감 독 확 인 란

제33회 데이터 분석 준전문가 자격검정 시험 복원 문제

- 검정일시 : 2022. 05. 21(토) / 10:00~11:30
- 수험번호 :
- 성　　명 :

과목 I 데이터 이해 * 문항 수(8문항), 배점(문항 당 2점)

01. 사물끼리 정보를 주고받는 사물인터넷 시대를 빅데이터의 관점에서 바라볼 때 다음 중 사물인터넷의 의미로 가장 적절한 것은?

① 모든 것의 데이터화(Datafication)
② 서비스 지능화(intelligent service)
③ 분석 고급화(advanced analytics)
④ 정보 공유화(information sharing)

02. 빅데이터의 위기 요인과 통제방안을 서로 연결한 것 중 잘못 연결된 것은?

아래

가. 사생활 침해 - 동의제에서 책임제로 변화
나. 책임원칙 훼손 - 알고리즘 접근 허용
다. 데이터 오용 - 정보선택 옵션 제공

① 가, 나
② 가, 다
③ 나, 다
④ 가, 나, 다

03. 다음 중 NoSQL 데이터베이스가 아닌 것은?

① HBase
② MongoDB
③ MySQL
④ Cassandra

04. 다음 중 데이터베이스의 일반적인 특징에 대한 설명으로 가장 부적절한 것은?

① 한 조직의 다수 사용자가 공동으로 이용하고 유지하는 공용 데이터이다.
② 동일한 내용의 데이터가 중복되지 않는 통합 데이터이다.
③ USB, HDD 또는 SSD와 같은 컴퓨터가 접근할 수 있는 매체에 저장된 데이터이다.
④ 저장, 검색, 분석이 용이하게 수치로 명확하게 표현되는 정량 데이터이다.

05. 다음 중 데이터 사이언티스트에게 요구되는 소프트 스킬으로 가장 적절하지 않은 것은?

① 이론적 지식
② 창의적 사고
③ 커뮤니케이션 기술
④ 시각화를 활용한 설득력

06. 다음 중 데이터의 양을 표현하는 단위를 작은 것에서 큰 것 순으로 나열한 것으로 가장 적절한 것은?

① 엑사바이트 < 페타바이트 < 요타바이트 < 제타바이트
② 페타바이트 < 엑사바이트 < 제타바이트 < 요타바이트
③ 페타바이트 < 요타바이트 < 엑사바이트 < 제타바이트
④ 요타바이트 < 제타바이트 < 엑사바이트 < 페타바이트

07. 다음 중 빅데이터 분석 활용의 효과로 가장 적절하지 않은 것은?

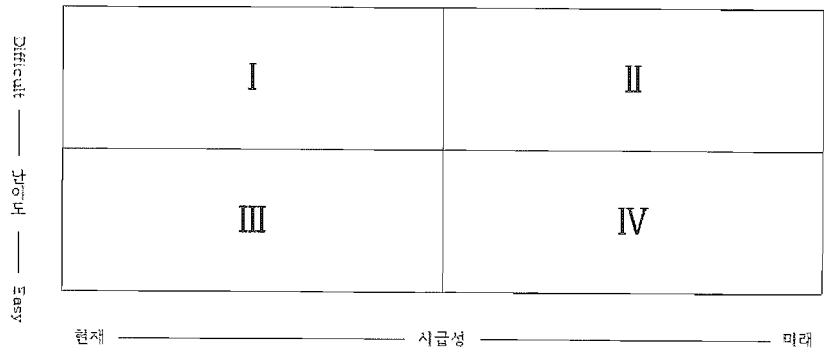
① 서비스 산업의 확대와 제조업의 축소
② 상품 개발과 조립 비용의 절감
③ 운송 비용의 절감
④ 새로운 수익원의 발굴 및 활용

08. 다음 중 빅데이터에 대한 설명으로 가장 적절하지 않은 것은?

① 빅데이터 환경에서는 필요한 정보만을 추출하기 위해 표본조사의 중요성이 더욱 대두되고 있다.
② 빅데이터를 통해 기존 방식으로는 얻을 수 없었던 새로운 통찰이나 가치 창출이 가능하다.
③ 빅데이터의 출현배경으로 SNS의 확산, 클라우드 컴퓨팅의 발전, 저장장치의 가격하락 등이 있다.
④ 4차 산업혁명 시대에 과거 석탄과 철과 같은 역할을 하게 될 것으로 기대한다.

과목 Ⅱ 데이터 분석 기획 * 문항 수(8문항), 배점(문항 당 2점)

09. 아래는 분석 과제 우선순위 선정 매트릭스이다. 분석과제의 적용 우선순위를 시급성에 두었을 때 결정해야 할 우선순위로 적절한 것은?



① Ⅲ - Ⅱ - Ⅰ ② Ⅲ - Ⅳ - Ⅱ ③ Ⅲ - Ⅱ - Ⅳ ④ Ⅲ - Ⅰ - Ⅱ

10. 빅데이터의 특징을 고려한 분석 ROI 요소와 분석우선순위 평가기준에 대한 설명으로 가장 부적절한 것은?

① 분석과제의 우선순위 평가에서 시급성은 전략적 중요도, 데이터 수집비용 등을 평가하고 난이도는 분석 수준과 복잡도가 평가요소이다.

② 분석 난이도는 분석 준비도와 성숙도 진단 결과에 따라 해당 기업의 분석 수준을 파악하고 이를 바탕으로 결정된다.

③ 시급성이 높고 난이도가 높은 분석과제는 경영진 또는 실무 담당자의 의사결정에 따라 적용 우선순위를 조정할 수 있다.

④ 시급성이 높고 난이도가 낮은 분석과제는 우선순위가 높다.

11. 다음 중 아래의 하향식 접근법(Top Down Approach)데이터 분석 기획 단계를 순서대로 나열한 것으로 적절한 것은?

아래

가. 문제 탐색(Problem Discovery)
나. 문제 정의(Problem Definition)
다. 해결방안탐색(Solution Search)
라. 타당성 검토(Feasibility study)

① 나 → 가 → 라 → 다
② 나 → 가 → 다 → 라
③ 가 → 나 → 라 → 다
④ 가 → 나 → 다 → 라

12. 다음 중 데이터 분석에서 정확도(Accuracy)와 정밀도(Precision)에 대한 설명으로 가장 적절하지 않은 것은?

① 정확도는 True로 예측한 것 중 실제 True인 비율, 정밀도는 실제 True인 경우에서 True로 예측한 비율이다.
② 정확도는 모델의 실제 값 사이의 차이이고, 정밀도는 모델을 지속적으로 반복했을 때 편차의 수준이다.
③ 모형의 활용측면에서는 정확도가, 모델의 안정성측면에서는 정밀도가 중요하다.
④ 정확도와 정밀도는 트레이드-오프(Trade-off) 관계가 되는 경우가 많다.

13. 다음 중 데이터 분석 마스터 플랜 수립시 분석과제의 우선순위를 결정할 때 고려해야 할 요소로서 가장 적절하지 않은 것은?

① 전략적 중요도
② 비즈니스 성과 및 ROI
③ 실행 용이성
④ 데이터 필요 우선순위

14. 다음 중 계층적 데이터 분석 프로세스 모델에 대한 설명으로 가장 적절하지 않은 것은?

① 최상위 계층은 단계(Phase)로 구성되고 마지막 계층은 태스크(Task)로 구성된다.
② 각 단계는 보통 기준선(Baseline)을 설정하여 관리되고 버전 관리를 통하여 통제가 이루어져야한다.
③ 마지막 단계인 스텝(step)은 입력(Input)과 출력(Output) 등으로 구성된 단위 프로세스이다.
④ 데이터 분석 프로세스는 동료간 평가(Peer Review) 수행이 적절하지 않다.

15. 다음 중 데이터 분석 기획 단계에서 수행하는 주요 과제(Task)로 가장 적절하지 않은 것은?

① 필요 데이터의 정의
② 프로젝트 범위 설정
③ 프로젝트 정의
④ 위험 식별

16. 다음 중 데이터 분석 상향식 접근(Bottom Up Approach)에 대한 설명으로 가장 적절하지 않은 것은?

① 문제를 정의하기 어려운 경우에 사용한다.
② 다양한 원천 데이터를 대상으로 분석을 수행하여 가치 있는 문제를 도출하는 일련의 과정이다.
③ 일반적으로 지도 학습(Supervised Learning) 방식을 수행한다.
④ 하향식 접근 방식과는 달리 복잡하고 다양한 환경에서 발생하는 문제 해결에도 적합하다.

과목 III 데이터 분석 * 문항 수(24문항), 배점(문항 당 2점)

17. 확률변수 X가 확률질량함수 f(x)를 갖는 이산형 확률변수인 경우 그 기댓값으로 옳은 식은?

① $E(X)=\sum xf(x)$
② $E(X)=\int xf(x)dx$
③ $E(X)=\sum x^2f(x)$
④ $E(X)=\int x^2f(x)dx$

18. 다음 중 시계열 데이터를 조정하여 예측하는 평활법(Smoothing method)에 대한 설명으로 적절하지 않은 것은?

① 이동평균법이란 시계열 데이터가 일정한 주기를 갖고 비슷한 패턴으로 움직이고 있는 경우에 적용시킬 수 있는 방법이다.
② 이동평균법은 시계열자료에서 계절변동과 추세변동을 제거하여 순환변동만 가진 시계열자료로 변환하는 방법이다.
③ 단순지수평활법은 추세나 계절성이 없어 평균이 변화하는 시계열에 사용하는 방법이다.
④ 이중지수평활법은 평균을 평활하는 모수와 함께 추세를 나타내는 식을 다른 모수로 평활하는 방법이다.

19. 스피어만 상관계수를 계산할 때 대상이 되는 자료의 종류는 무엇이어야 하는가?

① 서열척도
② 명목척도
③ 비율척도
④ 등간척도

20. 아래는 근로자의 임금 등에 대한 데이터에 대한 분석 결과이다. 다음 중 유의수준 0.05에서 이에 대한 설명으로 가장 적절하지 않은 것은?

아래

```
> summary(Wage[,c("wage", "age", "jobclass")])
      wage             age                    jobclass
 Min.   : 20.09   Min.   :18.00   1. Industrial :1544
 1st Qu.: 85.38   1st Qu.:33.75   2. Information:1456
 Median :104.92   Median :42.00
 Mean   :111.70   Mean   :42.41
 3rd Qu.:128.68   3rd Qu.:51.00
 Max.   :318.34   Max.   :80.00

> model<-lm(wage~age+jobclass+age*jobclass,data=Wage)
> summary(model)

Call:
lm(formula = wage ~ age + jobclass + age * jobclass, data = Wage)

Residuals:
     Min       1Q   Median       3Q      Max
-105.656  -24.568   -6.104   16.433  196.810

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 73.52831    3.76133  19.548  < 2e-16 ***
age                          0.71966    0.08744   8.230 2.75e-16 ***
jobclass2. Information      22.73086    5.63141   4.036 5.56e-05 ***
age:jobclass2. Information -0.16017    0.12785  -1.253     0.21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.16 on 2996 degrees of freedom
Multiple R-squared:  0.07483,Adjusted R-squared:  0.07391
F-statistic: 80.78 on 3 and 2996 DF,  p-value: < 2.2e-16
```

① 직업군이 동일할 때, 나이가 많을수록 임금이 올라가는 경향이 있다.

② 나이가 동일할 때, Information 직군이 Industrial 직군에 비해 평균적으로 임금이 높다.

③ 나이에 따라 두 직군 간의 임금의 평균 차이가 유의하게 변하지 않는다.

④ 위의 회귀식은 유의수준 0.05에서 임금의 변동성을 설명하는데 유의하지 않다.

21. 다음 중 신경망 모형에서 입력받은 데이터를 다음 층으로 어떻게 출력할지를 결정하는 함수로 가장 적절한 것은?

① 로짓함수
② 활성화함수
③ CHAID 함수
④ 오즈비 함수

22. 모집단내에서 모집단의 특성을 잘 타나낼 수 있는 일부를 추출하여 이들로부터 자료를 수집하고 수집된 자료를 토대로 모집단의 특성을 추정하게 된다. 이 때 조사하는 모집단의 일부분을 표본(sample)이라 한다. 다음 중 표본조사에 대한 설명으로 가장 부적절한 것은?

① 표본오차(sampling error)는 모집단을 대표할 수 있는 표본 단위들이 조사대상으로 추출되지 못함으로서 발생하는 오차를 말한다.
② 표본편의(sampling bias)는 모수를 작게 또는 크게 할 때 추정하는 것과 같이 표본추출방법에서 기인하는 오차를 의미한다.
③ 표본편의는 확률화(randomization)에 의해 최소화하거나 없앨 수 있다.
④ 비표본오차(non-sampling error)는 표본오차를 제외한 모든 오차로 조사 과정에서 발생하는 모든 부주의나 실수, 알 수 없는 원인 등 모든 오차를 의미하며 조사대상이 증가한다고 해서 오차가 커지지는 않는다.

23. 다음 중 주성분분석에 대한 설명으로 적절한 것은?

① 상관관계가 있는 고차원의 데이터를 저차원 데이터로 축소하는 방법이므로 독립변수들간 다중공선성 문제를 해결할 수 있다.
② 여러 대상 간의 거리가 주어졌을 때, 대상들을 동일한 상대적 거리를 가진 실수 공간의 값들로 배치하여 자료들의 상대의 관계를 이해하는 시각화 방법의 근간으로 주로 사용된다.
③ 비슷한 특징을 가지는 소집단으로 특이 패턴을 찾는 것으로 고객 세분화 등에 많이 활용된다.
④ 항목 간의 "조건-결과"식으로 표현되는 유용한 패턴을 발견할 수 있으며, 흔히 장바구니 분석이라고도 불린다.

24. 다음 중 아래 오분류표를 이용하여 구한 F1 값은 얼마인가?

		예측치		합계
		True	False	
실제값	True	200	300	500
	False	300	200	500
합계		500	500	1000

① 0.15 ② 0.3 ③ 0.4 ④ 0.55

25. 아래 오분류표에서 재현율(Recall)로 가장 적절한 것은?

		예측치		합계
		True	False	
실제값	True	30	70	100
	False	60	40	100
합계		90	110	200

① $\frac{3}{10}$ ② $\frac{2}{5}$ ③ $\frac{1}{3}$ ④ $\frac{7}{11}$

30 DAY

26. 다음 중 k-means 군집의 단점으로 가장 부적절한 것은?

① 볼록한 형태가 아닌 군집이 존재하면 성능이 떨어진다.
② 사전에 주어진 목적이 없으므로 결과 해석이 어렵다.
③ 잡음이나 이상값에 영향을 많이 받는다.
④ 한번 군집이 형성되면 군집내 객체들은 다른 군집으로 이동 할 수 없다.

27. 다음 중 연관규칙의 단점으로 가장 적절하지 않은 것은?

① 품목수가 증가하면 분석에 필요한 계산이 기하급수적으로 증가한다.
② 지나치게 세분화된 품목으로 연관규칙을 찾으려고 하면 의미 없는 분석 결과가 나올 수 있다.
③ 상대적으로 거래량이 적은 품목은 당연히 포함된 거래수가 적어 규칙 발견이 제외되기 쉽다.
④ 품목 간에 구체적으로 어떠한 영향을 줄 수 있는지 해석하기 어렵다.

28. 아래의 확률을 알고 있다고 가정할 때, 질병을 가지고 진단한 사람이 실제로 질병을 가진 사람일 확률은?

아래

- 전체 인구 중 해당 질병을 가지고 있는 사람은 10%
- 진단 결과 전체 인구 중 20%가 해당 질병을 가지고 있다고 진단됨
- 해당 질병을 가지고 있는 사람의 90%는 질병을 가지고 있는 것으로 진단됨

① 0.9 ② 0.8 ③ 0.45 ④ 0.3

29. 아래 데이터 셋 A, B 의 유클리드 거리(Euclidean distance)를 계산하시오.

	A	B
키	185	180
앉은 키	70	75

① 0 ② $\sqrt{10}$ ③ $\sqrt{25}$ ④ $\sqrt{50}$

30. 다음 중 주성분 회귀 분석에 대한 설명으로 가장 적절하지 않은 것은?

① 차원이 축소된 주성분으로 회귀분석에 적용하는 방법으로 자료의 시각화에 도움을 줄 수 있다.
② 변수들의 선형결합으로 이루어진 주성분은 서로 직교하며, 기존 자료보다 적은 수의 주성분들을 회귀분석의 독립변수로 설정할 수 있다.
③ 주성분의 개수는 기존보다 큰 고유값(Eigenvalue)의 계수로 정할 수 있다.
④ 개별 고유치의 분해 가능 여부를 판단하여 주성분의 개수를 정한다.

31. SOM은 비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도 형태로 형상화하는 방법이다. 다음 중 SOM 방법에 대한 설명으로 부적절한 것은?

① SOM은 입력변수의 위치 관계를 그대로 보존한다는 특징으로 인해 입력 변수의 정보와 그들의 관계가 지도상에 그대로 나타난다.
② SOM을 이용한 군집분석은 인공신경망의 역전파 알고리즘을 사용함으로써 수행 속도가 빠르고 군집의 성능이 매우 우수하다.
③ SOM 알고리즘은 고차원의 데이터를 저차원의 지도 형태로 형상화하기 때문에 시각적으로 이해하기 쉽다.
④ SOM은 경쟁 학습으로 각각의 뉴런이 입력 벡터와 얼마나 가까운가를 계산하여 연결강도를 반복적으로 재조정하여 학습한다.

32. 다음 중 의사결정 나무 모형의 학습 방법에 대한 설명으로 부족한 것은 무엇인가?

① 이익도표 또는 검정용 자료에 의한 교차타당성 등을 이용해 의사결정나무를 평가한다.

② 분리 변수의 P차원 공간에 대한 현재 분할은 이전 분할에 영향을 받지 않고 이루어지며, 공간을 분할하는 모든 직사각형들이 가능한 순수하게 되도록 만든다.

③ 각 마디에서의 최적 분리규칙은 분리변수의 선택과 분리기준에 의해 결정된다.

④ 가지치기는 분류 오류를 크게 할 위험이 높거나 부적절한 규칙을 가지고 있는 가지를 제거하는 작업이다.

33. 앙상블모형(Ensemble)이란 주어진 자료로부터 여러 개의 예측모형을 만든 후 이러한 예측모형들을 결합하여 하나의 최종 예측모형을 만드는 방법을 말한다. 다음 중 앙상블모형에 대한 설명으로 적절하지 않은 것은?

① 배깅은 주어진 자료에서 여러 개의 붓스트랩(Bootstrap) 자료를 생성하고 각 붓스트랩 자료에 예측모형을 만든 후 결합하여 최종 모형을 만드는 방법이다.

② 부스팅은 배깅의 과정과 유사하여 재표본 과정에서 각 자료에 동일한 확률을 부여하여 여러 모형을 만들어 결합하는 방식이다.

③ 랜덤 포레스트(Random Forrest)는 의사결정나무모형의 특징인 분산이 크다는 점을 고려하여 배깅보다 더 많은 무작위성을 추가한 방법으로 약한 학습기들을 생성하고 이를 선형결합해 최종 학습기를 만드는 방법이다.

④ 앙상블모형은 훈련을 한 뒤 예측을 하는데 사용하므로 교사학습법(Supervised Learning)이다.

34. 아래는 피자와 햄버거의 거래 관계를 나타낸 표로, Pizza/Hamburgers는 피자/햄버거를 포함하는 거래수를 의미하고 (Pizza)/(Hamburgers)는 피자/햄버거를 포함하지 않은 거래 수를 의미한다. 아래 표에서 피자 구매와 햄버거 구매에 대해 설명한 것으로 가장 적절한 것은 무엇인가?

	Pizza	(Pizza)	합계
Hamburgers	2,000	500	2,500
(Hamburgers)	1,000	1,500	2,500
합계	3,000	2,000	5,000

① 지지도가 0.6로 전체 구매 중 햄버거와 피자가 같이 구매되는 경향이 높다.

② 정확도가 0.7로 햄버거와 피자의 구매 관련성은 높다.

③ 향상도가 1보다 크므로 햄버거와 피자 사이에 연관성이 높다고 할 수 있다.

④ 연관규칙 중 "햄버거->피자" 보다 "피자->햄버거" 의 신뢰도가 더 높다.

35. 다음 중 회귀분석의 변수 선택법에 대한 설명으로 가장 적절하지 않은 것은?

① 전진 선택법은 중요하다고 생각되는 설명 변수부터 차례로 선택하는 방법이다.
② 전진 선택법으로 변수를 추가할 때 기존 변수들의 중요도에 영향을 받지 않는다.
③ 후진 제거법은 변수의 개수가 많은 경우에 사용하기가 어렵다.
④ 전진 선택법은 변수값의 작은 변동에도 결과가 크게 달라지는 단점이 있다.

36. 과대적합(overfitting)은 통계나 기계학습에서 모델의 변수가 너무 많아 모델이 복잡하고 과대하게 학습될 때 주로 발생한다. 다음 중 과대 적합에 대한 설명으로 가장 부적절한 것은?

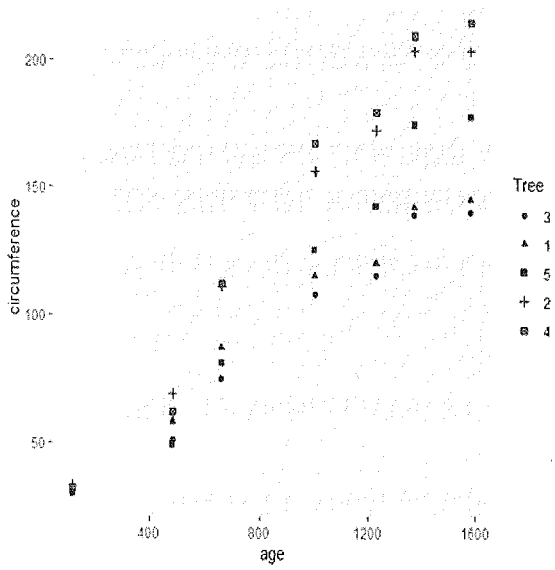
① 생성된 모델이 훈련 데이터에 너무 최적화되어 학습하여 테스트 데이터의 작은 변화에 민감하게 반응하는 경우는 발생하지 않는다.
② 변수가 너무 많아 모형이 복잡할 때 생긴다.
③ 과대적합이 발생할 것으로 예상되면 학습을 종료하고 업데이트하는 과정을 반복해 과대적합을 방지할 수 있다.
④ 학습데이터가 모집단의 특성을 충분히 설명하지 못할 때 자주 발생한다.

37. 다음 중 통계적 추론에 대한 설명으로 가장 적절하지 않은 것은?

① 구간추정은 모수의 참값이 포함되어 있다고 추정되는 구간을 결정하는 것이며, 실제 모집단의 모수는 신뢰구간에 포함되어야 한다.
② 점추정은 표본의 정보로부터 모집단의 모수를 하나의 값으로 추정하는 것이다.
③ 통계적 추론은 제한된 표본을 바탕으로 모집단에 대한 일반적인 결론을 유도하려는 시도이므로 본질적으로 불확실성을 수반한다.
④ 전수조사가 불가능하면 모집단에서 표본을 추출하고 표본을 근거로 확률론을 활용하여 모집단의 모수들에 대해 추론하는 것을 추정이라 한다.

38. 아래는 다섯 종류의 오렌지 나무에 대한 연령(age)와 둘레(circumstence)를 측정한 자료이다. 다음 중 아래에 대한 설명 중 가장 적절하지 않은 것은?

아래




```
> head(Orange)
  Tree  age circumference
1    1  118            30
2    1  484            58
3    1  664            87
4    1 1004           115
5    1 1231           120
6    1 1372           142

> 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
```

① 연령이 증가할수록 둘레가 증가하는 경향이 있다.

② 나무 연령의 평균값은 922.1이다.

③ 나무 종류별로 둘레에 유의한 차이가 있다.

④ 나무 둘레의 평균값은 115.9이다.

39. 다음 headsize 데이터는 25개 가구에서 첫 번째와 두 번째 성인 아들의 머리길이(head)와 머리폭(breadth)을 보여준다. 이에 대한 설명 중 가장 부적절한 것은?

아래

```
> head(headsize)
     head1 breadth1 head2 breadth2
[1,]   191      155   179      145
[2,]   195      149   201      152
[3,]   181      148   185      149
[4,]   183      153   188      149
[5,]   176      144   171      142
[6,]   208      157   192      152
> str(headsize)
 num [1:25, 1:4] 191 195 181 183 176 208 189 197 188 192 ...
 - attr(*, "dimnames")=List of 2
  ..$ : NULL
  ..$ : chr [1:4] "head1" "breadth1" "head2" "breadth2"
> out<-princomp(headsize)
> print(summary(out),loadings=TRUE)
Importance of components:
                          Comp.1  Comp.2  Comp.3  Comp.4
Standard deviation        15.1    5.42    4.12    3.000
Proportion of Variance    0.8     0.10    0.06    0.032
Cumulative Proportion     0.8     0.91    0.97    1.000

Loadings:
          Comp.1  Comp.2  Comp.3  Comp.4
head1     0.570   0.693  -0.442
breadth1  0.406   0.219   0.870  -0.173
head2     0.601  -0.633  -0.209  -0.441
breadth2  0.386  -0.267          0.881
```

① 주성분분석의 결과를 보여준다.

② 앞의 두 개 주성분으로 전체 데이터 분산의 91% 설명할 수 있다.

③ 두 번째 주성분은 네 개의 변수와 양의 상관관계를 가진다.

④ 네 개의 주성분을 사용하면 전체 데이터 분산을 모두 설명할 수 있다.

40. 모집단이 정규분포를 따르고 분산이 알려져 있으며 모평균에 대한 95% 신뢰수준하에서 신뢰구간이 "$50 \pm 1.96\frac{1}{\sqrt{100}} = (49.804, 50.196)$ 로 도출되었을 때, 다음 중 이에 대한 해석으로 가장 적절하지 않은 것은?

① 모집단의 표준편차는 1이다.
② 표본의 개수는 100개이고, 그 표본평균은 50이다.
③ 신뢰구간 추정값(신뢰구간)의 구간 내에 실제 평균값이 포함되어 있지 않을 수도 있다.
④ 동일 모집단에서 동일한 방법과 개수로 다시 표본을 추출하면, 새로운 표본의 신뢰구간 추정값도(신뢰구간)으로 동일하다.

단 답 형

* 문항 수(10문항), 배점(문항 당 2점, 부분점수 없음)

01. 아래 데이터 분석과 관련된 기술을 설명한 것이다. (가) 에 들어갈 용어를 기입하시오.

아래

기업의 의사결정 과정을 지원하기 위한 주제 중심적이고 통합적이며 시간성을 가지는 비휘발성 데이터의 집합을 (가)라고 한다.

(　　)

02. 아래 (㉠) 안에 공통적으로 들어갈 용어는 ?

아래

(㉠)(이)란 데이터로부터 의미있는 정보를 추출해 내는 학문으로, 통계학과는 달리 정형 또는 비정형을 막론하고 다양한 유형의 데이터를 분석 대상으로 한다. 또한 분석에 초점으로 두는 데이터 마이닝과는 달리 (㉠)는 분석 뿐만 아니라 이를 효과적으로 구현하고 전달하는 과정까지 포함하는 포괄적인 개념이다.

(　　)

03. 분석은 분석의 대상(What)과 분석의 방법(How)에 따라 아래와 같이 분류한다. 다음 중 아래의 빈칸에 들어갈 용어로 가장 적절한 것은?

		분석의 대상(What)	
		Known	Un-Known
분석의 방식 (How)	Known	Optimization	( )
	Un-Known	Solution	Discovery

( )

04. 아래는 여러 분석 방법론 중 하나에 대한 설명이다. 이것으로 적절한 용어는?

아래

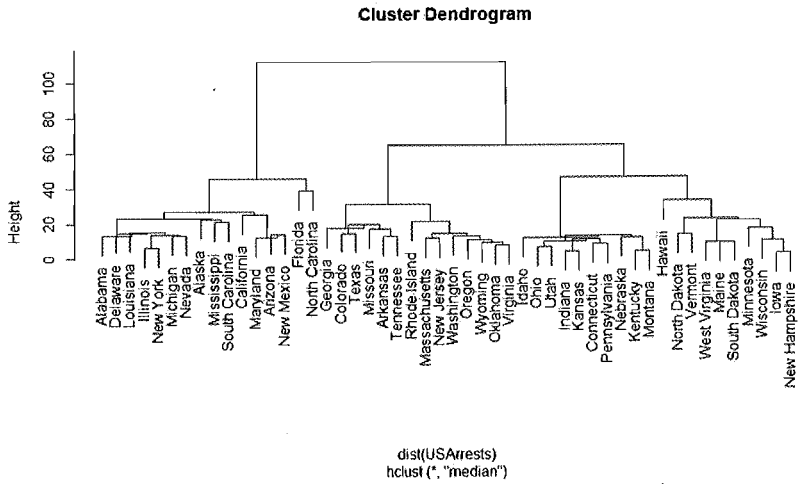
이것은 반복을 통하여 점진적으로 개발하는 방법으로서, 처음 시도하는 프로젝트에 적용이 용이하지만 관리 체계를 효과적으로 갖추지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있다.

( )

05. 오분류표(Confusion Matrix)를 활용하여 모형을 평가하는 지표 중 실제값이 FALSE인 관측치 중 예측치가 적중한 정도를 나타내는 지표는?

( )

06. 계층적 군집분석 결과를 아래와 같이 덴드로그램으로 시각화하였다고 할 때 Tree의 높이(height)가 60일 경우 나타나는 군집의 수를 쓰시오.



( )

07. 가설검정 결과에서 귀무가설이 옳은데도 귀무가설을 기각하게 되는 오류는?

( )

08. 로지스틱 회귀분석에서는 이산형(Binary) 종속변수가 1일 확률을 모형화한다. 설명변수가 한 단위 증가할 때 종속변수가 1인 확률과 0인 확률 비의 증가율을 나타내는 것은?

( )

09. 신경망 모형에서 출력값 z가 여러 개로 주어지고 목표치가 다범주인 경우 각 범주에 속할 사후 확률을 제공하여 출력노드에 주로 사용되는 함수는?

( )

10. 신경망 모형의 학습을 위한 역전파 과정에서 오차를 더 줄일 수 있는 가중치가 존재함에도 기울기가 0이 되어버려 더 이상 학습이 진행되지 않는 문제를 나타내는 용어는?

( )