

47회 ADsP 합격을 완성하는 마무리 1%

DATA
ANALYSIS
Level 1.5

DATA ANALYSIS

AI EDU AIEDU 대표 : 김 계 철

한 권으로 끝내기 ADsP

ADSP 코어프렙 앱 활용 가이드

AI EDU AIEDU



앱 설치 (QR or 스토어 검색)



회원가입



휴대폰 인증



도서 구매 인증

1. 앱 설치 (QR or 스토어 검색)

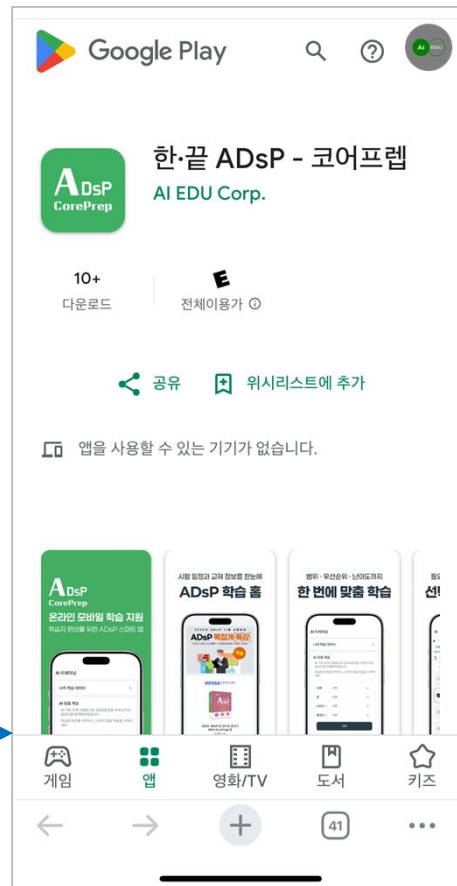
Android

1) QR 코드 촬영하여 설치



또는

2) 구글 스토어에서
"한·끝 Adsp" 검색 후
앱 다운로드



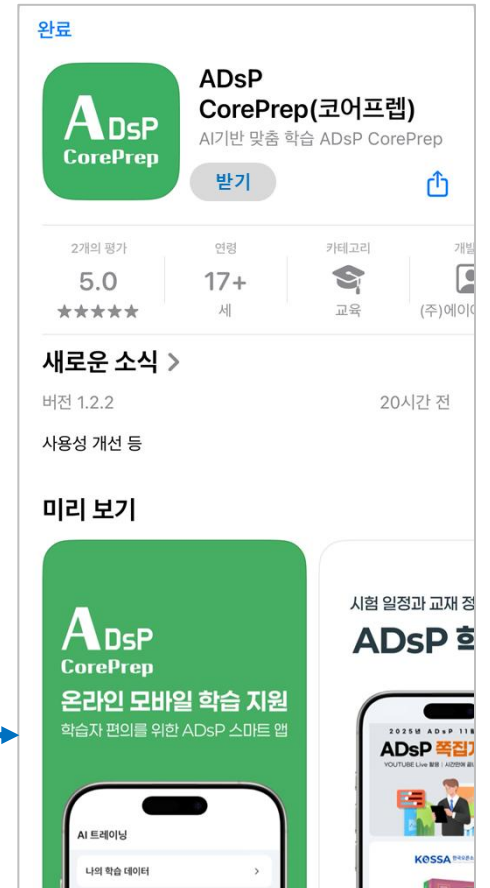
iOS

1) QR 코드 촬영하여 설치



또는

2) 앱 스토어에서
"adsp 코어프렘" 검색 후
앱 다운로드



앱 설치 (QR or 스토어 검색)



회원가입

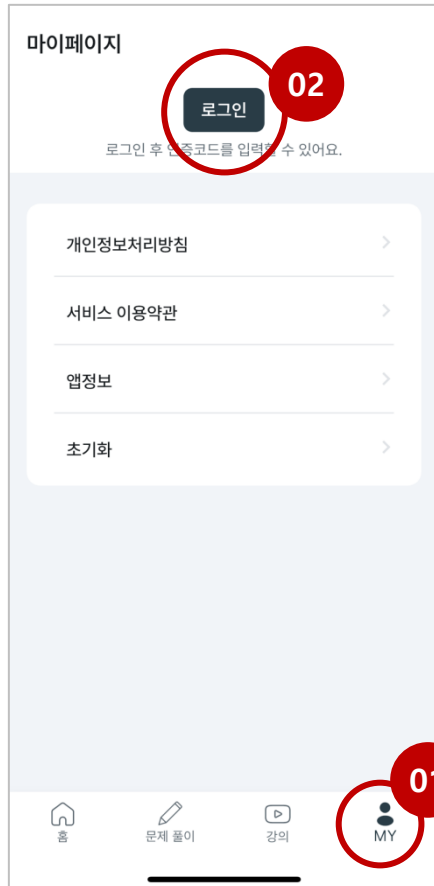


휴대폰 인증

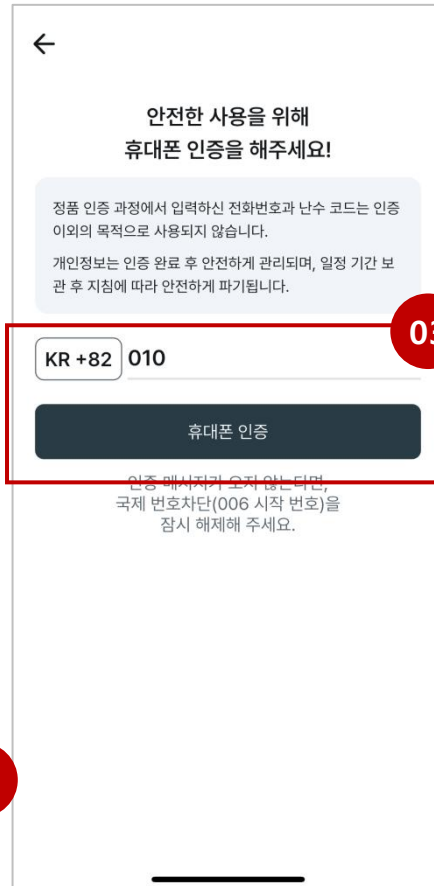


도서 구매 인증

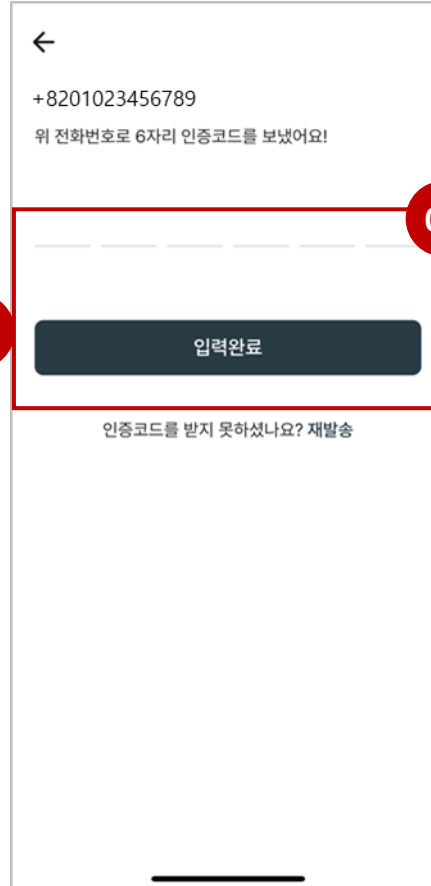
회원가입



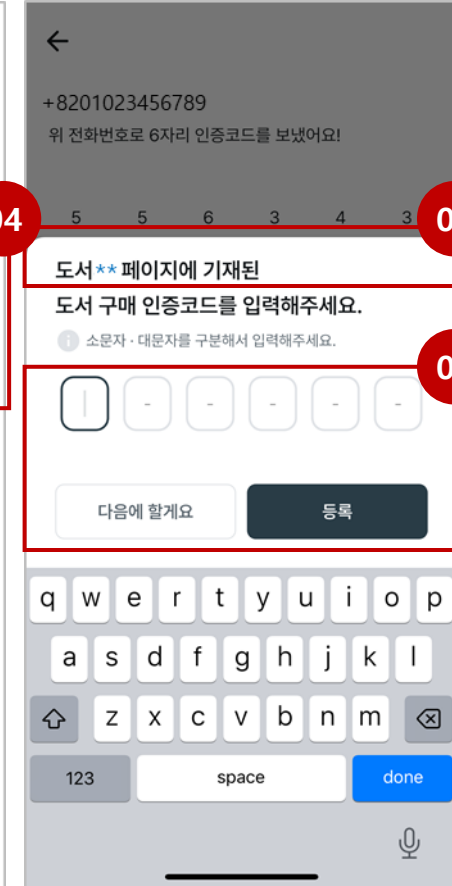
휴대폰으로 로그인



휴대폰 인증



도서 구매 인증



01.회원가입

메인에서 MY 클릭해서
MY페이지 이동 후
로그인 버튼 클릭

02.휴대폰으로 로그인

본인 휴대폰으로 입력 후 휴대폰 인증

03.본인 휴대폰 인증

006~으로 시작되는 번호로 발신된
인증코드 인증

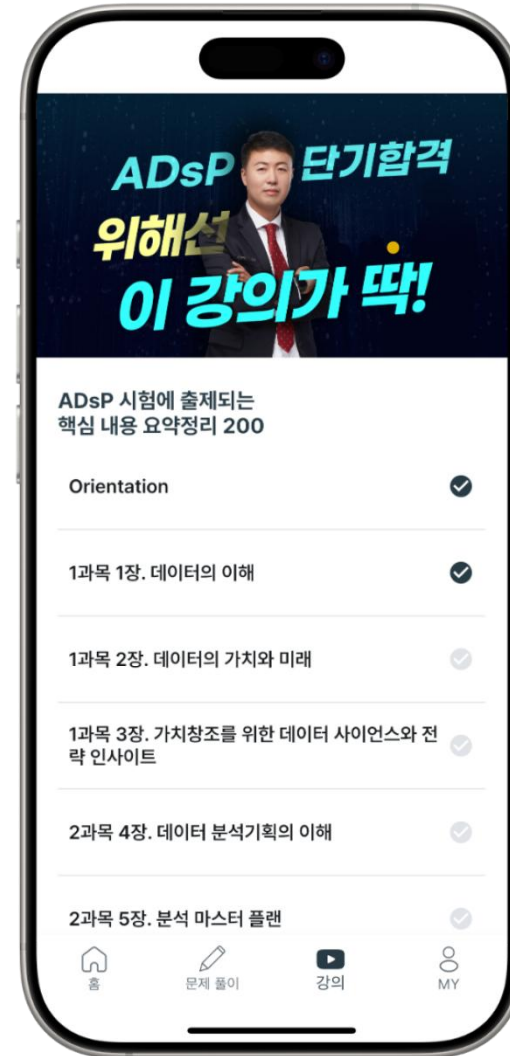
04.도서 구매 인증

휴대폰 인증 후 팝업 상단에
도서**(페이지 랜덤하게 표시됨)
표시된 페이지에 기재된 인증코드 등록

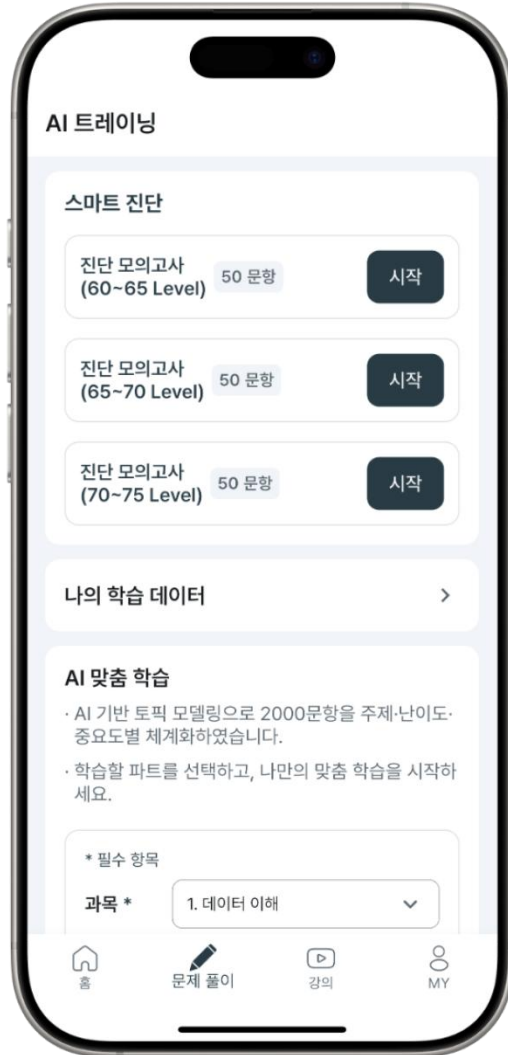
1. 요약본 PDF 파일 다운로드



2. 강의 학습

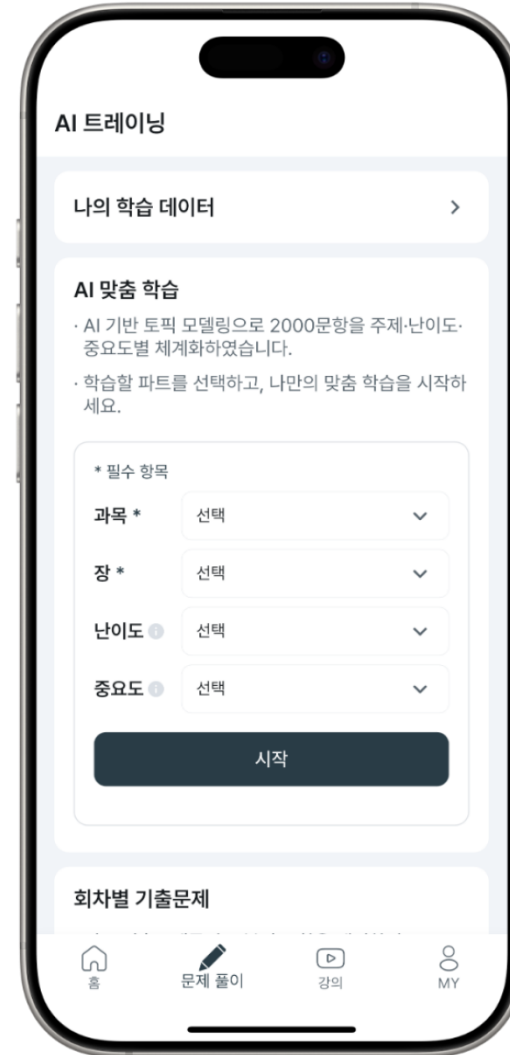


3. 스마트 진단



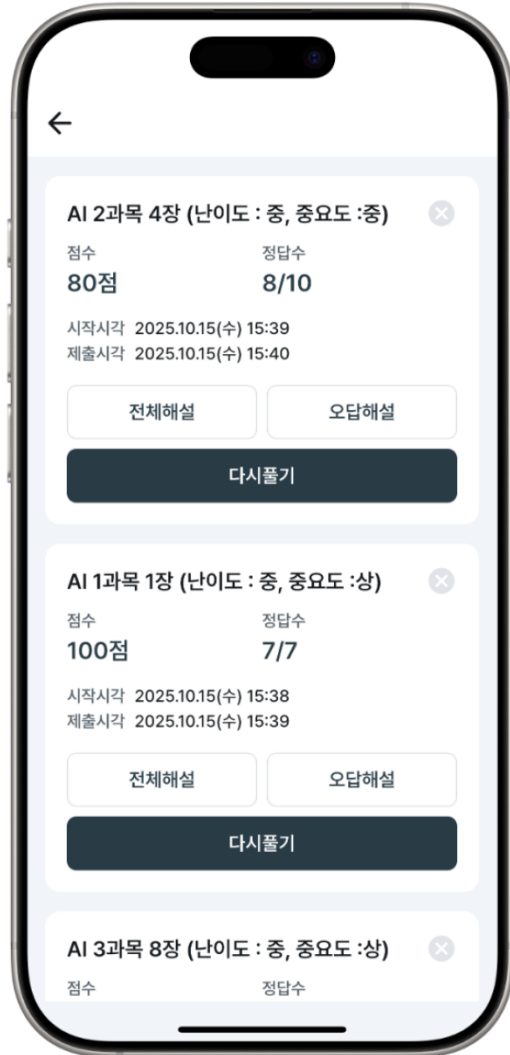
스마트 진단 학습 후
수준별 진단고사로 보완 파트 확인

4. 문제 풀이



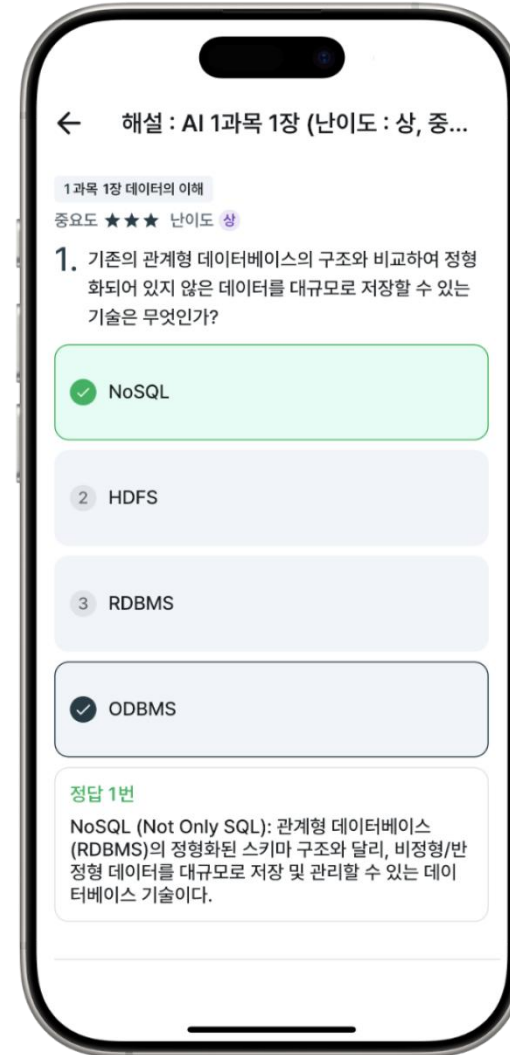
과목·파트·중요도·난이도별
보완 파트 집중 학습

5. 이력



학습 오답 노트 반복 학습

6. 문제 풀이



상세한 문제 풀이

1. 데이터 분석 파트의 출제 난이도 증가

2. 데이터 이해, 데이터 분석 기획은 전략적 과목으로

3. 합격을 위한 내용의 단권화가 필요

1과목 : 데이터 이해

1장. 데이터의 이해

01. 데이터와 정보

02. 데이터베이스 정의와 특징

03. 데이터베이스 활용

004

암묵지와 형식지의 상호 작용



AI

TIP: 출제유형

암묵지와 형식지의 상호작용 구분

- 암묵지와 형식지의 상호작용은 지식을 창출하고 공유하며, 이를 조직의 학습 및 혁신으로 연결하는 핵심 과정이다. 이 과정은 다음 네 가지 단계로 구분된다.
 - ⊖ 공통화(Socialization)
 - 암묵지를 다른 사람과 공유하여 새로운 암묵지를 생성하는 과정
 - ⊖ 표출화(Externalization)
 - 암묵지를 형식지로 전환하는 과정. 개인의 경험, 직감을 문서화하거나 시각적 자료로 표현
 - ⊗ 연결화(Combination)
 - 기존의 형식지를 조합하거나 새로운 지식을 추가하여 더 발전된 형식지로 확장하는 과정
 - ④ 내면화(Internalization)
 - 형식지를 학습하여 개인의 암묵지로 습득하는 과정
 - 문서, 매뉴얼, 교육 자료 등을 기반으로 학습하여 실무에 적용하고 경험을 통해 내재화

006 데이터베이스 정의



AI

TIP: 출제유형

DB와 DBMS 개념 구분

⊖ 데이터베이스의 정의

- 데이터베이스는 동시에 복수의 적용 업무를 지원하며, 다수 이용자의 요구를 수용하고 저장 및 공급하기 위해 일정한 구조에 따라 **조직된 데이터의 집합**이다.
- **(독립성)** 응용프로그램 독립성은 데이터베이스가 **응용프로그램으로부터 독립되어** 설계되고 관리된다는 것을 의미한다.

⊖ 데이터베이스 관리시스템(DBMS) 정의

- **DBMS**는 데이터베이스를 효율적으로 관리하고 활용할 수 있는 **소프트웨어 도구**로, 데이터 검색, 저장, 보안, 장애 복구 등의 기능을 제공한다.
- **(중개 역할)** 데이터베이스와 사용자(또는 응용 프로그램) 사이에서 데이터를 요청하고 처리하는 역할을 수행한다.
- 데이터 검색 및 수정
- 사용자 권한 관리 및 보안
- **데이터 무결성 유지** 및 장애 복구

010 데이터베이스 특징



AI

TIP: 출제유형

데이터베이스 특성 구분

⊖ 정보의 축적 및 전달 측면

- 대량의 정보를 기계가 읽고 쓸 수 있는 **기계 가독성**을 가지며, 필요한 정보를 쉽게 찾을 수 있는 **검색 가능성**과 원거리에서도 온라인으로 활용할 수 있는 **원격 조작성**을 갖는다.

⊖ 정보 이용 측면

- 이용자의 요구에 따라 다양한 정보를 **신속하고 경제적으로** 획득할 수 있다.

⊗ 정보 관리 측면

- 방대한 양의 정보를 **체계적으로 축적**하고, 새로운 내용을 **추가·갱신**하기가 쉽다.
- 데이터베이스는 정보처리, 검색, 관리 소프트웨어 등과 함께 **네트워크 기술 발전**을 견인하는 역할을 한다.

④ 경제.산업적 측면

- 데이터베이스는 **사회.산업 인프라**로서 기능하며, 경제.산업.사회 활동의 **효율성**을 **제고**하고 국민의 **편익**을 증진한다.

012 데이터 웨어하우스(Data Warehouse)와 데이터 마트(Data Mart)★★★



AI

TIP: 출제유형

데이터 웨어하우스와 데이터 마트의 특징

- 데이터 웨어하우스와 데이터 마트는 데이터 관리와 분석을 위한 저장소라는 공통점이 있지만, 규모와 목적에서 차이가 있다.

⊖ 데이터 웨어하우스(Data Warehouse)

- 기업의 모든 데이터를 중앙 집중화·통합 저장하고, 의사결정을 지원하기 위한 대규모 데이터 저장소
- 통합된 데이터: 여러 출처에서 수집한 데이터를 하나의 시스템에 통합
- 주제 중심적: 조직 전체에서 사용할 수 있도록 데이터가 설계되어 있다.
- 시간 가변적: 과거 데이터와 현재 데이터를 모두 포함하여 분석 가능하다.
- 비휘발성: 데이터를 저장한 후에는 변경하지 않으며, 분석을 목적으로만 사용한다.

⊖ 데이터 마트(Data Mart)

- 데이터 웨어하우스의 데이터를 **특정 부서나 비즈니스 영역에 최적화하여** 제공하는 소규모 데이터 저장소
- 단일 주제:**특정 부서나 비즈니스 영역**(마케팅, 판매)에 초점
- 독립적 운영 가능:데이터 웨어하우스 없이 독립적으로 운영되기도 한다.
- 규모:데이터 웨어하우스보다 규모가 작고 구조가 간단하다.
- 사용자:**특정 부서나 팀의 실무자**

015 ERP와 BI와 BA★★

- ERP, BI, BA는 기업의 데이터 관리와 의사결정을 지원하는 핵심 기술이다.
- 각각의 목적과 기능은 다르지만, 상호 보완적으로 사용되어 조직의 효율성과 경쟁력을 높이는 데 기여한다.

⊖ ERP(Enterprise Resource Planning)

- 기업의 핵심 비즈니스 프로세스를 통합적으로 관리하는 시스템이다.
- 제조, 구매, 재고, 주문, 공급망, 고객 서비스 등 다양한 프로세스를 지원한다.

⊖ BI(Business Intelligence)

- 데이터를 기반으로 의사결정을 지원하기 위해 설계된 리포트 중심 도구이다.
- 과거 데이터를 분석하여 현재의 상태를 시각적으로 이해하고 판단을 돕는 데 초점을 둔다.

⊗ BA(Business Analytics)

- BI보다 고급 분석 기술을 적용하여 미래를 예측하거나, 특정 전략의 결과를 예측하는 도구이다.
- 통계적 및 수학적 모델링과 머신러닝을 활용한 데이터 분석을 수행한다.

016 기타 기업 내부 데이터베이스 솔루션

⊖ EAI(Enterprise Application Integration)

- 기업 애플리케이션 통합 솔루션으로, 기업 내 다양한 시스템(ERP, CRM, SCM 등)을 상호 연동하여 통합적으로 운영할 수 있도록 지원한다.

⊖ 블록체인(Blockchain)

- 데이터 분산 처리 기술로, 네트워크에 참여하는 모든 사용자가 거래 데이터를 분산 저장하여 데이터의 위·변조를 방지한다.
- 블록과 체인 구조로 이루어져, 데이터의 투명성과 보안을 보장한다.
- 중앙 서버가 아닌 네트워크 참여자들이 데이터를 공유하고 저장한다.
- 중앙 서버가 없어 특정 서버 해킹으로 인한 데이터 손실이 방지된다.

017 사회기반 데이터 베이스 솔루션



AI TIP: 출제유형 사회 기반 데이터베이스 솔루션 유형

㊤ 종합물류정보망

- 운행 차량의 위치 및 상태를 실시간으로 추적하여 운송회사와 서비스 가입자의 합리적 의사결정을 지원한다.

㊤ 부가가치통신망(VAN, Value Added Network)

- 통신회선에 부가가치를 첨가한 정보 전송 네트워크로, 정보를 축적, 가공, 변환하여 제공한다.

⊛ 국가지리정보체계(NGIS, National Geographic Information System)

- 국가 및 지역별 지리 데이터를 관리하고 활용하는 정보 체계이다.

④ 지능형교통시스템(ITS, Intelligent Transportation System)

- 교통 상황 데이터를 실시간으로 수집하고, 이를 활용해 교통 흐름을 최적화하는 시스템이다.

⑤ 의료 EDI(Electronic Data Interchange)

- 병원, 약국, 보험사 간 전자적으로 의료 정보를 교환하는 시스템이다.

⑥ 교육행정정보시스템(NEIS, National Education Information System)

- 전국 초.중.고등학교의 교육 행정 데이터를 통합 관리하는 시스템이다.

1과목 : 데이터 이해

2장. 데이터의 가치와 미래

01.빅데이터의 이해

02.빅데이터의 가치와 영향

03.비즈니스 모델

04.위기 요인과 통제방안

05.미래의 빅데이터

022

빅데이터가 만들어 내는 본질적 변화



AI

TIP: 출제유형

빅데이터 본질적인 변화

- 빅데이터의 등장과 활용은 데이터 처리 방식, 분석 방법, 그리고 의사결정 패러다임에 근본적인 변화를 가져왔다.
- ⊖ 정보의 사전처리에서 사후처리 시대로
 - 과거: 데이터를 분석하기 전에 필요한 정보를 선별 하고 처리하는 방식이 일반적이었다.
 - 현재: 데이터를 모두 수집한 후 분석하며, 저장 비용이 감소하면서 데이터의 선별 없이 대규모 데이터 분석이 가능
- ⊖ 표본조사에서 전수조사로
 - 과거: 표본조사를 통해 제한된 데이터만 분석하여 전체 경향성을 추정
 - 현재: 빅데이터 기술로 인해 전수조사가 가능해져 정확성과 신뢰도가 향상되었다.

⊛ 질보다 양으로

- 과거:데이터의 품질에 초점을 맞추어 제한된 데이터로 분석하였다.
- 현재:데이터의 양이 많아질수록 더 정밀한 결과를 도출할 수 있다는 패러다임 변화가 나타났다.

④ 인과관계에서 상관관계로

- 과거:데이터를 분석하여 인과관계를 도출하려는 시도가 중심이었다.
- 현재:빅데이터 분석에서는 상관관계를 찾아내는데 초점을 맞춘다.

023 빅데이터의 가치 산정이 어려운 이유



AI

TIP: 출제유형

빅데이터 가치 산정이 어려운 이유

- 빅데이터는 방대한 양, 다양한 형식, 높은 처리 속도 등의 특성으로 인해 가치 산정이 쉽지 않다.
- ⊖ 데이터 활용 방식
 - 재사용 및 다목적 활용: 데이터는 한 번 사용되고 끝나는 것이 아니라, 재사용되고 다양한 목적으로 활용된다.
- ⊖ 데이터의 기존에 없던 가치 창출
 - 데이터는 기존에 정의되지 않은 방식으로 새로운 가치를 창출한다.
- ⊗ 분석 기술 발달의 영향
 - 새로운 기술로 데이터의 가치가 지속적으로 창출된다.

025

빅데이터 활용 기법^{★★★}



AI

TIP: 출제유형

빅데이터 테크닉 기법의 정의 및 사례

- 빅데이터 분석에서는 다양한 기법이 활용되며, 각각의 기법은 특정한 문제를 해결하거나 패턴을 발견하는데 사용된다.
- 아래는 대표적인 빅데이터 활용 기법과 그 적용 사례이다.
 - ⊖ **연관규칙학습(Association Rule Learning)**
 - 데이터에서 **변수 간 상관관계**를 찾아내는 방법으로, 어떤 항목이 함께 발생할 가능성이 높은지를 분석한다.
예 마트에서 상관관계가 높은 상품을 함께 진열, 온라인 쇼핑몰에서 고객의 구매 패턴 분석한다.
 - ⊖ **유형 분석(Classification Tree Analysis)**
 - 데이터를 특정 기준에 따라 **그룹화하거나 분류하는데** 사용된다.
 - “사용자가 어떤 특성을 가진 집단에 속하는가?”와 같은 질문에 답을 제공한다.
예 온라인 수강생들의 특성에 따라 학습 그룹 분류한다.
 - ⊗ **유전 알고리즘(Genetic Algorithms)**
 - 자연선택과 유전학 원리를 활용하여 **최적화 문제**를 해결하는 기법이다.
예
 - 최대의 시청률을 얻기 위한 방송 시간대 최적화
 - 연료 효율적인 자동차 개발 설계
 - 응급실에서 의사 배치 최적화

④ 기계 학습(Machine Learning)

- 데이터에서 **학습하고 패턴을 발견하여** 미래를 예측하거나 의사결정을 지원하는 알고리즘
- 훈련 데이터를 기반으로 모델을 생성한다.
- 넷플릭스의 영화 추천 시스템, 전자상거래에서 고객 맞춤형 상품 추천한다.

⑤ 회귀분석(Regression Analysis)

- 독립변수와 종속변수 간의 관계를 파악하여 변수 간 상관성을 분석한다.

예 구매자의 나이가 구매 차량 타입에 미치는 영향 분석

⑥ 감성분석(Sentiment Analysis)

- 텍스트 데이터에서 긍정적, 부정적 감정을 파악하여 의견과 평가를 분석한다.

예 새로운 환불 정책에 대한 고객의 평가를 분석한다.

⑦ 소셜 네트워크 분석(Social Network Analysis)

- 소셜 네트워크 데이터를 분석하여 **연결성과 영향력** 있는 인물을 식별한다.

예 소셜커머스에서 고객 간의 관계와 영향력을 분석한다.

026

빅데이터 시대의 위기 요인과 통제 방안



AI

TIP: 출제유형

빅데이터 위기 요인과 통제 방안의 연계

- 빅데이터 시대는 정보와 기술의 발전으로 많은 기회를 제공하지만, 동시에 다양한 윤리적, 사회적 위기를 초래할 수 있다.

⊖ 사생활 침해(Privacy Violation)

• 위기 요인

- 정보 수집 센서(M2M)의 확산
- 데이터의 2차, 3차 활용 가능성 증가로 인한 통제 어려움
- 사생활 침해와 사회적 위협

• 통제 방안

- 동의에서 책임으로 전환: 개인이 매번 동의해야 하는 현재의 시스템은 비효율적이며, 데이터 사용자가 더 큰 책임을 지도록 하는 체계가 필요하다.

⊖ 책임 원칙의 훼손(Erosion of Accountability)

• 위기 요인

- 빅데이터 분석으로 개인의 행동, 성향, 위험 요인을 예측하는 알고리즘이 발전
- 예측 결과만으로 개인을 처벌하거나 사회적 제한을 가하는 방향으로 변질될 가능성
- 이는 개인의 기본권 침해와 민주주의 원칙 훼손으로 이어질 수 있다.

• 통제 방안

- **책임 원칙 강화**: 예측 결과만으로 처벌이나 불이익을 주는 것을 금지하고, 투명성과 정당성을 확보해야 한다.

⊗ 데이터의 오용(Misuse of Data)

• 위기 요인

- 빅데이터는 과거에 일어난 사건과 데이터를 기반으로 분석하고 예측한다.
- 그러나 과거 데이터만으로 미래를 예측하는 데는 한계가 있으며, 새로운 상황이나 변수를 반영하지 못 할 수 있다.
- 이는 개인의 기본권 침해와 민주주의 원칙 훼손으로 이어질 수 있다

• 통제 방안

- **알고리즘에 대한 접근권 허용** : 데이터 알고리즘의 설계와 작동 방식에 대한 투명성 확보
- **객관적 인증 방안 도입** : 알고리즘이 편향 없이 설계되었는지, 신뢰할 수 있는 결과를 도출하는지 검증하는 공인 인증 시스템 구축
- **알고리즘 전문가 역할(알고리즘미스트)** : 책임과 투명성 강화

③ 개인정보 비식별화 기술

비식별 기술	제거방법	예시
가명처리	식별요소를 다른 값으로 대체	홍길동, 35세, 서울 거주, 한국대 재학 → 임꺽정, 30대 서울 거주, 국제대 재학
총계처리 또는 평균값 대체	데이터를 총합으로 표시하여 개별 데이터 값을 보이지 않도록 함	임꺽정 180cm, 홍길동 170cm → 1-5반 학생 키 합 350cm, 평균 키 175cm
데이터 값 삭제	개인 식별을 인식할 수 있는 값을 삭제	홍길동, 35세, 서울 거주, 한국대 졸업 → 35세, 서울 거주
범주화	범주의 값으로 변환	홍길동, 35세 → 홍씨, 30-40세
데이터 마스킹	개인 식별자가 보이지 않도록 처리	홍길동, 35세 → 홍**, 35세

031

일차적인 분석 vs 전략 도출을 위한 가치 기반[★] 분석



AI

TIP: 출제유형

가치 기반 분석 특징

- 데이터 분석은 일차적인 분석에서 시작하지만, **기업의 성과와 지속 가능성을 높이기 위해서는 전략적**

⊖ **인사이드를 제공하는 가치 기반 분석으로** 나아가야 한다.
⊖ **일차적인 분석**

- 일차적인 분석은 주로 특정 부서나 팀의 단기적인 업무 개선을 목표로 하며, **비교적 좁은 범위와 내부 문제에 초점을 둔다.**

⊖ **전략 도출을 위한 가치 기반 분석**

- 가치 기반 분석은 조직의 경쟁력을 강화하기 위해 전략적 인사이트를 도출하는 데 중점을 둔 분석이다.
- 업계 내부 데이터 뿐만 아니라, **인구통계학적 변화, 경제사회 트렌드, 고객 니즈의 변화** 등 외부 요인까지 고려한다.
- **전략적 가치기반 분석은 기존 성과를 유지하는 데 초점이 아니라, 비즈니스 성과를 개선하고 혁신하는 것이 주된 목적이다.**

034

데이터 사이언티스트가 갖춰야 할 역량(가트너[★] 기준)



AI

TIP: 출제유형

데이터 사이언티스트의 역량

- 데이터 사이언티스트는 데이터 분석과 관련된 기술적 능력 뿐만 아니라, **비즈니스 이해와 소프트웨어 스킬을** 겸비해야 한다.
- 가트너(Gartner)가 제시한 데이터 사이언티스트의 주요 역량
 - ⊖ 데이터 관리
 - 데이터를 수집, 정리, 저장, 가공하는 능력으로, 데이터의 구조와 성격을 깊이 이해가 필요하다.
 - ⊖ 분석 모델링
 - 데이터를 분석하고 통계적.머신러닝 모델을 설계하고 구현하는 능력이다.
 - ⊗ 비즈니스 분석
 - 분석 결과를 활용해 비즈니스 문제를 해결하고, 조직의 전략적 목표 달성을 지원하는 능력이다.
 - ④ 소프트웨어 스킬
 - 데이터 분석 결과를 조직에 효과적으로 전달하고, 협업을 통해 프로젝트를 성공적으로 이끄는 능력이다.

035

데이터 사이언티스트 요구 역량 (Hard Skill & Soft Skill)^{***}



AI

TIP: 출제유형

하드 스킬과 소프트 스킬 개념 구분

- 데이터 사이언티스트는 하드 스킬(기술적 역량)과 소프트 스킬(비기술적 역량)을 모두 갖춰야 한다.
- 하드 스킬은 데이터 분석 기술의 전문성을, 소프트 스킬은 분석 결과를 활용해 조직 내 가치를 창출하는 능력을 나타낸다.

⊖ Hard Skill

- 빅데이터에 대한 이론적 지식
- 분석 기술에 대한 숙련
- 최적화 설계 및 노하우 축적

⊖ Soft Skill

- **통찰력 있는 분석:**창의적 사고,논리적 비판, 호기심
- **설득력 있는 전달:**분석 결과를 효과적으로 전달하고, 의사결정을 지원할 수 있는 스토리텔링과 시각화 능력
- **다분야 간 협력:**분석 과정에서 다양한 팀과 협력하고, 커뮤니케이션을 통해 조화를 이루는 능력

2과목 : 데이터 분석 기획

4장. 데이터 분석 기획의 이해

- 01. 분석 기획 방향성 도출
- 02. 분석 방법론
- 03. 분석 과제 발굴
- 04. 분석 프로젝트 관리 방안

042 분석 방법론의 구성 요소



AI

TIP: 출제유형

분석 방법론의 구성 요소

⊖ 상세한 절차(Procedure)

- 분석 수행을 위한 단계적 가이드라인으로, 작업의 체계성과 일관성을 유지

예 데이터 수집 → 전처리 → 이탈 예측 모델링 → 결과 해석

⊖ 방법(Methods)

- 분석 문제 해결을 위한 구체적인 접근 방식

예 통계적 분석, 기계학습

⊗ 도구와 기법(Tools & Techniques)

- 분석을 수행하기 위한 소프트웨어 및 알고리즘

예 Python, R, Tableau, Power BI

④ 템플릿과 산출물(Templates & Outputs)

- 분석 과정의 표준화를 돕는 템플릿과 결과물을 문서화한 자료

예 분석계획서

043

폭포수(Waterfall), 나선형(Spiral), 프로토타입(Prototype) 모델 비교^{★★}



AI

TIP: 출제유형

분석 모형 프로세스의 특징 구분

- 소프트웨어 개발 및 프로젝트 관리에서 사용되는 폭포수 모델, 나선형 모델, 프로토타입 모델은 각기 다른 특징과 적합한 상황을 가진다.

⊖ 폭포수 모델(Waterfall Model)

- 프로젝트를 단계별로 진행하며, **각 단계가 완료되고 승인된 후에** 다음 단계로 넘어가는 방식
- 장점
 - 각 단계가 명확히 정의되어 있어 체계적이고 이해하기 쉬움
- 단점
 - 단계별로만 진행되므로 변경사항 적용이 어렵고 비용이 증가
 - 반복과 피드백이 제한적임

⊖ 나선형 모델(Spiral Model)

- 여러 번의 개발 반복을 통해 점진적으로 프로젝트를 완성하는 방식
- 장점
 - 위험 관리가 용이하며, 반복을 통해 점진적으로 품질을 향상시킬 수 있음
 - 초기에 요구사항이 명확하지 않아도 점진적으로 구체화 가능
 - 대규모이거나 복잡한 프로젝트에서 유용
- 단점
 - 관리가 복잡하며, 반복을 제대로 관리하지 못하면 비용과 일정 지연 발생 가능

⊗ 프로토타입 모델(Prototype Model)

- 초기 단계에서 사용 가능한 프로토타입(시제품)을 개발하여 사용자로부터 피드백을 받고, 이를 바탕으로 반복적으로 개선하는 방식
- 장점
 - 사용자의 요구사항을 보다 명확히 파악할 수 있음
 - 초기 오류를 신속히 발견하고 개선 가능
 - 반복적으로 개선해 나가는 개발 방식으로 전체 개발 시간을 단축하는 데 유리
- 단점
 - 지속적인 피드백 과정으로 개발비용 및 시간 증가 우려

044

KDD 분석 절차 (Knowledge Discovery in Databases) ★★★

AI

TIP: 출제유형

KDD 프로세스 순서 및 데이터 전처리의 주요 내용

- KDD는 데이터베이스에서 유용한 지식을 발견하기 위한 체계적인 분석 절차
- 데이터 준비부터 분석 결과의 해석과 활용까지 포괄적인 단계를 포함하며, 데이터 마이닝 기법의
- ⊖ 효과적인 적용을 지원 **비즈니스 이해와 목표 설정**
 - 무엇을 분석하고, 어떤 결과를 원하는지 정의
- ⊖ **데이터 선택**
 - 목적에 맞는 데이터 셋을 선정
- ⊗ **데이터 전처리(Preprocessing)**
 - 분석의 정확성과 신뢰성을 높이기 위해 데이터를 정리하고 준비하는 과정
 - 잡음(noise) 제거: 데이터에 포함된 불필요한 정보 삭제
 - 결측값 처리: 삭제, 평균 대체, 예측값 대체 등 적절한 방식 적용
 - 이상값(Outlier) 처리: 데이터 입력 오류, 측정 오류 등을 대체

④ 데이터 변환(Transformation)

- 분석 목적에 맞게 데이터 셋을 조정하거나 변경하는 단계
- 변수 선택(Feature Selection): 분석 목적과 관련성이 높은 변수를 선택
- 차원 축소(Dimensionality Reduction): 데이터의 차원을 줄여 분석 속도를 높이고
과대적합(Overfitting)을 방지
- 데이터 스케일링(Scaling): 데이터 값의 범위를 조정하여 모델 학습 과정에서 특정 변수의
영향을 과도하게 반영되지 않도록 함

⑤ 데이터 마이닝(Data Mining)

- 변환된 데이터 셋을 활용하여 패턴을 발견, 데이터를 분류, 예측 작업을 수행하는 과정
- 데이터 마이닝은 다양한 기법과 알고리즘을 사용하여 비즈니스 목표에 맞는 분석 결과를 도출하는 데 중점

⑥ 데이터 마이닝 결과 평가(Interpretation / Evaluation)

- 데이터 마이닝 결과 평가 단계는 분석 결과를 해석하고, 평가하며, 실질적으로 활용하는 과정으로, 데이터 마이닝의 성공 여부를 결정짓는 중요한 단계

045 CRISP-DM 분석 방법론

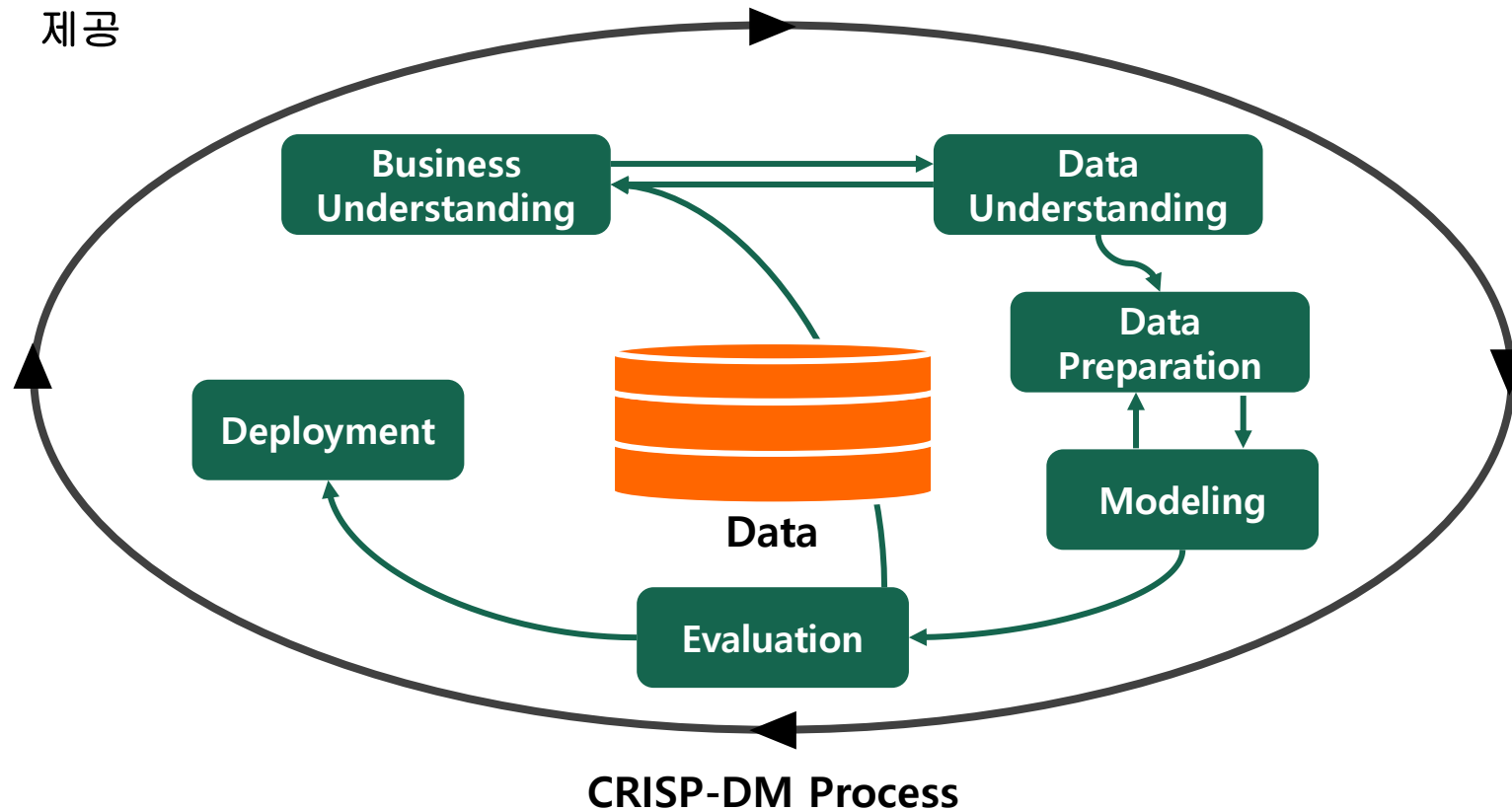


AI

TIP: 출제유형

CRISP-DM 피드백 구간

- CRISP-DM(Cross Industry Standard Process for Data Mining)은 다양한 산업 분야에서 데이터 마이닝 프로젝트를 체계적으로 수행하기 위해 설계된 표준 프로세스
- 6단계로 구성되어 있으며, 각 단계는 상호 피드백과 반복을 통해 완성도를 높이고 유연성을 제공



046

CRISP-DM 6단계 ★★★



AI

TIP: 출제유형

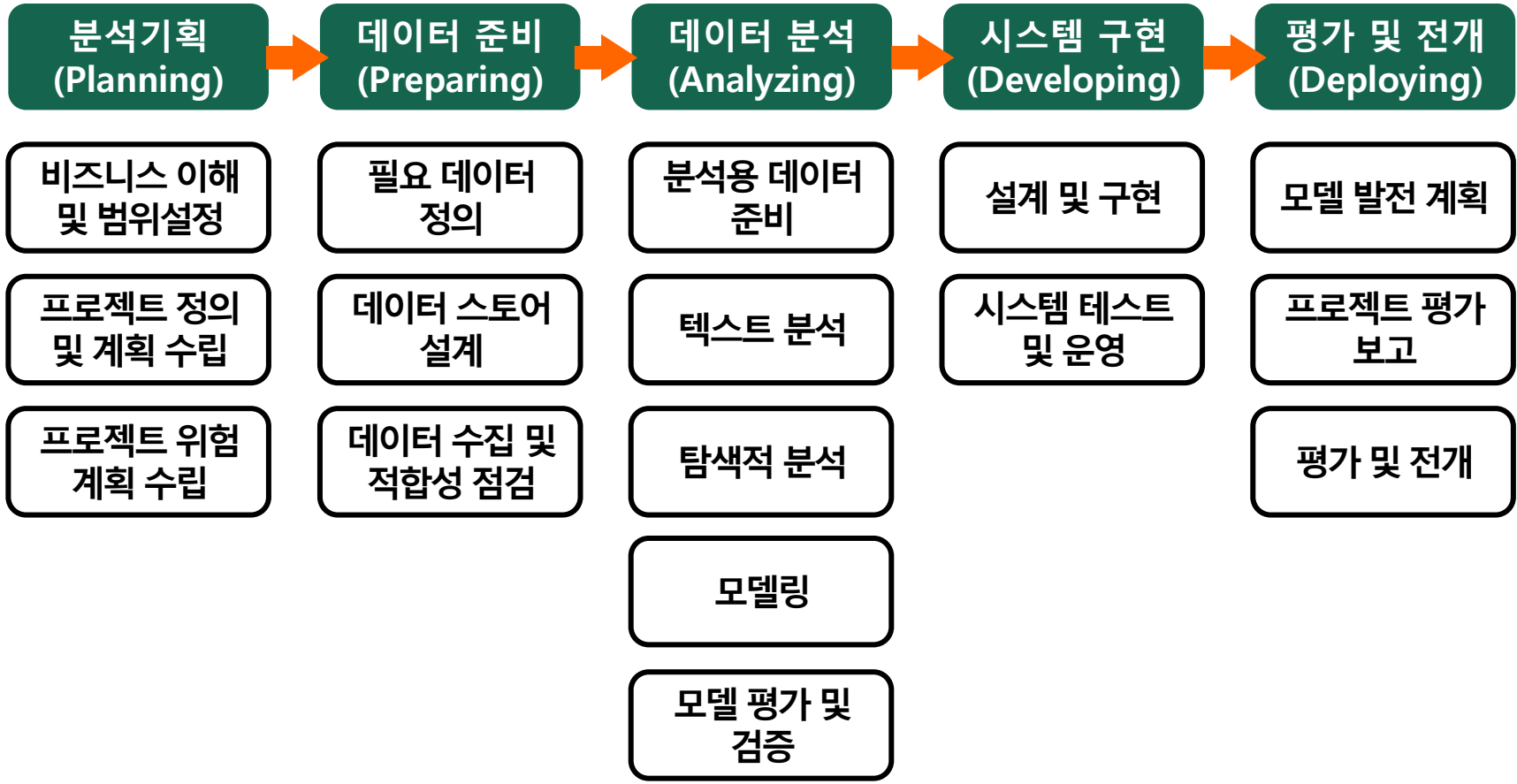
CRISP-DM 프로세스 순서 및 단계와 태스크의 구분

⊖ 업무 이해(Business Understanding), 단계(Phase) 계층

- 업무 이해 단계는 데이터 마이닝 프로젝트의 목적과 요구사항을 비즈니스 관점에서 이해하고, 이를 데이터 분석 문제로 정의하는 기초 단계
 - 업무 목적 파악-태스크(Task) 계층
 - 프로젝트의 비즈니스 목표를 명확히 정의
 - 상황 파악- 태스크(Task) 계층
 - 기존 프로세스 및 시스템 검토
 - 데이터 마이닝 목표 설정- 태스크(Task) 계층
 - 데이터 마이닝의 구체적 목표 정의(분류, 예측, 군집화 등)

048 빅데이터 분석 방법론

AI TIP: 출제유형 빅데이터 분석 방법론의 프로세스



050

분석 과제 도출



AI

TIP: 출제유형

하향식 및 상향식 접근방식 특징

- 빅데이터 분석 과제를 정의하고 수행하기 위해서는 문제를 **데이터 분석 문제로 변환**하는 과정이 필요
- 분석 과제 도출은 하향식(Top Down)과 상향식(Bottom Up) 접근 방법을 통해 이루어질 수 있으며, 이 두 가지 접근 방식은 상황에 따라 상호 보완적으로 활용된다.

⊖ 하향식 접근 방법(Top-Down Approach)

- 문제가 **명확히** 주어진 경우, 이를 체계적으로 분석하여 해결방안을 도출하는 방식.
- 각 과정이 체계적으로 단계적으로 정리되어 수행되며, 프로젝트 목표와 요구 사항이 **명확할 때 적합**

⊖ 상향식 접근 방법(Bottom-Up Approach)

- 문제정의 자체가 어려운 경우, 데이터를 기반으로 문제를 **재정의**하고 해결방안을 탐색하는 방식
- 시행착오를 통해 문제와 해결 방안을 점진적으로 개선하며, 새로운 인사이트 도출에 유용

053

하향식 접근 방식 (Top-Down Approach) ★★★



AI

TIP: 출제유형

하향식 접근방식의 프로세스

- 하향식 접근 방식은 명확히 주어진 문제나 전략적 목표를 기반으로 분석을 체계적으로 설계하고 실행하는 접근법이다.
 - 이 방식은 **현황 분석**에서 시작해, 문제를 데이터 문제로 변환하여 구체적이고 실행 가능한 해결방안을 도출하는 데 초점을 맞춘다.
 - 하향식 접근법(Top-Down Approach)은 문제가 명확히 정의된 상황에서 문제를 해결하는 데 적합하지만, 새로운 문제를 탐색하거나 발견하는 데는 몇 가지 한계를 가진다.
- ⊖ 문제탐색(Problem Discovery)
- **현황 분석**을 통해 확인된 문제나 기회를 심층적으로 탐색
 - 문제의 원인과 연관성을 파악하여 데이터를 기반으로 해결 가능성을 검토
 - 세부적인 구현 및 솔루션에 중점을 두는 것이 아니라, 문제를 해결했을 때 발생하는 가치에 집중
- ⊖ 문제정의(Problem Definition)
- 문제 정의 단계는 비즈니스 문제를 데이터 분석 문제로 변환하여, 필요한 데이터와 분석 기법을 명확히 정의하는 단계

062

분석 프로젝트 관리 방안



AI

TIP: 출제유형

분석 프로젝트의 5가지 속성

- 분석 프로젝트는 도출된 분석 기회를 통해 구체적인 가치를 증명하고 목표를 달성하기 위한 체계적인 관리가 필요하다.
- 일반적인 프로젝트 관리 요소(범위, 일정, 품질, 리스크 등) 외에도, **분석 프로젝트는 데이터와 분석 모델의 특수성을 반영한 5가지 주요 속성을 중심으로** 관리가 이루어져야 한다.

⊖ Data Size(데이터 크기)

- 데이터 양이 방대할수록 처리 속도와 자원 관리가 중요

⊖ Data Complexity(데이터 복잡성)

- 정형 데이터 뿐만 아니라 비정형·반정형 데이터(텍스트, 이미지 등)를 포함할 경우, 데이터 확보와
모델 선정이 더욱 까다로움

⊗ Speed(속도)

- 분석 결과를 도출하는 시간적 요구사항을 의미한다.

④ Analytic Complexity(분석 복잡성)

- 분석 모델의 복잡도가 증가하면 정확도(Accuracy)는 높아지지만, 모델 해석과 적용이 어려워진다.
- 복잡성과 정확도 간의 균형을 고려한 기준점 설정해야 한다.

⑤ Accuracy & Precision(정확도와 정밀도)

- Accuracy와 Precision 속성은 트레이드 오프(trade-off) 관계를 가지며, 분석 모델의 해석 및 적용 단계에서 사전 고려가 필요하다.
- 정확도(Accuracy) : 모델의 예측 값이 실제 값과 얼마나 가까운지 나타냄
- 정밀도(Precision) : 모델을 반복 적용했을 때 결과가 얼마나 일관되게 도출되는지를 나타냄
- Accuracy와 Precision을 고려한 모델 설계에서 활용이 중요하다면 Accuracy를, 안정성이 중요하다면 Precision을 우선 고려해야 한다.

2과목 : 데이터 분석 기획

5장. 분석 마스터플랜

01. 마스터 플랜 수립

02. 분석 거버넌스 체계 수립

065

마스터 플랜 수립 프레임 워크



AI

TIP: 출제유형

분석과제 우선순위 평가 기준

우선순위 고려 요소

전략적 중요도

비즈니스성과/ROI

실행 용이성

적용
우선순위
결정

적용 범위/방식 고려 요소

업무 내재화 적용수준

분석 데이터 적용수준

기술 적용 수준

분석 구현
로드맵 수립

072

분석 준비도(Readiness) ★★★



AI

TIP: 출제유형

분석 준비도 6개 영역과 세부 진단 항목

- 분석 준비도(Readiness)는 기업이 데이터 분석을 도입 하고 성공적으로 실행할 수 있는 환경과 역량을 갖추었는지 파악하기 위한 진단 방법이다.
- 주요 진단 영역은 분석 업무, 인력 및 조직, 분석 기법, 분석 데이터, 분석 문화, IT 인프라 등

6가지로 구성
[분석준비도 6개 영역]

⊖ 분석 업무 파악

- 발생한 사실 여부 파악
- 예측 분석업무
- 시뮬레이션 분석업무
- 최적화 분석업무
- 분석업무 정기적 개선

⊖ 인력 및 조직

- 분석전문가 직무 존재
- 분석전문가 교육 훈련 프로그램
- 관리자들의 기본적 분석 능력
- 전사 분석업무 총괄 조직 존재
- 경영진 분석업무 이해 능력

⊛ 분석기법

- 업무별 적합한 분석기법 사용
- 분석업무 도입 방법론
- 분석기법 라이브러리
- 분석기법 효과성 평가
- 분석기법 정기적 개선

⑤ 분석문화

- 사실에 근거한 의사결정
- 관리자의 데이터 중시
- 회의 등에서 데이터 활용
- 경영진의 직관보다 데이터
- 데이터 공유 및 협업문화

④ 분석 데이터

- 분석업무를 위한 데이터 충분성
- 분석업무를 위한 데이터 신뢰성
- 분석업무를 위한 데이터 적시성
- 비구조적 데이터 관리
- 외부 데이터 활용 체계
- 기준데이터 관리

⑥ IT 인프라

- 운영시스템 데이터 통합
- EAI, ETL 등 데이터 유통체계
- 분석 전용 서버 및 스토리지
- 빅데이터 분석 환경
- 통계 분석 환경
- 비주얼 분석 환경

073

분석 성숙도^{★★★}



AI

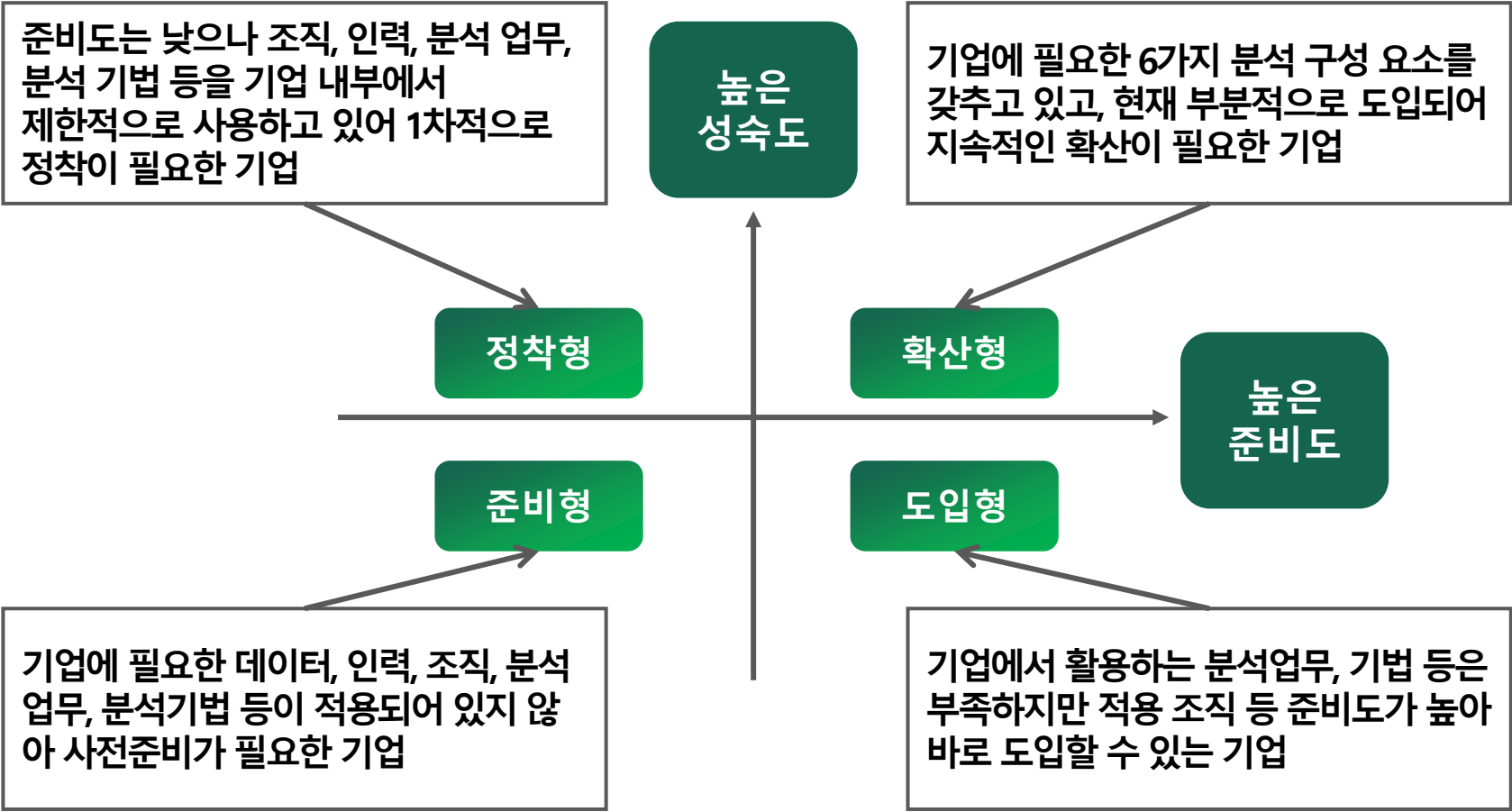
TIP: 출제유형

분석 성숙도의 3개 영역별 진단 항목

- 분석 성숙도 진단은 조직이 데이터 분석을 얼마나 효과적으로 도입하고 활용하고 있는지를 평가하기 위한 방법론이다.
- **CMMI(Capability Maturity Model Integration)** 모델 기반으로 개발되었으며, 조직의 데이터 분석 역량과 활용 수준을 평가하기 위해 **비즈니스 부문, 조직·역량 부문, IT 부문**을 기준으로 도입단계, 활용단계, 확산단계, 최적화 단계로 구분한다.

단계	도입	활용	확산	최적화
설명	분석을 시작하여 환경과 시스템을 구축	분석 결과를 실제 업무에 적용	전사 차원에서 분석을 관리하고 공유	분석을 진화 시켜서 혁신 및 성과 향상에 기여
비즈니스 부문	• 실적분석 및 통계 •정기보고 수행 •운영 데이터 기반	• 미래 결과 예측 •시뮬레이션 •운영 데이터기반	•전사 성과 실시간 분석 •프로세스혁신 3.0 •분석규칙 관리 •이벤트 관리	•외부환경 분석 활용 •최적화 업무 적용 •실시간 분석 •비즈니스 모델 진화
조직·역량 부문	• 일부 부서에서 수행 •담당자 역량에 의존	•전문 담당부서 수행 •분석 기법 도입 •관리자가 분석 수행	•전사 모든 부서 수행 •분석 CoE 조직 운영 •데이터 사이언티스트 확보	•데이터 사이언스 그룹 •경영진 분석 활용 •전략 연계
IT부문	•데이터웨어하우스 •데이터 마트 •ETL/EAI •OLAP	•실시간 대시보드 •통계분석 환경	•빅데이터 관리 환경 •시뮬레이션 최적화 •비주얼 분석 •분석 전용 서버	•분석 협업 환경 •분석 Sandbox •프로세스 내재화 •빅데이터 분석

분석 성숙도 모델



사분면 분석

077

데이터 거버넌스 체계 요소★★★



TIP: 출제유형

데이터 거버넌스 체계 요소

- 데이터 거버넌스 체계는 조직이 데이터를 효율적으로 관리하고 활용하기 위한 **표준화, 관리 체계, 저장소 관리, 표준화 활동**으로 구성된다.

⊖ 데이터 표준화

- **데이터 표준 용어 설명**:데이터의 명확한 정의와 용어 표준화
- **명명 규칙(Naming Rule)**:데이터 필드, 테이블, 파일 등에 대해 명명 규칙 수립
- **메타데이터 구축**:데이터 속성, 출처, 관계 등을 정의한 메타데이터 체계 수립
- **데이터 사전 구축**:데이터 구조와 속성을 체계적으로 정리한 데이터 사전 마련

⊖ 데이터 관리 체계

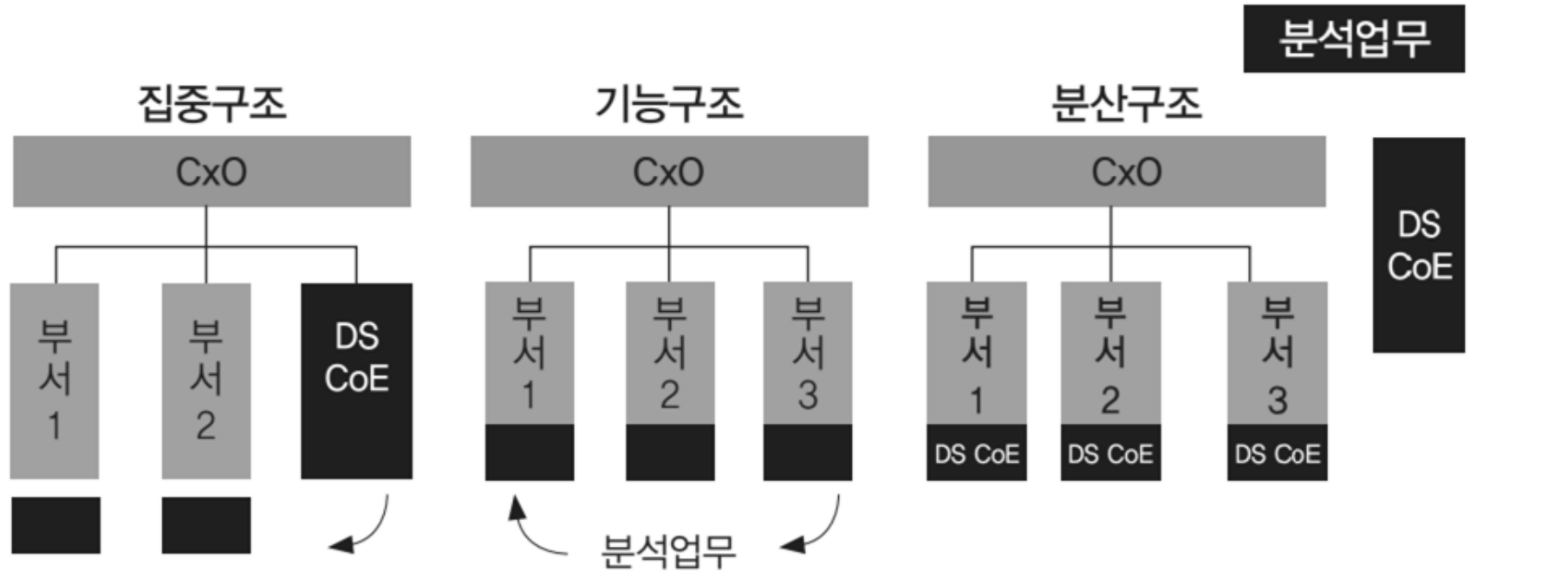
- **관리 원칙 수립**:메타데이터 관리 지침 및 데이터 사전 관리 기준 정의
- **데이터 품질 관리**:데이터의 정확성, 완전성, 일관성을 유지하기 위한 프로세스 수립
- **접근 권한 관리**:데이터 접근 권한 및 보안 정책 설정

⊛ 데이터 저장소 관리

- **저장소 구성**: 전사 차원의 데이터 저장소를 설계 및 운영
- **워크플로우 및 관리 소프트웨어 지원**: 데이터 저장, 변경, 접근 등에 대한 워크플로우와 관리 도구를 제공
- **시스템 간 인터페이스**: 데이터 관리 대상 시스템과의 연계 및 통제를 수행한다.
- **사전 영향 평가**: 데이터 구조 변경 시 발생할 수 있는 영향을 사전에 평가한다.

④ 표준화 활동

- **표준 준수 여부 점검**: 데이터 표준화 및 관리 체계가 잘 준수되고 있는지 정기적으로 평가한다.
- **모니터링**: 데이터 품질 및 보안 상태를 지속적으로 점검한다.
- **지속적 개선**: 점검 결과를 바탕으로 데이터 표준 및 관리 체계를 개선한다.



- 전사 분석 업무를 별도의 분석 전담 조직에서 담당
- 전략적 중요도에 따라 분석 조직이 우선순위를 정해서 진행 가능
- 현업 업무부서의 분석업무와 이중화이원화 가능성 높음

- 일반적인 분석 수행 구조
- 별도 분석조직이 없고 해당 업무 부서에서 분석수행
- 전사적 핵심 분석이 어려우며, 부서 현황 및 실적통계 등 과거 실적에 국한된 분석 수행 가능성 높음

- 분석조직 인력들을 현업부서로 직접 배치하여 분석 업무 수행
- 전사 차원의 우선순위 수행
- 분석 결과에 따른 신속한 Action 가능
- 베스트 프랙티스 공유가능
- 부서 분석 업무와 역할 분담 명확히 해야 함

분석업무 수행 주체에 따른 3가지 유형

3과목 : 데이터 분석

6장. R 기초와 데이터 마트

01. R 기초

02. 데이터 마트

03. 결측값 처리와 이상값 검색

084

기초 통계량^{★★★}



AI

TIP: 출제유형

summary() 함수의 해석 및 기초 통계량의 정의

- 기초 통계와 관련된 주요 개념과 함수들은 데이터를 요약하고 분포의 특성을 이해하는 데 중요한 역할을 한다.

⊖ summary() 함수

- summary() 함수는 수치형 데이터와 범주형 데이터의 특성을 요약하여 다르게 출력한다.
- 수치형 데이터(Numeric Data)
 - **Min** : 최솟값
 - **1st Qu** : 제1사분위수(하위 25%에 해당하는 값)
 - **Median** : 중앙값(50%에 해당하는 값)
 - **Mean** : 평균
 - **3rd Qu** : 제3사분위수(상위 25%에 해당하는 값)
 - **Max** : 최댓값
- 범주형 데이터(Categorical Data): 각 범주의 빈도수를 출력한다.

㊤ 변동계수(Coefficient of Variation, CV)

- 표준편차를 평균으로 나눈 값으로, 데이터의 상대적인 변동성을 나타낸다.
- 측정 단위가 다른 데이터를 비교할 때 유용하다.

⊗ IQR(Interquartile Range, 사분위수 범위)

- 데이터의 제3사분위수(75%)와 제1사분위수(25%)의 차이로, 중앙 50% 데이터가 퍼져 있는 범위를 나타낸다.

④ 범위(Range)

- 데이터의 최댓값과 최솟값의 차이를 의미한다.

⑤ 왜도(Skewness)

- 데이터 분포의 비대칭 정도를 나타낸다.
- 왜도 > 0 : 오른쪽으로 긴 꼬리(왼쪽으로 치우친 분포)
- 왜도 < 0 : 왼쪽으로 긴 꼬리(오른쪽으로 치우친 분포)
- 왜도 = 0 : 대칭적인 분포

⑥ 사분위수(Quantiles)

- 데이터를 특정 백분위수로 나누는 기준 값이다.

예 25%(1사분위수), 50%(중앙값), 75%(3사분위수)

085

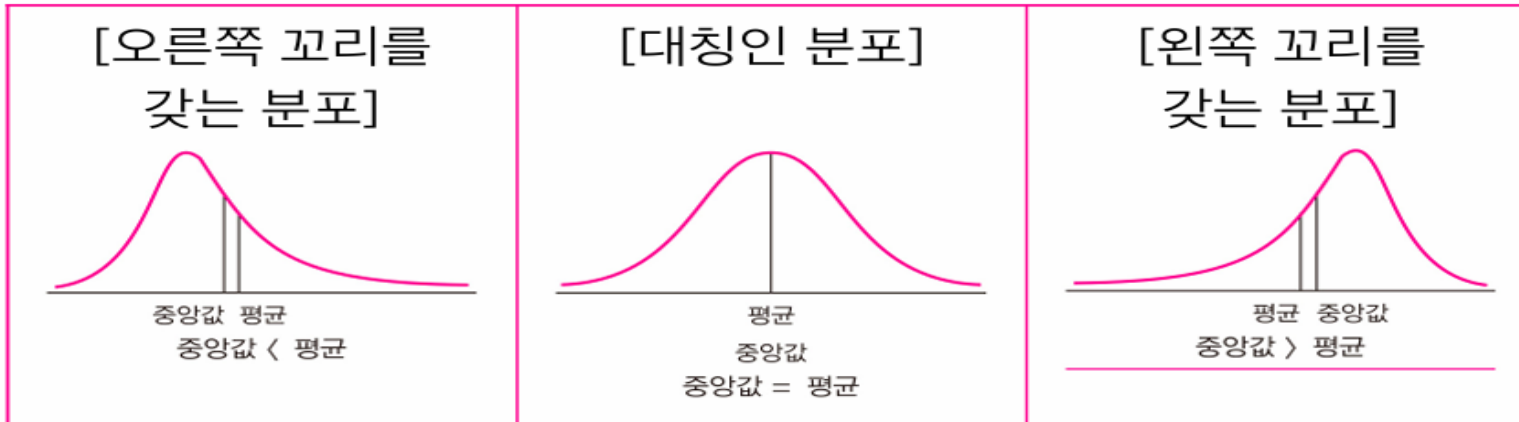
분포에 따른 대푯값의 관계

- 평균은 이상치에 민감하지만, 데이터의 전체적인 중심 경향을 잘 나타낸다.
- 중위수는 이상치에 덜 민감하며, 비대칭 분포에서 적합한 대푯값이다.
- 최빈치는 범주형 데이터나 빈도 분석에서 유용하다.
- 데이터의 특성을 고려하여 적합한 대푯값을 선택하는 것이 분석의 신뢰성을 높이는 핵심이다.

⊖ 좌우대칭 분포일 경우 예 정규분포): 평균 = 중앙값 = 최빈치

⊖ 좌측 비대칭 분포일 경우 (왼쪽 꼬리가 긴 분포): 평균 < 중앙값 < 최빈값

⊗ 우측 비대칭 분포일 경우 (오른쪽 꼬리가 긴 분포): 평균 > 중앙값 > 최빈값



088

상자 그림(Box plot)^{★★}



AI

TIP: 출제유형

상자 그림의 해석

- 상자 그림(Box plot)은 데이터의 분포를 시각적으로 요약하는 데 사용되는 그래프이다.
- 데이터의 중심, 산포, 그리고 이상치를 한눈에 파악할 수 있어 탐색적 데이터 분석(EDA)에서 중요한 도구이다.

⊖ 상자 그림의 구성 요소 사분위수(Quartiles)

- 제1사분위수(Q1) : 하위 25% 데이터의 경계
- 제2사분위수(Q2, Median) : 중앙값으로, 데이터의 중간값
- 제3사분위수(Q3) : 상위 25% 데이터의 경계

⊖ IQR(Interquartile Range, 사분위수 범위)

- $IQR = Q3 - Q1$
- 중앙 50% 데이터를 포함하는 범위를 나타낸다.

093

파생변수와 요약변수^{★★}



AI

TIP: 출제유형

파생변수와 요약변수 정의

⊖ 파생변수(Derived Variable)

- 기존의 변수를 조합하여 새롭게 생성된 변수이다.
- 주관적 판단, 특정 함수, 또는 규칙에 의해 만들어지며, 사용자의 목적에 따라 정의된다.
- 파생변수는 주관적이고 목적 중심으로 데이터를 변형하여 새로운 분석 관점을 제공한다.
- 예 BMI(체질량지수(BMI) 계산)= $\text{weight} / (\text{height}/100)^2$

⊖ 요약변수(Aggregated Variable)

- 원시 데이터를 요약·집계하여 생성된 변수이다.
- 데이터셋 전체나 특정 그룹을 대표하는 값을 나타낸다.
- 요약변수는 데이터의 크기를 줄이고, 분석 효율성을 높이기 위해 단순화된 정보를 제공한다.
- 두 변수는 상호 보완적으로 사용될 수 있으며, 적절한 활용은 데이터 분석의 효과를 극대화할 수 있다.
- 예 부서별 평균 연봉, 제품별 총 매출

095

결측치 처리★



AI

TIP: 출제유형

완전분석법의 정의

1. 삭제법(Deletion Methods): 결측값을 포함한 사례를 제거하는 방법

1) 완전분석법(Complete Case Analysis)

- 결측치가 포함된 행을 제거하고, 결측치가 없는 관측치만을 사용하여 분석한다.

2) 선택적 삭제법(Pairwise Deletion)

- 분석에 필요한 변수에서만 결측이 있으면 해당 값만 제외, 나머지는 활용

2. 대치법(Imputation Methods): 결측값을 적절한 값으로 대체하여 완전한 데이터셋을 만드는 방법

1) 단순 대치법(Simple Imputation): 결측값을 하나의 대체 값으로 채우는 방법으로, 단순하고 빠르게 적용

① 평균 대치(Mean Imputation): 평균값으로 결측치를 채움

⊖ 단순 확률 대치법(Single Stochastic Imputation): 결측치를 가진 변수의 분포를 기반으로 **랜덤하게** 값을 생성하여 대치

⊗ 회귀 대치법(Regression Imputation)

- 결측치가 있는 변수를 종속 변수로 설정하고, 결측치가 없는 변수들을 독립 변수로 사용하여 회귀 모델을 통해 결측값을 예측한다.

097 이상값 탐지★



AI

TIP: 출제유형

단변량 이상값 탐지 방법

⊖ 단변량 이상값 탐지 (Univariate Outlier Detection)

- 상자그림 (Boxplot): 사분위수 범위를 활용하여 이상치 판단
- Z-Score : 평균과 표준편차를 기준으로 극단적인 값 탐지
- ESD 알고리즘 (Extreme Studentized Deviate Test) : 정규분포를 가정한 반복적 이상값 탐지

⊖ 다변량 이상값 탐지 (Multivariate Outlier Detection)

- 마할라노비스 거리 (Mahalanobis Distance): 변수 간 공분산 구조를 반영한 거리 기반 이상값 탐지
- 주성분 분석 (Principal Component Analysis): 저차원 공간에서 이상 데이터 식별
- 산점도 (Scatter Plot): 2차 또는 3차원 상에서 변수 간 관계를 시각화하여 이상값을 시각적으로 탐지

3과목 : 데이터 분석

7장. 통계분석

- 01. 통계학 개론
- 02. 기초 통계분석
- 03. 다변량 분석
- 04. 시계열 예측

098

표본추출 방법 (Sampling Methods) ★★★

AI

TIP: 출제유형

확률적 표본추출 방법의 정의

- 표본추출(Sampling) : 모집단으로부터 일부 개체(표본)를 선택하여 분석 및 연구에 사용하는 과정
- 확률적 추출(Probability Sampling) : 모집단의 각 개체가 표본으로 **선택될 확률이 알려진 경우**

⊖ 단순 무작위 추출(Simple Random Sampling)

- 모집단의 각 개체가 **동일한 확률로** 선택됨
- 표본크기 n , 모집단 크기 N 인 경우, 각 개체의 선택 확률은 n/N
- 모집단의 전체 명부 필요

⊖ 계통 추출(Systematic Sampling)

- 모집단에 일련번호를 부여하고, 첫 표본을 **무작위로 선택한 뒤** 일정 간격 **k** 로 표본추출
- 모집단에 주기적 패턴이 있을 경, 추출 결과가 편향(Bias) 될 수 있음

⊛ 층화 추출(Stratified Sampling)

- 모집단을 층(strata)으로 나누고, 각 층에서 무작위로 표본을 추출
- 층 내는 동질적, 층 간은 이질적이어야 함
- 층에 대한 정확한 정보가 필요

④ 군집 추출(Cluster Sampling)

- 모집단을 여러 그룹(군집, Cluster)으로 나누고, 이 중 몇 개의 그룹을 무작위로 선택한 뒤, 선택된 그룹에서 전체 구성원 또는 일부를 추출하는 방법이다.
- 모집단 전체를 조사하는 대신 그룹 단위로 조사하여 시간과 비용을 줄이는 것이 목표이다.
- 군집 내는 이질적, 군집 간은 동질적이어야 일 때 적합

099 척도(Scale) ★★★



AI

TIP: 출제유형

척도의 종류 및 특징

- **척도**는 측정 대상에 부여된 숫자들 간의 관계를 나타내며, 자료의 속성을 설명함
- 척도의 유형에 따라 적용 가능한 자료 분석 방법이 달라짐
- 척도의 종류는 **명목척도**, **서열척도**, **등간척도**, **비율척도**로 구분됨

⊖ 명목 척도(Nominal Scale)

- 분류 또는 식별의 목적으로 숫자를 부여
- 숫자는 대상을 식별하기 위한 용도로 사용되며, 서열 관계나 양적 비교가 불가능

예 혈액형(A형 = 1, B형 = 2, O형 = 3, AB형 = 4)

⊖ 서열 척도(Ordinal Scale)

- 대상 간의 순위(대소, 높고 낮음)를 나타냄
- 순위는 의미 있지만, 간격의 크기는 알 수 없음
- 양적 비교는 불가능하나 순서 비교는 가능

예 만족도 조사(매우 불만족 = 1, 불만족 = 2, 보통 = 3, 만족 = 4, 매우 만족 = 5)

⊛ 등간 척도(Interval Scale)

- 대상의 순위 정보 제공, 간격의 크기가 동일함
- 0의 위치가 임의적이며, 절대적 0이 존재하지 않음
- 덧셈, 뺄셈 등의 산술적 연산은 가능하나, 비율 계산은 불가능

예 온도(섭씨, 화씨)

④ 비율 척도(Ratio Scale)

- 순위 정보 제공, 간격이 동일, 절대적 0이 존재
- 덧셈, 뺄셈, 곱셈, 나눗셈 등 모든 산술적 연산 가능
- 비율 계산이 가능하며, 가장 정교한 수학적으로 완전한 척도

예 키, 몸무게, 나이, 소득

104

두 사건의 독립성과 종속성일 때, 교집합(결합확률)^{★★}의 확률



AI

TIP: 출제유형

독립 사건의 정의 및 교집합 확률

⊖ 독립 사건(Independent Events)

- 한 사건의 발생 여부가 다른 사건의 발생에 영향을 미치지 않는 경우
- 독립 사건의 교집합 확률

$$P(A \cap B) = P(A) \cdot P(B)$$

⊖ 종속 사건(Dependent Events)

- 한 사건의 발생 여부가 다른 사건의 발생에 영향을 미치는 경우
- 종속 사건의 교집합 확률

$$P(A \cap B) = P(A) \cdot P(B|A)$$

107

이산 확률분포의 주요 유형과 특징

⊖ 이항분포(Binomial Distribution)



AI

TIP: 출제유형

이항분포와 포아송 분포의 조건

- 베르누이 시행을 기반으로 한 확률분포
- 베르누이 시행: 성공(1) 또는 실패(0)로 이루어진 단일 확률 실험
- n 번의 독립적인 베르누이 시행에서 **성공의 횟수**를 확률변수로 하는 분포
- 베르누이 분포는 이항분포의 특수한 경우로, **$n=1$ 일 때 이항분포가 베르누이 분포와 동일**
- **각 시행은 독립적**
- **각 시행은 두 가지 결과 중 하나로 나타남**
- 각 시행에서 성공할 확률 p 은 모든 시행에서 동일

⊖ 포아송 분포(Poisson Distribution)

- 단위 시간 또는 단위 공간에서 사건이 발생하는 횟수를 모델링하는 분포
- 사건은 독립적으로 발생
- 단위 시간 또는 공간 내 사건 발생 확률은 일정
- 동시에 두 사건이 발생할 확률은 매우 낮다.

108

연속확률분포의 주요 유형과 [★]특징



AI

TIP: 출제유형

표준화 공식 및 F분포와 카이제곱 분포 특징

⊖ 정규분포(Normal Distribution)

- 연속확률변수의 대표적인 분포로, 평균을 중심으로 대칭적인 종모양을 갖는 분포

⊖ 표준정규분포(Standard Normal Distribution)

- 평균이 0, 표준편차가 1인 정규분포
- 모든 정규분포는 표준화를 통해 표준정규분포로 변환 가능
- 표준화 공식

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

- Z : 표준화된 확률변수
- X : 분석 대상이 되는 확률변수
- μ : 확률변수 X 의 평균값
- σ : 확률변수 X 의 표준편차

109

카이제곱 통계량(적합도 검정)

AI

TIP: 출제유형

카이제곱 통계량(적합도 검정)의 의미

- 비모수 검정 중에서 카이제곱 검정(χ^2 test)은 크게 세 가지 유형으로 나뉘며, 각각의 통계량은 동일한 구조를 가지지만 검정 목적과 가정이 다르다.

1. 적합성 검정 (Goodness of Fit Test)

- 목적: 한 범주형 변수의 관측 분포가 어떤 이론적(기대) 분포에 잘 부합하는지를 검정한다.
- 귀무가설(H_0): 관측값은 기대되는 분포에 적합한다.
- 검정 통계량

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- O_i : 관측도수, E_i : 기대도수
- χ^2 값이 크다는 것은:
- 관측 도수와 기대 도수 간의 차이(편차)가 크다는 의미이다.
- 즉, 실제로 관측된 값이 우리가 가정한 분포에서 많이 벗어나 있다는 것을 의미
- 귀무가설(H_0 : 기대 분포에 적합)이 틀릴 가능성이 커짐

113

중심극한정리 (Central Limit Theorem, CLT)



AI

TIP: 출제유형

중심 극한 정리의 개념

- 중심극한정리는 표본 크기가 충분히 클 경우, 모집단의 분포 형태와 관계없이 표본평균의 분포가 정규분포에 가까워진다는 통계학의 기본 정리
- 조건
 - 모집단의 분포가 정규분포일 필요가 없음
 - 표본은 독립적으로 랜덤하게 추출되어야 함
 - 표본 크기 n 이 증가할수록 결과의 정확성과 정규분포 근사 정도가 더욱 향상 ($n \geq 30$)
- 중심극한정리는 모집단의 분포를 몰라도, 표본크기가 충분히 크면 정규분포를 가정한 통계 분석이 가능하게 해준다.

- 잘못된 해석 ⇒ "중심극한정리가 성립하려면 모집단이 정규분포여야 한다."
- 올바른 해석 ⇒ "중심극한정리는 모집단의 분포 형태에 관계없이 표본 크기가 충분히 크면 표본평균의 분포가 정규분포에 근사함"
- 잘못된 해석 ⇒ "중심극한정리는 표본평균 외에도 모든 통계량에 대해 적용된다."
- 올바른 해석 ⇒ "중심극한정리는 표본평균에 대해서만 적용되는 이론이다."

115

신뢰수준의 의미★



AI

TIP: 출제유형

신뢰수준의 해석

⊖ 신뢰수준($1-\alpha$)

- **정확한 의미** : 신뢰수준은 구간추정 방법의 신뢰성을 나타내는 빈도적 개념이다.
- **잘못된 해석** : “ 참모수를 포함할 확률 ”이라는 표현은 확률론적 해석으로, 신뢰수준의 의미와 다르다.

⊖ 올바른 해석

- 신뢰수준은 구간추정 과정을 무한히 반복했을 때 그 중 얼마나 많은 구간이 참모수를 포함할 것인지에 대한 비율을 의미한다.
- 신뢰수준 **95%**의 의미는 다음과 같다.
 - 동일한 조건에서 표본추출과 구간추정을 무한히 반복하면, 생성된 구간들 중 약 **95%**가 참모수를 포함하고, 나머지 약 **5%** 포함하지 않을 수 있다.

118

가설검정★



AI

TIP: 출제유형

귀무가설과 대립가설의 정의

- 가설검정은 표본 데이터를 기반으로 모집단에 대한 주장(가설)의 타당성을 검토하는 통계적 방법이다.
- 주어진 데이터로부터 귀무가설(H_0)을 기각하거나 채택할지 결정하는 과정이다.

⊖ 귀무가설(H_0)

- 현재의 상태나 기존의 주장을 나타내는 기본 가정한다.
- “차이 없음”, “효과 없음”, “변화 없음”을 주장한다.
- 검정의 목적은 귀무가설이 데이터를 통해 기각될 수 있는지를 판단하는 것이다.
- 귀무가설을 기각하지 못하면 기존의 주장을 유지

⊖ 대립가설(H_1)

- 귀무가설과 반대되는 주장으로, 연구자가 증명하고자 하는 주장
- “차이 있음”, “효과 있음”, “변화 있음”을 나타낸다.
- 귀무가설이 기각될 경우 대립가설을 채택

119 가설검정에서 유의수준(α)과 유의확률(p-값)의 의미



AI TIP: 출제유형 유의수준과 유의확률의 정의

⊖ 유의수준(α)

- 가설검정에서 귀무가설(H_0)이 참임에도 불구하고 이를 잘못 기각할 확률을 의미하며,
- 이는 제1종 오류(Type I Error)이다.
- 연구자가 사전에 정하는 기준값, 일반적으로 $\alpha = 0.05$ 또는 $\alpha = 0.01$ 사용된다.
- 유의수준은 검정 결과의 판단 기준으로, 유의확률(p-값)과 비교하여 귀무가설의 기각 여부를 판단한다.

⊖ 유의확률(p-값)

- 귀무가설(H_0)이 참이라고 가정했을 때, 현재 표본 데이터와 같거나 또는 더 극단적인 결과가 관찰될 확률이다.
- p-값이 작을수록, 귀무가설이 참이라는 가정 하에 관찰된 데이터가 발생할 가능성이 낮음을 의미한다.
- 유의수준(α)과 비교하여 가설검정의 결과를 해석한다.
 - $p \leq \alpha$: 귀무가설 기각 → 대립가설 지지 → 통계적으로 유의하다.
 - $p > \alpha$: 귀무가설 기각하지 않음 → 통계적으로 유의하지 않다.

122

검정의 오류^{★★★}



AI

TIP: 출제유형

제1종 오류와 제2종 오류의 정의

- 가설검정에서는 잘못된 결론을 내리는 두 가지 유형의 오류가 존재한다.

⊖ 제1종 오류(Type I Error)

- 귀무가설(H_0)이 참임에도 불구하고, 이를 잘못 기각하는 오류

⊖ 제2종 오류(Type II Error)

- 귀무가설(H_0)이 거짓임에도 불구하고, 이를 잘못 채택하는 오류

⊗ 검정력(Power of Test)

- 대립가설(H_1)이 참일 때 귀무가설(H_0)을 올바르게 기각할 확률
- 검정력 = $1 - \beta$

④ 제1종 오류(α)와 제2종 오류(β) 간의 균형

- α 를 낮추면 제1종 오류는 감소하지만, 제2종 오류(β)는 증가할 수 있음
- 반대로, α 를 높이면 제2종 오류는 감소하지만, 제1종 오류가 증가
- 일반적으로 α -오류(제1종 오류)가 β -오류(제2종 오류)보다 더 심각한 결과를 초래할 수 있기 때문에, 유의수준 α 를 통제함

124

독립된 두 집단의 평균 비교 검정★



AI

TIP: 출제유형

모수적 검정에 대응하는 비모수적 검정

1. 독립된 두 집단의 비교 검정이란?

- 두 개의 서로 독립적인 집단에서 관측된 데이터가 같은 분포 또는 동일한 중심 경향을 가지는지 여부를 통계적으로 검정하는 방법이다.

구분	모수적 검정	비모수적 검정
조건	• 정규분포	• 분포 가정 없음
대상	• 평균차이	• 중앙값
검정	• 독립 표본 t- 검정	• 윌콕슨의 순위합 검정 • 만 위트니 검정

125

짝지은(대응) 두 집단의 평균 비교 검정★



AI TIP: 출제유형

모수적 검정에 대응하는 비모수적 검정

1. 짝지은(대응) 집단의 비교 검정이란?

- 짝지은 집단의 비교 검정(Paired Sample Test)은 같은 대상에게 두 번 측정한 값 또는 유사한 특성을 지닌 쌍(pair)에서 수집한 두 데이터 간의 차이를 분석할 때 사용하는 통계 검정이다

구분	모수적 검정	비모수적 검정
조건	• 정규분포	• 분포 가정 없음
대상	• 평균차이	• 중앙값
검정	• 대응 표본 t- 검정	• 윌콕슨의 부호순위 검정 • 부호 검정

129

다중회귀분석(Multiple Regression)과 다항회귀분석(Polynomial Regression)[★]



AI

TIP: 출제유형

다중회귀분석과 다항회귀분석 차이

⊖ 다중회귀분석(Multiple Regression)

- 다중회귀분석은 다수의 독립변수가 하나의 종속변수에 미치는 영향을 분석하는 통계기법이다.
- 이 모델은 독립변수는 종속변수와 선형 관계를 가정한다.

⊖ 다항회귀분석(Polynomial Regression)

- 다항회귀분석은 하나의 독립변수와 종속변수 간의 비선형 관계를 분석하는 기법이다.
- 다항식 항을 새로운 독립변수로 변환하여 선형 회귀의 틀 안에서 비선형 관계를 추정한다.

134

단계적 변수 선택 (Stepwise Variable Selection)^{★★}



AI

TIP: 출제유형

전진 선택법과 후진 제거법의 정의

- 단계적 변수 선택은 회귀분석에서 최적의 독립변수 조합을 선택하기 위한 통계적 기법이다.
- 여러 독립변수 중에서 종속변수와 가장 연관성이 높은 독립변수 집합을 선택한다.
- 과대적합(overfitting)을 방지하고, 불필요한 변수를 배제하여 모델의 단순성을 유지할 수 있다.

⊖ 전진 선택(Forward Selection)

- 처음에는 아무 변수도 포함되지 않은 상태에서 시작한다.
- 가장 유의미한 변수를 하나씩 추가하며 모델을 확장해 나간다.
- 한 번 포함된 변수는 이후 제거되지 않는다.

⊖ 후진 제거(Backward Elimination)

- 모든 변수를 포함한 상태에서 시작한다.
- 가장 유의미하지 않은 변수를 하나씩 제거하며 모델을 축소해 나간다.
- 한번 제거된 변수는 다시 추가되지 않는다.

⊗ 단계별 방법(Stepwise Method)

- 단계별 방법(Stepwise Method)은 전진 선택법 (Forward Selection)과 후진 제거법(Backward Elimination)을 결합한 변수 선택 방법이다.
- 변수를 단계적으로 추가하면서도 동시에 기존에 포함된 변수의 중요성을 재검토하고, 필요에 따라 제거하는데, 이는 새로운 변수가 추가됨에 따라 기존 변수의 중요도가 달라질 수 있기 때문이다.

135

정규화 선형회귀 (Regularized Linear Regression)[★]



AI

TIP: 출제유형

릿지 회귀의 정의

- 정규화 선형회귀는 과대적합(Overfitting)을 방지하고 모델의 일반화 성능을 향상시키기 위해, 회귀계수에 제약(regularization)을 추가한 선형 회귀 모델이다.
- 주요 방법으로 릿지(Ridge), 라쏘(Lasso), 엘라스틱넷 (Elastic Net)이 있다.

⊖ 릿지 회귀(Ridge Regression)

- 릿지 회귀(Ridge Regression)는 회귀 계수의 크기를 제어하기 위해 **L2 정규화**를 사용한다.
- L2 정규화는 회귀 계수의 **제곱합을 최소화하는** 제약 조건이다.
- 변수 선택(완전히 제거)은 불가능하다.

137

공분산과 상관계수의 관계



TIP: 출제유형

확률변수가 독립일 때 공분산 관계

- 공분산은 두 변수 간의 선형적 관계를 나타내는 통계량으로, 한 변수가 증가하거나 감소할 때 다른 변수가 어떻게 변화하는지를 측정한다.
- 공분산이 양수이면 확률변수 X가 증가하면 Y도 증가하는 경향이 있다.
- 공분산이 음수이면 확률변수 X가 증가하면 Y는 감소하는 경향이 있다.
- 공분산이 0이면 확률변수 X,Y간에 선형적 관계가 없음을 의미한다.
- 확률변수 X,Y가 독립이면 공분산은 항상 0이지만, 공분산이 0이라고 해서 두 확률변수가 독립이라 할 수 없다.
- 이는 단순히 선형적 관계가 없다는 것만 의미하며, 비선형 관계는 존재할 수 있다.

	공분산	상관계수
정의	두 변수의 선형적 관계 측정	두 변수의 선형적 관계를 단위없이 표준화
범위	-무한대 ~ + 무한대	-1에서 1사이
해석의 용이성	단위에 의존하므로 해석이 어렵고 비교가 어려움	단위가 없으므로 해석과 비교가 용이함

⊖ 스피어만 상관계수(Spearman Rank Correlation Coefficient)

- 스피어만 상관계수는 두 변수 간의 **순위 기반 관계**를 측정하는 통계량이다.
- 값의 범위는 $-1 \leq r \leq 1$ 이며, 이는 두 변수 간의 **단조 (monotonic) 관계**를 평가한다.
- **단조관계란** 한 변수가 증가(또는 감소)하면, 다른 변수도 일정한 방향으로 증가(또는 감소)하는 관계를 의미한다.
- 스피어만 상관계수는 **두 변수 간의 순서 관계를 평가하므로**, 선형뿐만 아니라 **비선형 관계**도 반영할 수 있다.
- 서열척도처럼 변수 값 간의 순서만 의미가 있고, 간격이 일정하지 않은 경우에 적합하다.

139

상관계수의 유의성 검정(Significance Test for Correlation Coefficient)[★]



AI

TIP: 출제유형

상관계수의 유의성 검정 시 귀무가설과 대립가설

- 상관계수는 관계의 강도와 방향만 제공할 뿐, 해당 관계가 통계적으로 유의미한지 여부는 확인할 수 없다.
- 따라서 상관계수의 유의성 검정을 통해 관찰된 상관관계가 우연인지 아닌지를 검토해야 한다.
- 상관계수의 유의성 검정
 - 귀무가설: 상관관계가 없다(상관계수 = 0)
 - 대립가설: 상관관계가 있다(상관계수 $\neq$ 0)

146

주성분 적재량 (Loadings)



AI

TIP: 출제유형

주성분 적재량의 해석

⊖ 주성분 적재량(Loadings)의 정의

- 적재량(Loadings)은 각 원래 변수(Murder, Assault, UrbanPop, Rape)가 주성분 (Comp.1, Comp.2 등)을 형성하는데 기여하는 비율(가중치)을 나타낸다.
- 적재량은 원래 변수와 주성분 간의 상관관계를 나타내며, 값이 클수록 해당 변수가 주성분에 더 크게 기여함을 의미한다.
- 주성분은 원래 변수의 선형 결합으로 나타나며, 적재량이 각 변수의 가중치로 역할을 한다.

⊖ 주성분의 표현식

$$Y1 = -0.536 \cdot \text{Murder} - 0.583 \cdot \text{Assault} - 0.278 \cdot \text{UrbanPop} - 0.543 \cdot \text{Rape}$$

$$Y2 = 0.418 \cdot \text{Murder} + 0.188 \cdot \text{Assault} - 0.873 \cdot \text{UrbanPop} - 0.167 \cdot \text{Rape}$$

• 첫 번째 주성분(Y1)

- Murder(-0.536), Assault(-0.583), Rape(-0.543)가 비슷한 크기로 강하게 기여하고 있다.
- 첫 번째 주성분이 범죄율(특히 폭력 관련 범죄)과 관련된 데이터를 요약한다는 것을 의미

• 두 번째 주성분(Y2)

- Murder와 Assault는 양수로, UrbanPop은 음수로 적재된다.
- UrbanPop(-0.873)이 압도적으로 기여도가 크다.

- 두 번째 주성분이 도시화 수준(UrbanPop)과 관련된 데이터를 요약한다는 것을 의미한다.

149

평활법(Smoothing Method)★



AI

TIP: 출제유형

지수 평활법의 정의

- **평활법**은 시계열 데이터에서 불규칙한 변동을 제거하여 데이터의 전반적인 흐름(추세)을 파악하고 미래를 예측하기 위한 기법
 - 불필요한 변동성을 제거하여 시계열의 주요 패턴을 명확히 하고, 분석 및 예측의 정확도를 높일 수 있다. 주요 평활법으로 **이동평균법(Moving Average)**과 **지수평활법(Exponential Smoothing)** 이 있다.
- ⊖ **이동평균법 (Moving Average)**
- 일정한 기간(윈도우 크기)의 데이터 평균을 계산하여 이동하면서 시계열을 부드럽게 조정하는 방법이다.
 - **계절변동과 불규칙변동을 제거하여** 데이터의 주요 변동 패턴인 **추세변동과 순환변동**만 남기도록 하는 시계열 평활화 기법이다.
- ⊖ **지수평활법 (Exponential Smoothing)**
- **최근 데이터에 더 큰 가중치를 부여하여** 시계열을 평활화 방법이다.
 - 시간 흐름에 따라 과거 데이터의 영향이 지수적으로 감소하며, 최근 추세를 보다 민감하게 반영할 수 있다.

152

시계열 모형 자기회귀누적이동평균 모형(ARIMA)



AI

TIP: 출제유형

ARIMA 모형 식별

- 비정상 시계열 데이터를 차분(Differencing)을 통해 정상화한 후, 자기회귀(AR)와 이동평균(MA)의 조합으로 모델링하는 시계열 분석 방법이다.
- ARIMA는 다양한 시계열 패턴을 포괄하며, 대부분의 비정상 시계열 데이터를 모델링하는 데 사용한다.
- 대부분의 시계열 데이터는 비정상을 띠기 때문에, ARIMA 모형은 차분(d)을 통해 데이터를 정상화한 후, ARMA 모형 형태로 분석을 수행한다.

⊖ ARIMA 모형의 구성

▪ ARIMA(p, d, q)

- p : AR(자기회귀) 모형의 차수
- d : 차분(Differencing)의 횟수
- q : MA(이동평균) 모형의 차수

⊖ 특수 형태

- $p = 0$: IMA(d, q) 모형 → d번 차분한 후 MA(q) 모형으로 해석
- $q = 0$: ARI(p, d) 모형 → d번 차분한 후 AR(p) 모형으로 해석
- $d = 0$: ARMA(p, q) 모형

3과목 : 데이터 분석

8장. 정형 데이터 마이닝

01. 데이터 마이닝

02. 모형평가

03. 분류분석

04. 군집분석

05. 연관분석

156

데이터 마이닝 세부 추진 5단계



AI

TIP: 출제유형

데이터 마이닝 프로세스 순서

- 데이터 마이닝 프로젝트를 효율적으로 수행하기 위해 일반적으로 다음 5단계로 구성된 절차를 따른다.

⊖ 문제 정의(Problem Definition)

- 데이터 마이닝의 목표와 문제를 명확히 정의하는 단계

⊖ 데이터 준비(Data Preparation)

- 데이터 마이닝 수행에 필요한 데이터를 수집하고 정리하는 단계

⊗ 데이터 가공

- 수집된 데이터를 적합한 형식으로 가공하는 단계
- 목표변수를 정의하고, 입력 변수와 구분

④ 데이터 마이닝 기법의 적용

- 목적에 맞는 데이터 마이닝 기법을 적용하여 분석 수행하는 단계

⑤ 검증

- 데이터 마이닝 결과의 신뢰성과 타당성을 평가하는 단계

159

교차검증 방법 ★★★



AI

TIP: 출제유형

교차검증 기법 정의

⊖ 홀드아웃(Hold-out) 교차검증

- 홀드아웃(Hold-out)은 주어진 데이터를 랜덤하게 두 부분으로 분리하여 훈련용과 검증용에 사용하는 가장 단순한 교차검증 기법이다.
- 훈련용 데이터로 모델을 학습하고, 검증용 데이터로 모델 성능을 평가한다.
- 데이터 분할과 성능 평가 과정이 간단하고 빠르게 수행이 가능하다.
- 전체 데이터의 일부만 학습에 사용하므로, 데이터가 적을 경우 모델 성능이 불안정할 수 있음

⊖ k-폴드 교차검증(k-Fold Cross Validation)

- k-fold 교차검증은 주어진 데이터를 여러 개의 하부 집합으로 나누어 모델의 성능을 평가하는 방법이다.
- 데이터를 k개의 하부집합(subset)으로 분리하고, 각 하부집합을 한 번씩 검증 데이터로 사용하며, 나머지를 훈련 데이터로 사용하는 과정을 k번 반복한다.
- 각 반복에서 측정한 성능을 평균하여 모델의 최종 평가 점수로 사용한다.
- 데이터가 많을 경우 모델 훈련과 검증을 반복하므로 실행 시간이 길어질 수 있다.

③ 붓스트랩(Bootstrap) 교차검증

- 붓스트랩은 단순 랜덤 복원추출 방법을 사용하여, 주어진 데이터에서 동일한 크기의 표본을 여러 번 생성하여 모델을 훈련하고 검증하는 기법이다.
- 추출된 데이터는 훈련 데이터로 사용되고, 한 번도 추출되지 않은 데이터는 검증 데이터로 사용한다.
- 전체 n 개의 데이터에서 n 번 복원추출을 할 때 :
 - 한번이라도 샘플이 포함될 확률 : 63.2%.
 - 한 번도 포함되지 않을 확률 : 36.8%
 - 데이터 하나가 특정 추출에서 선택되지 않을 확률은 $1-1/n$
 - 데이터가 n 번의 추출에서 한 번도 선택되지 않을 확률은 $(1-1/n)^n = \text{약 } 36.8\%$
- 붓스트랩은 데이터가 작을 때도 모델 성능을 안정적으로 평가할 수 있는 강력한 도구이다.

160

오분류표를 활용한 평가지표

AI TIP: 출제유형 혼동행렬을 활용한 평가지표

평가지표	계산식	설명
정분류율 (Accuracy)	$\frac{TP + TN}{TP + TN + FP + FN}$	•전체 데이터 중에서 실제 분류 범주를 정확하게 예측한 비율
오분류율 (Error Rate)	$\frac{FP + FN}{TP + TN + FP + FN}$	•실제 분류 범주를 잘못 예측한 비율 •오분류율 = 1- Accuracy
민감도(Sensitivity) = 재현율(Recall)	$\frac{TP}{TP + FN}$	•실제 Positive 사례 중에서 모델이 올바르게 Positive로 예측한 비율

160 오분류표를 활용한 평가지표 ★★★

평가지표	계산식	설명
특이도 (Specificity)	$\frac{TN}{TN + FP}$	• 실제로 'Negative'인 범주 중에서 'Negative'으로 올바르게 예측(TN)한 비율 • 범주의 불균형 문제에 사용되는 지표
FP Rate	$\frac{FP}{TN + FP}$	• 실제로 'Negative'인 범주 중에서 'Positive'으로 잘못 예측(FP)한 비율 • FP Rate=1-(특이도)
정밀도 (Precision)	$\frac{TP}{TP + FP}$	• 'Positive'로 예측한 비율 중에서 실제로 'Positive' (TP)인 비율

평가지표	계산식	설명
F_1	$2 \times \frac{Precision \times Recall}{Precision + Recall}$	• 정확도와 재현율의 조화평균 • 값의 범위는 0~1사이 • 정밀도와 재현율 중 어느 하나라도 낮으면 F_1 Score도 낮아짐
F_β	$\frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}$	• $\beta > 1$: 재현율(Recall)에 가중치를 부여 • $\beta < 1$: 정밀도(Precision)에 가중치를 부여
카파 통계량	$K = \frac{P_o - P_e}{1 - P_e}$	• K: 카파 계수 • P_o : 관측된 일치율 • P_e : 우연에 의한 기대 일치율 • 분류 모델의 예측값과 실제값 간의 일치 정도를 우연 확률과 비교하여 보정한 지표 • 값의 범위는 -1~1

161

ROC(Receiver Operating Characteristic) 곡선★★



AI

TIP: 출제유형

ROC 곡선의 해석

- ROC 곡선은 이진 분류 모형의 성능을 평가하기 위한 시각화 도구로, 다양한 임계값에 따라 변화하는 민감도(Sensitivity)와 1-특이도(1-Specificity) 간의 관계를 좌표 평면에 곡선으로 나타낸 그래프이다.
- AUC(Area Under the Curve)는 ROC 곡선 아래의 면적으로, 모형의 성능을 숫자로 요약한 값이다.

⊖ ROC 곡선의 해석

- 완벽한 분류 성능을 보이는 모형은 ROC 곡선이 왼쪽 위 모서리($x = 0, y = 1$)에 가까움
- $AUC = 1$ (최상의 분류 성능)
- 분류 성능이 전혀 없는 무작위 예측의 경우, ROC 곡선은 대각선에 위치($AUC = 0.5$)

⊗ 로짓변환

- 로지스틱 회귀모형은 종속변수 y 가 이항값(0 또는 1)만을 가지므로, 선형회귀처럼 직접 예측값을 직선 형태로 나타내기 어렵다.
- 따라서 확률 p 를 오즈(odds)로 바꾸고, 이를 다시 로그 변환한 값을 사용한다.
- 로짓변환을 통해 비선형 관계를 선형식으로 바꿀 수 있어, **선형회귀처럼 회귀계수의 직관적 해석이 가능해진다.**

④ 로지스틱 회귀계수의 부호가 갖는 의미와 해석

- 로지스틱 회귀에서 계수 β_1 의 부호는 설명변수 x 가 증가함에 따라 사건이 발생할 확률이 증가하는지 감소하는지를 나타낸다.
- $\beta_1 > 0$: x 가 증가하면 사건의 확률도 증가 (**그래프는 S자 곡선**)
- $\beta_1 < 0$: x 가 증가하면 사건의 확률은 감소 (**그래프는 역S자 곡선**)
- x 를 1단위 증가시키게 되면 오즈의 예측값은 **$\exp(\beta_1)$ 만큼** 곱해지는 것으로 해석한다.

㉠ 용어

- **가중치(Weights)** : 신경망의 각 연결선에 부여된 값으로, **입력 값이 출력에 미치는 중요도를 나타낸다.**

- **파라미터(Parameter)** : 학습 과정에서 모델에 의해 자동으로 갱신되는 값.

예 가중치, 바이어스

- **하이퍼파라미터(Hyperparameter)** : 모델이 학습되기 이전에 사용자가 직접 설정해야 하는 값으로, 학습의 성능과 효율에 영향을 미친다.

예 학습률(Learning Rate), 은닉층 노드의 수, 배치 크기(Batch Size), 반복 횟수(Epoch)

167

인공신경망에서 은닉층과 은닉 노드 수를 정할 때 고려사항[★]



AI

TIP: 출제유형

과대적합과 인공 신경망의 노드 수

⊖ 노드 수와 과대적합

- 은닉층의 노드 수가 많을수록 모델의 표현력이 향상되어, 복잡한 데이터 패턴을 더 잘 학습할 수 있다.
- 과도하게 많은 노드는 과대적합(Overfitting)의 원인이 될 수 있다.

⊖ 복잡한 의사결정 경계

- 은닉층의 노드 수가 적을 경우, 복잡한 데이터의 패턴이나 의사결정 경계를 잘 학습하지 못할 수 있다.
- 적절한 노드 수를 선택해 데이터의 패턴을 충분히 표현할 수 있도록 조정해야 한다.

⊗ 출력층 노드 수

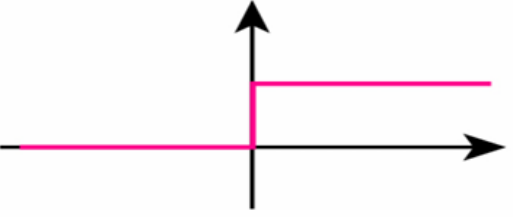
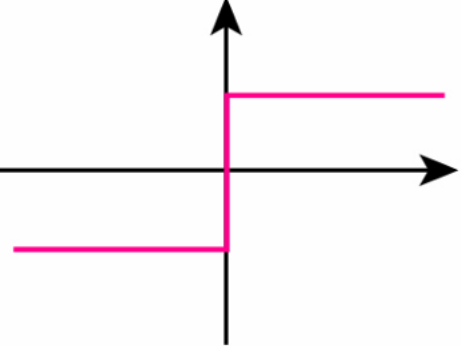
- 출력층의 노드 수는 출력 범주(Output Categories)의 수에 따라 결정한다.
- 입력층의 노드 수는 입력 데이터의 차원(Input Dimension)에 따라 결정한다.

169

인공신경망의 활성화 함수 (Activation Function)^{★★}

AI TIP: 출제유형 활성화함수의 역할

- 활성화 함수(Activation Function)는
인공신경망의 각 노드에서 입력값을 처리하여 비선형 변환을 적용하는 함수이다.
- 입력 데이터의 선형 조합에 비선형성을 부여하여 복잡한 문제(패턴)를 학습할 수 있게 만든다.

활성화 함수	그래프	설명
계단함수 (Step Function)		• 임계값 0.1을 기준으로 활성화 또는 비활성화가 된다.
부호함수 (Sign Function)		• 임계값을 기준으로 양의 부호(+1), • 0, 음의 부호(-1)를 출력

171

기울기 소실 문제 (Vanishing Gradient Problem)★



AI

TIP: 출제유형

기울기 소실 문제의 정의

- 기울기 소실 문제는 깊은 신경망(Deep Neural Network)에서 역전파 알고리즘을 수행할 때 발생하는 대표적인 학습 장애 현상이다.
- 역전파 과정에서 출력층에서 입력층 방향으로 이동할수록, 손실 함수의 기울기 값이 점점 작아지게 되어, 초반층의 가중치가 거의 갱신되지 않게 되는 문제를 의미한다.

175

의사결정나무의 형성 프로세스

AI

TIP: 출제유형

의사결정나무의 생성 프로세스 순서

- 의사결정나무는 데이터를 기반으로 분류 또는 예측하는 데 사용되는 도구이며, 전체 과정은 의사결정나무 생성→가지치기→타당성 평가→분류 및 예측 기준으로 나뉘 설명할 수 있다.

⊖ 의사결정나무 생성

- 각 분할 단계에서 데이터를 가장 효과적으로 구분할 수 있는 특징(Feature)을 선택한다.
- 분리 기준
- 데이터를 가장 효과적으로 나눌 수 있는 방법(지니 불순도, 엔트로피, 분산 감소 등)을 선택한다
- 정지 규칙
- 트리의 과도한 성장을 방지하기 위해 설정한다.
- 적절한 매개변수를 지정하여 모델 성능 최적화와 과대적합 방지를 위해 필요하다.
- 재귀적 분할
- 선택된 특징과 기준을 바탕으로 데이터를 분할하고, 각 자식 노드에 대해 동일한 과정을 반복한다.

177

의사결정나무의 분류와 회귀^{★★}



AI

TIP: 출제유형

분류나무와 회귀나무의 불순도 측정 지표

- 의사결정나무는 분류(Classification)와 회귀(Regression) 문제 모두 해결하는 데 활용될 수 있는 유연한 지도학습 모델이다.

⊖ 분류나무

- 목표변수가 이산형인 경우 사용
- 데이터를 특정 클래스(범주)로 나누기 위한 분류하는 역할을 수행한다.
- 카이제곱 통계량의 p값
 - p값이 작을수록 자식 노드 간의 이질성이 크므로, p값이 작아지는 방향으로 분할한다.
- 지니 지수(Gini Impurity)
 - 자식 노드 내의 불순도를 측정하며, 값이 작아질수록 자식 노드가 순수가 높아진다.
- 엔트로피 지수(Entropy)
 - 값이 작아질수록 데이터가 순수하므로, 엔트로피가 최소화되는 방향으로 분할한다.

⊖ 회귀나무

- 목표변수가 연속형인 경우 사용
- 입력 데이터를 값의 크기에 따라 분할하여 예측하는 회귀 기반 트리 모델이다.
- F통계량의 p값
 - 각 분할에서 자식 노드 간의 평균차이(이질성)를 평가하는데 사용한다.
 - F통계량이 클수록(p값이 작을수록) 자식 노드 간의 이질성이 크다는 것을 의미한다.
 - F통계량이 큰 방향으로 분할을 수행하는 것이 바람직하다.
- 분산 감소량(Variance Reduction)
 - 분할 전후의 분산 차이를 측정하여 분산이 크게 감소하는 방향으로 데이터를 분할한다.

179

의사결정나무의 장·단점



AI

TIP: 출제유형

의사결정나무의 장점 및 단점

⊖ 장점

- 직관적이고 해석이 용이
 - 트리 구조를 통해 변수 간의 관계와 분류.예측 기준을 명확히 드러나며
 - 시각적으로 결과를 직관적으로 이해할 수 있다.
- 교호작용(Interaction) 파악 가능
 - 특정 변수 조합(예 나이와 소득, 혈압과 BMI 등)이 목표 변수에 어떤 영향을 미치는지 트리 분기 구조를 통해 시각적으로 확인할 수 있다.
- 비모수적 모형
 - 데이터 분포에 대한 가정이 없이 작동하며, 다양한 유형의 변수에 유연하게 적용 가능하다.

⊖ 단점

- 분류 기준값 경계선 부근의 자료값에 대해 오차가 크다.
 - 경계값에 가까운 데이터 포인트는 분류 또는 예측이 불안정하거나 부정확할 가능성이 높다.
- 로지스틱 회귀와 같이 각 예측변수의 효과를 파악하기 어렵다.
 - 로지스틱 회귀나 선형 회귀와 달리, 변수의 계수(coefficient)를 제공하지 않으므로 변수의 효과를 정량적으로 해석하기가 어렵다

181

앙상블 모형의 주요 방법^{★★★}



AI

TIP: 출제유형

앙상블 모형의 종류 특징

⊖ 배깅(Bagging : Bootstrap Aggregating)

- 훈련 데이터에서 중복을 허용하여 여러 **부트스트랩**을 생성한다.
- 각 샘플로 독립적인 분류기를 학습한 후, 결과를 결합하여 최종 예측을 생성한다.
- 각 샘플에 대해 독립적인 학습기를 훈련한다.
- 학습기들은 병렬적으로 학습되며, 서로 간의 의존성이 없다.

⊖ 부스팅(Boosting)

- 학습기를 순차적으로 생성하며, 각 단계에서 이전 학습기의 오류를 개선하는 방식이다.
- 각 학습기에서 **잘못 분류된 데이터에 더 많은 가중치를 부여**하여 다음 학습기가 다음 학습기가 해당 데이터를 더 잘 학습하도록 유도한다.
- 학습기들이 순차적으로 연결되며, 오류를 누적적으로 보완한다.

⊗ 스택킹(Stacking)

- 서로 다른 알고리즘의 학습기들을 결합하여 성능을 향상시키는 앙상블 기법이다.
- 기본학습기(Base Learners)
 - 서로 다른 종류의 모델들이 동일한 훈련 데이터를 기반으로 각각 예측값을 생성한다.
- 메타 학습기
 - 기본 학습기들의 예측값을 새로운 입력으로 사용하여 학습되는 모델이다.

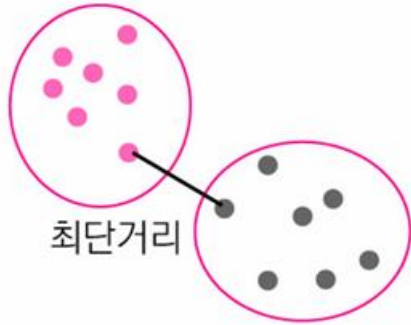
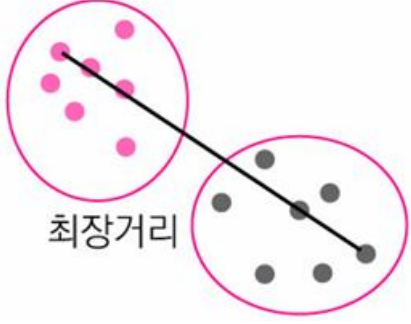
④ 랜덤 포레스트(Random Forest)

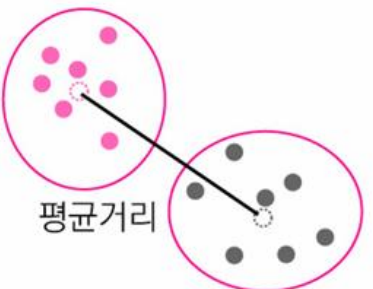
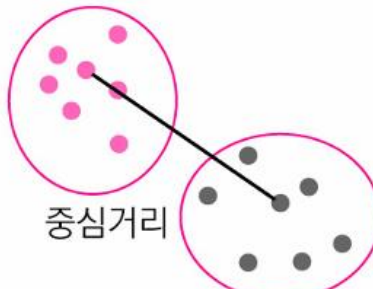
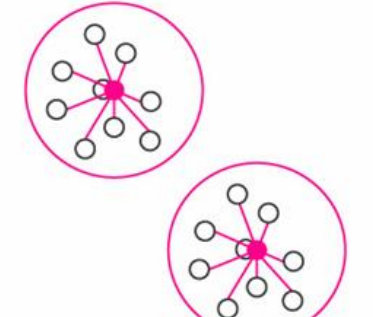
- 배깅의 확장 버전으로, 여러 개의 결정트리를 학습하고 그 결과를 집계하여 예측을 수행한다.
- 각 노드에서 분할할 때, 전체 특징 중 무작위로 선택된 일부 특징만 사용하여 모델 간 다양성을 유도한다.
- 과대적합을 효과적으로 방지 효과가 크며, 단일 결정트리보다 예측 성능이 뛰어나다.

186 계층적 군집분석의 거리 측정(병합방법)

AI TIP: 출제유형 계층적 군집분석의 병합방법

- 계층적 군집분석에서는 군집 간의 유사성 또는 거리를 측정하여 데이터를 병합하거나 분할한다.
- 이를 위해 다양한 거리 측정 방법이 사용되며, 각 방법은 군집의 정의와 결과에 영향을 미친다.

군집방법	거리 측정 비교	두 군집사이의 거리
단일연결법 (single linkage) 또는 최단 연결법		• 한 군집의 점과 다른 군집의 점 사이의 가장 짧은 거리 (Shortest Distance) • 사슬 모양으로 생길 수 있으며, 고립된 군집을 찾는데 중점을 둔 방법이다.
완전연결법 (complete linkage) 또는 최장 연결법		• 두 군집에 속한 모든 점들 간 거리 중 가장 먼 거리를 기준으로 군집 간 유사성을 측정한다. • 군집 내 응집도가 높은 군집을 형성한다. • 군집 간 경계가 명확하다.

군집방법	거리 측정 비교	두 군집사이의 거리
평균 연결법 (average linkage)		모든 항목에 대한 거리 평균을 구하면서 군집화를 하기 때문에 계산량이 불필요하게 많아질 수 있다.
중심 연결법 (centroid linkage)		- 각 군집의 중심 간 거리를 측정한다. - 두 군집이 결합할 때 새로운 군집의 평균은 가중평균을 통해 구해진다.
와드 연결법 (ward linkage)		군집 내의 오차제곱합의 증가를 최소화한다. - 가장 일반적이고 강력한 방법 중 하나로, 응집력 높은 군집을 형성한다.

190 K-평균 군집(K-Means Clustering) 작동 원리



AI

TIP: 출제유형

K-평균 군집의 프로세스 순서

⊖ 초기화

- k개의 초기 군집 중심(centroid)을 임의로 선택한다.
- 초기값은 무작위로 설정되며, 이는 군집 결과에 영향을 미칠 수 있다.

⊖ 데이터 할당(Assignment)

- 각 데이터 포인트를 가장 가까운 군집 중심에 할당한다.
- 유클리드 거리(Euclidean Distance)를 이용해 데이터와 군집 중심 간 거리를 계산한다.

⊗ 중심값 갱신(Update)

- 각 군집 내의 데이터 평균을 계산하여 군집 중심을 갱신한다.
- 갱신된 중심은 이전 중심과 다를 수 있다.

④ 반복(Iteration)

- 할당과 중심 갱신 단계를 반복한다.
- 군집 중심이 더 이상 변하지 않거나(수렴), 최대 반복 횟수에 도달하면 종료한다.

192

군집분석의 실루엣 계수 (Silhouette Coefficient)[★]



AI

TIP: 출제유형

실루엣 계수의 정의

- 실루엣 계수(Silhouette Coefficient)는 군집화의 품질을 평가하는 지표로, 각 데이터 포인트가 자신의 군집 내에서 얼마나 잘 속해 있는지 다른 군집과 얼마나 잘 분리되는지를 측정한다.
- 실루엣 계수는 군집의 응집도(cohesion)와 분리도(separation)를 동시에 고려한다.

⊖ 실루엣 계수의 해석

- 실루엣 계수는 -1 에서 1 사이의 값을 가진다.
- 1 : 데이터가 잘 맞는 군집에 속하며, 다른 군집과 잘 분리됨
- -1 : 데이터가 잘못된 군집에 속함
- 엘보우 기법과 함께 군집 수(k)를 결정하는 데 유용하게 사용된다.

195

EM 알고리즘

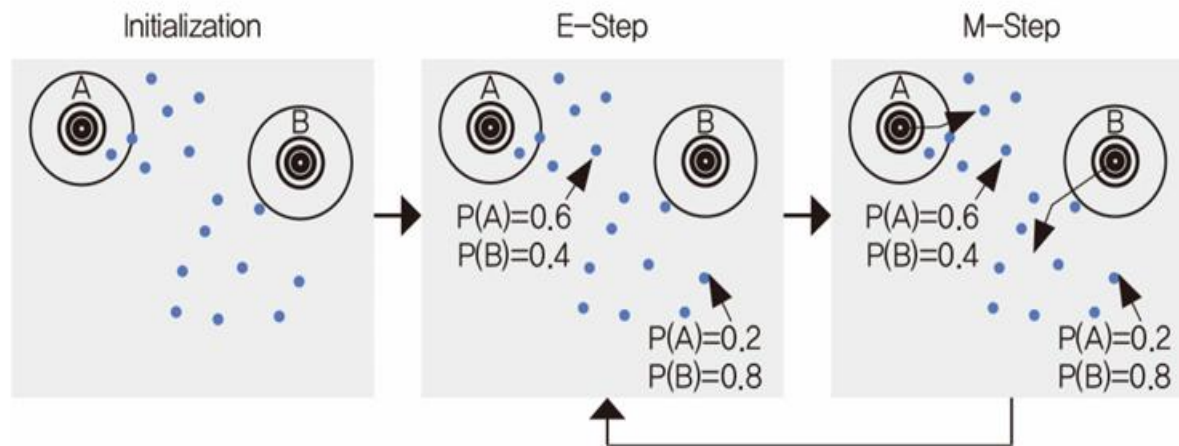


AI

TIP: 출제유형

EM 알고리즘의 정의

- A군집과 B군집이 혼합되어 있는 혼합분포 군집에서는, 초기에 K-평균 군집처럼 군집 중심이 랜덤하게 설정
- 각 관측치는 A군집에 속할 확률 또는 B군집에 속할 확률이 계산된다.
- 이때 각 데이터가 어느 집단에 속하는지에 대한 정보를 가지는 변수를 **잠재변수 (Latent variable)**라 한다.
- 이런 확률을 계산하는 단계를 **E-step(단계)**라 하고, 이후에는 이 확률을 이용해서 최대우도추정을 통해 모수를 다시 추정하는 **M-step (단계)**을 수행한다.
- 이 과정을 반복



!을 완성한다.

[EM 알고리즘]

197

SOM 경쟁 학습 (Competitive Learning)의 개념★



AI

TIP: 출제유형

경쟁학습과 SOM의 구조

- SOM(Self-Organizing Maps)의 경쟁 학습은 입력 벡터와 출력 뉴런(격자 내 신경세포) 간의 유사성을 기반으로 동작하며, **특정 뉴런이 승자 뉴런(Winner Neuron)이 되어 학습을 수행한다.**
- 입력 벡터가 여러 번 신경망에 제시되며, 출력 뉴런 간 경쟁을 통해 연결 가중치가 자율적으로 조정된다.

⊖ 알고리즘 특징

- 입력 벡터들을 신경회로망에 계속 제시하면서 자율적으로 연결가중치를 변경된다.

⊖ 단순구조

- 한 개의 입력층과 한 개의 출력층으로 구성된다.
- 입력층과 출력층이 **완전연결(fully-connected)** 되어 있다.
- 출력 뉴런들은 승자 뉴런이 되기 위해 경쟁하고 오직 승자만이 학습에 참여한다.

199

연관성 분석 (Association Analysis) ★★★



TIP: 출제유형

연관성 분석의 측정 지표

- 연관성 분석은 대규모 데이터셋에서 데이터 간의 숨겨진 패턴과 관계를 탐구하는 데이터 마이닝 기법
- 주로 장바구니 분석(Market Basket Analysis)에서 활용되며, 고객이 함께 구매하는 상품 간의 연관 규칙을 도출하는 데 사용

용어	개념 및 수식
지지도 (Support)	• 전체 거래 중 항목 A와 B를 동시에 포함하는 거래의 비율 • $P(A \cap B) = \frac{A와 B가 동시에 포함된 거래수}{전체거래수}$ • 지지도는 연관규칙 $A \rightarrow B, B \rightarrow A$가 같은 지지도를 갖는다.

용어	개념 및 수식
신뢰도 (Confidence)	A상품을 구매했을 때 B상품을 구매하게 되는 조건부 확률 $\frac{P(A \cap B)}{P(A)} = \frac{A와 B가 동시에 포함된 거래수}{A를 포함하는 거래수}$
향상도 (Lift)	- A와 B가 동시에 발생한 정도를, A와 B가 독립적일 때 기대되는 발생 정도와 비교한 비율 $\frac{P(B A)}{P(B)} = \frac{P(A \cap B)}{P(A) \times P(B)}$ - 향상도 = 1: 두 품목은 독립이다. - 향상도 > 1: 양의 관계 → 함께 발생할 확률이 더 높음 - 향상도 < 1: 음의 관계 → 함께 발생할 확률이 더 낮음

200

Apriori 알고리즘 프로세스



AI

TIP: 출제유형

Apriori 알고리즘 프로세스

- Apriori 알고리즘은 연관성 분석을 수행하기 위해 빈발 아이템 집합(Frequent Itemsets)과 연관 규칙 (Association Rules)을 생성하는 대표적인 1세대 알고리즘이다.
- ⊖ 최소 지지도 설정
 - 분석을 시작하기 전에 데이터에서 빈발한 아이템 집합을 정의하기 위한 최소 지지도 (Minimum Support)를 설정한다.
- ⊖ 단일 품목 중 최소 지지도를 넘는 품목 찾기
 - 각 개별 항목에 대해 지지도를 계산한다.
 - 계산된 지지도가 설정한 최소 지지도 이상인 항목만 선택
- ⊗ 최소 지지도를 넘는 두 가지 품목 집합 찾기
- ④ 최소 지지도를 넘는 세 가지 품목 집합 찾기
- ⑤ 반복적으로 수행하여 빈발 품목 집합 찾기
 - 이전 단계에서 빈발 아이템 집합이 더 이상 생성되지 않을 때까지 절차를 반복
 - 최종적으로 남은 모든 빈발 아이템 집합으로부터 신뢰도 기준을 적용하여 의미 있는 연관규칙을 생성한다

201

연관분석의 장·단점★★

AI

TIP: 출제유형

연관분석의 장점 및 단점

⊖ 장점

- 이해하기 쉬운 결과
 - 결과가 조건반응(if-then) 형식의 규칙으로 제시되어 직관적 이해가 용이하다.
- 비목적성 분석 기법
 - 특정 목적을 미리 설정하지 않아도 숨겨진 관계를 탐색할 수 있다.
- 계산의 단순성
 - 알고리즘 구조가 비교적 단순하여 구현 및 이해가 용이하다.

⊖ 단점

- 분석 대상 품목 수가 증가하면, 후보 집합의 수가 기하급수적으로 증가(조합 폭발)하여 계산량이 급격히 증가한다.
- 거래량이 적은 품목 **예** (희소 데이터)은 지지도가 낮아 분석에서 제외될 가능성이 높음
- 품목이 지나치게 세부화 되어 있을 경우, 의미 없는 연관 규칙이 도출될 수 있다.

여러분의 **ADsP**
합격을 기원합니다^^!!

DATA
ANALYSIS
Level 1.5

직장에서 사랑받는 프로 일잘러의 모든 것

epass biz