2020년 A형 - 인쇄용 (50문제)

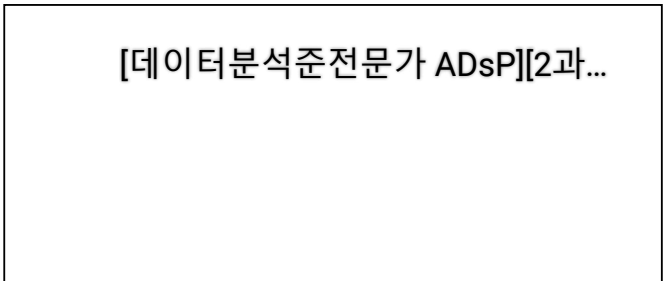
총점 48/50

✓ 1. 다음 빅데이터 거버넌스에 대한 설명으로 옳지 않은 것은? 1/1

- ○ ERD는 운영 중인 데이터베이스와 일치하기 위하여 철저한 변경관리가 필요하다
- ◉ 양질의 데이터가 중요하므로 수명주기보다 데이터 품질관리가 중요하다 ✓
- ○ 산업 분야별, 데이터 유형별, 정보 거버넌스 요소별로 구분하여 작성한다
- ○ 분석 조직 및 인력에 대해 지속적인 교육과 훈련을 실시해야 한다

의견 보내기

2-31. 빅데이터 거버넌스의 특징
*빅데이터 분석에서 품질관리도 중요하지만, 데이터 수명주기 관리방안을 수립하지 않으면 데이터 가용성 및 관리 비용 증대 문제에 직면할 수 있다*https://youtu.be/wekFi3Cl7FM



✓ 2. 다음 중 빅데이터 활용 테크닉 관계가 올바르게 짝지어 진 것은? 1/1

- 회귀분석 : 사용자의 만족도가 충성도에 어떤 영향을 미치는가?
- 연관규칙학습: 호텔에서 고객의 논평을 받아 서비스를 개선하기 위해 활용함
- 유전알고리즘: 응급실에서 의사를 어떻게 배치하는 것이 가장 효율적인가?
- 감정분석: 우유구매자가 기저귀도 같이 구매하는가?

- ○ 회귀분석, 연관규칙학습
- ○ 감정분석, 유전알고리즘
- ○ 감정분석, 연관규칙학습
- ◉ 회귀분석, 유전알고리즘 ✓

의견 보내기

1-15 (빅데이터 활용 기법)
연관규칙학습 : 변수간 주목할 만한 상관관계가 있는지 찾아내는 방법
감정분석 : 특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석함
소셜 미디어에 나타난 의견을 바탕으로 고객이 원하는 것을 찾아낼 때 활용
(2) 감정분석, (4) 연관규칙 학습

✓ 3. 다음 중 데이터베이스 설계 절차를 나열한 것으로 올바른 것은? 1/1

- ◉ 요구조건분석 - 개념적 설계 - 논리적 설계 - 물리적 설계 ✓
- ○ 개념적 설계 - 논리적 설계 - 요구조건분석 - 물리적 설계
- ○ 논리적 설계 - 개념적 설계 - 요구조건분석 - 물리적 설계
- ○ 물리적 설계 – 개념적 설계 – 요구사항분석 - 논리적설계

의견 보내기

1-07. 데이터베이스 설계
요구조건분석 : 데이터베이스 사용자, 사용목적, 사용범위, 제약조건 등을 정리, 명세서 작성
개념적 설계 : 정보를 추상적 개념으로 표현하는 과정, DBMS 독립적 E-R 다이어그램 작성
논리적 설계 : 자료를 컴퓨터가 이해할 수 있도록 특정 DBMS의 논리적 자료 구조로 변환
물리적 설계 : 논리적 구조로 표현된 데이터를 물리적 구조의 데이터로 변환하는 과정

✓ 4. 다음 중 사생활 침해에 따른 문제에 대한 해결책은 무엇인가? 1/1

- 알고리즘 접근 허용
- 결과기반 책임 원칙 강화
- 정보사용자의 책임제로 전환 ✓
- 알고리즈미스트의 육성

의견 보내기

1-16. 빅데이터 위기요인과 통제방안
사생활 침해 -> 동의제를 책임제로 전환
책임원칙의 훼손 -> 기존의 책임원칙 강화
데이터의 오용 -> 데이터 알고리즘에 대한 접근권 허용 및 객관적 인증방안 도입 필요성 제기

✓ 5. 다음 중 데이터베이스 구성요소에 설명 중 올바른 것은? 1/1

- 메타데이터 : 데이터에 관한 구조화된 데이터로, 다른 데이터를 설명해 주는 데이터
- 인덱스 : 데이터베이스 분야에 있어서 테이블에 대한 동작의 속도를 높여주는 자료 구조
- 데이터 사전 : 사용자의 의사결정에 도움을 주기 위해 기간시스템의 데이터베이스에 축적된 데이터를 공통의 형식으로 변환해서 관리하는 데이터베이스
- 어트리뷰트 : 테이블의 행을 의미함

- 메타데이터, 인덱스 ✓
- 어트리뷰트, 인덱스
- 인덱스, 데이터 사전
- 메타데이터, 데이터 사전

의견 보내기

데이터 사전(Data Dictionary) : 자료에 관한 정보를 모아 두는 저장소. 자료의 이름, 표현 방식, 자료의 의미와 사용 방식, 그리고 다른 자료와의 관계 저장
어트리뷰트(Attribute) : 테이블의 열을 의미함
예시의 데이터 사전 설명은 '데이터 웨어하우스에 대한 설명

✓ 6. DIKW 단계를 설명하는 것 중 성질이 다른 것은? 1/1

- A반 학생의 평균 점수는 80점, B반은 82점이다
- B마트의 소고기 600g의 가격은 54000원이다
- B마트의 가격이 더 싸다 ✓
- 개인의 구글 하루 평균 방문 빈도는 10회이다

의견 보내기

1, 2, 4는 데이터, 3은 정보(Information)이다
1-04. 데이터와 정보의 관계
데이터 : 존재 형식을 불문하고, 타 데이터와 상관관계가 없는 가공 전의 수치나 기호
정보 : 데이터의 가공 및 상관/연관 관계 속에서 의미가 도출된 것
지식 : 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물
지혜 : 근본 원리에 대한 깊은 이해를 바탕으로 도출되는 아이디어

✓ 7. 다음 중 빅데이터 본질적인 변화로 옳지 않은 것은? 1/1

- 표본조사에서 전수조사로 변화
- 상관관계에서 인과관계로 변화 ✓
- 사전처리에서 사후처리로의 변화
- 질보다는 양을 중시하게 되었음

의견 보내기

1-13. 빅데이터의 가치 산정, 본질적 변화
사전처리 -> 사후처리
표본조사 -> 전수조사
질(Quality) -> 양(Quantity)
인과관계 -> 상관관계

✓ 8. 다음 중 딥러닝(Deep Learning)과 가장 관련 없는 분석 기법은? 1/1

- LSTM
- Autoencoder
- KNN ✓
- CNN

의견 보내기

머신러닝 관련 분석 기법 : K-NN, LogisticRegression, DecisionTree, Ensemble, SVM
딥러닝 관련 분석 기법 : ANN, DNN, CNN, RNN, LSTM, Autoencoder

✓ 9. DIKW 피라미드 계층구조에서 데이터의 가공 및 상관/연관 관계 속에서 의미가 도출된 것을 무엇이라 하는가? 1/1

정보 ✓

의견 보내기

정보 / Information

1-04. 데이터와 정보의 관계
정보(Information) : 데이터의 가공 및 상관/연관 관계 속에서 의미가 도출된 것
지식(Knowledge) : 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물
지혜(Wisdom) : 근본 원리에 대한 깊은 이해를 바탕으로 도출되는 아이디어

✓ 10. 다음 보기에서 설명하는 빅데이터 활용 기법을 무엇이라 하는가? 1/1

- 구매를 많이 하는 집단에 속하는지 아닌 집단에 속하는지에 대한 문제 해결에 사용함
- 문서를 분류하거나 조직을 그룹으로 나눌 때 사용함

유형분석 ✓

의견 보내기

1-15. 빅데이터 활용 기법
유형 분석 : 사용자는 어떤 특성을 가진 집단에 속하는가? 와 같은 문제 해결에 사용함
문서를 분류하거나 조직을 그룹으로 나눌 때 사용함

✓ 1. 다음 중 CRISP-DM의 모델링 단계의 Task가 아닌 것은 무엇인가? 1/1

- ○ 모델링 기법 선택
- ○ 모델 테스트 계획 설계
- ○ 모델 평가
- ◉ 모델 적용성 평가 ✓

의견 보내기

2-09. CRISP-DM 분석 방법론
모델링 : 모델링 기법 선택, 모델링 작성, 모델 평가
평가 : 분석 결과 평가, 모델링 과정 평가, 모델 적용성 평가
평가가 붙은 것 중에 '모델 평가' 만 모델링에 포함되죠!

✓ 2. 다음 중 분석 과제 정의서에 대한 설명 중 옳지 않은 것은? * 1/1

- 분석 과제 정의서는 분석 모델에 적용될 알고리즘과 분석 모델의 기반이 되는 특성(Feature)이 포함될 필요는 없다.
- 분석 과제 정의서는 프로젝트를 수행하는 이해관계자가 프로젝트의 방향을 설정하지만, 성공여부를 판별에는 사용할 수 없는 자료이다. ✓
- 분석 과제 정의서는 소스데이터, 데이터 입수 및 분석의 난이도, 분석 방법 등에 대한 항목이 포함되어야 한다.
- 분석 과제 정의서는 프로젝트 계획서를 작성하기 위한 중간 결과로써 구성항목으로 도출할 필요가 있다.

의견 보내기

2-15. 분석 프로젝트의 특징 분석 과제 정의서
하향식, 상향식, 디자인 싱킹과 같은 분석 과제 도출 방법을 통해 도출된 분석 과제를 분석 과제 정의서로 정리함
필요한 소스 데이터, 분석 방법, 데이터 입수 난이도, 데이터 입수 사유, 분석 수행주기, 분석 결과에 대한 검증, 분석 과정 상세 등을 작성함
프로젝트 수행 계획의 입력물로 사용됨
이해관계자가 프로젝트의 방향을 설정하고, 성공 여부를 판별할 수 있는 중요한 자료로 명확하게 작성해야 함

✓ 3. 빅데이터 분석 방법론에서 단계 간 피드백이 반복적으로 많이 발생할 수 있는 단계는? *1/1

- 분석 기획 - 데이터 준비
- 데이터 준비 - 데이터 분석 ✓
- 시스템구현 - 평가 및 전개
- 평가 및 전개 – 분석 기획

의견 보내기

2-10-3. 데이터 분석 단계
데이터 준비 - 데이터 분석 :추가적 데이터 확보가 필요한 경우 " 반복적인 피드백을 수행하는 구간"

✓ 4. 다음 중 ROI 관점에서의 분석 과제에 대한 우선순위 평가 기준에 대한 설명 중 적절하지 않은 것은? *1/1

- 시급성이 높고 난이도가 낮은 분석과제를 일반적으로 우선순위가 높다.
- 시급성의 판단 기준은 전략적 중요도 및 목표가치가 핵심이다.
- 난이도는 경영진 또는 실무 담당자의 의사결정에 따라 적용 우선순위를 조정할 수 없다. ✓
- 난이도는 현 시점에서 과제를 추진하는 것이 비용측면과 범위측면에서 바로 적용하기 쉬운 것인지 또는 어려운 것인지에 대한 판단 기준이다.

의견 보내기

2-20. 분석 과제 우선순위 선정 기법
시급성이 높고 난이도가 높은 영역(1사분면)은 경영진 또는 실무 담당자의 의사결정에 따라 적용 우선순위를 조정할 수 있음

✓ 5. 다음 중 데이터 거버넌스 체계 요소 중 데이터 표준 용어 설정, 명명규칙 수립, 메타데이터 구축, 데이터 사전 구축 등의 업무 구성의 요소를 무엇이라 하는가? *1/1

- 데이터 표준화 ✓
- 데이터 관리 체계
- 데이터 저장소 관리
- 표준화 활동

의견 보내기

2-28. 데이터 거버넌스 체계 수립
데이터 표준화 단계 : 데이터 표준용어 설정, 명명규칙 수립, 메타 데이터 구축, 데이터 사전 구축
데이터 관리체계 : 메타데이터와 데이터 사전(Data Dictionary)의 관리 원칙 수립
데이터 저장소관리 : 메타데이터 및 표준 데이터를 관리하기 위한 전사 차원의 저장소를 구성
표준화 활동 : 데이터 거버넌스 체계 구축 후, 표준 준수 여부를 주기적으로 점검, 모니터링

✓ 6. 다음 데이터 유형 중 정형-반정형-비정형의 순서로 올바르게 연결된 것은? *1/1

- (●) Demand Forecast – Competitor Pricing – Email Record ✓
- ( ) Twitter Feeds – ERP – Web Log
- ( ) RFID – CRM – Email Record
- ( ) IoT – Facebook comment – Weather data

의견 보내기

2-05. 데이터 유형, 저장 방식
정형 : ERP, CRM Transaction data, Demand Forecast
반정형 : Competitor Pricing, Sensor, machine data
비정형 : email, SNS, voice, IoT, 보고서, news

✓ 7. 다음 프로토타이핑(Prototyping) 접근 방법에 관한 설명 중 옳은 것은? * 1/1

- (●) 신속하게 해결책 모형 제시, 상향식 접근방법에 활용 한다. ✓
- ( ) 빠른 결과보다 모델의 안정성에 중점을 둔 기법이다
- ( ) 폭포수 방식처럼 전체적인 플랜을 짜고 문서를 통해 개발한다
- ( ) 대표적인 하향식 접근방법이다

의견 보내기

2-07. 분석 방법론의 모델 3가지
프로토타입 모델
사용자 요구사항이나 데이터를 정확히 규정하기 어렵고 데이터 소스도 명확히 파악하기 어려운 상황에서 사용
일단 분석을 시도해보고 그 결과를 확인해가면서 반복적으로 개선해 나가는 방법
신속하게 해결책 모형제시, 상향식 접근방법에 활용

✓ 8. 분석기획 발굴의 범위 확장에 관한 설명 중 적절하지 않은 것은? * 1/1

- 역량의 재해석 관점에서는 현재 해당 조직 및 기업이 보유한 역량 뿐만 아니라 해당 조직의 비즈니스에 영향을 끼치는 파트너 네트워크를 포함한 활용 가능한 역량을 토대로 폭넓은 분석 기회를 탐색한다
- (선택됨) 경쟁자 확대 관점에서는 현재 수행하고 있는 사업 영역의 제품, 서비스에 대해서만 분석 기획 발굴의 폭을 넓혀서 탐색한다. ✓
- 시장의 니즈 탐색 관점에서는 현재 수행하고 있는 사업에서의 고객 뿐만 아니라 고객과 접촉하는 역할을 수행하는 채널 및 고객의 구매와 의사결정에 영향을 미치는 영향자들에 대한 폭넓은 관점을 바탕으로 분석 기회를 탐색한다
- 거시적 관점에서는 현재의 조직 및 해당 산업에 폭넓게 영향을 미치는 사회, 경제적 요인을 STEEP 영역으로 나누어 좀 더 폭넓게 기회 탐색을 수행한다

의견 보내기

2-12-1. 하향식 접근 방식 – 문제 탐색 단계
거시적 관심의 요인: STEEP - 사회, 기술, 경제, 환경, 정치 영역
경쟁자 확대 관점: 대체재 영역, 경쟁자 영역, 신규진입자 영역
시장의 니즈 탐색: 고객(소비자) 영역, 채널 영역, 영향자들 영역
역량의 재해석 관점: 내부역량 영역, 파트너 네트워크 영역

✗ 9. 빈 칸에 들어갈 용어는 무엇인가? * ···/1

분석 과제 관리 프로세스는 크게 과제 발굴과 (1) 으로 나누어진다. 분석 아이디어와 분석 과제가 확정이 되면 팀을 구성하고 (2) 하고 분석 과제 진행 관리 및 결과를 공유하고 개선하는 절차를 수행한다.

과제 수행, 분석 과제 실행 ✗

정답

(1) 과제 수행 (2) 분석 과제 실행

의견 보내기

2-30. 분석 과제 관리 프로세스
과제 발굴 : 분석 아이디어와 발굴, 분석 과제 후보 제안, 분석 과제 확장
과제 수행 : 팀 구성, 분석 과제 실행, 분석 과제 진행 관리, 결과 공유/개선

✓ 10. 합리적인 의사결정을 방해하는 요소로써 문제의 표현 방식에 따라 동일한 사건이나 상황임에도 불구하고 사람들의 선택이나 판단이 달라지는 현상을 무엇이라 하는가? *1/1

프레이밍 효과 ✓

의견 보내기

2-06. 분석 방법론 개요
기업의 합리적 의사결정 장애요소 : 고정관념, 편향된 생각, 프레이밍 효과(Framing Effect)

✓ 1. 두 집단의 분산이 같은지를 검정할 때 사용되는 검정 통계1.량은 어떤 분포를 활용하는 것이 가장 적절한가? *1/1

- ○ t-분포
- ◉ F-분포 ✓
- ○ z-분포
- ○ 카이제곱 분포

의견 보내기

3-61. 모수적 추론(inference)
모평균과 표본평균과의 차이 : z-분포, t-분포
모분산과 표본분산과의 차이 : F-분포(집단 2개), 카이제곱(χ^2)분포(집단 1개)

✓ 2. 다음 중 데이터의 집중경향치와 산포도에 관한 설명으로 틀린 것은? * 1/1

- (●) 중앙값은 대표적인 집중경향치로 이상값 및 다른 관측값에 의한 영향에 민감하다는 단점이 있다. ✓
- ( ) 평균은 데이터의 총 값을 총 개수로 나눈 값을 의미한다.
- ( ) 분산은 관측값에서 평균을 뺀 값을 제곱하고, 그것을 모두 더한 후 전체 개수로 나누어 구한다.
- ( ) 최빈값은 가장 많이 관측되는 수로, 주어진 자료에서 평균, 중앙값을 구하기 어려운 경우 특히 유용하다.

의견 보내기

3-42. 집중화 경향 측정
집중경향치(평균, 중앙값, 최빈값) 에서 이상값 및 다른 관측값에 의한 영향에 민감한 것은 " 평균 " 이다.

✓ 3. 자료의 척도에 대한 설명으로 부적절한 것은? * 1/1

- ( ) 명목척도 : 단순히 측정 대상의 특성을 분류하거나 확인하기 위한 목적으로 사용된다.
- ( ) 서열척도 : 대소 또는 높고 낮음 등의 순위만 제공할 뿐 양적인 비교는 할 수 없다.
- ( ) 등간척도 : 순위를 부여하되 순위 사이의 간격이 동일하여 양적인 비교가 가능하다.
- (●) 비율척도 : 절대 영점이 존재하지 않으며, 측정값 사이의 비율 계산이 가능한 척도이다. ✓

의견 보내기

3-41. 척도의 종류
비율척도(Ratio scale) 절대 0 점이 존재하여 측정값 사이의 비율 계산이 가능한 척도
등간척도(구간척도) : 순위를 부여하되 순위 사이의 간격이 동일하여 양적인 비교가 가능, 절대 0 점 존재하지 않음 (온도계 수치, 물가지수)

✓ 4. 다음은 어느 문구점의 판매품목 및 거래수에 대한 결과이다. 연필 --> 지 우개에 대한 향상도는? *1/1

항목	거래수
연필	200
지우개	300
노트	200
연필, 노트	100
지우개, 연필	150
연필, 지우개, 노트	50
전체 거래건수	1000

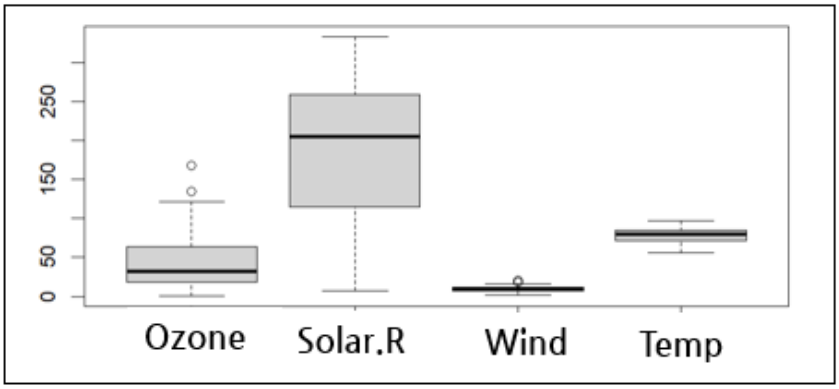
- (●) 0.8 ✓
- ( ) 3.3
- ( ) 0.2
- ( ) 1.0

의견 보내기

3-99. 연관규칙 측정지표

*향상도 = P(B|A)/P(B) = P(A∩B) / (P(A)*P(B))*
*(0.2) / (0.5 * 0.5) = 0.2 / 0.25 = 0.8*

✓ 5. 다음 airquality 데이터셋에 대한 설명 중 적절하지 않은 것은? * 1/1



- ○ Solar.R 중위수가 가장 크
- ○ Solar.R 분산이 가장 크다고 할 수 있다
- ○ Ozone은 이상값이 2개가 존재한다
- ◉ Temp의 데이터 수가 가장 많음을 알 수 있다 ✓

의견 보내기

3-26. 그래프 종류 – Boxplot
Boxplot은 데이터 수에 대한 정보를 표현하지 않는다

✓ 6. 오분류표를 활용하여 모형을 평가하는 지표 중 실제값이 True인 것에 대해 예측값이 True인 것을 무엇이라 하는가? *1/1

- ○ Accuracy
- ○ F1
- ○ Precision
- ◉ Sensitivity ✓

의견 보내기

3-91. 오분류표를 활용한 평가 지표
정밀도(Precision) : 예측값이 True 인 것에 대해 실제값이 True 인 지표
재현율, 민감도(Sensitivity) : 실제값이 True 인 것에 대해 예측값이 True 인 지표
Accuracy : 전체 예측에서 옳은 예측의 비율=> (TP + TN) / (TP + FP + FN + TN)
*F1 : 오분류표 중 정밀도와 재현율의 조화평균을 나타내며 정밀도와 재현율에 같은 가중치를 부여하여 평균한 지표 => 2 * (Precision * Recall) / (Precision + Recall)*

✓ 7. 다음 중 모델의 학습 과정에서 학습 및 검증데이터를 나눌 때, 단순히 한 번 나누는 게 아니라 K 번 나누고 각각의 학습 모델의 성능을 비교하여 평균값으로 분류분석 모형을 평가하는 방법을 무엇이라 하는가? *1/1

- (●) K-fold ✓
- ( ) 홀드아웃
- ( ) 붓스트랩
- ( ) 군집분석

의견 보내기

3-90. 모형 평가
K-fold 교차검증: 주어진 데이터를 가지고 반복적으로 성과를 측정하여 그 결과를 평균한 것으로 분류 분석 모형의 평가 방법
홀드아웃: 원천 데이터를 랜덤하게 두 분류로 분리하여 교차검정을 실시하는 방법
붓스트랩: 평가를 반복하는 측면에서 교차검증과 유사하지만, 훈련용 자료를 반복 재선정한다는 점에서 차이가 있는 평가 방법

✓ 8. 다음 두 점의 맨해튼 거리는 무엇인가? 1/1

A(10, 6) B(3, 4)

- ( ) 12
- ( ) 5
- (●) 9 ✓
- ( ) 10

의견 보내기

3-94. 계층적 군집의 거리
맨해튼 거리는 절대값의 합으로 구한다
|10-3|+|6-4|=9

✓ 9. 다음 시계열 데이터에 대한 설명 중 가장 옳지 않은 것은? * 1/1

- 시계열 데이터의 모델링은 다른 분석 모형과 같이 탐색 목적과 예측 목적으로 나눌 수 있다.
- 짧은 기간 동안의 주기적인 패턴을 계절변동이라 한다.
- (●) 잡음은 무작위적인 변동이지만 일반적으로 원인은 알려져 있다. ✓
- 시계열분석의 주목적은 외부인자와 관련해 계절적인 패턴 추세와 같은 요소를 설명할 수 있는 모델을 결정하는 것이다.

의견 보내기

본문 내용에 없었음
잡음은 무작위적인 변동이며 일반적으로 원인은 알려져 있지 않다.

✓ 10. 다음 주성분 분석 결과에 대한 설명 중 올바르지 않는 것은? * 1/1

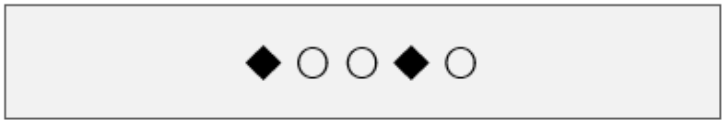
```
> fit <-prcomp(USArrests, scale=TRUE)
> summary(fit)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.5749 0.9949 0.59713 0.41645
Proportion of Variance 0.6201 0.2474 0.08914 0.04336
Cumulative Proportion  0.6201 0.8675 0.95664 1.00000

> fit$rotation
                PC1        PC2        PC3         PC4
Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

- ○ 주성분 분석에 대한 결과로 첫 번째 주성분으로 전체 데이터 분산의 62%를 설명할 수 있다
- ◉ 두 번째 주성분은 4개의 변수 모두와 양의 관계에 있음을 보여준다 ✓
- ○ 변수들이 선형결합으로 이루어진 주성분은 서로 독립이며 기존 자료보다 적은 수의 주성분들로 기존 자료의 변동을 설명한다.
- ○ 위 주성분 분석에서는 상관행렬을 사용하였다

의견 보내기

3-74. 주성분 분석(PCA)
두 번째 주성분과 UrbanPop, Rape는 음의 관계가 있음을 알 수 있다.

✓ 11. 아래 표는 불순도 측정 결과이다. 지니 지수는 얼마인가? * 1/1



- (●) 12/25 ✓
- ( ) 13/25
- ( ) 3/5
- ( ) 2/5

의견 보내기

3-82. 의사결정나무 모형

지니지수 식 : 1 - ∑(각 범주별수/전체수)2
*= 1 − ((2/5)2 + (3/5) 2) = 1 − (**4/25+9/25**)=**12/25***

✓ 12. 데이터마이닝의 상품에 관한 이해를 증가시키기 위한 것으로 데이터의 특징 및 의미를 표현 및 설명하는 기능을 무엇이라 하는가? *1/1

- ( ) 분류(Classification)
- ( ) 예측(Forecast)
- ( ) 추정(Estimate)
- (●) 기술(Description) ✓

의견 보내기

3-80. 대표적 데이터 마이닝 기법
분류 : 새롭게 나타난 현상을 검토하여 기존의 분류, 정의된 집합에 배정하는 것
추정 : 주어진 입력 데이터를 사용하여 알려지지 않은 결과의 값을 추정하는 것
예측 : 미래에 대한 것을 예측, 추정하는 것을 제외하면 분류나 추정과 동일한 의미
기술 : 데이터가 가진 특징 및 의미를 단순하게 설명하는 것

✓ 13. 다음 중 신뢰구간에 대한 설명 중 가장 적절하지 않은 것은? * 1/1

- ○ 구간추정은 모수의 참값이 포함되어 있으리라고 추정되는 구간을 결정하는 것이다
- ○ 관측치의 크기가 커지면 신뢰구간의 길이는 줄어든다
- ◉ 신뢰수준 95%란 신뢰구간에 미지의 수가 포함되지 않을 확률이 95%를 의미한다 ✓
- ○ 점추정의 정확성을 보완하는 방법이다

의견 보내기

3-59. 통계적 추정(Estimation)
신뢰수준 95% 의미는 실제 모수값이 신뢰구간에 존재할 확률이 95%라 할 수 있다

✓ 14. 다음은 Default 데이터를 사용한 고객에 대한 체납 여부(default)에 대한 분석 결과에 대한 설명으로 옳지 않은 것은? *1/1

```
> model = glm(default ~ ., data=Default, family=binomial)
> summary(model)

Call:
glm(formula = default ~ ., family = binomial, data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

default -> 고객의 채무불이행 확률
studentYes : 학생여부(Yes/No)
balance : 채무잔액, income : 연수입

- balance는 default에 통계적으로 유의미한 영향을 주는 변수이다
- 로지스틱 회귀분석을 사용했다.
- (●) income은 default에 통계적으로 유의미한 영향을 주는 변수이다 ✓
- studentYes의 값이 Yes 일 때, 채무불이행(default) 될 확률이 낮다.

의견 보내기

통계적으로 유의미한 영향을 주기 위해서는 0.05 보다 작은 p-value를 가지고 있어야 한다

✓ 15. 다음과 같은 방법을 무엇이라 하는가? * 1/1

데이터 전처리 방법 중 데이터를 일정범위로 Feature scaling 범위 0~1사이로 적용해주고 원 데이터의 분포를 유지하는 정규화 방법

- Robust Normalization
- (●) Min-Max Normalization ✓
- Standardization
- Smooting

의견 보내기

3-71. 스케일링
Min Max Nomalization : 값의 범위를 [0, 1]로 변환하는 것
Standardization : 평균 0, 표준편차 1 인 표준 정규분포로 변환하는 것

✓ 16. 다음 중 결측값(missing value) 처리에 대한 대치법(imputation)에 관한 설명으로 틀린 것은? * 1/1

- complete case analysis는 불완전 자료는 모두 무시하고 완전하게 관측된 자료만으로 표준적 통계기법에 의해 분석하는 방법을 말한다.
- 평균대치법(mean imputation)은 관측 또는 실험되어 얻어진 자료의 적절한 평균값으로 결측값을 대치해서 불완전한 자료를 완전한 자료로 만든 후, 완전한 자료를 마치 관측 또는 실험되어 얻어진 자료라 생각하고 분석하는 방법을 말한다.
- 단순확률대치법(single stochastic imputation)은 평균대치법에서 추정량 표준오차의 과소추정 문제를 보완하고자 고안된 방법이다.
- (●) 다중대치법은 추정량의 과소추정이나 계산의 난해성 문제를 보완하는 방법이다. ✓

의견 보내기

3-36. 결측치 대치법
다중대치법
단순 대치법을 한 번이 아닌 m 번 수행하여 m 개의 가상적 완전 자료를 만듦
추정량 표준오차의 과소추정 또는 계산의 난해성 문제를 가지고 있음

✓ 17. 다음 중 앙상블 모형이 아닌 것은? * 1/1

- ○ 배깅(bagging)
- ○ 랜덤 포레스트(Random Forest)
- ◉ 시그모이드(sigmoid) ✓
- ○ 부스팅(boosting)

의견 보내기

3-83. 앙상블(Ensemble) 모형
배깅(Bagging), 랜덤 포레스트(Random Forest), 부스팅(Boosting)
Sigmoid 는 로지스틱 회귀에서 비선형적 값을 얻기 위해 사용하는 Logistic 함수이고, 인공 신경망에서 사용되는 activation 함수이다.

✓ 18. 계층적 군집의 응집형 군집 방법과 관련이 없는 것은? 1/1

- ○ 최단연결법
- ○ 와드연결법
- ◉ k-중앙값군집 ✓
- ○ 평균연결법

의견 보내기

3-94. 계층적 군집(Hierarchical Clustering)
계층적 군집 - 응집형(병합 군집) 군집 방법
최단연결법, 최장연결법, 중심연결법, 와드연결법, 평균연결법

✓ 19. 귀무가설이 사실일 때 기각하는 1종 오류 시 우리가 내린 판정이 잘못 되었을 때 실제 확률은? *1/1

- ○ 유의수준
- ◉ p-value ✓
- ○ 기각역
- ○ 대립가설

의견 보내기

3-60. 가설검정 - 2
유의수준 : Significance level, 제 1종 오류의 최대 허용 한계
유의확률(p-value) : 귀무가설이 사실일 때 기각하는 1종 오류 시 우리가 내린 판정이 잘못 되었을 확률
기각역 : 가설검정에서 가설을 통해 얻어진 통계량이 어떤 범위 r 내에 들어오면 이 가정은 옳지 않다고 판단하며, 이때의 범위 r을 가리키는 말
대립가설 : 귀무가설이 기각되었을 때 채택되는 가설

✓ 20. 표본조사에 대한 설명이 부적절한 것은? * 1/1

- ○ 표본 오류(Sampling Error)는 모집단을 대표하지 못하는 표본을 추출하여 발생하는 오류이다.
- ○ 표본편의(Sampling Bias)는 표본추출방법에서 기인하는 오차를 의미한다.
- ○ 표본편의는 확률화(Randomization)에 의해 최소화하거나 없앨 수 있다.
- ◉ 비표본오차(non-sampling error)는 표본크기가 증가함에 따라 감소한다. ✓

의견 보내기

비표본오차는 학습 내용에 없었음
비표본오차 : 표본오차를 제외한 조사의 전체과정에서 발생할 수 있는 모든 오차로 표본의 크기에 비례하여 커지므로 표본의 크기가 크다고 반드시 좋은 것만은 아니다

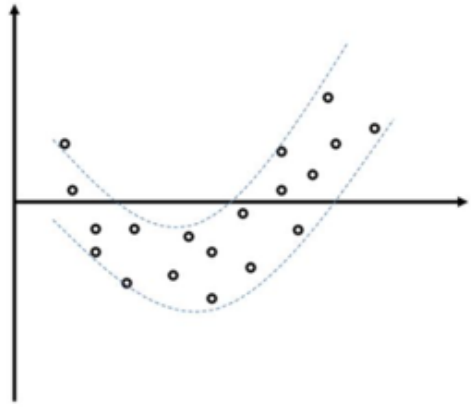
✓ 21. 독립변수간 상관관계가 높아 많은 문제점을 발생하는 현상으로 회귀계수의 분산을 증가시켜 불안정하고 해석하기 어렵게 만들게 되는 것을 무엇이라 하는가? 1/1

- (●) 다중공선성 ✓
- ( ) 공분산행렬
- ( ) 후진제거법
- ( ) 전진선택법

의견 보내기

3-66. 다중 공선성
모형의 일부 설명변수(=예측변수)가 다른 설명변수와 상관되어 있을 때 발생하는 조건
중대한 다중공선성은 회귀계수의 분산을 증가시켜 불안정하고 해석하기 어렵게 만들기 때문에 문제가 됨

✓ 22. 회귀분석의 잔차 분석 결과가 다음과 같을 때 해결 방안은? * 1/1



- (●) x^2 항을 모형에 추가 ✓
- ( ) 데이터를 좀 더 수집함
- ( ) 변수 제거
- ( ) 중요 변수 선택

의견 보내기

학습내용에는 없음
U자는 비선형을 의미하므로 제곱항을 모형에 추가해 본다
이미지 출처 : https://slidesplayer.org/slide/11708770/

https://colab.research.google.com/drive/1QDuCKk86lKTP8oxOTwOo1D2935Tu-5-e#scrollTo=9Sw_Jmm1Dylh 함께 학습해 두세요! 잔차에 대한 해석

✓ 23. Q1, Q3가 다음과 같을 때 이상값 판단을 위한 하한 및 상한은 얼마인가? 1/1

Q1= 4, Q3 = 12

- (●) 하한 = -8, 상한 = 24 ✓
- ( ) 하한 = 24, 상한 = -8
- ( ) 하한 = -8, 상한 = 16
- ( ) 하한 = -4, 상한 = 16

의견 보내기

IQR = 3사분위수 - 1사분위수 = 12 − 4 = 8
상한 : Q3 + 1.5 IQR = 12 + 1.5 * 8 = 24*
*하한 : Q1 − 1.5 * IQR = 4 − 1.5 * 8 = -8*

✓ 24. 연관분석에 대한 설명이 아닌 것은? * 1/1

- ( ) 연관규칙(Association rule)은 항목들 간의 '조건-결과' 식으로 표현되는 유용한 패턴이다
- ( ) 장바구니 분석이라고도 한다.
- ( ) Apriori 알고리즘과 FP Growth 가 대표적이다.
- (●) 연관규칙 측정지표에는 정확도, 정밀도, 민감도 등이 있다 ✓

의견 보내기

3-98. 연관분석(Association Analysis)
연관규칙 측정지표에는 지지도, 신뢰도, 향상도가 있다

✓ 25. 다음 중 다중회귀분석에서 종속변수(Fertility)에 가장 유의한 변수는? * 1/1

```
> temp <- lm(Fertility~., data=swiss)
> summary(temp)

Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,    Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

Education ✓

의견 보내기

3-65. 회귀모형 해석(평가방법)
p-value가 작을 수록 유의한 변수이다

대소문자 구분해서 적어주세요!

✓ 26. 아래 오분류표를 이용하여 특이도 값을 구하시오. * 1/1

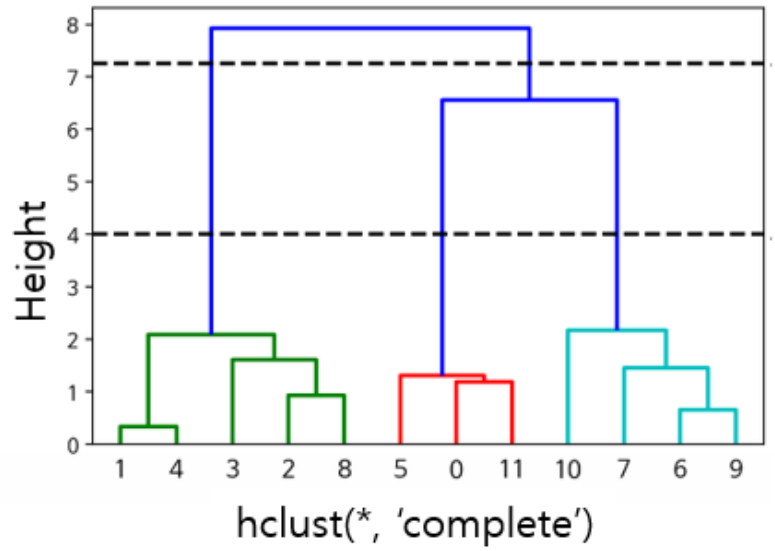
Confusion matrix		예측값	
		TRUE	FALSE
실제값	TRUE	40	60
	FALSE	60	40

0.4 ✓

의견 보내기

3-91. 오분류표를 활용한 평가 지표
특이도(Specificity) : 실제로 N 인 것들 중 예측이 N으로 된 경우의 비율
특이도 : TN / (TN + FP) = 40 / (40 + 60) = 0.4

✓ 27. 다음은 덴드로그램의 결과이다. Height가 4일 때 군집의 수는? * 1/1



3 ✓

의견 보내기

학습 내용에 없음
계층적 군집 분석 관련 내용임
이미지 : https://data-newbie.tistory.com/25

https://colab.research.google.com/drive/1QDuCKk86lKTP8oxOTwOo1D2935Tu-5-e#scrollTo=RHN1l6Jk2uhl 의 덴드로그램도 확인하세요. (주의 사항 포함)

✗ 28. 로지스틱 회귀분석에서 exp(x1) 의 의미는 x1이 한 단위 증가할 때마다 성공의 ( )가 몇 배 증가하는지 나타낸다. 빈 칸에 알맞은 용어는? *···/1

오즈 ✗

정답

오즈(odds)

odds

✓ 29. 시계열분석에서 시계열의 수준과 분산에 체계적인 변화가 없고, 주기적 변동이 없다는 것으로 미래는 확률적으로 과거와 동일하다는 것을 의미하는 용어는? *1/1

정상성 ✓

의견 보내기

3-75. 시계열 자료(time series)
시계열 자료의 정상성 - 시계열분석에서 시계열의 수준과 분산에 체계적인 변화가 없고, 주기적 변동이 없다는 것으로 미래는 확률적으로 과거와 동일하다는 것을 의미

✓ 30. 연관규칙 분석기법의 대표적 알고리즘으로 가장 빈번한 항목 집합을 찾기 위한 접근 방식으로, 이해하기 쉽고, 전체 데이터를 스캔한다. 1/1

Apriori ✓

의견 보내기

3-98. 연관분석(Association Analysis)
Apriori
연관규칙의 대표적 알고리즘으로 현재도 많이 사용됨
데이터들에 대한 발생 빈도를 기반으로 각 데이터 간의 연관관계를 밝히는 방법
데이터셋이 큰 경우 모든 후보 itemset에 대해 하나하나 검사하는 것이 비효율적임
FP Growth
Apriori 단점을 보완하기 위해 FP-tree와 node, link라는 특별한 자료 구조를 사용