감 독 확 인 란

제30회 데이터 분석 준전문가 자격검정 시험 복원 문제

- 검정일시 : 2021 . 08 . 29(일) / 10:00~11:30
- 수험번호 :
- 성　　명 :

과목 I 데이터 이해 * 문항 수(8문항), 배점(문항 당 2점)

01. 빅데이터가 만들어내는 본질적인 변화로 틀린 것은?

① 표본조사에서 전수조사로 변화했다.
② 데이터의 질에서 데이터의 양으로 변화했다.
③ 비정형 데이터에서 정형 데이터로 변화했다.
④ 인과관계에서 상관관계로 변화했다.

02. 다음 중 빅데이터 활용사례로 부적절한 것은?

① 구글, 애플 등의 기업에서는 정형화된 데이터만 수집하여 웹과 스마트폰의 서비스에 활용한다.
② NSA(National Security Agency)가 소셜미디어, 통화기록 등의 모니터링과 분석으로 국가안전을 확보한다.
③ 구매 패턴 데이터를 수집하고 분석하여 고객 맞춤형 가전제품을 추천한다.
④ 소셜 미디어를 통해 고객 소비 패턴을 분석하는 연구소를 운영한다.

03. 다음 중 미래 사회의 특성과 빅데이터 역할이 올바르게 연결되지 않는 것은?

① 융합 - 창조력
② 리스크 - 대응력
③ 불확실성 - 통찰력
④ 단순화 - 경쟁력

04. 다음 중 빅데이터의 활용으로 알맞지 않은 것은?

① 데이터 수집 및 저장
② 고객 맞춤형 서비스 제공
③ 교통패턴, 지역 인구기반 상권 분석
④ 물류 등 유통 효율성 제고

05. 다음 중 데이터 분석가에게 필요한 것 중 틀린 것을 고르시오.

① 문맥과 의미
② 통찰력
③ 이론적 지식
④ 천재적 직관력

06. 다음 중 데이터베이스의 일반적인 특징이 아닌 것은?

① Integrated Data
② Stored Data
③ Shared Data
④ Unchanged Data

07. 다음 중 빅데이터 가치 산정이 어려운 이유는 무엇인가?

① 데이터의 과다성
② 가치창출의 어려움
③ 분석기술의 한계
④ 기업 경영의 난이도 상승

08. 아래는 특정산업의 일차원적 분석 사례를 나열한 것이다. 다음 중 특정산업으로 적절한 것은?

아래

트레이딩, 공급/수요예측

① 소매업
② 에너지
③ 운송업
④ 금융서비스

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

09. 다음 중 아래의 데이터 거버넌스 체계가 설명하는 항목은?

아래

메타데이터 관리, 데이터 사전관리, 데이터 생명주기 관리

① 데이터 표준화
② 데이터 관리 체계
③ 데이터 저장소 관리
④ 표준화 활동

10. 다음 중 데이터 거버넌스의 구성 요소가 아닌 것은?

① 원칙(principle)
② 조직(organization)
③ 분석방법(method)
④ 절차(process)

11. 다음 중 하향식 접근법에서 문제 탐색단계에 대한 내용 중 틀린 것은?

① 과제 발굴단계에서는 세부적인 구현 및 솔루션에 중점을 둔다.
② 시장의 니즈 탐색 관점에서는 현재 수행하고 있는 사업에서의 직접 고객뿐만 아니라 고객과 접촉하는 역할을 수행하는 채널 및 고객의 구매와 의사결정에 영향을 미치는 영향자들에 대한 폭넓은 관점을 바탕으로 분석 기회를 탐색한다.
③ 현재 경쟁자는 아니지만, 향후 시장에 대해서 파괴적인 역할을 수행할 수 있는 잠재적 경쟁자에 대한 동향을 파악하여 이를 고려한 분석 기회를 도출한다.
④ 거시적 관점의 메가트렌드에서는 현재의 조직 및 해당 산업에 폭넓게 영향을 미치는 사회·경제적 요인을 사회·기술·경제·환경·정치 영역으로 나누어서 좀 더 폭넓게 기회 탐색을 수행한다.

12. 지속적인 분석 내재화를 위한 "장기적인 마스터 플랜 방식"에 비하여 "과제 중심적인 접근 방식"의 특징으로 가장 적절하지 못한 것은?

① Quick-Win
② Accuracy & Deploy
③ Problem Solving
④ Speed & Test

13. 분석 과제를 발굴하기 위한 접근법 중 상향식 접근방식의 특징으로 올바른 것은?

① 타당성 검토의 과정을 거치며 경제적, 데이터 및 기술적 타당도 등이 있다.
② 일반적으로 상향식 접근 방식의 데이터 분석은 지도학습 방법에 의해 수행된다.
③ Design thinking 중 Ideate 단계에 해당한다
④ 인사이트를 도출한 후 반복적인 시행착오를 통해서 수정하며 문제를 도출하는 일련의 과정이다.

14. 다음 중 분석 마스터 플랜 수립시 분석 과제 우선 순위를 결정하는 고려 요소로써 가장 부적절한 것은?

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

15. 아래는 분석 방안 구체화에 대한 설명이다. 알맞은 단계를 선택하면?

아래

- 정의된 의사결정 모형의 분석 컨텍스트별로 수행할 분석을 정리하여 의사결정을 위한 전체 분석 세트와 관계를 도출함
- 각 분석들의 관계와 집합은 의사결정을 위한 시그널 허브로 작동
- 중간단계의 분석 결과들도 의사결정자들에게 필요한 시그널로 제공
- 지속적으로 보완되는 과정을 거쳐 의사결정 모형의 분석체계 확정

① 의사결정 요소 모형화
② 분석 체계 도출
③ 분석 필요 데이터 정의
④ 분석 ROI 평가

16. 분석 준비도(Readiness)는 기업의 데이터 분석 도입의 수준을 파악하기 위한 진단방법으로 6가지 영역을 대상으로 파악한다. 아래 보기의 내용은 어떤 영역의 내용인가?

아래

- 업무별 적합한 분석기법 사용
- 분석업무 도입 방법론
- 분석기법 라이브러리
- 분석기법 효과성 평가
- 분석기법 정기적 개선

① 분석기법
② 분석 인력 및 조직
③ 분석 데이터
④ 분석업무 파악

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

17. 민코우스키 거리는 맨하탄 거리와 유클리디안 거리를 한번에 표현한 공식이다. 다음 중 민코우스키 거리를 나타내는 수식으로 올바른 것은?

① $d(x,y)=\sqrt{(x-y)'(x-y)}$

② $d(x,y)=\max_i \mid x_i-y_i \mid$

③ $d(x,y)=\sum_{i=1}^{p} \mid x_i-y_i \mid$

④ $d(x,y)=\left[\sum_{i=1}^{p} \mid x_i-y_i \mid^m\right]^{1/m}$

18. 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)

19. 앙상블(ensemble) 모형은 여러 모형의 결과를 결합함으로써 단일 모형으로 분석했을 때보다 신뢰성 높은 예측값을 얻을 수 있다. 다음 중 앙상블 모형의 특징으로 옳지 않은 것은?

① 이상값(outlier)에 대한 대응력이 높아진다.
② 전체적인 예측값의 분산을 감소시켜 정확도를 높일 수 있다.
③ 모형의 투명성이 떨어져 원인 분석에는 적합하지 않다.
④ 각 모형의 상호 연관성이 높을수록 정확도가 향상된다.

20. 다음 중 k평균 군집에 대한 설명으로 부적절한 것은?

① 한번 군집이 형성되면 군집에 속하는 개체들은 다른 군집으로 이동할 수 없다.

② 초기 군집의 중심을 임의로 선택해야한다.

③ 군집의 개수를 미리 선택해야 한다.

④ 이상값에 영향을 많이 받는다.

21. Hitters 데이터셋은 메이저리그의 선수 322명에 대한 타자 기록으로 20여개의 변수를 포함하고 있다. 아래 회귀모형에서 변수선택을 하기 위한 결과물의 일부이다. 다음 중 결과물에 대한 설명으로 부적절한 것은?

아래

```
> model<-lm(Salary~., data=Hitters)
> step(model, direction="backward")
Start:  AIC=3046.02
Salary ~ AtBat + Hits + HmRun + Runs + RBI + Walks + Years +
    CAtBat + CHits + CHmRun + CRuns + CRBI + CWalks + League +
    Division + PutOuts + Assists + Errors + NewLeague

            Df Sum of Sq      RSS    AIC
- CHmRun     1      1138 24201837 3044.0
- CHits      1      3930 24204629 3044.1
- Years      1      7869 24208569 3044.1
- NewLeague  1      9784 24210484 3044.1
- RBI        1     16076 24216776 3044.2
- HmRun      1     48572 24249272 3044.6
- Errors     1     58324 24259023 3044.7
- League     1     62121 24262821 3044.7
- Runs       1     63291 24263990 3044.7
- CRBI       1    135439 24336138 3045.5
- CAtBat     1    159864 24360564 3045.8
<none>                   24200700 3046.0
- Assists    1    280263 24480963 3047.1
- CRuns      1    374007 24574707 3048.1
- CWalks     1    609408 24810108 3050.6
- Division   1    834491 25035190 3052.9
- AtBat      1    971288 25171987 3054.4
- Hits       1    991242 25191941 3054.6
- Walks      1   1156606 25357305 3056.3
- PutOuts    1   1319628 25520328 3058.0
```

① 후진제거법을 통한 변수선택을 하고 있다.

② 모든 설명변수가 포함된 모형에서 시작한다.

③ Start AIC보다 작은 11개의 변수는 다음 Step에서 제외된다.

④ 한번 제거된 변수는 다시 모형에 포함될 수 없다.

22. Default 데이터셋은 10000명의 신용카드 고객에 대한 카드대금 연체여부(default=Yes/No), 카드 대금납입 후 남은 평균 카드잔고(Balance), 연봉(Income), 학생여부(student=Yes/No)를 포함한다. 아래는 연체 가능성을 모형화하기 위한 로지스틱 회귀분석 결과이다. 다음 중 유의수준 0.05하에서 아래에 대한 설명으로 가장 부적절한 것은?

아래

```
> model<-glm(default~., data=Default, family="binomial")
> summary(model)

Call:
glm(formula = default ~ ., family = "binomial", data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

① balance는 default를 설명하는 데 통계적으로 유의하다.

② income는 default를 설명하는 데 통계적으로 유의하다.

③ student는 default를 설명하는 데 통계적으로 유의하다.

④ balance는 income이 동일할 때 학생일수록 default 가능성이 낮다.

23. 로지스틱 회귀분석은 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는데 사용되는 통계기법이다. 다음 중 로지스틱 회귀모형의 모형 검정 방법으로 알맞은 것을 고르시오.

① 최소제곱법
② 양측검정
③ F-검정
④ 카이제곱 검정

24. 다음 중 주성분분석에서 변수의 중요도 기준이 되는 값은 무엇인가?

① 고윳값(Eigenvalue)
② 특이값(Singular Value)
③ 표준편차(Standard Deviation)
④ 스칼라(Scalar)

25. 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 고윳값(average eigenvalue)방법은 고윳값들의 평균을 구한 후 고윳값이 평균값 이상이 되는 주성분을 제거하는 방법이다.
③ Scree graph를 이용하는 방법은 고윳값의 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.

26. 다음 중 회귀분석의 결과 중 잔차분석에서 만족해야 하는 가정으로 맞는 것은?

① 독립성, 등분산성, 정규성
② 독립성, 등분산성, 유일성
③ 정규성, 효율성, 등분산성
④ 정규성, 불편성, 독립성

27. 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.

28. 확률이란 "특정사건이 일어날 가능성의 척도"라고 정의할 수 있다. 통계적 실험을 실시할 때 나타날 수 있는 모든 결과들의 집합을 표본공간이라고 하고, 사건이란 표본공간의 부분집합을 말한다. 다음 중 확률 및 확률분포에 대한 설명으로 가장 부적절한 것은?

① 모든 사건의 확률값은 0과 1사이에 있다.
② 서로 배반인 사건들의 합집합의 확률은 각 사건들의 확률의 합이다.
③ 두 사건 A, B가 독립이라면 사건 B의 확률은 A가 일어난다는 가정하에서의 B의 조건부확률과 동일하다.
④ 확률변수 X가 구간 또는 구간들의 모임인 숫자 값을 갖는 확률분포함수를 이산형 확률밀도함수라 한다.

29. 다음 중 연관성분석에 활용되는 측정지표 중에 전체 거래 중에서 품목 A와 품목 B가 동시에 포함된 거래의 비중을 나타내는 지표는 무엇인가?

① 신뢰도(confidence)
② 향상도(lift)
③ 지지도(support)
④ 순서도(flowchart)

30. SOM은 비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도의 형태로 형상화 한다. 다음 중 입력벡터의 특성에 따라 벡터가 한 점으로 클러스터링 되는 층은 어떤 층인가?

① 경쟁층(Competitive layer)
② 입력층(Input layer)
③ 은닉층(Hidden layer)
④ 출력층(Output layer)

31. 적합된 회귀모형의 안정성을 평가하기 위한 통계적 방법을 영향력 진단이라 한다. 자료에서 특정 관측치가 제외됨에 따라 분석 결과의 주요 부분에 많은 변동이 있다면 안정성이 약하다고 판단된다. 다음 중 각 개체의 영향력 진단에 대한 설명으로 가장 부적절한 것은?

① 쿡의 거리(Cook's distance)는 관측 개체 하나가 제외되었을 때, 최소제곱추정치 벡터의 변화를 표준화한 측도이다.
② 영향점은 비교할 대상이 있어 그 값들에 비해 값이 매우 크거나 작아 회귀 계수 추정값을 변화 시키는 관측개체를 말한다.
③ DFBETAS의 절대값이 유난히 큰 관측개체는 해당 회귀계수의 추정에 대하여 큰 영향력을 행사하는 것으로 간주한다.
④ DFFITS(Difference in fits)의 절대값이 매우 큰 관측개체는 y의 예측에 영향력이 크다고 간주한다.

32. 다음 중 데이터의 정규성을 확인하기 위한 방법으로 부적절한 것은?

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

33. 다음 제1종 오류에 대한 설명 중 올바른 것은?

① H_0가 사실일 때, H_0가 사실이라고 판정
② H_0가 사실이 아닐 때, H_0가 사실이라고 판정
③ H_0가 사실일 때, H_0가 사실이 아니라고 판정
④ H_0가 사실이 아닐 때, H_0가 사실이 아니라고 판정

34. 데이터 전처리 과정에서 이상치를 어떻게 처리할지 결정할 때 이상치를 판정하는 방법을 사용할 수 있다. 다음 중 상자그림을 이용하여 이상치를 판정하는 방법에 대한 설명으로 가장 부적절한 것은?

① IQR = Q3-Q1 이라고 할 때, Q1-1.5*IQR < x < Q3+1.5*IQR 을 벗어나는 x를 이상치라고 규정한다.
② 평균으로부터 3*표준편차 벗어나는 값들을 이상치라 규정하고 제거한다.
③ 이상치는 변수의 분포에서 벗어난 값으로 상자 그림을 통해 확인할 수 있다.
④ 이상치는 분포를 왜곡할 수 있으나 실제 오류인자에 대해서는 통계적으로 실행하지 못하기 때문에 제거여부는 실무자들을 통해서 결정하는 것이 바람직하다.

35. 70명의 실험자를 대상으로 A, B 두 종류의 수면 유도제 복용 전과 후의 평균 체중 비교에 대한 분석을 수행하고 있다. 90% 신뢰구간을 구하고자 할 때, 아래의 빈칸 (가), (나)에 순서대로 들어갈 숫자를 고르시오

아래

$$\bar{D} \pm t_{(\text{가})} \frac{S_D}{\sqrt{(\text{나})}}$$

① (가) : 0.1, (나) : 70
② (가) : 0.1, (나) : 71
③ (가) : 0.05, (나) : 70
④ (가) : 0.05, (나) : 71

36. 사회 관계망 모형에서 연결망 내 전체 구성원들이 서로 얼마나 많은 관계를 맺고 있는 지를 나타내며, SNS 내에서 존재하는 가능한 총 관계 수 중에서 실제로 맺어진 관계의 수를 비율로 계산하는 기법은?

① 밀도
② 중심성
③ 중심화
④ 구조적 틈새

37. 여섯 가지 종류의 닭 사료 첨가물의 효과를 비교하기 위한 데이터이다. 아래에 대한 설명으로 부적절한 것은 무엇인가?

아래

```
> summary(chickwts)
     weight             feed
 Min.   :108.0   casein   :12
 1st Qu.:204.5   horsebean:10
 Median :258.0   linseed  :12
 Mean   :261.3   meatmeal :11
 3rd Qu.:323.5   soybean  :14
 Max.   :423.0   sunflower:12
```

① Weight의 중앙값은 261.3이다.

② feed는 범주형 변수이다.

③ 약 25%의 닭의 weight가 204.5보다 작다.

④ weight의 범위는 315이다.

38. 아래는 R의 내장데이터인 cars에서 속도(speed)와 제동거리(dist)의 관계를 회귀모형으로 추정한 것이다. 아래의 내용 중 부적절한 것은 무엇인가?

아래

```
> out=lm(dist~speed, data=cars)
> anova(out)
Analysis of Variance Table

Response: dist
          Df Sum Sq Mean Sq F value    Pr(>F)
speed      1  21186 21185.5  89.567 1.49e-12 ***
Residuals 48  11354   236.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

① 회귀계수는 5% 수준에서 유의하다.

② 오차 분산 σ^2의 불편추정량은 236.5이다.

③ 전체 관측치 수가 49개이다.

④ 결정계수는 약 0.65이다.

39. 아래의 지문에서 말하고 있는 시계열의 종류는 무엇인가?

아래

- 현재의 충격은 미래의 y값에 관한 예측치에 아무런 영향을 미치지 못함
- 어느 시기에 충격이 발생하여 y값이 평균 이하로 감소하면 미래의 어느 기간에 걸쳐서 y의 증가율이 일시적으로 평균 수준보다 더 높아야 y가 평균수준을 회복하여 현재의 충격이 무한 미래의 y에 미치는 영향이 소멸됨

① 안정 시계열
② 표준자기함수
③ 불안정 시계열
④ 이동평균함수

40. College 데이터 프레임은 777개 미국 소재 대학의 각종 통계치를 포함하고 있고 Books 변수(단위:달러)는 평균적인 교재구입비용을 말한다. 미국 전체 대학의 평균 교재비용에 대해 추론하려 할 때, 아래의 결과에 대한 설명으로 다음 중 적절하지 않은 것은 무엇인가?

아래

```
>t.test(College$Books,mu=570)
 One Sample t-test
data: College$Books t = -3.4811, df = 776, p-value = 0.0005272
alternative hypothesis: true mean is not equal to 570
95 percent confidence interval:
537.7537 561.0082
sample estimates:
mean of x
 549.381
```

① 777개 대학의 평균 교재구입비용은 549.38 달러이다.
② 대학의 평균 교재구입비용에 대한 점추정량은 549.38 달러이다.
③ 대학의 평균 교재구입비용이 570 달러와 같다는 가설은 기각되지 않는다.
④ 대학의 평균 교재구입비용에 대한 95% 신뢰구간은 (537.75, 561.01)이다.

단 답 형

* 문항 수(10문항), 배점(문항 당 2점, 부분점수 없음)

01. 다음에 설명에 맞는 데이터 유형은 무엇인가?

아래

- 데이터의 가공 및 상관관계 간 이해를 통해 패턴을 인식하고 의미를 부여한 데이터
- 지식을 도출할 때 사용하는 데이터

( )

02. 아래에서 언급한 것은 무엇인가?

아래

기업내부 데이터베이스 중 기업 전체가 경영자원을 효과적으로 이용하기 위해 통합적으로 관리하고 경영의 효율화를 기하기 위한 수단으로 정보의 통합을 위해 기업의 모든 자원을 최적으로 관리하기 위한 기업 경영 정보시스템

( )

03. 다음 중 빈칸에 들어갈 알맞은 단어를 순서대로 적으시오.

아래

데이터 거버넌스란 전사차원의 모든 데이터에 대하여 정책 및 지침, 표준화, 운용조직 및 책임 등의 표준화된 관리체계를 수립하고 운영을 위한 프레임워크(Framework) 및 저장소(Repository)를 구축하는 것을 말한다. 특히 (a), (b), (c)는 데이터 거버넌스의 중요한 관리 대상이다.

( )

04. 다음 중 빈칸에 들어갈 알맞은 단어를 적으시오.

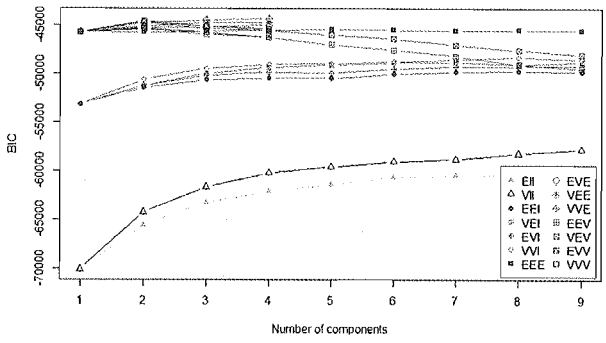
아래

(a)은(는) 전략적 중요도가 핵심이며, 이는 현재의 관점에서 전략적 가치를 둘 것인지, 미래의 중장기적 관점에 전략적인 가치를 둘 것인지를 고려하고, 분석 과제의 목표가치(KPI)를 함께 고려하여 (a)의 여부를 판단할 수 있다.

( )

05. 아래 Hitters 데이터프레임은 1966~1967년 시즌 메이저리그 야구선수 322명에 대한 데이터이다. 이 데이터를 표준화 한 후 mclust 패키지를 사용해 혼합분포 군집 방법으로 군집분석을 시행한 결과물(공분산 형태의 BIC)이다. 아래의 그래프를 보고 최적의 군집 수는 몇 개인지 쓰시오.

아래



( )

06. 다수 모델의 예측을 관리하고 조합하는 기술을 메타 학습(meta learning)이라 한다. 여러 분류기(classifier)들의 예측을 조합함으로써 분류 정확성을 향상시키는 기법은?

( )

07. 아래 데이터의 Terminal변수는 약 몇 %가 92 보다 큰 값을 가지는가?

아래

```
> summary(College$Terminal)

   Min.  1st Qu.  Median    Mean   3rd Qu.    Max.
   24.0    71.0    82.0    79.7    92.0     100.0
```

( )

08. 이산형 확률분포 중 주어진 시간 또는 영역에서 어떤 사건의 발생 횟수를 나타내는 확률 분포는 무엇인가?

( )

09. 오분류표에서 실제/예측 True와 실제/예측 False가 100으로 동일하다고 한다. 민감도가 0.8 이라고 할 때, 정확도(precision)은 얼마인가?

( )

10. 아래에서 언급한 것은 무엇인가?

아래

- 데이터의 패턴을 발견하고 데이터 모델의 매개 변수를 자동으로 학습한다.
- 자체 알고리즘을 사용하여 시간이 경과함에 따라서 경험을 축적하면서 작업 성능이 향상된다.

( )

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)

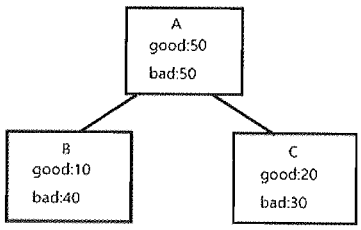
아래

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

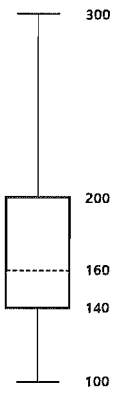
② Single linkage method

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

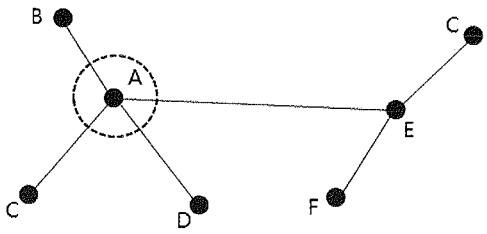
② 피어슨 상관계수

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?

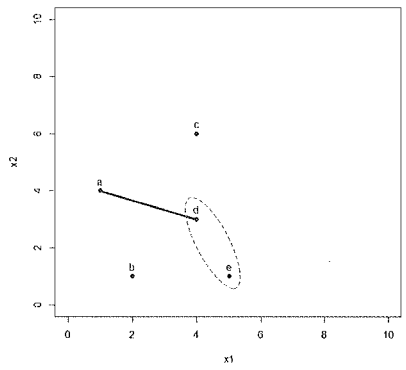
( )

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?

( )

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?

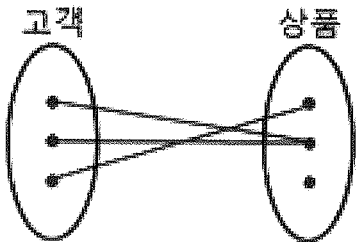
아래

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.

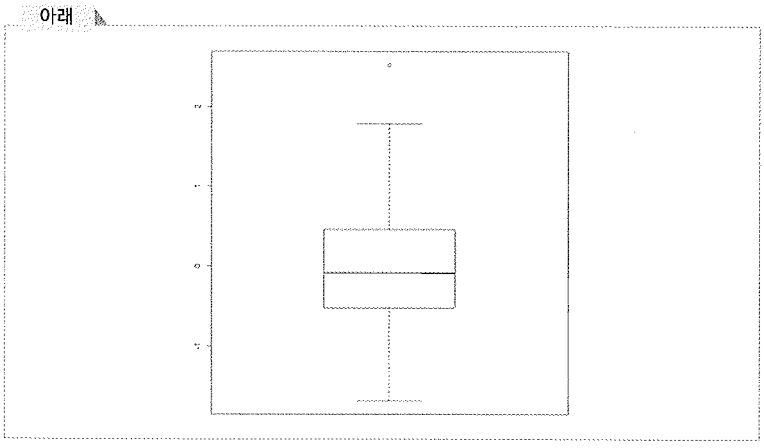
29. 다음 중 시간의 흐름에 따라 관측된 데이터에 관한 것으로 적절한 것은?

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.

( )

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?

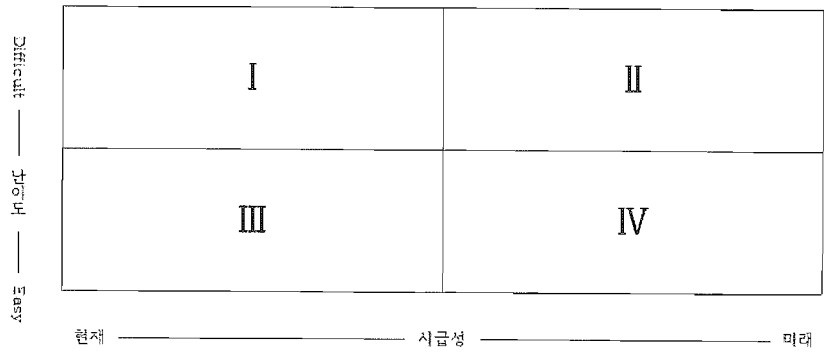
① 서비스 산업의 확대와 제조업의 축소
② 상품 개발과 조립 비용의 절감
③ 운송 비용의 절감
④ 새로운 수익원의 발굴 및 활용

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?

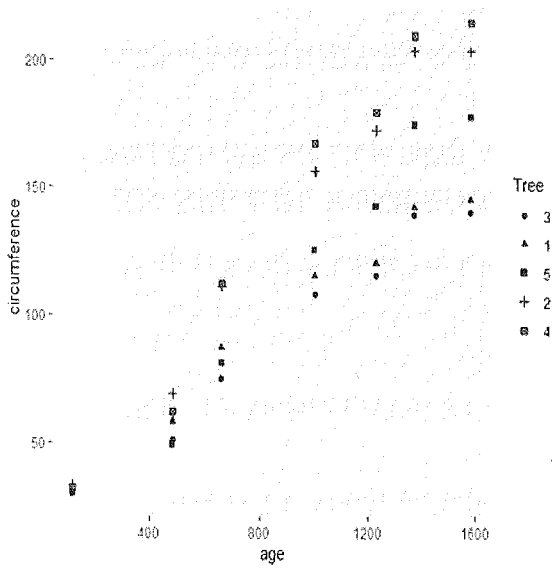
① 생성된 모델이 훈련 데이터에 너무 최적화되어 학습하여 테스트 데이터의 작은 변화에 민감하게 반응하는 경우는 발생하지 않는다.
② 변수가 너무 많아 모형이 복잡할 때 생긴다.
③ 과대적합이 발생할 것으로 예상되면 학습을 종료하고 업데이트하는 과정을 반복해 과대적합을 방지할 수 있다.
④ 학습데이터가 모집단의 특성을 충분히 설명하지 못할 때 자주 발생한다.

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

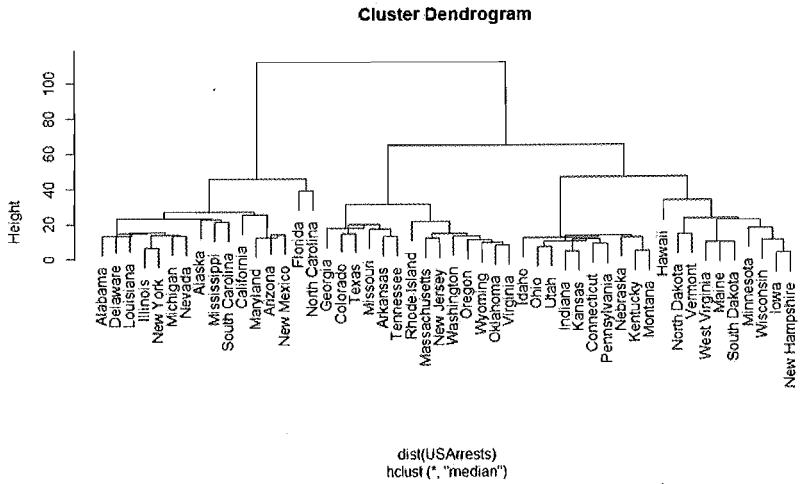
이것은 반복을 통하여 점진적으로 개발하는 방법으로서, 처음 시도하는 프로젝트에 적용이 용이하지만 관리 체계를 효과적으로 갖추지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있다.

( )

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?

( )

감독확인란

제34회 데이터 분석 준전문가 자격검정 시험 복원 문제

- 검정일시 : 2022. 08. 28(일) / 10:00~11:30
- 수험번호 :
- 성　　명 :

과목 I 데이터 이해 * 문항 수(8문항), 배점(문항 당 2점)

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.

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

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 오너십

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 → B

② A → C

③ C → A

④ B → 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=0일 때, IMA(d,q) 모형이라고 부르고, d번 차분하면 MA(q)모형을 따른다.

20. ROC(Receeiver 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 안전성이 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

```
heptathion_pca <-prcomp(heptathion2[,-score], scales=TRUE)
 Summary(heptathion_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
```

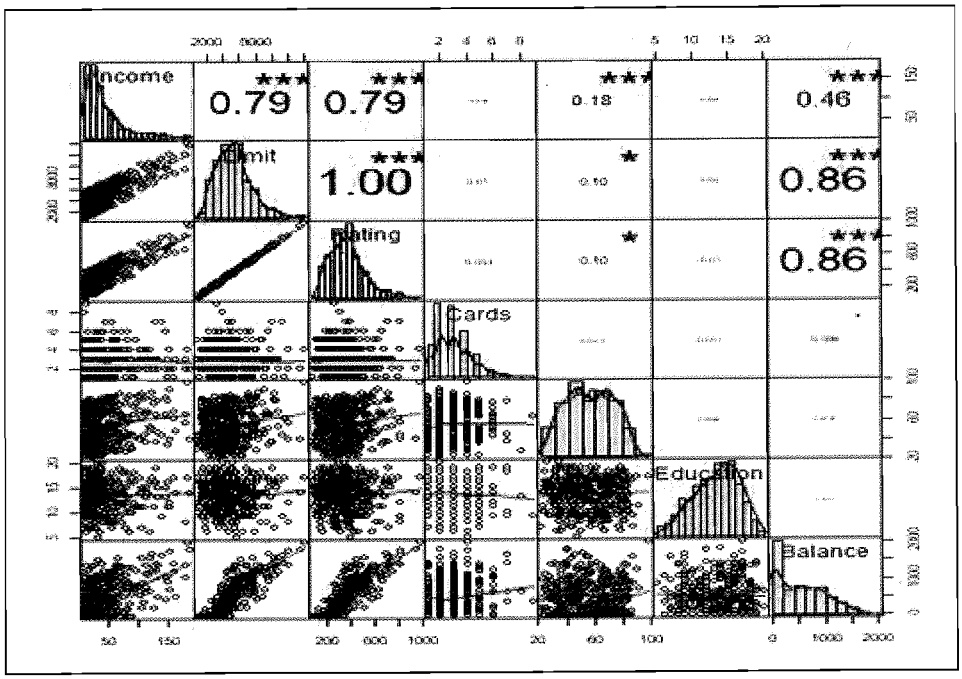
① 한 개의 주성분으로 자료를 축약할 때 전체 분산의 61.8%가 설명 가능하다.

② 두 개의 주성분으로 자료를 축약할 때 전체 분산의 12.8%가 설명 가능하다.

③ 정보손실율 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 Ratting은 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 p개의 변수들을 중요한 m(p)개의 주성분으로 표현하여 전체 변동을 설명하는 것으로 m개의 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 재학하는데 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  1.0000000 0.127962970 -0.1994282
Books        0.001060894  0.03885487  0.1279630 1.000000000  0.1792948
Personal    -0.269343964 -0.29908690 -0.1994282 0.179294764  1.0000000
```

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 배깅에 랜덤과정을 추가한 방법이다. 원 자료로부터 붓스트랩 샘플을 추출하고, 각 붓스트랩 샘플에 대해 트리를 형성해 나가는 과정은 배깅과 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. 인공신경망에서 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?

( )

감 독 확 인 란

제35회 데이터 분석 준전문가 자격검정 시험 복원 문제

- 검정일시 : 2022. 10. 29(토) / 10:00~11:30
- 수험번호 :
- 성　　명 :

과목 I 데이터 이해 * 문항 수(8문항), 배점(문항 당 2점)

01. 다음 중 데이터베이스와의 통신을 위해 고안된 언어로 가장 적절한 것은?

① Java
② R
③ Python
④ SQL

02. 다음 중 데이터 사이언티스트의 필요 역량으로 가장 부적절한 것은?

① 설득력 있는 스토리텔링
② 통찰력 있는 분석
③ 네트워크 최적화
④ 다분야 간 협력을 위한 커뮤니케이션

03. 다음 중 빅데이터 위기 요인 중 사생활 침해를 막기 위한 방지 기술로 적절한 것은 무엇인가?

① 익명화
② 일반화
③ 정규화
④ 표준화

04. 다음 중 빅데이터가 발생시키는 문제를 중간자 입장에서 중재하고 해결하는 역할을 하는 직업은 무엇인가?

① 데이터 관리자
② 알고리즈미스트
③ 정보보안 전문가
④ 애널리스트

05. 다음 중 빅데이터 및 데이터 사이언스 등의 기술이 가져올 변화로 가장 적절하지 않은 것은?

① 해당 기술은 비용절감, 고객 서비스 향상, 내부 의사결정 지원 등에서 엄청난 가치를 발견할 것이다.
② 급변하는 환경에서 예측하지 못했던 전환이나 위기에 빨리 적응할 수 있게 할 것이다.
③ 사물인터넷의 적용으로 사람의 개입이 최대화 되어 실시간으로 데이터를 수집할 것이다.
④ 디지털화된 정보와 대상들이 서로 연결되기 때문에 연결이 얼마나 원활할 지가 중요해질 것이다.

06. 다음 중 빅데이터 기술의 활용에 대한 설명으로 가장 적절하지 않은 것은?

① 기업 활용 사례로서 구글 검색 기능, 월마트 매출 향상 등이 있다.
② 정부 활용 사례로서 실시간 교통 정보 제공, 기후 정보 제공, 각종 지원 활동 예측 등이 있다.
③ 정부는 이익을 목적으로 개인의 정보를 활용할 수 있는 방안을 모색한다.
④ 가수는 팬들의 음악 청취 기록을 분석해 공연의 음악 순서 방안을 모색한다.

07. 빅데이터 시대 위기 요인으로 가장 부적절한 것은?

① 데이터 오용
② 사생활 침해
③ 데이터 분석 예측
④ 책임원칙 훼손

08. 다음 중 사용자와 데이터베이스 사이에서 사용자의 요구에 따라 정보를 처리해주고 데이터베이스를 관리해 주는 소프트웨어는 무엇인가?

① SQL
② ERD
③ Data Dictionary
④ DBMS

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

09. 다음 중 아래의 데이터 거버넌스 체계가 설명하는 항목은?

아래

메타데이터 관리, 데이터 사전관리, 데이터 생명주기 관리

① 데이터 표준화
② 데이터 관리 체계
③ 데이터 저장소 관리
④ 표준화 활동

10. 다음 중 데이터 거버넌스의 구성 요소가 아닌 것은?

① 원칙(principle)
② 조직(organization)
③ 데이터 매니지먼트(Data Management)
④ 절차(process)

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

① 가치는 투자비용 요소이다.
② 전략적 중요도, ROI, 실행 용이성은 분석 과제 우선순위 결정에 고려할 사항이다.
③ 시급성과 전략적 필요성은 전략적 중요도의 평가 요소이다.
④ 적용 기술의 안전성 검증은 기술 용이성의 평가 요소이다.

12. 기업의 데이터 분석 도입 수준을 명확하게 파악하기 위한 방법으로 분석 준비도(readiness)를 진단 할 수 있다. 다음 중 분석 준비도를 측정하기 위한 요소가 아닌 것은?

① 분석 목표 및 전략
② 분석 기법
③ 분석 데이터
④ 분석 인력 및 조직

13. 다음 중 난이도와 시급성을 고려하였을 때 우선적으로 추진해야 하는 분석 과제로 적절한 것은?

① 난이도 : 쉬움(Easy), 시급성 : 현재
② 난이도 : 어려움(Difficult), 시급성 : 미래
③ 난이도 : 쉬움(Easy), 시급성 : 미래
④ 난이도 : 어려움(Difficult), 시급성 : 현재

14. 아래에서 설명하는 데이터 분석 조직 구조로 가장 적절한 것은?

아래

분석 조직 인력을 현업 부서에 배치하여 분석 업무를 수행하는 형태로서, 전사 차원에서 분석 과제의 우선순위를 선정하여 수행할 수 있고, 분석 결과를 신속하게 실무에 적용할 수 있다.

① 집중 구조
② 기능 구조
③ 분산 구조
④ 확산 구조

15. 분석 과제 발굴의 상향식 접근법에서 프로세스 분석을 통한 절차로 가장 적절한 것은?

① 분석 요건 정의 → 분석 요건 식별 → 프로세스 분류 → 프로세스 흐름 분석
② 분석 요건 식별 → 프로세스 흐름 분석 → 프로세스 분류 → 분석 요건 정의
③ 프로세스 흐름 분석 → 프로세스 분류 → 분석 요건 정의 → 분석 요건 식별
④ 프로세스 분류 → 프로세스 흐름 분석 → 분석 요건 식별 → 분석 요건 정의

16. 다음 중 빅데이터 분석 방법론의 분석 기획 단계에서 프로젝트 위험 대응 계획을 수립할 때 예상되는 위험에 대한 대응 방법의 구분으로 부적절한 것은?

① 회피(Avoid)
② 관리(Manage)
③ 완화(Mitigate)
④ 수용(Accept)

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

17. 다음 중 lasso 회귀모형에 대한 설명으로 부적절한 것은?

① 모형에 포함된 회귀계수들의 절대값의 크기가 클수록 penalty를 부여하는 방식이다.
② 자동적으로 변수선택을 하는 효과가 있다.
③ Lpenalty의 정도를 조정하는 모수가 있다.
④ L2 penalty를 사용한다.

18. 아래 데이터 셋(data set) A, B간의 유사성을 맨하탄 거리로 계산하면?

물 품	A	B
키	180	175
몸무게	65	70

① 0
② 10
③ $\sqrt{10}$
④ $\sqrt{50}$

19. 혼합분포군집 모형의 특징으로 적절하지 않은 것은?

① 확률분포를 도입하여 군집을 수행하는 모형 기반 군집 방법이다.
② 군집을 몇 개의 모수로 표현할 수 있다.
③ 모수 추정에서 데이터가 커지면 수행하는 데 시간이 걸릴 수 있다.
④ 군집의 크기가 작을수록 추정의 정도가 쉽다.

20. 다음 중 분류(Classification) 모델링에 대한 설명으로 가장 적절한 것은?

① 데이터의 이해를 더 쉽게하기 위해 데이터를 특정 기준으로 분류 및 범주화하고 등급화하는 방법을 말한다.
② 같이 팔리는 물건과 같이 유사 아이템을 분류하는 것을 의미한다.
③ 군집분석과 동일하게 레코드 자체가 먼저 분류 되어있지 않아도 적용할 수 있다.
④ 대표적인 분석 방법으로 장바구니 분석 기법이 존재한다.

21. 다음 중 아래의 표가 나타내는 확률질량함수를 가진 확률변수 x의 기댓값 E(x)로 가장 적절한 것은?

X	1	2	3	4
f(x)	0.5	0.3	0.2	0

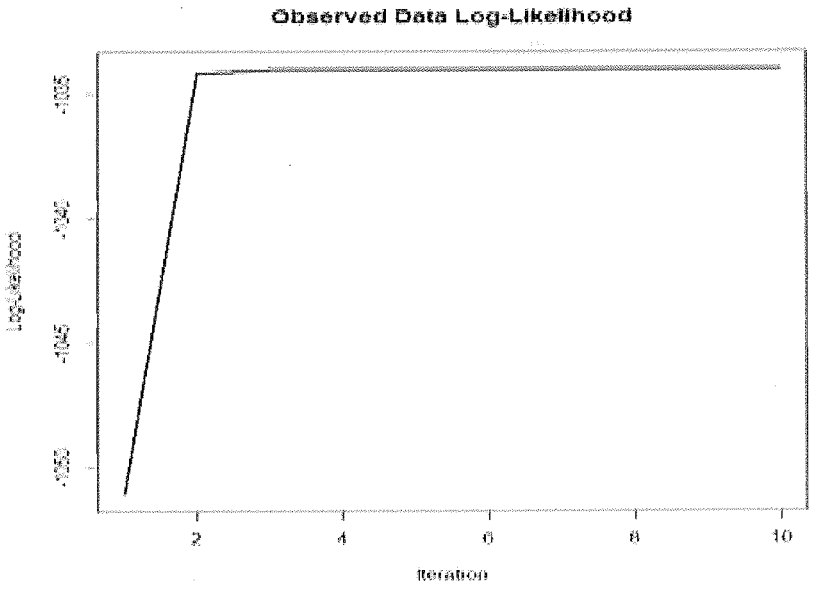
① 1

② 1.7

③ 2.5

④ 10

22. EM 알고리즘을 사용하여 혼합분포 모형을 추정하고자 한다. 아래와 같은 그래프가 도출되었을 때, 다음 중 가장 적절한 해석은?



① 반복횟수 2회만에 로그-가능도 함수가 최대가 되었다.

② 정규혼합분포가 2가지로 관찰되었다.

③ 모수의 추정을 위해 10회 이상의 반복횟수가 필요하다.

④ 로그-가능도 함수의 최소값이 -1040이다.

23. 확률변수 X의 확률은 아래와 같이 나타낼 수 있다. 다음 중 옳은 것은?

아래

$$P(X=1)=\frac{1}{3},\ P(X=2)=\frac{1}{6},\ P(X=3)=\frac{1}{2}$$

① X의 기댓값은 13/6이다.

② X가 1 혹은 2일 확률은 1/2 보다 크다.

③ X가 4일 확률은 0보다 크다.

④ X가 1, 2, 3 중 하나의 값을 가질 확률은 1보다 작다.

24. 아래 오분류표를 이용하여 계산된 정밀도는 무엇인가?

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

① 3/10

② 4/10

③ 3/9

④ 7/11

25. Credit 데이터는 400명의 신용카드 고객에 대해 신용카드 대금(balance)과 소득(income), 학생여부(student=Y/N)를 포함한다. Balance를 종속변수로 하는 아래의 모형 적합 결과 중 가장 부적절한 것은?

아래

```
Call:
lm(formula = Balance ~ (Income + Student)^2, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-773.39 -325.70  -41.13  321.65  814.04

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       200.6232    33.6984   5.953 5.79e-09 ***
Income              6.2182     0.5921  10.502  < 2e-16 ***
StudentYes        476.6758   104.3512   4.568 6.59e-06 ***
Income:StudentYes  -1.9992     1.7313  -1.155    0.249
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 391.6 on 396 degrees of freedom
Multiple R-squared:  0.2799,Adjusted R-squared:  0.2744
F-statistic:  51.3 on 3 and 396 DF,  p-value: < 2.2e-16
```

① 위의 모형은 Balance를 설명하는데 통계적으로 유의하다.

② Income이 증가할수록 Balance가 증가하는 경향이 있다.

③ Income과 StudentYes의 교호작용은 유의하지 않다.

④ Income이 증가함에 따라 커지는 Balance의 증가분이 학생 여부에 따라 유의적인 차이가 있다.

26. 다음 중 회귀분석에서 모형의 설명력을 확인하기 위해 사용되는 결정계수의 특성으로 부적절한 것은?

① 결정계수는 0에서 1의 값을 가진다.

② 높은 값을 가질수록 측정된 회귀식의 설명력이 높다.

③ 종속변수와 독립변수 사이의 표본상관 계수값과 같다.

④ 총 변동에서 추정된 회귀식에 의해 설명되는 변동의 비율로 나타낼 수 있다.

27. 다음 중 목표변수가 연속형인 회귀나무에서 분류 기준값의 선택 방법으로 가장 적절한 것은?

① 카이제곱 통계량, 지니지수
② 지니지수, F-통계량
③ F-통계량, 분산 감소량
④ 분산 감소량, 엔트로피 지수

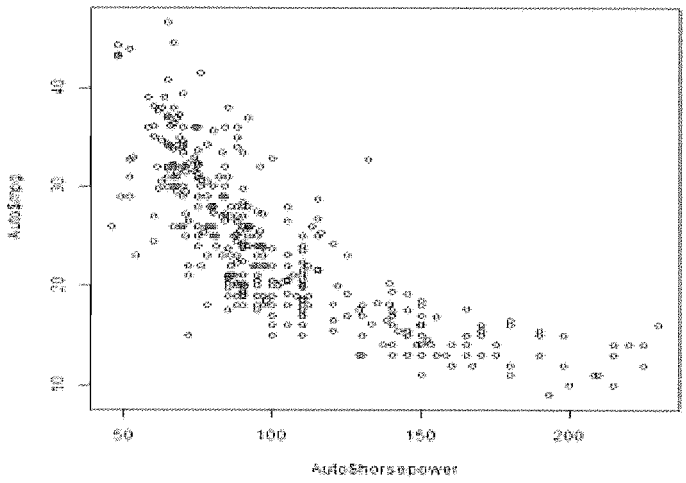
28. 다음 중 자기조직화지도(Self-Organizing Maps, SOM)에 대한 것으로 옳지 않은 것은?

① SOM 모델은 입력층과 경쟁층으로 구성되어 있다.
② 입력층의 뉴런은 경쟁층에 있는 뉴런들과 부분적으로(locally) 연결되어 있다.
③ 단 하나의 전방 패스를 사용함으로써 속도가 매우 빠르다.
④ 경쟁 학습으로 연결 강도를 반복적으로 재조정하여 학습한다.

29. 다음 중 선형회귀모형이 통계적으로 유의미한지 평가하는 통계량으로 가장 적절한 것은?

① F-statistics
② T-statistics
③ Chi-statistics
④ R-square

30. 아래 그래프는 392대의 자동차에 대한 연비(mpg)와 엔진 마력(horsepower)를 포함하고 있다. 다음 중 이에 대한 설명으로 가장 적절하지 않은 것은?



① mpg를 설명하기 위해 horsepower를 설명변수로 하는 단순선형회귀모형은 적절하다.
② horsepower가 증가할수록 mpg가 감소하는 경향이 있다.
③ mpg와 horsepower 간의 피어슨 상관계수는 두 변수의 관계를 잘 설명하지 못할 수도 있다.
④ mpg와 horsepower는 음의 상관관계를 가진다.

31. 다음 중 연관규칙의 측정 지표 중 품목 A, B에 대한 지지도를 구하기 위한 식으로 적절한 것은?

① (A 또는 B가 포함된 거래 수)/(전체 거래 수)
② (A와 B가 동시에 포함된 거래 수)/(전체 거래 수)
③ (A와 B가 동시에 포함된 거래 수)/(A를 포함하는 거래 수)
④ (A와 B가 동시에 포함된 거래 수)/(A 또는 B가 포함된 거래 수)

32. 다음 중 아래 데이터 마이닝 추진 단계를 순서대로 나열한 것은?

아래

가. 목적 정의
나. 데이터 준비
다. 데이터 가공
라. 데이터 마이닝 기법 적용
마. 검증

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

33. 아래에서 설명하는 활성화 함수로 가장 적절한 것은?

아래

입력층이 직접 출력층에 연결되는 단층 신경망에서 이 활성화 함수를 사용하면 로지스틱 회귀모형의 작동원리가 유사해진다.

① 계단 함수
② tanh 함수
③ ReLU 함수
④ 시그모이드 함수

34. 다음 중 로지스틱 회귀모형에 대한 설명으로 가장 적절한 것은?

① 일반적으로 반응변수가 범주형인 경우에 적용되는 모형이다.
② 시계열 예측에서 가장 많이 활용되는 모형 중 하나이다.
③ 반응변수가 비율 척도일 때, 많이 활용되는 모형 중 하나이다.
④ 로지스틱 회귀모형은 오즈의 관점에서 해석할 수 없다.

35. 다음 중 군집분석에 대한 설명으로 가장 적절하지 않은 것은?

① 분할적 군집은 모든 데이터를 단일 군집에 속한다고 정의하고 시작하는 방법으로 상위 군집에서 잘못된 결정을 하면 하위 군집에 파급되는 정도가 크다는 단점이 있다.

② k-평균법은 중심으로부터 거리를 기반으로 군집화하기 때문에 구형으로 뭉쳐져 있는 볼록(Convex)한 데이터 세트에서는 비교적 잘 작동되나 오목한(non-convex) 형태의 군집 모델은 특성을 구별해내는 데 성능이 떨어진다.

③ k-medoid 모델은 실제 데이터에 있는 값을 중심점으로 하기 때문에 이상값이나 잡음(noise) 처리에 매우 우수하나, k-평균법에 비해 계산량이 많다는 단점이 있다.

④ 밀도 기반 클러스터링(DBSCAN) 모델은 밀도 있게 연결된 데이터 집합을 동일한 군집으로 판단하는 방법이지만 k-평균법 모델처럼 오목한 형태의 데이터 세트에서는 군집 특성을 잘 찾아내지 못한다.

36. 다음 중 예측모형의 과적합을 방지하기 위해 활용되는 자료 추출 방법으로 가장 적절하지 않은 것은?

① 홀드아웃 방법

② 교차검증

③ 부스트랩

④ 의사결정나무

37. 아래 오분류표를 이용하여 계산된 특이도는 무엇인가?

		예측치		합계
		True	False	
실제값	True	200	400	600
	False	300	100	400
합계		500	500	1000

① 0.20

② 0.25

③ 0.75

④ 0.80

38. 다음 중 군집의 개수를 미리 정하지 않아도 되어 탐색적 분석에 사용하는 군집 모형으로 적절한 것은?

① k-평균군집 모형
② SOM 모형
③ 계층적군집 모형
④ 혼합분포군집 모형

39. 다음 중 연관분석의 설명으로 가장 적절한 것은?

① 품목 수와 상관없이 분석에 필요한 계산은 일정하다.
② 세분화된 품목에 대해 연관 규칙을 찾으려 할 때 적절한 방법이다.
③ 상대적으로 거래량이 적은 품목에 대해서 적용하기 좋은 방법이다.
④ 조건 반응(if-then)으로 표현되는 연관분석의 결과를 이해하기 쉽다.

40. 다음 중 시계열 데이터의 정상성(stationary)에 대한 설명으로 가장 적절하지 않은 것은?

① 비정상 시계열 자료는 정상성을 만족하도록 데이터를 정상 시계열로 만든 후에 시계열 분석을 수행한다.
② 정상 시계열은 어떤 일정한 값을 중심으로 일정한 변동 폭을 가진다.
③ 시계열 자료가 추세를 보이는 경우에는 차분(differencing)을 통해 비정상 시계열을 정상 시계열로 바꿀 수 있다.
④ 시계열 자료가 정상성을 만족하는지 판단하기 위해 시계열 자료 그림을 통해 자료의 이상점 등을 살핀다.

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

01. 아래에서 설명하는 것은 무엇인가?

아래

문자, 기호, 음성, 화상, 영상 등 상호 연관괸 다수의 콘텐츠를 정보 처리 및 정보통신 기기에 의하여 체계적으로 수집·축적하여 다양한 용도와 방법으로 이용할 수 있도록 정리한 정보의 집합체

( )

02. 아래에서 설명하고 있는 빅데이터 활용 기본 테크닉은 무엇인가?

아래

가) 생명의 진화를 모방하여 최적해(Optimal Solution)를 구하는 알고리즘으로 존 홀랜드(John Holland)가 1975년에 개발하였다.
나) '최대의 시청률을 얻으려면 어떤 시간대에 방송해야 하는가?'와 같은 문제를 해결할 때 사용된다.
다) 어떤 미지의 함수 Y=f(x)를 최적화하는 해 x를 찾기 위해, 진화를 모방한(Simulated evolution) 탐색 알고리즘이라고 말할 수 있다.

( )

03. 문제가 주어지고 이에 대한 해법을 찾기 위하여 각 과정이 체계적으로 단계화되어 수행하는 분석 과제 발굴 방식을 무엇이라고 하는가?

( )

04. 아래에서 설명하는 데이터 분석 조직 구조는 무엇인가?

아래

- 전사 분석업무를 별도의 분석 전담 조직에서 담당
- 전략적 중요도에 따라 분석조직이 우선순위를 정해서 진행 가능
- 현업 업무부서의 분석업무와 이중화/이원화 가능성 높음

( )

05. 아래에서 설명하는 시각화 방법은?

아래

여러 대상 간의 관계에 관한 수치적 자료를 이용해 유사성에 대한 측정치를 상대적 거리로 시각화하는 방법이다.

( )

06. 최적방정식을 선택하기 위한 방법 중 모든 독립변수 후보를 포함한 모형에서 시작하여 가장 적은 영향을 주는 변수를 하나씩 제거하면서 더 이상 유의하지 않은 변수가 없을 때까지 설명변수를 제거하는 방법은 무엇인가?

( )

07. 모형 평가방법 중 주어진 원천 데이터를 랜덤하게 두 분류로 분리하여 교차 검정을 실시하는 방법으로 하나는 모형의 학습 및 구축을 위한 훈련용 자료로, 다른 하나는 성과 평가를 위한 검증용 자료로 사용하는 방법은 무엇인가?

( )

08. P(A)=0.3, P(B)=0.4이다. 두 사건 A와 B가 독립일 경우 P(B|A)는 얼마인가?

( )

09. 이산형 확률분포 중 주어진 시간 또는 영역에서 어떤 사건의 발생 횟수를 나타내는 확률 분포는 무엇인가?

( )

10. 아래 빈칸에 들어갈 용어는?

아래

( )은/는 계층적 군집분석 방법 중 하나로 군집과 군집, 또는 데이터와의 거리계산 시, 최단거리를 계산하여 거리가 가까운 데이터, 또는 군집을 새로운 군집으로 형성하는 방법이다. 이 방법은 사슬 구조의 군집이 생길 수 있다.

( )

ADsP 30회 기출문제 답안

【 객관식 정답 】

01	③	11	①	21	③	31	④
02	①	12	②	22	②	32	④
03	④	13	④	23	④	33	③
04	①	14	④	24	①	34	②
05	④	15	②	25	②	35	③
06	④	16	①	26	①	36	①
07	②	17	④	27	③	37	①
08	②	18	③	28	④	38	③
09	②	19	④	29	③	39	①
10	③	20	①	30	①	40	③

【 주관식 정답 】

01	정보
02	ERP
03	a : 마스터 데이터, b : 메타데이터, c : 데이터사전
04	시급성
05	4개
06	앙상블 기법
07	25%
08	포아송분포(Poisson Distribution)
09	0.8
10	머신러닝 or 기계학습

01. 빅데이터가 만들어내는 본질적인 변화는 사전처리에서 사후처리, 표본조사에서 전수조사, 질에서 양, 인과관계에서 상관관계이다.

02. 구글, 애플 등의 기업에서는 정형화된 데이터 이외에 비정형 데이터를 수집하여 서비스에 활용하고 있다.

03. 미래 사회의 특성과 빅데이터의 역할은 불확실성 - 통찰력, 리스크 - 대응력, 스마트 - 경쟁력, 융합 - 창조력의 관계를 맺고 있다.

04. 빅데이터를 활용해 기업의 혁신, 경쟁력제고, 생산성이 향상되었으며, 정부입장에서는 환경탐색, 상황분석, 미래대응에 활용할 수 있게 되었다. 데이터 수집 및 저장은 일반적인 데이터를 활용해서도 가능하다.

05. 데이터 분석가는 데이터를 기반으로 의사결정을 하기 때문에 천재적 직관력은 필요한 능력이 아니다.

06. 데이터베이스의 일반적인 특징은 통합된 데이터(Integrated Data), 저장된 데이터(Stored Data), 공용 데이터(Shared Data), 변화되는 데이터(Changable Data)이다.

07. 빅데이터 가치 산정이 어려운 이유는 새로운 가치 창출, 데이터 활용 방식, 분석 기술의 발전이다.

08. 일차원적 분석 어플리케이션 중 에너지는 트레이딩, 공급/수요 예측이 있다.

09. 데이터 거버넌스 체계 중 데이터 관리 체계는 데이터 정합성 및 활용의 효율성을 위하여 표준데이터를 포함한 메타 데이터와 데이터 사전의 관리 원칙을 수립한다. 빅데이터의 경우 데이터 양의 급증으로 데이터의 생명 주기 관리 방안을 수립하지 않으면 데이터 가용성 및 관리비용 증대 문제에 직면하게 될 수 있다.

10. 데이터 거버넌스의 구성요소는 원칙(Principle), 조직(Organization), 프로세스(Process)는 유기적으로 조합하고 효과적으로 관리하여, 데이터를 비즈니스 목적에 부합하도록 하고 최적의 정보 서비스를 제공할 수 있도록 한다.

11. 과제 발굴 단계에서는 세부적인 구현 및 솔루션에 초점을 맞추는게 아니라, 문제를 해결함으로써 발생하는 가치에 중점을 두는 것이 중요하다.

12. 과제 중심적인 접근 방식의 특징에는 Speed & Test, Quick-Win, Problem Solving이 해당하며 Accuracy & Deploy는 장기적인 마스터 플랜 방식에 해당하는 내용이다.

13. 상향식 접근 방식은 프로토 타이핑 등을 활용하여 반복적으로 결과를 개선해 나간다. Design Thinking의 Ideate 단계에는 diverge 단계에 해당한다.

14. 분석과제 우선순위 고려요소는 전략적 중요도, 비즈니스 성과/ROI, 실행 용이성이 있다.

15. 분석 방안 구체화에는 의사결정 요소 모형화, 분석 체계 도출, 분석 필요 데이터 정의, 분석 ROI 평가, 분석 활용 시나리오 정의, 분석 정의서 작성, 전사관점 분석적용이 있다.

16. 분석 준비도는 분석업무파악, 인력 및 조직, 분석기법, 분석 데이터, 분석문화, IT 인프라 총 6가지로 구성되어 있다.

17. 민코우스키거리는 맨하탄 거리와 유클리디안 거리를 한번에 표현한 공식으로 L1 거리(맨하탄거리), L2 거리(유클리디안 거리)라 불리고 있다.

18. 마할라노비스 거리는 통계적 개념이 포함된 거리이며 변수들의 산포를 고려하여 이를 표준화한 거리이다.
두 벡터 사이의 거리는 산포를 의미하는 표본 공분산으로 나눠주어야 하며, 그룹에 대한 사전 지식 없이는 표본 공분산 S를 계산할 수 없으므로 사용하기 곤란하다.

19. 각 모형의 상호 연관성이 높을수록 정확도는 떨어진다.

20. k평균군집은 한번 군집이 형성되더라도 다른 군집으로 이동이 가능하다.

21. Start AIC보다 작은 값들 중에 가장 작은 변수를 제외한 나머지 설명변수로 다음 Step을 이어 나간다.

22. income은 p-value 값이 0.71152로 나타나 default를 설명하는데 통계적으로 유의하지 않게 나타나고 있다.

23. 로지스틱 회귀분석의 모형 검정방법은 카이제곱 검정으로 이루어지며, 선형회귀분석은 F-검정, T-검정으로 이루어진다.

24. 주성분 분석은 분산을 최대화하는 차원을 찾는 방법으로 공분산행렬에서 고윳값을 통해 찾을 수 있다.

25. 평균 고유값 방법은 고유값들의 평균을 구한 후 고유값이 평균값 이상이 되는 주성분을 제거하는 것이 아니라 설정하는 것이다.

26. 잔차분석에서 잔차는 독립성, 등분산성, 정규성을 만족해야 한다.

27. 순환변동은 경제적이나 자연적인 이유 없이 알려지지 않은 주기를 가지고 변화하는 자료를 의미한다.

28. 4번의 설명은 연속형 확률밀도함수에 대한 설명이다.

29. 지지도는 전체 거래 중 항목 A와 항목 B를 동시에 포함하는 거래의 비율이다.

30. 경쟁층은 입력벡터의 특성에 따라 벡터가 한 점으로 클러스터링 되는 층이다.

31. DFFITS는 절대값이 공식에 대입한 값보다 큰 값이 나타나야 높은 영향력으로 간주한다.

32. Durbin-Watson test는 회귀 모형 오차항의 자기상관이 있는지에 대한 검정이다. 히스토그램, Q-Q plot, Shaprio-Wilk 검정 등을 활용하여 데이터의 정규성을 확인한다.

33. 1종 오류는 귀무가설이 옳은데도 귀무가설을 기각하게 되는 오류이다.

34. 이상치는 이상치 자체의 의미가 있을 수 있어 평균으로부터 3*표준편차를 벗어나더라도 제거하면 안된다.

35. (가)는 α/2인데, 여기서 α는 0.1이므로 α/2=0.05이다. (나)에는 표본수이므로 70이다.

36. 사회연결망 분석은 여러 개의 기법들로 구성되어 있고, 국내의 사회연결망 연구에서 많이 활용되고 있는 기법은 중심성, 밀도, 중심화 등이 있다. 그 중 밀도는 연결망 내에서 전체 구성원들이 서로 간에 얼마나 많은 관계를 맺고 있는가를 나타낸다.

37. Weight의 중앙값은 258.0이다.

38. 전체 관측치 수는 50개이다.

39. 안정적 시계열은 시간의 추이와 관계없이 평균, 분산이 불변하여, 변화했다고 해도 다시 평균으로 회귀하는 경향을 보인다.

40. 유의수준 0.05와 p-value를 비교했을 때, p-value 값이 현저하게 작게 나타나 귀무가설이 기각되어 '대학의 평균 교재 구입비용이 570달러와 같다'는 기각된다

【단답형】

단답형 01. 정보는 데이터의 가공, 처리와 데이터간 연관관계 속에서 의미가 도출된 것이다.

단답형 02. ERP는 기업내부 데이터베이스 중 기업 전체가 경영자원을 효과적으로 이용하기 위해서 기업의 모든 자원을 통합하여 관리하기 위한 기업 경영정보 시스템이다.

단답형 03. 데이터 거버넌스에서 마스터 데이터, 메타데이터, 데이터 사전은 중요한 관리대상이다.

단답형 04. 시급성은 전략적 중요도와 목표가치(KPI)이 핵심이다.

단답형 05. BIC 값이 가장 큰 지점을 찾았을 때, 4가 가장 크므로 최적의 군집의 수는 4개라고 할 수 있다.

단답형 06. 앙상블 기법은 주어진 자료로부터 여러 개의 예측모형들을 만든 후 예측모형들을 조합하여 하나의 최종 예측모형을 만드는 방법이다.

단답형 07. 데이터를 크기순으로 나열했을 때, 3사분위수가 92이라고함은 데이터의 75%위치에 있는 값이 92이기 때문에 92보다 작은 값이 75%정도 있을 것이며, 나머지 25%는 92보다 큰 값을 가진다.

단답형 08. 포아송 분포(Poisson Distribution)는 단위 시간 안에 어떤 사건이 몇번 발생할 것인지를 표현하는 이산확률분포이다.

단답형 09. 민감도가 0.8이면 TP=4FN이다. 실제/예측 TRUE와 실제/예측 FALSE가 100으로 통일이라면 아래와 같이 된다. TP+FP=100, FN+TN=100, TP+FN=100, FP+TN=100 이 중 TP+FN=100에서 FN=20 이므로 TP=80이다. 그러면 FP=20, TN=80이므로 Precision을 구하면 TP/TP+FP=80/100=0.8이다.

단답형 10. 머신러닝은 학습과 개선을 위해 명시적으로 컴퓨터를 프로그래밍하는 대신, 컴퓨터가 데이터로 학습하고 경험을 통해 개선하도록 훈련하는 데 중점을 둔다.

ADsP 31회 기출문제 답안

【 객관식 정답 】

01	③	11	②	21	①	31	③
02	①	12	④	22	①	32	③
03	④	13	①	23	②	33	④
04	①	14	②	24	②	34	③
05	②	15	③	25	④	35	④
06	④	16	②	26	③	36	①
07	①	17	②	27	④	37	④
08	③	18	①	28	④	38	③
09	③	19	③	29	④	39	③
10	③	20	②	30	④	40	②

【 주관식 정답 】

01	머신러닝 또는 기계학습
02	KMS(지식관리시스템)
03	ISP(정보전략계획)
04	IT 인프라
05	나이브 베이즈 분류
06	최단연결법(단일연결법)
07	스테밍(Stemming) 또는 어간 추출
08	차분
09	57.5%
10	배깅(Bagging)

01. 사물인터넷(Internet of Things)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.

02. 기능구조는 별도 분석조직이 없고 해당업무부서에서 분석 수행한다.

03. 통합된 데이터는 동일한 내용의 데이터가 중복되어 있지 않다는 것을 의미한다.

04. (a)는 가공하기 전의 순수한 수치나 기호인 데이터를 의미하며, (b)는 데이터의 가공 및 상관관계 이해를 통해 패턴 인식하고 의미를 부여하는 정보를 의미한다. (c)는 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물인 지식을 의미한다.

05. 재무관리분야의 분석 유즈 케이스에는 일별로 예정된 자금 지출과 입금을 추정하는 자금시재예측 등이 있다.

06. 다) 개인정보 사용자의 정보사용에 대한 무한책임의 한계로 개인정보 사용 동의제보다 책임체로 더욱 강화시켜야 한다.
라) 민주주의 국가의 형사 처벌과 같이 잠재적 위협이 아닌 명확하게 행동한 결과에 대해 책임을 묻기 때문에 빅데이터 사전 성향 분석을 실시한다면 책임 원칙을 훼손한다.

07. 데이터베이스는 종속성과 중복성을 배제한다. 데이터 종속성이란 응용프로그램별로 데이터를 별도 관리한다.

08. 빅데이터의 경우 데이터양의 급증으로 데이터의 생명 주기(수명주기) 관리방안을 수립하지 않으면 데이터 가용성 및 관리비용 증대되는 문제에 직면하게 될 수 있다.

09. 분산구조는 별도 분석전담조직, 분석조직 인력을 현업부서로 직접 배치한다.

10. 솔루션은 분석대상은 알고 있지만, 분석의 방법을 모를 때 나타나는 분석 주제의 유형이다.

11. 분석 유즈 케이스는 기업의 전사 또는 개별 업무별 주요 의사결정 포인트에 활용할 수 있는 분석의 후보들을 의미한다.

12. 반복적으로 위험분석을 수행하여 위험을 관리하며 순환적으로 개선하는 것은 나선형 모델이다.

13. 분석과제 발굴 방법론에서 상향식 접근 방식은 문제의 정의 자체가 어려운 경우, 데이터를 기반으로 문제의 재정의 및 해결방안을 탐색하고 이를 지속적으로 개선하는 방식이다.

14. (A) 단순히 대용량 데이터를 수집·축적하는 것보다는 어떤 목적으로 어떤 데이터를 어떻게 분석에 활용할 것인가가 더욱 중요하다.
    (B) 빅데이터의 경우 데이터양의 급증으로 데이터의 생명 주기 관리방안을 수립하지 않으면 데이터 가용성 및 관리비용 증대 문제에 직면할 수 있다.

15. 전략적 통찰력을 얻기 위해서는 내부뿐만 아니라 외부환경을 같이 분석해야 한다.

16. 분석 마스터플랜의 우선순위 고려요소는 전략적 중요도, 비즈니스 성과/ROI, 실행 용이성이 있으며, 적용 범위/방식의 고려요소는 업무 내재화 적용 수준, 분석데이터 적용 수준, 기술적용 수준이 있다.

17. 데이터 마트는 데이터웨어하우스와 사용자 사이에 위치한 것으로, 하나의 주제 또는 하나의 부서 중심의 데이터웨어하우스라고 할 수 있다.

18. 지지도는 전체 거래 중 항목 A와 항목 B를 동시에 포함하는 거래의 비율이다.

19. 향상도는 $\frac{P(B \mid A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)}$ 로 계산할 수 있다.

20. C의 지니지수는 $1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2 = 0.48$ 로 구할 수 있다.

21. MAPE는 MAE를 퍼센트로 변환한 것으로 실제값(A_t)과 예측값(F_t)으로 구할 수 있다.

$$MAPE = \left(\sum_{t=1}^{n} \frac{|A_t - F_t|}{A_t}\right) \div n \times 100 = 0.4/4 \times 100 = 10\%$$

22. 평균은 이상치에 민감한데 이를 보완하기 위해 군집의 가장 중심에 있는 값(메도이드(medoid))을 사용하여 군집을 찾는 방법이 k-medoids 군집화 방식이다.

23. 부스팅은 붓스트랩 표본을 구성하는 재표본 과정에서 분류가 잘못된 데이터에 더 큰 가중치를 주어 표본을 추출하는 기법/예측력이 약한 모형들을 결합하여 강한 예측모형을 만드는 방법이다.

24. 일반적으로 학습 모형의 유연성이 클수록 분산은 높고 편향은 낮다.

25. 데이터 마이닝은 대용량 데이터에서 의미있는 패턴을 파악하거나 예측하여 의사결정에 활용하는 방법이다.

26. Estimate값인 회귀계수는 Intercept의 회귀계수와의 차이를 의미하고 종속변수 wage의 평균과 차이는 아니다.

27. SOM은 비지도 학습에 해당하고 나머지 항목은 지도학습에 해당한다.

28. k-means clustering, Single linkage method, DBSCAN은 군집분석 방법이다.

29. Q1 - 1.5 * IQR ≤ 데이터 ≤ Q3 + 1.5* IQR 범위를 벗어난 데이터를 이상치라고 한다. 160은 중앙값(Q2)이고 백분율을 4등분 하였으므로 각각의 범위에는 25%의 데이터가 있다. IQR은 Q3-Q1으로 200-140=60이다.

30. 층화추출법은 이질적인 원소들로 구성된 모집단에서 각 계층을 고루 대표할 수 있도록 표본을 추출하는 방법.

31. 연관분석은 흔히 장바구니분석 또는 서열분석이라고 불리며, 상품의 구매, 서비스 등 일련의 거래 또는 사건들 간의 규칙을 발견하기 위해 적용한다.

32. 스피어만 상관계수는 서열척도인 두 변수들의 상관관계 측정방식으로 순위를 기준으로 상관관계를 측정하는 비모수적 방법이다.

33. 연결정도는 해당 노드에 직접 연결되어 있는 노드 또는 링크의 수이다.

34. 마할라노비스 거리는 통계적 개념이 포함된 거리이며 변수들의 산포를 고려하여 이를 표준화한 거리이다.

35. 분류분석은 레코드의 특정 속성의 값이 범주형으로 정해져 있으며 데이터의 실체가 어떤 그룹에 속하는지 예측하는데 사용되는 기법으로 사기방지모형, 이탈모형, 고객 세분화 모형 등을 개발할 때 활용한다.

36. matrix 함수는 행렬을 만드는 함수로 c(1,2,3,4,5,6)를 통해 행렬을 구성하는 원소를 입력한다.
그리고 ncol=2는 컬럼의 개수이며, byrow=T는 행방향(가로축)으로 채워넣는 인자이다.

37. 가, 다, 마는 지도학습이며, 나, 라는 비지도 학습이다.

38. 변수 k의 분산팽창요인 $\mathrm{VIF}_k = \frac{1}{1-R_k^2}$ 으로 결정계수에 영향을 받는다. 결정계수는 회귀모델에서 독립변수가 종속변수를 얼마나 잘 설명하는지를 나타내는 것으로 회귀식의 기울기와는 관계가 없다.

39. 순환변동은 경제적이나 자연적인 이유 없이 알려지지 않은 주기를 가지고 변화하는 자료를 의미한다.

40. 평균 고유값 방법은 고유값들의 평균을 구한 후 고유값이 평균값 이상이 되는 주성분을 제거하는 것이 아니라 설정하는 것이다.

【 단답형 】

단답형 01. 머신러닝은 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야이다.

단답형 02. 지식관리시스템은 기업의 환경이 물품을 주로 생산하던 산업사회에서 지적 재산의 중요성이 커지는 지식사회로 급격히 이동함에 따라, 기업 경영을 지식이라는 관점에서 새롭게 조명하는 접근방식이다.

단답형 03. 정보기술 또는 정보시스템을 전략적으로 활용하기 위하여 조직 내/외부 환경을 분석하여 기회나 문제점을 도출하고 사용자의 요구사항을 분석하여 시스템 구축 우선순위를 결정하는 등 중장기 마스터 플랜을 수립하는 절차이다.

단답형 04. 분석 준비도 중 IT인프라에는 운영시스템 데이터 통합, EAI, ETL 등 데이터 유통 체계, 분석 전용 서버 및 스토리지, 빅데이터 분석 환경, 통계 분석 환경, 비쥬얼 분석 환경이 있다.

단답형 05. 나이브 베이즈 분류는 특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기의 일종으로 1950년대 이후 광범위하게 연구되고 있다.

단답형 06. 최단연결법은 가장 가까운 데이터를 묶어서 군집을 형성한다.

단답형 07. 스테밍(Stemming)은 어형이 변형된 단어로부터 접사 등을 제거하고 그 단어의 어간을 분리해내는 것을 의미한다.

단답형 08. 차분은 시계열의 수준에서 나타내는 변화를 제거하여 시계열의 평균 변화를 일정하게 만드는 것을 돕는다.

단답형 09. 첫 번째 분산(Proportion of Variance)은 0.5748331로 나타났다.

단답형 10. 배깅(Bagging)은 원 데이터 집합으로부터 크기가 같은 표본을 여러 번 단순임의 복원추출하여 각 표본에 대해 분류기를 생성한 후 그 결과를 앙상블하는 방법이다.

ADsP 32회 기출문제 답안

【 객관식 정답 】

01	③	11	②	21	③	31	③
02	②	12	④	22	③	32	④
03	①	13	④	23	③	33	②
04	③	14	①	24	④	34	④
05	①	15	④	25	②	35	④
06	①	16	①	26	②	36	④
07	②	17	④	27	④	37	②
08	②	18	③	28	④	38	③
09	④	19	①	29	②	39	①
10	②	20	④	30	③	40	①

【 주관식 정답 】

01	사물인터넷 (IoT, Internet of Things)
02	SCM(Supply Chain Management)
03	프레이밍 효과(Framing effect)
04	활용
05	분해 시계열
06	AR 모형
07	과대적합(Overfitting)
08	인공 신경망
09	이익도표
10	데이터 마트

01. SQL 명령어 중 DML(Data Manipulation Language)은 데이터 조작어로 SELECT, INSERT, UPDATE, DELETE가 있다. CREATE는 데이터 정의어인 DDL(Data Definition Language)에 포함된다.

02. 고객이 제품 1개에 대해 여러 명이 구매할 수 있다. 그러므로 N:1이 정답이다.

03. ERP는 회사의 모든 정보 뿐만 아니라, 공급 사슬관리, 고객의 주문정보까지 포함하여 통합적으로 관리하는 시스템이다. 경영, 인사, 재무, 생산 등 기업의 전반적 시스템을 하나로 통합함으로써 효율성을 극대화하는 경영 전략이다.

04. SVM은 분류분석의 기법 중 하나로 딥러닝과 관련 없는 분석 기법이다.

05. 판별분석, 회귀분석, 분류분석은 지도학습의 방법이며, 군집분석은 비지도학습의 방법이다.

06. 빅데이터 활용의 기본 3요소는 데이터, 기술, 인력이다.

07. 컨버전스에서 디버전스로의 변화, 생산에서 서비스로의 변화, 생산에서 시장창조로의 변화가 인문학 열풍을 가져오게 한 외부환경 요소이다.

08. 알고리즈미스트는 알고리즘에 의해 부당하게 피해입은 사람을 구제한다.

09. 조직 및 해당 산업에 폭넓게 영향을 미치는 사회·경제적 요인을 STEEP으로 요약되는 Social(사회), Technological(기술), Economic(경제), Environmental(환경), Political(정치) 영역으로 폭넓게 나눈다.

10. SoW(Statement of Works)는 프로젝트 관리분야에서 서비스를 제공하기 위한 활동, 산출물, 작업 시간 등을 포함하는 기술서이다.

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

12. 분석 ROI 요소에서 투자비용 요소는 크기(Volume), 다양성(Variety), 속도(Velocity)이다.

13. 데이터 거버넌스는 전사 차원의 모든 데이터에 대하여 정책 및 지침, 표준화, 운영조직 및 책임 등의 표준화된 관리체계를 수립하고 운영을 위한 프레임워크(Framework) 및 저장소(Repository)를 구축하는 것을 말한다.

14. 빅데이터 분석 방법의 5단계는 분석기획, 데이터 준비, 데이터 분석, 시스템 구현, 평가 및 전개 순으로 이루어진다.

15. 분석 기회 선별 방식 중 톱다운은 기업의 비즈니스 모델을 분석하여 경쟁력 강화를 위한 핵심 분석기회를 식별하며, 보텀업 경로 접근 방식은 특정 대상 프로세스를 선정한 후 주제별로 분석기회를 식별한다. 마지막으로 분석 유즈 케이스 벤치마킹 접근 방식은 제공되는 산업별, 업무 서비스별 분석 테마 후보 풀을 벤치마킹을 통한 분석 기회를 식별한다.

16. 분석 수준 진단의 대상은 분석 업무, 분석 인력/조직, 분석 기법, 분석 데이터, 분석 문화, 분석 인프라가 있다.

17. 유의수준은 귀무가설의 기각 여부를 결정하는데 사용하는 기준이 되는 확률로 귀무가설이 옳은데도 이를 기각하는 확률을 크기라고도 한다.

18. 리스트는 여러 자료형의 원소들이 포함될 수 있다.

19. Precision은 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율이다.

20. A→B의 신뢰도를 구하면 $\frac{P(A \cap B)}{P(A)} = \frac{300/1000}{600/1000} = 0.5$ 이다.

21. 잡은은 무작위적 변동이며 일반적인 원인이 알려져 있지 않다.

22. 민감도는 $\frac{TP}{TP+FN} = \frac{40}{40+60} = 0.4$ 이다.

23. 향상도는 $\frac{\frac{300}{1000}}{\frac{600}{1000} \times \frac{600}{1000}} = \frac{5}{6} = 0.833...$이므로 83%이다.

24. 연관분석은 상품의 구매, 서비스 등 일련의 거래 또는 사건들 간의 규칙을 발견하기 위해 적용하며, 카탈로그 배열, 교차판매 등의 마케팅에 활용된다.

25. 특이도= TN/(TN+FP)이므로 여기서 TN+FP은 N이다. 즉, 민감도= TN÷N 이다.

26. 혼합분포군집은 모형기반의 군집 방법이며, 데이터가 k개의 모수적 모형의 가중합으로 표현되는 모집단 모형으로부터 나왔다는 가정하에서 모수와 함께 가중치를 자료로부터 추정하는 방법을 사용한다.

27. 카이제곱 검정은 모수적 검정 방법이다.

28. 마할라노비스 거리는 통계적 개념이 포함된 거리이며 변수들의 산포를 고려하여 표준화한 거리이다. 두 벡터 사이의 거리를 산포를 의미하는 표본 공분산으로 나눠주어야 하며, 그룹에 대한 사전 지식 없이는 표본 공분산을 계산 할 수 없으므로 사용하기 곤란하다.

29. 시계열 자료는 시간의 흐름에 따라 관찰된 값을 의미한다.

30. 군집분석은 각 객체의 유사성을 측정하여 유사성이 높은 대상 집단을 분류하고, 군집에 속한 객체들의 유사성과 서로 다른 군집에 속한 객체간의 상이성을 규명하는 분석 방법이다.

31. 상자수염그림 안에 그어진 선은 중앙값을 의미한다.

32. 비표본오차는 표본오차를 제외한 모든 오차로서 조사 과정에서 발생하는 모든 부주의나 실수, 알 수 없는 원인 등 모든 오차를 의미하며 조사대상이 증가하면 오차가 커진다.

33. LOOCV는 전체 관측치(n) 중 단 하나의 관측값만을 Validation set으로 사용하고 나머지 n-1개 관측값은 Train set으로 사용하므로 k=n인 경우이다.

34. 제1주성분은 변동을 최대로 설명해주는 방향으로의 변수들의 선형결합식이다.

35. k-평균군집은 먼저 원하는 군집의 개수와 초기 값들을 정해 seed 중심으로 군집을 형성하고 각 데이터를 거리가 가까운 seed가 있는 군집으로 분류한 후 각 군집내의 자료들의 평균을 계산하며 모든 개체가 군집으로 할당될 때까지 과정을 반복한다.

36. 사분위수를 이용하여 (Q1 - 1.5*IQR , Q3 + 1.5*IQR) = (4-1.5*(12-4) , 12+1.5*(12-4)) = (-8 , 24)로 이상값을 판단하는 하한선과 상한선을 구할 수 있다.

37. 사분위수 범위는 3사분위수에서 1사분위수를 뺀 값으로 전체 자료의 중간에 있는 절반의 자료들이 지니는 값의 범위를 말하며, 중간에 50%의 데이터들이 흩어진 정도를 의미한다.

38. 주성분 분석은 여러 변수들의 변량을 주성분이라는 서로 상관성이 높은 변수들의 선형 결합으로 만들어 기존의 상관성이 높은 변수들을 요약, 축소하는 기법이다.

39. 많은 모형에서 공통적으로 사용될 수 있는 변수는 요약변수이다.

40. 정상성의 기준은 모든 시점에서 일정한 평균을 가지며, 분산도 일정, 마지막으로 공분산은 시차에만 의존한다.

【 단답형 】

단답형 01. 사물인터넷(IoT, Internet of Things)

단답형 02. SCM(Supply Chain Management)

단답형 03. 프레이밍 효과(Framing effect)

단답형 04. 활용

단답형 05. 분해 시계열은 시계열 데이터를 요인에 따라 분해하는 기법으로 추세, 순환, 계절, 불규칙요인이 있다.

단답형 06. AR모형은 p시점 전의 자료가 현재 자료에 영향을 주는 모형이다.

단답형 07. 너무 큰 나무모형은 자료를 과대적합하고 너무 작은 나무모형은 과소적합할 위험이 있다.

단답형 08. 인공신경망은 인간의 신경세포를 통한 학습방법에서 아이디어를 얻어 이를 디지털 네트워크 모형으로 구현한 것이다.

단답형 09. 이익도표는 분류모형의 성능의 평가하기 위한 척도로, 분류된 관측치에 대해 얼마나 예측이 잘 이루어졌는지를 나타내기 위해 임의로 나눈 각 등급별로 반응검출율, 반응률, 리프트 등의 정보를 산출하여 나타내는 도표이다.

단답형 10. 데이터 마트는 데이터 웨어하우스와 사용자 사이의 중간층에 위치한 것으로, 하나의 주제 또는 하나의 부서 중심의 데이터 웨어하우스라고 할 수 있다.

ADsP 33회 기출문제 답안

【 객관식 정답 】

01	①	11	④	21	②	31	②
02	③	12	①	22	④	32	②
03	③	13	④	23	①	33	②
04	④	14	①	24	③	34	③
05	①	15	①	25	①	35	②
06	②	16	③	26	④	36	①
07	①	17	①	27	④	37	①
08	①	18	②	28	③	38	③
09	②	19	①	29	④	39	③
10	①	20	④	30	④	40	④

【 주관식 정답 】

01	데이터 웨어하우스 (Data Warehouse)
02	데이터 사이언스 (Data Science)
03	통찰(Insight)
04	나선형 모델(Spiral Model)
05	특이도(Specificity)
06	3개
07	제 1종 오류
08	오즈(Odds)
09	소프트맥스(Softmax) 함수
10	기울기 소실(Gradient Vanishing)

01. 사물인터넷(IoT)은 인터넷을 기반으로 모든 사물을 연결하여 사물과 사물, 사물과 사람 간의 정보를 상호 소통하는 지능형 기술 및 서비스이다. 빅데이터 관점에서 사물인터넷은 사물에서 나오는 데이터를 활용해 더욱 지능화 된 기기 활용을 할 수 있도록 데이터를 수집하여야 되므로 모든 사물에서 데이터를 추출할 수 있어야 된다.

02. 책임원칙 훼손의 통제방안은 결과 기반 책임 원칙 고수이며 데이터 오용은 알고리즘 접근의 허용이다.

03. NoSQL(비관계형) 데이터베이스에는 HBase, MongoDB, Redis 등이 속한다. MySQL은 관계형 데이터베이스에 해당한다.

04. 데이터베이스의 일반적인 특징으로는 통합된 데이터, 저장된 데이터, 공용 데이터, 변화되는 데이터로 변화되는 데이터는 데이터베이스에 저장된 내용은 곧 데이터베이스의 현 시점에서의 상태를 나타낸다.

05. 소프트 스킬에는 통찰력 있는 분석, 설득력 있는 전달, 다분야간 협력 등이 있다. 이론적 지식은 하드 스킬에 포함이 된다.

06. 데이터의 크기는 작은 것부터 큰 것까지 순서대로 페타바이트(PB), 엑사바이트(EB), 제타바이트(ZB), 요타바이트(YB)이다.

07. 빅데이터 분석 활용의 효과로 서비스 산업의 확대되지만, 제조업의 경우에도 생산성 향상, 불량률 감소 등에 대한 성과를 거둘 수 있기에 제조업이 축소되는 것은 아니다.

08. 빅데이터 시대에는 과거에는 표본조사였으나 현재는 전수조사로 바뀌었다.

09. 분석과제의 적용 우선순위 기준을 '시급성'에 둔다면 Ⅲ - Ⅳ - Ⅱ 이다.

10. 분석과제의 우선순위 평가에서 시급성은 전략적 중요도와 목표가치를 평가하고 난이도는 데이터획득/저장/가공비용과 분석 적용 비용, 분석수준이 평가요소이다.

11. 하향식 접근법의 데이터 분석 기획 단계는 문제 탐색(Problem Discovery) → 문제 정의(Problem Definition) → 해결 방안탐색(Solution Search) → 타당성 검토(Feasibility study)이다.

12. 정확도(Accuracy)는 True를 True, False를 False라고 예측하는 지표이며, 정밀도(Precision)는 True라고 예측한 것 중에서 실제 True인 것의 비율이다.

13. 분석과제 우선순위 고려요소는 전략적 중요도, 비즈니스 성과/ROI, 실행 용이성이 있다.

14. 최상위 계층은 단계로 구성되고 마지막 계층은 스탭으로 구성된다.

15. 분석 방법론의 분석 기획 단계에서는 비즈니스 이해 및 범위 설정, 프로젝트 정의 및 계획수립, 프로젝트 위험계획 수립을 주요 업무로 한다. 필요데이터 정의는 데이터 준비 단계에서 수행하는 Task이다.

16. 상향식 접근법은 비지도 학습(Unsupervised Learning)방식을 수행한다.

17. 이산형 확률변수의 기댓값은 $E(X) = \sum xf(x)$ 이다.

18. 이동평균법은 시계열자료에서 계절변동과 불규칙변동을 제거하여 추세변동과 순환변동만 가진 시계열 자료로 변환하는 평활법이다.

19. 스피어만 상관계수는 서열척도의 자료를 대상으로 한다.

20. F-statistic 우측의 p-value를 보면 회귀식의 유의성을 확인할 수 있다. p-value<2.2e-16으로 매우 작으므로 회귀식이 유의하다고 할 수 있다.

21. 활성화 함수는 입력 신호의 총합을 출력 신호로 변환하는 함수로 종류로는 계단, 시그모이드, ReLU, Softmax 등이 있다.

22. 비표본오차는 표본오차를 제외한 모든 오차로서 조사 과정에서 발생하는 모든 부주의나 실수, 알 수 없는 원인 등 모든 오차를 의미하며 조사대상이 증가하면 오차가 커진다.

23. 2번은 다차원척도법, 3번은 군집분석, 4번은 연관규칙분석에 대한 설명이다.

24. F1의 정의는 $2 \times \frac{\mathrm{Re}call \times \mathrm{Pr}ecision}{\mathrm{Re}call + \mathrm{Pr}ecision}$ 이다. Recall과 Precision을 대입하면 $2 \times \frac{\frac{2}{5} \times \frac{2}{5}}{\frac{2}{5} + \frac{2}{5}} = 0.40$이다.

25. $\mathrm{Re}call = \frac{TP}{TP + FN} = \frac{30}{30 + 70} = \frac{3}{10}$ 이다.

26. k개의 초기 중심값은 임의로 선택이 가능하므로 한번 군집을 형성되어도 군집 내 객체들은 다른 군집으로 이동이 될 수 있다.

27. 연관규칙의 분석 결과로는 품목 간 구체적인 영향을 줄수 있는지 알 수 없다.

28. P(질병)=0.1, 진단 시 질병 가진 사람=P(양성)=0.2, 질병을 가지고 있는 사람 중 질병이라 진단 받은 사람=P(양성|질병)=0.9이며, 구해야할 값은 P(질병|양성)으로 베이즈 정리에 따라 P(양성∩질병)=0.09이다. 그러므로 P(질병|양성)=P(양성∩질병)/P(양성)=0.09/0.2=0.45이다.

29. A, B의 유클리드 거리는 $\sqrt{(185-180)^2+(70-75)^2}=\sqrt{50}$ 이다.

30. 주성분의 개수는 고윳값, 누적 기여율, Scree Plot를 통해 확인할 수 있다. 고유치 분해 가능여부는 주성분 개수를 정하는 방법이 아니다.

31. SOM은 역전파 알고리즘 등을 이용하는 인공신경망과 달리 단 하나의 전방 패스를 사용함으로써 속도가 매우 빠르다. 따라서 실시간 학습처리를 할 수 있는 모형이다.

32. 분리변수의 P차원 공간에 대한 현재 분할은 이전 분할에 영향을 받는다.

33. 부스팅은 예측력이 약한 모형들을 결합하여 강한 예측모형을 만드는 방법이다.

34. 향상도는 $\frac{\frac{2000}{5000}}{\frac{3000}{5000}\times\frac{2500}{5000}}=\frac{4}{3}$ 로 1보다 큰 값이 나와 연관성이 높다고 할 수 있다.

35. 새로운 변수가 추가될 때 기존 변수의 중요도가 약해질 수 있으므로 그러한 변수를 제거하는 방법이 단계별 선택법(Stepwise selection)이다.

36. 생성된 모델이 훈련 데이터에 최적화되어 있기 때문에 테스트 데이터의 작은 변화에 민감하게 반응한다.

37. 구간추정은 모수의 참값이 포함되어 있다고 추정되는 구간을 결정하는 것이지만, 실제 모집단의 모수가 신뢰구간에 꼭 포함되어 있는 것은 아니다.

38. 산점도와 head, summary값으로 나무 종류별로 둘레에 유의한 차이가 있는지 알 수 없다.

39. Loadings에서 Comp.2의 값을 보면 head2와 breath2는 음의 값을 가짐을 보아 양의 상관관계를 가진다고 할 수 없다.

40. 동일 모집단에서 동일한 방법과 개수로 표본을 다시 추출하더라도 동일한 신뢰구간이 나오지 않을 수 있다.

【단답형】

단답형 01. 데이터 웨어하우스(Data Warehouse)

단답형 02. 데이터 사이언스(Data Science)

단답형 03. 통찰(Insight)

단답형 04. 나선형 모델(Spiral Model)

단답형 05. 특이도(Specificity)

단답형 06. height가 60에서 직선을 긋게 되면 총 3개의 군집으로 분류된다.

단답형 07. 제 1종 오류는 귀무가설이 실제로 참이어서 채택해야 함에도 불구하고 표본의 오차 때문에 이를 채택하지 않는 오류를 말한다.

단답형 08. 오즈(Odds)는 성공할 확률이 실패할 확률의 몇배인지를 나타낸다.

단답형 09. 소프트맥스(Softmax) 함수는 세 개 이상으로 분류하는 다중 클래스 분류에서 사용되는 활성화 함수이다.

단답형 10. 기울기 소실(Gradient Vanishing) 문제는 신경망의 활성함수의 도함수 값이 계속 곱해지다 보면 가중치에 따른 결과값의 기울기가 0이 되어 버려서, Gradient Descent를 이용할 수 없게 되는 문제이다.

ADsP 34회 기출문제 답안

【 객관식 정답 】

01	④	11	④	21	③	31	②
02	②	12	②	22	④	32	②
03	③	13	④	23	③	33	③
04	②	14	③	24	③	34	③
05	④	15	①	25	④	35	③
06	②	16	②	26	①	36	④
07	①	17	②	27	③	37	④
08	③	18	①,④	28	③	38	①
09	③	19	④	29	③	39	④
10	④	20	②	30	④	40	③

【 주관식 정답 】

01	정보(Information)
02	사물인터넷 (IoT, Internet of Things)
03	문제 정의
04	데이터 거버넌스
05	랜덤 포레스트 (Random Forest)
06	부스팅(Boosting)
07	역전파 알고리즘
08	실루엣(Shilouette)
09	점 추정
10	지니 지수

01. 데이터가 커진다고 해서 분석에 많이 사용되는 것이 아니라 데이터에 따른 적정한 분석 방법이 경쟁우위를 가져다준다고 할 수 있다.

02. 데이터의 가치를 측정하기 어려운 이유는 다음과 같다.
- 데이터 활용 방식: 재사용, 재조합(mashup), 다목적용 개발
- 새로운 가치 창출
- 분석 기술 발전

03. 데이터베이스에 있는 데이터 중 분석이 불가한 데이터들도 있으며, 그런 데이터는 처리 등을 통해 분석에 활용할 수 있다.

04. 일차적인 분석을 통해서도 해당 부서나 업무 영역에서는 상당한 효과를 얻어낼 수 있다.

05. 데이터 오용은 베트남 전쟁 시 적군 사망자 수를 과장해 보고하는 것을 통해 알 수 있었다.

06. 데이터 사이언스는 통찰력 있는 분석에 초점을 두고 진행한다.

07. 데이터 마트는 데이터웨어하우스로부터 구축된 데이터 속에서 한 가지 주제 또는 한 부서 중심으로 구축된 소규모, 단일 주제의 웨어하우스로 예측 가능한 질의에 대해서 매우 빠르게 응답할 수 있도록 데이터를 제공하는 시스템이다.

08. 데이터가 많다고 무조건 더 많은 가치가 창출되지는 않는다.

09. 타당성 검토 단계에서는 효과적으로 평가하기 위해서 비즈니스 지식과 기술적 지식이 요구되기 때문에 비즈니스 분석가, 데이터 분석가, 시스템 엔지니어 등과의 협업이 수반되어야 한다.

10. Precision은 모델을 지속적으로 반복했을 때의 편차의 수준으로써 동일한 결과를 제시한다는 것을 의미한다.

11. 문제 탐색 단계에서 현재의 비즈니스 모델 및 유사·동종사례 탐색을 통해서 도출한 분석 기회들을 구체적인 과제로 만들기 전에 분석 유즈케이스로 표기하는 것이 필요하다.

12. 데이터 분석 준비 프레임워크에서 분석 업무 파악 영역에는 발생한 사실 분석 업무, 예측 분석 업무, 시뮬레이션 분석 업무, 최적화 분석 업무, 분석 업무 정기적 개선이 있다.

13. 시스템 구현 단계에서 정보보안영역과 코딩은 주요 고려사항이 아니다. 시스템 설계 및 구현, 테스트 및 운영이 주요 고려사항이다.

14. 분석과제 정의서를 통해 분석별로 필요한 소스 데이터, 분석방법, 데이터 입수 및 분석의 난이도, 분석 수행주기, 분석결과에 대한 검증 오너십, 상세 분석 과정 등을 정의한다. 분석 데이터 소스는 내·외부의 비구조적인 데이터와 소셜 미디어 및 오픈 데이터까지 범위를 확장하여 고려하고 분석방법 또한 상세하게 정의한다.

15. 분석 프로젝트 관리방안에서 시간관리는 프로젝트의 활동 일정을 수립하고 일정 통제의 진척상황을 관찰하는데 요구되는 프로세스이다.

16. 데이터 준비 단계에서는 분석용 데이터 셋 선택, 데이터 정제, 분석용 데이터 셋 편성, 데이터 통합, 데이터 포맷팅 수행업무가 있다.

17. 상자 그림으로는 이상치를 확인할 수 있다.

18. 1~4번 중 지지도가 25%이상인 규칙은 1번과 4번이다. 그 중 신뢰도가 50%이상인 규칙은 A→B, B→C 이다.

19. p, d, q에 따라서 각각 0 이면 IMA(d,q), ARMA(p,q), ARI(p,d)모형으로 부를 수 있다. 이 중 IMA(d,q)를 d번 차분하면 MA(q) 모형을 따른다.

20. ROC곡선의 좌표는 (1-특이도, 민감도)로 x축이 낮고 y축이 높을수록 분류정확도가 높다는 것을 의미하므로 이상적으로 완벽히 분류한 모형의 좌표는 (0,1)이다.

21. 적절한 세분화로 인한 품목 결정이 장점이지만 너무 세분화된 품목은 의미 없는 결과를 도출한다.

22. K 값이 작을수록 과대적합(Overfitting) 문제가 발생한다.

23. 시그모이드 함수를 단층신경망에서 활성화함수로 사용하면 로지스틱 회귀모형과 작동원리가 유사하다.

24. 군집의 분리에 대해 안정성도 중요 하지만 해당 군집에 대한 분리가 논리적으로 설명이 되는 부분이 더 중요하다고 할 수 있다.

25. 일반적으로 정상성을 만족하지 않을 때는 log, root를 취하여 정규분포를 취하도록 만든다.

26. 의사결정나무 모형은 지도학습 모형으로 하향식 의사결정에 가깝다고 생각할 수 있다.

27. 앙상블은 주어진 자료로부터 여러 개의 예측모형들을 만든 후 예측모형들을 조합하여 하나의 최종 예측 모형을 만드는 방법으로 Bagging, Boosting, Random Forest, Stacking 등이 있다.

28. 연관분석은 실시간 상품추천을 통한 교차판매 등에 활용할 수 있다.

29. 모형에서 종속변수와 독립변수 간의 상관계수가 유의한지는 상관관계 분석을 통해 확인한다.

30. 지수평활법은 시간의 흐름에 따라 최근 시계열에 더 많은 가중치를 부여하여 미래를 예측하는 방법이다.

31. 회귀분석 결과에서 분석이 잘 되었다면 잔차는 더 이상 독립변수와 상관관계를 가지지 않는다.

32. 2개의 주성분으로 자료를 축약할 때 전체 분산의 74.6%가 설명 가능하다.

33. Balance와 가장 상관관계가 높은 변수는 Limit와 Rating이다.

34. Recall은 $\frac{TP}{TP+FN}$ 이므로 $\frac{40}{40+60}=0.4$ 이다.

35. Apriori 알고리즘은 최소 지지도를 설정하고 개별 품목 중 최소 지지도가 넘는 모든 품목을 먼저 찾는다. 그리고 개별 품목만으로 최소 지지도가 넘는 2가지 품목을 찾고, 이것들을 결합해 3가지 품목집합을 찾으며 반복해 빈발품목집합을 찾는다.

36. m개의 주성분은 원래 변수들 중 서로 상관성이 높은 변수들의 선형결합으로 만들어진 것이다.

37. 피어슨 상관계수는 연속형 변수에 사용하며 정규성을 가정한다. 스피어만 상관계수는 순서형 변수에 사용하며 비모수적 방법이다.

38. 배깅(Bagging)은 주어진 자료에서 여러 개의 붓스트랩(bootstrap) 자료를 생성하고 각 붓스트랩 자료에 예측모형을 만든 후 결합하여 최종 예측모형을 만드는 방법이다.

39. 가설검정은 어떤 모수의 값 또는 확률분포에 대하여 가설을 세우고 이 가설이 맞다고 주 장해도 이상이 없는지를 표본 데이터의 통계적 확률에 의해 결정하는 것을 말한다.

40. 상관관계 분석으로는 상관관계를 파악할 수 있다. 인과관계는 회귀분석에서 확인할 수 있다.

【단답형】

단답형 01. 정보(Information)

단답형 02. 사물 인터넷(IoT, Internet of Thing)

단답형 03. 문제 정의

단답형 04. 데이터 거버넌스

단답형 05. 랜덤 포레스트(Random Forest)는 의사결정 나무 여러 개로 만들어진 모델이다.

단답형 06. 부스팅(Boosting)은 예측력이 약한 모형들을 결합하여 강한 예측모형을 만드는 방법으로 GBM, XgBoost, LightGBM 등이 있다.

단답형 07. 출력값에 대한 입력값의 기울기(미분값)을 출력층 layer에서부터 계산하여 거꾸로 전파시키는 것이다.

단답형 08. 실루엣(Shilouette) 계수는 군집 모형 평가 기준 중 군집의 밀집 정도를 계산하는 방법으로 군집 내의 거리와 군집간의 거리를 기준으로 군집 분할의 성과를 평가하는 것이다.

단답형 09. 점추정이란 추정하고자 하는 하나의 모수에 대하여 모집단에서 임의로 추출된 n개 표본의 확률변수로 하나의 통계량을 만들고 주어진 표본으로부터 그 값을 계산하여 하나의 수치를 제시하려고 하는 것이다.

단답형 10. 지니지수는 노드의 불순도를 나타내는 값으로 지니지수의 값이 클수록 이질적이며 순수도가 낮다.

ADsP 35회 기출문제 답안

【 객관식 정답 】

01	④	11	①	21	②	31	②
02	③	12	①	22	①	32	①
03	①	13	①	23	①	33	④
04	②	14	③	24	③	34	①
05	③	15	④	25	④	35	④
06	③	16	②	26	③	36	④
07	③	17	④	27	③	37	②
08	④	18	②	28	②	38	③
09	②	19	④	29	①	39	④
10	③	20	①	30	③	40	④

【 주관식 정답 】

01	데이터베이스(Database)
02	유전자 알고리즘 (Generic Algorithm)
03	하향식 접근 방식
04	집중 구조
05	다차원 척도법
06	후진 제거법 (Backward Elimination)
07	홀드아웃 방법
08	0.4
09	포아송 분포Poisson Distribution)
10	최단 연결법

01. SQL은 데이터베이스를 사용할 때 데이터베이스에 접근할 수 있는 데이터베이스의 하부 언어이다.

02. 데이터 사이언티스트의 필요 역량은 하드 스킬과 소프트 스킬이 있으며, 소프트 스킬 중 통찰력 있는 분석, 설득력 있는 전달, 다분야간 협력이 있다.

03. 사생활 침해와 관련하여 익명화(anonymization) 기술 발전이 필요하다.

04. 알고리즘으로 부당한 피해를 보는 사람을 방지하기 위해서 생겨난 직업이다.

05. 사물인터넷의 적용으로 사람의 개입이 최소화 되고 기기에서 수집하는 데이터가 실시간으로 수집될 것이다.

06. 정부는 이익을 목적으로 하지 않고 공익을 목적으로 한다. 개인 정보를 활용하여 이익을 목적으로 하는 것은 사기업이다.

07. 빅데이터 시대 위기 요인은 사생활 침해, 책임 원칙 훼손, 데이터 오용이 있다.

08. DBMS는 다수의 사용자들이 DB 내의 데이터를 접근할 수 있도록 해주는 소프트웨어이다.

09. 데이터 거버넌스 체계 중 데이터 관리 체계는 데이터 정합성 및 활용의 효율성을 위하여 표준데이터를 포함한 메타 데이터와 데이터 사전의 관리 원칙을 수립한다. 빅데이터의 경우 데이터 양의 급증으로 데이터의 생명 주기 관리방안을 수립하지 않으면 데이터 가용성 및 관리비용 증대 문제에 직면하게 될 수 있다.

10. 데이터 거버넌스의 구성 요소인 원칙(Principle), 조직(Organization), 프로세스(Process)는 유기적으로 조합하고 효과적으로 관리하여, 데이터를 비즈니스 목적에 부합하도록 하고 최적의 정보 서비스를 제공할 수 있도록 한다.

11. 가치는 비즈니스 효과이다.

12. 분석 준비도를 측정하기 위한 요소는 분석업부, 인력 및 조직, 분석기법, 분석 데이터, 분석 문화, IT 인프라가 있다.

13. 사분면 영역에서 난이도와 시급성을 모두 고려할 때 가장 우선적인 분석 과제 적용이 필요한 영역은 난이도 : 쉬움, 시급성 : 현재를 나타내는 3사분면이다.

14. 분산구조는 분석조직 인력들을 현업부서로 직접 배치하여 분석업무를 수행하며, 전사차원의 우선순위에 따라 수행한다. 분석결과에 따른 신속한 Action이 가능하며, 베스트 프랙티스 공유가 가능하다.

15. 분석을 통한 상향식 접근법(Bottom Up Approach)의 프로세스는 프로세스 분류 -> 프로세스 흐름 분석 -> 분석 요건 식별 -> 분석 요건 정의이다.

16. 분석 방법론 기획단계에서 프로젝트 위험 대응 계획을 수립할 때 예상되는 위험에 대해 회피(Avoid), 전이(Transfer), 완화(Mitigate), 수용(Accept)으로 구분하여 위험관리 계획서를 작성한다.

17. lasso 회귀모형에서는 사용하는 규제 방식을 L1 규제(Penalty)라고 한다.

18. 맨하탄 거리를 구하면 |180-175|+|65-70|=10이다.

19. 혼합분포 군집모형은 군집의 크기가 너무 작으면 추정의 정도가 떨어지거나 어려울 수 있다.

20. 분류 모델링은 데이터가 어느 종류에 속하는지를 판별하는 모델이다.

21. 1*0.5 + 2*0.3 + 3*0.2 + 4*0 = 1.7이다.

22. 분석 결과를 보면 EM 알고리즘을 통해 모수를 추정하는 과정에서 반복횟수 2회 만에 로그-가능도 함수가 최대가 됨을 알 수 있다.

23. X의 기댓값은 $\frac{1}{3}\times 1+\frac{1}{6}\times 2+\frac{1}{2}\times 3=\frac{13}{6}$ 이다.

24. 정밀도는 $\frac{TP}{TP+FP}=\frac{30}{30+60}=\frac{3}{9}$ 이다.

25. 학생여부에 따른 Balance가 커지는거에 대한 유의적인 차이는 회귀분석 결과에서 확인하기 어렵다.

26. 단순회귀분석에서 결정계수는 상관계수의 제곱과 같다.

27. 연속형 목표변수의 분리 기준은 분산분석에서의 F-통계량, 분산 감소량이 있다.

28. SOM에서 입력층의 뉴런은 경쟁층에 있는 뉴런들과 완전연결(Fully Connected)되어 있다.

29. 모형의 적합성을 파악하기 위해서는 F-통계량으로 파악한다.

30. 산점도를 확인했을 때, hosepower이 커질수록 mpg가 작아지는 걸로 보아 상관계수가 음의 상관을 가질 것으로 말할 수 있어 관계를 파악하기에 유용하게 사용할 수 있다.

31. 지지도는 전체 거래 중 항목 A와 항목 B를 동시에 포함하는 거래의 비율로 정의한다.

32. 데이터 마이닝 추진단계는 순서대로 목적 설정, 데이터 준비, 가공, 기법적용, 검증이다.

33. 시그모이드 함수는 S자형 곡선 또는 시그모이드 곡선을 갖는 수학 함수로 로지스틱 회귀분석에서 사용된다.

34. 로지스틱 회귀모형은 반응변수가 범주형인 경우에 적용되는 모형이다.

35. DBSCAN은 데이터 형태가 오목하거나 다른 임의의 모양일 때 k-means보다 성능이 더 좋다.

36. 데이터 분할 방법으로는 홀드아웃방법, 교차검증, 부스트랩 등이 있다.

37. 특이도는 $\frac{TN}{TN+FP}=\frac{100}{300+100}=0.25$ 이다.

38. 계층적 군집모형은 군집의 개수가 정해지지 않았을 때 사용하며, 군집의 개수를 모를 때 사용하기 때문에 몇 개의 군집으로 나누어야 하는지 결정하기 위해 사용하기도 한다.

39. 연관규칙분석은 상품의 구매, 서비스 등 일련의 거래 또는 사건들 간의 규칙을 발견하기 위해 적용하며, 조건과 반응의 형태(if-then)로 이루어져 있다.

40. 정상성을 만족하는지 판단하기 위해서는 평균과 분산이 일정한지 등에 대해 판단해야 한다.

【 단답형 】

단답형 01. 데이터베이스(Database)

단답형 02. 유전자 알고리즘(Generic Algorithm)

단답형 03. 하향식 접근 방식

단답형 04. 집중 구조

단답형 05. 다차원 척도법은 객체간 접근성을 시각화하는 통계기법으로 군집분석과 같이 개체들을 대상으로 변수들을 측정한 후에 개체들 사이의 유사성/비유사성을 측정하여 개체들을 2차원 공간상에 점으로 표현하는 분석방법이다.

단답형 06. 후진 제거법은 독립변수 후보 모두를 포함한 모형에서 출발해 가장 적은 영향을 주는 변수부터 하나씩 제거하면서 더 이상 제거할 변수가 없을 때의 모형을 선택한다.

단답형 07. 홀드아웃 방법은 주어진 데이터를 랜덤하게 두 분류로 분리하는 방법으로 보통 70:30 또는 80:20으로 분리한다.

단답형 08. 사건 A와 B가 독립이면 $P(A \cap B)=P(A)P(B)$이므로 $P(B|A)=P(A \cap B)/P(A)=P(A)P(B)/P(A)=P(B)$이다. 즉, 0.4가 된다.

단답형 09. 포아송 분포(Poisson Distribution)는 단위 시간 안에 어떤 사건이 몇 번 발생할 것인지를 표현하는 이산 확률분포이다.

단답형 10. 최단연결법

초판 1쇄 발행 2015년 03월 20일
9판 1쇄 발행 2023년 01월 20일
3쇄 발행 2023년 07월 07일

발행인 윤종식
저자 윤종식
편집디자인 윤미선 | 인쇄제본 정민피앤피
펴낸곳 (주)데이터에듀
출판등록번호 제2020-000003호
주소 부산시 해운대구 센텀1로 28, 102동 3003호
대표전화 051-523-4566 | 팩스 0303-0955-4566
이메일 books@dataedu.co.kr | 홈페이지 www.dataedu.kr

ISBN 979-11-978895-8-5
가격 31,000원