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.

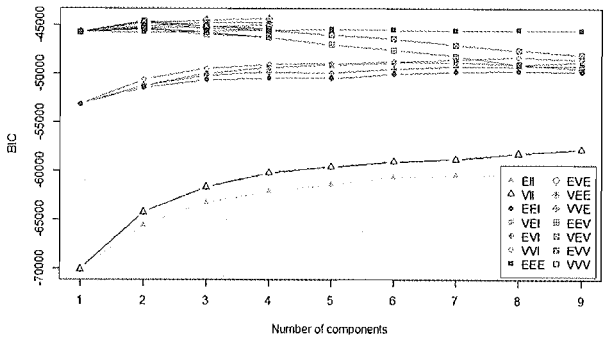
아래

(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.

( )