1과목 데이터 이해

1장 데이터의 이해

1절 데이터와 정보

정성적 데이터		정량적 데이터
비정형 데이터, 주관적 내용 통계분석 어려움	↔	정형 데이터, 객관적 내용 통계분석 용이

암묵지	- 학습과 경험을 통해 개인에게 체화되어있지만 겉으로 드러나지 않는 지식 - 사회적으로 중요하지만 공유 어려움	공통화, 내면화
형식지	- 문서나 매뉴얼처럼 형상화된 지식 - 전달과 공유가 용이	표출화, 연결화

데이터 Data	개별 데이터 자체로는 의미가 중요하지 않은 객관적인 사실 ex) 연필은 A마트에서 100원, B마트에서 200원에 판매되고 있다.
정보 Information	데이터의 가공, 처리와 데이터간 연관관계 속에서 의미가 도출된 것 ex) A마트의 연필이 더 싸다.
지식 Knowledge	데이터를 통해 도출된 다양한 정보를 구조화하여 유의미한 정보를 분류하고 개인적인 경험을 결합시켜 고유의 지식으로 내재화된 것 ex) 상대적으로 저렴한 A마트에서 연필을 사야겠다.
지혜 Wisdom	지식의 축적과 아이디어가 결합된 창의적 산물 ex) A마트의 다른 상품들도 B마트 보다 쌀 것이다.

2절 데이터베이스 정의와 특징

▶ 데이터베이스의 일반적인 특징

① 통합된 데이터 : 동일한 내용의 데이터가 중복되어 있지 않음

② 저장된 데이터 : 컴퓨터가 접근할 수 있는 저장 매체에 저장되는 것

③ 공용 데이터 : 여러 사용자가 서로 다른 목적으로 데이터를 공동으로 이용

④ 변화되는 데이터 : 새로운 데이터의 삽입, 기존 데이터의 삭제, 갱신으로 항상 변화하면서도 현재의 정확한 데이터를 유지

▶ 데이터베이스의 다양한 측면에서의 특징

① 정보의 축적 및 전달 측면

- 기계가독성 : 일정한 형식에 따라 컴퓨터 등의 정보처리기기가 읽고 쓸 수 있음
- 검색가독성 : 다양한 방법으로 필요한 정보를 검색
- 원격조작성 : 정보통신망을 통하여 원거리에서도 즉시 온라인을 이용

② 정보 이용 측면

- 이용자의 정보 요구에 따라 다양한 정보를 신속하게 획득
- 원하는 정보를 정확하고 경제적으로 찾아낼 수 있음

③ 정보 관리 측면

- 정보를 일정한 질서와 구조에 따라 정리, 저장, 검색, 관리할 수 있도록 하여 방대한 양의 정보를 체계적으로 축적하고 새로운 내용의 추가나 갱신이 용이

④ 정보기술 발전 측면

- 정보처리, 검색 · 관리 소프트웨어, 관련 하드웨어, 정보 전송을 위한 네트워크 기술 발전을 견인할 수 있음

⑤ 경제 · 산업 측면

- 다양한 정보를 필요에 따라 신속하게 제공 · 이용할 수 있는 인프라로, 경제, 산업, 활동의 효율성을 제고하고 국민의 편의를 증진

3절 데이터베이스의 활용

▶ 1980년대 기업내부 데이터베이스

OLTP	On-Line Transaction Processing 호스트 컴퓨터와 온라인으로 접속된 여러 단말 간의 처리 형태 호스트 컴퓨터가 데이터베이스를 액세스하고, 처리 결과를 돌려보내는 형태
OLAP	On-Line Analytical Processing 정보 위주의 분석 처리로, 다양한 비즈니스 관점에서 쉽고 빠르게 다차원적인 데이터에 접근하여 의사결정에 활용할 수 있는 정보를 얻을 수 있게 해주는 기술

▶ 각 분야별 내부 데이터베이스

DW(Data Warehouse) : 기업의 의사결정 과정을 지원하기 위한 주제 중심적으로 통합적이며 시간성을 가지는 비휘발성 데이터의 집합

CRM(Customer Relationship Management) : 고객관계관리. 고객과 관련된 내·외부 자료를 분석·통합해 고객 중심 자원을 극대화하고 이를 토대로 고객특성에 맞게 마케팅 활동을 계획·지원·평가하는 과정

SCM(Supply Chain Management) : 공급망 관리. 기업에서 원재료의 생산·유통 등 모든 공급망 단계를 최적화해 수요자가 원하는 시간과 장소에 제품 제공

ERP(Enterprise Resource Planning) : 독립적으로 운영되던 각종 관리 시스템의 경영자원을 하나의 통합 시스템으로 재구축

BI(Business Intelligence) : 기업이 보유하고 있는 수많은 데이터를 정리하고 분석해 의사결정에 활용

RTE(Real-Time Enterprise) : 회사 전 부문의 정보를 하나로 통합함으로써 경영자의 빠른 의사결정을 끌어냄

EAI(Enterprise Application Inegration) : 개업 내 상호 연관된 모든 애플리케이션을 유기적으로 연동하여 필요한 정보를 중앙 집중적으로 통합, 관리, 사용할 수 있는 환경을 구서

EDW(Enterprise Data Warehouse) : 기존 DW를 전사적으로 확장한 모델로, BPR, CRM, BSC같은 다양한 분석 애플리케이션들을 위한 원천

KMS(Knowledge Management System) : 지식관리 시스템. 지적재산의 중요성이 커지며 기업 경영을 지식이라는 관점애서 새롭게 조명하는 접근방식

RFID(RF, Radio Frequency) : 주파수를 이용해 ID를 식별하는 시스템으로, 일명 전자태그

EDI(Electronic Data Interchange) : 각종 서류를 표준화된 양식을 통해 전자적 신호로 바꿔 거래처에 전송

VAN(Value Added Network) : 부가가치통신망. 통신회선을 차용하여 독자적인 네트워크 형성

CALS(Commerce At Light Speed) : 전자상거래 구축을 위해 기업 내에서 비용 절감과 생산성 향상을 추구할 목적으로 시작된, 제품의 라이프 사이클 전반에 관련된 데이터를 통합하고 공유·교환할 수 있도록 한 경영통합정보시스템

2장 데이터의 가치와 미래

1절 빅데이터의 이해

▶

양(Volume)	다양성(Variety)	속도(Velocity)
데이터의 규모 측면	데이터의 유형과 소스 측면	데이터의 수집과 처리 측면

+

4V
가치(Value) 시각화(Visualization) 정확성(Veracity)

빅데이터가 만들어내는 본질적인 변화

- 사전처리 → 사후처리, 표본조사 → 전수조사, 질 → 양, 인과관계 → 상관관계

2절 빅데이터의 가치와 영향

▶ 빅데이터 가치 산정이 어려운 이유

- 데이터 재사용, 재조합, 다목정 데이터 개발 등이 일반화되면서 특정 데이터를 누가 활용할지 알 수 없음
- 데이터가 기존에 없던 가치를 창출
- 추후에 새로운 분석기법이 등장하면 데이터의 가치가 변동

3절 비즈니스 모델

연관규칙 학습	변인들 간에 주목할만한 상관관계
유형분석	새로운 사건이 속하게 될 범주
유전 알고리즘	최적화가 필요한 문제를 자연선택, 돌연변이 등의 메커니즘으로 진화시킴
기계학습	훈련데이터로부터 학습한 알려진 특성을 활용해 예측
회귀분석	독립변수를 조작하며, 종속변수가 어떻게 변하는지 보며 두 변인의 관계 파악
감정분석	글을 쓴 사람의 감정을 분석
소셜네트워크 분석	고객들 간 소셜 관계 파악

4절 위기 요인과 통제 방안

위기 요인 : 사생활 침해, 책임 원칙 훼손, 데이터 오용

통제 방안 : 동의에서 책임으로, 결과 기반 책임 원칙 고수, 알고리즘 접근 허용

3장 가치창조를 위한 데이터 사이언스와 전략 인사이트

1절 빅데이터 분석과 인사이트

▶ 빅데이터 회의론의 원인 및 진단

- 투자효과를 거두지 못했던 부정적 학습효과 → CRM: 도입만 하면 모든 문제를 한번에 해소할 것처럼 강조
- 기존 분석 프로젝트를 빅데이터 분석으로 과대포장

2절 전략 인사이트 도출을 위한 필요 역량

▶ 데이터사이언티스트의 필요 역량

① Hard Skill : 빅데이터에 대한 이론적 지식, 분석 기술에 대한 숙련

② Soft Skill : 통찰력 있는 분석, 설득력있는 전달, 다분야간 협력

- Analytics: 수학, 확률모델, 머신러닝, 분석학, 패턴인식과 학습, 불확실성 모델링 등
- IT(Data Managing): 시그널 프로세싱, 프로그래밍, 데이터 엔지니어링, 데이터 웨어하우징, 고성능 컴퓨팅
- 비즈니스 분석: 커뮤니테이션, 프리젠테이션, 스토리텔링, 시각화

추가 최신 빅데이터 상식

DBMS(Data Base Management System) : 데이터베이스를 구축하는 틀을 제공하며, 효율적인 데이터 검색, 저장 기능 등을 제공

▶ 개인정보 비식별 기술

데이터 마스킹	데이터의 길이, 유형, 형식과 같은 속성을 유지한 채, 새롭고 읽기 쉬운 데이터를 익명으로 생성하는 기술
가명처리	개인정보 주체의 이름을 다른 이름으로 변경하는 기술, 다른 값으로 대체할 시 일정한 규칙이 노출되지 않도록 주의
총계처리	데이터의 총합 값을 보임으로서 개별 데이터의 값을 보이지 않도록 함
데이터값 삭제	데이터 공유, 개방 목적에 따라 데이터셋에 구성된 값 중에 필요 없는 값 또는 개인식별에 중요한 값을 삭제, 개인과 관련된 날짜 정보 연단위 처리
데이터 범주화	데이터의 값을 범주의 값으로 변환하여 값을 숨김

데이터 무결성(Data integrity) : 데이터베이스 내의 데이터에 대한 정확한 일관성, 유효성, 신뢰성을 보장하기 위해 데이터 변경/수정 시 여러 가지 제한을 두어 데이터의 정확성을 보증

데이터 레이크(Data Lake) : 수많은 정보 속에서 의미 있는 내용을 찾기 위해 방식에 상관없이 데이터를 저장하는 시스템으로, 대용량의 정형 및 비정형 데이터를 저장할 뿐만 아니라 접근도 쉽게 할 수 있는 대규모의 저장소

3과목 데이터 분석 기획

1장 데이터 분석 기획의 이해

1절 분석기획 방향성 도출

▶ 분석 대상과 방법

분석의 대상 (What)			분석의 방법 (How)
Known	Un-Known		
Optimization(최적화)	Insight(통찰)	Known	
Solution(솔루션)	Discovery(발견)	Un-Known	

▶ 분석 기획시 고려사항 : 가용데이터, 적절한 활용방안과 유즈케이스, 장애요소들에 대한 사전계획 수립

종류	정형 데이터	반정형 데이터	비정형 데이터
특징	- 데이터 자체로 분석 가능 - RDB구조의 데이터 - 데이터베이스로 관리	- 데이터로 분석 가능하지만 해석 불가능 - 메타정보를 활용해 해석 가능	- 데이터 자체로 분석 불가능 - 분석데이터로 변경 후 분석
유형	ERP, CRM, SCM 등 정보시스템	로그데이터, 모바일데이터, 센싱 데이터	파일형태로 저장, 관리 영상, 음성, 문자 등

2절 분석 방법론

▶ 기업의 합리적 의사결정을 가로막는 장애요소 : 고정관념, 편향된 생각, 프레이밍 효과

▶ 방법론 적용 업무의 특성에 따른 모델

① 폭포수 모델 : 단계를 순차적으로 진행

② 프로토타입 모델 : 점진적 개발. 일부분을 우선 개발하여 제공하고 그 결과를 통한 개선작업을 시행하는 모델

③ 나선형 모델 : 반복을 통해 점증적 개발. 관리체계를 효과적으로 갖추지 못한 경우 복잡도 상승

KDD	CRISP-DM	빅데이터 분석
분석대상 비즈니스 이해	업무 이해	분석 기획
데이터셋 선택	데이터의 이해	데이터 준비
데이터 전처리		
데이터 변환	데이터 준비	데이터 분석
데이터 마이닝	모델링	
데이터마이닝 결과 평가	평가	시스템 구현
데이터 마이닝 활용	전개	평가 및 전개

▶ KDD 분석 절차

① 데이터 선택 : 분석 대상의 비즈니스 도메인에 대한 이해와 프로젝트 목표 설정 필수. 데이터마이닝에 필요한 목표데이터 구성하여 분석에 활용

② 데이터 전처리 : 잡음, 이상치, 결측치 식별하고 제거. 추가로 요구되는 데이터셋 필요시 선택 프로세스 재실행

③ 데이터 변환 : 분석 목적에 맞게 변수 생성·선택, 데이터 차원 축소, train/test 데이터 분리

④ 데이터마이닝 : 데이터마이닝 기법 선택 및 실행. 필요에 따라 전처리와 변환 추가 실행

⑤ 데이터마이닝 결과 평가 : 결과에 대한 해석과 평가, 분석 목적과의 일치성 확인

▶ CRISP-DM 분석 방법론

① 업무이해 : 프로젝트의 목적과 요구사항 이해. 데이터 분석을 위한 문제정의. 프로젝트 계획 수립

② 데이터 이해 : 데이터를 수집, 속성을 이해. 문제점 식별 및 숨겨진 인사이트 발견

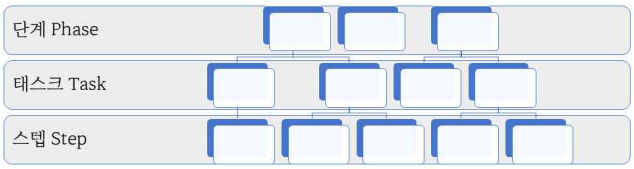
③ 데이터 준비 : 분석용 데이터셋 선택, 데이터 정제

④ 모델링 : 다양한 모델링 기법과 알고리즘 선택하고 최적화. 모델 평가

⑤ 평가 : 모델링 결과가 프로젝트 목적에 부합하는지 평가. 모델 적용성 평가

⑥ 전개 : 완성된 모델을 실 업무에 적용하기 위한 계획 수립. 유지보수 계획 마련. 프로젝트 종료 보고서 작성

▶ 빅데이터 분석 방법론



- 그룹을 통하여 완성된 단계별 산출물 생성
- 단계를 구성하는 단위활동으로 품질 검토의 항목
- WBS 워크 패키지에 해당. 입력자료, 처리 및 도구, 출력자료로 구성된 단위 프로세스

① 분석기획 : 비즈니스 이해 및 범위 설정, 프로젝트 정의 및 계획 수립, 프로젝트 위험 계획 수립
- 출력 자료 : 프로젝트 범위 정의서(SOW), 프로젝트 정의서, 모델 운영 이미지 설계서, 모델 평가 기준서, 프로젝트 수행 계획서, WBS, 위험관리 계획서

② 데이터 준비 : 필요 데이터 정의, 데이터 스토어 설계, 데이터 수집 및 정합성 점검
- 데이터 정의서, 데이터 획득 계획서, 데이터 스토어 설계서, 데이터 매핑 정의서, 정합성 점검 보고서

③ 데이터 분석 : 분석용 데이터 준비, 모델링, 모델 평가 및 검증, 모델 적용 및 운영방안 수립
- 비즈니스 룰, 분석용 데이터 셋, 데이터 탐색 보고서, 분석 보고서, 시각화 보고서, 모델링 결과 보고서, 알고리즘 설명서, 모델 평가 보고서, 모델 검증 보고서

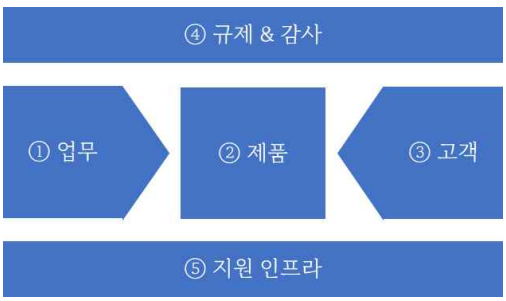
④ 시스템 구현 : 설계 및 구현, 시스템 테스트 및 운영
- 시스템 분석 및 설계서, 구현 시스템, 시스템 테스트 결과 보고서, 매뉴얼, 시스템 운영 계획서

⑤ 평가 및 전개 : 모델 발전 계획 수립, 프로젝트 평가 보고
- 모델 발전 계획서, 프로젝트 성과 평가서, 프로젝트 최종 보고서

3절 분석 과제 발굴

▶ 하향식 접근법 : 문제 탐색 → 문제 정의 → 해결방안 탐색 → 타당성 검토

① 문제 탐색 단계 : 세부적인 구현 및 솔루션이 아닌 문제를 해결함으로써 발생하는 가치에 중점
- 비즈니스 모델 기반 문제 탐색



업무 : 제품 및 서비스를 생산하기 위해서 운영하는 내부 프로세스 및 주요 자원 관점
제품 : 생산 및 제공하는 제품·서비스를 개선 관점
: 제품·서비스를 제공받는 사용자 및 고객, 이를 제공하는 채널의 관점
감사 : 제품 생산 및 전달 프로세스 중에서 발생하는 규제 및 보안 관점
지원 인프라 : 분석을 수행하는 시스템 영역 및 이를 운영·관리하는 인력 관점

- 분석 기회 발굴의 범위 확장
  - 거시적 관점 : 사회, 기술, 경제, 환경, 정치
  - 경쟁자 확대 : 경쟁사의 동향 - 대체제, 경쟁자, 신규 진입자
  - 시장니즈 탐색 : 고객, 채널, 영향자들
  - 역량의 재해석 : 역량의 변화 - 내부역량, 파트너 네트워크
- 외부 참조 모델 기반 문제 탐색 : 유사·동종 사례 벤치마킹을 통한 분석 기회 발굴
- 분석 유즈 케이스 : 현재의 비즈니스 모델 및 유사 동종사례 탐색을 통해 도출한 분석 기회들을 구체적으로 과제화 하기 전에 분석 유즈 케이스로 표기하는 것 필요

② 문제 정의 단계 : 비즈니스 문제를 데이터의 문제로 변환하여 정의

③ 해결방안 탐색 단계 : 정의된 데이터 분석 문제를 해결하기 위한 방안 모색. 분석 역량 파악

④ 타당성 검토
- 경제적 타당성 : 비용대비 편익 분석 관점의 접근
- 데이터 및 기술적 타당성 : 데이터 존재 여부, 분석 시스템 환경 및 분석 역량
  : 우월한 대안 선택 → 데이터 분석 문제 및 선정된 솔루션 방안 포함 → 분석과제 정의서의 형태로 명시 → 프로젝트 계획의 입력물로 활용

▶ 상향식 접근법 : 기업에서 보유하고 있는 다양한 원천 데이터로부터의 분석을 총하여 통찰력과 지식을 얻음
- 비지도 학습 : 데이터 자체의 결합, 연관성, 유사성 등을 중심으로 표현. 상향식 접근법에서 많이 사용
  ex) 장바구니 분석, 군집 분석, 기술 통계 및 프로파일링
- 지도 학습 : 명확한 목적 하에 데이터 분석을 실시하는 것. 분류, 추측, 예측, 최적화를 통해 지식 도출

▶ 빅데이터 분석 환경에서 프로토타이핑의 필요성

① 문제에 대한 인식 수준 : 문제 정의가 불명확하거나 새로운 경우, 문제 이해와 구체화에 도움

② 필요 데이터 존재 여부의 불확실성 : 데이터 존재하지 않을 경우 사용자와 분석가 간의 반복적이고 순환적인 협의 과정이 필요. 대체 불가능한 데이터 있다면 수행하기 전에 리스크 방지 가능

③ 데이터 사용 목적의 가변성 : 기존의 데이터 정의를 재검토하여 데이터 사용 목적과 범위 확대 가능

4절 분석 프로젝트 관리 방안

▶ 분석 과제 관리를 위한 5가지 주요 영역

① Data Size : 분석하고자 하는 데이터의 양

② Data Complexity : 데이터에 잘 적용될 수 있는 분석 모델 선정 사전에 고려

③ Speed : 시나리오 측면에서의 속도 고려. 실시간 탐지 vs 주 단위 실적

④ Analytic Complexity : 복잡도가 높아지면 고객에게 설명이 어려움. 해석이 가능한 최적모델 필요

⑤ Accuracy & Precision : 활용 측면에서 Accuracy, 안정성 측면에서 Precision 중요

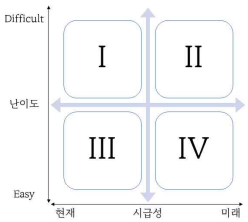
▶ 분석 프로젝트 영역별 주요 관리 항목 : 범위, 시간, 원가, 품질, 통합, 조달, 자원, 리스크, 의사소통, 이해관계자

2장 분석 마스터 플랜

1절 마스터 플랜 수립 프레임워크

▶ 적용 우선순위 설정
- 우선순위 고려요소 : 전략적 중요도, 비즈니스 성과/ROI, 실행 용이성
  - 전략적 중요도 : 전략적 필요성 & 시급성
  - 실행 용이성 : 투자 용이성, 기술 용이성
  - ROI : 투자비용(Investment) 요소 - Volume, Variety, Velocity / 비즈니스 효과(Return) 요소 - Value
- 적용범위/방식 고려요소 : 업무 내재화 적용 수준, 분석 데이터 적용 수준, 기술 적용 수준

▶ 포트폴리오 사분면 분석을 통한 과제 우선순위 선정



- 시급성 우선 : III → IV → II
- 난이도 우선 : III → I → II

▶ 로드맵 수립 : 포트폴리오 사분면 분석을 통해 1차 우선순위 결정 → 분석 과제별 적용범위 및 방식 고려하여 우선순위 결정 후 로드맵 수립 → 단계별 추진 목표 정의 → 추진 과제별 선·후행 관계 고려하여 단계별 추진 내용 정렬

2절 분석 거버넌스 체계 수립

▶ 구성요소 : 분석기획 및 관리 조직, 과제 기획 및 운영 프로세스, 분석 관련 시스템, 데이터, 관련 교육 및 마인드 육성 체계

▶ 데이터 분석 수준진단
- 분석 준비도 : 분석 업무, 분석인력·조직, 분석 기법, 분석 데이터, 분석 문화, 분석 인프라
- 분석 성숙도 : 도입 → 활용 → 확산 → 최적화
  ① 도입 : 분석을 시작하여 환경과 시스템 구축
  ② 활용 : 분석 결과를 실제 업무에 적용
  ③ 확산 : 전사 차원에서 분석을 관리하고 공유
  ④ 최적화 : 분석을 진화시켜서 혁신 및 성과 향상에 기여

▶ 데이터 거버넌스 구성 3요소

① 원칙 : 유지·관리를 위한 지침과 가이드

② 조직 : 데이터를 관리할 조직의 역할과 책임

③ 프로세스 : 데이터 관리를 위한 활동과 체계

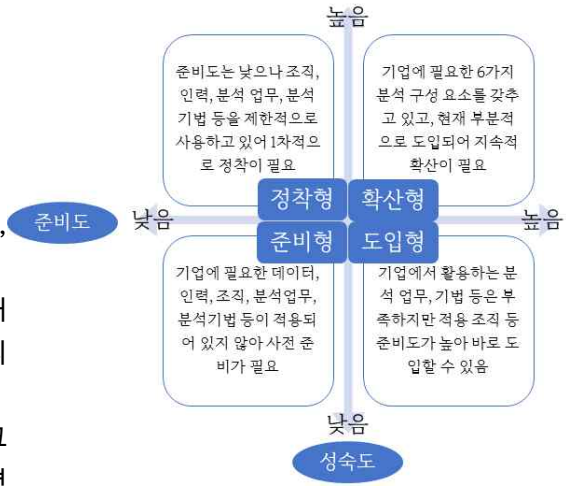
▶ 데이터 거버넌스 체계

① 데이터 표준화 : 데이터 표준 용어 설정, 명명 규칙 수립, 메타 데이터 구축, 데이터 사전 구축

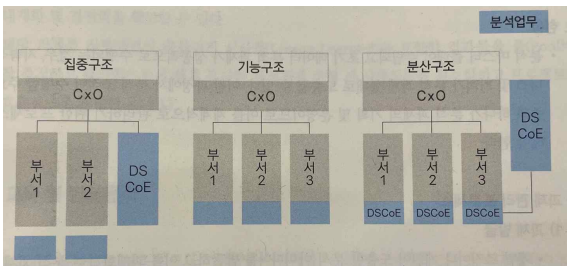
② 데이터 관리 체계 : 데이터 정합성 및 활용 효율성을 위해 표준 데이터를 포함한 메타 데이터와 데이터 사전의 관리 원칙 수립

③ 데이터 저장소 관리 : 데이터 관리 체계 지원을 위한 워크플로우 및 관리용 응용 소프트웨어 지원, 데이터 구조 변경에 따른 사전영향평가

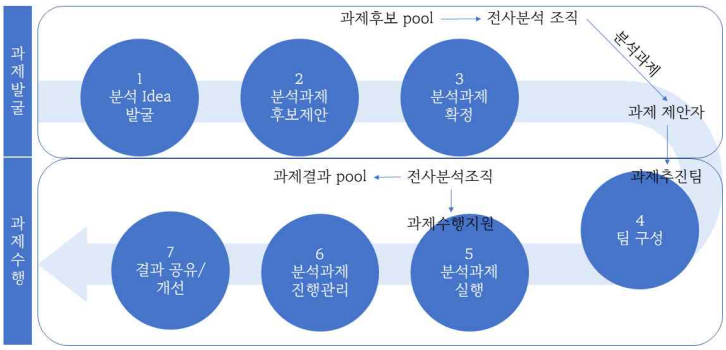
④ 표준화 활동 : 표준 준수 여부 점검 및 모니터링, 변화 관리 및 주기적 교육 진행



▶ 분석 조직 구조



▶ 분석과제 관리 프로세스



4과목 데이터 분석

1장 데이터 분석 개요

1절 데이터 분석 기법의 이해

시각화 : 빅데이터 분석 및 탐색적 분석에 필수. 복잡한 분석보다 더 효율적일 수 있음. SNA(사회연결망)분석에 자주 활용

공간분석(GIS) : 공간적 차원과 관련된 속성들을 시각화

탐색적 자료 분석(EDA) : 데이터 특징과 내재하는 구조적 관계를 알아내기 위한 기법들의 통칭

- 4가지 주제 : 저항성의 강조, 잔차 계산, 자료변수의 재표현, 그래프를 통한 현시성
- 데이터 이해, 변수생성, 변수 선택 단계에서 활용

기술통계: 모집단으로부터 표본을 추출하고 표본이 가지고 있는 정보를 쉽게 파악할 수 있도록 표현

추론통계 : 표본의 표본통계량으로부터 모수에 관해 통계적으로 추론

▶ 데이터마이닝 : 대표적인 고급 데이터 분석법으로 대용량의 자료로부터 관계, 패턴, 규칙 등을 탐색하고 모형화하여 유용한 지식을 추출

- 방법론 : 데이터베이스에서의 지식탐색, 기계학습(인공신경망, 의사결정나무, 군집화, 베이지안분류, SVM 등), 패턴인식(장바구니 분석, 연관규칙 등)

수행방안의 최종 산출물 : 분석계획서와 WBS(Work Breakdown Structure)

▶ 모델링 성능평가

- 데이터마이닝 : 정확도, 정밀도, detect rate, 향상도(lift)
- 시뮬레이션 : Throughput, Average Wating Time, Average Que Length, Time in System
- 최적화 : 최적화 이전 Object Function Value와 최적화 이후 값의 차이

2장 R 프로그래밍 기초

▶ R 특징
- 오픈소스 프로그램
- 그래픽 성능이 상용 프로그램과 대등하거나 월등
- 각 세션 사이마다 시스템에 데이터셋을 저장하므로 매번 로딩할 필요 없음
- S통계 언어 기반 구현
- 객체지향 언어이며 함수형 언어

▶ 헷갈리는 문법

apply(df, 1or2, func) : 함수 적용 / 1:행, 2:열
sapply(var, func) : 함수 적용 / 벡터 반환
lapply(var, func) : 함수 적용 / 리스트 반환
tapply(vec, fac, func) : 요소별로 함수 적용
subset(df, select=var, subset=조건) = df[df$col_name="col_name] ??

3장 데이터 마트

1절 데이터 변경 및 요약

▶ 데이터 마트 : 데이터 웨어하우스와 사용자 사이의 중간층
- 요약변수 : 분석에 맞게 종합한 변수로 많은 모델에 공통으로 사용될 수 있어 재활용성 높음
- 파생변수 : 특정 조건을 만족하거나 특정함수에 의해 만들어 의미 부여한 변수로 주관적이기 때문에 논리적 타당성을 갖추어 개발해야 함

▶ reshape 활용
- melt(df, id=c(col_names)) : 기준이 될 column 골라 원데이터 형태로 만드는 함수
- cast(df, A~B) : A를 기준으로 잡고 B에 대해 요약형태로 만드는 함수. 여러 개는 +로 이어줌

▶ sqldf 활용 : sql 명령어 이용 가능 - sqldf("sql문 작성“)

▶ plyr 활용 : split-apply-combine
: 데이터 형태의 이니셜 따서 함수 사용(?) ex) df+df : ddply / list+df : ldply

▶ data.table : data.frame보다 빠름, 빠른 그루핑과 ordering, 짧은 문장 지원 측면에서 유용

2절 데이터 가공

summary(dataset)
: 수치형 변수-최대값, 최소값, 평균, 1사분위수, 2사분위수, 3사분위수 / 명목형 변수 - 명목값, 데이터 개수

3절 기초 분석 및 데이터 관리

▶ 결측값 처리 방법
- completes analysis : 결측값이 존재하는 레코드 삭제 - R : complete.cases(), is.na()
- 평균대치법 : 비조건부 평균 - 관측데이터의 평균으로 대치 / 조건부 평균 - 회귀분석 활용
  - R : centralImputation() 숫자는 중위값, factor는 최빈값
- 단순확률대치법 : Hot-deck, nearest neighbor - R : knnImputation()
- 다중대치법 : m번의 대치를 통해 m개의 가상 자료 만듦. 대치 → 분석 → 결합
- ect : rfImpute() 랜덤포레스트에서 사용

▶ 이상값 인식과 처리
- ESD : 평균으로부터 3표준편차 떨어진 값
- 기하평균-2.5x표준편차 〈 data 〈 기하평균+2.5x표준편차
- boxplot outer fence 바깥 제거
- 극단값 절단 : 기하평균 이용, 상하단 절단
- 극단값 조정 : 이상치를 상한값/하한값으로 바꿈

4장 통계분석

1절 통계분석의 이해

▶ 표본 추출 방법

- 단순랜덤 추출방법 : 임의의 n개
- 계통추출법 : 임의 위치에서 매 k번째 항목 추출
- 집락추출법 : 군집 분류하고 군집별로 단순랜덤 추출
- 층화추출법 : 유사한 원소끼리 몇 개의 층으로 나누어 랜덤 추출

기술통계 : 평균, 표준편차, 중위수, 최빈값, 그래프, 왜도, 첨도 등

통계적 추론 : 모수추정 → 가설검정 → 예측

▶ 확률변수 : 특정값이 나타날 가능성이 확률적으로 주어지는 변수

- 기댓값 : $X) = \begin{cases} x\, f(x_i) \\ x\, f(x)dx \end{cases}$
- k차 적률 : $E(X^k) = \begin{cases} \sum x_i f(x_i) \\ \int x_i\, f(x)dx \end{cases}$
- k차 중심적률 : $E[(X-\mu)^k] = \begin{cases} \sum (x_i-\mu)^k f(x_i) \\ \int (x-\mu)^k f(x)dx \end{cases}$

▶ 이산형 확률변수

- 베르누이 확률분포 : 결과가 2개
- 이항분포 : 베르누이를 n번 반복했을 때 k번 성공할 확률
- 기하분포 : 성공확률이 p인 베르누이 시행에서 첫 성공이 있기까지 x번 실패할 확률
- 다항분포 : 이항분포의 확장으로, 세 가지 이상의 결과를 가지는 시행에서 발생하는 확률분포
- 포아송분포 : 시간과 공간 내에서 발생하는 사건의 발생횟수에 대한 확률분포

▶ 연속형 확률변수

- 균일분포
- 정규분포 : 평균이 μ, 표준편차가 σ인 x의 확률밀도함수
- 지수분포 : 어떤 사건이 발생할 때까지 경과 시간에 대한 확률분포
- t-분포 : 표본이 커져서(30이상) 자유도 증가하면 표준정규분포에 근접. 두 집단의 평균이 동일한지 알고자 할 때 검정통계량으로 활용
- x^2분포 : 모평균과 모분산이 알려지지 않은 집단에 대한 가설검정에 사용. 두 집단 간의 동질성 검정에 활용
- F-분포 : 두 집단 간 분산의 동일성 검정에 활용, 자유도 2개, 자유도 커질수록 정규분포에 근접

▶ 가설검정

- 귀무가설(H0) : 증명하고자 하는 가설
- 대립가설(H1) : H0에 반대되는 가설
- 검정통계량 : 관찰된 표본으로부터 구하는 통계량, 검정시 가설의 진위를 판단하는 기준
- 유의수준 : 귀무가설이 옳은데도 이를 기각하는 확률의 크기 = α
- 기각역 : 귀무가설이 옳다는 전제 하에 구한 검정통계량의 분포에서 확률이 유의수준 α인 부분

	사실이라고 판정	거짓으로 판정
H0은 사실	옳은 결정	제1종오류 α
H0은 거짓	제2종오류 β	옳은 결정

▶ 비모수 검정 : 추출된 모집단의 분포에 대한 아무 제약을 가하지 않고 검정. 자료 수가 적거나 서열관계인 경우

- 가정된 분포가 없으므로 분포의 형태에 대해 설정
- 관측값들의 순위나 두 관측값 차이의 부호 등을 이용

2절 기초 통계분석

중심위치 측도 : 자료, 표본평균, 중앙값

산포 측도 : 분산, 표준편차, 사분위수범위, 사분위수, 백분위수, 변동계수, 평균의 표준오차

분포의 형태에 관한 측도 : 왜도, 첨도

▶ 그래프

- 막대그래프 : 범주형 데이터, 순서 의도에 따라 바꿀 수 있음
- 히스토그램 : 연속형 데이터, 임의로 순서 바꿀 수 없고, 막대의 간격 없음
- 줄기-잎 그림 : 데이터를 줄기와 잎 모양으로 그림
- 상자그림 : 사분위수 범위(Q1-Q3) 상자, 안울타리(Q1+-1.5 x IQR), 바깥울타리(Q1+-3 x IQR)
  - 보통 이상점 : 바깥울타리와 안울타리 사이 / 극단이상점 : 바깥울타리 밖의 자료
- 산점도 : 좌표평면 위에 점들로 표현한 그래프

독립변수 : 설명변수, x / 종속변수 : 반응변수, y

공분산 : 두 확률변수 X, Y의 방향의 조합(선형성) - $Cov(X, Y) = E[(X-\mu_X)(Y-\mu_Y)]$

▶ 상관분석

- 1~0.7 : 강한 상관 / 0.7~0.3 : 약한 상관 / + : 양의 상관 / - : 음의 상관
- 피어슨 : 연속형 변수, 정규성 가정 / 스피어만 : 순서형 변수, 비모수적 방법, 순위 기준

3절 회귀분석

회귀분석의 가정 : 선형성, 등분산성, 독립성, 비상관성, 정상성(정규성)

종류 : 단순회귀, 다중회귀, 로지스틱 회귀(종속변수가 범주형 2진 변수), 다항회귀, 곡선회귀, 비선형회귀

▶ 단순선형회귀분석

- 검토사항 : t분포 p값이 .05보다 작으면 유의함 / 결정계수(R^2) 높을수록 설명력 높음 / 잔차그래프로 회귀진단
- 추정 : 최소제곱법(=최소자승법)

▶ 다중선형분석

- F통계량의 p값이 .05보다 작으면 유의함 / 결정계수 혹은 수정된 결정계수 확인 / 잔차와 종속변수 산점도 확인

▶ 최적회귀방정식 : 설명변수 선택 → 모형 선택 → 단계적 변수선택

- 벌점화된 선택 기준 : 모든 후보 모형들에 AIC, BIC 적용하고 값이 최소가 되는 모형 선택
- 전진선택법 : 상수부터 시작해 중요한 설명변수 추가
- 후진제거법 : 영향 적은 변수부터 제거
- 단계선택법 : 전진선택법으로 추가하다가 영향 적어지는 변수 제거

4절 시계열 분석

▶ 정상성

- 평균이 일정 : 차분을 통해 정상화 가능
- 분산이 일정 : 변환을 통해 정상화 가능
- 공분산도 단지 시차에만 의존, 실제 특정 시점 t, s에는 의존하지 않음
- 어떤 시점에서 자기공분산을 측정하더라도 동일한 값을 갖는다.
- 평균으로 회귀하려는 경향이 있으며, 변동은 대체로 일정한 폭을 갖는다.
- 정상시계열이 아닌 경우 다른 시기로 일반화할 수 없다.

▶ 분석방법

- 수학적 이론모형 : 회귀분석방법, Box-Jenkins 방법 / 직관적 방법 : 지수평활법, 시계열분해법
- 장기예측 : 회귀분석방법 / 단기예측 : Box-Jenkins 방법, 지수평활법, 시계열 분해법

▶ 자기회귀모형(AR 모형)

- p시점 전의 자료가 현재 자료에 영향을 주는 모형
- AR(1) 모형 : 직전 시점 데이터로만 분석
- AR(2) 모형 : 연속된 3시점 정도의 데이터로 분석
- 자기상관함수(ACF)는 빠르게 감소, 부분자기함수(PACF)는 어느 시점에서 절단점

▶ 이동평균모형(MA 모형)
- 유한개수의 백색잡음의 결합 → 정상성 만족
- MA1 모형 : 같은 시점의 백색잡음과 바로 전 시점의 백색잡음의 결합
- MA2 모형 : 바로 전 시점의 백색잡음과 시차가 2인 백색잡음의 결합
- ACF에서 절단점, PACF 빠르게 감소

▶ 자기회귀누적이동평균모형(ARIMA)
- 비정상 시계열 모형 → 차분, 변환을 통해 AR모형, MA모형 혹은 둘을 합친 ARMA모형으로 정상화 가능
- ARIMA(p,d,q) : p는 AR 모형 / q는 MA 모형과 관련 / d는 시계열 $\{Z_t\}$의 차분 횟수
  - d=0 : ARMA(p,q) / p=0 : IMA(d,q) / q=0 : ARI(d,p)

▶ 분해시계열: Z = f(T, S, C, I) T=경향요인, S=계절요인, C=순환요인, I=불규칙요인 - 회귀분석적 방법 사용

5절 다차원척도법

- 객체간 근접성을 시각화하는 통계기법으로 군집분석처럼 유사성/비유사성을 측정하여 2차원 공간에 점으로 표현
- 데이터 속에 잠재해있는 패턴, 구조 찾고 기하학적으로 표현
- 데이터 축소의 목적
- 유클리드 거리행렬 사용
- 적합 정도를 Stress Value로 나타내며, 부적합도 기준으로 STRESS나 S-STRESS사용
  - 0 : 완벽 / .05 : 매우 좋은 / .1 : 만족 / .15 : 보통 / .15 이상 : 나쁨
- 계량적 MDS : 구간척도나 비율척도에 활용, 유클리드 거리 사용
- 비계량적 MDS : 순서척도에 활용, 거리의 속성과 같도록 변환하여 거리 생성

6절 주성분분석

- 서로 상관성이 높은 변수들의 선형결합으로 만들어 상관성이 높은 변수들을 요약, 축소하는 기법
- 다중공선성이 존재하는 경우 상관성이 없는 주성분으로 변수들 축소하여 모형 개발에 활용
- 군집화 결과와 연산속도 개선 가능
- 누적기여율 85% 이상이면 주성분의 수로 결정
- scree plot을 활용해 주성분 개수 결정

5장 정형 데이터 마이닝

1절 데이터마이닝의 개요

예측 Predict	분류규칙 Classification	회귀분석, 판별분석, 신경망, 의사결정나무
설명 Descriptive	연관규칙 Association	동시발생 매트릭스
	연속규칙 Sequence	동시발생 매트릭스
	데이터 군집화 Clustering	K-Means Clustering

데이터마이닝 추진 단계 : 목적 설정(모델/데이터 정의) → 데이터 준비 → 가공 → 기법 적용 → 검증

데이터 분할 : training 50%, validation 30%, test 20% / hold-out / k-fold cross-validation

▶ 혼동행렬(Confusion Matrix)

		답	
		Positive	Negative
예측	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

- Accuracy(정분류율) : $\frac{TN+TP}{N \quad TP+FN+FP} = \frac{\text{정답}}{\text{체}}$
- Error Rate(오분류율) : 1 - Accuracy
- Recall(재현율) = Sensitivity(민감도) : $\frac{TP}{TP+FN}$ = 답이 Positive인 것 중 정답
- Precision(정확도) : $\frac{TP}{TP+FP}$ = Positive로 예측한 것 중 정답
- Specificity(특이도) : $\frac{TN}{TN+FP}$ = 답이 Negative인 것 중 정답
- F1 Score = $2\times\frac{\mathrm{P}recision\times Recall}{\mathrm{P}recision+Recall}$

▶ ROC Curve

- 가로축 : FP Rate(=1-Specificity), 세로축 : TP Rate(Sensitivity)
- 좌상단에 가까울수록(AUROC, Area Under ROC 넓을수록) 성능 좋음

이익도표 : Lift 향상도 = 반응률 → 빠르게 감소할수록 좋음 (%Captured Response = 해당등급 빈도/전체)

2절 분류분석

분류모델링 : 신용평가모형, 사기방지모형, 이탈모형, 고객세분화

분류기법 : 회귀분석, 의사결정나무, CART, C5.0, 베이지안 분류, 인공신경망, SVM, KNN, Case-Based

▶ 로지스틱 회귀분석 : 새로운 독립변수가 주어질 때 종속변수의 각 범주에 속할 확률 추정하여 분류

- 종속변수 : 범주형 → 추정된 확률 : 사후확률
- β 〉 0 : S모양 / β 〈 0 : 역 S모양
- 오즈비(odds ratio) : 성공할 확률이 실패할 확률의 몇 배인지를 나타내는 확률인 오즈(odds)의 비율
- glm(종속변수 ~ 독립변수, family=binomial, data=data_name)

	선형회귀분석	로지스틱 회귀분석
종속변수	연속형 변수	0, 1
계수 추정법	최소제곱법	최대우도추정법
모형검정	F검정, T검정	카이제곱 검정

▶ 의사결정나무 : 분류함수를 의사결정 규칙으로 이루어진 나무모양으로 시각화

- 예측이 중요할 경우 예측력에 치중하고, 이유의 설명이 중요할 경우 해석력에 치중해야 함
- 활용 : 세분화, 분류, 예측, 차원축소 및 변수선택, 교호작용효과 파악
- 장점 : 설명용이, 계산 복잡x, 대용량 데이터에도 빠르게 만들 수 있음, 잡음데이터에 민감x, 한 변수와 상관성이 높은 불필요한 변수에 영향 크게 받지 않음, 수치형/범주형 변수 모두 사용 가능, 모형분류 정확도 높음
- 단점 : 과대적합 가능성 높음, 분류 경계선 부근 오차 큼, 설명변수 간 중요도 판단 어려움
- 분석 과정 : 성장(최적의 분리 규칙 찾고, 정지규칙 만족하면 중단) → 가지치기(불필요 가지 제거) → 타당성 평가(이익도표, 위험도표, 시험자료 이용) → 해석 및 예측
- 분리 기준: 이산형- 카이제곱 p값, 지니 지수, 엔트로피 지수 / 연속형- 분산분석에서 F 통계량, 분산의 감소량
- party : ctree(____ ~ ., data=data_name)

3절 앙상블 분석

▶ 배깅 : 여러 개의 bootstrap 자료 생성하고 예측모형 결합하여 최종 예측모형 만드는 방법

- bootstrap : 주어진 자료에서 동일한 크기의 표본을 랜덤 복원추출로 뽑은 자료
- voting : 여러 모형에서 산출된 결과를 다수결에 의해 최종 결과 산정
- 배깅에서는 pruning 하지 않고 최대로 성장한 의사결정나무 활용
- 훈련자료를 모집단으로 생각하고 평균예측모형을 구하여 분산을 줄이고 예측력 상승

▶ 부스팅 : 예측력 약한 모형들을 결합하여 강한 예측모형을 만드는 방법

- Adaboost : 이진분류에서 랜덤분류기보다 조금 더 좋은 분류기 n개에 가중치 설정하고 결합하여 최종 분류기
- 배깅보다 Adaboost가 뛰어난 경우 많음

▶ 랜덤포레스트 : 약한 학습기들을 생성한 후 선형결합하여 최종 학습기 만드는 방법

- randomForest(____ ~ ., data=data_name, ntree=num, proximity=TRUE)

4절 인공신경망 분석

- 인간의 뇌 기반, 뉴런이 기본 정보처리 단위
- 활성화 함수 사용 : sigmoid(0~1), softmax(출력이 여러개, 범주에 속할 사후확률), ReLU(max(0, a))
- 적합한 입력변수 : 범주형- 모든 범주에서 일정 빈도 이상 & 빈도 일정 / 연속형- 변수 간 큰 범위차이x - 변환/범주화로 분포를 변환시켜줌
- 역전파 알고리즘은 초기값에 따라 결과 많이 달라짐. 가중치 0이면 근사적으로 선형 모형이 됨
- 온라인 학습모드 : 순차로 하나씩 신경망에 투입 / 확률적 학습모드 : 투입 순서 랜덤 / 배치 학습 모드 : 전체 훈련자료를 동시에 투입
- 은닉층과 은닉노드가 많으면 과적합 문제 발생 / 적으면 과소적합 문제 발생

5절 군집분석

- 유사성을 바탕으로 집단 분류
- 차이점 : 요인분석 - 유사한 변수 묶음 / 판별분석 - 사전에 나누어진 자료 바탕으로 새로운 데이터를 할당

▶ 거리

- 연속형 변수 : 유클리디안 거리(일반적 거리), 표준화 거리(표준편차로 척도변환 후 유클리디안), 마할라노비스 거리(표준편차와 변수 간 상관성 고려), 체비셰프 거리, 맨하탄 거리(절대값), 캔버라 거리, 민코우스키 거리(L1:맨하탄, L2:유클리디안)
- 범주형 변수 : 자카드 거리, 자카드 계수, 코사인 거리, 코사인 유사도

▶ 계층적 군집분석

- 최단연결법 : 가장 가까운 데이터 묶어 군집형성 → 최단거리로 계산하여 수정 → 가까운 데이터를 새 군집으로
- 최장연결법 : 최장거리를 거리로 계산하여 거리행렬 수정
- 평균연결법 : 평균을 거리로 계산하여 거리행렬 수정
- 와드연결법 : 군집내 편차들의 제곱합 고려, 정보손실 최소화
- 군집화 단계 : 거리행렬 기준 덴드로그램 → 최상단부터 가로선을 그어 군집 개수 선택 → 적당한 군집 수 선정

▶ 비계층적 군집분석 : K-means clustering

- 각 클러스터와 거리 차이의 분산을 최소화하여 k개의 클러스터로 묶는 알고리즘
- 과정 : seed 정함 → seed 중심으로 군집 형성 → 가장 가까운 seed 있는 군집에 분류 → seed 다시 계산 → 모두 할당될 때까지 반복
- 연속형 변수에 활용 가능
- 초기 중심값 임의로 선택 가능, 초기 중심값에 따라 결과 달라짐
- 탐욕적 알고리즘이라 안정된 군집이지만 최적은 보장할 수 없음
- 장점 : 알고리즘 단순, 빠름, 많은 양의 데이터 다룰 수 있음, 사전정보 없어도 가능, 다양한 형태에 적용 가능
- 단점 : 군집 수/가중치/거리 정의 어려움, 사전 목적 없으므로 결과해석 어려움, 볼록하지 않은(non-convex) 군집이 존재하면 성능 떨어짐, 초기 군집 수 결정 어려움

▶ 혼합 분포 군집
- 모수와 가중치의 추정(최대우도추정)에 EM 알고리즘 사용
- EM(Expectation-Maximization) 알고리즘 진행과정
  : E단계(잠재변수 Z 기대치 계산) → M단계(기대치 이용하여 파라미터 추정)
- 확률분포를 도입하여 군집 수행, K-means처럼 이상치에 민감
- 데이터 커지면 수렴에 시간 걸림, 크기 너무 작으면 추정 어려움

▶ SOM(Self Organizing Map)
- 비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도의 형태로 형상화
- 구성 : 입력층(입력변수의 개수와 동일한 뉴런 수) - 경쟁층(2차원 격자, 승자 독식 구조)
- 지도 형태로 형상화하기 때문에 시각적으로 이해 쉬움
- 단 하나의 전방 패스(feed-forward flow)를 사용하여 속도 빠름 → 실시간 학습처리 가능

6절 연관분석
- 장바구니분석, 서열분석
- 측도 : 지지도 = P(A∩B) , 신뢰도 = 지지도/P(A) , 향상도 = 신뢰도/P(B)
- 절차 : 최소 지지도 결정 → 최소 지지도 넘는 품목 분류 → 반복적으로 수행해 빈발품목 집합 찾음
- Apriori 알고리즘 : 최소지지도 이상의 빈발항목집합을 찾은 후 그것에 대해서만 연관 규칙 계산
- FP-Growth 알고리즘 : 후보 빈발항목집합 생성하지 않고 Frequent Pattern Tree 만든 후 분할 정복