감독확인란

제32회 데이터 분석 준전문가 자격검정 시험 복원 문제

- 검정일시 : 2022. 02. 26(토) / 10:00~11:30
- 수험번호 :
- 성　　명 :

과목 I 데이터 이해 * 문항 수(8문항), 배점(문항 당 2점)

01. 아래 SQL 명령 중 DML에 해당하는 항목은?

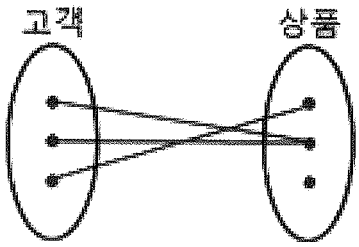
아래

A. UPDATE
B. SELECT
C. INSERT
D. DELETE
E. CREATE

① A, B
② A, B, C
③ A, B, C, D
④ A, B, C, D, E

02. 아래는 고객과 상품의 대응관계를 도식화한 것이다. 대응비(cardinality Ratio) 관점에서 둘 간의 관계가 옳은 것은?

아래



① 1 : 1
② N : 1
③ 1 : N
④ N : N

03. 아래는 데이터베이스를 기반으로 기업 내 구축되는 주요 정보시스템 중 하나를 설명한 것이다. 아래의 보기에서 가장 적합한 것을 고르시오.

아래

기업 전체를 경영자원의 효과적 이용이라는 관점에서 통합적으로 관리하고 경영의 효율화를 기하기 위한 시스템

① ERP
② CRM
③ SCM
④ KMS

04. 다음 중 딥러닝과 가장 관련 없는 분석 기법은?

① CNN(Convolutional Neural Network)
② LSTM(Long Short-Term Memory)
③ SVM(Support Vector Machine)
④ Autoencorder

05. 머신러닝 알고리즘은 크게 지도학습(Supervised learning)과 비지도학습(Unsupervised learning)으로 나눌 수 있다. 이러한 측면에서 보기 중 나머지와 성격이 다른 것은?

① 군집분석
② 판별분석
③ 회귀분석
④ 분류분석

06. 빅데이터 활용에 필요한 기본적인 3요소로 가장 적절한 것은?

① 데이터, 기술, 인력
② 데이터, 기술, 프로세스
③ 기술, 인력, 프로세스
④ 데이터, 인력, 프로세스

07. 데이터 사이언스에서 인문학적 사고는 '전략 인사이트 도출'을 위해 반드시 필요한 요소이다. 다음 중 인문학 열풍을 가져오게 한 외부 환경 요소로 가장 부적절한 것은?

① 디버전스 동역학이 작용하는 복잡한 세계화
② 빅데이터 분석 기법의 이해와 분석 방법론 확대
③ 경제의 논리가 생산에서 최근 패러다임인 시장 창조로 변화
④ 비즈니스 중심이 제품생산에서 체험 경제를 기초로 한 서비스로 이동

08. 다음 중 데이터 사이언티스트가 하는 일로 가장 적절하지 않은 것은?

① 다분야 간 협력을 통해 빅데이터의 가치를 실현한다.
② 알고리즘에 의해 부당하게 피해입은 사람을 구제한다.
③ 데이터를 시각화해 설득력을 높여 정보를 전달한다.
④ 빅데이터를 다각적으로 분석하여 인사이트를 도출한다.

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

09. 다음 중 거시적 관점의 메가트랜드에 해당하지 않는 것은?

① 사회(Social)
② 기술(Technological)
③ 환경(Environmental)
④ 채널(Channel)

10. 다음 중 분석 기획 단계의 비즈니스 이해 및 범위설정 태스크에서 프로젝트 범위 설정의 산출물은 무엇인가?

① WBS(Work Breakdown Structure)
② SoW(Statement of Works)
③ Phase
④ ERD(Entity Relationship Diagram)

11. 데이터 표준화에 대한 설명으로 가장 적절한 것은?

① 데이터 표준화란 데이터 정합성 및 활용의 효율성을 위하여 표준 데이터를 포함한 메타 데이터와 데이터 사전의 관리 원칙을 수립하는 것이다.
② 데이터 표준 용어 설정, 명명 규칙 수립, 메타 데이터 구축, 데이터 사전 구축 등의 업무로 구성된다.
③ 메타데이터 및 표준데이터를 관리하기 위한 전사 차원의 저장소를 구축하는 것이다.
④ 데이터 거버넌스 체계를 구축한 후 표준 준수 여부를 주기적으로 점검하고 모니터링 하는 것이다.

12. 빅데이터의 특성을 고려한 분석 ROI 요소에서 투자비용 요소로 적절하지 않은 것은 무엇인가?

① Volume
② Variety
③ Velocity
④ Value

13. 전사 차원의 모든 데이터에 대하여 정책 및 지침, 표준화, 운영조직 및 책임 등의 표준화된 관리체계를 수립하고 운영을 위한 프레임 워크 및 저장소를 구축하는 것을 말하는 것은 무엇인가?

① 데이터 관리 체계
② 분석 마스터 플랜
③ 데이터 저장소
④ 데이터 거버넌스

14. 다음 중 빅데이터 분석 방법의 절차 5단계를 순서대로 나타낸 것은?

① 분석 기획 → 데이터 준비 → 데이터 분석 → 시스템 구현 → 평가 및 전개
② 분석 기획 → 데이터 준비 → 시스템 구현 → 데이터 분석 → 평가 및 전개
③ 분석 기획 → 데이터 준비 → 데이터 분석 → 평가 및 전개 → 시스템 구현
④ 분석 기획 → 데이터 모델링 → 데이터 준비 → 데이터 분석 → 평가 및 전개

15. 다음 중 데이터 분석 기회 선별 방식으로 틀린 것은?

① 톱다운 접근 기반의 특징과 경쟁력에 따른 후보 기회를 선택
② 유즈 케이스 벤치마킹 산업별, 업무별 벤치마킹을 통한 기회 선택
③ 유즈 케이스 벤치마킹 동일 업종의 비교분석을 통해 기회 선택
④ 톱다운 접근 기반의 특정 주제별로 분석 기회를 선택

16. 분석 수준 진단의 대상으로 적절하지 않은 것은?

① 분석 성과에 대한 조사
② 분석 업무 수행에 대한 조사
③ 분석 인력 및 조직에 대한 조사
④ 분석 인프라에 대한 조사

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

17. 다음 가설검정 용어 중 '귀무가설이 옳은데도 이를 기각하는 확률의 크기'는 어느 용어인가?

① 제 2종 오류
② 검정통계량
③ 기각역
④ 유의수준

18. R에서 데이터 타입이 같지 않은 객체들을 하나의 객체로 묶어놓을 수 있는 자료구조는 어떤 것인가?

① 행렬(Matrix)
② 배열(Array)
③ 리스트(List)
④ 문자열(String)

19. 다음 중 오분류표의 평가지표 중 True로 예측한 관측치 중 실제 True인 지표를 무엇이라고 하는가?

① Precision
② Specificity
③ Recall
④ Sensitivity

20. 아래 거래 전표에서 연관규칙 'A→B'의 신뢰도(Confidence)는?

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

① 20% ② 30% ③ 40% ④ 50%

21. 다음 중 시계열 데이터에 대한 설명으로 가장 부적절한 것은?

① 시계열 데이터의 모델링은 다른 분석모형과 같이 탐색 목적과 예측 목적으로 나눌 수 있다.
② 짧은 기간 동안의 주기적인 패턴을 계절변동이라 한다.
③ 잡음(noise)은 무작위적인 변동이지만 일반적으로 원인은 알려져 있다.
④ 시계열분석의 주목적은 외부인자와 관련해 계절적인 패턴, 추세와 같은 요소를 설명할 수 있는 모델을 결정하는 것이다.

22. 아래의 오분류표를 이용하여 민감도(Sensitivity)를 구하시오.

		예측치		합계
		True	False	
실제값	True	40	60	100
	False	60	40	100
합계		100	100	200

① 0.25
② 0.3
③ 0.4
④ 0.55

23. 아래 거래 전표에서 연관규칙 '커피→우유'의 향상도(Lift)는?(단, 나누어 떨어지지 않을 경우 소수점 첫째 자리에서 반올림)

물품	거래건수
{커피}	100
{우유}	100
{녹차}	100
{커피, 우유, 녹차}	50
{우유, 녹차}	200
{커피, 우유}	250
{커피, 녹차}	200

① 30%　② 50%　③ 83%　④ 100%

24. 카탈로그 배열, 교차 판매 등의 마케팅을 계획할 때 적절한 데이터 마이닝 기법은 무엇인가?

① 분류
② 추정
③ 군집
④ 연관분석

25. 분류모형의 성과 분석 중 ROC Curve는 x축에 FP Ratio, y축에는 민감도를 나타낸다. 아래와 같은 오분류표가 있을 때 특이도를 계산하는 방식으로 가장 적절한 것은?

		예측치		합계
		True	False	
실제값	True	TP	FN	P
	False	FP	TN	N
합계		P'	N'	P+N

① (TP+TN)÷(P+N)
② TN÷N
③ TP÷(TP+FP)
④ TP÷P

26. 모형기반(Model-based)의 군집방법으로 데이터가 k개의 모수적 모형의 가중합으로 표현되는 모집단 모형으로부터 나왔다는 가정 하에서 모수와 함께 가중치를 자료로부터 추정하는 방법으로 사용하는 군집 방법은 무엇인가?

① k-평균군집(k-Means Clustering)
② 혼합 분포 군집(Mixture Distribution Clustering)
③ 계층적 군집(Hierarchical Clustering)
④ 분리 군집(Partitioning Clustering)

27. 다음 중 비모수 검정 방법으로 부적절한 것은?

① 맨-휘트니 U검정
② 런 검정
③ 윌콕슨의 순위합 검정
④ 카이제곱검정

28. 거리를 이용하여 데이터 간 유사도를 측정할 수 있는 척도는 데이터의 속성과 구조에 따라 적합한 것을 사용해야 한다. 다음 중 유사도 측도에 대한 설명으로 부적절한 것은?

① 유클리드 거리는 두 점을 잇는 가장 짧은 직선거리이다. 공통으로 점수를 매긴 항목의 거리를 통해 판단하는 측도이다.
② 맨하튼 거리는 각 방향 직각의 이동 거리 합으로 계산된다.
③ 표준화 거리는 각 변수를 해당 변수의 표준편차로 변환한 후 유클리드 거리를 계산한 거리이다. 표준화를 하게 되면 척도의 차이, 분산의 차이로 인해 왜곡을 피할 수 있다.
④ 마할라노비스 거리는 변수의 표준편차를 고려한 거리 측도이나 변수 간에 상관성이 있는 경우에는 표준화 거리 사용을 검토해야 한다.

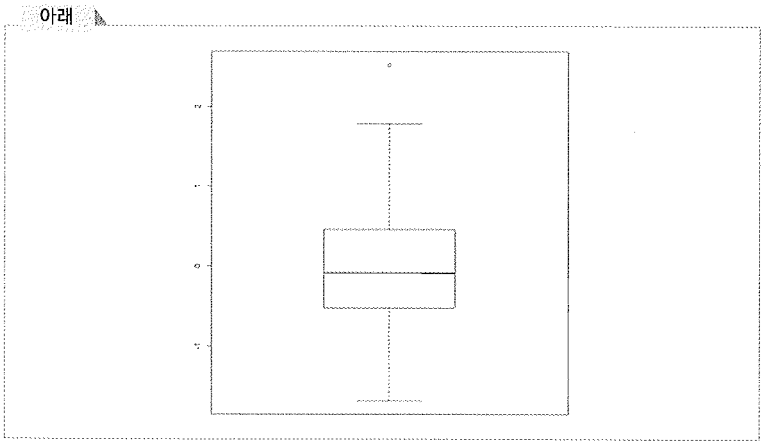
29. 다음 중 시간의 흐름에 따라 관측된 데이터에 관한 것으로 적절한 것은?

① 질적 자료
② 시계열 자료
③ 양적 자료
④ 횡단면 자료

30. 다음 데이터 마이닝의 대표적인 기능 중 이질적인 모집단을 세분화하는 기능으로 적절한 것은?

① 분류분석
② 모수추정
③ 군집분석
④ 연관분석

31. 아래의 상자수염 그림에서 상자 안에 그어진 선이 의미하는 것은 무엇인가?



① Minimum
② Mean
③ Median
④ Maximum

32. 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. 확률화란 모집단으로부터 편의되지 않은 표본을 추출하는 절차를 의미하며 확률화 절차에 의해 추출된 표본을 확률표본(random sample)이라 한다.

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.

33. 다음 중 k-폴드 교차검증(k-fold Cross Validation)에 대한 설명으로 가정 적절하지 않은 것은?

① 모형이 데이터에 과적합하는 문제를 해결하기 위한 방법이다.

② K=2인 경우, LOOCV(Leave-One-Out Cross-Validation)이라고 한다.

③ 하나의 그룹을 검증용 셋(Validation set)으로, K-1개 그룹을 훈련용 셋(Train set)으로 사용하여 K번 반복 측정하고 결과를 평균 낸 값을 최종 평가로 사용한다.

④ 데이터 셋을 K개의 그룹으로 분할한다.

34. 다음 중 주성분분석에 대한 설명으로 부적절한 것은?

① 차원축소 방법 중 하나이다.

② 비지도학습(unsupervised learning)에 해당한다.

③ 이론적으로 주성분 간 상관관계가 없다.

④ 원변수의 선형결합 중 가장 분산이 작은 것을 제1주성분(PC1)으로 설정한다.

35. 아래는 k-평균군집을 수행하는 절차를 단계별로 기술한 것이다. 다음 중 k-평균군집 수행 절차로 가장 올바른 것은?

아래

가. 각 자료를 가장 가까운 군집 중심에 할당한다.
나. 군집 중심의 변화가 거의 없을 때(또는 최대 반복 수)까지 단계2와 단계3를 반복한다.
다. 초기 (군집의) 중심으로 k개의 객체를 임의로 선택한다.
라. 각 군집 내의 자료들의 평균을 계산하여 군집의 중심을 업데이트한다.

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

36. 이상값 탐색을 위해 상자그림(boxplot)을 사용하려 한다. 아래와 같은 데이터 요약 결과가 있을 때, 다음 중 이상값을 판단하는 하한선, 상한선으로 옳은 것은?

아래

```
>summary(x)
Min.  1st Qu.  Median  Mean   3rd Qu.   Max.
0     4        7       9.615  12        39
```

① (-12, 36)
② (4, 12)
③ (-2, 30)
④ (-8, 24)

37. 다음 중 중앙 50%의 데이터들이 흩어진 정도를 의미하는 것은?

① 중앙값(median)
② 사분위수 범위(Interquantile Range)
③ 표준편차(Standard Deviation)
④ 평균(Mean)

38. 아래에서 설명하는 통계분석의 방법은 무엇인가?

아래

- 고차원의 데이터를 저차원의 데이터로 변환시키는 통계적 기법
- 원래의 변수들을 선형결합으로 새로운 변수들을 생성함
- 전체 변수의 사용 대신 도출되는 몇 개의 새로운 변수만의 사용으로 분석을 대신할 수 있음

① 카이제곱 분석

② 회귀 분석

③ 주성분 분석

④ 분산 분석

39. 다음 중 파생변수에 대한 설명 중 부적절한 것은?

① 많은 모형에서 공통적으로 사용될 수 있다.

② 주관적일 수 있으므로 논리적 타당성을 갖추어야한다.

③ 세분화, 고객 행동 예측, 캠페인 반응예측에 잘 활용된다.

④ 특정 상황에서만 유의미하지 않게 대표성을 갖도록 해야 한다.

40. 시계열 분석에서 정상성 기준에 대한 설명 중 적절하지 않은 것은?

① 시계열 자료 간에 독립성 조건을 충족한다.

② 모든 시점에 일정한 평균을 갖는다.

③ 분산이 시점에 의존하지 않고 일정하다.

④ 공분산이 시점 s에 의존하지 않고 단지 시차에만 의존한다.

단 답 형 * 문항 수(10문항), 배점(문항 당 2점, 부분점수 없음)

01. 아래에 설명하는 (가)는 무엇인가?

아래

이것은 인터넷에 연결된 기기가 사람의 개입 없이 상호간에 알아서 정보를 주고 받아 처리한다. 구글의 Google Glass, 나이키의 Fuel band 등이 있다.

( )

02. 아래는 기업 내부에서 활용되는 데이터베이스의 활용에 대한 설명이다. (가) 에 들어갈 말로 적절한 것은 무엇인가?

아래

(가)은 기업이 외부 공급업체 또는 제휴업체와 통합된 정보시스템으로 연계하여 시간과 비용을 최적화시키기 위한 것으로, 자재 구매, 생산, 제고, 유통, 판매, 고객 데이터로 구성된다.

( )

03. 합리적 의사결정을 방해하는 요소로 표현방식 및 발표자에 따라 동일한 사실(Fact)에도 판단을 달리하는 현상을 이르는 말은?

( )

04. 아래의 (㉠)에 들어갈 용어로 적절한 것은?

아래

분석적 기업으로 도약을 위해서는 가장 먼저 조직의 분석(Analytics) 도입 여부 및 활 수준에 대한 명확한 진단이 요구된다. 특히 분석 수준 진단 방법 중 조직의 분석 및 활용을 위한 역량수준을 파악하기 위해 '도입 → (㉠)→ 확산 → 최적화' 의 분석 성숙도(Maturity) 단계 포지셔닝을 파악한다.

( )

05. 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법은 무엇인가?

( )

06. 다음 내용이 설명하고 있는 것을 적으시오.

아래

- 시계열 모델 중 자기 자신의 과거 값을 사용하여 설명하는 모형임
- 백색 잡음의 현재값과 자리 자신의 과거값의 선형 가중합으로 이루어진 정상확률 모형
- 모형에 사용하는 시계열 자료의 시점에 따라 1차, 2차, ..., p차 등을 사용하나 정상시계열 모형에서는 주로 1, 2차를 사용함

( )

07. 아래 ( )에 들어갈 적절한 용어는?

아래

의사결정나무에서 끝마디가 너무 많으면 모형에 ( )인 상태로 현실문제에 적용될 수 있는 적절한 규칙이 나오지 않게 된다. 따라서 분류된 관측치의 비율 또는 MSE(Mean Square Error) 등을 고려하여 적절한 수준의 가지치기 규칙을 제공해야 한다.

( )

08. 데이터 마이닝 기법 중 동물의 뇌신경계를 모방하여 분류(또는 예측)을 위해 만들어진 모형은?

( )

09. 분류 분석 모형을 사용하여 분류된 관측치가 각 등급별로 얼마나 포함되는지를 나타내는 도표는?

( )

10. 아래에서 설명하는 이것은?

아래

이것은 데이터 웨어하우스 환경에서 정의된 접근 계층으로, 데이터 웨어하우스에서 데이터를 꺼내 사용자에게 제공하는 역할을 한다. 보통 특정한 조직 혹은 팀에서 사용하는 것을 목적으로 한다.

( )