AIEDU 대표 : 김 계 철

ADsP 49회 합격을 위한 교육 자료



KT 통신사의 ADsP 어플 인증 수신에 대한 공지사항

오늘 라이브 강의 중 앱 이용 안내

현재 일부 KT 및 KT망 알뜰폰 회선에서
ADsP 어플 인증번호 문자가 수신되지 않는 사례가 확인되었습니다.

해당 건은 KT 기술부서에 확인을 의뢰한 상태입니다.

수험생 여러분의 원활한 앱 사용을 위해 아래 임시 인증 방법을 안내드립니다.

현재 확인된 상황

SKT·LG U+는 정상 수신, KT 및 일부 KT망 알뜰폰에서 수신 지연 또는 미수신 사례 발생

정상 수신

SKT / LG U+ 회선

수신 오류 확인

KT 회선 및 일부 KT망 알뜰폰

조치 진행

KT 기술부서 확인 의뢰 중

안내 핵심: 앱 설치는 가능하며, 인증번호 수신 문제만 일부 회선에서 발생하고 있습니다.

오늘 강의 중 임시 인증 방법 안내

인증번호가 오지 않는 수험생은 아래 방법으로 임시 인증을 진행해 주세요.

1

앱 설치

ADsP 어플을
먼저 설치합니다.

→

2

인증번호 요청

본인 동의하에
가족·지인 번호로
인증번호를 받습니다.

→

3

코드 전달

받은 인증코드를
수험생 본인이
앱에 입력합니다.

→

4

앱 사용

인증 완료 후
강의에 맞춰
앱을 사용합니다.

※ 개인정보 보호를 위해 인증번호는 수험생 본인이 직접 입력하고, 임시 인증 후 앱 이용 안내에 따라 진행해 주세요.

강의 중 안내 멘트

대표 또는 강사가 바로 읽을 수 있는 문구

현재 KT 통신사 및 일부 KT망 알뜰폰 회선에서 ADsP 어플 인증번호가
수신되지 않는 사례가 확인되어 KT 기술부서에 확인을 의뢰한 상태입니다.

오늘 강의 중 앱 사용이 필요한 경우, 앱 설치 후 인증번호가 오지 않는 수험생께서는
가족 또는 지인의 휴대폰 번호로 인증코드를 받아 임시로 인증을 진행해 주시기 바랍니다.

불편을 드려 죄송하며, 안정적인 이용을 위해 계속 확인하겠습니다.

출제 예상

1 과목 데이터 이해

01장 데이터의 이해

01 데이터와 정보

① 정성적 및 정량적 데이터 정의

② 암묵지와 형식지 상호작용(공통화→표출화→연결화→내면화)

③ DIKW 정의 및 사례★

02 데이터 베이스 정의와 특징

④ DB vs DBMS(객체지향 DBMS)

⑤ 데이터베이스 vs ODS vs 데이터 레이크 vs 데이터웨어하우스 vs 데이터 마트 개념 구분★

⑥ DB 설계 프로세스, ETL 기능

⑦ 데이터베이스 특성(정보의 축적 및 절달, 정보 이용, 정보 관리, 정보 기술 발전) ★★

03 데이터베이스 활용

⑧ 기업 내부데이터 베이스 솔루션 7+1 정의★

⑨ 사회기반 데이터베이스 솔루션

출제 예상

1 과목 데이터 이해

02장 데이터의 가치와 미래

01 빅데이터의 이해

① 빅데이터의 특징 3V

② 빅데이터의 본질적인 변화

③ 빅데이터 플랫폼 역할

02 빅데이터의 가치와 영향

④ 빅데이터의 가치 산정이 어려운 이유

⑤ 빅데이터의 영향

03 데이터베이스 활용

⑦ 빅데이터 활용 테크닉 ★

출제 예상

1 과목 데이터 이해

02장 데이터의 가치와 미래

04 위기요인과 통제방안

① 빅데이터 시대의 위기요인과 통제방안 ★

② 개인정보 비식별화 기술 ★

③ 개인정보보호법

05 미래의 빅데이터

④ 빅데이터 활용 3요소

출제 예상

1 과목 데이터 이해

03장 가치 창조를 위한 데이터 사이언스와 전략 인사이트

01 빅데이터 분석과 전략 인사이트

① 전략적 통찰이 없는 분석의 함정 (일차원적 분석 vs 전략 도출 위한 가치 기반 분석) ★

02 전략 인사이트 도출을 위한 필요 역량

② 데이터 사이언스(vs 통계학 vs 데이터마이닝) vs 데이터 사이언티스트(하드 vs 소프트)★

03 빅데이터 그리고 데이터 사이언스의 미래

③ 가치 패러다임의 변화

1. 다음 중 암묵지와 형식지의 상호작용 과정으로 옳은 것은(47회)?

② 공통화 → 표출화 → 연결화 → 내면화

- 정성 데이터(기상특보)와 정량 데이터 비교
- 정량적 데이터는 수치 계산이 가능
- 정성적 데이터는 범주·속성 구분만 가능
- 정형 데이터 반정형 데이터 비정형 데이터

1. SECI 모델의 지식 변환 과정에 대한 설명이다. 이 중 옳지 않은 것은 무엇인가(48회)?

② 연결화(Combination)는 암묵지를 기반으로 기존 형식지를 분석·편집·결합하여 새로운 암묵지를 창출하는 과정이다.

- 암묵지를 기반으로 새로운 암묵지를 창출하는 과정은 공동화 (Socialization)이다.

2. DIKW 계층 구조에 따라 분류할 때, 유형이 다른 하나는 무엇인가(47회)?

③ 1~8월 매출 추이를 고려할 때 다음 분기 매출이 증가할 것으로 예상된다.

- DIKW 정의 및 사례
- 정보와 지식의 차이점
- 정보→ 데이터에 의미를 부여
- 지식→ 정보를 기반으로 판단이나 예측

2. 정성적 데이터, 정량적 데이터 및 데이터의 일반적 특성에 대한 설명이다. 이 중 옳은 것 만을 모두 고른 것은 무엇인가(48회)?

(가) 정량적 데이터는 정성적 데이터에 비해 분석 및 활용에 더 많은 비용과 기술이 필요하다.
(다) 정성적 데이터는 수치 형태로 표현되며, 정량적 데이터에 비해 저장과 분석이 용이하다

- (가).(다) 올바르지 않은 설명

3. 다음 보기가 설명하는 데이터베이스의 특징으로 가장 적절한 것은?(47회)

대량의 정보를 컴퓨터가 읽고 쓸 수 있는 기계 가독성을 가지며, 필요한 정보를 신속하게 찾을 수 있는 검색 가능성, 그리고 정보 통신망을 통해 원거리에서도 즉시 이용 가능한 원격 조작성을 갖는다.

① 정보의 축적 및 전달 측면

- 데이터베이스 특성별 사례 올바르게 연결하기
- 정보의 축적 및 전달
- 정보이용
- 정보관리 측면
- 경제.산업적 측면

3. 다음은 데이터베이스(Database)와 DBMS에 대한 설명이다. 이 중 옳지 않은 것은 무엇인가?(48회)

① 데이터베이스는 데이터, 정보, 지식, 저작물 등의 인식 가능한 모든 자료를 의미하는 포괄적 개념이다.

- **데이터베이스(DB)**는 단순히 "모든 자료"를 의미하는 개념이 아니라,
  여러 자료를 **일정한 기준과 구조에 따라 체계적으로 저장·관리하여 검색과 활용이 가능하도록 한 집합**을 의미

4. 개인정보를 타인이 알아볼 수 없도록 식별 정보를 제거하거나 변형하는 것을 의미하는 용어는 무엇인가(47회)?

③ 익명성(Anonymity)

- 개인정보 비식별화 기술 가명처리·데이터값 삭제·데이터 마스킹·범주화 정의
- 가명처리와 익명처리 차이
- 데이터 3법 관련 문제, 민감정보(유전정보, 범죄 이력 포함), 고유 식별 정보(주민번호, 여권번호)

4. 데이터베이스의 목적 및 특성에 대한 설명이다. 이 중 가장 적절하지 않은 것은?(48회)?

④ 데이터베이스는 여러 사용자가 서로 다른 목적을 가지고 데이터를 공동으로 이용할 수 없다.

- 데이터베이스는 여러 사용자가 공동으로 이용할 수 있도록 설계된다.

5. 빅데이터의 특징으로 가장 적절하지 않은 것은(47회)?

④ 데이터의 가용성(Availability)을 확보하기 위해 안정적인 접근 환경이 필요하다.

- 데이터의 크기 TB→PB→EB→ZB→YB는 각 단계마다 1024배씩 증가.
- 인접한 단위 간의 차이는 모두 1024배
- 데이터 크기를 1024배로 구분하는 이유는 컴퓨터가 2진수 기반으로 동작하며, 메모리와 저장공간이 2의 제곱수 단위로 관리되기 때문이다.

5. 시험 공부 시간에 따라 성적의 변화를 분석하거나 예측하기 위해 가장 적절한 분석 방법은 무엇인가(48회)?

④ 회귀 분석

6. 다음 중 연관분석에 대한 설명으로 적절하지 않은 것은(47회)?

④ 연관분석은 변수 간 선형관계를 파악하는 분석 방법이다.

- 유형분석·유전알고리즘·기계학습·회귀분석·감정분석·소셜네트워크 분석 정의

6. 빅데이터의 활용 과정에서 발생하는 위기 요인에 해당 하는 것 만을 모두 고른 것은(48회)?

(가) 알고리즘 접근 허용 (나) 사생활 침해 (다) 책임 원칙 훼손 (라) 데이터 오용 (마) 익명화

- (나), (다), (라)

7. 다음 중 연관분석에 대한 설명으로 적절하지 않은 것은(47회)?

④ 연관분석은 변수 간 선형관계를 파악하는 분석 방법이다.

- 유형분석·유전알고리즘·기계학습·회귀분석·감정분석·소셜네트워크 분석 정의

7. 빅데이터(데이터 분석) 활용을 위한 기본 3요소에 해당하지 않는 것은 무엇인가(48회)?

① 프로세스

- 빅데이터 활용 기본 3요소: 데이터, 기술, 인력

8. 빅데이터의 위기 요인과 통제 방안에 대한 설명으로 잘못된 것을 고르시오(47회).

가. 사생활 침해 – 제공자 동의에서 사용자 책임제로 전환

나. 책임원칙 훼손 – 알고리즘 접근 허용

다. 데이터 오용 – 결과 기반 책임 부과

③ 나, 다

- 빅데이터의 위기요인과 통제방안 연계
- 빅브라더(Big Brother)는 개인의 행동과 정보를 광범위하게 감시·수집함으로써 사생활 침해를 초래할 수 있는 대표적인 위기요인이다

8. 데이터 사이언티스트의 역할에 대한 설명이다. 이 중 적절하지 않은 것은(48회).

③ 알고리즘으로 인해 부당한 피해를 입은 개인이나 집단을 구제하고, 법적·윤리적 책 임을 직접 수행한다.

- **알고리즈미스트(Algorithmist)**는 데이터 분석 알고리즘이나 AI 알고리즘으로 인해 **부당한 피해를 입은 개인이나 집단을 보호·구제하고**, 알고리즘의 문제점을 해석·검토하는 전문가를 말함

9. 아래 보기가 설명하는 데이터베이스 시스템을 무엇이라 하는가(47회)?

기업에서 발생하는 거래 데이터를 실시간으로 처리하고, 데이터의 입력·수정·삭제가 수시로 이루어지는 시스템

① OLTP

- 대표적 기업내부 데이터베이스 솔루션
- OLTP·OLAP·CRM·SCM·ERP·BI·AI ·KMS
- 블록체인 정의
- 사회기반 데이터베이스 솔루션과 기업내부 데이터베이스 솔루션 구분

9. 빅데이터와 데이터 사이언스에 대한 설명이다. 이 중 가장 적절하지 않은 것은(48회)?

① 데이터 사이언티스트의 소프트 스킬에는 빅데이터에 대한 이론적 지식과 분석 기술 에 대한 숙련도가 포함된다.

- **빅데이터 이론 지식과 분석 기술은 하드 스킬이다.**

10. 데이터 사이언티스트에 대한 설명으로 적절하지 않은 것은(47회)?

② 데이터 사이언티스트는 분석 능력과 기술적 역량이 가장 중요하며, 의사소통 능력이나 협업 역량은 상대적으로 중요하지 않다.

- 데이터 사이언티스트 하드 스킬·소프트 스킬
- 데이터 사이언스·데이터 마이닝·통계학

10. 빅데이터의 가치 패러다임 변화 과정을 올바르게 나열한 것은(48회)?

② Digitalization → Connection → Agency

출제 예상

2과목 데이터 분석 기획

4장 데이터 분석 기획의 이해

01 분석 기획 방향성 도출

① 분석 주제 유형 ★

② 과제단위와 마스터 플랜 단위

③ 분석 기획 시 고려사항- 정형, 반정형 구분/ 데이터 저장방식(RDB vs NoSQL)

02 분석 방법론

③ 분석 방법론의 구성요소

④ 폭포수(Waterfall) 모델 vs 나선형(Spiral) 모델 vs 프로토타입(Prototype) 모델

⑤ KDD vs CRISP-DM vs 빅데이터 분석 방법론(단계, 태스크 스텝) ★

출제 예상

2과목 데이터 분석 기획

4장 데이터 분석 기획의 이해

03 분석과제 발굴

⑥ 하향식과 상향식, 디자인 사고 ★

⑦ 하향식· 상향식 접근 방식 프로세스 ★

⑧ 하향식 접근방식의 혁신적 관점

⑨ 지도학습과 비지도학습 차이

⑩ 프로토타이핑 프로세스 필요성

⑪ 분석과제 정의서

04 분석 프로젝트 관리 방안

⑫ 분석 프로젝트 5가지 주요 특성 ★

⑬ 분석 프로젝트 주요 10개 관리 항목

출제 예상

2과목 데이터 분석 기획

5장 분석 마스터 플랜

01 마스터 플랜 수립

① 마스터 플랜 수립 우선순위 및 적용범위 고려요소

② 시급성과 난이도 구분★

③ 포트폴리오 사분면 분석을 통한 과제 우선순위를 선정

02 분석 거버넌스 체계 수립

④ 분석 거버넌스 체계 구성요소

⑤ 분석준비도와 성숙도 ★

⑥ 데이터 거버넌스의 구성 요소

⑦ 데이터 거버넌스의 체계 요소 ★

⑧ 데이터 분석 업무 주체에 따른 유형

⑨ 분석 과제 관리 프로세스

11. 다음 중 분석 기획 단계에 대한 설명으로 옳지 않은 것은(47회)?

① 복잡한 분석 모형을 설정한다.

- 분석 기획 시 고려사항: 가용한 데이터,적절한 유스 케이스, 장애 요소에 대한 사전 계획 수립
- 복잡하고 정교한 모형이 반드시 효과적인 것은 아니다. 분석 결과를 활용할 수 있도록 단순하고 직관적인 모델을 설계하는 것이 중요하다.

11. 분석 방법은 알고 있으나, 분석해야 할 대상이나 문제 정의가 명확하지 않은 경우에 해 당하는 분석 과제 도출 유형은(48회)?

① 통찰(Insight)

12. 다음 중 데이터 형태에 대한 설명으로 적절하지 않은 것은(47회)?

② IoT 기기에서 생성되는 로그 데이터는 정형 데이터에 해당한다.

- 정형데이터와 반정형데이터의 특징
- 반정형데이터와 비정형데이터의 종류

12. CRISP-DM 분석 방법론의 '데이터 준비(Data Preparation)' 단계에 해당하지 않는 것은 (48회)?

② 데이터 탐색

13. 데이터 탐색 과정의 단계에 대한 설명으로 적절하지 않은 것은(48회)?

① 상세한 문제 해결 방안까지 구체적으로 설계해야 한다.

- 탐색적 분석 과정은 기초 통계량 산출 데이터 분포와 변수 간 관계 파악 데이터 특성 파악해서 모델링 기초 자료로 활용
- KDD·CRISP-DM·빅데이터 분석 방법론

13. 상향식(Bottom-up) 접근 방식으로 데이터 분석을 수행할 때 가장 적절한 분석 방법은(48회)?

① 비지도학습

14. 분석 과제를 발굴할 때, 분석 대상은 명확하나 분석 방법이 명확하지 않은 분석 주제 유형을 무엇이라 하는가(47회)?

③ 솔루션(Solution)

- 분석 주제 유형은 분석 대상과 분석 방법에 존재 여부에 따라 최적화·솔루션·통찰·발견 구분

14. 데이터 분석 프로젝트에서의 품질관리(Quality Management)에 대한 설명이다. 가장 적절한 것은(48회)?

③ 품질관리는 품질검토와 품질보증 활동을 포함하는 개념이다.

- 품질관리는 프로젝트 산출물·절차의 품질을 확보하기 위한 활동
- 품질검토(Quality Review): 산출물 중심
- 품질보증(Quality Assurance): 프로세스 중심

15. 상향식 접근 방식에 대한 설명으로 옳지 않은 것은(47회)?

① 문제가 명확할 때 사용하는 접근 방식이다.

- 분석 과제 발굴 하향식과 상향식 개념 구분
- 상향식 접근 방식의 프로세스 : 프로세스 분류→ 프로세스 흐름 분석→분석요건 분석→분석요건 정의
- 하향식 접근 방식의 프로세스

15. 데이터 분석 과제의 우선순위 설정 기준에 대한 설명이다. 옳은 것만을 모두 고른 것은(48회)?

(가) 데이터의 양과 범위를 조절하면 분석 과제의 시급성을 조절할 수 있다.
(나) 새로운 분석 기술이나 도구의 등장은 분석 과제의 난이도에 영향을 미칠 수 있다.
(다) 분석 수준의 고도화와 비용 증가는 분석 과제의 시급성을 높이는 요인이다.

- **(가) 데이터 양·범위는 난이도와 관련이 더 큼**
- **(나) 새로운 분석 기술·도구의 등장은 난이도에 영향을 줄 수 있음 — 옳음**
- **(다) 비용 증가는 시급성보다는 실행 가능성·난이도와 관련**

16. 분석 마스터 플랜을 수립할 때 우선적으로 고려해야 할 사항을 모두 고르시오(47회).

가. 전략적 중요도
나. 비즈니스 성과
다. 실행 용이성

④ 가, 나, 다

- 분석과제 적용 범위 : 업무 내재화, 분석 데이터 적용 수준, 기술 적용 수준

16. 분석 성숙도(Analytics Maturity)에 대한 설명이다. 가장 적절하지 않은 것은 **(48회).**

④ 분석 성숙도 평가는 유사 업종이나 경쟁사와의 비교 분석을 포함한다.

- 분석 성숙도 평가는 조직 내부의 분석 업무·인력·기법·데이터·문화·IT 인프라 수준을 진단하는 데 초점 유사 업종·경쟁사 비교는 별개의 벤치마킹 영역

17. 다음 중 아래 보기에 해당하는 분석 성숙도 단계를 무엇이라 하는가(47회)?

분석이 시작되는 단계로, 일부 부서에서만 분석이 활용되며
분석 환경과 시스템을 구축하기 시작하는 단계이다.

① 도입

- 분석 성숙도는 CMMI(Capability Maturity Model Integration) 모델 등을 활용하여, 현재 조직의 분석 활용 수준과 프로세스가 어느 단계에 와 있는지 평가하는 것
- 분석 성숙도(Maturity) 단계는 일반적으로 아래와 같은 4단계를 거친다:
- 도입(Initiation)
- 활용(Use)
- 확산(Expansion)
- 최적화(Optimization)

17. 분석 마스터플랜 수립 시 우선순위 결정 요소로 적절하지 않은 것은? (48회)?

① 기술 적용 수준

- 분석 과제 우선순위: 전략적 중요도, 비즈니스 성과, 실행 가능성

18. 분석 업무를 별도의 전담 조직에서 담당하여, 회사 차원의 우선순위에 따라 일괄 수행하는 분석 조직 유형을 무엇이라 하는가(47회)?

① 집중형 조직구조
② 기능 중심형 조직구조
③ 분산형 조직구조
④ 협업형 조직구조

- 분석 조직 구조 중 집중구조와 분산구조의 차이

18. 분석 수준 진단(Level Assessment)을 위한 조사 대상으로 적절하지 않은 것은 **(48회)?**

① 분석 성과에 대한 조사

- **분석 수준 진단은 조직의 분석 준비도와 성숙도를 파악하기 위한 것 분석 성과 자체보다는 역량과 기반을 진단**

19. 데이터 거버넌스에 대한 설명으로 적절하지 않은 것은(47회)?

④ 데이터 거버넌스는 독립적으로 운영되어야 한다.

- 데이터 거버넌스 조직은 기업의 규모나 상황에 따라 독립적인 조직(Council 등)으로 운영될 수도 있고, IT 부서나 전략 부서 내의 하위 조직으로 통합되어 운영될 수도 있다.
- 데이터 거버넌스 체계요소
- 데이터 표준화
- 데이터 관리체계
- 데이터 저장소 관리
- 표준화 활동

19. 분석 조직 구조 유형에 대한 설명이다. 적절하지 않은 것은(48회)?

④ 분산형 분석 조직은 분석 인력이 각 부서에 전담 배치되어 있기 때문에 분석 결과를 실무에 적용하기 어렵다.

- **분산형 분석 조직은 현업 부서에 분석 인력이 배치되어 실무 적용이 쉬움.**

20. 분석 준비도 평가 요소에 포함되지 않는 것은 무엇인가(47회)?

③ 분석에 투입할 비용 및 예산 확보 수준

- 분석 준비도는 기업이 데이터 분석을 도입하고 수행할 수 있는 기반(인프라, 문화, 데이터 등)이 얼마나 갖춰져 있는지 파악하는 것.
- 이때 비용 및 예산은 준비도 평가가 아닌, 분석 과제의 우선순위 선정(Priority) 단계에서 투자 효율성(ROI)이나 시급성을 따질 때 주로 고려되는 요소

20. 데이터 거버넌스 관리 체계 및 관련 업무에 대한 설명으로 적절하지 않은 것은(48회)?

③ 데이터 관리 비용 관리

- 데이터 거버넌스는 데이터 표준화, 품질관리, 메타데이터 관리, 보안, 생명주기 관리 등 조직 차원의 **데이터 관리 체계. 비용 관리는 일반 재무 영역에 해당**

출제 예상

3과목 데이터 분석

6장 R 기초와 데이터 마트

01 R기초

① 데이터 형태(벡터 vs 행렬 vs 데이터 프레임 vs 리스트)

② 상자그림(boxplot)과 히스토그램 해석 ★

③ 데이터 비대칭 분포(중앙값과 평균 크기 비교)

02 데이터 마트

④ reshape, plyr 패키지

03 결측값 처리와 이상값 검색

⑤ 단순 대치법(완전분석법, 평균,중앙값,최빈값, 단순확률,회귀) ★

⑥ summary( )해석 ★

⑦ 단변량 이상치 검색 (ESD)

출제 예상

3과목 데이터 분석

7장 통계분석

01 통계학 개론

① 확률적 표본 추출 ★

② 자료의 척도 ★

③ 조건부 확률(독립 vs 종속)

④ 베이지안 정리

⑤ 이산확률과 연속확률 기댓값 ★

⑥ 이산확률분포(이항,포아송,초기하) 연속확률분포(z분포,t분포,카이제곱,F분포) ★

⑦ 중심극한정리 개념

⑧ 구간추정과 신뢰수준 의미 ★

⑨ 제1종 오류와 제2종 오류 ★

⑩ 유의확률과 유의수준 ★

⑪ 모수와 비모수 특징

출제 예상

3과목 데이터 분석

7장 통계분석

02 기초 통계분석 ★★★

① 회귀분석의 가정

② 잔차분석

③ 다중회귀분석 해석

④ 최적회귀방정식(전진 vs 후진)

⑤ 다중공선성

⑥ 다중회귀 vs 다항회귀

⑦ 회귀분석의 분산분석표

출제 예상

3과목 데이터 분석

7장 통계분석

04 시계열 예측

① 정상성 ★

② 모형 식별(AR,MA,ARIMA) ★

③ 이동평균법 vs 지수평활법

④ 분해시계열 ★

⑤ 비정상 시계열 전환

출제 예상

3과목 데이터 분석

7장 통계분석

03 다변량분석

① 상관계수 정의 및 유의성 검정

② 피어슨 vs 스피어만 ★

③ 다차원척도

④ 주성분 분석 개념 및 결과해석 ★

⑤ 공분산 개념

출제 예상

3과목 데이터 분석

8장 정형 데이터마이닝

03 분류분석

① 로지스틱 회귀분석 vs 일반회귀분석

② 인공신경망의 역전파 알고리즘

③ 활성화 함수 역할

④ 의사결정나무의 불순도 측정 지표

⑤ 분류나무 vs 회귀나무

⑥ CHAID vs CART

⑦ 의사결정나무 특징

⑧ 앙상블 모형(배깅 vs 부스팅 vs 랜덤 포레스트 vs 스태킹)

출제 예상

3과목 데이터 분석

8장 정형 데이터 마이닝

01 데이터마이닝

① 데이터 마이닝 프로세스

② 데이터 마이닝 기능

02 모형평가

③ 오분류표를 활용한 평가지표 ★

④ ROC Curve ★

⑤ 이익도표와 향상도 곡선

⑥ 교차검증(홀드아웃 vs k-fold vs 붓스트랩) ★

출제 예상

3과목 데이터 분석

8장 정형 데이터마이닝

04 군집분석

① 계층 군집 vs 비계층 군집

② k-means 프로세스 및 단점

③ EM 알고리즘

④ SOM vs ANN

⑤ 밀도기반군집

⑥ 실루엣 계수, 엘보우 기법, 덴드로 그램

출제 예상

3과목 데이터 분석

8장 정형 데이터마이닝

05 연관분석

① 지지도 vs 신뢰도 vs 향상도

② 연관규칙의 장점과 단점

③ apriori 알고리즘

④ 순차패턴분석

21. 다음은 주어진 데이터의 사분위수(Q1, Q3)이다.
IQR을 이용하여 이상치를 판단하는 기준 범위(하한,상한)를 구하시오(47회)

$$Q_1 = 4, Q_3 = 12$$

① -8, 24

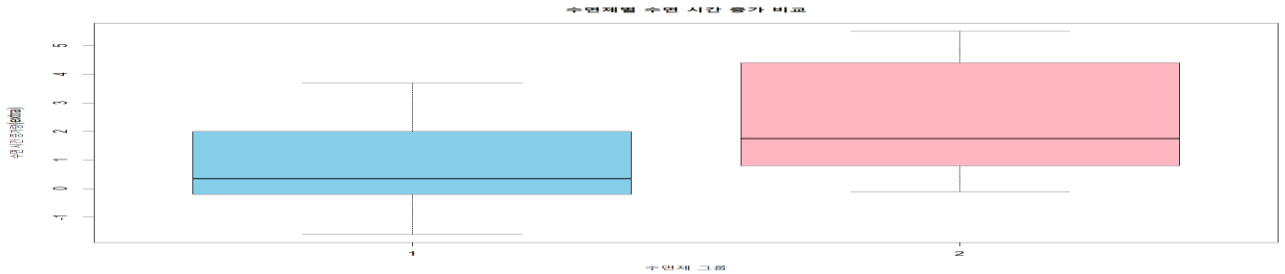
- 상한·하한은 “정상 범위”의 경계선이며, 이를 벗어난 값은 이상치로 간주하기 위한 기준선이다.
- 상자그림만으로는 데이터의 수(표본 크기)를 파악할 수 없다.
- 상자그림은 탐색적 데이터 분석(EDA) 도구
- 가설검정, p-value, 신뢰구간 등을 제공하지 않음

21. 데이터의 이상값(outlier) 탐지 및 처리 방법에 대한 설명으로 가장 적절하지 않은 것은**(48회)?**

① 상자그림(Box Plot)을 활용한 이상값 탐지는 평균과 표준편차를 기준으로 이상 여부를 판단한다.

- Box Plot 기반 이상값 탐지는 평균·표준편차가 아니라 사분위수(Q1, Q3)와 IQR을 기준 평균·표준편차 기준은 ESD(z-score) 방법이다.

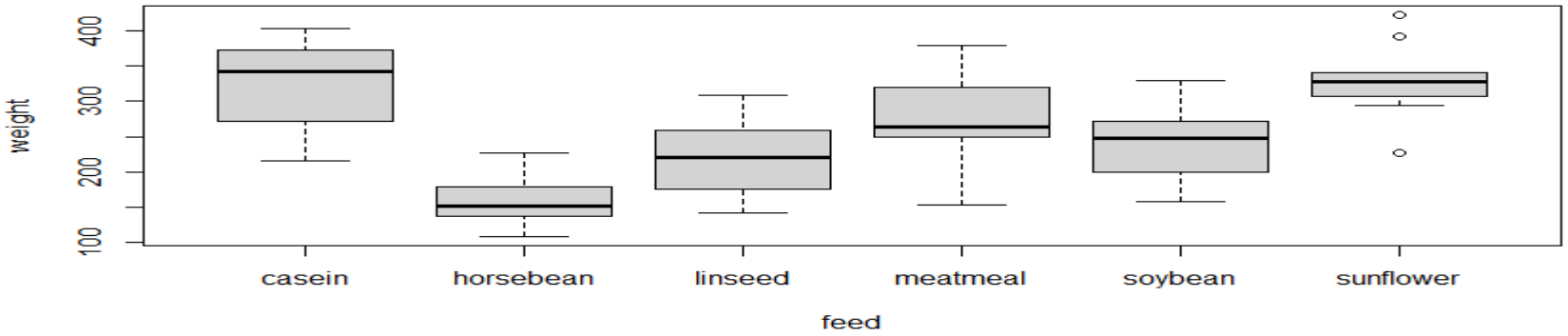
22. R의 내장 데이터셋 sleep 에 대한 박스플롯(boxplot) 해석으로 옳지 않은 것은(47회)?



① 그룹 2는 왼쪽으로 긴 꼬리(long left tail)를 가진 분포를 보인다..

- 상자그림에서 중앙값이 Q1에 가까이 위치하고, Q3 이후의 수염이 더 길게 나타나면 오른쪽 꼬리 분포

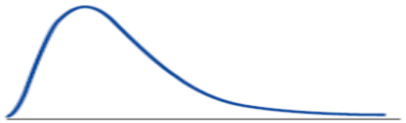
22. chickwts 데이터(여섯 가지 닭 사료별 무게)의 상자그림(Box Plot) 해석 중 가장 적절하지 않은 것은(48회)?



① meatmeal 그룹에 속한 관측치의 수가 가장 많다.

- 관측치 수(표본 크기)는 상자그림만으로 판단 불가함

23. 다음 그림은 오른쪽 꼬리 분포(Right-skewed distribution)를 나타낸 것이다.
이러한 분포의 특징으로 평균, 중앙값, 최빈값의 크기 관계를 바르게 나타낸 것은(47회)?



① 최빈값 < 중앙값 < 평균

- R에서 summary() 함수는 다음과 같은 통계량을 제공한다.
- summary(x)
- Min. 1st Qu. Median Mean 3rd Qu. Max.
- 1사분위수(Q1)의 의미 :전체 데이터를 오름차순으로 정렬했을 때 하위 25% 지점의 값
- 3사분위수(Q3)의 의미: 전체 데이터를 오름차순으로 정렬했을 때 하위 75% 지점의 값
- 데이터 범주형 자료의 예로 성별 변수는 남성 20명, 여성 20명과 같이 빈도(count)로 표현된다.
- 소득이 1사분위수(Q1)가 100만원일 때, 전체 데이터의 약 25%가 100만원 이하임을 의미

23. 교차분석(cross-tabulation analysis)에 대한 설명으로 부적절한 것은(48회)?

① 교차분석은 두 변수 모두 범주형 변수가 아니어도 적용할 수 있으며, 변수 간 관계를 분석하는 데 사용된다.

- 교차분석은 범주형 변수 간 관계를 빈도표 형태로 분석
- 수치형 변수 관계는 상관분석·회귀분석 활용
- 유의성은 카이제곱 검정으로 확인

24. 절대 영점이 존재하는 척도는 무엇인가(47회)?

④ 비율척도

- 명목척도의 가장 중요한 특징은 값의 크기나 순서에 의미가 없고, 단순히 범주를 구분하기 위한 분류용 척도
- 예: 성별, 혈액형, 지역, 종교
- 서열척도의 핵심 특징은 범주 간의 '순서'는 의미가 있으나, 순서 간 '간격의 크기'는 알 수 없다는 점이다.
- 예 만족도(매우 불만족 < 불만족 < 보통 < 만족 < 매우 만족)
- 비율척도의 핵심 특징은 순서·간격이 모두 의미 있으며, '절대적 0'이 존재하여 비율 비교가 가능하다는 점
- 비율척도는 절대적 0을 기준으로 순서·간격·비율 비교가 모두 가능한 측정척도이다.

24. 온도(섭씨, 화씨)나 각종 지수와 같이 절대 영점이 존재하지 않는 데이터를 분류하는 데 가장 적절한 척도는(48회)?

④ 구간척도

25. 표본추출방법에 대한 설명으로 옳지 않은 것은(47회)?

② 계통추출법: 모집단을 일정한 간격으로 나누고, 각 구간 안에서 무작위로 추출한다.

- 군집추출은 군집 내 이질성, 군집 간 동질성을 전제로 한다.
- 군집(cluster)은 표본을 대신해 뽑히는 단위이다.
- 따라서 하나의 군집만 봐도 모집단의 다양한 특성이 포함되어야 한다.
- 군집 간은 '동질적'이어야 할까?
- 서로 다른 군집들이 비슷한 구조와 특성을 가져야 어떤 군집을 뽑아도 결과가 크게 달라지지 않는다.
- 층화추출은 층 내 동질성, 층 간 이질성을 전제로 한다..

25. 상관계수에 대한 설명 중 옳은 것은(48회)?

② X+0.3과 Y+0.2의 상관계수는 0.5이다.

- 상관계수는 직선 관계만 측정한다. (선형 상관)
- 상수 더하기/빼기/양수 곱하기 = 상관계수 (변화 없음)
- 제곱, 로그, 루트, 역수,음수 곱하기 = 상관계수(값이 바뀜)

26. 다음 중 표본오차와 비표본오차에 대한 설명으로 옳지 않은 것은(47회)?

① 표본추출로 비표본 오류를 최소화하거나 없앨 수 있다.

- 표본오차(Sampling Error)
  → 표본통계량(표본평균)과 모수(모평균) 간의 차이.
- 중심극한정리에 의해 표본오차 자체도 분포를 갖는다.
- 중심극한정리는 표본평균의 분포를 통해 표본오차가 평균 0의 정규분포를 따르며, 표본 크기가 커질수록 표본오차가 감소함을 설명한다.
- 중심극한정리(Central Limit Theorem, CLT)란,
  모집단의 분포 형태와 상관없이 표본의 크기가 충분히 크면, 표본평균의 분포는 정규분포에 가까워진다는 이론이다.
- 모집단 분포는 중요하지 않다. → 정규분포가 아니어도 됨
- 표본평균의 분포가 정규분포를 따른다

26. 선형 회귀 모델의 잔차 분석 결과, 잔차 산점도(잔차 vs 예측값)에서 U자형 또는 곡선 패턴이 나타났다. 이 모델의 성능을 개선하기 위한 가장 적절한 조치는(48회)?

① 독립변수와 종속변수 간의 관계를 검토하여 비선형 모델(다항 회귀 등) 도입을 고려한다.

잔차의 곡선 패턴 = "선형성 가정이 위배되었다"는 신호
이때 가장 본질적인 해결책은 모델의 형태 자체를 바꾸는 것이다.
다항 회귀 또는 변수 변환 후 선형회귀

27. 다음 중 가설검정에서 제1종 오류와 제2종 오류에 대한 설명으로 적절하지 않은 것은(47회)?

② 유의수준이 커질수록 귀무가설을 기각할 가능성이 높아져, 대립가설이 채택될 가능성도 높아진다.

- 올바르게 교정한 표현(정답용)
- "귀무가설을 기각한다 / 기각하지 못한다"로만 표현한다.
- ② 유의수준(α)이 커질수록 귀무가설을 기각할 가능성이 높아지므로, 대립가설을 지지하는 결론에 도달할 가능성도 증가한다.
- 가설검정의 오류란

  표본을 기반으로 한 판단이므로, 현실(참·거짓)과 다른 결론을 내릴 수 있는 가능성을 말한다.
- 1종 오류는 귀무가설이 참인데 기각하는 오류이며, 그 확률이 유의수준(α)이다.
- 2종 오류는 귀무가설이 거짓인데 기각하지 못하는 오류
- 검정력(Power)이란 귀무가설이 거짓일 때, 이를 올바르게 기각할 확률이다.
- 검정력 = 실제로 효과가 있을 때, 효과가 있다고 검정이 맞게 판단할 확률

27. 가설검정에 대한 설명이다. 이 중 가장 적절하지 않은 것은(48회)?

② 귀무가설이 참이라는 가정하에 검정통계량의 값이 나타날 가능성이 작을수록 귀무가설을 채택한다.

- 관측 결과가 나타날 가능성이 작으면 귀무가설을 기각할 근거가 커짐 (채택이 아님)

28. 다음 중 가설검정의 오류 해석에 관한 설명으로 옳지 않은 것은(47회)?

① 유의수준(α)은 제1종 오류의 최소 허용 확률이다.

- 가설검정에서 유의수준과 유의확률은 "내가 얻은 데이터가 우연인가, 아니면 정말 의미 있는 결과인가?"를 판단하는 두 가지 핵심 척도입니다.
- 유의수준은 실제 제1종 오류 발생 확률이 아니라,
  연구자가 사전에 설정하는 제1종 오류의 최대 허용 기준이기 때문이다.
- p-value = 귀무가설(H_0)이 참일 때,
  현재 관측된 통계량 이상으로 극단적인 결과가 나올 확률
- p-value < α 가 의미하는 것
- 이 결과는 우연이라고 보기엔 너무 드물다”
- 귀무가설이 참이라면 이런 결과가 나올 확률이 허용 기준보다 작다”
- 귀무가설을 기각할 수 있으며, 결과는 통계적으로 유의하다.

28. 다음 중 가설검정의 오류 해석에 관한 설명으로 옳지 않은 것은(48회)?

① 가설검정에서 사용되는 검정통계량은 항상 하나의 값으로 결정된다.

- 검정통계량은 가설검정의 목적과 조건에 따라 달라진다.
- 예를 들어 평균 차이 검정에서는 t통계량이나 z통계량을 사용할 수 있고, 분산 검정에서는 카이제곱통계량이나 F통계량을 사용할 수 있다.

29.다음 중 탐색적 데이터 분석(EDA: Exploratory Data Analysis)의 특성으로 가장 적절하지 않은 것은(47회)?

③ 동일한 분석 결과를 얻기 위해 재현성(reproducibility)을 검증하는 단계이다.

- 탐색적 데이터 분석(EDA)은 본격적인 모델링 전에 데이터의 구조·특징·패턴·이상치·분포 등을 파악하는 과정이다.
- 변수 단위 조정, 이상치·결측치 탐색 → EDA의 핵심 과정
- 분포·중심·산포 파악 → 요약통계 기반 EDA의 대표 내용
- 그래프 활용하여 직관적 이해 → EDA의 주요 목적

29. 통계적 추론(Statistical Inference)에 대한 설명으로 가장 적절하지 않은 것은(48회)?

③ 구간추정은 모수의 참값이 포함되어 있으리라고 추정되는 구간을 결정하는 것이며, 신뢰수준은 주어진 신뢰구간이 모수를 포함할 확률이다.

- "모수가 이 구간 안에 있을 확률이 95%다" ❌ (모수는 고정값이라 확률이 아님)

30.단순회귀모형에서 SSE가 20이고, 표본 크기가 10일 때 MSE는 얼마인가(47회)?

① 0.20
② 2.50
③ 2.22
④ 5.67

- 회귀분석에서 SSE를 자유도로 나누어 MSE를 계산하는 문제의 의도는 모델의 적합도 평가와 F-검정을 위한 통계적 표준화를 이해시키기 위함
- 모델의 적합도(Model Fit)는 회귀모델이 데이터를 얼마나 잘 설명하고 예측하는지를 평가하는 지표
- SSE는 회귀분석에서 실제 관측값과 예측값 간 차이(잔차)의 제곱을 모두 합한 값입니다.
- MSE = SSE / (n -2), n: 표본 크기, 절편+기울기=2
- 표본크기에 따른 자유도 조정을 통해 서로 다른 모델을 공정히 비교하는 방법

30. sleep 데이터의 두 집단 평균 비교 결과에 대한 설명 중 가장 적절하지 않은 것은(48회)?

```
> t.test(extra ~ group, data = sleep, var.equal = TRUE)

        Two Sample t-test

data:  extra by group
t = -1.8608, df = 18, p-value = 0.07919
alternative hypothesis: true difference in means between group 1 and group 2 is not
95 percent confidence interval:
 -3.363874  0.203874
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

④ 유의수준 1%에서 수면유도제 2가 수면유도제 1보다 평균 수면시간을 유의하게 증가 시킨다고 결론 내릴 수 있다.

- **p-값(0.07919)이 유의수준 1%(0.01)보다 훨씬 크기 때문에, 1% 유의수준에서는 귀무가설을 기각할 수 없다.**

31. 다중회귀모형에서 다중공선성을 해결하기 위한 방안으로 옳지 않은 것은(47회)?

④ 로지스틱 회귀분석(Logistic Regression)

- 다중공선성(multicollinearity)은
- 독립변수들끼리 높은 상관관계를 가져 회귀계수 추정이 불안정해지는 문제를 의미한다.
- 이 문제를 해결하거나 완화하기 위해서는 정규화(regularization) 또는 규제라고하는 사용하는 회귀 기법이 효과적이다.
- 규제(Regularization)란 무엇인가?
- 회귀계수(β)가 너무 커지면 모델이 과적합(overfitting) 발생
- 회귀계수가 커진다는 것은 모델이 데이터의 작은 변동(노이즈)까지 과하게 반응한다.
- L1 규제 (라쏘, Lasso) 회귀계수의 절댓값을 더한 크기에 벌점을 부과
- L2 규제 (릿지, Ridge) 회귀계수의 제곱을 더한 크기에 벌점을 부과

31. 주성분 분석(PCA, Principal Component Analysis)에 대한 설명이다. 이 중 옳지 않은 것은?(48회)?

④ 제1주성분에서 제2주성분으로 갈수록 각 주성분이 설명하는 분산은 증가한다.

제1주성분에서 제2주성분, 제3주성분으로 갈수록 분산은 감소한다"
즉, PC1 ≥ PC2 ≥ PC3 ≥ … ≥ PCp 의 순서로 분산이 정렬됨

32. 다음 중 회귀분석 결과의 해석으로 옳지 않은 것은(47회)?

```
Coefficients:
             Estimate      Std. Error    t value     Pr(>|t|)
(Intercept)  29.394934    6.156303      4.775       5.13e-05 ***
Hp           -0.032230    0.008925      -3.611      0.001178 **
drat          1.615049    1.226983       1.316      0.198755
Wt           -3.227954     0.796398     -4.053      0.000364 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.561 on 28 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-squared: 0.8194
F-statistic: 47.88 on 3 and 28 DF, p-value: 3.768e-11
```

④ drat 변수가 통계적으로 유의하지 않으므로 제거 후 예측모형은

mpg = 29.394934 − 0.032230·hp − 3.227954·wt 이라고 할 수 있다.

- **drat 변수를 제거하면 회귀식을 다시 추정해야 하며, 기존 hp, wt의 계수도 모두 새롭게 계산되어야 한다.**
- **다중회귀에서는 하나의 변수를 제거하면**
  → 나머지 변수들의 계수도 달라지는 것이 정상이다.

32. 회귀분석 모형의 적절성을 진단하기 위한 기준으로 가장 적절하지 않은 것은(48회)?

④ 모형이 데이터에 적합한가? → 설명변수 간 상관계수를 확인한다.

- 모형 적합성: 결정계수(R^2), F-검정, 잔차 분석으로 확인
- 설명변수 간 상관계수는 다중공선성 점검에 사용

33. 다중회귀분석에서 다중공선성(Multicollinearity)에 대한 설명으로 적절하지 않은 것은(47회)?

③ 다중공선성이 존재하더라도 회귀계수의 분산은 항상 같다.

- 다중공선성(Multicollinearity)은 독립변수들 간의 상관관계가 높을 때 발생하며, 회귀계수의 추정값을 불안정하게 만들고 분산을 증가
- 두 개의 독립변수가 서로 너무 비슷하면 모델은 “누가 진짜 영향을 준 것인지” 구분하기 어렵다.
- 그래서 회귀계수를 정확하게 추정할 수 없게 되고 추정값이 조금만 데이터가 변해도 크게 흔들리며(=불안정)
  ➡️ 회귀계수의 분산이 커진다

33. 메이저리그 야구선수 데이터인 Hitters를 이용해 연봉(Salary)을 종속변수의 다중회귀분석 결과에 대한 설명 중 적절하지 않은 것은(48회)?

③ 회귀계수의 크기가 가장 큰 변수는 PutOuts(0.28)이므로, 통계적으로 가장 유의한 설명변수는 PutOuts이다.

```
> fit <- lm(Salary ~ ., data = Hitters)
> summary(fit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  163.10359   90.77854   1.797 0.073622 .
AtBat         -1.97987    0.63398  -3.123 0.002008 **
Hits           7.50077    2.37753   3.155 0.001808 **
HmRun          4.33088    6.20145   0.698 0.485616
Runs          -2.37621    2.98076  -0.797 0.426122
RBI           -1.04496    2.60088  -0.402 0.688204
Walks          6.23129    1.82850   3.408 0.000766 ***
Years         -3.48905   12.41219  -0.281 0.778874
CAtBat        -0.17134    0.13524  -1.267 0.206380
CHits          0.13399    0.67455   0.199 0.842713
CHmRun        -0.17286    1.61724  -0.107 0.914967
CRuns          1.45430    0.75046   1.938 0.053795 .
CRBI           0.80771    0.69262   1.166 0.244691
CWalks        -0.81157    0.32808  -2.474 0.014057 *
LeagueN       62.59942   79.26140   0.790 0.430424
DivisionW   -116.84925   40.36695  -2.895 0.004141 **
PutOuts        0.28189    0.07744   3.640 0.000333 ***
Assists        0.37107    0.22120   1.678 0.094723 .
Errors        -3.36076    4.39163  -0.765 0.444857
NewLeagueN   -24.76233   79.00263  -0.313 0.754218
```

"가장 유의한 변수" 는 p-값(또는 t-값) 으로 판단한다.
*표준화 회귀계수(β)*는 "가장 영향력이 큰 변수" 를 비교할 때 쓰는 도구이다.
➡ "유의성"과 "영향력"은 서로 다른 개념이다.

34. Default 데이터셋을 이용해 소득(income)을 종속변수로 한 다중회귀분석 결과에 대한 설명으로 적절하지 않은 것은(47회)?

```
income1 <- lm(income ~ balance + student + default, data = Default)
summary(income1)
Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  39005.12      150.47     259.2   <2e-16 ***
balance         10.27        0.21      48.5   <2e-16 ***
studentYes   -3835.11      189.55     -20.2   <2e-16 ***
defaultYes   -1525.76      225.32      -6.8   <2e-16 ***
```

④ 독립변수(balance, student, default)는 서로 독립이다.

- 나머지 변수가 동일할 때, 카드 잔액(balance)이 1단위 증가하면 소득이 평균 10.27만큼 증가하는 방향
- student 변수가 Yes인 경우, 평균 소득은 No인 경우보다 낮다
- Yes인 경우 소득이 평균 약 3,835만큼 낮아지는 효과
- default가 Yes인 경우, 평균 소득이 낮게 나타난다.
- 주어진 회귀 결과만으로 변수들 간의 독립성을 알 수 없음

34 다음은 로지스틱 회귀분석(Logistic Regression) 에 대한 설명이다. 이 중 가장 적절하지 않은 것은(48회)?

④ 로지스틱 회귀모형의 연결함수(link function)는 시그모이드(sigmoid) 함수이다.

모형을 만들 때: 0~1 사이의 확률을 → 로짓변환으로 펼쳐서 → 회귀식으로 표현
예측할 때: 회귀식의 결과(실수)를 → 시그모이드로 압축해서 → 0~1 확률로 변환
그래서 시험에서 '연결함수가 무엇이냐' 고 물으면, 모형을 만들 때 사용하는 로짓이 정답이다.

35. College 데이터셋을 이용해 사립학교(Private) 여부에 따른 졸업률(Grad.Rate)을 비교한 회귀분석 결과에 대한 설명으로 올바르지 않은 것은? *(Private: Yes = 사립학교, No = 공립학교)* ***47회***

```
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)     56.042       1.112    50.406    <2e-16 ***
PrivateYes      12.956       1.304     9.937    <2e-16 ***

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

Residual standard error: 16.19 on 775 degrees of freedom
Multiple R-squared: 0.113,     Adjusted R-squared: 0.1119
F-statistic: 98.74 on 1 and 775 DF,  p-value: < 2.2e-16
```

④ 사립학교가 공립학교보다 졸업률이 낮다.

- **사립학교의 계수가 양수이므로 사립학교가 공립학교보다 졸업률이 높다, 낮다는 설명(④)은 잘못된 설명이다**

35. 상관계수에 대한 설명 중 적절하지 않은 것은?*48회*

④ 상관계수의 부호가 음수이면, 두 변수는 반비례 관계라고 해석할 수 있다.

▶ '음의 상관': "X가 늘면 Y가 줄어드는 직선적 경향" → 일정한 비율로 감소
(예: 매일 1시간 더 놀면 점수 5점씩 떨어짐)
▶ '반비례': "Y = k/X" 형태의 곡선 관계 → 비율이 변함 (예: 속도가 2배면 시간이 1/2배)

"상관계수의 부호가 음수이면, 두 변수는 음의 상관(역의 선형 관계) 이라고 해석할 수 있다"
"반비례"는 수학적 정비례/반비례 관계를 뜻하는 다른 개념이다.

36. 다음 중 변수 간의 비선형 관계를 측정할 수 있는 상관계수로 가장 적절한 것은(47회)?

① 스피어만 상관계수(Spearman's rank correlation)

- 피어슨 상관계수(Pearson Correlation)
→ 두 변수 간의 선형(linear) 관계만 측정 가능
→ 곡선 형태(비선형) 관계는 잘 포착하지 못함
- 스피어만 상관계수(Spearman Rank Correlation)
→ 데이터의 순위(rank)를 기반으로 계산
→ 관계가 직선이 아니어도 단조(monotonic) 관계라면 상관성 측정 가능
→ 즉, 비선형 관계도 측정 가능한 상관계수
- 확률변수가 독립이면 공분산은 0이지만, 공분산이 0이라고 해서 반드시 독립인 것은 아니다.
- 공분산이 0이라는 것은 선형 관계가 없다는 뜻일 뿐, 비선형 관계까지 없다는 의미는 아니다.

36. 다음 중 **분해 시계열(decomposition time series)** 을 구성하는 요소로 **적절하지 않은 것은?(48회)?**

① 규칙요인(Regular)

"시계열 분해 4대 요소: T-S-C-I"
추세(T) - 계절(S) - 순환(C) - 불규칙(I)
▪ .

37. 주성분 분석(PCA)에 대한 설명으로 옳지 않은 것은(47회)?

① 제1주성분과 제2주성분은 서로 아무 관계가 없다. 따라서 상관관계가 존재할 수 있다.

- 주성분 분석(PCA)은 고차원의 정량적 데이터를 정보 손실을 최소화하면서 서로 상관관계가 없는 소수의 핵심 변수(주성분)로 요약하는 차원 축소 기법
- 이를 통해 복잡한 데이터 구조를 단순화하여 시각화와 분석의 효율성을 높이고, 변수 간의 다중공선성 문제를 해결
- PCA는 '분산을 최대로 설명하는 직교(orthogonal) 축'을 생성하는 방식이기 때문
- PC1: 데이터 분산을 가장 많이 설명하는 방향
- PC2: PC1과 직교(orthogonal)하면서 그다음으로 분산을 크게 설명하는 방향
- 직교한다 = 내적이 0 = 상관계수가 0 → 서로 독립적인 방향

37. ChickWeight 데이터의 단순선형회귀(Weight ~ Time) 결과에 대한 설명 중 올바르지 않은 것은(48회)?

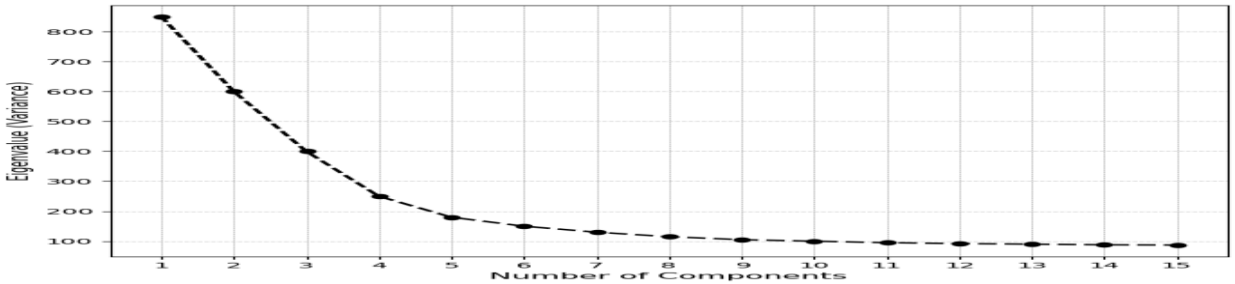
① 회귀식의 절편이 0인지 여부를 뚜렷하게 알 수 없다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.4674     3.0365   9.046   <2e-16 ***
Time          8.8030     0.2397  36.725   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.91 on 576 degrees of freedom
Multiple R-squared:  0.7007,    Adjusted R-squared:  0.7002
F-statistic:  1349 on 1 and 576 DF,  p-value: < 2.2e-16
```

- 회귀계수의 통계적 유의성 검토 시 **절편도 동일한 방식**으로 판단함
- p-값 < 0.05이면 **귀무가설(절편 = 0) 기각**

38. 다음 중 아래에 제시된 주성분 분석(PCA)의 Scree Plot을 통해 유의미한 주성분의 수로 가장 타당한 선택은 (47회)?



① 2 ② 4

③ 6 ④ 8

- Scree Plot은 각 주성분(PC)이 설명하는 고유값(eigenvalue) 또는 분산 기여도를 시각화하여 '몇 개의 주성분을 선택할지'를 판단하는 데 사용하는 그래프이다.
- Scree Plot의 꺾이는 지점(팔꿈치 모양)을 찾아 적절한 주성분 개수를 결정
- 이 지점 이후의 주성분은 설명력이 급격히 떨어짐→ 추가할 필요가 없음

38 다음 중 추세(trend)가 존재하는 비정상 시계열을 정상 시계열로 변환하는 방법으로 가장 적절한 것은(48회)?

② 차분(differencing)

39. 시계열 자료에서 정상성을 판단하기 위한 조건으로 적절하지 않은 것은(47회)?

③ 자기상관성이 시간에 따라 감소한다.

- 시계열 분석에서 정상성(Stationarity)이란 시간이 흐르더라도 시계열 데이터의 통계적 특성이 변하지 않는 상태를 의미
- 즉, 데이터의 평균과 분산이 일정하고, 두 시점 사이의 공분산이 시점 자체가 아닌 '시간 간격'에만 의존하는 성질을 말함
- 시계열 데이터에서 공분산이 시점(Time)'에 의존하지 않고 '시차(Lag)'에만 의존한다는 것은,
  두 데이터 사이의 상관관계가 "언제 측정했느냐"가 아니라 "두 지점 사이의 간격이 얼마냐"에 의해서만 결정된다는 뜻입니다.
- 과거의 통계적 특성이 미래에도 유지된다는 가정하에 예측이 가능해지므로, AR(자기회귀), MA(이동평균) 등의 전통적인 시계열 모델은 정상성을 전제

39. 다음은 머신러닝 모델 개발 시 데이터 분할(학습 · 검증 · 테스트) 에 대한 설명이다. 옳지 않은 것은(48회)?

③ 검증 데이터는 최종 모델의 성능을 평가하기 위해 사용된다.

- 검증 데이터는 하이퍼파라미터 튜닝과 모델 선택에 사용되며, 최종 모델의 성능 평가는 테스트 데이터가 담당한다.

40. ARIMA(p, d, q) 모형을 구축하는 과정에 대한 설명으로 옳지 않은 것은(47회)?

① 자기회귀 차수를 먼저 정하고, 차분 차수를 정한다.

- ARIMA 모형 구축 순서
① 시계열 정상성 확인
- 시계열 그래프
② 차분(d) 수행 → 정상 시계열로 변환
- 1차 차분, 필요 시 2차 차분
③ 모형 차수(p, q) 식별
- ACF / PACF 패턴을 이용
- AR(p): PACF 절단, ACF 점진적 감소
- MA(q): ACF 절단, PACF 점진적 감소
- ARMA(p, q): 둘 다 점진적 감소

40. 아래 데이터 마이닝 추진 단계 를 순서대로 나열한 것 중 가장 옳은 것은(48회)?

```
(가) 데이터 준비
(나) 목적 정의
(다) 데이터 마이닝 기법 적용
(라) 데이터 가공
(마) 검증
```

① (가) → (나) → (다) → (라) → (마)

41. 시계열분석에 대한 설명으로 옳지 않은 것은(47회)?

④ 백색소음(white noise) 과정은 비정상 시계열의 한 형태이다.

- 백색소음은 평균과 분산이 일정하고, 모든 시차에서 자기상관이 0인 완전한 정상 시계열이다.
- AR, MA, ARMA 모형은 모두 정상성을 가정하며, 비정상 시계열의 경우 차분을 통해 정상화한 후 ARMA 구조를 적용한 것이 ARIMA 모형이다.
- AR 모형 ; 현재 시점의 값이 과거 시점 값들의 선형결합으로 설명되는 모형(과거 자신의 값에 의존)
- MA 모형: 현재 시점의 값이 과거 오차(백색잡음)의 선형결합으로 설명되는 모형(과거 오차에 의존)

41. 다음은 **의사결정나무(Decision Tree)** 모형에 대한 설명이다. 이 중 **가장 적절하지 않은 것은(48회)?**

④ 반응변수가 연속형 변수인 경우, 의사결정나무 모형은 학습 데이터에 대해 항상 예측 정확도 100%의 모형을 구축할 수 있다.

- **반응변수가 연속형 변수인 경우, 의사결정나무는 회귀나무로 사용되며, 말단 노드의 평균값 등을 이용해 예측한다. 따라서 학습 데이터에 대해서도 항상 완벽한 예측을 보장하는 것은 아니며, 평가는 정확도보다 RMSE, MAE, MSE 등의 회귀 평가지표를 사용한다.**

42. 다음 표는 분류 모형의 혼동행렬(confusion matrix)이다. 민감도(sensitivity)를 구하시오(47회)

	실제값: Positive	실제값: Negative
예측값: Positive	200 (TP)	100 (FP)
예측값: Negative	400 (FN)	300 (TN)

① 0.33

- 민감도(Sensitivity)란
- 실제로 '참(True)'인 것 중에서 모델이 '참(True)'으로 올바르게 예측한 비율
- 200/600=0.33

42. 다음 빈칸에 들어갈 가장 적절한 용어는? (48회)

정밀도(precision)와 재현율(recall)은 한 지표의 값이 높아지면 다른 지표의 값이 낮아질 가능성이 높은 **트레이드오프(trade-off)** 관계를 지니고 있다.
( )는 이러한 효과를 보정하기 위해 정밀도와 재현율의 **조화평균(harmonic mean)** 으로 계산한 지표이다.

① F1-score

43. 인공신경망에서 은닉층(hidden layer)에 노드가 하나 뿐일 때, 출력층과 동일한 기능을 수행하는 활성화 함수로 가장 적절한 것은(47회)?

① Step
② tanh
③ sigmoid
④ ReLU

- 은닉층 노드가 하나 뿐일 때는 그 출력이 최종 결정과 직결되므로, 이진 분류 시 0과 1 사이의 확률값으로 변환해 주는 시그모이드(Sigmoid)를 활성화 함수로 선택
- 활성화 함수는 신경망에 비선형성(Non-linearity)을 부여하여 선형 결합만으로는 해결할 수 없는 복잡한 패턴을 모델이 학습이 가능
- 만약 활성화 함수를 사용하지 않거나 선형 함수만 사용한다면, 층을 아무리 깊게 쌓아도 결국 하나의 선형 모델과 동일해져 딥러닝의 장점을 잃게 됨
- 적절한 활성화 함수 선택은 학습 과정에서 기울기 소실(Vanishing Gradient) 문제를 방지하고 수렴 속도를 결정하므로 모델의 성능과 직결

43. 인공신경망(Artificial Neural Network) 모형에서, 입력층에서 전달된 신호의 가중합을 출력값으로 변환하여 다음 층으로 전달하는 함수는(48회)?

③ 활성화 함수 (Activation Function)

- 은닉층 노드가 하나 뿐일 때는 그 출력이 최종 결정과 직결되므로, 이진 분류 시 0과 1 사이의 확률값으로 변환해 주는 시그모이드(Sigmoid)를 활성화 함수로 선택
- 활성화 함수는 신경망에 비선형성(Non-linearity)을 부여하여 선형 결합만으로는 해결할 수 없는 복잡한 패턴을 모델이 학습이 가능
- 만약 활성화 함수를 사용하지 않거나 선형 함수만 사용한다면, 층을 아무리 깊게 쌓아도 결국 하나의 선형 모델과 동일해져 딥러닝의 장점을 잃게 됨
- 적절한 활성화 함수 선택은 학습 과정에서 기울기 소실(Vanishing Gradient) 문제를 방지하고 수렴 속도를 결정하므로 모델의 성능과 직결

44. 다음 중 소프트맥스(Softmax) 함수의 수식으로 가장 적절한 것은(47회) ?

① $f(x) = \frac{1}{1+e^{-x}}$

② $f(x) = e^{-x}$

③ $f(x) = \frac{e^{xi}}{\sum_{j=1}^{k} e^{xi}}$

- 소프트맥스 함수는 각 범주의 출력값을 지수화하여 상대적 차이를 극대화한 뒤, 전체 합이 1이 되도록 정규화하여 다범주 분류를 위한 사후 확률을 제공
- k개의 클래스가 있을 때, i번째 출력 노드의 값(xi)에 대한 소프트맥스 함수값
- i번째 점수/전체 점수합, 가장 큰 확률이 모델이 선택한 클래스
- Softmax는 이를 확률처럼 해석할 수 있도록 변환 모든 클래스 확률 합 = 1

44. 다음 중 변수 간 상관관계와 분산을 고려하여 거리 측정을 수행하는 방법 (48회) ?

③ 마할라노비스 거리 (Mahalanobis Distance)

45. 앙상블(Ensemble) 분석에 대한 설명으로 옳지 않은 것은(47회) ?

① 배깅(Bagging)은 부트스트랩(Bootstrap)으로 생성된 여러 표본의 결과를 종합하여 예측한다.
② 부스팅(Boosting)은 모든 약한 학습기에 동일한 가중치를 부여하여 모델을 학습한다.
③ 랜덤 포레스트(Random Forest)는 배깅에 무작위성(Randomness)을 추가한 기법이다.
④ 부스팅(Boosting)은 이전 모델의 예측 오류를 다음 모델이 보완하도록 순차적으로 학습한다.

- 배깅은 독립적인 모델들의 평균으로 분산을 줄이며, 랜덤 포레스트는 여기에 데이터 샘플링뿐만 아니라 변수(Feature) 선택까지 무작위성을 추가하여 트리 간 상관관계를 낮추고 일반화 성능을 극대화
- 반면 부스팅은 무작위성보다는 이전 모델의 오차를 순차적으로 보완하는 방식에 집중하여 편향을 줄이고 예측 정확도를 높이는 데 주력
- 배깅은 무작위성을 활용해 개별 모델의 독립성을 확보한 뒤, 이들의 결과를 합쳐 분산을 감소시키는 데 목적이 있습니다.
- 반면 부스팅은 오답을 보완하는 순차적 학습을 통해 모델의 정교함을 높여 편향을 제거하는 데 초점

45. 연관규칙 A → B 에 대한 지지도(Support) 의 계산식으로 올바른 것은 (48회) ?

② (A와 B가 동시에 포함된 거래 수) / (전체 거래 수).

46. 다음 중 '군집 내의 오차 제곱합(SSE, Sum of Squared Errors)'을 최소화하는 방식으로 군집을 수행하는 것은 (47회)?

① 와드 연결법(Ward's linkage)
② 중심 연결법(Centroid linkage)
③ 평균 연결법(Average linkage)
④ 완전 연결법(Complete linkage)

- 오차제곱합(SSE)이란
  각 데이터가 자신이 속한 군집의 중심(centroid)으로부터 떨어진 거리를 제곱하여 모두 합한 값이다.
- 군집 내부에서 데이터들이 얼마나 가까이 모여 있는지
- 군집의 응집도(cohesion)가 얼마나 높은지를 나타내는 지표이다.
- 와드 연결법은 두 군집을 합칠 때 군집 내 오차제곱합(SSE)의 증가량이 최소가 되도록 군집을 병합하는 계층적 군집화

**46. 어느 회사의 고객 데이터는 이름과 성별 두 개 항목으로 구성되어 있다.
각 항목은 10% 의 확률로 결측값(missing value)을 가지며, 두 항목의 결측 발생은 서로 독립이다.
이 데이터에서 두 항목 중 하나라도 결측이 있으면 해당 행을 삭제 하는 완전 케이스 분석(Listwise Deletion)을 수행한다고 할 때, 임의의 1건이 삭제될 확률은 얼마인가 (48회)?**

① 0.19 (19%)

$$P(\text{삭제}) = P(A \cup B) = 1 - P(A^c \cap B^c)$$

독립이므로:

$$P(A^c \cap B^c) = P(A^c) \times P(B^c) = 0.90 \times 0.90 = 0.81$$

따라서:

$$P(\text{삭제}) = 1 - 0.81 = \mathbf{0.19}$$

47. 군집화 기법에 대한 설명으로 옳지 않은 것은(47회)?

① DBSCAN은 밀도 기반 군집화 알고리즘으로, 초기 중심값 설정이 불필요하다.
② 혼합분포모델(GMM)은 K-평균과 유사하게 초기 중심값 설정에 따라 결과가 달라질 수 있다.
③ 자기조직화지도(SOM)는 고차원 데이터를 2차원 공간에 시각화하여 표현할 수 있다.
④ 자기조직화지도(SOM)는 모든 노드에 동일한 수의 데이터가 자동으로 할당된다.

- SOM(Self-Organizing Map)은 경쟁 학습(Competitive Learning) 기반의 비지도학습 알고리즘이다.
- 각 데이터는 가장 유사한 노드(BMU, Best Matching Unit)에 할당된다.
- 노드별 데이터 개수는 균등하지 않으며 데이터 분포와 패턴에 따라 특정 노드에 데이터가 집중될 수 있고,
- 일부 노드는 거의 선택되지 않거나 비어 있을 수도 있다.

47. 다음 중 두 군집을 병합할 때 군집 내 오차 제곱합(SSE; Sum of Squared Errors) 의 증가량을 최소화하는 방향으로 군집을 형성하는 계층적 군집분석 연결법은(48회)?

④와드 연결법 (Ward's Method)

48. 혼합분포모형(Gaussian Mixture Model)에서 잠재변수(latent variable)가 존재하여 최대가능도추정(MLE)을 직접 수행하기 어렵다. 이를 해결하기 위해 사용되는 알고리즘은 무엇인가(47회)?

③ EM 알고리즘(Expectation–Maximization Algorithm)

- 혼합분포모형에서는 데이터가 어느 그룹에서 나왔는지(군집 소속)를 알 수 없다.
- 이 보이지 않는 정보가 잠재변수이다.
- 그런데 최대가능도추정(MLE)은 "이 데이터가 이렇게 나올 확률을 가장 크게 만드는 값은 무엇인가?"를 계산하는 방법
- 누가 어느 그룹 소속인지 모르는 상태에서는 한 번에 바로 계산하기가 어렵다.
- 그래서 EM 알고리즘을 사용한다.
- E-step:
  "이 데이터는 A그룹일 확률 70%, B그룹일 확률 30%쯤 되겠네" 하고 소속을 확률로 가정
- M-step:
  그 확률을 이용해서 평균·분산 등을 다시 계산

48. 정형 데이터 마이닝 기법 중 반응변수(종속변수)가 없는 데이터 의 구조를 파악하는 데 활용되는 비지도학습(Unsupervised Learning) 기법은(48회)?

③ K-평균 군집분석 (K-means Clustering

49. 다음 중 군집분석 기법에 대한 설명으로 적절하지 않은 것은(47회)?

② k-평균 군집분석은 초기 중심값 k의 설정은 군집분석 결과에 영향을 주지 않는다.

- 계층적 군집분석은 초기값을 무작위로 설정하지 않고, 거리와 병합 규칙에 따라 결정적으로 군집을 형성하므로 실행할 때마다 동일한 결과가 나온다.
- 계층적 군집분석: 초기값 없음 → 결과 항상 동일
- K-평균: 초기 중심 랜덤 → 결과 달라질 수 있음

49. 다음 거래 데이터에서 연관규칙 **A → B**의 신뢰도(Confidence)는 얼마인가**(48회)?**

거래 항목	거래 수
{A}	100
{C}	100
{A, C}	200
{B, D}	100 + 200 = 300 ⭐
{A, B, D}	250
{A, B, C, D}	50
전체 (N)	1,000

② 50%

신뢰도는 **선행항 A가 포함된 거래 중에서 후행항 B도 함께 포함된 거래의 비율**이다.

$$\text{신뢰도}(A \rightarrow B) = \frac{A\text{와}B\text{를모두포함한거래수}}{A\text{를포함한거래수}}$$

먼저 **A를 포함한 거래**는 다음과 같다.

$$[A] + [A, B, C, D] + [A, B, D] + [A, C]$$

$$100 + 50 + 250 + 200 = 600$$

다음으로 **A와 B를 모두 포함한 거래**는 다음과 같다.

$$[A, B, C, D] + [A, B, D]$$

50. 아래는 품목별 거래 내역을 통해 우유 → 커피의 지지도를 구하시오(47회).

거래품목	거래 건수
우유	20
커피	50
우유·커피	30
전체 거래수	100

① 0.3 ② 0.4

③ 0.5 ④ 0.6

50. 연관규칙 분석에서 **향상도(Lift)** 에 대한 설명으로 가장 적절한 것은?**(48회)**.

① Lift = Confidence(A→B) / P(B)