2024.01.13

# 14기 신입교육세션

데 이 터  분 석  소 개  및  E D A

B.a.f

# CONTEST

1 데이터 분석과 소개

2 데이터 수집

3 EDA 실습 및 팀별 토의

4 과제 안내 - Github

# 데이터 분석과 소개

1. 분석 목적

2. 데이터/변수 분류 및 분석 방법

3. 데이터 분석 순서

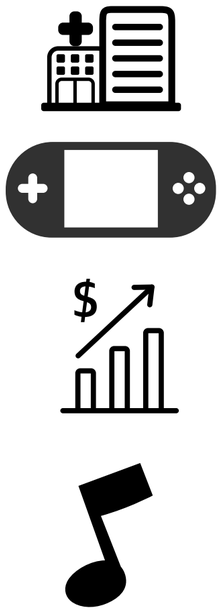
4. 관련 프로그램

B.a.f

# 1. 분석 목적

## 데이터만 있다면 뭐든 할 수 있다!

빅데이터 등장 -> 다양한 분야의 데이터 플랫폼이 존재



온라인 쇼핑몰의 판매량 예측
충무로역 배달 수요 예측
특정 조건을 가진 고객이 물건을 살지 말지 예측
게임의 승패 예측
고객의 성향과 니즈 파악, 맞춤 서비스 제공
연구 결과 해석
의약품의 효과 유의성 파악
병원 최적 장소 파악
손글씨 인식
보행자 및 장애물 인식
영상 조회수 예측

게임에서 승리 전략 짜기
시니어 맞춤 여행 상품 및 관광 코스 개발
구내식당 메뉴 선정
다음 시즌에 유행할 색상과 스타일 분석
공장 관리 상황 파악
.
.
.

# 2. 데이터/변수 분류 및 분석 방법

## 데이터 분류

- **정형 데이터** ex) csv 파일
- **비정형 데이터** ex) 텍스트, 영상, 음성, 이미지 파일
  정형데이터와 다른 전처리 방법을 사용

## 변수 종류

- **독립변수** ex) 성별, 강수량
  = feature, 설명변수, column
- **종속변수** ex) 매출액, 생존 여부
  = 반응변수, target 변수

- **질적변수**
  - **명목형** ex) 색깔
  - **순서형** ex) 건강상태(건강, 양호, 심각)
- **양적변수**
  - **이산형** ex) 자동차 등록대 수
  - **연속형** ex) 몸무게

* 변수 종류별로 EDA, 전처리 방법이 다름
* 변수 타입은 도메인과 EDA 후 변환하기도 함

# 2. 데이터 분류 및 분석 방법

## 데이터 분석 방법

**지도학습 : 정답이 있는 데이터**
정답을 맞추는 것이 중요

- **분류** 종속변수가 질적 변수 ex) 게임의 승패 예측
- **회귀** 종속변수가 양적 변수 ex) 판매량 예측

**비지도학습 : 정답이 없는 데이터** ex) 클러스터링(군집화), 차원축소
패턴&형태를 찾아내 의미를 부여하는 것이 중요
상황과 목적에 맞는 분석 방법 채택

**강화학습 : 높은 점수를 낼 때마다 보상 제공** ex) DQN, A3C
많은 학습 데이터 필요, 최적의 방법을 찾는 것 중요

# 2. 데이터 분류 및 분석 방법

## 지도학습 - 분류

특정 조건을 가진 고객이 물건을 살지 말지 예측

게임의 승패 예측

의약품의 효과 유의성 파악

손글씨 인식

보행자 및 장애물 인식

## 지도학습 - 회귀

온라인 쇼핑몰의 판매량 예측

충무로역 배달 수요 예측

영상 조회수 예측

## 비지도학습

시니어 맞춤 여행 상품 및 관광 코스 개발

다음 시즌에 유행할 색상과 스타일 분석

공장 관리 상황 파악

병원 최적의 장소 파악

## 강화학습

게임에서 승리 전략 짜기

# 2. 데이터 분류 및 분석 방법

## 지도학습 - 분류

특정 고객이 해당 식단을 먹을지 여부 예측

## 지도학습 - 회귀

식자재 가격 예측

요일별/시간별 수요 예측

## 구내식당 메뉴 선정

고객 군집화
대상별 선호 메뉴 조사
수요 예측
식자재 가격 조사/예측
최근 음식 트렌드 조사
메뉴 군집화
.
.

## 비지도학습

고객 성향에 따른 군집화

- 나이, 성별, 직업, 체질, 특이사항, 선호 음식 등

메뉴별 군집화

- 맵기, 당도, 나트륨수치, 식자재 가격 등

최근 음식 트렌드 키워드 군집화

# 3. 데이터 분석 순서

## 데이터 수집

- 주제 선정 / 변수 구체화
- 데이터 수집 및 추출

<수집 시 고려사항>
데이터 개수, 필요성, 신뢰성

도메인 지식 사전 조사 필수

## EDA (데이터 탐색)

- 데이터 기본 정보 확인
- 결측치 및 이상치 확인
- 변수 간의 상관관계 확인

- 데이터 시각화 필수
- 가설 검정

간단한 인사이트를 얻을 수 있음

## 데이터 전처리

- 결측/이상치, 중복값 처리
- 데이터 연계, 통합
- 변수 선택 및 변환
- 파생 변수 생성

새롭게 생성된 데이터
-> 새로운 EDA 필요

## 분석 및 모델링

- 통계 분석 및 모델링
- 머신러닝/딥러닝 이용
- 패턴 인식
- 유의미한 결과 도출
- 성능을 높임

머신러닝 및 딥러닝 기법

## 인사이트 도출

**최종 목적은 결국**
**의사결정**

결과 요약 시각화,
스토리텔링 능력 필요

데이터 엔지니어

서베이 리서치

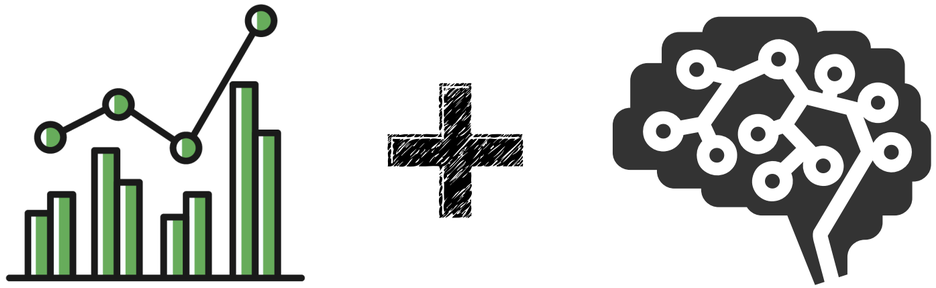
데이터 분석가 (기획자적 성향)

머신러닝 엔지니어 (개발자적 성향)

데이터 사이언티스트 (연구적 성향)

# 3. 데이터 분석 순서

## 분석 및 모델링 방법



**통계학적 접근**

**모르는 모수를 추정하는 데 있어서 가능성을 높이는 방법 탐색**

#수학 #추론 #분포가정

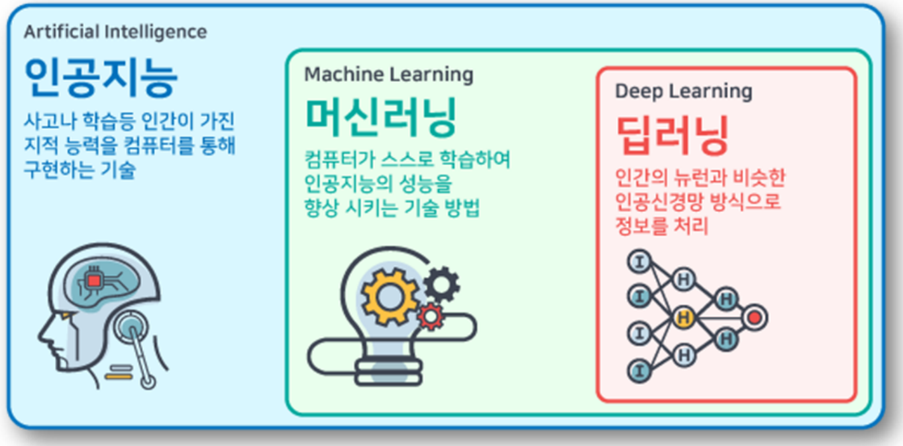
**머신러닝/딥러닝 이용**

**오차를 줄이고 손실함수를 최소화할 수 있는 일반화 모델을 만드는 방법 탐색**

#컴퓨터과학 #예측

## 머신러닝/딥러닝이란?

**인간의 학습능력을 컴퓨터가 갖게 환경을 만들어 주고 학습시키는 것!**



# 4. 관련 프로그램






# 데이터 수집

1. 데이터 사이트 소개

2. 캐글 타이타닉 data 다운로드

3. 파이썬에서 데이터 불러오기

B.a.f

# 1. 데이터 사이트 소개

## 데이터 사이트

공공데이터포털
서울 열린데이터 광장
sk 빅데이터 허브(통화 데이터)
kaggle
한국 소비자원
국가통계포털
마이크로데이터 통합서비스
문화데이터포털
기상청
통계청
구글트렌드/네이버 데이터랩
.
.

## kaggle

- 분석 대회 플랫폼
- 다양한 데이터셋
- 공유된 코드로 공부 가능

# 2. 캐글 타이타닉 데이터 다운로드

## 타이타닉 데이터

타이타닉 침몰에서 생존/사망한 사람들의 data
주어진 승객들의 **생존/사망 여부를 예측**하는 것이 주제

- PassengerId : id
- Pclass : 티켓클래스 (1:1등석, 2:2등석, 3:3등석)
- Name : 승객 이름
- Sex : 성별
- Age : 나이
- Sibsp : 형제 자매 수
- Parch : 부모 자식 수
- Ticket : 티켓 넘버
- Fare : 티켓 요금
- Cabin : 선실
- Embarked : 승선한 항
- Survived : 생존 여부 (0:사망, 1:생존)

# 2. 캐글 타이타닉 데이터 다운로드

## 1) 캐글 회원 가입

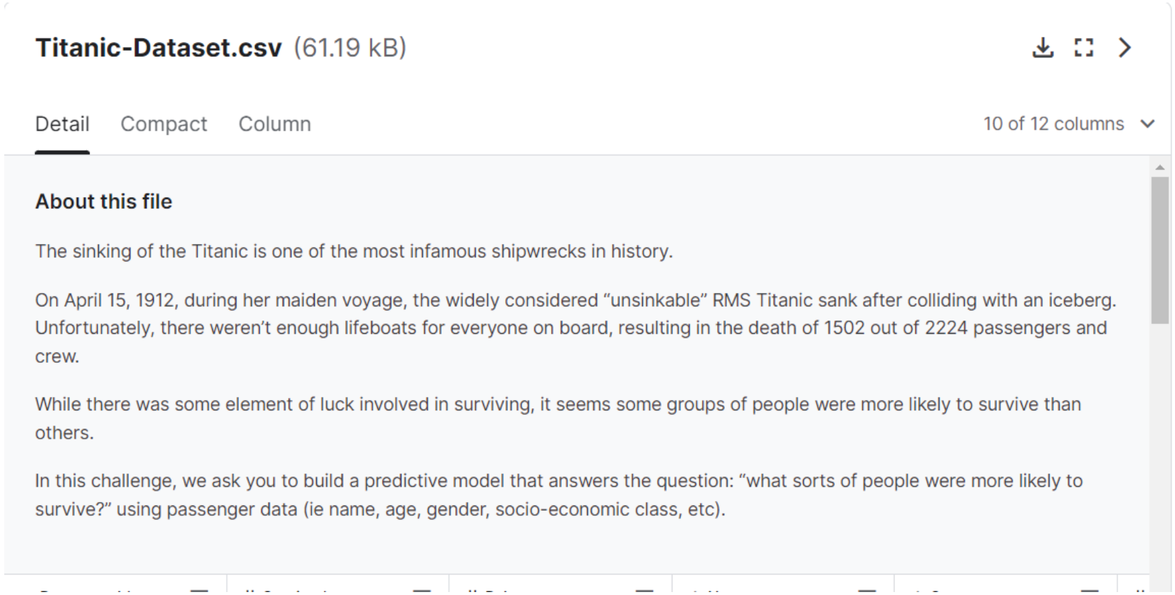
## 2) 타이타닉 데이터 검색

**Titanic Dataset**

Dataset · 2y ago · by M Yasser H

**Titanic** Survival Prediction Dataset

## 3) 적절한 데이터 선택 후 다운로드



# 3. 파이썬에서 데이터 불러오기




```
# pandas 라이브러리 불러오기
import pandas as pd

# 데이터 불러오기
data = pd.read_csv('C:/류수민/3-2/비어플/Titanic-Dataset.csv')
```

파일 경로 파일 이름. 확장자명

# EDA 실습

## 1. 타이타닉 데이터 EDA 실습

B.a.f

# 과제 안내 - Github

1. 과제 안내

2. 깃허브 제출 방법 안내

3. 팀별 토의

B.a.f

# 1. 과제 안내

1. 관심 있는 주제와 이유를 간단히 쓰고, 관련 데이터를 찾아서 깃허브에 업로드 해주세요!

-> 링크로 업로드 해주셔도 되고, 직접 다운로드 후 파일을 첨부하셔도 됩니다

2. 따릉이 데이터로 EDA 한 후 ipnyb 파일을 깃허브에 올려주세요!

-> 이 때 코드에는 간단히 주석을 달아주세요

-> 마크다운을 활용해주셔도 좋습니다

다음주 금요일 낮 12시까지 깃허브에 public으로 업로드 해주세요

# 2. 깃허브 제출 방법 안내

1. 레포지토리 만들기

   깃허브 레포지토리 이름 : BAF-14-Fresh-Edu

2. 관심 있는 주제, 이유 / 데이터 및 ipnyb 파일 첨부하기

# 3. 팀별 토의

팀별 톡방을 개설해 주세요 ! (추후 원활한 토의를 위함입니다)

오늘 다뤘던 내용 / 과제 관련하여 12시까지 자유롭게 토의해주세요

# 3. 팀별 토의

1조

김민열, 김승원, 오윤지, 현민영

2조

서가은, 심고은, 정호원, 홍민화

3조

성지수, 이선재, 장윤서

4조

김민지

# 감사합니다

B.a.f