<15회 ADSP>

1. 데이터에 관한 구조화된 데이터로, 다른 데이터를 설명해 주는 데이터는?

답 : 메타데이터 / 보기 : 데이터마트

2. 정량적 데이터와 정성적 데이터 중에 정성적 데이터는?

답 : 기상특보 / 보기 : 습도, 강수량

3. 암묵지와 형식지의 상호작용이 아닌 것은?

답 : 추상화 / 보기 : 공통화, 내면화, 표준화, 연결화

4. DIKW 피라미드 - 지식(Knowledge)에 해당하는 것은?

답 : USB를 A마트가 아니라 B마트에서 사야겠다 / 보기 : A마트는 USB를 10000원 B마트는 USB를 5000원에 판다, A마트보다 B마트가 USB를 싸게판다, B마트의 다른 물건들이 A마트의 물건들보다 쌀 것이다.

5. 빅데이터가 만들어내는 본질적인 변화에서 틀린것은?

답 : 전수조사 -> 표본조사 / 보기 : 사전처리 -> 사후처리, 질 -> 양, 인과관계 -> 상관관계

6. 데이터 사이언스의 영역은 분석적 영역, 비지니스 컨설팅 영역, 데이터 처리와 관련된 IT영역이 있는데 보기 중 다른 영역에 속하는 것은?

답 : 데이터 시각화 / 보기 : 데이터 웨어하우스, 프로그래밍, 데이터 엔지니어링

주관식 1. 데이터 사이언티스트의 요구 역량 두 가지. 이론적 지식이나 분석 기술에 대한 숙련은 ____ skill이고 통찰력있는 분석이나 설득력은 ____ skill이다.

답 : hard, soft

주관식 2. SQL 언어 문제. A와 B 사이에 있는 값을 구하는 그런 문제였음.

답 : BETWEEN

2과목 데이터 분석 기획

1. 단계를 순차적으로 진행하는 방법으로, 이전 단계가 완료되어야 다음 단계로 진행될 수 있으며 문제가 발견되면 피드백 과정이 수행되기도 한다.

답 : 폭포수 모델 / 보기 : 나선형 모델, 프로토타입 모델

2. 분석 방법론에서 문제 하나 더 나왔어요. 자세히 기억은 안나는데 폭포수, 나선형, 프로토타입 모델에 관련된 내용으로 보기가 주어졌었고 적절하지 않은 것 고르는 문제였습니다. 제 생각에 답은 '나선형 모델은 표본추출부터 계속 반복해야한다(?)' 이런 보기 골랐습니다.

3. 하향식 접근법(Top Down Approach) - 타당성 검토에

해당하지 않는 것은?

답 : 절차적 타당성 / 보기 : 경제적 타당성, 기술적 타당성, 데이터 타당성

4. 하향식 접근법의 타당성 검토에서 문제 하나 더 나왔어요. 경제적, 데이터, 기술적 타당성의 세부내용을 보기에 주고 적절하지 않는 것 고르는 문제였는데, 답은 아마도 데이터 타당성에서 '데이터를 어떻게 모을 것인가를 고려해야한다'. 이게 정답인것 같아요.
보기에 어자일 프로젝트 관리방식이 나왔던 것 같은데 여기는 공부를 자세히 안해서 잘 모르겠습니다.(데이터에듀 150page에 있어요.)

5. 분석 거버넌스 - 분석관점에서의 사분면 분석 - 기업에서 활용하는 분석업무, 기법 등은 부족하지만 적용조직 등 준비도가 높아 바로 도입할 수 있는 기업 단계는?

답 : 도입형 / 보기 : 정착형, 확산형, 준비형 (도입,활용,확산,최적화단계로 이루어진 분석 성숙도 모델하고 헷갈리시면 안됩니다.)

6. SAA(Self Automatic? Analysis)에 관한 것 중에 틀린 것은? 이 문제도 아예 처음 보는 문제였어요.
답도 잘 모르겠습니다. 보기로는 구성요소로 BI, 머신러닝 등이 있다, R이나 파이썬같은 프로그래밍을 잘해야한다, 오픈소스로 되어있다,

주관식 1. 구축된 시스템의 검증을 위하여 단위 테스트, 통합 테스트, 시스템 테스트 등을 실시한다. O테스트는 품질 관리 차원에서 진행함으로써 적용된 시스템의 객관성과 완전성을 확보한다. O은?

답 : 시스템

주관식 2. 데이터 거버넌스에서 데이터 저장소 관리는 메타데이터 및 표준 데이터를 관리하기 위한 전사 차원의 저장소를 구성한다. 저장소는 데이터 관리 체계 지원을 위한 워크플로우 및 관리용 응용 소프트웨어를 지원하고 관리 대상 시스템과의 인터페이스를 통한 통제가 이루어져야 한다. 데이터 구조 변경에 따른 [ ]도 수행되어야 효율적인 활용이 가능하다. [ ]은?

답 : 사전 영향 평가

3과목 데이터 분석

1. R프로그래밍 문제 - 표준편차 구하는 방법이 아닌것은?

답 : stdev(a) / 보기 : sd(a), var(a)^(1/2), sqrt(var(a))

2. 이산형 변수의 기댓값 구하는 방법은?

답 : E(X) = 시그마 x*p(x) / 보기 : 시그마 x^2*p(x), 인테그랄 x*p(x), 인테그랄 x^2*p(x)

3. 확률밀도함수 관련 문제, 적절치 않은 것은?

답 : 이산형 확률변수의 확률밀도함수는 이산형 확률밀도함수라고 한다. (확률질량함수가 맞는 표현인것 같습니다.)

4. 구간추정 관련 내용 출제되었습니다. 보기는 기억이 나지 않네요. 아마 정답은 '정규분포 외에서는 구간추정을 사용할 수 없다' 이런 뉘앙스였습니다.

5. 데이터 시각화에서 틀린 것은?

답 : 줄기잎 그림은 계산이 복잡하다. (답이 아닐 수도 있습니다.) / 보기로는 히스토그램, 상자그림 등의 설명이 나왔습니다.

6. 피어슨 상관계수와 스피어만 상관계수에서 틀린 것은?

답 : 피어슨 상관계수는 변수들을 순서척도로 바꾼 후 스피어만 상관계수를 구하는 방법으로 계산한다.

7. 회귀분석 R로 돌린 결과값 보고 해석하는 문제가 나왔습니다. 유의한 변수에서 계수에 해당하는 값보고 그 변수가 증가하면 결과가 어떻게 달라지는지에 대한 관계를 알고 있어야 풀 수 있었습니다. 변수가 연속형이 아니라 두가지 정도의 명목척도일때도 어떻게 나오는지 알아야했어요.

8. 회귀분석 결과로 나온 변수 대 잔차그림에서 얘가 y=x^2 그래프처럼 모양이 나왔는데(p346 20번 그림) 이때 어떤 조치를 해줘야하는가 문제가 나왔는데, 저는 후진제거법으로 유의하지 않은 변수를 줄여준다를 선택했습니다. 이건 잘 모르겠네요..

9. 단계적 변수선택 문제가 나왔습니다. 전진선택법, 후진제거법, 단계별방법에 관한 보기가 나와 틀린거 고르는 거 였어요. 보기는 기억이 나지 않습니다.

10. 시계열 분석에서 데이터만 가지고 분석하는 것과 모델을 만들어 분석하는 것중에 데이터만 가지고 분석하는 방법이 아닌 것은?

답 : 이동평균 모형, 자기회귀 모형에 관한 내용이 있는 것을 선택했습니다.

11. 시계열 분석 문제가 하나 더 나온 것같아요.

12. 의사결정나무에서 지니지수(지니계수) 구하는 문제. [A]-[B],[C], [A노드 Good:50 Bad:50], [B노드 Good:10 Bad:40], [C노드 Good:20 Bad:30] 여기서 B노드의 지니지수를 구해라.

답 : 0.32 (정확하지 않습니다.)

13. 의사결정나무에서 오차를 크게 할 위험이 높거나 부적절한 추론규칙을 가지고 있는 가지 또는 불필요한

가지를 제거하는 단계

답 : 가지치기(pruning)

14. 앙상블 기법에서 배깅에 관련된 문제가 나왔습니다. 옳은 거 고르는거 였는데 제가 고른 답은 '배깅은 랜덤 복원추출을 여러번 하기 때문에 추출되지 않은 원소도 있을 수 있다.' 이거 골랐습니다.

15. 인공신경망 모델에서 은닉노드가 적으면 무슨 문제가 발생하는가?

답 : 기억이 안나요ㅠ

16. 반응범수가 범주형인 경우에 적용하는 회귀분석 모형은?

답 : 로지스틱 회귀분석

17. 연속형 변수의 거리 중에서 표준화와 공분산을 동시에 만족하는 거리는?

답 : 마할라노비스 거리 / 보기 : 유클리드 거리, 표준화 거리, 민코우스키 거리

18. 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도의 형태로 형상화 하는 것은?

답 : SOM (Self-Organizing Map)

19. 연관분석에서 지지도, 신뢰도, 향상도 계산하는 간단한 문제 출제되었습니다.

20. EM 알고리즘 그림 보고 해석하는 문제가 나왔는데 이게 데이터에듀 책에는 없네요ㅠ

주관식 1. 독립변수 후보 모두를 포함한 모형에서 출발해 가장 적은 영향을 주는 변수부터 하나씩 제거하면서 더 이상 제거할 변수가 없을 때의 모형을 선택한다.

답 : 후진제거법

주관식 2. 상자그림(box-plot) 결과값 보여주고 어떤 값(3/4분위수)보다 큰 자료는 몇 %가 있냐?

답 : 25%

주관식 3. 민감도와 1-특이도를 활용하여 모형을 평가하는 그림은 무엇인가?

답 : ROC 커브 향상도 곡선이라고 합니다.

주관식 4. 아이스크림 판매량 회귀분석 R로 돌린 결과값 보여주고 신뢰수준 0.01을 만족하는 회귀모형을 만들어라.

답 : CONSUME = 0.08 + 0.0076INC + 0.003TEMP (유의한 변수들만 적어야하는 점 주의)

주관식 5. 데이터마이닝 성과분석 중 오분류에 대한 추

정치에서 TN/(FP+TN)은 뭐라고 하는가?

답 : Specificity

주관식 6. mclust 문제가 출제되었습니다. 책에서 내용을 찾기가 어렵네요. R코드와 결과 그래프를 보여주고 최적의 군집 개수를 구하라는 문제였습니다.

<16회 ADSP>

1번) 글로벌기업의 빅데이터 활용사례 중 잘못 연결된 것?

-->페이스북 (상)

2번) 빅데이터의 위기 요인과 통제방안을 서로 잘못 연결한 것?

--> 책임원칙회손, 데이터오용(하)

3번) 데이터에 대한 설명이 잘못된 것?

--> 개별 데이터 자체로 의미가 중요하다.(하)

4번) 데이터 사이언스와 통계학이 다른 점?

--> 기존 통계학에 데이터마이닝을 접목한 학문(중)

5번) 빅데이터 활용의 3요소?

-->데이터, 기술, 인력(하)

6번) 일차원 분석의 사례로 적절한 것?

--> 에너지 (상)

7번) 사회기반구조로써의 데이터베이스?

-->NEIS(중)

8번)빅데이터에 대한 설명이 잘못된 것?

-->표본조사의 중요성이 대두된다. (하)

단1) 기존 금융회사의 중앙집중형 서버에 거래 기록을 보관하는 방식에서 거래에 참여하는 모든 사용자에게 거래 내역을 보내주며 거래 때마다 이를 대조하는 데이터 위조 방지 기술?

-->블록체인(상)

단2) 수치로 명확하게 표현되는 데이터

-->정량적데이터(하)

9번) 분석과제 발굴에 대해 잘못 설명한 것?

-->분석 대상이 명확하면 상향식 접근방식이 적절하다. (하)

10번) 분석과제 우선 순위 결정 요소가 아닌 것?

--> 데이터 필요 우선 (하)

11번) 분석 거버넌스 체계가 아닌 것은?

--> 과제 예산 및 비용집행 (상)

12번) 경쟁자 확대 관점의 분석기회 발굴 영역이 아닌 것?

--> 경쟁채널(하)

13번) 분석 ROI에서 비즈니스 효과?

-->value(하)

14번) 빅데이터 분석 방법론의 피드백이 반복해서 발생하는 단계?

--> 데이터 준비, 데이터 분석(중)

15번) 거버넌스 체계에 대한 설명?

-->데이터 관리 체계(중)

16번) 분석 우선순위 평가 기준에 대해 잘못 설명한 것?

-->시급성은 전략적 중요도와 데이터 수집비용으로 평가한다. (중)

단3) CRISP-DM에서 잡음, 이상치, 결측치 식별 및 제거와 데이터셋을 선택하는 단계?

-->데이터 준비(상)

단4) 데이터 분석 기획에서 데이터 분석가에게 요구되는 기술?

-->소프트스킬(하)

17번) 근로자 임금과 교육수준 관련 그래프 해석이 부적절한 것?

-->각 막대의 높이가 임금수준이다. (하)

18번) 상자그림에 대한 설명이 부적절한 것?

--> 수염은 이상치를 제외하고 데이터의 하위 상위 25% 범위이다.(하)

19번) Hitters 데이터의 상관관계표를 잘못 설명한 것?

--> 상관계수가 통계적으로 유의하다. (중)

20번) lasso 회귀모형에 대해 잘 못 설명한 것?

--> Lasso는 L1 penalty를 사용한다.(상)

21번) default 데이터의 모자이크 플랏에 대한 설명이 잘못된 것?

-->학생인 고객이 많다.(하)

22번) Hitters 데이터에서 train set과 test set 에 대한 설명이 잘못된 것?

--> test set 결과과 일반적으로 train set 결과보다 좋다.(중)

23번) default 데이터의 로지스틱 회귀분석 결과에서 유의수준 0.05.에서 설명이 잘못 된 것?

-->income과 default는 통계적으로 유의하다.(하)

24번) 의사결정나무모형에 대한 설명이 잘못된 것?

-->의사결정나무는 상향식 의사결정의 흐름을 따른다.(중)

25번) nci.data의 계층적 군집분석에 대한 설명이 잘못된 것?

-->최단 연결법은 평균연결법에 비해 계산 연산시간이 빠르다.(상)

26번) 결측값 처리 방법에 대한 설명이 잘못된 것?

-->다중대치법은 추정량의 과소추정이나 계산의 난해성 문제를 보완하는 방법이다.(상)

27번)시간의 흐름에 따라 관측한 데이터는?

-->시계열 자료.(하)

28번)시계열의 요소분해법 중 분해 요소에 대한 설명이 잘못된 것?

-->순환요인은 명백한 경제적이나 자연적인 이유가 없이 알려지지 않은 주기를 갖고 변화하는 자료이다.(하)

29번)연관성 분석에 대한 설명이 잘못된 것?

-->시차 연관분석은 원인과 결과의 형태로 해석이 가능하다.(상)

30번)통계적 추론에 대한 설명이 잘못된 것?

-->비모수적 추론은 모집단에 대한 분포를 가정하지 않지만, 분포의 특성을 결정하는 모수를 추론하는 방법은 아니다.(중)

31번)원천 데이터를 기반으로 감춰진 지식 등을 발견하고 의사결정 등에 유용한 정보로 활용하고자 하는 작업은?

-->데이터 마이닝은 대용량 데이터에서 의미 있는 패턴을 파악하거나 예측하여 의사결정에 활용한다.(하)

32번)R의 데이터 구조 중 벡터에 대한 설명이 맞는 것은?

-->R에서 벡터는 하나 또는 그 이상의 스칼라 원소들을 갖는 집합이다.(하)

33번)R에서 연속변수의 최대, 최소, 중앙값 등과 범주변수의 범주 빈도를 출력해주는 함수는?

-->R에서 summary함수는 요약통계를 나타내주는 함수이다.(하)

34번)앙상블 기법이 아닌 것은?

-->시그모이드는 인공신경망의 활성화 함수의 하나이다.(하)

35번)사과-->딸기에 대한 향상도는?

-->0.3/(0.7 × 0.45)(하)

36번)군집화 방법중 DBSCAN 기법 등 군집탐색에 가장 효과적인 방법은?

-->DBSCAN(Density-based spatial clustering of applications with noise) 기법은 밀도 기반 군집의 하나이다.(상)

37번)세분화하는 기능이 있는 데이터 마이닝 분석 방법은?

-->집단을 세분화하는 기능은 분류분석이다.(하)

38번)붓스트랩을 사용하고 전체 관측치 중 훈련용 자료로 사용되는 비율은?

-->붓스트랩 복원 추출방법으로 훈련용 자료로 사용되는 비율은 63.2%이다.(상)

39번)kmeans 군집의 단점을 보완하기 위해 평균 대신 사용하는 것은?

-->중앙값 (하)

40번)k-means 군집의 장점은?

-->k-means 군집은 계층적 군집보다 많은 양의 자료를 다룰 수 있다.(중)

단5) 로지스틱 회귀모형에서 단위가 증가할 때마다 성공의 ()이/가 몇 배 증가 하는가?

-->오즈(odds), 승산비 (중)

단6) 주성분분석에서 3차원으로 축소할 경우 잃게 되는 정보량은?

-->1 – 0.7966195=0.2033805 ==> 20.3% (하)

단7) 베이즈 정리를 활용한 분류방법 알고리즘은?

-->나이브 베이지안 분류(Naive Bayesian classification) (상)

단8) 오분류표의 F1 값은?

-->F1=2*(재현율*정확도)/(재현율+정확도)=

2*(0.15*0.3)/(0.15+0.3) = 0.2 (상)

단9) 시점에 상관없이 시계열의 특성이 일정하다는 용어는?

--> 정상 시계열(하)

단10) 군집분석의 품질 평가 지표로 응집도와 분리도를 계산하는 지표는?

--> 실루엣(shilouette) (상)

<17회 ADSP>

1번) 데이터베이스와 통신을 위해 고안된 언어?

-->SQL (하)

2번) 빅데이터 활용 기본 테크닉 중 사례(맥주사는 사람이 기저귀 산다.)을 묻는 문제?

--> 연관성분석(하)

3번) 멀티미디어 등 복잡한 데이터 구조를 관리하는 DBMS?

--> 객체지향 DBMS (상)

4번) 데이터 사이언스에 대한 부적절한 설명?

--> 주로 분석 정확성에 초점 (하)

5번) 개인정보를 무작위 처리하는 등 사생활침해를 막기위한 데이터 가공을 방지하는 기술?

--> 난수화(중)

6번) 방대한 조직내 분산된 데이터베이스 관리시스템을 통합, 운영하는 방법?

--> 데이터웨어하우스 (하)

7번) 암묵지와 형식지의 상호작용?

-->공통화->표준화->연결화->내면화(중)

8번)빅데이터가 만드는 본질적인 변화가 부적절한 것?

--> 상관관계 -> 인과관계 (하)

단1) 기업이 외부 공급업체 또는 제휴업체와 통합된 정보시스템으로 연계하여, 시간과 비용을 최적화시키는 데이터베이스 시스템?

--> SCM (상)

단2) 데이터 가공 및 상관관계를 통한 패턴을 인식하고 그 의미를 부여하는 것?

--> 정보(하)

9번) 분석 프로세스 특징 중 부적절한 것?

--> 분석 프로젝트는 일반적인 프로젝트 관리와는 상이하다. (중)

10번) 비즈니스 모델 캠버스의 구성 단위?

--> 규제 – 업무 – 제품 – 고객 - 인프라 (하)

11번) 빅데이터 분석 방법론의 분석 기획 단계의 순서?

--> 범위설정 - 프로젝트 정의 - 수행계획 수립 - 위험식별 (상)

12번) 프로토타이핑 접근법의 프로세스로 부적절한 것?

--> 순환적 문제 탐색(상)

13번) 분석 준비도에서 분석업무 파악 항목으로 부적절한 것?

--> 분석 기법 라이브러리(중)

14번) 분석 조직 구조의 설명이 부적절한 것?

--> 분산구조는 분석이 집중되지 못해 신속한 실무 적용이 어렵다. (하)

15번) 분석 마스터 플랜 설명이 부적절한 것?

--> 전체를 반복적이고 순환적인 계획작성(중)

16번) 분석방법 알고, 대상 모를경우 분석 기법?

--> Insight (하)

단3) 소프트웨어와 시스템공학의 역량 성숙도를 측정하기 위한 모델?

--> 능력 성숙도 통합모델(Capability Maturity Model Integration, CMMI) (상)

단4) 도일한 사실에도 판단이 달라지는 현상?

--> 프레이밍 효과(상)

17번) summary 함수에 대한 설명이 부적절한 것?

--> wage는 범주형 변수 (하)

18번) K 평균군집에 대한 설명이 부적절한 것?

--> 객체들이 다른 군집으로 이동할 수 없다. (하)

19번) 스피어만 상관계수에 대한 설명이 부적절한 것?

--> 비선형적인 상관관계는 나타내지 못한다. (상)

20번) Hitters 데이터의 회귀분석 결과에 대한 설명이 부적절한 것?

--> 전지선택법을 사용했다.(하)

21번) Credit 데이터의 변수간 산점도와 피어슨 상관관계에 대한 설명이 부적절한 것?

--> balance와 가장 상관관계가 높은 변수는 Income.(하)

22번) 비모수 검정 방법으로 부적절한 ?

--> 카이제곱 검정 (중)

23번) Auto 데이터에서 단순선형회귀분석시 발생가능한 것

--> 등분산성의 위배 (상)

24번) Wage 데이터에서 t-test에 대한 설명이 부적절한 것?

--> 귀무가설에서 설정한 평균이 참값을 포함한다. (중)

25번) Default 데이터에서 로지스틱 회귀분석 시 설명이 잘못된 것?

--> income이 높을수록 default 가능성이 낮다. (중)

26번) Hitters 데이터에서 데이터 분할에 대한 설명이 부적절한 것?

--> 반복 시행시 이전과 다른 데이터를 얻는다 (하)

27번)교차분석에 대한 설명이 부적절한 것?

-->교차분석은 두 문항 모두 범주형 변수가 아니어도 사용할 수 있다.(상)

28번)산포측도에 대한 설명으로 부적절한 것?

-->평균절대편차는 데이터가 얼마나 퍼져 있는가를 나타내는 값 (하)

29번)히스토그램에 대한 설명으로 부적절한 것?

-->히스토그램은 표본의 크기가 작아도 분포의 형상을 잘 표현해 낸다.(중)

30번)추정에 대한 설명으로 부적절한 것?

-->신뢰수준 95%의 의미는 추정값이 신뢰구간 내에 존재할 확률이 95%(상)

31번)신경망 모형에 대한 설명으로 부적절한 것?

-->은닉층의 뉴런 수와 개수는 자동으로 설정된다.(상)

32번)의사결정나무에 대한 설명으로 부적절한 것?

-->분리변수의 P차원 공간에 대한 현재 분할은 이전 분할에 영향을 받는다.(상)

33번)군집분석에 대한 설명으로 부적절한 것?

-->군집 결과에 대한 안정성을 검토하는 방법은 교차타당성을 이용하는 방법 (상)

34번)데이터를 실험데이터와 평가데이터로 구분하는 방법?

-->홀드아웃 방법(상)

35번)특정 기준에 따라 서로 상이한 소집단에서 무작위로 추출하는 표본추출법?

-->층화추출법(하)

36번) 재현율을 계산하는 산식?

-->TP/(TP+FN)(중)

37번)사회연결망 분석에서 행과 열에 다른 개체가 배열되는 매트릭스?

-->2원모드 매트릭스.(상)

38번)간접적으로 연결된 모든 노드 간 거리를 합산해 중심성을 측정?

-->근접중심성(상)

39번)Apriori 알고리즘의약점을 보완하기 위해 고안된 알고리즘?

-->FP-성장 (상)

40번)오분류표를 사용한 평가 지표에 대한 것?

-->F1(중)

단5) 유사한 클러스터끼리 합쳐가면서 원하는 클러스터 개수가 될 때까지 진행하는 방식?

-->병합적 군집(상향적 군집, agglomerative clustering)(상)

단6) (관찰된 일치율-우연에 의한 일치율)/(1-우연한 일치율)?

-->카파(kappa) (상)

단7) 민감도와 특이도가 어떤 관계를 갖고 변하는지 이차원 평면상에 표현?

--> ROC(Receiver Operating Characteristic) (중)

단8) 시계열에서 분리해서 분석하는 방법?

--> 분해 시계열 (하)

단9) SOM의 경쟁층의 프로토타입 벡터와의 거리를 계산하고 가장 가까운 프로토타입 벡터?

--> BMU(Best-Matching Neighbors)(상)

단10) 목표치가 다범주인 경우 각 범주에 속할 사후 확률을 제공?

--> 소프트맥스 함수 (상)

<18회 ADSP>

1번) 빅데이터 출현 배경 설명으로 부적절한 것?

--> 4) 데이터 구조의 정형화 (하)

2번) 통찰력을 제공하는 분석 기술이 아닌 것?

--> 4) 추출 (하)

3번) 데이터 사이언티스트의 역량에 관해 묻는 문제

--> 하드 스킬(Hard Skill)(상)

4번) 내재된 경험을 문서나 매체로 저장하는 과정?

--> 표출화 (하)

5번) 빅데이터 시대 위기 요인으로 부적절한 것?

--> 1)익명화 (하)

6번) 인문학 열풍의 외부환경 요소가 아닌 것?

--> 4)빅데이터 분석 기법과 방법론의 확대 (하)

7번) 데이터사이언스에 대한 설명이 부적절한 것?

--> 1) 통계학은 더욱 확장된 유형의 데이터를 근간으로 한다 (중)

8번) DIKW 중 다른 하나?

--> 8월 매출액은 3000만원으로 예상한다. (중)

단1) (A)인과관계, (B)상관관계 (상)

단2) 플렛폼 (중)

9번) 데이터 거버넌스 체계의 내용?

아래) 데이터 표준 용어 설정, 명명규칙 수립 등.....

--> 1)데이터 표준화 (중)

10번) 문제 탐색 단계에 대한 설명 중 부적절한 것?

--> 4)유스케이스 활용보다는 새로운 이슈탐색을 우선 (하)

11번) 과제 우선순위 결정 내용 중 부적절한 것?

--> 가치는 투자비용 요소(하)

12번) 분석 기획 고려사항 중 장애요소에 대한 부적절한 설명?

--> 복잡하고 정교한 모형 (중)

13번) 분석 대상과 분석 방법의 4가지 유형?

--> 통찰 - 발견 (하)

14번) 분석 과제 수행시 고려해야할 5가지 속성이 아닌 것?

--> 데이터 분석 방법 (하)

15번) 데이터 유형으로 적절한 것?

--> 1) Demand Forecasts – Competitor pricing- Email records (상)

16번) 프로토타이핑 으로 적절한 것?

--> 2)신속하게 해결책이나 모형을 제시함으로써.... 유용한 상향식 접근 방법이다. (중)

단1) 하향식 접근방식 (하)

단2) 모델링 과정 (상)

17번) 회귀모형의 변수선택법으로 부적절한 것?

--> 주성분분석 (중)

18번) Credit 데이터를 통한 회귀 그래프를 보고 적절한 R명령어?

--> lm(Balance~Income+Student, data=Credit) (중)

19번) 거리를 활용한 유사도 측도에 대한 설명으로 부적절한 것?

--> 마할라노비스 거리는~~~(상)

20번) 데이터 정규성의 확인하기 위한 방법이 아닌 것?

--> Durbin Watson test (상)

21번) 상관분석의 설명이 잘못된 것?

--> 종속변수 값을 예측하는 선형모형을 추출하는 방법 (하)

22번) 데이터마이닝 단계 중 목적변수 등 데이터를 준비하는 단계?

--> 데이터 가공 (하)

23번) 연관규칙의 향상도의 설명이 적절한 것?

--> 향상도가 1보다 크면 결과가 우수하다. (중)

24번) R 데이터의 저장 형식 설명으로 부적절한 것?

--> 행력을 as.vector함수를 적용하면 1행부터 차례로 원소로 생성된다.(중)

25번) 의사결정나무의 해석으로 부적절한 것?

--> 뿌리마디에서 아래로 내려갈수록 각 마디에서의 불순도는 점차 증가한다.(중)

26번) 각 열이 서로 다른 타입의 R 데이터 구조?

--> 데이터 프레임(하)

27번) 비모수 검정에 대한 특징이 잘못된 것?

--> 평균, 분산 등을 이용한 검정을 실시(중)

28번) SOM 에 관한 설명으로 부적절한 것?

--> SOM은 역전파 알고리즘을 사용하여 수행(상)

29번) 추정과 가설검정에 대해 부적절한 것?

--> 귀무가설이 사실일 때, ... 검정통계량이 나올 확률을 p 값이라고 한다.(중)

30번) Bias-variance tradeoff 내용?

--> (상)

31번) Chickwts에서 71마리 병아리 데이터로 설명이 가장 부적절한 것?

--> 첨가물의 개수가 5개(중)

32번) 보험사에서 고객 데이터를 갱신해서 예측하려고 할 때 적절한 분석기법?

--> 랜덤포레스트(하)

33번) 광고 채널을 통한 매출정도의 상관분석으로 부적절한 설명?

--> TV광고와 Sales는 증가하는 인과관계를 가진다.(하)

34번) 회귀분석의 결정계수의 설명이 부적절한 것?

--> 1)총변동과 오차에 의한 변동 비율(중)

35번) 귀무가설이 사실인데 우리가 내린 판정이 잘못되었을 실제 확률?

--> p-value(하)

36번) 회귀분석에 대한 설명으로 부적절한 것?

--> 2)기울기가 0이 아닌 것 가장을 귀무가설이라 한다.(하)

37번) 교차 판매 등에 적용하는 기법은?

--> 연관분석(중)

38번) R패키지의 설명으로 부적절한 것?

-->1) data.table 패키지는 ddply함수를 제공한다. (중)

39번) R명령어 중 3*y의 결과?

--> 3 6 9 NA(하)

40번) 기법의 활용분야가 다른 하나?

--> SOM(하)

단 5번) 오분류 표의 정확도 표기

--> (a+d)/(a+b+c+d) (하)

단 6번)맨하탄 거리 구하기

--> 2 (중)

단 7번) 분류가 잘못된 데이터에 큰 가중치를 부여하는 기법

--> 부스팅(하)

단 8번) 현시점에서 전시점 자료를 빼는 것

--> 차분(하)

단 9번) 회귀모형의 계수 추정 방식

--> 최소제곱법(하)

단 10번) 군집의 개수 구하기

--> 3개 (하)

<19회 ADSP>

1번) 암묵지의 예로 부적절한 것?

--> 4) 대차대조표에 요구되는 지식 (하)

2번) 데이터 분석 용어와 의미가 잘못된 것끼리 묶인 것?

--> 3) ABC (중) -> 데이터마이닝 만 맞음

3번) 기업내 구축된 정보시스템의 설명으로 적절한 것?

--> 1) ERP (하)

4번) 데이터에 대한 설명으로 부적절한 것?

--> 3) 형식지란 개인에 체화된 비밀스러운 지식(하)

5번) 설명에 해당하는 구성요소?

--> 1) A-메타데이터, B - 인덱스 (상)

6번) 책임원칙 훼손의 사례?

--> 1)범죄 예측 프로그램을 통한 체포 (하)

7번) 정량 데이터가 아닌 것??

--> 4) 문자 (하)

8번) 빅데이터의 위기와 통제가 적절하게 묶인 것?

--> 4)나, 마 (중) 데이터오용 -> 알고리즈미스트

단1) 데이터사이언스 (하)

단2) 유전자 알고리즘 (중)

9번) 분석과제 기획시 고려할 사항이 아닌 것?

--> 1)데이터 분석을 위해서는 데이터의 정형화가 필수 (하)

10번) 문제 담색 단계의 도구가 아닌 것?

--> 2)탐색적 문제 발견 (하)

11번) 분석 프로젝트 관리 영역이 아닌 것?

--> 3)프로세스 관리(상) ->시험에 처음으로 나온 유형

12번) 빅데이터 거버넌스로 옳은 것끼리 묶인 것?

--> 2)C, D(중)

풀이) "B. 빅데이터분석에서는 수명주기관리보다 품질관리가 중요하다"는 잘못된 내용이며 B가 포함되지 않은 보기는 2번 뿐임

13번) 과제 중심적인 접근 방식의 특징이 잘못된 것?

--> 2)Accuracy와 Deploy (하)

14번) 난이도와 시급성에서 우선적인 분석 과제는?

--> 1) 난이도 쉬움, 시급성 현재 (하)

15번) 데이터 표준화 설명으로 적절한 것?

--> 2) 표준용어 설정, 명명규칙 수립, 메타 데이터 구축, 데이터 사전 구축 (중)

16번) 데이터 분석 조직의 설명을 옳은 것?

--> <아래> 별도의 분석 전담조직, 현업부서의 분석업무 이원화/이중화 가능성

1) 집중구조(하)

단3) 나선형 모델 (상)

단4) 정보전략계획(ISP) (상)

17번) 시계열의 구성요소가 아닌 것?

--> 2) 교호요인 (하)

18번) R 프로그램에 대한 설명으로 부적절한 것?

--> 3) xy[,1]은 y와 동일하다 (중)

<해설> rbind(x,y) 한 경우 2개의 벡터가 행으로 연결되어

>xy

1 2 3 4 5

10 20 30 40 50

19번) 연관 규칙의 "A->B"의 향상도?

--> 3) 83% (중)

<해설> 지지도 : 3/10, P(A) : 6/10, P(B) : 6/10

20번) 유아용 카시트 상관분석 결과가 부적절한 것?

--> 4) sales와 price는 선형관계이다.(하)

21번) 주성분분석 결과로 부적절한 것?

--> 4) 정규화 전의 데이터로 주성분분석 결과와 정규화 후 주성분분석 결과는 다르다.(중)

22번) 연관규칙 지표에 대한 설명?

<아래> 1보다 크면 해당 규칙이 결과가 우수하다

--> 1) 향상도 (하)

23번) 이산형 확률변수의 기대값?

--> 2) E(x)= SUM(xf(x)) (중)

24번) 메이저리그 데이터 주성분 결과로 설명이 잘 못된 것?

--> 3) 공분산 행렬을 사용하여 주성분분석을 했다.(하)

25번) 확률 및 확률분포에 관한 설명으로 부적절한 것?

--> 3) 두 사건 A, B가 독립일 때 조건부확률(중)

26번) 수면유도제 데이터를 통한 t-test결과가 잘못된

것?

--> 1)수면유도제 2가 수면유도제 1보다 효과적이다.(하)

27번) 주성분분석에 대한 설명으로 부적절한 것?

--> (상)

28번) 3개의 군집으로 나누어라?

--> 4)(A,B),(C),(D,E)(하)

29번) 연관분석의 장점으로 부적절한 것?

--> 3) 품목세분화에 의미 없이 규칙 발견 가능(중)

30번) 앙상블 모형을 고르시오?

--> 1)배깅 (하)

31번) 계층적 군집분석에서 dist() 함수에서 지원하지 않는 거리?

--> 3) binary(하)

32번) 계층적 군집분석에서 군집의 오차제곱합을 활용한 군집분석 방법?

--> 와드연결법(중)

33번) Cars 데이터에서 설명이 잘못된 것?

--> 4) speed의 75% 백분위 값을 알 수 없다.(하)

34번) 오분류표에서 민감도와 같은 지표?

--> 4)재현율 recall(하)

35번) 신경망모형에서 설명이 잘못된 것?

--> 2) 입력변수의 속성에 따라 활성화 함수의 선택이 달라진다.(상)

36번) 표본조사에 대한 설명이 부적절한 것?

--> 4)비표본오차는 조사 대상이 증가한다고 해서 오차가 커지지 않는다.(상)

37번) 자료의 척도에 관한 설명이 잘못 된 것?

--> 3) 구간척도는 순서와 간격에 의미가 있고 0이 절대적 의미가 있다.(하)

38번) 다중회귀모형의 통계적 유의성 확인 방법?

--> 1) F 통계량을 확인한다. (중)

39번) 연관규칙의 측정지표 중 A와 B의 지지도 공식?

--> 2) A와 B가 동시에 거래되는 건수/전체거래건수(하)

40번) R에서 패키지 설치 순서?

--> 1,2 모두 정답 (하)

1) install.packages("패키지명") -> library("패키지명")

2) install.packages("패키지명") -> library(패키지명)

단 5번) 모든변수 적용 후 하나씩 변수 제거 방법?

--> 후진제거법 (하)

단 6번) 의사결정나무에서 MSE 등을 고려한 과적합 문제 해결?

--> 가지치기 (중)

단 7번) 랜덤모델과 비교하여 모델 성능평가?

--> 향상도곡선 (lift curve) (상)--> 이익도표(gain chart)는 오답

단 8번) A와 B의 지지도 구하기

--> 1/5 (하)

단 9번) 시계열 모형 중 백색잡음의 선형 가중합 모형

--> MA 모형(중)

단 10번) 선형회귀모형에서 결정계수 구하기

--> SSR/SST = 3163/7178=0.441 (상)

<20회 ADSP>

1-1. 빅데이터 분석에 경제적 효과를 제공해준 결정적 기술로 가장 적절한 것은?

1 텍스트 마이닝

2 클라우드 컴퓨팅

3 저장장치 비용의 지속적인 하락

4 스마트폰의 급속한 확산"

1-2. 기업내부 데이터베이스 활용과 관련 없는 것은?

1 CRM 2 ERP 3 ITS 4 KMS"

1-3. 데이터 저장방식에는 RDB, NoSQL, 분산파일시스템 저장방 식이 있다. 다음 중 NoSQL 관련이 없는 도구는?

1 MongoDB 2 Hbase 3 Redis 4 mySQL"

1-4. 데이터웨어하우스 고유 특성이 아닌 것은?

1 데이터웨어하우스는 기업내의 의사결정지원 어플리케이션

을 위한 정보 기반을 제공하는 하나의 통합된 데이터 저장

공간을 말한다.

2 ETL은 주기적으로 내부 및 외부 데이터베이스로부터 정보

를 추출하고 정해진 규약에 따라 정보를 변환한 후에 데 이터웨어하우스에 정보를 적재한다.

3 데이터웨어하우스에서 관리하는 데이터들은 시간적 흐름 에 따라 변화하는 값을 유지한다.

4 일반적으로 데이터웨어하우스는 전사적 차원에서 접근하 기 보다는 재무,생산, 운영과 같이 특정 조직의 특정의 업 무 분야에 초점을 두고 있다."

1-5 다음 중 빅 데이터 활용 테크닉에 관한 설명이다. 적절하 지 않은 것은?

1 유형분석 : 택배차량을 어떻게 배치하는 것이 비용에 효 율적인가?

2 유전알고리즘 : 응급실에서 의사를 어떻게 배치하는 것이 가장 효율적인가?

3 연관분석 : 시스템 로그데이터를 분석해 침입자나 유해 행위자를 색출할 수 있는가?

4 회귀분석 : 사용자의 만족도가 충성도에 어떤 영향을 미 치는가?"

1-6. 데이터베이스 특징에 대한 설명 중 부적절한 것은?

1 데이터베이스는 통합된 데이터이다. 이것은 데이터베이스에서

동일한 내용의 데이터가 중복되어 있지 않다는 것을 의미한다.

2 데이터베이스는 저장된 데이터이다. 이것은 자기디스크나 자기 테이프 등과 같이 컴퓨터가 접근할 수 있는 저장 매체에 저장되

는 것을 의미한다.

3 데이터베이스는 공용 데이터이다. 이것은 여러 사용자가 서로 다른 목적으로 데이터베이스의 데이터를 공동으로 이용되는 것을 의미한다.

4 데이터베이스는 변화되는 데이터이다. 데이터베이스가 저장되는 내용은 정량적데이터 상태로만 유지됨을 의미한다."

1-7.정성적 데이터에 속하는 것은?

1강수량

2온도

3기상특보

4풍속

1-8. 다른 이해 관계자들이 보완적인 상품, 서비스를 제공하는 생태계를 구축하고자 하는 비즈니스 모델을 무엇이라 하는 가?

1 플랫폼 비즈니스 모델

2 가치사슬형 비즈니스 모델

3 상거래형 비즈니스 모델

4 대리인형 비즈니스 모델

9. 다음은 무엇을 의미하는가?

이것은 데이터를 통합/분석하여 기업 활동에 연관된 의사결정을 돕는 프로세스를 말합니다. 이와 관련해 가트너는 이것은 '여러 곳에 산재되어 있는 데이터를 수집하여 체계적이고 일목요연하게 정리함으로써 사용자가 필요로 하는 정보를 정확한 시간에 제공할 수 있는 환경으로 정의하였다.

BI

10. 다음 중 빈칸에 알맞은 용어는?

데이터 사이언티스트들은 주로 데이터 처리나 분석 기술과관련된(1)만을요구받고있는것처럼보인 다. 하지만 이러한 (1) 은 훌륭한 데이터 사이언티 스트가 갖춰야 하는 능력의 절반에 불과하다. 나머지 절반은 통찰력 있는 분석, 설득력 있는 전달, 협력 등 (2)이다.

하드스킬, 소프트스킬

2-1.다음 중 분석방법론의 구성요소가 아닌 것은?

1 목적

2 절차

3 방법

4 도구와 기법

2-2.프로젝트 위험 계획 수립 시 예상되는 위험에 대한 대응방 법이 아닌 것은?

1 회피(Avoid)

2 전이(Transfer)

3 완화(Mitigate)

4 관리(Management)

2-3. 다음 중 하향식 접근법의 데이터분석 기획 단계는?

1 Problem discovery-> Problem Definition-> Solution Search-> Feasibility Study

2 ProblemDefinition-> Problem discovery-> Solution Search-> Feasibility Study

3 Solution Search-> Feasibility Study-> Problem discovery-> Problem Definition

4 Feasibility Study -> Problem discovery-> Problem Definition-> Solution Search

2-4. 다음은 분석기획 발굴의 범위 확장에 관한 설명 중 적절 하지 않은 것은?

1 거시적 관점에서는 현재의 조직 및 해당 산업에 폭넓게 영향을 미치는 사회,경제적 요인을 STEEP로 요약되는 Social(사회),Technology(기술), Economic(경제), Environmental(환경), Political(정치) 영역으로 나누어서 기회 탐색을 수행한다.

2 경쟁자 관점에서는 현재 수행하고 있는 사업 영역의 제 품, 서비스에 대해서만 분석 기회 발굴의 폭을 넓혀서 탐 색한다.

3 시장의 니즈 탐색 관점에서는 현재 수행하고 있는 사업에 서의 직접 고객 뿐만 아니라 고객과 접촉하는 역할을 수행 하는 채널 및 고객의 의사결정에 영향을 미치는 영향자들 에 대한 폭 넓은 관점을 바탕으로 분석 기회를 탐색한다.

4 역량의 재해석 관점에서는 현재 해당 조직 및 기업이 보 유한 역량뿐만 아니라 해당 조직의 비즈니스에 영향을 끼 치는 파트너 네트워크를 포함한 활용 가능한 역량을 토대 로 폭넓은 분석 기회를 탐색한다.

2-5.포트폴리오 사분면 분석을 통한 과제 우선순위를 선정하는 기법 중 분석과제의 적용 우선순위를 '시급성'에 둔다면

결정해야 할 우선순위는? (교재 그림 129페이지 참조)

1 III->IV->II

2 I->II->III

3 II->IV->I

4 III->I->II

2-6. 마스터플랜 수립 때 적용 범위 및 방식의 고려요소가 아닌 것은?

1 업무 내재화 적용 수준

2 분석 데이터 적용 수준

3 투자 비용 수준

4 기술적용수준

2-7. 분석기획 단계에서의 task가 아닌 것은?

1 비즈니스의 이해

2 프로젝트 정의 및 계획 수립

3 필요 데이터 정의

4 프로젝트 위험 계획 수립

2-8. 빅데이터 4V에 해당되지 않은 것은?

1 Volume

2 Variety

3 Velocity

4 Visuality

9. 다음과 같은 현상을 무엇이라 하는가?

동일한 사안이라고 해도 제시되는 방법에 따라 그에 관한 해석이나 의사결정이 달라지는 왜곡현상

프레이밍효과

10 다음 빈칸에 알맞은 용어는?

데이터 거버넌스에서 데이터 저장소 관리는 메타데이터 및 표준 데이터를 관리하기 위한 전사 차원의 저장소를 구성한다. 저장소는 데이터 관리 체계 지원을 위한 워크플로우 및 관리용 응용 소프트웨어를 지원하고 관리 대상 시스템과의 인터페이스를 통한 통제가 이루어져야 한 다. 데이터 구조 변경에 따른 [ ]도 수행되어야 효율적인 활용이 가능하다.

사전영향평가

3-2. 자료의 척도에 설명으로 부적절한 것은?

1 명목척도- 단순히 측정 대상의 특성을 분류하거나 확인하

기 위한 목적으로 숫자를 부여

2 서열척도- 대소 또는 높고 낮음 등의 순위만 제공할 뿐

양적인 비교는 할 수 없다.

3 등간척도-순위를 부여하되 순위 사이의 간격이 동일하여

양적인 비교가 가능하다.

4 비율척도- 측정값 사이의 비율 계산이 가능한 척도이며,절대 영점이 존재하지 않는다.

3-3. 아래의 R 코딩 출력값은 무엇인가?

DF<-c("Monday","Tuesday","Wednesday") substr(DF,1,2)

1 "Mo", "Tu", "We"

2 "ay","ay","ay"

3"Monday","Tuesday","Wednesday"

4 "Mon","Tue","Wed

3-4.회귀모형에 관한 설명 중 올바르지 않은 것은?

1 오차의 표준편차에 대한 추정값은 7.165이다.

2 설명변수 모두가 출산율 변동의 원인임을 보여준다.

3 Fertility 변동성을 설명하는데 가장 유의한 설명변수는

Education이다.

4 회귀식은 Fertility변동성을 70.67% 설명한다.

3-5. 다음 R의 연산 결과이다. 올바르지 않은 것은?

[x<-c(1,2,3,4) y<-c(""apple"",""banana"",""orange"") xy<-c(x,y)]

1 xy결과값은 문자형이다.

2 xy 결과값은 ""1"" ""2"" ""3"" ""4"" ""apple"" ""banana"" ""orange"" 이다.

3 xy[1]+xy[2]의 값은 3이다

4 xy[c(5,7)], y[c(1,3)] 결과값은 서로 같다.

x[1]+x[2]=3.

3-6.타입이 다른 데이터 타입을 하나의 객체로 묶는 구조는?

1 행렬

2 벡터

3 스칼라

4 리스트

3-7. 다음 중 데이터마이닝 기법 사례와 분석방법이 올바른 것 은?

1 카탈로그 배열 및 교차판매, 공격적 판촉행사 등의 마켓 팅 계획-연관분석

2 부모가 있는 어린이의 수를 추정, 가족구성원의 총수입 추정-의사결정나무

3 생물을 문, 종, 속으로 나누는 것이나, 물질을 요소별로 나누는 것-장바구니분석

4 시장 세분화의 첫 단계로서, 판촉 활동에 가장 반응률이 높은 고객 선별-회귀분석

3-8. 데이터마이닝의 목적 중 사람, 상품에 관한 이해를 증가 시키기 위한 것으로 데이터의 특징 및 의미를 표현 및 설 명하는 기능을 무엇이라 하는가?

1분류

2추정

3예측

4기술

3-9.다음 중 오분류표의 평가지표 중 True로 예측한 것 중 실 제 True인 지표를 무엇이라 하는가?

1 Precision

2 Accuracy

3 F1

4 Kappa

3-10. 같은 모집단 내의 다른 데이터에 적용하는 경우에도 안 정적인 결과를 제공하는 것을 의미하며 데이터를 확장하여 적용할 수 있는지에 대한 모형 평가 기준을 무엇이라 하는 가?

1 일반화의 가능성

2 효율성

3 예측의 정확성

4 분류의 정확성

3-11. 의사결정나무에서 이산형 목표변수는 지니지수, 연속형 목표변수는 분산 감소량을 사용하는 알고리즘은 무엇인가?

1 CHAID

2 CART

3 C4.5

4 C5.0

3-12. 다음 중 주성분 분석의 해석으로 올바르지 않은 것은?

정답: (2)공분산행렬을 사용하여 주성분했다.

3-13. 다음 중 결과 값이 다른 것은?(유사문제 응용)

[fruit<-c(5,10,1,2) names(fruit)<-c(""orange"",""banana"",""apple"",""peach"")]

1 fruit[c(""apple"",""banana"")] 2 fruit[3:2]

3 fruit[c(3,2)]

4 fruit[-c(2,3)]

3-14. apply 함수와 multi-core 사용함수를 이용하면 for loop 를 사용하지 않고 매우 간단하게 처리할 수 있는 있고, apply 함수에 기반해 데이터와 출력변수를 동시에 배열로 치환하여 처리하는 패키지는 무엇인가?

1 reshape

2 plyr

3 데이터 데이블

4 outlier

3-15. 다음 중 신용카드 고객 파산여부를 예측하는 모형이 아 닌 것은?

1 로지스틱회귀분석

2 선형회귀분석

3 의사결정나무

4 앙상블모형

3-16. 성별을 구별하는데 사용되는 척도는?

1 명목척도

2 서열척도

3 구간척도

4 비율척도

3-17. 보기에서 사용되는 척도는 무엇인가?

[매우불만족-불만족-보통-만족-매우만족]

1 명목척도

2 구간척도

3 순서척도

4 비율척도

3-18. 시그모이드 함수의 범위는?

1 0~1

2 -1~1

3 -1~0

4 0~1

3-19. 다음 중 주성분 분석에 설명 중 적절하지 않은 것은?

1 가장 분산이 적은 것을 제1주성분으로 설정한다.

2 주성분 분석은 상관관계가 있는 변수들을 결합해 상관관

계가 없는 변수로 분산을 극대화하는 변수로 선형결합을

해 변수를 축약하는데 사용하는 방법이다.

3 공분산 행렬을 사용하는 경우 고유값이 1보다 큰 주성분

의 개수를 이용한다.

4 공분산행렬을 이용한 분석의 경우 변수들의 측정단위에

민감하다.

3-20.다음 중 k 평균군집분석의 분석절차 순서는?

a초기군집중심으로 k개의 객체로 임의 선택

b 각 자료를 가장 가까운 군집 중심에 할당

c 각 군집내의 자료들을 평균을 계산하여 군집의 중심 갱신

d 군집 중심의 변화가 없을 때 까지 b,c 반복한다.

1 a->b->c->d

2 c->a->b->d

3 a->b->d->c

4 b->a->c->d

3-21. 빵-> 우유의 신뢰도는(주관식)?

1 - 빵,맥주,우유

2 - 빵,우유,계란

3 - 맥주,우유

4 - 빵,맥주,계란

5- 빵, 맥주, 우유, 계란

0.75

3-22. 종속변수가 성공 또는 실패인 이항변수 되어 있을 때 종 속변수와 독립 변수간의 관계식을 이용하여 두 집단 또는 그 이상의 집단을 분류하고자 할 때 사용되는 분석기법을 무엇이라 하는가?

1 로지스틱 회귀분석

2 다중 회귀분석

3 의사결정나무

4 앙상블 모형

3-23. 조건-결과(if-then) 유형의 패턴을 발견하는데 사용하는 데이터마이닝 기법은?

1 som

2 연관규칙

3 다차원척도

4 의사결정나무

3-24. (유사문제응용) 다음은 음주와 비음주와 사고와 비사고 확률 분할표로 작성하였다. 다음 중 조건부확률 P(음주/사 고)는 얼마인가?

0.54

3-25. 다음 값은 얼마인가?

P(A)=0.3, P(B)=0.4 이며 서로 독립이다. P(B|A)?

0.4

3-26. 다음은 어떤 군집방법에 대한 설명인가?

[두 군집사이의 거리를 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 거리의 최소값으로 측정하며, 고립 된 군집을 찾는데 중점을 둔 방법이다.]

최단연결법(단일연결법)

3-27. 회귀모형에 대한 가정에 대한 설명이다. 빈칸에 알맞은 용어는? 잔차항

· 잔차와 독립변수이 값이 관련되어 있지 않다. - 독립성

· 잔차들끼리 상관이 없어야 한다. - 비상관성

· ( )이 정규분포를 이뤄야 한다. -정상성

3-28. 다음을 무엇이라고 하는가?

모집단을 성격에 따라 몇 개의 집단 또는 층으로 나누고, 각 집단내에서 원하는 크기의 표본을 무작위로 추출하는 확률적 표본추출방법

층화추출

3-29. 다음을 무엇이라고 하는가?

원 데이터 집합으로부터 크기가 같은 표본을 여러 번 단순 임의 복원추출하여 각 표본에 대해 분류기를 생성하는 앙상블 기법

배깅

3-30. 다음을 무엇이라고 하는가?

주어진 원천 데이터를 랜덤하게 두 분류로 분리하여 교차 검정을 실시하는 방법으로 하나의 모형 학습 및 구축을 위한 훈련용 자료로 하나는 성과평가를 위한 검증용 자 료로 사용하는 방법

홀드아웃

<21회 ADSP>

1번) 빅데이터가 만든 변화와 거리가 먼 것?

--> 3) 표본조사 기법의 중요성 (하)

2번) 빅데이터 출현 배경과 거리가 먼 것?

--> 1) 공공정보 개방 (중)

3번) 데이터 베이스 설계 절차?

--> 1) 요구분석->개념->논리->물리 (상)

4번) "커피-> 탄산수"의 빅데이터 분석 기법?

--> 2) 연관규칙 (하)

5번) 감성분석의 설명이 부적절한 것?

--> 1) 사회적 관계를 알고자 할 때 (하)

6번) 고객과의 관계를 지속적으로 하기 위한 정보시스템?

--> 1) CRM (하)

7번) 개인 정보 비식별화 기법이 부적절한 것?

--> 3)데이터 마스킹- 특정 값 삭제처리 (상)

8번) 데이터유형이 다른 것?

--> 2) 소음에 대한 센서 데이터(상)

단1) 데이터 사이언스(Data Science) (하)

단2) 분류분석, 유형분석(classification)(중)

9번) CRISP-DM에서 모델링 단계가 아닌 것?

--> 4) 모델 적용성 평가 (하)

10번) 분석 프로젝트에 관한 설명이 부적절한 것?

--> 3) 데이터 수집에 대한 철저한 통제와 관리가 필요 (상)

11번) Accuracy와 Precision에 대한 설명이 부적절한 것?

--> 1) 활용측면은 Precision, 안정성 측명은 Accuracy(하)

12번) BI와 비교하여 빅데이터 분석의 키워드로 적절한 것?

--> (상)

13번) 대상은 명확하고 방식이 명확하지 않은 경우?

--> 4) solution (하)

14번) 분석 우선순위 평가 기준의 설명이 부적절한 것?

--> 1) 시급성은 전략적 중요도와 데이터 수집비용 등으로 평가, 난이도는 분석수준과 복잡도로 평가 (중)

15번) CRISP-DM의 업무이해에 해당하는 것은?

--> 2) 업무파악-> 상황파악-> 데이터마이닝 목표 설정 ->프로젝트 계획수립(중)

16번) 분석 조직에 대한 설명으로 해당하는 것?

--> 3) 집중구조 (하)

단3) ㄱ).과제수행, ㄴ.ㄷ.)팀구성, 분석과제 실행, 분석과제 진행관리(상)

단4) 디자인 띵킹(Design Thinking) (상)

17번) 두 집단의 평균 차이 검정?

--> 3) t-분포 (하)

18번) 이산형 확률분포 중 두가지 경우만 존재하며, 성공할 확률이 일정하고, 전후 사건에 독립적인 확률분포?

--> 4) 베르누이 확률분포 (중)

19번) 증명하고 싶은 가설?

--> 1) 대립가설(하)

20번) 비모수 검정 중 짝지어진 두 개의 관찰치의 크고 작음에 대한 가설 검증?

--> 3) 부호검정 (상)

21번) 오른쪽으로 꼬리가 긴 분포의 평균과 중앙값의 관계?

--> 4) 평균이 중앙값보다 크다(하)

22번) 50%의 데이터의 흩어진 정도?

--> 1) 사분위 범위(하)

23번) 비선형적인 관계도 파악할 수 있는 상관계수?

--> 3) 스피어만 상관계수(상)

24번) lazzy learning에 해당하는 기법?

--> 2) 최근접 이웃 모형(중)

25번) 활성화 함수의 설명으로 적절한 것?

--> 4) 시그모이드 함수 (중)

26번) 앙상블의 특징으로 옳지 않은 것?

--> 4) 상호 연관성이 높으면 정확도가 향상 (상)

27번) 앙상블 방법론 중 부분집합엣 모형을 생성하여 결합하는 방식?

--> 3) 랜덤포레스트 (상)

28번) 분류모형에서 관측치가 현저히 부족하여 모형이 학습하기 힘든 문제?

--> 3) 범주 불균형 문제 (상)

29번) F2 의 설명?

--> 4) 재현율에 2배 가중치를 부여하여 조화평균 (상)

30번) 신경망에서 가중치의 절대값이 커져 조정이 더 이상 힘든 과소적합이 발생하는 문제? --> 3) 포화문제 (상)

31번) 임의적 모양의 군집 탐색에 효과적인 기법?

--> 3) 밀도기반 군집 기법 (상)

32번) 군집내 거리와 군집간의 거리를 기준으로 군집 분할 성과를 측정하는 방식?

--> 4) 실루엣 계수 (상)

33번) 상자수염그림에서 하한과 상한을 구하시오?(1Q-4, 3Q-12)

--> 4) -8,24 (하)

34번) 군집의 개수를 미리 정하지 않아도 되는 탐색적 모형?

--> 3) 계층적 군집 (중)

35번) 두 벡터 사이의 각도를 이용하여 벡터간의 유사 정도를 측정하는 방식?

--> 3) 코사인 유사도 (상)

36번) SOM 에 대한 설명으로 옳은 것은?

--> 4) 승자 독점의 학습 규칙에 따라 학습 (상)

37번) 정규화 방법 중 원데이터의 분포를 유지하면서 정규화하는 방법?

--> 3) min-max정규화 (상)

38번) 지지도는 얼마인가?

--> 3) 0.3(하)

39번) F1은 얼마인가?

--> 2) 18/57 (중)

40번) 지니지수 구하기?

--> 2) 0.48 (상)

단 5번) 57.4% (하)

단 6번) 다차원 척도법(상)

단 7번) 와드연결법(중)

단 8번) 정상성(중)

단 9번) 정지규칙(중)

단 10번) 향상도 (상)

<22회 ADSP>

1번) 데이터 분석 기반 가치 창출의 설명으로 부적절한 것?

--> 2) 복잡한 최적화 능력은 최고의 가치를 창출한다 (중)

2번) 효과적인 분석 모델 개발의 고려사항 중 부적절한 것?

--> 4) 모델 범위 바깥 요인을 판단하기 위해 가능한 많은 과거 상황 데이터를 모델에 포함한다. (상)

종식 의견: 신용평가 모형을 생각해 보면 모든 부실 가능성을 판단해서 모형을 개발할 수 없고 경기가 급격하게 변경되는 경우, 모형에 추가적인 평가를 적용하여 모형을 업그레이드 하면서 모형을 변경한다.

3번) 빅데이터의 정의로 거리가 먼 것?

--> 4) 작은 데이터도 하둡의 대용량 분선 처리 기술을 통해 가치 창출(하)

4번) 데이터 가치 측정이 어려운 이유?

--> 4) 전문 인력의 증가로 다양한 곳에서 빅데이터 활용 (중)

5번) 다각적 분석 차원의 데이터 사이언티스트의 역량이 아닌 것?

--> 3) 뉴럴네트워크 최적화 능력 (하)

6번) 머신러닝 학습 방법이 나머지와 성격이 다른 것?

--> 1) 군집분석 (중)

7번) 데이터를 가공 및 처리하여 얻을 수 있는 것이 아닌 것?

--> 4) 기호 (하)

8번) 복잡한 데이터 구조를 표현 및 관리하는 DBMS는 ?

--> 2) 객체지향 DBMS (상)

단1) SCM(Supply Chain Management) (중)

단2) 유전자 알고리즘 (상)

9번) 데이터 분석 방법론의 구성 요소가 아닌 것?

--> 3) 목적 (중)

10번) 빅데이터 분석 방법론의 분석기획의 위험에 대한 대응으로 부적절한 것?

--> 2) 관리 (상)

11번) 하향식 접근법의 과제 도출 단계?

--> 1) 문제탐색->문제정의->해결방안탐색->타당성검토

(하)

12번) 분석과제의 적용에서 시급성을 둔 우선순위?

--> 2) 3->4->2 (중)

13번) 빅데이터 분석 방법론의 분석기획 단계가 아닌 것은?

--> 1) 필요데이터 정의 (하)

14번) 분석 마스터 플랜 수립시 적용범위 및 방식의 고려요소가 아닌 것?

--> 3) 투입비용 수준(하)

15번) 모델캔버스를 5가지 요소로 줄인 것?

--> 2) 업무, 가치, 고객, 규제 및 감사, 지원 인프라(중)

16번) 분석 거버넌스 체계의 구성요소라 아닌 것?

--> 4) 과제 예산 및 비용집행 (하)

단3) 빅데이터 기획 전문가 ??? (상) --> 정답인지 확실하지 않지만 DBguide.net에서 교육하는 과정에 나옴.

단4) ISP(Information Strategy Planning) (중)

17번) 신뢰구간에 대한 설명이 옳지 않은 것?

--> 2) 95% 신뢰구간은 미지의 수가 포함되지 않을 확률이 95%다. (하)

18번) 작은 규모의 데이터 웨어하우스?

--> 2) 데이터마트(하)

19번) 출생지(서울특별시, 부산광역시, 경기도 등)의 척도?

--> 1) 명목척도(하)

20번) 시계열의 정상성을 만족하는 의미?

--> 4) 분산이 시점에 의존하지 않는다. (상)

21번) 응답자 1과 2의 피어슨 상관계수?

--> 1) 1 (중)

22번) 남학생일 때 사과를 좋아할 확률?

--> 3) 3/7 (중)

23번) ARIMA(1,2,3)의에서 차분은 몇 번 한 것인가?

--> 2) 2번 (하)

24번) 맨하튼 거리 계산?

--> 1) 25(중)

25번) 특이도 계산? 실제 F 100 개 중 40개를 F로 예측

--> 2) 4/10 (하)

26번) 지니지수 계산? (◇ - 2개, ○ - 2개)

--> 4) 1/2 (상)

27번) 의사결정나무에서 알고리즘의 설명이 잘못된 것?

--> 2) ㄴ- 엔트로피 (상)

28번) k- means 군집분석의 수행 순서?

--> 4) 다, 가, 라, 나 (중)

29번) 반응변수가 범주형일 때 회귀분석?

--> 3) 로지스틱 회귀분석 (하)

30번) 구매 순서가 고려된 연관성 기법?

--> 3) 순차패턴 (하)

31번) 숫자, 문자, 논리연산자가 포함된 벡터의 형식은?

--> 3) 문자형 벡터 (하)

32번) m<-matrix(c(1,2,3,4,5,6),ncol=2, byrow=T)?

--> 1) 1 2

3 4

5 6 (중)

33번) 입력은 리스트, 출력은 데이터프레임인 plyr 함수는?

--> 3) ldply() (상)

34번) 표본의 값에 대한 용어?

--> 1) 통계량 (하)

35번) 다중회귀분석에서 독립 변수 간의 상관관계가 존재하는 경우?

--> 1) 다중공선성 (중)

36번) 피어슨 상관계수에서 두 변수가 상관관계가 존재하지 않은 경우?

--> 4) 0 (하)

37번) 딥러닝 기법의 기반이 되는 모형은?

--> 2) 신경망 모형 (중)

38번) 다층 신경망에서 은닉층이 많아 학습이 이루어지지 않는 문제?

--> 1) 기울기 소실 문제 (상)

39번) ROC에서 완벽한 모형인 경우 x, y 값은?

--> 2) 0,1 (중)

40번) 이상치 판정 방법 중 가장 부적절하 것?

--> 3)Q2+1.5*IQR 보다 크거나 작으면 이상치로 인식 (중)

단 5번) 1일 확률이 0.3인 경우의 기댓값? 0.3 (중)

단 6번) 배깅(하)

단 7번) 홀드아웃(hold-out)(상)

단 8번) ESD(Extream Studentized Deviation) (중)

단 9번) 포아송분포 (상)

단 10번) ROC (하)