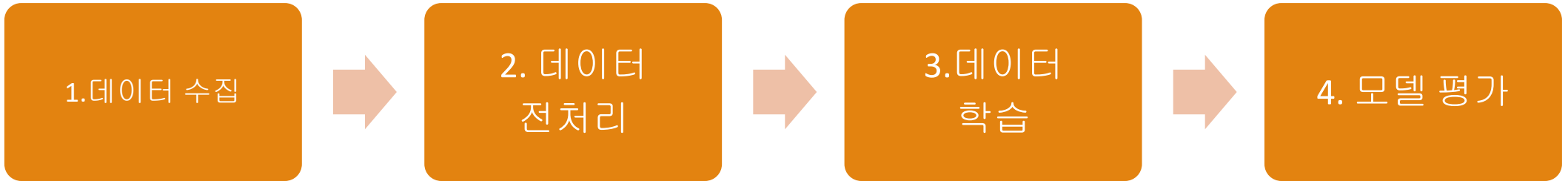
머신 러닝 개념

박승원

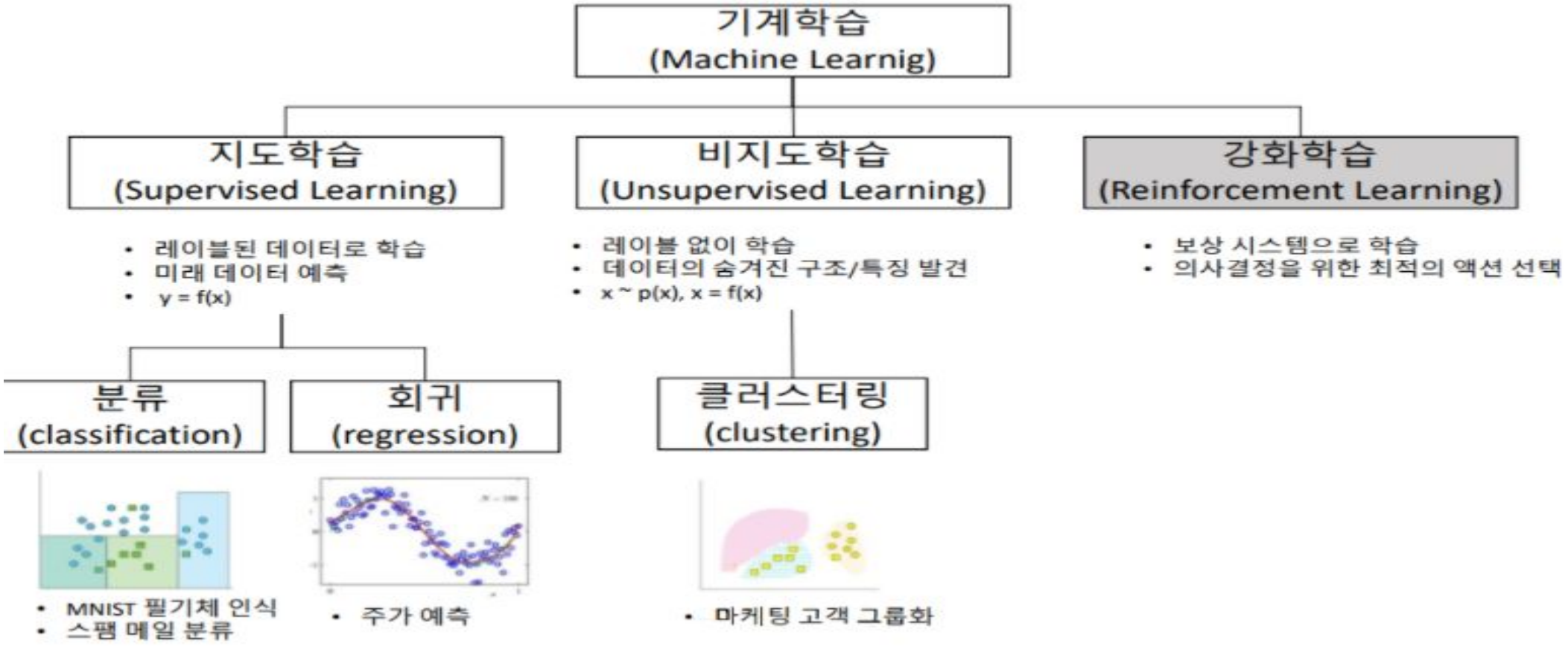
1.머신 러닝 훈련 과정



1.머신 러닝 훈련 과정

단계	설명
1.데이터 수집	1. 머신 러닝의 첫번째 단계는 문제 정의 단계 2. 데이터를 직접 수집 또는 공개 되어 있는 소스를 사용
2.데이터 전 처리	1. 데이터를 정리하는 작업 2. 데이터 충분: 누락된 데이터 제거 3. 데이터 제한적:누락된 값 채 워줌
3.데이터 학습	1.학습 알고리즘 먼저 선택 2.학습 데이터와 검증 데이터로 나누기
4. 모델 평가	1.데이터가 충분하지 못할 경우에 교차 평가

1.머신 러닝 훈련 과정-데이터 학습 알고리즘



2.알고리즘 종류-(1)분류

1. 분류

- 1.이진 분류
- 2. 다중 분류

2.알고리즘 종류-(1)회귀

1. 이진 분류

- 두개의 클래스로 분류
- 예) 예/ 아니요 구분

2. 다중 분류

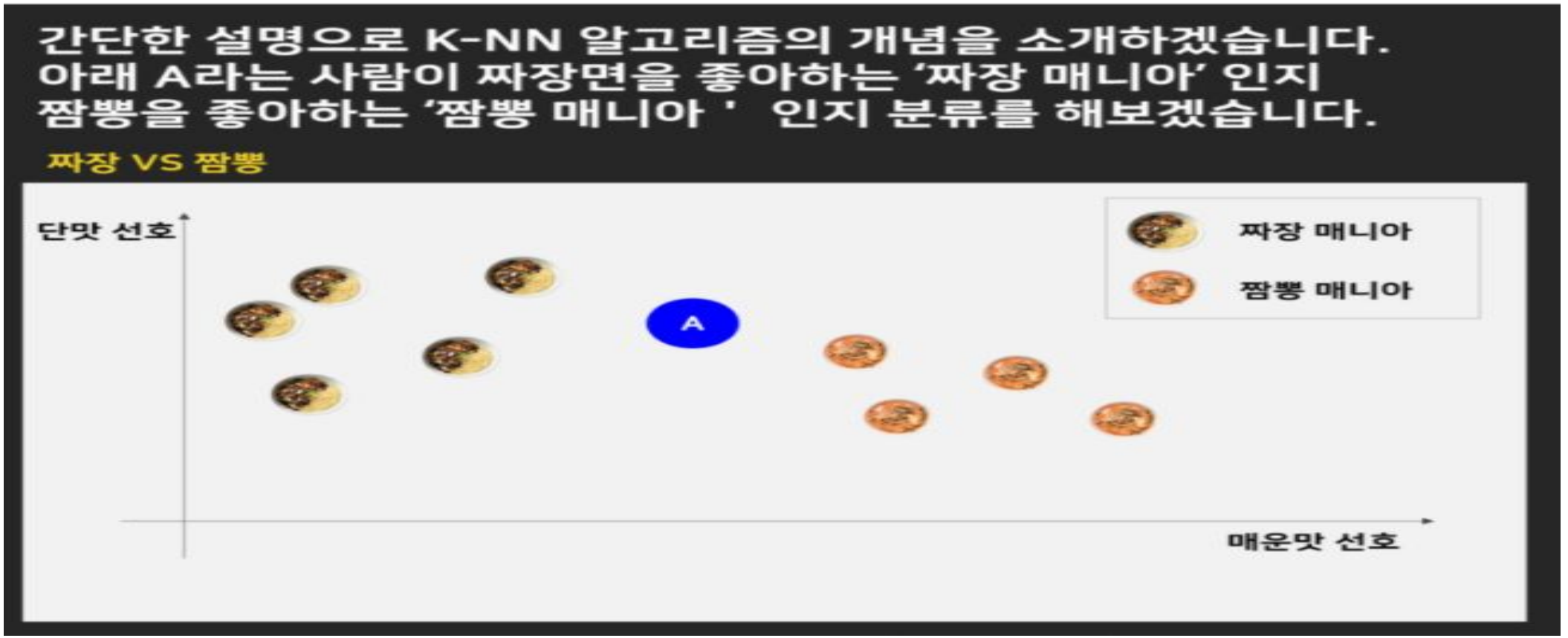
- 셋 이상의 클래스로 분류
- 예) 부모와 자녀의 키의 연관성

2.알고리즘 종류-(1)분류 알고리즘 종류

knn	1.유유상종 “같은 날개를 가진 새들 끼리 모인다”란 뜻의 속담처럼 머신 러닝 데이터를 가장 가까운 유사속성에 따라 분류
	2.지도 학습에 한 종류로 거리기반의 분류 모델
	3.데이터로 부터 거리가 가까운 ‘k’개의 다른 데이터의 레이 불을 참조하여 분류하는 알고리즘 으로 거리를 측정 할때 유클 리디안 거리 계산법을 사용

2.알고리즘 종류-(1)분류 알고리즘 종류

간단한 설명으로 K-NN 알고리즘의 개념을 소개하겠습니다.
아래 A라는 사람이 짜장면을 좋아하는 '짜장 매니아' 인지
짬뽕을 좋아하는 '짬뽕 매니아 ' 인지 분류를 해보겠습니다.



2.알고리즘 종류-(1)분류 알고리즘 종류

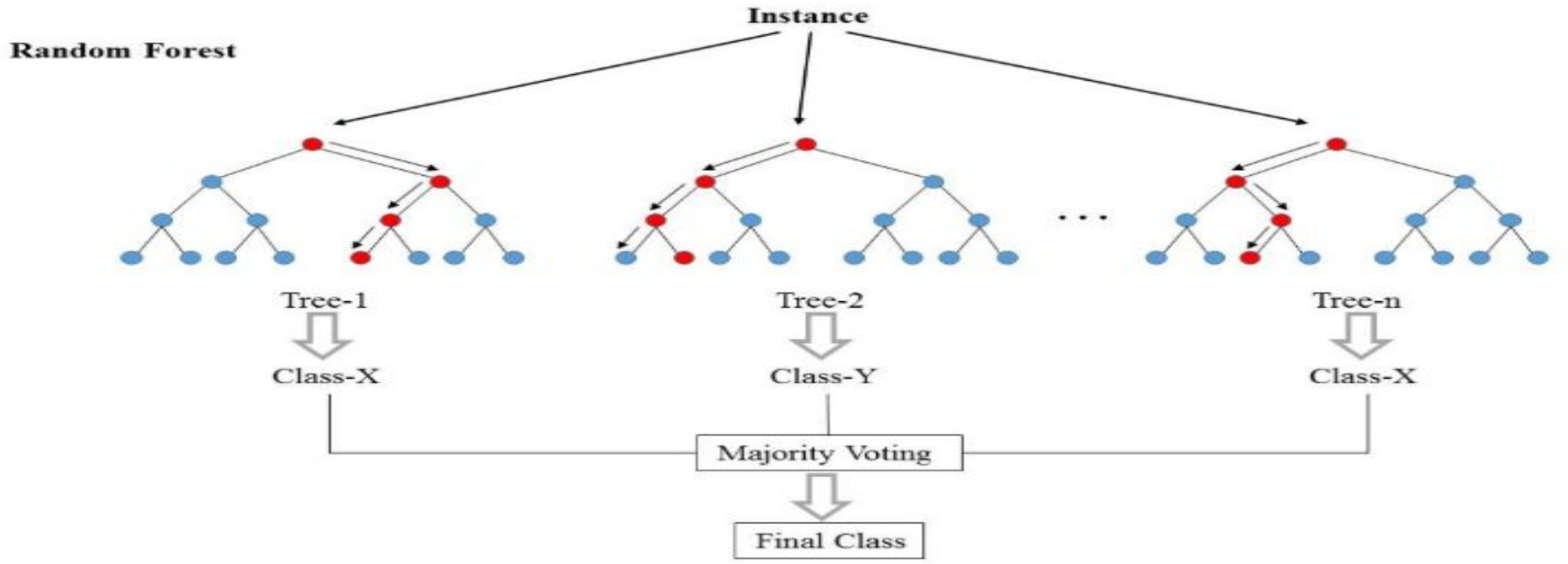
Decision Tree(의사 결정 트리)	1. 가장 단순한 classifier 중 하나로, decision tree와 같은 도구를 활용하여 모델을 그래프로 그리는 매우 단순한 구조로 되어 있다. 이 방식은 root에서부터 적절한 node를 선택하면서 진행하다가 최종 결정을 내리게 되는 model이다.
	2. 이 트리의 장점은 누구나 쉽게 이해할 수 있고, 결과를 해석 할수있다. -예를 들어 yes를 선택했던 것을 no로 바꾸기만 하면 간단하게 로직을 바꿀 수 있다.
	3. 가지고 있는 데이터의 Feature를 분석해서 Tree를 Build하는 과정이 제일 중요하다.

2.알고리즘 종류-(1)분류 알고리즘 종류

2.알고리즘 종류-(1)분류 알고리즘 종류

Random Forest	1. Decision tree가 여러개 모여 Forest를 이룬 것이다.
	2. Decision tree보다 작은 Tree가 여러개 모이게 되어, 모든 트리의 결과들을 합하여 더많은 값을 최종결과로 본다.
	3. 여러 의사 결정 나무를 생성한 후에 다수결 또는 평균에 따라 출력 변수를 예측하는 알고리즘입니다. 즉 의사 결정 나무와 bagging을 혼합한 형태라고 볼 수 있습니다.

2.알고리즘 종류-(1)분류 알고리즘 종류



2.알고리즘 종류-(2)회귀

1. 정의

- 연속적인 숫자 또는 부동소수점 를 예측

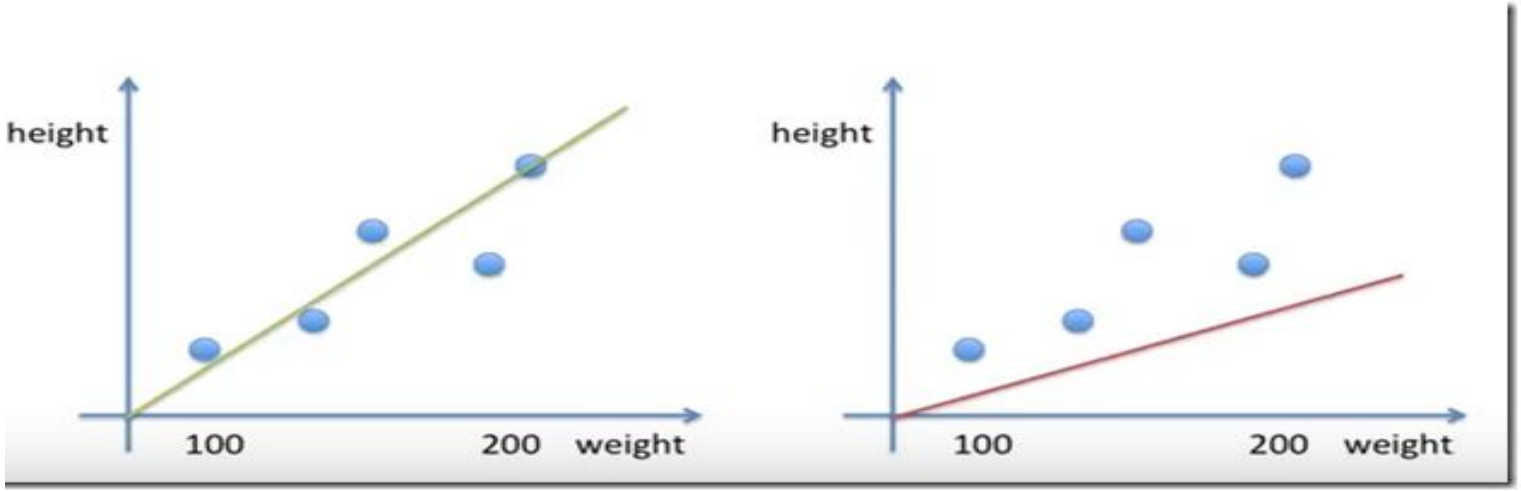
2. 사례

- 어떤 사람의 교육 수준,나이,주거지를 바탕으로 연간 소득을 예측

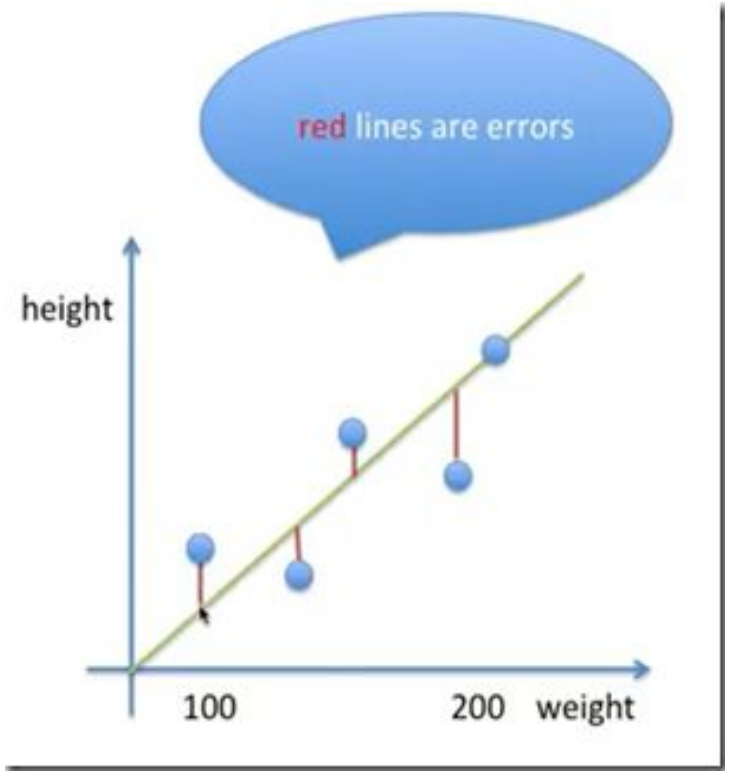
2.알고리즘 종류-(2)회귀 알고리즘 종류

1. Linear Regression

- 일차 함수 개념인 y=ax+b 직선을 임의로 그려 놓고 그 직선을 바탕 으로 예측하는 것이 선형 회귀

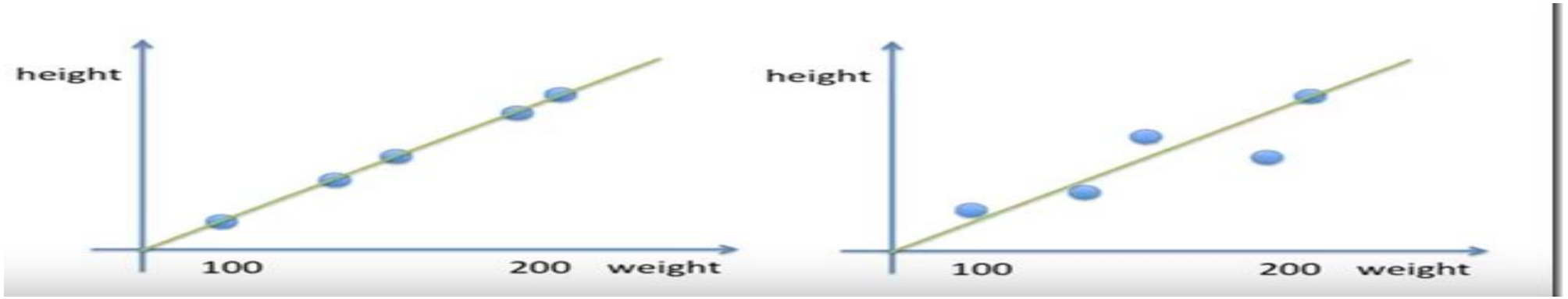


2.알고리즘 종류-(1)회귀 알고리즘 종류



예측 하기 위해 만든 모델인 y=ax+b 직선 과 실제 찍어 놓은 점들의 y값 차이를 error

2.알고리즘 종류-(2)회귀 알고리즘 종류

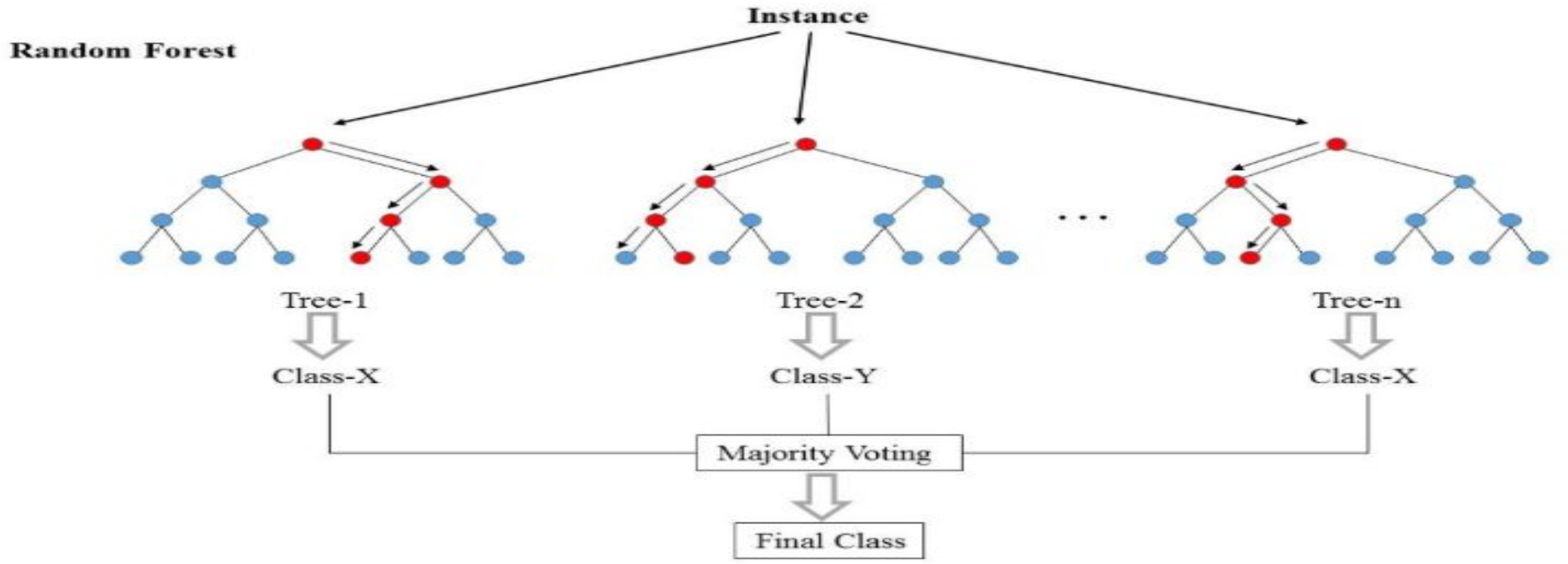


1.왼쪽이 더 예측 하기 쉬운 모델
-왼쪽의 직선 모델 에는 에러가 없다

2.알고리즘 종류-(2)분류 알고리즘 종류

Random Forest	1. Decision tree가 여러개 모여 Forest를 이룬 것이다.
	2. Decision tree보다 작은 Tree가 여러개 모이게 되어, 모든 트리의 결과들을 합하여 더많은 값을 최종결과로 본다.
	3. 여러 의사 결정 나무를 생성한 후에 다수결 또는 평균에 따라 출력 변수를 예측하는 알고리즘입니다. 즉 의사 결정 나무와 bagging을 혼합한 형태라고 볼 수 있습니다.

2.알고리즘 종류-(2)분류 알고리즘 종류



2.알고리즘 종류-(2)분류 알고리즘 종류

1. 부 스팅 알고리즘 (Boosting Algorithm)

- 부스팅은 머신러닝 앙상블 기법 중 하나로 약한 학습기(weak learner)들을 순차적으로 여러 개 결합하여 예측 혹은 분류 성능을 높이는 알고리즘이다

-즉, 약한 학습기들을 결합하여 강한 예측 모형을 만드는 것이다.

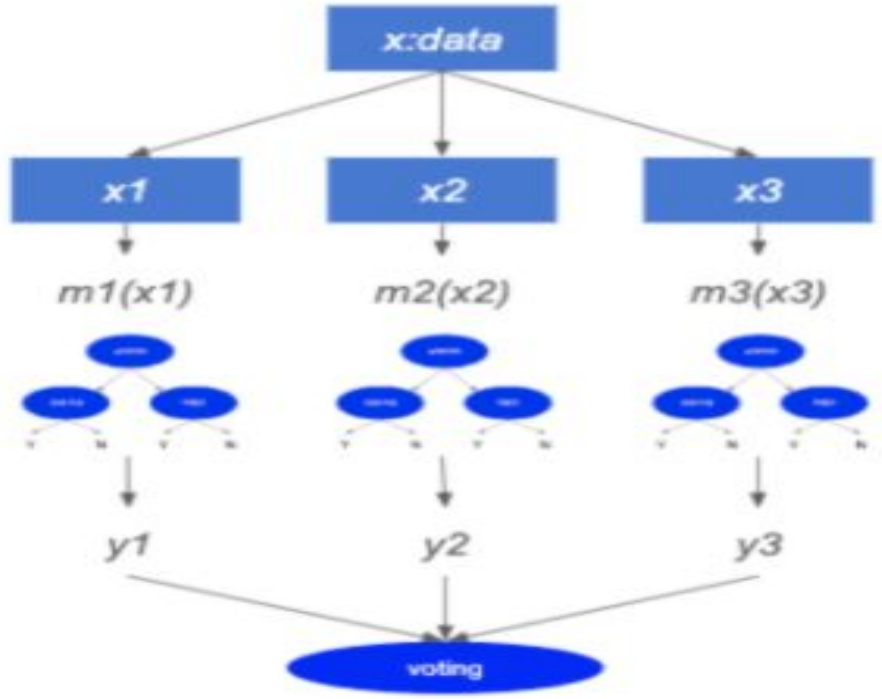
2. 부 스팅 알고리즘 (Boosting Algorithm) 원리

- 여러 개의 알고리즘이 순차적으로 학습-예측을 하면서 이전에 학습한 알고리즘의 예측이 틀린 데이터를 올바르게 예측할 수 있도록, 다음 알고리즘에, 가중치를 부여하여 학습과 예측을 진행하는 방식입니다.

2.알고리즘 종류-(2)분류 알고리즘 종류

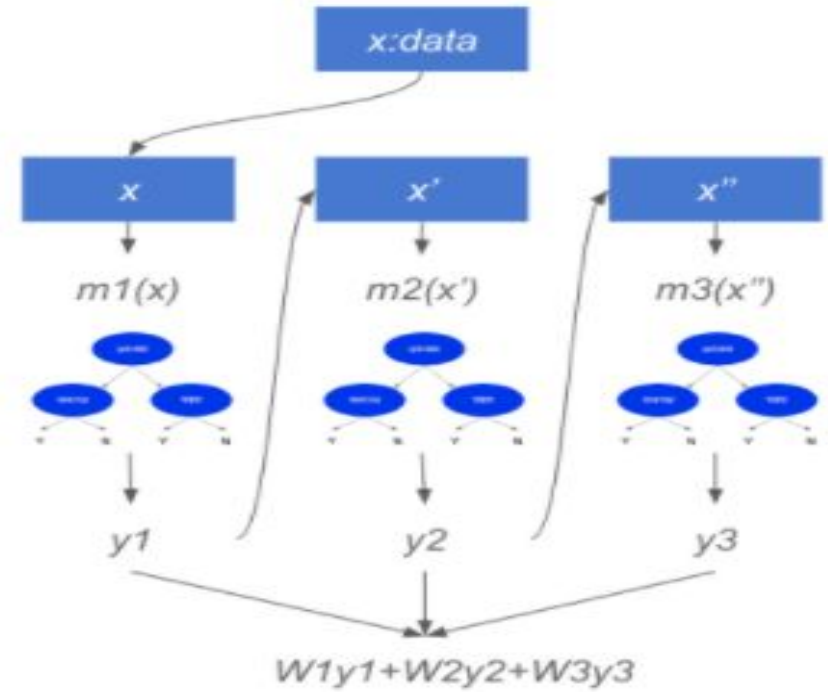
Bagging

1) Bagging은 여러 모델을 사용 할때, 각 모델에서 나온 값을 계산하여, 최종 결과값을 내는 방식이다.예를 들어 아래 그림과 같이 모델 m1,m2,m3 3개의 모델이 있 을 때, 입력 데이터 X를 모델 m1~3에 넣고, 그 결과값을 받아서 합산 (또는 평 균등 여러가지 방법이 있음) 해서, 최종 결과를 취하는 방식이다.



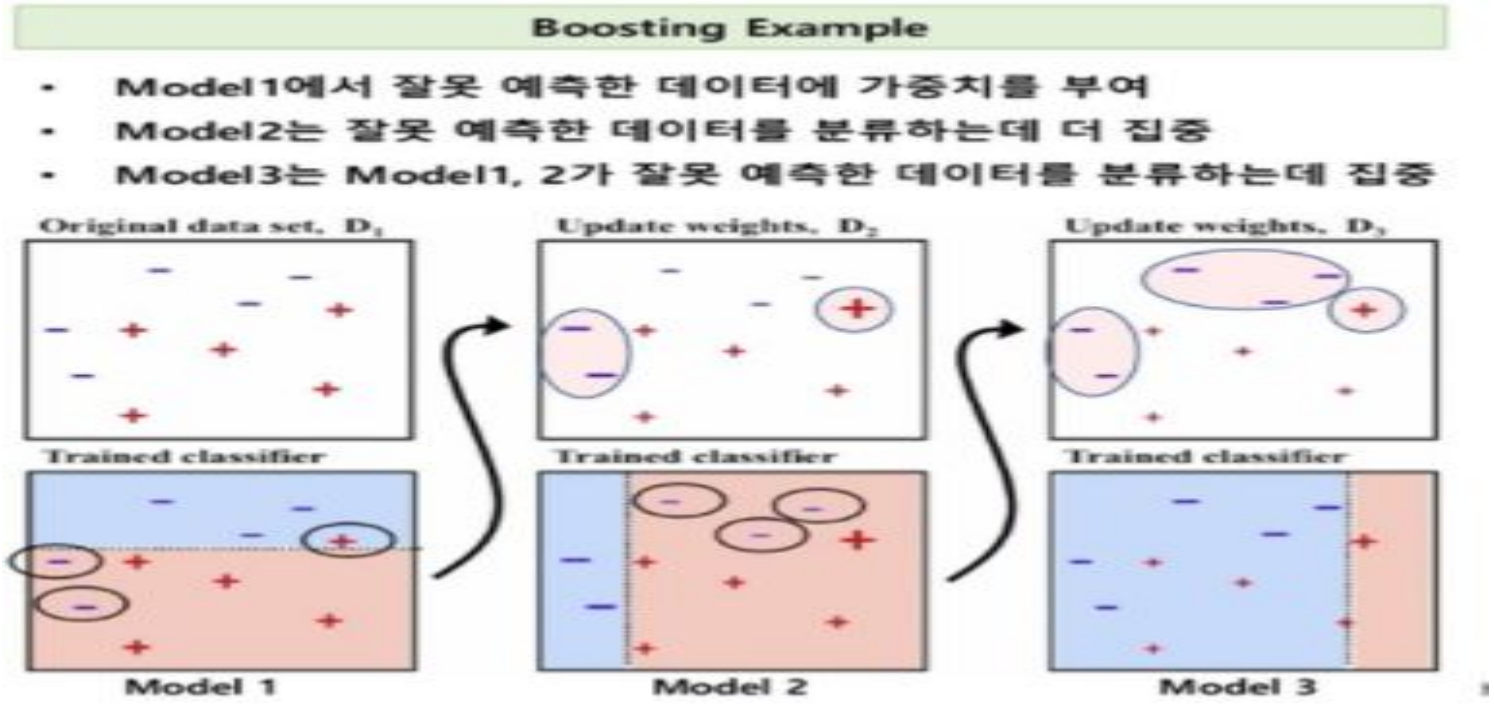
2.알고리즘 종류-(2)분류 알고리즘 종류

Boosting은 원리가 다른데 먼저 m1~3 모델이 있을때, m1에는 x에서 샘플링된 데이터를 넣는다. 그리고, 나온 결과중에서, 예측이 잘못된 x중의 값들에 가중치를 반영해서 다음 모델인 m2에 넣는다. 마찬가지로 y2 결과에서 예측이 잘못된 x'에 값들에 가중치를 반영해서 m3에 넣는다.
그리고, 각 모델의 성능이 다르기 때문에, 각 모델에 가중치 W를 반영한다. 이를 개념적으로 표현하면 다음과 같은 그림이 된다.



2.알고리즘 종류-(2)분류 알고리즘 종류

부 스 팅 알고리즘 원리



2.알고리즘 종류-(2)분류 알고리즘 종류

1. XGBOOST

- 약한 분류기를 세트로 묶어서 정확도를 예측하는 기법이다
- 욕심쟁이(Greedy Algorithm)을 사용하여 분류기를 발견하고 분산처리를 사용하여 빠른 속도로 적합한 비중 파라미터를 찾는 알고리즘이다.
- 욕심쟁이(Greedy Algorithm)-미래를 생각하지 않고 각 단계에서 가장 최선의 선택을 하는 기법

2. XGBOOST 장점

- 병렬 처리를 사용하기에 학습과 분류가 빠르다
- 유연성이 좋다. 커스텀 최적화 옵션을 제공한다
- 욕심쟁이(Greedy-algorithm)을 사용한 자동 가지치기가 가능하다. 과적합이 잘일 어나지 않는다.
- 다른 알고리즘과 연계하여 앙상블 학습이 가능하다.

2.알고리즘 종류-(2)분류 알고리즘 종류

1. LightGBM 장점

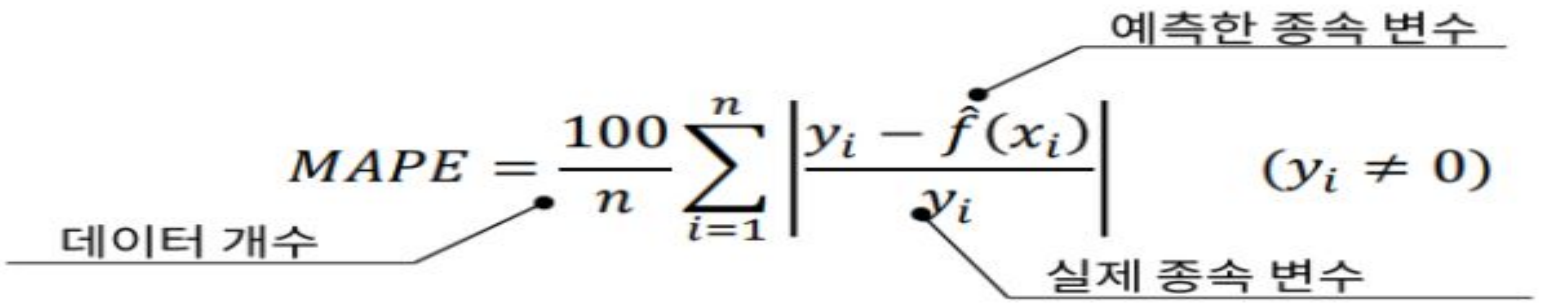
- 학습하는데 걸리는 시간이 적다, 빠른 속도
- 메모리 사용량이 상대적으로 적은 편이다
- categorical feature들의 자동 변환과 최적 분할
- GPU 학습 지원

2. LightGBM 단점

- 작은 dataset을 사용할 경우 과 적합 가능성이 크다
- 일반적으로 10,000개 이하의 데이터를 적다고 한다

3.회귀 알고 리 즘 평가 방법

MAPE(mean absolute percentage error)

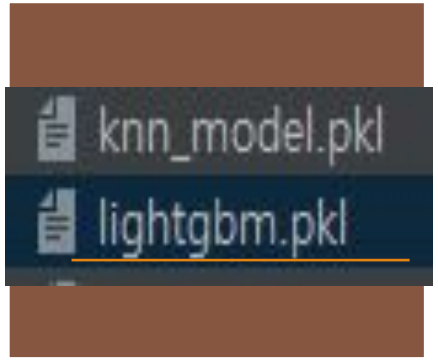
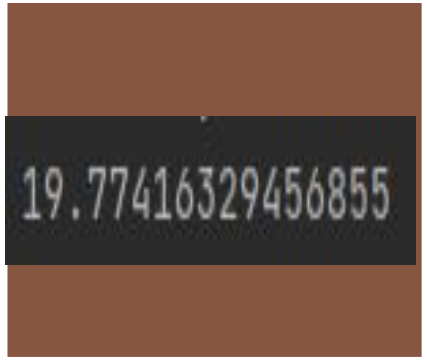
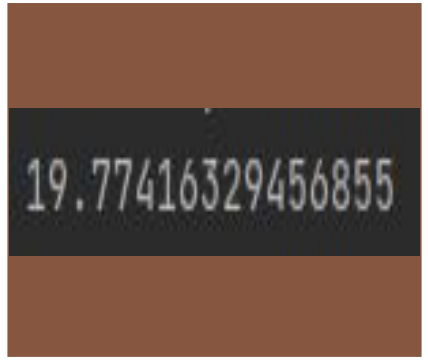


1. MAPE는 **퍼센트 값을 가지며** 0에 가까울수록 회귀 모형의 성능이 좋다고 해석할 수 있음

2. 0~100% 사이의 값을 가져 이해하기 쉬우므로 성능 비교 해석이 가능

MSE가 100이다 했을 때 이 모형이 좋은지 판단하기가 어려움 그래서 MAPE의 퍼센트 값을 통해 성능 평가

3.회귀 알고 리 즘 평가 방법- 평가 실제 사례

1.모델 생성	2.테스트	2.테스트 결과 해석
		
1.모델 생성 4개의 모델 중 점수 가 높은 모델로 저장 (lightgbm.pkl)	2.점수 출력 lightgbm.pkl 파일을 불러온 다음 테스트 데이터 로 성능 평가	3.점수 해석 1)100-19=81 2)최저가 예측에 알맞은 모델

3.회귀 알고 리 즘 평가 방법

평균 제곱 오차(Mean squared error)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

1. 오차의 제곱을 평균으로 나눈 것이다.

2. MSE가 0에 가까울수록 추측한 값이 원본에 가까운 것이기 때문에 정확도가 높다고 할 수 있다.

3. 예측 값과 실제 값 차이의 면적의 평균이라고 할 수 있다