21회

01. 다음 중 빅데이터가 만들어 내는 변화와 가장 거리가 먼 것은? ③

① 가치가 있을 것이라고 예상되는 특정한 정보만 모아서 처리하는 것이 아니라 가능한 한 많은 데이터를 모으고 그 데이터를 다양한 방식으로 조합해 숨은 정보를 찾아내는 방식이 중요해진다.

② 데이터의 규모가 증가함에 따라 사소한 몇 개의 오류 데이터는 분석결과에 영향을 미치지 않기 때문에 데이터세트에 포함하여 분석해도 상관없는 경우가 많아진다.

③ 데이터의 양이 증가하고 유형이 복잡해짐에 따라 수많은 데이터 중에서 분석에 필요한 데이터를 선정하기 위해 정교한 표본조사 기법의 중요성이 대두되고 있다.

④ 인과관계의 규명 없이 상관관계 분석 결과만으로도 인사이트를 얻고 이를 바탕으로 수익을 창출할 수 있는 기회가 점차 늘어나고 있다.

02. 다음 중 빅데이터 현상이 출현하게 된 배경과 가장 거리가 먼 것은? ①

① 의료정보 등 공공데이터의 개방 가속화

② M2M, Iot와 같은 통신 기술의 발전

③ 하둡 등 분산처리 기술의 발전

④ 트위터, 페이스북 등 SNS의 급격한 확산

03. 다음 중 데이터베이스 설계 절차가 적절하게 배치된 것은? ②

① 요구사항 분석 _ 개념적 설계 — 논리적 설계 _ 물리적 설계

② 요구사항 분석 — 객체적 설계 — 논리적 설계 — 물리적 설계

③ 요구사항 분석 — 물리적 설계 — 개념적 설계 — 논리적 설계

④ 요구사항 분석 — 개념적 설계 — 객체적 설계 — 논리적 설계

04. 다음 중 "커피를 구매하는 사람이 탄산음료를 더 많이 구입하는가?"에 대한 문제를 해결한 빅데이터 분석 기법은 무엇인가? ②

① 유전알고리즘

② 연관 규칙 학습

③ 회귀 분석

④ 감성 분석

05. 다음 중 감성분석에 대한 설명으로 부적절한 것은? ①

① 사용자간의 사회적 관계를 알아내고자 할 때 이용됨

② 고객의 주관적 평가를 측정하고자 할 때 수행됨

③ 특정 주제에 대해 사용자의 긍정 부정 의견을 분석

④ 주로 문장이나 단어가 분석 대상이 됨

06. 기업 내부 데이터베이스를 기반으로 다양한 정보시스템이 구축•활용되고 있다. 고객 관련 데이터 베이스를 분석하여 고객 개개인에게 적합한 차별적 제품 및 서비스를 제공함으로써 고객과의 관계를 지속적으로 강화해 나가기 위해 구축하는 정보시스템은 다음 중 무엇인가? ①

① CRM 시스템

② SCM 시스템

③ ERP 시스템

④ KMS 시스템

07. 다음 중 개인정보 비식별화 기법을 설명한 것으로 부적절한 것은? ③

① 가명처리 - 개인 식별이 가능한 데이터에 대하여 직접적으로 식별 할 수 없는 다른 값으로 대체

② 범주화 - 단일 식별 정보를 해당 그룹의 대표 값으로 변환

③ 데이터마스킹 - 개인 정보 식별이 가능한 특정 데이터 값 삭제 처리

④ 총계처리 - 개별 데이터 값을 총합 또는 평균값으로 대체하는 것

08. 다음 중 데이터의 유형이 다른 하나는? ②

① 개인 페이스북에 올린 어느 회사 제품에 대한 사용 후기글

② 어느 기계에서 작동하는 동안 발생한 소음을 데시벨 단위로 기록한 센서 데이터

③ 어느 포털 사이트에서 하루 동안 언급된 모든 검색어

④ 콜센터에 접수된 어느 고객의 제품 불만사항을 녹음한 음성파일

09. 다음 중 CRISP-DM 방법론의 모델링 단계에서 수행하는 태스크가 아닌 것은? ④

① 모델 테스트 계획 설계

② 모델링 기법 선택

③ 모델 평가

④ 모델 적용성 평가

10. 다음 중 분석 프로젝트 관리에 대한 설명으로 가장 부적절한 것은? ③

① 분석 프로젝트 관리는 KSA ISO 21500:2013을 가이드로 활용할 수 있다.

② 데이터 분석 모델의 품질을 평가하기 위해서 SPICE를 활용할 수 있다.

③ 분석 프로젝트의 일정계획 수립 시 데이터 수집에 대한 철저한 통제와 관리가 필요하다.

④ 분석 프로젝트의 최종 산출물이 보고서 또는 시스템인지에 따라 프로젝트 관리에 차이가 있다.

11. 분석과제의 특징 중 Accuracy와 Precision에 대한 설명으로 가장 부적절한 것은? ④

① 분석의 활용적인 측면에서는 Accuracy가 중요하며 안정성 측면에서는 Precision0] 중요

② Accuracy는 모델과 실제 값과의 차이를 평가하는 정확도를 의미

③ Precision은 모델을 지속적으로 반복했을 때의 편차의 수준으로써 , 일관적으로 동일한 결과를 제시한다는 의미

④ Accuracy와 Precision은 Trade-Off 관계가 없음

12. 업에서 기존에 수행했던 데이터 분석 또는 BKBusiness Intelligence) 와 비교하여 , 빅데이터 분석에 대한 키워드를 가장 적절하게 표현한 것은? ④

① Clean Data, Statistical Analysis, Forecast, Predict

② Optimize, Predict, Forecast, Statistical Analysis

③ Alerts, Predict, Optimize, Ad hoc Report

④ Information, Ad hoc Report, Alerts, Clean Data

13. 분석의 대상 및 방식에 따라서 분석의 주제는 크게 4가지 유형으로 나뉜다. 이 중 분석 대상은 명확하지만 분석 방식이 명확하지 않은 경우 수행하는 주제 유형으로 가장 적절한 것은? ④

① Optimization 유형

② Discovery 유형

③ Insight 유형

④ Solution 유형

14. 빅데이터의 특징을 고려한 분석 ROI 요소와 분석우선순위 평가기준에 대한 설명으로 가장 부적절한것은? ①

① 분석과제의 우선순위 평가에서 시급성은 전략적 중요도 , 데이터 수집비용 등을 평가하고 난이

도는 분석 수준과 복잡도가 평가요소이다.

② 분석 난이도는 분석 준비도와 성숙도 진단 결과에 따라 해당 기업의 분석 수준을 파악하고 이

를 바탕으로 결정된다.

③ 시급성이 높고 난이도가 높은 분석과제는 경영진 또는 실무 담당자의 의사결정에 따라 적용 우선순위를 조정할 수 있다.

④ 시급성이 높고 난이도가 낮은 분석과제는 우선순위가 높다.

15. 다음 중 CRISP-DM 분석 방법론에서 업무 이해(Business 니nderstanding)에 해당하는 것은? ②

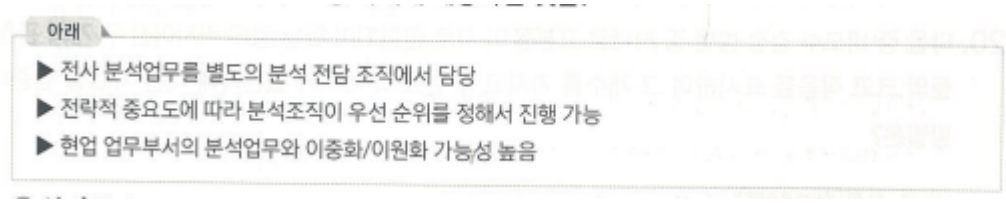
① 초기 데이터 수집 - 데이터 기술 분석 - 데이터 탐색 - 데이터 품질 확인

② 업무 목적 파악 - 상황 파악 - 데이터 마이닝 목표 설정 - 프로젝트 계획 수립

③ 모델링 기법 선택 - 모델 테스트 계획 설계 - 모델 작성 - 모델 평가

④ 분석 결과 평가 - 모델링 과정 평가 - 모델 적용성 평가

16. 데이터 분석을 위한 조직 구조 중 아래에 해당하는 것은? ③



① 분산구조

② 기능구조

③ 집중구조

④ 복합구조

17. 다음 중 연속형 확률 변수의 분포 중 정규분포로부터 유도되었으며 , 정규 분포의 평균을 측정할 때 주로 사용되는 분포로 두 집단의 평균 차이 검증 등에 활용되는 분포는? ③

① 균일분포(uniform distribution)

② 지수분포(exponential distribution)

③ t-분포 (t-distribution)

④ F_ 분포 (F-distribution)

18. 이산형 확률 분포 중 하나로 개별 사건이 두 가지 경우만 존재하며 , 각 사건이 성공할 확률이 일정하고 전후 사건에 독립적인 특수한 상황의 확률 분포를 나타내는 것은? ④

① 포아송분포

② 지수분포

③ 다항분포

④ 베르누이 확률분포

19. 다음 중 확실하게 증명하고 싶은 가설 , 뚜렷한 증거가 있어야 채택할 수 있는 가설(hypothesis)은? ①

① 대립가설

② 영가설

③ 귀무가설

④ 기각가설

20. 다음 중 비모수 검정 방법 중 하나로 표본들이 서로 관련되어 있는 경우 짝지어진 두 개의 관찰치들의 크고 작음을 표시하여 그 개수를 가지고 두 분포의 차이가 있는지에 대한 가설을 검증하는 방법은? ③

① 런 검정 (xun test)

② 만-위트니의 U검정

③ 부호 검정 (sign test)

④ 스피어만 순위상관계수

21. 다음 중 소득 수준과 같이 정규 분포를 따르지 않고 오른쪽 꼬리가 긴(right-skewed)분포를 나타내는 자료의 평균과 중앙값의 관계로 옳은 것은 무엇인가? ④

① 자료의 크기(scale)에 따라 달라진다.

② 평균이 중앙값보다 작은 경향을 보인다.

③ 평균과 중앙값이 일치하는 경향을 보인다.

④ 평균이 중앙값보다 큰 경향을 보인다.

22. 다음 중 중앙 50%의 데이터들이 흩어진 정도를 의미하는 것은 무엇인가? ①

① 사분위범위 (interquartile range)

② 중앙값 (median)

③ 표준편차 (standard deviation)

④ 평균(mean)

23. 다음 중 한 변수를 단조 증가 함수로 변환하여 다른 변수를 나타낼 수 있는 정도를 나타내며 두

변수의 선형 관계의 크기 뿐만 아니라 비선형적인 관계도 나타낼 수 있는 상관계수는 무엇인가? ③

① 코사인 유사도

② 피어슨 상관계수

③ 스피어만 상관계수

④ 자카드 인덱스

24. 다음 분류 분석 모형 중 훈련용 데이터 집합으로부터 미리 모형을 학습하는 것이 아니라 새로운 자료에 대한 예측 및 분류를 수행할 때 모형을 구성하는 lazy learning기법을 사용하는 것은 무엇인가? ②

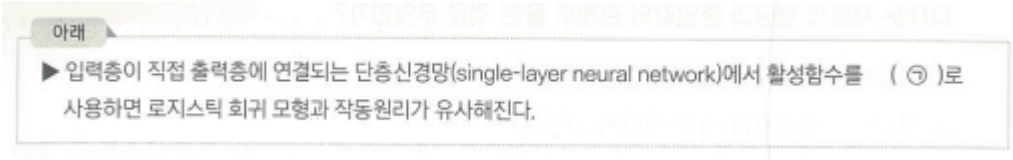
① 유전자 알고리즘(genetic algorithm)

② 최근접 이웃(nearest neighbor) 모형

③ 신경망(artificial neural network) 모형

④ 서포트 벡터 기 계 (support vector machine)

25. 다음 중 아래 (㉠)에서 설명하는 활성함수로 가장 적절한 것은? ④



① 계단(step) 함수

② tanh 함수

③ ReLU 함수

④ 시그모이드(sigmoid) 함수

26. 앙상블(ensemble) 모형은 여러 모형의 결과를 결합함으로써 단일 모형으로 분석했을 때보다 신뢰성 높은 예측값을 얻을 수 있다. 다음 중 앙상블 모형의 특징으로 옳지 않은 것은? ④

① 이상값(outlier)에 대한 대응력이 높아진다.

② 전체적인 예측값의 분산을 감소시켜 정확도를 높일 수 있다.

③ 모형의 투명성이 떨어져 원인 분석에는 적합하지 않다.

④ 각 모형의 상호 연관성이 높을수록 정확도가 향상된다.

27. 다음 중 의사결정나무를 앙상블(ensemble)하는 방법 중 전체 변수 집합에서 부분 변수 집합을 선택하여 각각의 데이터 집합에 대해 모형을 생성한 후 결합하는 방식은? ③

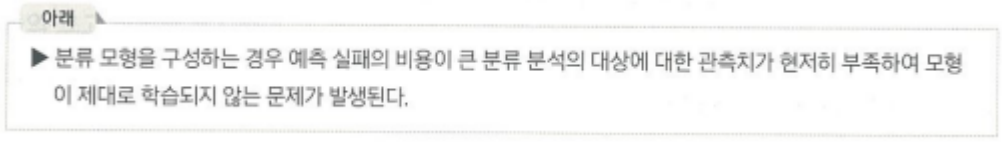
① 부스팅 (boosting)

② 배깅(bagging)

③ 랜덤포레스트(random forest)

④ 붓스트랩 (bootstrap)

28. 다음 중 아래에서 설명하는 문제를 나타내는 용어로 적절한 것은? ③



① 과대적합 문제(overfitting problem)

② 과소적합 문제(underfitting problem)

③ 범주 불균형 문제(case imbalance problem)

④ 정보과부하 문제(information overload problem)

29. 분류 모형의 평가 기준 중 정확도(precision)와 재현율(recall)은 한 지표의 값이 높아지면 다른 지표의 값이 낮아질 가능성이 높다. 이러한 효과를 보정하여 하나의 지표로 나타낸 Fp 지표에서 P=2일 경우 다음 설명 중 옳은 것은? ④

① 정확도(precision)에 2배만큼의 가중치를 부여하여 조화 평균한다.

② 정확도(precision)와 재현율(recall)을 조화 평균한 뒤 0.5배한다.

③ 정확도(precision)와 재현율(recall)을 조화 평균한 뒤 2배한다.

④ 재현율(recall)에 2배만큼의 가중치를 부여하여 조화 평균한다.

30. 다음 중 인공 신경망 모형에서 역전파를 진행함에 따라 각 노드를 연결하는 가중치의 절대값이 커져 조정이 더 이상 이루어지지 않아 과소적합(underfitting)이 발생되는 문제는? ③

① 비선형 문제

② 전역최적화 문제

③ 포화 문제

④ 수용 영역 국소화 문제

31. 군집화 기법 중 특정 공간에서 가까이 있는 데이터가 많은 지역을 중심으로 클러스터를 구성하며

비교적 비어 있는 지역을 경계로 하는 군집 기법으로 임의적인 모양의 군집 탐색에 효과적인 기법은 무엇인가? ③

① 계층적 군집 기법

② 분리 군집 기법

③ 밀도 기반군집 기법

④ 격자 기반 군집 기법

32. 군집 모형 평가 기준 중 하나이며 군집의 밀집정도를 계산하는 방법으로 군집 내의 거리와 군집 간의 거리를 기준으로 군집 분할의 성과를 평가하는 것은 다음 중 무엇인가? ④

① 피어슨 상관 계수(Pearson Correlation Coefficient)

② ARI(Adjusted Rand Index)

③ NMI(Normalized Mutual Information)

④ 실루엣 계수(Silhouette Coefficient)

33. 이상값 탐색을 위해 상자그림(boxplot)을 사용하려 한다. 아래와 같은 데이터 요약 결과가 있을 때 , 다음 중 이상값을 판단하는 하한선, 상한선으로 옳은 것은? ④

아래

```
> summary( x )
   Min.   1st Qu.  Median    Mean   3rd Qu.  Max.
   0       4       7       9.615   12       39
```

① (-12, 36)

② (4, 12)

③ (-2, 30)

④ (-8, 24)

34. 다음 군집 모형 중 군집의 개수를 미리 지정하지 않아도 되는 장점으로 탐색적 분석에 사용하는 모형은 무엇인가? ③

① k-평균군집모형

② SOM(Self-Organizing Maps) 모형

③ 계층적 군집

④ 혼합분포군집 모형

35. 계층적 군집은 두 개체 간의 거리에 기반하여 군집을 형성해나간다. 두 개체간의 거리 측도 중 , 두 벡터 사이의 각도를 이용하여 벡터간의 유사 정도를 측정하는 측도는 다음 중 무엇인가? ③

① 자카드 유사도

② 피어슨 유사도

③ 코사인 유사도

④ 캔버라 거리

36. 다음 중 자기조직화지도(Self-Organizing Maps, SOM)에 대한 것으로 옳은 것은? ④

① 군집 분할을 위해 역전파 알고리즘을 사용한다.

② 지도(map) 형태로 형상화가 이루어지지만 입력 변수의 위치 관계를 보존하지는 않는다.

③ 학습횟수(epochs)와 군집내 거리는 반비례한다.

④ 승자 독점의 학습 규칙에 따라 입력 패턴과 가장 유사한 경쟁층 뉴런이 승자가 된다.

37. k-평균군집은 단순하고 빠르게 수행될 수 있지만 변수의 크기(scale)에 영향을 받음에 따라 군집 분석을 수행하기 전에 정규화(normalization) 과정이 필수적이다. 다음 정규화 방법 중 원(raw) 데이터의 분포를 유지하면서 정규화가 가능한 방법은 무엇인가? ③

① z-score 정규화

② 로그 정규화

③ min-max 정규화

④ 벡터 정규화

38. 아래는쇼핑몰의 거래내역이다. 연관규칙 "우유ㅡ 커피"에 대한 지지도(Support)는 얼마인가? ③

아래

항목	거래수
우유	10
커피	20
{우유, 커피}	30
{커피, 초코렛}	40
전체 거래 수	100

① 0.1

② 0.2

③ 0.3

④ 0.4

39. 다음 중 아래 오분류표에 대한 F1 은 얼마인가? ③

아래

		예측치		합계
		True	False	
실제값	True	30	70	100
	False	60	40	100
합계		90	110	200

① 4/10

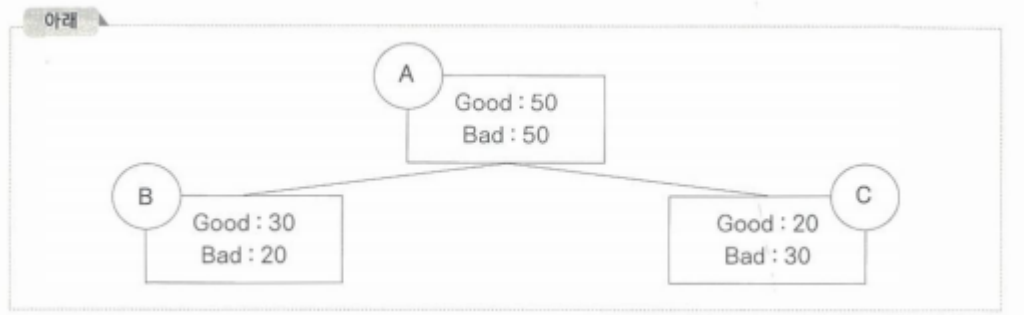
② 18/57

③ 3/8

④ 7/11

40. 아래는 의사결정나무를 나타낸 것이다. C의 지니 지수(gini index)는 얼마인가? ②

아래



① 0.2

② 0.48

③ 0.4

④ 0.32

01. 아래에서(㉠)안에공통적으로 들어갈말로 적절한 것은

아래

▶ (㉠) (이)란 데이터로부터 의미있는 정보를 추출해 내는 학문으로, 통계학과는 달리 정형 또는 비정형을 막론하고 다양한 유형의 데이터를 분석 대상으로 한다. 또한 분석에 초점을 두는 데이터마이닝과는 달리 (㉠) (은)는 분석 뿐 아니라 이를 효과적으로 구현하고 전달하는 과정까지 포함하는 포괄적인 개념이다.

(데이터사이언스)

02. 아래는 빅데이터 활용 기본 테크닉 중 어떤 분석에 관한 설명인가?

아래

▶ 은행에서 대출 심사를 할 때, 소득, 카드 사용액, 나이 등 해당 고객의 개인적인 정보를 바탕으로 그 고객이 대출 상환을 잘 하는 집단에 속할지 그렇지 않은 집단에 속할지를 예측할 수 있다.

(분류분석, 유형분석)

03. 아래 ( ) 안에 각각 들어갈 용어로 적절한 것은?

아래

▶ 분석 과제 관리 프로세스는 크게 과제 발굴과 (㉠)(으)로 나누어진다. 조직이나 개인이 도출한 분석 아이디어를 발굴하고 이를 과제화하여 분석 과제 풀(Pool)로 관리하면서 분석과제가 확정되면 (㉡), (㉢), 분석과제 결과 공유/개선의 분석과제 관리 프로세스를 수행하게 된다.

(㉠ 과제수행, ㉡㉢ 탐구성, 분석과제실행, 분석과제 진행관리)

04. 상향식 접근 방식의 발산단계와 도출된 옵션을 분석하고 검증하는 하향식 접근 방식의 수렴단계를 반복하여 과제를 발굴하는 방법을 무엇이라고 하는가?

(디자인 사고(Design Thinking))

05. 아래는 주성분 분석을 수행한 결과이다. 첫 번째 분산은 전체 분산의 몇 %를 설명하고 있는가?

(소수점 첫째자리까지 표시하시오)

아래

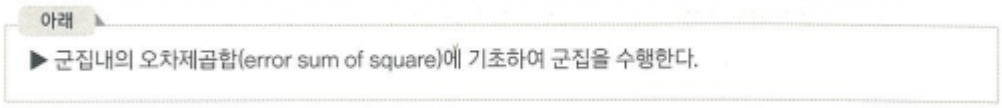
	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.5574873	0.9943214	0.5943221	0.4123679
Proportion of Variance	0.5748331	0.2321003	0.1834561	0.0096105
Cumulative Proportion	0.5748331	0.8069334	0.9903895	1.0000000

(57.4%)

06. 여러 대상 간의 거리가 주어져 있을 때 , 대상들을 동일한 상대적 거리를 가진 실수 공간의 점들로 배치시키는 방법을 무엇이라 하는가?

(다차원척도법)

07. 계층적군집을 수행할 때 두 군집간의 거리를 측정하는 방법 중 아래에서 설명하는 방법은?



(와드연결법)

08. 회귀 모형의 가정 중 잔차항이 정규분포를 이루어야 하는 가정을 의미하는 용어는 무엇인가?

(정상성(정규성))

09. 의사결정 나무에서 더 이상 분기가 되지 않고 현재의 마디가 끝마디(leaf node, terminal node)가 되도록 하는 규칙을 나타내는 용어는 무엇인가?

(정지규칙)

10. 연관규칙의 측정 지표 중 도출된 규칙의 우수성을 평가하는 기준으로 두 품목의 상관관계를 기준으로 도출된 규칙의 예측력을 평가하는 지표는 무엇인가?

(향상도)