빅분기 단답형 예상문제

1. 기존 데이터베이스 관리도구의 데이터 수집, 저장, 관리, 분석 역랑을 넘어서는 대량의 정형 또는 비정형 데이터 세트 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술을 뭐라고 하는가?
2. 데이터의 가공 및 상관관계 간 이해를 통해 패턴을 인식하고 그 의미를 부여한 데이터를 지칭하는 용어는?
3. 빅데이터의 변화 중 데이터 수집 비용의 감소와 클라우드 컴퓨팅 기술의 발전으로 인한 데이터 처리비용의 감소로 인해 생긴 변화는?
4. 분석 조직 구조 중 전사 분석 업무를 별도의 분석전담 조직에서 담당 하는 구조는?
5. ( )는(은) 직역하면 묘사적 통계이다. 데이터를 계량화한 수치값으로 표현한다. 평균, 최빈값, 중앙값을 구하고 분산을 구하는 것이 바로 ()이다.
6. ( )은 표본을 수치 계랑한 내용으로 표본의 특징을 설명해주는 기능을 한다.
7. 데이터가 얼마나 ‘퍼져있는지’를 표현하는 통계랑이다. ( )이 클수록 데이터 서로 간의 퍼짐이 크고, 작을수록 데이터 퍼짐 정도가 작다.
8. 모집단 내에 나누어 놓은 클래스 중 하나를 선택하는 추출방식은?
9. 귀무가설이 실제로 참이라 채택해야 하나 반대로 기각해버린 오류를 무엇이라고 하는가
10. 서로 다른 두 집단의 평균에 대한 통계 검정에 주로 사용된다. 주로 모집단이 정규분포라는 정도만 알고 모표준편차는 모를 때, 모평균 추정을 위해 표준정규분포를 대신해 표본이 작은 분포은 ( )를 사용한다.
11. 역학에서 어떤 객체의 움직임을 나타낼 수 있는 최소한의 독립변수(x)의 수를 의미하는 것은?
12. 서로 다른 두개 이상의 모집단의 분산이 서로 같은지를 확인할 때 사용된다. 즉, 분산분석과 회귀분석의 결과를 해석할 때 주로 사용된다. 이는 무엇인가?
13. 두 개 이상의 서로 다른 범주에 대해 가설 검정, 모분산의 추정, 분포 간의 차이 등에 많이 사용되는 분포는 무엇인가?
14. 대량의 트랜잭션 정보로부터 개별 데이터(변수) 사이에서 연관규칙(x면 y가 발생)을 찾는 것을 말한다. 가 령 슈퍼마켓의 구매내역에서 특정 물건의 판매 발생 빈도를 기반으로 ‘A물건을 구매하는 사람들은 B물건 을 구매하는 경향이 있다.’라는 규칙을 찾을 수 있다 이는 무엇인가?
15. 비교 대상이 되는 항목들의 빈도를 이용하여 자료 간 관계의 유의성을 파악할 때 사용한다. 주로 ‘범주형’ 자료 간의 관계를 확인하는 데 쓰이며, 전체 비율을 통해 예상빈도를 구하여 실제빈도와의 차이를 대조하는 방식은?.
16. 자료 간의 차이를 대조시키는 분석기법이다. 여기서 분산은 평균을 중심으로 데이터가 얼마나 퍼져있는지를 표현하는 통계량이다. 만약 데이터의 퍼짐이 없고 모든 개별 데이터값들이 동일하다면, 편차가 없으므로 분산은 ‘0’이 될 것이다. 이는 무엇인가?
17. ()은 𝑥와 𝑦변수 간에 관계가 어떤 선형적인 관계를 갖고 있는지를 파악한다. 두 변수 간의 관계의 강도도 계산할 수 있다. 두 변수가 변하는 패턴이 얼마나 비슷한가를 확인하는 과정이 ()이다.
18. ()는 확률변수의 절대 크기(=구간)에 영향을 받지 않도록 각 확률변수에 표준편차를 나누어 표준화 시킨 것이다. 빈칸에 들어갈 용어는?
19. 변숫값 대신 순위를 이용하여 상관계수를 구하는 경우 사용된다. 데이터를 작은 것부터 차례로 순위를 매겨 서열 순서로 바꾼 뒤 순위를 이용하는 상관계수는?
20. 여러 변수 간의 상관관계를 공통의 분산 으로 밝히고 일정한 기준(=상관계수)으로 우선순위를 설정하는 과정, 즉 종속변수 y에 영향도가 높은 1개 이상의 독립변수 𝑥𝑛를 찾는 것을 무엇이라고 하는가?
21. 회귀분석에서 최초엔 임의의 추세선(=가설초기화)을 기준으로 분포한 x, y의 좌푯값의 차이(=잔차)를 제곱하여 모두 더한 값이 최소가 되는 지점들을 연결하는 방법은?
22. 데이터가 어떤 범주에 속할 확률을 0에서 1 사이의 값으로 예측하고, 그 확률에 따라 가능성이 더 높은 범주에 속하는 것으로 분류하는 기법은?
23. 사건이 발생할 확률이 발생하지 않을 확률에 비해 몇 배 더 높은가를 설명하는 개념은 무엇인가?
24. 귀무가설이 진실일 때 적어도 그 정도의 극단적인 표본값이 나올 확률, 즉 귀무가설이 참임에도 이를 기각할 확률은?
25. 시계열 분석에서 직전 값에서 현재 값을 빼는 과정을 무엇이라고 하는가?
26. 과거 데이터에 기반하여 미래를 예측하는 모형으로 추세선을 통한 예측 과정은 선형 회귀모델과 동일한 모형은?
27. 모든 시계열 자료를 사용해 평균을 구하고 시간 흐름에 따라 최근 시계열에 더 높은 가중치를 부여해 추세를 계산하는 방법은?
28. 우리가 예측하려고 하는 임의의 데이터와 가장 가까운 거리의 데이터 K개를 찾아 다수결에 의해 데이터를 예측하는 방법은?
29. 결정트리에서 불순도를 측정하는 지표는 무엇인가
30. 결정트리를 독립병렬 결합하는 방식을 ( )이라고 한다. 원 자료에서 하위 데이터 세트를 만드는 부트스트랩 과정을 포함한다. 데이터 건수는 동일하게 그리고 데이터 무작위 추출 시 중복은 허용(복원추출)하여 각 트리 모형들이 학습할 데이터 세트를 그 수에 맞게 준비해야 하는 용어로 빈칸에 알맞은것은?
31. 결정트리를 연결병렬 결합하는 방식이며, 첫 번째 결정트리의 결과가 그 다음 결정트리의 학습 데이터 구성에 영향을 미치는 순환구조인 것은?
32. 계층적 군집의 결과는 ()의 형태로 표현된다.
33. 사전에 결정된 군집 수 K에 기초하여 전체 데이터를 상대적으로 유사한 K개의 군집으로 구분하는 방법은?
34. ‘밀도 기반’으로 군집을 할당한다. 여기서 밀도 기반이란, 어떤 데이터 포인트에서의 반지름 x내에 n개 이상의 포인트를 갖는 것을 하나의 군집으로 구분하는 것은?
35. 활성화함수 중 0보다 작은 구간에서 기울기가 0이고 나머지 0보다 큰 구간에서는 항상 일정한 크기의 기울기를 가지는 특성을 갖고 있는 함수는?
36. 모델(입력층→은닉층→출력층)을 통해 나온 결괏값이 실제로 우리가 원하는 값(Labeling)과 얼마나 다른지 정도를 측정하기 위해 사용하는 함수는?

<정답>

1. 빅데이터

2. 정보

3. 전수조사

4. 집중구조

5. 기술통계

6. 추론통계

7. 분산

8. 집락추출, 군집추출

9. 1종오류

10. T분포

11. 자유도

12. F분포

13. 카이제곱분포

14. 연관규칙분석

15. 교차분석

16. ANOVA

17. 상관분석

18. 상관계수

19. 스피어만상관계수

20. 주성분분석

21. 최소제곱법, 최소자승법

22. 로지스틱회귀분석

23. 오즈비

24. 유의확률

25. 차분

26. AR모형 자기회귀모형

27. 지수평활법

28. KNN

29. 엔트로피

30. 배깅

31. 부스팅

32. 덴드로그램

33. K 평균군집

34. DBSCAN

35. ReLU 함수

36. 손실함수