머신러닝(Machine Learning)

목 차

01 machine learning(기계학습)이란?

▶ 머신러닝은 실험적인 데이터를 기반으로 프로그램의 실행 동작을 진화시키는 알고리즘의 설계와 개발을 고려하는 분야이다.

(위키 피디아)

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

(dictionary.com)

02 머신러닝으로 무엇이 가능한가?

(가) 편지 봉투에 손으로 쓴 우편번호 숫자 판별

(나) 의료 영상 이미지에 기반한 종양 판단

(다) 의심되는 신용카드 거래 감지

=> 신용카드 거래 내역이 **입력**이 되고 부정 거래인지가 **출력**이 된다.

(라) 블로그 글의 주제 구분

=> 많은 양의 **텍스트 데이터를 요약**하고 그 안에 담긴 **핵심 주제**를 찾기.

02 머신러닝으로 무엇이 가능한가?

(마) 고객들을 취향이 비슷한 그룹으로 묶기

=> 어떤 고객들의 취향이 비슷한지 비슷한 취향의 고객을 그룹으로 묶고 싶을 때,

(바) 비정상적인 웹 사이트 접근 탐지

=> 정상 패턴과 비정상 패턴을 찾아본다.

03 용어 이해하기

▶ 모델, 모형, 모델링(Model, Modeling)

(가) 정보 시스템 모델링

가. 데이터 모델링 : 현실세계의 복잡한 데이터들을 컴퓨터 정보 구조로 변환시키는 과정

(나) 수학적 모델링

가. 시스템의 변화을 나타내는(예측하는) 수학적 모델이 방정식으로 표현된다.

(정보통신기술용어해설 참조)

03 용어 이해하기

▶ 데이터 마이닝

(가) **대용량의 데이터**로부터 데이터 내에 존재하는 **관계, 패턴, 규칙 등을 탐색하고 모형화**하여 유용한 지식을 추출하는 일련의 과정.

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

(정보통신기술용어해설 참조)

03 용어 이해하기

▶ 샘플(sample), 데이터 포인트(data point), 특성(feature)

(가) 샘플(sample) 또는 데이터 포인트

하나의 개체 또는 행을 샘플이라고 말한다.

(나) 특징(feature or variable)

샘플의 속성, 즉 열을 말한다.

(다) 특성 추출(feature extraction) or feature engineering

좋은 입력 데이터를 만들어 내는 것.

03 용어 이해하기

▶ 클래스(class)

(가) 레이블(label)

원하는 답

(나) 클래스(class)

레이블의 범주를 클래스라 한다.

04 machine learning(기계학습)이란?

- 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 존재한다.

학습하고자 하는 데이터의 정답이 있다.

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

▶ 비지도학습(unsupervised learning) or 자율학습

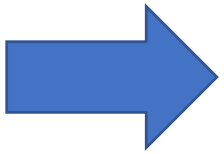
(나) 목표(Target)가 존재하지 않는다. 교사 역할이 없음.

05 머신러닝의 구분

▶ 지도학습(supervised learning) 구분

(가) Regression – (회귀) – 수치형 변수

(나) Classification – (분류) – 범주형 변수



공통점 : 입력 및 **특성(feature) 값**을 이용하여

주어진 **입력변수에 대한** 타깃(target, 목표변수)**의 값을 예측**하는

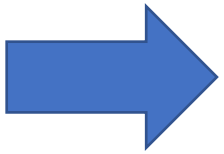
모델을 구축한다.

05 머신러닝의 구분

▶ 지도학습(supervised learning)

(가) Regression – (회귀) – 수치형 변수

(나) Classification – (분류) – 범주형 변수



차이점 :

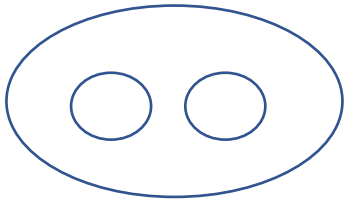
A. 목표 변수의 형태가 회귀의 경우 연속형이다.

B. 분류의 경우는 범주형이다.(고정되어 있음)

05 머신러닝의 구분

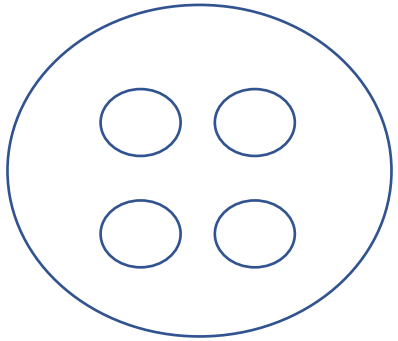
▶ Classification(분류)의 구분 - 이항분류와 다항분류

분류
모델



이항분류

그룹이 2개



다항분류

그룹이 3개 이상

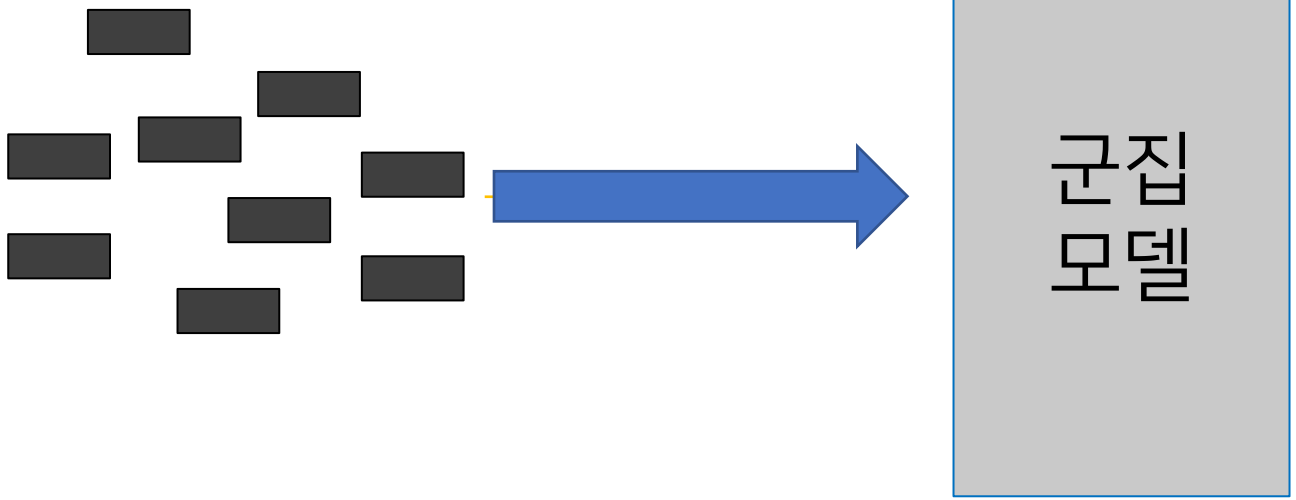
05 머신러닝의 구분 - 비지도학습(unsupervised learning)

군집은 레이블이 없다.

레이블(목표 변수)가 없다.

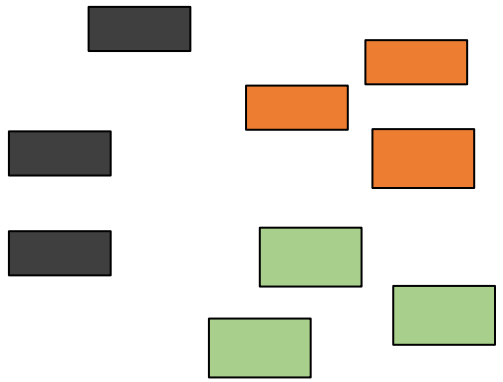
군집은 레이블 없이 확보된

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)
- knn - 최근접이웃기법
- Decision Tree
- 앙상블(RandomForest)
- 앙상블(GradientBoosting)

Regression

- Regression
- knn - k-최근접이웃기법
- Neural Net(신경망)
- Decision Tree
- 앙상블(RandomForest)
- 앙상블(GradientBoosting)

비지도학습 (unsupervised learning)

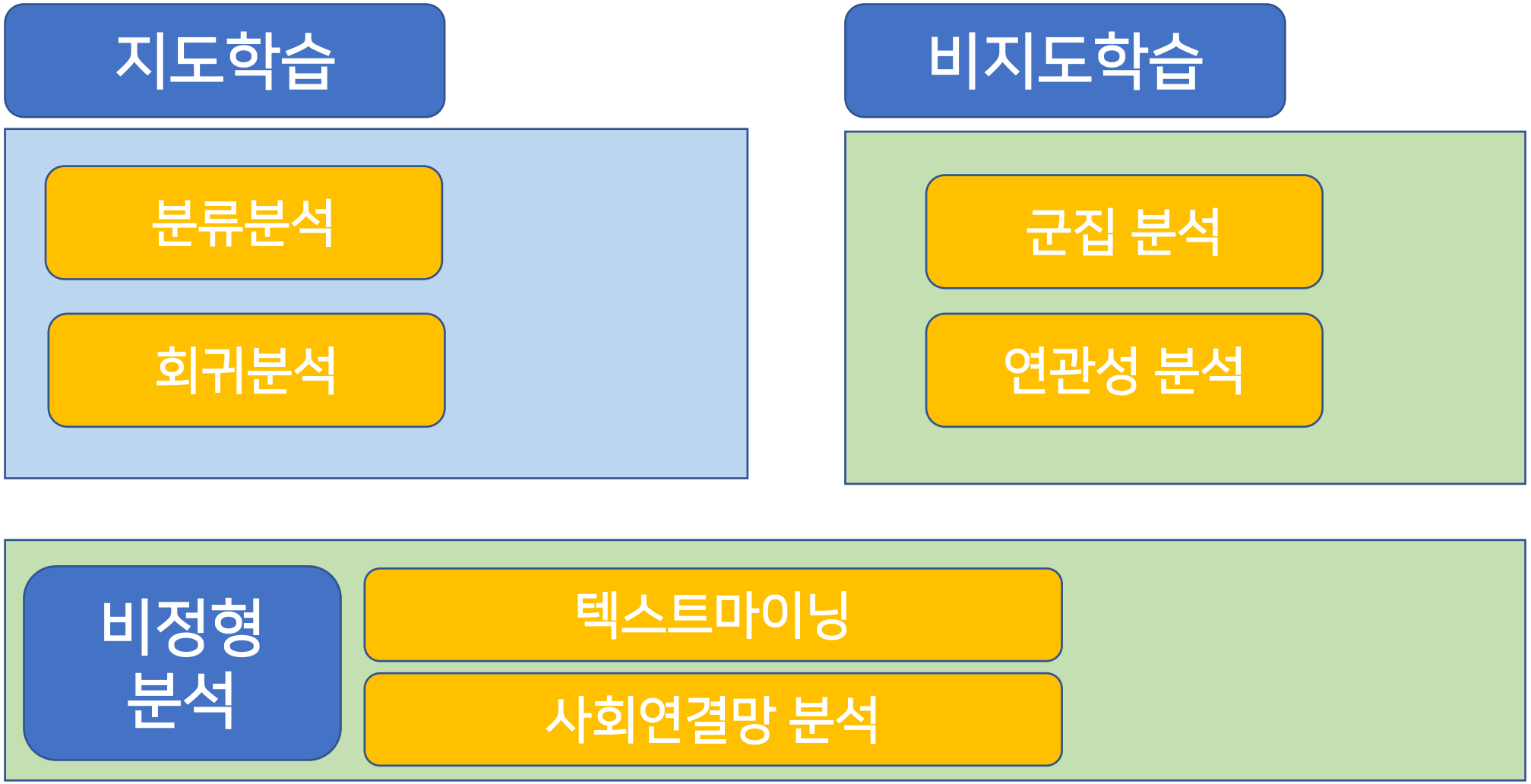
Clustering

- k-mean(K 평균)
- 계층적 군집분석
- DBSCAN

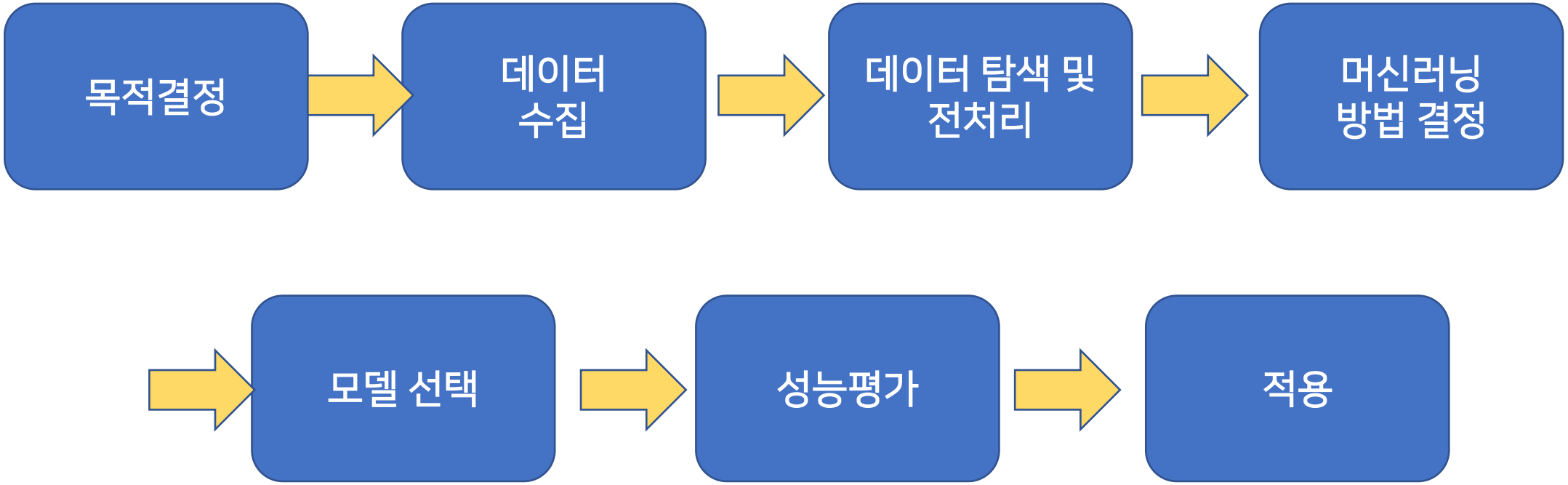
연관성 분석

- 장바구니 분석
- 서열 분석
- 트랜잭션 데이터분석

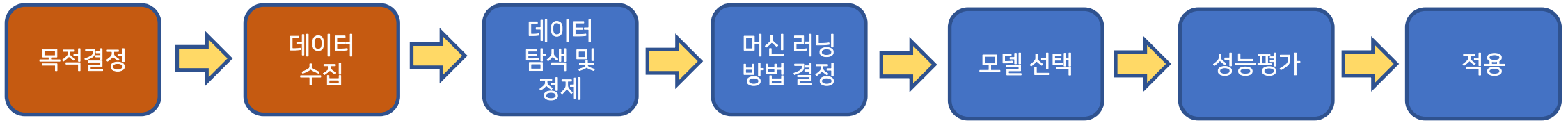
06 지도학습, 비지도학습, 비정형분석



07 데이터 마이닝의 수행단계



07 데이터 마이닝의 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 데이터 마이닝의 수행단계

데이터탐색 및 정제

▶ 데이터 표준화 및 점검 (시각화)

(1) 데이터에 결측치가 존재하는지,

(2) 모든 값에 상식적인 범위 내에 있는지,

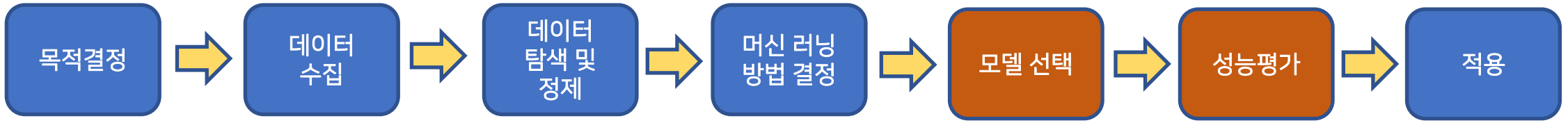
(3) 이상치가 존재하는지,

머신 러닝 방법 결정

▶ 머신 러닝 도전 과제(분류, 회귀, 군집화 등)을 결정

▶ 머신 러닝 기법(로지스틱 회귀, 신경망, 계층 군집 등)을 선택하는 단계

07 데이터 마이닝의 수행단계



모델 선택

▶ 가장 좋은 모형을 찾는 단계

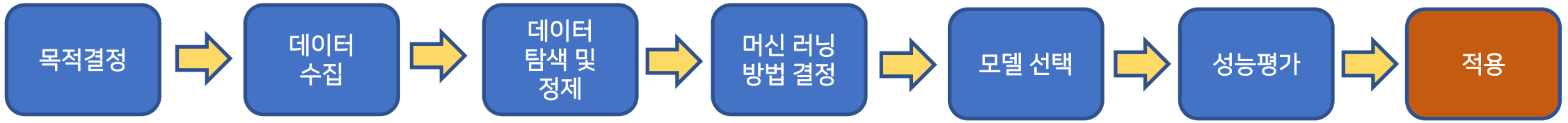
머신러닝 프로세스의 여러 단계를 반복적으로 수행하여 가장 좋은 모형을 찾는 단계.

성능평가

▶ 검증 데이터(테스트 데이터)를 이용하여 구축된 모형의 성능을 평가하여 **효율적인 모형을 찾는다.**

▶ 예측력이 가장 우수한 것을 선택하여 **최종 모형 선정**

07 데이터 마이닝의 수행단계



적 용

▶ 구축된 모형을 운용 시스템에 탐재 후, 실제 의사결정에 적용.

(예제) **구축된 모형을 적용**하여 구매가능성이 높은 고객을 결정하고 해당 고객에게 구매권유 메일을 보내어 수익창출 가능성을 높인다.