한양대학교
HANYANG UNIVERSITY

Introduction to Machine Learning

정소희

기계가 학습을 한다???

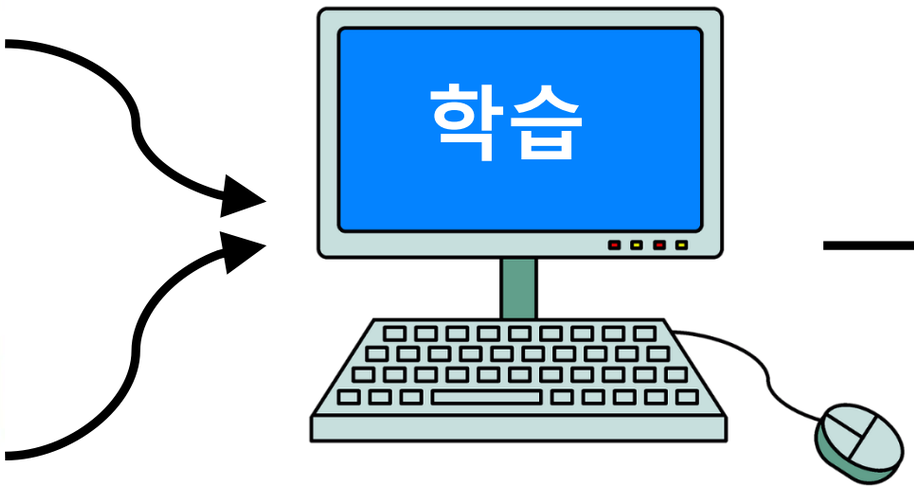
1. 강아지와 고양이 분류하기

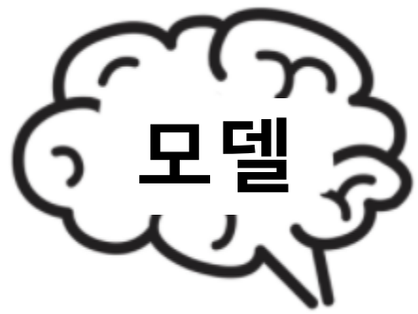
머신러닝을 이용한 강아지와 고양이 분류

머신러닝 과정

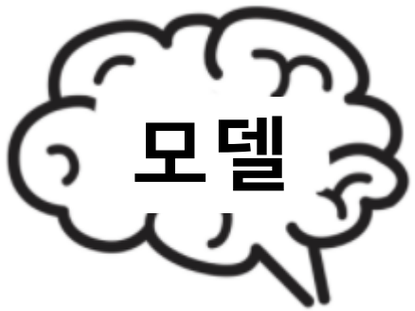
강아지 데이터

고양이 데이터





모델을 이용한 예측



강아지

2. 공부시간에 따른 성적 예측하기

원인 → 결과

시험 날짜	공부시간	성적
2022-05-01	2	40
2022-05-08	3	60
2022-05-15	4	80
2022-05-22	1	20
2022-05-29	3	60
2022-06-05	**5**	**?**

과거의 데이터

예측

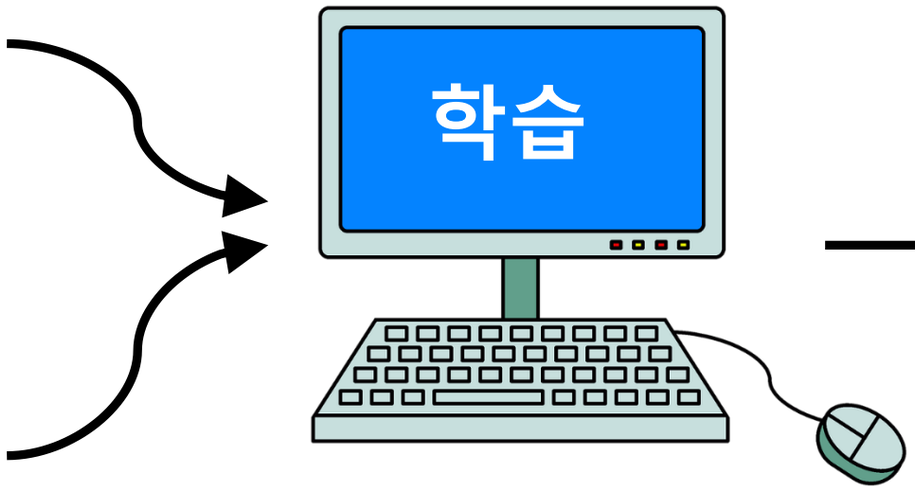
머신러닝 과정

한양대학교
HANYANG UNIVERSITY

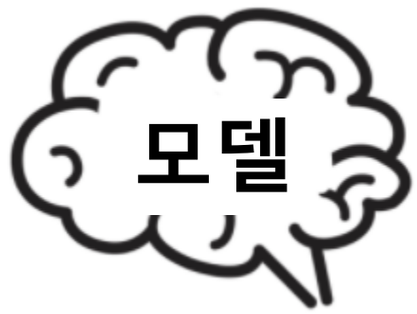
과거의 데이터

원인(독립변수) 결과(종속변수)

시험 날짜	공부시간	성적
2022-05-01	2	40
2022-05-08	3	60
2022-05-15	4	80
2022-05-22	1	20
2022-05-29	3	60



공부시간 X 20 = 성적

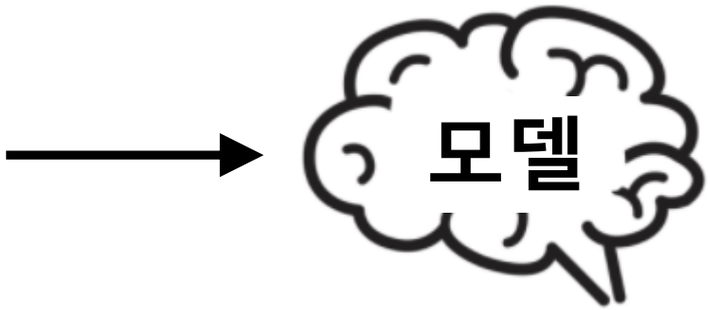


모델을 이용한 예측

한양대학교
HANYANG UNIVERSITY

공부시간 X 20 = 성적

시험 날짜	공부시간	성적
2022-06-05	5	?



100

3. 머신러닝(기계학습)

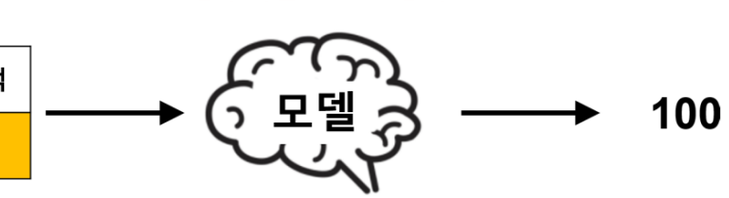
분류 vs 회귀

분류
Classification
예측하고 싶은 결과가 범주형일때

회귀
Regression
예측하고 싶은 결과가 수치형일때

공부시간 X 20 = 성적

시험 날짜	공부시간	성적
2022-06-05	5	?



분류 예 vs. 회귀 예

한양대학교
HANYANG UNIVERSITY

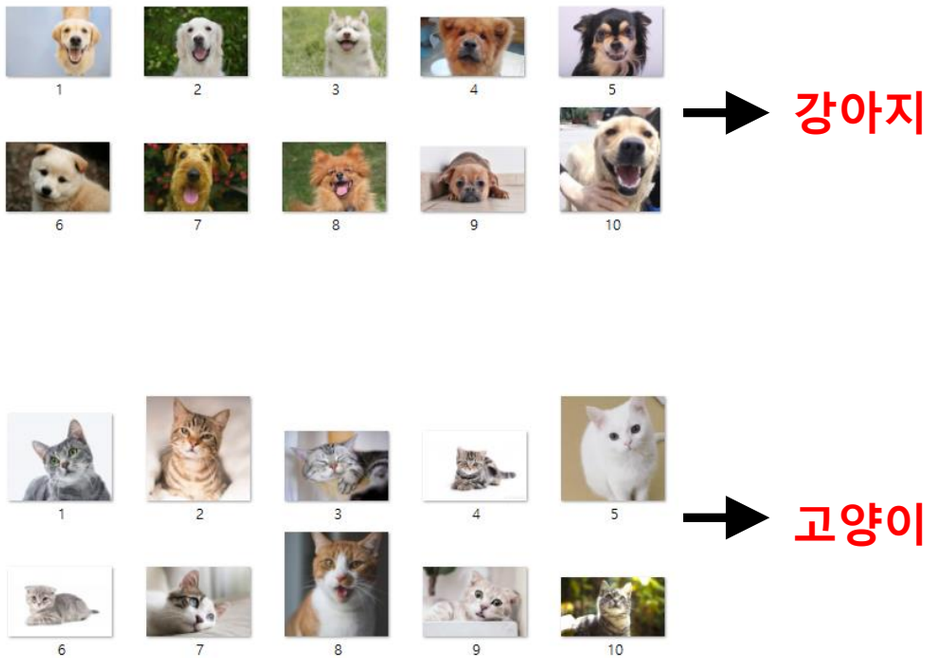
원인(독립변수)	결과(종속변수)
공부시간	합격 여부
종양의 크기, 두께	악성 종양의 여부
품종, 산도, 당도, 지역, 연도 등	와인의 등급
키, 몸무게, 시력, 지병	현역, 공익, 면제
메일 발신인, 제목, 본문 내용 (사용된 단어, 이모티콘 등)	스팸 메일 여부
고기의 지방함량, 지방색, 성숙도, 육색	소고기 등급

원인(독립변수)	결과(종속변수)
공부시간	시험점수
온도	레모네이드 판매량
역세권, 조망 등	집 값
온실 기체량	기온 변화량
자동차 속도	충돌 시 사망 확률
나이	키

지도학습(Supervised Learning)

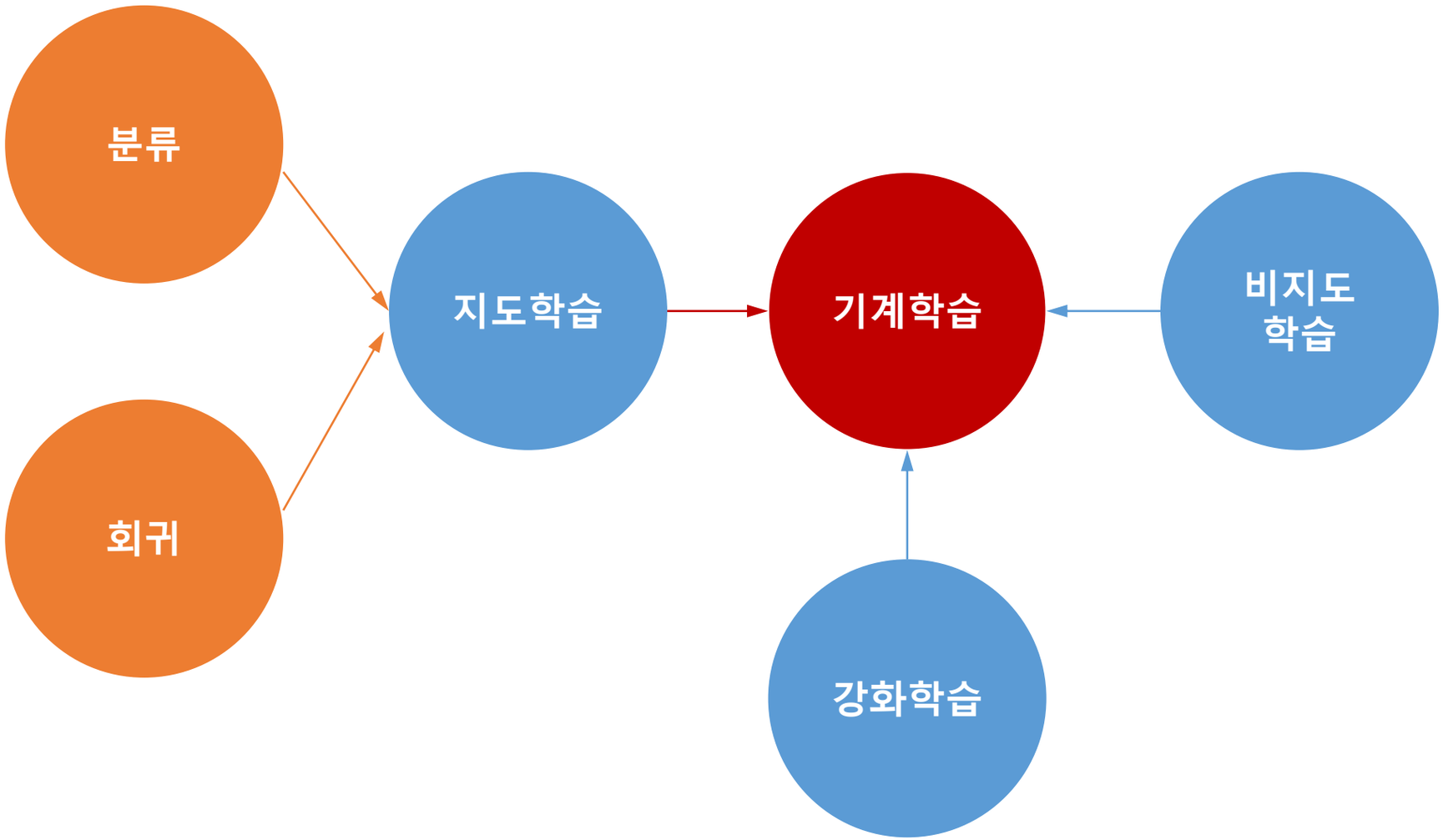
한양대학교 HANYANG UNIVERSITY

❖ 학습하기 위한 데이터와 정답이 존재함

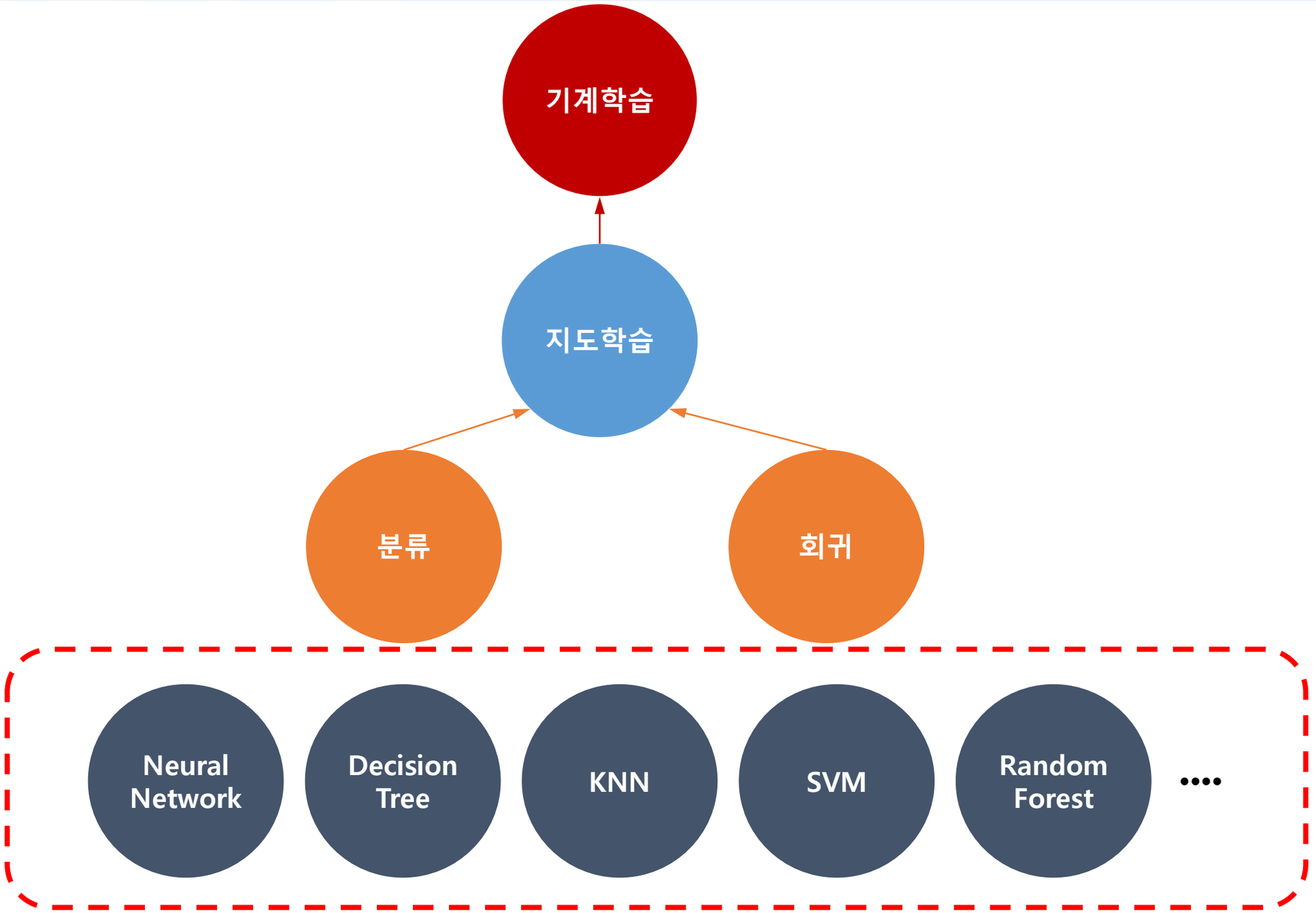


시험 날짜	공부시간	성적
2022-05-01	2 →	40
2022-05-08	3 →	60
2022-05-15	4 →	80
2022-05-22	1 →	20
2022-05-29	3 →	60

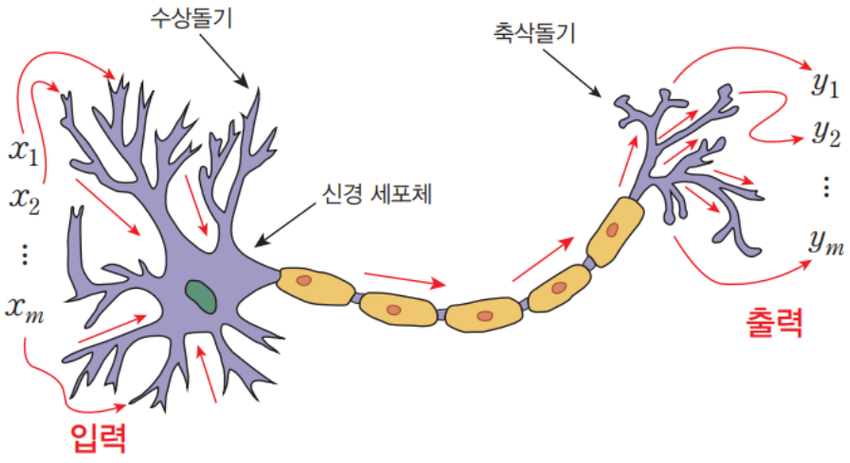
머신러닝(기계학습) 분류

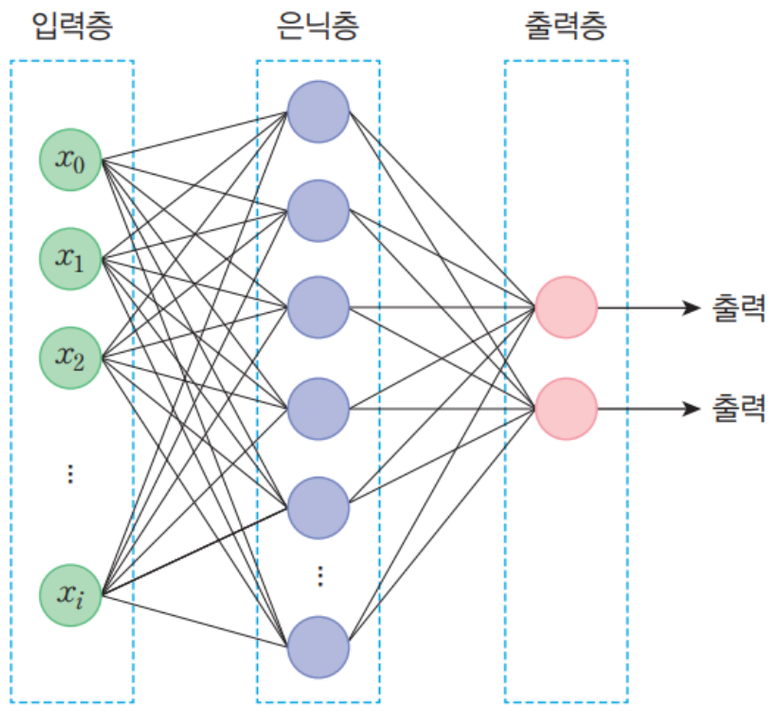


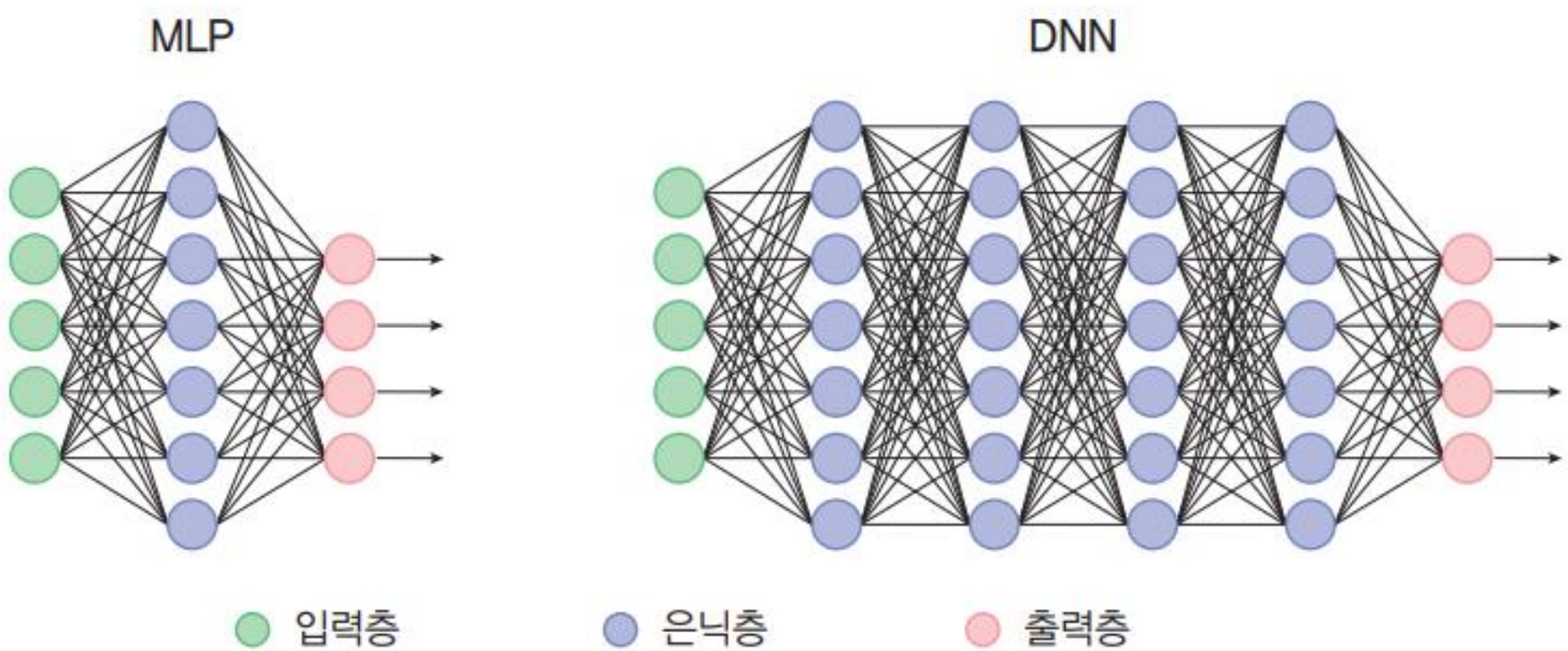
머신러닝 알고리즘


기계학습
지도학습
분류
회귀
Neural Network
Decision Tree
KNN
SVM
Random Forest
....

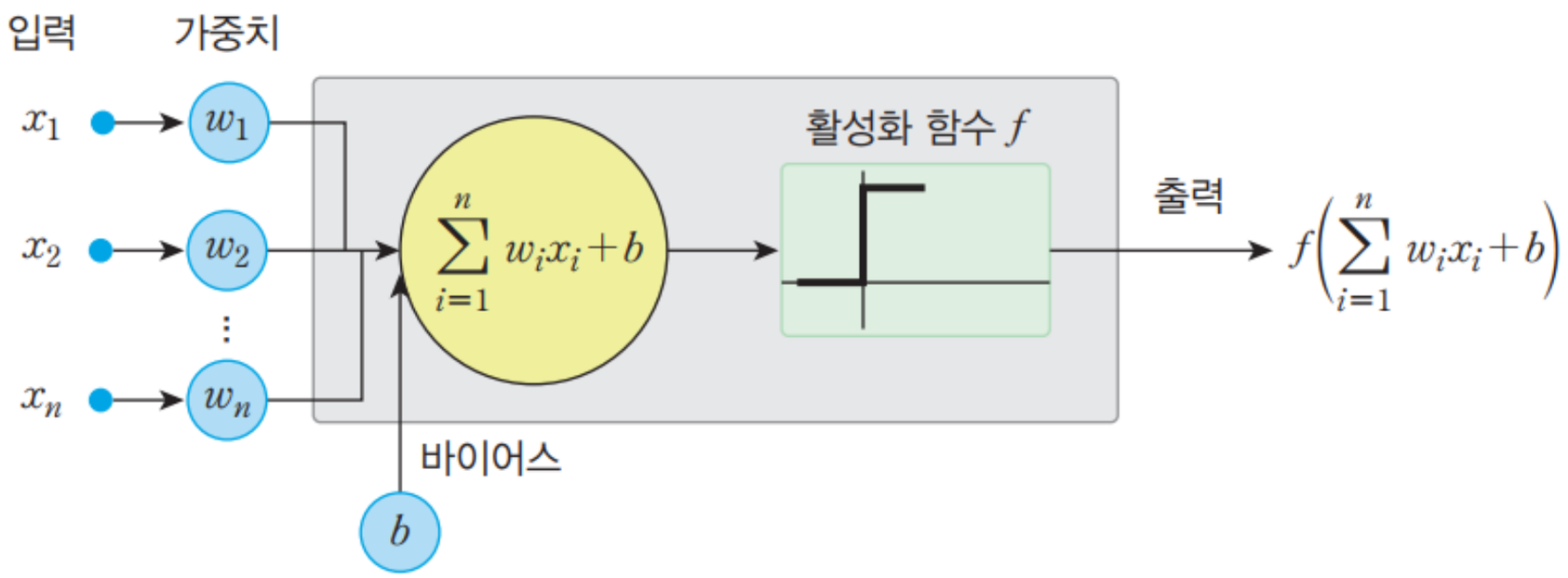
신경망(Neural Network)





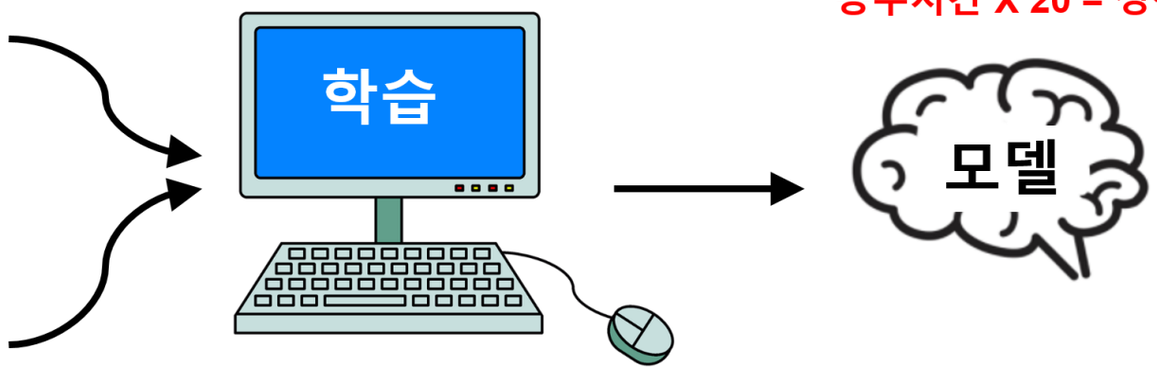

MLP
DNN
입력층
은닉층
출력층

딥러닝 학습 개념



과거의 데이터

시험 날짜	공부시간	성적
2022-05-01	2	40
2022-05-08	3	60
2022-05-15	4	80
2022-05-22	1	20
2022-05-29	3	60



공부시간 X 20 = 성적

머신러닝 프로세스 1

과거의 데이터

시험 날짜	공부시간	성적
2022-05-01	2	40
2022-05-08	3	60
2022-05-15	4	80
2022-05-22	1	20
2022-05-29	3	60

1. 데이터 준비

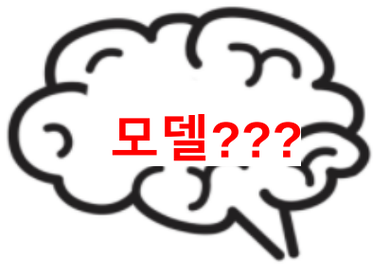
2. 모델의 구조를 만듦

3. 데이터로 모델을 학습시킴

4. 학습된 모델이 만들어짐

공부시간 X 20 = 성적

5. 만들어진 모델 평가



6. 모델의 구조를 변경

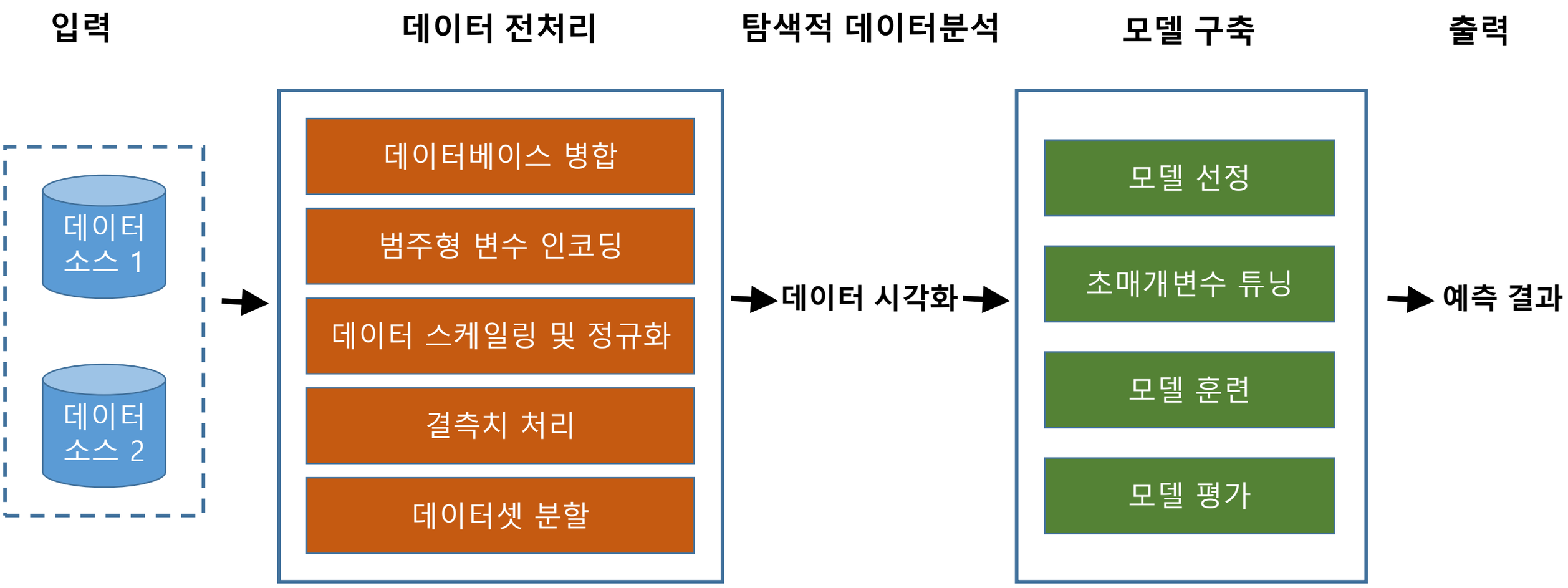
7. 모델을 이용한 예측

공부시간 X 20 = 성적

시험 날짜	공부시간	성적
2022-06-05	5	?

100

머신러닝 프로세스 2


입력
데이터 전처리
탐색적 데이터분석
모델 구축
출력
데이터 소스 1
데이터 소스 2
데이터베이스 병합
범주형 변수 인코딩
데이터 스케일링 및 정규화
결측치 처리
데이터셋 분할
데이터 시각화
모델 선정
초매개변수 튜닝
모델 훈련
모델 평가
예측 결과

4. 딥러닝을 이용한 당뇨병 예측

한양대학교 HANYANG UNIVERSITY

❖ 피마 인디언 당뇨병 데이터셋

컬럼명	의미
Pregnancies	과거 임신 횟수
Glucose	혈장 혈당
BloodPressure	이완기 혈압
SkinThickness	삼두근에서 측정한 피부두겹두께
Insulin	혈청 인슐린 농도
BMI	체질량 지수
DiabetesPedigreeFunction	환자가 당뇨에 얼마나 취약한지(유전적 소인이 어느 정도인지) 요약한 점수, 환자의 당뇨 가족력을 바탕으로 추정함
Age	햇수로 계산한 나이
Outcome	예측 목표 변수, 최초 측정 이후 5년 내 당뇨가 발병하면 값이 1이며, 반대로 미발병하면 0임.

❖ 미국 뉴욕의 택시 요금 데이터셋

컬럼명	의미
key	pickup_datetime 커럼과 동일한 값으로 고유 ID로 사용
fare_amount	운행을 마친 후 지불한 요금 모델이 예측할 목표 변수임
pickup_datetime	승객이 승차한 날짜(년, 월, 일)와 시간(시, 분, 초)
pickup_longitude	승차 위치(위도)
pickup_latitude	승차 위치(경도)
dropoff_longitude	하차 위치(위도)
dropoff_latitude	하차 위치(경도
passenger_count	승객 수

Thank you!

Beyond The Engine of Korea

HANYANG UNIVERSITY