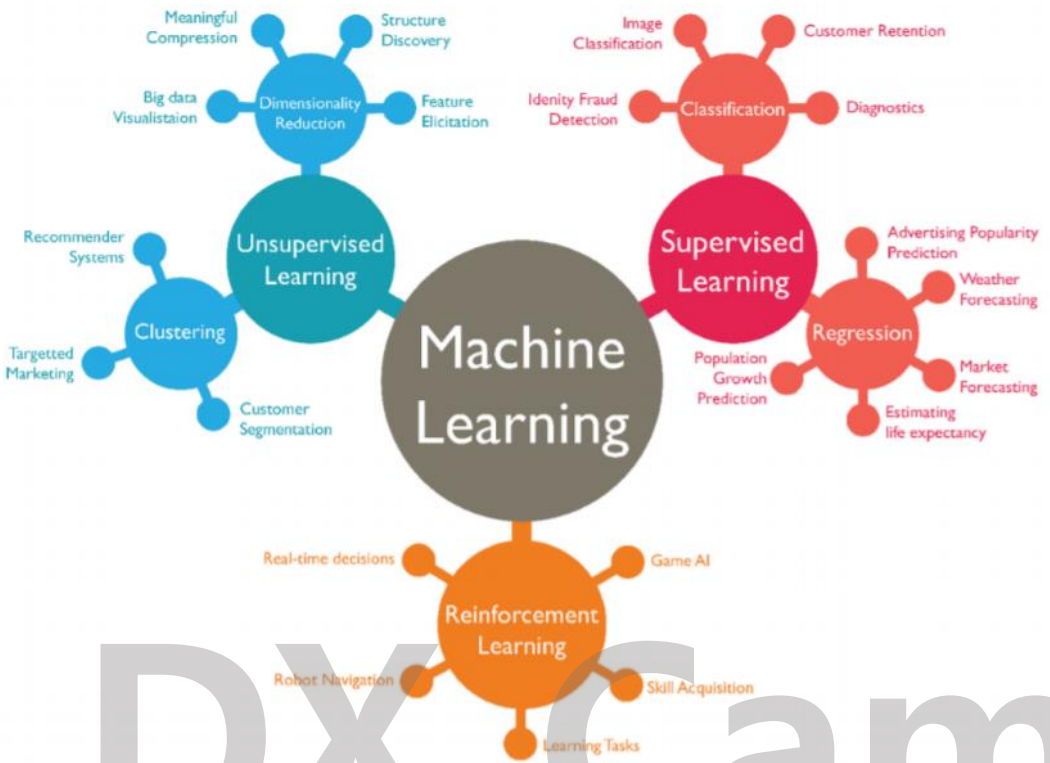
X1	X2
4	2

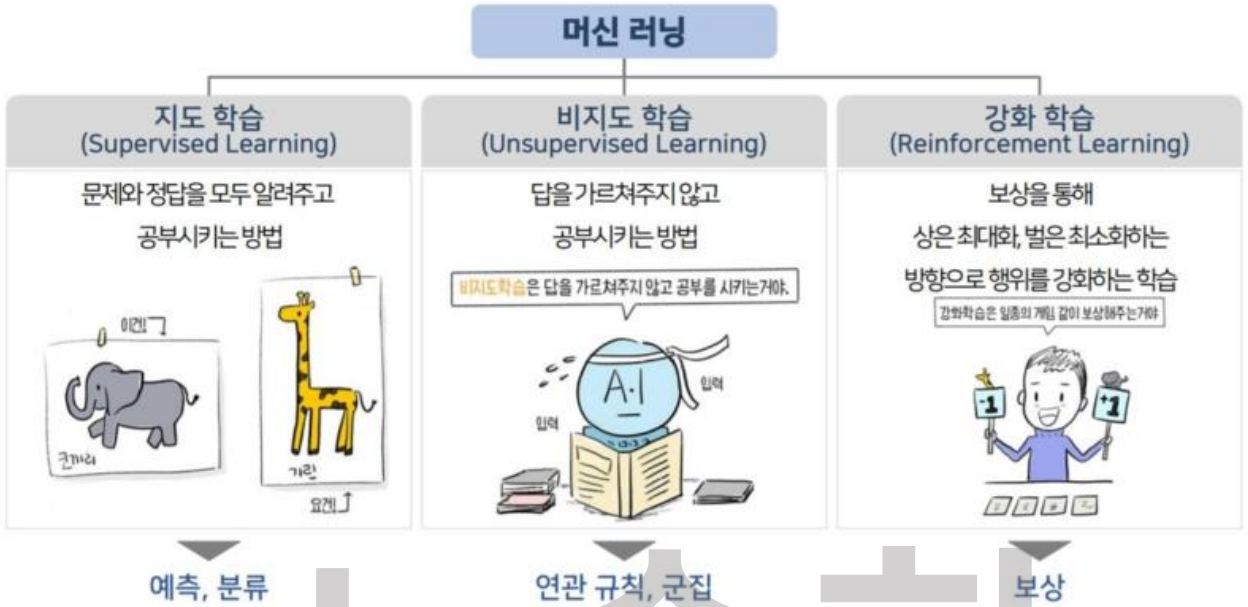
추론

(Y = 5X1 + 2X2)

Y = 24

머신러닝의 종류



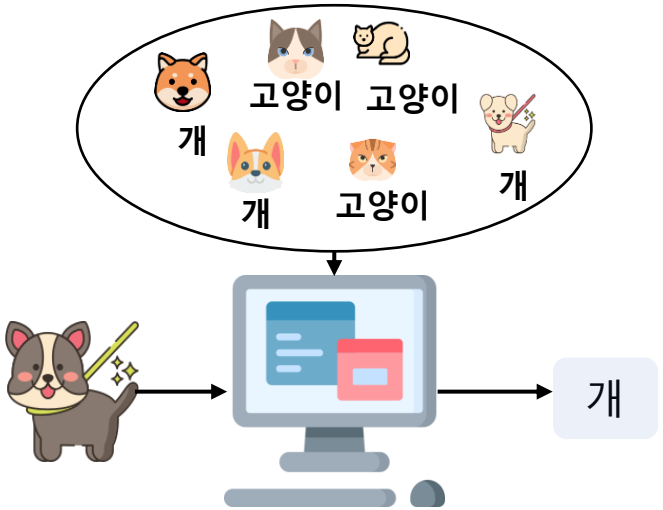


https://www.mactores.com

지도 학습supervised learning

: **정답 (label)** 이 있는 데이터들을 학습한다.

* 사람에 의해 데이터와 정답의 역할을 하는 **레이블**label을 제공받는다.



- 지도학습의 목표는 레이블링 된 데이터를 통해 카테고리별로 일반적인 규칙을 학습하는 것.
- 예를 들어서 개와 고양이를 구분할 때, 정답을 알려 준 뒤에 학습한 컴퓨터가 새로운 이미지를 보고 개인지 고양이인지 맞추도록 하는 것.

지도학습 예시

분류 모델

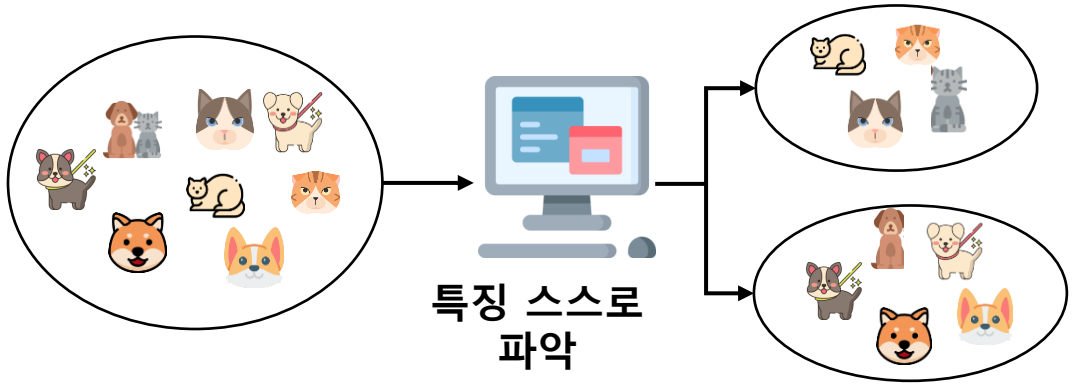
- ✓ 신용카드 사기 거래 탐지
- ✓ 채무 불이행 판별
- ✓ 스팸 메일 판별

회귀 모델

- ✓ 전력 사용량 예측
- ✓ 태양광/풍력 발전량 예측
- ✓ 주차수요 예측
- ✓ 자전거 대여량 예측

비지도 학습unsupervised learning

: 지도 학습과는 달리 외부에서 **정답(label)**이 주어지지 않고 학습 알고리즘이 스스로 입력으로 부터 어떤 구조를 발견하는 학습이다.



주어진 데이터를 이용하여
패턴을 발견하여 분류하는 군집화

- 인간의 개입이 없는 데이터를 스스로 학습하여 그 속의 **패턴(pattern)** 또는 각 데이터 간의 **유사도(similarity)**를 학습.
- 데이터에서 숨겨진 패턴을 발견 가능

  * 대표적 예 : **군집화**clustering , 이 방법은 주어진 데이터를 특성에 따라 둘 이상의 그룹으로 나누는 것.

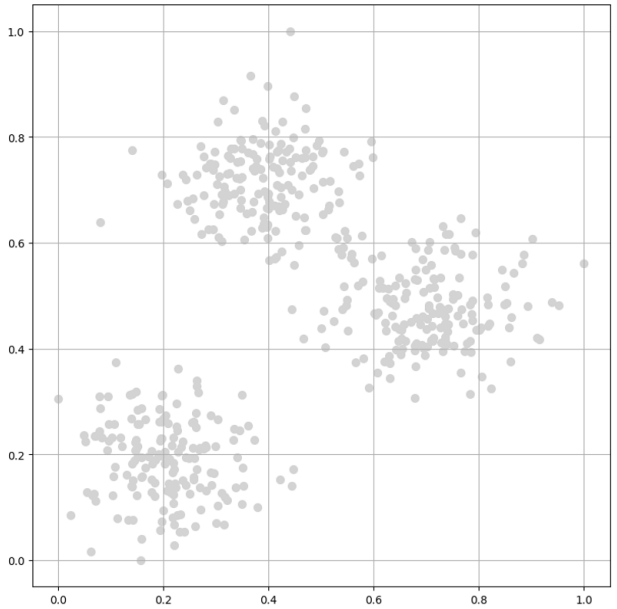
비지도학습 예시

군집화

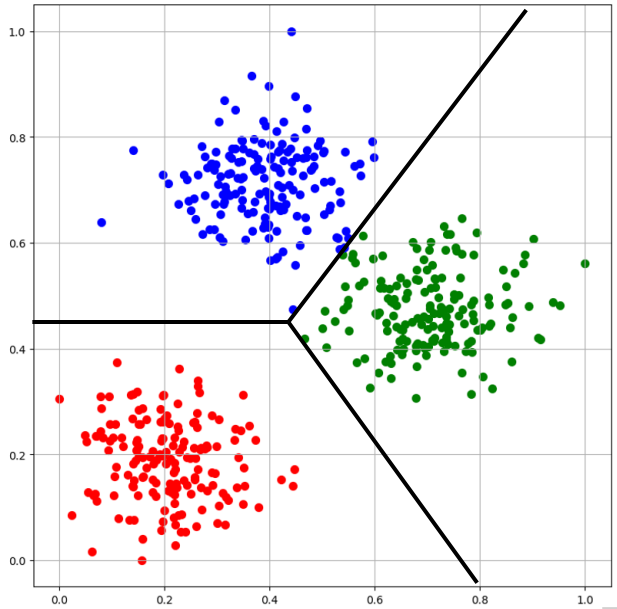
이상 감지

차원 축소

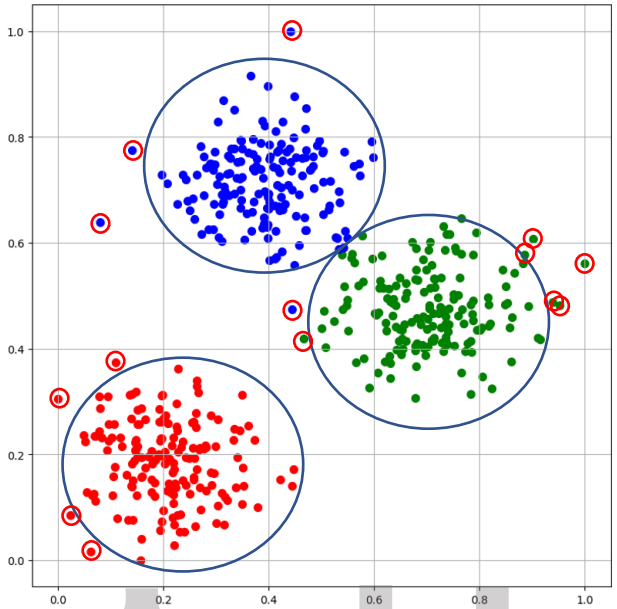
군집화 vs 이상감지



Data

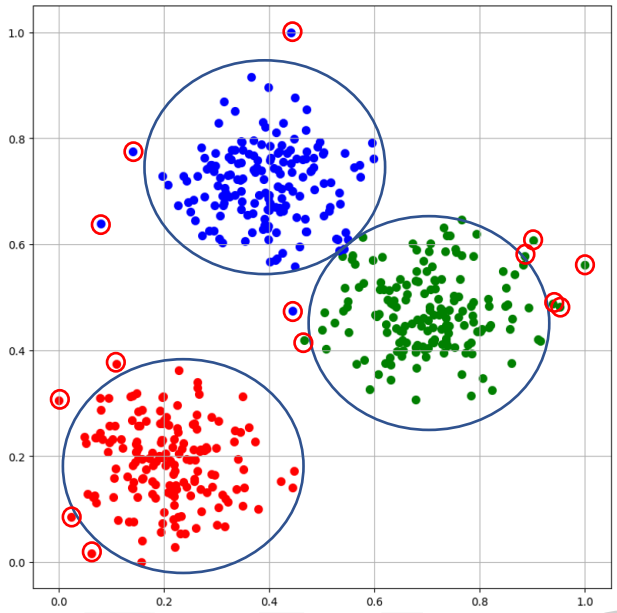


Clustering



Anomaly Detection

이상 감지 활용 분야

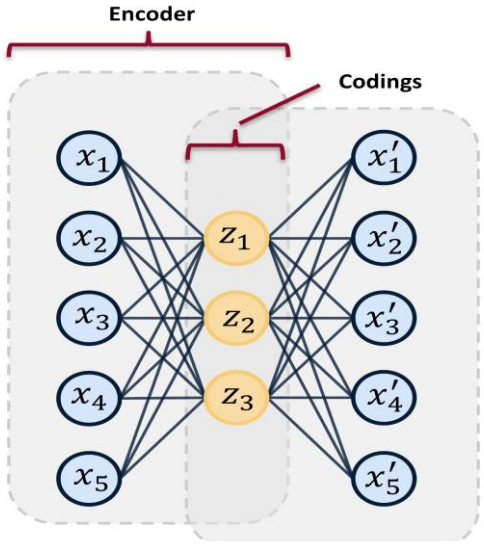


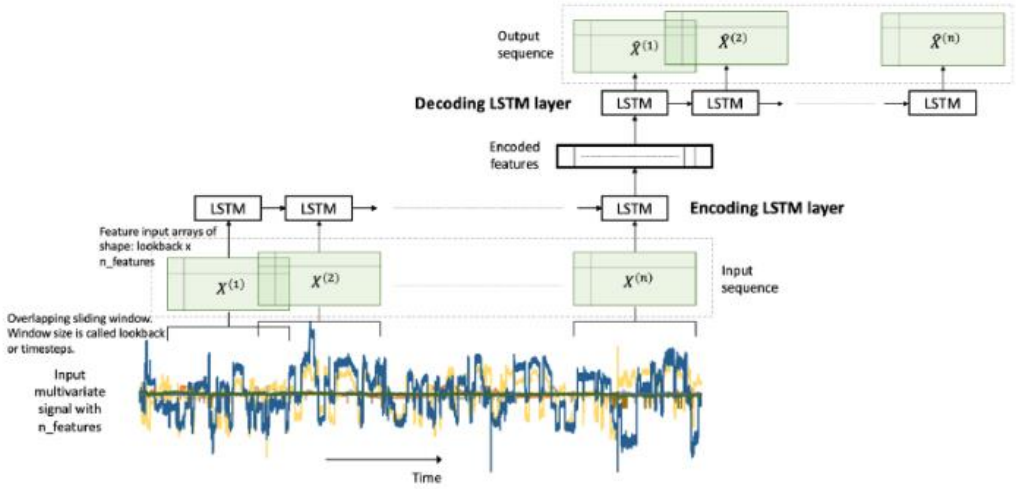
Anomaly Detection

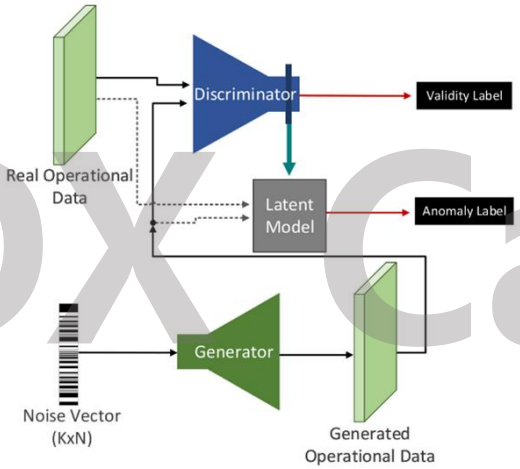
- **Cyber-Intrusion Detection**: 컴퓨터 시스템 상에 침입을 탐지.
- **Fraud Detection**: 보험, 신용, 금융 관련 데이터에서 불법 행위를 검출
- **Malware Detection: Malware**(악성코드)를 검출
- **Medical Anomaly Detection**: 의료 영상, 뇌파 기록 등의 의학 데이터에 대한 이상치 탐지
- **Social Networks Anomaly Detection**: Social Network 상의 이상치들을 검출
- **Log Anomaly Detection**: 시스템이 기록한 log를 보고 실패 원인을 추적
- **IoT Big-Data Anomaly Detection**: 사물 인터넷에 주로 사용되는 장치, 센서들로 부터 생성된 데이터에 대해 이상치를 탐지
- **Industrial Anomaly Detection**: 산업 속 제조업 데이터에 대한 이상치를 탐지

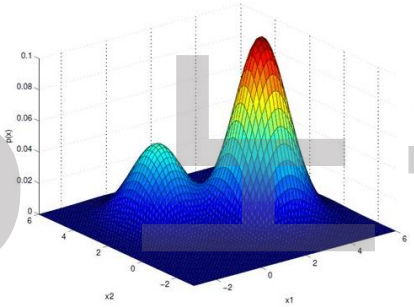
https://hoya012.github.io/blog/anomaly-detection-overview-1/

이상 감지 모델









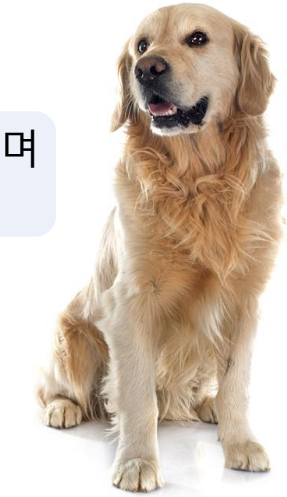
GAUSSIAN MIXTURE MODEL

CronJ

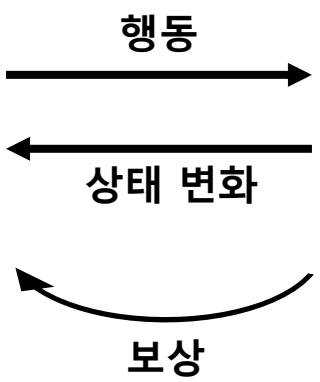
강화 학습reinforcement learning

: 강화(Reinforcement)는 시행착오(Trial and Error)를 통해 학습하는 방법으로 이러한 강화를 바탕으로 강화학습은 실수와 보상을 통해 학습을 하여 목표를 찾아가는 알고리즘.

에이전트(개)가 환경(훈련)과 상호작용하며 보상(먹을 것)을 통해 행동을 결정(정책)



에이전트

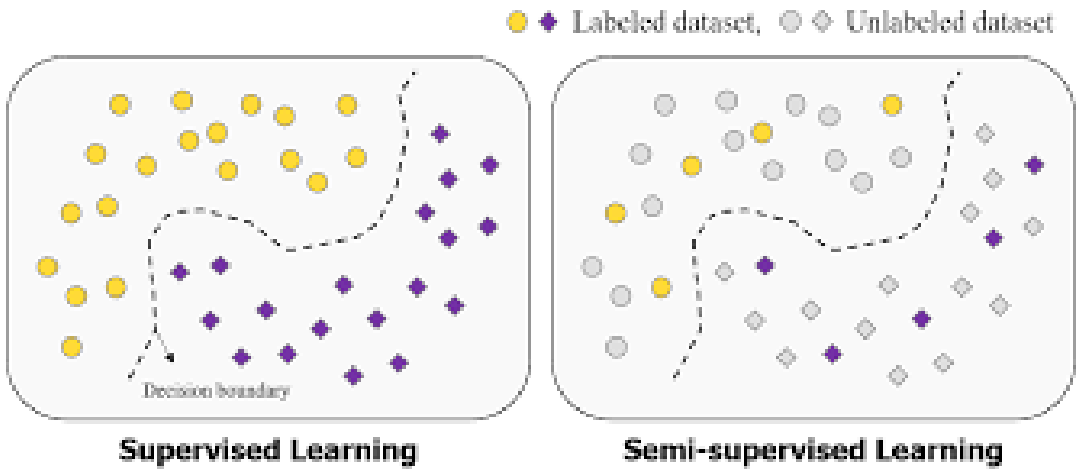


환경

- 보상 및 처벌의 형태로 학습 데이터가 주어짐.
  * 주로 차량 운전이나 상대방과의 경기 같은 동적인 환경에서 프로그램의 행동에 대한 피드백만 제공되는 경우 사용
- 학습을 수행하고 행동을 하는 에이전트agent 가 환경과 상호작용을 한 뒤 보상에 따라 행동을 결정하는 정책policy을 바꾸어 나가는 방식이다.

준지도 학습semi-supervised learning

: 소량의 레이블링 된 데이터에는 지도학습을 적용하고 대용량 레이블링 되지 않은 데이터에는 비지도 학습을 적용하여 추가적인 성능향상을 목표로 하는 방법.



- 우리가 실제로 구할 수 있는 데이터에는 레이블이 없는 경우가 많고, 레이블은 사람이 부여하는 경우가 많다.
  * 대규모 데이터에 레이블을 모두 부여하는 일은 매우 어려운 일이 될 수 있다.
- 이런 경우 일부 데이터에만 레이블을 부여하여도 레이블이 전혀 없는 것보다 전체적인 학습의 정확도를 높일 수 있다.

그 외에 이론

- **전이학습(Transfer Learning):** 특정 분야에서 학습된 신경망의 일부 능력을 유사하거나 전혀 새로운 분야에서 사용되는 신경망의 학습에 이용하는 것.

- **연합학습(Federated Learning):** 모바일 기기에 저장되어 있는 모델을 학습한 후 해당 모델을 서버로 보내고 중앙 서버에서 학습하는 방식.

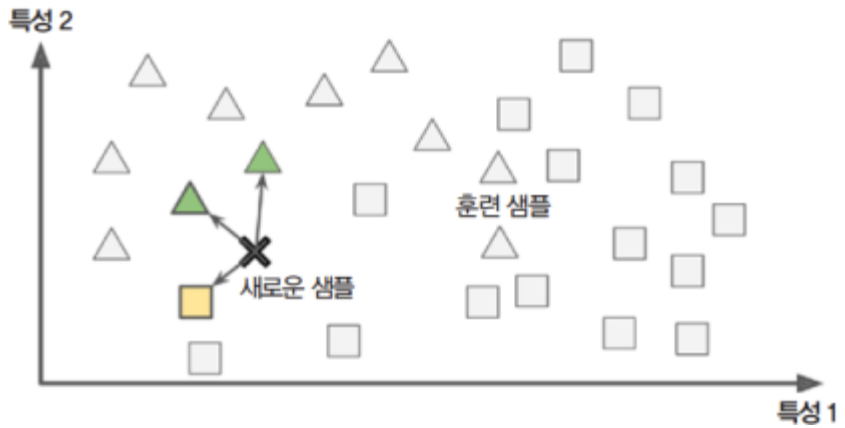
- **메타러닝(Meta Learning):** 다른 Task를 위해 학습된 머신러닝 모델을 이용해서, 적은 데이터셋을 가지는 다른 Task도 잘 수행할 수 있도록 학습시키는 방식.

- **설명가능한 인공지능(eXplainable Artificial Intelligence: XAI)**: 머신러닝, 딥러닝 모델의 결과값에 대한 이유를 인간이 이해할 수 있도록 블랙박스 성향을 분해하고 파악하여 설명 가능성을 제공하는 방식.

- **자기지도학습(Self-Supervised Learning):** 데이터 자체에 스스로 레이블을 생성하여 학습에 이용하는 방법으로 다량의 레이블이 없는 데이터로부터 데이터 부분들의 관계를 통해 레이블 자동 생성을 통해 지도학습에 이용하는 비지도 학습 방식.

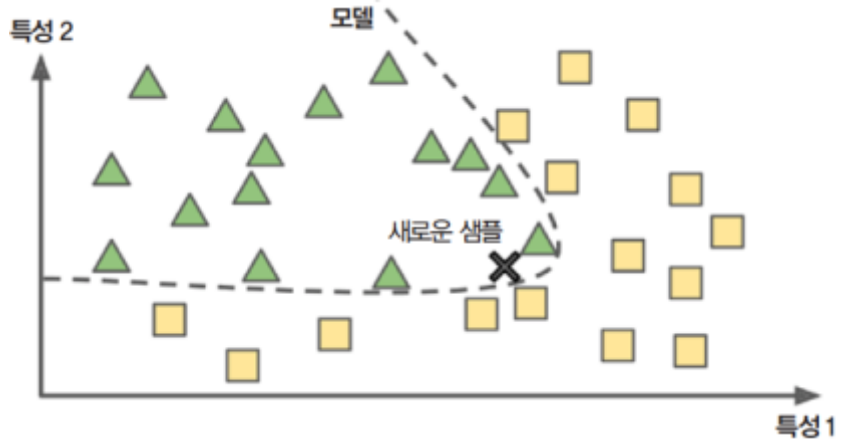
모델 학습 방법

사례 기반 학습



- 샘플을 기억하는 것이 훈련
- 예측을 위해 기존 샘플과의 유사도 측정
- 새로운 데이터가 들어올 때마다 재 훈련해야 함.

모델 기반 학습



- 샘플을 사용하여 모델을 만들고 훈련시키는 벙법.
- 새로운 데이터가 들어왔을 때 훈련된 모델을 통해 예측.

인공지능(AI) 기술의 응용 분야

1. 전문가 시스템(Expert System)
인공지능 기술 응용 분야 중에서 가장 활발한 영역. 특정 문제에 대한 전문적인 지식을 컴퓨터에 기억시키고, 시스템화하여 비전문가도 전문지식을 활용할 수 있도록 하는 시스템. 대표적인 예시로 의료 진단 시스템, 설계 시스템을 들 수 있음.

2. 자연어 처리(Natural Language Processing)
인간의 언어, 억양 및 맥락을 컴퓨터가 이해할 수 있도록 돕는 인공지능의 한 분야. 딥러닝 기반의 자연어처리 기술은 대량의 텍스트로부터 의미 있는 정보를 추출하고 활용할 수 있음. 번역기, 챗봇, 텍스트 자동 완성 등이 자연어 처리 기술을 활용하는 분야.

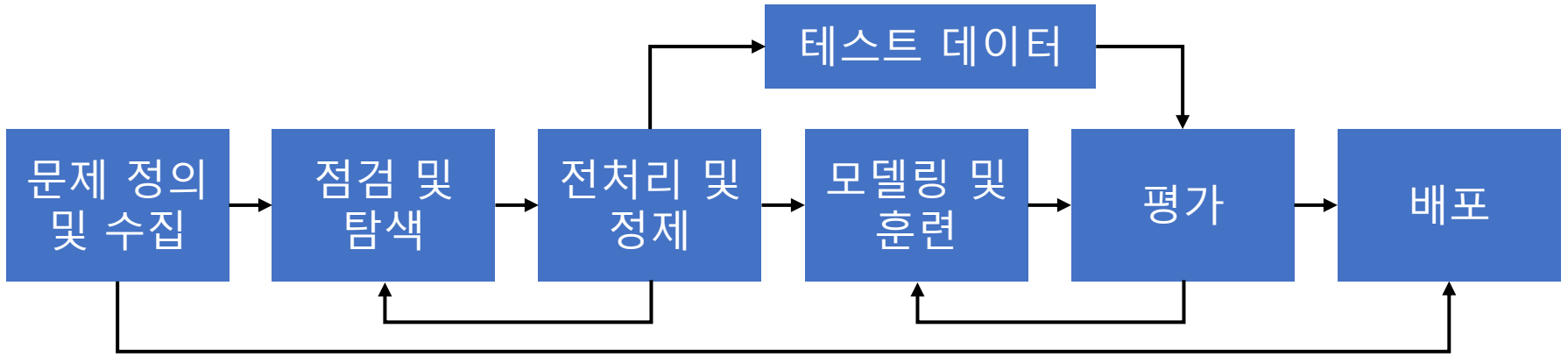
3. 데이터 마이닝(Data Mining)
보유한 데이터를 분석하여 유용한 정보를 추출해 조합하는 기술. 방대한 양의 데이터 속에서 특정 패턴을 뽑아내고 통계적인 방식으로 가치를 부여. 해당 기술은 위험 및 생산성 관리, 시장 분석, 시스템 설계 등에 활용.

4. 컴퓨터 비전(Computer Vision)
컴퓨터의 시각적인 부분을 연구하여 디지털 이미지, 비디오 등에서 의미 있는 정보를 추출하는 기술. 컴퓨터 비전은 우리에게 익숙한 안면 인식에 활용되고 있음. 인간의 시각으로는 판단하기 어려운 부분을 컴퓨터가 분석하는 영역.

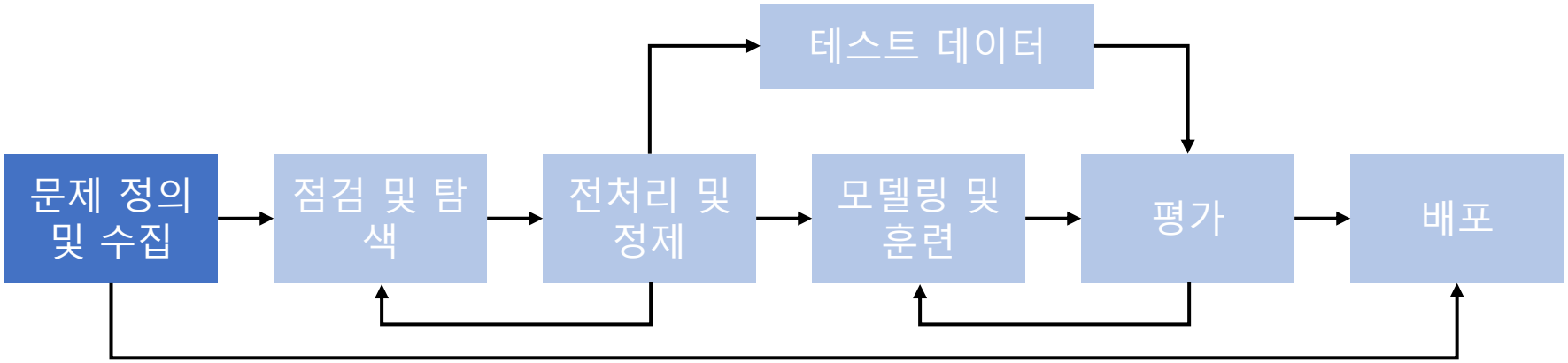
5. 지능로봇(Intelligent Robots)
인공지능 기술을 활용한 로봇을 통칭. 외부환경을 인식하여 스스로 상황을 판단하고 자율적으로 움직이는 기계. 우리가 상상하는 인공지능의 대표적인 이미지가 지능로봇에 해당.

데이터 분석 라이브러리 코드 실습

머신러닝 워크플로우

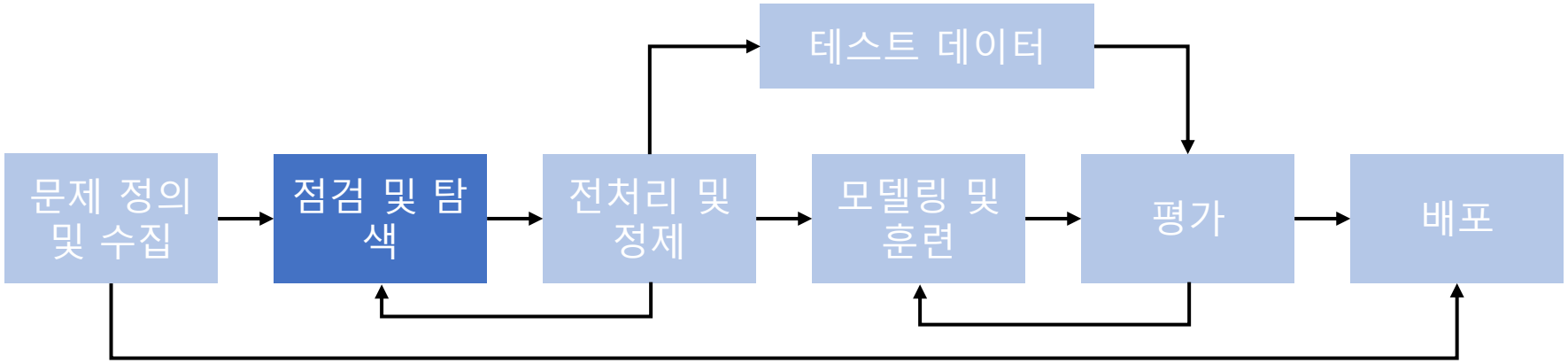


1. 문제 정의 및 데이터 수집



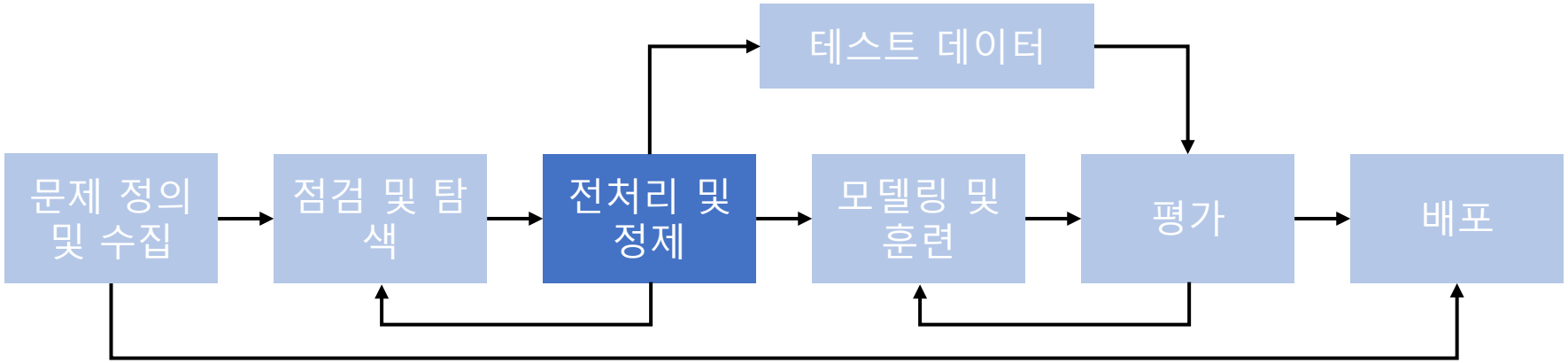
- 전체 머신러닝 프로세스 중 가장 중요한 단계
- 명확한 목적 의식을 가지고 프로세스를 시작
- 모델의 종류 결정 및 탐색할 데이터의 종류를 결정
- 문제 정의 후 모델을 학습할 데이터를 수집(크롤링, 센서 활용, 구글링, 데이터 API 활용 등)

2. 점검 및 탐색



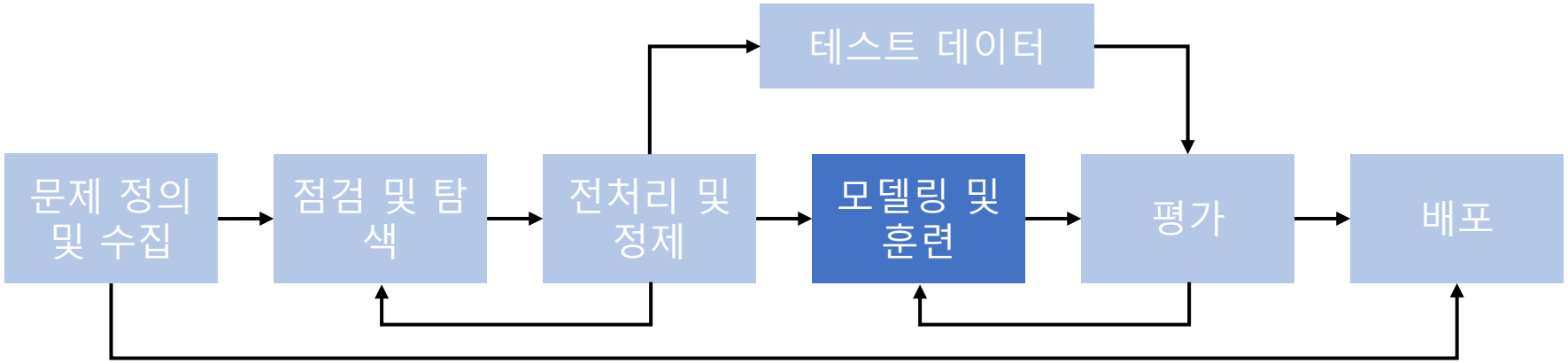
- 데이터를 점검하고 탐색하는 단계
- 데이터의 구조, 노이즈 데이터, 머신 러닝 적용을 위해서 데이터를 어떻게 정제해야 하는지 등을 파악
- 독립 변수, 종속 변수, 변수 유형, 변수의 데이터 타입 등을 점검
- 데이터의 특징과 내재하는 구조적 관계를 알아내는 과정을 의미
- 이 과정에서 시각화와 간단한 통계 테스트를 진행

3. 전처리 및 정제



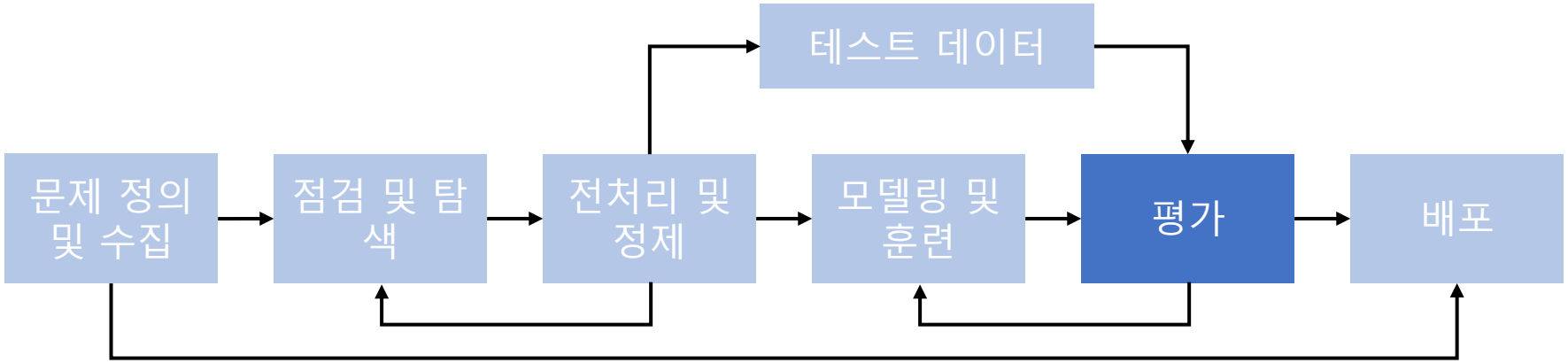
- 머신 러닝 워크플로우에서 가장 까다로운 작업 중 하나이다
- 빠르고 정확한 데이터 전처리를 하기 위해서는 사용하고 있는 툴에 대한 다양한 라이브러리 지식이 필요
- 데이터의 결측치 및 이상치 확인, 제거, 일관성 있는 데이터의 형태로 전환하는 과정
- 전처리의 종류 : 데이터 클리닝(cleaning), 데이터 통합(integration) , 데이터 변환(transformation), 데이터 축소(reduction), 데이터 이산화(discretization) 등

4. 모델링 및 훈련



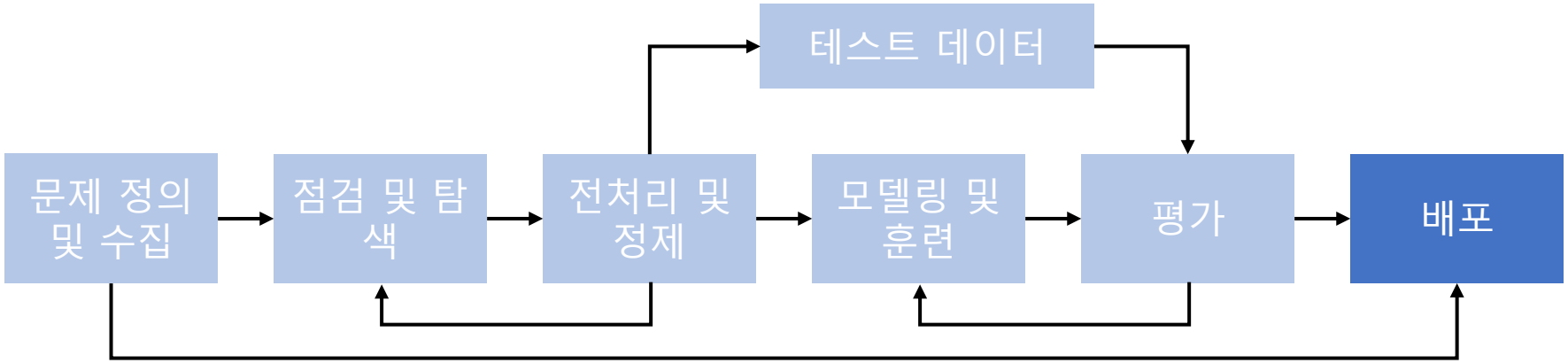
- 데이터에 적합한 머신러닝 모델을 선택 후 모델링
- 전처리가 완료된 데이터를 머신러닝 모델학습
- 학습 후 훈련이 제대로 되었다면 우리가 원하는 태스크(task)를 수행 가능
- 주의해야 할 점 : 데이터 훈련하기 전 훈련용, 테스트용 데이터를 나누어 모델 학습 (성능 테스트 필요)

5. 평가



- 학습이 된 모델을 테스트용 데이터로 성능 평가하는 과정
- 기계가 예측한 데이터가 테스트용 데이터의 실제 정답과 얼마나 가까운지를 측정

6. 배포



- 평가 단계에서 기계가 성공적으로 훈련이 된 것으로 판단된다면 완성된 모델이 배포
- 다만, 여기서 완성된 모델에 대한 전체적인 피드백으로 인해 모델을 업데이트 해야 하는 상황이 온다면 수집 단계로 돌아감

머신러닝을 위한 도구

캘리포니아 주택 가격 예측 코드 실습

- 머신러닝 프로젝트 처음부터 끝까지 실습하기(쉬운 버전)