



유튜브 강의로 완벽 정복

IT 전문가가 직접 정리한

# 빅데이터 분석기사 필기 완벽대비 요약노트

시험에 나오는 핵심내용  
+ 저절로 외워지는 쉬운 암기법



아답터

IT의 답을 터득하다

- 아답터(민 기술사)

개정 버전 : 2025.03.07



# 빅데이터 분석기사 필기 완벽대비 요약노트

이 자료의 내용은 아답터교육의 소중한 자산이며, 특정 단체나 커뮤니티, 학원에서  
의 허가 받지 않은 사용 및 공유 시 법적 처벌을 받을 수 있습니다.

# **1과목**

## **빅데이터 분석 기획**

## - 1 과목(빅데이터 분석 기획) 주요 내용

빅데이터의 이해	빅데이터 개요 및 활용	빅데이터의 특징
		빅데이터의 가치
		데이터 산업의 이해
		빅데이터 조직 및 인력
	빅데이터 기술 및 제도	빅데이터 플랫폼
		빅데이터와 인공지능
		개인정보 법·제도
		개인정보 활용
데이터분석 계획	분석방안수립	분석 로드맵 설정
		분석 문제 정의
		데이터 분석 방안
	분석 작업 계획	데이터 확보 계획
		분석 절차 및 작업 계획
데이터 수집 및 저장 계획	데이터 수집 및 전환	데이터 수집
		데이터 유형 및 속성 파악
		데이터 변환
		데이터 비식별화
		데이터 품질 검증
	데이터 적재 및 저장	데이터 적재
		데이터 저장

## 빅데이터의 이해

### 1. 빅데이터 개요 및 활용

#### - 빅데이터의 특징

##### ● 데이터의 정의

- 데이터 : 있는 그대로의 객관적 사실, 가공되지 않은 상태 (주문수량)
- 정보 : 데이터로부터 가공된 자료 (베스트셀러)

##### ● 빅데이터 출현 배경

- 인터넷 확산, 스마트폰 보급, 클라우드 컴퓨팅으로 인한 경제성 확보, 저장매체 가격하락, 하둡을 활용한 분산 컴퓨팅, 비정형 데이터 확산

##### ● 빅데이터 활용 위한 3대 요소

- 인력, 자원(데이터), 기술

☞ '인자기'

##### ● 빅데이터의 3V (가트너 정의)

- (1) Volume(규모) : 데이터 양 증가
- (2) Variety(다양성) : 데이터 유형 증가
- (3) Velocity(속도) : 데이터 생성, 처리 속도 증가
- (4) 그 외 5V에 포함되는 요소
  - Value(가치) : 숨겨진 가치 발견이 중요
  - Veracity(신뢰성) : 고품질 데이터

##### ● DIKW 피라미드

- (1) 데이터(Data) : 있는 그대로의 사실 (A대리점 핸드폰 100만원, B대리점 핸드폰 200만원)
- (2) 정보(Information) : Data를 통해 패턴 인식 (A대리점이 핸드폰이 싸다)
- (3) 지식(Knowledge) : 패턴을 통해 예측 (A에서 핸드폰을 사면 이득을 보겠다)
- (4) 지혜(Wisdom) : 창의적인 산물 (A대리점의 다른 기기들도 B대리점보다 저렴 할 것이다)

##### ● 암묵지, 형식지간 상호작용

- 암묵지 : 개인에게 습득되고 겉으로 드러나지 않음
- 형식지 : 문서, 매뉴얼 등의 형상화된 지식
  - 1) 공통화 : 암묵지 지식을 다른 사람에게 알려줌
  - 2) 표출화 : 암묵지 지식을 매뉴얼이나 문서로 전환
  - 3) 연결화 : 교재, 매뉴얼에 새로운 지식 추가
  - 4) 내면화 : 만들어진 교재, 매뉴얼에서 다른 사람의 암묵지를 터득

☞ '공표연내'

## ● 데이터베이스의 개념

- (1) DB : 일정 구조에 맞게 조직화된 **데이터의 집합**
  - **스키마** : DB의 구조와 제약조건에 관한 전반적 명세 (외부스키마, 개념스키마, 내부스키마)
  - 인스턴스 : 데이터 개체를 구성하는 속성에 대한 데이터 타입과 값
  - **메타데이터** : 데이터를 설명하는 데이터, 데이터 구조를 설명하고 검색하는데 활용
  - 인덱스 : 정렬, 탐색을 위한 데이터의 이름
- (2) DBMS : **DB를 관리**, 접근 환경 제공하는 소프트웨어
  - 1) 관계형 DBMS : 테이블(표)로 정리
  - 2) **NoSQL DBMS** : **비정형 데이터**를 저장하고 처리
- (3) SQL : 데이터 베이스에 접근할 수 있는 하부언어
  - 1) 정의언어(DDL) : CREATE, ALTER, DROP
  - 2) 조작언어(DML) : SELECT, INSERT, DELETE, UPDATE
  - 3) 제어언어(DCL) : COMMIT, ROLLBACK, GRANT, REVOKE

## ● 데이터베이스의 특징

- (1) 공유 데이터 : 여러 사용자가 **다른 목적**으로 데이터 **공동 이용**
  - (2) 통합된 데이터 : 동일한 데이터 **중복되어 있지 않음**
  - (3) 저장된 데이터 : **저장매체**에 저장
  - (4) 변화되는 데이터 : 새로운 데이터 추가, 수정, 삭제에도 **현재의 정확한 데이터 유지**
- ☞ **'공동저본'**

## ● 기업 활용 데이터베이스

- OLTP : 데이터를 수시로 갱신 (거래단위)
- OLAP : 다차원 데이터를 **대화식**으로 분석
- CRM : **고객**과 관련 자료 분석, 마케팅 활용
- SCM : **공급망 연결** 최적화
- ERP : 기업 경영 **자원**을 효율화
- RTE : 최신 정보로 **빠른** 의사결정 지원
- BI : 기업 보유 데이터 정리, 분석하는 **리포트** 중심 도구
- BA : 통계 기반 비즈니스 통찰력
- Block Chain : **네트워크**에 **참여**한 모든 사용자가 정보를 분산, 저장
- KMS : 기업의 모든 **지식**을 포함

## - 빅데이터의 가치

### ● 빅데이터 가치 산정이 어려운 이유

- (1) 특정 데이터를 언제, 어디서, 누가 활용할지 알 수 없음
- (2) 기존에 가치 없는 데이터도 새로운 분석기법으로 **가치를 창출**

## ● 빅데이터가 만들어내는 변화

- (1) 표본조사 → **전수조사**
- (2) 사전처리 → **사후처리**
- (3) 질 → **양**
- (4) 인과관계 → **상관관계**

☞ '전후양상'

## - 데이터 산업의 이해

### ● 데이터 산업의 발전

- 처리 → 통합 → 분석 → 연결 → 권리

- 1) 처리 : 프로그래밍 언어를 활용한 데이터의 처리
- 2) 통합 : DBMS의 등장
- 3) 분석 : 빅데이터 분석 기술의 발전
- 4) 연결 : API를 활용한 모듈들의 연결
- 5) 권리 : **마이데이터(MyData)**를 활용한 데이터의 주권 행사

\* 마이데이터 : 자신의 신용 정보를 다른 제3자에게 제공하여 서비스를 제공받는 제도

### ● 데이터 사이언스

- 데이터와 관련된 모든 분야의 전문지식을 종합한 학문
- 정형/비정형 데이터를 막론하고 데이터를 분석 (**총체적 접근법**)

### ● 데이터 사이언스 핵심 구성요소

- (1) Aalytics : 이론적 지식
- (2) IT : 프로그래밍적 지식
- (3) Business 분석 : 비즈니스적 능력

☞ '**AI비**'

### ● 데이터 사이언티스트의 필요역량

- (1) **하드 스킬(Hard Skill)** : 이론적 지식(수학, 통계학, 가설검정 등), 가트너 제시 역량에 미포함
- (2) **소프트 스킬(Soft Skill)** : 스토리텔링, 리더십, 창의력, 분석 등

☞ **하드스킬은 이과적, 소프트 스킬은 문과적인 느낌**

### ● 하둡(Hadoop)

- **여러 컴퓨터를 하나로 묶어** 대용량 데이터를 처리하는 오픈 소스 빅데이터 솔루션

- (1) 하둡 코어프로젝트

- 1) **HDFS** : 분산 파일 시스템

## 2) MapReduce

- 분산된 데이터를 병렬로 처리
- 클라이언트, 잡 트래커, 태스크 트래커로 구성
- 패턴 종류 : 조인 패턴, 그룹화 패턴, 단어카운트 패턴, 통계 패턴, 필터링 패턴

## (2) 하둡 에코시스템

- 다양한 서브 프로젝트들의 모임

1) 분산 코디네이터 : Zookeeper

2) 분산 리소스관리 : YARN, Mesos

3) 데이터 저장 : HBase, HDFS, Kudu

4) 데이터 수집 : Chukwa, Flume, Scribe, Kafka

5) 데이터 처리 : Pig, Mahout, Spark, Impale, Hive, MapReduce

## ● 데이터 단위

- KB < MB < GB < TB < PB < EB < ZB < YB (Peta < Exa < Zetta < Yotta)

☞ '패지요!'

## ● 빅데이터 가치 패러다임 변화

- Digitalization → Connection → Agency

(1) Digitalization : 아날로그 세상을 디지털화

(2) Connection : 디지털화된 정보들의 연결

(3) Agency : 연결을 효과적으로 관리

☞ 'DigitalCA메라'

## - 빅데이터 조직 및 인력

### ● 조직 및 인력방안 수립 (DSCoE : 분석조직)

- 집중 구조 : 독립적인 전담 조직 구성 (중복 업무 가능성 존재)
- 기능 구조 : 해당 부서에서 직접 분석 (DSCoE가 없음)
- 분산 구조 : 분석 조직 인력을 현업 부서에 배치

☞ '집기분'

## 2. 빅데이터 기술 및 제도

### - 빅데이터

### ● 빅데이터 플랫폼

- 데이터의 수집부터 저장, 처리, 분석 등의 파이프라인 전 과정을 통합적으로 제공하는 환경



## ● 빅데이터 플랫폼의 계층 구조

### (1) 소프트웨어 계층

- 데이터 수집 및 정제, 데이터 처리 및 분석, 사용자/서비스 관리

### (2) 플랫폼 계층

- 데이터 및 자원의 관리, 작업 스케줄링, 프로파일링

### (3) 인프라스트럭처 계층

- 자원의 배치 및 관리, 저장장치 및 네트워크 관리

☞ 소프트웨어가 상위 계층, 인프라스트럭처가 하위 계층

## - 빅데이터와 인공지능

## ● 인공지능, 머신러닝, 딥러닝의 관계

- 딥러닝 < 머신러닝 < 인공지능

## ● 머신러닝의 종류

- 지도학습 : 정답을 알려주고 학습시키는 방법
- 비지도학습 : 정답을 가르쳐주지 않고 학습시키는 방법
- 준지도학습 : 정답이 있는 데이터와 정답이 없는 데이터를 모두 활용하는 방법
- 강화학습 : 에이전트가 보상을 받기 위해 학습하는 방법

## ● 약인공지능, 강인공지능

- 약인공지능 : 주어진 조건에서만 동작하는 인공지능
- 강인공지능 : 인간과 동일한 사고가 가능한 인공지능
- 초인공지능 : 기술적 특이점을 뛰어넘어 인간을 초월한 인공지능

## ● 경량 딥러닝 학습 기법

- 전이학습 : 사전에 훈련된 모델을 재사용하는 학습 방식
- Fine-Tuning : 학습된 모델을 특정 타겟에 맞게 재조정하는 방법
- 지식증류 : Teacher Network의 지식을 Student Network에 전달하는 방법

## ● 최신 인공지능 기술동향

- (1) AutoML : 머신러닝 프로세스를 자동화
- (2) MLOps : DevOps를 머신러닝 프로세스에 적용
- (3) XAI(eXplainable AI) : 설명가능한 인공지능 기술
- (4) 생성형 AI : 학습한 데이터를 바탕으로 새로운 콘텐츠를 생성하는 기술
  - 텍스트 생성(ChatGPT, BERT), 이미지 생성(Midjourney, Stable Diffusion), 비디오 생성
  - LLM(Large Language Model) : 수십억 개의 파라미터가 학습된 거대 언어 모델
  - Diffusion Models : 확률적 프로세스를 이용하여 새로운 데이터를 생성하는 방법

## - 개인정보 법제도

### ● 데이터 3법

- (1) 개인정보보호법
- (2) 정보통신망 이용 촉진 및 정보보호 등에 관한 법률(정보통신망법)
- (3) 신용정보의 이용 및 보호에 관한 법률(신용정보법)

☞ '개정판'

### ● 데이터 3법 주요 특징

- **가명정보**의 개념 도입 (통계 작성, 연구, 공익적 기록보존 목적 하에 **동의 없이 활용 가능**)
- 개인정보보호 거버넌스 체계 효율화
- 개인정보처리자 책임 강화
- 개인정보의 판단기준 명확화

### ● 개인정보, 가명정보, 익명정보

- (1) 개인정보 : 개인을 알아볼 수 있는 정보, 동의를 받아 활용 가능 (홍길동, 33세)
- (2) 가명정보 : 가명처리를 통해 추가정보 없이 특정 불가 (홍OO, 30대 초반)
- (3) 익명정보 : 더 이상 개인을 알아볼 수 없는 정보, **제한 없이 자유롭게 활용** (OOO, 30대)

### ● 개인정보 비식별 조치 가이드라인

- 사전검토 → 비식별조치(총계,삭제,마스킹) → 적정성평가(k-익명성,l-다양성,t-근접성) → 사후관리

## - 개인정보 활용

### ● 위기 요인과 통제방안

- (1) 사생활 침해 : SNS 올린 데이터가 사생활 침해  
→ 제공자에서 **사용자 책임**으로 전환
- (2) 책임 원칙 훼손 : 범죄 예측 프로그램으로 예측하여 체포하는 문제  
→ **결과에 대해서만 책임**
- (3) 데이터의 오용 : 분석 결과가 항상 옳은 것은 아님  
→ 알고리즘을 해석가능한 **알고리즘미스트** 필요  
\* 알고리즘미스트 : 부당하게 피해가 발생한 사람들을 **구제**하는 전문인력

## 데이터분석 계획

### 1. 분석방안수립

#### - 분석 로드맵 설정

##### ● 분석 대상과 방법

- 4가지 유형을 **넘나들며 분석**을 수행

대상 방법	Known	UnKnown
Known	<b>최적화</b> (Optimization)	<b>통찰</b> (Insight)
Un-Known	<b>솔루션</b> (Solution)	<b>발견</b> (Discovery)

##### ● 분석 기획 방안

	과제 중심적 접근	장기적 마스터 플랜
목적	<b>빠르게</b> 해결	<b>지속적 분석 원인</b> 해결
1차 목표	Speed & Test	Accuracy & Deploy
과제유형	Quick & Win	Long Term View
접근방식	Problem Solving	Problem Definition

##### ● IT 프로젝트의 우선순위 선정 기준

- (1) 전략적 중요도 : 전략적 필요성, 시급성
- (2) 실행 용이성 : 투자 용이성, 기술 용이성

##### ● 데이터 분석 프로젝트의 우선순위 선정 기준

- (1) **시급성** 관점 : 비즈니스 효과(Return), **KPI**(핵심성과지표) - **Value**
- (2) **난이도** 관점 : 투자비용 요소(Investment) - Volume, Variety, Velocity

(어려움)	1	2
난이도		
(쉬움)	3	4
	(현재)	시급성 (미래)

- 시급성 중요시 : 3 → 4 → 2

- 난이도 중요시 : 3 → 1 → 2

☞ 3과 2는 앞 뒤로 고정하고 가운데만 변경

## ● 의사결정을 가로막는 요소

- 고정 관념, 편향된 생각
- **프레임링 효과** : 동일 상황임에도 개인의 판단, 결정이 달라짐

## - 분석 문제 정의

### ● 하향식 접근 방법

- 문제가 주어지고 해답을 찾기 위해 진행
- **문제 탐색** → 문제 정의 → 해결방안 → **타당성 검토**
- (1) **문제탐색**
  - 1) 빠짐없이 문제를 도출하고 식별하며, 솔루션 초점 보다는 가치에 초점
  - 2) 비즈니스 모델 캔버스 단순화 측면 : **업무, 제품, 고객, 규제와 감사, 지원인프라**  
 ☞ **'지원인프라 업무 중에 고객이 제품을 규제와 감사 했다.'**
  - 3) **관점**
    - 거시적 관점 : **STEEP**(사회, 기술, 경제, 환경, 정치)
    - 경쟁자 확대 관점 : 대체자, 경쟁자, 신규 진입자
    - 시장의 니즈 탐색 관점 : 고객, 채널, 영향자
- (2) **문제 정의**
  - 비즈니스 문제를 데이터 문제로 변환하여 정의
- (3) **해결 방안**
  - 기존 시스템 활용, 시스템 고도화, 인적 자원 확보, 아웃소싱 등
- (4) **타당성 검토**
  - 경제적 타당성 : **비용대비 편익** 분석관점 접근
  - 데이터 타당성 : **데이터 존재여부, 분석역량**이 필요
  - 기술적 타당성 : 역량 확보 방안 사전에 수립

### ● 상향식 접근 방법

- 문제 정의 자체가 어려울 때, 사물을 그대로 인식하는 **What 관점**
- 주로 **비지도 학습** 활용

### ● 혼합 접근 방법

- (1) 발산 단계 : **상향식 접근 방법**으로서, 가능한 방안들을 도출
- (2) 수렴 단계 : **하향식 접근 방법**으로서, 도출된 방안들을 분석

### ● 디자인 싱킹

- 사용자에게 **공감**으로 시작해서 아이디어 **발산/수렴** 과정을 통한 **피드백**으로 발전하는 과정
- 공감하기 → 문제정의 → 아이디어 도출 → 프로토타입 → 테스트

## - 데이터 분석 방안

### ● 분석 방법론의 구성요소

- 절차, 방법, 도구와 기법, 템플릿과 산출물

### ● 분석 과제에서 고려해야할 5가지 요소

- 데이터 크기, 속도, 데이터 복잡도, 분석 복잡도, 정확도/정밀도
  - \* 정확도(Accuracy)와 정밀도(Precision)는 Trade-Off 관계
  - ☞ 여기에서의 정확도와 정밀도는 4과목의 오분류표에서의 평가지표와는 다른 개념

### ● 프로젝트 관리 지식 체계 10가지 영역

- 통합, 범위, 시간(일정), 원가, 품질, 인적자원, 의사소통, 리스크(위험), 조달(아웃소싱), 이해관계자
  - ☞ '이범통이 의자에서 시원한 조리품을 먹었다.'

### ● 분석 방법론 모델

- (1) 계층적 프로세스 모델 : 단계(Baseline으로 관리) → 태스크 → 스텝(단기간 수행 WorkPackage)
- (2) 폭포수 모델 : 이전 단계 완료되어야 다음 단계 진행 (Top-Down)
- (3) 나선형 모델 : 여러 개발과정 거쳐 점진적으로 완성, 위험요소 제거 초점
- (4) 프로토타입 모델 : 일부분(프로토타입)을 우선 개발하고 보완
- (5) 반복적 모델
  - 증분형 모형 : 전체 시스템을 작은 기능 단위로 나누어 개발
  - 진화형 모형 : 핵심 부분을 개발한 후 요구사항을 반영하여 진화
- (6) 애자일 : 짧은 개발 주기를 가지고 고객 피드백을 지속적으로 반영하여 반복적인 개발

### ● KDD 분석 방법론

- 데이터선택 → 전처리 → 변환 → 마이닝 → 결과 평가
  - 1) 데이터선택 : 원시데이터(Raw Data)나 DB에서 필요한 데이터 선택
  - 2) 전처리 : 이상값, 잡음 식별 및 데이터 가공
  - 3) 변환 : 변수 선택 및 차원축소
  - 4) 마이닝 : 알고리즘을 선택하여 분석 수행
  - 5) 결과 평가 : 결과에 대한 해석, 결과가 충족되지 않으면 절차를 반복 수행

### ● Crisp-DM 분석 방법론

- 업무 이해 → 데이터 이해 → 데이터 준비 → 모델링 → 평가 → 전개
  - ☞ '업데데이트모델평가전'
  - 1) 업무 이해 : 업무 목적 파악, 상황파악, 목표 설정, 프로젝트 계획 수립
  - 2) 데이터 이해 : 초기 데이터 수집, 기술 분석, EDA, 데이터 품질 확인

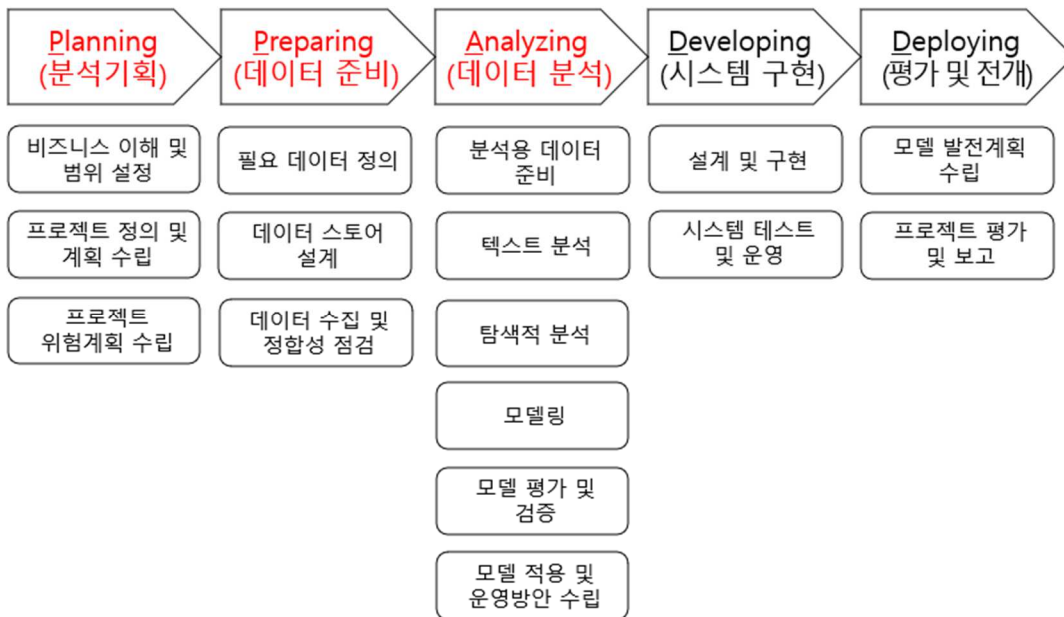
- 3) 데이터 준비 : 데이터 셋 선택 및 정제, 통합
  - 4) 모델링 : 모델링 기법 선택, 테스트 계획 설계, **모델 작성 및 평가**
  - 5) 평가 : **분석결과 평가, 모델링 과정 평가, 모델 적용성 평가**
  - 6) 전개 : 전개 계획, 모니터링 및 유지보수 계획 수립, 프로젝트 종료 보고서 작성, 프로젝트 리뷰
- \* 평가 → 전개에서 **위대한 실패**(업무 이해로 다시 돌아감) 발생 가능

## ● SEMMA 분석 방법론

- Sample → Explore → Modify → Model → Assess

- 1) Sample : 분석 대상 데이터 추출
- 2) Explore : 탐색하고 오류 확인
- 3) Modify : 데이터의 변환
- 4) Model : 알고리즘 적용
- 5) Assess : 모델의 평가 및 검증

## ● 빅데이터 분석 방법론



### ☞ 'PPADD'

#### 1) 분석 기획

- 프로젝트 위험계획 수립 : **회피, 전이, 완화, 수용**

### ☞ '회전완수'

#### 2) 데이터 준비

- 데이터 스토어 설계 : 정형, 비정형, 반정형 데이터에 따른 효율적 저장소를 설계

#### 3) 데이터 분석

- 분석용 데이터 준비 : 추가적인 데이터 확보 필요 시, 데이터 준비 단계로 다시 진행
- 모델링 : 알고리즘 설명서는 상세히 작성
- 모델 평가 및 검증 : 성능이 저조한 모델은 튜닝 작업 수행

## ● 분석 거버넌스 체계 구성요소

- 조직, 프로세스, 시스템, 데이터, 분석관련 교육 및 마인드 육성체계

☞ ‘시조프로마인드데’

## ● 데이터 분석 수준 진단

### (1) 분석 준비도

분석적 업무판악	인력 및 조직	분석기법
발생한 사실 분석업무 예측분석 업무 시뮬레이션 분석업무 분석업무 정기적 개선	분석전문가 직무 존재 분석전문가 교육훈련 프로그램 관리자들의 기본적 분석능력 전사 분석업무 총괄 조직 존재 경영진 분석업무 이해능력	업무별 적합한 분석기법 사용 분석업무 도입 방법론 분석기법 라이브러리 분석기법 효과성 평가 분석기법 정기적 개선
분석 데이터	분석 문화	IT 인프라
분석업무를 위한 데이터 충분성 분석업무를 위한 데이터 신뢰성 분석업무를 위한 데이터 적시성 비구조적 데이터 관리 외부 데이터 활용 체계 마스터데이터 관리(MDM)	사실에 근거한 의사결정 관리자의 데이터 중시 회의 등에서 데이터 활용 경영진의 직관보다 데이터 데이터 공유 및 협업 문화	운영시스템 데이터 통합 EAL, ETL 등 데이터 유통체계 분석전용 서버 및 분석환경 빅데이터 분석환경 통계분석 환경 비주얼분석 환경

☞ ‘IT문데기인파’

### (2) 분석 성숙도

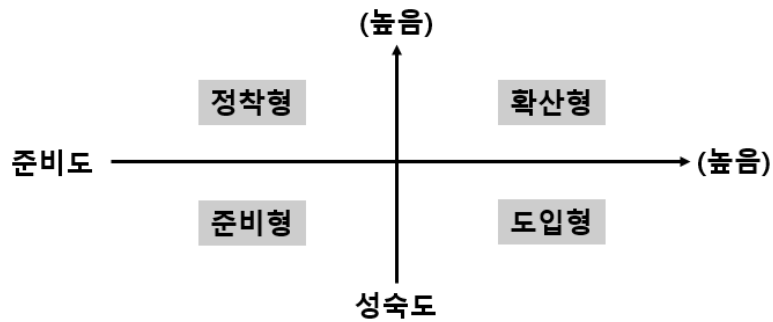
\* CMMI 모델 기반 (1~5단계)

- 1) 도입 : 환경, 시스템 구축
- 2) 활용 : 업무에 적용
- 3) 확산 : 전사 차원 관리, 공유
- 4) 최적화 : 혁신, 성과향상에 기여

☞ ‘도활확최’

단계	비즈니스 부문	조직 및 역량부분	IT 부문
도입	실적 분석 및 통계 정기 보고 수행 운영 데이터 기반	일부 부서에서 수행 담당자 역량에 의존	데이터 웨어하우스 데이터 마트 ETL/EAI OLAP
활용	미래결과 예측 시뮬레이션 운영 데이터 기반	전문담당부서 수행 분석 기법 도입 관리자가 분석 수행	실시간 대시보드 통계분석 환경
확산	전사성과 실시간 분석 프로세스 혁신 3.0 분석규칙 관리 이벤트 관리	전사 모든 부서 수행 분석 COE 운영 데이터 사이언티스트 확보	빅데이터 관리 환경 시뮬레이션/최적화 비주얼 분석 분석 전용 서버
최적화	외부 환경 분석 활용 최적화 업무 적용 실시간 분석 비즈니스 모델 진화	데이터 사이언스 그룹 경영진 분석 활용 전략 연계	분석 협업 환경 분석 SandBox 프로세스 내재화 빅데이터 분석

## ● 데이터 분석 성숙도 모델



- 1) **준비형** : 낮은 준비도, 낮은 성숙도
  - 데이터, 인력, 조직, 분석업무, 분석기법 적용 안되어 사전 준비 필요
- 2) **정착형** : 낮은 준비도, 높은 성숙도
  - 인력, 조직, 분석업무, 분석기법 등을 제한적으로 사용
- 3) **도입형** : 높은 준비도, 낮은 성숙도
  - 조직 및 인력 등 준비도는 높으나, 분석업무 및 기법 부족
- 4) **확산형** : 높은 준비도, 높은 성숙도
  - 6가지 분석 구성요소가 모두 갖추고 있으며, 지속적 확산이 가능
  - ☞ '도준정확' 4사분면부터 시계방향(역순)으로 암기

## ● 분석 지원 인프라 방안 수립

- 확장성을 고려한 플랫폼 구조 적용 (중앙집중적 관리)
- (1) 분석 플랫폼 구성요소
  - 1) 광의의 분석 플랫폼 : 분석 서비스 제공엔진, 분석 어플리케이션, 분석 서비스 API, 하드웨어, OS
  - 2) 협의의 분석 플랫폼 : 데이터 처리 프레임워크, 분석엔진, 분석 라이브러리
 ☞ 광의의 분석 플랫폼은 협의의 분석 플랫폼 요소들을 포함하는 개념

## ● 데이터 거버넌스

- (1) 데이터 거버넌스
  - 전사 차원에서 데이터 대해 표준화된 관리 체계 수립
  - 1) 구성요소 : 원칙, 조직, 프로세스
    - ☞ '원조프'
  - 2) 중요 관리대상
    - 마스터 데이터 : 자료 처리에 기준이 되는 자료
    - 메타데이터 : 다른 데이터를 설명해 주는 데이터
    - 데이터 사전 : DB에 저장된 정보를 요약
- (2) 데이터 거버넌스 체계
  - 1) 데이터 표준화 : 메타데이터 및 사전 구축
  - 2) 데이터 관리 체계 : 효율성을 위함
  - 3) 데이터 저장소 관리 : 저장소 구성
  - 4) 표준화 활동 : 모니터링, 표준 개선 활동



● 빅데이터 거버넌스

- 데이터 거버넌스 체계 + 빅데이터 효율적 관리, 데이터 최적화, 정보보호, 데이터 카테고리 별 관리책임자 지정 등을 포함

## 데이터 수집 및 저장 계획

### 1. 데이터 수집 및 전환

#### - 데이터 수집

##### ● 데이터 수집 기술

- (1) ETL : **Extraction, Transformation, Load** 3단계를 통하여 DB 시스템에 데이터 적재
- (2) FTP : TCP/IP 네트워크에서 **컴퓨터들 간의 파일을 교환**하기 위한 통신 규약
- (3) API : 응용 프로그램에서 다른 프로그램으로 데이터를 전송할 수 있는 인터페이스
- (4) 아파치 스콧(Sqoop) : RDBMS와 하둡간의 대용량 데이터를 전송하는 도구
- (5) 아파치 플럼(Flume) : 대량의 로그 데이터를 효율적으로 수집, 전송하는 분산형 서비스 시스템
- (6) 웹 크롤링 : **웹 상의 데이터를 탐색하고 수집**하는 기법

#### - 데이터 유형 및 속성 파악

##### ● 데이터의 유형

- (1) 정성적, 정량적
  - 정량적 데이터 : 자료를 수치화 - 수치, 기호 (온도, 풍속)
  - 정성적 데이터 : 자료의 특징을 풀어 설명 - 언어, 문자 (기상특보, 주관식 설문 응답)
- (2) 정형, 반정형, 비정형
  - 정형 데이터 : 정보 형태가 정해짐 (관계형DB, 엑셀-스프레드시트, CSV)
  - **반정형 데이터** : 데이터를 설명하는 **메타데이터**를 포함 (HTML, XML, JSON, RDF)
  - 비정형 데이터 : 형태가 정해지지 않음 (SNS, 유튜브, 음원)
- (3) 가역, 불가역 데이터
  - 가역 데이터 : 원본으로 일정 수준으로 복원이 가능
  - 비가역 데이터 : 원본으로 복원이 불가능
- (4) 내부, 외부 데이터
  - 내부 데이터 : 운영, 비즈니스 활동, 고객 상호작용 등에서 발생하는 조직 내부에서 생성된 데이터
  - 외부 데이터 : 외부 기관이나 정부에서 데이터를 수집해야 하며, 검증이나 가공이 필요

##### ● 데이터의 척도 구분

- (1) 질적 척도
  - 명목척도 : 어느 집단에 속하는지 나타내는 자료 (대학교, 성별)
  - 순서척도(서열척도) : 서열관계가 존재하는 자료 (학년, 순위)
- (2) 양적 척도
  - 등간척도(구간척도) : **구간 사이 간격이 의미**가 있으며 덧셈과 뺄셈만 가능 (온도, 지수 등)
  - 비율척도 : **절대적 기준 0이 존재**하고 사칙연산 가능한 자료 (무게, 나이 등)

## - 데이터 비식별화

### ● 개인정보 비식별화

- (1) 가명처리 (홍길동, 35세 → 임걱정, 30세)
  - 휴리스틱 가명화(사람의 판단), 암호화, 교환 방법
- (2) 총계처리 (홍길동 170cm, 임걱정 180cm → 평균 키 175cm)
  - 부분총계, 라운딩(올림, 내림, 반올림), 재배열(타인 정보와 섞어 전체 정보 손상 없음)
- (3) 데이터 삭제 (주민등록번호 901111-1234567 → 90년대 생, 남자)
  - 식별자 삭제, 식별자 부분삭제, 레코드 삭제, 식별요소 전부삭제
- (4) 데이터 범주화 (홍길동, 35세 → 홍길동, 30~40세)
  - 감추기(평균, 범주로 변환), 랜덤 라운딩, 범위 방법, 제어 라운딩
- (5) 데이터 마스킹 (홍길동, 35세 → 홍OO, 35세)
  - 임의 **잡음(노이즈)** 추가, 공백과 대체

### ● 프라이버시 보호 모델

- (1) k-익명성
  - 같은 값이 존재하도록 하여 다른 정보로 결합할 수 없도록 함
  - **연결공격**으로부터 보호
- (2) l-다양성
  - 민감한 정보의 다양성을 높여 추론 가능성을 낮춤
  - **동질성 공격**과 **배경 지식에 의한 공격**으로부터 보호
- (3) t-근접성
  - 민감 정보의 분포를 낮추어 추론 가능성을 더욱 낮춤
  - **쓸림 공격**이나 **유사성 공격**으로부터 보호

### ● 차등정보보호

- 개인 정보를 **다른 수많은 데이터와 조합(노이즈)**하여, 개인정보를 침해하지 않고 정보 패턴 발견
- **섭동**(오차를 활용한 교란), 이중지수분포 활용 등의 기법 적용

## - 데이터 품질 검증

### ● 데이터 품질 기준

- (1) 완전성 : 필요한 데이터가 빠짐없이 존재하는가?
- (2) 정확성 : 데이터가 실제 값이랑 일치하는가?
- (3) 일관성 : 여러 시스템에서 데이터가 동일한가?
- (4) 최신성 : 데이터가 최신 상태를 반영하는가?
- (5) 유효성 : 사전에 약속된 형식을 만족하는가?
- (6) 접근성 : 필요한 사용자가 쉽게 접근이 가능한가?
- (7) 보안성 : 데이터가 안전하게 보호되고 있는가?

## ● 데이터 품질 진단 및 개선 절차

- (1) **진단 절차** : 진단 대상 정의 → 품질 진단 실시 → 진단 결과 분석
  - 1) 진단 대상 정의 : 품질 진단 수요조사, 품질 진단 현황분석, 품질 진단 추진계획 수립
  - 2) 품질 진단 실시 : 품질 진단 준비, 품질 진단 수행
  - 3) 진단 결과 분석 : 오류 원인 분석, 업무 영향 분석, 개선 기회 도출
- (2) **개선 절차** : 개선 계획 수립 → 개선 수행 → 품질 통제
  - 1) 개선 계획 수립 : 품질개선 방향 정의, 품질개선 추진계획 수립
  - 2) 개선 수행 : 품질개선 준비, 품질개선 수행
  - 3) 품질 통제 : 결과 평가, 품질 목표 관리, 품질 통제 실시

## 2. 데이터 적재 및 저장

### ● 분산 파일 시스템

- (1) **HDFS**(하둡 분산파일 시스템)
  - 슬레이브 관리하는 마스터노드(=네임노드)와 데이터 처리하는 슬레이브노드(=데이터노드)로 구성
  - 64MB의 여러 블록으로 분산 저장
  - 분산 처리를 통하여 시스템의 과부하 및 병목 현상 해소
  - 오픈 소스로서 무료 사용 가능
- (2) GFS
  - 구글의 데이터 처리를 위해 설계된 분산 파일 시스템
  - 마스터(관리/통제), 청크 서버(물리 저장소 처리), 클라이언트(어플리케이션)로 구성
- (3) 그 외 파일 시스템
  - Ceph, 아마존 S3 등

☞ 분산 파일 시스템과 분산 데이터베이스는 구분되는 개념임

### ● 데이터베이스

- (1) 관계형 데이터베이스 : 정형 데이터의 처리, 데이터의 **무결성 보장** (MySQL, MariaDB, Oracle)
  - \* 무결성 : 데이터가 정확하고 신뢰할 수 있으며, 조작되지 않고 원본 상태로 보존되는 것
- (2) **NoSQL 데이터베이스** : 비정형 데이터의 처리, 확장성과 가용성에 이점
  - 1) 키-값 데이터베이스 : 가장 단순한 구조로, 속도가 빠름 (Redis, VoldeMorte)
  - 2) 열 데이터베이스 : 열 단위로 저장되므로, 압축 및 확장성에 유리 (Hbase, BigTable, Cassandra)
  - 3) 문서 데이터베이스 : 문서를 트리 구조로 저장하거나 검색 (MongoDB, CouchDB)
  - 4) 그래프 데이터베이스 : 노드로 표현되며 유연하고 유지보수 용이 (Neo4j, OrientDB)

### ● 저장소(스토리지)

- (1) DAS(Direct Attached Storage) : 서버와 저장소가 물리적으로 직접 연결
- (2) NAS(Network Attached Storage) : LAN 네트워크를 통해 여러 컴퓨터가 동일 저장소 공유
- (3) SAN(Storage Area Network) : 특수 목적용 고속 네트워크와 스위치를 통해 연결

## ● 병렬 DBMS

- 데이터 병렬 처리를 활용하여 성능 개선
- 공유 메모리, 공유 디스크, Shared Nothing 구조를 가짐
- 종류 : VoltDB, SAP HANA, Vertica 등

## ● 데이터웨어하우스, 데이터마트, 데이터레이크

- (1) 데이터 웨어하우스(DW)의 특징
  - 주제지향성 : 분석목적 설정이 중요
  - 데이터 통합 : 일관화 된 형식으로 저장
  - 시계열성 : 히스토리를 가진 데이터
  - **비휘발성** : 읽기전용 - 수시로 변하지 않음
- (2) 데이터 웨어하우스의 구성요소
  - ETL(Extraction, Transform, Load)
  - **ODS(Operational Data Store)** : 다양한 DBMS에서 추출한 데이터를 임시 저장
- (3) 데이터 마트(DM)
  - 데이터 웨어하우스의 한 분야로 **특정 목적을 위해 사용** (소규모 데이터웨어하우스)
- (4) 데이터 레이크
  - 비정형 데이터를 저장하며 하둡과 연계하여 처리

## 2과목 빅데이터 탐색

## - 2 과목(빅데이터 탐색) 주요 내용

데이터 전처리	데이터 정제	데이터 정제
		데이터 결측값 처리
		데이터 이상값 처리
	분석 변수 처리	변수 선택
		차원축소
		파생변수 생성
		변수 변환
		불균형 데이터 처리
데이터 탐색	데이터 탐색 기초	데이터 탐색 개요
		상관관계 분석
		기초통계량 추출 및 이해
		시각적 데이터 탐색
	고급 데이터 탐색	시공간 데이터 탐색
		다변량 데이터 탐색
		비정형 데이터 탐색
통계기법 이해	기술통계	데이터요약
		표본추출
		확률분포
		표본분포
	추론통계	점추정
		구간추정
		가설검정

## 데이터 전처리

### 1. 데이터 정제

#### - 데이터 정제

##### ● 데이터의 종류

- (1) 단변량 데이터 : 데이터의 특성이 하나인 데이터
- (2) 다변량 데이터 : 데이터의 특성이 두 개 이상인 데이터
- (3) **시계열 데이터** : 시간 순서에 따라 관측된 데이터 (=종단면적 데이터)
  - \* **종단면적 데이터**(여러 시점에 측정) ↔ **횡단면적 데이터**(한 시점에 측정) / 패널데이터(종단+횡단)

##### ● 데이터 정제

- (1) 집계 : 데이터를 요약 (합계, 평균, 분산, 개수, 최대/최소)
- (2) 일반화 : 데이터의 일반적인 특성 추출
- (3) 정규화 : 데이터를 정해진 구간으로 조정하여 상대적 차이 제거
- (4) **평활화** : 잡음을 제거하여 추세를 부드럽게 만듦 (이동평균법, 지수평활법)

#### - 데이터 결측값 처리

##### ● 결측치 종류

- 존재하지 않는 데이터, null/NA로 표시
- (1) 완전 무작위 결측(MCRA)
  - 다른 변수들과 아무런 상관없는 경우
  - 예) 입력 실수, 전산 오류
- (2) 무작위 결측(MAR)
  - 특정 변수와 관련되어 발생하였지만, 결과와는 관계가 없는 경우
  - 예) 특정 정치 성향 유권자들의 응답률이 낮으나 정당의 득표율이 낮은 것은 아님
- (3) 비 무작위 결측(MNAR)
  - 결측치가 변수의 결과에 상관이 있는 경우
  - 예) 소득이 낮은 응답자들의 응답률이 낮음

##### ● 결측값 처리

- (1) 완전분석법 : 결측값 가지는 데이터 삭제
- (2) 평균 대체법(=비조건부 평균 대체) : 단순 평균으로 대체
- (3) 회귀 대체법(=조건부 평균 대체) : 회귀분석의 결과로 대체
- (4) 단순 확률 대체법 : 확률적으로 선택하여 대체
  - Nearest Neighbor : 바로 가까운 응답으로 대체
  - Hot-Deck : 현재 데이터 셋에서 비슷한 성향으로 대체
  - Cold-Deck : 유사한 외부 출처에서 비슷한 성향으로 대체
- (5) 다중 대체법 : 여러 번 대체 (**대치** → **분석** → **결합**)



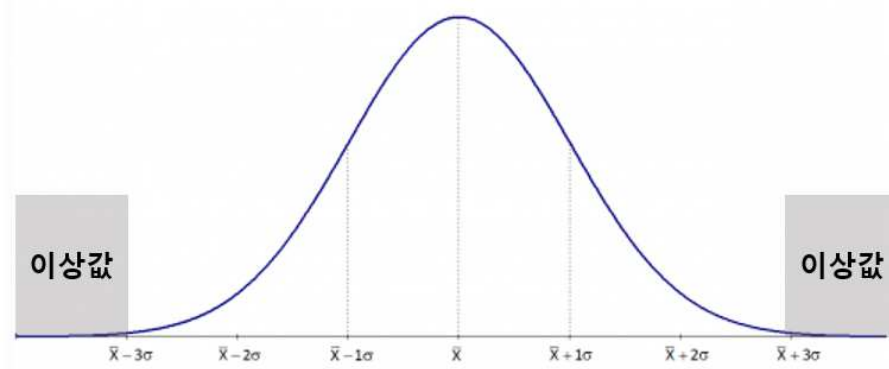
## - 데이터 이상값 처리

### ● 이상값 처리

- 극단적으로 크거나 작은 값이며, 의미 있는 데이터 일수도 있음 (체중 3kg)
- 이상값을 항상 제거하는 것은 아님

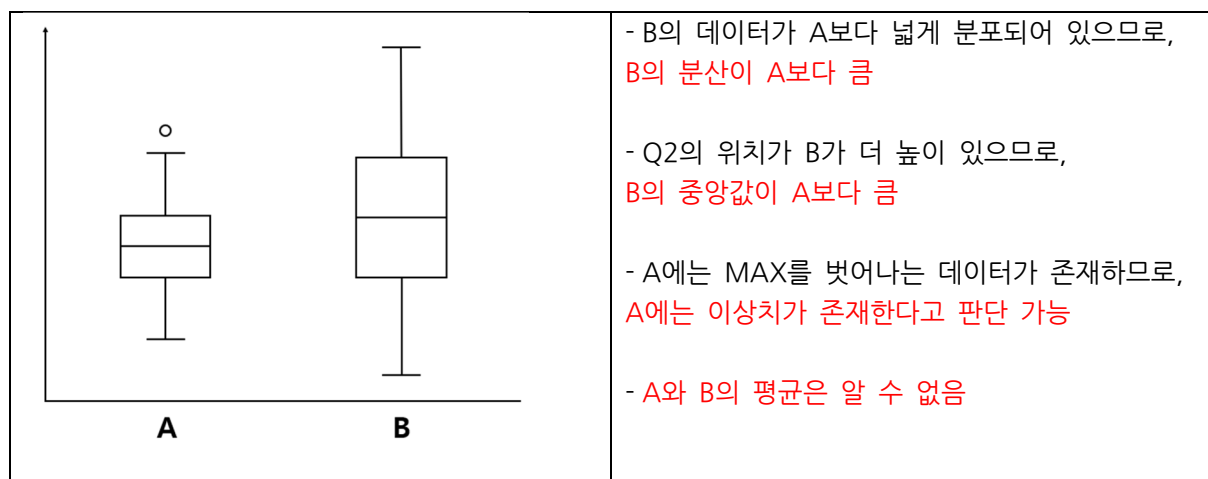
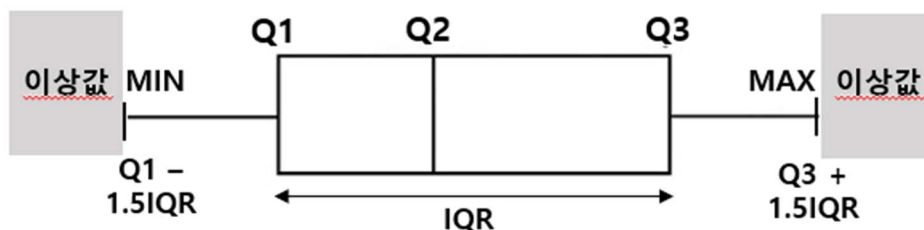
#### (1) ESD(Extreme Studentized Deviation)

- 평균으로부터 표준편차의 3배 넘어가는 데이터는 이상값으로 판단



#### (2) 사분위수

- $Q1 - 1.5IQR$ 보다 작거나,  $Q3 + 1.5IQR$ 보다 크면 이상값으로 판단
- 최솟값, 1~3사분위값, 최댓값 등을 표현하며, 평균값은 표현하지 않음



#### (3) Z-Score

- 데이터를 정규화(평균 0, 표준편차 1) 후, 일정 임계 값을 초과할 경우 이상값으로 판단

#### (4) DBScan

- 밀도를 이용하여 밀도가 적은 부분의 데이터를 이상값으로 판단

## 2. 분석 변수 처리

### - 변수 선택

#### ● 변수 선택 방법

- (1) 전진선택법 : 변수를 하나씩 추가해 나가는 방법
- (2) 후진제거법 : 변수를 하나씩 제거해 나가는 방법
- (3) 단계별 선택법 : 전진선택법 + 후진선택법으로 변수 추가할 때 벌점을 고려

상관계수 매트릭스 분석

	X1	X2	X3	X4
X1	1.0	0.21	-0.31	0.95
X2	0.21	1.0	0.02	0.13
X3	-0.31	0.02	1.0	-0.05
X4	0.95	0.13	-0.05	1.0

- X1과 X4가 높은 상관계수를 가지므로, 둘 중 하나의 변수 제거

☞ 상관계수는 -1 ~ 1사이의 값을 가지며, -1은 음의상관관계, 1은 양의상관관계를 의미

### - 차원 축소

#### ● 차원의 저주

- 데이터 학습 시 차원이 높아질수록 알고리즘의 성능이 저하되는 현상

#### ● 차원 축소의 효과

- 차원의 저주 해소, **데이터 시각화**, **노이즈 제거**, **데이터 압축**, 성능 향상, 특징 추출, 계산 비용 절감

#### ● 차원 축소 기법

- (1) 선형 차원 축소 기법

- 1) 주성분 분석 (PCA)

- **분산이 최대화**되는 방향으로 차원을 축소

- 2) LDA(Linear Discriminant Analysis)

- 분류 에서 클래스 간 **분산을 최대화**하면서 차원을 축소하는 기법

- 3) ICA(Independent Component Analysis)

- 서로 **독립적인 성분**을 찾아 차원 축소

- 4) SVD(Singular Value Decomposition)

- $m \times n$  크기의 **비정방행렬**  $A$ 를 행렬 분해 ( $A = U \Sigma V^T$ )

- $U$ 는  $m \times m$  직교행렬,  $\Sigma$ 는  $m \times n$  대각행렬,  $V^T$ 는  $n \times n$  직교행렬

- 5) 요인분석(Factor Analysis)

- 관측된 변수들을 몇 개의 **잠재 요인**으로 **축소**하는 방법

- 심리학, 사회과학 등에서 잠재 변수의 해석이 필요한 경우

## (2) 비선형 차원 축소 기법

### 1) MDS(Multi-Dimensional Scaling)

- 데이터 간 **거리정보의 근접성을 보존**하는 차원 축소
- Stress 함수의 값이 최소가 되도록 구성

### 2) t-SNE (t-Distributed Stochastic Neighbor Embedding)

- 고차원 **데이터간 거리 정보를 확률적으로 유지하여** 차원 축소
- 데이터의 복잡한 비선형 패턴을 시각화 하며, 이미지나 텍스트 데이터에 적합

### 3) UMAP(Uniform Manifold Approximation and Projection)

- t-SNE와 유사한 비선형 차원 축소 기법으로, 구조적 패턴을 보존
- 대규모 데이터에서 빠르게 적용 가능

### 4) AutoEncoder

- 신경망을 활용한 차원 축소 기법으로 데이터를 압축 후 다시 복원하는 학습

## - 파생변수 생성

### ● 요약변수와 파생변수

- (1) 요약변수 : 수집된 정보를 종합한 변수로서 **재활용성**이 높음 (1개월간 수입)
- (2) 파생변수 : 의미를 부여한 변수, **논리적 타당성** 필요 (고객구매등급)
  - 파생변수 생성방법 : 특징 추출, 결합, 부가적 정보 결합, 수학적 변환, **교호작용**의 반영
  - \* 교호작용 : **두 개 이상의 독립변수가 상호작용**을 하여, 종속변수에 영향을 미치는 경우

## - 변수 변환

### ● 수치형 자료와 범주형 자료

- (1) 수치형 자료 : 키, 몸무게 → **회귀분석**
- (2) 범주형 자료 : 혈액형, 성별 → **분류분석**

### ● 수치형 변수 변환

- (1) Z-Score 정규화 : **평균 0, 표준편차 1**로 변환 ~ N(0,1)

$$Z = \frac{X - \mu}{\sigma}$$

- (2) 최소-최대 정규화 : 0에서 1사이로 변환

$$X_{new} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- (3) 로그 변환 : 로그를 취한 값으로 변환, **데이터가 한쪽으로 치우쳐 있을 시 활용**

$$X_{new} = \log X$$

- (4) 그 외 변환 기법들 : 지수변환, 제곱근 변환, Box-Cox 변환(**양수 데이터의 비대칭 분포** 시 활용)

## ● 범주형 변수 변환

### (1) 레이블 인코딩

- 데이터를 정수로 변환
- 오렌지, 바나나, 포도 → 0, 1, 2

### (2) 원-핫 인코딩

- 고유 값에 해당하는 컬럼만 1로 표시하고 나머지는 0으로 표시
- 오렌지, 바나나, 포도 → [1, 0, 0], [0, 1, 0], [0, 0, 1]

### (3) 타깃 인코딩

- 타깃 변수를 평균값으로 변환

## ● 날짜/시간 변수 변환

- (1) 분할 : 날짜/시간 데이터를 년, 월, 일, 시, 분 등으로 분할
- (2) 파생 : 시간대(오전, 오후), 요일 또는 계절 등의 파생변수 생성

## - 불균형 데이터 처리

### ● 불균형 데이터의 처리 방법

#### (1) 가중치 균형 적용(Weighted Balance)

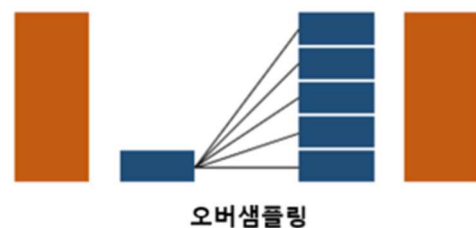
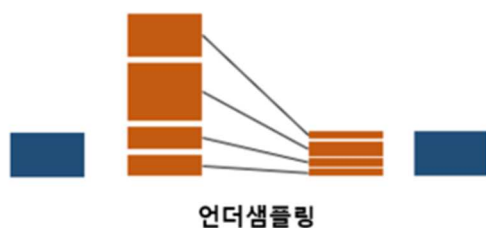
- 불균형 데이터에 가중치를 주는 방법

#### (2) 언더샘플링

- 다수 데이터의 일부만 선택
- 방법 : 랜덤 추출법, 계통 추출법, 집락 추출법, 층화 추출법

#### (3) 오버샘플링

- 소수 데이터를 복사하거나 유사한 데이터를 만드는 방식
- 방법 : SMOTE, ADSYN, ROS



## 데이터 탐색

### 1. 데이터 탐색 기초

#### - 데이터 탐색 개요

##### ● EDA (탐색적 자료 분석)

- 데이터의 의미를 찾기 위해 **통계**, **시각화**를 통해 파악
- EDA의 4가지 주제
  - 1) **전향성의 강조** : 자료 변동에 민감하지 않음
  - 2) **잔차 계산** : 값들이 주경향으로부터 얼마나 벗어나 있는지 확인하는 척도
  - 3) **자료변수의 재표현** : 원래 변수를 적당한 척도로 변환
  - 4) **그래프를 통한 현시성** : 시각화를 통하여 효율적으로 파악

☞ '저잔재현'

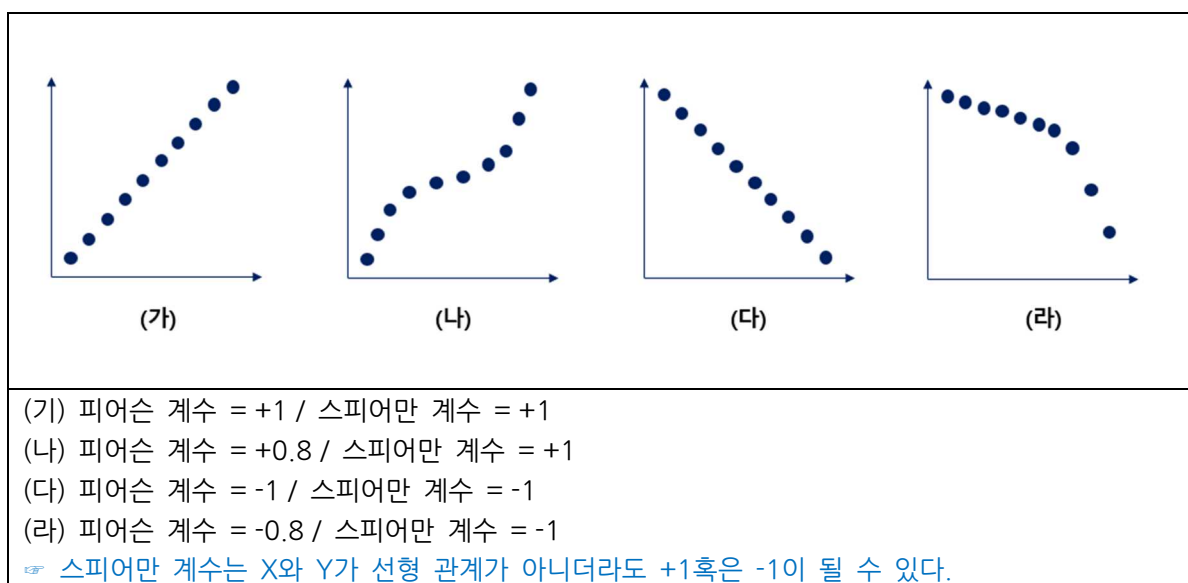
#### - 상관관계 분석

##### ● 상관분석

- 두 변수간의 선형적 관계가 존재하는 파악하는 분석
- (1) **단순상관분석** : 두 개의 변수가 어느정도 강한관계에 있는가를 측정
- (2) **다중상관분석** : 3 개 이상의 변수 간 관계를 측정
- (3) **편상관관계분석** : 제 3의 변수를 통제한 상태에서 두 변수의 상관관계를 분석

##### ● 상관분석 방법

- (1) **피어슨 상관분석** : 양적 척도, 연속형 변수, 선형관계 크기 측정
- (2) **스피어만 상관분석** : 서열 척도, 순서형 변수, 선형/비선형적 관계 나타냄



## - 기초통계량 추출 및 이해

### ● 기초 통계량

#### (1) 중심경향성 측면

- 산술평균 : 일반적인 평균 개념으로, 모든 값을 더한 후 데이터 개수로 나눈 값
- 기하평균 : 모든 값들을 곱하고,  $n$  제곱근을 구하는 방식 (비율적 증가율)
- 조화평균 : 역수의 산술평균을 구한 후, 다시 역수를 취하는 방식 (비율 계산)
- 중앙값 : 데이터를 크기 순서로 나열했을 때 중간에 위치한 값
- 최빈값 : 데이터에서 가장 자주 나타나는 값

#### (2) 분산 정도 측면

- 범위 : 데이터의 최댓값과 최솟값 사이, 전체적인 퍼짐을 나타냄
- 분산 : 각 데이터가 평균과 얼마나 떨어져 있는지 나타내는 지표
- 표준편차 : 분산에 제곱근을 취한 값
- 사분위수(IQR) : 데이터의 상위 75%와 하위 25%의 중간 범위
- 변동계수(CV) : 평균대비 표준편차의 비율

#### (3) 관계 측면

##### 1) 공분산 : 두 확률변수의 상관정도

- 공분산 = 0 : 상관이 전혀 없는 상태
- 공분산  $> 0$  : 양의 상관관계
- 공분산  $< 0$  : 음의 상관관계
- 최소, 최대값이 없어 강약 판단 불가

##### 2) 상관계수 : 상관정도를 -1 ~ 1값으로 표현

- 상관계수 = 1 : 정비례 관계
- 상관계수 = 0 : 상관없음
- 상관계수 = -1 : 반비례 관계

##### 3) 공분산과 독립성의 관계

- 두 변수가 독립이면 공분산은 0이지만, 공분산이 0이라고 두 변수가 독립이라고 할 수는 없음

### ● 기댓값과 분산의 특성

서로 독립인  $X, Y$ 가 각각 정규분포  $N(20, 2^2)$  와  $N(27, 1^2)$ 을 따른다고 할 때, 확률변수  $Z = 5X - 7Y + 15$ 일 경우의  $Z$ 의 기댓값과 분산의 계산

#### 1) $Z$ 의 기댓값

$$: E(Z) = E(5X - 7Y + 15) = 5 \times 20 - 7 \times 27 + 15 = -74$$

#### 2) $Z$ 의 분산

$$: V(Z) = V(5X - 7Y + 15) = V(5X - 7Y) = 25 \times 2^2 + 49 \times 1^2 = 149$$

☞  $V(\text{상수}) = 0$  으로 취급

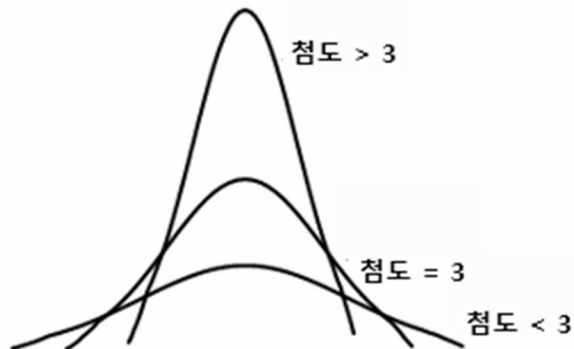
## ● 첨도와 왜도

(1) 첨도 : 자료의 분포가 얼마나 뾰족한 지 나타내는 척도

- 첨도 = 3 : 정규 분포 형태

☞ 3을 빼서 0을 기준으로 정규분포 형태를 판단하기도 함

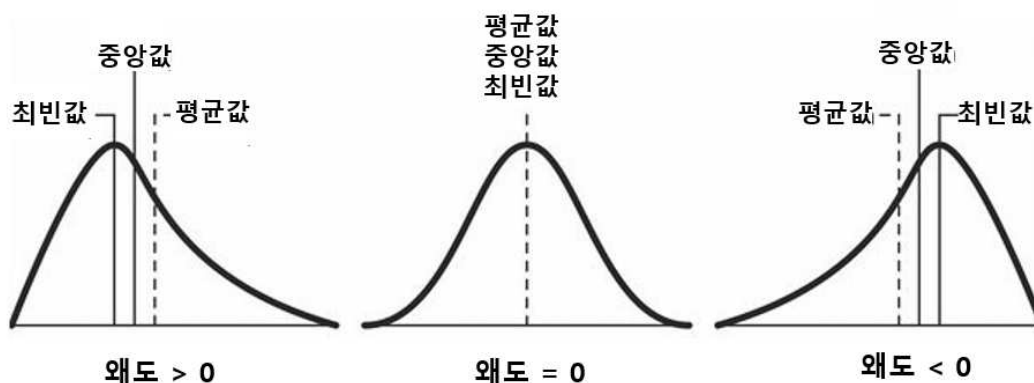
- 값이 클수록 뾰족한 모양



(2) 왜도 : 자료 분포의 비대칭 정도 (0일 때 대칭)

- 왜도 < 0 : 최빈값 > 중앙값 > 평균값

- 왜도 > 0 : 최빈값 < 중앙값 < 평균값



☞ 평균값은 꼬리를 따라감

## ● Summary함수 결과의 해석

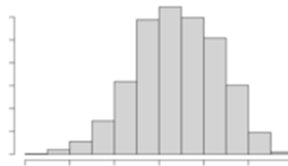
- 통계 요약 해석			1) Age 변수 - Mean, Median 등 존재 → 수치형 변수 - 25% 지점 : 21.00, 75% 지점 : 39.00 - Median < Mean → 왜도 > 0 - 결측치(NA's) 개수 : 86개
	Age	Survived	
	Min. : 0.17	0:266	2) Survived 변수 - 집단의 빈도 수 → 범주형 변수 - 범주 0과 1이 클래스 불균형 상태 → 0을 다운샘플링 하거나 1을 오버샘플링
	1st Qu.:21.00	1:152	
	Median :27.00		
	Mean :30.27		
	3rd Qu.:39.00		
	Max. :76.00		
	NA's :86		

## - 시각적 데이터 탐색

Line Plot



Histogram



Bar Plot



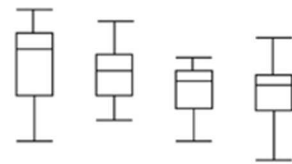
Pie Plot



Scatter plot



Box plot



## 2. 고급 데이터 탐색

### - 시공간 데이터 탐색

#### ● 시공간 데이터

- 공간적 정보에 시간의 흐름이 결합된 데이터
  - 활용 : 시공간 **패턴**을 통한 예측, **지도**를 통한 위치정보, 지리공간의 **결자** 차트 결합
- ☞ '패지력'

### - 다변량 데이터 탐색

#### ● 다변량 데이터 탐색

- (1) 목표 : 변수 간 관계, 패턴 분석, 이상치 탐지, 데이터 요약
- (2) 방법 : 상관관계분석, PCA, 다차원척도법, 다중선형회귀, 군집 분석 등

### - 비정형 데이터 탐색

#### ● 자연어 전처리

- (1) **토큰화(Tokenization)** : 의미 있는 말뭉치로 문자를 나누는 작업
  - (2) **불용어(Stop Words) 처리** : 조사, 접속사 등의 의미 없는 정보를 제거
  - (3) 정규화 : 같은 의미이면서 표현이 다른 단어를 통합
  - (4) **어간추출(stemming)** : 접사를 제거하여 기본 형태를 찾아내는 작업
  - (5) 표제어 추출(lemmatization) : 사전에 기반하여 단어를 원형으로 변환
- ☞ 한국어는 띄어쓰기가 일정하지 않고, 형태소 단위로 분리해야 되어 토큰화가 어려움



## 통계기법 이해

### 1 기술통계

#### - 표본추출

##### ● 전수조사와 표본조사

- 전수조사 : 전체를 다 조사, 시간과 비용 많이 소모
- 표본조사 : 일부만 추출하여 모집단을 분석

##### ● 확률적 표본 추출 방법

- (1) 랜덤 추출법 : 무작위로 표본 추출
- (2) **계통 추출법**(=계층적 추출법) : 번호 부여하여 일정 간격으로 추출
- (3) **집락 추출법**(=군집 추출법)
  - 여러 군집으로 나눈 뒤 군집을 선택하여 랜덤 추출
  - 군집 내 이질적 특징, 군집 간 동질적 특징
- (4) **층화 추출법**
  - 군집 내 동질적 특징, 군집 간 이질적 특징
  - 같은 비율로 추출 시, **비례 층화 추출법**
- (5) 복원, 비복원 추출
  - 복원 추출 : 추출되었던 데이터를 **다시 포함**시켜 표본 추출
  - 비복원 추출 : 추출되었던 데이터는 제외하고 표본 추출

##### ● 비확률적 표본 추출 방법

- (1) 편의 추출법 : 연구자가 쉽게 접근 가능한 대상으로 표본을 추출
- (2) 의도적 추출법 : 연구자가 특정 기준을 정하고, 이에 맞는 표본을 추출
- (3) 할당 추출법 : 특정 기준으로 나눈 후, 그 그룹에서 할당된 수 만큼 추출
- (4) 눈덩이 추출법 : 초기 응답자로부터 새로운 응답자를 추천 받는 방식
- (5) 자기선택 추출법 : 응답자가 스스로 조사에 참여할지 결정

#### - 확률분포

##### ● 기초 확률 용어

- (1) 확률 : 통계적 현상의 확실함을 나타내는 척도로 수학적 확률과 통계적 확률로 구분
- (2) 사건 : 여러 반복된 시행을 통해 결과로서 나타나는 표본공간의 부분 집합
- (3) 표본공간 : 통계적 실험에 의하여 일어날 수 있는 모든 가능한 결과
  - 예) 동전 두 개를 던질 때 표본공간  $S = \{(앞, 앞), (앞, 뒤), (뒤, 앞), (뒤, 뒤)\}$
- (4) 확률변수 : 표본공간의 각 원소에 해당하는 값(확률)을 대응하는 함수
  - 예) 확률변수  $X$ 가 어떤 집합의 키를 나타낼 때 기가 160~170 확률은  $P(160 \leq X \leq 170)$
- (5) 조건부 확률 : 특정 사건  $B$ 가 발생했을 때  $A$ 가 발생할 확률
  - $P(A|B) = P(A \cap B)/P(B)$  (백신을 맞았을 때 감기에 걸릴 확률)

- (6) **독립사건** : A, B가 서로 영향을 주지 않는 사건 (  $P(A|B) = P(A)$  )  
 -  $P(A \cap B) = P(A)P(B)$  (주사위 A가 3이 나왔을 때, 주사위 B가 3이 나올 확률)  
 (7) **배반사건** : A, B가 서로 동시에 일어나지 않는 사건  
 -  $P(A \cap B) = \emptyset$  (동전을 던졌을 때 앞면과 뒷면이 동시에 나올 확률)  
 (8) **베이즈 정리** : 두 확률 변수의 사전 확률과 사후 확률 사이의 관계를 나타내는 정리  
 -  $P(A|B) = P(B|A)P(A)/P(B)$

## ● 확률분포

- 확률 변수의 개별 값들이 가지는 확률 값의 분포

### (1) 이산 확률분포

- 값을 셀 수 있는 분포, 확률질량함수로 표현
- 1) 이산균등분포 : 모든 곳에서 값이 일정한 분포
  - 예) 주사위의 각 면이 나오는 확률은 모두 동일
- 2) 베르누이분포 : 결과가 두 가지 중 한가지로 나타나는 베르누이시행으로 나타나는 분포
  - 예) 동전 던지기, 시험의 합격/불합격
- 3) 이항분포 : N번의 베르누이시행 중 K번 성공할 확률의 분포
  - 예) 동전을 20번 던져 앞면이 나오는 횟수
- 4) 기하분포 : 성공확률이 p인 베르누이시행에서 처음으로 성공할 때까지 시행횟수의 분포
  - 예) 동전을 던져 처음으로 앞면이 나오기까지 던진 횟수
- 5) 음이항분포 : 성공확률이 p인 베르누이시행을 r번 성공할 때까지 반복 시행횟수의 분포
  - 예) 동전을 던져 앞면이 5번 나오기까지 던진 횟수
- 6) 초기하분포 : N개 중 비복원추출로 n번 추출했을 때 원하는 결과가 k번 나올 확률의 분포
  - 예) 10개 구슬 중 4개의 구슬이 당첨 구슬일 때, 4번 뽑았을 때 당첨 구슬을 2번 뽑을 확률
- 7) 다항분포 : N번 시행에서 각 시행이 여러 개의 결과를 가질 수 있는 확률 분포
  - 예) 주사위를 20번 던져 각 면이 나오는 횟수
- 8) 포아송분포 : 단위 시간 내 발생할 수 있는 사건의 발생 횟수에 대한 분포
  - 예) 하루동안 발생하는 출생자 수, 한 시간 동안 사무실에 걸려온 전화의 수

☞ '베포항항하'

### (2) 연속 확률분포

- 값을 셀 수 없는 분포, 확률밀도함수로 표현
- 1) **정규분포** : 우리가 일상생활에서 흔히 보는 확률변수의 평균 분포를 근사한 분포 (Z검정 활용)
  - 예) 사람들의 키 혹은 IQ 점수의 분포, 시험 성적의 분포
- 2) **t분포** : 정규분포와 유사하지만, 꼬리 부분이 더 두껍고 긴 분포
  - (T검정 활용) 표본이 30개 보다 작은 집단에 대한 평균 검정
- 3) 카이제곱분포 : 독립적인 정규분포를 따르는 변수들의 제곱합으로 구성된 분포
  - (카이제곱 검정 활용) 두 집단의 동질성 검정, 단일 집단의 모분산 검정
- 4) F분포 : 두 개의 서로 다른 카이제곱 분포의 비율
  - (F검정 활용) 두 집단의 분산 동질성 검정

### ● 확률분포의 기댓값

- 확률변수  $X$ 의  $f(x)$  확률분포의 대한 기댓값( $E(X)$ )

1) 이산적 확률변수 :  $E(X) = \sum xf(x)$

2) 연속적 확률변수 :  $E(X) = \int xf(x)$

(1) 동전을 3개 던지는 확률실험을 할 때, 확률변수  $X$ (앞면F의 개수)의 기댓값은?

(2) 1~12의 숫자가 표시된 원형시계에서, 확률변수  $X$ (시계 바늘이 가르키는 시간)의 기댓값은?

(1)

$$- P(X = 0) = P(\{BBB\}) = \frac{1}{8}$$

$$- P(X = 1) = P(\{FBB, BFB, BBF\}) = \frac{3}{8},$$

$$- P(X = 2) = P(\{FFB, FBF, BFF\}) = \frac{3}{8}$$

$$- P(X = 3) = P(\{FFF\}) = \frac{1}{8}$$

$$\therefore E(X) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 1.5$$

(2)



$$\therefore E(X) = \int_0^{12} xf(x)dx = \int_0^{12} x \left(\frac{1}{12}\right) dx = 6$$

### - 표본분포

#### ● 중심극한정리

- 임의의 모집단으로부터 추출된 표본분포는 표본크기가 충분히 크면(30개 이상) 정규분포
- 모집단의 분포에 상관없이 표본분포가 정규분포를 이룸

#### ● 표본평균의 표본분포

(1) 표본평균의 표본분포의 평균 :  $E(\bar{X}) = \mu$

(2) 표본평균의 표본분포의 분산 :  $V(\bar{X}) = \sigma^2/n$

(3) 표본평균의 표준오차 :  $SE(\bar{X}) = \sigma/\sqrt{n}$

(4) 표본평균의 표준화 :  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

( $\mu$  : 모집단의 평균,  $\sigma$  : 모집단의 표준편차,  $\bar{X}$  : 표본평균,  $n$  : 표본의 크기)

#### ● 표본비율의 표본분포

(1) 표본비율 :  $\hat{P} = X/n$

(2) 표본비율의 평균 :  $E(\hat{P}) = P$

(3) 표본비율의 표준오차 :  $SE(\hat{P}) = \sqrt{pq/n}$

(4) 표본비율의 표준화 :  $Z = \frac{\hat{P} - p}{\sqrt{pq/n}} \sim N(0,1)$

## 2. 추론통계

### - 점추정

#### ● 점추정

- 모집단이 특정한 값으로 추정하며, **추정량(Estimator)**으로 모수를 추정

(1) 추정량의 조건

- 1) 불편성(Unbiasedness) : 추정량의 기댓값이 실제 모수와 같음 (**편향**이 0이 되는 경우)
- 2) 효율성(Efficiency) : 여러 추정량 중 분산이 작은 것이 더 효율적인 추정량
- 3) 일치성(Consistency) : 표본 크기가 증가할수록 추정량이 모수에 가까워짐
- 4) 충족성(Sufficiency) : 추정량이 모집단의 정보를 최대한 반영

☞ **'불효일충'**

(2) 대표적인 추정량

- 1) 모집단의 평균  $\mu \rightarrow$  표본평균  $\bar{X} = \frac{1}{n} \sum X$
- 2) 모집단의 분산  $\sigma^2 \rightarrow$  표본분산  $s^2 = \frac{1}{n-1} \sum (X - \bar{X})^2$
- 3) 모집단의 비율  $p \rightarrow$  표본비율  $\hat{p} = X/n$

(3) 평균제곱오차(MSE) : 점추정에서 추정량과 실제 모수 간의 평균적인 오차 크기

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2$$

### - 구간추정

#### ● 구간추정

- 모집단이 특정한 구간으로 추정 (95%, 99%를 가장 많이 사용)

#### ● 모평균의 구간추정(신뢰구간)

(1) 모집단의 분산을 알고 있는 경우

$$\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

- 신뢰수준 95% :  $Z_{\frac{\alpha}{2}} = 1.960$

- 신뢰수준 99% :  $Z_{\frac{\alpha}{2}} = 2.576$

(2) 모집단의 분산을 모르는 경우

- **자유도가 n-1인 t분포**를 이용하여 신뢰구간을 추정

$$\bar{X} - t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} \quad (S : \text{표본표준편차})$$

(3) 비율에 대한 신뢰 구간

$$\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

## - 가설검정

### ● 가설검정

- 모집단의 특성에 대한 주장을 가설로 세우고 표본조사로 가설의 채택여부를 판정
- (1) 귀무가설( $H_0$ ) : 일반적으로 생각하는 가설 (차이가 없다)
- (2) 대립가설( $H_1$ ) : 귀무가설을 기각하는 가설, **증명하고자 하는 가설** (차이가 있다, 크다/작다)
- (3) 검정통계량 : 귀무가설을 기각할지 결정하기 위해 표본으로 산출되는 통계량
- (4) 유의수준( $\alpha$ ) : 귀무가설이 참일 때 기각하는 **1종 오류**를 범할 확률의 허용 한계 (일반적 0.05)
- (5) 기각역 : 귀무가설이 기각되고 대립가설이 채택되는 **검정통계량**의 영역

검정결과 실제	$H_0$ 가 사실이라고 판정	$H_0$ 가 거짓이라고 판정
$H_0$ 가 사실	옳은 결정	<b>1종 오류(<math>\alpha</math>)</b>
$H_0$ 가 거짓	<b>2종 오류(<math>\beta</math>)</b>	옳은 결정

### ● 가설 검정 문제 풀이 방법

- (1) 귀무가설 / 대립가설 설정
  - ‘차이가 없다’ 혹은 ‘동일하다’ → 귀무가설
- (2) 양측 혹은 단측검정 확인
  - 대립가설의 값이 ‘같지 않다’ → 양측검정
  - ‘값이 크다’, ‘값이 작다’ → 단측검정
- (3) 일표본 혹은 이표본 확인
  - 하나의 모집단 → 일표본
  - 두개의 모집단 → 이표본
- (4) 검정통계량 계산과 기각역 판단
  - 검정통계량이 기각역에 존재 → 귀무가설 기각
  - 검정통계량이 채택역에 존재 → 귀무가설 채택
- (5) t 검정인 경우 - 단일표본, 대응표본, 독립표본 확인
  - 모집단에 대한 평균검정 → 단일표본
  - 동일 모집단에 대한 평균비교 검정 → 대응표본
  - 서로 다른 모집단에 대한 평균비교 검정 → 독립표본

A 공장 제품의 불량률은 1.7%, 모표준편차 0.3%로 관리하고있다. 이 공장에서 9개의 샘플을 활용하여 분석하였더니 불량률이 평균 1.5%였다. 이 공장의 불량률을 1.7%로 관리하고 있다는 주장에 대해서 의문을 제기할 때, 5% 유의수준으로 가설검정을 수행 ( $Z_{0.05} = 1.645, Z_{0.025} = 1.96$ )

- 1) 귀무가설/대립가설 설정
  - ‘차이가 없다’ 혹은 ‘동일하다’ → 귀무가설로 설정
  - **A공장의 불량률은 1.7%이다.**
- 2) 양측 혹은 단측검정 확인
  - 대립가설의 값이 같지 않다 → **양측검정**
- 3) 일표본 혹은 이표본 확인
  - 한개의 모집단 → **일표본**

4) 검정통계량 계산과 기각역 판단

- 검정통계량 :  $Z_o = (\bar{X} - \mu)/(\delta/\sqrt{n}) = (1.5 - 1.7)/(0.3/\sqrt{9}) = -2.0$
- 기각역 :  $\pm Z_{0.025} = \pm 1.96$
- 판정 :  $Z_o < -1.96 \rightarrow$  귀무가설 기각

(심화) 모표준편차를 모르는 경우 (표본표준편차가 0.3% 일 때)  $\rightarrow$  t검정을 사용

1), 2), 3)은 동일

4) 검정통계량 계산과 기각역 판단

- 검정통계량 :  $t_o = (\bar{X} - \mu)/(S/\sqrt{n}) = (1.5 - 1.7)/(0.3/\sqrt{9}) = -2.0$
- 기각역 :  $\pm t_{0.025, 8} = \pm 2.18$
- 판정 :  $t_o > -2.18 \rightarrow$  귀무가설 채택

5) 단일표본, 대응표본, 독립표본 확인

- 모집단에 대한 평균검정  $\rightarrow$  단일표본

## 3과목 빅데이터 모델링

### - 3 과목(빅데이터 모델링) 주요 내용

분석모형 설계	분석 절차 수립	분석모형 선정
		분석모형 정의
		분석모형 구축 절차
	분석 환경 구축	분석 도구 선정
		데이터 분할
분석기법 적용	분석기법	회귀분석
		로지스틱 회귀분석
		의사결정나무
		인공신경망
		서포트벡터머신
		연관성분석
		군집분석
	고급 분석기법	범주형 자료 분석
		다변량 분석
		시계열 분석
		베이지안 기법
		딥러닝 분석
		비정형 데이터 분석
		앙상블 분석
		비모수 통계



## 분석모형 설계

### 1. 분석 절차 수립

#### - 분석모형 선정

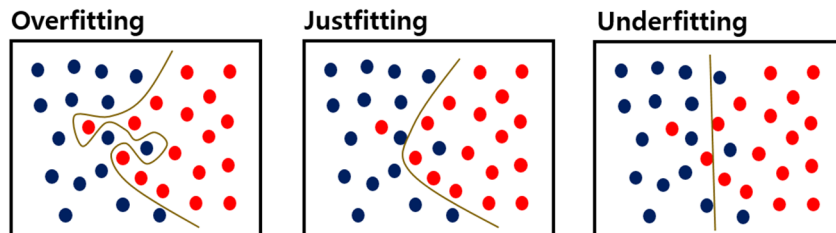
##### ● 분석모형 선정

- (1) 통계 기반 : 회귀분석, 상관분석, 주성분분석(PCA), 분산분석(ANOVA), 판별분석
- (2) 데이터마이닝 기반 : 분류모델, 예측모델, 군집모델, 연관규칙모델
- (3) 머신러닝 기반 : 지도학습, 비지도학습, 준지도학습, 강화학습
- (4) 비정형데이터 기반 : 텍스트 마이닝, 오피니언 마이닝, 소셜네트워크 분석

#### - 분석모형 정의

##### ● 분석모형 정의 시 고려사항

- (1) **과대적합** : 모델이 지나치게 데이터를 학습하여 매우 복잡해진 모델
- (2) **과소적합** : 데이터를 충분히 설명하지 못하는 단순한 모델
- (3) 모형 선택의 오류 : 적합하지 않은 모델 선택
- (4) 변수 선택의 오류 : 관련 있는 변수가 누락되거나 관련 없는 변수가 포함된 모델
- (5) 데이터 편향 : 대표성 없는 훈련 데이터로 인한 편향



#### - 분석모형 구축 절차

##### ● 분석모형 구축 절차

- (1) 요건정의 : 분석요건 도출, 수행방안 설계, 요건 확정
- (2) 모델링 : 설계 및 구축, 탐색적 분석 및 유의변수 도출, 모델링, 모델링 성능평가
- (3) 검증 및 테스트 : 실제 테스트 수행, 비즈니스 영향도 평가
- (4) 적용 : 운영시스템에 적용, 주기적 리모델링

### 2. 분석 환경 구축

#### - 분석 도구 선정

##### ● 분석 도구 선정

- (1) R : 통계 분석에 특화, 처리속도 느림, 강력한 시각화, SPSS/SAS등 연동 가능
- (2) Python : 간결함과 높은 가독성, R보다 속도 빠름, C구현된 모듈과 연동, R보다 약한 시각화

## - 데이터 분할

### ● 데이터 분할

- 과대적합과 과소적합을 방지하고, 데이터가 불균형한 문제를 해결하기 위해 사용

(1) 훈련용(Training Set) : 모델을 학습하는데 활용

(2) 검증용(Validation Set) : 모델의 과대,과소 적합을 조정하는데 활용

(3) 평가용(Test Set) : 모델을 평가하는데 활용

## 분석기법 적용

### 1. 분석기법

#### - 회귀분석

##### ● 회귀분석

(1) 개념 : 독립변수들이 종속변수에 영향을 미치는 파악하는 분석방법

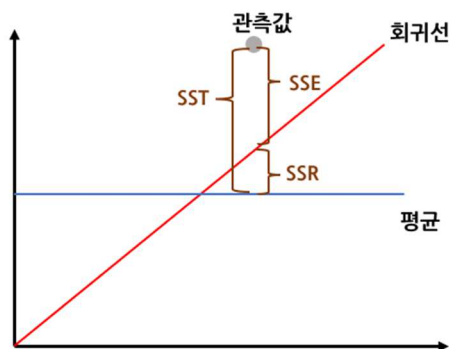
- 1) 독립변수 : 원인을 나타내는 변수 (x)
- 2) 종속변수 : 결과를 나타내는 변수 (y)
- 3) 잔차 : 계산값과 예측값의 차이

(2) 회귀계수 추정방법

- **최소제곱법(=최소자승법)** : 잔차의 제곱합(SSE)이 최소가 되는 회귀계수와 절편을 구하는 방법

(3) 회귀모형 평가

- **R-square** : 총 변동 중에서 회귀모형에 의하여 설명되는 변동이 차지하는 비율 (0 ~ 1)



##### ● SST, SSE, SSR

- (1) SST : 전체의 변동
- (2) SSE : 모형에 의해 설명되지 않는 변동
- (3) SSR : 모형에 의해 설명되는 변동
- (4)  $R^2 = SSR/SST = 1 - SSE/SST$

☞ SST : Sum of Squares Total / SSR : Sum of Squares Regression / SSE : Sum of Squares Error

##### ● 선형회귀분석의 가정

- (1) 선형성 : 종속변수와 독립변수는 선형관계
  - (2) 등분산성 : 잔차의 분산이 고르게 분포
  - (3) 정상성(정규성) : 잔차가 정규분포의 특성을 지님
  - (4) 독립성 : 독립변수들간 상관관계가 없음
- \* **다중공선성** : 독립변수들간 강한 상관관계가 나타나는 문제
- **VIF(분산팽창인수)** 값이 10 이상이면 다중공선성 존재한다고 판단

☞ '선분정독'

##### ● 회귀분석 종류

- (1) 단순회귀 : 1 개의 독립변수와 종속변수의 선형관계
- (2) **다중회귀** : 2 개 이상의 독립변수와 종속변수의 선형관계
- (3) **다항회귀** : 2 개 이상의 독립변수와 종속변수가 2 차 함수 이상의 관계
- (4) 릿지회귀(L2 규제) / 라쏘회귀(L1 규제) : 규제를 포함하는 회귀 모형
- (5) **교호항이 포함된 회귀** : 독립변수들의 교호작용이 포함된 회귀 모형

## ● 회귀 모형의 구축절차

- (1) 독립변수, 종속변수의 설정
- (2) 회귀 계수의 추정
- (3) 회귀 계수들의 유의성 검정
- (4) 모형의 유의성 검정

## ● 회귀 모형의 변수 선택 방법

- (1) 전진선택법 : 변수를 하나씩 추가하면서 최적의 회귀방정식을 찾아내는 방법
  - (2) 후진제거법 : 변수를 하나씩 제거하면서 최적의 회귀방정식을 찾아내는 방법
  - (3) 단계별 선택법 : 전진선택법 + 후진선택법으로 변수를 추가할 할 때 벌점을 고려
    - 1) AIC (아카이케 정보 기준) : 편향과 분산이 최적화 되는 지점 탐색, 자료가 많을수록 부정확
    - 2) BIC (베이즈 정보 기준) : AIC를 보완했지만 AIC보다 큰 패널티를 가지는 단점
- ☞ AIC와 BIC 모두 작을수록 좋음

## - 로지스틱 회귀분석

### ● 로지스틱 회귀분석

- 종속변수가 **범주형 데이터**를 대상으로 성공과 실패 2개의 집단을 분류하는 문제에 활용
- (1) 오즈(Odds)
  - 성공할 확률과 실패할 확률의 비
  - $\text{Odds} = P / (1 - P) = \text{성공확률} / \text{실패확률}$
- (2) 로짓(logit)변환
  - 오즈에 자연로그(자연상수 e가 밑)를 취하여 선형 관계로 변환
  - $\log(P/(1 - P)) = \alpha + \beta x$
- (3) 시그모이드 함수
  - 로짓 함수의 역함수를 통하여, 0~1사이 확률을 도출하는 함수
  - 독립변수  $x$ 가 **n증가하면 확률이  $e^n$  만큼 증가**

## - 의사결정나무

### ● 의사결정나무(Decision Tree)

- **노드 내 동질성이 커지고, 노드 간 이질성이 커지는 방향으로 분리**

### ● 의사결정나무(Decision Tree) 분할 방법


- (1) 분류(범주형)에서의 분할 방법
  - 1) CHAID 알고리즘 : 카이제곱 통계량
  - 2) CART 알고리즘 : **지니지수** 활용 ( $1 - \sum P^2$ )
  - 3) C4.5/C5.0 알고리즘 : **엔트로피지수** 활용 ( $-\sum P(\log P)$ )

- (2) 회귀(연속형)에서의 분할 방법
- 1) CHAID 알고리즘 : ANOVA, F-통계량
  - 2) CART 알고리즘 : 분산감소량

● 의사결정나무의 과적합 방지 방안

- (1) 정지규칙 : 분리를 더 이상 수행하지 않고 나무의 성장을 멈춤
- (2) 가지치기 : 일부 가지를 제거하여 과대적합을 방지

● 지니지수와 엔트로피지수의 계산



- 앞면 확률 =  $\frac{3}{5}$ , 뒷면 확률 =  $\frac{2}{5}$

- 지니지수 :  $1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = \frac{12}{25}$

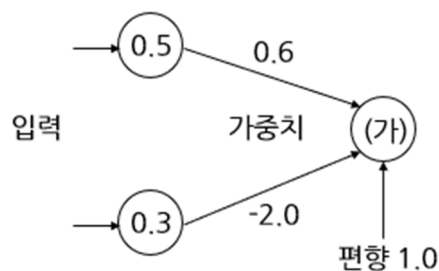
- 엔트로피지수 :  $-\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right)$

- 인공지능망

● 인공지능망

- 인간의 뇌 구조를 모방한 퍼셉트론을 활용한 추론모델
- 1) 단층 신경망 : 입력층과 출력층으로 구성 (단일 퍼셉트론)
- 2) 다층 신경망 : 입력층과 출력층 사이에 1개 이상의 은닉층 보유 (다층 퍼셉트론)
  - 은닉층 수는 사용자가 직접 설정

다층 퍼셉트론 구조에서 출력 값 계산



- (가) 퍼셉트론의 출력 :  $W_1X_1 + W_2X_2 + b = (0.6) \times (0.5) + (-2.0) \times (0.3) + (1.0) = 0.7$

☞ 가중치는 각 퍼셉트론 간의 연결 강도를 의미

## ● 활성화 함수와 손실함수

- (1) 은닉층에서의 활성화함수
  - 인공신경망의 선형성을 극복 (XOR 문제 해결)
    - 1) 시그모이드 함수 : 0 ~ 1 사이의 확률 값을 가지며, 로지스틱 회귀 분석과 유사
    - 2) 하이퍼볼릭 탄젠트(Tanh) 함수 : -1 ~ 1 사이 값, 시그모이드 함수의 최적화 지연을 해결
    - 3) ReLU 함수 : 기울기 소실문제를 극복,  $\max(0, x)$
    - 4) 그 외 활성화함수 : Leaky RELU, GELU, ELU 등
- (2) 출력층에서의 활성화함수
  - 1) 시그모이드 함수 : 이진분류 모델 (0~1 사이 확률)
  - 2) 소프트맥스 함수 : 다중 분류 모델 (확률의 총합이 1)
- (3) 손실함수
  - 예측값과 실제값의 차이를 측정하는 함수
    - 1) MSE(Mean Square Error) : 회귀 모델
    - 2) 크로스 엔트로피(Cross-Entropy) : 분류 모델

## ● 인공신경망 학습 방법

- (1) 순전파(피드포워드) : 정보가 전방으로 전달
- (2) 역전파 알고리즘 : 가중치를 수정하여 손실함수의 값을 줄임 (합성함수의 곱 활용)
- (3) 경사하강법
  - 경사의 내리막길로 이동하여 오차가 최소가 되는 최적의 해를 찾는 기법 (편미분 활용)
- (4) 기울기 소실 문제
  - 다수의 은닉층에서 시그모이드 함수 사용 시, 학습이 제대로 되지 않는 문제

## ● 인공신경망의 과적합 방지방안

- (1) 규제 : 라쏘(L1) 규제, 릿지(L2) 규제
- (2) 드롭아웃(Dropout) : 일부 퍼셉트론을 비활성화시켜 학습
- (3) 조기종료 : 특정 지점에서 학습 미리 종료
- (4) 모델의 복잡도 줄이기 : 은닉층의 퍼셉트론 수 감소
- (5) 데이터 증강 : 데이터에 변형을 주어 데이터 수 증가
- (6) 배치정규화 : 각 출력마다 정규화

## - 서포트벡터머신

### ● 서포트벡터머신(SVM)

- 마진이 최대가 되는 초평면을 찾아 선형이나 비선형 이진 분류, 회귀에서 활용 가능한 다목적 모델
- (1) 하이퍼플레인(초평면) : 데이터를 구분하는 기준이 되는 경계, 가중치벡터와 편향으로 결정
- (2) 서포트벡터 : 클래스를 나누는 하이퍼플레인과 가까운 위치의 샘플
- (3) 마진 : 하이퍼플레인과 서포트벡터 사이의 거리
- (4) 커널함수 : 저차원 데이터를 고차원 데이터로 변경하는 함수

## ● 서포트벡터머신(SVM)의 유형

- (1) 하드마진분류 : 오류 비허용
- (2) 소프트마진분류 : 마진 내 어느 정도 오류 허용

## - 연관성분석

### ● 연관분석

- 항목들간의 조건-결과로 이루어지는 패턴을 발견하는 기법 (**장바구니 분석**)
- (1) 특징
  - 결과가 단순하고 분명 (IF~THEN~)
  - 강력한 **비목적성 분석**기법
  - 품목 수가 증가할수록 계산량이 기하급수적으로 증가
  - **Apriori 알고리즘(최소 지지도 활용 빈발항목집합 추출)**을 활용 후, 연관분석을 수행
- (2) 순차패턴
  - : 연관분석에 **시간 개념을 추가**하여 품목과 시간에 대한 규칙 찾는 기법

### ● 연관분석의 지표

- (1) 지지도 :  $\frac{N(A \cap B)}{\text{전체}} = P(A \cap B)$ 
  - A와 B 두 품목이 동시에 포함된 거래 비율
- (2) 신뢰도 :  $\frac{P(A \cap B)}{P(A)}$ 
  - A 품목이 거래 될 때 B품목도 거래될 확률 (**조건부 확률**)
- (3) 향상도 :  $\frac{P(A \cap B)}{P(A)P(B)}$ 
  - A 품목과 B 품목의 상관성
  - (향상도 > 1 : 양의상관관계, 향상도 = 1 : 상관없음, 향상도 < 1 : 음의상관관계)

☞ **'지신향'**

- 맥주를 구매할 때 치킨을 구매하는 확률에 대한 신뢰도와 향상도

거래코드	품목	거래 횟수
1	맥주	10
2	치킨	20
3	햄버거	70
4	맥주, 치킨	20
5	맥주, 햄버거	30
6	치킨, 햄버거	10
7	맥주, 치킨, 햄버거	40

- (1) 맥주의 구매 확률 =  $(10 + 20 + 30 + 40) / 200 = 0.5$
- (2) 치킨의 구매 확률 =  $(20 + 20 + 10 + 40) / 200 = 0.45$

- (3) 맥주와 치킨의 지지도 =  $(20 + 40) / 200 = 0.3$   
 (4) 맥주 → 치킨의 신뢰도 =  $0.3 / 0.5 = 0.6$   
 (5) 맥주와 치킨의 향상도 =  $0.3 / (0.5 * 0.45) = 1.33$   
 - 맥주와 치킨의 향상도가 1보다 크므로 양의 상관관계를 가짐

## - 군집분석

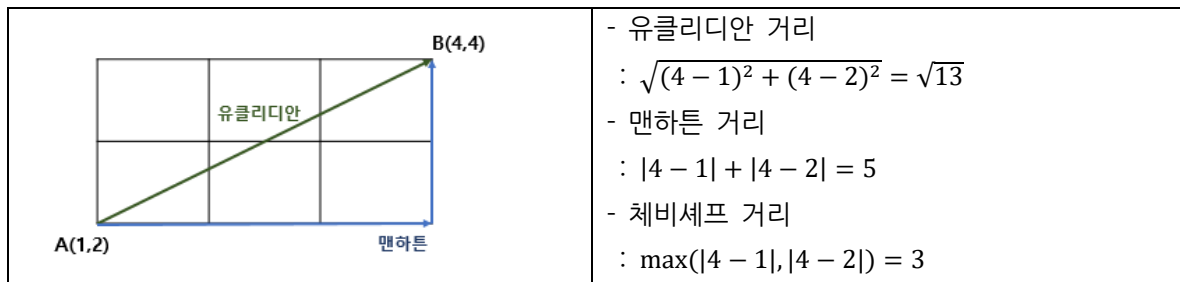
### ● 군집분석

- 비지도 학습으로 데이터들 간 거리나 유사성을 기준으로 군집을 나누는 분석

### ● 거리측도

#### (1) 연속형 변수

- 유클리디안 거리 : 두 점 사이의 직선 거리
- 맨하튼 거리 : 각 변수들의 차이의 단순 합
- 체비셰프 거리 : 변수 거리 차 중 최댓값
- 표준화 거리 : 데이터를 표준화 후에 거리를 계산
- 민코우스키 거리 : 유클리드, 맨하튼 거리를 일반화한 거리
- 마할라노비스 거리 : 표준화 거리에서 변수의 상관성 고려



#### (2) 범주형 변수

- 자카드 유사도 : 두 집합이 겹쳐져 있는 정도
- 코사인 유사도 : 두 집합간의 각도의 유사성
- ☞ '맨체스터 유나이티드' '자코'

### ● 계층적 군집분석

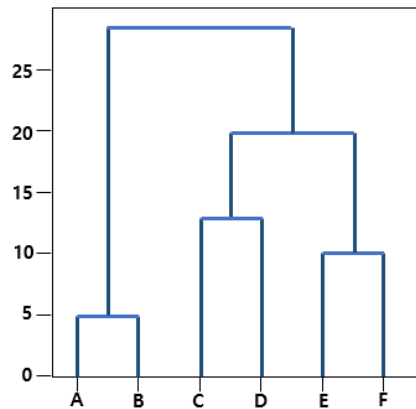
#### (1) 거리측정 방법

- 1) 최단 연결법(단일 연결법) : 군집간 가장 가까운 데이터
- 2) 최장 연결법(완전 연결법) : 군집간 가장 먼 데이터
- 3) 평균 연결법 : 군집의 모든 데이터들의 평균
- 4) 중심 연결법 : 두 군집의 중심
- 5) 와드 연결법 : 두 군집의 편차 제곱합이 최소가 되는 위치

#### (2) 덴드로그램

- 계층적 군집화를 시각적으로 나타내는 Tree모양의 그래프





- 거리를 15에서 나누면 3개의 클러스터, 25에서 나누면 2개의 클러스터로 나눌 수 있음

### ● K평균 군집화(K-means Clustering)

- 비계층적 군집화 방법으로 **거리기반**

#### (1) 특징

- 안정된 군집은 보장하나 최적의 보장은 어려움
- 한번 군집에 속한 데이터는 **중심점이 변경되면 군집이 변할 수 있음**
- 초기 중심 값에 따라 결과가 달라짐

#### (2) 과정

- 1) 군집의 개수 K개 설정 (**Elbow Method를 활용** 최적의 K 설정)
- 2) 초기 중심점 설정
- 3) 데이터들을 가장 가까운 군집에 할당
- 4) 데이터의 평균으로 중심점 재설정
- 5) 중심점 위치가 변하지 않을 때까지 3), 4)번 과정 반복

(3) K-medoids 군집화 (=PAM) : 평균 중심점이 아닌, 실제 데이터 중 하나인 대표(Medoids)를 설정

### ● DBSCAN

- 비계층적 군집화 방법으로 **밀도기반**

- **군집 개수 K는 지정할 필요 없음**, 노이즈와 이상치에 강함

### ● 기타 비계층적 군집분석

#### (1) 퍼지군집화 - 확률 기반

- 각 데이터가 특정 군집에 속할 확률을 각각 계산해가며 군집화

#### (2) EM알고리즘 - 분포 기반

- Likelihood의 기댓값을 계산하는 E단계와 기댓값 최대화 추정값을 계산하는 M단계 반복

#### (3) 자기조직화지도(SOM) - 그래프 기반

- **신경망을 활용**하여 차원축소를 통해 지도로 형상화하여 군집화하는 방법
- 완전연결의 형태를 가지며, **순전파 방식**만 사용

## 2. 고급 분석기법

### - 범주형 자료 분석

#### ● 분할표

- 여러 개의 범주형 변수를 기준으로 관측치를 기록하는 표
- 상대 위험도(RR) =  $\pi_1 / \pi_2$
- 오즈비(Odds Ratio) =  $Odds_{P_1} / Odds_{P_2} = \left( \frac{P_1}{1-P_1} \right) / \left( \frac{P_2}{1-P_2} \right)$

	심장질환 True	심장질환 False	합
알콜중독 True	14 (0.7)	6 (0.3)	20
알콜중독 False	10 (0.2)	40 (0.8)	50

- 상대 위험도(RR) =  $\pi_1 / \pi_2 = 0.7 / 0.2 = 3.5$   
 - Odds Ratio =  $Odds_{P_1} / Odds_{P_2} = (0.7/0.3) / (0.2/0.8) = 9.33$

#### ● KNN(K-Nearest Neighbors)

- 거리 기반으로 이웃에 더 많은 데이터가 포함되어 있는 범주로 분류
- 단순하고 효율적이며, 훈련이 따로 필요 없는 Lazy Model
- K에 따라 결과가 달라짐

### - 다변량 분석

#### ● 공분산 분석

- (1) 분산분석(ANOVA) : 여러 집단의 평균을 비교하여, 차이가 통계적으로 유의미한지 확인
  - 일원분산분석(One-Way ANOVA) : 하나의 요인과 여러 집단간의 분석
  - 이원분산분석(Two-Way ANOVA) : 두 개의 요인과 여러 집단간의 분석
- (2) 다변량분산분석(MANOVA) : 다수의 요인과 여러 집단간의 분석

#### ● 요인분석 (Factor Analysis)

- 다수 변수들을 상관관계를 분석하여 소수의 요인으로 축약하는 기법
- (1) 요인추출방법
  - 주성분분석(PCA) : 전체 분산을 토대로 요인을 추출, 가장 많이 사용
  - 주축요인분석 : 공통 분산만을 토대로 요인을 추출
  - 최소제곱요인분석 : 잔차를 최소화하여 요인을 추출
- (2) 요인회전
  - 직각회전 : VARIMAX, QUARTIMAX, EQUIMAX
  - 사각회전 : OBLIMIN, PROMAX

## - 시계열 분석

### ● 시계열 분석

- 시간의 흐름에 따라 관찰된 자료의 특성을 파악하여 미래를 예측 (주가데이터, 기온데이터)

### ● 정상성

- 시계열 예측을 위해서는 **모든 시점에 일정한 평균과 분산**을 가지는 정상성을 만족해야 함
- 정상시계열로 변환 방법
  - 1) **차분** : 현 시점의 자료를 이전 값으로 빼는 방법
  - 2) 이동평균법 : 일정 기간의 평균
  - 3) 지수평활법 : 최근 시간 데이터에 가중치를 부여
  - 4) 그 외 정상화 방법 : 지수변환, 로그변환, Box-Cox 변환 등

### ● 백색 잡음

- 시계열 모형의 오차항을 의미 (평균 및 분산 일정, 자기상관 없음)
- 평균이 0 이면 가우시안 백색잡음

### ● 시계열 모형

- (1) 자기회귀(AR) 모형
  - 자기자신의 과거 값이 미래를 결정하는 모형
  - 부분자기상관함수(PACF)를 활용하여  $p+1$  시점 이후 급격 감소하면 AR( $p$ ) 모형 선정
- (2) 이동평균(MA) 모형
  - 이전 백색잡음들의 선형결합으로 표현되는 모형
  - 자기상관함수(ACF)를 활용하여  $q+1$  시차 이후 급격히 감소하면 MA( $q$ ) 모형 선정
- (3) 자기회귀누적이동평균(ARIMA) 모형
  - AR 모형과 MA 모형의 결합
  - **ARIMA( $p, d, q$ )**
    - 1)  $p$ 와  $q$ 는 AR 모형과 MA 모형이 관련 있는 차수
    - 2)  $d$ 는 **정상화시에 차분 몇 번 했는지** 의미
    - 3)  $d = 0$  : ARMA 모델 /  $p = 0$  : IMA 모델 /  $q = 0$  : ARI 모델

### ● 분해시계열

- 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법
  - (1) **추세** 요인 : 장기적으로 증가, 감소하는 추세
  - (2) **계절** 요인 : 계절과 같이 고정된 주기에 따라 변화
  - (3) **순환** 요인 : 알려지지 않은 주기를 갖고 변화 (**경제 전반, 특정 산업**)
  - (4) **불규칙** 요인 : 위 3 가지로 설명 불가한 요인
- ☞ '추운 계절의 순환이 불규칙하다.'

## - 베이지안 기법

### ● 베이지 정리

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)} \quad (P(A|B): \text{사후확률}, P(B|A): \text{우도}, P(A): \text{사전확률}, P(B): \text{주변우도})$$

A대학 입시에 응시한 남학생과 여학생의 비율이 60%와 40%이고 남학생의 합격률은 30%, 여학생의 합격률은 50%이다. 이때, A대학에 합격한 신입생 중 남학생을 고를 확률은?

우도표	합격률	불합격률	
남학생	$0.6 \times 0.3$ = 0.18	$0.6 \times 0.7$ = 0.42	0.6
여학생	$0.4 \times 0.5$ = 0.20	$0.4 \times 0.5$ = 0.20	0.4

$$P(A|B) = P(\text{남학생} | \text{합격한 신입생}) = \frac{P(A \cap B)}{P(B)} = \frac{0.18}{0.38} = 0.47$$

### ● 나이브베이지 분류

- 나이브(독립) + 베이지 정리를 기반으로 계산을 단순화하여 범주에 속할 확률 계산
- 서로 독립적이라는 가정이 필요
- 과거의 경험을 활용하는 귀납적인 추론 방법

## - 딥러닝 분석

### ● DNN (심층 신경망)

- 은닉층이 2개 이상으로 구성된 인공신경망 (입력층 - 은닉층 - 출력층)

### ● CNN (합성곱 신경망)

- Convolution Layer와 Pooling Layer를 활용하여 이미지에서 패턴을 찾는 신경망
- (1) 구조 : Input - Convolution Layer - Pooling Layer - Flatten - Fully Connected Layer
  - 1) Convolution Layer : Convolution 연산을 이미지의 특징을 추출 (Filtering)
  - 2) Pooling Layer : 데이터의 공간적 특성은 유지하면서 크기를 줄여 연산량을 감소
  - 3) Flatten : 2D/3D의 데이터를 1D로 변환
  - 4) Fully Connected Layer : DNN구조를 가지며, 분류 또는 회귀를 수행
- (2) 주요 모델
  - 1) 분류 : LeNet, AlexNet, VGG Nets, GoogLeNet, ResNet, EfficientNet
  - 2) 객체탐지 : RCNN, SPP Net, YOLO, Attention Net, EfficientDet
  - 3) 분할(세그멘테이션) : FCN, DeepLab, U-net, Segnet

## ● RNN (순환 신경망)

- 순차적인 데이터 학습에 특화된 순환구조를 가지는 신경망

### (1) 장기 의존성 문제

- 은닉층의 과거 정보가 전달되지 못하는 현상

### (2) 장기 의존성 극복 모델

1) **LSTM** : Forget Gate, Input Gate, Output Gate

2) **GRU** : Reset Gate, Update Gate

## ● 오토인코더

- 입력 데이터를 인코더로 압축한 후에 디코더로 형태를 재구성하는 비지도 학습 신경망

(1) 구조 : Encoder - **Context Vector**(=Latent Space) - Decoder

### (2) 활용

#### 1) 생성모델

- VAE : **확률분포**를 학습하여 데이터 생성

- GAN : 생성기와 판별기의 경쟁을 유사한 데이터를 생성 (**적대적 훈련**)

- DCGAN : GAN + CNN으로 안정적으로 학습

☞ **오토 인코더는 생성형 AI의 기반 모델**

2) 이상탐지 : 정상 데이터만 학습하여, 비정상 데이터를 판단

## - 비정형 데이터 분석

### ● 워드 임베딩

#### (1) 희소표현

1) 원핫 인코딩 : 메모리가 낭비되며, 단어 상호간 의미를 담지 못함

#### (2) 카운트 기반

1) BOW : 단어의 등장 개수 기반의 표현

2) TDM : 문서에서 등장하는 단어들의 빈도를 행렬로 표현

3) TF-IDF

- TF : 1개의 문서 내에서 특정 단어의 출현 빈도

- IDF : 특정 단어가 전체 문서에 등장하는 정도

4) LSA : SVD(특이값 분해)를 활용하여 잠재적인 의미 반영

#### (3) 단어 수준 기반

1) **Word2Vec** : 거리를 기반으로 벡터로 표현

- CBOW : 앞, 뒤 단어로 주어진 단어를 유추

- Skip-Gram : 중심단어에서 주변단어 예측

2) Glove : 전체 문장의 통계정보를 반영

3) FastText : 하나의 단어를 여러 개로 잘라서 벡터로 계산

4) ELMo : 양방향 언어 모델을 적용

## ● Seq2Seq

- 인코더와 디코더를 활용하여 한 문장을 다른 문장으로 번역하는 모델
- **입력과 출력의 길이가 달라도 변환이 가능**
- **긴 문장에서 정보의 손실이 발생**

## ● 트랜스포머

- 느린 속도와 병렬처리 불가 및 정보 손실의 단점을 개선한 **Attention 모델**
- (1) 구성요소 : Positional Encoding, Self-Attention, Feed Forward Network
- (2) 주요 모델
  - 1) BERT : 구글 개발, 인코더 구조, 문장 중간 빈칸 학습, 양방향
  - 2) GPT : OpenAI 개발, 디코더 구조, 이전 단어로 다음 단어 예측, 일방향

## ● 기타 비정형 데이터 분석

- (1) 유전자 알고리즘 : **최적화** 필요한 문제의 해결책
  - 활용 : 택배차량 어떻게 배치, 최대 시청률 얻기 위한 프로그램 방송 배치
  - 처리과정 : 초기화 → 적합도 → 선택 → 교차 → 변이 → 대체
    - 1) 초기화 : 여러 개의 해를 염색체로 표현하여 초기 개체군 생성
    - 2) 적합도 : 각 염색체의 적합도를 평가
    - 3) 선택 : 적합도가 높은 개체를 선택 (룰렛 휠 선택, 순위 선택, 토너먼트 선택)
    - 4) 교차 : 부모의 유전자를 섞어 새로운 자손 생성 (단일점 교차, 두점 교차, 균등 교차)
    - 5) 변이 : 유전자를 변경하여 다양성을 유지 (이진 인코딩, 순열 인코딩, 실수값 인코딩)
    - 6) 대체 : 새로운 염색체를 기존 염색체와 교체하면서 세대를 반복
- (2) 소셜 네트워크 분석 : 노드와 엣지간의 관계로 표현
  - 활용 : 사람들간의 관계 SNS상 사용자들 관계 속 영향력 높은 사람 찾기
  - 연결 중심성 : 연결 중심성, 매개 중심성, 위세 중심성, 근접 중심성
- (3) 감정분석 : 텍스트 데이터에서 감정(긍정/부정)을 분석

## - 앙상블 분석

### ● 앙상블

- 여러 개의 예측 모형들을 조합하는 기법으로 **전체적인 분산을 감소**시켜 성능 향상이 가능
- (1) 보팅(Voting)
  - 다수결 방식으로 최종 모델을 선택
- (2) 배깅(Bagging)
  - 복원추출에 기반을 둔 **붓스트랩**을 생성하여 모델을 학습 후에 보팅으로 결합
  - 복원추출을 무한히 반복할 때 특정 하나의 데이터가 선택되지 않을 확률
 
$$\rightarrow \lim_{N \rightarrow \infty} (1 - \frac{1}{N})^N = 36.8\%$$
- (3) 부스팅(Boosting)
  - 잘못된 분류 데이터에 큰 가중치를 주는 방법, 이상치에 민감
  - 종류 : AdaBoost, GBM, XGBoost(**GBM보다 빠르고 규제 포함**), Light GBM(학습속도 개선)

(4) 스택킹(Stacking)

- 각각의 모델에서 학습한 예측 결과를 다시 학습

(5) 랜덤포레스트

- 배깅에 의사결정트리를 추가하는 기법으로 성능이 좋고 이상치에 강한 모델

☞ 보팅, 배깅, 랜덤포레스트는 병렬처리가 가능하며, 부스팅은 병렬처리가 불가

## - 비모수 통계

### ● 비모수 검정

- 모집단에 대한 아무런 정보 없어, 관측 자료가 특정 분포를 따른다고 가정 불가 시 검정
- 두 관측 값의 순위나 차이로 검정
- 부호검정, 순위합검정, 만-휘트니 U검정, 크러스칼-월리스 검정, 프리먼드 검정, 카이제곱 검정

## 4과목 빅데이터 결과해석



## - 4 과목(빅데이터 결과해석) 주요 내용

분석모형 평가 및 개선	분석모형 평가	평가 지표
		분석모형 진단
		교차 검증
		모수 유의성 검정
		적합도 검정
	분석모형 개선	과대적합 방지
		매개변수 최적화
		분석모형 융합
		최종모형 선정
분석결과 해석 및 활용	분석결과 해석	분석모형 해석
		비즈니스 기여도 평가
	분석결과 시각화	시공간 시각화
		관계 시각화
		비교 시각화
		인포그래픽
	분석결과 활용	분석모형 전개
		분석결과 활용 시나리오 개발
		분석모형 모니터링
		분석모형 리모델링

## 분석모형 평가 및 개선

### 1. 분석모형 평가

#### - 평가 지표

##### ● 분류모델 평가지표

##### (1) 혼동행렬(오분류표)

		실제	
		TRUE	FALSE
예측	TRUE	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)
	FALSE	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)

예측과 실제가 같으면 TRUE, 예측이 TRUE면 POSITIVE

##### (2) 평가지표

지표	계산식
정밀도(Precision)	$\frac{TP}{TP + FP}$
재현율(Recall)	$\frac{TP}{TP + FN}$
특이도(Specificity)	$\frac{TN}{FP + TN}$
정확도(Accuracy)	$\frac{TP + TN}{TP + FP + FN + TN}$
FP Rate (False Alarm Rate)	$\frac{FP}{FP + TN}$
F-1 Score	$2 \times \frac{Precision \cdot recall}{Precision + recall}$
F-β Score	$(1 + \beta^2) \times \frac{Precision \cdot recall}{(\beta^2 \cdot Precision) + recall}$

1) 재현율(Recall)은 민감도(Sensitivity), TP Rate, Hit Rate라고도 함

2) F-1 Score는 Precision과 Recall의 조화평균

3) Precision과 Recall은 Trade-Off 관계

4) F-β Score

- $\beta > 1$  : 재현율(Recall)에 큰 비중
- $\beta < 1$  : 정밀도(Precision)에 큰 비중
- $\beta = 1$  : F-1 Score와 동일

##### (3) ROC 커브

- 가로축을 1-특이도(FPR), 세로축을 민감도(TPR)로 두어 시각화한 그래프
- 그래프 면적(AUC)은 0.5~1사이이며, 1에 가까울수록 모델의 성능이 좋다고 평가

##### (4) 이익도표(Lift chart)

- 임의로 나눈 각 등급별로 반응검출율, 반응률, 리프트 등의 정보를 산출하여 나타내는 도표
- 향상도 곡선 : 이익도표를 시각화한 곡선

## ● 회귀모델 평가지표

### (1) 손실함수(비용함수)

- 1) MSE(Mean Squared Error) =  $\frac{1}{n} \sum (y - \hat{y})^2$
- 2) MAE(Mean Absolute Error) =  $\frac{1}{n} \sum |y - \hat{y}|$
- 3) RMSE(Root-Mean Squared Error) =  $\sqrt{MSE}$
- 4) MAPE(Mean Absolute Percentage Error) =  $\frac{1}{n} \sum \left| \frac{y - \hat{y}}{y} \right| \times 100$
- 5) MPE(Mean Percentage Error) =  $\frac{1}{n} \sum \left( \frac{y - \hat{y}}{y} \right) \times 100$

### (2) 결정계수

- 1) 결정계수(R-Square) =  $\frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- 2) 수정된 결정계수(Adjusted R-square) =  $1 - (n - 1) \frac{MSE}{SST}$

## ● 군집분석의 평가지표

### (1) 실루엣 계수

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (-1 \sim 1)$$

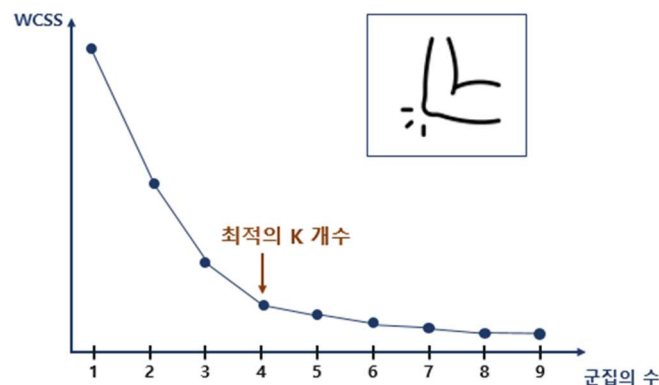
- $a(i)$  : i번째 데이터에서 자신이 속한 군집 내 다른 데이터 포인트 들과의 평균 거리
- $b(i)$  : i번째 데이터에서 가장 가까운 다른 군집 내 다른 데이터 포인트 들과의 평균 거리

### (2) WCSS (Within Clusters Sum of Squares)

$$WCSS = \sum_{C_k}^n \left( \sum_{d_i \in C_k}^{d_m} distance(d_i, C_k)^2 \right)$$

- C : 군집의 중심 값
- d : 클러스터 내에 있는 데이터

- **Elbow 기법** : 최적의 군집 개수(K) 선택 방법



- 경사가 완만해지는 지점이 최적의 K 개수

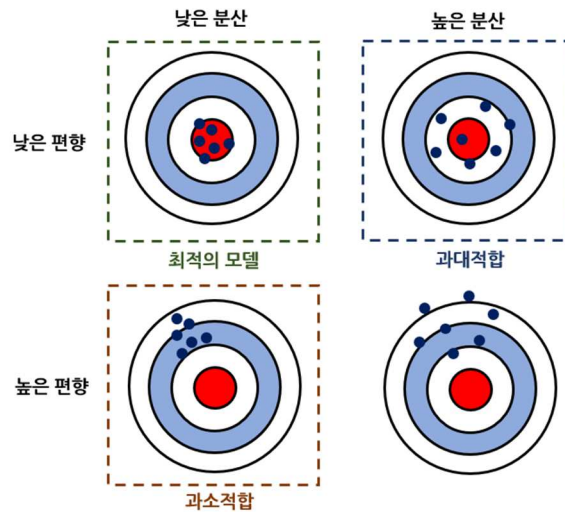
☞ 같은 군집간에는 응집도가 높고, 다른 군집간에는 응집도가 낮아야 좋은 모델

## - 분석모형 진단

### ● 과대적합과 과소적합

- (1) 과대적합(Overfitting) : 높은 분산, 낮은 편향
- (2) 과소적합(Underfitting) : 낮은 분산, 높은 편향

## ☞ 분산과 편향이 모두 낮은 모델이 최적의 모델



## - 교차 검증

### ● 교차 검증

- (1) 홀드아웃 : 훈련용과 평가용 **2개의 셋**으로 분할
- (2) K-fold 교차검증 : 데이터를 k개의 집단으로 구분하여 **k-1개 학습, 나머지 1개로 평가**
- (3) LOOCV : **1개의 데이터로만 평가**, 나머지로 학습
- (4) 부트스트래핑 : 복원추출을 활용하여 데이터 셋을 생성, 데이터 부족, 불균형 문제 해소

## - 적합도 검정

### ● 적합도 검정

- (1) Q-Q plot : 데이터의 정규성을 **분포의 분위수를 활용**하여 시각적으로 파악  
- 정규성 존재한다면, 점들이 대각선(45도) 위에 분포
- (2) 카이제곱 검정 : 변수를 **범주**로 묶어 적합성에 대한 검정
- (3) 샤피로 윌크 검정 : 데이터의 표준정규점수와 선형상관관계를 측정하여 검정
- (4) 콜모고로프-스미르노프 검정 : **누적 분포함수**를 비교하여 **연속형** 데이터의 검정

## - 모수 유의성 검정

### ● 모집단 유의성 검정

- (1) Z-검정 : 표본이 모집단에 속하는 검정
- (2) T-검정 : 두 집단의 평균치 차이의 비교 검정 시 사용
- (3) ANOVA 검정 : 셋 이상의 집단에 대한 평균 검정
- (4) 카이제곱 검정 : 두 집단의 동질성 검정, 혹은 단일 집단 모분산에 대한 검정
- (5) F-검정 : 두 집단의 분산의 동일성 검정 시 사용

## ※ 두 학교의 학생들의 수학 점수에 대한 t검정

```
> t.test(schoolA, schoolB, conf.level=0.95)

Welch Two Sample t-test

data:  schoolA and schoolB
t = -0.59758, df = 97.409, p-value = 0.5515

alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.293157  5.528998

sample estimates:
mean of x      mean of y 
 61.91373     64.29581
```

- 귀무가설 : 두 학교의 성적의 평균은 동일하다
- 대립가설 : 두 학교 성적의 평균이 다르다
- 대립가설의 값이 같지 않다 → 양측검정
- 두개의 모집단 → 이표본
- p-value : 0.5515 > 유의수준( $\alpha$ ) : 0.05 → 귀무가설 채택
- 신뢰구간(95%) → [-10.293157, 5.528998]
- 서로 다른 모집단에 대한 평균비교 검정 → 독립표본

## 2. 분석모형 개선

### - 과대적합 방지

#### ● 과대적합 방지방안

- (1) 모델의 복잡도 감소 : 경량화된 모델의 활용, 드롭아웃(Dropout)
- (2) 규제
  - 1) 릿지(L2) 규제 -  $L2 : \sum W^2$  (유클리디안 거리 기반)
  - 2) 라쏘(L1) 규제 -  $L1 : \sum |W|$  (맨하탄 거리 기반)
- (3) 편향-분산의 트레이드오프 확인

### - 매개변수 최적화

#### ● 파라미터(매개변수)와 하이퍼파라미터(초매개변수)

- (1) 파라미터 : 최적화된 모델 구현이 목적이며, 수동으로 설정 불가
- (2) 하이퍼파라미터 : 최적의 파라미터 도출이 목적이며, 사용자가 수동으로 설정

## ● 경사하강법

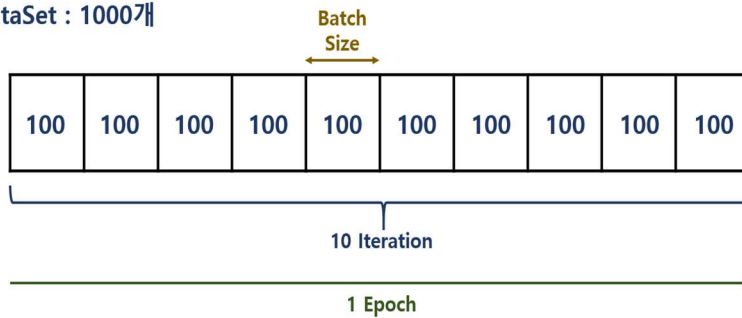
- 모델 최적화를 위해 손실함수의 최솟값을 찾아가는 반복 알고리즘

$$w_{new} = w_{old} - \alpha \frac{d}{dw} cost(W)$$

- (1)  $\alpha$ 가 너무 클 때 : 이전보다 높은 곳으로 발산
- (2)  $\alpha$ 가 너무 작을 때 : 수렴하는데 시간이 오래 걸림

## ● Batch Size, Epoch, Iteration

DataSet : 1000개



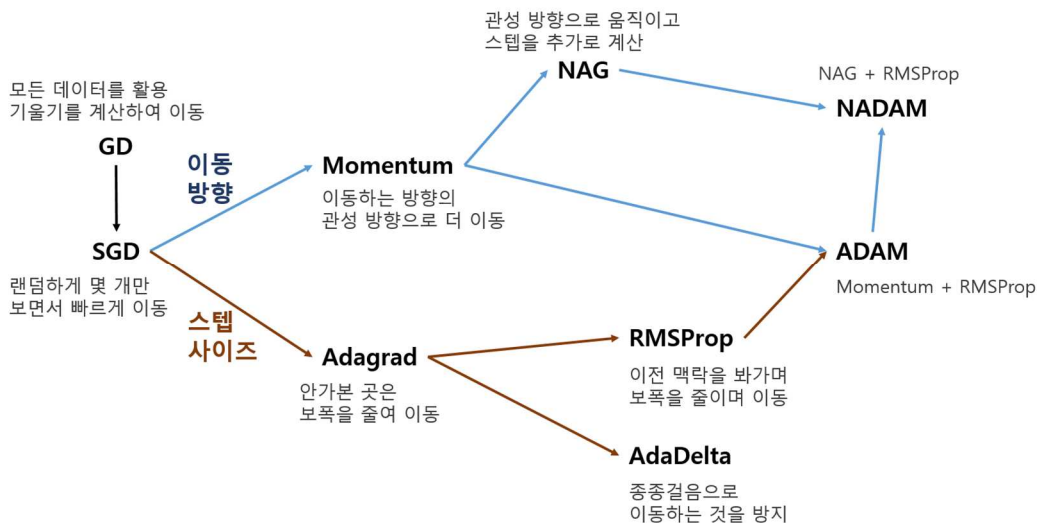
- Batchsize : 하나의 소그룹에 속하는 데이터 수
- Epoch : 모든 데이터셋을 학습하는 횟수
- Iteration : 1 Epoch를 마치는데 필요한 배치

## ● 하이퍼파라미터 튜닝

- (1) Manual Search : 경험 또는 감으로 설정
- (2) Grid Search : 모든 조합을 탐색
- (3) Random Search : 임의로 탐색
- (4) Bayesian Optimization : 모델링을 통한 최적의 조합 발견

## ● 경사하강법 옵티마이저

- 옵티마이저를 통하여 경사하강법의 학습속도 및 지역 극솟값 수렴 문제 해결 가능



## 분석결과 해석 및 활용

### 1. 분석결과 해석

#### - 분석모형 해석

#### ● 주성분 분석 (PCA)

##### (1) 주성분 분석의 결과 해석

- 상관성 높은 변수들의 선형 결합으로 차원을 축소하여 새로운 변수를 생성
- 자료의 **분산이 가장 큰 축이 첫 번째 주성분**
- **70 ~ 90%의 설명력**을 갖는 개수로 결정

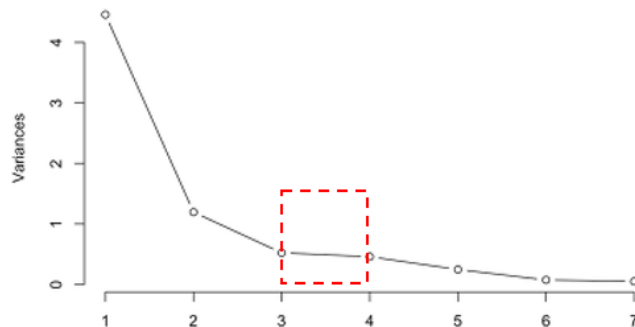
```
> result<-prcomp(data,center=T,scale.=T)
> summary(result)
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.1119	1.0928	0.72181	0.67614	0.49524	0.27010	0.2214
Proportion of Variance	0.6372	0.1706	0.07443	0.06531	0.03504	0.01042	0.0070
Cumulative Proportion	0.6372	0.8078	0.88223	0.94754	0.98258	0.99300	1.0000

- Center=T: 평균을 0, Scale.=T: 데이터의 표준화 수행
- 첫번째 주성분(PC1)의 분산(0.6372)이 가장 큼
- 두 개의 주성분(PC1, PC2)을 적용하면 **전체 데이터의 약 80%를 설명**

##### (2) 스크리플롯(Screplot)

- 주성분의 개수를 선택하는데 도움이 되는 그래프 (x축 주성분 개수, y축 분산변화)
- **수평을 이루기 바로 전 단계 개수로 선택**



- 기울기가 3-4 구간에서 완만해지므로 **주성분 개수는 2개로 선택**

☞ 주성분 개수의 선택은 절대적인 것은 없으며, 연구자의 판단 (3개로 선택도 가능)

#### ● 회귀분석의 분산분석(ANOVA)표

요인	제곱합	자유도	제곱평균	F비
회귀	$SSR = \sum(\hat{Y} - \bar{Y})^2$	p(회귀계수 수)	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
잔차	$SSE = \sum(Y - \hat{Y})^2$	n(전체 데이터 수) - p - 1	$MSE = \frac{SSE}{n - p - 1}$	
총	$SST = SSR + SSE$	n - 1		

- ANOVA 검정 : 3 개 이상의 그룹의 평균을 비교하는 검정 (회귀모형의 유의성 분석 시 활용)
- 전체 데이터 수 = 자유도 + 1
- 결정계수(R-Square) =  $SSR/SST$       - 수정된 R-square =  $1 - (n - 1)(MSE/SST)$
- ☞ 다중 회귀에서는 수정된 R-square 값을 일반적으로 사용

## ● 회귀 모형의 검정결과

- 1) 모형이 통계적으로 유의미한가 : 모형에 대한 F통계량, p-value
  - 귀무가설 : '모든 회귀계수는 0이다'
- 2) 회귀계수들이 유의미한가 : 회귀계수들의 t통계량, p-value
  - 각각의 회귀계수에 대한 귀무가설 : '회귀 계수는 0이다'
- 3) 위 1), 2) 모두를 기각하면 해당 모델을 활용
- 4) 모형이 설명력을 갖는가 : 결정계수(R square) 값

## ※ 일반적인 회귀 모형의 검정결과 해석

```
Call:
lm(formula = height ~ age + no_siblings, data = ageandheight)

Residuals:
    Min       1Q   Median       3Q      Max
-0.28029 -0.22490 -0.02219  0.14418  0.48350

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    64.95872    0.55752   116.515 1.28e-15 ***
age             0.63516    0.02254    28.180 4.34e-10 ***
no_siblings    -0.01137    0.05893    -0.193  0.851
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2693 on 9 degrees of freedom
Multiple R-squared:  0.9888, Adjusted R-squared:  0.9863
F-statistic: 397.7 on 2 and 9 DF, p-value: 1.658e-09
```

- 종속변수 height / 독립변수 age, no\_siblings
- 회귀모형 F분포의 P-value(1.658e-09)가 0.05보다 작으므로 모형이 유의미
- age의 p-value(4.34e-10)가 0.05보다 작으므로 회귀계수 유의미
- no\_siblings의 p-value(0.851)가 0.05보다 크므로 제외하고 회귀분석 재수행을 권장
- 위 모형은 다중회귀 모형
- R-square : 0.9888, Adjusted R-square : 0.9863 (모형은 전체 데이터의 98% 이상을 설명)
- 회귀 자유도 : 2, 잔차의 자유도 : 9 → 총 2 + 9 + 1 = 12개의 데이터를 활용하여 분석
- 모델 회귀 식 :  $Y_{height} = 0.63516X_{age} - 0.01137X_{no\_siblings} + 64.95872$

☞ no\_siblings 변수가 유의하지 않기에, 제거하고 검정을 다시 수행하는 것은 연구자의 판단



※ (심화) 교호항이 포함된 회귀 모형의 검정결과 해석

```
> summary(Wage[,c("wage", "age", "jobclass")])
```

wage		age		jobclass	
Min.	: 20.09	Min.	: 18.00	1. Industrial:	1544
1st Qu.:	85.38	1st Qu.:	33.75	2. Information:	1456
Median	: 104.92	Median	: 42.00		
Mean	: 111.70	Mean	: 42.41		
3rd Qu.:	128.68	3rd Qu.:	51.00		
Max.	: 318.34	Max.	: 80.00		

```
model <- lm(wage ~ age + jobclass + age * jobclass, data = Wage)

summary(model)
```

Call:

```
lm(formula = wage ~ age + jobclass + age * jobclass, data = Wage)
```

Residuals:

Min	1Q	Median	3Q	Max
-105.656	-24.568	-6.104	16.433	196.810

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	73.52831	3.76133	19.548	< 2e-16 ***
age	0.71966	0.08744	8.230	2.75e-16 ***
jobclass2. Information	22.73086	5.63141	4.036	5.56e-05 ***
age:jobclass2. Information	-0.16017	0.12785	-1.253	0.21

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.16 on 2996 degrees of freedom

Multiple R-squared: 0.07483, Adjusted R-squared: 0.07391

F-statistic: 80.78 on 3 and 2996 DF, p-value: < 2.2e-16

- 종속변수 Wage / 독립변수 age, jobclass, age\*jobclass(교호항)
- jobclass는 Information과 Industrial 2개의 클래스를 가진 범주형 변수
- Jobclass2. Information의 회귀계수 22.73086 : Information이 Industrial보다 임금 높음
- age:jobclass2. Information의 p-value(0.21)가 0.05보다 크므로 교호작용은 유의하지 않음

## - 비즈니스 기여도 평가

### ● 비즈니스 기여도 평가

#### (1) 재무적 평가

- 1) 투자 대비 효과(ROI) : 투자로 얻을 수 있는 순 효과를 비용으로 나눈 값
- 2) 순현재가치(NPV) : 미래의 현금 흐름을 현재가치로 계산
- 3) 내부 수익률(IRR) : NPV = 0일때의 수익률
- 4) 총 소유 비용(TCO) : 자산의 매입 가격과 운용 원가를 더한 금액

#### (2) 비즈니스 성과 평가

- KPI 기반 평가, 업무 자동화율, 생산성 증가율, 고객 만족도 등

#### (3) 비즈니스 기여도 평가 프로세스

- 분석 목표 설정 → 성과 지표(KPI) 선정 → 데이터 수집 및 분석 → 평가 결과 도출 및 피드백

## 2. 분석결과 시각화

### - 시공간 시각화

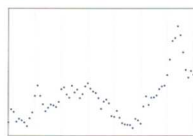
#### ● 시간 시각화

- 시간에 따른 변화를 표현 (x축 시간, y축 값)

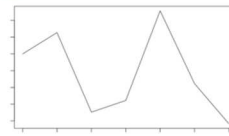
막대그래프



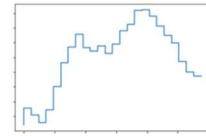
점그래프



선그래프



계단식 그래프



#### ● 공간 시각화

- 장소나 지역에 대한 데이터의 분포를 표현

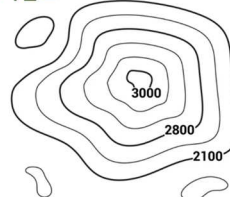
등치지역도



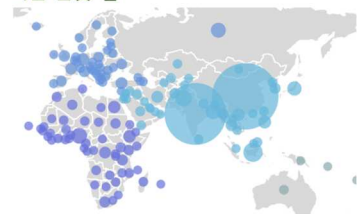
카토그램



등치선도



버블 플롯 맵

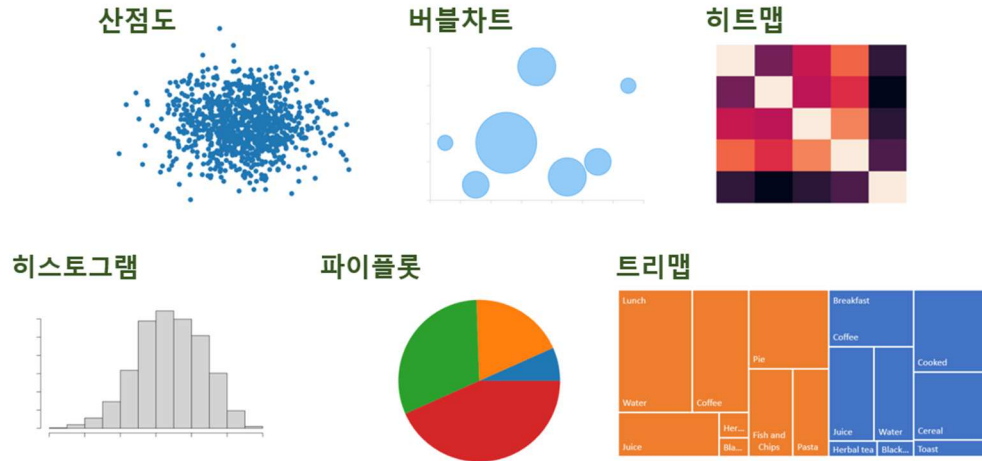


- (1) 등치지역도 : 영역에 따라 색의 채도를 이용하여 자료를 계급으로 집단화
- (2) 카토그램 : 데이터 값을 강조위해 면적을 통한 특정 지역 왜곡
- (3) 버블 플롯 맵 : 지도 위에 버블의 크기를 이용하여 데이터 값을 표현

## - 관계 시각화

### ● 관계 시각화

- 데이터 사이의 관계나 분포, 패턴을 표현

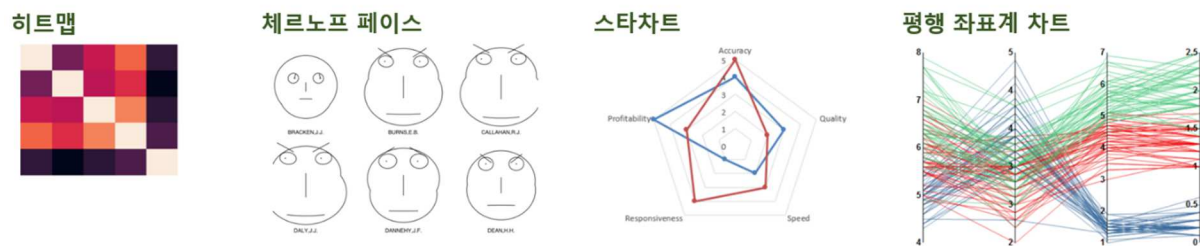


- (1) 버블차트 : 3개의 변수를 활용하며, 마지막 변수는 원의 크기로 표현
  - 데이터 좁은 공간에 표현 가능하나, 겹치면 이해하기 어려움
  - 면적으로 수량을 비교하여 시각적 왜곡 발생 가능
- (2) 히트맵 : 색상변화로 데이터 간의 관계를 표현
- (3) 트리맵 : 데이터의 비율과 계층구조를 시각적으로 표현
  - 음수 값을 표현하고 어렵고, 이웃하지 않은 사각형 사이 비교가 어려움

## - 비교 시각화

### ● 비교 시각화

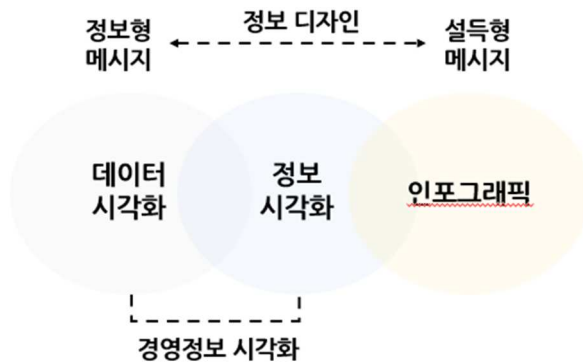
- 여러 변수들 간의 차이나 유사성을 비교



- (1) 체르노프 페이스 : 사람의 얼굴 형태로 변환하여 시각적으로 표현
- (2) 스타차트 : 3개 이상의 변수의 균형과 경향을 직관적으로 비교 가능
- (3) 평행 좌표계 차트 : 다차원 데이터에 대한 시각화

## - 인포그래픽

### ● 인포그래픽



- 정보(Information) + 시각적 형상(Graphic)
- 목적에 따라 정보를 시각적으로 표현
- 분석 시각화와는 달리 패턴을 발견하는 것 보다 일반인에게 설득형 메시지 전달 목적
- 데이터 가공에 대한 노력이 들어가며 원 데이터(Raw Data) 취급하지 않음

### ● 인포그래픽의 유형

- (1) 지도형 : 지도 활용
- (2) 도표형 : 표와 그래프 활용
- (3) 타임라인형 : 시간 순서로 나열
- (4) 스토리텔링형 : 이야기를 구성
- (5) 만화형 : 만화적 요소를 활용
- (6) 비교분석형 : 두 가지 이상의 내용을 비교
- (7) 컨셉 맵 : 내용 간의 연관성

### ● 인포그래픽의 원리

- (1) 기본원리 : 단순성, 명확성, 중요성, 일관성, 가독성, 효과성, 대상 독자
- (2) 오컴의 면도날 : 복잡한 설명보다 단순한 설명이 선호되어야 함

## 3. 분석결과 활용

### ● 분석결과 활용

- (1) 분석모형 전개 : 분석결과를 확장 및 적용
- (2) 분석결과 활용 시나리오 개발 : 인사이트 발굴, 결과를 업무에 반영
- (3) 분석모형 모니터링 : 변화를 지속적 반영 위해 모니터링
- (4) 분석모형 리모델링 : 새로운 알고리즘 및 데이터 반영