1과목 빅데이터 분석 기획

01 분석 과제 우선순위 평가 기준에서 전략적 중요도, 목표 가치와 관련이 있는 빅데이터 특성은 무엇인가?

① Value
② Volume
③ Variety
④ Velocity

02 다음 중 빅데이터 유형의 사례를 설명한 것으로 가장 부적절한 것은 무엇인가?

① 정형 – 관계형 데이터베이스
② 정형 – HTML
③ 반정형 – JSON, XML
④ 비정형 – 텍스트 문서, 이미지

03 다음이 설명하는 빅데이터의 유형으로 가장 올바른 것은 무엇인가?

- 수집 데이터 각각이 데이터 객체로 구분
- 고정 필드 및 메타데이터(스키마 포함)가 정의되지 않음
- 텍스트 문서, 이진 파일, 이미지, 동영상 등

① 정형 데이터 ② 반정형 데이터
③ 비정형 데이터 ④ 정수형 데이터

04 다음 중 데이터 사이언티스트의 일반적인 요구 역량으로 가장 적합하지 않은 것은?

① 통찰력 있는 분석
② 다분야 간 협력
③ 빅데이터에 대한 이론적 지식
④ 높은 지능과 과학적 지식

05 다음 중 데이터 사이언티스트에게 요구되는 역량을 설명한 것으로 가장 부적절한 것은 무엇인가?

① 스토리텔링, 데이터 시각화를 사용한 등 설득력 있는 전달을 위해 Soft Skill이 필요하다.
② 빅데이터에 대한 이론적 지식인 Soft Skill이 필요하다.
③ 최적의 분석 설계 및 노하우 축적하는 등 분석기술에 대한 숙련을 위해 Hard Skill이 필요하다.
④ 창의적 사고, 호기심, 논리적 비판하는 Soft Skill이 필요하다.

06 빅데이터 플랫폼은 원천 데이터에서 정형, 반정형, 비정형 데이터를 수집하고 저장한다. 다음 중 빅데이터 수집 기술로 가장 부적절한 기법은 무엇인가?

① NoSQL
② ETL
③ EAI
④ 크롤러(Crawler)

07 다음 중 분석 가치 에스컬레이터에 대한 설명으로 가장 올바르지 않은 것은?

① 묘사 분석은 과거에 어떤 일이 일어났고, 현재는 무슨 일이 일어나고 있는지 확인하는 분석이다.
② 진단 분석은 데이터를 기반으로 왜 발생했는지 이유를 확인하는 분석이다.
③ 예측 분석은 무엇을 해야 할 것인지를 확인하는 분석이다.
④ 분석 가치 에스컬레이터에서는 높은 난도를 수반하는 데이터 분석은 더 많은 가치를 창출한다.

08 다음 중 대용량 파일을 저장하고 처리하기 위해서 개발된 파일 시스템으로 네임 노드(Master)와 데이터 노드(Slave)로 구성된 것은?

① 아파치 스파크(Apache Spark)
② 얀(YARN)
③ 맵리듀스(Map Reduce)
④ 하둡 분산 파일 시스템(HDFS)

09 다음의 하둡 에코시스템 중에서 비정형 데이터 수집을 위한 시스템으로 가장 부적절한 것은 무엇인가?

① 척와(Chukwa)
② 플럼(Flume)
③ 스크라이브(Scribe)
④ 피그(Pig)

10 데이터 거버넌스 체계에 대한 설명으로 가장 올바르지 않은 것은?

① 데이터 표준 용어 설명, 명명 규칙, 메타데이터 구축, 데이터 사전 구축 등 데이터 표준화를 관리해야 한다.
② 메타데이터와 데이터 사전의 관리 원칙 수립을 해야 한다.
③ 메타데이터 및 표준 데이터를 관리하기 위한 별도의 저장소 구축은 필요 없다.
④ 데이터 거버넌스 체계 구축 이후 표준 준수 여부를 주기적으로 점검 및 모니터링을 실시해야 한다.

11 다음 중 기업의 데이터 분석 수준을 파악하기 위한 조직 평가 성숙도 단계에 대한 설명으로 적절하지 않은 것은?

① 도입 단계는 분석을 시작하는 단계로 환경과 시스템을 구축하고, 전문 담당 부서에서 분석을 수행하는 단계이다.
② 활용 단계는 분석 결과를 실제 업무에 적용하는 단계로 분석 기법을 도입하는 단계이다.
③ 확산 단계는 전사 차원에서 분석을 관리하고 공유하는 단계이다.
④ 최적화 단계는 분석을 진화시켜서 혁신 및 성과 향상에 기여하는 단계이다.

12 다음 중 개인정보 비식별 조치 방법으로 가장 올바르게 설명한 것은 무엇인가?

① 데이터 마스킹: 정약용, 21세 → 박 씨, 20~30세
② 데이터 범주화: 정약용, 21세 → 정 씨, 평균 20세
③ 가명처리: 정약용, 21세 → 장길산, 20대
④ 총계처리: 장길산 160cm, 정약용 180cm → 학생 키 150~200cm

13 다음 중 개인정보의 수집·이용을 위해 정보주체의 동의를 받을 때 고지사항으로 가장 올바르지 않은 것은?

① 개인정보의 수집·이용 목적
② 동의를 거부할 권리가 있다는 사실 및 동의 거부에 따른 불이익이 있는 경우에는 그 불이익의 내용
③ 개인정보를 수집하는 기관, 담당자 연락처
④ 수집하려는 개인정보의 항목

14 다음 중 분석의 대상이 무엇인지를 인지하고 있는 경우(Known), 즉 해결해야 할 문제를 알고 있고 이미 분석의 방법도 알고 있는 경우(Known)에 사용하는 분석 유형은 무엇인가?

① 최적화(Optimization)
② 솔루션(Solution)
③ 통찰(Insight)
④ 발견(Discovery)

15 다음 중 CRISP-DM 분석 방법론에서의 데이터 준비 과정을 설명한 것으로 가장 적절한 것은 무엇인가?

① 다양한 모델링 기법과 알고리즘을 선택하고 파라미터를 최적화한다.
② 분석 결과 평가, 모델링 과정을 평가한다.
③ 전개 계획 수립, 모니터링과 유지보수 계획을 수립한다.
④ 분석용 데이터 세트 선택, 데이터 정제, 데이터 통합 등을 수행한다.

16 다음 중 개인정보를 목적 외의 용도로 이용하거나 제3자에게 제공이 가능한 경우로 옳지 않은 것은?

① 정보주체로부터 별도의 동의를 받은 경우
② 데이터 이용 활성화를 위한 통계작성에 이용해야 할 경우
③ 다른 법률에 특별한 규정이 있는 경우
④ 범죄의 수사와 공소의 제기 및 유지를 위하여 필요한 경우

17 다음 중 인터넷상에서 제공되는 다양한 웹 사이트로부터 소셜 네트워크 정보, 뉴스, 게시판 등의 웹 문서 및 콘텐츠를 수집하는 기술은 무엇인가?

① RSS(Rich Site Summary)
② Open API
③ 아파치 카프카(Apache Kafka)
④ 크롤링(Crawling)

18 빅데이터 수집 시스템에서 수집 대상이 되는 데이터를 시간 관점(활용 주기)에서 분류하면 실시간 데이터, 비실시간 데이터로 나눌 수 있다. 다음 중 실시간 데이터로 가장 부적절한 것은 무엇인가?

① IoT 센서 데이터
② 네트워크 장비 로그
③ 구매 정보
④ 알람

19 다음 중 가명처리 4단계 절차에 대한 설명으로 가장 올바르지 않은 것은?

① 사전준비 단계 – 가명처리 대상 항목 및 처리수준을 정의하기 위해서는 처리 목적이 적합한지 여부를 확인하고 사전 계획을 수립한다.
② 가명처리 단계 – 가명 정보처리 시에도 목적에 부합되면 개인정보를 최대한 제공할 수 있도록 처리해야 하며, 가명처리 방법을 정할 때에는 처리목적, 처리(이용 또는 제공)환경, 정보의 성격 등을 종합적으로 고려한다.
③ 적정성 검토 및 추가처리 – 목적달성을 위해 적절한 수준으로 가명처리가 이루어졌는지, 재식별 가능성은 없는지 등에 대한 최종적인 판단절차를 수행한다.
④ 사후관리 – 적정성 검토 결과 가명처리가 적정하다고 판단되면 가명 정보를 본래 활용 목적을 위해서 처리할 수 있으며, 법령에 따라 기술적·관리적 · 물리적 안전조치를 이행한다.

20 다음 중 l-다양성의 쏠림 공격, 유사성 공격을 보완하기 위한 프라이버시 보호 모델로 동질 집합에서 특정 정보의 분포와 전체 데이터 집합에서 정보의 기준값 이하의 차이를 보여야 하는 모델은?

① k-익명성 ② k-가명성
③ m-유일성 ④ t-근접성

2과목 빅데이터 탐색

21 다음 중 아래와 같은 수학적 정의를 갖는 이론은 무엇인가?

- 사건 A가 조건으로 일어났을 때 사건 B의 확률은 $P(B|A)=\frac{P(A\cap B)}{P(A)}$, $P(A)\neq 0$으로 정의할 수 있다.
- 사건 B가 조건으로 일어났을 때 사건 A의 확률은 $P(A|B)=\frac{P(A\cap B)}{P(B)}$, $P(B)\neq 0$으로 정의할 수 있다.

① 조건부 확률 ② 전 확률의 정리
③ 베이즈 정리 ④ 나이브 베이즈 정리

22 데이터 결측값 처리 방법에서 단순 확률 대치법이란 평균 대치법에서 관측된 자료를 토대로 추정된 통계량으로 결측값을 대치할 때 어떤 적절한 확률값을 부여한 후 대치하는 방법이다. 다음 중 단순 확률 대치법의 유형으로 가장 적절한 것은 무엇인가?

① 평균 대치법
② 핫덱(Hot-Deck) 대체
③ 완전 분석법
④ 다중 대치법

23 다음 중 데이터 이상값 검출 방법이 아닌 것은 무엇인가?

① 시각화
② 다중 대치법
③ 머신러닝 기법
④ 마할라노비스 거리 활용

24 성인 남성의 평균 보폭을 측정하기 위하여 임의로 성인 남성 81명을 추출하여 조사한 결과 평균 보폭은 60cm, 분산은 9cm이었다. 성인 남성의 평균 보폭에 대한 90% 신뢰구간의 하한과 상한은 다음 중 무엇인가? (단, $Z_{0.05}=1.65$, $Z_{0.1}=1.28$로 계산하며 계산 결과는 소수 3번째 자리에서 반올림한다.)

① 58.35, 61.65 ② 59.57, 60.43
③ 58.72, 61.28 ④ 59.45, 60.55

25 다음 중 변수를 변환하는 방법을 설명한 것으로 가장 부적절한 것은 무엇인가?

① 변수의 분포를 변경하기 위해서 로그 변환 기법을 사용한다.
② 기존 데이터를 범주화하기 위해서 비닝(Bining) 기법을 사용한다.
③ 데이터를 특정 구간으로 바꾸는 정규화 기법을 사용한다.
④ 무작위로 정상 데이터의 일부만 선택하는 과소표집 기법을 사용한다.

26 다음이 설명하는 데이터 이상값 발생 원인은 무엇인가?

> 100미터 달리기를 하는데, 한 선수가 '출발' 신호를 못 듣고 늦게 출발했다면 그 선수의 기록은 다른 선수들보다 늦을 것이고, 그의 경기 시간은 이상값이 될 수 있음

① 고의적인 이상값 ② 표본추출 오류
③ 실험 오류 ④ 측정 오류

27 사건 A, B가 있다. x가 발생했을 때, A가 일어날 확률인 $P(A|x)$를 구하는 공식으로 옳은 것은?

① $P(A|x)=\dfrac{P(A)\cdot P(x|A)}{P(A)\cdot P(x|A)+P(B)\cdot P(x|B)}$

② $P(A|x)=\dfrac{P(A)\cdot P(A|x)}{P(A)\cdot P(A|x)+P(B)\cdot P(B|x)}$

③ $P(B|x)=\dfrac{P(x)\cdot P(B|x)}{P(x)\cdot P(A|x)+P(x)\cdot P(B|x)}$

④ $P(A|x)=\dfrac{P(x)\cdot P(x|A)}{P(x)\cdot P(x|A)+P(x)\cdot P(x|B)}$

28 다음 중 데이터를 탐색하기 위한 시각화 기법을 설명한 것으로 가장 부적절한 기법은 무엇인가?

① 자료 분포의 형태를 직사각형 형태로 보여주기 위해 히스토그램을 사용한다.

② 많은 데이터를 그림을 이용하여 집합의 범위와 중앙값을 빠르게 확인할 수 있으며, 또한 통계적으로 이상값이 있는지 빠르게 확인하기 위해 박스 플롯을 사용한다.

③ 데이터값이 큰 지역의 면적을 시각적으로 더 크게 표시하여 데이터를 직관적으로 보기 위해 버블 플롯맵을 사용한다.

④ 가로축과 세로축의 좌표평면상에서 각각의 관찰점들을 표시하여 2개의 연속형 변수 간의 관계를 보기 위하여 산점도를 사용한다.

29 다음 중 박스 플롯을 통해 알 수 없는 것은?

① 이상값 ② 최댓값

③ 중위수 ④ 분산

30 동전 던지기를 했을 때 앞면이 나오면 성공이고, 뒷면이 나오면 실패이다. 동전을 1번 던졌을 때 확률분포의 기댓값과 분산은 얼마인가?

① $E(X)=\dfrac{1}{2},\ V(X)=\dfrac{1}{4}$

② $E(X)=\dfrac{1}{2},\ V(X)=\dfrac{1}{2}$

③ $E(X)=1,\ V(X)=1$

④ $E(X)=1,\ V(X)=2$

31 다음 중 연속확률분포를 설명한 것으로 가장 적절한 것은 무엇인가?

① 정규 분포 함수에서 X를 Z로 정규화한 분포를 카이제곱 분포라고 한다.

② 모집단이 정규 분포이고, 모 표준편차(σ)는 모를 때 Z-분포를 사용한다.

③ 독립적인 χ^2-분포가 있을 때, 두 확률변수의 비는 F-분포이다.

④ 모평균이 μ, 모분산이 σ^2이라고 할 때, 종 모양의 분포는 T-분포이다.

32 데이터의 크기가 커지면 그 데이터가 어떠한 형태이던 그 데이터 표본의 분포는 최종적으로 정규 분포를 따른다는 원칙은 무엇인가?

① 큰 수의 법칙(Law Large Number)

② 중심극한정리(Central Limit Theorem)

③ 체비세프의 정리

④ 마르코프 부등식

33 다음 중 표본의 정보로부터 모집단의 모수를 하나의 값으로 추정하는 점 추정의 조건으로 가장 부적절한 것은 무엇인가?

① 불편성(Unbiasedness)
② 사용성(Usability)
③ 일치성(Consistency)
④ 충족성(Sufficient)

34 가설검정에 대한 설명으로 가장 옳지 않은 것은 무엇인가?

① 대립 가설은 H_0으로 표기하고, 귀무가설은 H_1으로 표기한다.
② 귀무가설은 현재까지 주장되어 온 것이거나 기존과 비교하여 변화 혹은 차이가 없음을 나타내는 가설이다.
③ 대립가설을 연구가설이라고 한다.
④ 표본을 통해 확실한 근거를 가지고 입증하고자 하는 가설은 대립가설이다.

35 다음 중 추정과 가설검정에 대한 설명으로 가장 부적절한 것은?

① 점 추정은 표본의 정보로부터 모집단의 모수를 하나의 값으로 추정하는 것이다.
② 구간 추정은 추정량의 분포에 대한 전제가 주어져야 하고, 구해진 구간 안에 모수가 있을 가능성의 크기(신뢰수준)가 주어져야 한다.
③ 귀무가설이 사실일 때, 관측된 검정 통계량의 값보다 더 대립가설을 지지하는 검정 통계량이 나올 확률을 p-값이라고 한다.
④ 신뢰수준이란 추정값이 존재하는 구간에 모수가 포함될 확률을 의미한다.

36 다음 중 제1종 오류를 설명한 것으로 가장 적절한 것은 무엇인가?

① 귀무가설이 참인데 이를 채택하는 결정
② 귀무가설이 참이 아닌데 이를 채택하지 않는 결정
③ 귀무가설이 참인데 이를 기각하는 결정
④ 귀무가설이 참이 아닌데 이를 채택하는 결정

37 우리나라 고등학생의 영어성적을 추정하려고 한다. 16명의 고등학생을 임의로 조사한 결과 평균이 80점이었다. 우리나라 고등학생의 영어성적의 95% 신뢰구간은 다음 중 무엇인가? (단, 모집단의 분포를 정규분포라고 가정하고 모분산은 16, $Z_{0.025} = 1.96$, $Z_{0.05} = 1.645$이다.)

① $78.36 \leq \mu \leq 81.65$
② $78.04 \leq \mu \leq 81.96$
③ $79.02 \leq \mu \leq 80.98$
④ $79.18 \leq \mu \leq 80.82$

38 다음 중 T-분포를 설명한 것으로 가장 부적절한 것은 무엇인가?

① 정규 분포의 평균(μ)의 해석에 많이 쓰이는 분포이다.
② 모집단이 정규 분포라는 정도만 알고, 모 표준편차(σ)는 모를 때 사용한다.
③ 독립적인 카이제곱 분포가 있을 때, 두 확률변수의 비이다.
④ T-분포에서는 자유도가 표본의 수인 n보다 1 적은 $n-1$이 된다.

39 표본평균의 표준오차에 대한 설명으로 가장 옳지 않은 것은 무엇인가?

① 표준오차는 0 이상의 값을 가진다.
② 표본평균의 표준편차이다.
③ 모집단의 표준편차가 클수록 표본평균의 표준오차는 작아진다.
④ 표본의 크기가 커질수록 표본평균의 표준오차는 작아진다.

40 다음 중 표본추출 기법에 대하여 설명한 것으로 가장 부적절한 것은 무엇인가?

① 단순 무작위 추출: 200개의 구슬에서 무작위로 20개의 구슬을 추출
② 계통 추출: 100명의 교육 참석자에게 이벤트 쿠폰을 나눠주고 자리가 2로 끝나는 사람들을 선정
③ 층화 추출: 연령별 여론 조사를 위해 연령대를 나누고, 각 연령대에서 무작위로 50명씩 선정
④ 군집 추출: 검은색, 흰색, 빨간색 구슬을 무작위로 추출

3과목 빅데이터 모델링

41 시계열 데이터는 관측치가 시간적 순서를 가지며 이러한 데이터를 통해 미래의 값을 예측하는 기법을 시계열 분석이라고 한다. 다음 중 시계열 데이터 분석 기법의 종류가 아닌 것은 무엇인가?

① 분해법
② 지수 평활법
③ ARIMA 모델
④ 응집분석법

42 다음 중 분석 기법에 따른 활용 사례를 설명한 것으로 가장 부적절한 것은 무엇인가?

① 연관규칙학습: 우유를 구매하는 사람이 사과를 더 많이 사는가?
② 의사결정나무: 구매자의 나이가 디지털 가전의 구매 유형에 어떤 영향을 미치는가?
③ 유전자 알고리즘: 물류비 절감을 위해 최소 배송 경로를 구하려면?
④ 분류 분석: 이 사용자는 어떤 특성을 가진 집단에 속하는가?

43 다음 중 분석 모형 구축 절차로 가장 적합한 것은 무엇인가?

① 요건 정의 → 모델링 → 적용 → 검증 및 테스트
② 요건 정의 → 모델링 → 검증 및 테스트 → 적용
③ 모델링 → 검증 및 테스트 → 요건 정의 → 적용
④ 모델링 → 요건 정의 → 적용 → 검증 및 테스트

44 다음 중 모든 고객에게 광고물을 발송하지 않고 특정 고객에게 광고물을 발송하여, 우편 비용을 줄일 때 사용할 수 있는 분석 기법으로 가장 적절한 것은 무엇인가?

① 분류 ② 예측
③ 연관성 분석 ④ 군집

45 CNN에서 원본 이미지가 4*4에서 스트라이드(Stride)가 1이고, 필터가 3*3일 때 Feature Map은 무엇인가?

① (1, 1) ② (2, 2)
③ (3, 3) ④ (4, 4)

46 다음 중에서 RNN(Recurrent Neural Network) 알고리즘에 대한 설명으로 가장 옳지 않은 것은 무엇인가?

① 은닉층에서 재귀적인 신경망을 갖는 알고리즘이다.

② 음성신호, 연속적 시계열 데이터 분석에 적합하다.

③ 확률적 경사 하강법, 시간 기반 오차 역전파를 사용해서 가중치를 업데이트한다.

④ 필터 기능을 이용하여 입력 이미지로부터 특징을 추출한 뒤 신경망에서 분류작업을 수행한다.

47 다음 중에서 지도 학습 유형이 아닌 것은 무엇인가?

① 로지스틱 회귀(Logistic Regression)

② 인공신경망 분석(Artificial Neural Network)

③ 서포트 벡터 머신(Support Vector Machine)

④ 자기 조직화 지도(Self-Organizing Map)

48 다음 중 의사결정나무의 구성요소를 설명한 것으로 옳지 않은 것은 무엇인가?

① 뿌리 마디(Root Node)는 시작되는 마디로 전체 자료를 포함한다.

② 가지(Branch)는 뿌리 마디로부터 끝마디까지 연결된 상태의 마디들이다.

③ 깊이(Depth)는 뿌리 마디부터 끝마디까지의 부모 마디들의 수이다.

④ 자식 마디(Child Node)는 하나의 마디로부터 분리되어 나간 2개 이상의 마디들이다.

49 다음 중에서 초매개변수는 무엇인가?

① 인공신경망에서의 가중치

② KNN에서의 K의 개수

③ 서포트 벡터 머신에서의 서포트 벡터

④ 로지스틱 회귀 분석에서의 결정계수

50 다음 중에서 편향(Bias)을 발생시키지는 않으나 과대적합을 발생시켜 예측 성능을 저하시키는 부적합 모형 현상은 무엇인가?

① 모형 선택 오류 ② 변수 누락

③ 부적합 변수 생성 ④ 동시 편향

51 인공신경망은 입력값을 받아서 출력값을 만들기 위해 활성화 함수를 사용한다. 다음 중 인공신경망의 활성화 함수를 설명한 것으로 가장 부적절한 것은 무엇인가?

① 활성화 함수는 순 입력함수로부터 전달받은 값을 출력값으로 변환해 주는 함수이다.

② 활성화 함수에는 계단함수, 부호함수, 선형함수, 시그모이드 함수, tanh 함수, ReLU 함수가 있다.

③ 인공신경망은 입력값을 받아서 출력값을 만들기 위해 활성화 함수를 사용한다.

④ ReLU는 x값이 증가하면 y값도 지속적으로 증가한다.

52 다음 중에서 회귀 모형의 가정이 아닌 것은 무엇인가?

① 상관성 ② 선형성

③ 독립성 ④ 정상성

53 서포트 벡터 머신(SVM; Support Vector Machine)은 기계학습의 한 분야로 사물 인식, 패턴 인식, 손 글씨 숫자 인식 등 다양한 분야에서 활용되고 있는 지도 학습 모델이다. 다음 중 서포트 벡터 머신을 설명한 것으로 가장 적절하지 않은 것은 무엇인가?

① 서포트 벡터(Support Vector)는 훈련 데이터 중에서 결정 경계와 가장 가까이에 있는 데이터들의 집합이다.
② 결정 경계(Decision Boundary)는 데이터 분류의 기준이 되는 경계이다.
③ SVM은 훈련 시간이 상대적으로 빠르고 다른 방법보다 과대 적합의 가능성이 높은 모델이다.
④ SVM은 공간상에서 최적의 분리 초평면(Hyperplane)을 찾아서 분류 및 회귀를 수행한다.

54 연관성 분석으로 어떤 상품을 고객에게 판매해야 하는지를 예측하려고 한다. 다음은 고객의 영수증 데이터를 확인하여 집계한 결과이다. 사과를 구매한 고객 중에서 우유를 구매한 고객의 신뢰도는 얼마인가?

판매품목	판매 건수
사과	4,000
우유	2,000
사과, 우유 동시 구매	1,000
커피	3,000
전체 거래량	10,000

① 10%　② 20%
③ 25%　④ 30%

55 다음 중 시계열 분석에서 시점에 상관없이 시계열의 특성이 일정하다는 것은 무엇인가?

① 신속성　② 적합성
③ 유의성　④ 정상성

56 다음 중에서 로지스틱 회귀 분석에 대한 설명으로 가장 바르지 않은 것은?

① 로지스틱 회귀 분석은 반응변수가 범주형인 경우 적용되는 회귀 분석 모형이다.
② 모형의 적합을 통해 추정된 확률을 사후 확률(Posterior Probability)로도 부른다.
③ 로지스틱 회귀 분석은 분류하는 목적으로 사용될 수 있다.
④ 독립변수가 한 개인 경우 회귀계수의 부호와 상관 없이 그래프의 형태는 S자 모양을 가진다.

57 다음이 설명하는 딥러닝 알고리즘은 무엇인가?

> • 시각적 이미지를 분석하는 데 사용되는 심층신경망으로 합성곱 신경망이라고도 한다.
> • 기존 영상처리의 필터 기능과 신경망을 결합하여 성능을 발휘하도록 만든 구조이다.

① DNN　② CNN
③ RNN　④ GAN

58 다음 중 비정형 데이터 분석 기법에 대한 설명으로 가장 부적절한 것은 무엇인가?

① 사회 연결망 분석(SNA)은 그룹에 속한 사람들 간의 네트워크 특성과 구조를 분석하고 시각화하는 분석 기법이다.
② 오피니언 마이닝(Opinion Mining)은 비정형화된 로그 데이터에서 정보를 수집하는 기법이다.
③ 웹 마이닝(Web Mining)은 웹에서 발생하는 고객의 행위 분석과 특성 데이터를 추출, 정제하여 의사결정에 활용하기 위한 기법이다.
④ 텍스트 마이닝(Text Mining)은 텍스트 형태로 이루어진 비정형 데이터들을 자연어 처리방식을 이용해 정보를 추출하는 기법이다.

59 다음 앙상블 기법 중 훈련 데이터에서 다수의 부트스트랩(Bootstrap) 자료를 생성하고, 각 자료를 모델링한 후 결합하여 최종 예측 모형을 만드는 것은 무엇인가?

① 배깅(Bagging)
② 부스팅(Boosting)
③ 랜덤 포레스트(Random Forest)
④ 스태킹(Stacking)

60 다음 중 비모수 통계에 대한 설명으로 가장 부적절한 것은 무엇인가?

① 비모수 통계분석에서는 빈도(Frequency), 부호(Sign), 순위(Rank) 등의 통계량을 사용한다.
② 평균이나 분산 같은 모집단의 분포에 대한 모수성을 가정하지 않고 분석하는 통계적 방법이다.
③ 표본의 크기가 커질수록 간편하고 반복적인 계산이 없어서 효율적이다.
④ 모집단의 분포에 무관하게 사용할 수 있다.

4과목 빅데이터 결과 해석

61 다음 혼동 행렬의 평가지표 중에서 실제로 '부정'인 범주 중에서 '부정'으로 올바르게 예측(TN)한 비율은 무엇인가?

① 특이도(Specificity)
② 정확도(Accuracy)
③ 민감도(Sensitivity)
④ 정밀도(Precision)

62 1학년부터 3학년까지 3개의 학년 50명을 대상으로 각 학년별 국어 평균을 구하여 일원 배치 분산 분석을 하려고 한다. 다음의 일원 분산 분석표의 ㉠과 ㉡에 들어갈 값은 얼마인가?

요인	제곱합	자유도	제곱평균	F
집단 간	SSR	㉠	MSR	MSR/MSE
집단 내	SSE	㉡	MSE	
총	SST	49		

① 3, 47
② 2, 47
③ 2, 48
④ 3, 46

63 교차 검증은 모델의 일반화 오차에 대해 신뢰한 추정치를 구하기 위해 훈련, 평가 데이터를 기반으로 하는 검증 기법이다. 다음 중 홀드 아웃 교차 검증을 설명한 것으로 가장 적합한 것은 무엇인가?

① 전체 데이터에서 평가 데이터를 학습에도 사용하므로 데이터 손실이 발생하지 않는다.
② 전체 데이터를 비복원추출 방식을 이용하여 랜덤하게 훈련 데이터(Training Set)와 평가 데이터(Test Set)로 나눠 검증하는 기법이다.
③ 검증 데이터(Validation Set)는 최적화된 분류기의 성능을 평가할 때 사용하는 데이터이다.
④ 데이터 집합을 무작위로 동일 크기를 갖는 K개의 부분 집합으로 나누고, 그중 1개를 평가 데이터(Test Set)로, 나머지 (K-1)개를 훈련 데이터(Training Set)로 선정하여 분석 모형을 평가하는 기법이다.

64 다음 중 두 개 이상의 집단 간 비교를 수행하고자 할 때 집단 내의 분산, 총 평균과 각 집단의 평균 차이에 의해 생긴 집단 간 분산 비교로 얻은 F-분포를 이용하여 가설검정을 수행하는 방법은 무엇인가?

① Z-검정 ② T-검정
③ 분산 분석 ④ 카이제곱 검정

65 다음 중 적합도 검정 방법에서 카이제곱 검정에 대한 설명으로 가장 부적절한 것은 무엇인가?

① 관측된 데이터가 가정된(알려진) 확률을 따르는지 확인한다.
② 가정된 확률이 정해져 있을 경우에 사용하는 검정 방법이다.
③ 데이터가 가정된 확률을 따르는 경우 대립 가설(H_1)을 채택한다.
④ R 언어의 chisq.test() 함수를 이용하여 검정이 가능하다.

66 다음 중 과대 적합(Over-Fitting)을 방지하기 위한 드롭아웃을 설명한 것으로 가장 부적절한 것은 무엇인가?

① 개별 가중치 값을 제한하여 복잡한 모델을 좀 더 간단하게 하는 방법이다.
② 드롭아웃은 신경망 학습 시에만 사용하고, 예측 시에는 사용하지 않는다.
③ 학습 시에 인공신경망이 특정 뉴런 또는 특정 조합에 너무 의존적이게 되는 것을 방지해준다.
④ 드롭아웃은 학습 과정에서 신경망의 일부를 사용하지 않는 방법이다.

67 다음 중 주어진 데이터로부터 학습을 통해 모델 내부에서 결정되는 매개변수를 최적화하는 기법으로 가장 부적절한 것은 무엇인가?

① 확률적 경사 하강법
② 모멘텀
③ AdaGrad
④ 드롭아웃

68 다음 중 데이터 시각화 기법을 설명한 것으로 가장 적합한 것은 무엇인가?

① 시간의 변화에 따른 경향(트렌드)을 파악하기 위해 산점도를 사용한다.
② 집단 간의 상관관계를 확인하여 다른 수치의 변화를 예측하기 위해 히트맵을 사용한다.
③ 지도를 통해 시점에 따른 경향, 차이 등을 확인하기 위해 카토그램을 사용한다.
④ 전체에서 부분 간 관계를 설명하기 위해 막대그래프를 사용한다.

69 다음이 설명하는 관계 시각화 유형으로 가장 올바른 것은?

> • x축과 y축 각각에 두 변숫값의 순서쌍을 한 점으로 표시하여 변수의 관계를 나타낸 그래프
> • 상관관계, 군집화, 이상값 패턴을 파악하기에 유용한 그래프

① 산점도(Scatter Plot)
② 버블 차트(Bubble Chart)
③ 히스토그램(Histogram)
④ 히트맵(Heat Map)

70 다음 혼동 행렬(Confusion Matrix)에서 참이 0이고 거짓이 1일 때, 특이도(Specificity)와 정밀도(Precision)는 무엇인가?

		실제 레코드		
		0	1	총합
예측 실적	0	25	15	40
	1	30	70	100
총합		55	85	140

① 특이도: 7/10, 정밀도: 5/11
② 특이도: 5/7, 정밀도: 5/11
③ 특이도: 7/10, 정밀도: 5/9
④ 특이도: 5/7, 정밀도: 5/9

71 다음 중 공간 시각화에서 사용되는 기법으로 가장 부적합한 것은 무엇인가?

① 카토그램
② 버블 플롯맵
③ 도트맵
④ 히스토그램

72 다음 중 인포그래픽 유형의 예시를 설명한 것으로 가장 부적절한 것은 무엇인가?

① 매장분포 – 지도형
② 유명인사 – 스토리텔링형
③ 심리정보 – 만화형
④ 주요 제품 비교 – 타임라인형

73 다음 중 ROC 곡선에 대한 설명으로 가장 옳지 않은 것은 무엇인가

① AUC(Area Under ROC; AUROC)는 ROC 곡선 아래의 면적으로 면적을 모형의 평가지표로 삼는다.
② ROC 곡선은 가로축을 특이도(Specificity), 세로축을 참 긍정률(TP Rate)로 두어 시각화한 그래프이다.
③ AUC의 값은 항상 0.5~1의 값을 가지며, 1에 가까울수록 좋은 모형이다.
④ 거짓 긍정률(FP Rate)과 참 긍정률(TP Rate)은 어느 정도 비례 관계에 있다.

74 다음 중 버블 차트(Bubble Chart)의 시각화 유형과 같은 것은 무엇인가?

① 체르노프 페이스(Chernoff Faces)
② 산점도(Scatter Plot)
③ 트리맵(Tree Map)
④ 히트맵(Heat Map)

75 다음 중 빅데이터 모형 모니터링을 수행하는 방법으로 가장 부적절한 것은 무엇인가?

① 성능 모니터링을 위한 주요 성능 측정 항목을 정의한다.
② 개발된 모델의 성능을 확인하기 위해 모니터링을 수작업으로 수행한다.
③ 이벤트 등급별로 알람을 통해 이벤트 모니터링에서 성능을 관리하도록 한다.
④ 측정 항목별 임계치를 설정한다.

76 다음 중 빅데이터 활용 분야를 검토할 때 아이디어 개발 관점의 분류로 가장 적합하지 않은 것은 무엇인가?

① 마인드맵 방식
② 친화 도표 방식
③ 피라미드 방식
④ 버블차트 방식

77 다음 중 여러 가지 변수를 비교할 수 있는 시각화 그래프로 칸 별로 색상을 구분하여 데이터값을 표현한 비교 시각화 기법은 무엇인가?

① 체르노프 페이스
② 히트맵
③ 스타 차트
④ 평행 좌표 그래프

78 다음 중 지도상의 위도와 경도에 해당하는 좌표점에 산점도와 같이 점을 찍어서 표현하고, 시간의 경과에 따라 점진적으로 확산을 나타내는 경우에 사용하는 공간 시각화 유형은 무엇인가?

① 등치지역도 ② 등치선도
③ 도트 플롯맵 ④ 버블 플롯맵

79 다음 중 부트스트랩을 설명한 것으로 가장 부적합한 것은 무엇인가?

① 무작위로 복원추출하는 기법이다.
② 전체 데이터에서 중복을 허용한다.
③ 데이터 크기만큼 샘플을 추출하고 이를 훈련 데이터(Training Set)로 한다.
④ 샘플에 한 번도 선택되지 않는 원 데이터가 발생할 수 없다.

80 다음 중 혼동 행렬을 설명한 것으로 가장 부적합한 것은 무엇인가?

① 정확도(Accuracy) – 전체 예측에서 참 긍정(TP)과 참 부정(TN)이 차지하는 비율
② 민감도(Sensitivity) – 실제로 '긍정'인 범주 중에서 '긍정'으로 올바르게 예측(TP)한 비율
③ 정밀도(Precision) – 실제로 '부정'인 범주 중에서 '부정'으로 올바르게 예측(TN)한 비율
④ 거짓 긍정률(FP Rate) – 실제로 '부정'인 범주 중에서 '긍정'으로 잘못 예측(FP)한 비율

1과목 빅데이터 분석 기획

01 다음 중 데이터에 양을 측정하는 바이트의 크기로 잘못 연결된 것은 무엇인가?

① 킬로바이트(KB): 10^3Bytes
② 페타바이트(PB): 10^{15}Bytes
③ 테라바이트(TB): 10^{12}Bytes
④ 메가바이트(MB): 10^9Bytes

02 다음이 설명하는 개념으로 적합한 용어는 무엇인가?

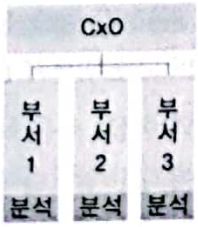
> 개인이 정보 관리의 주체가 되어 능동적으로 본인의 정보를 관리하고, 본인의 의지에 따라 신용 및 자산관리 등에 정보를 활용하는 일련의 과정이다.

① YourData ② MyData
③ OurData ④ ItsData

03 다음 중 데이터 거버넌스의 구성요소로 올바르지 않은 것은?

① 원칙 ② 플랫폼
③ 조직 ④ 프로세스

04 다음 중 다음 그림과 같은 빅데이터 조직 구조로 가장 적절한 것은 무엇인가?



① 집중 구조 ② 기능 구조
③ 분산 구조 ④ 협업 구조

05 다음이 설명하는 개인정보 비식별화 절차는 어떤 단계에 속하는가?

> 비식별 정보 안전조치, 재식별 가능성 모니터링 등 비식별 정보 활용 과정에서 재식별 방지를 위해 필요한 조치 수행

① 사전검토 ② 비식별 조치
③ 적정성 평가 ④ 사후관리

06 다음이 설명하는 CRISP-DM 분석 방법론의 절차는 무엇인가?

> • 분석을 위한 데이터를 수집 및 속성을 이해하고, 문제점을 식별하며 숨겨져 있는 인사이트를 발견하는 단계
> • 초기 데이터 수집, 데이터 기술 분석, 데이터 탐색, 데이터 품질 확인

① 업무 이해 (Business Understanding)
② 데이터 이해 (Data Understanding)
③ 데이터 준비(Data Preparation)
④ 모델링(Modeling)

07 다음 중 아래에서 설명하는 프라이버시 보호 모델은 무엇인가?

> • 동질성 공격, 배경 지식에 의한 공격을 방어하기 위한 프라이버시 모델이다.
> • 주어진 데이터 집합에서 함께 비식별 되는 레코드들은 (동질 집합에서) 적어도 몇 개의 서로 다른 민감한 정보를 가져와야 하는 프라이버시 모델이다.
> • 비식별 조치 과정에서 충분히 다양한 서로 다른 민감한 정보를 갖도록 동질 집합을 구성한다.

① k-익명성 ② l-다양성
③ m-유일성 ④ t-근접성

08 다음 중 개인정보 비식별 조치 방법으로 가장 적절한 것은 무엇인가?

조치 전	주민등록번호 901212-1234567
조치 후	90년대 생, 남자

① 가명처리 ② 총계처리
③ 데이터 삭제 ④ 데이터 마스킹

09 다음 중 분석 로드맵 단계로 가장 적절한 것은 무엇인가?

① 데이터 분석체계 도입 → 데이터 분석 유효성 검증 → 데이터 분석 확산 및 고도화
② 데이터 분석체계 도입 → 데이터 분석 확산 및 고도화 → 데이터 분석 유효성 검증
③ 데이터 분석 유효성 검증 → 데이터 분석 확산 및 고도화 → 데이터 분석체계 도입
④ 데이터 분석 유효성 검증 → 데이터 분석체계 도입 → 데이터 분석 확산 및 고도화

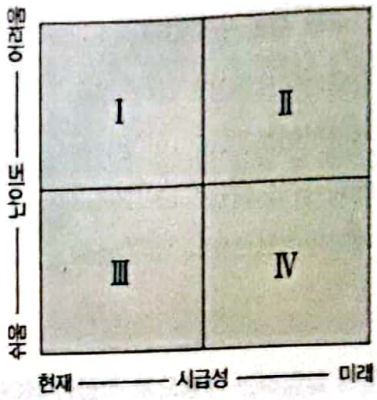
10 다음 중 분석의 대상이 명확하게 무엇인지 모르는 경우 기존 분석 방식을 활용하여 새로운 지식을 도출하는 것으로 가장 적절한 것은 무엇인가?

① 최적화(Optimization)
② 솔루션(Solution)
③ 통찰(Insight)
④ 발견(Discovery)

11 빅데이터의 3V(Volume, Variety, Velocity)의 특징에 해당하지 않는 것은?

① 가치(Value)
② 다양성(Variety)
③ 속도(Velocity)
④ 규모(Volume)

12 다음 중 분석과제 우선순위 선정 매트릭스에서 분석과제의 적용 우선순위를 "난이도"에 둘 때 가장 올바른 우선순위는?



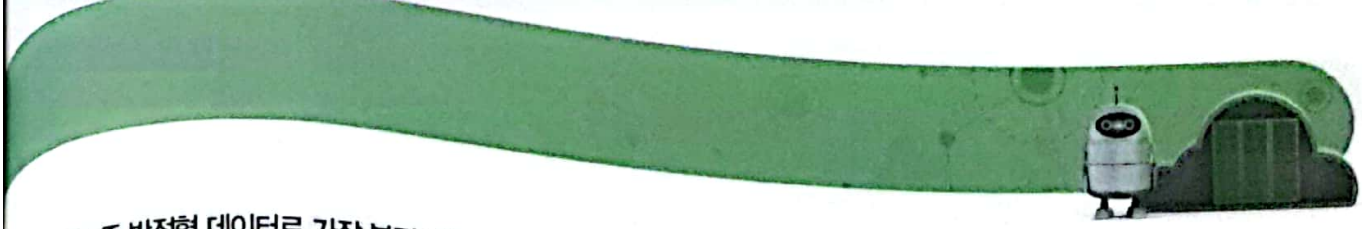
① III → I → II ② III → IV → II
③ III → II → IV ④ III → II → I

13 다음 중 CRISP-DM 분석 방법론 단계로 올바른 것은?

① 데이터 이해 → 데이터 준비 → 업무 이해 → 모델링 → 평가 → 전개
② 업무 이해 → 데이터 이해 → 데이터 준비 → 모델링 → 평가 → 전개
③ 데이터 이해 → 데이터 준비 → 업무 이해 → 모델링 → 전개 → 평가
④ 업무 이해 → 데이터 이해 → 데이터 준비 → 모델링 → 전개 → 평가

14 개인정보를 제공하기 위해 정보주체의 동의를 받을 때 고지 사항으로 올바르지 않은 것은?

① 개인정보의 수집 · 이용 목적
② 개인정보에 대한 암호화 여부 및 안전성 확보 조치 여부
③ 개인정보를 제공받는 자
④ 수집하려는 개인정보의 항목



15 다음 중 반정형 데이터로 가장 부적절한 것은 무엇인가?

① XML
② JSON
③ HTML
④ 오디오

16 다음 중 데이터로부터 잡음을 제거하기 위해 데이터 추세에 벗어나는 값들을 변환하는 기법으로 구간화, 군집화 등의 기법을 적용하는 데이터 변환 기술은 무엇인가?

① 정규화
② 평활화
③ 집계
④ 일반화

17 다음 중 빅데이터 위기 요인의 통제 방안에 대한 설명으로 가장 옳지 않은 것은?

① 개인정보 유출 및 사용으로 발생하는 피해에 대해 사용자가 책임을 지게 한다.
② 예측 알고리즘을 통해 범죄를 일으킬 가능성이 있는 사람에 대하여 사전에 구속, 접근 금지 등의 조치를 취한다.
③ 예측 알고리즘의 부당함을 반증할 수 있도록 알고리즘에 대한 접근권을 제공한다.
④ 알고리즈미스트를 통하여 불이익을 당한 사람들을 대변한다.

18 다음 중에서 암묵지와 형식지 간의 4단계 지식 전환 단계를 순서대로 가장 바르게 나타낸 것은 무엇인가?

① 공통화 → 표출화 → 연결화 → 내면화
② 내면화 → 연결화 → 공통화 → 표출화
③ 표출화 → 공통화 → 연결화 → 내면화
④ 연결화 → 표출화 → 공통화 → 내면화

19 상향식 접근 방식(Bottom Up Approach)으로서 시행 착오를 통한 문제 해결을 위해 사용되며 가설의 생성(Hypotheses), 디자인에 대한 실험(Design Experiments), 실제 환경에서의 테스트(Test), 테스트 결과에서의 통찰(Insight) 도출 및 가설 확인의 프로세스로 구성되는 접근법은 다음 중 무엇인가?

① 프로토타이핑(Prototyping)
② 최적화(Optimization)
③ 디자인 사고(Design Thinking)
④ 지도 학습(Supervised Learning)

20 개인정보가 유출되었음을 알게 되었을 때에 개인정보처리자가 지체 없이 해당 정보주체에게 알려야 하는 사실로 올바르지 않은 것은?

① 유출된 개인정보의 항목
② 유출로 인하여 발생할 수 있는 피해를 최소화하기 위하여 정보주체가 할 수 있는 방법 등에 관한 정보
③ 개인정보처리자의 대응조치 및 피해 구제절차
④ 개인정보 유출 사고의 실시간 수사 진행 상황

2과목 빅데이터 탐색

21 다음 중 실시간으로 발생하는 이벤트 처리에 대한 결괏값을 수집하고 처리하는 기술은 무엇인가?

① CEP
② 맵리듀스
③ ETL
④ 피그

22 다음 중 불완전 자료는 모두 무시하고 완전하게 관측된 자료만 사용하여 분석하는 방법으로 가장 적절한 것은 무엇인가?

① 평균 대치법
② 핫덱(Hot-Deck) 대체
③ 완전 분석법
④ 다중 대치법

23 다음 중 데이터값을 몇 개의 버킷으로 분할하여 계산하는 방법은 무엇인가?

① 단순 기능 변환
② 정규화
③ 로그 변환
④ 비닝

24 다음 중 특정 모델링 기법에 의존하지 않고 데이터의 통계적 특성으로부터 변수를 택하는 기법으로 적절한 것은 무엇인가?

① 필터 기법
② 임베디드 기법
③ 라쏘
④ 릿지

25 다음 중 변수들의 공분산 행렬이나 상관행렬을 이용하고 원래 데이터 특징을 잘 설명해주는 성분을 추출하기 위하여 고차원 공간의 표본들을 선형 연관성이 없는 저차원 공간으로 변환하는 기법으로 가장 적절한 것은?

① 주성분 분석
② 특이값 분해
③ 상관 분석
④ 다차원 척도법

26 다수 클래스에 밀집된 데이터가 없을 때까지 데이터를 제거하여 데이터 분포에서 대표적인 데이터만 남도록 하는 방법으로 가장 적절한 것은?

① 랜덤 과소 표집(Random Under-Sampling)
② 토멕 링크 방법(Tomek Link Method)
③ CNN(Condensed Nearest Neighbor)
④ ENN(Edited Nearest Neighbours)

27 포아송 분포에서 사건 발생 확률이 λ이고, 사건이 일어나는 횟수를 n이라고 할 때, 기댓값과 분산은 얼마인가?

① 기댓값: λ, 분산: λ
② 기댓값: $1/\lambda$, 분산: np
③ 기댓값: λ, 분산: np
④ 기댓값: $1/\lambda$, 분산: λ

28 다음 중 다수 클래스의 데이터를 일부만 선택하여 데이터의 비율을 맞추는 방법은 무엇인가?

① 앙상블 기법
② 과대 표집
③ 과소 표집
④ 임곗값 이동

29 다음 중 변수의 속성에 따른 분석 방법에 대한 설명 중 올바르지 않은 것은?

① 수치로 표현을 할 수 있는 측정 가능한 데이터 변수는 피어슨(Pearson) 상관계수를 통해서 분석한다.
② 데이터의 순서에 의미를 부여한 데이터 변수는 T-분포를 통해서 분석한다.
③ 명목적 데이터는 카이제곱 검정을 통해서 분석한다.
④ 데이터에 대한 분류의 의미를 지닌 명목적 데이터 변수 사이의 상관계수를 계산하는 것은 큰 의미가 없다.

30 확률 변수 X와 확률 질량 함수 $P(X)$가 다음과 같이 주어질 때, 확률 변수 X의 기댓값은 얼마인가?

X	1	2	3	4
$P(X)$	1/6	2/6	1/6	2/6

① 1
② $\frac{4}{3}$
③ $\frac{5}{3}$
④ $\frac{8}{3}$

31 다음 데이터 중 최빈수으로 가장 적절한 것은 무엇인가?

3, 5, 6, 4, 5

① 3
② 4
③ 5
④ 6

32 컴퓨터를 소유하고 있는 집단 A는 전체 학생의 80%이고, 그중에 50%가 League Of Legend를 플레이해본 적이 있다. 컴퓨터를 소유하고 있지 않은 집단 B는 전체 학생의 20%이고, 그중에 20% 학생이 League Of Legend를 플레이해본 적이 있다. League Of Legend를 플레이해본 학생을 임의로 선택했을 때 학생이 컴퓨터를 소유하지 않는 집단 B에 속할 확률은 얼마인가?

① $\frac{1}{3}$
② $\frac{1}{5}$
③ $\frac{1}{11}$
④ $\frac{3}{4}$

33 크기가 100인 표본으로 95% 신뢰수준을 가지도록 모평균을 추정하였는데, 신뢰구간의 길이가 20이었다. 동일한 조건에서 크기가 400인 표본으로 95% 신뢰수준을 가지도록 모평균을 추정할 경우에 표본의 길이는 얼마인가?

① 10 ② 20
③ 40 ④ 80

34 다음 일변량 데이터 탐색 방법으로 가장 알맞은 것은 무엇인가?

① 기술 통계량 ② 산점도 행렬
③ 별 그림 ④ 등고선 그림

35 다음 중 확률변수의 분산에 대한 성질로서 바르지 않은 것은? (단, X, Y는 확률변수이고 서로 독립이며, a는 상수이다.)

① $V(X-Y)=V(X)-V(Y)$
② $V(a)=0$
③ $V(aX)=a^2V(X)$
④ $V(X+Y)=V(X)+V(Y)$

36 다음 중 제1종 오류를 범할 최대 허용확률을 의미하는 용어는 무엇인가?

① 신뢰수준(Level of Confidence)
② 유의수준(Level of Significance)
③ 베타 수준(β Level)
④ 검정력

37 평균이 10이고 분산이 25인 정규 모집단에서 크기가 9인 표본을 추출하였을 경우 표본평균의 표준편차는 얼마인가?

① $\frac{5}{2}$ ② $\frac{5}{3}$
③ $\frac{25}{9}$ ④ $\frac{10}{9}$

38 다음 중 모집단이 정규 분포라는 정도만 알고, 모 표준편차(σ)는 모를 때 사용하는 분포로 가장 알맞은 것은 무엇인가?

① 정규 분포 ② F-분포
③ T-분포 ④ χ^2-분포

39 동전을 세 번 던졌을 때 앞면이 한 번 나올 확률은 얼마인가?

① 0.125 ② 0.375
③ 0.5 ④ 0.625

40 고등학교 남학생의 키를 추정하기 위하여 100명을 임의로 선택하여 평균 키를 측정하였더니 175cm, 분산은 25였다. 고등학교 남학생의 평균 키에 대한 95% 신뢰구간은 다음 중 무엇인가? (단, $Z_{0.025}=1.96$, $Z_{0.05}=1.645$)

① $174.18 \le \mu \le 175.82$
② $174.02 \le \mu \le 175.98$
③ $173.36 \le \mu \le 176.65$
④ $173.04 \le \mu \le 176.96$

3과목 빅데이터 모델링

41 다음 중에서 의사결정나무의 알고리즘에 대한 설명으로 가장 옳지 않은 것은?

① CART는 목적변수가 이산형일 경우에 불순도의 측도로 엔트로피 지수를 이용한다.
② C4.5와 C5.0은 각 마디에서 다지 분리(Multiple Split)가 가능하다.
③ CHAID에서는 불순도의 측도로 카이제곱 통계량을 이용한다.
④ QUEST에서 분리규칙은 분리변수 선택과 분리점 선택의 두 단계로 나누어 시행한다.

42 다음 중 매개변수(Parameter)의 예시로 가장 알맞지 않은 것은?

① 신경망 학습에서 학습률(Learning Rate)
② 인공신경망에서의 가중치
③ 서포트 벡터 머신에서의 서포트 벡터
④ 선형 회귀나 로지스틱 회귀 분석에서의 결정계수

43 다음 중 분석 모형 구축 절차 중 요건 정의에 따라 상세 분석 기법을 적용해 모델을 개발하는 과정으로 가장 적합한 것은 무엇인가?

① 요건 정의
② 모델링
③ 적용
④ 검증 및 테스트

44 다음 중 R과 거의 같은 작업 수행이 가능한 C언어 기반의 오픈 소스 프로그래밍 언어는 무엇인가?

① R-Studio
② S-PLUS
③ 파이썬
④ JAVA

45 다음 중 데이터 분할에 대한 설명으로 가장 알맞지 않은 것은?

① 데이터 분할은 데이터를 훈련 데이터, 검증 데이터, 평가 데이터로 분할하는 작업이다.
② 데이터 분할을 하는 이유는 모형이 주어진 데이터에 대해서만 높은 성능을 보이는 과대 적합의 문제를 예방하여 1종 오류인 잘못된 귀무가설을 채택하는 오류를 방지하는 데 목적이 있다.
③ 훈련 데이터와 검증 데이터는 학습 과정에서 사용하며 평가 데이터는 학습 과정에 사용되지 않고 오로지 모형의 평가를 위한 과정에만 사용된다.
④ 검증 데이터를 사용하여 모형의 학습 과정에서 모형이 제대로 학습되었는지 중간에 검증을 실시하고, 과대 적합과 과소 적합의 발생 여부 등을 확인하여 모형의 튜닝에도 사용한다.

46 다음 중 다중 회귀 모형의 수식으로 가장 알맞은 것은?

① $Y = \beta_0 + \beta_1 X + \epsilon$
② $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$
③ $Y = \alpha e^{-\beta X} + \epsilon$
④ $Y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \epsilon$

47 다음 중 로지스틱 회귀 분석에 대한 설명으로 가장 알맞지 않은 것은?

① 반응변수가 범주형인 경우 적용되는 회귀 분석 모형이다.

② 새로운 설명변수의 값이 주어질 때 반응변수의 각 범주에 속할 확률이 얼마인지를 추정하여 추적 확률을 기준치에 따라 분류하는 목적으로 사용될 수 있다.

③ 승산비는 $\frac{1-p}{p}$로 계산한다.

④ R을 사용하여 로지스틱 회귀 분석을 수행하고 결과를 해석할 수 있다.

48 다음 중 의사결정나무의 분석 과정으로 가장 알맞은 것은 무엇인가?

① 의사결정나무 성장 → 가지치기 → 타당성 평가 → 해석 및 예측

② 의사결정나무 성장 → 타당성 평가 → 가지치기 → 해석 및 예측

③ 타당성 평가 → 가지치기 → 의사결정나무 성장 → 해석 및 예측

④ 타당성 평가 → 의사결정나무 성장 → 가지치기 → 해석 및 예측

49 다음 중 시간이 지날수록 관측치의 평균값이 지속적으로 증가하거나 감소하는 시계열 모형으로 가장 알맞은 것은?

① 자기 회귀 모형

② 이동평균 모형

③ 백색잡음

④ 분해 시계열

50 다음 중 서포트 벡터 머신의 구성요소로 가장 알맞지 않은 것은?

① 초평면 ② 활성화 함수

③ 서포트 벡터 ④ 슬랙 변수

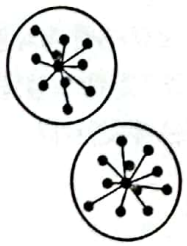
51 다음은 수제비 쇼핑몰의 거래내역이다. 연관 규칙 "커피 → 빵"에 대한 지지도(Support)는 얼마인가?

항목	거래 수
커피	10
빵	20
커피, 빵 동시 구매	50
기타	20
전체 거래 수	100

① $\frac{3}{5}$ ② $\frac{1}{3}$

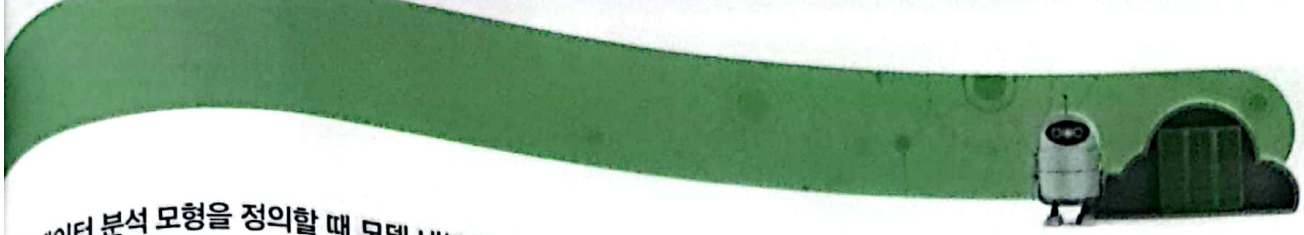
③ $\frac{1}{2}$ ④ $\frac{2}{5}$

52 다음 중 아래 그림과 같이 군집 내의 오차 제곱합(Error Sum of Square)에 기초하여 군집을 수행하는 기법은 무엇인가?



① 최장연결법 ② 중심연결법

③ 평균연결법 ④ 와드연결법



53 데이터 분석 모형을 정의할 때 모델 내부에서 확인이 가능한 변수로 데이터를 통해서 산출이 가능한 값은 무엇인가?

① 매개변수(Parameter)
② 편향(Bias)
③ 신경망 학습에서 학습률(Learning Rate)
④ KNN에서의 K의 개수

54 다음이 설명하는 의사결정나무 분석 과정 단계는 무엇인가?

> 분석 목적과 자료구조에 따라 적절한 분리 규칙(Splitting Rule) 및 정지 규칙(Stopping Rule)을 지정함

① 의사결정나무 성장(Growing)
② 가지치기(Pruning)
③ 해석 및 예측
④ 타당성 평가

55 다음 중 입력층과 출력층으로만 구성된 최초의 인공 신경망은 무엇인가?

① 퍼셉트론(Perceptron)
② 순방향 신경망(Feed Forward Neural Network)
③ 합성곱 신경망(Convolutional Neural Network)
④ 다층 퍼셉트론(Multi-Layer Perceptrons; MLP)

56 다음 중 활성화 함수로 옳지 않은 것은?

① 계단함수
② 시그모이드 함수
③ ReLU 함수
④ 우도 함수

57 다음에서 설명하는 변수 거리의 측정 방법은 무엇인가?

> • 명목형 변수의 거리 측정 방법
> • 두 집합 사이의 유사도를 측정하는 방법
> • 0과 1 사이의 값을 가지며 두 집합이 동일하면 1의 값, 공통의 원소가 하나도 없으면 0의 값을 가짐

① 단순 일치계수
② 자카드 계수
③ 순위 상관계수
④ 맨하튼 거리

58 다음 중 텍스트 형태로 이루어진 비정형 데이터들을 자연어 처리 방식을 이용해 정보를 추출하는 기법으로 가장 알맞은 것은 무엇인가?

① 배깅(Bagging)
② 서포트 백터 머신(SVM)
③ ADASYN
④ 텍스트 마이닝(Text Mining)

59 다음 중 서포트 벡터 머신(SVM)에 대한 설명으로 옳지 않은 것은?

① 서포트 벡터 머신에서 서포트 벡터는 여러 개 일 수도 있다.
② 서포트 벡터 머신은 기계학습의 한 분야로 사물 인식, 패턴 인식, 손 글씨 숫자 인식 등 다양한 분야에서 활용되고 있는 지도 학습 모델이다.
③ 최대 마진(Margin; 여유 공간)을 가지는 비확률적 선형 판별에 기초한 이진 분류기이다.
④ SVM은 훈련 시간이 상대적으로 빠르지만, 정확성이 낮고 다른 방법보다 과대 적합의 가능성이 높은 모델이다.

60 다음 중 잘못 분류된 개체들에 가중치를 적용, 새로운 분류 규칙을 만들고, 이 과정을 반복해 최종 모형을 만드는 알고리즘으로 가장 알맞은 것은?

① 배깅(Bagging)
② 보팅(Voting)
③ 랜덤 포레스트(Random Forest)
④ 부스팅(Boosting)

4과목 빅데이터 결과 해석

61 다음 예측값과 실젯값의 차이의 제곱합으로 가장 알맞은 것은?

① SSE ② SST
③ SSR ④ AE

62 다음 중 데이터 분석 모형의 오류에 대한 설명으로 가장 알맞지 않은 것은?

① 일반화 오류는 분석 모형을 만들 때 주어진 데이터 집합의 특성을 지나치게 반영하여 발생하는 오류이다.
② 일반화 오류는 과소 적합되었다고 한다.
③ 일반화 오류는 주어진 데이터 집합은 모집단 일부분임에도 불구하고 그것이 가지고 있는 주변적인 특성, 단순 잡음 등을 모두 묘사하기 때문에 일반화 오류가 발생한다.
④ 학습오류는 주어진 데이터 집합에 부차적인 특성과 잡음이 있다는 점을 고려하여 그것의 특성을 덜 반영하도록 분석 모형을 만들어 생기는 오류이다.

63 데이터 집합을 무작위로 동일 크기를 갖는 K개의 부분 집합으로 나누고, 그중 1개 부분 집합을 평가 데이터(Test Set)로, 나머지 (K-1)개의 부분 집합을 훈련 데이터(Training Set)로 선정하여 분석 모형을 평가하는 기법으로 가장 알맞은 것은?

① 랜덤 서브샘플링(Random Sub-Sampling)
② K-fold Cross Validation
③ 홀드 아웃(Holdout)
④ LOOCV

64 다음 중 범주에 따라 분류된 변수가 정규 분포되어 있다면 빈도가 실제 기대되는 값으로부터 유의미한 차이가 관찰되는가를 보기 위한 검증으로 가장 알맞은 것은?

① Z-검정
② 카이제곱 검정
③ 분산 분석
④ T-검정

65 다음 적합도 검정 방법 중에서 정규성 검정에 사용되지 않는 검정 방법은?

① 샤피로-윌크 검정
② 콜모고로프-스미르노프 적합성 검정
③ 카이제곱 검정
④ Q-Q Plot

66 다음 중 과대 적합(Over-fitting)을 방지하는 방법으로 가장 알맞지 않은 것은?

① 데이터 세트 감소
② 모델 복잡도 감소
③ 가중치 규제
④ 드롭아웃

67 매개변수 중 하나의 뉴런에 입력된 모든 값을 다 더한 값(가중합)에 더해주는 상수는 무엇인가?

① 손실 함수 ② 가중치
③ 학습률 ④ 편향

68 다음 중 훈련 데이터를 중복하여 사용하지 않고 훈련 데이터 세트를 나누는 기법으로 가장 알맞은 것은?

① 직접 투표
② 배깅
③ 페이스팅
④ 랜덤 서브스페이스

69 다음 중 경사 하강법(Gradient Descent)을 이용하여 가중치 업데이트하여 최적화된 결과를 얻는 기법은 무엇인가?

① 랜덤 패치
② 랜덤 포레스트
③ 에이다 부스트
④ 그레디언트 부스트

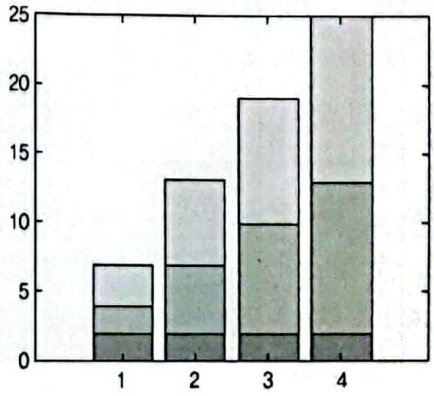
70 다음 중 관계 시각화 기법으로 가장 알맞지 않은 것은?

① 산점도
② 도넛 차트
③ 버블 차트
④ 히스토그램

71 다음 중 하나의 자산을 획득하려 할 때 주어진 기간 동안 모든 연관 비용을 고려할 수 있도록 확인하기 위해 사용되는 평가 기법으로 가장 알맞은 것은?

① TCO ② ROI
③ NPV ④ IRR

72 다음 중 막대를 사용하여 전체 비율을 보여주면서 여러 가지 범주를 동시에 차트로 표현 가능한 그래프로 가장 알맞은 것은 무엇인가?

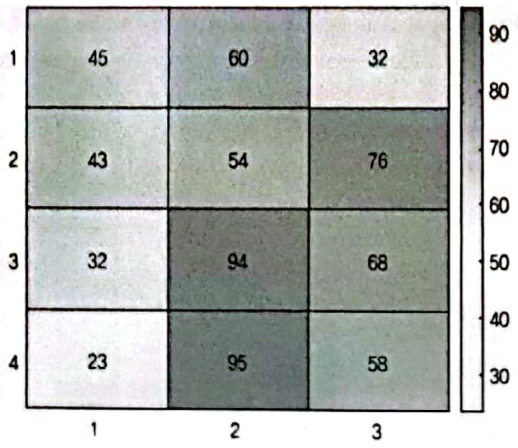


① 막대그래프
② 선 그래프
③ 영역 차트
④ 누적 막대그래프

73 다음 중 관계 시각화에 대한 설명으로 가장 알맞지 않은 것은?

① 다변량 데이터 사이에 존재하는 변수 사이의 연관성, 분포와 패턴을 찾는 시각화 방법이다.
② 변수 사이의 연관성인 상관관계는 한 가지 요소의 변화가 다른 요소의 변화와 관련이 있는지를 표현하는 시각화 기법이다.
③ 산점도 행렬은 산점도에서 데이터값을 나타내는 점 또는 마크에 여러 가지 의미를 부여하여 확장된 차트이다.
④ 관계 시각화의 유형으로 산점도, 산점도 행렬, 버블 차트, 히스토그램 등이 있다.

74 다음 중 여러 가지 변수를 비교할 수 있는 시각화 그래프로 가장 알맞은 것은?



① 히트맵
② 플로팅 바 차트
③ 체르노프 페이스
④ 스타 차트

75 다음 중 인포그래픽 유형으로 가장 알맞지 않은 것은?

① 지도형은 특정 국가나 지역의 지도 안에 정보를 담는 방식이다.
② 스토리텔링형은 캐릭터 등의 만화적 요소를 활용한 방식이다.
③ 도표형은 다양한 표와 그래프를 사용해 정보를 담는 방식이다.
④ 타임라인형은 주제를 선정하여 관련된 히스토리를 타임라인 형태로 나타내는 방식이다.

76 다음 혼동 행렬(Confusion Matrix)에서 특이도(Specificity)와 정밀도(Precision)는 무엇인가?

		예측		총합
		Positive	Negative	
실제	Positive	5	35	40
	Negative	15	45	60
총합		20	80	100

① 특이도: 3/4, 정밀도: 2/4
② 특이도: 3/4, 정밀도: 1/4
③ 특이도: 1/4, 정밀도: 3/4
④ 특이도: 1/4, 정밀도: 2/4

77 초기 아이디어 개발 관점 분류 중 생각하고 있는 것, 기억하고 있는 내용을 마음속에 지도를 그리듯이 줄거리를 이해하며 정리하는 방법으로 가장 알맞은 것은?

① 친화 도표 방식
② 마인드맵 방식
③ 피라미드 방식
④ 평행 좌표 그래프

78 다음 중 초매개변수(Hyper Parameter)로 설정이 가능하지 않은 것은?

① 학습률(Learning Rate)
② 가중치(Weight)
③ 은닉층(Hidden Layer)의 수
④ 의사결정 나무(Decision Tree)의 깊이

79 다음 중 개선 데이터 선정 시 고려 사항으로 가장 알맞지 않은 것은 무엇인가?

① 최신 데이터 적용이나 변수 추가 방식으로 분석 모형을 재조정한다.
② 업무 프로세스 KPI의 변경 또는 주요 시스템 원칙 변경, 발생 이벤트의 건수 증가에 따라 성능 평가를 하고 필요하면 재조정한다.
③ 조건 변화나 가중치 변화 시 계수 값 조정 또는 제약조건 추가로 재조정한다.
④ 최근 데이터 위주로만 오류율을 점검하고 기존 데이터 집합에 대한 데이터 오류율은 점검하지 않는다.

80 다음 중 혼동 행렬에 대한 설명으로 적절하지 않은 것은?

		실제	
		Positive	Negative
예측	Positive	TP	FN
	Negative	FP	TN

① 실제로 부정인 범주 중 부정으로 올바르게 예측(True Negative)한 비율을 정확도(Accuracy)라고 하며, (TP+TN)/(TP+TN+FP+FN)라고 표기한다.
② 카파 값(Kappa value)은 0~1 사이의 값을 가지며, 1에 가까울수록 예측값과 실제값이 일치하지 않는 것을 의미한다.
③ 정밀도(Precision)는 '긍정'으로 예측한 비율 중에서 실제로 긍정인 비율로 TP/(TP+FP)라고 표기한다.
④ 머신러닝 성능 평가지표 중 오차 비율(Error Rate)을 표기하는 식은 (FP+FN)/(TP+TN+FP+FN)이다.

1과목 빅데이터 분석 기획

01 다음 중 DIKW 피라미드에서 아래 설명에 해당하는 피라미드 요소는?

> 근본 원리에 대한 깊은 이해를 바탕으로 도출되는 창의적 아이디어
>
> 예 A 사이트의 다른 상품들도 B 사이트보다 쌀 것이라 판단

① 데이터(Data)
② 지식(Knowledge)
③ 지혜(Wisdom)
④ 정보(Information)

02 다음 중 빅데이터의 특징에 대한 설명으로 올바르지 않은 것은?

① 휘발성(Volatility): 데이터가 얼마나 오래 저장될 수 있고, 타당하여 오랫동안 쓰일 수 있을지에 관한 특징
② 규모(Volume): 정형 데이터뿐만 아니라 비정형, 반정형 데이터를 포함하는 특징
③ 속도(Velocity): 사물 정보(센서, 모니터링), 스트리밍 정보 등 실시간성 정보의 생성 속도 증가에 따라 처리 속도의 가속화가 요구되는 특징
④ 정확성(Validity): 데이터의 규모가 아무리 크더라도 질 높은 데이터를 활용한 정확한 분석 수행이 없다면 의미가 없다는 특징

03 다음 중 데이터 지식경영의 암묵지, 형식지의 상호작용에 대한 설명으로 올바르지 않은 것은?

① 내재화는 형식지가 상호결합하면서 새로운 형식지를 창출하는 과정이다.
② 공통화는 다른 사람과의 대화 등 상호작용을 통해 개인이 암묵지를 습득하는 단계이다.
③ 내면화는 행동과 실천 교육 등을 통해 형식지가 개인의 암묵지로 체화되는 단계이다.
④ 표출화는 형식지 요소 중 하나로 개인에게 내재된 경험을 객관적인 데이터로 문서나 매체에 저장, 가공, 분석하는 과정이다.

04 1996년 Fayyad가 프로파일링 기술을 기반으로 통계적 패턴이나 지식을 찾기 위해 체계적으로 정리한 방법론으로 분석 절차가 데이터 세트 선택, 데이터 전처리, 데이터 변환, 데이터 마이닝, 데이터 마이닝 결과 평가 단계로 이루어진 분석 방법론은 무엇인가?

① KDD 분석 방법론
② CRISP-DM 분석 방법론
③ SEMMA 분석 방법론
④ SAS 분석 방법론

05 다음 중 빅데이터 조직 구조에 대한 설명으로 가장 올바르지 않은 것은?

① 집중 구조는 전사 분석 업무를 별도의 분석 전담 조직에서 담당하고, 전략적 중요도에 따라 분석 조직이 우선순위를 정해서 진행이 가능한 조직이다.
② 기능 구조는 일반적인 형태로 별도 분석조직이 없고 해당 부서에서 분석을 수행하는 조직이다.
③ 혼합 구조는 전사적 핵심 분석이 어려우며 과거에 국한된 분석을 수행하는 구조이다.
④ 분산 구조는 분석조직 인력들을 현업 부서로 직접 배치해 분석 업무를 수행하는 조직이다.

06 다음 중 데이터 사이언티스트에 대한 설명으로 가장 올바르지 않은 것은?

① 데이터 사이언티스트는 복잡한 비즈니스 문제를 모델링하고 인사이트를 도출하며 통계학, 알고리즘, 데이터 마이닝 그리고 시각화 기법 등을 통해 그 속에서 가치를 찾아내는 사람이다.
② 데이터 사이언티스트의 요구역량에는 Hard Skill과 Soft Skill이 있다.
③ 데이터 사이언티스트의 Soft Skill에는 통찰력 있는 분석과 설득력 있는 전달이 있다.
④ 데이터 사이언티스트의 Hard Skill에는 다분야 간 협력이 있다.

07 다음은 빅데이터 플랫폼 기술에 대한 설명이다. 괄호 ( )안에 들어갈 용어로 맞는 것은?

- 데이터 수집 기술에는 수집 대상 데이터를 추출, 가공(변환, 정제)하여 데이터 웨어 하우스 및 데이터 마트에 저장하는 기술인 (Ⓐ)이/가 있다.
- 데이터 저장 기술에는 2차원 테이블인 데이터 모델에 기초를 둔 관계형 데이터베이스를 생성하고 수정하고 관리할 수 있는 소프트웨어인 (Ⓑ)이/가 있다.

① Ⓐ: ETL, Ⓑ: NoSQL
② Ⓐ: 데이터 레이크, Ⓑ: NoSQL
③ Ⓐ: ETL, Ⓑ: RDBMS
④ Ⓐ: 데이터 레이크, Ⓑ: RDBMS

08 개인정보처리자는 정보주체의 동의를 받은 경우에는 정보주체의 개인정보를 제3자에게 제공(공유 포함)할 수 있다. 개인정보를 제공하기 위해 정보주체의 동의를 받을 때 고지해야 할 사항으로 옳지 않은 것은?

① 개인정보 폐기 사유
② 개인정보를 제공받는 자의 개인정보 이용 목적
③ 제공하는 개인정보의 항목
④ 개인정보를 제공받는 자의 개인정보 보유 및 이용 기간

09 다음 중 가명처리 가이드라인에 따라 개인정보처리자가 정당한 처리 범위 내에서 정보주체의 동의 없이 가명정보를 처리할 수 있는 분야로 올바르지 않은 것은?

① 시장조사와 같은 상업적 목적의 통계 작성
② 기술의 개발과 실증, 기초 연구, 응용 연구뿐만 아니라 새로운 기술 · 제품 · 서비스 개발 등 산업적 목적을 위한 연구
③ 선거관리위원회에 정식으로 등록된 정당에서 지역 주민의 정치 성향 분석 연구
④ 민간기업, 단체 등이 일반적인 공익을 위한 기록 보존

10 다음 중 하향식 접근 방식에 대한 설명으로 가장 올바르지 않은 것은?

① 업무, 제품, 고객, 규제와 감사, 지원 인프라 5가지 영역으로 기업 비즈니스를 분석한다.
② 문제 탐색 시 분석 유스케이스를 정의한다.
③ 절차는 문제 탐색, 문제 정의, 해결방안 탐색, 타당성 검토, 선택 순이다.
④ 문제에 대한 비지도 학습 방법 및 프로토타이핑 접근법을 사용해서 분석한다.

11 빅데이터 분석 방법론의 계층 중에서 입력자료(Input), 처리 및 도구(Process & Tool), 출력자료(Output)로 구성된 단위 프로세스(Unit Process)는 무엇인가?

① 단계(Phase)
② 태스크(Task)
③ 스텝(Step)
④ 프로세스 그룹(Process Group)

12 다음 중 추가정보의 사용 없이는 특정 개인을 알아볼 수 없게 조치한 정보를 무엇이라고 하는가?

① 개인 정보
② 가명 정보
③ 익명 정보
④ 신용 정보

13 다음 중 SEMMA 분석 방법론에 대한 설명으로 가장 올바르지 않은 것은?

① SEMMA 분석 방법론의 분석 절차는 샘플링, 탐색, 수정, 최적화, 검증의 5단계로 되어 있다.
② 분석 솔루션 업체 SAS사가 주도한 통계 중심의 5단계 방법론이다.
③ 탐색 단계에서는 기초통계, 그래프 탐색, 요인별 분할표, 클러스터링, 변수 유의성 및 상관 분석을 통한 분석 데이터를 탐색한다.
④ 수정 단계에서는 수량화, 표준화, 각종 변환, 그룹화를 통한 분석 데이터 수정 및 변환을 한다.

14 다음 중 원천 데이터 수집 유형에 대한 설명으로 올바르지 않은 것은?

① 내부 데이터는 조직(인프라) 내부에 데이터가 위치하며, 데이터 담당자와 수집 주기 및 방법 등을 협의하여 데이터를 수집한다.
② 내부 데이터는 내부 조직 간 협의를 통한 데이터를 수집해야 한다.
③ 내부 데이터는 주로 수집이 용이한 정형 데이터이고, 서비스의 수명 주기 관리가 용이하다.
④ 내부 데이터의 유형으로는 센서 데이터, 장비 간 발생 로그, LOD 등이 있다.

15 익명화 기법 중 동일한 확률적 정보를 가지는 변형된 값에 대하여 원래 데이터를 대체하는 기법은 무엇인가?

① 가명처리(Pseudonym)
② 일반화(Generalization)
③ 치환(Permutation)
④ 섭동(Perturbation)

16 다음 중 대상별 분석 기획 유형 중 분석 대상(What)이 명확하게 무엇인지 모르는 경우 기존 분석 방식을 활용하여 새로운 지식을 도출해 내는 유형은?

① 최적화(Optimization)
② 솔루션(Solution)
③ 통찰(Insight)
④ 발견(Discovery)

17 다음 중 기업에서 사용하는 데이터의 가용성, 유용성, 통합성, 보안성을 관리하기 위한 정책과 프로세스를 다루며 프라이버시, 보안성, 데이터품질, 관리규정 준수를 강조하는 모델로 가장 적절한 것은 무엇인가?

① 데이터 거버넌스 ② 데이터 레이크
③ 데이터 마트 ④ 데이터 사이언스

18 다음 중 상향식 접근 방식 절차로 올바른 것은?

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

19 다음 중 개인정보의 파기에 관한 사항으로 올바르지 않은 것은?

① 개인정보처리자는 보유 기간의 경과, 개인정보의 처리 목적 달성 등 그 개인정보가 불필요하게 되었을 때는 지체 없이 그 개인정보를 파기하여야 한다.
② 개인정보처리자가 개인정보를 파기하지 아니하고 보존하여야 하는 경우에는 해당 개인정보 파일을 다른 개인정보와 함께 저장할 수 있다. 함께 저장할 때는 반드시 개인정보 파일을 암호화하여 저장·관리하여야 한다.
③ 개인정보처리자가 제1항에 따라 개인정보를 파기할 때에는 복구 또는 재생되지 아니하도록 조치하여야 한다.
④ 개인정보의 파기방법 및 절차 등에 필요한 사항은 대통령령으로 정한다.

20 다음 중 HDFS에 대한 설명으로 올바르지 않은 것은?

① HDFS는 수십 테라바이트 또는 페타바이트 이상의 대용량 파일을 분산된 서버에 저장하고, 그 저장된 데이터를 빠르게 처리할 수 있게 하는 파일 시스템이다.
② HDFS는 블록 구조의 파일 시스템으로 파일을 특정 크기의 블록으로 나누어 분산된 서버에 저장되는데, 블록 크기는 64MB에서 하둡 2.0부터는 128MB로 증가되었다.
③ HDFS의 유형에는 Key-Value Store, Column Family Data Store, Document Store, Graph Store가 있다.
④ HDFS는 하나의 네임 노드(Name Node)와 하나 이상의 보조 네임 노드, 다수의 데이터 노드(Data Node)로 구성된다.

2과목 빅데이터 탐색

21 다음 중 데이터 정제에 대한 설명으로 가장 올바르지 않은 것은?

① 데이터 정제는 결측값을 채우거나 이상값을 제거하는 과정을 통해 데이터의 신뢰도를 높이는 작업이다.
② 데이터 정제 절차는 데이터 오류 원인 분석, 데이터 정제 대상 선정, 데이터 정제 방법 결정 순으로 처리된다.
③ 데이터 오류 원인 중 결측값(Missing Value)은 실제는 입력되지 않았지만 입력되었다고 잘못 판단된 값으로 일정 간격으로 이동하면서 주변보다 높거나 낮으면 평균값으로 대체해서 처리한다.
④ 데이터 정제는 삭제, 대체, 예측값 삽입 등의 방법을 사용한다.

22 영향력이 가장 큰 변수를 하나씩 추가하는 변수 선택 기법은 다음 중 무엇인가?

① 후진 소거법
② 전진 선택법
③ 단계적 방법
④ 필터 기법

23 다음이 설명하는 데이터 이상값 발생 원인은 무엇인가?

> 음주량을 묻는 조사가 있다고 가정했을 때 10대 대부분은 자신들의 음주량을 적게 기입할 것이고, 오직 일부만 정확한 값을 적는 경우 발생

① 고의적인 이상값
② 표본추출 오류
③ 실험 오류
④ 데이터 처리 오류

24 다음 중 데이터 결측값에 대한 설명으로 가장 올바르지 않은 것은?

① 데이터 결측값이란 입력이 누락된 값을 의미하고, 결측값은 NA, 999999, Null 등으로 표현한다.
② 데이터 결측값의 종류에는 완전 무작위 결측(MCAR), 무작위 결측(MAR), 비 무작위 결측(MNAR)이 있다.
③ 무작위 결측(MAR; Missing At Random)은 변수상에서 발생한 결측값이 다른 변수들과 아무런 상관이 없는 결측값을 말한다.
④ 데이터 결측값은 결측값 식별, 결측값 부호화, 결측값 대체 절차로 처리된다.

25 다음 중 데이터 결측값을 처리하는 방법 중 단순 대치법(Single Imputation)에 대한 설명으로 가장 올바르지 않은 것은?

① 단순 대치법은 결측값을 그럴듯한 값으로 대체하는 통계적 기법이다.
② 단순 대치법의 종류에는 완전 분석법, 평균 대치법, 단순 확률 대치법이 있다.
③ 평균 대치법의 종류에는 핫덱 대체, 콜드덱 대체, 혼합방법이 있다.
④ 단순 확률 대치법은 평균 대치법에서 관측된 자료를 토대로 추정된 통계량으로 결측값을 대치할 때 어떤 적절한 확률값을 부여한 후 대치하는 방법이다.

26 다음의 임베디드 기법(Embedded Method)들에 대한 설명으로 가장 올바르지 않은 것은?

① 라쏘(LASSO): 가중치의 절댓값의 합을 최소화하는 것을 추가적인 제약조건으로 하는 방법이다.
② 릿지(Ridge): L1-norm을 통해 제약을 주는 방법이다.
③ 엘라스틱 넷(Elastic Net): 라쏘(LASSO)와 릿지(Ridge) 두 개를 선형 결합한 방법이다.
④ SelectFromModel: 의사결정나무 기반 알고리즘에서 피처를 추출하는 방법이다.

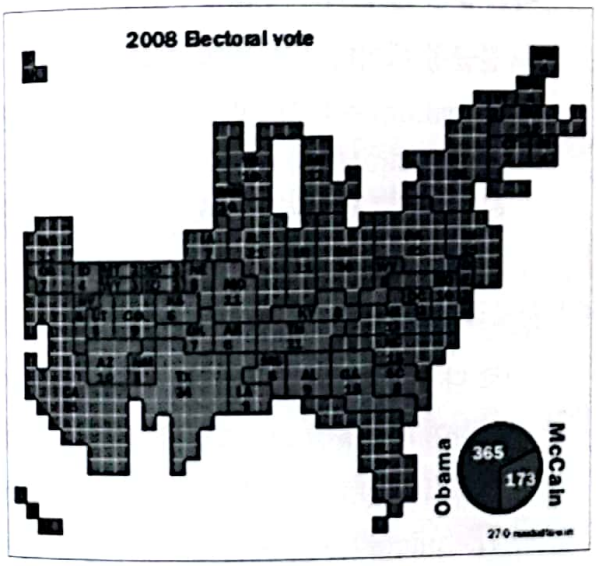
27 확률 변수 X, Y가 있을 때, $E(X)=2$, $E(X^2)=5$이고, $Y=2X+3$와 같이 주어질 경우에 $V(Y)$는 얼마인가?

① 1 ② 2
③ 4 ④ 6

28 데이터 분포의 모양이 왼쪽 편포(왼쪽 꼬리 분포)일 경우에 평균(Mean)과 중위수(Median), 최빈수(Mode)의 크기를 가장 바르게 설명한 것은 무엇인가?

① 평균(Mean) < 최빈수(Mode) < 중위수(Median)
② 평균(Mean) < 중위수(Median) < 최빈수(Mode)
③ 중위수(Median) < 평균(Mean) < 최빈수(Mode)
④ 중위수(Median) = 평균(Mean) = 최빈수(Mode)

29 아래의 그림은 선거인단에 따른 미국 대선 지형도이다. 이와 같이 특정한 데이터값의 변화에 따라 지도의 면적이 왜곡되는 지도를 무엇이라고 하는가?



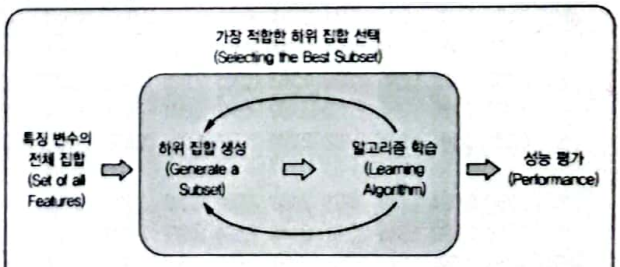
① 코로플레스 지도(Choropleth Map)
② 카토그램(Cartogram)
③ 버블 차트(Bubble Chart)
④ 도트맵(Dot map)

30 표본추출 기법 중에서 다음이 설명하는 기법으로 가장 옳은 것은?

> 100명의 사람에게 번호표를 나눠주고 끝자리가 7로 끝나는 사람들을 대상으로 설문 조사를 실시하였다.

① 단순 무작위 추출(Simple Random Sampling)
② 계통 추출(Systematic Sampling)
③ 층화 추출(Stratified Random Sampling)
④ 군집 추출(Cluster Random Sampling)

31 다음 중 아래에서 설명하고 있는 변수 선택 세부 기법으로 올바르지 않은 것은?



모델의 정확도에 기여하는 변수를 학습하고, 좀 더 적은 계수를 가지는 회귀식을 찾는 방향으로 제약조건을 주어 이를 제어하는 방법이다.

① 라쏘(LASSO)
② 릿지(Ridge)
③ 엘라스틱 넷(Elastic Net)
④ RFE(Recursive Feature Elimination)

32 A 보험회사에서 가입자를 세 그룹인 고위험군, 중위험군, 저위험군으로 나누고 있다. 고위험군은 전체 가입자의 20%, 중위험군은 전체 가입자의 30%, 저위험군은 전체 가입자의 50%를 차지하고 있다. 고위험군, 중위험군, 저위험군에 속한 가입자가 보험금을 청구할 확률은 각각 50%, 30%, 20%이다. 어느 가입자가 보험금을 청구했을 때, 이 가입자가 고위험군에 속한 가입자일 확률은 얼마인가?

① $\frac{1}{10}$ ② $\frac{1}{5}$

③ $\frac{5}{29}$ ④ $\frac{10}{29}$

33 건전지를 대량 생산하는 제조 회사의 건전지 16개를 표본추출하여 수명을 추출하였더니 평균이 25시간이고 표준편차가 2시간이었다. 모집단이 정규분포를 따른다고 가정하였을 때 이 제조회사 건전지의 평균 수명에 대한 95% 신뢰 수준은 다음 중 무엇인가? (t-분포표는 다음의 표와 같으며 계산 결과는 소수 3째 자리에서 반올림하여라.)

df \ α	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
⋮										
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850

① $23.93 \leq \mu \leq 26.07$

② $23.93 \leq \mu \leq 25.98$

③ $24.12 \leq \mu \leq 25.88$

④ $24.12 \leq \mu \leq 26.07$

34 다음 중 불균형 데이터 처리 방법으로 가장 올바르지 않은 것은?

① 불균형 데이터 처리 방법으로는 임곗값 이동, 과소 표집, 과대 표집, 앙상블 기법이 있다.

② 임곗값 이동은 같거나 서로 다른 여러 가지 모형들의 예측/분류 결과를 종합하여 최종적인 의사결정에 활용하는 기법이다.

③ 과소 표집은 무작위로 정상 데이터의 일부만 선택하는 방법으로 유의미한 데이터만을 남기는 방식으로 데이터의 소실이 매우 크고, 때로는 중요한 정상 데이터를 잃게 될 수 있다.

④ 과대 표집은 무작위로 소수의 데이터를 복제하는 방법으로 정보가 손실되지 않는다는 장점이 있으나, 복제된 관측치를 원래 데이터 세트에 추가하면 여러 유형의 관측치를 다수 추가하여 과적합(Over-fitting)을 초래할 수 있다.

35 정확한 데이터 분석을 위해서는 불균형 데이터를 처리하는 것이 필요하다. 다음 중 불균형 데이터 처리에 대한 설명 중 올바르지 않은 것은?

① 불균형 데이터 처리 방법 중 과소 표집(Under-Sampling)은 데이터량을 감소시켜서 불균형 데이터를 처리하는 방법이고, 과대 표집(Over-Sampling)은 데이터량을 증가시켜서 불균형 데이터를 처리하는 방법이다.

② 앙상블 기법(Ensemble Technique)은 같거나 서로 다른 여러 가지 모형들의 예측·분류 결과를 종합하여 최종적인 의사결정에 활용하는 기법이다.

③ SMOTE(Synthetic Minority Over-sampling TEchnique)는 소수 클래스에서 중심이 되는 데이터와 주변 데이터 사이에 가상의 직선을 만든 후, 그 위에 데이터를 추가하는 방법이다.

④ 임곗값 이동(Cut-Off Value Moving)은 임곗값을 데이터가 많은 쪽으로 이동시키는 방법으로 학습 단계에서부터 임곗값을 이동한다.

36 A 버스 정류장에서 4분에 2명씩 승객이 온다. A 버스 정류장에 2분 동안 1명 이내로 올 확률을 구하시오. (e는 자연상수)

① $\frac{1}{e}$ ② $\frac{2}{e}$

③ e ④ $2e^2$

37 검정 통계량 및 이의 확률분포에 근거하여 귀무가설이 참일 때 귀무가설을 기각하게 되는 제1종 오류를 범할 확률은 다음 중 무엇인가?

① p-값 ② $1-\alpha$

③ α ④ β

38 다음 중 이산확률분포(Discrete Probability Distribution)에 대한 설명으로 올바르지 않은 것은?

① 이산확률분포는 확률변수 X가 0, 1, 2, 3, …와 같이 하나씩 셀 수 있는 값을 갖는 이산확률변수 X가 가지는 확률분포이다.

② 확률 질량 함수는 이산확률변수가 특정 값보다 작거나 같을 확률을 나타내는 함수이다.

③ 포아송 분포는 이산형 확률분포 중 주어진 시간 또는 영역에서 어떤 사건의 발송 횟수를 나타내는 확률분포이다.

④ 베르누이 분포는 특정 실험의 결과가 성공 또는 실패로 두 가지의 결과 중 하나를 얻는 확률분포이다.

39 모표준편차 $\sigma = 16$인 정규분포를 따르는 모집단에서 표본의 크기가 16인 표본을 추출하였을 때 표본평균($\overline{X}$)는 52이다. 모평균 μ에 대한 95% 신뢰구간을 구하여라. (단, $Z_{0.05} = 1.645$, $Z_{0.025} = 1.96$이다.)

① $44.16 \leq \mu \leq 59.84$

② $45.42 \leq \mu \leq 58.58$

③ $50.04 \leq \mu \leq 53.96$

④ $50.355 \leq \mu \leq 53.645$

40 다음 중 추론통계에 대한 설명으로 올바르지 않은 것은?

① 점 추정(Point Estimation)은 표본의 정보로부터 모집단의 모수를 하나의 값으로 추정하는 것으로 표본의 평균, 중위수, 최빈수 등을 사용한다.

② 점 추정에 사용되는 통계는 표본평균, 표본분산, 중위수, 최빈수이 있다.

③ 구간 추정(Interval Estimate) 은 신뢰도를 제시하면서 범위로 모수를 추정하는 방법이다.

④ 가설의 종류에는 귀무가설과 대립가설이 있고, 귀무가설은 표본을 통해 확실한 근거를 가지고 입증하고자 하는 가설이다.

3과목 빅데이터 모델링

41 다음 중 데이터 마이닝(Data Mining) 기반 분석 모형 선정에 대한 설명으로 올바르지 않은 것은?

① 데이터 마이닝 기반 분석 모델은 분류(Classification), 예측(Prediction), 군집화(Clustering), 연관규칙(Association Rule) 모델이 있다.

② 분류 모델(Classification Model)은 범주형 변수 혹은 이산형 변수 등의 범주를 예측하는 것으로, 다수의 속성 혹은 변수를 가지는 객체들을 사전에 정해진 그룹이나 범주 중의 하나로 분류하는 모델이다.

③ 예측 모델(Prediction Model)은 데이터에 숨어 있는, 동시에 발생하는 사건 혹은 항목 간의 규칙을 수치화하는 것이다.

④ 군집화 모델(Clustering Model)은 이질적인 집단을 몇 개의 동질적인 소집단으로 세분화하는 모델로 크게 계층적 방법과 비 계층적 방법으로 구분한다.

42 다음 중 분석 모형의 활용 사례에 대한 설명으로 가장 올바르지 않은 것은?

① 연관규칙학습 – 햄버거를 구매하는 사람이 탄산음료를 더 많이 사는가?

② 유전 알고리즘 – 최소의 비용을 위한 최적의 배송 경로는 무엇인가?

③ 분류 분석 – 구매자의 나이가 구매 차량의 유형에 어떤 영향을 미치는가?

④ 소셜 네트워크 분석 – 고객들 간 관계망은 어떻게 구성되어 있나?

43 다음 중 지도 학습(Supervised Learning)에 대한 설명으로 가장 올바르지 않은 것은?

① 지도 학습은 정답인 레이블(Label)이 포함되어 있는 훈련 데이터를 통해 컴퓨터를 학습시키는 방법으로 설명변수와 목적변수 간의 관계성을 표현해내거나 미래 관측을 예측해 내는 것에 많이 활용된다.

② 지도 학습 유형에는 로지스틱 회귀, 인공신경망 분석(ANN), 의사결정나무, 서포트 벡터 머신(SVM), Q-Learning 등이 있다.

③ 지도 학습은 분석하고자 하는 목적변수(혹은 반응변수, 종속변수)의 형태가 수치형(양적 변수)인가 범주형(질적 변수)인가에 따라 분류와 수치예측 방법으로 다시 나눌 수 있다.

④ 지도 학습 유형 중 서포트 벡터 머신(Support Vector Machine)은 주어진 훈련 데이터를 회귀분석을 이용해서 2개의 그룹으로 분류하는 지도학습 모델이다.

44 다음이 설명하는 연관성 분석의 주요용어로 올바른 것은?

> 전체 거래 중 항목 A와 B를 동시에 포함하는 거래의 비율

① 지지도(Support)

② 신뢰도(Confidence)

③ 향상도(Lift)

④ 결합도(Coupling)

45 다음 중 군집 분석(Cluster Analysis)에 대한 설명으로 가장 옳지 않은 것은?

① 관측된 여러 개의 변숫값들로부터 유사성에만 기초하여 n개의 군집으로 집단화하고, 형성된 집단의 특성으로부터 관계를 분석하는 다변량 분석 기법이다.

② 군집 간의 거리측정 방법으로는 최단연결법, 최장연결법, 중심연결법 등이 있다.

③ 군집 간의 거리 계산을 위해 다익스트라(Dijkstra) 알고리즘을 활용한다.

④ 순위상관계수(Rank Correlation Coefficient)를 이용하여 거리를 측정한다.

46 다음 중 주성분 분석(PCA)에 대한 설명으로 옳은 것은?

① 상관관계가 있는 저차원 자료를 자료의 변동을 최대한 보존하는 고차원 자료로 변환하는 차원축소 방법이다.

② 차원축소는 고윳값이 낮은 순으로 정렬해서, 낮은 고윳값을 가진 고유벡터만으로 데이터를 복원한다.

③ 분석을 통해 나타나는 주성분으로 변수들 사이의 구조를 이해하기는 매우 쉽다.

④ 주성분 분석은 서로 상관성이 높은 변수들의 선형 결합으로 만들어 기존의 상관성이 높은 변수들을 요약, 축소하는 기법이다.

47 다음 중 회귀 분석(Regression Analysis)에 대한 설명으로 올바르지 않은 것은?

① 회귀 분석은 데이터들이 가진 속성들로부터 분할 기준 속성을 판별하고, 분할 기준 속성에 따라 트리 형태로 모델링하는 분류 예측 모델이다.

② 회귀 모형에서는 선형성, 독립성, 등분산성, 비상관성, 정상성의 가정을 만족시킬 수 있어야 한다.

③ 회귀 분석에서 단순선형 회귀 모형은 회귀 모형 중에서 가장 단순한 모형으로 독립변수와 종속변수가 각각 한 개이며 오차항이 있는 선형관계로 이뤄져 있다.

④ 로지스틱 회귀 분석은 반응변수가 범주형인 경우 적용되는 회귀 분석 모형으로 새로운 설명변수의 값이 주어질 때 반응변수의 각 범주에 속할 확률이 얼마인지를 추정하여 추적 확률을 기준치에 따라 분류하는 목적으로 사용될 수 있다.

48 다음 중 아래에서 설명하는 용어는 무엇인가?

> • 회귀계수를 추정하는 데 사용한다.
> • 측정값을 기초로 하여 제곱합을 만들고 그것을 최소로 하는 값을 구하여 측정결과를 처리하는 방법으로 오차 제곱의 합이 가장 작은 해를 구하는 것을 의미한다.

① 전체 제곱법

② 지수 평활법

③ 최소 제곱법

④ 오차 제곱법

49 다음 중 다중선형 회귀 분석(다변량 회귀 분석; Multi Linear Regression Analysis)에 대한 설명으로 가장 올바르지 않은 것은?

① 다중선형 회귀 분석 회귀식은 $Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+\epsilon$이다.

② 모형의 통계적 유의성은 P-통계량으로 확인한다.

③ 다중 선형 회귀 분석에서 다중공선성의 문제가 발생하면 문제가 있는 변수를 제거하거나 주성분 회귀, 능형 회귀 모형을 적용하여 문제를 해결한다.

④ 다중선형 회귀 분석의 결정계수(R^2)는 전체 데이터를 회귀 모형이 얼마나 잘 설명하고 있는지를 보여주는 지표로 회귀선의 정확도를 평가하는 데 사용한다.

50 다음 중 의사결정나무(Decision Tree)에 대한 설명으로 가장 올바르지 않은 것은?

① 의사결정나무는 주어진 입력값에 대하여 출력값을 예측하는 모형으로 예측나무 모형과 군집나무 모형이 있다.

② 의사결정나무 알고리즘 중 CART는 가장 성취도가 좋은 변수 및 수준을 찾는 것에 중점을 둔 알고리즘으로 개별 입력변수뿐만 아니라 입력변수들의 선형 결합 중에서 최적의 분리를 구할 수 있다.

③ 의사결정나무의 분석 과정은 의사결정나무 성장, 가지치기, 타당성 평가, 해석 및 예측 순으로 되어 있다.

④ 의사결정나무는 데이터들이 가진 속성들로부터 분할 기준 속성을 판별하고, 분할 기준 속성에 따라 트리 형태로 모델링하는 분류 예측 모델이다.

51 다음 중 인공신경망(Artificial Neural Network; ANN)에 대한 설명으로 가장 올바르지 않은 것은?

① 인공신경망은 인간의 뉴런 구조를 모방하여 만든 기계학습 모델로 입력값을 받아서 출력값을 만들기 위해 활성화 함수를 사용한다.

② 기초 형태의 인공신경망인 퍼셉트론(Perceptron)의 구조는 입력값, 가중치, 순 입력함수, 활성 함수, 예측값(출력값)으로 되어 있다.

③ 퍼셉트론에 은닉층(Hidden Layer)을 다층으로 하여 만든 인공신경망인 다층 퍼셉트론은 과대적합(Over-fitting)과 기울기 소멸의 문제점이 있다.

④ 활성화 함수 중 ReLU 함수는 기울기 소실의 원인이었지만, 시그모이드 함수 또는 tanh 함수를 통해 기울기 소실의 문제를 해결하였다.

52 다음 중 군집 분석(Cluster Analysis)에 대한 설명으로 가장 올바르지 않은 것은?

① 군집 분석 중 계층적 군집을 형성하는 방법에는 병합적 방법과 분할적 방법이 있고, 분할적 방법은 큰 군집으로부터 출발하여 군집을 분리해 나가는 방법으로 R의 {cluster} 패키지의 diana(), mona() 함수가 있다.

② 군집 간의 연결법에는 최단연결법, 최장연결법, 평균 연결법, 중심연결법, 와드연결법이 있다.

③ 군집 간의 거리 계산에 사용되는 연속형 변수 거리로는 유클리드 거리, 맨하튼 거리, 민코프스키 거리, 표준화 거리, 자카드(Jaccard) 계수 등이 있다.

④ 군집 분석 종류 중 혼합 분포 군집은 데이터가 K개의 모수적 모형의 가중합으로 표현되는 모집단 모형으로부터 나왔다는 가정하에서 모수와 함께 가중치를 자료로부터 추정하는 방법이다.

53 다음 중 교차 분석(카이제곱 검정; Chi-Squared Test)에 대한 설명으로 가장 올바르지 않은 것은?

① 교차 분석은 적합도 검정(Goodness Fitness Test), 독립성 검정(Test of Independence), 동질성 검정(Test of Homogeneity)의 3가지로 분류할 수 있다.

② 카이제곱 검정 공식은

$$\theta = \frac{\sum(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum(r_i - \bar{r})^2}\sqrt{\sum(s_i - \bar{s})^2}} (-1 \le \theta \le 1)$$

이다.

③ 교차 분석에서 적합도 검정은 1개의 요인을 대상으로 표본 집단의 분포가 주어진 특정 이론을 따르고 있는지를 검정하는 기법이다.

④ 교차 분석에서 독립성 검정은 여러 범주를 가지는 2개의 요인이 독립적인지, 서로 연관성이 있는지를 검정하는 기법이다.

54 다음 중 시계열 구성요소로 가장 옳지 않은 것은 무엇인가?

① 추세 요인(Trend Factor)

② 순환 요인(Cyclical Factor)

③ 계절 요인(Seasonal Factor)

④ 규칙 요인(Regular Factor)

55 다음이 설명하는 딥러닝 알고리즘에 해당하는 것은?

- 입력층, 은닉층, 출력층으로 구성되며 은닉층에서 재귀적인 신경망을 갖는 알고리즘이다.
- 음성신호, 연속적 시계열 데이터 분석에 적합하다.
- 장기 의존성 문제와 기울기 소실문제가 발생하여 학습이 이루어지지 않을 수 있다.

① RNN ② DNN

③ CNN ④ GAN

56 다음 중 딥러닝(Deep Learning) 분석에 대한 설명으로 가장 올바르지 않은 것은?

① 딥러닝 알고리즘에는 DNN, CNN, RNN, GAN 등 다양한 알고리즘이 존재한다.

② CNN 알고리즘은 시각적 이미지를 분석하는 데 사용되는 심층신경망으로 기존 영상처리의 필터 기능(Convolution)과 신경망(Neural Network)을 결합하여 성능을 발휘하도록 만든 구조이다.

③ RNN 알고리즘은 입력층에서 가중치가 곱해져서 은닉층으로 이동시키고, 은닉층에서도 가중치가 곱해지면서 다음 계층으로 이동한다.

④ DNN 알고리즘은 은닉층(Hidden Layer)을 심층(Deep) 구성한 신경망(Neural Network)으로 학습하는 알고리즘으로 입력층, 다수의 은닉층, 출력층으로 구성되어 있다.

57 다음 중 비모수 통계 검정 방법에 대한 설명으로 가장 올바르지 않은 것은?

① 부호 검정(Sign Test)은 차이의 부호와 상대적인 크기를 고려한 검정 방법이다.

② 윌콕슨 순위 합 검정(Wilcoxon Rank Sum Test)은 두 표본의 혼합 표본에서 순위 합을 이용한 검정 방법으로 자료의 분포가 연속적이고 독립적인 분포에서 나온 것이라는 기본 가정 외에 자료의 분포에 대한 대칭성 가정이 필요하다.

③ 대응 표본 검정(Paired Sample Test)은 하나의 모집단에서 두 가지 처리를 적용하여 관찰 값을 얻은 후 각 쌍의 차이를 이용하여 두 중위수의 차이를 검정하는 방법이다.

④ 크루스칼 왈리스 검정(Kruscal-Wallis Test)은 세 집단 이상의 분포를 비교하는 검정 방법으로 모수적 방법에서의 One-Way ANOVA와 같은 목적으로 쓰이고, 그룹별 평균이 아닌 중위수가 같은지를 검정한다.

58 다음 중 런 검정(Run Test)에 대한 설명으로 가장 올바르지 않은 것은?

① 런 검정은 두 개의 값을 가지는 연속적인 측정값들이 어떤 패턴이나 경향이 없이 임의적으로 나타난 것인지를 검정하는 방법이다.

② 런 검정은 이분화된 자료가 아닌 경우는 이분화된 자료로 변환시켜야 하고, 평균, 중위수, 최빈수 또는 사용자가 정의한 숫자 등의 기준값을 이용하여 이분화한다.

③ 동전의 앞면과 뒷면이 각각 1, 0이라고 할 때 '101001'이 나타났을 경우 3개의 연속적인 런(Run)이라고 한다.

④ 검정 통계량 계산 공식은 $\mu = \frac{2n_1n_2}{n_1+n2}+1$, $\sigma^2 = \frac{2n_1n_2(2n_1n_2-n1-n2)}{(n_1+n_2)^2(n_1+n_2-1)}$, $z = \frac{r-\mu}{\sigma}$이다.

59 다음 중 앙상블 분석에 대한 설명으로 가장 올바르지 않은 것은?

① 앙상블 알고리즘은 주어진 자료로부터 여러 개의 예측 모형을 만든 후 예측 모형들을 조합하여 하나의 최종 예측 모형을 만드는 방법으로 다중 모델 조합(Combining Multiple Models), 분류기 조합(Classifier Combination)이 있다.

② 앙상블 기법 중 배깅(Bagging)은 잘못 분류된 개체들에 가중치를 적용, 새로운 분류 규칙을 만들고, 이 과정을 반복해 최종 모형을 만드는 알고리즘이다.

③ 앙상블 기법 중 랜덤 포레스트는 의사결정나무의 특징인 분산이 크다는 점을 고려하여 배깅과 부스팅보다 더 많은 무작위성을 주어 약한 학습기들을 생성한 후 이를 선형 결합하여 최종 학습기를 만드는 방법이다.

④ 앙상블 기법 중 랜덤 포레스트의 주요기법에는 배깅 이용한 포레스트 구성, 임의노드 최적화 등이 있다.

60 다음 중 회귀 분석 유형 중 독립변수가 K개이며 종속변수와의 관계가 선형(1차 함수)인 것은?

① 단순선형 회귀 ② 다중선형 회귀
③ 곡선 회귀 ④ 로지스틱 회귀

4과목 빅데이터 결과 해석

61 다음 중 분포 시각화의 유형으로, 여러 개의 영역 차트를 겹겹이 쌓아놓은 모양의 시각화 방법은 무엇인가?

① 플로팅 바 차트(Floating Bar Chart)
② 누적 영역 차트(Stacked Area Graph)
③ 스타 차트(Star Chart)
④ 파이 차트(Pie Chart)

62 다음 중 혼동 행렬(Confusion Matrix; 정오 행렬)에 대한 설명으로 가장 올바르지 않은 것은?

① 혼동 행렬은 분석 모델에서 구한 분류의 예측 범주와 데이터의 실제 분류 범주를 교차 표(Cross Table) 형태로 정리한 행렬이다.

② 혼동 행렬에서 True/False는 예측한 값, Positive/Negative는 예측한 값과 실젯값의 비교 결과이다.

③ 혼동 행렬을 통한 분류 모형의 평가지표에서 정확도(Accuracy)에 대한 계산식은 $\frac{TP+TN}{TP+TN+FP+FN}$이다.

④ 혼동 행렬에서 F1-Score는 정밀도와 민감도(재현율)를 하나로 합한 성능평가지표이고, 정밀도와 민감도 양쪽 다 클 때 F1-Score도 큰 값을 가진다.

63 다음은 혼동 행렬을 통한 분류 모형의 평가지표에 대한 설명이다. 괄호 안에 들어갈 계산식으로 올바른 것은?

> 혼동 행렬을 통한 분류 모형의 평가지표에서 민감도(Sensitivity)의 계산식은 (Ⓐ)이고, 정밀도(Precision)의 계산식은 (Ⓑ)이다.

① Ⓐ: $\frac{TN}{TN+FP}$, Ⓑ: $\frac{TP}{TP+FP}$

② Ⓐ: $\frac{TN}{TN+FP}$, Ⓑ: $\frac{FP}{TN+FP}$

③ Ⓐ: $\frac{TP}{TP+FN}$, Ⓑ: $\frac{FP}{TN+FP}$

④ Ⓐ: $\frac{TP}{TP+FN}$, Ⓑ: $\frac{TP}{TP+FP}$

64 다음 중 교차 검증(Cross Validation)에 대한 설명으로 가장 올바르지 않은 것은?

① 홀드 아웃 교차 검증은 전체 데이터를 비복원 추출 방식을 이용하여 랜덤하게 훈련 데이터(Training Set)와 평가 데이터(Test Set)로 나눠 검증하는 기법이다.

② 홀드 아웃 교차 검증에서 훈련 데이터는 분류기를 만들 때 사용하는 데이터이고, 검증 데이터는 분류기들의 매개변수를 최적화하기 위해 사용하는 데이터이다.

③ 랜덤 서브샘플링은 모집단으로부터 조사의 대상이 되는 표본을 무작위로 추출하는 기법으로 모든 데이터를 학습(Training)과 평가(Test)에 사용할 수 있으나, K값이 증가하면 수행 시간과 계산량도 많아진다.

④ 부트스트랩은 주어진 자료에서 단순 랜덤 복원추출 방법을 활용하여 동일한 크기의 표본을 여러 개 생성하는 샘플링 방법이다.

65 다음 중 관계 시각화에 대한 설명으로 옳지 않은 것은?

① 다변량 데이터 사이에 존재하는 변수 사이의 연관성, 분포와 패턴을 찾는 시각화 방법이다..

② 관계 시각화의 주요 유형으로 히트맵, 체르노프 페이스, 스타차트 등이 있다.

③ 변수 사이의 연관성인 상관관계는 한 가지 요소의 변화가 다른 요소의 변화와 관련이 있는지를 표현하는 시각화 기법이다.

④ 정보를 SNS상에 쉽고 빠르게 전달할 수 있다.

66 다음 중 적합도 검정(Goodness of Fit Test)에 대한 설명으로 가장 올바르지 않은 것은?

① 적합도 검정은 표본 집단의 분포가 주어진 특정 이론을 따르고 있는지를 검정하는 기법이다.

② 적합도 검정 기법으로는 카이제곱 검정, 샤피로-윌크 검정, K-S 검정, Q-Q Plot이 있다.

③ 카이제곱 검정에서는 R 언어에서 chisq.test() 함수를 이용하여 나온 결과의 p-value 값이 0.05보다 클 경우 관측된 데이터가 가정된 확률을 따른다고 할 수 있다.

④ 정규성 검정(Normality Test)은 R에서 sharpirowilk.test() 함수를 이용하여 검정할 수 있으며, 이때 귀무가설은 "표본은 정규 분포를 따른다."이다.

67 다음 중 아래에서 설명하고 있는 검정 방법은 무엇인가?

> • 데이터가 어떤 특정한 분포를 따르는가를 비교하는 검정 기법이고, 비교 기준이 되는 데이터를 정규 분포를 가진 데이터로 두어서 정규성 검정을 실시할 수 있다.
> • *R*에서 ks.test() 함수를 이용하여 검정을 실시한다. (인자는 x, y, alternative 등이 있음)
> • x는 검정할 데이터, y는 비교 검정할 데이터이거나 이론적 분포이다.

① 콜모고로프-스미르노프 적합성 검정(Kolmogorov-Smirnov Goodness of Fit Test; K-S 검정)
② 정규성 검정(Normality Test)
③ 샤피로-윌크 검정(Shapiro-Wilk Test)
④ F-검정(F-test)

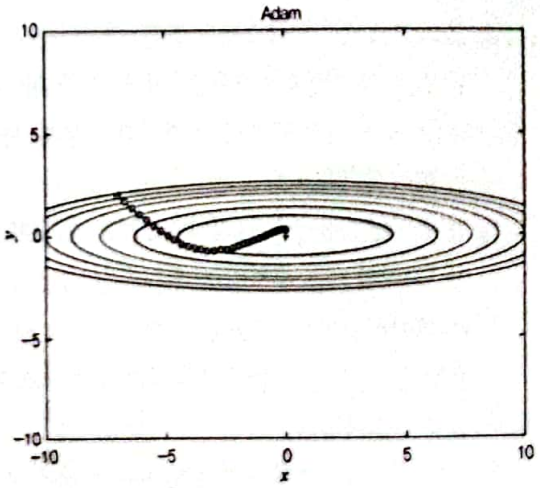
68 다음 중 과대 적합(Over-fitting)에 대한 설명으로 가장 올바르지 않은 것은?

① 과대 적합은 제한된 훈련 데이터 세트에 너무 과하게 특화되어 새로운 데이터에 대한 오차가 매우 커지는 현상이다.
② 과대 적합을 방지하기 위해 데이터 세트 증가, 모델 복잡도 감소, 가중치 규제, 드롭아웃 방법을 적용한다.
③ 가중치 규제는 개별 가중치 값을 제한하여 복잡한 모델을 좀 더 간단하게 하는 방법으로 종류에는 P1 규제와 P2 규제가 있다.
④ 드롭아웃은 학습 과정에서 신경망 일부를 사용하지 않는 방법이다.

69 다음 중 매개변수 최적화에 대한 설명으로 가장 올바르지 않은 것은?

① 매개변수 최적화는 학습 모델과 실제 레이블과 차이는 손실 함수로 표현되며, 학습의 목적은 오차, 손실 함수의 값을 최대한 작게 하도록 하는 매개변수(가중치, 편향)를 찾는 것이다.
② 매개변수의 종류에는 하나의 뉴런에 입력된 모든 값을 다 더한 값(가중합)에 더해주는 상수인 가중치(Weight)와 각 입력값에 각기 다르게 곱해지는 수치인 편향(Bias)이 있다.
③ 매개변수 최적화 기법 중 확률적 경사 하강법이란 손실 함수의 기울기를 구하여, 그 기울기를 따라 조금씩 아래로 내려가 최종적으로는 손실 함수가 가장 작은 지점에 도달하도록 하는 알고리즘이다.
④ 매개변수 최적화 기법 중 모멘텀은 기울기 방향으로 힘을 받으면 물체가 가속된다는 물리 법칙을 적용한 알고리즘이다.

70 다음 중 아래에서 설명하는 매개변수 최적화 기법은 무엇인가?



① AdaGrad ② Adam
③ 확률적 경사 하강법 ④ 드롭아웃

71 다음 중 AUC(Area Under ROC)에 대한 설명으로 옳지 않은 것은?

① ROC 곡선 아래의 면적을 모형의 평가지표로 삼는다.
② AUC는 진단의 정확도를 측정할 때 사용한다.
③ AUC의 값은 항상 0.5~1의 값을 가진다.
④ AUC의 값이 0.5인 경우 우수한 모형으로 판단한다.

72 다음 중 데이터 시각화 유형 중 가장 올바르지 않은 것은?

① 시간 시각화는 시간 흐름에 따른 변화를 통해 경향(트렌드) 파악하는 방법으로 막대그래프 기법과 점그래프 기법이 있다.
② 분포 시각화는 전체에서 부분 간 관계를 설명하는 방법으로 파이 차트 기법, 도넛 차트 기법, 트리 차트 기법이 있다.
③ 관계 시각화는 집단 간의 상관관계를 확인하여 다른 수치의 변화를 예측하는 방법으로 산점도 기법, 버블 차트 기법, 히스토그램 기법이 있다.
④ 비교 시각화는 각각의 데이터 간의 차이점과 유사성 관계도 확인 가능한 방법으로 등치선도 기법, 도트맵 기법, 카토그램 기법이 있다.

73 아래의 데이터 시각화 유형에 대한 설명 중 괄호( ) 안에 들어갈 가장 올바른 용어는 무엇인가?

> (Ⓐ)은/는 직교 좌표계를 이용해 두 개 변수 간의 관계를 나타내는 방법이고, (Ⓑ)은/는 자료 분포의 형태를 직사각형 형태로 시각화하여 보여주는 차트로, 수평축에는 각 계급을 나타내고, 수직축에는 도수 또는 상대도수를 나타낸다.

① Ⓐ: 산점도, Ⓑ: 히스토그램
② Ⓐ: 산점도, Ⓑ: 히트맵
③ Ⓐ: 등치선도, Ⓑ: 히스토그램
④ Ⓐ: 등치선도, Ⓑ: 히트맵

74 다음 중 빅데이터 시각화 도구 중 코딩 없이 스프레드시트, 데이터베이스 형태 데이터를 쉽게 가시화하는 시각화 도구는 무엇인가?

① 태블로(Tableau)
② 차트 블록(Chart Blocks)
③ 인포그램(Infogram)
④ 데이터 래퍼(Data Wrapper)

75 다음 혼동 행렬(Confusion Matrix)에서 참이 0이고 거짓이 1일 때, 특이도(Specificity)와 정밀도(Precision)는 무엇인가?

		실제 데이터		총합
		0	1	
예측 데이터	0	55	45	100
	1	20	40	60
총합		75	85	160

① 특이도: 1/3, 정밀도: 7/15
② 특이도: 1/3, 정밀도: 7/15
③ 특이도: 2/3, 정밀도: 11/15
④ 특이도: 2/5, 정밀도: 11/15

76 다음 중 공간 시각화 유형으로 가장 올바르지 않은 것은?

① 등치지역도 ② 도트맵
③ 산점도 ④ 카토그램

77 다음 중 응용 프로그램 성능 측정 항목의 측정 주기에 대한 설명으로 가장 올바르지 않은 것은?

① 응답시간/트랜잭션 처리량은 실시간 측정을 한다.
② 메모리 사용은 정기적 측정을 한다.
③ 데이터베이스 처리는 실시간 측정을 한다.
④ 오류 및 예외 발생 여부는 정기적 측정을 한다.

78 다음 중 카파 통계량(Kappa Statistic)에 대한 설명으로 가장 옳지 않은 것은?

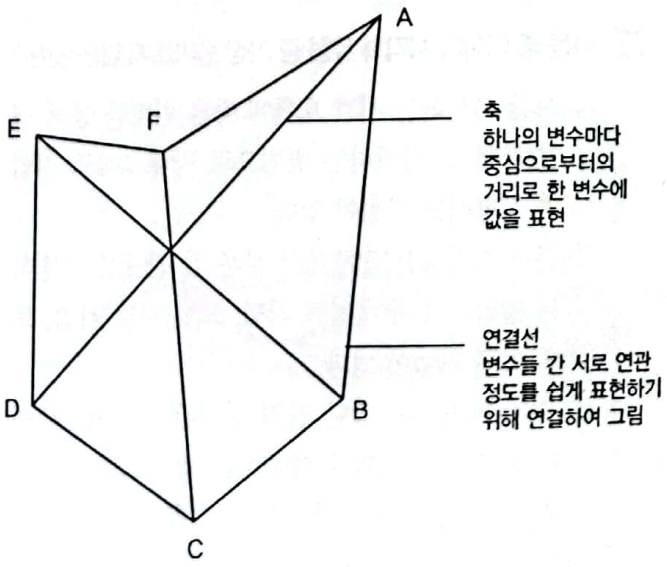
① 두 관찰자가 측정한 범주 값에 대한 일치도를 측정하는 방법이다.
② 0~1의 값을 가지며 1에 가까울수록 모델의 예측값과 실젯값이 정확히 일치하며, 0에 가까울수록 모델의 예측값과 실젯값이 불일치한다.
③ 정확도 외에 카파 통계량을 통해 모형의 평가 결과가 우연히 나온 결과가 아니라는 것을 설명
④ 카파 통계량의 계산식은 $2\times\frac{Precision\times Recall}{Precision+Recall}$ 이다.

79 다음 중 아래에서 설명하는 기법은 무엇인가?

> 두 개 이상의 집단 간 비교를 수행하고자 할 때 집단 내의 분산, 총 평균과 각 집단의 평균 차이에 의해 생긴 집단 간 분산 비교로 얻은 F-분포를 이용하여 가설검정을 수행하는 방법

① 분산 분석 ② 카이제곱 검정
③ Z-검정 ④ T-검정

80 다음 중 아래에서 설명하고 있는 비교 시각화 유형은 무엇인가?



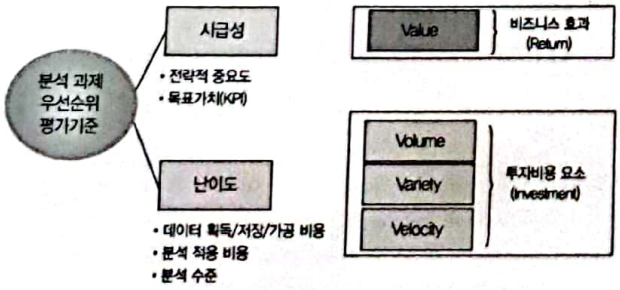
① 플로팅 바 차트 ② 스타 차트
③ 히트맵 ④ 평행 좌표

최종모의고사 정답 및 해설

1회 정답

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

01 **해설** 분석 과제 우선순위 평가기준에서 전략적 중요도, 목표가치와 관련이 있는 빅데이터 특성은 Value이다.



02 **해설** 반정형은 고정된 필드에 저장되어 있지만, 메타데이터나 데이터 스키마 정보를 포함하는 데이터로 XML, HTML, JSON 등이 있다.

03 **해설**

정형	• 정형화된 스키마 구조, DBMS에 내용이 저장될 수 있는 구조 • 고정된 필드(속성)에 저장된 데이터 • 관계형 데이터베이스(Oracle, MS-SQL 등)
반정형	• 데이터 내부에 데이터 구조에 대한 메타 정보 포함된 구조 • 고정된 필드에 저장되어 있지만, 메타데이터나 데이터 스키마 정보를 포함하는 데이터 • XML, HTML, JSON 등
비정형	• 수집 데이터 각각이 데이터 객체로 구분 • 고정 필드 및 메타데이터(스키마 포함)가 정의되지 않음 • 텍스트 문서, 이진 파일, 이미지, 동영상 등

04 **해설** • 높은 지능과 과학적 지식은 데이터 사이언티스트의 일반적인 요구 역량으로 올바르지 않다.
• 가트너(Gartner)는 데이터 사이언티스트가 갖추어야 할 역량으로 분석 모델링, 데이터 관리, 소프트 스킬, 비즈니스 분석을 제시했다.

05 **해설** 데이터 사이언티스트는 빅데이터에 대한 이론적 지식인 Hard Skill이 필요하다.

데이터 사이언티스트의 요구 역량	
협통전 숙지	(소프트 스킬) 협력 능력 / 통찰력 / 전달력 (하드 스킬) 숙련도 / 지식

06 **해설** • 빅데이터 플랫폼은 크게 수집, 저장, 분석, 활용 단계로 구성된다.
NoSQL은 빅데이터 저장 기술이다.

07 **해설** • 무엇을 해야 할 것인지를 확인하는 분석은 처방 분석이다.
• 가트너의 분석 가치 에스컬레이터는 아래와 같다.

묘사 분석 (Descriptive Analysis)	• 분석의 가장 기본적인 지표 • 과거에 어떤 일이 일어났고, 현재는 무슨 일이 일어나고 있는지 확인
진단 분석 (Diagnostic Analysis)	• 묘사 단계에서 찾아낸 분석의 원인을 이해하는 과정 • 데이터를 기반으로 왜 발생했는지 이유를 확인
예측 분석 (Predictive Analysis)	• 데이터를 통해 기업 혹은 조직의 미래, 고객의 행동 등을 예측하는 과정 • 무슨 일이 일어날 것인지를 예측
처방 분석 (Prescriptive Analysis)	• 예측을 바탕으로 최적화하는 과정 • 무엇을 해야 할 것인지를 확인 • 무엇을 해야 할 것인지를 확인하는 분석은 처방 분석이다.

08 **해설** 네임 노드(Master)와 데이터 노드(Slave)로 구성되어 있고 대용량 파일을 저장하고 처리하기 위해서 개발된 파일 시스템은 하둡 분산 파일 시스템(HDFS)이다.

09 **해설** • 비정형 데이터 수집을 위한 시스템에는 척와, 플럼, 스크라이브가 있다.
• 피그(Pig)는 데이터 가공을 위해 맵리듀스 API를 매우 단순화시키고, SQL과 유사한 형태로 설계된 시스템으로 데이터 정제 기술이다.

10 **해설** 데이터 거버넌스 체계는 데이터 표준화, 데이터 관리체계, 데이터 저장소 관리, 표준화 활동으로 구분된다.

데이터 표준화	• 데이터 표준 용어 설명, 명명 규칙, 메타데이터 구축, 데이터 사전 구축 • 데이터 표준 준수 진단, 논리·물리 모델 표준에 맞는지 검증
데이터 관리 체계	• 메타데이터와 데이터 사전의 관리 원칙 수립
데이터 저장소 관리	• 메타데이터 및 표준 데이터를 관리하기 위한 전사 차원의 저장소 구성
표준화 활동	• 데이터 거버넌스 체계 구축 이후 표준 준수 여부를 주기적으로 점검 및 모니터링 실시

11 **해설** • 도입 단계는 분석을 시작하는 단계로 환경과 시스템을 구축하고, 일부 부서에서 분석을 수행하여, 담당자 역량에 의존하는 단계이다.
• 기업의 데이터 분석 수준을 파악하기 위한 조직 평가 성숙도 단계는 아래와 같다.

도입 단계	분석을 시작해 환경과 시스템을 구축	• 일부 부서에서 수행 • 담당자 역량에 의존
활용 단계	분석 결과를 실제 업무에 적용	• 전문 담당 부서에서 수행 • 분석 기법 도입 • 관리자가 분석 수행
확산 단계	전사 차원에서 분석을 관리하고 공유	• 전사 모든 부서 수행 • 분석 COE 조직 운영 • 데이터 사이언티스트 확보
최적화 단계	분석을 진화시켜서 혁신 및 성과 향상에 기여	• 데이터 사이언스그룹 • 경영진 분석 활용 • 전략 연계

12 **해설** • 가명처리는 개인 식별이 가능한 데이터를 직접적으로 식별할 수 없는 다른 값으로 대체하는 기법으로 휴리스틱 가명화, 암호화, 교환 방법이 있다.
• 데이터 범주화는 단일 식별 정보를 해당 그룹의 대푯값으로 변환(범주화)하거나 구간 값으로 변환(범위화)하여 고유 정보 추적 및 식별 방지하는 것이다.
• 데이터 마스킹은 개인 식별 정보에 대하여 전체 또는 부분적으로 대체 값(공백, '*', 노이즈 등)으로 변환하는 것이다.
• 총계처리는 개인정보에 대하여 통계값을 적용하여 특정 개인을 판단할 수 없도록 하는 것이다.

13 **해설** • 개인정보의 수집·이용을 위해 정보주체의 동의를 받을 때 고지사항(개인정보보호법 15조 2항)

개인정보의 수집·이용을 위해 정보주체의 동의를 받을 때 고지사항	
목항기불	개인정보의 수집·이용 목적 / 수집하려는 개인정보의 항목 / 개인정보의 보유 및 이용 기간 / 동의를 거부할 권리가 있다는 사실 및 동의 거부에 따른 불이익이 있는 경우에는 그 불이익의 내용

14 **해설** 분석 대상과 방법을 모두 알고 있는 경우(Known)에는 최적화 기법을 사용한다.

분석의 방법(How) \ 분석의 대상(What)	Known	Un-Known
Known	Optimization	Insight
Un-Known	Solution	Discovery

15 **해설** • 데이터 정제, 새로운 데이터 생성 등 자료를 분석 가능한 상태로 만드는 것은 데이터 준비 단계이다.
• 데이터 준비는 많은 시간이 소요되며 분석용 데이터 세트 선택, 데이터 정제, 데이터 통합, 학습/검증 데이터 분리 등을 수행한다.

16 **해설** • 개인정보를 목적 외의 용도로 이용하거나 제3자에게 제공이 가능한 경우는 아래와 같다. (개인정보보호법 18조 2항)

1. 정보주체로부터 별도의 동의를 받은 경우
2. 다른 법률에 특별한 규정이 있는 경우
3. 정보주체 또는 그 법정대리인이 의사표시를 할 수 없는 상태에 있거나 주소불명 등으로 사전 동의를 받을 수 없는 경우로서 명백히 정보주체 또는 제3자의 급박한 생명, 신체, 재산의 이익을 위하여 필요하다고 인정되는 경우
4. 삭제
5. 개인정보를 목적 외의 용도로 이용하거나 이를 제3자에게 제공하지 아니하면 다른 법률에서 정하는 소관 업무를 수행할 수 없는 경우로서 보호위원회의 심의·의결을 거친 경우
6. 조약, 그 밖의 국제협정의 이행을 위하여 외국정부 또는 국제기구에 제공하기 위하여 필요한 경우
7. 범죄의 수사와 공소의 제기 및 유지를 위하여 필요한 경우
8. 법원의 재판업무 수행을 위하여 필요한 경우
9. 형 및 감호, 보호처분의 집행을 위하여 필요한 경우

17 **해설**

RSS	XML 기반으로 정보를 배포하는 프로토콜을 활용하여 데이터를 수집하는 기술
Open API	공개된 API를 이용하여 데이터를 수집하는 기술
아파치 카프카	레코드 스트림을 발행(Publish), 구독(Scribe)하는 방식의 분산 스트리밍 플랫폼 기술
크롤링	인터넷상에서 제공되는 다양한 웹 사이트로부터 소셜 네트워크 정보, 뉴스, 게시판 등의 웹 문서 및 콘텐츠 수집 기술

18 **해설** • 실시간 데이터는 생성된 이후 수 초~수 분 이내에 처리되어야 의미가 있는 현재 데이터이다.
• 센서 데이터, 시스템 로그, 네트워크 장비 로그, 알람, 보안 장비 로그가 있다.
• 과거 정보는 비실시간 데이터로, 생성된 데이터가 수 시간 또는 수 주 이후에 처리되어야 의미가 있는 과거 데이터이다.

19 **해설** • 가명 정보처리 시에도 개인정보의 최소처리원칙을 준수해야 한다.

사전준비	가명처리 대상 항목 및 처리수준을 정의하기 위해서는 처리 목적이 적합한지 여부를 확인하고 사전 계획을 수립함
가명처리	가명 정보처리 시에도 개인정보의 최소처리원칙을 준수하여야 하며, 가명처리 방법을 정할 때에는 처리목적, 처리(이용 또는 제공)환경, 정보의 성격 등을 종합적으로 고려함
적정성 검토 및 추가처리	목적달성을 위해 적절한 수준으로 가명처리가 이루어졌는지, 재식별 가능성은 없는지 등에 대한 최종적인 판단절차를 수행함
사후관리	적정성 검토 결과 가명처리가 적정하다고 판단되면 가명 정보를 본래 활용목적을 위해서 처리할 수 있으며, 법령에 따라 기술적 · 관리적 · 물리적 안전조치를 이행함

20 **해설** • 프라이버시 보호 모델은 다음과 같다.

k-익명성 (k-Anonymity)	• 주어진 데이터 집합에서 같은 값이 적어도 k개 이상 존재하도록 하여 쉽게 다른 정보로 결합할 수 없도록 하는 모델 • 공개된 데이터에 대한 연결 공격 취약점을 방어하기 위한 모델
l-다양성 (l-Diversity)	• 주어진 데이터 집합에서 함께 비식별 되는 레코드들은(동질 집합에서) 적어도 l개의 서로 다른 민감한 정보를 가져야 하는 프라이버시 모델 • 비식별 조치 과정에서 충분히 다양한(l개 이상) 서로 다른 민감한 정보를 갖도록 동질 집합을 구성 • k-익명성에 대한 두 가지 취약점 공격인 동질성 공격, 배경 지식에 의한 공격을 방어하기 위한 프라이버시 모델
t-근접성 (t-Closeness)	• 동질 집합에서 특정 정보의 분포와 전체 데이터 집합에서 정보의 분포가 t 이하의 차이를 보여야 하는 모델 • l-다양성의 쏠림 공격, 유사성 공격을 보완하기 위해 제안된 모델
m-유일성 (m-Uniqueness)	• 원본 데이터와 동일한 속성 값의 조합이 비식별 결과 데이터에 최소 m개 이상 존재하도록 하여 재식별 가능성 위험을 낮춘 모델

21 **해설** • $P(B|A)=\frac{P(A\cap B)}{P(A)}$의 위의 수학적 정의를 갖는 이론은 조건부 확률(Conditional Probability)이다.

• 조건부 확률은 어떤 사건이 일어난다는 조건에서 다른 사건이 일어날 확률로 두 개의 사건 A와 B에 대하여 사건 A가 일어난다는 선행조건 아래에 사건 B가 일어날 확률이다.

22 **해설** • 단순 확률 대치법에는 핫덱(Hot-Deck) 대체, 콜드덱(Cold-Deck) 대체, 혼합방법이 있다.

• 핫덱(Hot-Deck) 대체는 무응답을 현재 진행 중인 연구에서 "비슷한" 성향을 가진 응답자의 자료로 대체하는 방법이며 표본조사에서 주로 사용되는 기법이다.

23 **해설** 다중 대치법은 단순 대치법을 한 번 하지 않고 m번 대치를 통해 m개의 가상적 완전한 자료를 만들어서 분석하는 데이터 결측값 처리 기법이다.

데이터 이상값 검출 방법	
개통시 머마엘아	개별 데이터 관찰 / 통계값 / 시각화 / 머신러닝 기법 / 마할라노비스 거리 활용 / LOF / iForest

24 **해설** • 모평균을 모르는 대표본일 경우 평균의 $100\times(1-\alpha)\%$ 신뢰구간은 표본 분산이 s인 Z-분포를 이용하며 공식은 다음과 같다.

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

• $\bar{X}=60, n=81$이고 표본 분산(s^2) = 9이므로 표본 표준편차(s) = 3이 된다. 또한, 90% 신뢰구간이므로 $\alpha=0.1$이고, $\frac{\alpha}{2}=0.05$이므로 $Z_{\frac{\alpha}{2}}=Z_{0.05}=1.65$이다.

• 공식에 대입하면 $60-1.65\frac{3}{\sqrt{81}}\le\mu\le60+1.65\frac{3}{\sqrt{81}}$ 이므로 $59.45\le\mu\le60.55$이다.

• 따라서, 신뢰 구간의 하한은 59.45, 상한은 60.55가 된다.

25 **해설** 무작위로 정상 데이터의 일부만 선택하는 과소 표집 기법은 변수를 변환하는 방법이 아니고 불균형 데이터 처리기법이다.

26 **해설** • 데이터 이상값 발생 원인은 다음과 같다.

표본추출 오류	데이터를 샘플링하는 과정에서 나타나는 오류
고의적인 이상값	자기보고식 측정에서 나타나는 오류
데이터 입력 오류	데이터를 수집, 기록 또는 입력하는 과정에서 발생할 수 있는 오류
실험 오류	실험조건이 동일하지 않은 경우 발생하는 오류
측정 오류	데이터를 측정하는 과정에서 발생하는 오류
데이터 처리 오류	여러 개의 데이터에서 필요한 데이터를 추출하거나, 조합해서 사용하는 경우에 발생하는 오류
자연 오류	인위적이 아닌, 자연스럽게 발생하는 이상 값

27 **해설** • 전 확률의 정리 공식에 따르면 $P(x)$
$=P(A)P(x|A)+P(B)P(x|B)$이다.

• 베이즈 정리는 $P(B|x)$
$=\frac{P(B\cap x)}{P(x)}=\frac{P(A)P(x|A)}{P(A)P(x|A)+P(B)P(x|B)}$

29 해설 • 데이터값이 큰 지역의 면적을 시각적으로 더 크게 표시하여 데이터를 직관적으로 보기 위해 사용하는 기법은 카토그램이다.
• 시공간 데이터 탐색에서 버블 플롯맵은 위도와 경도를 사용하여 좌표를 원으로 정의하는 차트이다.

30 해설

하위 경계	• 제1 사분위에서 1.5 IQR을 뺀 위치
최솟값	• 하위 경계 내의 관측치의 최솟값
제1 사분위(Q_1)	• 자료들의 하위 25%의 위치를 의미
제2 사분위(중위수)	• 자료들의 50%의 위치로 중앙값(Median)을 의미
제3 사분위(Q_3)	• 자료들의 하위 75%의 위치를 의미
최댓값	• 상위 경계 내의 관측치의 최댓값
상위 경계	• 제3 사분위에서 IQR의 1.5배 위치
수염	• Q_1, Q_3로부터 IQR의 1.5배 내에 있는 가장 멀리 떨어진 데이터까지 이어진 선
이상값	• 수염보다 바깥쪽에 데이터가 존재한다면, 이것은 이상값으로 분류

31 해설 • 성공/실패로 두 가지 결과 중 하나를 얻는 확률분포로 베르누이 분포를 따른다.
• 성공 확률이 p는 $\frac{1}{2}$이고, 기댓값 $E(X)=p=\frac{1}{2}$이고, $V(X)=p(1-p)=\frac{1}{2}\left(1-\frac{1}{2}\right)=\frac{1}{4}$이다.

32 해설 • 정규분포 함수에서 X를 Z로 정규화한 분포는 표준 정규분포이다.
• 모집단이 정규분포라는 정도만 알고, 모 표준편차(σ)는 모를 때에는 T-분포를 사용한다.
• 모평균이 μ, 모분산이 σ^2이라고 할 때, 종 모양의 분포는 정규분포이다.
• 독립적인 χ^2-분포가 있을 때, 두 확률변수의 비는 F-분포이다.

33 해설 표본수가 무한히 크면 표본의 분포와 관련 없이 표본합 또는 표본평균은 정규분포를 따른다는 것은 중심극한정리이다.

34 해설 점 추정은 표본의 정보로부터 모집단의 모수를 하나의 값으로 추정하는 것이다.

점 추정 조건	
불효일충	불편성 / 효율성 / 일치성 / 충족성

35 해설 귀무가설은 H_0으로 표기하고, 대립가설은 H_1으로 표기한다.

36 해설 p-값은 귀무가설이 옳다는 가정하에 얻은 통계량이 귀무가설을 얼마나 지지하는지를 나타낸 확률이다.

37 해설 귀무가설이 참인데 잘못하여 기각하게 되는 것은 제1종 오류이다.

38 해설 • 모분산을 알고 있는 경우 모평균에 대한 $100\times(1-\alpha)$% 신뢰구간을 구하는 공식은 $\bar{X}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\le\mu\le\bar{X}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$이다.
• 표본평균 $\bar{X}=80$, 표본의 크기 $n=16$이다.
• 95% 신뢰구간이므로 $\alpha=0.05$, $\frac{\alpha}{2}=0.025$이다. 또한, 모분산 $\sigma^2=16$이므로 모표준편차 $\sigma=\sqrt{16}=4$이다.
• 공식에 값들을 대입하면 $80-1.96\frac{4}{\sqrt{16}}\le\mu\le80+1.96\frac{4}{\sqrt{16}}$, 따라서 모평균에 대한 신뢰구간은 $78.04\le\mu\le81.96$이 된다.

39 해설 독립적인 카이제곱 분포가 있을 때, 두 확률변수의 비는 F-분포이다.

40 해설 • 표준오차는 $\left(\frac{\sigma}{\sqrt{n}}\right)$이고, $n>0$, $\sigma\ge0$이므로 항상 0 이상의 값을 가진다.
• 표본의 크기가 커지면 표준오차는 작아지고 모집단의 표준편차가 클수록 표준오차도 커진다.

41 해설 • 군집 추출은 모집단을 여러 군집으로 나누고, 일부 군집의 전체를 추출하는 방식이다.
• 100개의 구슬에 무작위로 검은색, 흰색, 빨간색을 칠하고 빨간색의 구슬을 모두 추출하는 기법이다.

42 해설 응집분석법은 각 객체를 하나의 소집단으로 간주하고 단계적으로 유사한 소집단들을 합쳐 새로운 소집단을 구성하는 것으로 군집 방법이다.

43 해설 • 독립변수의 조작에 따른 종속변수의 변화를 확인하여 두 변수 간의 관계를 파악할 때 사용하는 것은 회귀 분석이다.
• "구매자의 나이가 디지털 가전의 구매 유형에 어떤 영향을 미치는가?"는 회귀 분석을 사용한다.

44 해설 분석 모형 구축은 요건 정의, 모델링, 검증 및 테스트, 적용 단계로 진행한다.

분석 모형 구축 절차	
요모검적	요건 정의 / 모델링 / 검증 및 테스트 / 적용

12 해설 분류 모델은 다음과 같은 경우에 사용 가능하다.
• 신용평점자들에 대해서 저신용, 중간, 고신용 등과 같은 분류, 개별 고객의 재무 배경 및 구매 내역에 대한 데이터를 평가하는 경우 이를 "낮음", "중간" 또는 "높음" 신용 위험으로 분류할 때 사용한다.
• 모든 고객에게 광고물을 발송하지 않고 특정 고객만을 분류해서 광고물을 발송하여, 우편 비용을 줄일 때 사용한다.
• 고객이 신용카드를 사용할 때 분실 및 복제 등의 오용을 분류하여 탐지할 때 사용한다.

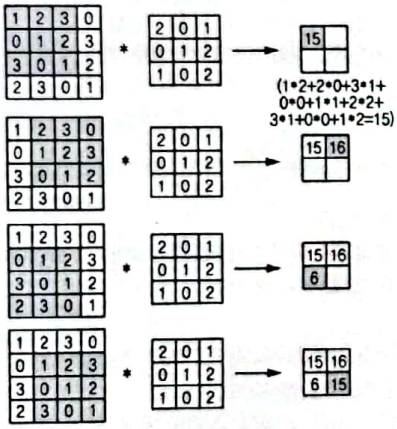
45 **해설** • 원본 이미지 필터 축의 크기가 4이고 필터 축의 크기가 3, 스트라이드가 1이므로 공식에 의해 Feature Map은 2x2이다.

$$\text{Feature Map}=\left(\frac{n+2p-f}{s}+1\right)\times\left(\frac{n+2p-f}{s}+1\right)$$

(원본 이미지 축의 크기 n, 필터 축의 크기 f, 패딩 p, 스트라이드 s)

$$\text{Feature Map}=\left(\frac{4+2\times0-3}{1}+1\right)\times\left(\frac{4+2\times0-3}{1}+1\right)$$
$$=2\times2$$

- 예를 들어, 아래와 같이 4*4 이미지를 3*3 필터로 계산하면 Feature Map은 2*2가 된다.
- 그림과 같이 입력과 필터에서 대응하는 원소끼리 곱한 후 그 총합을 구한다.



46 **해설** 필터 기능을 이용하여 입력 이미지로부터 특징을 추출한 뒤 신경망에서 분류작업을 수행하는 알고리즘은 CNN(Convolution Neural Network) 알고리즘이다.

47 **해설** 자기 조직화 지도(Self-Organizing Map)는 비지도 학습의 유형이다.

48 **해설** 깊이(Depth)는 뿌리 마디부터 끝마디까지의 중간 마디들의 수이다.

49 **해설** • 초매개변수는 모델에서 외적인 요소로 데이터 분석을 통해 얻어지는 값이 아니라 사용자가 직접 설정해주는 값이다.
• KNN에서 K는 사용자가 직접 설정해주는 값이므로 초매개변수이다.

50 **해설** • 부적합 변수 생성은 편향(Bias)을 발생시키지는 않으나 과대 적합을 발생시켜 예측 성능을 저하시킨다.

51 **해설** ReLU는 x값이 0보다 큰 경우에만 y값도 지속적으로 증가한다.

52 **해설**

회귀 모형 가정	
선독등분비상정	선형성 / 독립성 / 등분산성 / 비상관성 / 정상성

53 **해설** 서포트 벡터 머신(SVM: Support Vector Machine)은 훈련 시간이 상대적으로 느리지만, 정확성이 뛰어나며 다른 방법보다 과대적합의 가능성이 낮은 모델이다.

54 **해설**

지지도	전체 거래 중 사과, 우유를 모두 구매한 고객의 비율 $\frac{\text{사과}\cap\text{우유}}{\text{전체거래수}}=\frac{1,000}{10,000}=10\%$
신뢰도	사과를 구매한 고객 중 우유를 구매한 고객의 비율 $\frac{\text{사과}\cap\text{우유}}{\text{사과}}=\frac{1,000}{4,000+1,000}=20\%$

55 **해설** • 시계열 분석을 위해서는 정상성을 만족해야 한다.
• 정상성(Stationary)은 시점에 상관없이 시계열의 특성이 일정하다는 의미이다.

56 **해설** 독립변수가 한 개(x_1)인 경우 회귀계수 β_1부호에 따라 그래프의 형태는 S자 모양($\beta_1>0$) 또는 역 S자 모양($\beta_1<0$)을 가진다.

57 **해설** 시각적 이미지를 분석하는 데 사용되는 심층신경망으로 합성곱 신경망은 CNN(Convolution Neural Network)이다.

58 **해설** • 오피니언 마이닝(Opinion Mining)은 주관적인 의견이 포함된 데이터에서 사용자가 게재한 의견과 감정을 나타내는 패턴을 분석하는 기법이다.
• 긍정, 부정, 중립으로 선호도를 판별할 때 사용된다.

59 **해설** • 부트스트랩은 주어진 자료에서 동일한 크기의 표본을 랜덤 복원추출로 뽑은 자료를 의미한다.
• 훈련 데이터에서 다수의 부트스트랩 자료를 생성하고, 각 자료를 모델링한 후 결합하여 최종 예측 모형을 만드는 알고리즘은 배깅이다.

60 **해설** • 모수 통계로 검정이 가능한 데이터를 비모수 통계를 이용하면 효율성이 떨어진다.
• 비모수 통계 기법은 표본의 크기가 커질수록 간편하지만 지루한 반복 계산을 요구한다.

61 **해설** 혼동 행렬에서 실제로 '부정'인 범주 중에서 '부정'으로 올바르게 예측(TN)한 비율은 특이도이다.

$$\text{특이도}=\frac{TN}{TN+FP}$$

62 **해설** 일원 배치 분산 분석표는 다음의 표와 같다.
전체 n = 500이고 k = 3이므로 ㉠은 2 ㉡은 470이 된다.

요인	제곱합	자유도	제곱평균	F
집단 간	SSR	k − 1	MSR	MSR/MSE
집단 내	SSE	n − k	MSE	
총	SST	n − 1		

63 해설 • 계산량이 많지 않아 모형을 쉽게 평가할 수 있으나 전체 데이터에서 평가 데이터만큼은 학습에 사용할 수 없으므로 데이터 손실이 발생한다.
• 검증 데이터는 분류기들의 매개변수를 최적화하기 위해 사용하는 데이터이다.
• 최적화된 분류기의 성능을 평가할 때 사용하는 데이터는 평가 데이터이다.
• 데이터 집합을 무작위로 동일 크기를 갖는 K개의 부분 집합으로 나누고, 그중 1개를 평가 데이터(Test Set)로, 나머지 (K-1)개를 훈련 데이터(Training Set)로 선정하여 분석 모형을 평가하는 기법은 K-fold Cross Validation이다.
• 홀드 아웃 교차 검증은 전체 데이터를 비복원추출 방식을 이용하여 랜덤하게 훈련 데이터(Training Set)와 평가 데이터(Test Set)로 나눠 검증하는 기법이다.

64 해설 분산 분석(ANOVA: Analysis of Variance)은 두 개 이상의 집단 간 비교를 수행하고자 할 때 집단 내의 분산, 총 평균과 각 집단의 평균 차이에 의해 생긴 집단 간 분산 비교로 얻은 F-분포를 이용하여 가설검정을 수행하는 방법이다.

65 해설 • 카이제곱 검정은 가정된 확률을 검정하는 것이다.
• 데이터가 가정된 확률을 따르는 경우 귀무가설(H_0)을 채택한다.

66 해설 개별 가중치 값을 제한하여 복잡한 모델을 좀 더 간단하게 하는 방법은 가중치 규제(Weight Regularization)이다.

67 해설 드롭아웃은 과대 적합(Over-fitting)을 방지하기 위한 기법이다.

68 해설 데이터 시각화 기법은 다음과 같다.

시간 시각화	• 시간 흐름에 따른 변화를 통해 경향(트렌드) 파악	• 막대그래프 • 점그래프
분포 시각화	• 분류에 따른 변화를 최대, 최소, 전체 분포 등으로 구분 • 전체에서 부분 간 관계를 설명	• 파이 차트 • 도넛 차트 • 트리맵
관계 시각화	• 집단 간의 상관관계를 확인하여 다른 수치의 변화 예측	• 산점도 • 버블 차트 • 히스토그램
비교 시각화	• 각각의 데이터 간의 차이점과 유사성 관계도 확인 가능	• 히트맵 • 평행 좌표 그래프 • 체르노프 페이스
공간 시각화	• 지도를 통해 시점에 따른 경향, 차이 등을 확인 가능	• 등치선도 • 도트맵 • 카토그램

69 해설 • 버블 차트는 산점도에서 데이터값을 나타내는 점 또는 마크에 여러 가지 의미를 부여하여 확장된 차트이다.
• 히스토그램은 표로 되어 있는 도수 분포를 정보 그림으로 나타낸 그래프이다.
• 히트맵은 여러 가지 변수를 비교할 수 있는 시각화 그래프이다.

70 해설 • 특이도(Specificity)=TN/(TN+FP)=70/(70+30)=70/100=7/10
• 정밀도(Precision)= TP/(TP+FP) = 25/(25+30)=25/55=5/11

71 해설 • 공간 시각화의 유형으로 등치지역도, 등치선도, 도트 플롯맵, 버블 플롯맵, 카토그램 등이 있다.
• 히스토그램은 다변량 데이터 사이에 존재하는 변수 사이의 연관성, 분포와 패턴을 찾는 관계 시각화 방법이다.

72 해설 • 타임라인형은 주제를 선정하여 관련된 히스토리를 타임라인 형태로 나타내는 방식이다.
• 특정 제품군의 주요 제품 비교 등에는 비교분석형을 사용한다.

73 해설 ROC곡선은 가로축을 거짓 긍정률(FP Rate, 1-Specificity)로, 세로축을 참 긍정률(TP Rate)로 두어 시각화한 그래프이다.

74 해설 • 버블 차트는 산점도에서 데이터값을 나타내는 점 또는 마크에 여러 가지 의미를 부여하여 확장된 차트이다.

관계 시각화 유형	
산점버하네	산점도 / 산점도 행렬 / 버블 차트 / 히스토그램 / 네트워크 그래프

75 해설 모니터링을 수작업으로 하게 되면 개발된 모델이 많아질수록 과업이 늘어날 수 있으니 DBMS에 성과자료를 누적하여 자동으로 모니터링하고 이상 시에만 확인하는 프로세스를 수립한다.

76 해설 빅데이터 활용 분야를 검토하기 위해 초기 아이디어 개발 관점을 분류할 때에는 마인드맵, 친화 도표, 피라미드 구조와 같은 시각적인 방법을 이용해서 아이디어를 분류할 수 있다.

77 해설 여러 가지 변수를 비교할 수 있는 시각화 그래프이며, 칸 별로 색상을 구분하여 데이터값을 표현한 것은 히트맵이다.

78 해설 지도상의 위도와 경도에 해당하는 좌표점에 산점도와 같이 점을 찍어서 표현하고, 시간의 경과에 따라 점진적으로 확산을 나타내는 경우에 사용하는 공간 시각화는 도트 플롯맵이다.

79 해설 부트스트랩을 통해 100개의 샘플을 추출하더라도 샘플에 한 번도 선택되지 않는 원 데이터가 발생할 수 있다.

80 해설 • 실제로 '부정'인 범주 중에서 '부정'으로 올바르게 예측(TN)한 비율은 특이도(Specificity)이다.
• 정밀도(Precision)는 '긍정'으로 예측한 비율 중에서 실제로 '긍정'(TP)인 비율이다.

2회 정답

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

01 **해설** 10^9Bytes는 기가바이트(GB)에 해당한다.

킬로바이트(KB)	10^3Bytes
메가바이트(MB)	10^3KB = 10^6Bytes
기가바이트(GB)	10^3MB = 10^9Bytes
테라바이트(TB)	10^3GB = 10^{12}Bytes
페타바이트(PB)	10^3TB = 10^{15}Bytes
엑사바이트(EB)	10^3PB = 10^{18}Bytes
제타바이트(ZB)	10^3EB = 10^{21}Bytes
요타바이트(YB)	10^3ZB = 10^{24}Bytes

02 **해설** 마이데이터란 정보주체가 개인 데이터에 대한 열람, 제공 범위, 접근 승인 등을 직접 결정함으로써 개인의 정보 활용 권한을 보장, 데이터 주권을 확립하는 패러다임, 서비스 등을 통칭한다.

03 **해설** 데이터 거버넌스 구성요소는 아래와 같다.

원칙 (Principle)	• 데이터를 유지 · 관리하기 위한 지침과 가이드 • 품질기준, 보안, 변경관리
조직 (Organization)	• 데이터를 관리할 조직의 역할과 책임(R&R) • 데이터 관리자, 데이터베이스 관리자(DBA), 데이터 아키텍트 등
프로세스 (Process)	• 데이터 관리를 위한 활동과 체계 • 작업 절차, 모니터링 활동, 측정 활동 등

04 **해설** • 조직 구조에는 집중, 기능, 분산 구조가 있다.
3 • 기능 구조는 일반적인 형태로 별도 분석 조직이 없고, 해당 부서에서 분석 수행한다.

05 **해설** 개인정보 비식별화 절차는 다음과 같다
123

사전검토	• 데이터가 개인정보에 해당하는지 검토 • 개인정보가 아닐 경우 법적 규제 없이 자유롭게 활용
비식별 조치	• 데이터 집합에서 개인을 식별할 수 있는 요소를 전부 또는 일부 삭제하거나 대체하는 등의 방법을 활용해 개인을 알아볼 수 없도록 하는 조치
적정성 평가	• 다른 정보와 쉽게 결합하여 개인을 식별할 수 있는지를 비식별 조치, 적정성 평가단을 통해 평가
사후관리	• 비식별 정보 안전조치, 재식별 가능성 모니터링 등 비식별 정보 활용 과정에서 재식별 방지를 위해 필요한 조치 수행

06 **해설** CRISP-DM 절차는 다음과 같다.
12

업무 이해 (Business Understanding)	• 각종 참고 자료와 현업 책임자와의 커뮤니케이션을 통해 비즈니스를 이해하는 단계
데이터 이해 (Data Understanding)	• 분석을 위한 데이터를 수집 및 속성을 이해하고, 문제점을 식별하며 숨겨져 있는 인사이트를 발견하는 단계
데이터 준비 (Data Preparation)	• 데이터 정제, 새로운 데이터 생성 등 자료를 분석 가능한 상태로 만드는 단계
모델링 (Modeling)	• 다양한 모델링 기법과 알고리즘을 선택하고 파라미터를 최적화하는 단계
평가 (Evaluation)	• 데이터 정제, 새로운 데이터 생성 등 자료를 분석 가능한 상태로 만드는 단계 • 평가에 많은 시간이 소요를 수행
전개 (Deployment)	• 데이터 정제, 새로운 데이터 생성 등 자료를 분석 가능한 상태로 만드는 단계 • 전개에 많은 시간이 소요

07 **해설** • 주어진 데이터 집합에서 함께 비식별 되는 레코드들은(동질 집합에서) 적어도 몇 개의 서로 다른 민감한 정보를 가져와야 하는 프라이버시 모델은 *l*-다양성이다.
• 프라이버시 보호 모델에는 k-익명성, *l*-다양성, t-근접성, m-유일성 등이 있다.

k-익명성 (k-Anonymity)	• 주어진 데이터 집합에서 같은 값이 적어도 k개 이상 존재하도록 하여 쉽게 다른 정보로 결합할 수 없도록 하는 모델 • 공개된 데이터에 대한 연결 공격 취약점을 방어하기 위한 모델
l-다양성 (*l*-Diversity)	• 주어진 데이터 집합에서 함께 비식별 되는 레코드들은(동질 집합에서) 적어도 *l*개의 서로 다른 민감한 정보를 가져야 하는 프라이버시 모델 • 비식별 조치 과정에서 충분히 다양한(*l*개 이상) 서로 다른 민감한 정보를 갖도록 동질 집합을 구성

	• k-익명성에 대한 두 가지 취약점 공격인 동질성 공격, 배경 지식에 의한 공격을 방어하기 위한 프라이버시 모델
l-근접성 (l-Closeness)	• 동질 집합에서 특정 정보의 분포와 전체 데이터 집합에서 정보의 분포가 l 이하의 차이를 보여야 하는 모델 • l-다양성의 쏠림 공격, 유사성 공격을 보완하기 위해 제안된 모델
m-유일성 (m-Uniqueness)	• 원본 데이터와 동일한 속성 값의 조합이 비식별 결과 데이터에 최소 m개 이상 존재하도록 하여 재식별 가능성 위험을 낮춘 모델

8 **해설** • 출생년도, 성별 외에 개인식별에 중요한 나머지 값을 삭제하였으므로 데이터 삭제에 해당한다.
• 주민등록번호에서 연도 정보와 성별(남자) 정보만 남기고 주민등록번호는 삭제처리한다.

9 **해설** 분석 로드맵은 3단계로서 데이터 분석체계 도입, 데이터 분석 유효성 검증, 데이터 분석 확산 및 고도화의 단계로 이루어진다.

10 **해설** • 빅데이터 분석은 분석의 대상과 방법에 따라 최적화, 솔루션, 통찰, 발견의 4가지로 분류한다.
• 분석의 대상이 명확하게 무엇인지 모르는 경우 기존 분석 방식을 활용하여 새로운 지식을 도출하는 것은 통찰이다.

11 **해설** 빅데이터는 전통적으로 3V(Volume, Variety, Velocity)의 특징이 있지만, 최근에는 4V(Value 추가), 5V(Veracity, Value 추가), 7V(Validity, Volatility 추가)로 확장되고 있다.

12 **해설** 분석과제의 적용 우선순위 기준을 '시급성'에 둔다면 Ⅲ → Ⅳ → Ⅱ 영역 순이며, 우선순위 기준을 '난이도'에 둔다면 Ⅲ → Ⅰ → Ⅱ 영역 순으로 의사결정을 할 수 있다.

13 **해설** CRISP-DM 분석 방법론 단계는 업무 이해 → 데이터 이해 → 데이터 준비 → 모델링 → 평가 → 전개이다.

14 **해설** 개인정보를 제공하기 위해 정보주체의 동의를 받을 때 고지사항(개인정보보호법 17조 2항)

개인정보를 제공하기 위해 정보주체의 동의를 받을 때 고지사항	
자목항기불	개인정보를 제공받는 자 / 개인정보의 수집·이용 목적 / 수집하려는 개인정보의 항목 / 개인정보의 보유 및 이용 기간 / 동의를 거부할 권리가 있다는 사실 및 동의 거부에 따른 불이익이 있는 경우에는 그 불이익의 내용

15 **해설** 반정형 데이터는 스키마(형태) 구조 형태를 가지고 메타데이터를 포함하며, 값과 형식에서 일관성을 가지지 않는 데이터로 XML, HTML, JSON 등이 있다.

16 **해설** 명확화는 데이터 집합에 존재하는 잡음으로 인해 거칠게 분포된 데이터를 매끄럽게 만들기 위해 군집화, 군집화 등의 기법 적용된다.

17 **해설** 예측 알고리즘을 통한 판단을 근거로 불이익을 줄 수 없다.

18 **해설** 다른 사람과의 대화 등 상호 작용을 통해 개인이 암묵지를 습득하는 단계는 공통화 단계이다.

19 **해설** 시행착오를 통한 문제 해결을 위해 사용되는 상향식 접근법은 프로토타이핑 접근법이다.

20 **해설** • 개인정보 유출 시 정보주체에게 고지해야 할 사항(개인정보보호법 34조 1항)

개인정보 유출 시 정보주체에게 고지해야 할 사항	
항시주대부	유출된 개인정보의 항목 / 유출된 시점과 그 경위 / 유출로 인하여 발생할 수 있는 피해를 최소화하기 위하여 정보주체가 할 수 있는 방법 등에 관한 정보 / 개인정보처리자의 대응조치 및 피해 구제절차 / 정보주체에게 피해가 발생한 경우 신고 등을 접수할 수 있는 담당부서 및 연락처

21 **해설** • 실시간 이벤트 처리 기술에는 CEP(Complex Event Processing)가 있다.
• IoT 센싱 데이터, 로그, 음성 데이터 등 실시간 데이터 처리가 가능하다.

22 **해설** • 단순 대치법은 결측값을 그럴듯한 값으로 대체하는 통계적 기법으로 종류에는 완전 분석법, 평균 대치법, 단순 확률 대치법이 있다.
• 불완전 자료는 모두 무시하고 완전하게 관측된 자료만 사용하여 분석하는 방법은 완전 분석법이다.

23 **해설** 로그 변환은 단순 기능 변환 중 한 가지 방법이다.

단순 기능 변환	• 한쪽으로 치우친 변수를 변환하여 분석 모형을 적합하게 하는 방법
비닝	• 데이터값을 몇 개의 버킷으로 분할하여 계산하는 방법
정규화	• 데이터를 특정 구간으로 바꾸는 척도법

24 **해설** • 특정 모델링 기법에 의존하지 않고 데이터의 통계적 특성으로부터 변수를 택하는 기법은 필터 기법(Filter Method)이다.

25 **해설** • 변수들의 공분산 행렬이나 상관행렬을 이용하고 행의 수와 열의 수가 같은 정방행렬에서만 사용한다.
• 원래의 데이터 세트의 변수들을 선형 변환하여 서로 직교하도록 선택된 새로운 변수들(주성분)을 생성, 이를 통해 원래 변수를 설명하고자 하는 기법은 주성분 분석이다.

26 **해설** 과소 표집 기법은 다음과 같다.

랜덤 과소 표집 (Random Under-Sampling)	• 무작위로 다수 클래스 데이터의 일부만 선택하는 방법
ENN (Edited Nearest Neighbours)	• 소수 클래스 주위에 인접한 다수 클래스 데이터를 제거하여 데이터의 비율을 맞추는 방법
토멕 링크 방법 (Tomek Link Method)	• 토멕 링크(Tomek Link)는 클래스를 구분하는 경계선 가까이에 존재하는 데이터 • 다수 클래스에 속한 토멕 링크를 제거하는 방법
CNN (Condensed Nearest Neighbor)	• 다수 클래스에 밀집된 데이터가 없을 때까지 데이터를 제거하여 데이터 분포에서 대표적인 데이터만 남도록 하는 방법
OSS (One Sided Selection)	• 토멕 링크 방법과 Condensed Nearest Neighbor 기법의 장점을 섞은 방법 • 다수 클래스의 데이터를 토멕 링크 방법으로 제거한 후 Condensed Nearest Neighbor를 이용하여 밀집된 데이터 제거

27 **해설** 포아송 분포는 다음과 같다.

$P=\frac{\lambda^n e^{-\lambda}}{n!}$	• λ: 정해진 시간/영역 안에 어떤 사건이 일어날 횟수에 대한 기댓값 • n: 정해진 시간/영역 안에 사건이 일어나는 횟수
기댓값: $E(X)=\lambda$ 분산: $V(X)=\lambda$	

28 **해설** 과소 표집은 다수 클래스의 데이터를 일부만 선택하여 데이터의 비율을 맞추는 방법이다.

29 **해설** • 데이터의 순서에 의미를 부여한 데이터 변수는 스피어만(Spearman) 순위 상관 분석을 통해서 분석한다.
• 변수의 속성에 따른 상관성 분석 방법의 분류는 다음과 같다.

수치적 데이터	피어슨 상관계수
순서적 데이터	스피어만 순위 상관 분석
명목적 데이터	카이제곱(χ^2) 검정(교차분석)

30 **해설** 기댓값의 공식에 의해서 다음과 같이 계산한다.

$$E(X)=\sum xf(x)=\left(1\times\frac{1}{6}\right)+\left(2\times\frac{2}{6}\right)+\left(3\times\frac{1}{6}\right)+\left(4\times\frac{2}{6}\right)$$
$$=\frac{16}{6}=\frac{8}{3}$$

31 **해설** 최빈수는 가장 많이 관측되는 수로, 5가 2회 관측되어 최빈수에 해당한다.

32 **해설** • $P(L)$: League Of Legend 플레이 경험
• $P(A)$: A 집단의 학생 = 80%
• $P(L|A)$: A 집단에서 League Of Legend 플레이 경험이 있는 확률 = 50%,
• $P(B)$: B 집단의 학생 = 20%
• $P(L|B)$: B 집단에서 League Of Legend 플레이 경험이 있는 확률 = 20%,일 때 $P(B|L)$를 구하는 문제이다.
• 베이즈 정리에 의해서

$$P(B|L)=\frac{P(L|B)\times P(B)}{P(L|A)\times P(A)+P(L|B)\times P(B)}$$
$$=\frac{(20\%\times 20\%)}{(80\%\times 50\%)+(20\%\times 20\%)}=\frac{0.04}{0.4+0.04}=\frac{1}{11}$$

33 **해설** • 모평균 추정 시 신뢰구간의 길이는 표준오차에 비례하고 표본의 크기의 제곱근에 반비례한다.
• 표본의 크기를 100에서 400으로 4배 증가시켰으므로 신뢰구간의 길이는 $\frac{1}{\sqrt{4}}=\frac{1}{2}$배 감소한다. 따라서 신뢰구간의 길이는 $20\times\frac{1}{2}=10$이 된다.

34 **해설** • 일변량 데이터 탐색 방법에는 기술 통계량, 그래프 통계량 두 가지 종류가 있다.
• 기술 통계량에는 평균, 분산, 표준편차 등이 있고 그래프 통계량에는 히스토그램, 상자 그림 등이 있다.
• 다변량 데이터 탐색 도구로는 산점도 행렬, 별 그림, 겨냥도 그림이다.

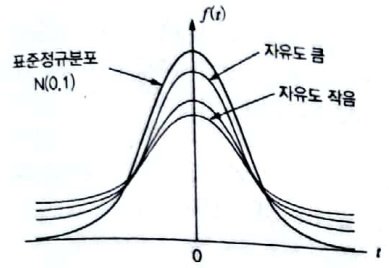
35 **해설** 확률변수에 대한 분산의 성질에서
$V(X-Y)=V(X)+V(Y)-2Cov(X,Y)$이며,
서로 독립일 경우
$Cov(X,Y)=0$이므로 $V(X-Y)=V(X)+V(Y)$이다.

36 **해설**

유의수준 (Level of Significance)	• 제1종 오류를 범할 최대 허용확률을 의미 • α로 표기
신뢰수준 (Level of Confidence)	• 귀무가설이 참일 때 이를 참이라고 판단하는 확률($1-\alpha$)
베타 수준 (β Level)	• 제2종 오류를 범할 최대 허용확률을 의미 • β로 표기
검정력	• 귀무가설이 참이 아닌 경우 이를 기각할 수 있는 확률($1-\beta$)

37 **해설** 표본평균의 표준편차는 표준오차이다.
$n=9, \sigma=\sqrt{25}=5$이므로, 표준오차는 $\frac{\sigma}{\sqrt{n}}=\frac{5}{\sqrt{9}}=\frac{5}{3}$이다.

38 **해설** T-분포는 정규분포의 평균(μ)의 해석에 많이 쓰이는 분포이다.



▲ T-분포

39 **해설** • 세 번(n번) 시행 중에 한 번(k번) 성공할 확률이므로 이항분포이다.
• 동전을 세 번 던지기 때문에 $n=3$, 동전 앞면이 나올 확률 $p=0.5$, 앞면이 한 번 나오므로 $k=1$이다.
• $P=\binom{n}{1}p^k(1-p)^{n-k}=\binom{3}{2}p^1(1-0.5)^{3-1}=\frac{3!}{1!(3-1)!}0.5\times 0.5^2=3\times 0.5^3=0.375$이다.

40 **해설** • 모분산을 모르는 대표본($n\geq 30$)일 경우 $100\times(1-\alpha)$ 신뢰구간을 구하는 공식은 $\overline{X}-Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\leq\mu\leq\overline{X}+Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$이다.
• 표본평균 $\overline{X}=175$, 표본의 크기 $n=100$이다.
• 95% 신뢰구간이므로 $\alpha=0.05$, $\frac{\alpha}{2}=0.025$이다.
• 또한, 표본 표준편차 $s=\sqrt{25}=5$이므로 공식에 값들을 대입하면 $175-1.96\frac{5}{\sqrt{100}}\leq\mu\leq 175+1.96\frac{5}{\sqrt{100}}$
• 따라서 모평균에 대한 신뢰구간은 $174.02\leq\mu\leq 175.98$이 된다.

41 **해설** CART는 목적변수가 이산형일 경우에 불순도의 측도로 지니 지수를 이용한다.

42 **해설** • 매개변수는 모델 내부에서 확인이 가능한 변수로 데이터를 통해서 산출이 가능한 값이다.
• 매개변수의 예시로 인공신경망에서의 가중치, 서포트 벡터 머신에서의 서포트 벡터, 선형 회귀나 로지스틱 회귀 분석에서의 결정계수가 있다.
• 신경망 학습에서 학습률(Learning Rate)은 초매개변수 예시이다.

43 **해설** • 요건 정의는 기획단계의 분석 과제 정의를 통해 도출된 내용을 구체화하는 과정이다.
• 검증 및 테스트는 분석용 데이터를 학습용과 테스트용으로 분리한 다음 분석용 데이터를 이용해 자체 검증 후 실제 테스트에서는 신규 데이터 모델을 적용해 결과를 도출하는 단계이다.
• 적용은 분석결과를 업무 프로세스에 완전히 통합해 실제 일, 주, 월 단위로 운영하는 단계이다.
• 요건 정의에 따라 상세 분석 기법을 적용해 모델을 개발하는 과정은 모델링이다.

44 **해설** 파이썬은 고급 프로그래밍 언어로 플랫폼에 독립적이며, 객체지향적 인터프리터식 언어이다.

45 **해설** 2종 오류인 잘못된 귀무가설을 채택하는 오류를 방지하는 데 목적이 있다.

46 **해설**

단순선형회귀	$Y=\beta_0+\beta_1X+\epsilon$
다중선형회귀	$Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k+\epsilon$
다항 회귀	K=2이고 2차 함수인 경우 $Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_{11}X_1^2+\ldots+\beta_{22}X_2^2+\beta_{12}X_1X_2+\epsilon$
곡선 회귀	2차 곡선인 경우: $Y=\beta_0+\beta_1X+\beta_2X^2+\epsilon$ 3차 곡선인 경우: $Y=\beta_0+\beta_1X+\beta_2X^2+\beta_3X^3+\epsilon$
로지스틱 회귀	$\pi(x)=\frac{\exp(\alpha+\beta_1x_1+\cdots+\beta_kx_k)}{1+\exp(\alpha+\beta_1x_1+\cdots+\beta_kx_k)}$
비선형 회귀	$Y=\alpha e^{-\beta X}+\epsilon$

47 **해설** 승산비는 교차비라고도 하며 $\text{odds}=\frac{p}{1-p}$로 계산한다.

48 **해설**

의사결정나무의 분석 과정	
성가타해	의사결정나무 성장 / 가지치기 / 타당성 평가 / 해석 및 예측

49 **해설** • 자기 회귀 모형은 현시점의 자료가 p 시점 전의 유한개의 과거 자료로 설명될 수 있다는 의미이다.
• 백색잡음은 모든 개별 확률변수들이 서로 독립이고 동일한 확률분포를 따르는 확률 과정을 말한다.
• 분해 시계열은 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법이다.

50 **해설** 활성화 함수는 인공신경망에서 순 입력함수로부터 전달받은 값을 출력값으로 변환해 주는 함수이다.

정답과 해설

서포트 벡터 머신의 구성요소	
결초마서슬	결정 경계 / 초평면 / 마진 / 서포트 벡터 / 슬랙 변수

51 **해설** 지지도는 전체 거래 중 항목 A와 B를 동시에 포함하는 거래의 비율이다.

따라서 $\frac{\text{커피, 빵 동시구매 거래 수}}{\text{전체 거래 수}} = \frac{50}{100} = \frac{1}{2}$이다.

52 **해설** 군집 내의 오차 제곱합(Error Sum of Square)에 기초하여 군집을 수행하는 기법은 와드 연결법이다.

최단연결법	두 군집 사이의 거리를 각 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 거리의 최솟값으로 측정
최장연결법	두 군집 사이의 거리를 각 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 거리의 최댓값으로 측정
중심연결법	두 군집이 결합될 때 새로운 군집의 평균은 가중 평균을 통해 구함
평균연결법	모든 항목에 대한 거리 평균을 구하면서 군집화

53 **해설** • 매개변수는 경험에 의해 정해지기도 하며, 예측 알고리즘 모델링의 문제점을 위해 조절한다.

54 **해설**

의사결정나무 성장(Growing)	분석의 목적과 자료구조에 따라서 적절한 분리 규칙을 찾아서 나무를 성장시키는 과정
가지치기 (Pruning)	분류 오류를 크게 할 위험이 높거나 부적절한 추론규칙을 가지고 있는 가지 또는 불필요한 가지를 제거하는 단계
타당성 평가	이익 도표, 위험 도표 또는 평가 데이터를 이용하여 교차 타당성 등을 이용한 평가 수행 단계
해석 및 예측	구축된 의사결정나무 모형을 해석하고, 분류 및 예측 모형을 설정하여 데이터의 분류 및 예측에 활용하는 단계

55 **해설** 퍼셉트론은 인공신경망의 한 종류로서, 1957년에 코넬 항공 연구소의 프랑크 로젠블라트에 의해 고안되었다.

56 **해설** 활성화 함수에는 계단함수, 부호함수, 선형함수, 시그모이드 함수, tanh 함수, ReLU 함수가 있다.

57 **해설** 자카드(Jaccard) 계수에 대한 설명으로 자카드 계수는 두 집합에 대한 합집합과 교집합에 대한 비(Proportion)이다.

58 **해설** • 텍스트 마이닝은 인간이 이해할 수 있는 언어를 기계가 이해할 수 있도록 하는 자연어 처리 기술에 기반한다.
• 텍스트 마이닝 절차는 텍스트 수집, 의미추출, 패턴 분석, 정보 생성이다.

59 **해설** SVM은 훈련 시간이 상대적으로 느리지만, 정확성이 뛰어나며 다른 방법보다 과대 적합의 가능성이 낮은 모델이다.

60 **해설** • 배깅은 훈련 데이터에서 다수의 부트스트랩(Bootstrap) 자료를 생성하고, 각 자료를 모델링한 후 결합하여 최종 예측 모형을 만드는 알고리즘이다.
• 보팅은 여러 개의 머신러닝 알고리즘 모델을 학습시킨 후 새로운 데이터에 대해 각 모델의 예측값을 가지고 다수결 투표를 통해 최종 클래스를 예측하는 기법이다.
• 랜덤 포레스트는 의사결정나무의 특징인 분산이 크다는 점을 고려하여 배깅과 부스팅보다 더 많은 무작위성을 주어 약한 학습기들을 생성한 후 이를 선형 결합하여 최종 학습기를 만드는 방법이다.

61 **해설** • SST는 전체 제곱합으로 실젯값과 평균값의 차이의 제곱합이다.
• SSR은 제곱 잔차 합계로 예측값과 평균값의 차이(잔차) 제곱합이다.
• AE는 평균 오차로 예측한 결괏값의 오류 평균이다.
• 예측값과 실젯값 차이(오차)의 제곱합은 SSE이다.

62 **해설** 일반화 오류는 과대 적합, 학습오류는 과소 적합되었다고 한다.

63 **해설** • 랜덤 서브샘플링은 모집단으로부터 조사의 대상이 되는 표본을 무작위로 추출하는 기법이다.
• 홀드 아웃은 전체 데이터를 비복원추출 방식을 이용하여 랜덤하게 훈련 데이터(Training Set)와 평가 데이터(Test Set)로 나눠 검증하는 기법이다.
• LOOCV은 전체 데이터에서 1개 샘플 만을 Test에 사용하고 나머지 (N-1)개는 학습에 사용하고, 이 과정을 N번 반복하는 기법이다.
• 데이터 집합을 무작위로 동일 크기를 갖는 K개의 부분 집합으로 나누고, 그중 1개를 평가 데이터(Test Set)로, 나머지 (K-1)개를 훈련 데이터(Training Set)로 선정하여 분석 모형을 평가하는 기법은 K-fold Cross Validation이다.

64 **해설** • Z-검정은 귀무가설에서 검정 통계량의 분포를 정규분포로 근사할 수 있는 통계 검정이다.
• T-검정은 두 집단 간의 평균을 비교하는 모수적 통계 방법으로서 표본이 정규성, 등분산성, 독립성 등을 만족할 경우 적용한다.
• 분산 분석(ANOVA)은 두 개 이상의 집단 간 비교를 수행하고자 할 때 집단 내의 분산, 총평균과 각 집단의 평균 차이에 의해 생긴 집단 간 분산 비교로 얻은 F-분포를 이용하여 가설검정을 수행하는 방법이다.
• 카이제곱 검정은 범주에 따라 분류된 변수가 정규분포되어 있다면 빈도가 실제 기대되는 값으로부터 유의미한 차이가 관찰되는가를 보기 위한 검증 방법이다.

65 **해설** • 카이제곱 검정은 가정된 확률을 검정하는 것이다.
• 적합도 검정기법은 다음과 같다.

가정된 확률 검정	카이제곱 검정
정규성 검정	샤피로-윌크 검정, 콜모고로프-스미르노프 적합성 검정, Q-Q Plot

66 **해설** • 과대 적합은 모델이 훈련 데이터에 너무 잘 맞지만, 일반화가 떨어지는 현상이다.
• 과대 적합 방지는 데이터 세트를 증가시켜서 방지한다.

67 **해설** • 매개변수에는 가중치와 편향이 있다.
• 가중치는 각 입력값에 각기 다르게 곱해지는 수치이다.
• 매개변수 중 하나의 뉴런에 입력된 모든 값을 더한 값(가중합)에 더해주는 상수는 편향이다.

68 **해설** • 직접 투표(Hard Voting)는 단순 투표 방식으로 개별 모형의 결과 기준이다.
• 배깅은 훈련 데이터의 중복을 허용하며, 훈련 데이터 세트를 나누는 기법으로 복원추출 방식이다.
• 랜덤 서브스페이스는 훈련 데이터를 모두 사용하고 특성은 샘플링하는 방식이다.

69 **해설** • 부스팅 방법론에는 에이다 부스트와 그레디언트 부스트가 있다.
• 에이다 부스트는 약한 모형들을 순차적으로 적용해 나가는 과정에서 잘 분류된 샘플의 가중치는 낮추고 잘못 분류된 샘플의 가중치는 상대적으로 높여주면서 샘플 분포를 변화시키는 기법이다.

70 **해설** 도넛 차트는 분포 시각화 기법이다.

71 **해설** • TCO(Total Cost of Ownership)는 총 소유 비용이다.
• TCO는 하나의 자산을 획득하려 할 때 주어진 기간 동안 모든 연관 비용을 고려할 수 있도록 확인하기 위해 사용되는 평가 기법이다.

72 **해설**

막대그래프	동일한 너비의 여러 막대를 사용하여 데이터를 표시하며, 각 막대는 특정 범주를 나타내는 그래프
선 그래프	수량을 점으로 표시하고, 점들을 선분으로 이어 그린 그래프
영역 차트	선 그래프와 같이 시간에 값에 따라 크기 변화를 보여주는 그래프
누적 막대그래프	막대를 사용하여 전체 비율을 보여주면서 여러 가지 범주를 동시에 차트로 표현 가능한 그래프

73 **해설** • 산점도 행렬은 다변량 변수를 갖는 데이터에서 가능한 모든 변수 쌍에 대한 산점도를 행렬 형태로 표현한 그래프이다.
• 버블 차트는 산점도에서 데이터값을 나타내는 점 또는 마크에 여러 가지 의미를 부여하여 확장된 차트이다.

74 **해설** • 플로팅 바 차트는 막대가 가장 낮은 수치부터 가장 높은 수치까지 걸쳐있게 표현한 차트이다.
• 체르노프 페이스는 데이터를 눈, 코, 귀, 입 등과 일대일 대응하여 얼굴 하나로 표현하는 방법이다.
• 스타 차트는 각 변수를 표시 지점을 연결선을 통해 그려 별 모양의 도형으로 나타낸 차트이다.
• 여러 가지 변수를 비교할 수 있는 시각화 그래프는 히트맵이다.

75 **해설** • 스토리텔링형은 하나의 사건이나 주제에 대해 이야기를 들려주는 구성방식이다.
• 만화형이 캐릭터 등의 만화적 요소를 활용한 방식이다.

76 **해설** • 특이도(Specificity)=TN/(TN+FP)=45/(45+15)=45/60 =3/4
• 정밀도(Precision)= TP/(TP+FP) = 5/(5+15)=5/20=1/4

77 **해설** 마인드맵은 줄거리를 이해하며 정리하는 방법으로 많이 이용된다.

78 **해설** 초매개변수로 설정 가능한 예시로는 학습률(Learning Rate), 의사결정나무의 깊이(Depth), 신경망에서 은닉층(Hidden Layer)의 개수 등이 있다.

79 **해설** 기존 데이터 집합에 대한 데이터 오류율도 점검한다.

80 **해설** 혼동행렬 관련 주요 평가지표는 다음과 같다.

특이도 (Specificity)	TN/(TN+FP)	실제로 '부정'인 범주 중에서 '부정'으로 올바르게 예측(TN)한 비율
민감도 (Sensitivity)	TP/(TP+FN)	실제로 '긍정'인 범주 중에서 '긍정'으로 올바르게 예측(TP)한 비율
거짓 긍정률 (FP Rate)	FP/(TN+FP)	실제로 '부정'인 범주 중에서 '긍정'으로 잘못 예측(FP)한 비율
정밀도 (Precision)	TP/(TP+FP)	'긍정'으로 예측한 비율 중에서 실제로 '긍정'(TP)인 비율
정확도 (Accuracy)	(TP+TN)/(TP+TN+FP+FN)	전체 예측에서 실제로 '긍정'(TP)와 '부정'으로 올바르게 예측(TN)이 차지하는 비율
오차비율 (Error Rate)	(FP+FN)/(TP+TN+FP+FN)	실제 분류 범주를 잘못 분류한 비율

정답 및 해설

3회 정답

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

01 **해설** DIKW 피라미드에서 근본 원리에 대한 깊은 이해를 바탕으로 도출되는 창의적 아이디어는 지혜(Wisdom)이다.

02 **해설** 정형 데이터뿐만 아니라 비정형, 반정형 데이터를 포함하는 특징은 다양성(Variety)에 대한 특징이다.

03 **해설** 형식지가 상호결합하면서 새로운 형식지를 창출하는 과정은 연결화이다.

04 **해설**

KDD 분석 방법론 분석 절차	
선전변마평	데이터 셋트 선택 / 데이터 전처리 / 데이터 변환 / 데이터 마이닝 / 데이터 마이닝 결과 평가

05 **해설** 전사적 핵심 분석이 어려우며 과거에 국한된 분석을 수행하는 구조는 기능 구조이다.

06 **해설**

데이터 사이언티스트의 요구 역량	
협통전 숙지	(소프트 스킬) 협력 능력 / 통찰력 / 전달력 (하드 스킬) 숙련도 / 지식

07 **해설** • ETL은 수집 대상 데이터를 추출, 가공(변환, 정제)하여 데이터 웨어 하우스 및 데이터 마트에 저장하는 데이터 수집 기술이다.

• PDEMS는 2차원 테이블인 데이터 모델에 기초를 둔 관계형 데이터베이스를 생성하고 수정하고 관리할 수 있는 소프트웨어이다.

08 **해설**

개인정보를 제공하기 위해 정보주체의 동의를 받을 때 고지사항	
「자목항기불」	개인정보를 제공받는 자 / 개인정보를 제공받는 자의 개인정보 이용 목적 / 제공하는 개인정보의 항목 / 개인정보를 제공받는 자의 개인정보 보유 및 이용 기간 / 동의를 거부할 권리가 있다는 사실 및 동의 거부에 따른 불이익이 있는 경우에는 그 불이익의 내용

09 **해설** 개정된 가명처리가이드 라인에 따라 개인정보처리자는 정당한 처리 범위 내에서 통계작성, 과학적 연구, 공익적 기록보존 등의 목적으로 정보주체의 동의 없이 가명정보를 처리할 수 있다.

통계 작성	• 통계란 특정 집단이나 대상 등에 관하여 작성한 수량적인 정보 • 시장조사와 같은 상업적 목적의 통계 처리도 포함됨
과학적 연구	• 과학적 연구는 기술의 개발과 실증, 기초연구, 응용연구 및 민간 투자 연구 등 과학적 방법을 적용하는 연구
공익적 기록보존	• 공공의 이익을 위하여 지속적으로 열람할 가치가 있는 정보를 기록하여 보존하는 것

10 **해설** 비지도 학습 방법 및 프로토타이핑 접근법을 사용해서 분석하는 접근 방식은 상향식 접근 방식이다.

11 **해설** 스톰은 빅데이터 분석 방법론 계층에서 입력자료 처리 및 도구, 출력자료로 구성된 단위 프로세스이다.

12 **해설**

개인 정보	• 특정 개인에 관한 정보 • 개인을 알아볼 수 있게 하는 정보
가명 정보	• 추가정보의 사용 없이는 특정 개인을 알아볼 수 없게 조치한 정보
익명 정보	• 더 이상 개인을 알아볼 수 없게(복원 불가능한 정도로) 조치한 정보

13 **해설** SEMMA 분석 방법론의 분석 절차는 샘플링, 탐색, 수정, 모델링, 검증의 5단계로 되어 있다.

14 **해설** 센서 데이터, 장비 간 발생 로그, LOD 등은 외부 데이터이다.

15 해설 개인정보 익명처리 기법은 아래와 같다.

가명 (Pseudonym)	개인 식별이 가능한 데이터에 대하여 직접 식별할 수 없는 다른 값으로 대체하는 기법
일반화 (Generalization)	더 일반화된 값으로 대체하는 것으로 숫자 데이터의 경우 구간으로 정의하고, 범주화된 속성은 트리의 계층적 구조에 의해 대체하는 기법
섭동 (Perturbation)	동일한 확률적 정보를 가지는 변형된 값에 대하여 원래 데이터를 대체하는 기법
치환 (Permutation)	속성 값을 수정하지 않고 레코드 간에 속성 값의 위치를 바꾸는 기법

16 해설 대상별 분석 기획 유형은 아래와 같다.

최적화 (Optimization)	• 분석의 대상이 무엇인지를 인지하고 있는 경우(Known), 즉 해결해야 할 문제를 알고 있고 이미 분석의 방법도 알고 있는 경우(Known) 사용 • 개선을 통한 최적화 형태로 분석을 수행
솔루션 (Solution)	• 분석의 대상은 인지(Known)하고 있으나 방법을 모르는 경우(Un-Known)에는 해당 분석 주제에 대한 솔루션을 찾아냄
통찰 (Insight)	• 분석의 대상이 명확하게 무엇인지 모르는 경우(Un-Known)에는 기존 분석 방식을 활용(Known)하여 새로운 지식인 통찰을 도출
발견 (Discovery)	• 분석의 대상과 방법을 모르는 경우(Un-Known)에는 발견 접근법으로 분석의 대상 자체를 새롭게 도출

17 해설 • 데이터 마트는 전사적으로 구축된 데이터 속의 특정 주제, 부서 중심으로 구축된 소규모 단위 주제의 데이터 웨어하우스이다.
• 데이터 레이크는 정형, 반정형, 비정형 데이터를 비롯한 모든 가공되지 않은 다양한 종류의 데이터(Raw Data)를 저장할 수 있는 시스템 또는 중앙 집중식 데이터 저장소이다.
• 데이터 사이언스란 데이터 공학, 수학, 통계학, 컴퓨터공학, 시각화, 해커의 사고방식, 해당 분야의 전문지식을 종합한 학문이다.

18 해설 상향식 접근방식 절차는 아래와 같다.

프로세스 분류	전사 업무 프로세스를 가치사슬, 메가 프로세스, 메이저 프로세스, 프로세스 단계로 구조화해 업무 프로세스 정의
프로세스 흐름 분석	프로세스 맵을 통해 프로세스별로 업무 흐름을 상세히 표현
분석 요건 식별	각 프로세스 맵상의 주요 의사결정 포인트 식별
분석 요건 정의	각 의사결정 시점에 무엇을 알아야만 의사결정을 할 수 있는지 정의

19 해설 개인정보 파기(개인정보보호법 제21조)

①항 개인정보처리자는 보유 기간의 경과, 개인정보의 처리 목적 달성 등 그 개인정보가 불필요하게 되었을 때는 지체 없이 그 개인정보를 파기하여야 한다. 다만, 다른 법령에 따라 보존하여야 하는 경우에는 그러하지 아니하다.

②항 개인정보처리자가 제1항에 따라 개인정보를 파기할 때에는 복구 또는 재생되지 아니하도록 조치하여야 한다.

③항 개인정보처리자가 제1항 단서에 따라 개인정보를 파기하지 아니하고 보존하여야 하는 경우에는 해당 개인정보 또는 개인정보파일을 다른 개인정보와 분리하여서 저장·관리하여야 한다.

④항 개인정보의 파기방법 및 절차 등에 필요한 사항은 대통령령으로 정한다.

20 해설 • Key-Value Store, Column Family Data Store, Document Store, Graph Store는 NoSQL의 유형이다.
• NoSQL은 전통적인 RDBMS와 다른 DBMS를 지칭하기 위한 용어로 데이터 저장에 고정된 테이블 스키마가 필요하지 않고 조인(Join) 연산을 사용할 수 없으며, 수평적으로 확장이 가능한 DBMS이다.

21 해설 • 실제는 입력되지 않았지만 입력되었다고 잘못 판단된 값은 노이즈(Noise)이다.
• 노이즈는 일정 간격으로 이동하면서 주변보다 높거나 낮으면 평균값으로 대체하거나 일정 범위 중간값으로 대체한다.
• 결측값은 필수적인 데이터가 입력되지 않고 누락된 값이다.
• 결측값은 중심 경향값 넣기(평균값, 중위수, 최빈수), 분포기반(랜덤에 의하여 자주 나타나는 값 넣기)으로 넣기 등을 통해 처리한다.

22 해설 변수 선택 방법은 다음과 같다

전진 선택법	영향력이 가장 큰 변수를 하나씩 추가하는 변수 선택 기법
후진 소거법	모든 변수가 포함된 모형에서 시작하여 영향력이 가장 작은 변수를 하나씩 삭제하는 변수 선택 기법
단계적 방법	후진 소거법과 전진 선택법의 절충적인 형태의 기법

23 해설 데이터 이상값 발생 원인은 다음과 같다.

표본추출 오류	데이터를 샘플링하는 과정에서 나타나는 오류
고의적인 이상값	자기보고식 측정에서 나타나는 오류
데이터 입력 오류	데이터를 수집, 기록 또는 입력하는 과정에서 발생할 수 있는 오류
실험 오류	실험조건이 동일하지 않은 경우 발생하는 오류
측정 오류	데이터를 측정하는 과정에서 발생하는 오류
데이터 처리 오류	여러 개의 데이터에서 필요한 데이터를 추출하거나, 조합해서 사용하는 경우에 발생하는 오류
자연 오류	인위적이 아닌, 자연스럽게 발생하는 이상값

24 해설 • 변수상에서 발생한 결측값이 다른 변수들과 아무런 상관이 없는 결측값은 완전 무작위 결측(MCAR)이다.
• 무작위 결측(MAR)은 누락된 자료가 특정 변수와 관련되어 일어나지만, 그 변수의 결과는 관계가 없는 결측값이다.
• 비 무작위 결측(MNAR)은 누락된 값(변수의 결과)이 다른 변수와 연관 있는 결측값이다.

25 해설 • 핫덱 대체, 콜드덱 대체, 혼합방법은 단순 확률 대치법이다.
• 평균 대치법의 종류에는 비 조건부 평균 대치법과 조건부 평균 대치법 등이 있다.

26 해설 릿지(Ridge)는 L2-norm을 통해 제약을 주는 방법이다.

27 해설 확률변수의 분산, 기댓값 특징에 의해 다음과 같이 계산한다.

$$V(Y)=V(2X+3)=2^2V(X)=4V(X)$$
$$=4\{E(X^2)-E(X)^2\}=4(5-2^2)=4$$

28 해설 • 왼쪽 편포(왼쪽 꼬리 분포)일 경우 평균(Mean)<중위수(Median)<최빈수(Mode)이다.
• 편포에 상관없이 항상 중위수는 가운데 값임을 기억하자.

29 해설 카토그램은 특정한 데이터값의 변화에 따라 지도의 면적이 왜곡되는 지도로 변량비례도라고도 한다.

30 해설 계통 추출은 모집단을 일정한 간격으로 추출하는 방식이다.

31 해설 • 모델의 정확도에 기여하는 변수를 학습하고, 좀 더 적은 계수를 가지는 회귀식을 찾는 방향으로 제약조건을 주어 이를 제어하는 방법은 임베디드 방법이다.
• 변수 선택 기법 중 임베디드 방법의 세부 기법으로는 라쏘, 릿지, 엘라스틱 넷, SelectFromModel이 있다.
• RFE(Recursive Feature Elimination)는 래퍼(Wrapper) 세부 기법이다.

임베디드 방법	
라릿엘셀	라쏘 / 릿지 / 엘라스틱 넷 / SelectFromModel

32 해설 • $P(E)$: 보험금을 청구할 확률
• $P(A)$: 전체 가입자 중 고위험군에 속한 가입자의 비율 = 20%
• $P(E|A)$: 고위험군에 속한 가입자가 보험금을 청구할 확률 = 50%
• $P(B)$: 전체 가입자 중 중위험군에 속한 가입자의 비율 = 30%
• $P(E|B)$: 중위험군에 속한 가입자가 보험금을 청구할 확률 = 30%
• $P(C)$: 전체 가입자 중 저위험군에 속한 가입자의 비율 = 50%
• $P(E|C)$: 저위험군에 속한 가입자가 보험금을 청구할 확률 = 20%
• 문제는 보험금을 청구했을 때, 가입자가 고위험군에 속한 가입자일 확률 $P(A|E)$를 구하는 문제이다.
• 베이즈 정리에 의해서

$$P(A|E)=\frac{P(E|A)\times P(A)}{P(E|A)\times P(A)+P(E|B)\times P(B)+P(E|C)\times P(C)}$$
$$=\frac{20\%\times 50\%}{20\%\times 50\%+30\%\times 30\%+50\%\times 20\%}=\frac{10}{29}$$

33 해설 • 모분산을 모르는 소표본($n<30$)일 경우 자유도가 $n-1$인 t-분포를 따르며, $100\times(1-\alpha)$ 신뢰구간을 구하는 공식은 $\bar{X}-t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\le\mu\le\bar{X}+t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$이다.
• 표본평균 $\bar{X}=25$, 표본의 크기 $n=16$, 표본 표준편차 $s=20$이다.
• 95% 신뢰구간이므로 $\alpha=0.05$, $\frac{\alpha}{2}=0.025$이다.
• 따라서 $t_{\frac{\alpha}{2},n-1}=t_{0.025,15}$이다.
• t-분포표에서 df = 15 $\alpha=0.025$인 교차지점을 찾으면 $t_{0.025,15}=2.131$이다.

df \ α	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
⋮										
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850

• 공식에 각각의 값들을 대입하면, $25-2.131\frac{2}{\sqrt{16}}\le\mu\le 25+2.131\frac{2}{\sqrt{16}}$
• 따라서 모평균에 대한 신뢰구간은 $23.93\le\mu\le26.07$이 된다.

34 해설 • 같거나 서로 다른 여러 가지 모형들의 예측/분류 결과를 종합하여 최종적인 의사결정에 활용하는 기법은 앙상블 기법(Ensemble Technique)이다.
• 임곗값 이동(Cut-Off Value Moving)은 임곗값(Threshold)을 데이터가 많은 쪽으로 이동시키는 방법으로 학습 단계에서는 변화 없이 학습하고, 테스트 단계에서 임곗값을 이동하는 방법이다.

35 해설 임곗값 이동(Cut-Off Value Moving)은 임곗값을 데이터가 많은 쪽으로 이동시키는 방법으로 학습 단계에서는 변화 없이 학습하고 테스트 단계에서 임곗값을 이동한다.

36 해설 • 주어진 시간 또는 영역에서 어떤 사건의 발생 횟수를 나타내는 확률 분포이므로 포아송 분포를 사용한다.

• 4분에 2명씩 오면 2분에 1명씩 오기 때문에 사건 발생 확률은 $\lambda=1$이다.
• 2분 동안 0명 또는 1명이 온다고 했으므로 $n=0$일 때, $n=1$일 때 포아송 값을 합쳐야 한다.

$$P=\sum_{n=0}^{1}\frac{\lambda^n e^{-\lambda}}{n!}=\frac{1^0\times e^{-1}}{0!}+\frac{1^1\times e^{-1}}{1!}$$
$$=e^{-1}+e^{-1}=2e^{-1}$$
$$=\frac{2}{e}$$

37 **해설** 검정 통계량 및 이의 확률분포에 근거하여 귀무가설이 참일 때 귀무가설을 기각하게 되는 제1종 오류를 범할 확률은 p-값(p-value)이다.

38 **해설** 확률 질량 함수와 누적 질량 함수의 개념은 아래와 같다.

확률 질량 함수	이산확률변수에서 특정 값에 대한 확률을 나타내는 함수
누적 질량 함수	이산확률변수가 특정 값보다 작거나 같을 확률을 나타내는 함수

39 **해설** • 정규분포를 따르는 모집단에서 모표준편차가 알려져 있으므로 Z-분포를 이용한다.
• 95% 신뢰구간이므로 $\alpha=0.05$이고, 따라서 $Z_{\frac{\alpha}{2}}=Z_{0.025}$이다.

$$\bar{X}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\le\mu\le\bar{X}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$
$$=52-1.96\frac{16}{\sqrt{16}}\le\mu\le52+1.96\frac{16}{\sqrt{16}}$$
$$=44.16\le\mu\le59.84$$

40 **해설** • 귀무가설은 현재까지 주장되어 온 것이거나 기존과 비교하여 변화 혹은 차이가 없음을 나타내는 가설이다.
• 대립가설은 표본을 통해 확실한 근거를 가지고 입증하고자 하는 가설이다.

41 **해설** • 데이터에 숨어있는, 동시에 발생하는 사건 혹은 항목 간의 규칙을 수치화하는 모델은 연관규칙 모델(Association Rule Model)이다.
• 예측 모델(Prediction Model)은 범주형 및 수치형 등의 과거 데이터로부터 특성을 분석하여 다른 데이터의 결괏값을 예측하는 기법이다.
• 예측 모델 기법으로는 회귀 분석, 의사결정나무, 인공신경망 모델, 시계열 분석 등이 있다.

42 **해설** • 구매자의 나이가 구매 차량의 유형에 어떤 영향을 미치는가? 분석에 활용되는 분석 모형은 회귀 분석이다.
• 분류 분석은 '이 사용자는 어떤 특성을 가진 집단에 속하는가?'라는 분석에 활용된다.

43 **해설** • 지도 학습 유형에는 로지스틱 회귀, 인공신경망 분석(ANN), 의사결정나무, 서포트 벡터 머신(SVM), 랜덤 포레스트 등이 있다.
• Q-Learning은 강화 학습의 유형이다.

44 **해설**

지지도	전체 거래 중 항목 A와 B를 동시에 포함하는 거래의 비율
신뢰도	A 상품을 샀을 때 B 상품을 살 조건부 확률에 대한 척도
향상도	규칙이 우연히 일어날 경우 대비 얼마나 나은 효과를 보이는지에 대한 척도

45 **해설** 군집 간의 거리 측정을 위해 유클리드 거리, 맨하튼 거리, 민코프스키 거리, 표준화 거리, 마할라노비스 거리 등을 활용한다.

46 **해설** • 상관관계가 있는 고차원 자료를 자료의 변동을 최대한 보존하는 저차원 자료로 변환하는 차원축소 방법이다.
• 차원축소는 고윳값이 높은 순으로 정렬해서, 높은 고윳값을 가진 고유벡터만으로 데이터를 복원한다.
• 분석을 통해 나타나는 주성분으로 변수들 사이의 구조를 쉽게 이해하기는 어렵다.

47 **해설** • 회귀 분석은 하나 이상의 독립변수들이 종속변수에 미치는 영향을 추정할 수 있는 통계 기법이다.
• 데이터들이 가진 속성들로부터 분할 기준 속성을 판별하고, 분할 기준 속성에 따라 트리 형태로 모델링하는 분류 예측 모델은 의사결정나무이다.

48 **해설** 측정값을 기초로 하여 제곱합을 만들고 그것을 최소로 하는 값을 구하여 측정결과를 처리하는 방법으로 오차 제곱의 합이 가장 작은 해를 구하는 것을 의미하는 것은 최소 제곱법이다

49 **해설** • 다중 선형 회귀 분석에서 모형의 통계적 유의성은 F-통계량으로 확인한다.
• 유의수준 5% 이하에서 F-통계량의 p-값이 0.05보다 작으면 추정된 회귀식은 통계적으로 유의하다고 볼 수 있다.

50 **해설** 의사결정나무는 주어진 입력값에 대하여 출력값을 예측하는 모형으로 분류나무와 회귀나무 모형이 있다.

51 **해설** 활성화 함수 중 시그모이드 함수는 기울기 소실의 원인이었지만, ReLU 함수 또는 tanh 함수를 통해 기울기 소실의 문제를 해결하였다

52 **해설** • 군집 간의 거리 계산에 사용되는 연속형 변수 거리로는 유클리드 거리, 맨하튼 거리, 민코프스키 거리, 표준화 거리, 마할라노비스 거리 등이 있다.
• 군집 간의 거리 계산에 사용되는 명목형 변수 거리로는 단순 일치 계수(Simple Matching Coefficient), 자카드(Jaccard) 계수 등이 있다.

53 **해설** 카이제곱 검정 공식은 $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$이다.

$\theta = \frac{\sum(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum(r_i - \bar{r})^2}\sqrt{\sum(s_i - \bar{s})^2}} (-1 \le \theta \le 1)$는 피어슨 상관계수 공식이다.

54 **해설**

시계열 구성요소	
추계불순	추세 / 계절 / 불규칙 / 순환

55 **해설**
- DNN은 은닉층(Hidden Layer)을 심층(Deep) 구성한 신경망(Neural Network)으로 학습하는 알고리즘이다.
- CNN은 시각적 이미지를 분석하는 데 사용되는 심층신경망으로 합성곱 신경망이라고도 한다.
- GAN은 생성자(Generator)와 구분자(Discriminator)를 경쟁적으로 학습시키는 적대적 학습 알고리즘이다.
- 입력층, 은닉층, 출력층으로 구성되며 은닉층에서 재귀적인 신경망을 갖는 알고리즘은 RNN이다.

56 **해설**
- 입력층에서 가중치가 곱해져서 은닉층으로 이동시키고, 은닉층에서도 가중치가 곱해지면서 다음 계층으로 이동하는 딥러닝 알고리즘은 DNN 알고리즘이다.
- RNN(Recurrent Neural Network) 알고리즘은 입력층, 은닉층, 출력층으로 구성되며, 은닉층에서 재귀적인 신경망을 갖고, 장기 의존성 문제와 기울기 소실문제가 발생하여 학습이 이루어지지 않을 수 있다.

57 **해설** 부호 검정(Sign Test)은 차이의 크기는 무시하고 차이의 부호만을 이용한 중위수(Median)의 위치에 대한 검정 방법으로 자료를 중위수와 차이의 부호인 +와 −의 부호로 전환한 다음 부호들의 수를 근거로 검정한다.

58 **해설**
- 런(Run)은 동일한 측정값들이 시작하여 끝날 때까지의 덩어리를 말한다.
- 동전의 앞면과 뒷면이 각각 1, 0이라고 할 때 '101001'이 나타났을 경우, 1/0/1/00/1로서 5개의 연속적인 런이라고 한다.

59 **해설**
- 잘못 분류된 개체들에 가중치를 적용, 새로운 분류 규칙을 만들고, 이 과정을 반복해 최종 모형을 만드는 알고리즘은 앙상블 기법 중 부스팅(Boosting)이다.
- 앙상블 기법 중 배깅(Bagging)은 훈련 데이터에서 다수의 부트스트랩(Bootstrap) 자료를 생성하고, 각 자료를 모델링한 후 결합하여 최종 예측 모형을 만드는 알고리즘이다.

60 **해설**
- 회귀 분석 유형 중 독립변수가 K개이며 종속변수와의 관계가 선형(1차 함수)인 것은 다중선형 회귀이다.
- 단순선형 회귀는 독립변수가 1개이며 종속변수와의 관계가 직선이다.

61 **해설**
- 누적 영역 차트는 분포 시각화의 유형으로, 여러 개의 영역 차트를 겹겹이 쌓아놓은 모양의 시각화 방법이다.
- 가로축은 시간을 나타내고 세로축은 데이터를 나타낸다.

62 **해설** 혼동 행렬에서 Positive/Negative는 예측한 값, True/False는 예측한 값과 실젯값의 비교 결과이다.

63 **해설** 민감도(Sensitivity)의 계산식은 $\frac{TP}{TP+FN}$이고, 정밀도(Precision)의 계산식은 $\frac{TP}{TP+FP}$이다.

64 **해설** 모든 데이터를 학습(Training)과 평가(Test)에 사용할 수 있으나, K값이 증가하면 수행 시간과 계산량도 많아지는 교차 검증 기법은 K-fold Cross Validation이다.

65 **해설** 관계 시각화의 유형으로는 산점도, 산점도 행렬, 버블차트, 히스토그램, 네트워크 그래프 등이 있다.

66 **해설** 샤피로-윌크 검정은 R에서 sharpiro.test() 함수를 이용하여 검정하며, 이때 귀무가설은 "표본은 정규분포를 따른다."이다.

67 **해설** 데이터가 어떤 특정한 분포를 따르는가를 비교하는 검정 기법이고, 비교 기준이 되는 데이터를 정규분포를 가진 데이터로 두어서 정규성 검정을 실시할 수 있는 것은 콜모고로프-스미르노프 적합성 검정(Kolmogorov-Smirnov Goodness of Fit Test; K-S 검정)이다.

68 **해설** 가중치 규제는 개별 가중치 값을 제한하여 복잡한 모델을 좀 더 간단하게 하는 방법으로 종류에는 L1 규제와 L2 규제가 있다.

69 **해설** 매개변수의 종류에는 하나의 뉴런에 입력된 모든 값을 다 더한 값(가중합)에 더해주는 상수인 편향(Bias)과 각 입력값에 각기 다르게 곱해지는 수치인 가중치(Weight)가 있다.

70 **해설** 매개변수 최적화 기법인 Adam은 탐색 경로의 전체적인 경향은 모멘텀 방식처럼 공이 굴러가는 듯하고, 모멘텀 방식보다 좌우 흔들림이 덜 한 특징이 있다.

71 **해설** AUC의 값은 1에 가까울수록 우수한 모형으로 판단한다.

72 **해설**
- 공간 시각화에는 등치선도 기법, 도트맵 기법, 카토그램 기법이 있다.
- 비교 시각화에는 히트맵 기법, 평행 좌표 그래프 기법, 체르노프 페이스 기법이 있다.

73 **해설** 산점도는 직교 좌표계를 이용해 두 개 변수 간의 관계를 나타내는 방법이고, 히스토그램은 자료 분포의 형태를 직사각형 형태로 시각화하여 보여주는 차트로, 수평축에는 각 계급을 나타내고, 수직축에는 도수 또는 상대도수를 나타낸다.

74 **해설** • 빅데이터 시각화 도구 중 코딩 없이 스프레드시트 데이터베이스 형태 데이터를 쉽게 가시화하는 시각화 도구는 차트 블록(Chart Blocks)이다.
• 차트 블록은 웹 기반 차트 구현(트위터, 페이스북 등 공유 가능)한다.

75 **해설** • 특이도(Specificity)=TN/(TN+FP)=40/(40+20)=40/60=2/3
• 정밀도(Precision)= TP/(TP+FP) = 55/(55+20)=55/75=11/15

76 **해설** • 산점도는 관계 시각화 유형이다.
• 공간 시각화 유형에는 등치지역도, 등치선도, 도트 플롯맵, 버블 플롯맵, 카토그램 등이 있다.

77 **해설** 오류 및 예외 발생 여부는 실시간 측정을 한다.

78 **해설** 카파 통계량의 계산식은 $K=\frac{\Pr(a)-\Pr(e)}{1-\Pr(e)}$ 이다.

K: 카파 상관계수
$\Pr(a)$: 예측이 일치할 확률
$\Pr(e)$: 예측이 우연히 일치할 확률

79 **해설** 분산 분석(ANOVA; Analysis of Variance)은 두 개 이상의 집단 간 비교를 수행하고자 할 때 집단 내의 분산, 총 평균과 각 집단의 평균 차이에 의해 생긴 집단 간 분산 비교로 얻은 F-분포를 이용하여 가설검정을 수행하는 방법이다.

80 **해설** 각 변수를 표시 지점을 연결선을 통해 그려 별 모양의 도형으로 나타낸 차트는 스타 차트이다.