

## 제34회 데이터 분석 준전문가 자격검정 시험 복원 문제

• 검정일시 : 2022. 08. 28(일) / 10:00~11:30

• 수험번호 :

• 성 명 :

**01. 다음 중 빅데이터 분석의 특성에 대한 설명으로 가장 부적절한 것은?**

- ① 더 많은 정보가 더 많은 가치를 창출하는 것은 아니다.
- ② 비즈니스 핵심에 대해 보다 객관적이고 종합적인 통찰력을 줄 수 있는 데이터를 찾는 것이 중요하다.
- ③ 빅데이터 과제와 관련된 주된 걸림돌은 비용이 아니다.
- ④ 데이터가 커질수록 분석에 많이 사용되고 이것이 경쟁우위를 가져다주는 원천이다.

**02. 다음 중 데이터의 가치 측정이 어려운 이유로 적절하지 않은 것은 무엇인가?**

- ① 데이터의 재사용의 일반화로 특정 데이터를 언제 누가 사용했는지 알기 힘들기 때문이다.
- ② 빅데이터 전문인력의 증가로 다양한 곳에서 빅데이터가 활용되고 있기 때문이다.
- ③ 분석기술의 발전으로 과거에 분석이 불가능 했던 데이터를 분석할 수 있게 되었기 때문이다.
- ④ 빅데이터는 기존에 존재하지 않던 새로운 가치를 창출하기 때문이다.

**03. 다음 중 DBMS(Database Management System)에 관한 설명 중 틀린 것은?**

- ① 데이터베이스는 정의, 조작, 제어라는 3가지 필수 기능이 있다.
- ② 데이터베이스를 관리하고 운영하는 소프트웨어를 말한다.
- ③ 데이터베이스에 있는 모든 데이터는 분석이 가능하다.
- ④ 계층형(Hierarchical), 망형(Network), 관계형(Relational), 객체지향형(Object-Oriented), 객체관계형(Object-Relational) 등으로 분류된다.

**04. 빅데이터의 특성에 대한 설명으로 부적절한 것은?**

- ① 비즈니스 핵심에 대해 보다 객관적이고 종합적인 통찰력을 줄 수 있는 데이터를 찾는 것이 중요하다.
- ② 빅데이터 분석은 일차적인 분석으로는 불충분하다.
- ③ 기업에서의 빅데이터 분석은 기업의 분석 문화에 결정적으로 영향을 받는다.
- ④ 더 많은 정보가 더 많은 가치를 창출하는 것은 아니다.

05. 빅데이터 시대에 발생할 수 있는 위기요인과 예시로 적절하지 않은 것은?

아래

- (가) 사생활침해 : 카드사의 개인정보가 유출되 SMS, email 등으로 관련없는 광고정보 전송
- (나) 책임원칙훼손 : 범죄예측 프로그램에 의해 은행에서 대출자의 신용도와 무관하고 부당하게 대출을 거절
- (다) 책임원칙훼손 : 구글은 이미 서비스 이용자가 1시간 뒤에 어떤 일을 할지 87% 정확도로 예측할 수 있음
- (라) 데이터오용 : 개인정보를 무단으로 크롤링하여 활용

- ① (가), (나)
- ② (나), (라)
- ③ (가), (다)
- ④ (다), (라)

06. 데이터 사이언스에 대한 설명으로 가장 부적절한 것은?

- ① 데이터 사이언스는 데이터로부터 의미있는 정보를 추출하는 학문이다.
- ② 주로 분석의 정확성에 초점을 두고 진행한다.
- ③ 정형데이터 뿐만 아니라 다양한 데이터를 대상으로 한다.
- ④ 기존의 통계학과는 달리 총체적 접근법을 사용한다.

07. 다음 중 데이터 웨어하우스와 데이터 마트에 대한 설명으로 부적절한 것은?

- ① 데이터 마트는 모든 사용자 그룹에 서비스를 제공하는 데이터 웨어하우스 논리 모델을 지향한다.
- ② 데이터웨어하우스에서 관리하는 데이터들은 시간의 흐름에 따라 변화하는 값을 저장한다.
- ③ 데이터 마트는 특정 분야에 집중하고 있기 때문에 해당 분야에 대한 전문성만 갖추고 있다면 구축하는 것이 용이하다.
- ④ 데이터웨어하우스는 사용자의 의사결정에 도움을 주기 위해 정보를 기반으로 제공하는 하나의 통합된 데이터 저장 공간을 말한다.

08. 다음 중 데이터에 대한 설명으로 부적절한 것은?

- ① 1바이트는 256 종류의 서로 다른 값을 표현할 수 있는 데이터의 크기를 의미한다.
- ② 수치 데이터는 용량이 증가하더라도 텍스트 데이터에 비해 DBMS에 관리하기 용이하다.
- ③ 데이터가 많을수록 더 많은 가치가 창출된다.
- ④ 인터넷 댓글은 그 형태와 형식이 정해져 있지 않아 비정형 데이터라고 한다.

09. 다음 중 하향식 접근법의 내용으로 적절한 것은?

- ① 문제탐색 단계에서는 발생하는 가치에 중점을 두는 것이 아니라 세부적인 구현 및 솔루션에 초점을 둔다.
- ② 분석 역량을 확보하였으며, 기존의 분석 기법 및 시스템이 존재하지 않는다면 전문업체 Sourcing이 필요하다.
- ③ 타당성 검토 단계에서는 복잡한 문제이기 때문에 다양한 사람들의 의견 조합이 필요하다.
- ④ 분석 유즈 케이스는 분석 기회들을 구체적인 과제로 만들고 난 뒤에 표기한다.

10. 아래의 분석과제 관리를 위한 5가지 주요 영역의 내용 중 옳은 것은?

아래

- 가) 분석과제 관리를 위한 5가지 주요 영역은 Size, Complexity, Speed, Analytic Complexity, Accuracy&Precision이다.
- 나) 초기 데이터의 확보와 통합 뿐 아니라 해당 데이터에 잘 적용될 수 있는 분석 모델의 선정을 고려해야 한다.
- 다) Precision은 모델을 지속적으로 반복했을 때의 편차의 수준으로써 정확도를 의미한다.
- 라) 분석 모델의 정확도와 복잡도는 트레이드 오프(Trade-off)관계가 존재한다.

- ① 가
- ② 가, 나
- ③ 가, 나, 다
- ④ 가, 나, 라

11. 하향식 데이터 분석기획에서 문제 탐색 단계에 대한 설명으로 가장 부적절한 것은?

- ① 빠짐없이 문제를 도출하고 식별하는 것이 중요
- ② 문제를 해결함으로써 발생하는 가치에 중점을 두는 것이 중요
- ③ 비즈니스 모델 캔버스는 문제 탐색 도구로 활용
- ④ 문제 탐색은 유스케이스 활용보다는 새로운 이슈 탐색이 우선

12. 기업의 데이터 분석과제 수행을 위한 수준을 평가하기 위하여 분석 준비도(Readiness)를 파악하게 된다. 다음 중 데이터 분석 준비도 프레임워크에서 분석 업무 파악 영역으로 가장 부적절한 것은?

- ① 최적한 분석 업무
- ② 업무별 적합한 분석 기법
- ③ 예측 분석 업무
- ④ 발생한 사실 분석 업무

13. 다음 중 빅데이터 분석 방법론 중 시스템 구현에 대한 설명 중 가장 적절하지 않은 것은?

- ① 시스템 구현단계에는 설계 및 구현, 시스템 테스트 및 운영으로 이루어져 있다.
- ② 시스템 설계서를 바탕으로 BI 패키지를 활용하거나 새롭게 프로그램 코딩을 통하여 시스템을 구축한다.
- ③ 정보 보호 및 시스템 성능은 시스템 구현 단계에 해당된다.
- ④ 정보보안영역과 코딩은 시스템 구현 단계에서 주요 고려사항이다.

14. 다음 중 분석과제 정의서에 포함되지 않는 것은?

- ① 분석 수행주기
- ② 데이터 수집 난이도
- ③ 상세 알고리즘
- ④ 분석결과 검증 요녀십

15. 다음 중 데이터 분석 과제에서 프로젝트 관리에 대한 설명으로 가장 부적절한 것은?

- ① 분석 과제는 분석 전문가의 상상력을 요구하므로 일정을 제한하는 일정계획은 적절하지 못하다.
- ② 분석 과제는 많은 위험이 있어 사전에 위험을 식별하고 대응방안을 수립해야 한다.
- ③ 분석 과제는 적용되는 알고리즘에 따라 범위가 변할 수 있어 범위관리가 중요하다.
- ④ 분석 과제에서 다양한 데이터를 확보하는 경우가 있어 조달관리 또한 중요하다.

### 16. 다음 중 CRISP-DM의 설명으로 부적절한 것은?

- ① CRISP-DM 프로세스 중 Business Understanding, Data understanding 단계 간에는 피드백이 가능하다.
- ② 데이터 준비 단계에서는 데이터 정제, 데이터 탐색, 데이터 셋 편성 등의 수행업무가 있다.
- ③ 모델링 단계에서는 테스트용 데이터 셋으로 평가하여 모델의 과적합 문제를 확인한다.
- ④ CRISP-DM은 계층적 프로세스 모델로써 4개의 레벨로 구성되며, 6단계의 프로세스를 가진다.

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

### 17. 자료의 특징이나 분포를 한 눈에 보기 쉽도록 시각화하는 작업은 매우 중요하다. 다음 중 상자 그림(box plot)에 대한 설명으로 가장 부적절한 것은?

- ① 자료의 크기 순서를 나타내는 5가지 통계량(최소값, 최대값, 1사분위수, 중앙값, 3사분위값)을 이용하여 시각화하는 방법이다.
- ② 이상치를 판단하기에는 적합하지 않다.
- ③ 사분위수를 한 눈에 볼 수 있다.
- ④ 자료의 범위를 개량적으로 알 수 있다.

### 18. 아래의 거래 내역에서 지지도가 25%, 신뢰도가 50% 이상인 규칙은?

물 품	거래건수
{A}	10
{B}	5
{C}	25
{A, B, C}	5
{B, C}	20
{A, B}	20
{A, C}	15

- ①  $A \rightarrow B$
- ②  $A \rightarrow C$
- ③  $C \rightarrow A$
- ④  $B \rightarrow C$

19. 다음 중 시계열 모형에 대한 설명 중 옳은 것은?

- ① ARIMA의 약어는 AutoRegressive Improved Moving Average이다.
- ② 분해시계열은 일반적인 요인을 분리하여 분석하는 방법으로 회귀분석적인 방법과는 다르게 사용한다.
- ③ ARIMA 모형에서는 정상성을 확인할 필요가 없다.
- ④ ARIMA 모형에서  $p=0$ 일 때, IMA( $d, q$ ) 모형이라고 부르고,  $d$ 번 차분하면 MA( $q$ )모형을 따른다.

20. ROC(Receiver Operating Characteristic) 그래프에서 이상적으로 완벽히 분류한 모형의 x축과 y축 값으로 옳은 것은?

- ① (0, 0)
- ② (0, 1)
- ③ (1, 0)
- ④ (1, 1)

21. 다음 중 연관 분석의 장점으로 가장 부적절한 것은?

- ① 조건 반응(if-then)으로 표현되어 결과를 이해하기 쉽다.
- ② 목적변수가 없어 분석 방향이나 목적이 없어도 적용이 가능하다.
- ③ 품목 세분화에 관계없이 의미 있는 규칙 발견이 가능하다.
- ④ 분석을 위한 계산이 상당히 간단하다.

22. K-Nearest Neighbor 방법에 대한 설명으로 틀린 것은?

- ① 훈련 데이터에서 미리 모형을 학습하지 않고 새로운 자료에 대한 예측 및 분류를 수행할 때 모형을 구성하는 lazy learning 기법을 사용한다.
- ② 주변의 가장 가까운 K개의 데이터를 보고 데이터가 속한 그룹을 판단하는 알고리즘이다.
- ③ 그룹을 모르는 데이터 P에 대해 이미 그룹이 알려진 데이터 중 P와 가장 가까이 있는 k개의 데이터를 수집하여 그룹을 예측한다.
- ④ K값이 커질수록 과대적합(Overfitting)의 문제가 발생한다.

23. 아래의 수식에 알맞은 함수는 무엇인가?

아래

$$y = \frac{1}{1 + \exp(-x)}$$

- ① tanh 함수
- ② softmax 함수
- ③ sigmoid 함수
- ④ ReLU 함수

24. 다음 중 군집분석에 대한 설명으로 부적절한 것은?

- ① 군집분석에서는 군집의 개수나 구조에 대한 가정없이 다변량 데이터로부터 거리 기준에 의한 자발적인 군집화를 유도한다.
- ② 군집 결과에 대한 안정성을 검토하는 방법은 교차타당성을 이용하는 방법을 생각할 수 있다. 데이터를 두 집단으로 나누어 각 집단에서 군집분석을 한 후 합쳐서 군집분석한 결과와 비교하여 비슷하면 결과에 대한 안정성이 있다고 할 수 있다.
- ③ 군집의 분리가 논리적인가를 살펴보기 보다는 군집의 안정성이 더 중요하다고 할 수 있다.
- ④ 개체를 분류하기 위한 명확한 기준이 존재하지 않거나 기준이 밝혀지지 않은 상태에서 유용하게 이용할 수 있다.

25. 다음 중 잔차분석의 오차 정규성 검정에서 옳지 않은 것?

- ① Q-Q Plot으로 대략적인 확인이 가능하다.
- ② 잔차의 히스토그램이나 점도표를 그려서 정규성 문제를 검토하기도 한다
- ③ 정규성을 검정하는 방법으로 Shapiro-Wilk test, anderson-darling test 등을 이용할 수 있다.
- ④ 정상성을 만족하지 않을 때는 종속변수와 상관계수가 높은 독립변수를 제거한다.



**26. 다음 중 의사 결정나무모형에 대한 설명으로 부적절한 것은?**

- ① 의사결정나무 모형은 지도학습 모형으로 상향식 의사결정 흐름을 가지고 있다는 특징을 가지고 있다.
- ② 이익도표 또는 검정용 자료에 의한 교차타당성 등을 이용해 의사결정나무를 평가한다.
- ③ 가지치기는 분류 오류를 크게 할 위험이 높거나 부적절한 규칙을 가지고 있는 가지를 제거하는 작업이다.
- ④ 대표적인 적용 사례는 대출신용평가, 환자 증상 유추, 채무 불이행 가능성 예측 등이 있다.

**27. 다음 중 앙상블 기법에 대한 설명으로 적절한 것은?**

- ① 앙상블 기법을 사용하게 되면 각 모형의 상호 연관성이 높을수록 정확도가 향상된다.
- ② 전체적인 예측값의 분산을 유지하여 정확도를 높일 수 있다.
- ③ 대표적인 앙상블 기법은 배깅, 부스팅이 있다.
- ④ 랜덤 포레스트는 앙상블 기법 중 유일한 비지도학습 기법이다.

**28. 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)

**29. 다음 중 회귀모형을 해석하는 방법으로 옳지 않은 것은?**

- ① 모형이 통계적으로 유의미한가?
- ② 모형이 데이터를 잘 적합하고 있는가?
- ③ 모형의 종속변수, 독립변수 간의 상관계수가 유의한가?
- ④ 모형이 선형성, 정상성, 독립성을 만족하는가?

### 30. 시계열 분석에 관한 설명 중 틀린 것은?

- ① AR모형은 과거의 값이 현재의 값에 영향을 줄 때 사용하며, MA모형은 오차를 이용해 회귀식을 만드는 방법이다.
- ② ARMA모형은 약한 정상성을 가진 확률적 시계열을 표현하는데 사용한다.
- ③ 대부분의 시계열은 비정상 자료이다. 그러므로 비정상 자료를 정상성 조건에 만족시켜 정상 시계열로 만든 후 시계열 분석을 한다.
- ④ 지수평활법은 특정 기간 안에 속하는 시계열에 대해서는 동일한 가중치를 부여한다.

### 31. 다음 중 회귀분석에 대한 설명으로 가장 부적절한 것은?

- ① 독립변수의 수가 많아지면 모델의 설명력이 증가하지만 모형이 복잡해지고, 독립변수들 간에 서로 영향을 미치는 다중공선성의 문제가 발생하므로 상대적인 조정이 필요하다.
- ② 잔차와 독립변수는 상관관계가 있다면 분석이 잘 된 모형이라고 할 수 있다.
- ③ 명목형 변수는 회귀분석에서 더미변수화 하여 사용할 수 있다.
- ④ 총변동에서 추정된 회귀식에 의해 설명되는 변동의 비율로 나타낼 수 있다.

### 32. 아래는 1988년 서울올림픽에서의 여자 육상 7종 경기의 기록 데이터를 사용한 주성분분석 결과이다. 다음의 설명 중 가장 부적절한 것은?

아래

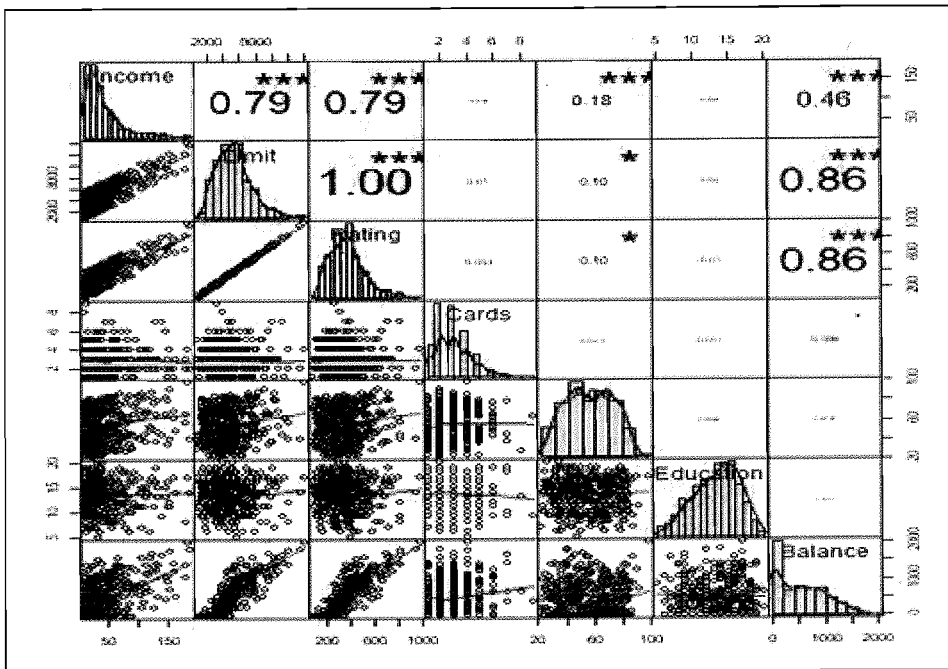
```
heptathlon_pca <- prcomp(heptathlon2[, -score], scales=TRUE)
Summary(heptathlon_pca)
```

importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.079	0.948	0.911	0.641	0.544	0.317	0.242
Portion of Variance	0.618	0.128	0.119	0.044	0.042	0.016	0.009
Cumulative preportion	0.618	0.746	0.865	0.931	0.973	0.990	1.000

- ① 한 개의 주성분으로 자료를 축약할 때 전체 분산의 61.8%가 설명 가능하다.
- ② 두 개의 주성분으로 자료를 축약할 때 전체 분산의 12.8%가 설명 가능하다.
- ③ 정보손실을 20% 이하로 변수 축약을 한다면 세 개의 주성분을 사용하는 것이 적당하다.
- ④ 첫번째 주성분의 분산이 가장 크다.

33. Credit 데이터는 400명의 신용카드 고객에 대한 신용카드와 관련된 변수들이 포함되어 있다. 아래 변수간의 산점도와 피어슨 상관계수를 나타내고 있다. 아래 그림을 보고 설명이 부적절한 것은?



- ① Income의 분포는 아래쪽으로 꼬리가 긴 분포를 가진다.
- ② Limit와 Rating은 거의 완벽한 선형관계를 가진다.
- ③ Balance와 가장 상관관계가 높은 변수는 Income이다.
- ④ Age와 Balance는 거의 상관관계가 없다.

34. 아래 오분류표에서 재현율(Recall)로 가장 적절한 것은?

		예측치		합계
		True	False	
실제값	True	40	60	100
	False	60	40	100
합계		100	110	200

- ① 0.15
- ② 0.3
- ③ 0.4
- ④ 0.55

35. 아래는 Apriori 알고리즘의 분석 순서이다. 다음 중 수행 순서를 순서대로 올바르게 나열한 것은?

아래

- 가. 최소 지지도를 설정한다.
- 나. 반복적으로 수행하여 최소지지도가 넘는 빈발품목집합을 찾는다.
- 다. 찾은 개별 품목만을 이용해 최소 지지도가 넘는 2가지 품목을 찾는다.
- 라. 찾은 품목 집합을 결합하여 최소 지지도를 넘는 3가지 품목집합을 찾는다.
- 마. 개별 품목 중에서 최소 지지도가 넘는 모든 품목을 찾는다.

- ① 가-나-다-라-마
- ② 가-나-마-다-라
- ③ 가-마-다-라-나
- ④ 가-마-나-다-라

36. 주성분분석은 차원의 단순화를 통해 서로 상관되어 있는 변수들 간의 복잡한 구조를 분석하는 것이 목적이다. 다음 중 주성분분석에 대한 설명으로 적절하지 않은 것은 무엇인가?

- ① 표본의 크기가 작거나 순서형 자료를 포함하는 범주형 자료에 적용이 가능하다.
- ② 변수들끼리 상관성이 있는 경우, 해석상의 복잡한 구조적 문제가 발생하는데 이를 해결하기 위해 사용한다.
- ③ 다변량 자료를 저차원의 그래프로 표시하여 이상치(Outlier) 탐색에 사용한다.
- ④  $p$ 개의 변수들을 중요한  $m(p)$ 개의 주성분으로 표현하여 전체 변동을 설명하는 것으로  $m$ 개의 주성분은 원래 변수와는 관계없이 생성된 변수들이다.

37. 다음 중 상관계수에 대한 설명으로 가장 부적절한 것은?

- ① 피어슨 상관계수는 두 변수 간의 선형관계의 크기를 측정한다.
- ② 스피어만 상관계수는 두 변수 간의 비선형적인 관계도 측정 가능하다.
- ③ 피어슨 상관계수와 스피어만 상관계수는 -1과 1사이의 값을 가진다.
- ④ 피어슨 상관계수는 두 변수를 순위로 변환시킨 후 두 순위 사이의 스피어만 상관계수로 정의된다.

38. 원데이터 집합으로부터 크기가 같은 표본을 여러 번 단순임의 복원 추출하여 각 표본에 대한 분류기를 생성한 후 그 결과를 앙상블하는 방법으로 다음 중 가장 적절한 것은?

- ① 배깅(bagging)
- ② 의사결정나무(decision tree)
- ③ 서포트 벡터 머신(support vector machine)
- ④ 유전자 알고리즘(genetic algorithm)

39. 모집단에 대한 어떤 가설을 설정한 뒤에 표본관찰을 통해 그 가설의 채택여부를 결정하는 분석 방법은 무엇인가?

- ① 구간추정
- ② 점추정
- ③ 신뢰수준
- ④ 가설검정

40. College 데이터프레임은 777개의 미국 소재 대학의 각종 통계치를 포함하고 있다. 각 대학에 재학 하는데 필요한 비용이 졸업률(Grad.Rate)에 미치는 영향을 알아보기 위해 등록금(Outstate), 기숙사비(Room.board), 교재구입비(Books), 그 외 개인지출비용(Personal)을 활용하기로 했다. 다음 주 아래의 결과물에 대한 설명으로 가장 부적절한 것은?

아래

```
> cor(College)
```

		Grad.Rate	Outstate	Room.Board	Books	Personal
Grad.Rate	1.000000000	0.57128993	0.4249415	0.001060894	-0.2693440	
Outstate	0.571289928	1.00000000	0.6542564	0.038854868	-0.2990869	
Room.Board	0.424941541	0.65425640	1.00000000	0.127962970	-0.1994282	
Books	0.001060894	0.03885487	0.1279630	1.000000000	0.1792948	
Personal	-0.269343964	-0.29908690	-0.1994282	0.179294764	1.00000000	

- ① Room.Board와 Outstate 간의 상관관계는 있다고 할 수 있다.
- ② Personal과 Grad.Rate, Outstate, Room.Board는 음의 상관계수를 가진다.
- ③ 위의 결과로 각 변수 간의 인과관계 알 수 있다.
- ④ Grde.Rate의 값이 커짐에 따라 Books의 값이 커지는 원인을 알 수 없다.

단 답 형

\* 문항 수(10문항), 배점(문항 당 2점, 부분점수 없음)

01. 데이터 가공 및 상관관계의 이해를 통해 패턴을 인식하고 그 의미를 부여하는 데이터를 무엇이라고 하는가?

( )

02. 아래에 설명하는 (가)는 무엇인가?

아래

(가)는 인터넷을 기반으로 모든 사물을 연결해 사람과 사물, 사물과 사물 간의 정보를 상호 소통하는 지능형 기술 및 서비스이며, 사물에서 생성되는 Data를 활용한 분석을 통해 마케팅 등에 활용할 수 있다.

( )

03. 아래에서 설명하는 (가)는 무엇인가?

아래

(가)는 식별된 비즈니스 문제를 데이터의 문제로 변환하여 정의하는 단계

( )

04. 다음 중 빈칸에 공통으로 들어갈 알맞은 단어를 적으시오.

아래

( )란 전사차원의 모든 데이터에 대하여 정책 및 지침, 표준화, 운용조직 및 책임등의 표준화된 관리체계를 수립하고 운영을 위한 프레임워크(Framework) 및 저장소(Repository)를 구축하는 것을 말한다. 특히 마스터 데이터(Master Data), 메타 데이터(Meta Data), 데이터 사전(Data Dictionary)은 ( )의 중요한 관리 대상이다.

( )

05. 다음 내용이 설명하고 있는 단어를 적으시오.

아래

이것은 배경에 랜덤과정을 추가한 방법이다. 원 자료로부터 붓스트랩 샘플을 추출하고, 각 붓스트랩 샘플에 대해 트리를 형성해 나가는 과정은 배경과 유사하나, 각 노드마다 모두 예측변수 안에서 최적의 분할을 선택하는 방법 대신 예측변수를 임의로 추출하고 추출된 변수 내에서 최적의 분할을 만들어 나가는 방법을 사용한다.

( )

06. 앙상블 기법 중 붓스트랩 표본을 구성하는 재표본 과정에서 분류가 잘못된 데이터에 더 큰 가중치를 주어 표본을 추출하는 기법은?

( )

07. 인공지능망에서 동일 입력층에 대해 원하는 값이 출력되도록 개개의 가중치(weight)를 조정하는 방법은 무엇인가?

( )

08. 군집분석의 품질을 정량적으로 평가하는 대표적인 지표로 군집 내의 데이터 응집도(cohesion)와 군집간 분리도(separation)를 계산하여 군집 내의 데이터의 거리가 짧을수록, 군집 간 거리가 멀수록 값이 커지며 완벽한 분리일 경우 1의 값을 가지는 지표는?

( )

09. 통계분석 개념 중 모집단의 특성을 단일한 값으로 추정하는 방법은 무엇인가?

( )

10. 불순도를 측정하는 지표로 노드의 불순도를 나타내는 값이다. 클수록 이질적이며 순수도가 낮다고 볼 수 있으며, CART에서 목적변수가 범주형일 경우 사용하는 이 지표는 무엇인가?

( )

