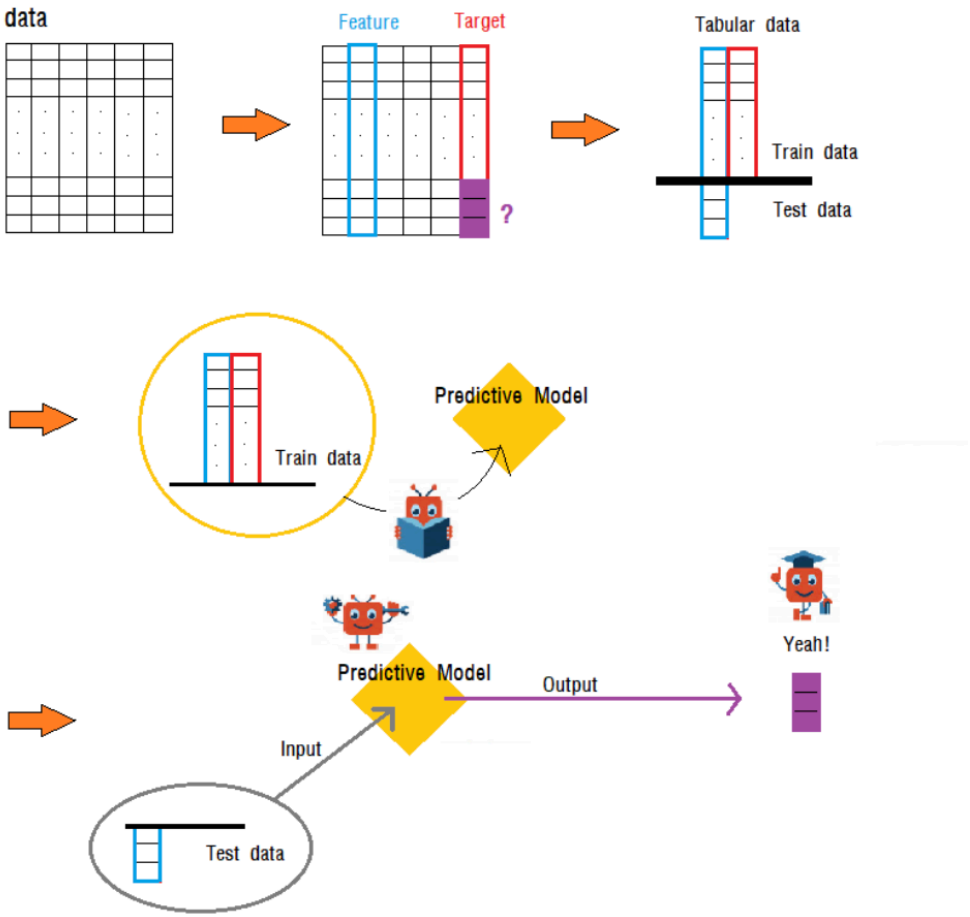
Classification & Regression Basics

Machine Learning (기계 학습)

- All about **predicting** based on the data we have

머신러닝 Process:

1. 특성 데이터(feature)와 타겟 데이터 (target)을 식벽한 후 데이터 전처리를 진행한다.
2. 가지고 있는 데이터를 학습 데이터 (train data)와 테스트 데이터(train data)로 나눈다.
3. 학습 데이터 (train data)로 예측 모델을 (fit) 만든다.
4. 테스트 데이터(train data)의 특성 데이터(feature)를 예측 모델에 넣어 fit 시킨 후 미지의 타겟 데이터(tager) 값을 예측한다.



분류(Classification)와 회귀(Regression)의 차이점은 Label(=Target, 예측할 값)의 상태이다.

- Label이 이산값(이진분류(0또는1, True또는False), 멀티분류)인 경우는 **분류알고리즘**모델로 예측한다.

- Label이 연속값인 경우는 **회귀알고리즘**모델로 예측한다.

Classification (분류 학습)

Target이 **이산값**(이진분류(0또는1, True또는False), 멀티분류)인 경우

Feature들을 통해 **과일인지 야채인지 분류**해보기

Train Data

idx	sweetness	Color	Taste	Type
1	4	Green	Bitter	Veggie
2	8	Red	Sweet	Fruit
…				
200	7	Red	Sweet	Fruit

Test Data

idx	sweetness	Color	Taste
1	4	Green	Bitter
2	8	Red	Sweet
…			
200	7	Red	Sweet

Regression (회귀 학습)

Target이 **연속값**인 경우는 **회귀알고리즘**모델로 예측한다.

Feature들을 통해 **몸무게 예측**해보기

Train Data

idx	Height	Sex	Age	Weight
1	160	Women	18	51
2	181	Men	23	69
…				
200	153	Women	27	49

Test Data

idx	Height	Sex	Age
1	160	Women	18
2	181	Men	23
…			
200	153	Women	27

Classification/Regression Models

분류 알고리즘	설명
로지스틱 회귀 (Logistic Regression)	독립변수와 종속변수의 선형 관계성에 기반한 알고리즘
결정트리 (Decision Tree)	데이터 균일도에 따른 규칙 기반의 알고리즘
나이브 베이즈 (Naive Bayes)	베이즈(Bayes) 통계와 생성 모델에 기반한 알고리즘
서포트 벡터 머신 (Support Vector Machine)	개별 클래스 간의 최대 분류 마진을 효과적으로 찾아주는 알고리즘
최소 근접 알고리즘 (Nearest Neighbor)	근접 거리를 기준으로 하는 알고리즘
신경망 (Neural Network)	심층 연결 기반의 알고리즘
앙상블 (Ensemble)	서로 다른(또는 같은) 머신러닝 알고리즘을 결합한 알고리즘