

2025.01.18

16기 신입교육세션

데이터 분석 소개 및 모델링



B . a . f



CONTEST

1

데이터 분석과 소개

2

데이터 모델링

3

실습 및 팀별 토의

4

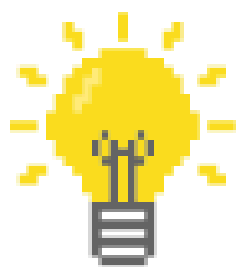
과제 안내 - Github

데이터 분석 소개

1. 데이터 분석 목적
2. 데이터 / 변수 분류
3. 데이터 분석 순서
4. 관련 프로그램

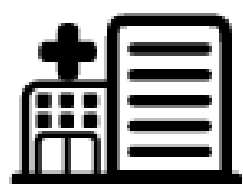


1. 분석 목적



데이터만 있다면 뭐든 할 수 있다!

빅데이터 등장 -> 다양한 분야의 데이터 플랫폼이 존재



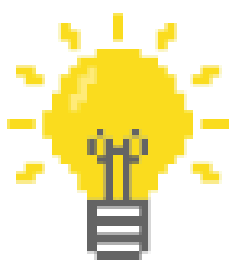
온라인 쇼핑몰의 판매량 예측
충무로역 배달 수요 예측
특정 조건을 가진 고객이 물건을 살지 말지 예측
게임의 승패 예측
고객의 성향과 니즈 파악, 맞춤 서비스 제공
연구 결과 해석
의약품의 효과 유의성 파악
병원 최적 장소 파악
손글씨 인식
보행자 및 장애물 인식
영상 조회수 예측



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



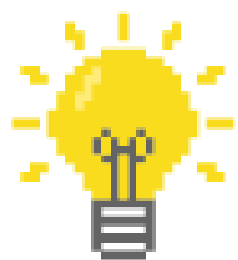
2. 데이터 / 변수 분류 및 분석 방법



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

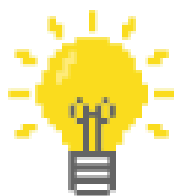
종속변수 ex) 매출액, 생존 여부
= 반응변수, target 변수

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

3. 데이터 분석 순서

데이터 수집	EDA (데이터 탐색)	데이터 전처리	분석 및 모델링	인사이트 도출
<div><ul style="list-style-type: none">- 주제 선정 / 변수 구체화- 데이터 수집 및 추출<p><수집 시 고려사항> 데이터 개수, 필요성, 신뢰성</p><p>도메인 지식 사전 조사 필수</p></div>	<div><ul style="list-style-type: none">- 데이터 기본 정보 확인- 결측치 및 이상치 확인- 변수 간의 상관관계 확인<p>- 데이터 시각화 필수</p><p>- 가설 검정</p><p>간단한 인사이트를 얻을 수 있음</p></div>	<div><ul style="list-style-type: none">- 결측/이상치, 중복값 처리- 데이터 연계, 통합- 변수 선택 및 변환- 파생 변수 생성<p>새롭게 생성된 데이터 -> 새로운 EDA 필요</p></div>	<div><ul style="list-style-type: none">- 통계 분석 및 모델링- 머신러닝/딥러닝 이용- 패턴 인식- 유의미한 결과 도출- 성능을 높임<p>머신러닝 및 딥러닝 기법</p></div>	<div><p>최종 목적은 결국 의사결정</p><p>결과 요약 시각화, 스토리텔링 능력 필요</p></div>

데이터 엔지니어

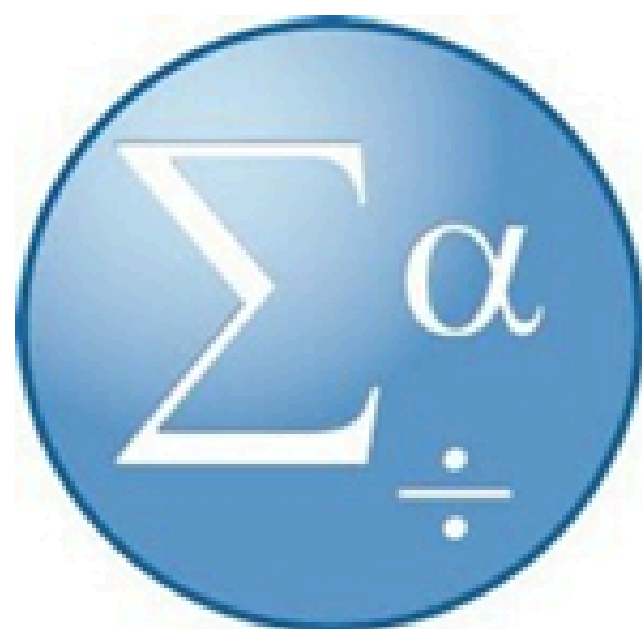
서베이 리서치

데이터 분석가 (기획자적 성향)

머신러닝 엔지니어 (개발자적 성향)

데이터 사이언티스트 (연구적 성향)

4. 관련 프로그램



데이터 모델링

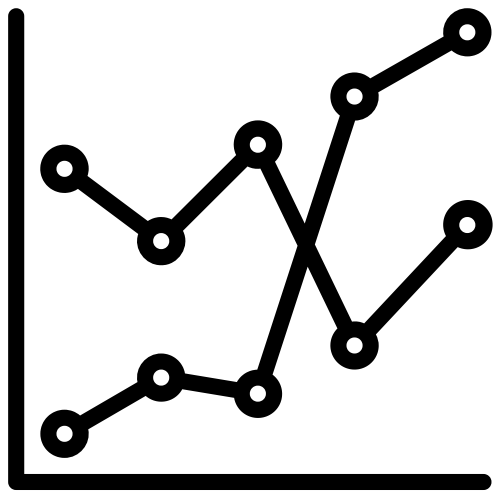
1. 데이터 모델링의 목적
2. 통계와 데이터마이닝의 차이
3. 데이터마이닝 모델링



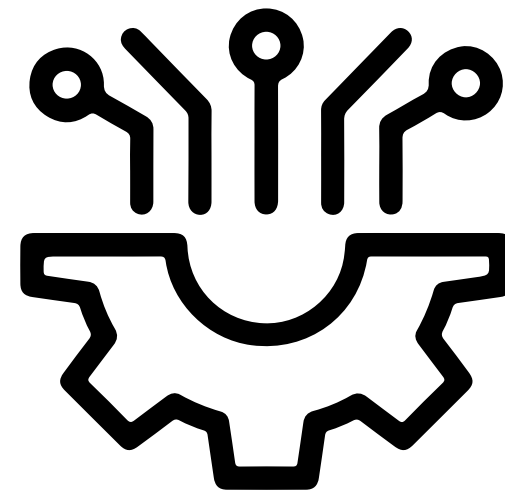
1. 데이터 모델링의 목적

데이터에서 가치를 창출하고, 의사결정을 지원하는 핵심 도구

- 데이터를 통해 최적의 결정을 내릴 수 있도록 지원
- 비즈니스 문제 해결을 위한 데이터 활용 기법
- 과거 데이터를 기반으로 미래 예측, 패턴 발견



통계 기반 모델

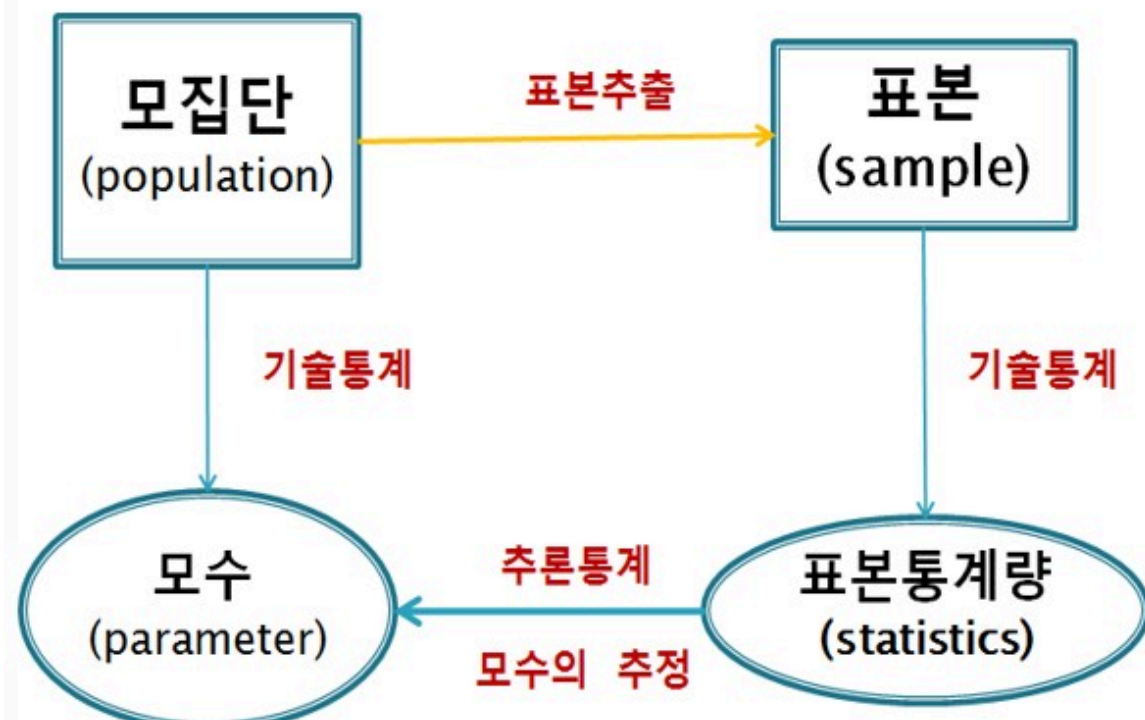


마이닝 기반 모델

전통적인 통계

💡 모집단과 표본의 개념이 중요
-> 샘플링을 통해 모집단을 추론

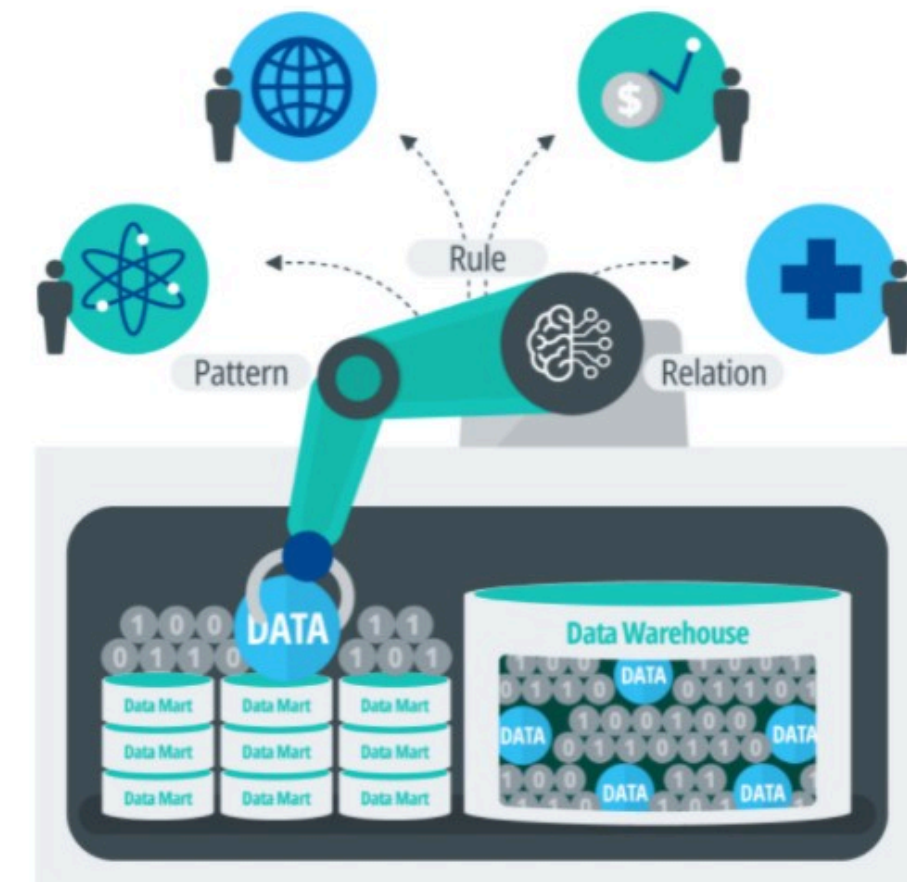
- 가정(Assumption) 설정
 - (정규분포/선형성/등분산성 등)
- 가설에 대한 검증이 목적



데이터 마이닝

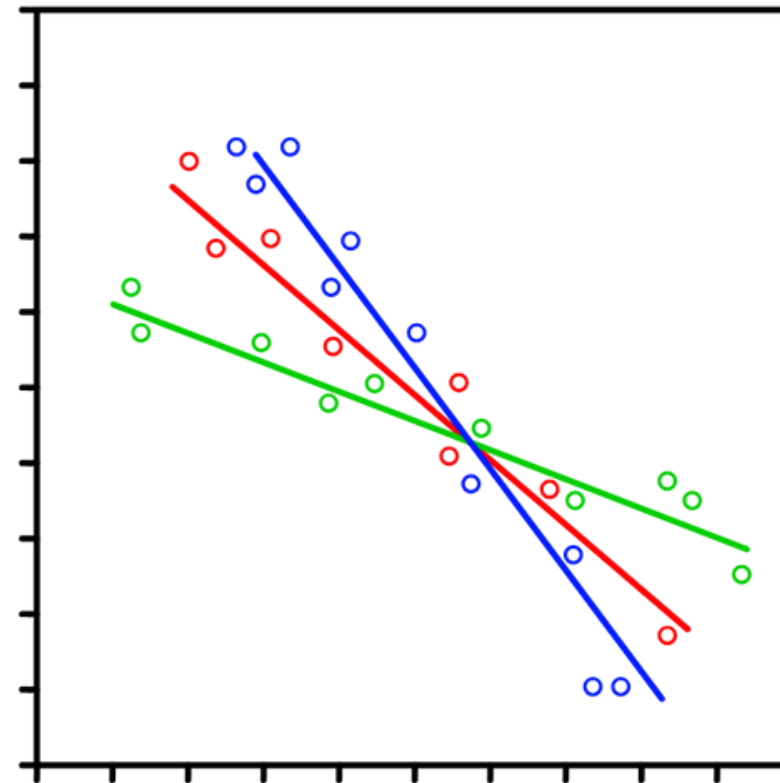
💡 모집단과 표본의 개념이 덜 중요!
-> 데이터 전체를 가지고 분석

- 데이터에 대한 가정 X
- 비선형성에 기반을 둔 알고리즘
- 복잡한 데이터 예측 및 분류를 위한 작업



'통계'를 공부하는 이유

데이터 마이닝 모델링 기법들은 "통계"를 기반으로 구축된 모델링이 많다.
더불어 최적화를 위해 통계를 기반으로 하는 기법을 사용하는 모델링도 많다.
대표적으로 선형회귀부터 딥러닝까지



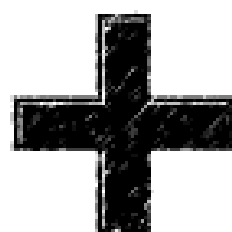
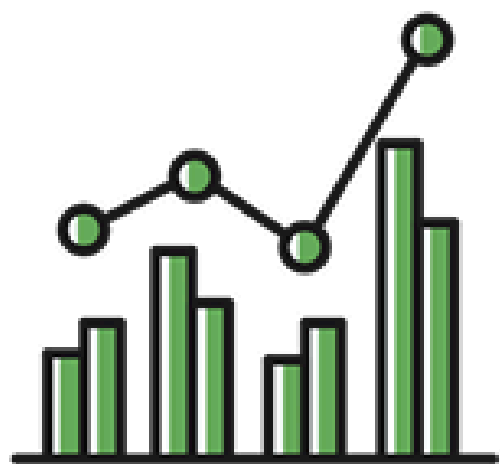
데이터마이닝 모델링

- 다양한 모델링 기법 존재 이유

- 1) 각각의 기법들이 서로 다른 목적과 장/단점을 가지고 있음
- 2) 기법의 유용성은 데이터 크기, 데이터 형태, 데이터에 존재하는 패턴의 유형, 기법이 요구하는 기본 가정 충족 여부, 분석 목적등 다양한 요인들에 의해 영향을 받음

알고리즘	특정 모델링 기법 (EX.분류나무, 판별분석 등)을 실행하기 위해 사용되는 특정 절차
독립변수	보통 X로 표기되며 속성, 특성, 예측변수, 입력변수
종속변수	보통 Y로 표기되며 예측되는 변수, 반응 변수, 목표변수
변수	입력변수(X)와 출력변수(Y)를 모두 포함하는 레코드의 측정치
차원	(독립)변수의 개수

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/딥러닝 이란?

인간의 학습능력을 컴퓨터가 갖게 환경을 만들어 주고
학습시키는 것!



데이터 분류 및 분석 방법



지도학습 : 정답이 있는 데이터

정답을 맞추는 것이 중요

[분류 종속변수가 질적 변수
회귀 종속변수가 양적 변수



비지도학습 : 정답이 없는 데이터

패턴&형태를 찾아내 의미를 부여하는 것이 중요
상황과 목적에 맞는 분석 방법 채택

ex) 클러스터링(군집화), 차원축소



강화학습 : 높은 점수를 낼 때마다 보상 제공 ex) DQN, A3C

대량의 학습 데이터 필요, 최적의 방법을 찾는 것이 중요

데이터마이닝 모델링 종류

지도학습 - 분류

- 특정 조건 가진 고객이 물건을 살지 말지 예측
- 게임의 승패 예측
- 의약품 효과 유의성 파악

지도학습 - 회귀

- 온라인 쇼핑몰 판매량 예측
- 충무로역 배달 수요 예측

비지도학습

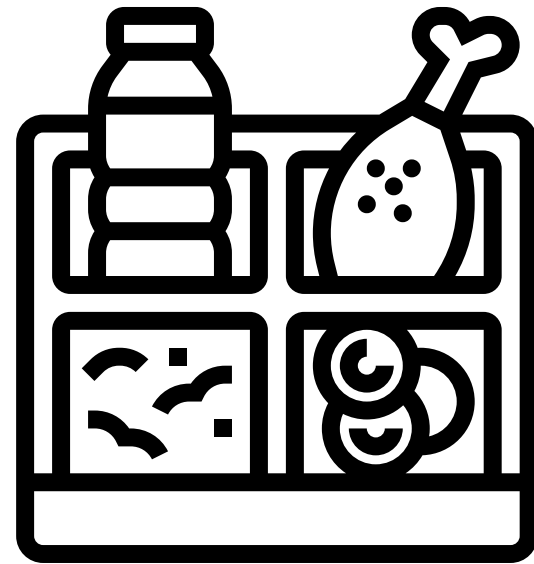
- 시니어 맞춤 여행 상품 및 관광 코스 개발
- 다음 시즌에 유행할 색상과 스타일 분석
- 공장 관리 상황 파악
- 병원 최적의 장소 파악

강화학습

- 게임에서 승리 전략 짜기
- 자율 주행 자동차

지도학습 - 분류

특정 고객이 해당 음식을 먹을 지 여부 예측



구내식당 메뉴 선정

- 고객/메뉴 군집화
- 대상별 선호 메뉴 조사
- 식사 수요 예측
- 식자재 가격 조사/예측
- 음식 트렌드 조사

지도학습 - 회귀

식자재 가격 예측

요일/시간별 수요 예측

비지도학습

고객 성향에 따른 군집화

- 나이, 성별, 직업, 체질, 선호 음식 등

메뉴 별 군집화

- 맵기, 당도, 나트륨수치, 식자재 가격 등

최근 음식 트렌드 키워드 군집화

과제 안내 - Github

1. 과제 안내
2. 깃허브 제출 방법
3. 팀별 토의
4. 추가 안내 사항



1. 과제 안내

1. 관심 있는 주제와 이유를 간단히 쓰고, 관련 데이터를 찾아서 깃허브에 업로드 해주세요!
 -> 관심 있는 주제와 이유는 **"마크다운"**을 사용해서 (2)번 파일에 함께 제출해주세요.
 -> 링크로 업로드 해주셔도 되고, 직접 다운로드 후 파일을 첨부하셔도 됩니다.
2. "타이타닉 데이터"를 가지고 EDA + 전처리 후 ipynb 파일을 깃허브에 올려주세요. (복습)
 -> 코드에는 간단히 주석을 달아주세요.
 -> 마크다운을 활용해주셔도 좋습니다.
 -> 데이터셋은 노션에서 다운받으실 수 있습니다.
3. 네이버 부스트코스 - 프로젝트로 배우는 데이터사이언스 - "1. 분류모델 기초" 수강 (예습)

다음주 금요일 낮 12시까지 깃허브에 public으로 업로드 해주세요.

2. 깃허브 제출 방법 안내

1. 레포지토리 만들기

-> 깃허브 레포지토리 이름 : [BAF-16-Fresh-Edu](#)

2. 관심 있는 주제, 이유 / 데이터 및 EDA&전처리 과제 ipynb 파일 첨부하기

3. 팀별 토의

팀별 독방을 개설해 주세요! (추후 원활한 토의를 위함입니다.)
오늘 다뤘던 내용 / 과제 관련하여 12시까지 자유롭게 토의해주세요.

1조 : 김현, 안재혁, 최연식

2조 : 서정유, 유영우, 지승우

3조 : 김민정, 박서연, 이용혁, 함주헌

4. 추가 안내 사항



신입 세션 우수상

모델 성능기준으로
2명에게 수상



과제 불성실

과제 불성실 인원에게
나머지 학습부여



가독성 유의

마크다운을 활용하여
가독성 좋은 코드 작성

감사합니다



B . a . f