감 독 확 인 란

제31회 데이터 분석 준전문가 자격검정 시험 복원 문제

- 검정일시 : 2021 . 11 . 06(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 사물인터넷과 관련이 가장 큰 것은?

① 인공지능(AI)
② 스마트 데이터(Smart Data)
③ 데이터화(Datafication)
④ 지능적 서비스(Intelligent Service)

02. 다음 데이터 분석 조직의 유형 중 별도의 분석 조직이 없고 해당 업무부서에서 분석을 수행하는 방식에 해당하는 것은?

① 기능형
② 분산형
③ 복합형
④ 집중형

03. 데이터베이스의 일반적인 특징으로 가장 부적절한 것은?

① 데이터베이스는 여러 사용자가 서로 다른 목적으로 데이터를 공동으로 이용할 수 있도록 구성되어 있다.
② 데이터베이스는 자기 디스크나 자기 테이프 등과 같이 컴퓨터가 접근할 수 있는 저장 매체에 저장된 데이터이다.
③ 데이터베이스는 변화하는 데이터로 데이터의 삽입, 삭제, 갱신을 한다고 하더라도 항상 현재의 정확한 데이터를 유지해야 한다.
④ 데이터베이스는 한곳에 통합된 데이터(integrated data)이므로 동일한 내용이더라도 데이터의 중복을 허용한다.

04. 데이터에서 가치를 찾아내는 과정을 피라미드의 계층구조로 나타낸다. 다음 예시를 알맞게 설명한 것을 고르시오.

아래

(a) : A마트는 100원에, B마트는 200원에 연필을 판매한다.
(b) : A마트의 연필이 더 싸다.
(c) : 상대적으로 저렴한 A마트에서 연필을 사야겠다.

① (a) : 데이터, (b) : 정보, (c) : 지식
② (a) : 데이터, (b) : 지식, (c) : 지혜
③ (a) : 데이터, (b) : 정보, (c) : 지혜
④ (a) : 정보, (b) : 지식, (c): 지혜

05. 일차원적 분석을 통해서도 해당 부서나 업무 영역에서는 상당한 효과를 얻을 수 있다. 다음 중 업무 영역과 분석 사례의 연결이 가장 부적절한 것은?

① 마케팅 관리 - 상점과 가게 위치 선정
② 재무 관리 - 거래처 선정
③ 공급체인 관리 - 적정 재고량 결정
④ 인력 관리 - 이직 인력 예측

06. 아래에서 빅데이터 시대의 위기와 통제에 대한 설명으로 가장 타당한 것끼리 묶은 것은?

아래

가) 데이터 익명화(anonymization)는 사생활 침해에 대한 근본요인을 차단할 수 있어 빠른 기술발전이 필요하다.
나) 빅데이터 분석은 일어난 일에 대한 데이터에 의존하므로 예측의 정확도는 높지만 항상 맞을 수는 없어 데이터 오용의 피해가 발생할 수 있다.
다) 개인정보 사용자의 정보사용에 대한 무한책임의 한계로 개인정보사용 책임제 보다 동의제를 더욱 강화해야 한다.
라) 민주주의에서 '행동결과'에 따른 처벌의 모순을 교훈삼아 빅데이터 사전 '성향' 분석을 통한 통제가 강화될 필요가 있다.
마) 빅데이터가 발생시키는 문제를 중간자 입장에서 중재하며 해결해 주는 알고리즈미스트(algorithmist)도 새로운 직업으로 부상하게 될 것이다.

① 가, 다　② 나, 다　③ 가, 라　④ 나, 마

07. 다음 중 데이터베이스의 특징과 가장 거리가 먼 것은?

① 응용프로그램 종속성
② 데이터의 무결성 유지
③ 프로그래밍 생산성 향상
④ 데이터 중복성 최소화

08. 다음 중 데이터 관리 체계에 대한 설명으로 가장 거리가 먼 것은?

① ERD는 운영 중인 데이터베이스와 일치하기 위하여 철저한 변경관리가 필요하다.
② 빅데이터 거버넌스는 산업 분야별, 데이터 유형별, 정보 거버넌스 요소별로 구분하여 작성한다.
③ 빅데이터는 고품질의 데이터 확보가 필요하므로 데이터 수명주기 관리보다는 품질관리가 중요하다.
④ 데이터 정합성 및 활용의 효율성을 위하여 표준 데이터를 포함한 메타데이터와 데이터 사전의 관리 원칙을 수립해야 한다.

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

09. 데이터 분석 조직구조의 설명으로 가장 부적절한 것은?

① 집중형 조직구조는 조직 내 별도의 분석 전담조직을 독립적으로 구성하는 것으로서 분석업무의 중복 또는 이원화의 이슈가 있다.

② 기능 중심의 조직구조는 별도의 분석전담조직을 구성하지 않고 해당 부처에서 직접 분석을 수행함으로써 국한된 분석 수행 이슈가 존재한다.

③ 분산구조는 분석 조직의 인력을 현업부서에 배치하여 분석업무를 수행함으로써 분석이 집중되지 못해 신속한 실무적용이 어렵다.

④ 분석 조직은 분석 전문인력뿐만 아니라 도메인 전문가, IT 인력, 변화관리 및 교육담당 인력으로 구성되어야 효율적인 운영이 가능하다.

10. 다음 중 분석 주체 유형을 분류할 때 조직 내 분석 대상이 무엇인지 인지하고 있으나 데이터 분석 방법과 다양한 분석 구조를 이해하지 못하는 유형은 무엇인가?

① 발견

② 통찰

③ 솔루션

④ 최적화

11. 아래 ( )안에 들어갈 용어로 적절한 것은?

아래

현재의 비즈니스 모델 및 유사/동종사례 탐색을 통해서 빠짐없이 도출한 분석 기회들을 구체적인 과제로 만들기 전에 ( )로 표기하는 것이 필요하다. 풀어야 할 문제에 대한 상세설명 및 해당 문제 해결했을 때 발생하는 효과를 명시함으로써 향후 데이터 분석 문제로의 전환 및 적합성 평가에 ( )를 활용하도록 한다.

① 분석과제 정의서

② 분석 유즈 케이스

③ 분석 주제 풀(Pool)

④ 프로젝트 계획서

12. 다음 중 프로토타입 방법론의 기본적인 프로세스와 가장 관련이 없는 것은?

① 가설 생성
② 디자인에 대한 실험
③ 실제 환경 테스트 결과에서 통찰 도출 및 가설 확인
④ 반복적으로 위험분석을 수행하여 위험을 관리하며 순환적으로 개선

13. 복잡하고 다양한 환경으로 인해 분석 대상이 무엇인지 모르거나, 문제의 정의 자체가 어려운 경우에 답을 미리 내는 것이 아니라 사물을 있는 그대로 인식하는 "What" 관점에서 접근하는 분석과제 발굴 방식은 무엇인가?

① 상향식
② 하향식
③ 하이브리드
④ 단계선택

14. 아래에서 빅데이터 거버넌스에 대한 설명으로 올바른 것끼리 묶은 것은?

아래

(A) 빅데이터 분석은 다양한 데이터를 활용하기 위하여 회사 내 모든 데이터를 활용해야 한다.
(B) 빅데이터 분석은 고품질의 데이터 확보가 필요하므로 수명주기관리보다는 품질관리가 중요하다.
(C) ERD는 운영중인 데이터베이스와 일치하기 위하여 철저한 변경관리가 필요하다.
(D) 빅데이터 거버넌스 산업분야별, 데이터 유형별, 정보 거버넌스 요소별로 구분하여 작성한다.

① A, B
② C, D
③ A, B, C
④ B, C, D

15. 분석을 사용하여 전략적 통찰력을 얻기 위한 방법으로 부적절한 것은?

① 경영의 본질을 제대로 바라볼 수 있도록 분석한다.
② 경영진은 직관적 결정을 지양하고 데이터 기반의 객관적 의사결정을 한다.
③ 사업 상황을 확인하기 위해 사업 내부의 문제들을 집중하여 분석을 이용한다.
④ 비즈니스의 핵심가치와 관련된 분석 프레임워크와 평가지표를 개발한다.

16. 다음 중 마스터 플랜을 수립할 때 우선순위 고려요소로 가장 적절하지 않은 것은?

① 전략적 중요도
② 데이터 우선 순위
③ 실행 용이성
④ 비즈니스 성과 / ROI

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

17. 데이터의 한 부분으로 특정 사용자가 관심을 갖고 있는 데이터를 담은 비교적 작은 규모의 데이터 웨어하우스는 무엇이라고 하는가?

① 데이터베이스
② 데이터 마트
③ 데이터 마이닝
④ 데이터 프레임

18. 연관성분석에서 유의미한 규칙을 찾아내기 위해 사용되는 측도(criterion) 중 아래의 설명이 가리키는 것으로 가장 적절한 것은?

아래

전체 항목 중 A와 B가 동시에 포함되는 항목수의 비율

① 지지도
② 민감도
③ 향상도
④ 신뢰도

19. 아래 거래 전표에서 연관 규칙 "A→B"의 향상도는 얼마인가?(소수점 첫째자리에서 반올림)

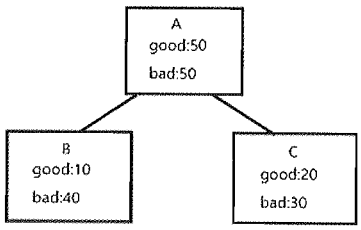
아래

물품	거래건수
{A}	100
{B, C}	100
{C}	100
{A, B, C, D}	50
{B, C}	200
{A, B, D}	250
{A, C}	200

① 30% ② 50% ③ 83% ④ 100%

20. 다음 중 아래 의사결정나무에서 C의 지니지수를 계산한 결과로 적절한 것은?

아래



① 0.5 ② 0.48 ③ 0.38 ④ 0.32

21. 아래는 회귀 모델의 예측 결과이다. 모델 성능을 MAPE로 계산했을 때 맞는 것은?

아래

Actual	1	2	5	10
Forecast	0.9	1.8	4.5	11

① 10% ② 15% ③ 32.5% ④ 45%

22. 비계층적 군집방법의 기법인 k-means clustering의 경우 이상값(outlier)에 민감하여 군집 경계의 설정이 어렵다는 단점이 존재한다. 이러한 단점을 극복하기 위해 등장한 비계층적 군집 방법으로 가장 적절한 것은?

① k-medoids Clustering
② 혼합 분포 군집(mixture distribution clustering)
③ Density based Clustering
④ Fuzzy Clustering

23. 붓스트랩 표본을 구성하는 재표본 과정에서 분류가 잘못된 데이터에 더 큰 가중치를 주어 표본을 추출하는 기법을 무엇이라고 하는가?

① 배깅(Bagging)
② 부스팅(Boosting)
③ 랜덤포레스트(Random Forest)
④ 시그모이드

24. 다음 중 Bias-variance trade off에 대한 아래 문장의 빈 칸에 들어갈 말로 순서대로 연결된 것은?

아래

일반적으로 학습모형의 유연성이 클수록 분산(variance)은 ( ), 편향(bias)은 ( ).

① 높고, 높다.　② 높고, 낮다.　③ 낮고, 높다.　④ 낮고, 낮다.

25. 다음 중 대용량 데이터 속에서 숨겨진 지식 또는 새로운 규칙을 추출해 내는 과정을 일컫는 것은?

① 지식경영　② 의사결정지원시스템

③ 데이터웨어하우징　④ 데이터 마이닝

26. 다음은 wage 데이터의 회귀분석 결과이다. 다음 설명 중 가장 옳지 않은 것을 고르시오.

아래

```
> summary(wage)
                education        wage
 1. < HS Grad      : 268     Min.   : 20.09
 2. HS Grad        : 971     1st Qu.: 85.38
 3. Some College   : 650     Median :104.92
 4. College Grad   : 685      Mean   :111.70
 5. Advanced Degree: 426     3rd Qu.:128.68
                              Max.   :318.34

> summary(lm(wage~.,wage))

Call:
lm(formula = wage ~ ., data = wage)

Residuals:
    Min      1Q  Median      3Q     Max
-112.31  -19.94   -3.09   15.33  222.56

Coefficients:
                                Estimate Std. Error t value  Pr(>|t|)
(Intercept)                      84.104      2.231   37.695  < 2e-16 ***
education2. HS Grad             11.679      2.520    4.634   3.74e-06 ***
education3. Some College        23.651      2.652    8.920  < 2e-16 ***
education4. College Grad         40.323      2.632   15.322  < 2e-16 ***
education5. Advanced Degree     66.813     2.848   23.462  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.53 on 2995 degrees of freedom
Multiple R-squared:  0.2348, Adjusted R-squared:  0.2338
F-statistic: 229.8 on 4 and 2995 DF,  p-value: < 2.2e-16
```

① education의 더미변수는 4개 이다.

② 회귀분석 결과를 회귀식으로 나타냈을 때, y절편은 84.104이다.

③ 회귀계수는 종속변수 wage 평균과의 차이를 의미하므로 "Advanced Degree" 그룹이 wage의 평균에 추가되는 값이 가장 크다.

④ 회귀식의 모든 변수가 통계적으로 유의미하다.

27. 다음 중 분석기법의 활용 분야가 나머지와 다른 하나를 고르시오.

① 로지스틱 회귀 분석

② 인공신경망

③ 의사결정나무

④ SOM

28. 다음 중 나머지와 분석 방법이 다른 것은?

① k-means clustering

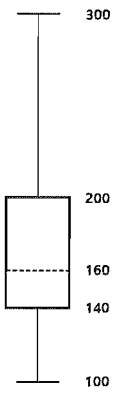
② Single linkage method

③ DBSCAN

④ 주성분 분석

29. 상품의 가격을 조사한 데이터를 나타낸 다음의 Box Plot에 대한 설명으로 맞는 것은?

아래



① 평균-1.5 * IQR ≤ 데이터 ≤ 평균+1.5* IQR 범위를 벗어난 데이터를 이상치라고 한다.

② 평균(mean)은 160이다.

③ 3사분위수보다 높은 가격 데이터가 약 50%이상이 있다.

④ 가격의 IQR(Interquartile Range)은 60 이다.

30. 모집단을 특정한 기준에 따라 서로 상이한 소집단으로 나누고 각각의 소집단으로부터 일정한 표본을 무작위로 추출하는 표본추출방법으로 적절한 것은?

① 단순랜덤추출법

② 계층추출법

③ 집락추출법

④ 층화추출법

31. 소매점에서 물건을 배열하거나 카탈로그 및 교차판매 등에 적용하기 적합한 데이터 마이닝 기법은 무엇인가?

① 분류(classification)

② 예측(prediction)

③ 연관분석(association analysis)

④ 군집(clustering)

32. 다음 중 한 변수를 단조 증가 함수로 변환하여 다른 변수를 나타낼 수 있는 정도를 나타내며, 두 변수의 선형 관계의 크기뿐만 아니라 비선형적인 관계도 나타낼 수 있는 상관계수는 무엇인가?

① 코사인 유사도

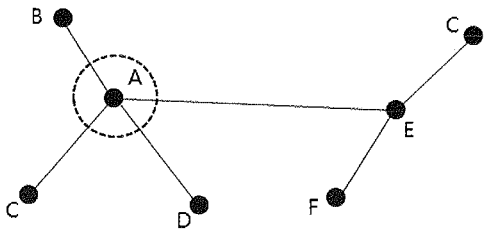
② 피어슨 상관계수

③ 스피어만 상관계수

④ 자카드 인덱스

33. 아래 사회연결망에서 노드 A의 연결정도 중심성은?

아래



① 1 ② 2 ③ 3 ④ 4

34. 계층적 군집방법은 두 개체(또는 군집) 간의 거리(또는 비유사성)에 기반하여 군집을 형성해 나가므로 거리에 대한 정의가 필요한데, 다음 중 변수의 표준화와 변수 간의 상관성을 동시에 고려한 통계적 거리로 적절한 것은?

① 표준화 거리(Standardized distance)
② 민코우스키 거리(Minkowski distance)
③ 마할라노비스 거리(Mahalanobis distance)
④ 자카드 계수(Jaccard coefficient)

35. 에어컨 회사에서 지역별 온도, 습도에 따라 고객군을 나눠서 마케팅전략을 수립할 때 적합한 분석방법은?

① 연관분석
② 회귀분석
③ 군집분석
④ 분류분석

36. R에서 matrix 명령어를 활용하여 벡터를 행렬로 아래와 같이 변환하였다고 할 때 생성된 mx의 결과로 옳은 것은 ?

아래

```
mx = matrix(c(1,2,3,4,5,6), ncol=2, byrow=T)
```

①
```
     [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
```

②
```
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

③
```
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
```

④
```
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

37. 아래의 데이터 마이닝 분석 예제 중 비지도 학습을 수행해야 하는 예제는?

아래

가) 우편물에 인쇄된 우편번호 판별 분석을 통해 우편물을 자동으로 분류
나) 고객의 과거 거래 구매 패턴을 분석하여 고객이 구매하지 않은 상품을 추천
다) 동일 차종의 수리 보고서 데이터를 분석하여 차량 수리에 소요되는 시간을 예측
라) 상품을 구매할 때 그와 유사한 상품을 구매한 고객들의 구매 데이터를 분석하여 쿠폰을 발행
마) 고장난 비행기들의 수리 이력 데이터를 분석하여 수리시간을 추정

① 나, 다
② 가, 라
③ 가, 다
④ 나, 라

38. 다음 중 다중공선성(multicollinearity)에 대한 설명으로 가장 부적절한 것은?

① 다중공선성 문제를 해결하기 위해 중요하지 않으면서 다른 변수와 상관성이 높은 변수를 제거한다.

② 표본수가 증가해도 VIF에서 일반 결정계수는 크게 변하지 않는다.

③ 두 변수의 VIF값이 "1"에 가까우면 회귀식의 기울기는 완만하다.

④ 구조적 다중공선성의 문제가 있는 경우에는 데이터의 평균 중심을 변화한다.

39. 시계열의 요소분해법은 시계열 자료가 몇 가지 변동들의 결합으로 이루어져 있다고 보고 변동요소별로 분해하여 쉽게 분석하기 위한 것이다. 다음 중 분해 요소에 대한 설명이 부적절한 것은?

① 추세분석은 장기적으로 변해가는 큰 흐름을 나타내는 것으로 자료가 장기적으로 커지거나 작아지는 변화를 나타내는 요소이다.

② 계절변동은 일정한 주기를 가지고 반복적으로 같은 패턴을 보이는 변화를 나타내는 요소이다.

③ 순환변동은 경제 전반이나 특정 산업의 부침을 나타내 주는 것을 말한다.

④ 불규칙변동은 불규칙하게 변동하는 급격한 환경변화, 천재지변 같은 것으로 발생하는 변동을 말한다.

40. 주성분분석은 p개의 변수들을 중요한 m(p)개의 주성분으로 표현하여 전체 변동을 설명하는 방법을 사용한다. 다음 중 주성분 개수(m)를 선택 방법에 대한 설명으로 가장 부적절한 것은?

① 전체 변이 공헌도(percentage of total variance) 방법은 전체 변이의 70~90% 정도가 되도록 주성분의 수를 결정한다.

② 평균 고유값(average eigenvalue)방법은 고유값들의 평균을 구한 후 고유값이 평균값 이상이 되는 주성분을 제거하는 방법이다.

③ Scree graph를 이용하는 방법은 고유값의 크기순으로 산점도를 그린 그래프에서 감소하는 추세가 원만해지는 지점에서 1을 뺀 개수를 주성분의 개수로 선택한다.

④ 주성분은 주성분을 구성하는 변수들의 계수 구조를 파악하여 적절하게 해석되어야 하며, 명확하게 정의된 해석 방법이 있는 것은 아니다.

단 답 형

* 문항 수(10문항), 배점(문항 당 2점, 부분점수 없음)

01. 인공지능의 한 분야로, 컴퓨터가 스스로 많은 데이터를 분석해서 패턴과 규칙을 찾아내고, 학습된 패턴과 규칙을 활용하여 분류나 예측을 하는 것을 무엇이라고 하는가?

( )

02. 조직 내 구성원들이 축적하고 있는 노하우 등 암묵적 지식을 형식지로 표출화 될 수 있도록 지원하는 등, 조직의 경쟁력 향상을 위해 지식자원을 체계화하고 원활하게 공유가 될 수 있도록 지원하는 시스템을 무엇이라고 하는가?

( )

03. 아래 ( ) 안에 공통적으로 들어갈 용어는?

아래

기업 및 공공기관에서는 시스템의 중장기 로드맵을 정의하기 위한 ( )을(를) 수행한다. ( )은(는) 정보기술 또는 정보시스템을 전략적으로 활용하기 위하여 조직 내·외부 환경을 분석하여 기회나 문제점을 도출하고 사용자의 요구사항을 분석하여 시스템 구축 우선순위를 결정하는 등 중장기 마스터플랜을 수립하는 절차이다.

( )

04. 데이터 분석 도입의 수준을 파악하기 위한 분석 준비도의 6가지 구성요소 중 하나로서 운영시스템 데이터 통합, 빅데이터 분석 환경, 통계분석 환경 등을 진단하는 구성요소는 무엇인가?

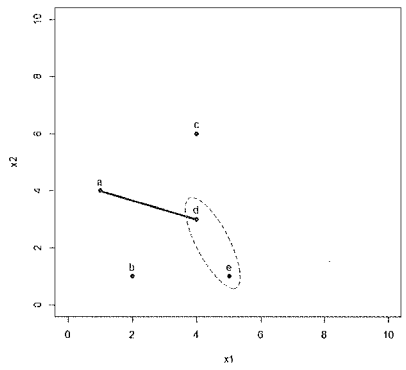
( )

05. 베이즈 정리와 특징에 대한 조건부 독립을 가설로 하는 알고리즘으로 클래스에 대한 사전 정보와 데이터로부터 추출된 정보를 결합하고 베이즈 정리를 이용하여 특정 데이터가 특정 클래스에 속하는지를 분류하는 알고리즘은 무엇인가?

( )

06. 계층적 군집분석에서 두 군집을 병합하는 방법 중, 군집과 군집, 또는 데이터와의 거리계산 시 최단거리를 계산하여 거리가 가까운 데이터, 또는 군집을 새로운 군집으로 형성하는 방법을 무엇이라고 하는가?

아래



( )

07. 텍스트 마이닝에서 어근에 차이가 있더라도 관련이 있는 단어들을 동일한 어간으로 매핑이 될 수 있도록 정해진 규칙에 따라 단어에서 어간을 분리하여 공통 어간을 가지는 단어를 묶는 작업을 무엇이라고 하는가?

( )

08. 시계열 분석을 위해서는 정상성을 만족해야 한다. 따라서 주어진 자료가 정상성을 만족하는지 판단하는 과정이 필요하다. 자료가 추세를 보이는 경우에는 현 시점의 자료값에서 전 시점의 자료를 빼는 방법을 통해 비정상시계열을 정상시계열로 바꾸어 준다. 이 방법은 무엇인가?

( )

09. 아래는 주성분 분석을 수행한 결과이다. 첫 번째 분산은 전체 분산의 몇 %를 설명하고 있는가? (소수점 첫째자리까지 표시하시오)

아래

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.5574873	0.9943214	0.5943221	0.4123679
Proportion of Variance	0.5748331	0.2321003	0.1834561	0.0096105
Cumulative Proportion	0.5748331	0.8069334	0.0096105	1.0000000

( )

10. 원 데이터 집합으로부터 크기가 같은 표본을 여러 번 단순 임의 복원추출하여 각 표본에 대해 분류기를 생성한 후 그 결과를 앙상블하는 방법을 무엇이라 하는가?

( )