모의고사 1회 ▶ 472p

01 ①	02 ②	03 ③	04 ②	05 ①
06 ④	07 ③	08 ④	09 ①	10 ①
11 ③	12 ④	13 ①	14 ③	15 ③
16 ④	17 ①	18 ①	19 ③	20 ④
21 ②	22 ③	23 ③	24 ②	25 ②
26 ③	27 ①	28 ③	29 ②	30 ②
31 ③	32 ③	33 ②	34 ④	35 ③
36 ④	37 ④	38 ④	39 ③	40 ②
41 ①	42 ②	43 ①	44 ④	45 ②
46 ④	47 ①	48 ②	49 ①	50 ②
51 ④	52 ②	53 ②	54 ②	55 ④
56 ①	57 ①	58 ②	59 ③	60 ③
61 ④	62 ①	63 ④	64 ②	65 ④
66 ②	67 ④	68 ④	69 ③	70 ②
71 ④	72 ①	73 ③	74 ②	75 ③
76 ③	77 ②	78 ①	79 ①	80 ①

1 과목 | 빅데이터 분석 기획

01 ①

ETL 프로세스는 데이터의 추출(Extract), 변환(Transform), 적재(Load)의 약어로, 다양한 원천 데이터를 취합해 추출하고 공통된 형식으로 변환하여 적재하는 과정이다.

02 ②

최근 해석 가능한 인공지능 기술에 대한 연구가 진행되고 있지만 딥러닝의 경우 이론적 근거가 부족하고 해석이 어렵다는 것이 다수의 견해이다.

03 ③

빅데이터 분석 방법론은 분석 기획, 데이터 준비, 데이터 분석, 시스템 구현, 평가 및 전개 5단계로 구성되어 있다.

04 ②

지도학습은 학습 데이터로부터 하나의 함수를 유추해내기 위한 방법으로 분류모형이나 회귀모형에 적합하다.

05 ①

민감한 정보의 분포를 낮추어 추론 가능성을 더욱 낮추는 기법은 t-근접성에 대한 설명이다.
m-유일성은 비식별 데이터셋의 속성을 조합했을 때 동일한 값이 m개 이상이 되도록 한다.

06 ④

개인정보 비식별화 방법으로 가명 처리, 총계 처리, 데이터 삭제, 데이터 범주화, 데이터 마스킹 기법이 있다.

07 ③

탐색적 데이터 분석
분석용 데이터셋에 대한 정합성 검토, 데이터 요약, 데이터 특성을 파악하고 모델링에 필요한 데이터를 편성한다. 다양한 관점으로 평균, 분산 등 기초 통계량을 산출하여 데이터의 분포와 변수간의 관계 등 데이터 자체의 특성과 통계적 특성을 파악한다. 또한 시각화를 탐색적 데이터 분석을 위한 도구로 활용하여 데이터의 가독성을 명확히 하고 데이터의 형상 및 분포 등 데이터 특성을 파악한다.

08 ④

Insight : 분석 대상을 모르는 경우

오답 피하기
- Discovery : 둘 다 모르는 경우
- Solution : 분석 방법을 모르는 경우
- Optimization : 분석 대상과 분석 방법을 모두 알고 있는 경우

09 ①

입사 지원자에 대하여 해당 기관에서 당사자의 범죄 이력을 조회하기 위해 정보주체의 사전 동의가 필요하다.

10 ①

정형 데이터 품질 진단 방법으로 메타데이터 수집 및 분석, 칼럼 속성 분석, 누락 값 분석, 값의 허용 범위 분석, 허용 값 목록 분석, 문자열 패턴 분석, 날짜 유형 분석, 기타 특수 도메인 분석, 유일 값 분석, 구조 분석 등이 있다.

11 ③

EDA는 모형을 선정하기 위한 과정이 아니라 모형에 적합한 데이터를 마련(가공)하는 과정 즉, 모델링에 필요한 데이터를 편성한다는 것에 주의한다.

12 ④

모형화는 복잡한 문제를 논리적이면서도 단순화하는 과정으로 많은 변수가 포함된 현실 문제를 특징적 변수로 정의한다. 문제를 변수들 간의 관계로 정의한다.

13 ①

진단(Diagnostic) 분석 : 원인은 무엇인가?

오답 피하기
- **기술(Descriptive) 분석** : 무엇이 일어났는가?
- **예측(Predictive) 분석** : 앞으로 어떻게 될 것인가?
- **처방(Prescriptive) 분석** : 어떻게 대처해야 하는가?

14 ③

이상치
변수의 분포에서 비정상적으로 분포를 벗어난 값이다. 각 변수의 분포에서 비정상적으로 극단값을 갖는 경우나 자료에 타당도가 없는 경우, 비현실적인 변수값들이 이에 해당한다. 이상치가 포함된 자료의 분석결과는 추정치가 이상점의 방향으로 편파성을 일으키는 문제, 타당도가 결여된 자료를 분석에 포함하여 발생하는 추정치의 타당도 문제가 발생한다.

15 ③

DBMS는 DBtoDB 방식으로 DBMS간 동기화나 데이터에 대한 전송을 할 수 있다.

16 ④

데이터 분석 성숙도 모델은 성숙도 수준에 따라 도입단계, 활용단계, 확산단계, 최적화단계로 구분한다.

17 ①

개인정보 수집 시 정보주체에게 수집 목적 및 출처, 이용 기간, 정보 활용 거부권 행사 방법 등을 투명하게 알려야 한다.

18 ①

상향식 접근 방식은 다량의 데이터 분석을 통해 왜(why) 그러한 일이 발생하는지 역으로 추적하면서 문제를 도출하거나 재정의할 수 있는 방식으로 데이터를 활용하여 생각지도 못했던 인사이트 도출 및 시행착오를 통한 개선이 가능하다.

19 ③

정확성은 실세계에 존재하는 객체의 표현 값이 정확히 반영되어야 한다는 것으로, 세부 품질 기준으로는 선후 관계 정확성, 계산/집계 정확성, 최신성, 업무규칙 정확성이 있다.

20 ④

데이터 거버넌스는 전사 차원의 모든 데이터에 대하여 정책 및 지침, 표준화, 운영조직 및 책임 등의 표준화된 관리 체계를 수립하고 운영하기 위한 프레임워크와 저장소를 구축하는 것이다.

2과목 | 빅데이터 탐색

21 ②

박스플롯은 수치적 자료를 표현하는 그래프이다. 이 그래프는 가공하지 않은 자료 그대로를 이용하여 그린 것이 아니라, 자료로부터 얻어 낸 통계량인 5가지 요약 수치(다섯 숫자 요약, Five-number Summary)를 가지고 그린다.

- **5가지 요약 수치** : 최솟값, 제 1사분위(Q1), 제2사분위(Q2), 제3사분위(Q3), 최댓값

최댓값과 최솟값을 통해 이상값이 존재하는지 파악할 수 있다. 분산은 데이터의 퍼짐정도를 나타내는 것으로 박스플롯을 통해 파악하기 힘들다.

22 ③

단계적 선택법(Stepwise Selection)

- 전진 선택법과 후진 선택법(래퍼기법)의 보완방법이다.
- 전진 선택법을 통해 가장 유의한 변수를 모형에 포함 후 나머지 변수들에 대해 후진 선택법을 적용하여 새롭게 유의하지 않은 변수들을 제거한다.
- 제거된 변수는 다시 모형에 포함하지 않으며 유의한 설명변수가 존재하지 않을 때까지 과정을 반복한다.

기본적으로 단계적 선택법은 전진 선택법과 후진 선택법의 결합으로 각각의 기본 룰을 지킴에 유의해야 한다.

23 ③

특정상황에만 유의미하지 않게 대표성을 나타나게 할 필요가 있다.

24 ②

오버샘플링 : 소수클래스의 복사본을 만들어, 대표클래스의 수만큼 데이터를 만들어 주는 것이다. 똑같은 데이터를 그대로 복사하는 것이기 때문에 새로운 데이터는 기존 데이터와 같은 성질을 갖게 된다.

오답 피하기

- **언더샘플링** : 대표클래스(Majority Class)의 일부만을 선택하고, 소수클래스(Minority Class)는 최대한 많은 데이터를 사용하는 방법이다. 이때 언더샘플링된 대표클래스 데이터가 원본 데이터와 비교해 대표성이 있어야 한다.
- 음수 미포함 행렬분해와 특이값분해는 데이터 축소에 관련한 기법이다.

25 ②

P(A)=0.5는 A공장 생산품일 확률, P(B)=0.3은 B공장 생산품일 확률, P(C)=0.2는 C공장 생산품일 확률이고 P(F)는 불량품이 나올 확률이라고 하자.

그럼 문제의 조건에서

P(F|A) : A공장 생산품 중 불량품이 나올 확률이고 값은 0.01
P(F|B) : B공장 생산품 중 불량품이 나올 확률이고 값은 0.02
P(F|C) : C공장 생산품 중 불량품이 나올 확률이고 값은 0.03

우리가 구하고자 하는 확률은 불량품인데 A공장 제품일 확률이므로 P(A|F)로 정의 될 수 있고, 베이지안 정리에 의해

$$P(A|F)=\frac{P(A\cap F)}{P(F)}=\frac{P(F|A)P(A)}{P(F|A)P(A)+P(F|B)P(B)+P(F|C)P(C)}$$

이 된다. 정리하면,

$$\frac{0.01\times 0.5}{0.01\times 0.5+0.02\times 0.3+0.03\times 0.2}=\frac{0.005}{0.017}=\frac{5}{17}$$

26 ③

한 학생이 80점에서 85점 사이의 점수를 받을 확률은

$$Z_1=\frac{X-\mu}{\sigma}=\frac{80-80}{10}=0\ ,\quad Z_2=\frac{X-\mu}{\sigma}=\frac{85-80}{10}=0.5$$

그러므로

$$P(80\le X\le 85)=P\left(0\le\frac{X-\mu}{\sigma}\le 0.5\right)=P(0\le Z\le 0.5)$$
$$=P(Z\le 0.5)-P(Z\le 0.0)=0.6915-0.5=0.1915$$

27 ①

최대우도에 의한 모수추정의 방법을 이용하여 (로그우도추정)

$$f(t;\theta)=\theta e^{-\theta t}\ (t\ge 0)$$
$$L(x_1,x_2,\dots,x_n;\theta)=\prod_{i=1}^{n}f(x_i;\theta)=\prod_{i=1}^{n}\theta e^{-\theta x_i}$$

여기서 x=3, 1, 2, 3, 3이므로

$$L(x_1,x_2,\dots,x_n;\theta)=\theta e^{-3\theta}\theta e^{-\theta}\theta e^{-2\theta}\theta e^{-3\theta}\theta e^{-3\theta}$$

정리하면

$$L(x_1,x_2,\dots,x_n;\theta)=\theta^5 e^{-12\theta}$$

양변에 로그를 취하면

$$ln[L(x_1,x_2,\dots,x_n;\theta)]=\ln(\theta^5 e^{-12\theta})=\ln\theta^5-12\theta$$

여기서 $(\ln(e^{-12\theta})=-12\theta\ln e=-12\theta)$ 식을 미분하면

$$\frac{d\,ln[L(x_1,x_2,\dots,x_n;\theta)]}{d\theta}=\frac{5\theta^4}{\theta^5}-12$$

상기식을 0 되게 하는 값, 즉 미분값이 0이 되는 값은

$$\frac{5\theta^4}{\theta^5}-12=0\ \rightarrow\ \frac{5}{\theta}=12$$

정리하면 θ=5/12

28 ③

음의 상관관계를 나타내주는 그래프 개형으로 피어슨 상관계수는 $-1 < \rho < 0$ 사이 값으로 나타내어질 수 있다.

29 ②

스피어만 상관계수에 대한 설명이다

오답 피하기

크론비흐 알파(Cronbach's alpha)계수인 신뢰도(reliability) 계수 α는 검사의 내적 일관성 신뢰도를 나타내는 값으로서 한 검사 내에서 변수들 간의 평균상관관계에 근거해 검사문항들이 동질적인 요소로 구성되어 있는지를 분석하는 것이다. 동일한 개념이라면 서로 다른 독립된 측정 방법으로 측정했을 때 결과가 비슷하게 나타날 것이라는 가정을 바탕으로 한다.

30 ②

스타차트(Star Chart)

하나의 공간에 각각의 변수를 표현하는 몇 개의 축을 그리고, 축에 표시된 해당 변수의 값들을 연결하여 별 모양(또는 거미줄 모양)으로 표현하는 그래프이다.

- 하나의 변수마다 축이 시작되는 시작점(중점)은 최소값을, 가장 먼 끝점은 최대값을 나타낸다.
- 값이 적은 축에 해당하는 부분이 다른 부분에 비해 들어가 보이기 때문에, 여러 변수 값들을 비교하여 부족하거나 넘치는 변수를 표현하는데 적합하다.
- 연결된 선의 모양이나 색을 다르게 하는 경우 여러 속성을 한번에 표현할 수 있다.

오답 피하기

버블차트(Bubble Chart)

x, y값의 위치를 표시하는 산점도에 점의 위치에 해당하는 제3의 변수값을 원의 크기로 표현한 그래프로 한 번에 3개의 변수를 비교해볼 수 있다.

- 제3의 값을 표시하는 원(버블)은 면적으로 표현되어야 하며, 반지름이나 지름으로 표현되면 실제 값보다 너무 크게 원이 그려질 수 있어서 주의해야 한다.
- 도시별 인구밀집도, 도시별 우유 판매량 등 국가나 지역에 따른 값의 분포를 표현하는데 매우 유리하다.

히트맵(Heat Map)

데이터 분포와 관계에 대한 정보를 색(Heat)으로 표현한 그래프이다. 데이터를 식별하기 위해 각각의 칸마다 색으로 수치의 정도를 표현한다.

산점도(스캐터플롯: Scatter Plot)

두 변수의 값을 2차원(또는 3차원) 좌표계를 활용하여 점으로 표시한 것으로 점들의 집합이 모여서 두 변수 사이의 관계를 표현한다.

- 점들의 분포에 따라 집중도(강도, 영향력)를 확인할 수 있으며, 관계 추정을 위해 추세선을 추가할 수 있다.
- 점의 크기, 형태, 색상 등을 다르게 하여 하나의 산점도에 다양한 데이터의 특징을 표현할 수 있다.

31 ③

모집단의 분산을 모르고 표본의 크기가 작은 경우이므로 t-분포에 의한 신뢰구간를 구하여 보면

$$\bar{X} - t_{\frac{\alpha}{2},n-1} \cdot \frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\frac{\alpha}{2},n-1} \cdot \frac{S}{\sqrt{n}}$$

표본 평균 $\bar{X}$=170, 분산이 S^2=25 이고 자유도는 25-1= 24

자유도가 24이고 $t_{\frac{0.05}{2},24}$=+2.064 이므로

(신뢰도 95% 이므로 1-0.95=0.05)

$$170 - 2.064 \cdot \frac{5}{\sqrt{25}} \le \mu \le 170 + 2.064 \cdot \frac{5}{\sqrt{25}}$$

정리하면

$$167.936 \le \mu \le 172.064$$

32 ③

기댓값을 나타내는 다음의 두 추정량을 추정량의 선택기준인 불편성과 효율성측면에서 비교하여 보자.

$$E(\hat{\theta}_1) = \frac{1}{4}E(X_1) + \frac{1}{4}E(X_2) + \frac{1}{4}E(X_3) + \frac{1}{4}E(X_4) = \frac{1}{4}4\mu = \mu$$

$$E(\hat{\theta}_2) = \frac{1}{4}E(X_1) + \frac{1}{2}E(X_2) + \frac{1}{4}E(X_3) = \frac{1}{4}\mu + \frac{1}{2}\mu + \frac{1}{4}\mu = \mu$$

이므로 둘다 불편 추정량이다. 그러나 분산을 비교하여 보면

$$Var(\hat{\theta}_1) = \frac{1}{16}Var(X_1) + \frac{1}{16}Var(X_2) + \frac{1}{16}Var(X_3) + \frac{1}{16}Var(X_4) = \frac{4}{16}\sigma^2 = \frac{1}{4}\sigma^2$$

$$Var(\hat{\theta}_2) = \frac{1}{16}Var(X_1) + \frac{1}{4}Var(X_2) + \frac{1}{16}Var(X_3) = \frac{6}{16}\sigma^2 = \frac{3}{8}\sigma^2$$

$\therefore Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$이 된다. 즉, θ_1이 θ_2보다 더 효율적이라고 말할 수 있다.

33 ②

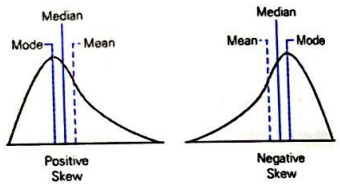
- **제1종 오류(Type I Error)** : 귀무가설이 참일 때 귀무가설을 기각하도록 결정하는 오류
- **제2종 오류(Type II Error)** : 귀무가설이 거짓인데 귀무가설을 채택할 오류

34 ④

차원의 저주(Curse of Dimensionality)

- 데이터분석 및 알고리즘을 통한 학습을 위해 차원이 증가하면서 학습데이터의 수가 차원의 수보다 적어져 성능이 저하되는 현상이다.
- 해결을 위해서 차원을 줄이거나 데이터의 수를 늘리는 방법을 이용해야 한다.

35 ③



- 프로스포츠 구단의 경우는 Positive Skew의 형태로 중심성의 경향으로 봤을 때는 중앙값이 중심성 경향의 통계량으로 가장 적절하다.
- 기본적으로 Positive Skew 경우는 최빈값(mode) < 중앙값(median) < 평균(Average) 순이다.
- 분산의 중심화경향이 아닌 산포성 퍼짐정도에 대한 통계량이다

36 ④

- **최적 배분법** : 추정량의 분산을 최소화 시키거나 주어진 분산의 범위 하에서 비용을 최소화 시키는 방법

오답 피하기

- **비례 배분법** : 각 층 내의 추출단위 수에 비례하여 표본크기를 배분하는 방법
- **네이만 배분법** : 각 층의 크기와 층별 변동의 정도를 동시에 고려한 표본배정 방법

37 ④

어떤 데이터에서 각 클래스(주로 범주형 반응 변수)가 갖고 있는 데이터의 양에 차이가 큰 경우, 클래스 불균형이 있다고 말한다. 데이터 클래스 비율이 너무 차이가 나면(Highly-imbalanced Data) 단순히 우세한 클래스를 택하는 모형의 정확도가 높아지므로 모형의 성능판별이 어려워진다.

불균형 데이터의 처리

- 가중치 균형방법(Weighted Balancing)
- 언더샘플링(Undersampling)
- 오버샘플링(Oversampling)

38 ④

표본의 크기가 커질수록 표준오차 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$는 점점 줄어든다.

39 ③

지수분포 정규분포 F –분포는 연속확률분포이나 이항분포는 이산확률분포이다.

40 ②

모집단이 정규분포여도 모분산 값을 아는지 여부에 따라 달라지는데 현문제는 모분산을 모르는 상태이므로 표본의 크기에 따라서 표본이 따르는 분포는 달라진다. 30을 기준으로 30 이상인 경우는 정규분포, 30 미만의 경우는 t–분포를 따르며 표준오차의 경우는 표본의 크기가 커질수록 줄어든다.

3 과목 | 빅데이터 모델링

41 ①

- **후진소거법** : 후진 선택법이라고도 하며 전체모델에서 시작, 모든 독립변수 중 종속변수와 단순상관계수의 절댓값이 가장 작은 변수를 분석모형에서 제외시킨다.

오답 피하기

- **전진선택법** : 영 모형에서 시작, 모든 독립변수 중 종속변수와 단순상관계수의 절댓값이 가장 큰 변수를 분석모형에 포함시키는 것을 말한다.
- **차원축소** : 어떤 목적에 따라서 변수(데이터의 종류)의 양을 줄이는 것이다.
- **요인분석** : 다수의 변수들 간의 관계(상관관계)를 분석하여 공통차원을 축약하는 통계분석 과정이다.

42 ②

오차역전파는 오차를 출력층에서 입력층으로 전달하고 연쇄법칙을 통해 가중치와 편향을 업데이트한다.

43 ①

특징맵의 출력 크기는 너비와 폭이 같은 정방형으로 (입력 높이(또는 너비) + 2 × 패딩값–필터 높이(또는 너비))/스트라이드값 + 1 로 계산한다. 따라서 (5 + 2 × 0 – 3) / 1 + 1 = 3으로 (3, 3)이 된다.

44 ④

다중 공선성 진단은 3개 이상의 독립변수 간 상관관계로 인한 문제가 없어야 한다.

45 ②

SVM은 선형 또는 비선형 분류로 이진분류만 가능하며 예측 정확도가 높은 편이나 데이터가 많을 시 모델 학습 시간이 오래 소요된다.

46 ④

다차원 척도법 : 객체간 근접성을 시각화한 통계기법으로 객체들 간 유사성/비유사성을 측정하여 2차원/3차원공간상에 점으로 표현한다.

47 ①

선형 회귀모델에 L1 규제를 추가한 것을 Lasso(라쏘)라고 한다.

오답 피하기

L2 규제를 추가한 것은 Ridge(리지)이며 elasticNet은 Ridge의 L2와 Lasso의 L1 정규화혼합 모델이다.

48 ②

SVM의 주요 요소들로 벡터, 결정영역, 초평면, 서포트 벡터, 마진 등이 있다.

49 ①

- 자료의 형태에 따른 범주형 자료 분석 방법

독립변수	종속변수	분석방법	예제
범주형	범주형	빈도분석, 카이제곱 검정 로그선형모형	지역별 선호정당 (지역별정당선호도)
연속형	범주형	로지스틱 회귀분석	소득에 따른 결혼의 선호도
범주형	연속형	T 검정(2그룹) 분산분석(2그룹 이상)	지역별 가계수입의 차이
연속형	연속형	상관분석, 회귀분석	

50 ②

		실제값	
		일반인	암환자
예측값	일반인	60 TN	0 FN
	암환자	10 FP	30 TP

정확도(Accuracy) = (TP+TN)/(TP+FP+TN+FN) =0.9
정밀도(Precision) = TP/(TP+FP) =0.75
재현율(Recall) = TP/(TP+FN) =1

- 데이터셋의 label 값이 불균형적으로 적은 경우 정확도만으로 예측 모델 성능을 평가하는 데는 한계가 있다.

51 ④

Adaboost 알고리즘은 부스팅 기법에 해당된다.

52 ②

P(A)는 A 사건 확률, P(B)는 B 사건확률, P(C)는 C 사건 확률, P(x)는 어떤 새로운 사건에 대한 확률
문제의 조건에서
P(x|A)= A 사건 중 x 나올 확률
P(x|B)= B 사건 중 x 나올 확률
P(x|C)= C 사건 중 x 나올 확률
B 사건 하에서 나올 조건부확률 P(B|x)로 정의될 수 있고, 베이지안 정리에 의해

$$P(B|x) = \frac{P(B \cap x)}{P(x)} = \frac{P(x|B)P(B)}{P(x|A)P(A) + P(x|B)P(B) + P(x|C)P(C)}$$

53 ②

Holdout(홀드아웃) 교차검증은 훈련데이터, 검증데이터, 테스트데이터를 일정 비율로 지정한 뒤 먼저 훈련데이터로 학습하되 훈련데이터 내에서 일정부분 검증데이터를 두어 검증한다.

54 ②

SNS 기반 선호 브랜드 그룹 분석은 비지도학습 중 하나인 군집 분석에 해당된다.

55 ④

색상비율에 따라 특정 감정 그룹 레이블(이름)으로 지정할 수 있으므로 분류 분석에 해당된다.

56 ①

오답 피하기

- **다층판별분석** : 종속변수가 남/여와 같이 두 개의 범주로 나누어져 있거나, 상/중/하와 같이 두 개 이상의 범주로 나누어져 있을 경우, 즉 종속변수가 비계량적 변수일 경우 이용된다.
- **요인분석** : 많은 수의 변수들 간 상호관련성을 분석하고, 이들 변수들을 어떤 공통 요인들로 설명하고자 할 때 이용된다.
- **분산분석** : 독립변수가 범주형(두 개 이상 집단)이고 종속변수가 연속형인 경우 이용된다.

57 ①

오답 피하기

자기회귀모형 : 시계열이 시차값 사이에 선형관계를 보이는 것을 자기상관이라 하며, 이러한 자기 상관성을 기반으로 과거의 패턴이 지속된다면 시계열 데이터 관측치 Xt는 과거 관측치 X(t−1), X(t−2),…, X(t−p)에 의해 예측할 수 있다.
정상성 : 시계열 데이터가 평균과 분산이 일정한 경우를 지칭한다.
백색 잡음 : 화이트 노이즈 등으로 불리며 무작위의 패턴을 보여주기 때문에 랜덤 노이즈라고도 한다.
이동평균법 : 시계열 자료를 대상으로 일정기간을 이동하면서 평균을 계산하여 추세를 파악하는 방법이다.

58 ②

비정형 데이터란 고정된 필드에 저장되지 않는 데이터로 텍스트, 이미지, 동영상, 음성, GPS 데이터 등이 있다.

59 ③

데이터 수가 많아지면 일반 의사결정나무에 비해 정확도는 높아지나 수행속도가 떨어진다.

60 ③

k-폴드 교차검증은 반복횟수 증가에 따른 모델 훈련과 평가/검증 시간이 오래 걸릴 수 있다.

4과목 | 빅데이터 결과 해석

61 ④

군집추출은 모집단을 여러 개의 군집으로 나누고, 특정 군집의 일부 또는 전체에 대한 분석을 시행한다. 표본크기가 같은 경우 단순 임의 추출에 비해 표본 오차가 증대할 가능성이 있다.

62 ①

매개변수는 데이터로부터 결정되는 학습의 대상으로 알고리즘을 통해 자동으로 학습하게 되며 가중치, 편향 등이 있다.

63 ④

막대그래프는 특정 변수의 시간에 따른 값의 변화를 보여주는 데 적합하며, 파이차트와 도넛차트는 특정변수값의 비율을 보여주는데 사용된다. 막대그래프와 도넛차트는 여러 변수를 표현할 수 있지만, 변수 사이의 관계를 표현하는 데 적합하지는 않다. 스캐터 플롯(산점도)은 2개 이상의 변수에 대한 상호 관계성을 표현하는데 적합하다.

64 ②

오답 피하기

막대그래프는 특정 변수의 시간에 따른 값의 변화를 보여주는 데 적합하며, 시간에 따른 변화를 표현하는 다른 도구로 꺾은선 그래프가 있다. 플로팅 차트는 X-Y축으로 값을 보여주며, 이때 x축을 시간축, y축을 값축으로 설정하는 경우 시간에 따른 값의 변화를 보여줄 수 있다.

65 ④

불균형 데이터 처리기법은 대표적으로 언더샘플링, 오버샘플링, 데이터 증강기법 등이 있다.

66 ②

ROC 곡선은 Y축 민감도(Sensitivity)와 X축 1−특이도(Specificity)로 그려지는 곡선이며 [0, 1] 범위로 하단 면적을 AUC(Area Under Curve)라고 한다. 이진 분류기의 성능을 평가하는 지표로 사용된다.

67 ④

Adjusted R^2, MAPE, RMSE는 회귀모델 평가지표에 해당된다.

68 ④

하이퍼파라미터는 최적의 딥러닝 모델을 구현하기 위해 사용자가 직접 설정하는 변수로 학습률, 배치크기, 은닉층의 뉴런개수, 훈련 반복 횟수 등이 있다.

69 ③

K-평균 군집분석은 군집 중심점(centroid), 즉 특정 임의지점을 선택하여 가까운 데이터들을 찾아서 묶어주는 대표적인 알고리즘이다.

70 ②

F1 = 2 / (1 / recall + 1 / precision)
= 2 × (precision−recall) / (precision + recall)
= 2 × ((0.95) × (0.9)) / ((0.95) + (0.9)) ≒ 92.4%

71 ④

다층 퍼셉트론은 입력층과 출력층 사이에 하나 이상의 은닉층이 존재하는 신경망이다.

72 ①

적합도 검정이란 데이터가 가정된 확률에 적합하게 따르는지를 검정하는 즉, 데이터 분포가 특정 분포함수와 얼마나 맞는지를 검정하는 방법이다.

73 ③

인포그래픽은 복잡한 데이터를 시각적으로 단순화 시켜서 제작한다.

74 ②

정답이 있는 데이터를 학습하는 지도 학습기법은 크게 분류, 회귀로 나눌 수 있다.

75 ③

전체 예측된 긍정에서 실제 긍정한 비율이 정밀도이다.

76 ③

드롭아웃은 학습시킬 때 무작위로 뉴런을 제외하여 뉴런의 특정 가중치에 덜 민감해지면서 과적합을 방지할 수 있다.

77 ②

분석결과에 대한 검증은 분석모델의 신뢰도를 높이기 위해 꼭 필요한 절차이다.

78 ①

엘보우 기법은 분산 비율의 증가분이 줄어드는 지점을 찾아 k값을 선택하며 실루엣 기법은 특정 객체와 속해 있는 군집 내 데이터들 간의 비유사성을 계산하여 k값을 증가시키면서 평균 실루엣 값이 최대가 되는 k를 선택한다.

79 ①

선형회귀분석은 잔차의 제곱의 합이 최소가 되게 하는 직선을 찾아가는 분석으로 잔차 분석이 있다. 로짓 변환은 로지스틱 회귀함수에서 승산(odds)으로 0과 1로 조정하는 과정을 통해 선형함수로 치환하는 것이며 크로스엔트로피는 분류모델에 대한 손실함수이다.

80 ①

적합도 검정은 추정된 회귀식이 표본의 실제값을 얼마나 잘 설명하는지에 대해 확인하는 절차이다.

모의고사 정답 ▶ 482p

01 ②	02 ②	03 ③	04 ①	05 ①
06 ②	07 ③	08 ④	09 ①	10 ④
11 ③	12 ④	13 ③	14 ②	15 ③
16 ④	17 ②	18 ②	19 ③	20 ④
21 ③	22 ①	23 ①	24 ③	25 ②
26 ④	27 ①	28 ②	29 ③	30 ②
31 ④	32 ③	33 ③	34 ②	35 ③
36 ①	37 ④	38 ①	39 ①	40 ③
41 ④	42 ③	43 ②	44 ③	45 ②
46 ③	47 ③	48 ④	49 ①	50 ②
51 ④	52 ③	53 ②	54 ③	55 ④
56 ②	57 ③	58 ③	59 ③	60 ①
61 ③	62 ④	63 ②	64 ③	65 ③
66 ①	67 ②	68 ④	69 ③	70 ②
71 ④	72 ②	73 ④	74 ③	75 ④
76 ②	77 ②	78 ②	79 ④	80 ④

1과목 | 빅데이터 분석 기획

01 ②

정형적 데이터는 비정형 데이터로 이루어져 있으며, 정형 데이터와 반정형 데이터는 정량적 데이터이다.

02 ②

형식지는 표출화(Externalization) 과정을 통해 암묵지가 구체화되어 외부로 표현된 것이다.

03 ③

다양한 비즈니스 관점에서 쉽고 빠르게 다차원적인 데이터에 접근하여 의사결정에 활용할 수 있는 정보를 얻을 수 있게 하는 기술은 OLAP에 대한 설명이다.

오답 피하기

데이터 웨어하우스는 사용자의 의사결정에 도움을 주기 위하여 기관 시스템의 데이터베이스에 축적된 데이터를 공통의 형식으로 변환해서 관리하는 데이터베이스이다.

04 ①

가치(Value)는 대용량의 데이터 안에 숨겨진 가치 발굴이 중요하며, 다른 데이터들과 연계 시 가치가 배로 증대한다. 단일 데이터라 하더라도 분석 방법론이나 분석가의 숙련도에 따라 높은 가치를 창출할 수 있다.

05 ①

연관규칙분석은 변인들 간 주목할 만한 상관관계가 있는지 찾아내는 방법이며, 독립변수가 종속변수에 미치는 영향을 분석할 때 사용하는 방법은 회귀분석이다.

06 ②

데이터 소비자(Data Consumer)의 역할과 활용 역량을 높이기 위한 데이터 리터러시(Data Literacy) 프로그램의 중요성이 커지고 있는 것은 데이터 분석 시대에 대한 설명이다.

07 ③

조직 내부에 있는 웹 서버나 시스템의 로그를 수집하는 기술은 로그 수집기이며, 센서 네트워크는(Sensor Network)는 유비쿼터스 컴퓨팅 구현을 위한 초경량 저전력의 많은 센서들로 구성된 유무선 네트워크이다.

08 ④

NoSQL(Not only SQL)은 전통적인 관계형 데이터베이스와는 다르게 데이터 모델을 단순화하여 설계된 비관계형 데이터베이스로 SQL을 사용하지 않는 DBMS와 데이터 저장장치이다. 기존의 RDBMS 트랜잭션 속성인 원자성(Atomicity), 일관성(Consistency), 독립성(Isolation), 지속성(Durability)을 포기한다.

09 ①

NoSQL은 고전적인 관계형데이터베이스의 주요 특성을 보장하는 ACID 특성 중 일부만을 지원하는 대신 성능과 확장성을 높이는 특성을 강조한다.

10 ④

문제에 대한 구체적 정의가 없다면 데이터 마이닝 기반으로 데이터를 분석하여 인사이트를 발굴하거나 일단 데이터 분석을 시도한 후 결과를 확인해 가면서 반복적으로 개선 결과를 도출해 볼 수 있다.
문제에 대한 구체적 정의가 가능하고, 필요 데이터를 보유하고 있으며, 분석역량을 갖추고 있다면 통계 기반의 전통적 데이터 분석을 수행할 수 있다.

11 ③

군집화는 특성이 비슷한 데이터를 하나의 그룹으로 분류하는 방법으로, 분류와 달리 학습 데이터 셋을 이용하지 않는 비지도학습 방법이다.

12 ④

강화학습

행동심리학에서 영감을 받았으며, 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 순서를 선택하는 방법이다. 강화학습의 초점은 학습 과정에서의 성능이며, 이는 탐색과 이용의 균형을 맞춤으로써 제고된다. 응용 영역으로는 게임 플레이어 생성, 로봇 학습 알고리즘, 공급망 최적화 등이 있다.

13 ③

개인정보의 제3자 제공은 해당 정보를 제공받는 자의 고유한 업무를 처리할 목적 및 이익을 위하여 개인정보가 이전되는 것이다. 개인정보와 제 3자에게 이전되거나 공동으로 처리하게 하는 것은 개인정보의 이전에 대한 개념이다.

14 ②

범주화는 데이터의 값을 범주의 값으로 변환하여 값을 숨기는 방법이다.

오답 피하기

총계처리는 데이터의 총합 값을 보여주고 개별 값을 보여주지 않는 방법으로, 특정 속성을 지닌 개인으로 구성된 단체의 속성 정보를 공개하는 것은 그 집단에 속한 개인의 정보를 공개하는 것과 마찬가지이므로 주의해야 한다.

15 ③

정형, 비정형, 반정형 등 모든 내외부 데이터를 대상으로 데이터의 속성, 오너, 관련 시스템 담당자 등을 포함한 데이터 정의서를 작성하는 것은 데이터 준비와 관련된 내용이다.

16 ④

메타 데이터 및 데이터 사전 구축은 데이터 표준화와 관련된 업무이며, 표준화 활동은 데이터 거버넌스 체계를 구축한 후 표준 준수 여부를 주기적으로 점검하는 것이다.

17 ②

총 6가지 영역을 대상으로 현재 수준을 파악하는 것은 분석 준비도이다. 분석 준비도는 조직 내 데이터 분석 업무 도입을 목적으로 현재 수준을 파악하기 위한 진단방법이다.

18 ②

프로타이핑 접근법은 상향식 접근 방식의 문제 해결 방법 중의 하나로 일단 먼저 분석을 시도해 보고 그 결과를 확인하면서 반복적으로 개선해 나가는 방식으로, 실험적 프로토타입보다는 진화적 프로토타입에 가깝다고 볼 수 있다.

19 ③

비즈니스 이해 및 범위 설정은 분석 기획 단계의 한 태스크로 향후 프로젝트 진행을 위한 방향을 설정하고, 프로젝트 목적에 부합한 범위를 설정하며, 프로젝트의 범위를 명확하게 파악하기 위해 구조화된 명세서를 작성한다.

20 ④

Trade off는 두 개의 목표 가운데 하나를 달성하려고 하면 다른 달성이 늦어지거나 희생되는 관계로, 정확도와 정밀도 또한 Trade off인 경우가 많지만 항상 그런 것은 아니다.

21 ③

- **평균 대치법**(Mean Imputation) : 관측 또는 실험으로 얻어진 데이터의 평균으로 결측치를 대치해서 사용한다. 평균에 의한 대치는 효율성의 향상 측면에는 장점이 있으나 통계량의 표준오차가 과소 추정되는 단점이 있다. 비조건부 평균 대치법이라고도 한다.
- **최근방 대치법**(Nearest-Neighbor Imputation) : 전체표본을 몇 개의 대체군으로 분류하여 각 층에서의 응답자료를 순서대로 정리한 후 결측값 바로 이전의 응답을 결측치로 대치한다. 응답값이 여러 번 사용될 가능성이 단점이다.

오답 피하기

- **회귀 대치법**(Regression Imputation) : 회귀분석에 의한 결측치를 대치하는 방법으로 조건부 평균 대치법이라고도 한다.
- **단순확률 대치법**(Single Stochastic Imputation) : 평균대치법에서 추정량 표준오차의 과소 추정을 보완하는 대치법으로 Hot-deck 방법이라고도 한다. 확률추출에 의해서 전체 데이터 중 무작위로 대치하는 방법이다.

22 ①

전진 선택법(Forward Selection)

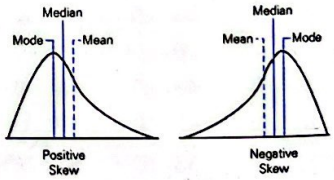
- 영 모형에서 시작, 모든 독립변수 중 종속변수와 단순상관계수의 절댓값이 가장 큰 변수를 분석모형에 포함시키는 것을 말한다.
- 부분 F 검정(F test)을 통해 유의성 검증을 시행, 유의한 경우는 가장 큰 F 통계량을 가지는 모형을 선택하고 유의하지 않은 경우는 변수선택 없이 과정을 중단한다.
- 한번 추가된 변수는 제거하지 않는 것이 원칙이다.

23 ①

PCA의 특징

- 차원 축소에 폭넓게 사용된다. 어떠한 사전적 분포 가정의 요구가 없다.
- 가장 큰 분산의 방향들이 주요 중심 관심으로 가정한다.
- 본래의 변수들의 선형결합으로만 고려한다.
- 차원의 축소는 본래의 변수들이 서로 상관이 있을 때만 가능하다.
- 스케일에 대한 영향이 크다. 즉 PCA 수행을 위해선 변수들 간의 스케일링이 필수이다.

24 ③



변수변환 전 분포	사용변수 변환식	변수변환 후 분포
좌로 치우침	X^3	정규분포화
좌로 약간 치우침	X^2	
우로 약간 치우침	X	
우로 치우침	ln(X)	
극단적 우로 치우침	1/X	

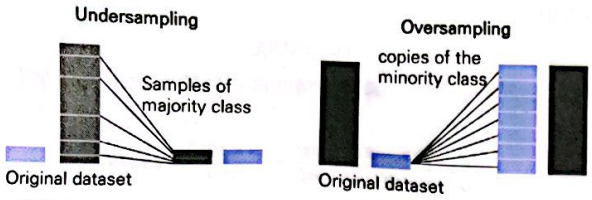
Negative Skew(우측 치우침) 경우로 ln(X)를 통한 변환을 이용한다.

오답 피하기

순위를 데이터로 범주를 나누어 상대비교로 나누어 정렬한다 : 범주형 데이터의 변환

모든 데이터를 최소값 0 최대값 1로 그리고 다른 값은 0과 1 사이 값으로 변환한다 : 데이터전체를 변환 모양과 상관없이 최대 최소 정규화 분포형태의 변화는 안한다.

25 ②



오버샘플링

소수클래스의 복사본을 만들어, 대표클래스의 수만큼 데이터를 만들어 주는 것이다. 똑같은 데이터를 그대로 복사하는 것이기 때문에 새로운 데이터는 기존 데이터와 같은 성질을 갖게 된다.

오답 피하기

언더샘플링

대표클래스의 일부만을 선택하고, 소수클래스는 최대한 많은 데이터를 사용하는 방법이다. 이때 언더샘플링된 대표클래스 데이터가 원본 데이터와 비교해 대표성이 있어야 한다.

26 ④

$$MAD(\text{Mean Absolute Deviation}) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

관측값에서 평균을 빼고, 그 차이값에 절댓값을 취하고, 그 값들을 모두 더하여 전체 데이터 개수로 나눠 준 것

평균은

$$\frac{12+20+23+25+30}{5} = 22$$

평균편차는

$$\frac{|12-22|+|20-22|+|23-22|+|25-22|+|30-22|}{5} = 4.8$$

그러나 절대 편차 형식의 최소값은 평균이 아닌 중앙값

Median = 23

$$\frac{|12-23|+|20-23|+|23-23|+|25-23|+|30-23|}{5} = 4.6$$

오답 피하기

B = 12 경우는 A =10

B = 30 경우는 A = 8

27 ①

판단추출법 (Judgement Sampling)

- 조사자가 나름의 지식과 경험에 의해 모집단을 가장 잘 대표한다고 여겨지는 표본을 주관적으로 선정하는 방법이다.
- 판단추출법에 의한 표본은 조사자의 주관적 판단에 의해서 표본이 추출되기 때문에 그 표본을 통해 얻은 추정치의 정확성에 대해 객관적으로 평가할 수 없다.
- 표본의 크기가 작은 경우에 조사의 오차를 좌우하는 요인은 추정량의 분산이 될 수 있다.

28 ②

A_1 : 간에 이상이 있을 사건

A_2 : 간에 이상이 없을 사건

B : 간기능 검사에서 이상이 나타날 사건

$P(A_1)=0.3,\ P(A_2)=0.7,\ P(B|A_1)=0.9\ P(B|A_2)=0.1$

이 된다.

여기서 총확률정리에 의해 임의의 직장인이 검사에서 이상반응을 보일 확률은

$$P(B)=P(B|A_1)P(A_1)+P(B|A_2)P(A_2)=0.9\times0.3+0.1\times0.7=0.34$$

이제 베이지안 정리에 의해

$P(A_2|B)$ = 실제 간기능에 문제가 없음에도 불구하고 이상이 있음을 나타내는 확률

$$P(A_2|B)=\frac{P(B|A_2)P(A_2)}{P(B)}=\frac{P(B|A_2)P(A_2)}{P(B|A_1)P(A_1)+P(B|A_2)P(A_2)}=\frac{0.1\times0.7}{0.34}=0.206=20.6\%$$

29 ③

확률밀도함수 : 확률 변수의 분포를 나타내는 함수이다.

모든 x에 대해서 $0 \le fx \le 1$

$$\int f(x)dx = 1$$

$$P(a<X<b)=\int_a^b f(x)dx$$

여기서

$$\int f(x)dx=\int_0^1 Ax^2dx = A\times\frac{1}{3}x^3]_0^1=\frac{A}{3}-0=1$$

$$A=3$$

P(X<1/2)일 확률은

$$P(X<1/2)=\int_0^{1/2}3x^2dx=3\times\frac{1}{3}x^3]_0^{1/2}=(\frac{1}{2})^3-0=\frac{1}{8}$$

30 ②

$$P(X>400)=\int_{400}^{\infty}\frac{1}{300}e^{-x/300}dx=\frac{1}{300}\int_{400}^{\infty}e^{-x/300}dx$$

$$=\frac{1}{300}\left[e^{-\frac{x}{300}}(-300)\right]_{400}^{\infty}=\lim_{t\to\infty}(e^{-4/3}-e^{-t})=e^{-4/3}$$

100시간동안 고장나지 않았을 때, 앞으로 400시간동안 고장나지 않고 작동할 활률은

P(X>100+400|X>100)=P(X>400)

과 같다는 것이 무기억성질(Memoryless Property) 인데

$$P(X>100+400|X>100)=\frac{P(X>500)}{P(X>100)}$$

$$=\frac{e^{-500/300}}{e^{-100/300}}=\frac{e^{-5/3}}{e^{-1/3}}=e^{-4/3}$$

31 ④

자유도가 1보다 클 때 스튜던트 t 분포에서 기대값은 0이다.

- 스튜던트 t 분포는 정규분포의 평균 측정 시 주로 사용하는 분포이다. 분포의 모양은 Z-분포와 유사하다. 종 모양으로서 t=0에 대하여 대칭을 이루는데 t-곡선의 모양을 결정하는 것은 자유도이다.
- 자유도가 클수록 정규분포에 모양이 수렴된다.

32 ③

- 표본의 크기가 큰 경우 근사적으로 정규분포를 따르게 된다는 것이 중심극한정리이다.
- 무작위로 뽑은 표본의 평균이 전체 모집단의 평균과 가까울 가능성이 높다는 것이 대수의 법칙이다.

33 ③

오답 피하기

- **편향** : 기대하는 추정량과 모수의 차이
- 표본평균은 불편추정량이나 표본분산은 불편추정량이 아니다.

34 ②

t 분포에서 자유도가 커지면 커질수록 분포의 형태는 정규분포를 따르게 되므로 평균=중앙값=최빈값으로 나타나는 분포의 모습을 그대로 유지하고 따르게 된다.

35 ③

모평균에 대한 신뢰구간를 구하는 방법 중 모집단의 분산을 모르는 경우(표본크기가 큰 경우)이므로

$$\bar{X} - Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}$$

95% 신뢰수준

$$90 - 1.960 \cdot \frac{4}{\sqrt{100}} \le \mu \le 90 + 1.960 \cdot \frac{4}{\sqrt{100}}$$

$89.22 \le \mu \le 90.78$(소수점 둘째 자리 반올림)

36 ①

각 나이대별 필요한 표본숫자는
비율에 대한 정보가 주어지지 않은 경우 p=1/2로 놓고 표본의 크기를 결정한다. 그래서

$$n = \frac{1}{4}\left(Z_{\frac{\alpha}{2}} \cdot \frac{1}{d}\right)^2$$

가 된다.

$$n = \frac{1}{4}\left(1.960 \cdot \frac{1}{0.01}\right)^2 = 9604$$

37 ④

④는 히트맵의 특징이다.

버블 차트의 특징

- x, y값의 위치를 표시하는 산점도에 점의 위치에 해당하는 제3의 변수값을 원의 크기로 표현한 그래프로 한 번에 3개의 변수를 비교해볼 수 있다.
- 제3의 값을 표시하는 원(버블)은 면적으로 표현되어야 하며, 반지름이나 지름으로 표현되면 실제 값보다 너무 크게 원이 그려질 수 있어서 주의해야 한다.
- 도시별 인구밀집도, 도시별 우유 판매량 등 국가나 지역에 따른 값의 분포를 표현하는데 매우 유리하다.

38 ①

서열자료인 두 변수들의 상관관계를 측정하는데 사용하는 것은 스피어만 상관계수에 대한 내용이다.

39 ①

모집단의 평균과 분산이 각각 μ, σ^2인 정규모집단 $N(\mu, \sigma^2)$에서 μ, σ^2가 미지인 경우 모분산 σ^2에 대한 가설검정은 점추정량인 s2을 이용하여 검정한다.

① **가설의 설정**

- **귀무가설** $H_0 : \sigma^2 = \sigma_0^2$
- **대립가설** $H_1 : \sigma^2 \neq \sigma_0^2$(양측검정), 또는 $H_1 : \sigma^2 > \sigma_0^2$(단측검정:우측검정), $H_1 : \sigma^2 < \sigma_0^2$(단측검정:좌측검정)

② 검정통계량 $\chi^2 = \frac{\sum(x_i - \bar{x})^2}{\sigma_0^2} = \frac{\phi s^2}{\sigma_0^2}$ (여기서 Ø=n-1 자유도)

③ 표본크기 n과 유의수준 α에 의해서 결정됨

귀무가설은

$$H_0 : \sigma^2 = 1200$$

대립가설은

$$H_1 : \sigma^2 < 1200$$

자유도는 Ø=n-1=30-1=29이고 이에 따른 검정 통계량은 다음 아래와 같다.

$$\chi^2 = \frac{\sum(x_i - \bar{x})^2}{\sigma_0^2} = \frac{\phi s^2}{\sigma_0^2} = \frac{29 \times 1050}{1200} = 25.375$$

이에 따른 기각역은 유의수준에 따라서 (x2분포표에 의해)

$$\chi^2 \ge \chi^2(\phi, 1-\alpha) = \chi^2(29, 0.95) = 17.71$$

25.375>17.71 이므로 H_0를 기각할 수 없다(채택). 그러므로 새로운 공정을 변경하더라도 제품수명의 변동은 적어지지 않는다.

40 ③

임계치는 주어진 유의수준 α에서 귀무가설의 채택과 기각에 관련된 의사결정을 할 때, 그 기준이 되는 점이다.

3과목 | 빅데이터 모델링

41 ④

로지스틱 회귀분석은 지도학습 분류 부문에 해당된다.

42 ③

GAN은 적대적 생성 모델로 생성모델과 판별모델이 존재한다.

43 ②

강화학습이란 주어진 환경에서 보상을 최대화하도록 에이전트를 학습하는 기법이다.

44 ③

의사결정나무에서의 뿌리마디는 나무가 시작되는 마디를 뜻한다.

45 ②

의사결정 분류나무에서 범주형 목표변수에 대해 분리를 수행할 시 카이제곱 검정을 적용하게 되면 관측도와 기대도수의 차이가 커질수록 순수도는 높아진다. 즉 카이제곱 검정 통계량이 가장 큰 예측 변수를 이용하여 자식마디를 형성하게 된다.

46 ③

정보획득 : 순도가 증가하고 불확실성이 감소

47 ③

부트스트래핑은 랜덤 샘플링으로 크기가 동일한 여러 개의 표본자료들을 생성한다.
복원 추출법은 샘플 추출 뒤 다시 표본자료에 포함시켜 추출하는 방식이다.

48 ④

의사결정나무는 구조가 복잡하게 되면 해석력이 떨어진다.

49 ①

노드는 신경계 뉴런, 가중치는 신경계 시냅스에 비유된다.

50 ②

가중치 매개변수 기울기를 미분을 통해 전달하는 것은 시간 소모가 크므로 이를 개선하기 위한 방법인 오차역전파는 실제 출력과 목표 출력값과의 오차를 출력층에서 입력층으로 전달, 연쇄법칙을 활용하여 가중치와 편향을 계산, 업데이트한다.

51 ④

딥러닝 모델 학습에서 가중치와 편향은 수동이 아닌 자동으로 설정되는 매개변수(파라미터)에 속한다.

52 ③

LSTM은 입력 게이트, 출력 게이트, 망각 게이트를 가진다.

53 ②

오토인코더는 다차원 데이터를 저차원으로 바꾸고 바꾼 저차원 데이터를 다시 고차원 데이터로 바꾸면서 특징점을 찾아내는 비지도학습 알고리즘이다.

54 ③

서포트 벡터는 두 클래스를 구분하는 경계선으로 각 서포트 벡터를 지나는 초평면의 거리가 초평면의 마진이다.

55 ④

맨하탄거리는 택시 거리, 시가지 거리로도 불리며 두 점의 좌표 값의 절대적 차이로 구한다.

56 ②

범주형 변수에 대해서 두변수간의 연관성 검증을 위해서 사용되는 분석기법은 교차분석이며
이때 통계량은 χ^2 이다.

57 ③

자기상관성(Autocorrelation)은 시차값 사이에 선형 상관관계를 보이는 것을 말한다.

58 ③

나이브 베이즈 모델은 $P(C_1|Doc)/P(Doc)$와 $P(C_2|Doc)/P(Doc)$를 비교해서 그 값이 큰 쪽으로 범주를 할당한다는 개념이다.

59 ③

60 ①

심층 신뢰 신경망

- 기계학습에서 사용되는 그래프 생성 모형이다.
- 딥러닝에서는 잠재변수의 다중계층으로 이루어진 심층신뢰 신경망을 의미한다. 계층 간에 연결이 있지만 계층 내 유닛 간에는 연결이 없다는 특징이 있다.

4과목 | 빅데이터 결과 해석

61 ③

분류 평가지표로 정확도, 재현율, 정밀도, F1 점수 등이 있다.

62 ④

모든 다양한 분류 임계값의 TPR 및 FPR을 나타내는 그래프는 ROC이며 AUC는 ROC 아래 면적을 뜻한다.

63 ②

MSE는 잔차(오차)의 제곱에 대한 평균을 취한 값으로 주요 회귀지표 중의 하나이다.

64 ③

k-평균군집 분석은 원하는 군집 수만큼(k개) 초기값을 지정하고, 각 개체를 가장 가까운 중심에 할당하여 군집을 생성한 뒤 각 군집 내 평균을 계산하여 중심점을 갱신한다. 해당 과정을 반복 진행하며 군집 중심의 변화가 없게 되면 최종 군집이 형성된다.

65 ③

K=5로 1가지 데이터셋을 5등분으로 Fold화 하며 각 Fold마다 한 번씩 평가(Validation) 데이터셋으로 사용하여 총 5회 훈련이 진행된다. 5회 평가, 최종 테스트 1회로 평가를 포함한 테스트 횟수는 총 6회이며 각 회당 학습 결과에 대한 전체 평균이 해당 모델의 성능으로 나타난다.

66 ①

드롭아웃은 훈련할 때 신경망의 뉴런을 부분적으로만 사용함으로써 학습이 덜 될 수 있으나 과적합을 예방할 수 있다.

67 ②

L1 규제기법은 규제 가중치의 절대값을 손실함수에 더해줌으로써 가중치를 작게 만들어 과적합을 방지할 수 있다.

68 ④

확률적 경사 하강법(SGD)은 손실함수를 가중치로 미분한 기울기에 학습률을 곱하여 현재의 매개변수인 가중치에서 뺀 값이 다시 손실함수가 계산되어 이를 통해 가중치를 갱신하는 과정이 반복된다.

69 ③

초매개변수는 사람이 직접 설정해주어야 하는 매개변수로 가중치는 직접 설정이 불가능하다. 또한 초매개변수 최적화는 임의로 범위 선정 후 무작위로 초매개변수 값을 샘플링하여 모델 정확도를 평가하면서 최적값의 범위를 줄여가는 과정으로 딥러닝 학습 시간이 오래 소요되므로 학습 에폭을 작게 검증/평가 시간을 단축시키는 것이 중요하다.

70 ②

배깅은 각각 독립적인 학습이 끝난 뒤 결과를 종합하는 기법이라면 부스팅은 이전 학습결과를 토대로 다음 학습 데이터의 샘플 가중치를 조정해 순차적으로 학습을 진행한다.

71 ④

스캐터 플롯(산점도), 히트맵, 버블차트는 비교시각화를 위한 도구이며, 파이차트는 하나의 변수에 대한 값의 분포를 보여주기에 적합한 분포시각화 도구이다.

72 ②

평행좌표계는 스타차트를 넓게 펼친 모양으로 여러 변수의 각 영역에 따른 값을 비교해서 보여주기에 적합하다.

01 ①	02 ②	03 ③	04 ④	05 ①
06 ②	07 ③	08 ④	09 ①	10 ②
11 ③	12 ④	13 ①	14 ②	15 ③
16 ④	17 ①	18 ②	19 ③	20 ④
21 ②	22 ②	23 ②	24 ④	25 ③
26 ③	27 ③	28 ④	29 ④	30 ③
31 ①	32 ④	33 ④	34 ①	35 ④
36 ①	37 ④	38 ④	39 ③	40 ①
41 ④	42 ③	43 ④	44 ③	45 ③
46 ②	47 ③	48 ③	49 ①	50 ④
51 ②	52 ①	53 ②	54 ①	55 ③
56 ②	57 ④	58 ③	59 ②	60 ③
61 ①	62 ②	63 ③	64 ①	65 ②
66 ④	67 ④	68 ②	69 ①	70 ④
71 ③	72 ④	73 ①	74 ②	75 ②
76 ④	77 ④	78 ④	79 ④	80 ②

1 과목 | 빅데이터 분석 기획

01 ①

정량적 데이터의 유형은 정형 데이터와 반정형 데이터이고, 정성적 데이터의 유형은 비정형 데이터이다.

02 ②

지식은 상호 연결된 정보를 구조화하여 유의미한 정보를 분류하고 개인적 경험을 결합시켜 내재화한 고유의 결과물이며, 이 경우 '텀블러를 저렴한 온라인 상점에서 구매할 것이다'는 표현으로 고쳐 쓰는 것이 더 적합하다.

03 ③

OLTP는 데이터 액세스 빈도가 높은 편이지만, OLAP는 데이터 액세스 빈도가 보통이다.

04 ④

아무리 데이터의 종류가 다양하다 하더라도 고품질의 데이터가 입력되어야 고수준의 인사이트 도출이 가능하다.

05 ①

빅데이터는 시장에 새롭게 진입하려는 잠재적 경쟁자에게는 진입장벽과도 같은 존재이다.

06 ②

데이터 산업은 데이터 처리 시대, 데이터 통합 시대, 데이터 분석 시대, 데이터 연결 시대, 데이터 권리 시대로 진화하고 있다.

07 ③

분석 수행의 일반적 구조는 기능형 조직구조이다.

오답 피하기

기능형 조직구조 : 각 현업 부서에서 분석 업무를 직접 수행한다. 전사적 관점에서 전략적 핵심 분석이 어려우며, 특정 협업 부서에 국한된 협소한 분석을 수행할 가능성이 높다.

오답 피하기

- 도넛 차트는 여러 변수(학생)를 보여줄 수 있지만, 과목별 점수를 직관적으로 비교하기는 어렵다.
- 산점도는 최대 3차원(3과목)까지 표현이 가능하다.

73 ④

탐색적 데이터분석에서 시각화를 이용한 데이터 분포를 파악한다. 연관분석은 장바구니 분석이라고도 하며 소비자의 구매패턴을 분석하는 기법이다. 히트맵과 인포그래픽은 시각화 도구이다.

74 ③

추세선 위에 존재하는 데이터들은 삭제해도 추세선의 모양에 영향을 주지 않는다.

오답 피하기

① 인플레이션과 실업률은 음의 상관관계가 있다.
② A는 동일한 실업률을 보이는 집단에서 인플레이션이 매우 높은 값을 보여주기 때문에 이상값으로 분류될 수 있다.
④ B와 C는 어느 한 시점의 인플레이션 지수를 표현한 것으로 특정 도시의 인플레이션 지수를 대표하지는 않는다.

75 ④

선유형 속성은 점선, 이중 점선 등 각각의 독립된 모양으로 데이터를 표현하는 것으로 연속형 데이터를 표현하기에 적합하지 않다.

76 ②

1월 1000, 2월 2000, 3월 3500, 4월 3000, 5월 2500, 6월 2000으로 총 합계는 14,000이다. 3월매출 / 총매출 = 3500 / 14000 = 0.25로 1/4을 차지한다. 360도×1/4은 90도이다.

77 ②

다차원척도법은 대상의 상대적인 거리를 표현하는 방법으로 관측대상의 x, y좌표값 실제값과 다르다.

78 ②

단계구분도는 면적이 넓은 지역의 값이 전체를 지배하는 것처럼 보이는 시각적 왜곡이 발생할 수 있다. 카토그램은 실제 데이터 값에 비례하여 지역을 확대하거나 축소함으로써 단계구분도의 시각적 왜곡을 보완한다.

79 ④

전개 단계에서는 1.분석결과 활용 계획 수립, 2.분석결과 적용과 보고서 작성, 3.분석모형 모니터링, 4.분석모형 리모델링의 업무가 진행된다.

80 ④

분석모형 리모델링은 현재 진행되고 있는 분석프로젝트의 성능을 유지, 개선하기 위한 활동을 말하며, 신규분석과제 발굴은 분석수요조사 및 기획과정에서 진행한다.

08 ④

빅데이터를 처리하는 과정에서는 생성 기술, 수집 기술, 저장(공유) 기술, 처리 기술, 분석 기술, 시각화 기술이 필요하다.

09 ①

강화학습은 행동심리학에서 영감을 받았으며, 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 순서를 선택하는 방법이다.

10 ②

개인정보는 생존하는 개인에 관한 정보여야 하며, 정보의 내용 및 형태 등은 제한이 없고, 개인을 알아볼 수 있는 정보여야 한다. 또한 다른 정보와 쉽게 결합하여 개인을 알아볼 수 있는 정보도 포함한다.

11 ③

데이터 활용에 따른 개인정보처리자의 책임을 강화한 것이지, 조직 대표자의 연대책임 여부까지 논한 것은 아니다.

12 ④

비록 분석 주제는 정의하지 못한 상태이지만 분석 방법을 알고 있다면 인사이트 발굴이 가능하다.

13 ①

IT 프로젝트의 과제 우선순위 평가기준으로는 전략적 필요성, 시급성, 투자용이성, 기술 용이성 항목이 있다.

14 ②

기존 시스템에 미치는 영향을 최소화하여 적용하는 방안이 이상적이기는 하지만 현실적으로 어려우므로, 기존 시스템과 별도로 시행하여 난이도 조율을 통한 우선순위를 조정할 수 있다.

15 ③

분석 역량을 확보하지 못하였고, 분석 기법이나 시스템을 보유하고 있지 않을 때 아웃소싱을 진행하며, 만일 분석 역량은 확보하고 있다면 시스템 고도화를 진행한다.

16 ④

폭포수 모형은 요구사항 도출이 어려우며, 원형 모형은 프로토타입의 폐기가 발생하고, 나선형 모형은 계획수립, 위험분석, 개발, 고객평가 순으로 진행된다.

17 ①

오답 피하기

② KDD 분석 방법론의 분석절차
③ SEMMA 분석 방법론의 분석절차
④ 빅데이터 분석 방법론의 개발절차

18 ②

분석 프로젝트 관리 시 데이터의 크기는 데이터 지속적으로 생성되어 증가하는 점을 고려한다.

19 ③

데이터의 종류는 정형 데이터, 반정형 데이터, 비정형 데이터 등을 한정하지 않고 모두 수용한다.

20 ④

임의 잡음 추가는 데이터 범주화 방법이 아닌 데이터 마스킹 방법이다.

2과목 | 빅데이터 탐색

21 ②

오답 피하기

① 데이터 정제는 수집된 데이터를 대상으로 분석에 필요한 데이터를 추출하고 통합하는 과정이다.
③ 데이터로부터 원하는 결과나 분석을 얻기 위해서 분석도구나 기법에 맞게 다듬는 과정이 필요하다.
④ 후처리는 데이터 저장 후의 처리를 지칭하며 저장데이터의 품질관리 등의 과정을 포함한다.

22 ②

서열자료(Ordinal Data) : 명목자료와 비슷하나 수치나 기호가 서열을 나타내는 자료이다.

오답 피하기

질적자료(Qualitative Data) : 정성적 자료라고도 하며 자료를 범주의 형태로 분류한다. 분류의 편리상 부여된 수치의 크기 자체에는 의미를 부여하지 않는 자료이며 명목자료, 서열자료 등 이질적자료로 분류된다.
명목자료(Nominal Data) : 측정대상이 범주나 종류에 대해 구분되어지는 것을 수치 또는 기호로 분류되는 자료이다.
④ 기본적으로 정형자료에 대한 분류 체계이다.

23 ②

완전 무작위 결측(MCAR) : 어떤 변수상에서 결측 데이터가 관측된 혹은 관측되지 않는 다른 변수와 아무런 연관이 없는 경우, 결측 데이터를 가진 모든 변수가 완전 무작위 결측이라면 대규모 데이터에서 단순 무작위 표본추출을 통해 처리 가능하다.

24 ④

단순확률 대치법(Single Stochastic Imputation) : 평균대치법에서 추정량 표준오차의 과소 추정을 보완하는 대치법으로 Hot-deck 방법이라고도 한다. 확률추출에 의해서 전체 데이터 중 무작위로 대치하는 방법이다.

25 ③

전진 선택법(Forward Selection)
- 영 모형에서 시작, 모든 독립변수 중 종속변수와 단순상관계수의 절댓값이 가장 큰 변수를 분석모형에 포함시키는 것을 말한다.
- 부분 F 검정을 통해 유의성 검증을 시행, 유의한 경우는 가장 큰 F 통계량을 가지는 모형을 선택하고 유의하지 않은 경우는 변수선택 없이 과정을 중단한다.
- 한번 추가된 변수는 제거하지 않는 것이 원칙이다.

26 ③

차원의 증가는 분석모델 파라메터의 증가 및 파라메터 간의 복잡한 관계의 증가로 분석결과의 과적합 발생의 가능성이 커진다. 이것은 분석모형의 정확도(신뢰도) 저하를 발생시킬 수 있다.

27 ③

오답 피하기

라) 차원 축소에 폭넓게 사용된다. 어떠한 사전적 분포 가정의 요구가 없다.
마) 차원의 축소는 본래의 변수들이 서로 상관이 있을 때만 가능하다.

28 ④

오답 피하기

① 데이터에서 각 클래스가 갖고 있는 데이터의 양에 차이가 큰 경우, 클래스 불균형이 있다고 말한다.
② 데이터 클래스 비율이 너무 차이가 나면(Highly-imbalanced Data) 단순히 우세한 클래스를 택하는 모형의 정확도가 높아지므로 모형의 성능 판별이 어려워진다. 즉, 정확도(accuracy)가 높아도 데이터 개수가 적은 클래스의 재현율(recall-rate)이 급격히 작아지는 현상이 발생할 수 있다.
③ 클래스 균형은 소수의 클래스에 특별히 더 큰 관심이 있는 경우에 필요하다.

29 ④

오답 피하기

중앙값은 전체변수의 범위에서 가운데가 아니라 관찰된 변수들 중의 가운데 값이므로 이상값의 영향을 받지 않는다.
첨도, 왜도는 데이터의 분포모양에 해당된다.

30 ③

피어슨 상관계수는 두 변수 X 와 Y 간의 선형 상관관계를 계량화한 수치이다.

오답 피하기

②, ④는 스피어만 상관계수에 대한 설명이다.

31 ①

$CV = \frac{\sigma}{\mu} \times 100(\%)$ (모집단의 변동계수) 이므로

체중에 대한 CV = 2.54/52.3×100=4.856%
신장에 대한 CV=2.28/152.3.3×100=1.493% 이므로
체중에 대한 CV가 더 큼 → 산포도가 넓으므로 개인차가 크다.

32 ④

위상적 타입 : 공간 객체간의 관계를 표현하며, 방위, 공간 객체간의 중첩, 포함, 교차, 분리 등과 같은 위치적 관계

33 ④

회귀분석의 경우 하나의 반응변수를 여러 개의 설명변수로 설명하고자 할 때, 가장 설명력이 높은 변수들의 선형결합을 찾아 이들 사이의 인과관계를 생각하는 반면에 정준분석에서는 이와 같은 인과성이 없다.

34 ①

최대대표라는 현상은 없다

오답 피하기

표본추출오차(Sampling Bias, Sampling Error) : 표본에서 선택된 대상이 모집단의 특성을 과잉 대표하거나 최소 대표할 때 발생한다.

35 ④

금융상품 가입 상담 건수 10회 중 실제 가입이 이루어진 수는 이항분포에 적용할 수 있다.

오답 피하기

포아송분포 : 단위 시간 안에 어떤 사건이 몇 번 발생할 것인지를 표현하는 이산확률분포

36 ①

오답 피하기

② 종 모양으로서 t=0에 대하여 대칭을 이루는데 t-곡선의 모양을 결정하는 것은 자유도이다.
③ 자유도가 클수록 정규분포에 모양이 수렴된다.
④ 자유도가 1보다 클 때만 스튜던트 t 분포에서 기대값은 0이다.

37 ④

표본평균은 불편추정량이나 표본분산은 불편추정량이 아니다.
(표본분산과 모분산의 계산 차이의 이유, n이 아닌 n-1로 나누는 이유)

38 ④

39 ③

오답 피하기

① 연구자에 의해 설정된 가설은 표본를 근거로 하여 채택여부를 결정짓게 되는데 이때 사용되는 통계량을 검정통계량이라 정의한다
② 귀무가설(Null Hypothesis, H_0)은 현재 통념적으로 믿어지고 있는 모수에 대한 주장 또는 원래의 기준이 되는 가설이다.
④ 대립가설(Alternative Hypothesis, H_1)은 연구자가 모수에 대해 새로운 통계적 입증을 이루어 내고자 하는 가설이다.

40 ①

두 독립표본의 평균차이 검정의 검정 통계량
검정 통계량 설정 : 위의 가설을 검정하는데 사용되는 검정통계량은 X-표본과 Y-표본의 표본평균인 $\bar{X}$와 $\bar{Y}$의 차이에 근거하여 구성한다.

검정 통계량 $T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$

여기서 $S_p^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$ 으로 공통분산(Common Variance) σ^2의 합동표본분산(Pooled Sample Variance)이며 S_1^2, S_2^2는 각각의 표본의 표본분산을 말한다. 검정 통계량 T는 자유도 m+n-2인 t 분포를 따른다.

3 과목 | 빅데이터 모델링

41 ④

다중회귀분석은 회귀(예측)모델로 분류된다.

42 ③

종속변수 : 범주형 변수
분포 : 이항분포

43 ④

다중공선성 진단 → 회귀계수 유의성 확인 → 수정된 결정계수 확인 → 모형의 적합도 평가

44 ③

정보 획득(Information Gain) : 순도가 증가하고 불확실성이 감소하는 것을 뜻한다. 정보의 가치를 반환하는 데 발생하는 사건의 확률이 작을수록 정보의 가치는 높아지며, 확률이 높을수록 가치는 낮아진다.

45 ③

의사결정나무의 대표적 알고리즘인 CART는 불순도 측도로 범주형 또는 이산형일 경우 지니지수를, 연속형인 경우 분산의 감소량을 이용한 이진분리를 활용한다

46 ②

배깅(Bagging) : 기계학습 알고리즘의 안정성과 정확도를 향상시키기 위해 고안되었다.

47 ③

분류 모델의 앙상블은 다수결로 0 또는 1로 분류한다.

48 ③

손실함수를 최소화하기 위해 가중치와 편향을 찾는 것이 인공신경망의 학습이며 일반적인 손실함수로는 평균제곱오차가 있다.

49 ①

데이터를 미니배치로 무작위 선정 뒤 손실함수 값을 줄이기 위해 각 가중치 매개변수 기울기를 구한다. 다음 가중치 매개변수 기울기 방향으로 조금씩 갱신하여 앞에서 진행한 단계들을 반복한다.

50 ④

오차역전파는 실제 출력과 목표 출력값과의 오차 산출, 비례한 가중치를 출력층에서 은닉층으로 갱신한다.

51 ②

$(4-1)^2+(4-3)^2$에 root(제곱근)을 적용, 계산한다.

52 ①

분류모델에 틀린 곳에 집중하여 새로운 분류규칙을 생성, 즉 weak classifier에 중점을 두는 지도학습 알고리즘은 부스팅이다.

53 ②

Relu 활성화 함수(이진 분류)는 Sigmoid의 Gradient Vanishing 문제를 해결하며 0보다 크면 입력값을 그대로 출력 0 이하의 값만 0으로 출력한다.

54 ①

랜덤포레스트는 여러 개의 의사결정 나무를 활용, 예측 결과를 투표 또는 다수결 방식으로 예측 결정한다.

55 ③

support(기저귀 → 맥주) = 3/5
confidence(기저귀 → 맥주) = 3/4
lift(기저귀 → 맥주) = 5/4

56 ②

오답 피하기

로지스틱 회귀분석 : 분석하고자 하는 대상들이 두 집단 또는 그 이상의 집단으로 나누어진 경우 개별관측치들이 어느 집단으로 분류될 수 있는지를 분석할 때 사용한다.
②는 카이제곱검정에 대한 내용이다.

57 ④

시계열 데이터가 분산이 일정하지 않으면 변환(transformation)을 통해 정상성을 가지도록 할 수 있다.

58 ③

딥러닝은 인공신경망의 학습수준을 높이기 위해 하나의 은닉층에 은닉노드를 3개가 아니라 10개, 100개 이런 식으로 동일레이어 내 수직으로 쭉 늘려 놓기만 했었는데, 그것보다는 은닉층 자체를 여러개로 만들어서 여러 단계를 거치도록 인공신경망을 구성하였더니 정확도가 훨씬 향상되는 원리이다.

59 ②

순환신경망(RNN: Recurrent Neural Network)의 정의와 특징에 대한 설명이다.

60 ③

앙상블(Ensemble) 기법은 동일한 학습 알고리즘을 사용해서 여러 모델을 학습하는 개념이다.

4과목 | 빅데이터 결과 해석

61 ①

정확도는 True인 데이터를 True로 False인 데이터를 False로 분류하는 정도를 말한다.

62 ②

평균제곱오차는 실제값과 예측값의 차이의 제곱에 대한 평균을 취한 값으로 다음과 같이 구할 수 있다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{4}((-1)^2 + (1)^2 + (-2)^2 + (-2)^2) = 2.5$$

63 ③

실루엣 계수는 같은 군집에 속한 요소들의 평균거리와 함께 가장 가까운 이웃군집까지의 거리도 함께 고려해서 계산한다.

64 ①

중심극한정리 : 동일한 확률분포를 가진 독립 확률 변수 n개의 평균의 분포는 n이 적당히 크다면 정규분포에 가까워진다는 이론으로 이때 표본분포의 평균은 모집단의 모평균과 동일하며 표준편차는 모집단의 모표준편차를 표본 크기의 제곱근으로 나눈 것과 같다.

65 ②

잔차의 등분산성 진단은 잔차의 분산이 특정 패턴이 없이 순서와 무관하게 일정한지를 진단한다.

66 ④

경사하강법은 매개변수 최적화에 사용되는 기법으로 손실함수의 값을 최소화하도록 하는 매개변수를 찾는 방법이다.

67 ④

군집분석은 유사성이 높은 요소들을 묶어주는 것으로 군집에 속한 요소들의 평균은 모델의 타당성을 검증하는 지표로 적절하지 않다.

68 ②

결합분석 모형은 두 종류 이상의 결과변수를 동시에 분석할 수 있는 방법으로 결과 변수 간의 유의성, 관련성을 설명할 수 있다.

69 ①

히트맵은 변수들 간의 관계를 표현하는데 적합하며 주로 회귀모델에서 사용된다.

70 ④

군집분석 모델은 군집그룹의 통계량을 요약하고 관측치의 공통점과 변동성을 확인한다. 요소 사이의 거리 평균은 모델의 성능을 평가할 때 사용하는 지표이다.

71 ③

파이차트는 데이터의 분포를 표현하는데 적합하며 연관분석 시각화 도구는 네트워크 다이어그램이 대표적이다.

72 ④

데이터 시각화는 수치정보뿐만 아니라 비정형 데이터인 텍스트나 지형정보의 표현도 모두 포함하는 개념이다.

73 ①

인포그래픽은 정보의 시각적 표현과 전달에 중심을 두며, 주로 뉴스 기사, 포스터 등에서 활용된다.

74 ②

범례는 차트에 표현되고 있는 기호나 선 등이 어떤 의미인지 설명하는 역할을 한다.

75 ②

누적막대그래프는 두 개 이상의 변수를 동시에 다루는 경우에 막대의 영역을 구분하여 나머지 변수의 값을 표현한다. 하나의 막대를 구성하는 세부항목 각각의 값과 전체의 합을 함께 표현할 때 유용하다.

76 ④

그래프의 추세선이 막대보다 위에 있는 경우는 이번달 매출보다 지난달 매출이 높다는 것을 의미한다.

77 ④

파이차트는 구성요소들이 차지하고 있는 비율을 표현하기에 적합하며, 시간에 따른 데이터의 변화를 표현하기 위해서는 적합하지 않다.

78 ④

데이터 시각화를 통해서 데이터의 이상치를 효율적으로 발견할 수 있으며, 결측치는 데이터가 비어있는 부분이라 시각화를 통해서 발견하기는 어렵다.

79 ④

CRISP-DM에서 분석모형 전개(Deploy)는 완성된 모델을 실제 업무 현장에 적용하는 단계로 전개 계획 수립, 모니터링과 유지보수 계획 수립, 프로젝트 종료 관련 프로세스로 구성된다.

80 ②

성과 평가의 결과를 바탕으로 필요한 경우 분석 모델을 리모델링한다.