ADsP 요약노트

1과목 데이터 이해

1. 데이터 이해
2. 데이터의 가치와 미래
3. 가치장조를 위한 데이터 사이언스와 전략인사이트

2과목 데이터분석 기획

1. 데이터분석 기획의 이해
2. 분석 마스터플랜

3과목 데이터 분석

1. R기초와 데이터마트
2. 통계분석
3. 정형데이터 마이닝

1과목 데이터 이해

1. 데이터의 이해

- 데이터와 정보

• 데이터의 정의

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

• 데이터의 유형

(1) 정성적, 정량적

- 정량적 데이터 : 자료를 수치화 - 수치, 기호(온도, 풍속)
- 정성적 데이터 : 자료의 특징을 풀어 설명 - 언어, 문자(기상특보, 주관식 설문응답)

(2) 정형, 반정형, 비정형

- 정형 데이터 : 정보 형태가 정해짐(관계형 DB, 엑셀 - 스프레드시트, CSV)
- 반정형 데이터 : 데이터를 설명하는 메타데이터를 포함(HTML, XML, JSON, RDF)
- 비정형 데이터 : 형태가 정해지지 않음(SNS, 유튜브, 음원)

• 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

• 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 저렴할 것이다)

• 데이터 단위

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

- 데이터베이스의 정의와 특징

• 데이터 베이스의 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(MySQL, MariaDB, Oracle)
2) NoSQL DBMS : 비정형데이터를 저장하고 처리(HBase, MongoDB, CouchDB, Redis, Cassandra)

(3) SQL : 데이터베이스에 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) 공용데이터 : 여러 사용자가 다른목적으로 데이터 공동이용
(2) 통합된 데이터 : 동일한 데이터 중복되어 있지 않음
(3) 저장된 데이터 : 저장매체에 저장
(4) 변화되는 데이터 : 새로운 데이터 추가, 수정, 삭제에도 현재의 정확한 데이터 유지

• **데이터베이스의 설계절차**
(1) 요구조건 분석
(2) 개념적 설계 : 개념적 스키마 생성
(3) 논리적 설계 : 개념적 ERD를 활용한 논리적 모델링
(4) 물리적 설계 : 저장 구조 설계

- 데이터베이스 활용

• **기업활용 데이터베이스**
- OLTP : 데이터를 수시로 갱신(거래단위)
- OLAP : 다차원 데이터를 대화식으로 분석
- CRM : 고객과 관련 자료 분석, 마케팅 활용
- SCM : 공급망 연결 최적화
- ERP : 기업경영 자원을 효율화
- RTE : 최신정보로 빠른 의사결정 지원
- BI : 기업 보유 데이터 정리, 분석하는 리포트 중심도구
- BA : 통계 기반 비즈니스 통찰력
- Block Chain : 네트워크에 참여한 모든 사용자가 정보를 분산, 저장
- KMS : 기업의 모든 지식을 포함

• **데이터웨어하우스(DW)**
(1) 특징 ★
- 주제지향성 : 분석목적 설정이 중요
- 데이터통합 : 일관화된 형식으로 저장
- 시계열성 : 히스트로리를 가진 데이터
- 비휘발성 : 읽기전용 - 수시로 변하지 않음
(2) 구성요소
- ETL(Extraction, Transform, Load)
- ODS(Operation Data Store) : 다양한 DBMS에서 추출한 데이터를 임시저장

• **데이터레이스(DataLake)**
- 비정형데이터를 저장하며 하둡과 연계하여 처리
※ 하둡 : 여러컴퓨터를 하나로 묶어 대용량 데이터를 처리하는 오픈 소스 빅데이터 솔루션
- HDFS : 분산형 파일 저장 시스템
- MapReduce : 분산된 데이터를 병렬로 처리

2. 데이터의 가치와 미래

- 빅데이터의 이해

• **빅데이터 출현 배경**

- 인터넷 확산, 스마트폰 보급, 클라우딩 컴퓨팅으로 인한 경제성확보, 저장매체 가격하락, 하둡을 활용한 분산 컴퓨팅, 비정형 데이터 확산

• **빅데이터의 3V(가트너 정의) ★**

(1) Volume(규모) : 데이터 양 증가(구글번역서비스)
(2) Variety(다양성) : 데이터 유형 증가
(3) Velocity(속도) : 데이터 생성, 처리속도 증가
(4) 그 외 5V/7V에 포함되는 요소
- Value(가치) : 숨겨진 가치발견이 중요
- Veracity(신뢰성) : 고품질 데이터
- Validity(정확성) : 데이터의 유효성 보장
- Volatility(휘발성) : 데이터의 의미 있는 기간

• **빅데이터에 대한 비유**

(1) 산업혁명의 석탄, 철 : 산업혁명에서의 석탄, 철 역할
(2) 원유 : 정보제공으로 생산성 향상
(3) 렌즈 : 현미경이 생물학 발전영향, 산업전반에 영향(구글 Ngram Viewer)
(4) 플랫폼 : 공동 활용 목적으로 구축된 구조물, 써드파티 비즈니스에 활용(페이스북)
※ 써드파티 : 원천기술을 활용한 파생상품 만드는 회사

• **빅데이터가 만들어내는 변화 ★**

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

- 빅데이터의 가치와 영향

• **빅데이터 가치산정이 어려운 이유**

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

- 비즈니스 모델

• **빅데이터 활용 위한 3대요소**

- 인력, 자원(데이터), 기술

• **빅데이터의 주요분석기법**

- 회귀분석 : 독립변수와 종속변수 관계, X가 Y에 어떤 영향을 미치는가?
(수도권에 거리가 가까울수록 부동산 가격이 비싼가?)
- 분류분석 : A와 B는 어디에 속하는 범주(고양이와 강아지의 이미지를 구분)
- 연관규칙 : 여러요소들 간의 규칙 상관관계 존재(마트에서 치킨과 맥주를 같이사는 관계)

- 유전자 알고리즘 ★ : 최적화 필요한 문제의 해결책(택배차량 어떻게 배치, 최대 시청률 얻으려면 어떤 프로그램을 어떤 시간대에 방송?)
- 기계학습 : 훈련 데이터로부터 컴퓨터가 학습하고 미래를 예측(넷플릭스 영화추천 시스템)
- 감정분석 : 텍스트 데이터에서 감정(긍정/부정)을 분석
- 소셜네트워크분석 : 사람간의 관계(SNS상 사용자들 관계 속 영향력 높은 사람 찾기)
- 텍스트마이닝 : 텍스트로부터 자연어처리(NLP)를 통한 숨겨진 의미발견(문서요약, 키워드추출)

- 위기요인과 통제방안

• 위기요인과 통제방안

(1) 사생활 침해 : SNS에 올린 데이터가 사생활 침해
  -> 제공자에서 사용자 책임으로 전환
(2) 책임원칙훼손 : 범죄예측프로그램으로 예측하여 체포하는 문제
  -> 결과에 대해서만 책임
(3) 데이터의 오용 : 분석결과가 항상 옳은 것은 아님
  -> 알고리즘을 해석가능한 알고리즈미스트 필요
※ 알고리즈미스트 : 부당하게 피해가 발생한 사람들을 구제하는 전문인력

• 데이터 3법

- 가명정보의 개념도입(통계작성, 연구, 공익적 기록보존 목적 하에 동의 없이 활용가능)
(1) 개인정보보호법
(2) 정보통신망 이용 촉진 및 정보보호 등에 관한 법률(정보통신망법)
(3) 신용정보의 이용 및 보호에 관한 법률(신용정보법)

• 개인정보, 가명정보, 익명정보

(1) 개인정보 : 개인을 알아볼 수 있는 정보, 동의를 받아 활용가능(홍길동, 33세)
(2) 가명정보 : 가명처리를 통해 추가정보 없이 특정불가(홍OO, 30대초반)
(3) 익명정보 : 더 이상 개인을 알아볼수 없는 정보, 제한없이 자유롭게 활용(OOO, 30대)

• 개인정보 비식별화

(1) 가명처리(홍길동, 35세 -> 임꺽정, 30세)
(2) 총계처리(홍길동 170cm, 임꺽정 180cm -> 평균 키 175cm)
(3) 데이터삭제(주민등록번호 901111-1234567 -> 90년대생, 남자)
(4) 데이터 범주화(홍길동, 35세 -> 홍길동 30 ~ 40세)
(5) 데이터 마스킹(홍길동, 35세 -> 홍OO, 35세)

• 프라이버시 보호 모델

(1) k-익명성 : 같은 값이 존재하도록 하여 다른정보로 결합할 수 없도록 함
(2) l-다양성 : 민감한 정보의 다양성을 높여 추론가능성을 낮춤
(3) t-근접성 : 민감 정보의 분포를 낮추어 추론 가능성을 더욱 낮춤

- 미래의 빅데이터

• 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 주권행사
※ 마이데이터 : 자신의 신용정보를 다른 제3자에게 제공하여 서비스를 제공받는 제도

3. 가치창조를 위한 데이터사이언스와 전략 인사이트

- 빅데이터분석과 전략 인사이트

• **전략 인사이트**
- 집중과 선택(많은 데이터나 다양한 대상에 분산보다는 현재 분석에 집중)
- 업계상황만 보지 말고 더 넓은 시야에서 봐야함
- 경영진의 전략적 인사이트에 기여

조직이 분석을 배우는 상태이거나 특정 문제의 범위를 해결할 때는 집중과 선택
사업상황들을 확인할 때는 넓은 시야

• **데이터 사이언스**
- 데이터와 관련된 모든 분야의 전문지식을 종합한 학문
- 정형/비정형데이터를 막론하고 데이터를 분석(총체적 접근법)

• **데이터 사이언스 핵심 구성요소**
(1) Analytics : 이론적 지식
(2) IT : 프로그래밍적 지식
(3) 비즈니스 분석 : 비즈니스적 능력

- 전략적 인사이트 도출을 위한 필요역량

• **데이터 사이언티스트의 필요역량 ★**
(1) 하드스킬(Hard Skill) : 이론적지식(수학, 통계학, 가설검정 등), 가트너제시역량에 미포함
(2) 소프트 스킬(Soft Skill) : 스토리텔링, 리더십, 창의력, 분석 등

하드스킬은 이과적, 소프트 스킬은 문과적인 느낌

- 빅데이터 그리고 데이터 사이언스의 미래

• **빅데이터 가치 패러다임변화**
- Digitalization -> Connection -> Agency
(1) Digitalization : 아날로그 세상을 디저털화
(2) Connection : 디지털화된 정보들의 연결
(3) Agency : 연결을 효과적으로 관리

2과목 - 데이터 분석기획

1. 데이터분석 기획의 이해

- 분석기획 방향성 도출

• 분석대상과 방법

- 4가지 유형을 넘나들며 분석을 수행

방법 \ 대상	Known	UnKnown
Known	최적화(Optimization)	통찰(Insight)
UnKnown	솔루션(Solution)	발견(Discovery)

• 분석기획방안

	과제중심적 접근	장기적 마스터플랜
목적	빠르게 해결	지속적 분석원인해결
1차목표	Speed & Test	Accuracy & Deploy
과제유형	Quick & Win	Long Term View
접근방식	Problem Solving	Problem Definition

• 분석기획시 고려사항

(1) 가용데이터 : 분석의 기본이 되는 데이터 확보 및 파악
(2) 적절한 유스케이스 탐색 : 기존에 잘 구현되어있는 유사시나리오 활용
(3) 장애요소에 대한 사전계획 수립 : 조직의 역량으로 내제화

• 의사결정을 가로막는 요소

- 고정관념, 편향된 생각
- 프레이밍 효과 : 동일 상황임에도 개인의 판단, 결정이 달라짐

- 분석방법론

• 분석방법론의 구성요소

- 절차, 방법, 도구와 기법, 템플릿과 산출물

• 분석방법론 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 모델 : 핵심 부분을 개발한 후 요구사항을 반영하여 진화
(6) 애자일 : 짧은 개발 주기를 가지고 고객 피드백을 지속적으로 반영하여 반복적인 개발

• KDD 분석방법론 ★

- 데이터 선택 -> 전치러 -> 변환 -> 마이닝 -> 결과 평가

1) 데이터 선택 : 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) 결과평가 : 결과에 대한 해석, 결과가 충족되지 않으면 절차가 반복 수행

• Crisp-DM 분석 방법론 ★

- 업무이해 -> 데이터 이해 -> 데이터준비 -> 모델링 -> 평가 -> 전개 : 업데데이트모델평가전

1) 업무이해 : 업무 목적 파악, 상황파악, 목표 설정, 프로젝트 계획 수립
2) 데이터 이해 : 초기 데이터 수집, 기술 분석, EDA, 데이터품질 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 : 전개계획, 모니터링 및 유지보수 계획 수립, 프로젝트 종료보고서 작성, 프로젝트 리뷰

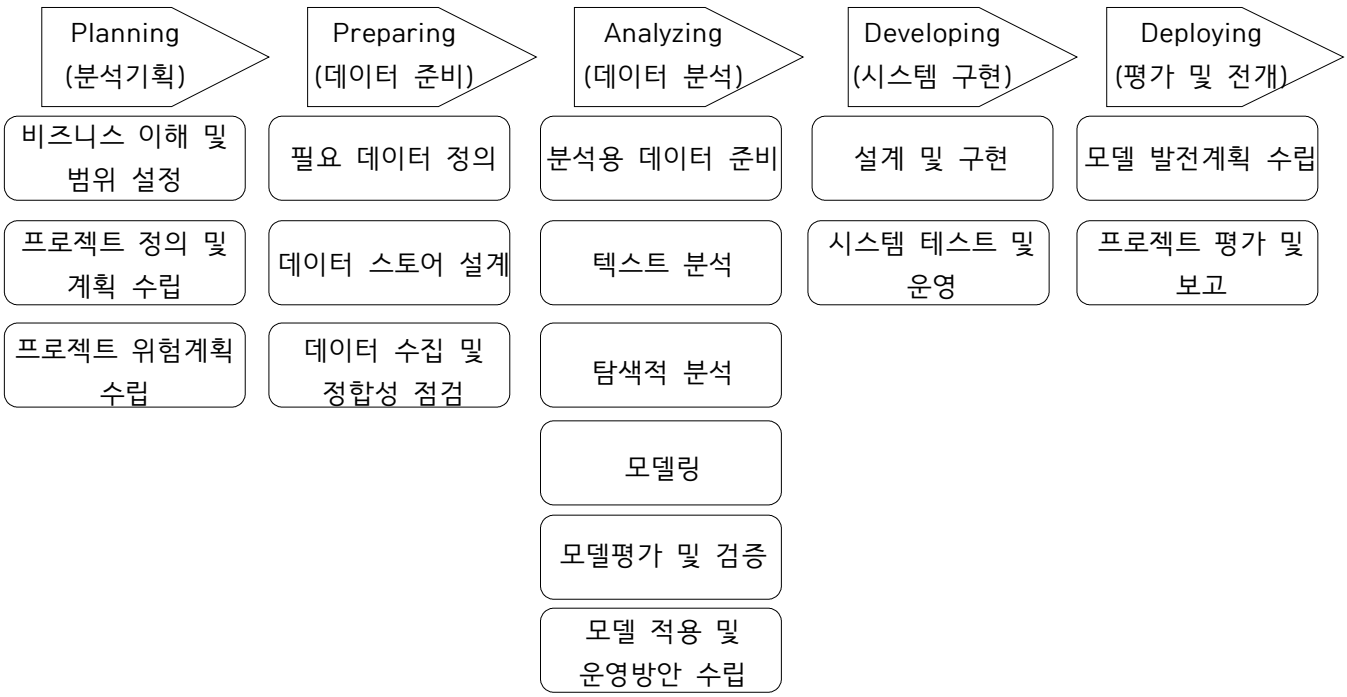
※ 평가 -> 전개에서 위대한 실패(업무 이해로 다시 돌아감) 발생 가능

• 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

• 빅데이터분석방법론 ★★★ PPADD



1) 분석기획

- 비즈니스 범위 설정 : SOW(Statement of Works) - 구조화된 프로젝트 정의서
- 프로젝트 위험계획 수립 : 회피, 전이, 완화, 수용

2) 데이터 준비
- 데이터 스토어 설계 : 정형, 비정형, 반정형 데이터에 따른 효율적 저장소를 설계
3) 데이터 분석
- 분석용 데이터 준비 : 추가적인 데이터 확보 필요시, 데이터 준비단계로 다시 진행
- 의사코드 : 일반적인 언어로 프로그래밍 언어의 알고리즘을 유사한 형식을 써 놓은 것
- 모델링 : 알고리즘 설명서는 상세히 작성
- 모델 평가 및 검증 : 성능이 저조한 모델은 튜닝 작업 수행

- 분석과제발굴

• 하향식 접근방법
- 문제가 주어지고 해답을 찾기 위해 진행
- 문제탐색 -> 문제정의 -> 해결방안 -> 타당성검토
(1) 문제탐색
1) 빠짐없이 문제를 도출하고 식별하여, 솔루션 초점보다는 가치에 초점
2) 비즈니스 모델 캔버스 단순화 측면 : 업무, 제품, 고객, 규제와 감사, 지원인프라
지원인프라 업무 중에 **고객**이 **제품**을 **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) 해결방안
- 기존 시스템 활용, 시스템 고도화, 인적 자원확보, 아웃소싱 등
(4) 타당성 검토
- 경제적 타당성 : 비용대비 편익 분석관점 접근
- 데이터 타당성 : 데이터 존재여부, 분석역량이 필요
- 기술적 타당성 : 역량 확보 방안 사전에 수립

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

• 혼합접근방법
(1) 발산단계 : 상향식 접근 방법으로서, 가능한 방안들을 도출
(2) 수렴단계 : 하향식 접근 방법으로서, 도출된 방안들을 분석

• 디자인싱킹
- 사용자가 공감으로 시작해서 아이디어 발산/수렴 과정을 통한 피드백으로 발전하는 과정
- 공감하기 -> 문제정의 -> 아이디어도출 -> 프로토타입 -> 테스트

• 지도학습, 비지도학습
(1) 지도학습 : 정답이 있는 데이터를 학습(하향식 접근법)
- 분류분석, 회귀분석, 의사결정트리, KNN, SVM
(2) 비지도학습 : 정답이 없는 데이터를 학습(상향식 접근법)

- 군집분석, 차원축소, 연관규칙분석

- 분석 프로젝트 관리방안

• 분석과제에서 고려해야할 5가지 요소

- 데이터 크기, 속도, 데이터 복잡도, 분석 복잡도, 정확도/정밀도

※ 정확도(Accuracy)와 정밀도(Precision)는 Trade-Off 관계

여기에서 정확도와 정밀도는 3과목의 오분류표에서의 평가지표와는 다른 개념

• 프로젝트관리지식체계 10가지 영역

- 통합, 범위, 시간(일정), 원가, 품질, 인적자원, 의사소통, 리스트(위험), 조달(아웃소싱), 이해관계자

이범통이 **의자**에서 **시원**한 **조리품**을 먹었다.

2. 분석 마스터 플랜

- 마스터플랜 수립

• it프로젝트의 우선순위 선정기준

- 중장기 마스터플랜을 수립위하여 ISP를 활용

(1) 전략적 중요도 : 전략적 필요성, 시급성

(2) 실행 용이성 : 투자 용이성, 기술 용이성

• 데이터분석프로젝트의 우선순위 선정 기준

(1) 시급성 관점 : 비즈니스 효과(Return) - Value

(2) 난이도 관점 : 투자비용 요소(Investment) - Volume, Variety, Velocity

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

- 시급성 중요시 : 3 -> 4 -> 2
- 난이도 중요시 : 3 -> 1 -> 2

3과 2는 앞 뒤로 고정하고 가운데만 변경

- 분석거버넌스 체계수립

• 분석거버넌스 체계 구성요소

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

시조프로마인드데

• **데이터분석수준진단 ★**

(1) 분석준비도

분석적 업무파악	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 조직존재 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 외부데이터 활용체계 마스터데이터 관리(MDM)	사실에 근거한 의사결정 관리자의 데이터중시 회의 등에서 데이터활용 경영진의 직관보다 데이터 데이터 공유 및 협업문화	운영시스템 데이터 통합 EAL, ETL 등 데이터 유통체계 분석전용 서버 및 분석환경 빅데이터 분석환경 통계분석 환경 비주얼분석 환경

IT문데기인파

(2) 분석 성숙도

- CMMI 모델기반(1 ~ 5단계)

1) 도입 : 환경, 시스템 구축

2) 활용 : 업무에 적용

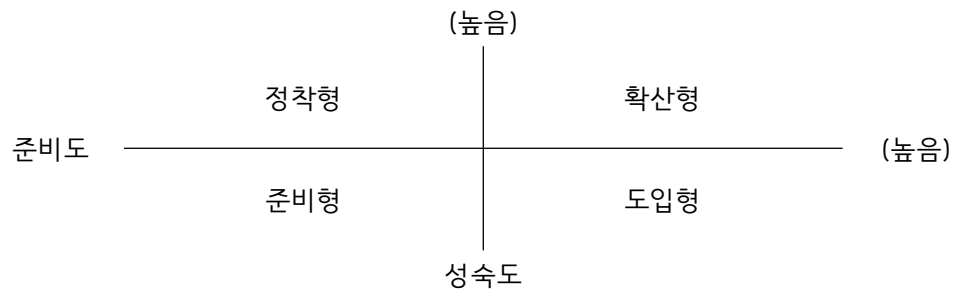
3) 확산 : 전사 차원관리, 공유

4) 최적화 : 혁신, 성과향상에 기여

도활확최

단계	비즈니스 부문	조직 및 역량부문	IT 부문
도입	실적 분석 및 통계 정기 보고 수행 운영데이터 기반	일부부서에서 수행 담당자 역량에 의존	데이터 웨어하우스 데이터 마트 ETL/EAI OLAP
활용	미래결과 예측 시뮬레이션 운영데이터기반	전문담당부서 수행 분석기법도입 관리자가 분석수행	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 비주얼분석 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

• **데이터분석성숙도 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 시계방향(역순)으로 암기

• **분석지원인프라방안수립**
- 확장성을 고려한 플랫폼 구조 적용(중앙집중적 관리)

(1) 분석 플랫폼 구성요소
1) 광의의 분석 플랫폼 : 분석 서비스 제공엔진, 분석 어플리케이션, 분석 서비스 API, 하드웨어
2) 협의의 분석 플랫폼 : 데이터 처리 프레임워크, 분석엔진, 분석 라이브러리

광의의 분석 플랫폼은 협의의 분석 플랫폼 요소들을 포함하는 개념

• **데이터 거버넌스**

(1) 데이터 거버넌스
1) 전사 차원에서 데이터에 대해 표준화된 관리 체계 수립
2) 구성요소 : 원칙, 조직, 프로세스 원조프
3) 중요 관리대상 : 마스터 데이터, 메타데이터, 데이터 사전 등
- 마스터 데이터 : 자료 처리에 기준이 되는 자료
- 메타데이터 : 다른 데이터를 설명해주는 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

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

- **조직 및 인력방안수립(DSCoE : 분석조직) ★★★**

- 집중구조 : 독립적인 전담 조직 구성(중복업무 가능성 존재)
- 기능구조 : 해당 부서에서 직접 분석(DSCoE가 없음)
- 분산구조 : 분석 조직 인력을 현업 부서에 배치

3과목 데이터 분석

1. R기초와 데이터 마트

- 데이터 마트

• **데이터 마트(DM)**

- 데이터 웨어하우스의 한 분야로 특정 목적을 위해 사용(소규모 데이터웨어하우스)

• **요약변수와 파생변수**

(1) 요약변수 : 수집된 정보를 종합한 변수로서 재활용성이 높음(1개월간 수입)

(2) 파생변수 : 의미를 부여한 변수, 논리적 타당성 필요(고객구매등급)

- 결측값과 이상값 검색

• **EDA(탐색적 자료분석)**

- 데이터의 의미를 찾기 위해 통계, 시각화를 통해 파악

- EDA의 4가지 주제

1) 저항성의 강조 : 자료 변동에 민감하지 않음

2) 잔차계산 : 값들이 주경향으로부터 얼마나 벗어나 있는지 확인하는 척도

3) 자료변수의 재표현 : 원래 변수를 적당한 척도로 변환

4) 그래프를 통한 현시성 : 시각화를 통하여 효율적으로 파악

• **결측값 처리**

- 존재하지 않는 데이터, null/NA로 표시

(1) 완전분석법 : 결측값 가지는 데이터 삭제

(2) 평균 대치법(=비조건부 평균 대치) : 단순 평균으로 대치

(3) 회귀 대치법(=조건부 평균 대치) : 회귀분석의 결과로 대치

(4) 단순 확률 대치법 : 확률적으로 선택하여 대치

- Nearest Neighbor : 바로 가까운 응답으로 대체

- Hot-Deck : 현재 데이터 셋에서 비슷한 성향으로 대체

- Cold-Deck : 유사한 외부 출처에서 비슷한 성향으로 대체

(5) 다중 대치법 : 여러번 대치(대치 -> 분석 -> 결합)

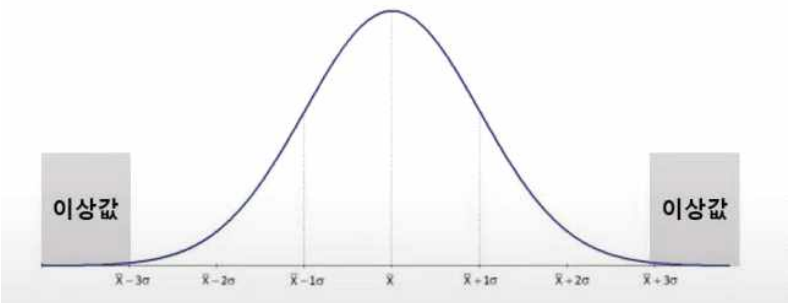
• **이상값 처리**

- 극단적으로 크거나 작은 값이며, 의미 있는 데이터 일수도 있음(체중 3kg)

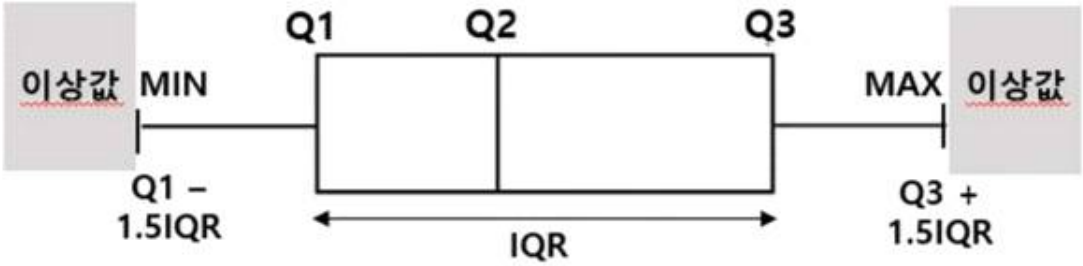
- 이상값을 항상 제거하는 것은 아님

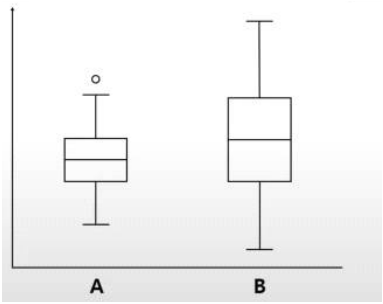
(1) ESD(Extreme Studentized Deviation)

- 평균으로부터 표준편차의 3배 넘어가는 데이터는 이상값으로 판단



(2) 사분위수
- Q1 - 1.5IQR보다 작거나, Q3 + 1.5IQR보다 크면 이상값으로 판단
- 최솟값, 1~3사분위값, 최댓값 등을 표현하며, 평균값은 표현하지 않음





- B의 데이터가 A보다 넓게 분포되어 있으므로, B의 분산이 A보다 큼
- Q2의 위치가 B가 더 높이 있으므로, B의 중앙값이 A보다 큼
- A에는 MAX를 벗어나는 데이터가 존재하므로, A에는 이상치가 존재한다고 판단 가능
- A와 B의 평균은 알 수 없음

(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. 통계분석

- 통계학 개론

• 전수조사와 표본조사
- 전수조사 : 전체를 다 조사, 시간과 비용 많이 소요
- 표본 조사 : 일부만 추출하여 모집단을 분석

• 자료의 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 간격의 의미가 있으며 덧셈과 뺄셈만 가능(온도, 지수 등)
- 비율척도 : 절대적 기준 0이 존재하고 사칙연산 가능한 자료(무게, 나이 등)

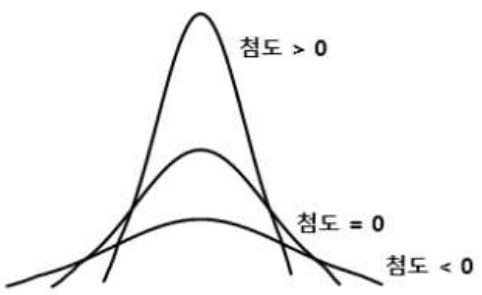
• 확률적 표본 추출방법
(1) 랜덤 추출법 : 무작위로 표본 추출
(2) 계통 추출법 : 번호 부여하여 일정 간격으로 추출
(3) 집락 추출법(=군집 추출법) ★
- 여러 군집으로 나눈 뒤 군집을 선택하여 랜덤 추출
- 군집 내 이질적 특징, 군집 간 동질적 특징

(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 포함시켜 표본 추출
- 비복원 추출 : 추출되었던 데이터는 제외하고 표본 추출

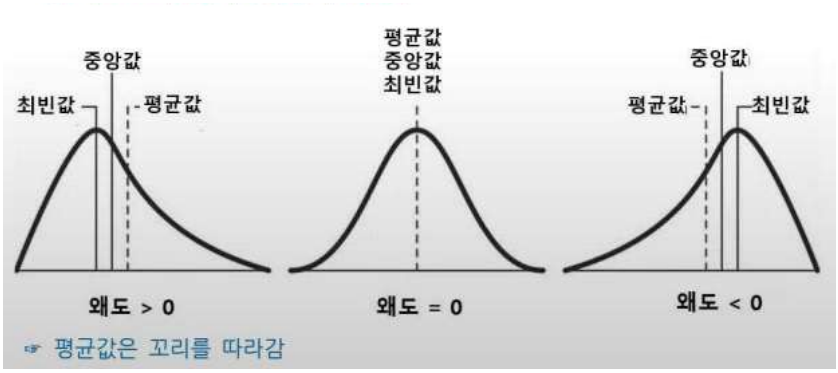
• 비확률적 표본 추출 방법
(1) 편의 추출법 : 연구자가 쉽게 접근 가능한 대상으로 표본을 추출
(2) 의도적 추출법 : 연구자가 특정 기준을 정하고 이에 맞는 표본을 추출
(3) 할당 추출법 : 특정 기준으로 나눈 후, 그 그룹에서 할당된 수 만큼 추출
(4) 눈덩이 추출법 : 초기 응답자로부터 새로운 응답자를 추천 받는 방식
(5) 자기선택 추출법 : 응답자가 스스로 조사에 참여할지 결정

• 기초 통계량
(1) 중심경향성 측면
- 산술평균 : 일반적인 평균 개념으로 모든 값을 더한 후 데이터 개수로 나눈 값
- 기하평균 : 모든 값들을 곱하고 n제곱근을 구하는 방식(비율적 증가율)
- 조화평균 : 역수의 산술평균을 구한 후, 다시 역수에 취하는 방식(비율계산)
- 중앙값 : 데이터를 크기 순서로 나열했을 때 중간에 위치한 값
- 최빈값 : 데이터에서 가장 자주 나타나는 값
(2) 분산 정도 측면
- 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
(3) 관계측면 ★★
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 : 정비례관계
- 상관계수 = 0 : 상관없음
- 상관계수 = -1 : 반비례관계
3) 공분산과 독립성의 관계
- 두 변수가 독립이면 공분산은 0이지만, 공분산이 0이라고 두 변수가 독립이라고 할 수는 없음

• 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



• **Summary함수 결과의 해석**

Age Survived Min. : 0.17 0:266 1st Qu.:21.00 1:152 Median :27.00 Mean :30.27 3rd Qu.:39.00 Max. :76.00 NA's :86	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 2) Survived 변수 - 집단의 빈도 수 -> 범주형 변수

• **기초확률이론**

(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 표본 공간 S = {(앞, 앞), (앞, 뒤), (뒤, 앞), (뒤, 뒤)}
(4) 확률변수 : 표본 공간의 각 원소에 해당하는 값(확률)을 대응하는 함수
- 예) 확률변수 X가 어떤 집합의 키를 나타낼 때 키가 160~170 확률은 $P(160 \leq X \leq 170)$
(5) 조건부 확률 : 특정 사건 B가 발생했을 때 A가 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
- $P(A \cap B) = \varnothing$(동전을 던졌을 때 앞면과 뒷면이 동시에 나올 확률)
(8) 베이즈 정리 : 두 확률 변수의 사전 확률과 사후 확률사이의 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) 베르노이분포 : 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 베르누이 시행을 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 뽑혔을 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 시간동안 사무실에 걸려온 전화의 수

베포항항하

(2) 연속확률분포
- 값을 셀 수 없는 분포, 확률밀도함수로 표현
1) 정규분포 : 우리가 일상생활에서 흔히 보는 확률변수의 평균분포를 근사한 분포(Z검정 활용) ★★★
 - 예) 사람들의 키 혹은 IQ 점수의 분포, 시험 성적의 분포
2) t분포 : 정규분포와 유사하지만, 꼬리부분이 더 두껍고 긴 분포 ★★★
 - (T검정 활용) 표본이 30개보다 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 X의 $f(x)$ 확률분포에 대한 기댓값($E(X)$)

1) 이산적 확률변수 : $E(X)=\sum xf(x)$ ★

2) 연속적 확률변수 : $E(X)=\int xf(x)$ ★

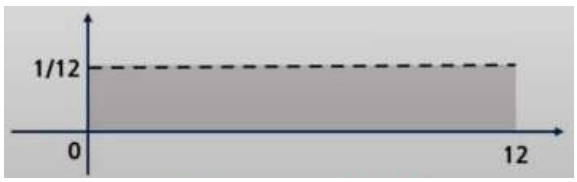
(1) 동전을 3개 던지는 확률실험을 할 때, 확률변수 X(앞면 F의 개수)의 기댓값은?
(2) 1~12의 숫자가 표시된 원형시계에서, 확률변수 X(시계바늘이 가르키는 시간)의 기댓값은?

(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)



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

• **중심극한 정리 ★★★**

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

• **표본평균의 표본분포**

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

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

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

(μ : 모집단의 평균, σ : 모집단의 표준편차, $\overline{X}$: 표본평균, n : 표본의 크기)

• **점추정**

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

(1) 추정량의 조건

1) 불편성(Unbiasedness) : 추정량의 기댓값이 실제 모수와 같음(편향이 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

(2) 대표적인 추정량

1) 모집단의 평균 μ -> 표본평균 $\overline{X} = \frac{1}{n}\sum X$

2) 모집단의 분산 σ^2 -> 표본분산 $s^2 = \frac{1}{n-1}\sum (X - \overline{X})^2$

• **구간추정(신뢰구간)**

- 모집단이 특정한 구간으로 추정(95%, 99를 가장 많이 사용)

(1) 모집단의 분산을 알고 있는 경우

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

- 신뢰수준 95% : $Z_{\frac{\sigma}{2}} = 1.960$
- 신뢰수준 99% : $Z_{\frac{\sigma}{2}} = 2.576$

(2) 모집단의 분산을 모르는 경우

- 자유도가 n-1인 t분포를 이용하여 신뢰구간을 추정

$$\overline{X} - t_{\frac{\sigma}{2}, n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\frac{\sigma}{2}, n-1} \frac{S}{\sqrt{n}}$$ (S : 표본표준편차)

• **가설검정**

- 모집단의 특성에 대한 주장을 가설로 세우고 표본조사로 가설의 채택여부를 판정

(1) 귀무가설(H0) : 일반적으로 생각하는 가설(차이가 없다)

(2) 대립가설(H1) : 귀무가설을 기각하는 가설, 증명하고자 하는 가설(차이가 있더, 크다/작다)

(3) 유의수준(α) : 귀무가설이 참일 때 기각하는 1종 오류를 범할 확률의 허용한계(일반적 0.05)

(4) 유의확률(p-value) : 귀무가설을 지지하는 정도를 나타내는 확률

실제 \ 검정결과	H0가 사실이라고 판정	H0가 거짓이라고 판정
H0가 사실	옳은 결정	1종 오류(α)
H0가 거짓	2종 오류(β)	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 / 두 개의 모집단 -> 이표본
4) 귀무가설 기각 혹은 채택
- p-value < 유의수준(α) -> 귀무가설 기각
- p-value > 유의수준(α) -> 귀무가설 채택
5) t검정인 경우 - 단일표본, 대응표본, 독립표본 확인
- 모집단에 대한 평균 검정 -> 단일표본
- 동일 모집단에 대한 평균비교 검정 -> 대응표본
- 서로 다른 모집단에 대한 평균비교 검정 -> 독립표본

※ 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
```

1) 귀무가설 / 대립가설 설정
- '차이가 없다' 혹은 '동일하다' -> 귀무가설로 설정 : 두학교의 성적은 동일하다
2) 양측 혹은 단측검정 확인
- 대립가설의 값이 '같지 않다' -> 양측검정
3) 일표본 혹은 이표본 확인
- 두 개의 모집단 -> 이표본
4) 귀무가설 기각 혹은 채택
- p-value : 0.5515 > 유의수준(α) : 0.05 -> 귀무가설 채택
5) 단일표본, 대응표본, 독립표본 확인
- 서로 다른 모집단에 대한 평균비교검정 -> 독립표본

• **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, 크러스칼윌리스 검정, 프리먼드 검정, 카이제곱 검정

- 기초 통계분석

• **회귀분석**

(1) 개념 : 독립변수들이 종속변수에 영향을 미치는지 파악하는 분석방법

1) 독립변수 : 원인을 나타내는 변수(x)

2) 종속변수 : 결과를 나타내는 변수(y)

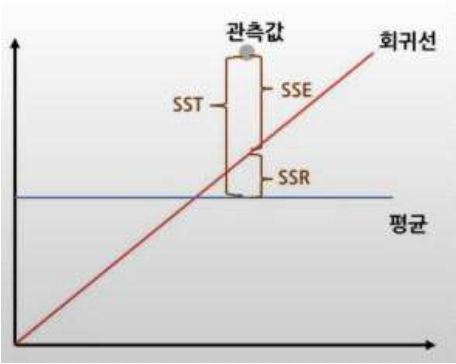
3) 잔차 : 계산값과 예측값의 차이(오차)

(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(Sum of Squares Total) : 전체의 변동

(2) SSE(Sum of Squares Error) : 모형에 의해 설명되지 않는 변동

(3) SSR(Sum of Squares Regression) : 모형에 의해 설명되는 변동

(4) $R^2 = SSR/SST = 1 - SSE/SST$

• **선형회귀분석의 가정**

(1) 선형성 : 종속변수와 독립변수는 선형관계

(2) 등분산성 : 잔차의 분산이 고르게 분포

(3) 정상성(정규성) : 잔차의 정규분포의 특성을 지님

(4) 독립성 : 독립변수들간 상관관계가 없음

※ 다중공선성 : 독립변수들간 강한 상관관계가 나타나는 문제 ★

- VIF(분산팽창인수) 값이 10 이상이면 다중공선성이 존재한다고 판단 $VIF = \frac{1}{1-R^2}$

• **회귀분석 종류**

(1) 단순회귀 : 1개의 독립변수와 종속변수의 선형관계

(2) 다중회귀 : 2개 이상의 독립변수와 종속변수의 선형관계

(3) 다항회귀 : 2개 이상의 독립변수와 종속변수가 2차함수 이상의 관계

(4) 릿지회귀(L2 규제) : L2 규제항을 포함 - $\sum W^2$ (유클리디안 거리 기반)

(5) 라쏘회귀(L1 규제) : L1 규제항을 포함 - $\sum |W|$ (맨하탄 거리 기반)

(6) 교호항이 포함된 회귀 : 독립변수들의 교호작용이 포함된 회귀 모형

※ 교호작용 : 두 개 이상의 독립변수가 상호작용을 하여, 종속변수에 영향을 미치는 경우

• 최적의 회귀방정식 탐색 방법

(1) 전진선택법 : 변수를 하나씩 추가하면서 최적의 회귀방정식을 찾아내는 방법

(2) 후진제거법 : 변수를 하나씩 제거하면서 최적의 회귀방정식을 찾아내는 방법

(3) 단계별 선택법 : 전진선택법 + 후진제거법으로 변수를 추가할 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

• 회귀분석의 분산분석(ANOVA)표

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

- ANOVA 검정 : 3개 이상의 그룹의 평균을 비교하는 검정(회귀모형의 유의성 분석시 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

• 회귀모형의 검정

1) 독립변수와 종속변수 설정

2) 회귀계수 값의 추정

3) 모형이 통계적으로 유의미한가 : 모형에 대한 F통계량, p-value

- 귀무가설 : '모든 회귀계수는 0이다'

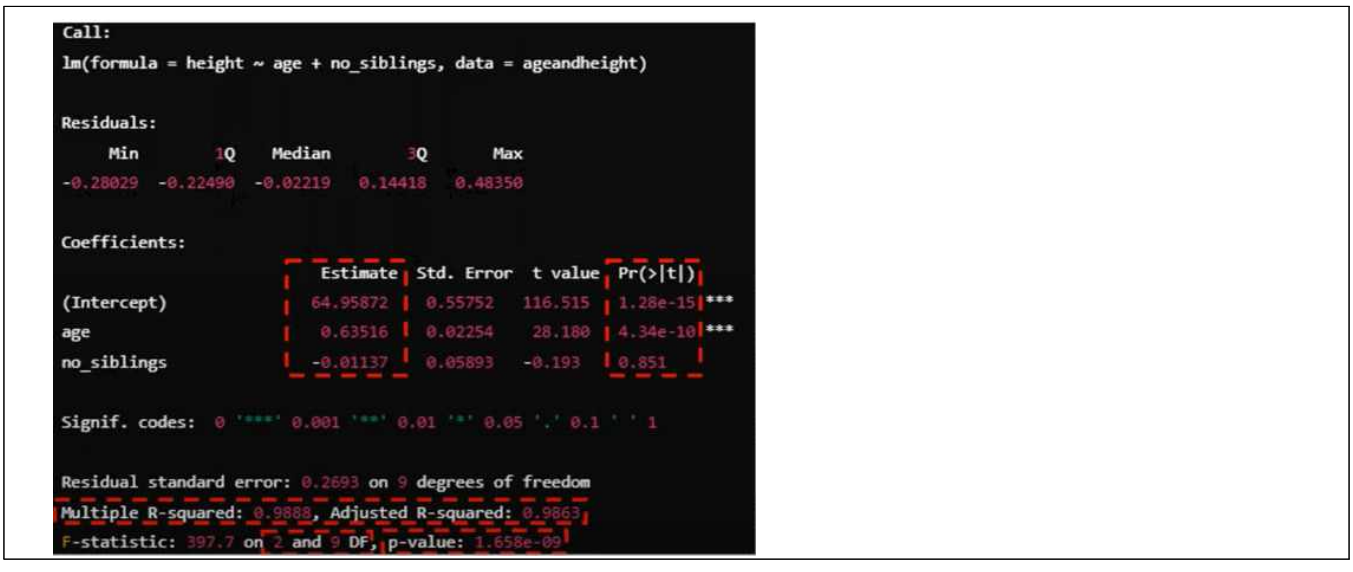
4) 회귀계수들이 유의미한가 : 회귀계수들의 t통계량, p-value

- 각각의 회귀계수에 대한 귀무가설 : '회귀계수는 0이다'

5) 위 1), 2) 모두를 기각하면 해당 모델을 활용

6) 모형이 설명력을 갖는가 : 결정계수(R-square) 값

⁕ 일반적인 회귀모형의 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 다중회귀모형
- 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 회귀모형의 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
- jobclasss2.Information의 회귀계수 22.73086 : Information이 Industrial보다 임금 높음
- age:jobclass2.Information의 p-value(0.21)가 0.05보다 크므로 교호작용은 유의하지 않음

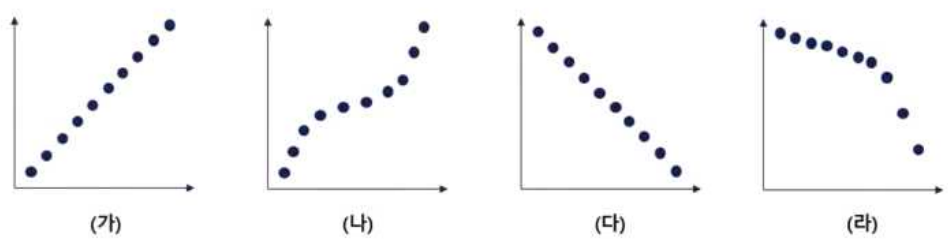
- 다변량 분석

• 상관분석 ★★★

- 두 변수간의 선형적 관계가 존재하는지 파악하는 분석

(1) 피어슨 상관분석 : 양적 척도, 연속형 변수, 선형관계 크기 측정

(2) 스피어만 상관분석 : 서열 척도, 순서형 변수, 선형/비선형 관계 나타냄



(가) 피어슨 계수 = +1 / 스피어만 계수 = +1

(나) 피어슨 계수 = +0.8 / 스피어만 계수 = +1

(다) 피어슨 계수 = -1 / 스피어만 계수 =-1

(라) 피어슨 계수 = -0.8 / 스피어만 계수 = -1

스피어만 계수는 X와 Y가 선형관계가 아니더라도 +1 혹은 -1이 될 수 있다.

• 주성분분석(PCA) ★★★

- 상관성 높은 변수들의 선형 결합으로 차원을 축소하여 새로운 변수를 생성
- 자료의 분산이 가장 큰 축이 첫 번째 주성분(고유값 고려)
- 70 ~ 90%의 설명력을 갖는 수를 결정

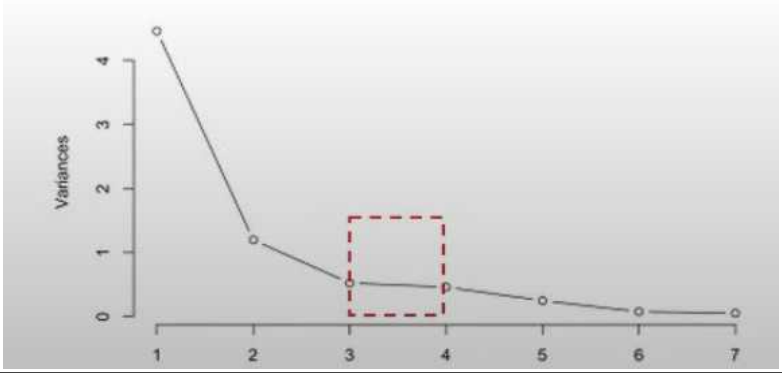
(1) 주성분 분석의 결과 해석

```
> 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(Screeplot)

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



- 기울기가 3~4구간에서 완만해지므로 주성분 개수는 2개로 선택

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

• **다차원 척도법(MDS : MultiDimensional Scaling)**
- 데이터 간 거리 정보의 근접성을 보존하는 방식으로 차원 축소하여 시각화
(1) 특징 : 데이터 축소 목적, Stress값이 0에 가까울수록 좋음, x/y축 해석이 불가
(2) 종류
1) 계량적 MDS : 양적척도 활용
2) 비계량적 MDS : 순서척도 활용

- 시계열 예측

• **시계열 분석**
- 시간의 흐름에 따라 관찰된 자료의 특성을 파악하여 미래를 예측(주가데이터, 기온데이터)

• **정상성 ★★**
- 시계열 예측을 위해서는 모든 시점에 일정한 평균과 분산을 가지는 정상성을 만족해야함
- 정상시계열로 변환 방법
1) 차분 : 현 시점의 자료를 이전 값으로 빼는 방법
2) 이동평균법 : 일정기간의 평균
3) 지수평활법 : 최근 시간 데이터에 가중치를 부여
4) 그 외 정상화 방법 : 지수변환, 로그변환, Box-Cox 변환 등

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

• **시계열 모형**
(1) 자기회귀(AR) 모형
- 자기자신의 과거 값이 미래를 결정하는 모형
- 부분자기상관함수(PACF)를 활용하여 p+1시점 이후 급격 감소하면 AR(p) 모형 선정
(2) 이동평균(MA) 모형
- 이전 백색잡음들의 선형결합으로 표현되는 모형
- 자기상관함수(ACF)를 활용하여 q+1시차 이후 급격히 감소하면 MA(q) 모형 선정
(3) 자기회귀누적이동평균(ARIMA)모형
- AR모형과 MA모형의 결합
- ARIMA(p,d,q)
1) p와 q는 AR모형과 MA 모형이 관련 있는 차수
2) d는 정상화시에 차분 몇 번 했는지 의미
3) d = 0 : ARIMA 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(경제전반, 특정 산업)
(4) 불규칙 요인 : 위 3가지로 설명 불가한 요인

3. 정형 데이터 마이닝
-데이터 마이닝 개요

• **데이터마이닝**

- 방대한 데이터 속에서 새로운 규칙, 패턴을 찾고 예측을 수행하는 분야

• **데이터 마이닝의 유형**

(1) 지도학습 : 정답이 있는 데이터를 활용

- 인공신경망, 의사결정트리, 회귀분석, 로지스틱회귀

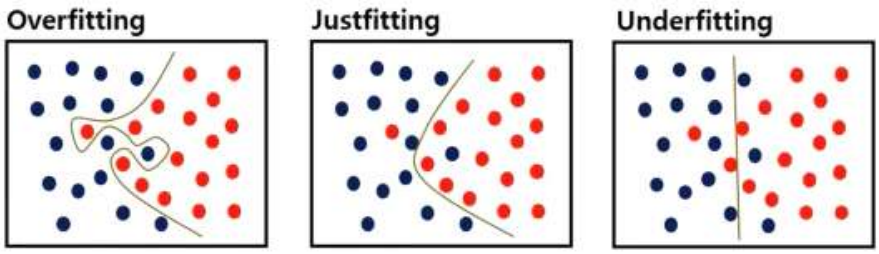
(2) 비지도학습 : 정답이 없는 데이터들 사이의 규칙을 파악

- 군집분석, SOM, 차원축소, 연관분석

• **과대적합과 과소적합**

(1) 과대적합 : 모델이 지나치게 데이터를 학습하여 매우 복잡해진 모델(높은 분산, 낮은 편향)

(2) 과소적합 : 데이터를 충분히 설명하지 못하는 단순한 모델(낮은 분산, 높은 편향)



• **데이터 분할**

- 과대적합과 과소적합을 방지하고, 데이터가 불균형한 문제를 해결하기 위해 사용

(1) 분할된 데이터셋 종류

1) 훈련용(Training Set) : 모델을 학습하는데 활용

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

3) 평가용(Test Set) : 모델을 평가하는데 활용

(2) 분할된 데이터의 학습 및 검증 방법

1) 홀드아웃 : 훈련용과 평가용 2개의 셋으로 분할

2) K-fold 교차검증 : 데이터를 k개의 집단으로 구분하여 k-1개 학습, 나머지 1개로 평가

3) LOOCV : 1개의 데이터로만 평가 나머지로 학습, 데이터 수가 부족할 때 적용

4) 부트스트래핑 : 복원추출을 활용하여 데이터셋을 생성, 데이터 부족, 불균형 문제 해소

- 분류분석

• **로지스틱 회귀분석**

- 종속변수가 범주형 데이터를 대상으로 성공과 실패 2개의 집단을 분류하는 문제에 활용

(1) 오즈(Odds)

- 성공할 확률과 실패할 확률의 비

- Odds = $P/(1-P)$ = 성공확률 / 실패확률

(2) 로짓(logit)변환

- 오즈에 자연로그(자연상수 e가 밑)을 취하여 선형관계로 변환

- $\log(P/(1-P)) = \alpha + \beta x$

(3) 시그모이드 함수

- 로짓함수의 역함수를 통하여, 0 ~ 1사이의 확률을 도출하는 함수

- 독립변수 x가 n증가하면 확률이 e^n만큼 증가

• KNN(K-Nearest Neighbors)

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

• 나이브베이즈 분류

(1) 베이즈 정리

$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$ ($P(A|B)$: 사후확률, $P(B|A)$: 우도, $P(A)$: 사전확률, $P(B)$: 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*0.3 = 0.18	0.6*0.7 = 0.42	0.6
여학생	0.4*0.5 = 0.20	0.4*0.5 = 0.20	0.4

$$P(A|B) = P(\textit{남학생}|\textit{합격한 신입 생}) = \frac{P(A \cap B)}{P(B)} = \frac{0.18}{0.38} = 0.47$$

(2) 나이브베이즈 분류

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

• 의사결정나무(Decision Tree)

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

(1) 분할 방법

1) 분류(범주형)에서의 분할 방법

- CHAID 알고리즘 : 카이제곱 통계량
- CART 알고리즘 : 지니지수 활용 ($1 - \sum P^2$)
- C4.5 / C5.0 알고리즘 : 엔트로피지수 활용($-\sum P(\log P)$)

2) 회귀(연속형)에서의 분할 방법

- CHAID 알고리즘 : ANOVA, F-통계량
- CART 알고리즘 : 분산감소량

- 지니지수와 엔트로피지수 계산

- 앞면 확률 = $\frac{3}{5}$, 뒷면 확률 = $\frac{2}{5}$
- 지니지수 : $1-(\frac{3}{5})^2-(\frac{2}{5})^2=\frac{12}{25}$
- 엔트로피지수 : $-\frac{3}{5}log(\frac{3}{5})-\frac{2}{5}log(\frac{2}{5})$

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

• 서포트벡터머신(SYM)
- 마진이 최대가 되는 초평명을 찾아 선형이나 비선형 이진 분류, 회귀에서 활용 가능한 다목적 모델

(1) 구성요소
- 하이퍼플레인(초평면) : 데이터를 구분하는 기준이 되는 경계, 가중치벡터와 편향으로 결정
- 서포트벡터 : 클래스를 나누는 하이퍼플레인과 가까운 위치의 샘플
- 마진 : 하이퍼플레인과 서포트벡터 사이의 거리
- 커널함수 : 저차원 데이터를 고차원 데이터로 변경하는 함수

(2) 유형
- 하드마진분류 : 오류 비허용
- 소프트마진분류 : 마진 내 어느정도 오류 허용

• 앙상블
- 여러 개의 예측 모형들을 조합하는 기법으로 전체적인 분산을 감소시켜 성능 향상이 가능

(1) 보팅(Voting)
- 다수결 방식으로 최종 모델을 선택

(2) 배깅(Bagging)
- 복원추출에 기반을 둔 붓스트랩을 생성하여 모델을 학습 후에 보팅으로 결합
- 복원추출을 무한히 반복할 때 특정 하나의 데이터가 선택되지 않을 확률

  → $\lim_{N\to\infty}(1-\frac{1}{N})^N=36.8\%$

(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) 스태킹(Stacking)
- 각각의 모델에서 학습한 예측 결과를 다시 학습

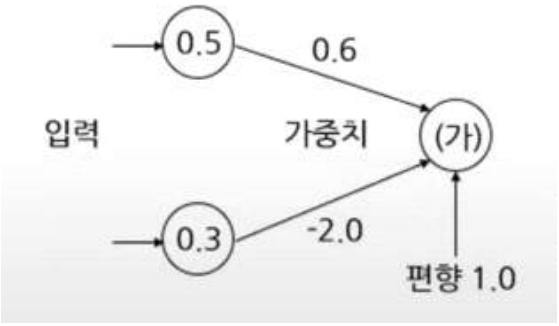
(5) 랜덤포레스트 ★★
- 배깅에 의사결정트리를 추가하는 기법으로 성능이 좋고 이상치에 강한 모델

보팅, 배깅, 랜덤포레스트는 병렬처리가 가능하며, 부스팅은 병렬처리가 불가

• 인공신경망
- 인간의 뇌 구조를 모방한 퍼셉트론을 활용한 추론모델

(1) 구조
1) 단층신경망 : 입력층과 출력층으로 구성(단일 퍼셉트론)
2) 다층 신경망 : 입력층과 출력층 사이에 1개 이상의 은닉층을 보유(다층 퍼셉트론)
- 은닉층 수는 사용자가 직접 설정하는 하이퍼라이터

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 퍼셉트론간의 연결강도를 의미

(2) 활성화 함수와 손실함수

1) 은닉층에서의 활성함수 : 인공신경망의 선형성을 극복(XOR 문제 해결)

- 시그모이드 함수 : 0 ~ 1 사이의 확률 값을 가지며, 로지스틱 회귀분석과 유사
- 하이퍼볼릭 탄젠트(Tanh) 함수 : -1 ~ 1 사이 값, 시그모이드 함수의 최적화 지연을 해결
- ReLU 함수 : 기울기 소실문제를 극복, $\max(0,x)$
- 그 외 활성함수 : Leaky RELU, GELU, ELU 등

2) 출력층에서의 활성함수

- 시그모이드 함수 : 이진분류 모델(0 ~ 1 사이 확률)
- 소프트맥수 함수 : 다중분류 모델(확률의 총합이 1)

3) 손실함수 : 예측값과 실제값의 차이를 측정하는 함수

- MSE(Mean Square Error) : 회귀 모델
- 크로스 엔트로피(Cross-Entropy) : 분류 모델

(3) 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 사용시, 학습이 제대로 되지 않는 문제

• 딥러닝

(1) DNN(심층 신경망) : 은닉층이 2개 이상으로 구성된 인공신경망(입력층 - 은닉층 - 출력층)

(2) CNN(합성곱 신경망) : Convolution Layer와 Pooling Layer를 활용, 이미지 패턴을 찾는 신경망

- 구조 : Input - Convolution Layer - Pooling Layer - Flatten - Fully Connected Layer

(3) RNN(순환 신경망) : 순차적인 데이터 학습에 특화된 순환구조를 가지는 신경망

- 과거 정보 전달되지 안흔ㄴ 장기의존성 문제 발생가능(극복모델 : LSTM, GRU)

(4) 오토인코더

- 입력 데이터를 인코더로 압축한 후에 디코더로 형태를 재구성하는 비지도 학습 신경망
- 구조 : Encoder - Context Vector(=Latent Space) - Decoder
- 오토인코더는 생성형 AI의 기반 모델

• **분류모델 평가지표**

(1) 오분류표(혼동행렬) ★★★★

		실제: TRUE	실제: FALSE
예측	TRUE	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)
예측	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) $F-\beta$ 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 ~ 1 사이이며, 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

- 군집분석

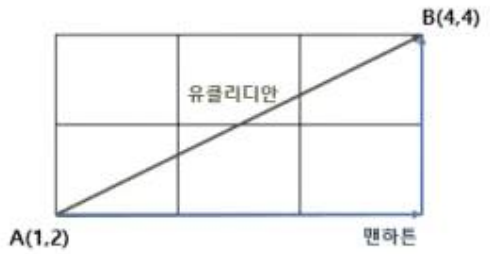
• 군집분석

- 비지도 학습으로 데이터들 간 거리나 유사성을 기준으로 군집을 나누는 분석

• 거리측도

(1) 연속형 변수

- 유클리디안 거리 : 두점 사이의 직선 거리
- 맨하튼 거리 : 각 변수들의 차이의 단순 합
- 체비셰프 거리 : 변수 거리 차 중 최댓값
- 표준화 거리 : 유클리디안 거리를 표준편차로 나눔
- 민코우스키 거리 : 유클리드, 맨하튼 거리를 일반화한 거리
- 마할라노비스 거리 : 표준화 거리에서 변수의 상관성 고려



- 유클리디안 거리
: $\sqrt{(4-1)^2+(4-2)^2}=\sqrt{13}$
- 맨하튼 거리
: $|4-1|+|4-2|=5$
- 체비셰프 거리
: $\max(|4-1|,|4-2|)=3$

(2) 범주형 변수

- 자카드 유사도(합집합과 교집합의 비율)
- 코사인 유사도(코사인 각도 활용)

맨체스터 유나이티드 자코

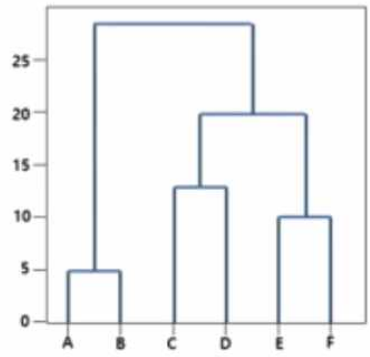
• 계층적 군집분석

(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 dnlc

(2) 덴드로그램

- 계층적 군집화를 시각적으로 나타내는 Tree모양의 그래프



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

• **K평균 군집화(K-mean Clustering)**
- 비계층적 군집화 방법으로 거리기반
(1) 특징
- 안정된 군집은 보장하나 최적의 보장은 어려움
- 한 번 군집에 속한 데이터는 중심점이 변경되면 군집이 변할 수 있음
- 초기 중심 값에 따라 결과가 달라짐
(2) 과정
1) 군집의 개수 K개 설정 (Elbow Method)를 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 없으며, 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 활용하여 차원축소를 통해 지도로 형상화하여 군집화하는 방법
- 완전연결의 형태를 가지며, 순전파 방식만 사용

• **실루엣 계수**
- 군집분석을 평가하는 지표로서 같은 군집간 가깝고, 다른 군집간 먼 정도를 판단(-1 ~ 1)

- 연관분석

• **연관분석**
- 항목들간의 조건-결과로 이루어지는 패턴을 발견하는 기법(장바구니 분석)
(1) 특징
- 결과가 단순하고 분명(IF ~ THEN ~)
- 강력한 비목적성 분석기법
- 품목 수가 증가할수록 계산량이 기하급수적으로 증가
- Apriori 알고리즘(최소 지지도 활용 빈발항목집합 추출)을 활용후, 연관분석을 수행
(2) 순차패턴
: 연관분석에 시간 개념을 추가하여 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 : $\frac{P(A \cap B)}{P(A)}$

- A품목이 거래될 때 B품목도 거래될 확률(조건부 확률)

(3) 향상도 : $\frac{P(A \cap B)}{P(A)P(B)}$

- A품목과 B품목의 상관성

(향상도 > 1 : 양의 상관관계, 향상도 = 1 : 상관없음, 향상도 < 1 : 음의 상관관계)

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 구매확률 = (10 + 20 + 30 + 40) / 200 = 0.5

(2) 치킨의 구매확률 = (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