2023년 36회 기출복원 인쇄용 (49문항)

총점 49/49

✓ **1. 다음 데이터에 대한 설명으로 가장 적절하지 않은 것은?** 1/1

- 추론, 예측, 전망, 추정을 위한 근거(basis)로 기능하는 특성을 갖는다.
- 데이터는 축적된 지식과 아이디어가 결합된 창의적인 산물이다 ✓
- 데이터는 개별 데이터 자체로는 의미가 중요하지 않은 객관적인 사실이다.
- 다른 객체와의 상호 관계 속에서 가치를 갖는다

의견 보내기

1-01. 데이터의 정의
- 데이터는 개별 데이터 자체로는 의미가 중요하지 않은 객관적인 사실(fact)
- 추론, 예측, 전망, 추정을 위한 근거(basis)로 기능하는 특성을 갖음
- 다른 객체와의 상호 관계 속에서 가치를 갖음
2번의 설명은 '지혜(Wisdom)'에 대한 것

✓ **2. 빅데이터가 가치 창출 측면에서 기업, 정부, 개인에게 미친 영향으로 옳지 않은 것은?** 1/1

- 산업 전체의 생산성이 향상되었다
- 개인은 맞춤형 서비스를 받거나 적시에 필요한 정보를 얻음으로써 기회비용을 절약하게 되었다.
- 기업활동의 투명성은 없어지지만 경쟁사보다 강한 경쟁력을 확보하는데 도움이 되었다 ✓
- 비즈니스 모델을 혁신하거나 신사업 발굴에 활용할 수 있게 되었다

의견 보내기

1-14. 빅데이터의 영향
- 기업: 비즈니스 모델 혁신, 신사업 발굴, 기업활동의 투명성 제공, 경쟁력 확보, 산업 전체의 생산성 향상, GDP 상승 효과
- 정부: 환경탐색, 상황분석, 미래 대응, 사회 변화를 추정, 각종 재해 관련 정보를 추출할 수 있음
- 개인: 맞춤형 서비스를 저렴한 비용으로 이용

✓ **3. 사생활 침해 문제를 해결하기 위한 방법으로 가장 적절한 것은 무엇인가?** 1/1

- (●) 개인정보 사용자 책임제로 전환 ✓
- ( ) 결과기반 책임 원칙 고수
- ( ) 알고리즘 접근 허용
- ( ) 사용자 동의제도 시행

의견 보내기

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

✓ **4. 암묵지와 형식지 상호작용의 과정 중 개인의 내재된 경험을 객관적인 데이터로 변환하여 문서나 매체에 저장·가공·분석하는 과정을 무엇이라고 하는가?** 1/1

- (●) 표출화 ✓
- ( ) 연결화
- ( ) 내재화
- ( ) 공통

의견 보내기

1-03. 암묵지와 형식지 상호작용
- 공통화 : 암묵적 지식 노하우를 다른 사람에게 알려주는 것
- 표출화 : 암묵적 지식 노하우를 책이나 교본 등 형식지로 만드는 것
- 연결화 : 책이나 교본(형식지)에 자신이 알고 있는 새로운 지식(형식지)를 추가하는 것
- 내면화 : 만들어진 책이나 교본(형식지)를 보고 다른 직원들이 암묵적 지식(노하우)을 습득

✓ 5. 다음 비식별화 기법에 대한 설명으로 틀린 것은? 1/1

- ○ 가명처리는 식별할 수 없는 다른 값으로 대체를 의미한다
- ◉ 데이터 마스킹은 개인 정보 식별이 가능한 특정 값을 삭제하는 것이다. ✓
- ○ 범주화는 단일 식별 정보 해당 그룹의 대푯값으로 변환을 한다.
- ○ 총계처리는 총합 또는 평균값으로 대체하여 개별 데이터의 값이 보이지 않도록 하는 것이다.

의견 보내기

1-17. 개인 정보 비식별화 기법
데이터 범주화 : 홍길동, 35세 -> 홍씨, 30~40세
데이터 마스킹 : 카드 뒤 4자리 숨기기, 주민등록 번호 뒤 6자리 숨기기
총계처리 : 데이터의 총합 값을 보여 개별 데이터의 값이 보이지 않도록 함

✓ 6. 데이터베이스에 대한 설명으로 적절하지 않은 것은? 1/1

- ○ 한 조직의 다수 사용자가 공동으로 이용하고 유지하는 공용데이터이다.
- ○ DBMS 소프트웨어를 사용하여 데이터베이스를 구축한다.
- ○ 법률적으로 데이터베이스는 기술을 기반으로 한 일종의 저작물로 인정한다.
- ◉ 데이터베이스내의 모든 데이터는 2차원 테이블로 표현된다. ✓

의견 보내기

- RDBMS의 경우 데이터베이스 내의 데이터가 2차원 테이블로 표현되기도 하지만
- ODBMS, NoSQL의 종류는 테이블 이외에 객체를 기반으로 한 계층 구조, Key와 Value를 사용한 방법도 존재함
- "모든" 데이터가 2차원 테이블로 표현되는 것은 아님

✓ **7. 데이터웨어하우스에 대한 설명으로 가장 적절하지 않은 것은 무엇인가?** 1/1

- ○ ETL은 주기적으로 내부 및 외부 데이터베이스로부터 정보를 추출하고 정해진 규약에 따라 정보를 변환한 후에 데이터웨어하우스에 정보를 적재한다
- ◉ 데이터웨어하우스는 전사적 차원보다는 특정 조직의 특정 업무 분야에 초점을 둔 것이다. ✓
- ○ 데이터웨어하우스에서 관리하는 데이터들은 시간적 흐름에 따라 변화하는 값을 유지한다.
- ○ 데이터웨어하우스는 기업 내의 의사결정 지원 애플리케이션을 위한 정보를 제공하는 하나의 통합된 데이터 저장 공간을 말한다

의견 보내기

1-09. 기업 내부 데이터베이스 솔루션
- 데이터웨어하우스: 사용자의 의사 결정에 도움을 주기 위하여, 다양한 운영 시스템에서 추출, 변환, 통합되고 요약된 데이터베이스이다
- 데이터 마트 : 전사적으로 구축된 데이터웨어하우스로부터 특정 주제, 부서 중심으로 구축된 소규모 단일 주제의 데이터웨어하우스로, 대게 특정 조직 혹은 팀 등 제한된 사용자 그룹에게 서비스가 제공된다.

✓ **8. 다음 데이터 사이언스에 대한 설명으로 가장 부적절한 것은?** 1/1

- ○ 데이터 사이언스란 데이터로부터 의미있는 정보를 추출해내는 학문이다
- ○ 분석 뿐 아니라 이를 효과적으로 구현하고 전달하는 과정까지 포함한 포괄적 개념이다
- ◉ 정형 데이터를 대상으로 총체적 접근법을 사용한다. ✓
- ○ 과학과 인문학의 교차로에 서 있다고 할 수 있다.

의견 보내기

1-21. 데이터 사이언스의 정의
- 데이터로부터 의미 있는 정보를 추출해내는 학문
- 정형, 반정형, 비정형의 다양한 유형의 데이터를 대상으로 함
- 분석 뿐 아니라 이를 효과적으로 구현하고 전달하는 과정까지 포함한 포괄적 개념
- 데이터 공학, 수학, 통계학, 컴퓨터 공학, 시각화, 해커의 사고방식, 해당 분야의 전문 지식을 종합한 학문 => 총체적(holistic) 접근법을 사용함
- 과학과 인문학의 교차로에 서 있다고 할 수 있음 => 스토리텔링, 커뮤니케이션, 창의력, 직관력 필요

✓ **9. 다음 빈 칸에 알맞은 단어는 무엇인가?** 1/1

( )는 거래정보를 하나의 덩어리로 보고 이를 차례로 연결한 거래장부다. 기존 금융회사의 경우 중앙 집중형 서버에 거래 기록을 보관하는 반면, ( )는 거래에 참여하는 모든 사용자에게 거래 내역을 보내주며 거래 때마다 이를 대조해 데이터 위조를 막는 방식을 사용한다.

블록체인 ✓

✓ **10. 다음이 설명하는 것은 무엇인가?** 1/1

인간의 개입을 최소화하여 인터넷 기반으로 모든 사물을 연결하여 상호 소통하는 지능형 기술로 허기스의 tweet pee, 구글의 google glass, 나이키의 fuel band, 삼성의 갤럭시 워치 등을 예로 들 수 있다.

IoT ✓

✓ **2-1. 다음 중 기업의 데이터 분석 도입의 수준 진단의 대상으로 가장 적절하지 않은 것?** 1/1

- ○ 분석 업무 파악
- ○ 분석 기법
- ○ 분석 인력 및 조직
- ◉ 분석 성과 평가 ✓

의견 보내기

2-24. 데이터 분석 수준 진단
- 분석 준비도 : 기업의 데이터 분석 도입의 수준을 파악하기 위한 진단방법, 6가지 영역을 대상으로 현 수준을 파악함
- 6가지 영역 : 분석 업무 파악, 인력 및 조직, 분석 기법, 분석 데이터, 분석 문화, IT 인프라(= 분석 인프라)

✓ **2-2. 분석 성숙도 모델 구성에서 고려하는 분석 성숙도 진단 부문으로 적절하지 않은 것은?** 1/1

- ○ 비즈니스 부문
- ◉ 기업 문화 부문 ✓
- ○ 조직의 역량 부문
- ○ IT 부문

의견 보내기

2-24. 데이터 분석 수준 진단
분석 준비도
- 기업의 데이터 분석 도입의 수준을 파악하기 위한 진단방법
- 분석 업무 파악, 인력 및 조직, 분석 기법, 분석 데이터, 분석 문화, IT 인프라(=분석 인프라)의 6가지 영역을 대상으로 현 수준을 파악함

분석 성숙도
- 시스템 개발 업무능력과 조직의 성숙도 파악을 위해 CMMI 모델을 기반으로 분석 성숙도를 평가함
- 비즈니스 부문, 조직/역량 부문, IT 부문을 대상으로 성숙도 수준에 따라 도입, 활용, 확산, 최적화 단계로 구분해 살펴 볼 수 있음

✓ **2-3. 다음 분석과제의 특징 중 Accuracy와 Precision에 대한 설명으로 틀린 것은?** 1/1

- ◉ 분석의 활용적인 측면에서는 Precision이 중요하며, 안정적인 측면에서는 Accuracy가 중요하다. ✓
- ○ Accuracy와 Precision의 관계는 트레이드 오프가 되는 경우가 많다.
- ○ Accuracy는 모델과 실제 값의 차이에 대한 것이다.
- ○ Precision은 모델을 반복했을 때의 편차를 의미한다.

의견 보내기

2-16. 분석 과제의 주요 5가지 특성 관리 영역 중 Accuracy, Precision
Accuracy : 분석의 활용적인 측면 (모델과 실제 값의 차이)
Precision : 분석의 안정성 측면 (모델을 반복했을 때의 편차)
Accuracy, Precision은 트레이드 오프인 경우가 많음
모델의 해석 및 적용 시 사전에 고려해야 함

✓ 2-4. 다음 중 분석 대상은 명확하지만 분석 방식이 명확하지 않은 경우 수행하는 분석 주제의 유형은 무엇인가? 1/1

- (●) 솔루션(Solution) ✓
- ( ) 통찰(Insight)
- ( ) 최적화(Optimization)
- ( ) 발견(Discovery)

의견 보내기

2-02. 분석 주제 유형 4가지
- Optimization : 분석 대상 및 분석 방법을 이해하고 현 문제를 최적화의 형태로 수행함
- Solution : 분석 과제는 수행되고, 분석 방법을 알지 못하는 경우 솔루션을 찾는 방식으로 분석 과제를 수행함
- Insight : 분석 대상이 불분명하고, 분석 방법을 알고 있는 경우 인사이트 도출
- Discovery : 분석 대상, 방법을 모른다면 발견을 통해 분석 대상 자체를 새롭게 도출함

✓ 2-5. 분석 과제 정의서에 대한 설명으로 가장 적절한 것은 무엇인가? 1/1

- ( ) 프로젝트 수행 계획 수립 단계에서 전체 업무를 분류하여 구성 요소로 만든 후 각 요소를 평가하고 일정별로 계획하며 그것을 완수할 수 있는 사람에게 할당해주는 역할을 한다.
- ( ) 분석 모델에 적용될 알고리즘과 분석모델의 기반이 되는 Feature가 포함되어야 한다.
- ( ) 이해관계자가 프로젝트의 방향을 설정하고, 성공 여부를 판별할 수 없는 자료이다.
- (●) 필요한 소스 데이터, 분석 방법, 데이터 입수 난이도, 분석 과정 상세 등의 항목이 포함되어야 한다. ✓

의견 보내기

2-15. 분석 과제 정의서(SOW, Statement Of Work)
- 다양한 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, 분석결과에 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
- 1) WBS(Work breakdown statement) 에 대한 설명이다.

✓ **2-6. 분석 마스터 플랜을 수립할 때 적용 범위 및 방식에 대한 고려요소가 아닌 것은 무엇인가?** 1/1

- (●) 투입 비용 수준 ✓
- ( ) 분석 데이터 적용 수준
- ( ) 업무 내재화 적용 수준
- ( ) 기술 적용 수준

의견 보내기

분석 마스터 플랜 수립 시 고려 요소
- 우선순위 고려 요소 : 전략적 중요도, ROI(투자자본수익률), 실행 용이성
- 적용 범위/방식 고려 요소 : 업무 내재화 적용 수준, 분석 데이터 적용 수준, 기술 적용 수준

✓ **2-7. 분석 마스터플랜의 세부 이행계획 수립 시 고려해야 할 데이터 분석 체계에 대한 설명으로 적절한 것은?** 1/1

- ( ) 분석 마스터플랜의 모든 단계를 반복한다.
- (●) 프로젝트의 세부 일정계획도 데이터 분석체계를 고려하여 작성한다. ✓
- ( ) 순차적인 정련 과정을 통해 프로젝트의 기간을 단축하는 방식을 주로 사용한다.
- ( ) 데이터 수집 및 확보와 분석 데이터를 준비하는 단계를 반복적으로 진행한다.

의견 보내기

2-21. 이행계획 수립
- 분석 마스터 플랜의 모든 단계를 반복하기보다 데이터수집 및 확보와 분석 데이터를 준비하는 단계를 순차적으로 진행하고, 모델링 단계는 반복적으로 수행하는 혼합형을 많이 적용함
- 반복적인 정련 과정을 통해 프로젝트의 성능을 높이는 방식을 주로 사용함
- 데이터 분석체계의 특징을 고려한 세부적인 일정계획을 수립해야 함

✓ **2-8. 분석 기획에 대한 설명으로 적절하지 않은 것은 무엇인가?** 1/1

- ○ 해당 문제 영역에 대한 전문성 역량 및 통계학적 지식을 활용한 분석 역량과 분석 도구인 데이터 및 프로그래밍 기술 역량에 대한 균형 잡힌 시각을 가지고 방향성 및 계획을 수립해야 한다
- ○ 성공적인 분석을 하기 전 중요 사전 작업이다.
- ◉ 상향식 분석은 분석 기획 전 탐색적 데이터 분석 수행을 한다 ✓
- ○ 실제 분석을 수행에 앞서 분석을 수행할 과제의 정의 및 의도했던 결과를 도출할 수 있도록 이를 적절하게 관리할 수 있는 방안을 사전에 계획하는 일련의 작업이다.

의견 보내기

2-01. 분석 기획의 정의 및 특징
어떤 목표(what)를 달성하기 위해 어떤 데이터를 가지고 어떤 방식(how)으로 수행할지에 대한 일련의 계획을 수립하는 작업

2-10-3. 데이터 분석 단계
탐색적 데이터 분석은 분석 기획 단계의 내용이 아니라 데이터 분석 단계에 포함된 내용임

✓ **2-9. 빅데이터의 4V는 빅데이터의 3V에 무엇이 추가된 것인가?** 1/1

Value ✓

의견 보내기

3V : Volume, Variety, Velocity
4V : 3V + Value

✓ **2-10. 다음 빈칸에 공통으로 들어갈 단어로 알맞은 것은 무엇인가?** 1/1

분석 과제 우선순위 평가 기준에는 ( )과 난이도가 있으며 ( )의 경우 전략적 중요도와 목표가치, 난이도의 경우 데이터 획득/저장/가공 비용, 분석 적용 비용, 분석 수준에 따라 판단하게 된다.

시급성 ✓

✓ 3-1. 다음 중 군집분석 기법으로 적절하지 않은 것은 무엇인가? 1/1

- PAM
- DBSCAN
- 실루엣 지수(Silhouette Coefficient) ✓
- 퍼지(Fuzzy) Clustering

의견 보내기

군집분석 기법의 종류
- Fuzzy clustering : 분할적 군집의 프로토타입 기반의 군집 분석 중 한가지로 Soft Clustering 이라고도 하며, 관측치가 여러 군집에 속할 수 있으며 이를 각 군집에 속할 가능성(possibility), 확률(probability)로 제시해 줌
- PAM(Partitioning Around Medoids) : K-means와 유사하지만 총비용을 계산해 이 값이 작을 때만 그룹의 medoid를 바꾸어 준다

✓ 3-2. Wage 데이터셋에 대한 anova 분석 결과 해석의 내용 중 틀린 것은? 1/1

```
> aov <-aov(wage ~ age, data=data)
> summary(aov)
              Df  Sum Sq Mean Sq F value Pr(>F)
age            1  199870  199870   119.3 <2e-16 ***
Residuals   2998 5022216    1675
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- age와 wage에 대한 선형관계를 알 수 없다.
- age와 wage는 양의 상관관계이다 ✓
- 관측치는 3000 개 이다
- MSE는 1675 이다

의견 보내기

- ANOVA 분석은 선형성을 가정하지 않으므로 양의 상관관계인지 알 수 없다.
- 관측치는 2998 + 2 = 3000 이다. (n = df + k + 1, k는 변수의 개수)
- MSE : Mean Sq와 Residuals의 교차 부분에 표기 되어 있으며 1675 이다.

✓ **3-3. 웹 데이터의 수집을 위해 웹페이지의 구조를 분석하여 데이터를 자동으로 수집하는 방법을 무엇이라고 하는가?** 1/1

- ○ FTP
- ◉ 웹 크롤링(Web Crawling) ✓
- ○ Streaming)
- ○ Open API

의견 보내기

데이터 수집 기술
- 웹 크롤링(Web Crawling) : 웹 데이터의 수집을 위해 웹페이지의 구조를 분석하여 데이터를 자동으로 수집하는 방법
- FTP(File Transfer Protocol) : 서버/클라이언트 사이의 파일 전송, TCP/IP 기반 빠른 데이터 송/수신
- Open API : 서비스, 정보, 데이터 등 오픈 된 정보로부터 API를 통해 실시간 데이터를 수집하는 기술 (API : 다수의 함수로 구성, 시스템 간 연동을 통한 실시간 데이터 송/수신)
- Streaming : 음성, 오디오, 비디오 등의 멀티미디어 데이터를 실시간 송/수신 하는 기술

✓ **3-4. 신경망 모형에서 출력값이 여러 개이고 목표치가 다범주인 경우에 사용하는 것으로 각 범주에 속할 사후 확률을(posterior probability) 제공하는 활성화 함수는 무엇인가?** 1/1

- ○ 항등 함수
- ○ ReLU
- ○ sigmoid
- ◉ softmax ✓

의견 보내기

활성화 함수(activation function)
- sigmoid : 결과가 두 개 범주인 경우 사용하는 활성화 함수
- softmax : sigmoid 함수의 일반화된 형태로 결과가 다 범주인 경우 사용, 각 범주에 속할 사후 확률을(posterior probability) 제공하는 활성화 함수
- ReLU : Hidden Layer에서 주로 사용되는 활성화 함수로 음수에 대해 0으로 양수는 그대로 내보냄
- 항등 함수 : 그대로 출력으로 내보내는 활성화 함수

✓ 3-5. 시그모이드 함수의 범위로 알맞은 것은? 1/1

- (●) 0 ~ 1 ✓
- ( ) -1 ~ 1
- ( ) -1 ~ 0
- ( ) 0.5 ~ 1

의견 보내기

sigmoid 함수
연속형 0~1, Logistic 함수라 불리기도 함
선형적인 Multi-Perceptron 에서 비선형 값을 얻기 위해 사용

✓ 3-6. 다음의 수식에 해당하는 데이터 간의 거리 계산 방식은 무엇인가? 1/1

$$\sum_{i=1}^{n} |x_i - y_i|$$

- ( ) 유클리드 거리
- (●) 맨해튼 거리 ✓
- ( ) 민코프스키 거리
- ( ) 마할라노비스 거리

의견 보내기

맨해튼 거리 : 거리 차의 절대값의 합
유클리드 거리 : 거리 차의 제곱의 합에 대한 제곱근

✓ **3-7. 연관분석에 대한 특징 틀린 것은?** 1/1

- ○ 조건반응(if then)으로 표현되는 연관분석의 결과를 이해하기 쉽다.
- ○ 비목적성 분석 기법이다.
- ○ 대표적인 알고리즘으로 Aprior가 있다.
- ◉ 분석을 위한 계산이 복잡하다는 단점이 있다. ✓

의견 보내기

3-98. 연관분석(Association Analysis)
- 너무 세분화된 품목을 가지고 연관규칙을 찾으려면 의미 없는 분석 결과가 도출됨
- 상대적 거래량이 적으면 규칙 발견 시 제외되기 쉬움
- 조건반응(if-then)으로 표현되는 연관 분석의 결과를 이해하기 쉬움
- 강력한 비목적성 분석 기법이며, 분석 계산이 간편함
- 분석 품목 수가 증가하면 분석 계산이 기하급수적으로 증가함

✓ **3-8. 다음 중 표본들이 서로 관련되어 있는 경우, 짝지어진 두 개의 관찰치들의 크고 작음을 +와 –로 표시하여 그 개수를 가지고 두 그룹의 분포 차이가 있는가에 대한 가설을 검증하는 방법은 무엇인가?** 1/1

- ◉ Sign Test ✓
- ○ Chi-Square Test
- ○ ANOVA Test
- ○ 스피어만 상관계

의견 보내기

3-63. 카이스퀘어, 부호 검정
카이스퀘어 검정
한 개 범주형 변수와 각 그룹 별 비율과 특정 상수비가 같은지 검정하는 적합도 검정
각 집단이 서로 유사한 성향을 갖는지 분석하는 동질성 검정
두 개 범주형 변수가 서로 독립인지 검정하는 독립성 검정

부호 검정
표본들이 서로 관련되어 있는 경우, 짝지어진 두 개의 관찰치들의 크고 작음을 +와 – 로 표시하여 그 개수를 가지고 두 그룹의 분포 차이가 있는가에 대한 가설을 검증하는 방법

✓ **3-9. 자료의 척도에 대한 설명으로 적절하지 않은 것은?** 1/1

- (●) 비율척도 – 사칙연산이 모두 가능하고, 혈액형, 학력 등이 해당된다. ✓
- ( ) 구간척도 – 덧셈, 뺄셈이 가능하고 절대 0점을 포함하지 않는 온도가 이에 해당된다.
- ( ) 서열척도 – 연산이 불가능하고 메달과 같이 범주간 순서가 있는 것이 이에 해당된다.
- ( ) 명목척도 - 단순히 측정 대상의 특성을 분류하거나 확인하기 위한 목적으로 사용된다.

의견 보내기

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

✓ **3-10. 소득순위처럼 정규분포가 아닌 오른쪽 꼬리가 긴 분포(Positive skewed)에서 평균과 중앙값의 관계로 알맞은 것은?** 1/1

- ( ) 중앙값이 평균보다 크다
- (●) 평균이 중앙값보다 크다 ✓
- ( ) 평균과 중앙값의 관계에 변화가 없다
- ( ) 평균은 중앙값의 제곱과 같다

의견 보내기

오른쪽 꼬리가 긴 분포 : 중앙값 < 평균
왼쪽 꼬리가 긴 분포 : 중앙값 > 평균
정규 분포 : 중강값 = 평균

✓ **3-11. R에서 숫자형, 문자형, 논리형 벡터를 하나로 합친 벡터를 구성하는 경우 합쳐진 벡터의 형식은 무엇인가?** 1/1

- ○ 숫자형 벡터
- ◉ 문자형 벡터 ✓
- ○ 논리형 벡터
- ○ 데이터프레임

의견 보내기

3-13. R의 데이터 구조-vector
- R에서는 서로 다른 종류 타입을 갖는 벡터를 하나로 합칠 때 문자형 벡터가 포함되어 있으면 문자형 벡터가 됨
- 만일 숫자형 벡터와 논리형 벡터를 하나로 합치면 숫자형 벡터가 됨

✓ **3-12. 다음 중 빅데이터 분석 프로세스에서 모델링 단계에 해당하지 않는 항목은 무엇인가?** 1/1

- ○ 데이터 분할
- ○ 데이터 모델링
- ○ 모델 적용 및 운영 방안
- ◉ 수행방안 설계 ✓

의견 보내기

모델링 단계 : 데이터 분할, 데이터 모델링, 모델 적용 및 운영 방안

✓ **3-13. 다음 중 모형 성과 평가 방법으로 적절하지 않은 것은?** 1/1

- ○ 결정계수
- ○ 실루엣 지수
- ◉ 엔트로피(Entropy) ✓
- ○ ROC 그래프

의견 보내기

회귀 모형 평가 : 결정계수, MAPE, MAE, MSE, MSLE, RMSLE 등
분류 모형 평가 : 오분류표를 활용한 정확도, 정밀도, F1, 특이도, 민감도, 오분류율 등과
ROC 그래프, 이익 도표, 향상도 그래프, Kappa 지수 등
군집 모형 평가 : 실루엣 지수, Dunn Index 등
Entropy – 불순도 측정에 사용함

✓ **3-14. 다음 중 분류 모형 평가에 활용하지 않는 것은 무엇인가?** 1/1

- ◉ 덴드로그램 ✓
- ○ 오분류표
- ○ ROC 그래프
- ○ Kappa 지수

의견 보내기

분류 모형 평가의 종류
오분류표를 활용한 정확도, 정밀도, F1, 특이도, 민감도, 오분류율 등
ROC 그래프, 이익 도표, 향상도 그래프, Kappa 지수 등

✓ **3-15. 다차원척도법에 대한 설명으로 가장 적절하지 않은 것은 무엇인가?** 1/1

- ○ 개체들의 거리는 유클리드(Euclidean) 거리와 유사도를 이용하여 구한다.
- ○ 관측 대상의 상대적 거리의 정확도를 높이기 위해 적합 정도를 스트레스 값(Stress Value)로 나타낸다.
- ○ 스트레스 값은 0에 가까울 수록 적합도가 좋음을 나타낸다.
- ◉ 개체들 사이의 유사성과 비유사성을 측정하여 차원을 축소하기 위해 사용한다. ✓

의견 보내기

3-73. 차원 축소 기법 – 다차원 척도법
개체들 사이의 유사성, 비유사성을 2차원 혹은 3차원 공간상에 점으로 표현하여 개체 사이의 군집을 시각적으로 표현하는 분석방법

✓ **3-16. 앙상블 모형의 특징으로 올바르지 않은 것은?** 1/1

- ○ 성능을 분산시키기 때문에 과대적합(overfitting) 감소 효과가 있다.
- ◉ 각 모형의 상호연관성이 높을수록 정확도 또한 높아진다. ✓
- ○ 여러 개의 모형의 결과를 종합하여 정확도를 높이는 방법이다.
- ○ Bagging, Boosting 등 다양한 방법의 앙상블 기법이 존재한다.

의견 보내기

3-83. 앙상블(Ensemble) 모형
각 모형의 상호연관성이 높을수록 정확도가 떨어진다.

✓ 3-17. 이상치 관련한 설명으로 가장 옳지 않은 것은? 1/1

- DBSCAN 군집을 실행해 군집에 포함되지 않은 것을 이상치로 한다.
- ESD 방법에서는 평균 – 3*표준편차보다 작거나, 평균 + 3*표준편차보다 큰 데이터를 이상치로 규정한다.
- 기하평균을 이용하는 경우 기하평균 – 2.5*표준편차 보다 작거나, 기하평균 + 2.5*표준편차보다 큰 데이터를 이상치로 규정한다.
- (선택됨) IQR을 사용하는 방식의 경우 Q2(중위수) + 1.5*IQR 보다 크거나 Q2(중위수) - 1.5*IQR 작은 데이터를 이상치로 규정한다. ✓

의견 보내기

IQR을 사용하는 이상치 탐색 방법
*- Q1 – 1.5 *IQR 보다 작거나, Q3 + 1.5*IQR 보다 큰 데이터로 규정*

✓ 3-18. 군집분석에 대한 설명으로 적절하지 않은 것은? 1/1

- 집단별 특성이 유사할 경우 안정성이 높다
- 유사성을 이용하여 몇 개의 집단으로 그룹화하는 분석이다
- (선택됨) 군집분석은 집단 간 이질성과 집단 내 동질성이 모두 낮아지는 방향으로 군집을 만든다 ✓
- 비계층적 군집분석 기법의 경우 사용자가 사전 지식 없이 그룹의 수를 정해주는 일이 많기 때문에 결과가 잘 나오지 않을 수 있다.

의견 보내기

군집분석
- 군집분석은 집단 간 이질성과 집단 내 동질성이 모두 높아지는 방향으로 군집을 만든다
- 집단별 특성이 유사할 경우 안정성이 높다
- 유사성을 이용하여 몇 개의 집단으로 그룹화하는 분석이다
- 비계층적 군집분석 기법의 경우 사용자가 사전 지식 없이 그룹의 수를 정해주는 일이 많기 때문에 결과가 잘 나오지 않을 수 있다.

✓ **3-19. 의사결정나무의 특징으로 알맞지 않은 것은?** 1/1

- ○ 연관성이 높은 변수가 있어도 영향을 받지 않는다.
- ◉ 비정상적인 잡음 데이터에 대해서는 민감하게 분류한다. ✓
- ○ 목적 변수가 이산형(범주형)인 경우와 연속형인 경우 모두 사용할 수 있다.
- ○ 설명력이 좋으며, 과대적합에 취약한 특징이 있다.

의견 보내기

3-82. 의사 결정 나무(Decision Tree) 모형
- 상관성이 높은 다른 불필요한 변수가 있는 경우에도 크게 영향을 받지 않음
- 비정상적인 잡음 데이터에 대해 민감하지 않음 : 가지치기(pruning)을 사용하여 잡음에 민감하지 않도록 할 수 있음
- 과대적합(Overfitting)에 취약함
- 설명력이 좋음

✓ **3-20. 데이터마이닝을 위한 데이터 분할과 관련된 설명 중 알맞지 않은 것은?** 1/1

- ○ 데이터는 학습용, 검증용, 평가용 데이터로 분할하여 사용할 수 있다.
- ◉ 검증용 데이터(validation data)는 학습과정에서 사용되지 않는다. ✓
- ○ 검증용 데이터는 훈련에 사용되지 않는다.
- ○ 데이터 수가 적을 때는 교차 검증을 사용한다

의견 보내기

Hold Out
- 데이터를 학습 세트, 검증 세트, 시험(테스트) 세트 세 가지로 분할하여 사용할 수 있음
- Training Data : 학습용 데이터, Test Data : 학습 종료 후 성능 확인용 데이터
- Validation Data : 학습 중 성능 확인용 데이터 (Overfitting 여부 확인, Early Stopping 등을 위해 사용)
- 데이터 수가 적을 때는 Hold Out 보다 교차 검증을 사용함

✓ **3-21. 확률적 표본추출 방법이 아닌 것은?** 1/1

- ○ 단순 무작위 추출법
- ○ 계통 추출법
- ◉ 집단 추출법 ✓
- ○ 층화 추출법

의견 보내기

3-40. 표본추출
- 단순 무작위 추출 : 모집단의 각 개체가 표본으로 선택될 확률이 동일하게 추출되는 경우
- 층화 추출 : 모집단을 서로 겹치지 않게 몇 개의 집단 또는 층(strata)으로 나누고, 각 집단 내에서 원하는 크기의 표본을 단순 무작위추출법으로 추출함
- 군집 추출(=집락 추출) : 모집단을 차이가 없는 여러 개의 집단(cluster)로 나눔, 이들 집단 중 몇 개를 선택 한 후, 선택된 집단 내에서 필요한 만큼의 표본을 임으로 선택함
- 계통 추출 : 모집단 개체에 1, 2,...,N 이라는 일련번호를 부여한 후, 첫 번째 표본을 임의로 선택하고 일정 간격으로 다음 표본을 선택함

✓ 3-22. 코드 실행 결과에 대한 설명으로 적절한 것은? 1/1

```
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
```

- (●) 회귀모형은 유의수준 5%하에서 통계적으로 유의미하다. ✓
- ( ) 모든 회귀계수들이 유의미한다.
- ( ) 설명력은 7.165 이다
- ( ) 데이터의 개수는 41개 이다

의견 보내기

- Examination의 회귀계수는 유의미하지 않음
- 설명력은 70.67%(Multiple R-squared) 임
- 데이터의 개수는 41+ 5+ 1=47

✓ **3-23. 데이터 전처리 과정에 대한 설명으로 올바른 것은 무엇인가?** 1/1

- ○ 결측치는 연산에 아무런 방해가 되지 않으므로 그대로 두어도 무방하다.
- ◉ 데이터 특성을 파악하고 통찰을 얻기 위한 방법을 데이터 EDA라고 한다. ✓
- ○ 모든 분석의 이상치는 시간이 오래 걸리더라도 모두 찾아내어 제거한다
- ○ 데이터 변환을 통해 정규분포 형태의 데이터로 만들면 데이터가 왜곡되어 올바른 학습이 되지 않는다.

의견 보내기

- 모든 분석의 이상치를 모두 제거해야 하는 것은 아니다.
- 결측치를 포함한 데이터는 연산이 불가하므로 반드시 해결해야 한다.
- 데이터 변환을 통해 정규분포 형태로 만드는 것은 학습에 도움이 될 수 있는 방법이다.

✓ **3-24. 변수 가공에 대한 설명으로 적절하지 않은 것은?** 1/1

- ◉ 구간화의 개수가 감소하면 정확도는 높아지지만 속도가 느려진다. ✓
- ○ log, sqrt를 취하면 큰 값을 작게 만들 수 있다 - 오른쪽 꼬리 긴 분포에 사용
- ○ 제곱, exp를 취하면 작은 값을 크게 만들 수 있다 - 왼쪽 꼬리 긴 분포에 사용
- ○ MinMax Normalization을 하면 값이 0 ~ 1 사이의 범위로 변경된다

의견 보내기

- 구간화의 개수가 감소하면 정확도가 낮아지고, 속도가 높아질 수 있다

✓ **3-25. 다음이 설명하는 시계열 모형은 무엇인가?** 1/1

"자기자신의 과거자료로 설명하는 모형으로 백색잡음의 현재값과 자기 자신의 과거값의 가중합으로 선형성을 표현하는 정상시계열 모형이다."

AR 모형 ✓

의견 보내기

3-76. 시계열 모형
- AR(p) : 현 시점의 자료를 p 시점 전의 유한 개의 자기 자신의 과거 값을 사용하여 설명
백색 잡음의 현재 값과 자기 자신의 과거 값의 선형 가중 값으로 이루어진 정상 확률 모형
- MA(q) : 과거 q 시점 이전 오차들에서 현재항의 상태를 추론한다
최근 데이터의 평균을 예측치로 사용하는 방법, 각 과거치는 동일 가중치가 주어짐

✓ **3-26. 다음이 설명하는 앙상블 모형의 종류는 무엇인가?** 1/1

"배깅(bagging)에 랜덤 과정을 추가한 방법으로 노드 내 데이터를 자식 노드로 나누는 기준을 정할 때 모든 예측 변수에서 최적의 분할을 선택하는 대신, 설명변수의 일부분만을 고려함으로 성능을 높이는 방법을 사용한다."

랜덤 포레스트 ✓

✓ **3-27. 다음 오분류표를 사용하여 F1-score 를 구하시오.** 1/1

		예측		
		TRUE	FALSE	합계
실제	TRUE	30	70	100
	FALSE	60	40	100
	합계	90	110	200

6/19 ✓

의견 보내기

F1 score 구하기
Precision : 30 / 90 = 1/3, Recall (=Sensitivity) : 30/ 100 = 3/10
*F1 score = (2 * Precision * Recall)/ (Precision + Recall) = (2 * 1/3 * 3/10) / (1/3 + 3/10) = 6/19*

✓ **3-28. 이산확률변수 X가 가능한 값으로 1, 2, 4가 있다. P(X=1) = 0.3 이고 기댓값이 2.7 일 때 P(X=2)는 무엇인가?** 1/1

0.2 ✓

의견 보내기

이산확률 변수의 기댓값 : $E(X) = \sum x \cdot f(x)$
*식1) 1 *0.3 + 2 * p1 + 4 * p2 = 2.7 # 기댓값 공식에 대입해서 구한 식*
식2) p1 + p2 = 0.7 # X=1 일때가 0.3 이므로 전체 확률의 합 1 − 0.3 = 0.7 이고, p1 + p2 = 0.7 이 됨
p2 = 0.7 − p1 를 식1에 대입

*0.3 + 2 * p1+ 4 * (0.7 − p1) = 2.7*
*0.3 + 2*p1 + 2.8 − 4*p1 = 2.7*
0.4 = 2p1, p1 = 0.2

✓ **3-29. 우등반에 들어가기 위해서는 어느 시험에서 상위 2% 안에 들어야 한다. 해당 시험 점수의 평균이 85점이고 표준편차가 5일 때, 우등반에 들어가기 위한 최소 시험 점수는? (단, P(Z<=2.05)=0.98)** 1/1

95.25 ✓

의견 보내기

*Z = 2.05 인 경우 0.98 확률을 갖게 되므로, 평균 + 표준편차 * 2.05 한 값을 구하면 됨.*
*>> 점수 = 85 + 5 * 2.05 = 95.25*
>> 그림은 영상 참조