18 회 기출

01. 다음 중 빅데이터 출현 배경에 관한 설명으로 부적절한 것은? ④

① 기업의 데이터 축적 및 데이터 활용에 대한 필요성 인지
② 인터넷, SNS, IoT 의 확산으로 데이터 생산량 증가
③ 데이터 저장 기술 발전과 저장 비용 감소
④ 분석 및 수집 관리에 편리한 형태로 데이터 구조의 정형화

02. 다음 중 통찰력을 제공하는 분석 기술로 부적절한 것은? ②

① 모델링
② 추출
③ 최적화
④ 예측

03. 다음 중 가트너가 제시한 데이터 사이언티스트가 갖춰야할 역량으로 부적절한 것은? ②

① 비즈니스 분석 (Business Analysis)
② 하드 스킬(Hard Skill)
③ 데이터 관리(Data Management)
④분석 모델링(Analytics Modeling)

04. 개인에게 축적된 경험을 언어나 기호 등의 객관적인 데이터로 문서나 매체에 저장, 가공, 분석하는 과정은? ③

① 내면화
② 공통화
③ 표출화
④ 연결화

05. 빅데이터 시대 위기 요인으로 가장 부적절한 것은? ①

① 익명화
② 사생활침해
③ 데이터 오용
④ 책임원칙 훼손

06. 다음 중 데이터 사이언스에서 인문학 열풍을 가져오게 한 외부 환경 요소로 가장 부적절한 것은? ④

① 단순 세계화인 컨버전스에서 복잡한 세계화인 디버전스로의 변화
② 비즈니스 중심이 제품생산에서 서비스로 이동
③ 경제와 산업의 논리가 생산에서 시장 창조로 변화
④ 빅데이터 분석 기법의 이해와 분석 방법론 확대

07. 데이터 사이언스와 데이터 사이언티스트에 대한 설명으로 부적절한 것은? ①

① 통계학과 데이터 사이언스는 “데이터를 다룬다.”는 것이 비슷하지만 통계학은 더욱 확장된 유형의 데이터를 다룬다.
② 대부분의 전문가들이 데이터 사이언티스트가 갖춰야할 역량으로 호기심을 언급한다.
③ 더 높은 가치 창출과 차별화를 가져오는 것은 전략적 통찰력과 관련된 소프트 스킬이다.
④ 뛰어난 데이터 사이언티스트는 정량적 분석이라는 과학과 인문학적 통찰을 근거로 합리적 추론을 한다.

08. 다음 DIKW 단계를 설명하는 것 중 다른 하나는 무엇인가? ③

① 7 월 A 상품을 구매하는 고객의 60%가 30 대 남성 고객이다.
② 작년 매출은 2 월에서 7 월까지 증가하였고, W 월에 다시 증가했다.
③ 날씨가 추워지고, 지점이 늘어나 11 월 매출액은 5000 만원으로 예상한다.
④ 작년 매출액의 70%는 2 월에 집중되어 있다.

09. 아래에서 설명하는 데이터 거버넌스 체계 항목은 무엇인가? ①

아래

데이터 표준 용어 설정, 명명 규칙 수립, 메타 데이터 구축, 데이터 사전 구축 등의 업무로 구성된다.

① 데이터 표준화

② 데이터 관리 체계

③ 데이터 저장소관리

④ 표준화 활동

10. 하향식 데이터 분석 기획에서 문제 탐색 단계에 대한 설명으로 가장 부적절한 것은? ④

① 비즈니스 모델 캔버스는 문제 탐색 도구로 활용

② 문제를 해결함으로써 발생하는 가치에 중점을 두는 것이 중요

③ 빠짐없이 문제를 도출하고 식별하는 것이 중요

④ 문제 탐색은 유스케이스 활용보다는 새로운 이슈탐색이 우선

11. 분석 마스터 플랜 수립에서 과제 우선순위 결정과 관련한 내용으로 부적절한 것은? ①

① 속도는 비즈니스 효과이다.

② 시급성과 전략적 필요성은 전략적 중요도의 평가 요소이다.

③ 적용 기술의 안전성 검증은 기술

④ 전략적 중요도, ROI, 실행 용이성은 분석과제 우선순위 결정에 고려할 사항이다.

12. 분석 기획 고려사항 중 장애요소에 대한 설명으로 부적절한 것은? ③

① 조직의 역량으로 내재화를 위한 변화 관리

② 유사 분석 시나리오 및 솔루션을 활용해 분석 모형의 안정적 성능 확보

③ 이해도 높은 모형보다는 복잡하고 정교한 모형

④ 비용대비 효과의 적정한 비용

13. 아래 (가) 와 (나) 에 순서.대로 들어갈 내용으로 적절한 것은? ①

> 아래
>
> 분석은 분석 대상(What) 및 분석 방법(How)에 따라서 4가지로 나눌 수 있다. 분석 대상이 명확하게 무엇인지 모르는 경우에는 기존 분석 방식을 활용하여 (가)을(를) 도출해냄으로써 문제의 도출 및 해결에 기여하거나 (나) 접근법으로 분석 대상 자체를 새롭게 도출할 수 있다.

① 통찰-발견

② 발견 - 솔루션

③ 최적화 - 통찰

④ 솔루션-통찰

14. 분석과제를 수행할 때 고려해야할 주요 5 가지 속성이 아닌 것은? ④

① 속도

② 분석 복잡도

③ 데이터 양

④ 데이터 분석방법

15. 다양한 데이터 유형 중 정형 데이터 - 반정형 데이터 - 비정형데이터 순서로 가장 적절한 것은? ①

① 물류 창고 재고 데이터 - XML - 이메일 전송 데이터

② 인스타그램 게시물 - 기상청 날씨 데이터 - 웹 로그 데이터

③ RFID - IoT 센서데이터 - 동영상 데이터

④ CRM 데이터 - 카카오톡 대화 데이터 - Twitter 상태메세지

16. 프로토타이핑 (Prototyping) 접근법에 대한 설명으로 가장 적절한 것은? ②

① 문제가 정형화되어 있고 문제해결을 위한 데이터가 완벽하게 조직에 존재하는 경우 효과적이다

② 신속하게 해결책이나 모형을 제시함으로써 이를 바탕으로 문제를 좀 더 명확하게 인식하고 필요한 데이터를 식별하여 구체화할 수 있게 하는 유용한 상향식 접근 방법이다.

③ 문제 정의가 불명확하거나 이전에 접해보지 못한 새로운 문제일 경우 적용하기 어렵다.

④ 문제가 주어지고 이에 대한 해법을 찾기 을하여 각 과정이 체계적으로 단계화되어 수행하는 방식이다.

17. 다음 중 회귀모형의 변수선택 방법으로 사용할 수 있는 것으로 부적절한 것은? ④

① 모든 조합의 회귀분석

② Lasso 회귀분석

③ 단계적 변수 선택 방법

④ 주성분분석

18. Credit 데이터는 신용카드 대금(balanced 소득(income), 학생여부(student=Y/N)를 포함한다. Balance 를 종속변수로 하는 보기의 명령어 중 아래의 그림과 같은 회귀식을 나타내는 것은? ③

① lm(Balance~Income, data=Credit)

② lm(Balance~Student, data=Credit)

③ lm(Balance~Income+ Student, data=Credit)

④ lm(Balance~Income+ Student+ Income*Student, data=Credit)

19. 다음 중 군집분석에서의 유사도 측도에 대한 설명으로 부적절한 것은? ④

① 표준화 거리는 각 변수를 해당 변수의 표준편차로 변환한 후 유클리드 거리를 계산한 거리이다.
② 맨하튼 거리는 각 방향 직각의 이동 거리 합으로 계산된다.
③ 유클리드 거리는 두 점을 잇는 가장 짧은 직선거리이다.
④ 마할라노비스 거리는 변수의 표준편차를 고려한 거리 측도이나 변수 간에 상관성이 있는 경우에는 표준화 거리 사용을 검토해야 한다.

20. 다음 중 데이터의 정규성을 확인하기 위한 방법으로 부적절한 것은? ④

① Shapiro-Wilks test
② 히스토그램
③ Q-Q plot
④ Durbin Watson test

21. 상관분석에 대한 설명으로 가장 부적절한 것은? ①

① 상관분석은 종속변수에 미치는 영향력의 크기를 파악하여 독립변수의 특정한 값에 대응하는 종속 변수값을 예측하는 선형모형을 산출하는 방법이다.
② 상관분석은 변수들 간의 연관성을 파악하기 위해 사용하는 분석 기법 중 하나로 변수 간의 선형 관계 정도를 분석하는 통계기법이다.
③ 서열 척도로 측정된 변수들 간의 상관계수는 스피어만 상관계수로 측정한다.
④ 등간 척도 및 비율척도로 측정된 변수들 간의 상관계수는 피어슨 상관계수로 측정한다.

22. 데이터 마이닝 단계 중 모델링 목적에 따라 목적변수를 정리하고 필요한 데이터를 데이터 마이닝 소프트웨어에 적용할 수 있도록 준비하는 단계는? ①

① 데이터 가공
② 데이터 준비
③ 분석 기법의 적용
④ 목적 설정

23. 다음 중 연관 규칙의 측정지표인 향상도에 대한 설명으로 가장 적절한 것은? ④

① 품목 A와 B의 구매가 서로 관련이 없는 경우 향상도는 0이다.
② 품목 B에 대한 품목 A의 조건부 확률로 나타낸다.
③ 전체 거리 중에서 품목 A, B가 동시에 포함된 거래의 비율이다.
④ 향상도가 1보다 크면 해당 규칙은 결과를 예측하는데 있어 우수하다.

24. R 의 데이터 구조와 저장형식에 관한 설명으로 가장 부적절한 것은? ④

① 데이터 프레임은 열별로 서로 다른 데이터 타입을 가질 수 있다.

② 숫자형 행렬에서 원소 중 하나를 문자형으로 변경하게 되면 해당 행렬의 모든 원소가 문자형으로 변경된다.

③ as.numeric 함수에 논리형 벡터를 입력하면 TRUE 에 1, FALSE 에 0 으로 대응되고 숫자형 벡터로 변형된다.

④ 행렬을 as.vector 함수에 입력하면 행 방향으로 1 행부터 차례로 원소를 나열하는 벡터가 생성된다.

25. 아래는 kyphosis 라는 자료를 이용하여 의사결정나무 분석을 수행한 결과이다. 결과에 대한 해석으로 부적절한 것은? ①

아래

```
> a<-rpart(Kyphosis~Age +Number+Start, data=kyphosis)
n= 81

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 81 17 absent (0.79012346 0.20987654)
   2) Start>=8.5 62  6 absent (0.90322581 0.09677419)
     4) Start>=14.5 29  0 absent (1.00000000 0.00000000) *
     5) Start< 14.5 33  6 absent (0.81818182 0.18181818)
      10) Age< 55 12  0 absent (1.00000000 0.00000000) *
      11) Age>=55 21  6 absent (0.71428571 0.28571429)
        22) Age>=111 14  2 absent (0.85714286 0.14285714) *
        23) Age< 111 7  3 present (0.42857143 0.57142857) *
   3) Start< 8.5 19  8 present (0.42105263 0.57894737) *
> plot(a)
> text(a, use.n=T)
```

① 뿌리마디에서 아래로 내려갈수록 각 마디에서의 불순도는 점차 증가한다.

② 이 자료에서 Start 변수의 값이 14.5 이상인 관찰치는 Kyphosis 변수의 값이 모두 absent 였을 것이다.

③ 위 결과의 단계에서 멈추지 않고 추가로 가지를 생성한다면, 새로운 자료에 대한 예측력은 떨어질 수도 있다.

④ 뿌리마디의 자료는 Start 변수를 이용하여 분리했을 때 present 와 absent 를 가장 잘 분리시킬 수 있다.

26. R 의 데이터 구조 중 2 차원 목록 데이터 구조이면서 각 열이 서로 다른 데이터 타입을 가질 수 있는 데이터 구조로 적절한 것은? ④

① 행렬

② 배열

③ 벡터

④ 데이터프레임

27. 다음 중 비모수적 방법의 특징으로 가장 부적절한 것은? ②

① 분포의 모수에 대한 가설을 설정하지 않고 분포의 형태에 대해 가설을 설정한다.

② 비모수 검정에서는 평균, 분산 등을 이용해 검정을 실시한다.

③ 비모수적 검정은 모집단의 분포에 대해 아무런 제약을 가하지 않는다.

④ 관측된 자료가 특정 분포를 따른다고 가정할 수 없는 경우에 이용된다.

28. 다음 중 자가 조직화 지도(Self-Organizing Map) 방법에 대한 설명으로 부적절한 것은? ②

① SOM 은 입력변수의 위치 관계를 그대로 보존하여 입력 변수의 정보와 그들의 관계가 지도상에 그대로 나타난다.

② SOM 을 이용한 군집분석은 역전파 알고리즘을 사용함으로써 군집의 성능이 우수하고 수행 속도가 빠르다.

③ SOM 은 경쟁 학습으로 각각의 뉴런이 입력 벡터와 얼마나 가까운가를 계산하여 연결강도를 반복적으로 재조정하여 학습한다. 이와 같은 과정을 거치면서 연결강도는 입력 패턴과 가장 유사한 경쟁층 뉴런이 승자가 된다.

④ SOM 알고리즘은 고차원의 데이터를 저차원의 지도 형태로 형상화하기 때문에 시각적으로 이해하기 쉬울 뿐 아니라 변수의 위치 관계 4 그대로 보존하기 때문에 실제 데이터가 유사하면 지도상 가깝게 표현된다.

29. 다음 중 추정과 가설검정에 대한 설명으로 가장 부적절한 것은? ③

① 점추정은 모수가 특정한 값일 것이라고 추정하는 것이다.

② 구간추정이란 일정한 크기의 신뢰구간으로 모수가 특정한 구간에 있을 것이라고 선언하는 것으로 구해진 구간을 신뢰구간이라고 한다.

③ 귀무가설이 사실일 때, 관측된 검정통계량의 값보다 더 대립가설을 지지하는 검정통계량이 나올 확률을 p 값이라고 한다.

④ 기각역이란 대립가설이 맞을 때 그것을 받아들이는 확률을 의미한다.

30. 다음 중 아래 문장의 빈 칸에 들어갈 말로 순서대로 연결된 것은? ②

아래

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

① 낮고, 낮다.

② 높고, 낮다.

③ 높고, 높다.

④ 낮고, 높다.

31. 아래는 Chickwts 데이터셋에 대해 첨가물 그룹 간 평균 무게에 차이가 있는지 검정하기 위해 분산분석을 한 결과 중 설명이 가장 부적절한 것은? ②

아래

```
> summary(aov(weight~feed, chickwts))
            Df Sum Sq Mean Sq F value   Pr(>F)
feed         5 231129   46226   15.37 5.94e-10 ***
Residuals   65 195556    3009
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

① 귀무가설은 "첨가물 그룹 간의 평균이 모두 동일하다"이다.

② 첨가물의 개수는 5 개다.

③ 위의 가설검정은 F 통계량을 기반으로 한다.

④ 유의수준 0.05 하에서 "첨가물 그룹 간의 무게 평균이 동일하지 않다"고 결론을 내릴 수 있다.

32. 한 보험회사에서는 자사 고객의 보험갱신 여부를 고객의 보험가입 채널 등의 정보를 사용하여 예측하려고 한다. 다음 중 가장 적절한 분석 기법은 무엇인가? ②

① 시계열분석
② 랜덤포레스트
③ k-means 군집분석
④ 주성분분석

33. 아래는 특정 제품의 sales 오|_ TV, Radio, Newspaper 광고예산 간의 피어슨 상관계수 행렬이다. 설명이 가장 부적절한 것은? ④

아래

	TV	Radio	Newspaper	Sales
TV	1.000	0.054	0.057	0.793
Radio	0.054	1.000	0.333	0.543
Newspaper	0.057	0.333	1.000	0.222
Sales	0.793	0.543	0.222	1.000

① Newspaper 광고예산이 증가할 때 Radio 광고 예산이 증가하는 경향이 있다.
② 3 가지 매체의 광고예산은 Sales 와 양의 상관관계를 가지고 있다.
③ Sales 와 가장 상관관계가 높은 변수는 TV 이다.
④ *TV* 광고 예산을 늘릴 경우 Sales 가 증가하는 인과관계를 가진다.

34. 회귀분석에서 결정계수(R^2)에 한 설명으로 부적절한 것은? ①

① 총 변동 중에서 설명이 되지 않는 오차에 의한 변동이 차지하는 비율이다.
② 회귀모형에서 입력 변수가 증가하면 결정계수도 증가한다.
③ 다중 회귀분석에서는 최적 모형의 선정기준으로 결정계수 값보다는 수정된 결정계수 값을 사용하는 것이 적절하다.
④ 수정된 결정계수는 유의하지 않은 독립변수들이 회귀식에 포함되었을 때 그 값이 감소한다.

35. 제 1 종 오류에서 '우리가 내린 판정이 잘못되었을 실제 확률'은 무엇으로 나타낼 수 있는가? ③
식

① 기각역
② 검정통계량
③ p-vlaue
④ 1-a

36. 다중 회귀분석에서 가장 적합한 회귀모형을 찾기 위한 과정의 설명으로 가장 부적절한 것은? ②

① 독립변수의 수가 많아지면 독립변수들 간에 서로 영향을 미치는 다중공선성의 문제가 발생하므로 상대적인 조정이 필요하다.

② 회귀식에 대한 검정은 독립변수의 기울기(회귀계수)가 o 이 아니라는 가정을 귀무가설, 기울기가 o 인 것을 대립가설로 놓는다.

③ 회귀분석의 가설검정에서 P 값이 0.05 보다 작은 값이 나와야 통계적으로 유의한 결과로 받아들일 수 있다.

④ 잔차의 독립성, 등분산성 그리고 정규성을 만족하는지 확인해야 한다.

37. 소매점에서 물건을 배열하거나 카탈로그 및 교차판매 등에 적용하기 적합한 데이터마이닝 기법은 무엇인가? ③

① 분류 (classification)

② 예측 (prediction)

③ 연관분석(association analysis)

④ 군집 (clustering)

38. R 에서 제공하는 데이터 가공, 처리를 위한 패키지의 설명으로 가장 부적절한 것은? ①

① data.table 패키지는 데이터 프레임 처리함수인 ddply 함수를 제공한다

② sqldf 패키지는 R 에서 표준 SQL 명령을 실행하고 결과를 가져올 수 있다.

③ plyr 패키지는 데이터의 분리, 결합 등 필수적인 데이터 처리 기능을 제공한다.

④ reshape 패키지는 melt 와 cast 를 이용하여 데이터를 재구성할 수 있다.

39. R 에서 y=c(3,4,5, NA)일 때 3*y 의 실행 결과는? ③

① 9 12 15 0

② 9 12 15 9

③ 9 12 15 NA

④ 에러가 발생하고 결과가 출력되지 않는다.

40. 다음 중 기법의 활용 분야가 나머지와 다른 하나를 고르시오. ④

① 랜덤 포레스트

② 인공신경망

③ Support Vector Machine

④ SOM

단답형

01. 아래는 빅데이터가 만들어 내는 본질적인 변화이다. (A)와 (B)에 들어갈 말을 쓰시오. (A: 인과관계, B: 상관관계)

> 아래
>
> (A)은(는) 어떤 현상에 대하여 현상을 발생시킨 원인과 그 결과 사이의 관계를 말하고,
> (B)은(는) 어떤 두 현상이 관계가 있음을 말하지만 어느 쪽이 원인인지 알 수 없다.

(　　)

02. 아래에서 빈칸에 공통적으로 들어갈 용어는? (플랫폼)

> 아래
>
> 하둡은 대규모 분산 병렬 처리의 업계 표준으로 맵리듀스 시스템과 분산 파일 시스템인 HDFS로 구성된 (　　) 기술이며, 선형적인 성능과 용량 확장성, 고장 감내성을 가지고 있다) 아마존(Amazon)은 S3 와 BC2 환경을 제공함으로써 (　　)을(를) 위한 클라우드 서비스를 최초로 실현하였다.

(　　)

03. 문제가 주어지고 이에 대한 해법을 찾기 위하여 각 과정이 체계적으로 단계화되어 수행하는 분석과제 발굴 방식을 무엇이라고 하는가? (하향식 접근 방식)

(　　)

04. 아래는 빅데이터 분석 프로세스에서 데이터 분석 단계 중 어떤 것에 대한 설명인가? (모델링)

> 아래
>
> 분석용 데이터를 이용한 가설 설정을 통하여 통계모델을 만들거나 기계학습을 이용한 데이터의 분류, 예측, 군집 등의 기능을 수행하는 모델을 만드는 과정

(　　)

05. 아래의 오분류표에서 정확도(accuracy)를 계산하는 산식을 a~d 를 사용하여 작성하시오. (a+ d)/(a+ b+ c+ d)

아래

		예측치	
		True	False
실제값	True	a	b
	False	c	d

( )

06. 아래는 학생들의 키와 몸무게를 정규화한 데이터이다. 맨하튼 거리를 이용하여 A 와 B 의 거리를 구하시오. (|2-1|+|4-5|=1+ 1=2)

아래

사람	(키, 몸무게)
A	(1, 5)
B	(2, 4)

( )

07. 앙상블 기법 중 붓스트랩 표본을 구성하는 재표본 과정에서 분류가 잘못된 데이터에 더 큰 가중치를 주어 표본을 추출하는 기법은? (부스팅(boosting))

( )

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? (차분(difference))

( )

09. 회귀모형의 계수를 추정하는 방법으로써 잔차제곱합 (SSR:아래 참조) 을 최소화하는 계수를 찾는 방법을 무엇이라고 하는가? (최소제곱법)

아래

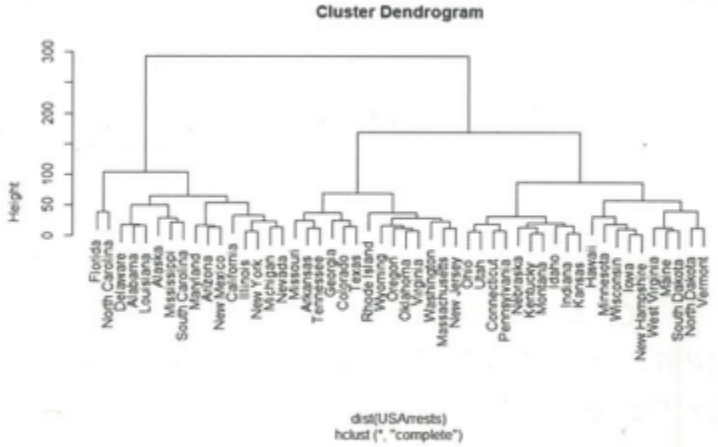
$$SSR = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)]^2$$

( )

( )

10. 아래는 미국 50 개 주의 범죄 유형으로 군집분석을 한 결과이다. height=150 에서 아래의 덴드로그램을 통해 군집 결과를 도출하면 총 군집의 수는 몇 개인가? (3 개)

아래



( )