

머신러닝(Machine Learning) /딥러닝(Deep Learning)

목 차

1-1 machine learning(기계학습)은 무엇일까요?

1-2 딥러닝(Deep Learning)은 무엇일까요?

1-3 머신러닝과 딥러닝의 차이점은 무엇인가요?

2-1 머신러닝/딥러닝으로 무엇이 가능한가요?

3-1 기본 용어 이해하기

4-1 머신러닝과 데이터 마이닝

5-1 머신러닝의 구분

6-1 지도학습, 비지도학습, 비정형분석

7-1 데이터 마이닝(머신러닝) 수행 단계

01 machine learning(기계학습)이란?

- ▶ 머신러닝은 컴퓨터가 데이터를 이용해 향후 적용 가능한 규칙을 생성해 낸다.
인공지능의 한 계열로 볼 수 있다.

(dictionary.com)

- ▶ 머신러닝은 데이터에 지식을 추출하는 작업
- ▶ 머신러닝은 통계학, 인공지능, 컴퓨터 과학이 얹혀 있는 분야이다.

01 딥러닝(Deep Learning)은 무엇일까요?

- ▶ 인공 신경망을 사용하여 데이터로부터 학습하는 머신러닝의 한 종류.
- ▶ 대량의 데이터로부터 패턴을 인식하고 예측 및 의사결정을 수행할 수 있다.
- ▶ 이미지, 음성, 자연어 처리 등 다양한 분야에 활용되고 있다.

01 머신러닝과 딥러닝의 차이점

	머신러닝	딥러닝
데이터 표현 방식	주로 수작업으로 특징을 설계하고 추출	데이터로부터 자동으로 특징을 학습하고 추출
데이터의 양	상대적으로 적은 양의 데이터로 학습이 가능	정확한 학습을 위해 대량의 데이터가 필요합니다.
장점	해석하기 쉬운 알고리즘, 상대적으로 적은 데이터로 학습 가능	뛰어난 패턴 인식 능력, 복잡한 작업 수행 가능
단점	데이터에 명시적으로 정의하고 추가된 특징이 필요. 딥러닝에 비해 성능이 떨어질 수 있음.	학습에 많은 데이터와 처리 능력 필요. 모델 해석의 어려움.
하드웨어 요구사항	상대적으로 적은 컴퓨팅 자원이 필요	대량의 데이터와 복잡한 모델로 인해 높은 컴퓨팅 파워가 필요
사용 분야	표 형식의 데이터에 많이 사용. (예) 신용 평가, 주가 예측, 사기 탐지, 마케팅, 고객 세분화 등의 비즈니스 분석 기타	비정형 데이터(이미지, 음성, 텍스트 등)에 많이 사용. 컴퓨터 비전(물체인식, 자율주행 등), 자연어 처리(기계 번역, 감정 분석 등), 음성 인식 등

02 머신러닝/딥러닝으로 무엇이 가능한가?

(1) 편지 봉투에 손으로 쓴 우편번호 숫자 판별

(2) 의료 영상 이미지에 기반한 종양 판단

=> 종양이 악성인지 양성인지 판단

(3) 의심되는 신용카드 거래 감지

=> 신용카드 거래 내역이 입력이 되고 부정 거래인지가 출력이 된다.

(4) 블로그 글의 주제 구분

=> 많은 양의 텍스트 데이터를 요약하고 그 안에 담긴 핵심 주제를 찾기.

02 머신러닝/딥러닝으로 무엇이 가능한가?

(5) 고객들을 취향이 비슷한 그룹으로 묶기

=> 고객 데이터로 고객들의 취향이 비슷한 고객을 그룹으로 묶고 싶을 때,

(6) 비정상적인 웹 사이트 접근 탐지

=> 정상 패턴과 비정상 패턴을 찾아본다.

(7) 객체 인식 - 얼굴, 눈 등 인식

=> 스마트폰 얼굴 인식. 이미지 픽셀 데이터를 이용한 학습을 통해 사람의 얼굴 확인 가능.

02 머신러닝/딥러닝으로 무엇이 가능한가?

(9) 생성형 AI

=> ChatGPT, Claude, Google Gemini, DALL-E 등 텍스트 입력으로 이미지, 영상, 음원을 생성.

(10) 멀티모달 AI(Multimodal AI)

=> 이미지, 텍스트, 음성 등 다양한 데이터를 통합하여 처리하는 기술이 발전

(11) 자율 주행

=> 딥러닝을 기반으로 한 자율 주행 기술은 차량이 스스로 주행 결정을 내릴 수 있도록 한다.
(카메라, 레이더, 센서 데이터를 활용한 주행 경로 예측, 객체 감지 및 추적, 신호 인식 등을 수행)

03 기본 용어 이해하기

▶ 모델, 모델링(Model, Modeling)

(가) 모델

- 데이터를 기반으로 만들어진 추상화된 수학적 또는 계산적 표현
- 머신러닝/딥러닝에서 모델은 **입력 데이터로부터 출력 값을 예측하거나 매핑**하는 수학적 함수 또는 알고리즘을 의미. (예) 선형 회귀, 의사결정트리, 심층 신경망

(나) 모델링

- 실제 문제를 해결하기 위해 가장 적절한 머신러닝 모델을 선택하고 구축하는 전체 과정.
- 모델링은 주어진 데이터를 분석하고 이해한 후, 그것을 수학적 또는 계산적 모델로 표현하는 과정.

03 기본 용어 이해하기

▶ 데이터 셋(Dataset), 학습 데이터, 테스트 데이터

(가) 데이터 셋

모델 학습에 사용되는 입력 데이터와 해당 데이터에 해당하는 출력의 집합을 말한다.

(나) 학습 데이터, 테스트 데이터

- 학습 데이터 : 머신러닝/딥러닝 모델을 학습 시키는데 사용되는 데이터 집합.
- 테스트 데이터 : 머신러닝/딥러닝 모델의 성능을 평가하는데 사용되는 데이터 집합

03 기본 용어 이해하기

▶ 샘플(sample), 데이터 포인트(data point), 레이블(Label)

(가) 샘플(sample) 또는 데이터 포인트

하나의 개체 또는 행을 샘플이라고 말한다.

(나) 특징 or 특성 (feature or variable)

학습 데이터에서 추출된 정보로 모델이 예측을 수행하는 데 사용되는 입력 데이터의 일부

(다) 레이블(Label)

학습 데이터의 원하는 출력값. 레이블은 숫자(0,1) 또는 범주(고양이, 개)와 같은 형태일 수 있음.

03 용어 이해하기

▶ 피처(features)

- 비슷한 다른 용어 :

예측 인자(Predictors), 독립변수(independent variables), 인풋(Input)

▶ 레이블(Label), 클래스(Class)

- 레이블은 원하는 답이며, 레이블의 범주를 클래스라고 한다.

- 비슷한 다른 용어 : 결과(outcome), 타겟(target), 종속변수(dependent variable)

03 기본 용어 이해하기

▶ 특성 추출, 특성 엔지니어링

(가) 특성 추출과 특성 엔지니어링

기계 학습 모델에 사용할 데이터의 특징을 만드는 과정을 의미.

(나) 특성 추출(feature extraction)

기존 데이터로부터 새로운 특징을 추출하는 과정. 예를 들어 이미지 데이터에서 모서리, 블록, 텍스처 등의 패턴을 찾는 것.

(다) 특성 엔지니어링(feature engineering)

기존의 데이터를 변형하거나 새로운 정보를 추가하여 나은 변수를 생성하는 작업으로

데이터 자체 뿐만이 아닌 문제와 관련된 지식과 경험을 바탕으로 한다.

03 용어 이해하기

▶ 데이터 마이닝

(가) 대용량의 데이터로부터 데이터 내에 존재하는 관계, 패턴, 규칙 등을 탐색하고 모형화하여 유용한 지식을 추출하는 일련의 과정으로 데이터 마이닝은 데이터 속에서 의미 있는 통찰력을 찾아내는 것에 초점을 둔다.

(나) 데이터 마이닝이 소개되기 전의 데이터 분석과 구분 짓는다면, 복잡성 높은 데이터 분석에 기계학습(machine learning) 이론이 적용되기 시작함.

(정보통신기술용어해설 참조)

04 머신러닝과 데이터 마이닝

- ▶ 머신 러닝과 데이터 마이닝은 종종 같은 방법을 사용.
- ▶ 머신 러닝은 훈련 데이터를 통해 학습된 알려진 속성을 기반으로 **예측에 초점**을 두고 있다.
- ▶ 데이터 마이닝은 데이터의 미처 **몰랐던 속성을 발견하는 것**에 집중한다.

(위키 백과 참조)

05 머신러닝의 구분

▶ 지도학습(supervised learning)

(가) 예측하고자 하는 **목표(Target)가 존재**한다.

학습하고자 하는 데이터의 정답이 있다.

우리는 이 정답을 레이블(label)이라 한다. 교사의 역할이 존재.

▶ 비지도학습(unsupervised learning) or 자율학습

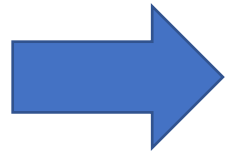
(나) **목표(Target)가 존재**하지 않는다. 교사 역할이 없음.

05 머신러닝의 구분

▶ 지도학습(supervised learning) 구분

(가) Regression – (회귀) – 레이블이 수치형 변수

(나) Classification – (분류) – 레이블이 범주형 변수



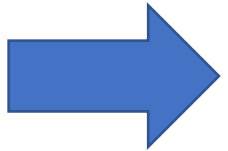
공통점 : 입력 및 특성(feature) 값을 이용하여
주어진 입력변수에 대한 **타겟(target, 목표변수)**의 값을 예측하는
모델을 구축한다.

05 머신러닝의 구분

▶ 지도학습(supervised learning)

(가) Regression – (회귀) – 레이블이 수치형 변수

(나) Classification – (분류) – 레이블이 범주형 변수



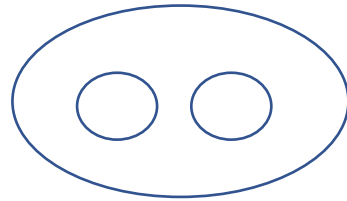
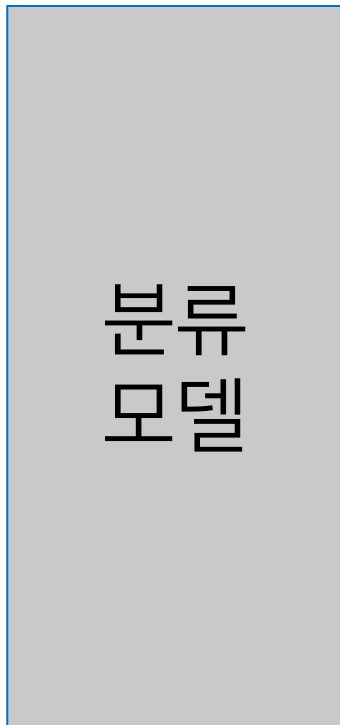
차이점 :

A. 목표 변수의 형태가 회귀의 경우 **연속형**이다.

B. 분류의 경우는 **범주형**이다.(고정되어 있음)

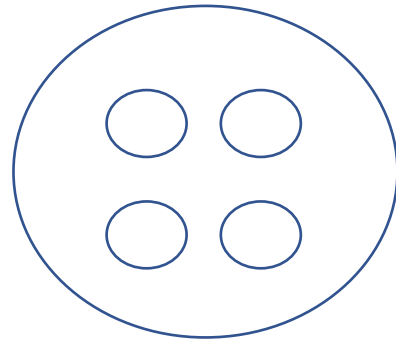
05 머신러닝의 구분

▶ Classification(분류)의 구분 - 이항분류와 다항분류



이항분류

목표값이 2개



다항(or 다중)분류

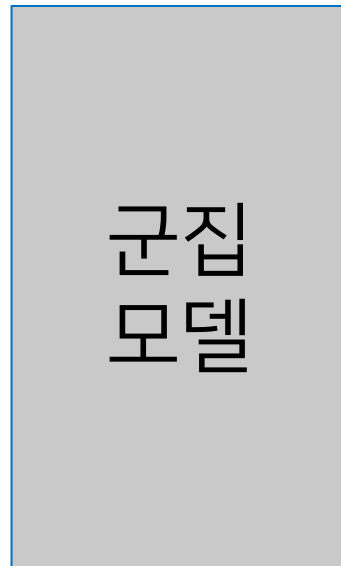
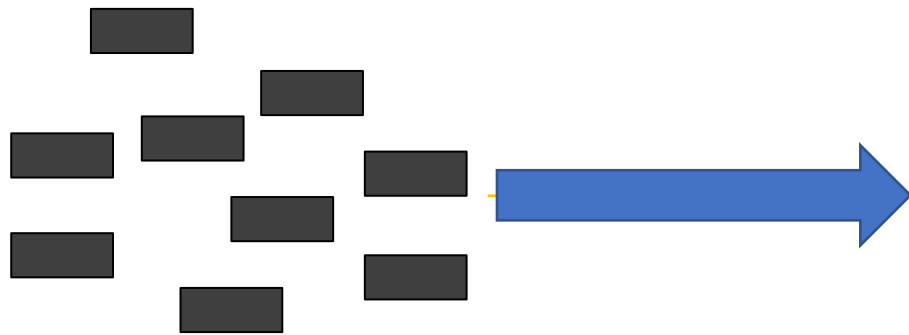
목표값이 3개 이상

05 머신러닝의 구분 - 비지도학습(unsupervised learning)

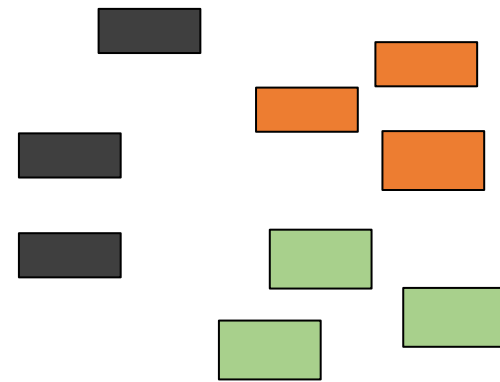
군집은 레이블이 없다.

레이블(목표 변수)가 없다.

군집은 레이블 없이 확보된
데이터의 특성을 분석



서로 유사한 특성을 가진
데이터끼리 그룹화



06 지도학습, 비지도학습, 비정형분석

지도학습(supervised learning)

Classification

Logistic Regression

SVM
(Support Vector machine)

Naïve Bayes Classification

Neural Net(신경망)

k-NN - 최근접이웃기법

Decision Tree

앙상블(RandomForest)

XGBoost LightGBM, CatBoost

Regression

Regression

knn - k-최근접이웃기법

Neural Net(신경망)

Decision Tree

앙상블(RandomForest)

XGBoost LightGBM,
CatBoost

비지도학습 (unsupervised learning)

Clustering(군집)

k-mean(K 평균)

DBSCAN

계층적 군집분석

연관성 분석

Apriori 알고리즘

Eclat 알고리즘

FP-Growth 알고리즘

06 지도학습, 비지도학습, 비정형분석

비정형 분석



텍스트 마이닝(Text mining)

토픽모델링(LDA), 감성 분석(LSTM, BERT),
텍스트 클러스터링



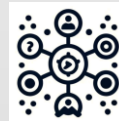
이미지/비디오 분석

객체 탐지(YOLO, Faster R-CNN등), 이미지
분류(VGG, ResNet, EfficientNet 등)



오디오/음성 분석

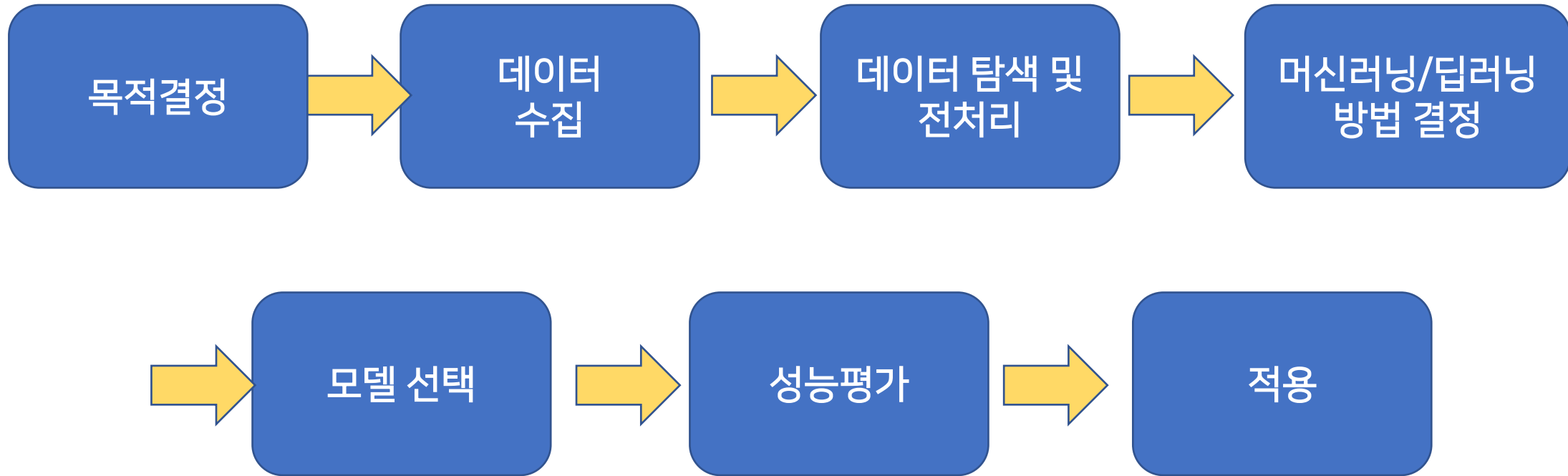
음성 인식(은닉 마르코프 모델-HMM), 트랜스포머
등, 음성 감정 분석(CNN, LSTM, 어텐션 메커니즘),
신호 처리 : 웨이블릿 변환, FFT 등



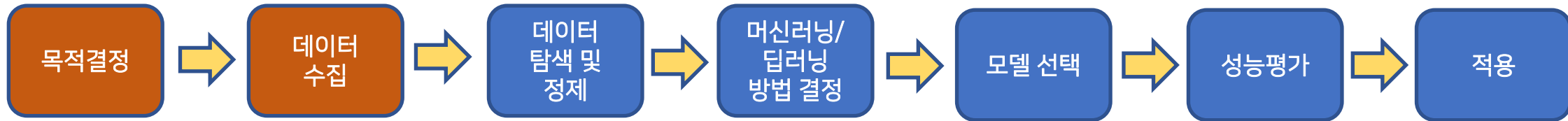
네트워크/그래프 분석

링크 예측 : 노드 임베딩, 그래프 신경망(GNN) 등,
커뮤니티 탐지, 중심성 분석

07 머신러닝의 수행단계



07 머신러닝의 수행단계



목적결정

▶ 프로젝트 목적을 계획하고 설정하는 단계

목적을 정하고 관련 데이터를 수집하기도 하지만,
때로는 데이터 수집 후 탐색 과정을 거쳐 문제가 설정되기도 한다.

데이터 수집

▶ 데이터 베이스 또는 분산된 데이터 베이스 이용

▶ 외부 데이터, 내부 데이터

07 머신러닝의 수행단계



데이터탐색 및 정제

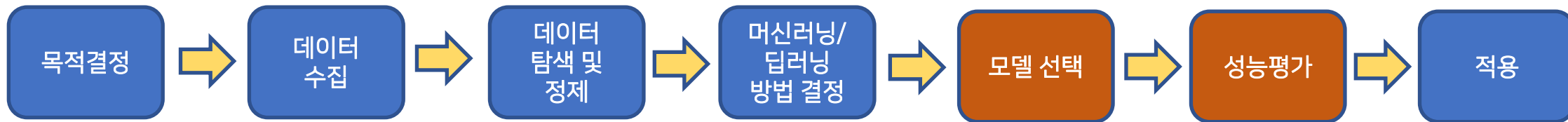
▶ 데이터 표준화 및 점검 (시각화)

- (1) 데이터에 결측치가 존재하는지, 이상치가 존재하는지.
- (2) 모든 값에 상식적인 범위 내에 있는지,
- (3) 머신러닝/딥러닝에 적용 가능한 데이터로 전처리

머신 러닝 방법 결정

- ▶ 머신 러닝/딥러닝 도전 과제(머신러닝/딥러닝)을 결정
- ▶ 알고리즘 선택(로지스틱 회귀, 신경망, 계층 군집 등)을 선택하는 단계

07 머신러닝의 수행단계



모델 선택

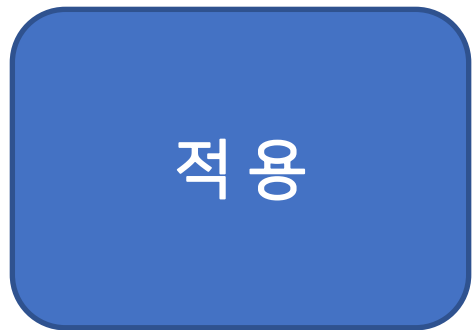
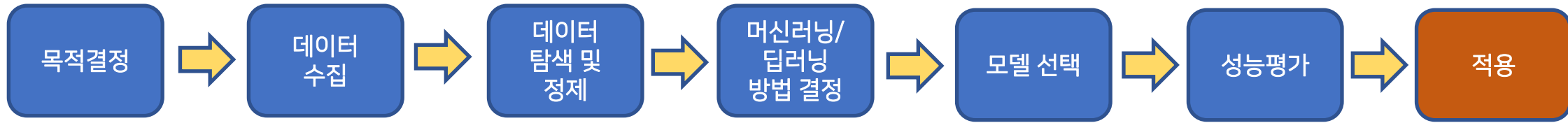
▶ 가장 좋은 모형을 찾는 단계

머신러닝 프로세스의 여러 단계를 반복적으로 수행하여 가장 좋은 모형을 찾는 단계.

성능평가

- ▶ 검증 데이터(테스트 데이터)를 이용하여 구축된 모형의 성능을 평가하여 효율적인 모형을 찾는다.
- ▶ 예측력이 가장 우수한 것을 선택하여 최종 모형 선정

07 머신러닝의 수행단계



▶ 구축된 모형을 운용 시스템에 탑재 후, 실제 의사결정에 적용.

(예제) **구축된 모형을 적용**하여 **구매가능성이 높은 고객을 결정**하고 해당 **고객에게 구매권유 메일**을 보내어 수익창출 가능성을 높인다.