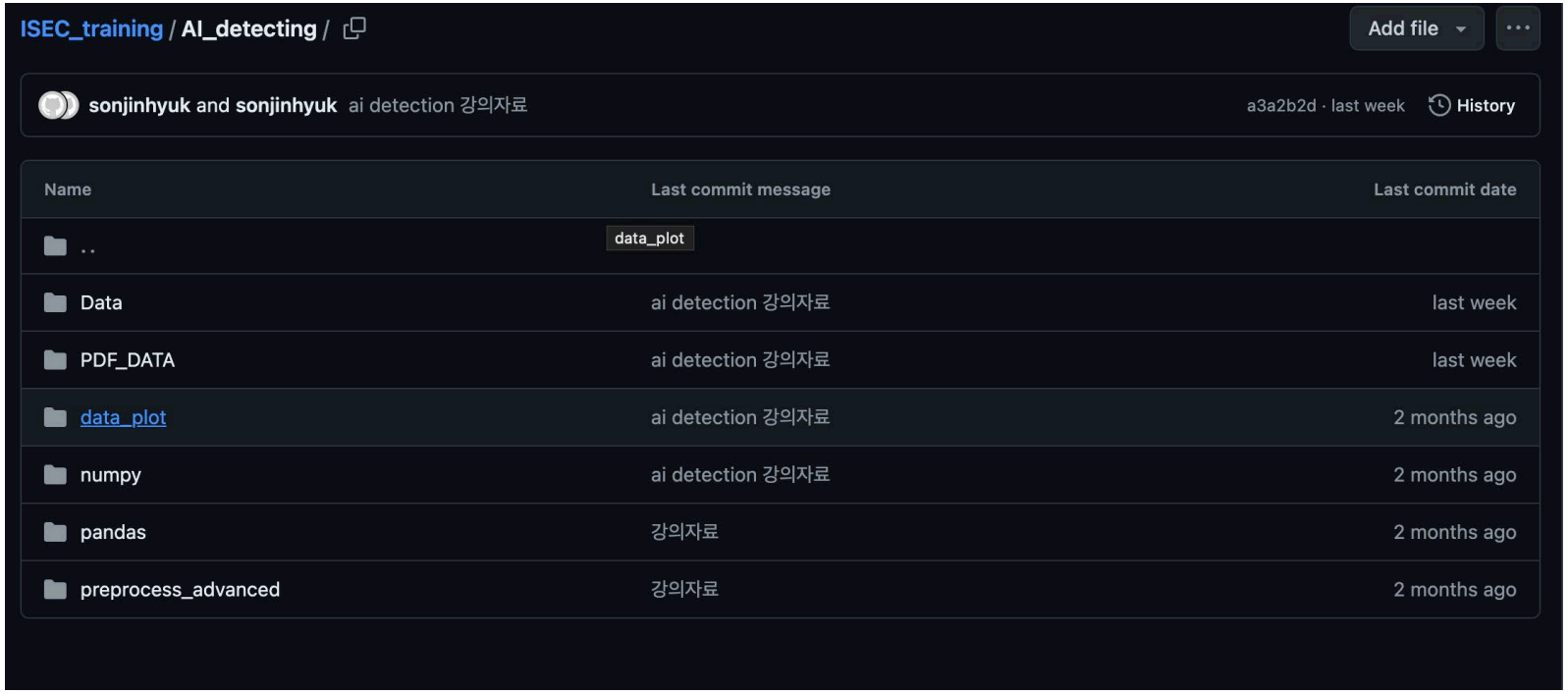
식별된 위협인자 기반 AI 기반 탐지 모델 생성 기술

교육 2일차

KAIST CSRC Cyber Security Research Center

CYBER SECURITY RESERCH CENTER

- 교육관련 자료 : Github (https://github.com/sonjinhyuk/ISEC_training/AI_detecting)
  - Python 3.10



일정

	시간	내용
2일차	10:00 ~ 11:20 (‘80)	전통적인 악성코드 탐지 방법론 통계적 분석 방법론 데이터 처리 - 1
	11:20 ~ 13:00 (‘100)	점심 시간
	13:00 ~ 14:00 (‘60)	데이터 처리 – 2
	14:00 ~ 14:10 (‘10)	Break Time
	14:10 ~ 15:10 (‘60)	악성코드 탐지 AI 모델 - 1
	15:10 ~ 15:20 (‘10)	Break Time
	15:20 ~ 16:30 (‘90)	악성코드 탐지 AI 모델 - 2
	16:30 ~ 16:50 (‘20)	Break Time
	16:50 ~ 18:00 (‘70)	문서형 악성코드 탐지 모델 해석

학습목표

- 식별된 위협 인자의 분석 방법론에 대해 알아본다.
- 식별된 PDF 문서형 악성코드 인자를 활용한 데이터 분석을 진행한다.
- 식별된 PDF 문서형 악성코드를 활용한 악성코드 탐지 모델을 생성한다.
- 생성한 악성코드 탐지 모델을 해석하기 위해 XAI 기법을 적용한다.

목차

- 전통적인 악성코드 탐지 방법론
- 통계적 분석 방법론
- 데이터 처리
- 문서형 악성코트 탐지 AI 모델

전통적인 악성코드 탐지 방법

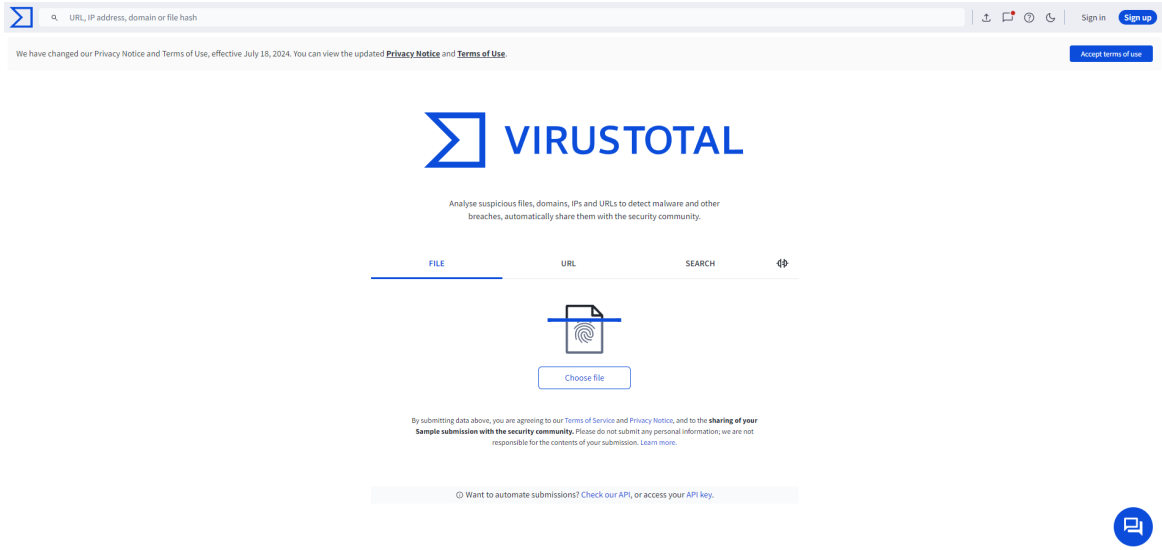
1. 악성코드 분석

- 정적 분석
  - 악성코드를 직접 실행하지 않고 그 자체가 갖고 있는 속성들을 통해 악의적인 여부를 진단하는 방법
  - 파일, 바이너리 코드, 문자열, 파일의 헤더, 리소스 등을 조사하여 악성코드를 식별
- 동적 분석
  - 악성코드를 직접 시스템에 실행시킨 후 변화를 모니터링하여 행동 패턴을 분석
  - 실제 악성코드가 실행되기 때문에 가상 환경이 필요
  - 프로세스->파일->레지스트리->네트워크 순으로 확인
  - 시스템의 변화가 악성코드가 원인인지를 파악하기 위해 반복적인 확인 필요

차이점	정적 분석	동적 분석
실행 여부	실행 하지 않고 악성코드 파일 자체를 분석	악성코드를 실행하여 동작을 관찰
복잡한 악성코드 분석	완전한 정보 제공의 어려움	정적 분석보다 더 많은 정보를 얻을 수 있음

2. 정적분석 도구

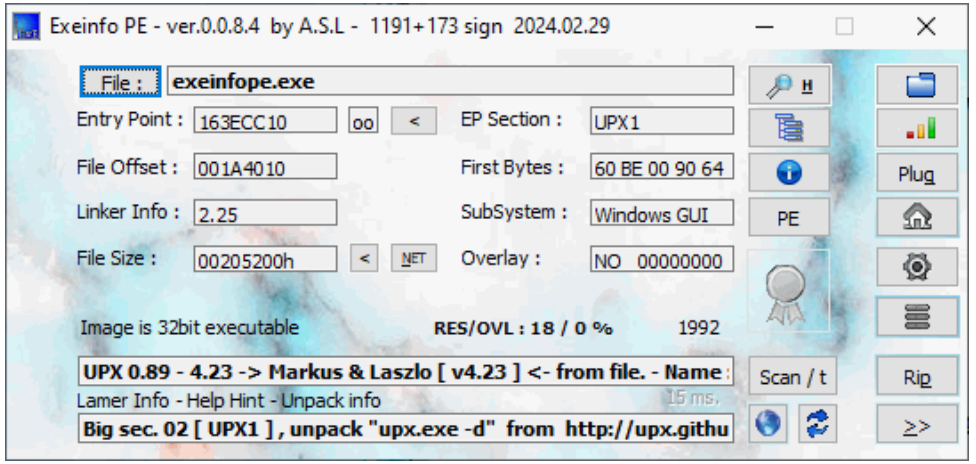
- Virustotal(기초 분석)
  - 백신에 따른 악성코드 확인과 해쉬값 조회를 통한 과거 검사 이력 확인
  - 기본적인 참고용 자료



출처: https://www.virustotal.com/gui/home/upload

2. 정적분석 도구(Cont.)

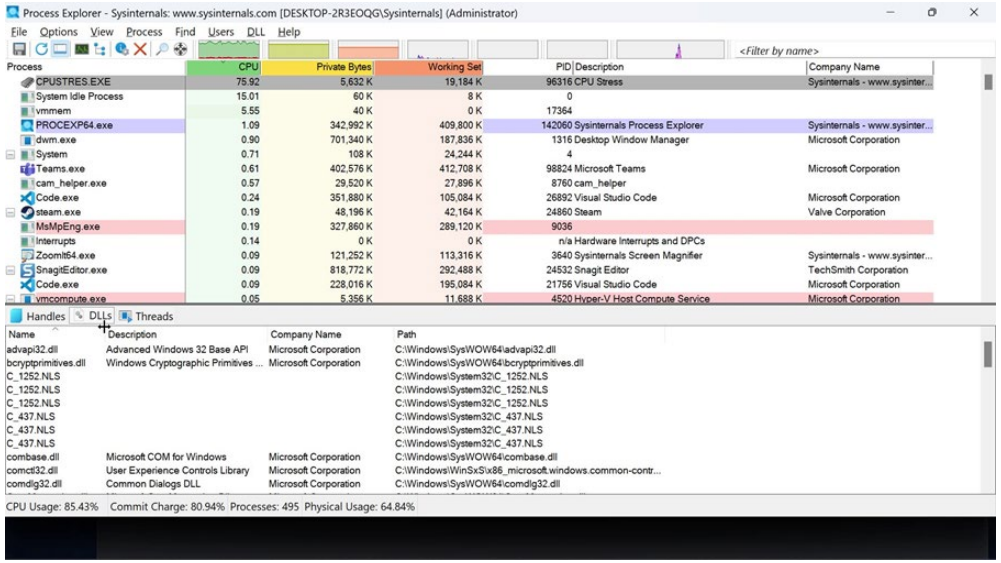
- Exeinfo PE(패킹 여부 확인)
  - 해커들은 악성코드의 크기를 줄여 빠르게 유포하고 분석하기 어렵게 악성코드를 암호화 및 압축하는 패킹 기법을 사용
  - 패킹 여부를 확인하고 패킹이 된 파일은 분석 전 패킹을 해제하여 평문 형태로 만든 후 정적 분석 진행



출처: https://github.com/ExeinfoASL/ASL/blob/master/exeinfo_screen.png

3. 동적분석 도구

- Process Explorer(프로세스 분석)
  - 악성코드 실행 후 나타나는 프로세스 모니터링
  - 프로그램의 행위, 시스템 호출 들을 관찰



출처: https://learn.microsoft.com/ko-kr/sysinternals/downloads/process-explorer

4. 정적, 동적분석 장단점

- 정적분석 및 동적분석에는 장·단점이 존재
- 전통적인 악성코드 탐지 방법의 장점을 활용하면서 단점을 상쇄하기 위해 최근에는 AI를 활용한 악성코드 탐지연구가 활발하게 진행

차이점	장점	단점
정적분석	1. 프로그램의 전체구조 파악에 용의 2. 분석 환경에 제약에서 자유로움 3. 악성 프로그램의 위협으로 안전	1. 난독화된 프로그램을 분석하기 어려움 2. 다양한 동적 요소를 고려하기 어려움 3. 변종 공격에 취약
동적분석	1. 프로그램의 개략적인 동작을 파악 2. 정적 분석보다 높은 정확도	1. 분석 환경 구축하기 어려움 2. 시간 및 자원 소모가 심함

5. AI를 활용한 PDF악성코드 탐지 논문

연번	제목	Feature/특이점
1	A Pattern Recognitioin System for Malicious PDF Files Detection(2012)	malware, benign file에서의 embedded keyword frequency를 vector 화 하여 사용
2	Lux0R: Detection of Malicious PDF-embedded JavaScript code through Discri minat Analysis of API References(2014)	Acrobat PDF API(JS)의 frequency 활용
3	A Structural and Content-based Approach for a Precise and Robust Detection of Malicious PDF Files(2015)	structure: 파일의크기, versions, indirect object 수, streams 수, compr essed object 수, object stream 수, x-ref stream 수, javscript를 포함한 object의 개수
4	Keeping pace with the creation of new malicious PDF files using an active-lea rning based detection framework(2016)	path와 name을 feature로 사용
5	Hidost: a tatic machine-learning-based detector of malicious files(2016)	Path와 name을 vectorization한 후 사용
6	When a Tree Falls: Using Diversity in Ensemble Calssifiers to Identify Evasion i n Malware Detectors(2016)	앙상블 모델처럼 계산하여 사용
7	Fepdf A robust feature extractor for malicious pdf Detection(2016)	javascript code의 extraction 고도화
8	Malware Detection in PDF Files Using Machine Learning(2018)	PDFiD에서 나오는 tag들
9	On Training Robust PDF Malware Classifiers(2018)	path와 name을 feature로 사용

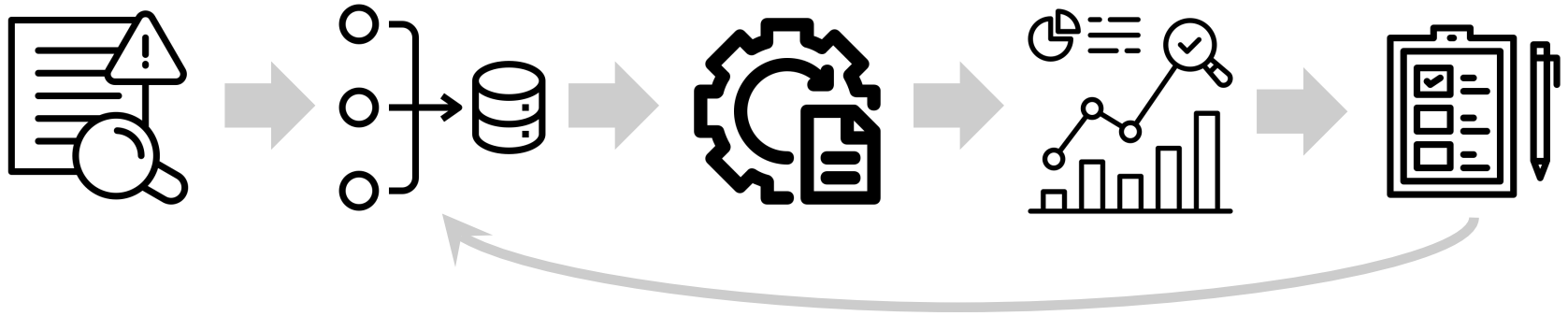
통계적 분석 방법론

1. 데이터 사이언스

- 분석 방법, 도메인 전문성 및 기술의 융합을 통해 데이터에서 패턴을 찾고, 추출하고, 표면화 하는 학문
  - 데이터 마이닝
  - 예측
  - 머신 러닝
  - 예측 분석
  - 통계 및 텍스트 분석
- 통계, 데이터 마이닝, 데이터 사이언스 모두 데이터로부터 의미 있는 정보를 추출하는 학문
  - 통계학: 정형화된 실험 데이터를 분석
  - 데이터 마이닝: 데이터 분석에 초점을 둠

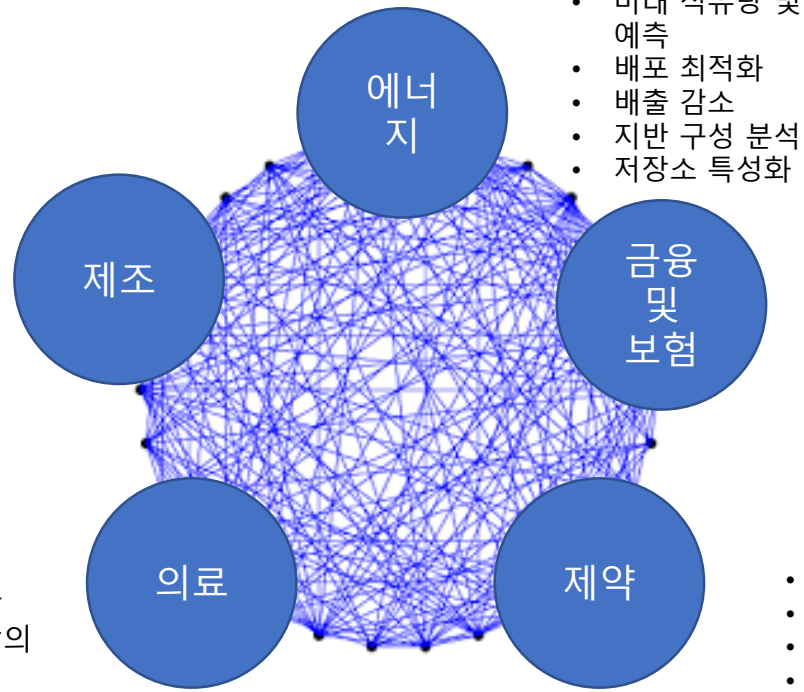
1. 데이터 사이언스(Cont.)

- 데이터 사이언스 프로세스
  - 문제 이해
  - 원시 데이터 수집 및 통합
  - 데이터 탐색, 변환, 정리 및 준비
  - 데이터를 기반으로 모델 생성 및 선택
  - 모델 테스트, 조정 및 배포
  - 모델 모니터링, 테스트, 재학습 및 관리



2. 데이터 사이언스 활용 분야

- 수율 향상
- 스크랩, 재작업 및 반품 감소
- 보증 사기 감지
- 규정 준수
- 장비 고장 예측 및 방지



- 장비 고장 예측
- 미래 석유량 및 가격 예측
- 배포 최적화
- 배출 감소
- 지반 구성 분석
- 저장소 특성화

- 신용 위험 예측
- 사기 감지
- 고객 분석
- 포트폴리오 위험 관리
- 고객 이탈 가능성 결정
- SOX, Basel II와 같은 규정의 준수

- 질병 위험 예측
- 사기성 주장 감지
- 맞춤형 약 복용량 처방
- 이미지 분석에 의한 암의 감지
- 청구 관리
- 환자 안전 향상
- 가장 위급한 환자 결정

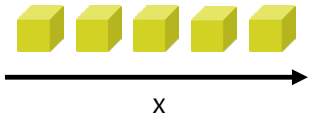
- 최적 배치 결정
- 임상 시험 분석
- 제품 추적
- 안정성 및 유통 기한 분석
- 규정 준수를 위한 보고 및 분석 검증
- 제조 프로세스, 데이터 분석

3. 통계의 용어 및 주요상식

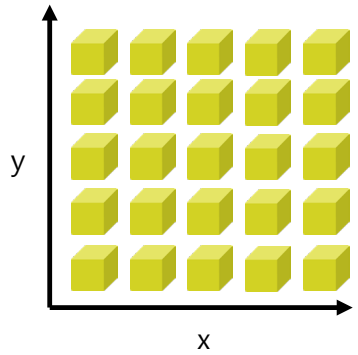
용어	내용
모집단(population)*	관심의 대상이 되는 모든 객체의 특성을 나타내는 관측 값, 측정값의 전체 집합
추출단위(sampling unit)	전체를 구성하는 각 개체들
특성 값(characteristic)	각 추출단위의 특성을 나타내는 값
표본(sample) *	통계적 분석을 위해 실제로 뽑힌 추출 단위의 집합 모집단은 규모가 크기 때문에 조사에 시간적, 공간적 제약이 따르기 때문에 표본을 통한 데이터 수집 및 분석이 이루어짐
관찰 값(observed values)	표본의 특성값, 관찰된 측정값, Sample의 값
모수(parameter) *	모집단의 특성을 나타내는 양적인 측도(고유의 상수)
통계량(statistic)	표본에 대한 특성을 나타내는 양적인 측도 표본을 통해 모집단의 특성을 추론하는 값

3. 통계의 용어 및 주요상식(Cont.)

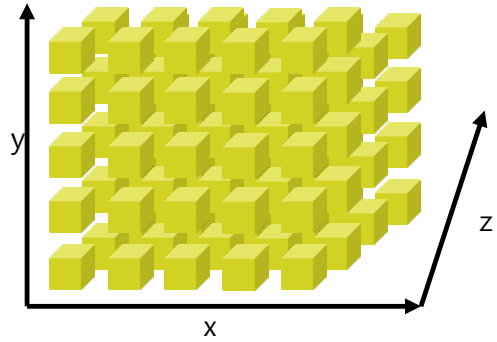
- 차원의 저주(Curse of Dimension)
- 변수가 증가
  - 표현하기 위한 데이터 양이 기하급수적으로 증가하며
  - 데이터의 밀도는 희박(sparse)해짐
  - 전체 공간에 있는 변수 양 동일, 찾고자 하는 공간에 있는 데이터의 양이 적어짐
- 일정 차원을 넘으면 분류기의 성능이 떨어짐 → overfitting
- Overfitting: 많은 연산이 쌓이면서 오차가 증가하고 예측력이 낮아짐



One cube contains N/5 samples.



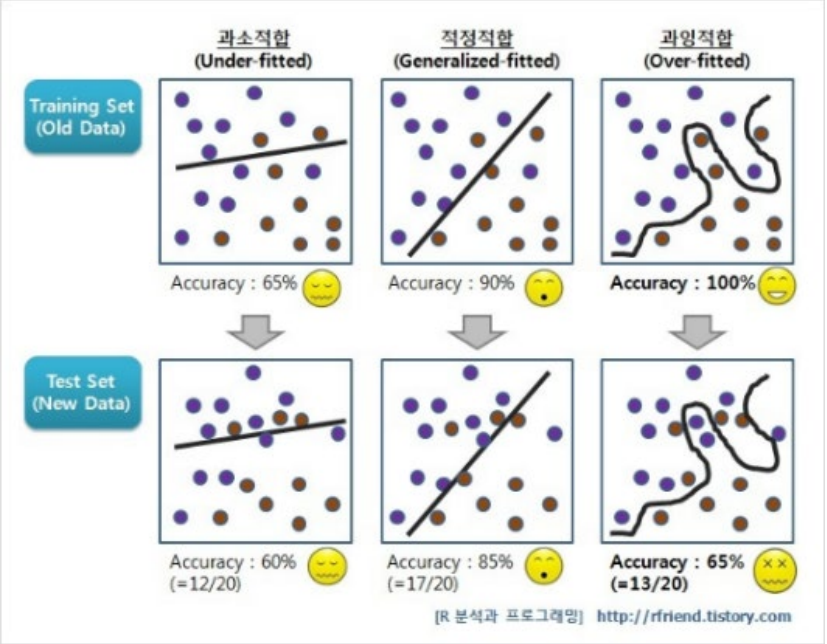
One cube contains N/25 samples.



One cube contains N/125 samples.

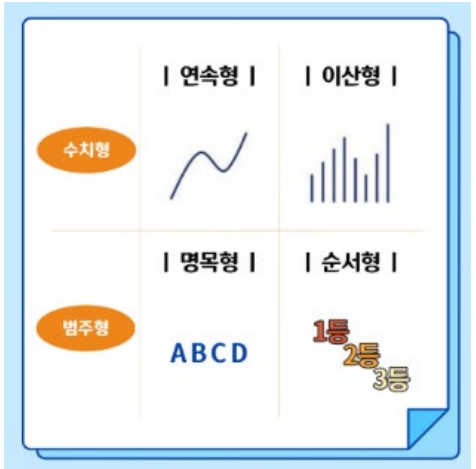
3. 통계의 용어 및 주요상식(Cont.)

- Overfitting
  - Under Fitting: 적정 수준의 학습을 하지 못하여 실제 성능이 떨어지는 경우
  - Right Fitting: 적정 수준의 학습으로 실제 적정한 일반화 수준을 나타냄



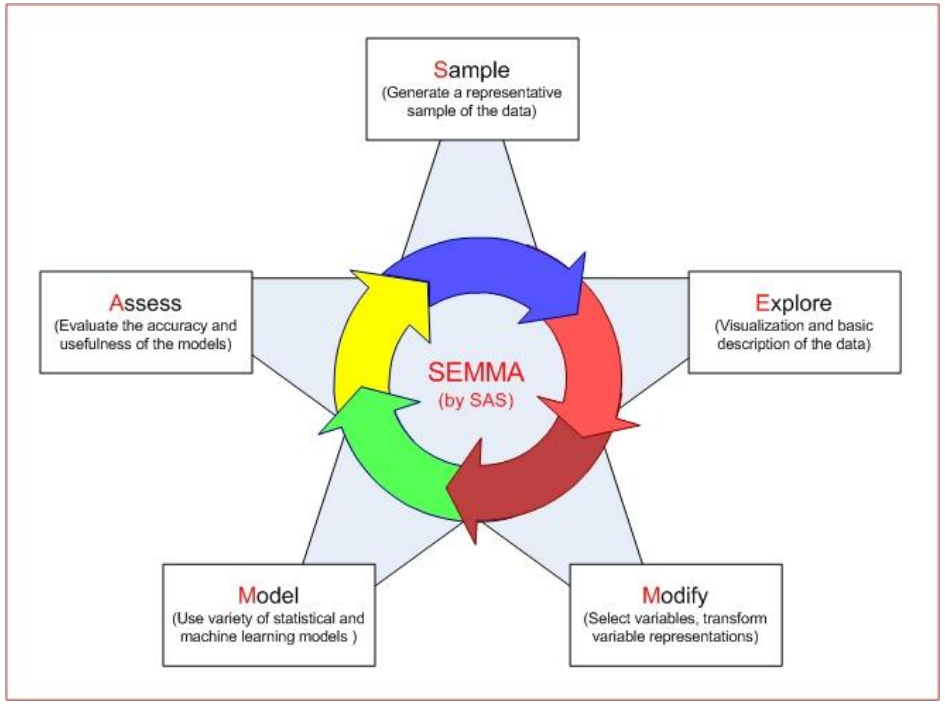
3. 통계의 용어 및 주요상식(Cont.)

- 자료의 분류
- 수치형 변수(Numberical Variable)
  - 연속형 변수(Continuous numbers): 키, 몸무게, 온도, 거리(정확한 값X, 연속된 수)
  - 이상현 변수(Discrete Variable): 수강생 수, 카페의 개수(정확한 숫자 값)
- 범주형 변수(Categorical Variable)
  - 명목형 변수(Nominal Variable): 혈액형, 성별, 통신사
  - 순위형 변수(Ordinal Variable): 학년, 등급, 설문지 척도



3. 통계의 용어 및 주요상식(Cont.)

- 통계 분석 프로세스
  - Sample: 샘플 선택
  - Explore: 데이터 변수
  - Modify: 유의미한 변수 찾기, 변수 변환 파생 변수 생성
  - Model: 통계 모델
  - Assess: 모델링 작동 평가

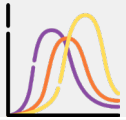


3. 통계의 용어 및 주요상식(Cont.)

통계 분석 도구

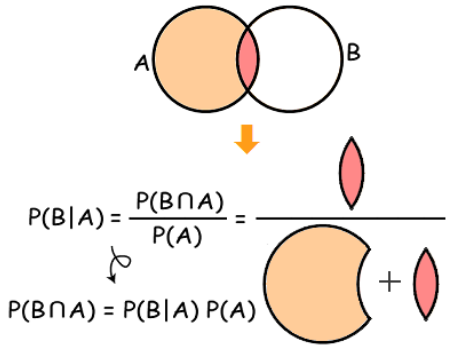
분석 도구	특징
R	자바로 만들어진 프로그램으로 스칼라 대신 벡터로만 처리 가능 빅데이터 처리에 단점을 보임
파이썬	**직관적이며 배우기 쉬움** **다양한 라이브러리 구현되어 있음**
엑셀	편리한 접근성 및 수식 계산
SSAS	내장된 데이터 마이닝 기능을 사용하여 데이터 패턴 검색에 용의 테이블 형식과 다차원 데이터간 병학이 불가능

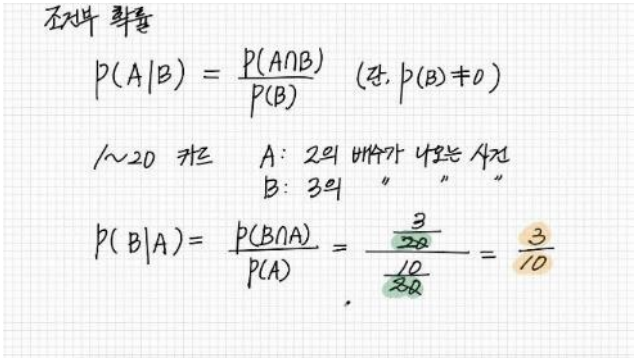
4. 통계 기법 유형

기법	종류
확률 	베이즈 정리 Monte Carlo Method
분포 	정규 분포
추정(예측) 	추정 이론 점추정 구간 추정
추론 	가설 검정 은닉 마르코프 모델
분석 	상관분석 회귀분석 시계열 분석 주성분 분석

4. 통계 기법 유형(Cont.)

- 베이즈 정리
  - 이전의 경험과 현재의 증거를 토대로 어떤 사건의 확률을 추론하는 조건부 확률의 이론
  - 조건부 확률: 사건 B가 일어났다는 조건 하에 사건 A가 일어날 확률





$P(A|B) = P(B|A)P(A)/P(B)$

- $P(A|B)$: 사후확률(posteriror)
- $P(A)$: 사전확률(prior)
- $P(B|A)$: 가능도(likelihood)
- $P(B)$: 정규화 상수(nomalizing constant) 또는 증거(evidence)

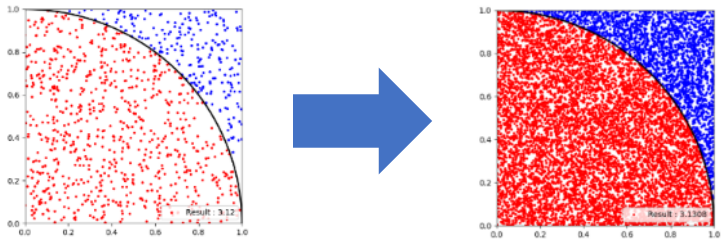
4. 통계 기법 유형(Cont.)

```
###사후 확률 구하기
#사전확률이 2개
#문제
#모 선출에서 2명의 후보자가 있다. A1, A2가 당선될 확률은 각각 0.7, 0.3이다.
#A1, a2가 당선되면 각각 0.2와 0.9의 예측 확률로 회비를 인상할 것으로 판단된다.
#회원의 회비가 인상될 확률은 얼마인가?

import pandas as pd
#P(A1)+P(A2)=1
a1=0.7      #P(A1)
a2=0.3      #P(A2)
b_a1=0.2    #P(B|A1)
b_a2 = 0.9  #P(B|A2)

bays = pd.DataFrame({"사건":['A1', 'A2'],
                     "사전확률_P(Ai)":[a1, a2],
                     "조건부확률_P(B|Ai)": [b_a1, b_a2],
                     "결합확률_P(Ai*B)":[a1*b_a1, a2*b_a2],
                     "사후확률_P(Ai|B)":[a1*b_a1/(a1*b_a1 + a2*b_a2),
                                         a2*b_a2/(a1*b_a1 + a2*b_a2)],
                     })
print(bays)
```

	사건	사전확률_P(Ai)	조건부확률_P(B|Ai)	결합확률_P(Ai*B)	사후확률_P(Ai|B)
0	A1	0.7	0.2	0.14	0.341463
1	A2	0.3	0.9	0.27	0.658537

4. 통계 기법 유형(Cont.)

- Monte Carlo Method
  - 난수(Random Number)를 이용하여 물리적, 수학적 시스템의 행동을 시뮬레이션 하기 위한 알고리즘
  - 범위 설정
    - 가능한 입력상수의 범위를 정의
  - 입력 상수 생성
    - 입력상수의 범위 안에서 확률분포를 통해 입력상수 발생
  - 연산
    - 생성된 입력상수에 대한 계산 수행
  - 종합평가
    - 계산 결과를 종합



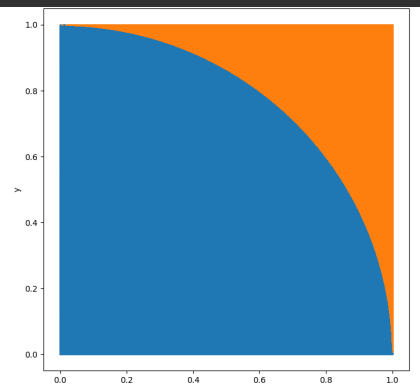
4. 통계 기법 유형(Cont.)

- Monte Carlo Method(예제)
  - 원주율(파이)값 계산하기

```python
from tqdm import tqdm
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# 변수 설정
n = 5000000
count = 0
x_list_under = []
y_list_under = []
x_list_upper = []
y_list_upper = []

for i in tqdm(range(n)):
    # 시드 부여 (난수)
    np.random.seed(i)
    # 0,1 사이의 난수 값 생성
    x = np.random.random()
    y = np.random.random()
    # 원 내부에 있는 점을 만족하는 경우, 그렇지 않은 경우 분리
    if (x ** 2 + y ** 2) <= 1:
        count += 1
        x_list_under.append(x)
        y_list_under.append(y)
    else:
        x_list_upper.append(x)
        y_list_upper.append(y)
```

```python
# 원 내부일 확률은 원 넓이의 1/4 이므로 곱하기 4로 π 계산
print('\u03C0 값', 4 * count / n)

under = pd.DataFrame(zip(x_list_under, y_list_under), columns=['x_under', 'y_under'])
upper = pd.DataFrame(zip(x_list_upper, y_list_upper), columns=['x_upper', 'y_upper'])

plt.figure(figsize=(8, 8))
plt.scatter(under['x_under'], under['y_under'], s=1)
plt.scatter(upper['x_upper'], upper['y_upper'], s=1)
plt.xlabel('x')
plt.ylabel('y')
plt.show()
```

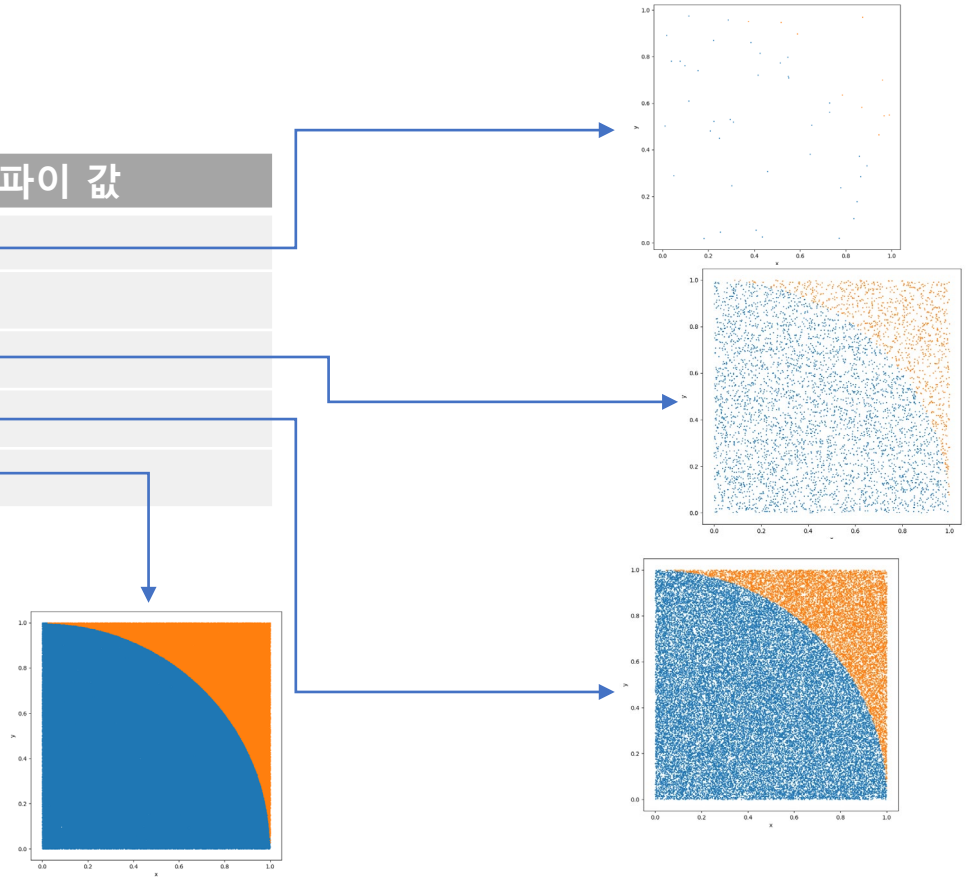
```
100%|██████████| 5000000/5000000 [00:28<00:00, 175130.55it/s]
π 값 3.1423632
```



4. 통계 기법 유형(Cont.)

- Monte Carlo Method(예제)
  - 원주율(파이)값 계산하기

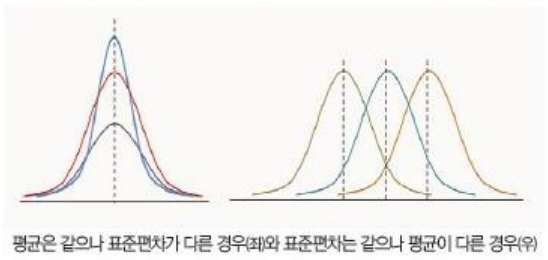
시행 횟수	파이 값
50	3.2
500	3.22
5,000	3.16
50,000	3.1352
500,000	3.1408



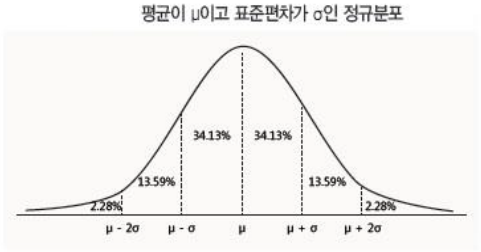
4. 통계 기법 유형(Cont.)

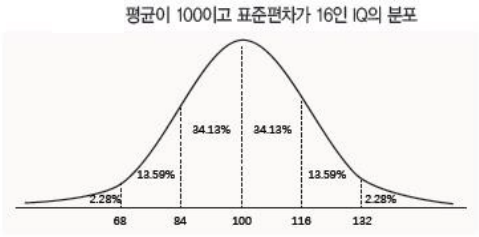
정규분포

- 평균을 중심으로 좌우대칭을 이루는 종 모양의 확률 분포(가우스 분포)
- 데이터의 분포를 확인할 때 사용
- 범주형 데이터의 경우 누적 밀도 함수와 같은 방식으로 정규분포를 확인



평균은 같으나 표준편차가 다른 경우(좌)와 표준편차는 같으나 평균이 다른 경우(우)

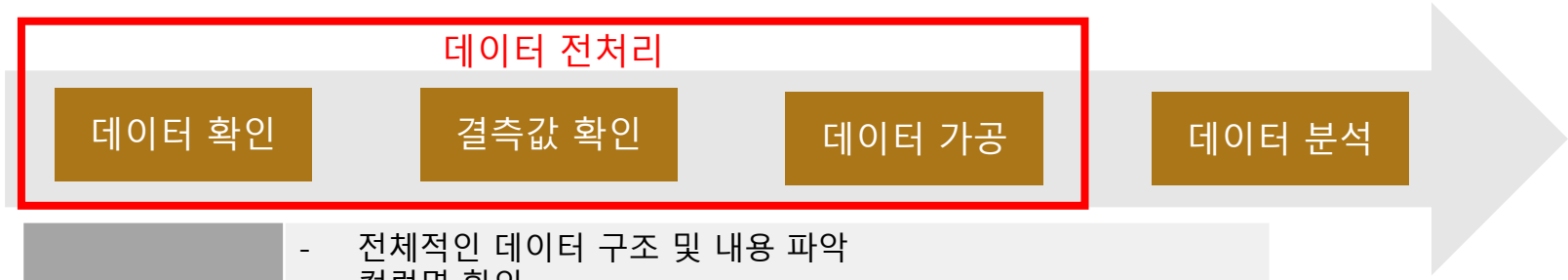




데이터 처리

1. 데이터 분석

- 데이터 분석 과정
  - 시각화 분석을 위해 데이터셋의 특징을 파악
  - 결측값 및 데이터 가공 과정의 전처리



데이터 확인	- 전체적인 데이터 구조 및 내용 파악 - 컬럼명 확인 - 행 / 열 개수 파악 - 데이터 유형 파악
결측값 확인	- 결측값 확인 및 수정 - 결측값 삭제 또는 대치
데이터 가공	- 데이터 분석을 위한 데이터 가공
데이터 분석	- 분석 목적에 맞는 차트로 시각화 - 데이터의 의미 있는 결과 해석

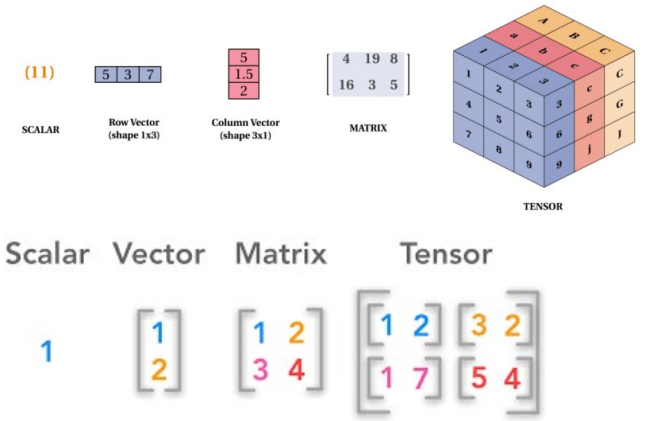
2. 데이터 처리 라이브러리

- Numpy(Numerical Python) 패키기
  - C언어로 구현된 파이썬 라이브러리
  - 고성능의 수치계산을 위해 제작
  - 벡터 및 행렬 연산에 있어 매우 편리한 기능 제공
  - Array 단위로 연산 수행

- 난수 생성 가능
- 선형대수 가능
- 각 열의 데이터 타입이 동일

2. 데이터 처리 라이브러리(Cont.)

- 데이터의 종류
  - 스칼라(Scalar)
    - 하나의 숫자로 이루어진 데이터
  - 벡터(vector)
    - 여러 개의 숫자들을 특정한 순서대로 모은 데이터 모음(데이터 레코드)
    - 1D Array(1차원 배열)
  - 행렬(Matrix)
    - 벡터의 집합
    - 2D Array(2차원 배열)
  - 텐서(Tensor)
    - 같은 크기의 행렬들의 집합
    - ND Array(다차원 배열)



2. 데이터 처리 라이브러리(Cont.)

- Numpy random 모듈

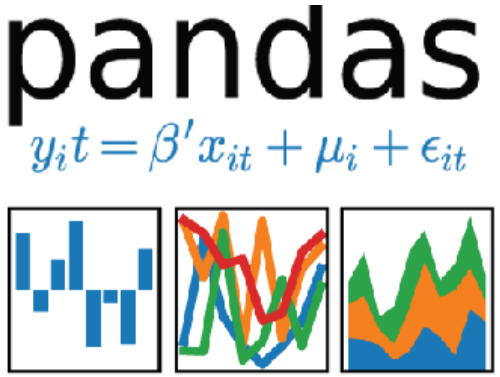
함수	내용
np.random.seed	Seed를 통한 난수 생성
np.random.rand	균일 분포의 정수 난수 1개 생성
np.random.randint	0~1 사이의 균일 분포에서 난수 생성
np.random.randn	가우시안 표준 정규분포에서 난수 생성
np.random.shuffle	기존 데이터 순서 바꾸기
np.random.choice	기존 데이터에서 샘플링
np.random.unique	데이터에서 중복된 값 제거

2. 데이터 처리 라이브러리(Cont.)

- 실습:AI_detecting/numpy/numpy.ipynb
  - 연습문제: AI_detecting/numpy/practice.ipynb
  - Numpy random 모듈
    - 문제 1) Array
      - 3 * 2 행렬 구조로 Array 형태 데이터 A, B를 생성하고, + 연산한 Array를 구하라
      - A, B데이터는 아래와 같다.
        - A: [2,3,4,5,6,7]
        - B: [1,1,3,3,5,5]
    - 문제 2) random
      - 1~100 사이의 정수 중 랜덤하게 한 개의 숫자 추출
    - 문제 3) random selection
      - 제공한 PDF Feature 데이터에서 랜덤한 feature valiue 추출

2. 데이터 처리 라이브러리(Cont.)

- PANDAS 패키지
  - 통계분석 툴인 R을 모티브로 만든 파이썬 라이브러리
  - 데이터 분석을 위한 수집, 전처리 등의 과정은 대부분 데이터프레임 형태로 사용



- 테이블을 수정하고 조작하는 다양한 기능 제공
- SQL처럼 두개의 데이터구조를 Join할 수 있음
- 각 열의 데이터 타입이 달라도 됨
- SQL, 엑셀파일을 읽을 수 있음
- 대부분의 데이터는 시계열(series)이나 표(table)의 형태로 나타낼 수 있는데, 이를 위해 시리즈(Series)와 데이터 프래임(DataFrame) 구조를 제공

2. 데이터 처리 라이브러리(Cont.)

- PANDAS 자료구조
  - 데이터를 다루기 위한 시리즈(Series) 클래스와 데이터프레임(DataFrame)구조를 제공

Series

1차원 데이터 배열

Kim	1
LEE	-5
CHOI	7
PARK	4
AHN	8
index	**values**

Data Frame

2차원 데이터 배열, 행과 열이 있음

columns

	val0	val1	val2	val3	val4	val5	val6
x1	a	10.0	110.0	1.0	1.0	0	1.0
x2	b	9.0	90.0	1.0	2.0	0.0	2.0
x3	c	10.0	80.0	1.0	3.0	0.0	3.0
x4	a	20.0	150.0	1.0	4.0	0.0	4.0
x5	c	7.0	90.0	1.0	5.0	0.0	5.0
x6	d	8.0	79.0	1.0	6.0	0.0	6.0
x7	e	10.0	96.0	1.0	7.0	0.0	7.0
x8	a	6.0	120.0	1.0	8.0	00.	8.0
index	**values**						

2. 데이터 처리 라이브러리(Cont.)

- 실습:AI_detecting/pandas/pandas.ipynb
- 연습문제: AI_detecting/ pandas /practice.ipynb
  - 문제 1) DataFrame
    - 아래 표를 참고하여 딕셔너리 방식으로 데이터프레임을 생성하라.

	/JS	/AA	Xref	size
0	0	3	12	35
1	0	4	55	44
2	1	0	12	55
3	1	0	4	23
4	5	1	4	15
5	3	15	20	23

  - 문제 2) DataFrame
    - 문제 1의 결과물에서 index 번호를 a~f로 변경하라

2. 데이터 처리 라이브러리(Cont.)

- PANDAS 패키지
  - 컬럼 선택의 기본 코드
    - 데이터프레임의 특정 컬럼을 선택할 때 두가지 방법이 존재
    - 단, 컬럼명이 '.'으로 끝나거나 띄어쓰기가 있는 경우에는 두번째 방식이 불가
    - 첫 번째 방법: 데이터명['컬럼명']
    - 두 번째 방법: 데이터명.컬럼명

3. 데이터 처리

결측치 처리

- 결측치란?
  - Missing feature, NA(Not Available), 결측치(결측값)라고 하며 값이 표기되지 않은 값을 말함
  - 데이터는 다양한 원인으로 누락 데이터를 포함
  - 데이터에서 None, NaN, 빈칸으로 표기되는 것은 누락 데이터
- 결측치 종류

MCAR, Missing completely at random (완전무작위 결측)	- X2의 결측치가 X1, X2, X3의 값과 아무런 상관관계가 없는 경우 - 대부분의 결측치 처리는 이러한 유형을 대상으로 하고 있음 - Ex) 전산상 오류, 입력 실수
MAR, Missing at random (무작위 결측)	- X1이 True일 때, X2는 결측치를 갖고, X1이 False일 때 X2가 값을 가지는 경우 - 다른 특성 값에 따라 결측치의 발생 확률이 계산되며, 이 때 X1과 X2의 상관관계를 알 수 없는 경우
MNAR Missing not at random (비무작위 결측)	- 위의 두가지 상황이 아닌 경우에 모두 해당 - 이 경우 결측치 특성 X2의 값이 다른 특성의 값과 상관관계가 있을 경우

3. 데이터 처리(Cont.)

- 결측치 예시
  - 성별 (X)를 사용해 체중(Y)를 예측하는 모델을 구축하기 위해, 설문조사를 통해 X와 Y를 DataFrame으로 구성
  - 이때, y열에 결측치가 있다고 가정

MCAR, Missing completely at random (완전무작위 결측)	- 별다른 이유 없이 체중을 응답하지 않은 경우 - Y가 누락된 이유는 성별(X) 및 본인의 체중(Y)값과 관련이 없음
MAR, Missing at random (무작위 결측)	- 성별(X)이 여성인 경우 체중에 잘 응답하지 않는 경향이 있다고 가정, - 즉 체중(Y)이 누락된 것은 성별(X)에 영향을 받음
MNAR Missing not at random (비무작위 결측)	- 체중이 무거운 사람들은 자신의 체중에 잘 응답하지 않는 경향이라 가정 - 즉 체중(Y)이 누락된 것은 체중(Y)자체에 영향을 받음

3. 데이터 처리(Cont.)

- 결측치 처리 방법
  - 대치(Imputation)
  - 결측치를 특정 값으로 대치
    - 최빈값(mode)
    - 중앙값(median)
    - 평균(mean)

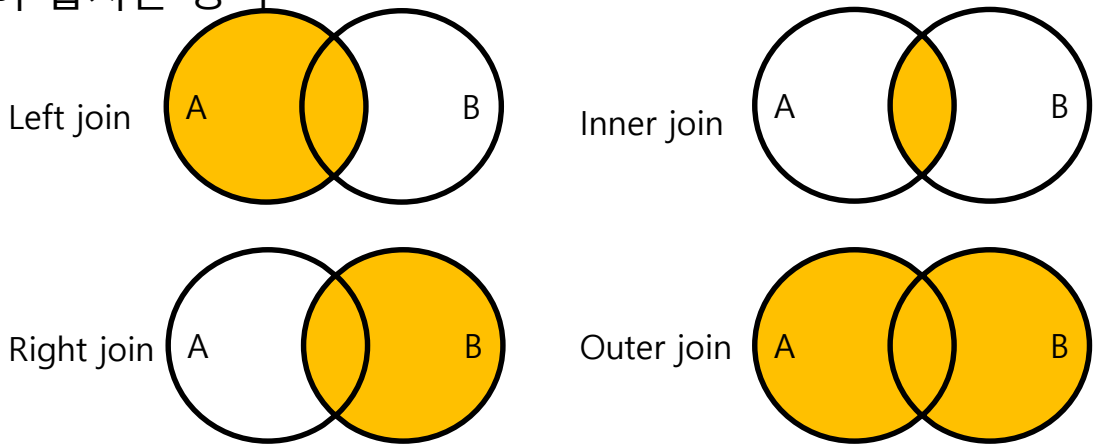
최빈값 (mode)	- 범주형에서 결측값이 발생 시, 범주별 빈도가 가장 높은 값으로 대치
중앙값 (median)	- 숫자형(연속형)에서 결측값을 제외한 중앙값으로 대치
평균 (mean)	- 숫자형(연속형)에서 결측값을 제외한 평균으로 대치

3. 데이터 처리(Cont.)

- 결측치 처리 방법
  - 예측 모델(Prediction model)
  - 예측 기법을 사용한 결측치 추정은 결측치들의 특성이 패턴을 가진다고 가정하고 진행
  - 결측값이 없는 컬럼들로 구성된 dataset으로 결측값이 있는 컬럼 예측
  - 회귀 분석 기술을 활용하거나 SVM과 같은 기계 학습 방법 또는 이러한 결측치를 채우는 데이터 마이닝 방법 등이 있음

3. 데이터 처리(Cont.)

- 데이터 프레임 결합(concat)
  - 두 개의 데이터를 이어 붙임
  - axis(축)에 따라 feature(columns)를 추가할지, 새로운 데이터(row)를 추가할지 결정
- 데이터프레임 Key 지정 합치기(merge)
  - 특정 조건을 활용하여 합치는 방식



3. 데이터 처리(Cont.)

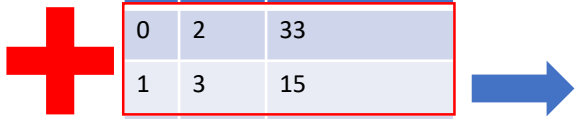
- Left join
  - 왼쪽 데이터프레임을 기준으로 조인
  - 오른쪽 데이터프레임에 없는 값은 NaN으로 표기
- Right join
  - 오른쪽 데이터프레임을 기준으로 조인
  - 왼쪽 데이터 프레임에 없는 값은 NaN으로 표기됨

Data A

	ID	Name
0	1	Kim
1	2	Choi
2	3	Lee

Data B

	ID	age
0	2	33
1	3	15
2	4	20



- Inner join
  - 교집합을 의미
  - 양쪽에 공통으로 있는 값만 조인 함
- Outer join
  - 합칩합을 의미
  - 왼쪽, 오른쪽 데이터프레임에 없는 값들은 NaN

Left Join

	ID	Name	age
0	1	Kim	NaN
1	2	Choi	33
2	3	Lee	15

Right Join

	ID	Name	age
0	2	Choi	33
1	3	Lee	15
2	4	NaN	20

inner Join

	ID	Name	age
1	2	Choi	33
2	3	Lee	15

outter Join

	ID	Name	age
0	1	Kim	NaN
1	2	Choi	33
2	3	Lee	15
3	4	NaN	20

3. 데이터 처리(Cont.)

- 실습: AI_detecting/pandas/pandas.ipynb
  - 실습 1) 주어진 PDF파일을 PDF parser 모듈을 활용하여 데이터 프레임 생성
  - 실습 2) 문제 1의 각 파일의 데이터프레임을 합치기
  - 실습 3) "class"가 0이고, '/AA'가 5 미만인 데이터프레임 저장
  - 실습 4) "class"가 1이고, '/JS'가 3 이상인 데이터프레임 저장
  - 실습 5) 실습 3, 실습 4의 데이터프레임 결합
  - 실습 6) '/JS' 컬럼의 평균값을 계산
  - 실습 7) 합친 데이터 프레임의 기본 통계 정보를 출력
  - 실습 8) 문제 2의 데이터프레임 저장

4. 데이터 시각화

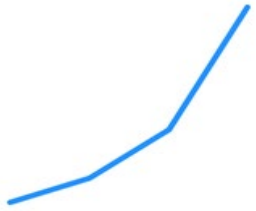
- 데이터 시각화의 목적
  - 커뮤니케이션의 목적
    - 데이터 분석 결과를 한눈에 알기 쉽게 표, 차트, 그래프 등의 이미지 형태로 정리하여 분석 결과의 설득력을 높이기 위한 도구로 사용
    - 데이터를 시각적으로 요약하면 자료에 내포된 정보를 보다 쉽고 빠르게 파악할 수 있음
  - 데이터 간의 숨겨진 관계와 패턴을 탐색
    - 불규칙하고 유의성을 찾기 어려운 데이터 속에서 일정한 패턴을 찾아 숨겨진 의미를 찾기 위하 사용
    - 시각적 데이터 탐색을 통해 수직으로 찾지 못하는 내용을 발견할 수 있음
  - 효율적인 의사결정
    - 데이터를 분석하여 효율적인 의사결정에 도움을 줄 수 있음

4. 데이터 시각화(Cont.)

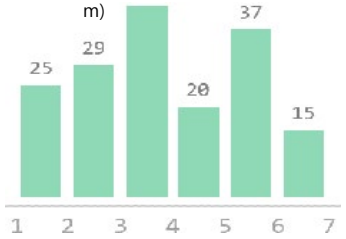
그래프 종류

- 그래프의 종류는 상당히 많음
- 그래프마다 특징과 목적이 다름
- 데이터 마다 내가 보고자 하는 목적에 따라, 그래프를 적절하게 사용할 필요가 있음

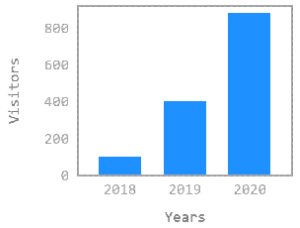
선형 그래프(Line Chart)



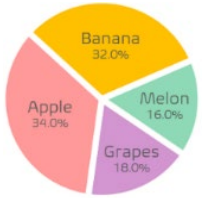
히스토그램(Histogram)



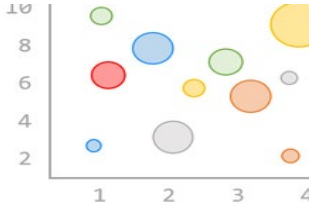
막대그래프(Bar Chart)



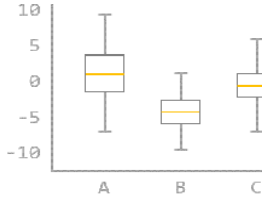
원형그래프(Pie Chart)



산점도(Scatter Plot)



상자그림(Box Plot)



4. 데이터 시각화(Cont.)

그래프별 특징

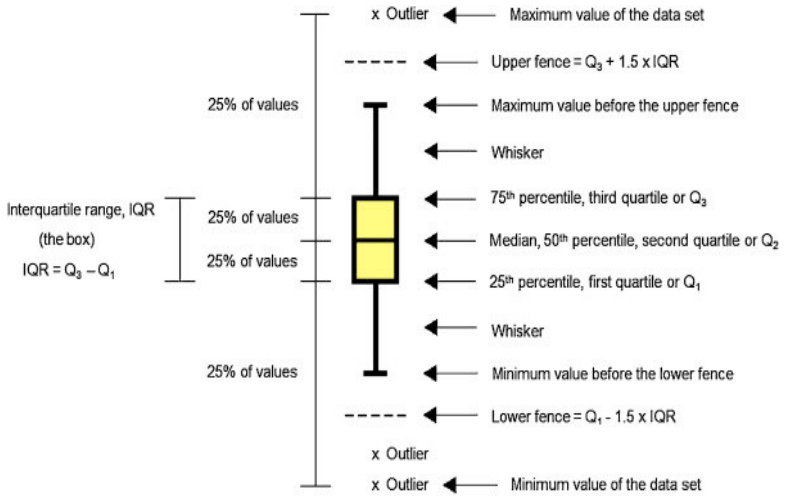
그래프	설명	예시
선형그래프 (Line Chart)	데이터를 시간의 흐름에 따른 변화를 확인	5년간 주가변동, 월간 웹페이지 유입수, 분기별 수익 변동 등
히스토그램 (histogram)	연속형 데이터의 구간별 빈도수 분포 확인	성적 구간별 성적 분포 확인 등
막대그래프 (Bar Chart)	범주형 데이터 통계량(평균, 개수 등) 확인	매장별 평균 매출, 설비별 생산 제품 수 등
원형그래프 (Pie Chart)	범주형자료별(카테고리별) 비율 확인	남녀 성비, 과일 재고 비율 등
산점도 (Scatter Plot)	두 개의 수치형 데이터의 관계를 확인	해변에서 아이스크림 판매에 따른 상어 사고 사건 수
상자그림(Box Plot)	수치형 데이터의 분포(최대값, 제1사분위수, 중앙값, 제2사분위수, 최소값)을 요약	데이터의 이상치 확인을 위한 모든 경우

4. 데이터 시각화(Cont.)

그래프 종류

- 상자그림(Box plot)
  - 많은 데이터를 눈으로 확인하기 어려울 때 데이터 집합의 범위와 중앙값을 빠르게 확인할 수 있는 목적으로 사용
  - 최소값, 제 1사분위 수(Q1), 중앙값(Q2), 제 3사분위 수(Q3), 최대값, 통계적으로 이상치(outlier)가 있는지 확인 가능

구분	설명
제 1사분위수(Q1) (First Quartile)	전체 데이터 중 하위 25%에 해당하는 값
제 2사분위수(Q2) (Second Quartile)	중앙값을 의미, 전체 데이터의 정 가운데 순위에 해당
제 3사분위수(Q3) (Third Quartile)	전체 데이터 중 상위 25%에 해당하는 값
IQR (Inter Quartile Range)	Q3-Q1을 의미. 데이터의 중간 50%
최소값(minimum)	Q1-α(1.5)*IQR 보다 큰 데이터 중 가장 작은 값
최대값(maximum)	Q3+α(1.5)*IQR 보다 작은 데이터 중 가장 큰 값
이상값(outlier)	최소값보다 작거나 최대값보다 큰 데이터



4. 데이터 시각화(실습)

- 그래프 종류

4. 데이터 시각화(실습)

- data_plot/data_plt1.ipynb

Character	Description	Character	Description
'.'	Point marker	's'	Square marker
','	Pixel marker	'p'	Pentagon marker
'o'	Circle marker	'P'	Plus(filled) marker
'v'	Triangle down marker	'*'	Star marker
'^'	Triangle up marker	'h'	Hexagon1 marker
'<'	Triangle left marker	'H'	Hexagon2 marker
'>'	Triangle right marker	'+'	Plus marker
'1'	Tri_down marker	'x'	X marker
'2'	Tri_up marker	'X'	X(filled) marker
'3'	Tri_left marker	'D'	Diamond marker
'4'	Tri_right marker	'd'	Thin_diamond marker
'8'	Octagon marker	'|'	Vline marker
		'_'	Hline marker

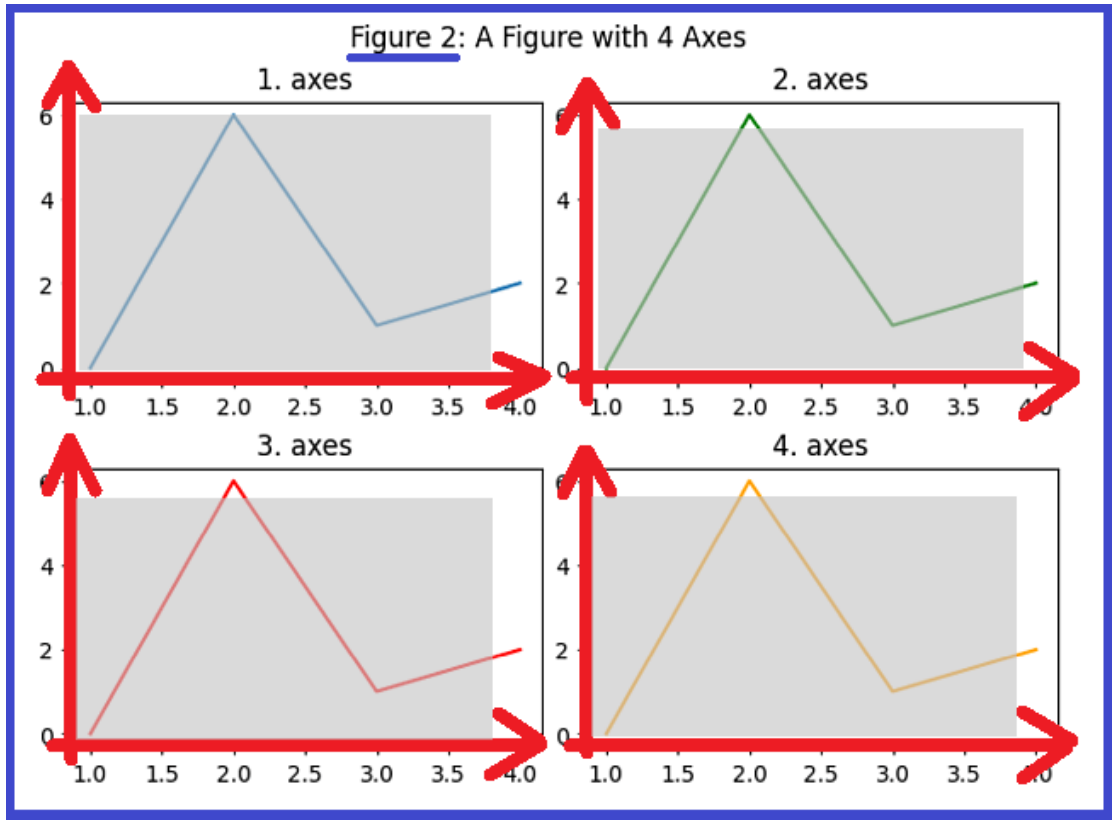
선 표시

Character	Description
'-'	Solid line style
'—'	Dashed line style
'-.'	Dash-dot line style
':'	Dotted line style

마크 표시

4. 데이터 시각화(실습)

- data_plot/data_plt1.ipynb



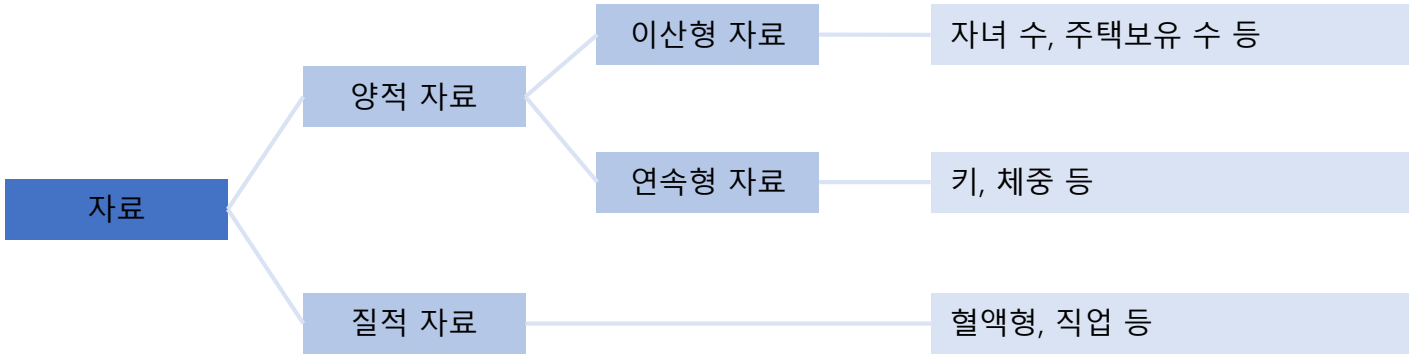
4. 데이터 시각화(실습)

- 실전데이터를 활용한 다양한 plot 및 데이터 전처리
  - PDF_DATA/pdf_plot/pdf_data_plt.ipynb

5. 데이터 유형

질적자료 vs 양적자료

질적 자료(qualitative data)	양적 자료(quantitative data)
범주형 데이터(혈액형, 학력, 직업 등) 원칙적으로 숫자로 표현할 수 없는 자료이며, 측정대상을 분류하기 위해 숫자를 부여하기도 함 빈도수, 최빈값, 비율, 백분율 등을 이용하여 데이터의 분포 특성을 알 수 있음 주로 원형 그래프, 막대형 그래프로 시각화	수치형 데이터(키, 체중, 자녀 수 등) 자료의 크기나 양을 숫자로 표현할 수 있음 셀 수 있는 정수값으로 표현되는 이산형 자료, 연속적인 숫자로 표현되는 연속형자료 평균, 분산, 표준편차, 첨도, 왜도 등을 이용하여 데이터의 분포 특성을 알 수 있음 주로 박스 플랏이나 히스토그램으로 시각화



문서형 악성코드 탐지 AI 모델

1. 머신러닝이란?

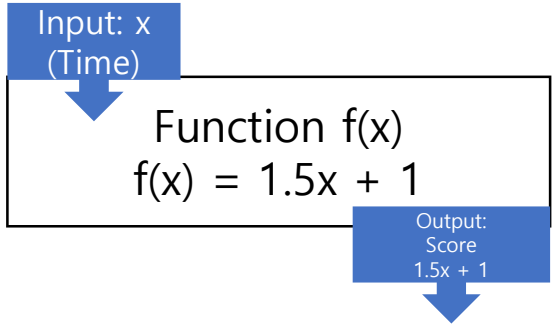
- What is Machine Learning
  - A field of study that lets computers have the ability to learn by themselves without being **explicitly** programmed
  - 전통적인 컴퓨터 프로그래밍은 주어진 데이터에 대해서 임의로 명시적으로 어떻게 프로그래밍할지를 지정했지만, 기계학습이라는 컴퓨터 프로그램은 주어진 Data하에서 스스로 관계를 추론하는 프로그래밍
  - A study of computer algorithms that allow computer program to automatically improve through experience
  - 어떤 경험을 통해서 자동적으로 컴퓨터가 학습할 수 있는 프로그램을 뜻하며, 경험은 Data를 뜻함

1. 머신러닝이란?(Cont.)

- What is Machine Learning
  - Given a set of time-score pairs, predict an exam score.
  - Time과 Score의 패턴을 찾아 함수 f(x)를 생성하여 Score라는 output의 값을 주는 방식이 일반적인 프로그래밍
  - Machine Learning 의 경우, input(time)과 output(score)만을 이용하여 w(1.5)의 값과 b()의 값을 찾아 가는 방식

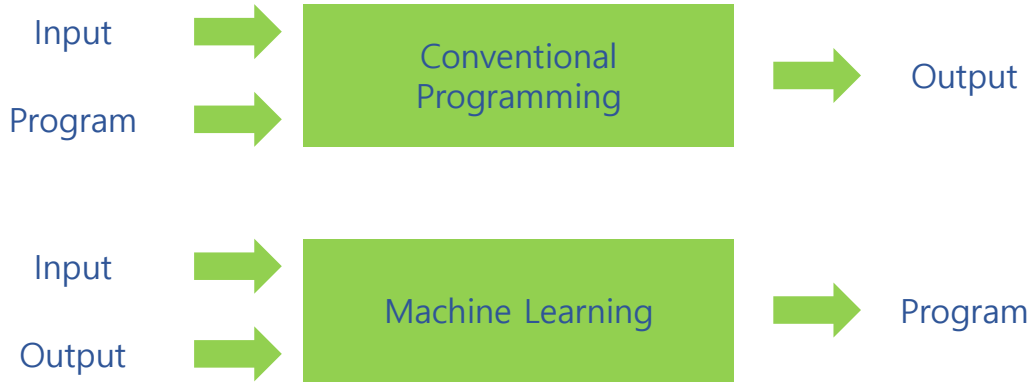
Time	Score
0	1
1	2
2	4
3	5
4	7
6	???

Score = f(Time) = w*Time + b



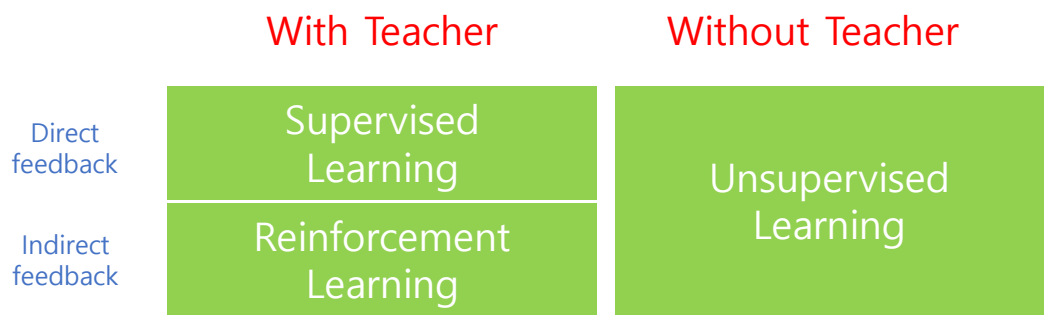
1. 머신러닝이란?(Cont.)

- What is Machine Learning
  - 전통적인 프로그래밍 방식은 함수(program)을 미리 정의하고 parameter(input)에 따라 output을 "계산"하는 방식
  - 머신 러닝이란, Input과 Output이 주어졌을 때 함수(program) 혹은 공식을 "계산"하는 방식

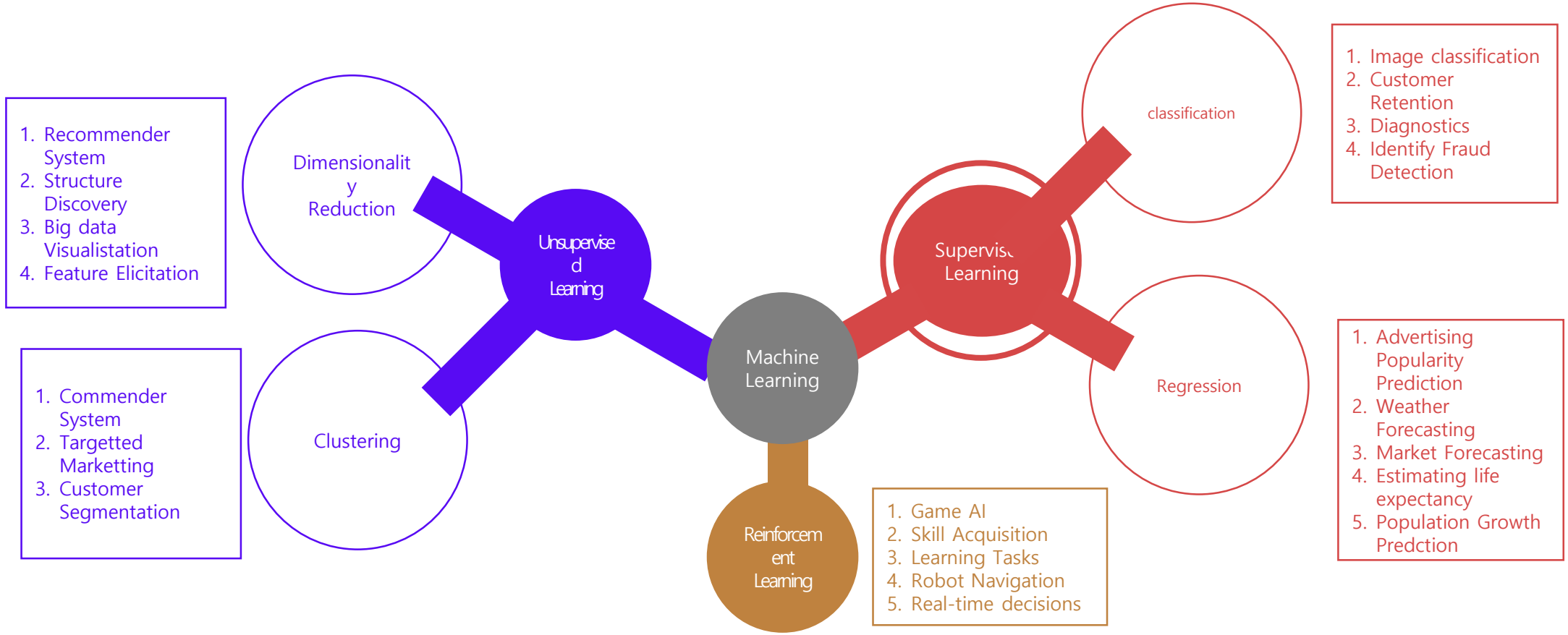


2. 머신러닝 종류

- Supervised Learning
  - 실제 현업에서 가장 많이 쓰는 방식
  - X와 Y의 pair를 활용
  - Y의 값은 직접 Labeling 해줘야 하는 수고가 있음
- **Reinforcement learning**
  - state, action의의 pair를 활용(f(s, a) = r)
  - r은 리워드로 어떠한 상태에서 어떠한 행동을 했을 때 주어지는 이점을 점수로 표현한 것
  - 직접으로 어떠한 action이 최적의 action인지를 직접적으로 확인할 수 없으며, 리워드 함수를 통해 간접적으로 파악해야 함



2. 머신러닝 종류(Cont.)



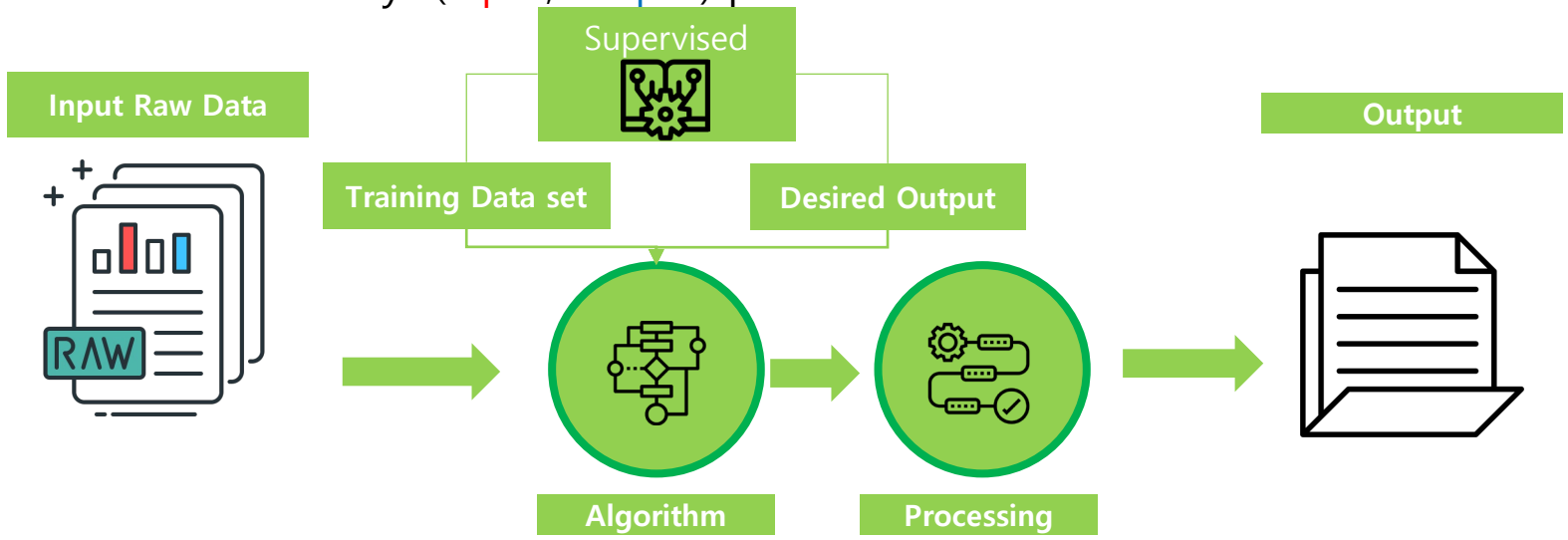
2. 머신러닝 종류

Given a set of input data X = {x1, x2, …. xn}

- Supervised Learning
  - Target outputs(labels, responsed)를 가지고 학습: {y1, y2,… yn}
  - 새로운 데이터에 대해 주어진 output과 같이 예측하는 것이 목적
- unsupervised Learning
  - 오직 input Data X만 가지고 있음
  - 데이터간 숨겨진 관계를 찾는것이 목적
- Reinforcement learning
  - 상태(s)와 모델(agent)이, 그리고 행동(a)을 가지고, 다음 상태 s'에 대한 r1, r2… rn을 가짐
  - 주어진 모델(agent)에서 최대 리워드를 가지는 행동을 찾는 것이 목적

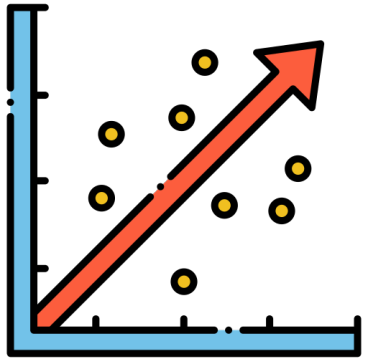
3. 지도학습

- Supervised Learning
  - input-output pairs로 구성
    - Input: convariates, predictors, and features
    - Output: variates, targets, labels
  - Let computers learn with many (input, output) pairs



3. 지도학습(Cont.)

- Supervised Learning
  - Regression vs Classification
  - Regression: Labels are continuous
  - Classification: Labels are discrete

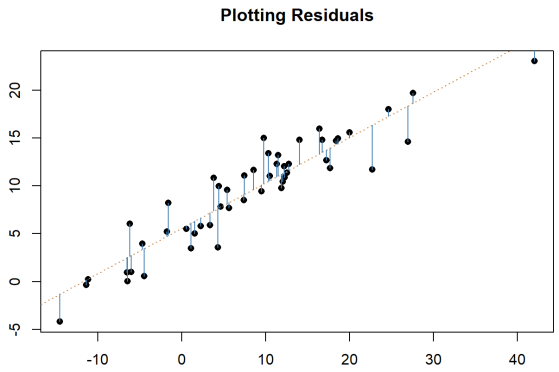


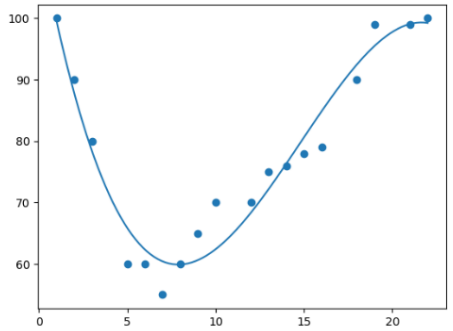
Regression

Classification

3. 지도학습(Cont.)

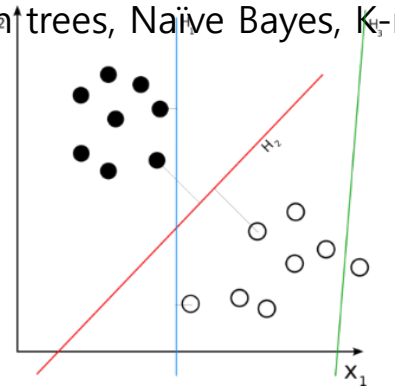
- Regression
  - Predicting real value (i.e., continuous labels)
  - Examples
    - Linear regression: 선형 관계
    - Polynomial regression: 다항식 관계를 가지는 관계(2차 함수, 3차함수 등)

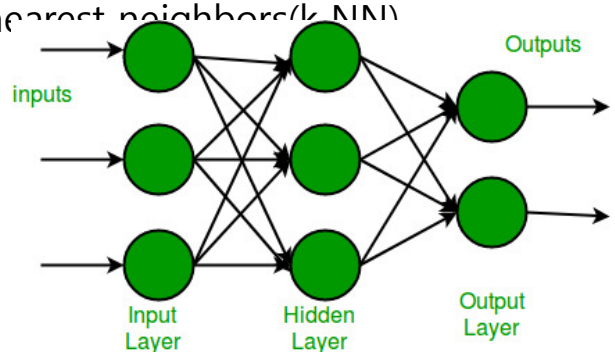


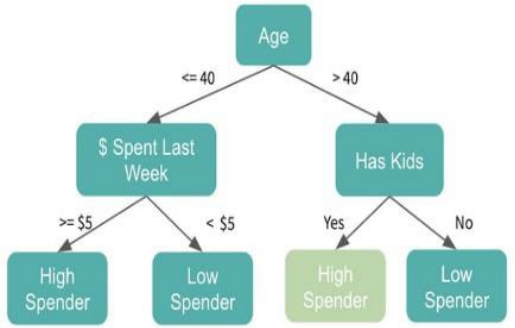


3. 지도학습(Cont.)

- Classification
  - Predicting categorical values (i.e., discrete labels)
    - 데이터 라벨간 차이의 경계를 결정하는 학습
  - Examples
    - Logistic regression, Support vector machine(SVM)
    - Neural Networks: Perceptron, Multilayer perceptron(MLP)
    - Decision trees, Naïve Bayes, K-nearest neighbors(k-NN)

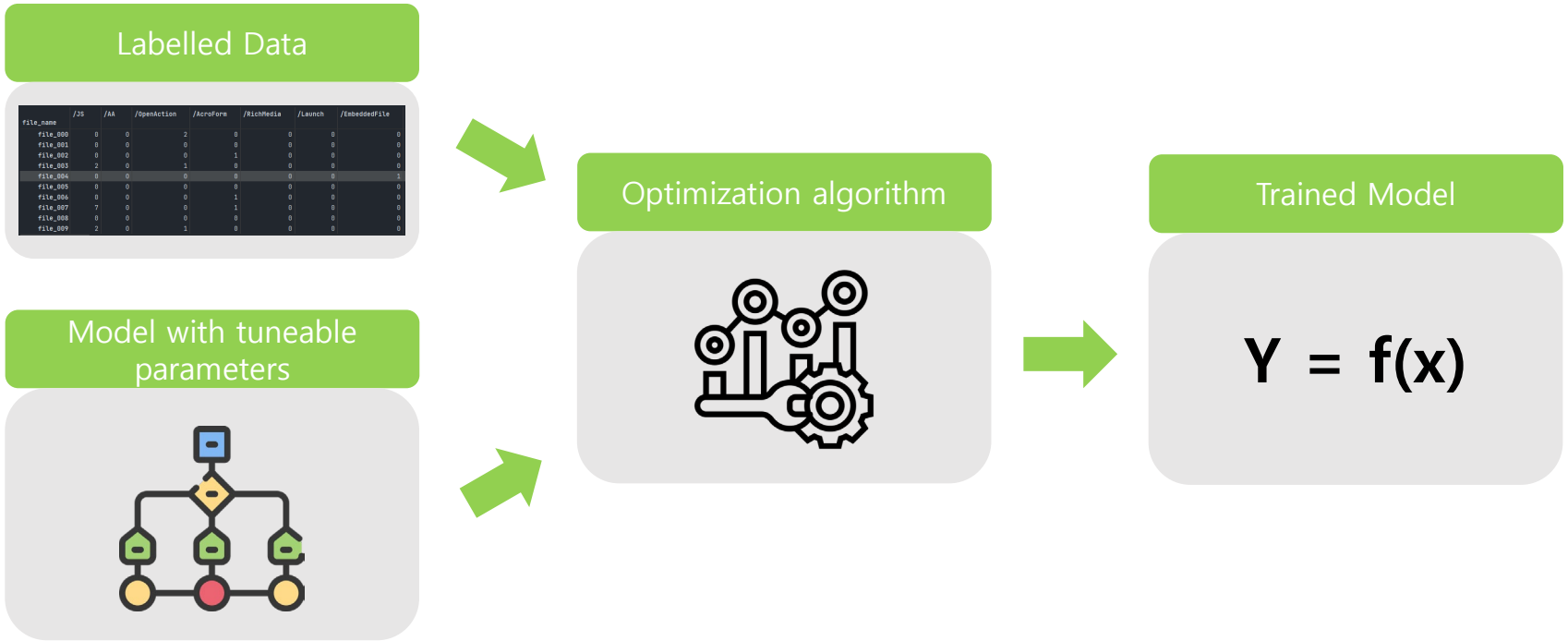






3. 지도학습(실습)

- PDF_DATA/AI/train.ipynb
  - PDF 악성 인자가 Feature로 포함된 파일과 XGBoost를 활용한 AI 모델 학습



4. 모델평가

- Regression models
  - Mean absolute error(MAE) $MAE = \frac{\Sigma|y-\hat{y}|}{n}$
  - Mean squared error(MSE) $MSE = \frac{\Sigma(y-\hat{y})^2}{n}$
  - Root mean squared error(RMSE) $RMSE = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n}}$
  - R2 score(R squard) $R^2 = 1 - \frac{\sum_i(y_i-\widehat{y_i})^2}{\sum_i(y_i-\bar{y}_i)^2}$
  - Error란 "예측 값 – 실제 값"을 의미하며 Loss라고 불리기도 함

4. 모델평가(Cont.)

- Classification Model
  - Confusion matrix
  - Actural Value란 실제 값을 의미, predict Value란 모델의 예측 값
  - TP, TN은 실제 값과 예측 값이 같을 경우
  - FN, FP는 모델의 예측 값과 실제 값이 틀린 경우를 뜻함

		Predict value	
		Positive	**Negative**
Actual Value	**Positive**	TP	FN
	Negative	FP	TN

4. 모델평가(Cont.)

dataset	actual
x1	+
x2	-
x3	-
x4	-
X5	-
X6	-

Model 1

predict
+
+
+
-
-
-

	Positive	Negative
Positive	1	0
Negative	2	3

Model 2

predict
-
-
-
-
-
-

	Positive	Negative
Positive	0	1
Negative	0	5

4. 모델평가(Cont.)

정확도

- 모델의 예측이 실제 정답을 얼마나 맞췄는지를 확인하는 지표
- (TP+TN)/(TP+FP+FN+TN)
- Error rate = 1 - accuracy

		Predict value	
		Positive	Negative
Actual Value	Positive	TP	FN
	Negative	FP	TN

4. 모델평가(Cont.)

- 정밀성(Precision)
  - 모델이 정답이라고 판단할 때 실제 값과 일치하는 비율
  - 모델이 악성코드로 판단할 때 악성코드의 비율
  - Precision = TP/(TP+FP)
- 민감도(Recall)
  - 실제 값 중 모델이 정답이라고 판단한 비율
  - 전체 악성코드 중 모델이 악성코드라고 판단한 비율
  - Recall = TP/(TP+FN)
- F-Score
  - 데이터의 불균형일땐, 데이터의 편향성 때문에 정확도보다 좋은 평가 지표로 사용
  - 정밀성(P)과 민감도(R)의 조화 평균을 활용
  - Fα = (α+1) * (P*R)/(P+R)

4. 모델평가(Cont.)

dataset	actual
x1	+
x2	-
x3	-
x4	+
X5	-
X6	-

Model 1

predict
+
+
-
+
+
+

	Positive	Negative
Positive	2	0
Negative	3	1

Precision	2/5
Recall	2/2
F-socre	0.571429

Model 2

predict
+
-
-
-
-
-

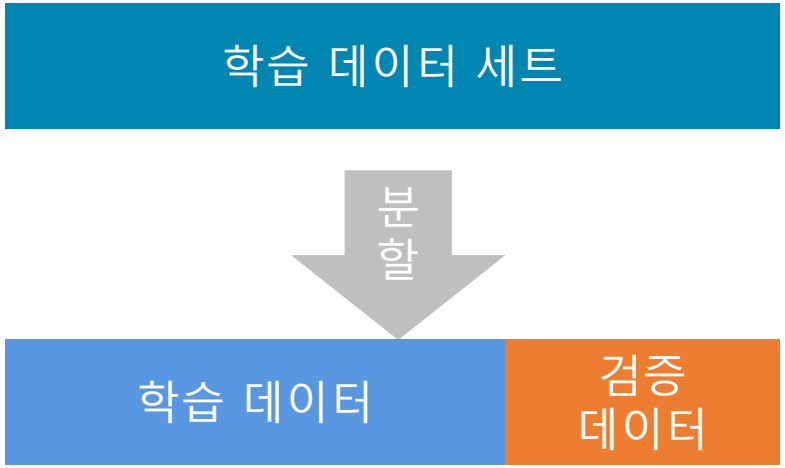
	Positive	Negative
Positive	1	1
Negative	0	4

Precision	1/1
Recall	1/2
F-socre	0.666667

5. 교차검증

- 교차 검증
  - 별도의 여러 세트로 구성된 학습 데이터와 검증 데이터 세트에서 학습과 평가를 수행
  - 수능을 보기 전 모의고사를 여러 번 보는 것과 같음

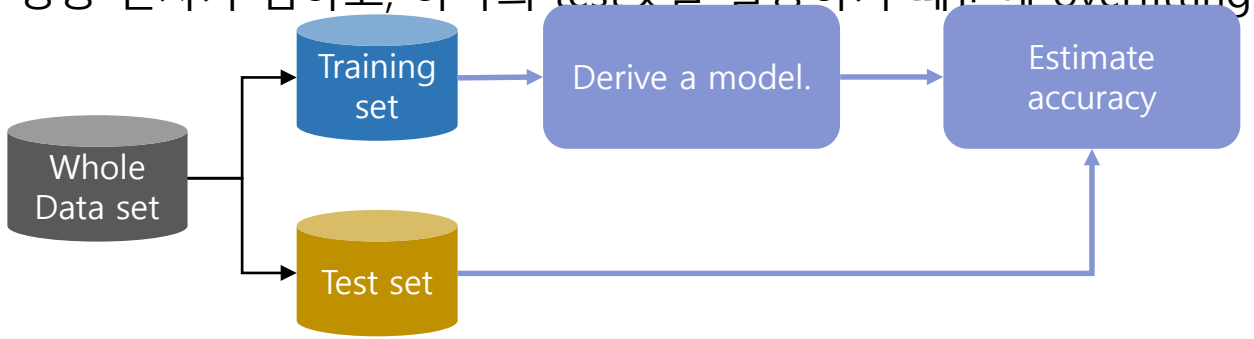
학습 데이터를 **학습 데이터**와 학습된 모델의 성능을 1차 평가하는 **검증 데이터**로 분할



모든 학습 및 검증 과정이 끝난 후 최종적으로 테스트 **데이터 세트**로 모델의 성능을 평가함

5. 교차검증

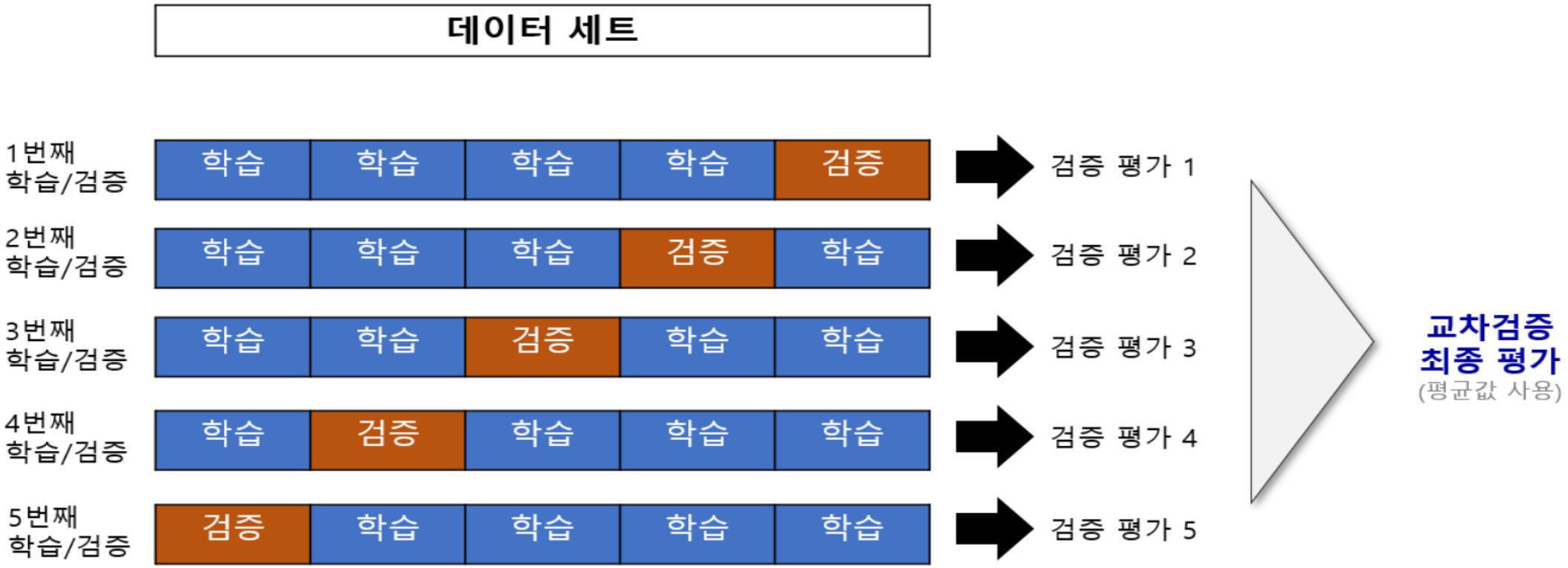
- 홀드 아웃(Hold-Out)
  - 전통적인 데이터 분할 방법
  - 데이터를 무작위로 학습데이터와 테스트 데이터 두 가지로 구분하고, 학습 데이터는 또다시 학습 데이터와 검증 데이터로 분할
  - 간단하고 쉽게 구현이 쉬우나, 랜덤 split이기 때문에 데이터 선택에 있어 공평함을 제공할 수 없고 모델마다의 성능 편차가 심하고, 하나의 test셋을 활용하기 때문에 overfitting 문제가 발생



5. 교차검증(Cont.)

- Cross Validation(K-fold)
  - K개의 Sub Data Set을 생성하여 모든 데이터가 적어도 한 번의 학습을 할 수 있는 환경 생성
  - 장점
    - 모든 데이터 셋을 평가에 활용
    - 평가에 사용되는 데이터 편중을 막을 수 있음
    - 평가 결과에 따라 좀 더 일반화된 모델을 만들 수 있음
    - 모든 데이터 셋을 훈련에 활용
    - 정확도 향상 가능
    - 데이터 부족으로 인한 underfitting 방지
  - 단점
    - Iteration횟수가 많기 때문에 모델 훈련/평가 시간이 오래 걸림

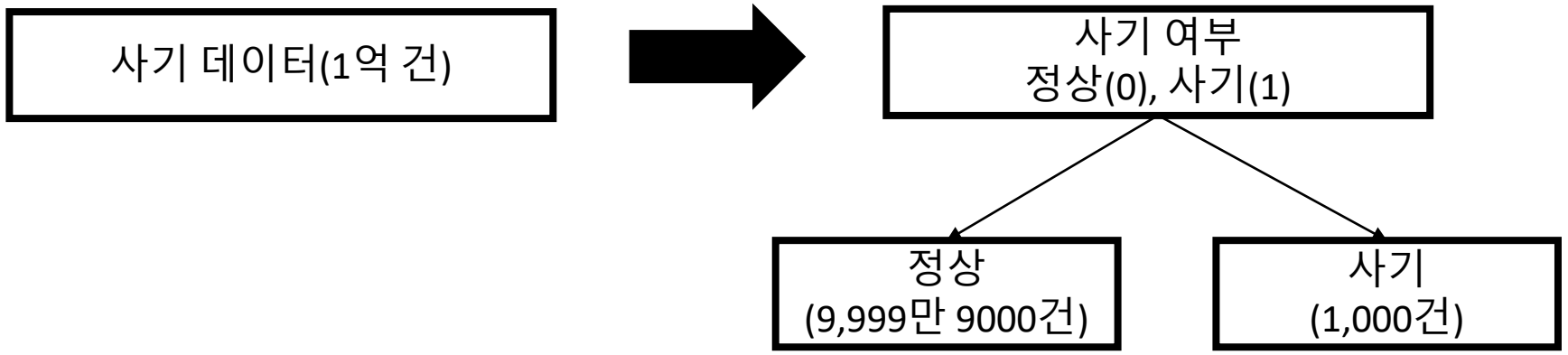
5. 교차검증(Cont.)



출처: https://velog.io/@jiazzang/교차검증-홀드아웃-K-Fold-Stratified-K-Fold-crossvalscore

5. 교차검증(Cont.)

- 계층적(Stratified) K-Fold 교차 검증
  - 일반적으로 회귀(Regression)문제가 아닌 분류(Classification) 문제에서 사용
  - K개의 Sub Data Set을 생성하여 모든 데이터가 적어도 한 번의 학습을 할 수 있는 환경 생성
  - 균등한 분포를 가진 레이블 데이터 집합을 위한 K-Fold 방식으로 각 Fold의 레이블 데이터의 비율이 같음
- 예시



5. 교차검증

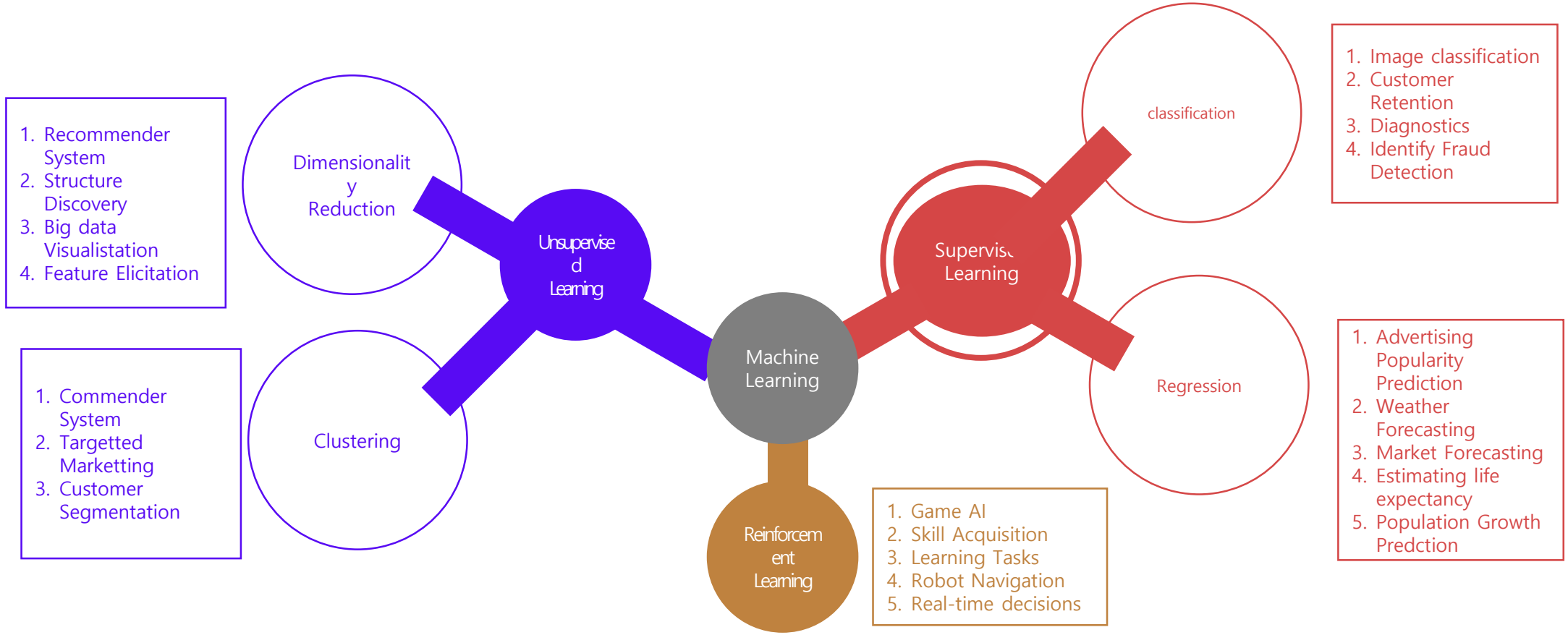
- PDF_DATA/AI/cross_validation.ipynb
  - PDF 악성 인자가 Feature로 포함된 파일과 XGBoost를 활용한 AI 모델 학습
  - Cross Validation을 통한 5-fold 학습 진행
  - 각 fold별 acc, recall, precision, f1-score 및 평균값 출력
  - F1-score가 가장 좋은 모델 저장

XAI를 활용한 AI모델 해석

1. XAI의 필요성

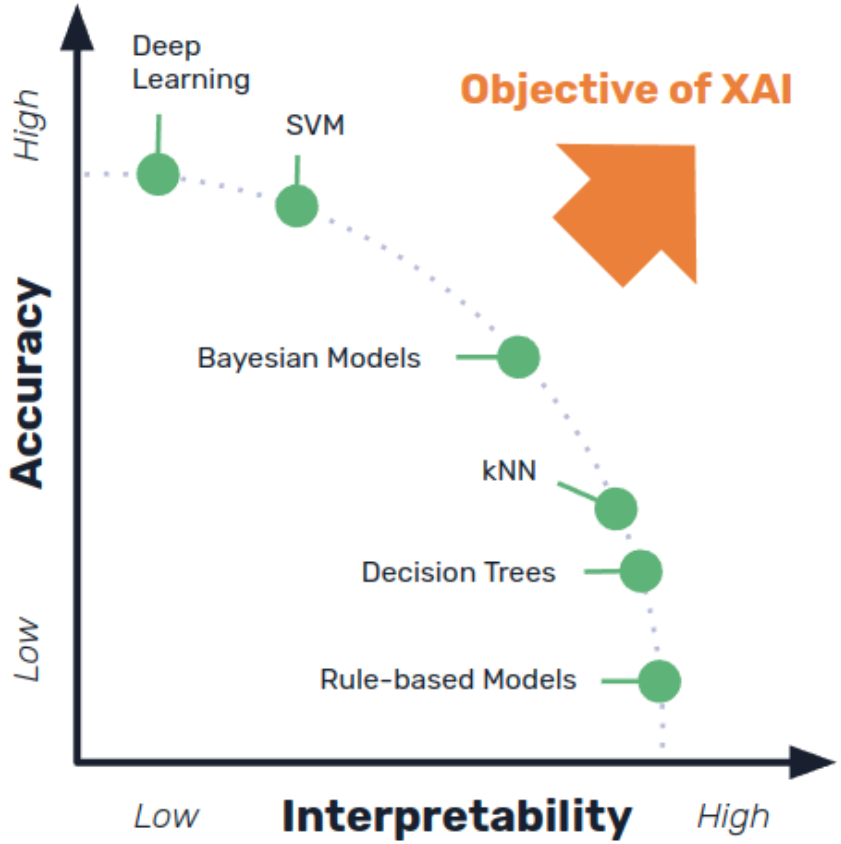
- 머신 러닝은 점점 더 우리 생활권역에 퍼져 있음
- 그러나 대다수의 머신 러닝 모델은 모델 판단 근거를 제시하기에 어려움이 존재
- 국가와 머신 러닝을 사용하는 영역에서는 모델이 판단하는 과정을 투명하게 보기 위한 노력을 진행
- 특히 의료, 보험 등과 같은 분야에서의 환자들을 위한 판단 근거는 환자들에게 설명을 요구할 필요 및 규약이 존재

1. XAI의 필요성(Cont.)

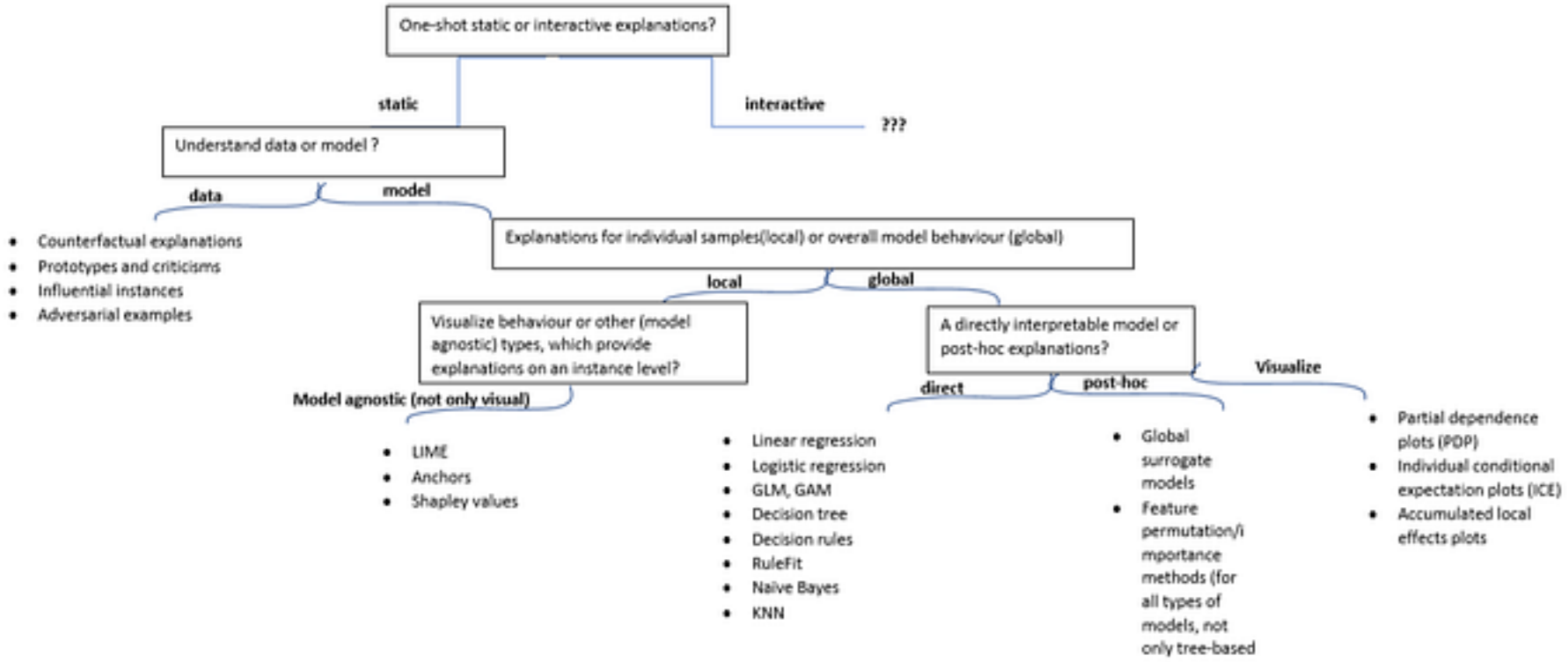


1. XAI의 필요성

- 모델의 정확성과 설명성은 서로 상충함
- 일반적으로 해석가능성이 높을수록 낮은 정확성을 보임
- 복잡한 모델인 Deep Learning의 경우 높은 성능을 보이지만, 방대한 파라미터로 인해 해석하기 어려움
- 반대로, Rule-based Model은 간단한 Rule로 이루어진 모델로 정확도는 낮지만 사용자가 해석하기 쉬움



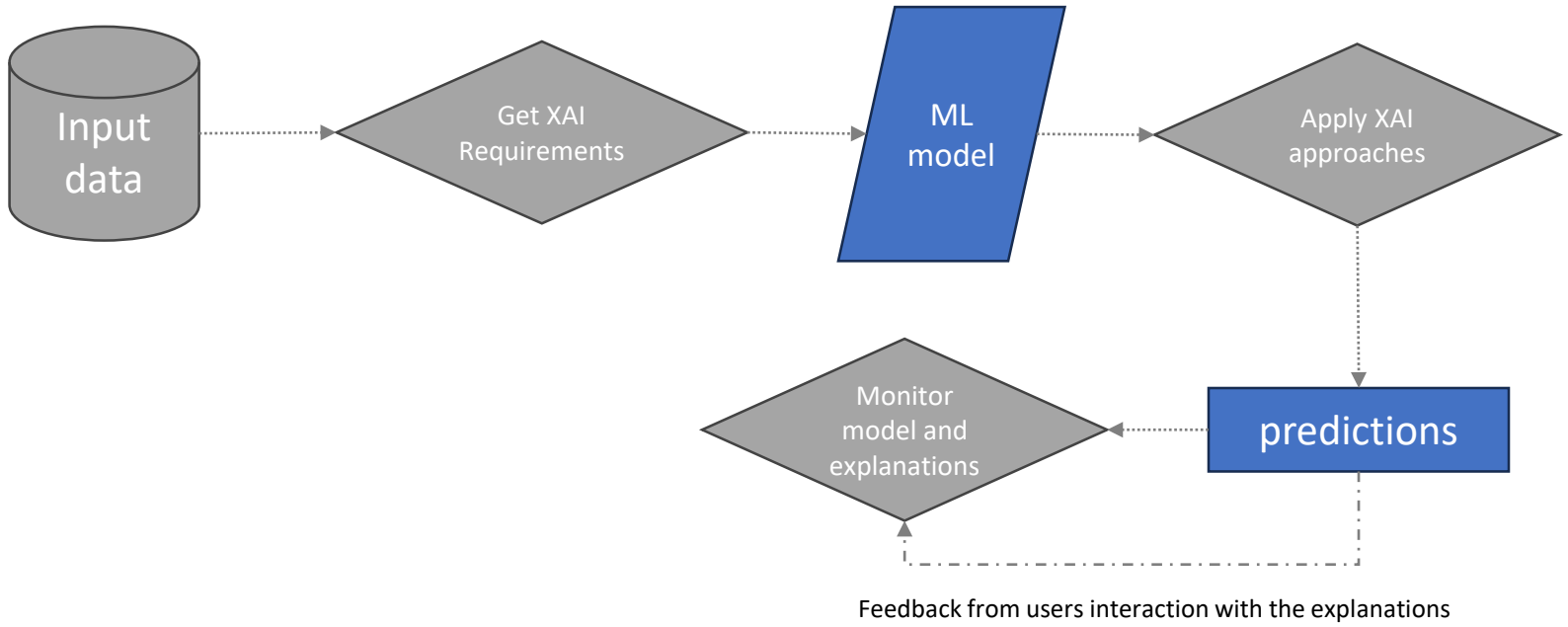
2. XAI의 종류 및 프로세스(Cont.)



2. XAI의 종류 및 프로세스(Cont.)

기술	특징	장점	단점
LIME	특정 예측에 대한 지역적 선형 모델을 사용하여 설명	모델에 독립적, 쉽게 구현 가능	높은 계산 비용, 설명의 일관성 문제
SHAP (Shapley Values)	**게임 이론을 기반으로 각 특징이 예측에 기여하는 정도를 계산**	**높은 설명력, 공정성**	**계산 비용이 높음, 복잡한 해석**
Anchors	특정 예측을 "고정"시키는 규칙 기반 설명 제공	직관적, 특정 예측에 대해 높은 설명력	복잡한 규칙 생성, 모델에 따라 다를 수 있음
Decision Trees	트리 구조를 사용하여 의사 결정을 설명	직관적 이해, 시각화 용이	복잡한 데이터에서 성능 저하, 과적합 가능성
Feature Permutation /Importance	특징의 중요도를 평가하기 위해 모델 출력 변화를 관찰	간단하고 직관적, 모든 모델에 적용 가능	중요한 특징 간 상호작용 무시
Attention	입력의 각 부분이 출력에 미치는 영향을 학습하여 시각화	직관적 시각화, 자연어 처리에 효과적	모델 복잡도 증가, 일부 모델에만 적용 가능
CAM (Class Activation Mapping)	CNN의 출력에 기여한 입력의 위치를 강조	이미지 데이터에 효과적, 시각적 설명	주로 CNN에 한정, 다른 모델에 적용 어려움
Integrated Gradient	입력과 기준 사이의 경로를 따라 그래디언트를 적분하여 중요도 계산	신경망 모델에 효과적, 계산 복잡도 적음	기울기 소실 문제, 기준 선택에 민감함

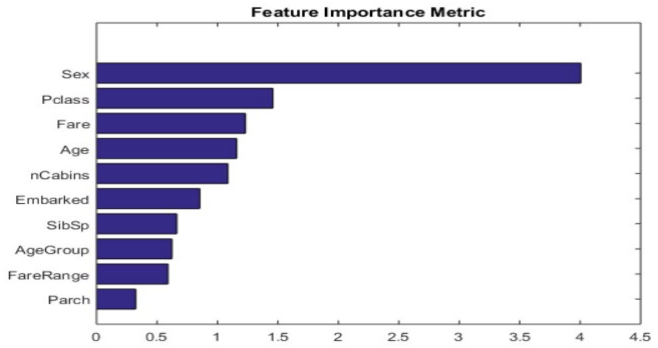
2. XAI의 종류 및 프로세스(Cont.)



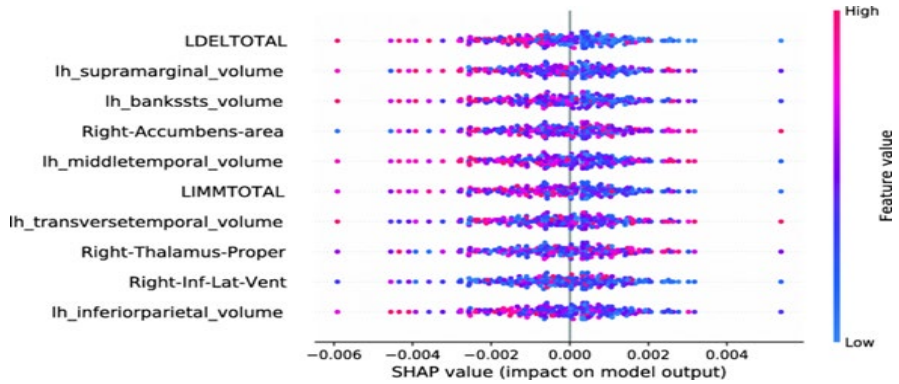
3.전역설명(Global)

- 'AI 모델의 전체적인 동작'을 이해하는 것을 목적
- 하나 하나의 사례(입력 데이터)에 대한 예측 과정을 설명하는 것이 아닌 다양한 사례에서의 예측을 통합해 보았을 때 'AI 내부의 지배적인 경향'을 설명
- 장점
  - **신뢰성**: 사용자가 모델의 동작 원리를 이해하고, 신뢰할 수 있음.
  - **투명성**: 모델의 작동 방식이 명확하게 설명되어 투명성을 제공
  - **규제 준수** :GDPR과 같은 법적 요구 사항을 준수하는 데 유용
  - **문제 발견**: 모델의 잠재적인 문제나 편향을 발견하는 데 유용

3.전역설명(Global)



출처:https://blogs.mathworks.com/loren/2015/06/18/getting-started-with-kaggle-data-science-competitions/



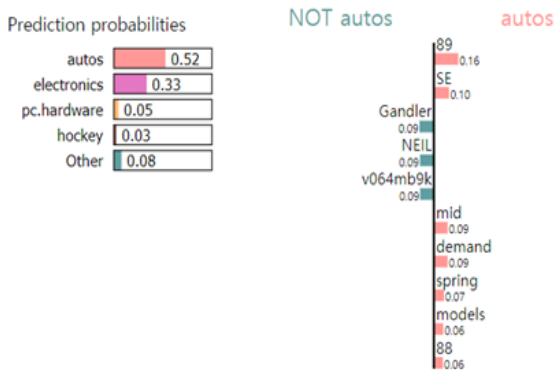
출처:https://alzres.biomedcentral.com/articles/10.1186/s13195-021-00879-4

설명 종류	특징
특성 중요도(Feature Importance)	모델이 예측을 위해 사용하는 각 특성이 얼마나 중요한지를 측정
Partial Dependence Plots	특성 값의 변화에 따라 예측 결과가 어떻게 변하는지를 시각화
Global Surrogate Models	복잡한 모델을 더 단순한 모델로 대체하여 전체적인 동작을 설명 (예: 선형 회귀, 의사 결정 트리)을 통해 원래 모델의 예측 방식을 이해
Shapley Values	모든 특성 조합을 고려하여 공정하게 특성 중요도를 평가합니다.

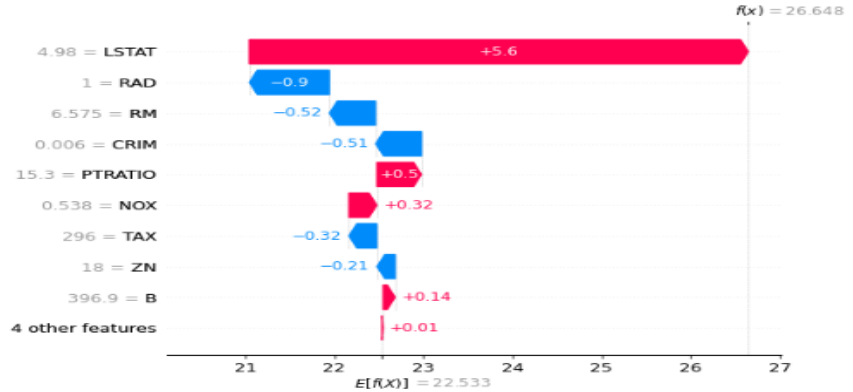
4.지역설명(Local)

- '각 예측 결과의 판단 사유를 이해하는 것'을 이해하는 것을 목적
- 하나 하나의 사례(입력 데이터)에 대한 예측 과정을 설명
- 장점
  - **투명성 향상**: 모델의 개별 예측이 어떻게 도출되었는지 명확하게 설명하여 사용자의 신뢰를 높임.
  - **사용자 피드백**: 비전문가도 모델의 예측 결과를 이해하고 피드백을 제공할 수 있음.
  - **오류 분석**: 모델이 잘못된 예측을 했을 때, 그 원인을 분석하고 수정할 수 있음.
  - **규제 준수**: GDPR 등 법적 요구 사항을 준수하는 데 도움이 됨.

4.지역설명(Local)



출처:https://velog.io/@tobigs_xai/1주차-대리분석LIME



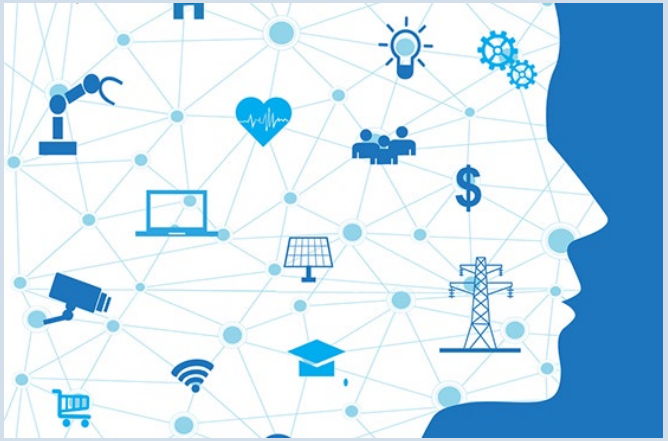
출처:https://csshark.tistory.com/70

설명 종류	특징
LIME (Local Interpretable Model-agnostic Explanations)	어떤 모델에도 적용 가능. 특정 예측에 대해 **단순한 모델**(예: 선형 회귀)을 사용해 해당 지역에서의 예측을 근사화하여 설명.
SHAP (SHapley Additive exPlanations)	게임 이론에 기반하여 각 특성이 예측에 얼마나 기여했는지를 설명. SHAP 값은 공정성을 보장하며, 전체 모델을 지역적으로 설명할 수 있음.
Decision Trees	모델이 예측을 위해 따르는 의사결정 과정을 시각적으로 나타냄. 간단하고 직관적인 방법으로 예측 과정을 설명할 수 있음.
Anchors	모델의 예측을 지배하는 "if-then" 규칙을 생성하여 설명. LIME과 유사하지만, 더 직관적이고 해석 가능한 규칙을 제공.

5. XAI 실습

- PDF_DATA/XAI/XAI.ipynb
  - 미리 학습한 모델을 활용하여 XAI(Shap Lib)적용
  - 전역 설명 플롯 저장
  - 지역 설명 플롯 저장
  - 전역+지역 설명 비교 분석

KAIST CSRC Cyber Security Research Center



감 사 합 니 다

CYBER SECURITY RESERCH CENTER