

인공지능 강의



AI 기술의 인플레이션

“더 이상 코딩을 할 필요가 없는 시대가 도래할 것”



*모든 자료에 대한 권한은 메타코드에 있으며, 무단으로 자료를 복제 및 배포 등 유료목적으로 활용하시면 별도의 조치가 들어갈 수 있습니다.

대작 (代作)

“아이디어와 산출물만 남는 시대”



*모든 자료에 대한 권한은 메타코드에 있으며, 무단으로 자료를 복제 및 배포 등 유료목적으로 활용하시면 별도의 조치가 들어갈 수 있습니다.

“Data Science를 통해 산출물을 만든다는 관점”

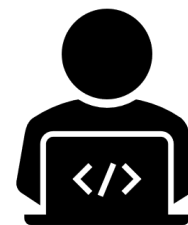
스토리작가

(만화의 뼈대와 스토리구성)



작화가

(그림을 그리는 사람)



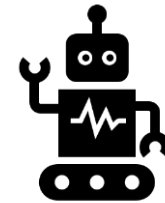
Data Scientist /
Machine Learning Engineer

“Data Science를 통해 산출물을 만든다는 관점”

스토리작가
(만화의 뼈대와 스토리구성)



작화가
(그림을 그리는 사람)




AI모형을
만드는 AI



Citizen Data Scientist의 덕목

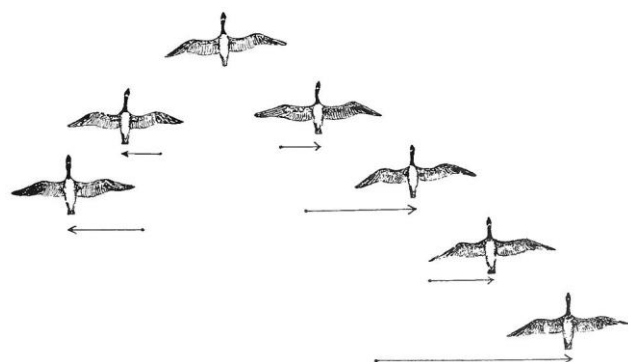
Data Scientist로 현실의 문제를 풀기 위해서
Context를 이해하고 적절한 '선택'을 할 수 있어야함



어떤 알고리즘?
어떤 모형?
어떻게 학습 & 제공
어떻게 평가 및 활용

창발성이란?

- 단순한 규칙들이 상호작용을 통해 놀라운 성능을 발휘함
- ML알고리즘은 생각보다 '간단'



기본적인 상호작용 (Agent)

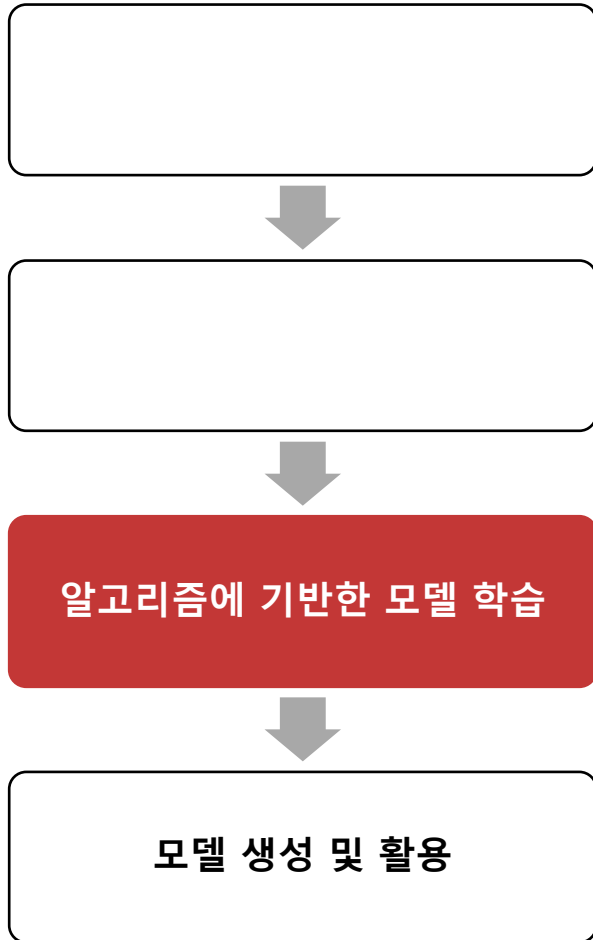
Bottom-up



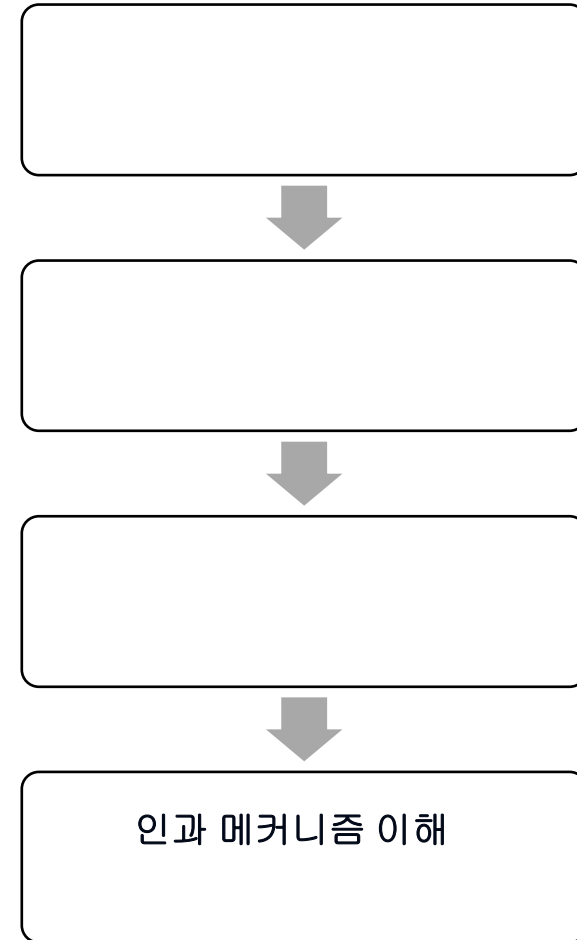
집단 행동

머신러닝의 특수성

머신러닝

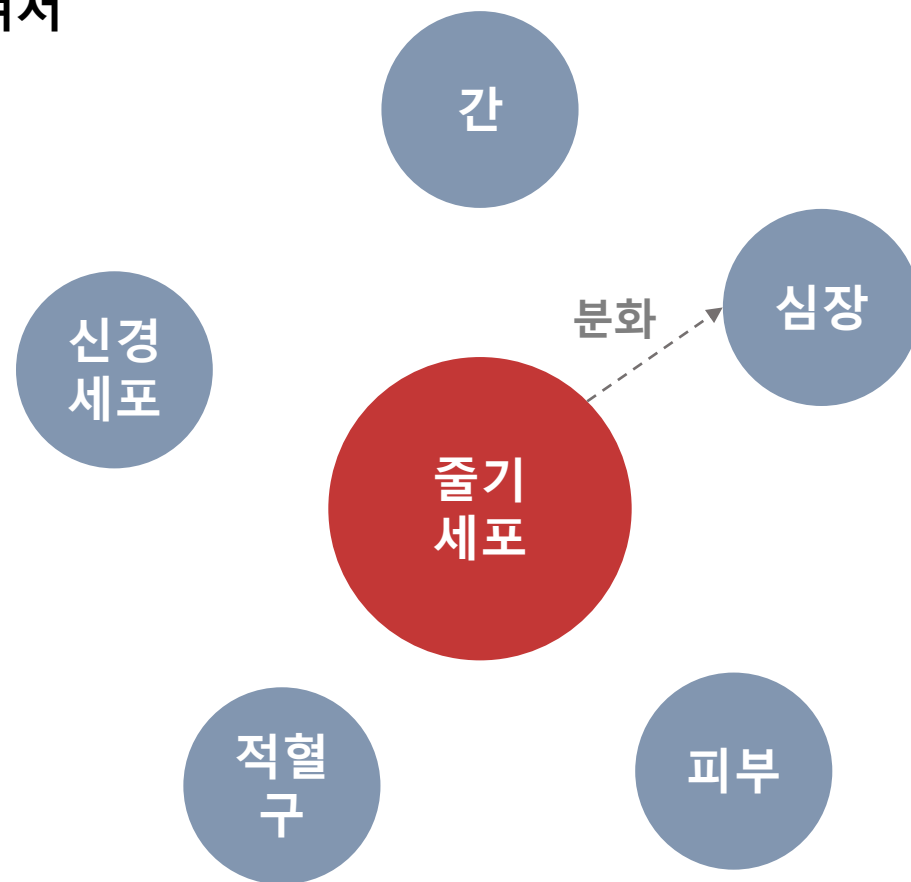


가설검증



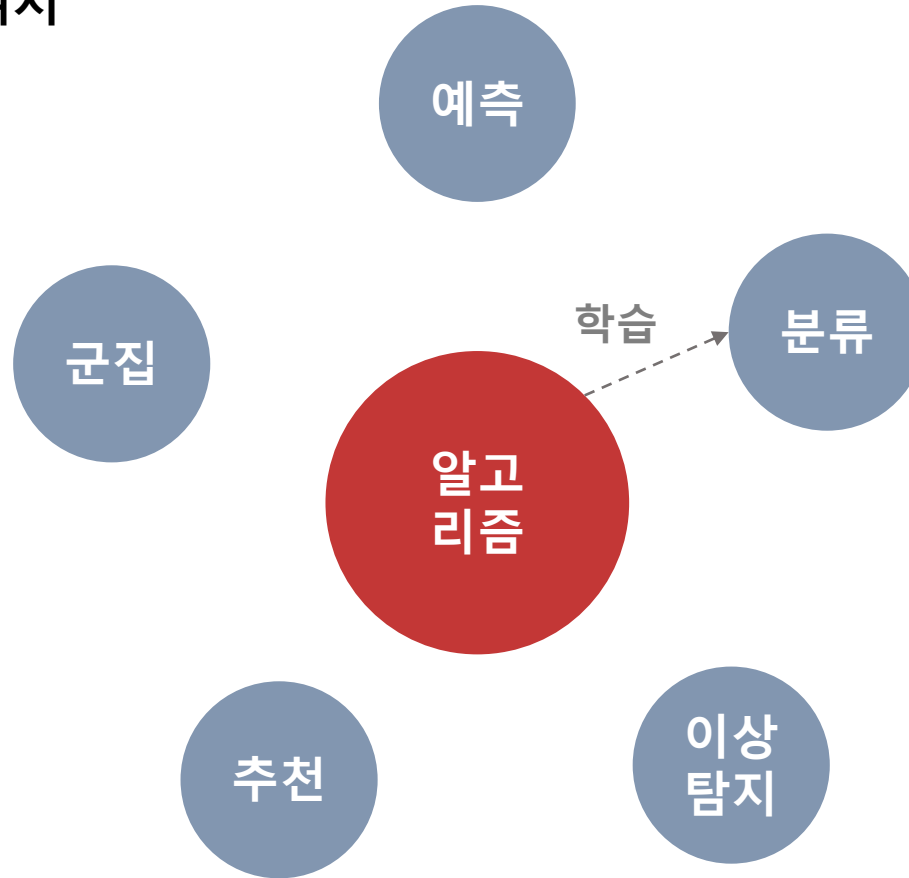
줄기세포의 사상

‘줄기세포’를 ‘분화’를 시켜서
세포/조직/장기 생성



머신러닝의 사상

'알고리즘'을 '학습'를 시켜서
'모델'을 생성

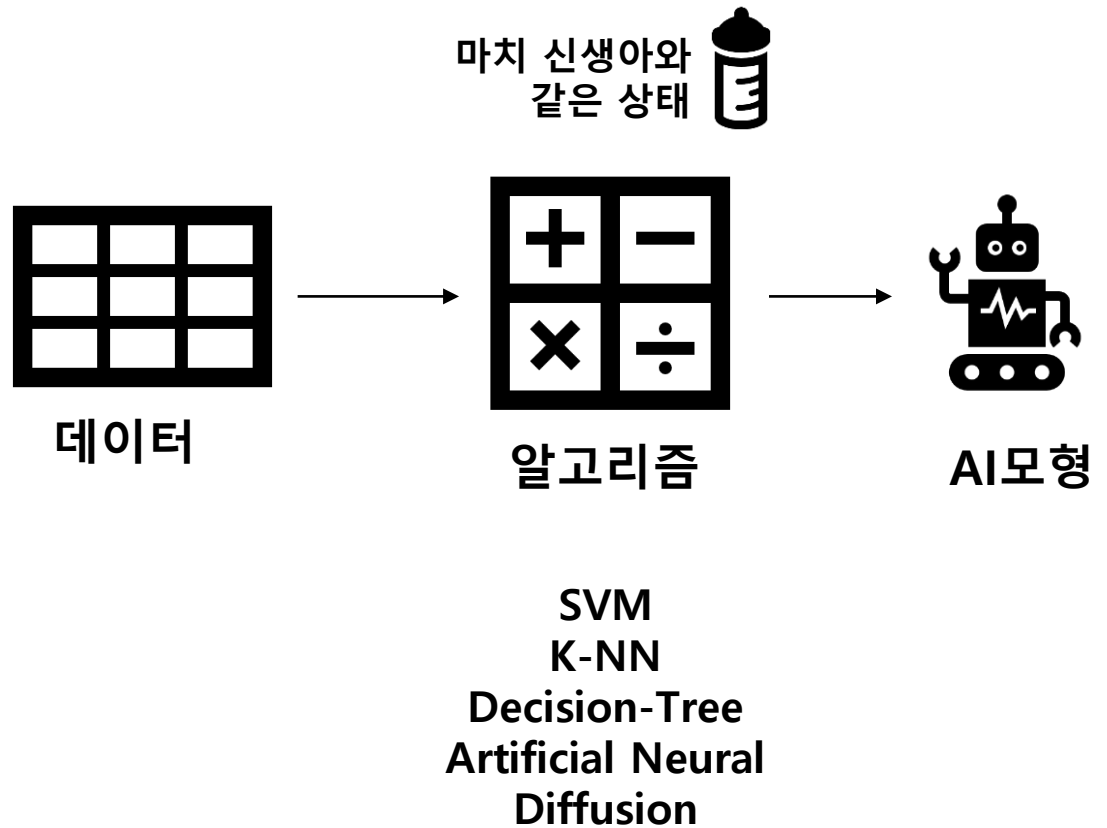


'학습'을 어떻게 할지가 우리의 핵심

- Objective(Loss) function
- Hyper-Parameter
- Train, Validation, Test Sets
- Evaluation Metrics

데이터 기반 접근 방식

- 불확실성
- 적응성과 진화
- 블랙박스
- 데이터 의존성
- 도메인 독립성



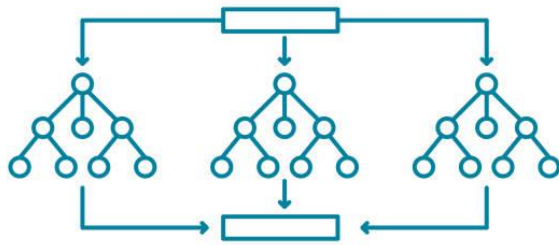
근간이 되는 알고리즘을 배워봅시다

‘알고리즘’ is 단순한 규칙

동작원리를 이해해야 판단과 최선의 의사결정을 할 수 있음

Decision-Tree

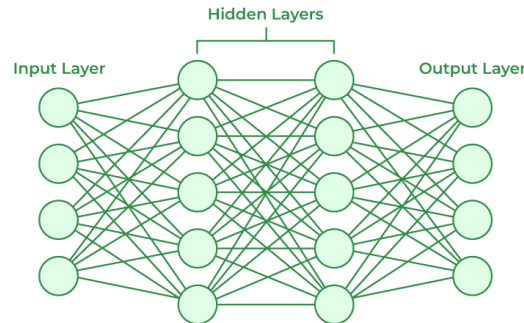
스무고개



#현업활용#작은데이터

Neural-Network

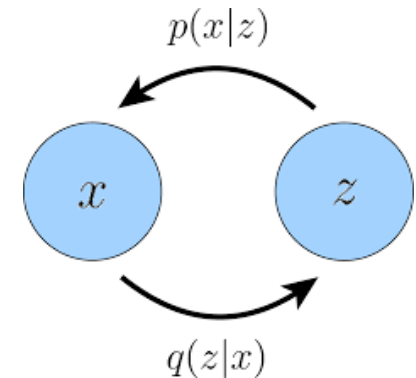
인간의뇌



#딥러닝#데이터과학핵심

Diffusion Algorithm

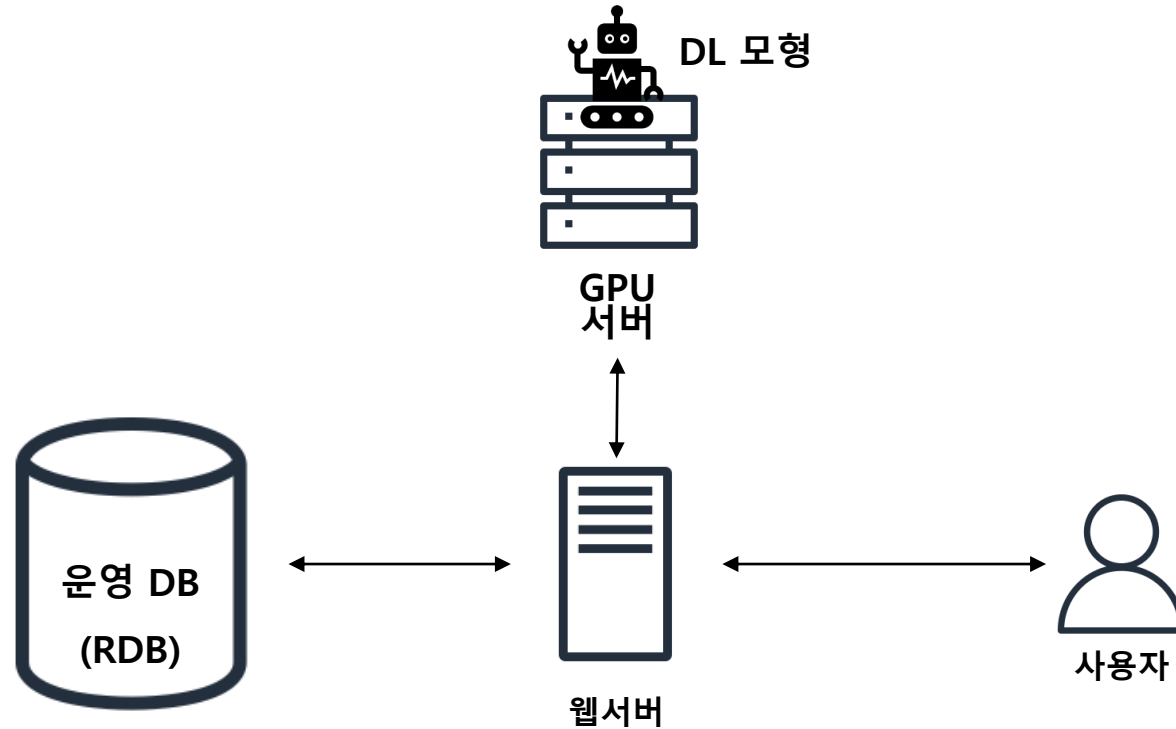
손실과복원



#생성형AI#Tex-to-Image

Decision Tree의 장점1

“가성비가 좋은 알고리즘”

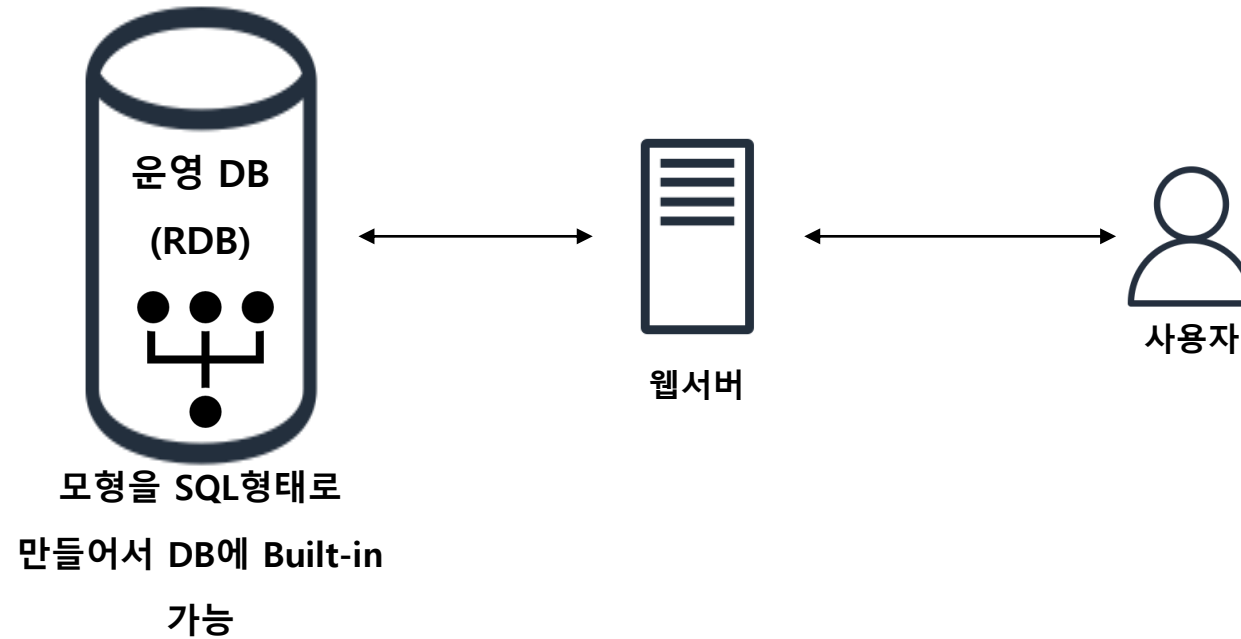


< DL모형 >

*모든 자료에 대한 권한은 메타코드에 있으며, 무단으로 자료를 복제 및 배포 등 유료목적으로 활용하시면 별도의 조치가 들어갈 수 있습니다.

Decision Tree의 장점1

“가성비가 좋은 알고리즘”



< Tree 계열 모형 >

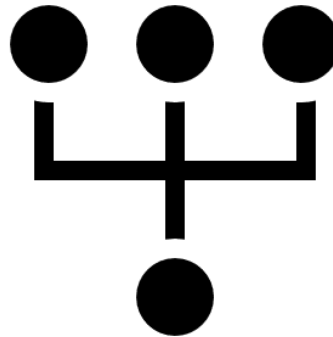
*모든 자료에 대한 권한은 메타코드에 있으며, 무단으로 자료를 복제 및 배포 등 유료목적으로 활용하시면 별도의 조치가 들어갈 수 있습니다.

Decision Tree의 장점2

“Excel을 다루는 업무에 최적화된 모델
작은 테이블 형태의 데이터(tabular data)에 특화”

장점

- 작은 데이터에서 강함
- 변수 타입에 유연성
- 스케일에 불변성
- 결측값 처리
- 해석가능성

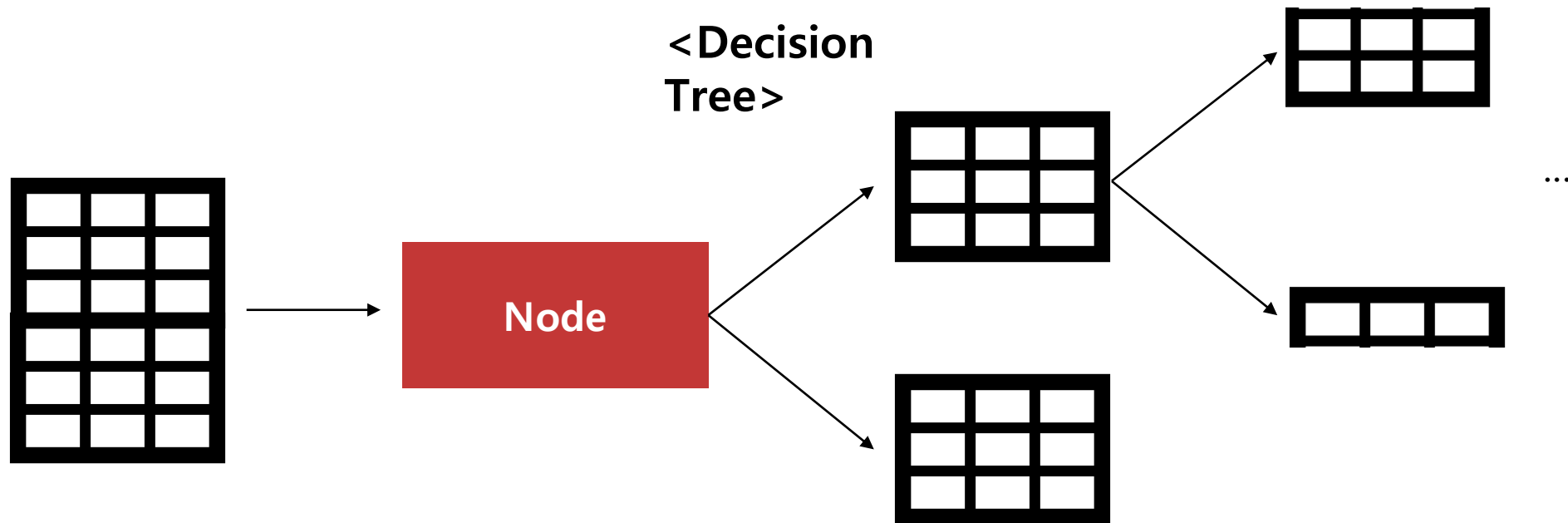


단점

- 과적합
- 이미지, 텍스트에 취약
- 낮은 표현력(expressiveness)

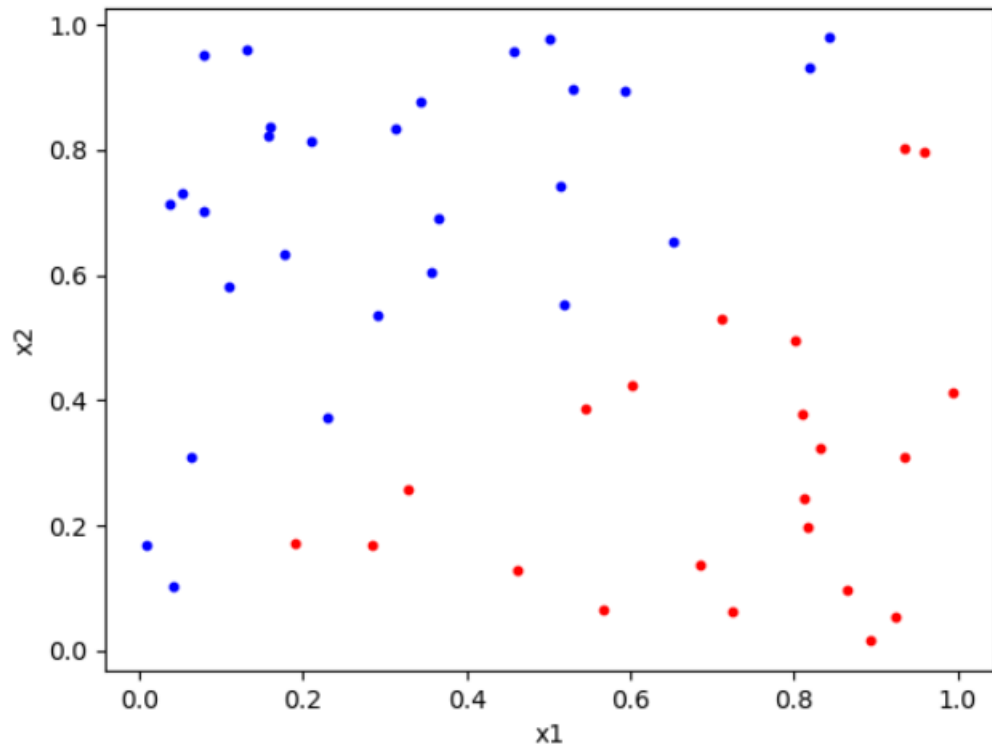
Decision-Tree 개념

- Key Idea: 데이터가 최대한 비슷해지도록 분할하는 것
- 비슷한 정도: 불순도라고 정의
 - 이것을 수학적으로 표현하기 위해 '**Entropy**'라는 물리학 공식을 빌려옴
 - **Entropy**: 비슷한게 많으면 값이 커지고, 비슷한게 적으면 값이 작아지는 특성



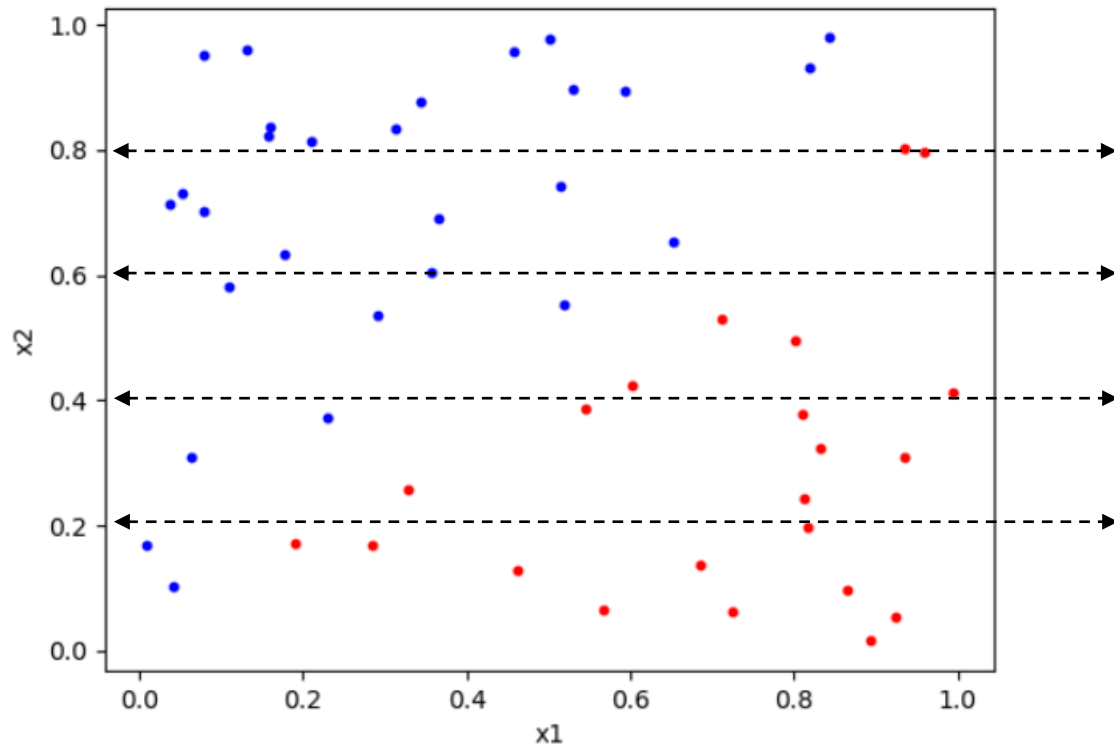
*모든 자료에 대한 권한은 메타코드에 있으며, 무단으로 자료를 복제 및 배포 등 유료목적으로 활용하시면 별도의 조치가 들어갈 수 있습니다.

분할여부 결정



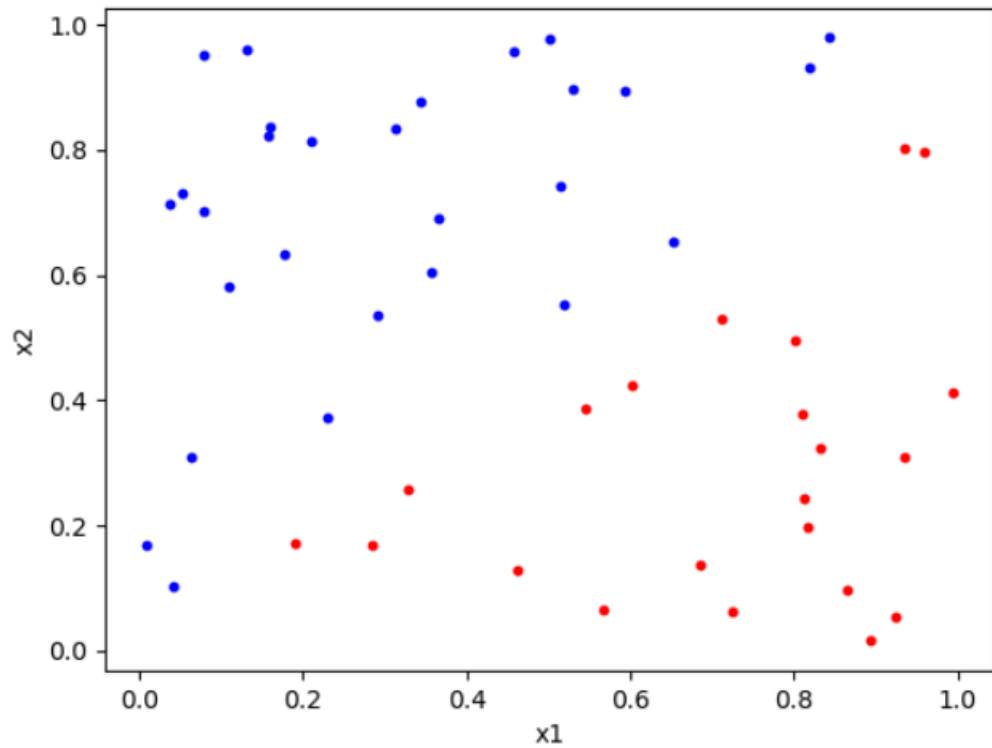
X2를 0.5
기준으로 분할

분할여부 결정

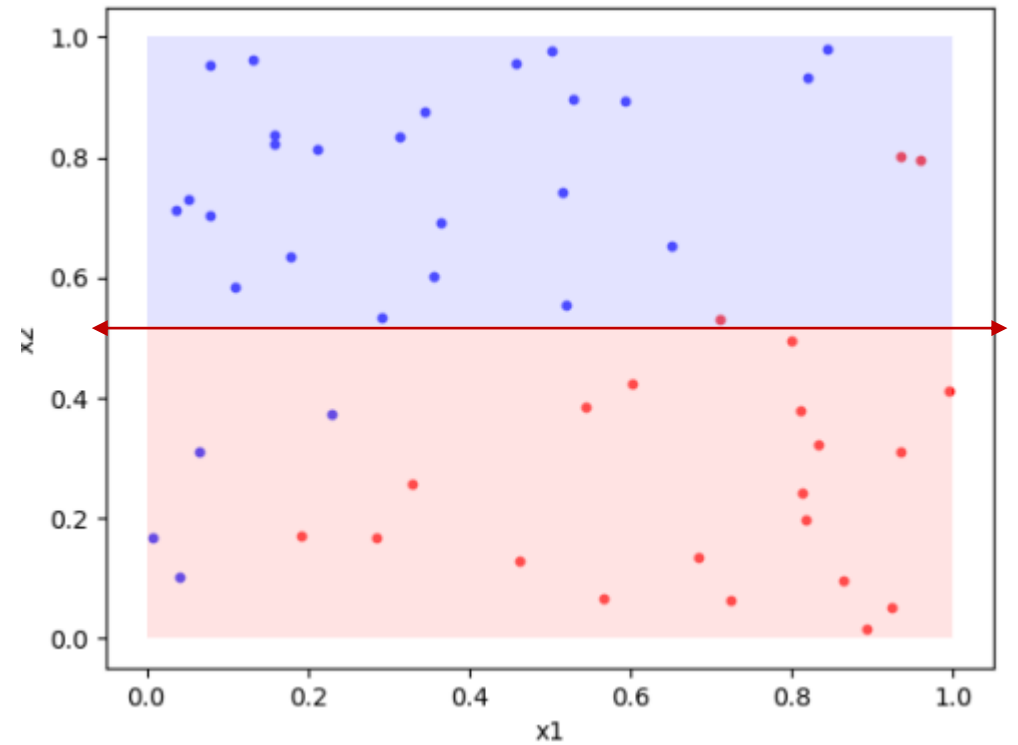


어떤 값을 기준으로 분할하는 것이
가장 불순도가 감소할까?

분할여부 결정

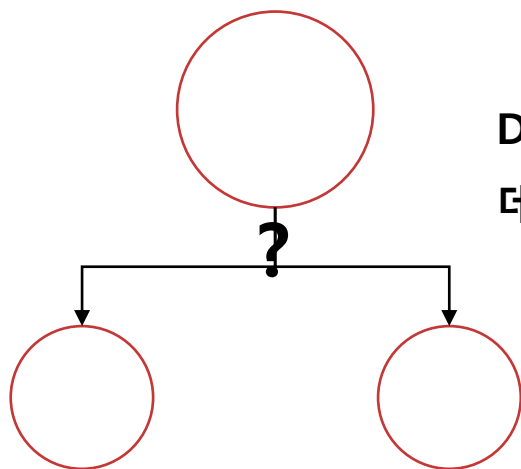


X2를 0.5
기준으로 분할



Entropy (Shannon)

물리학의 공식을 빌려와
불순도를 '수식'으로 표현



Decision-Tree의
데이터 분할여부 판단

$$H = - \sum p(x) \log p(x)$$

$P(x)$ 는 집단에서 'x'일 확률

최소값은: 0

최대값은 : $\log_2 n$

0은 모두 균일한 경우 (같은 값)

집단이 불순하면 값이 커지고
집단이 순수하면 값이 작아지는 특성

꼭 Shannon Entropy를 써야하나요?

집단이 불순하면 값이 커지고
집단이 순수하면 값이 작아지는
특성만 표현 가능하면 무엇이든지 가능

"gini", "entropy", "log_loss"

Gini

- 빠른 계산, 대용량 데이터셋
- 균형 잡힌 클래스 분포

Entropy

- 분포가 불균형한 데이터셋
- 과적합에 강함

log_loss

- 이진분류모형
- 해석의 편리함 (확률로 해석가능)

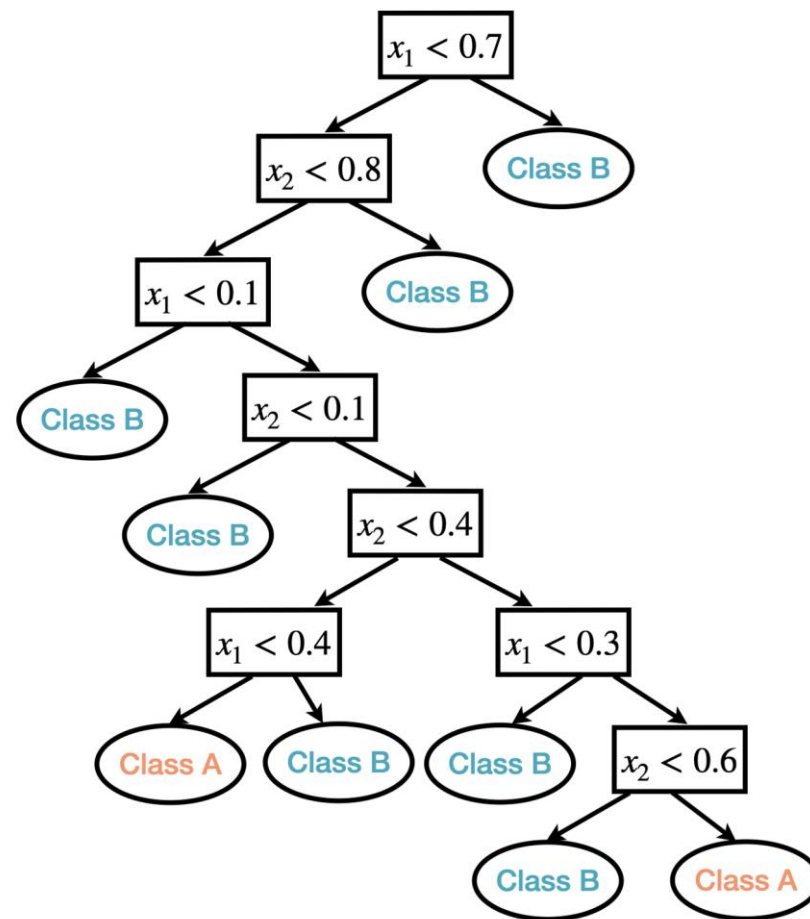
Decision-Tree Process

목적: 데이터를 순수한 부분집합으로 분할하는 것이 목표



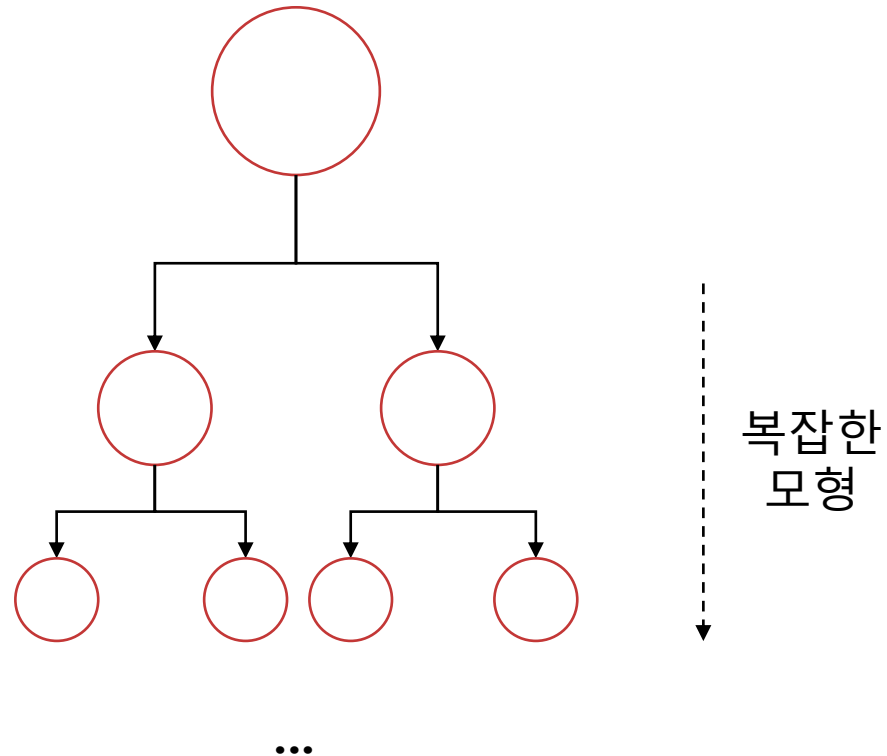
자식 노드에
대해서 반복

정보 획득: 분할 전 엔트로피 - 분할 후 엔트로피
최적 분할: 정보 획득이 가장 큰 분할 선택



언제까지 나무를 키워야 할까?

“데이터에 노이즈(불확실성)이 전혀 없는 이상적인 상황이라면,
Decision Tree를 크게 만드는 것이 성능을 극대화”

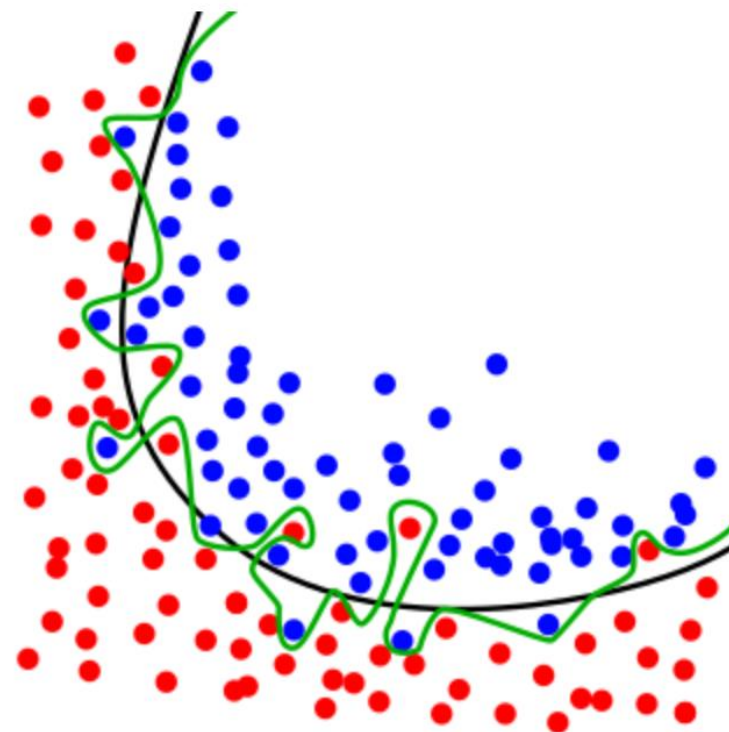


Overfitting

“현실의 데이터는 노이즈와 변동성이 가득함
의미 있는 패턴만 학습해야 함”

Decision Tree가 너무 깊으면 노이즈나
변동성까지
학습하여 과적합(**Overfitting**) 발생

과적합: 학습하면 안되는 것까지 학습한 상태



Hyperparameter

‘하이퍼파라미터는 모델의 학습 과정을 제어하는 사용자 설정 변수’

학습 옵션(**Option**)’이라고 이해
설정을 안하면 Default로

Criterion: 불순도 기준

Splitter: 각 노드에서 분할을 수행 방법

max_depth: 트리의 최대 깊이

min_samples_split: 노드의 최소 샘플 수

min_samples_leaf: 리프의 최소 샘플 수

max_features: 사용할 특성의 최대 개수

min_impurity_decrease: 최소 불순도 감소량

class_weight: 클래스 가중치

random_state: 난수 생성기의 시드

최적의 Option은?



Hyperparameter Tunner

‘최선의 옵션을 찾아주는 방법’

Criterion: 불순도 기준

Splitter: 각 노드에서 분할을 수행 방법

max_depth: 트리의 최대 깊이

min_samples_split: 노드의 최소 샘플 수

min_samples_leaf: 리프의 최소 샘플 수

max_features: 사용할 특성의 최대 개수

min_impurity_decrease: 최소 불순도 감소량

class_weight: 클래스 가중치

random_state: 난수 생성기의 시드

Hyperparameter
tunner

<Best Hyperparameter>

Criterion: 'gini'

Splitter: 'best'

max_depth: 5

min_samples_split: 10

min_samples_leaf: 5

max_features: 'sqrt'

min_impurity_decrease: 0.01

class_weight: None

random_state: 42

XAI (설명가능한 인공지능)

“인공지능 모델의 내부 작동 원리를 사람이 해석 가능한 형태로 제공하여
모델의 동작 방식을 파악할 수 있게 하는 것”

Feature Importance

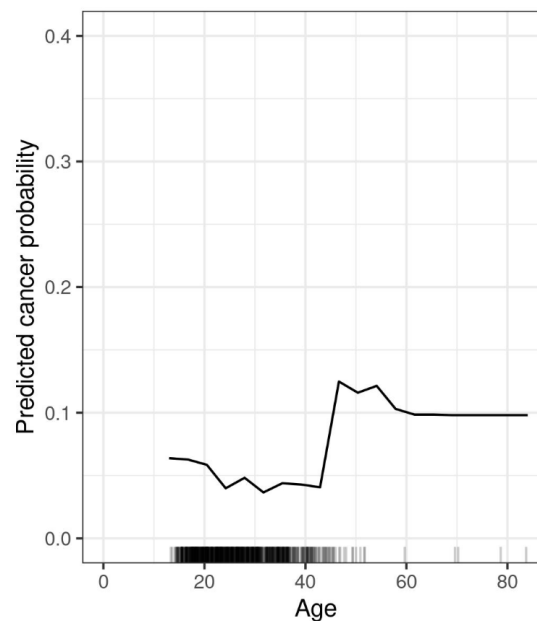
‘모델’에 대한 각 변수의 기여도



오렌지 & 사과 분류

Partial Dependence Plot

변수의 ‘구간별’ 기여도



Shapley value

‘예측’에 대한 각 변수의 기여도

