빅데이터 분석기사 기출 문제(6회, 2023년 4월 8일)

1과목 빅데이터 분석 기획

1. 맵리듀스 디자인 패턴 중 다른 데이터와 연결하여 분석하는 패턴은?
① 요약 패턴 ② 조인 패턴
③ 필터링 패턴 ④ 메타 패턴

2. 다음 중 데이터 탐색에 대한 설명으로 옳지 않은 것은?
① 데이터 탐색은 수집한 데이터를 분석하기 전에 통계적인 방법을 이용하여 다양한 각도에서 데이터의 특징을 파악하는 분석 방법이다.
② 탐색적 데이터 분석의 특징으로는 저항성, 잔차해석, 자료 재표현, 현시성이 있다.
③ 범주형↔범주형 데이터의 시각화는 막대형 그래프를 사용한다.
④ 데이터 탐색은 모형 해석 시에 필요하다.

3. 다음 중 외부 공공데이터 이용의 장점은?
① 데이터 제공자와 상호협약에 의한 의사소통이 가능하다.
② 제공되는 데이터의 범위가 넓다.
③ 주로 정형 데이터 형태로 수집이 용이하다.
④ 개인정보보호에 관한 문제점을 사전에 점검할 수 있다.

4. 다음 중 빅데이터 시대 위기 요인이 아닌 것은?
① 데이터 오용
② 책임 원칙 훼손
③ M2M시대 본격화
④ 사생활 침해

5. 다음 중 탐색적 데이터 분석에 대한 설명으로 옳은 것은?
① 탐색적 데이터 분석으로 데이터를 시각화할 수는 없다.
② 변숫값과 자료구조 간의 관계를 알 수 있다.
③ 범주형 데이터의 시각화는 주로 박스플롯을 사용한다.
④ 수치형 데이터의 시각화는 주로 막대형 그래프를 사용한다.

6. 다음 중 데이터 전처리 과정에 해당하는 분석 과정은?
① 데이터 시각화 ② 모델링
③ 적합도 검정 ④ 데이터 축소

7. 다음 중 데이터 사이언스에 대한 설명으로 옳은 것은?
① 인문, 사회, 공학 등 전반적인 영역에 골고루 퍼져 있다.
② 데이터 사이언스에는 딥러닝 기술이 활용되지 않는다.
③ 데이터 사이언스를 위해 활용되는 데이터는 주로 소규모 데이터이다.
④ 데이터 사이언스에 필요한 기술에 비즈니스 관련 기술은 포함되지 않는다.

8. 다음 중 분석 준비도의 척도가 아닌 것은?
① 분석 문화 ② 분석 업무
③ 분석 결과 활용 ④ 분석 인력

9. 다음 중 연속형 변수가 아닌 것은?
① 형광등 수명 ② 혈액형
③ 키 ④ 나이

10. 빅데이터를 정형, 비정형, 반정형으로 나눌 경우 빅데이터의 어떠한 특성을 기준으로 나눈 것인가?
① 저장 위치 ② 변수 개수
③ 수집 방법 ④ 다양성

11. 다음 중 데이터셋의 noise를 제거하거나 최소화하기 위한 알고리즘은?
① 일반화(generalization)
② 집계(aggregation)
③ 평활(smoothing)
④ 속성 생성(feature construction)

12. 다음 중 데이터 분석 조직 구조에 대한 설명으로 옳지 않은 것은?
① 빅데이터 조직 구조 유형에는 집중 구조, 기능 구조, 분산 구조가 있다.
② 집중 구조는 별도의 분석 조직이 존재하고, 협업 부서와 기능이 겹치지 않는다.
③ 기능 구조는 전사적 핵심 분석이 어려우며, 과거에 국한된 분석 수행 가능성이 높다.
④ 분산 구조는 업무 과다. 이원화 가능성이 존재할 수 있기 때문에 부서 분석 업무와 역할 분담이 명확해야 한다.

13. 다음 중 데이터 거버넌스의 3요소가 아닌 것은?
① 원칙 ② 조직

③ 시스템 ④ 프로세스

14. 다음 중 네트워크를 기반으로 파일의 수집 및 공유가 가능한 시스템은?
① 관계형 데이터베이스
② NoSQL
③ HBase
④ 분산 파일 시스템

15. 다음 중 데이터 분석 수행을 위한 현황 파악 및 분석을 통한 문제를 정의하는 단계는?
① 분석 목표 수립
② 프로젝트 계획 수립
③ 보유 데이터 자산 확인
④ 도메인 이슈 도출

16. 다음 중 분석 마스터 플랜에 대한 설명으로 옳은 것은?
① 전략적 중요도, 비즈니스 성과 및 ROI, 분석 과제의 실행 용이성을 고려하여 분석 구현 로드맵을 수립한다.
② 업무 내재화 적용 수준, 분석 데이터 적용수준, 기술 적용 수준을 고려하여 우선순위를 설정한다.
③ ISP는 정보기술 및 정보 시스템을 전략적으로 활용하기 위해 중장기 마스터플랜을 수립하는 절차이다.
④ 과제 우선순위 평가기준의 시급성에는 분석 수준, 분석 적용 비용이 포함된다.

17. 다음 중 기업의 분석 수준 진단에 대한 설명으로 옳지 않은 것은?
① 확산형은 기업에 필요한 분석 구성 요소를 갖추고 있고, 높은 성숙도를 갖는 유형이다.
② 정착형은 조직 및 인력, 분석 업무, 분석 기법이 내부에 오픈되어 있다.
③ 도입형은 기업에서 활용하는 분석 업무 및 기법은 부족하지만 준비도가 높아 바로 도입할 수 있는 유형이다.
④ 준비형은 기업에 필요한 구성 요소 등이 준비되지 않아 사전 준비가 필요한 유형이다.

18. 다음 설명하는 파생변수 생성 방법에 해당하는 것은?
타이타닉 생존자 데이터에서 형제, 부모 데이터를 가족 데이터로 결합
① 단위 변환
② 표현방식 변환
③ 요약 통계량 변환
④ 변수 결합

19. 다음 중 데이터 정제 방법이 아닌 것은?
① 삭제 ② 표준화
③ 대체 ④ 예측값 삽입

20. 다음 중 개인정보 비식별화 조치에 대한 설명으로 옳지 않은 것은?
① 가명처리는 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보 없이는 특정 개인을 알아볼 수 없도록 처리하는 방법이다.
② 데이터 범주화는 특정 정보를 해당 그룹의 대푯값으로 변환하거나 구간값으로 변환하여 특정 개인을 식별할 수 없도록 하는 방법이다.
③ 총계처리는 통계값을 적용하여 특정 개인을 식별할 수 없도록 하는 방법이다.
④ 데이터 마스킹은 민감 데이터 부분을 국소적으로 삭제하는 것이다.

2과목 빅데이터 탐색

21. 다음 중 주성분 분석(PCA)에 대한 설명으로 옳지 않은 것은?
① 비정방 행렬을 음상관 행렬의 곱으로 바꾼다.
② 가장 보편적으로 사용되는 차원 축소 기법 중 하나다.
③ 원본 데이터를 최대한 보존하면서 고차원 공간의 데이터를 저차원 공간 데이터로 변환하는 기법이다.
④ 기존 변수들을 조합하여 서로 연관성이 없는 새로운 변수를 생성한다.

22. 다음 설명하는 결측값 대치법에 해당하는 것은?

> 단순대치법을 한 번 하지 않고, n번 대치를 통해 n개의 완전한 자료를 만들어 분석하는 방법으로, 대치 → 분석 → 결합의 3단계로 구성된다.

① 핫-덱 대체 ② 콜드덱 대체

③ 다중대치법 ④ 혼합방법

23. 다음 중 표현하고 싶은 데이터를 1값으로, 그렇지 않은 데이터를 0값으로 표현하는 인코딩 방식은?
① 레이블 인코딩
② 대상 인코딩
③ 카운트 인코딩
④ 원-핫 인코딩

24. 다음 중 데이터 일관성 유지를 위한 방법이 아닌 것은?
① 삭제 ② 변환
③ 파싱 ④ 보강

25. 다음 중 이상값 처리에 대한 설명으로 옳지 않은 것은?
① 이상값 처리 방법에는 삭제, 대체, 변환이 있다.
② 평균값으로 이상값을 대체해도 데이터 변환 시에 신뢰도 문제가 발생하지 않는다.
③ ESD는 평균(μ)으로부터 3시그마(σ, 표준 편차) 떨어진 값을 이상치로 인식하는 방법으로, 양쪽 0.15%에 해당하는 값을 이상치로 인식한다.
④ 머신러닝 기법을 활용하여 이상값을 검출할 수 있다.

26. 다음과 같은 표본집단 데이터의 평균값과 분산은 얼마인가?

2, 4, 6, 8, 10

	평균	분산
①	5	10
②	5	8
③	6	10
④	6	8

27. 다음 중 데이터 정제(Data Cleansing)에 대한 설명으로 옳지 않은 것은?
① 데이터 정제는 원본 데이터를 다듬어서 데이터의 신뢰도를 높이는 작업이다.
② 데이터 정제의 목적은 데이터를 이해하기 쉽게 표현하는 것이다.
③ 데이터 정제 과정은 데이터 오류 원인 분석 -> 데이터 정제 대상 선정 -> 데이터 정제 방법 결정 순이다.
④ 데이터 정제 방법에는 삭제, 대체, 예측값 삽입이 있다.

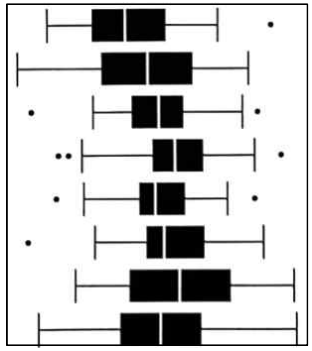
28. 다음 중 산포도 통계량에 대한 설명으로 옳지 않은 것은?
① 산포도 통계량은 데이터의 흩어진 정도를 나타내는 통계량이다.
② IQR은 사분위수 범위로 $Q_3 - Q_1$ 와 같이 연산된다.
③ 사분편차는 IQR의 절반 값이다.
④ 변동계수는 분산을 평균으로 나눈 값이다.

29. 다음 설명에 해당하는 확률 분포는?

> • 단위시간 또는 영역에서 어떤 사건의 발생횟수를 나타내는 확률 분포이다.
> • 수식은 $P=\frac{\lambda^n e^{-\lambda}}{n!}$ (λ : 평균, n : 발생횟수)와 같이 표현된다.

① 베르누이 분포
② 푸아송 분포
③ 이항분포
④ 연속확률분포

30. 다음과 같은 형태의 차트 이름은?



① Catogram
② Box-plot
③ Histogram
④ Heat Map

31. 시간 시각화 자료 중 일정 기간 동안 측정된 데이터들의 경향성을 보여주는 직선 또는 곡선은?
① 누적막대그래프 ② 추세선

③ 점그래프　　④ 계단그래프

32. 다음 중 표본분포에 대한 설명으로 옳지 않은 것은?
① 중심 극한 정리는 데이터의 크기가 작아지면 데이터의 표본분포는 최종적으로 정규분포의 형태를 따른다는 것이다.
② 표본분포는 모집단에서 추출한 일정한 크기의 표본에 대한 분포 상태를 의미한다.
③ 모수는 모집단 분포 특성을 규정짓는 척도로 관심의 대상이 되는 모집단의 대푯값이다.
④ 큰 수의 법칙은 데이터를 많이 선택할수록 표본평균의 분산은 0에 가까워진다는 것이다.

33. 다음 중 클래스 불균형에 대한 설명으로 옳지 않은 것은?
① 불균형 클래스 처리를 위해서 다수 클래스의 데이터 중 일부만 선택하여 사용하는 것을 과소표집이라고 한다.
② 가중치 균형(weight balancing)으로는 불균형 클래스를 처리할 수 없다.
③ 임곗값은 학습 단계에서는 변화 없이 학습하고, 테스트 단계에서 이동한다.
④ 과대표집 기법으로는 SMOTE, ADASYN 등이 있다.

34. 다음 중 기초 통계량에 대한 설명으로 옳지 않은 것은?
① 표준편차는 분산에 양의 제곱근을 취한 값이다.
② 사분편차는 사분위수 범위(IQR)의 절반 값이다.
③ 첨도는 데이터 분포의 뾰족한 정도를 나타내는 통계량이다.
④ 사분위수는 3분위수에서 1사분위수를 뺀 값이다.

35. 다음 중 파생변수 사용 예시로 옳지 않은 것은?
① 크루즈 탑승자 명단에서 형제, 부모 데이터를 가족 데이터로 변환하여 사용한다.
② A, B, O, AB 혈액형 데이터를 0, 1, 2, 3으로 변환하여 사용한다.
③ 화장품 업체의 분기별 매출 자료를 총 매출액으로 사용한다.
④ 차량 번호판에서 개인소유 혹은 렌터카 여부를 확인하여 사용한다.

36. 측정된 데이터들을 x축과 y축을 기반으로 점으로 표시한 그래프로, 측정된 데이터의 분포를 통해 변수간의 관계 파악이 가능한 그래프는?
① 점그래프　　② 산점도
③ 버블차트　　④ 네트워크그래프

37. 다음 중 차원 축소에 대한 설명으로 옳은 것은?
① 데이터가 많고 고차원일수록 모델의 정확도가 높다.
② 선형판별 분석은 다변량의 신호를 통계적으로 독립적인 하부 성분으로 분리하여 차원을 축소하는 기법이다.
③ 차원 축소는 분석에 활용되는 데이터의 변수 정보는 최대한 유지하면서 데이터셋 변수의 개수를 줄이는 데이터 분석 기법이다.
④ 주성분 분석(PCA)은 행과 열의 크기가 다른 임의의 M*N 차원의 행렬에서 특이값을 추출하여 효율적으로 차원을 축소하는 기법이다.

38. 세 학생의 중간고사 성적이 각각 60, 70, 80점이었다. 최소-최대 정규화를 했을 때, 세 학생의 성적의 합은 얼마인가?
① 1.5　　② 1
③ 0.5　　④ 2

39. 다음 중 다중회귀 분석의 가정이 아닌 것은?
① 잔차와 독립변수의 독립성
② 잔차와 종속변수의 선형성
③ 잔차의 분산이 독립변수와 무관한 등분산성
④ 잔차항의 정규성

40. 다음 설명에 해당되는 시스템은?

- 대규모 데이터를 저장하기 위한 데이터베이스 관리 시스템이다.
- 고정된 테이블 스키마가 없고, 조인(JOIN) 연산을 사용할 수 없다.
- 수평적 확장이 가능하다.
- 활용 예시로는 HBase, Cassandra, MongoDB 등이 있다.

① RDBMS
② MySQL
③ DFS
④ NoSQL

3과목 빅데이터 모델링

41. 다음 중 Causality Analysis에 대한 설명으로 옳은 것은?
① 하나 이상의 독립변수가 종속변수에 끼치는 영향을 추정하는 통계 방법이다.
② 두 개 이상의 변수 사이에 존재하는 상호 연관성을 분석하는 방법이다.
③ 독립변수와 종속변수 간의 인과관계를 분석하는 방법이다.
④ 서로 다른 집단의 평균에서 분산값을 비교하여 집단 간의 통계학적 차이를 확인하는 방법이다.

42. 다음 중 다중공선성을 진단하기 위한 지표는?
① 회귀계수(Regression Coefficient)
② 분산팽창지수(Variance Inflation Factor)
③ 자카드계수(Jaccard)
④ 순위상관계수(Rank Correlation Coeffecient)

43. 교차 검증 방법 중 N개 데이터 중 1개만 평가 데이터로 사용하고, 나머지 N-1개는 훈련 데이터로 사용하는 과정을 N번 반복하는 검증 방법은?
① K-fold 교차 검증
② Hold-out 교차 검증
③ LOOCV
④ LpOCV

44. 다음 중 인공신경망에 대한 설명으로 옳지 않은 것은?
① 머신러닝은 딥러닝의 일부이다.
② 인공신경망은 활성화 함수를 사용하고, 가중치를 알아내는 것이 목적이다.
③ 인공신경망의 활성화 함수는 입력 신호의 총합을 출력 신호로 변환하는 함수이다.
④ 퍼셉트론은 XOR 선형 분리 불가 문제가 발생하여 이를 보완하기 위해 다중 퍼셉트론이 개발되었다.

45. 다음과 같은 분할표에서 흡연 여부에 따른 폐암 발생률에 대한 오즈비는 얼마인가?

구분	폐암 발생	폐암 미발생	합계
흡연	6	5	11
비흡연	2	10	12
합계	8	15	23

① 8 ② 4 ③ 10 ④ 6

46. 다음 중 분석 모형 구축 절차로 옳은 것은?
① 비즈니스 영향도 평가 → 유의변수 도출 → 분석요건 확정 → 운영시스템 적용
② 유의변수 도출 → 비즈니스 영향도 평가 → 분석요건 확정 → 운영시스템 적용
③ 분석요건 확정 → 유의변수 도출 → 비즈니스 영향도 평가 → 운영시스템 적용
④ 비즈니스 영향도 평가 → 분석요건 확정 → 운영시스템 적용 → 유의변수 도출

47. 다음 중 시계열 데이터의 장기의존성 문제에 대한 LSTM기법을 보완한 방법은?
① SMOTE ② LOF
③ SEMMA ④ GRU

48. 다음 중 앙상블 분석에 대한 설명으로 옳지 않은 것은?
① 앙상블 분석 방법에는 배깅, 부스팅, 랜덤 포레스트, 보팅, 스태킹이 있다.
② 배깅(Bagging)은 데이터 사이즈가 크거나 결측값이 없는 경우에 사용하기 유리하다.
③ 부스팅 (Boosting)의 알고리즘에는 AdaBoost, GBM, XGBoost이 있다.
④ 간접투표(Soft Voting)는 각 모형의 클래스 확률값을 평균 내어 확률이 가장 높은 클래스를 최종 결과로 예측하는 방법이다.

49. 다음 중 기계학습과 통계분석에 대한 설명으로 옳지 않은 것은?
① 기계학습은 다양한 알고리즘을 활용한 학습 방법을 의미한다.
② 통계분석은 다양한 통계량을 활용한 분석방법으로 분석 결과를 시각화하여 표현할 수 있다.
③ 기계학습은 통계분석과 다르게 결과물에 대한 수식을 도출할 수 없다.
④ 기계학습을 위한 알고리즘 선정은 분석 대상에 따라 다르게 설정된다.

50. 다음 중 데이터 분할(split) 방법에 대한 설명으로 옳지 않은 것은?
① 데이터가 충분하지 않은 경우에는 학습 데이터와 검증 데이터로만 분할하여 분석하기도 한다.
② 훈련 데이터셋으로 학습한다.
③ 검증 데이터는 하이퍼파라미터의 성능을 평가하는 데 사용된다.
④ 테스트 데이터셋으로 성능을 확인한다.

51. 다음 중 과적합 방지 방법이 아닌 것은?
① 데이터 삭제 ② LASSO
③ 데이터 증강 ④ Drop Out

52. 다음 중 랜덤 포레스트에 대한 설명으로 옳지 않은 것은?
① 랜덤 포레스트는 의사결정나무 기반 앙상블 알고리즘이다.
② 이상치의 영향을 적게 받는다.
③ 분류기를 여러 개 사용할수록 예측편향이 줄어든다.
④ 랜덤 포레스트 모형에서는 모든 변수(Feature)를 학습시킨다.

53. 다음 중 변수의 성질이 다른 하나는?
① 결과변수
② 회귀변수
③ 실험변수
④ 통제변수

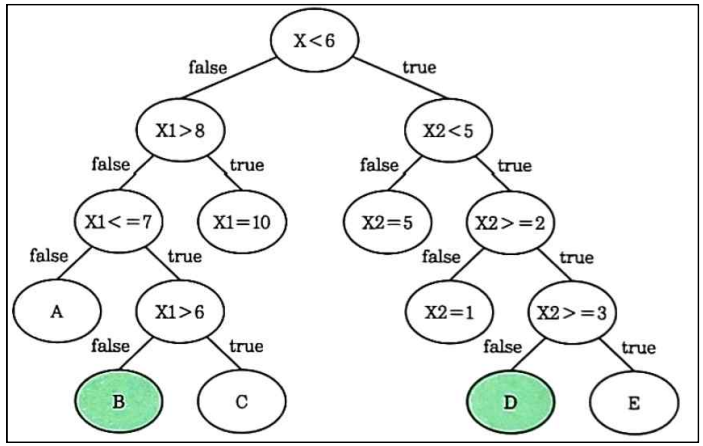
54. 다음 중 종속변수가 범주형일 때 사용되는 분석 기법이 아닌 것은?
① 판별 분석
② KNN
③ 다중선형 회귀 분석
④ 로지스틱 회귀 분석

55. 다음 중 다중선형 회귀 모형의 평가 지표는?
① ROC 곡선
② 결정계수(R^2)
③ 정밀도
④ 재현율

56. 다음 중 시계열 데이터의 공분산 기법은?
① 연관 분석
② 계절성 분석
③ 추세 분석
④ 자기상관 분석

57. 다음 중 시계열 데이터 예측 방법에 대한 설명으로 옳지 않은 것은?
① 시계열 데이터 예측 방법은 확률적 방법과 고전적 방법으로 나뉜다.
② 지수평활법은 과거 값에 가중치를 두고, 최근 값에 적은 비중을 두는 방법이다.
③ 이동평균법은 일정 기간의 관측치를 이용하여 평균을 구하고, 이를 이용해 예측하는 방법이다.
④ 확률적 방법은 주파수 영역과 시간 영역으로 나뉜다.

58. 다음과 같은 의사결정나무에서 B에 해당하는 X1 값과 D에 해당하는 X2값은?



	B(X1)	D(X2)
①	6	1
②	8	0
③	6	2
④	7	2

59. 다음 중 ReLU 함수의 뉴런이 죽는 현상(Dying ReLU)을 해결한 활성화 함수는?
① Sigmoid ② tanh
③ Leaky ReLU ④ Softmax

60. 다음과 같은 분석 방법에 해당하는 것은?

> • 독립변수가 종속변수에 얼마나 부정적인(−) 혹은 긍정적인(+) 영향을 주는지 확인하는 분석 방법으로 주로 의료통계 분야에서 많이 사용된다.
> • 종속변수(*Y*)가 이진 형태(남성 또는 여성, 성공 또는 실패, 증가 또는 감소)여야 하고, 독립변수(*X*)는 연속형 또는 범주형일 수 있다.

① 비선형 회귀 분석

② 다중선형 회귀 분석
③ 로지스틱 회귀 분석
④ 이항 로지스틱 회귀 분석

4과목 빅데이터 결과 해석

61. 다음 중 K-fold 교차 검증 학습 과정에 대한 설명으로 옳지 않은 것은?
① 데이터 학습과 검증 과정에서 테스트 데이터는 사용되지 않는다.
② K-1개의 검증 데이터를 만들고, 1개의 훈련 데이터를 만들어서 학습한다.
③ 데이터를 학습, 검증, 테스트 데이터로 나누어 교차 검증하는 방법이다.
④ 검증 데이터를 계속 바꾸어 사용하기 때문에 분할된 데이터는 한 번씩 검증 데이터로 사용된다.

62. 다음 중 시간 시각화에 대한 설명으로 옳지 않은 것은?
① 시간 시각화는 시간의 흐름에 따른 데이터의 변화를 나타낸 것을 의미한다.
② 추세선은 일정 기간 동안 측정된 데이터들의 경향성을 보여주는 직선 또는 곡선이다.
③ 일반적으로 y축은 시간을, x축은 데이터 값을 나타낸다.
④ 점그래프의 점들을 선으로 연결하면 선그래프로 표현할 수 있다.

63. 다음 설명에 해당하는 분석 방법은?

• 비계층적 군집분석 방법 중 하나로, 군집의 수를 지정하지 않아도 된다. • 밀도를 기반으로 군집을 이루기 때문에 기하학적인 모양의 군집도 찾을 수 있고, 이상값을 검출할 수 있다.

① K-means clustering
② DBSCAN
③ SOM
④ SVM

64. 다음 중 파라미터 최적화 방법으로 옳지 않은 것은?
① 손실함수 최소화
② AdaGrad
③ 확률적 경사하강법(SGD)
④ 베이지안 최적화(Bayseian Optimization)

65. 다음 중 ROC 곡선에 대한 설명으로 옳지 않은 것은?
① ROC 곡선의 x축은 1-Specificity이고, y축은 Sensitivity이다.
② ROC 곡선은 항상 0.5 이상의 값을 갖는다.
③ ROC 곡선은 가능한 모든 임곗값에 대한 참 긍정률과 거짓 긍정률을 확인한다.
④ ROC 곡선은 회귀 모형 평가 지표이다.

66. 다음 중 혼동행렬(Confusion Matrix)에 대한 설명으로 옳지 않은 것은?
① TPR은 $\frac{TP}{TP+FN}$와 같이 연산된다.
② F1-Score는 정밀도와 재현율의 기하평균이다.
③ Specificity는 실제 '부정' 범주 중 '부정'의 비율이다.
④ Precision은 $\frac{TP}{TP+FP}$ 와 같이 연산된다.

67. 다음 중 특정 사건 혹은 주제에 대한 정보를 이야기 들려주듯이 표현하는 인포그래픽 종류는?
① 비교분석형
② 만화형
③ 스토리텔링형
④ 타임라인형

68. 다음 중 역사적 사건이나 특정 주제와 관련된 히스토리를 시간 순서 형식으로 표현한 것으로 기업의 발전과정을 표현할 때 사용되는 인포그래픽 유형은?
① 통계
② 프로세스
③ 도표
④ 타임라인

69. 다음 중 스타차트에 대한 설명으로 옳지 않은 것은?
① 스타차트의 중요도는 별의 개수로 확인할 수 있다.
② 스타차트는 비교 시각화 유형에 속한다.
③ 스타차트의 축은 3개 이상이다.
④ 스타차트로 데이터의 이상값을 확인할 수 있다.

70. 다음과 같은 실젯값과 예측값 데이터가 있을 때 평균제곱오차(RMSE)는?

실젯값	10	20	15	8
예측값	8	18	13	6

① 1 ② 2
③ 3 ④ 4

71. 다음 중 ( ) 안에 알맞은 것은?

> • (㉠)은 학습 알고리즘에서 잘못된 가정으로 인한 오류를 의미하고, (㉡)은 학습 데이터의 내재된 작은 변동으로 발생하는 오차를 의미한다.
> • 이상적인 분석 모형은 낮은 (㉠)과 낮은 (㉡)으로 설정되어야 한다.

	㉠	㉡
①	오차	편향
②	잔차	분산
③	분산	편향
④	편향	분산

72. 다음과 같은 혼동행렬에서 정밀도는 얼마인가?

		예측 범줏값	
		Predicted Positive	Predicted Negative
실제 범줏값	Actual Positive	50	150
	Actual Negative	60	140

① 0.54 ② 0.45
③ 0.25 ④ 0.75

73. 다음 중 비즈니스 기여도 평가 기법에 대한 설명으로 옳지 않은 것은?
① 순 현재 가치(NPV)는 투자로부터 유입되는 미래 현금의 현재 가치와 해당 투자를 위해 투입된 비용의 차액으로 미래 시점의 순이익 규모이다.
② 투자대비효과(ROI)는 $\frac{\text{순이익}}{\text{투자비용}} * 100$ 으로 계산된다.
③ 투자회수기간(PP)은 누적투자금액과 매출금액의 합이 같아지는 기간으로 투자에 소요되는 모든 비용을 회수하는 데 걸리는 기간으로 보통 월(month) 단위로 기록한다.
④ 내부수익률(IRR)은 순 현재 가치를 '0'으로 만드는 할인율이다.

74. 다음 중 시간 시각화 유형에 속하지 않는 그래프는?
① 선그래프 ② 히스토그램
③ 계단식그래프 ④ 막대그래프

75. 정밀도가 80%이고, 재현율이 90%일 때 F1-Score는 얼마인가?
① 80.2% ② 83.1% ③ 84.7% ④ 85.3%

76. 다음 중 실젯값과 가장 오차가 작은 가설 함수를 도출하기 위해 사용되는 함수는?
① 손실 함수 ② 비용 함수
③ 활성화 함수 ④ 확률밀도함수

77. 다음 중 교차 검증에 대한 설명으로 옳지 않은 것은?
① Hold-Out 교차 검증은 가장 보편적으로 랜덤추출을 통해 데이터를 분할하는 방법으로 학습 데이터와 검증 데이터가 20~40%이고, 테스트 데이터가 60~80% 이다.
② Bootstrap은 주어진 자료에서 단순 랜덤 복원추출 방법을 활용해 동일한 크기의 표본을 여러 개 생성하는 방법이다.
③ LOOCV는 N개 데이터 중 1개만 평가 데이터로 사용하고, 나머지 N-1개는 훈련 데이터로 사용하는 과정을 N번 반복하는 방법이다.
④ K-fold 교차 검증은 데이터를 K개의 fold로 나누어 (K-1)개는 학습에, 나머지 하나는 검증에 사용하는 방법이다.

78. CNN에서 원본 이미지가 3×3, stride가 2, 필터가 5×5, padding의 크기가 2일 때 Feature Map은 얼마인가?
① (4, 4) ② (3, 3)
③ (1, 1) ④ (2, 2)

79. 다음 설명에 해당하는 오류는?

> 분석 모형을 만들 때 주어진 데이터의 특성이 지나치게 반영되어 발생하는 오류를 의미하고, 이를 과대적합(Over-Fitting)되었다고 표현한다.

① 분석 오류 ② 가정 오류
③ 일반화 오류 ④ 학습 오류

80. 다음 중 드롭아웃(DropOut)에 대한 설명으로 옳지 않은 것은?
① 드롭아웃은 학습과정에서 신경망의 일부를 사용하지 않는 기법이다.
② 제거되는 신경망의 종류와 개수는 랜덤하게 드롭아웃 확률에 의해 결정된다.
③ 드롭아웃은 서로 연결된 연결망에서 0~1사이의 확률(Drop Out Rate)로 뉴런을 제거하는 방법이다.
④ 드롭아웃은 신경망 예측 시에 사용하고, 학습 시에는 사용하지 않는다.