2020년 - B형 인쇄용 (50문제)

총점 48/50

✓ 1. 빅데이터의 처리비용을 낮춘 측면에서 결정적 기술로 가장 적절한 것은? 1/1

- ○ 스마트폰의 급속한 확산
- ○ 텍스트마이닝
- ○ 저장장치비용의 지속적인 하락
- ◉ 클라우드 컴퓨팅 ✓

의견 보내기

1-11. 빅데이터 출현 배경
클라우드 컴퓨팅: 빅 데이터 분석에 경제적 효과를 제공해준 결정적 기술

✓ 2. 다음 중 빅데이터 가치 산정에 대한 설명으로 옳지 않은 것은? 1/1

- ◉ 특정 데이터를 언제, 어디서, 누가 활용할지 알 수 있다 ✓
- ○ 빅데이터 시대에는 데이터가 기존에 없던 가치를 창출함에 따라 그 가치를 측정하기 어렵다
- ○ 데이터는 기존 사업자에게 경쟁우위를 제공하기도 한다
- ○ 분석 기술 발달로 지금은 가치 없는 데이터도 새로운 분석 기법의 등장으로 거대한 가치를 만들어 내는 재료가 될 가능성이 있다

의견 보내기

1-13. 빅데이터의 가치 산정
재사용이나 재조합, 다목적용 데이터 개발 등이 일반화되면서 특정 데이터를 언제,어디서, 누가 활용할지 알 수 없다
데이터가 기존에 없던 가치를 창출함에 따라 그 가치를 측정하기 어렵다
분석 기술의 발달로 지금은 가치 없는 데이터도 새로운 분석 기법의 등장으로 거대한 가치를 만들어내는 재료가 될 가능성이 있다

✓ 3. 다음 중 비식별 조치 방법에 대한 설명 중 올바른 것은? 1/1

- 총계처리는 카드 위 4자리를 숨기는 처리를 한 것이다
- 데이터 범주화는 임꺽정 180, 홍길동 170, 이콩쥐 160 에 대해 평균값 170으로 표시한 것이다
- (●) 가명처리는 홍길동, 35세를 임꺽정, 30대로 처리한다 ✓
- 데이터 마스킹은 홍길동, 35세를 홍씨, 30~40세로 처리한다

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
(1) 데이터 마스킹, (2) 총계처리, (4) 데이터 범주화

✓ 4. 일반적으로 데이터는 암묵지와 형식지의 상호작용에 있어 중요한 역할을 한다. 다음 중 암묵지와 형식지의 상호 순화 작용과 관련이 없는 것은? 1/1

- 공통화
- 연결화
- 내면화
- (●) 추상화 ✓

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 데이터 사이언스에 대한 설명 중 옳지 않은 것은? 1/1

- (●) 정형화된 실험데이터만을 분석 대상으로 한다 ✓
- ( ) 데이터로부터 의미 있는 정보를 추출해내는 학문이다
- ( ) 데이터 사이언스는 총체적 접근법을 사용한다
- ( ) 분석 뿐만 아니라 효과적으로 전달하는 과정까지 포함된 포괄적 개념이다

의견 보내기

1-21. 데이터 사이언스의 정의
데이터로부터 의미 있는 정보를 추출해내는 학문
정형, 반정형, 비정형의 다양한 유형의 데이터를 대상으로 함

✓ 6. 다음 빅데이터 활용 기술에 대한 설명 중 올바르지 않은 것은? 1/1

- ( ) 변수간 주목할 만한 상관관계가 있는지 찾아내는 방법은 연관관계분석이다
- ( ) 사용자는 어떤 특성을 가진 집단에 속하는가? 와 같은 문제 해결에 사용하는 것이 유형분석이다
- (●) 소셜 미디어에 나타난 의견을 바탕으로 고객이 원하는 것을 찾아낼 때 활용하는 것은 소셜 네트워크 분석이다 ✓
- ( ) 최근 핀테크 기업에서 소셜 네트워크 분석이 대출을 제공할 때 활용하고 있다

의견 보내기

1-15. 빅데이터 활용 기법

감정분석
특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석하는 것
소셜 미디어에 나타난 의견을 바탕으로 고객이 원하는 것을 찾아낼 때 활용함
호텔에서 고객의 논평을 받아 서비스를 개선하기 위해 활용함

✓ 7. 데이터 사이언티스트의 역량 중 소프트스킬(Soft skill)에 대한 설명이 아닌 것은? 1/1

- ○ 통찰력 있는 분석
- ○ 설득력 있는 전달
- ○ 다분야간 협력
- ◉ 머신러닝에 대한 지식 ✓

의견 보내기

1-23. 데이터 사이언티스트의 역량
데이터 사이언티스트들은 하드 스킬과 소프트 스킬 능력을 동시에 갖추고 있어야 한다
하드스킬 : Machine Learning, Modeling, Data Technical Skill
소프트 스킬 : 통찰력 있는 분석, 설득력 있는 전달, 다분야간 협력

✓ 8. 다음 데이터 사이언티스트가 갖춰야 할 역량 중 성격이 다른 것은? 1/1

- ○ Machine Learning
- ○ Modeling
- ◉ Data Visualization ✓
- ○ Data Technical Skill

의견 보내기

1-23. 데이터 사이언티스트의 역량 – 소프트 스킬
통찰력 있는 분석 : 창의적 사고, 호기심, 논리적 비판
설득력 있는 전달 : Storytelling, Visualization
다분야 간 협력 : Communication

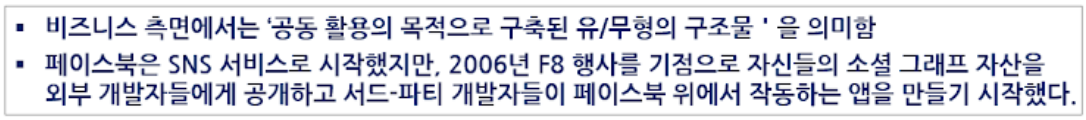
✓ 9. DIKW 피라미드 계층구조에서 데이터의 가공 및 상관/연관 관계 속에서 의미가 도출된 것을 의미하며, 이를 통하여 예측한 결과물을 지식이라고 한다. 1/1

정보 ✓

의견 보내기

1-04. 데이터와 정보의 관계
정보(Information) : 데이터의 가공 및 상관/연관 관계 속에서 의미가 도출된 것
지식 : 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물

✓ 10. 다음 보기에서 설명하는 빅데이터의 역할은 무엇인가? 1/1



플랫폼 ✓

의견 보내기

1-12. 빅데이터의 역할
빅데이터는 "석탄/철, 원유, 렌즈, 플랫폼 " 이다!

✓ 1. 다음 분석 과제 도출에 대한 접근 방법 설명 중 가장 적절하지 않은 것은? 1/1

- ○ 일반적으로 상향식 접근방식은 비지도학습 방법으로 수행된다
- ○ 상향식 접근과 하향식 접근이 반복적으로 수행되는 것은 디자인 씽킹이다
- ○ 문제가 주어진 상태에서 답을 구하는 경우 하향식 접근방식을 사용한다
- ◉ 문제의 정의 자체가 명확한 경우 상향식 접근방식을 사용한다 ✓

의견 보내기

2-11. 분석 과제 도출 방법
문제의 정의 자체가 어려운 경우 상향식 접근 방법을 사용함

✓ 2. 빅데이터 분석방법론에 대한 설명 중 올바르지 않은 것은? 1/1

- ○ 모델링에서는 모델의 과적합 발견과 일반화를 위해 데이터를 분할한다
- ◉ 시스템 구현단계에서 정보보안은 중요한 문제가 아니다 ✓
- ○ 시스템 구현단계는 설계 및 구현, 시스템 테스트 및 운영으로 구성된다
- ○ 프로젝트 위험계획 수립에 대응으로 회피, 전이, 완화, 수용이 있다

의견 보내기

시스템 구현단계에서 정보 시스템 개발방법론에 근거하여 소스코드 보안 약점 진단 및 개선을 진행함

✓ 3. 다음 중 데이터 표준화에 대한 설명으로 올바른 것은? 1/1

- ○ 메타데이터와 데이터 사전의 관리 원칙을 수립한다
- ◉ 데이터 표준 용어 설정, 명명 규칙수립, 메타데이터 구축, 데이터 사전 구축 등의 업무로 구성된다 ✓
- ○ 메타데이터 및 표준 데이터를 관리하기 위한 전사 차원의 저장소를 구성한다
- ○ 데이터 거버넌스 체계를 구축한 후 표준 준수 여부를 주기적으로 점검하고 모니터링을 실시한다

의견 보내기

2-28. 데이터 거버넌스 체계 수립
데이터 표준화 : 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

✓ 4. 다음 중 분석 마스터플랜 수립 시 우선순위 고려요소로 적절하지 않은 것은? 1/1

- (●) 기술 적용 수준 ✓
- ( ) 비즈니스 성과
- ( ) 실행 용이성
- ( ) 전략적 중요도

의견 보내기

2-18. 분석 마스터플랜 수립
분석 마스터플랜 수립 시
우선순위 고려요소
- 전략적 중요도, ROI(투자자본수익률), 실행 용이성

적용 범위/방식 고려요소
- 업무 내재화 적용 수준, 분석 데이터 적용 수준, 기술 적용 수준

✓ 5. 분석기획 발굴의 범위 확장 시 고려해야 하는 사항에 관한 설명 중 적절하지 않은 것은? 1/1

- ( ) 거시적 관점에서는 현재의 조직 및 해당 산업에 폭넓게 영향을 미치는 사회, 경제적 요인을 STEEP 영역으로 나누어 좀 더 폭넓게 기회 탐색을 수행한다
- ( ) 경쟁자 확대 관점에서는 현재 수행하고 있는 사업 영역의 제품, 서비스 뿐만 아니라 대체재와 신규진입자 등으로 확대하여 탐색한다
- (●) 역량의 재해석 관점에서는 파트너와 네트워크 영역은 주로 지적 재산권과 기술, 지식 등 인프라적인 유형 자산을 의미한다 ✓
- ( ) 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

의견 보내기

2-12-1. 하향식 접근 방식 – 분석 기회 발굴을 범위 확장
역량의 재해석 관점: 내부역량 영역, 파트너 네트워크 영역
지적 재산권과 기술, 지식 등 인프라적인 유형 자산을 의미하는 것 ‘내부역량 영역’ 설명

✓ 6. 다음 중 분석 기회 발굴 범위 확장시 경쟁자확대 관점의 영역이 아닌 것은? 1/1

- ○ 대체재 영역
- ○ 경쟁자 영역
- ○ 신규진입자 영역
- ◉ 고객 영역 ✓

의견 보내기

2-12-1. 분석 기회 발굴의 범위 확장

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
교재 160 페이지, 하단의 그림으로 암기하세요!

✓ 7. 분석 프로젝트 관리에 대한 설명 중 가장 적절하지 않은 것은? 1/1

- ○ 분석 모델의 성능 및 속도를 고려한 개발 및 테스트가 수행되어야 한다
- ○ 분석 프로젝트는 다른 프로젝트 유형처럼 범위, 일정, 품질, 리스크, 의사소통 등 영역별 관리가 수행되어야 한다
- ○ 분석하고자 하는 데이터의 양을 고려하는 관리방안 수립이 필요하다
- ◉ 분석 프로젝트는 지속적인 변경으로 인해 일정을 제한하는 계획은 적절하지 못하다 ✓

의견 보내기

2-17. 10개 주제별 프로젝트 관리 체계
분석 프로젝트의 경우 관리 영역에서 일반 프로젝트와 다르게 유의해야 할 요소 존재
시간, 범위, 품질, 통합, 이해관계자, 자원, 원가, 리스크, 조달, 의사소통
시간 프로젝트 활동의 일정을 수립, 일정 통제의 진척 상황 관찰

✓ 8. 다음 중 데이터 분석을 위한 조직 구조 중 집중형 조직 구조의 특징으로 가장 부적절한 것은? 1/1

- 조직내에 별도의 독립적인 분석 전담 조직 구성이다.
- (●) 일반적인 분석 수행구조, 전사적 핵심 분석이 어렵다. ✓
- 분석 전담조직에서 회사의 모든 분석 업무를 담당한다.
- 현업 업무부서의 분석 업무와 이중화 또는 이원화될 가능성이 높다

의견 보내기

2-29. 데이터 분석을 위한 조직 구조
기능중심 조직 구조 : 일반적인 분석 수행구조, 전사적 핵심 분석이 어려움

✓ 9. 데이터에 관한 구조화된 데이터로, 다른 데이터를 설명해 주는 데이터이며, 대량의 정보 가운데에서 찾고 있는 정보를 효율적으로 찾아내서 이용하기 위해 일정한 규칙에 따라 콘텐츠에 대하여 부여되는 데이터를 무엇이라고 하는가? *1/1

메타데이터 ✓

의견 보내기

메타데이터 : 데이터에 관한 구조화된 데이터로, 다른 데이터를 설명해 주는 데이터이며, 대량의 정보 가운데에서 찾고 있는 정보를 효율적으로 찾아내서 이용하기 위해 일정한 규칙에 따라 콘텐츠에 대하여 부여되는 데이터

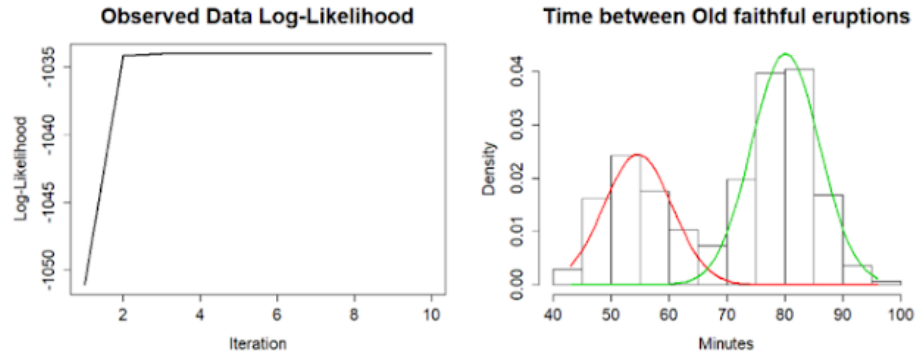
✓ 10. 기업에서 사용하는 데이터의 가용성, 유용성, 통합성, 보안성을 관리하기 위한 정책과 프로세스를 다루며 프라이버시, 보안성, 데이터품질, 관리규정 준수를 강조하는 것을 무엇이라고 하는가? 1/1

데이터 거버넌스 ✓

의견 보내기

데이터 거버넌스
전사 차원의 모든 데이터에 대하여 정책 및 지침, 표준화, 운영조직 및 책임 등의 표준화된 관리체계를 수립하고 운영을 위한 프레임워크 및 저장소 구축을 하는 것을 말함

✓ 1. EM알고리즘을 사용한 혼합분포 모형의 결과 해석에 대한 설명으로 적절한 것은 무엇인가? 1/1



이미지 출처 : https://post.naver.com/viewer/postView.nhn?volumeNo=28935019&memberNo=2534901

- (●) 반복횟수 2회 만에 로그가능도 함수가 최대가 됨을 알 수 있다 ✓
- ( ) 로그 가능도 함수의 최대값은 -1050이다
- ( ) 결과적으로 3개의 정규분포가 혼합된 것을 알 수 있다
- ( ) 모수 추정을 위해 8회 이상의 반복이 필요함을 알 수 있다

의견 보내기

로그가능도 함수의 최대값 : -1035
2개 정규분포 혼합
모수 추정을 위해 2회 반복 필요

✓ 2. 어떤 슈퍼마켓에서 고객 5명의 장바구니 구입품목이 다음과 같다고 한다. 1/1
연관규칙 빵→우유에 대한 신뢰도는?

장바구니	구입품목
1	(빵, 맥주, 과자)
2	(빵, 우유, 계란)
3	(과자, 우유)
4	(빵, 피자)
5	(빵, 우유, 아이스크림)

- (●) 50% ✓
- ( ) 25%
- ( ) 40%
- ( ) 75%

의견 보내기

3-99. 연관규칙 측정지표
신뢰도 = 빵과 우유 동시에 포함된 거래 수 / 빵을 포함하는 거래 수
= 2 / 4 = 0.5

✓ 3. 어떤 슈퍼마켓에서 고객 5명의 장바구니 구입품목이 다음과 같다고 한다. 연관규칙 빵→우유에 대한 향상도는? 1/1

장바구니	구입품목
1	(빵, 맥주, 과자)
2	(빵, 우유, 계란)
3	(과자, 아이스크림)
4	(빵, 우유)
5	(빵, 우유, 아이스크림)

- ○ 1.5
- ◉ 1.25 ✓
- ○ 1.8
- ○ 1.75

의견 보내기

3-99. 연관규칙 측정지표

*향상도 = 빵과 우유 동시에 포함된 확률/ (빵을 포함하는 거래 확률*우유를 포함하는 거래 확률)*

*= P(B|A)/P(B) = P(A∩B) / (P(A)*P(B))*
*= (3/5) / (4/5*3/5) = 0.6 / (0.8 * 0.6) = 1.25*

✓ 4. 다음 중 주성분 분석의 주성분 결정 기준에 대한 설명으로 옳지 않은 것은? 1/1

- ○ 고윳값은 분산의 크기를 나타내며, 고윳값이 1보다 큰 주성분만 사용한다
- ○ 누적 분산 비율이 70 ~ 90%가 되는 주성분 개수를 선택한다
- ○ Scree Plot은 고윳값을 가장 큰 값에서 가장 작은 값을 순서로 정렬해 보여준다
- ◉ 평균 고윳값 방법은 고윳값들의 평균을 구하고, 고윳값이 평균값 이상이 되는 주성분을 제거한다 ✓

의견 보내기

평균 고윳값 방법은 고윳값의 평균값 이상이 되는 주성분을 선택한다.

✓ 5. 다음 회귀분석에 대한 결과 해석 중 옳지 않은 것은? 1/1

```
> step(lm(Fertility~Agriculture+Examination+Education+Catholic+Infant.Mortality,
 swiss), direction='both')
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality

                  Df Sum of Sq    RSS    AIC
- Examination      1     53.03 2158.1 189.86
<none>                         2105.0 190.69
- Agriculture      1    307.72 2412.8 195.10
- Infant.Mortality 1    408.75 2513.8 197.03
- Catholic         1    447.71 2552.8 197.75
- Education        1   1162.56 3267.6 209.36

Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality

                  Df Sum of Sq    RSS    AIC
<none>                         2158.1 189.86
+ Examination      1     53.03 2105.0 190.69
- Agriculture      1    264.18 2422.2 193.29
- Infant.Mortality 1    409.81 2567.9 196.03
- Catholic         1    956.57 3114.6 205.10
- Education        1   2249.97 4408.0 221.43

Call:
lm(formula = Fertility ~ Agriculture + Education + Catholic +
    Infant.Mortality, data = swiss)

Coefficients:
     (Intercept)       Agriculture         Education          Catholic
         62.1013           -0.1546           -0.9803            0.1247
Infant.Mortality
          1.0784
```

- (●) 전진제거법을 이용하였다 ✓
- ( ) 최종 결과의 독립변수는 4개이다
- ( ) 독립변수 중 Examination이 제거되었다
- ( ) 제거 이전보다 제거 후의 AIC의 값이 작아지면 제거한다

의견 보내기

3-67. 설명 변수 선택 방법
step 함수에 direction='both'이므로 단계적 선택법을 사용한 것이다

✓ 6. 다음 One Sample t-test의 결과 해석으로 틀린 것은? 1/1

```
> t.test(x=Default$income, mu=33000)

        One Sample t-test

data:  Default$income
t = 3.8764, df = 9999, p-value = 0.0001067
alternative hypothesis: true mean is not equal to 33000
95 percent confidence interval:
 33255.56 33778.41
sample estimates:
mean of x
 33516.98
```

- ○ 평균이 33000과 같다는 것이 귀무가설이다
- ◉ 관측치의 개수는 9999 이다 ✓
- ○ 귀무가설은 기각되어 대립가설이 채택된다
- ○ 95% 신뢰구간은 33255.56 ~ 33778.41 이다

의견 보내기

관측치의 개수는 df + 1 로 10000 이다

✓ 7. 다음 두 점의 유클리드 거리는 무엇인가? 1/1

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

$\sqrt{17}$

옵션 1

$\sqrt{34}$

옵션 2

$\sqrt{53}$

옵션 3 ✓

$\sqrt{43}$

옵션 4

의견 보내기

3-94. 계층적 군집의 거리
유클리드 거리는 차이의 제곱의 합에 대한 제곱근이다
아래 루트가 있는 것입니다

$\sqrt{((10-3)^2+(6-4)^2)} = \sqrt{(49+4)} = \sqrt{53}$

✓ 8. 다음 중 Sigmoid 함수에 대한 식으로 올바른 것은? 1/1

$$y = \frac{1}{1 + e^{-x}}$$

- $y = (e^x - e^{-x})/(e^x + e^{-x})$)
- $y = 1/(1 + e^{-x})$ ✓
- $y=exp(z^2/2)$
- $y=-1$ 또는 y = 1

의견 보내기

3-81. 로지스틱 회귀분석

$Ysigmoid = 1/(1+e^{-x})) =1/(1+exp(-x))$

✓ 9. 다음이 설명하는 확률적 표본 추출 방법은 무엇인가? 1/1

모집단 개체에 1, 2,⋯,N 이라는 일련번호를 부여한 후, 첫 번째 표본을 임의로 선택하고 일정 간격(k)으로 다음 표본을 선택하는 방식이다.

- 단순 무작위 추출
- 계통 추출 ✓
- 층화 추출
- 군집 추출

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

✓ 10. 다음 시계열 분석에 대한 설명으로 옳지 않은 것은? 1/1

- 현 시점의 자료가 p 시점 전의 유한 개의 과거 자료로 설명될 수 있는 모형은 AR 모형이다
- 비정상 시계열은 차분, 변환을 통해 AR, MA, ARMA 모형으로 정상화 할 수 있다
- (●) 정상성을 만족하지 않는 비정상 시계열 자료는 시계열 분석을 할 수 없다 ✓
- MA 모형은 항상 정상성을 만족한다

의견 보내기

3-76. 시계열 모형
정상성을 만족하지 않는 비정상 시계열 자료를 정상 시계열로 변환 한 뒤 시계열 분석을 할 수 있다.

✓ 11. 두 개 변수, 1000개 Sample로 구성된 데이터에서 결측값을 제거하려고 한다. 결측치 비율이 변수 각각 5%이며, 두 변수가 독립일 때, 삭제되는 데이터 비율은? 1/1

- (●) 9.75% ✓
- 20%
- 2.5%
- 25%

의견 보내기

3-48. 사건의 종류
- 독립사건인 경우 $P(A \cap B) = P(A) \cdot P(B)$ 성립
- 즉, 각 변수를 A, B라고 하고, 결측치 비율을 더한 뒤, 교집합을 제외하면 삭제 데이터 비율을 구할 수 있음
$= P(A) + P(B) - P(A \cap B)$
$= 0.05+0.05 - 0.0025 = 0.0975 * 100 = 9.75\%$

독립사건에 대한 이해를 묻는 문제입니다. (두 변수가 독립일 때 라고 해서요)

✓ 12. 다음 중 지도 학습이며 종속변수가 범주형인 경우 사용되는 것은 무엇인가? 1/1

- (●) 분류분석 ✓
- ( ) 회귀분석
- ( ) 군집분석
- ( ) 연관분석

의견 보내기

지도학습 : 회귀분석(종속변수 - 연속형), 분류분석(종속변수 - 범주형)
비지도학습 : 군집분석, 연관분석

✓ 13. 다음 과대적합에 대한 설명 중 옳지 않은 것은? 1/1

- ( ) 과대적합을 피하기 위해 Ridge, Lasso 등의 규제 모형을 사용할 수 있다
- (●) 학습 데이터(train data)에 최적화 되어 평가 데이터(test data)의 작은 변화에는 민감하게 반응하지는 않는다 ✓
- ( ) 과대적합의 경우 학습 데이터에 대한 성능이 매우 높다
- ( ) 과대적합을 피하는 방법으로 앙상블 방법을 사용할 수 있다

의견 보내기

과대적합(Overfitting)
학습 데이터에 너무 잘 맞게 학습되어 학습 데이터에 대한 성능은 매우 높지만 평가 데이터에 대한 성능은 낮음
규제 모델, 앙상블 등의 방법으로 과대적합을 해결하거나 피할 수 있음
평가 데이터(test data)의 작은 변화에도 민감하게 반응 함

✓ 14. 다음 로지스틱 회귀 모형에 대한 설명으로 옳지 않은 것은? 1/1

- 종속변수가 혈액형, 생존여부 처럼 범주형인 경우 사용한다
- 모형 탐색 방법으로 최대우도법(MLE)를 사용한다
- 종속변수를 전체 실수 범위로 확장하여 분석하고, sigmoid 함수를 사용해 연속형 0~1값으로 변경하며 이는 선형적 값을 얻기 위해 사용한다. ✓
- odds는 성공률/실패율을 의미하는 것으로 log(odds)를 사용해 값의 범위를 전체 실수 범위로 확장한다

의견 보내기

3-81. 로지스틱 회귀분석
종속변수를 전체 실수 범위로 확장하여 분석하고, sigmoid 함수를 사용해 연속형 0~1 값으로 변경
sigmoid 함수는 Logistic 함수라 불리며 y 값을 [0, 1]의 비선형적 값을 얻기 위해 사용함

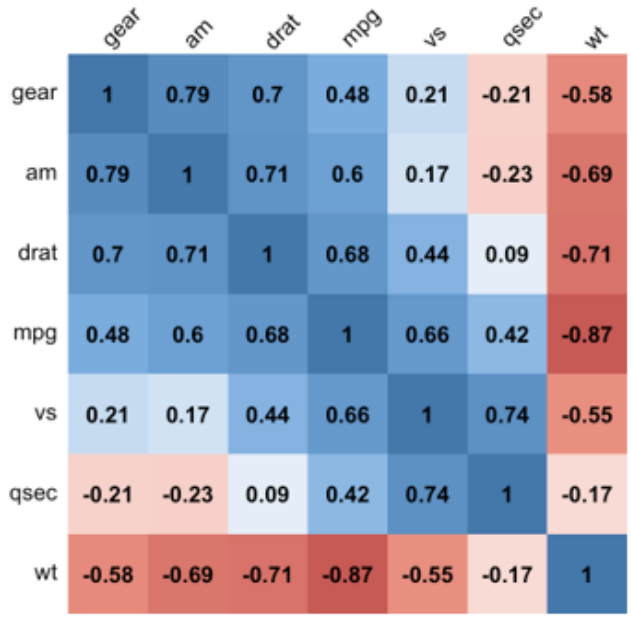
✓ 15. K-평균군집 분석은 군집 개수를 사전에 설정해야 한다. 다음 중 군집 개수 결정에 활용할 수 있는 그래프로 가장 적절한 것은 무엇인가? 1/1

- 실루엣
- 집단 내 제곱합(inertia) ✓
- 덴드로그램
- 히트맵

의견 보내기

K-평균 군집(K-means)
K-means의 경우 inertia를 사용하여 그래프를 그리고 elbow 기법으로 최적의 K를 정하는 것이 일반적인 방법이 됩니다.
- https://scikit-learn.org/stable/modules/clustering.html (2.3.2 K-means, inertia 참조)

✓ 16. 다음 변수간 상관분석의 결과를 그래프로 나타낸 것이다. 이에 대한 설명으로 옳지 않은 것은? 1/1



- wt와 mpg 는 가장 높은 상관 관계를 갖는다
- gear와 wt는 양의 상관 관계가 있다 ✓
- mpg와 drat는 양의 상관 관계가 있다
- drat가 높아지면 wt는 낮아진다

의견 보내기

3-72. 상관 분석
음수는 음의 상관 관계, 양수는 양의 상관 관계를 나타낸다
-1, 1에 가까울 수록 높은 상관 관계이다
양의 상관은 한쪽이 높아지면 다른 쪽도 높아지는 관계이고, 음의 상관은 한쪽이 높아질 때 다른 쪽은 낮아지는 관계이다

✓ 17. 다음 다섯 종류의 오렌지 나무에 대한 summary 결과에 대한 해석으로 틀린 것은? 1/1

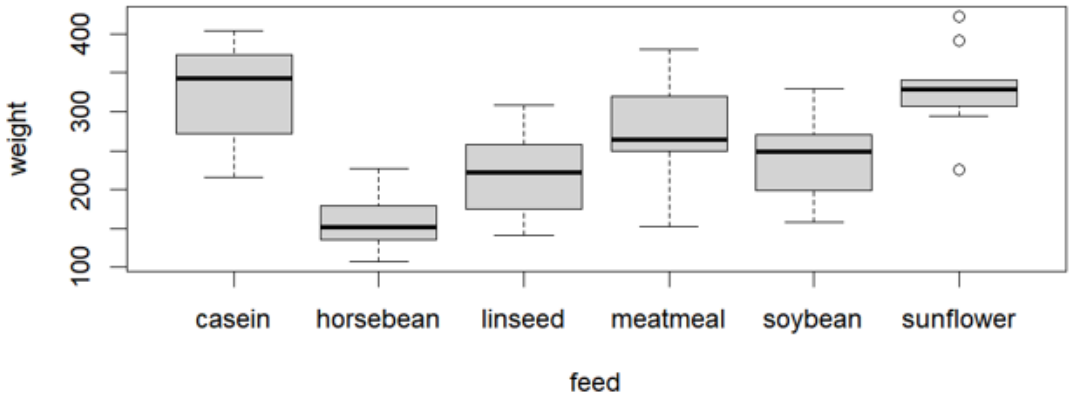
```
> summary(Orange)
 Tree       age         circumference
 3:7   Min.   : 118.0   Min.   : 30.0
 1:7   1st Qu.: 484.0   1st Qu.: 65.5
 5:7   Median :1004.0   Median :115.0
 2:7   Mean   : 922.1   Mean   :115.9
 4:7   3rd Qu.:1372.0   3rd Qu.:161.5
       Max.   :1582.0   Max.   :214.0
```

- circumference의 Median은 115이다
- Tree의 종류는 5가지이며 각 종류당 7개의 sample이 존재한다
- (●) Tree의 종류에 상관 없이 age가 높을 수록 circumference가 큰 것을 알 수 있다 ✓
- age의 IQR은 888이다

의견 보내기

age, circumference 사이의 관계를 알 수 없음

✓ 18. 다음 닭 사료의 종류(feed)와 닭의 성장에 대한 boxplot결과이다. 옳지 않은 것은? 1/1

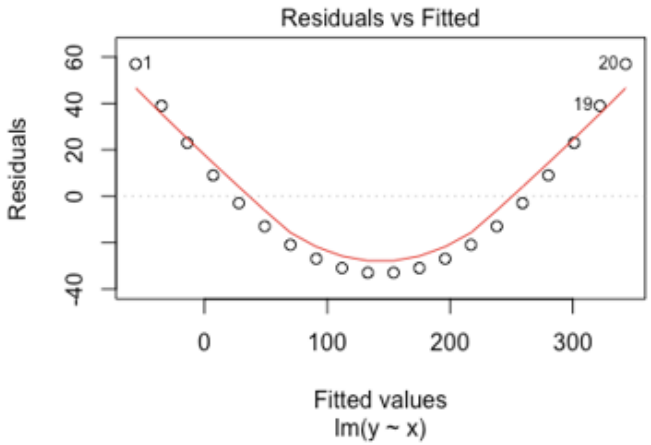


- (●) 이상치가 존재하지 않는 것을 알 수 있다 ✓
- ( ) casein의 경우 horsebean 보다 중위수가 크다
- ( ) soybean의 경우 meatmeal 보다 최소값은 크지만, 최대값이 작다
- ( ) horsebean 사료를 먹은 닭의 무게가 가장 작은 쪽에 분포해 있다

의견 보내기

sunflower의 경우 이상치가 존재한다

✓ 19. 다음 그림은 회귀분석의 가정 중 어떤 것을 위배하고 있다고 판단할 수 있는가? 1/1



- ○ 정상성
- ○ 비상관성
- ○ 독립성
- ◉ 선형성 ✓

의견 보내기

3-64. 회귀 모형의 가정
X와 y의 관계가 비선형이면 잔차도가 비선형인 모습으로 표현되며, 오차는 평균이 0이고 분산이 일정하다는 가정을 만족하지 않고 있으므로, 등분산성과 선형성을 만족하지 않고 있습니다.
이런 경우 제곱항을 추가하거나 변수 변환을 통해 모형 변환을 해볼 수 있습니다.

✓ 20.다음 중 교차분석(Cross Tabulation)에 관한 설명 중 올바르지 않은 것은? 1/1

- ○ 두 변수 간의 연관 관계를 볼 때 교차표를 작성하여 변수들 간 관계를 분석하게 된다.
- ○ 교차 분석에 사용되는 검정 통계량이 카이스퀘어 분포를 따르기 때문에 카이스퀘어 검정이라 한다.
- ◉ 교차 분석은 두 변수 부류가 범주형 변수가 아니어도 사용할 수 있다. ✓
- ○ 교차표로 두 변수의 값이 공유하고 있는 빈도수가 몇 개인지 파악할 수 있다.

의견 보내기

교차 분석은 두 변수 부류가 범주형 변수이어야 한다

✓ 21. 다음 연속형 확률 분포 관련 설명으로 옳지 않은 것은? 1/1

- 정규분포는 평균과 표준편차에 의해 모양이 결정되고, 평균 0, 표준편차 1인 정규분포를 z분포라 한다
- t-분포는 분산의 특징을 확률분포로 만든 것이다 ✓
- 확률밀도 함수는 $\int_{-\infty}^{\infty} f(x)dx = 1$을 만족한다
- 표본의 크기가 N인 확률표본의 표본평균은 N이 충분히 크면 근사적으로 정규분포를 따르게 된다
- 다음 이미지 참조 (^^ 적분 표시가 안되네요)

의견 보내기

3-54. 연속형 확률분포 - 카이제곱 분포($\chi 2$)
분산의 특징을 확률분포로 만든 것으로, 카이(χ)는 평균 0, 분산 1인 표준정규분포를 의미함

✓ 22. 다중회귀모형의 통계적 유의성을 확인하는 방법은? 1/1

- F 통계량을 확인한다 ✓
- 결정계수를 확인한다
- 잔차통계량을 확인한다
- 회귀계수의 t값을 확인한다

의견 보내기

3-65. 회귀모형 해석(평가방법)
F 통계량
모델의 통계적 유의성을 검정하기 위한 검정 통계량 (분산 분석)
F통계량 = 회귀제곱평균(MSR) / 잔차제곱평균(MSE)
F통계량이 클수록 회귀 모형은 통계적으로 유의하다

✓ 23. 다음 보기에 해당하는 척도를 무엇이라고 하는가? 1/1

(1) 서울특별시, (2) 경기도, (3) 부산광역시 (4) 그 외 지역

- (●) 명목척도 ✓
- ( ) 서열척도
- ( ) 구간척도
- ( ) 비율척도

의견 보내기

3-41. 척도의 종류

명목척도
단순히 측정 대상의 특성을 분류하거나 확인하기 위한 목적
숫자로 바꾸어도 그 값이 크고 작음을 나타내지 않고 범주를 표시함

✓ 24. 다음 중 보기에서 비지도 학습에 해당하는 것끼리 짝지어진 것은? 1/1

가. 고객의 과거 거래 구매 패턴을 분석하여 고객이 구매하지 않은 상품 추천
나. 우편물에 인쇄된 우편번호 판별분석을 통해 우편물 분류
다. 동일 차종의 수리 보고서 데이터를 분석하여 차량 수리 소요시간 예측
라. 상품 구매시 유사한 상품을 구매한 고객들의 구매 데이터를 분석하여 쿠폰 발행

- ( ) 가, 나
- ( ) 가, 다
- (●) 가, 라 ✓
- ( ) 나, 다

의견 보내기

비지도 학습의 경우 종속변수가 존재하지 않는 독립변수만으로 이루어진 학습
가, 라는 비지도 학습 중 연관분석에 해당함

✗ 25. 아래 오분류표를 이용하여 Accuracy를 구하는 식을 작성하시오. ···/1

confusion matrix		예측값	
		TRUE	FALSE
실제값	TRUE	a	b
	FALSE	c	d

(a + d) / (a + b + c + d) ✗

정답

(a + d) / (a+b+c+d)

의견 보내기

3-91. 오분류표를 활용한 평가 지표
정확도(accuracy) : 전체 예측에서 옳은 예측의 비율
(a + d) / (a+b+c+d)

✓ 26. 시계열 모형 중 현 시점의 자료가 p 시점 전의 유한 개의 과거 자료로 설명될 수 있는 모형을 무엇이라고 하는가? 1/1

AR 모형 ✓

의견 보내기

3-76. 시계열 모형
AR(p) : 현 시점의 자료가 p 시점 전의 유한 개의 과거 자료로 설명될 수 있는 모델
MA(q) : 최근 데이터의 평균을 예측치로 사용하는 방법, 현시점의 자료가 유한 개의 과거 백색잡음(정상시계열)의 선형결합으로 표현된 모형

✗ 27. 계층적 군집 방법에서 사용되는 측도 중 두 벡터의 내적의 코사인 값을 이용하여 측정된 벡터 간의 유사한 정도를 측정하는 방법은 무엇인가? ···/1

코사인(cosine) 유사도 ✗

정답

코사인 유사도

cosine similarity

의견 보내기

3-94. 계층적 군집의 거리
코사인 유사도(cosine similarity) : 각도가 0° 일 때의 코사인값은 1 이며, 다른 모든 각도의 코사인 값은 1 보다 작다. 따라서 이 값은 벡터의 크기가 아닌 방향의 유사도를 판단하는 목적으로 사용되며, 두 벡터의 방향이 완전히 같을 경우 1, 90° 의 각을 이룰 경우 0, 180° 로 완전히 반대 방향인 경우 -1 의 값을 갖는다.

✓ 28. 비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도의 형태로 형상화 하는 알고리즘은? 1/1

SOM ✓

의견 보내기

3-97. SOM(Self-Organizing Maps)
인공신경망의 한 종류로, 차원축소와 군집화를 동시에 수행하는 기법
비지도 학습(Unsupervised Learning)의 한 가지 방법
고차원으로 표현된 데이터를 저차원으로 변환해서 보는데 유용함
입력층과 2차원의 격자 형태의 경쟁층(=출력층) 으로 이루어져 있음(2개의 층으로 구성)

✓ 29. 배깅(bagging)에 랜덤 과정을 추가한 방법으로, 노드 내 데이터를 자식 노드로 나누는 기준을 정할 때 모든 예측변수에서 최적의 분할을 선택하는 대신, 설명변수의 일부분만을 고려함으로 성능을 높이는 방법을 사용하는 것은 무엇인가? 1/1

랜덤포레스트 ✓

의견 보내기

3-83. 앙상블(Ensemble) 모형
랜덤포레스트(Random Forest)
배깅(Bagging)에 랜덤 과정을 추가한 방법
노드 내 데이터를 자식 노드로 나누는 기준을 정할 때 모든 예측변수에서 최적의 분할을 선택하는 대신, 설명변수의 일부분만을 고려함으로 성능을 높이는 방법 사용
여러 개 의사결정 나무를 사용해, 하나의 나무를 사용할 때보다 과적합 문제를 피할 수 있음

✓ 30. 분류 모형 성능 평가에 사용되며, X 축은 FP Rate(1-특이도), Y축은 Sensitivity를 나타내는 이 두 평가 값의 관계로 모형을 평가하는 것으로 이것의 밑 부분의 면적이 넓을수록 좋은 모형으로 평가되는 그래프는 무엇인가? 1/1

ROC Curve ✓

의견 보내기

3-92. 분류 모형 성능 평가
ROC(Receiver Operating Characteristic) Curve
X축은 FP Rate, Y축은 민감도(Sensitivity)를 나타내 이 두 평가 값의 관계로 모형을 평가함
ROC 그래프의 밑부분의 면적(AUC, Area Under the Curve)이 넓을수록 좋은 모형으로 평가함