ADsP 34회 기출문제 답안

【 객관식 정답 】

01	④	11	④	21	③	31	②
02	②	12	②	22	④	32	②
03	③	13	④	23	③	33	③
04	②	14	③	24	③	34	③
05	④	15	①	25	④	35	③
06	②	16	②	26	①	36	④
07	①	17	②	27	③	37	④
08	③	18	①,④	28	③	38	①
09	③	19	④	29	③	39	④
10	④	20	②	30	④	40	③

【 주관식 정답 】

01	정보(Information)
02	사물인터넷 (IoT, Internet of Things)
03	문제 정의
04	데이터 거버넌스
05	랜덤 포레스트 (Random Forest)
06	부스팅(Boosting)
07	역전파 알고리즘
08	실루엣(Shilouette)
09	점 추정
10	지니 지수

01. 데이터가 커진다고 해서 분석에 많이 사용되는 것이 아니라 데이터에 따른 적정한 분석 방법이 경쟁우위를 가져다준다고 할 수 있다.

02. 데이터의 가치를 측정하기 어려운 이유는 다음과 같다.
- 데이터 활용 방식: 재사용, 재조합(mashup), 다목적용 개발
- 새로운 가치 창출
- 분석 기술 발전

03. 데이터베이스에 있는 데이터 중 분석이 불가한 데이터들도 있으며, 그런 데이터는 처리 등을 통해 분석에 활용할 수 있다.

04. 일차적인 분석을 통해서도 해당 부서나 업무 영역에서는 상당한 효과를 얻어낼 수 있다.

05. 데이터 오용은 베트남 전쟁 시 적군 사망자 수를 과장해 보고하는 것을 통해 알 수 있었다.

06. 데이터 사이언스는 통찰력 있는 분석에 초점을 두고 진행한다.

07. 데이터 마트는 데이터웨어하우스로부터 구축된 데이터 속에서 한 가지 주제 또는 한 부서 중심으로 구축된 소규모, 단일 주제의 웨어하우스로 예측 가능한 질의에 대해서 매우 빠르게 응답할 수 있도록 데이터를 제공하는 시스템이다.

08. 데이터가 많다고 무조건 더 많은 가치가 창출되지는 않는다.

09. 타당성 검토 단계에서는 효과적으로 평가하기 위해서 비즈니스 지식과 기술적 지식이 요구되기 때문에 비즈니스 분석가, 데이터 분석가, 시스템 엔지니어 등과의 협업이 수반되어야 한다.

10. Precision은 모델을 지속적으로 반복했을 때의 편차의 수준으로써 동일한 결과를 제시한다는 것을 의미한다.

11. 문제 탐색 단계에서 현재의 비즈니스 모델 및 유사·동종사례 탐색을 통해서 도출한 분석 기회들을 구체적인 과제로 만들기 전에 분석 유즈케이스로 표기하는 것이 필요하다.

12. 데이터 분석 준비 프레임워크에서 분석 업무 파악 영역에는 발생한 사실 분석 업무, 예측 분석 업무, 시뮬레이션 분석 업무, 최적화 분석 업무, 분석 업무 정기적 개선이 있다.

13. 시스템 구현 단계에서 정보보안영역과 코딩은 주요 고려사항이 아니다. 시스템 설계 및 구현, 테스트 및 운영이 주요 고려사항이다.

14. 분석과제 정의서를 통해 분석별로 필요한 소스 데이터, 분석방법, 데이터 입수 및 분석의 난이도, 분석 수행주기, 분석결과에 대한 검증 오너십, 상세 분석 과정 등을 정의한다. 분석 데이터 소스는 내·외부의 비구조적인 데이터와 소셜 미디어 및 오픈 데이터까지 범위를 확장하여 고려하고 분석방법 또한 상세하게 정의한다.

15. 분석 프로젝트 관리방안에서 시간관리는 프로젝트의 활동 일정을 수립하고 일정 통제의 진척상황을 관찰하는데 요구되는 프로세스이다.

16. 데이터 준비 단계에서는 분석용 데이터 셋 선택, 데이터 정제, 분석용 데이터 셋 편성, 데이터 통합, 데이터 포맷팅 수행업무가 있다.

17. 상자 그림으로는 이상치를 확인할 수 있다.

18. 1~4번 중 지지도가 25%이상인 규칙은 1번과 4번이다. 그 중 신뢰도가 50%이상인 규칙은 A→B, B→C 이다.

19. p, d, q에 따라서 각각 0 이면 IMA(d,q), ARMA(p,q), ARI(p,d)모형으로 부를 수 있다. 이 중 IMA(d,q)를 d번 차분하면 MA(q) 모형을 따른다.

20. ROC곡선의 좌표는 (1-특이도, 민감도)로 x축이 낮고 y축이 높을수록 분류정확도가 높다는 것을 의미하므로 이상적으로 완벽히 분류한 모형의 좌표는 (0,1)이다.

21. 적절한 세분화로 인한 품목 결정이 장점이지만 너무 세분화된 품목은 의미 없는 결과를 도출한다.

22. K 값이 작을수록 과대적합(Overfitting) 문제가 발생한다.

23. 시그모이드 함수를 단층신경망에서 활성화함수로 사용하면 로지스틱 회귀모형과 작동원리가 유사하다.

24. 군집의 분리에 대해 안정성도 중요 하지만 해당 군집에 대한 분리가 논리적으로 설명이 되는 부분이 더 중요하다고 할 수 있다.

25. 일반적으로 정상성을 만족하지 않을 때는 log, root를 취하여 정규분포를 취하도록 만든다.

26. 의사결정나무 모형은 지도학습 모형으로 하향식 의사결정에 가깝다고 생각할 수 있다.

27. 앙상블은 주어진 자료로부터 여러 개의 예측모형들을 만든 후 예측모형들을 조합하여 하나의 최종 예측 모형을 만드는 방법으로 Bagging, Boosting, Random Forest, Stacking 등이 있다.

28. 연관분석은 실시간 상품추천을 통한 교차판매 등에 활용할 수 있다.

29. 모형에서 종속변수와 독립변수 간의 상관계수가 유의한지는 상관관계 분석을 통해 확인한다.

30. 지수평활법은 시간의 흐름에 따라 최근 시계열에 더 많은 가중치를 부여하여 미래를 예측하는 방법이다.

31. 회귀분석 결과에서 분석이 잘 되었다면 잔차는 더 이상 독립변수와 상관관계를 가지지 않는다.

32. 2개의 주성분으로 자료를 축약할 때 전체 분산의 74.6%가 설명 가능하다.

33. Balance와 가장 상관관계가 높은 변수는 Limit와 Rating이다.

34. Recall은 $\frac{TP}{TP+FN}$ 이므로 $\frac{40}{40+60}=0.4$ 이다.

35. Apriori 알고리즘은 최소 지지도를 설정하고 개별 품목 중 최소 지지도가 넘는 모든 품목을 먼저 찾는다. 그리고 개별 품목만으로 최소 지지도가 넘는 2가지 품목을 찾고, 이것들을 결합해 3가지 품목집합을 찾으며 반복해 빈발품목집합을 찾는다.

36. m개의 주성분은 원래 변수들 중 서로 상관성이 높은 변수들의 선형결합으로 만들어진 것이다.

37. 피어슨 상관계수는 연속형 변수에 사용하며 정규성을 가정한다. 스피어만 상관계수는 순서형 변수에 사용하며 비모수적 방법이다.

38. 배깅(Bagging)은 주어진 자료에서 여러 개의 붓스트랩(bootstrap) 자료를 생성하고 각 붓스트랩 자료에 예측모형을 만든 후 결합하여 최종 예측모형을 만드는 방법이다.

39. 가설검정은 어떤 모수의 값 또는 확률분포에 대하여 가설을 세우고 이 가설이 맞다고 주 장해도 이상이 없는지를 표본 데이터의 통계적 확률에 의해 결정하는 것을 말한다.

40. 상관관계 분석으로는 상관관계를 파악할 수 있다. 인과관계는 회귀분석에서 확인할 수 있다.

【단답형】

단답형 01. 정보(Information)

단답형 02. 사물 인터넷(IoT, Internet of Thing)

단답형 03. 문제 정의

단답형 04. 데이터 거버넌스

단답형 05. 랜덤 포레스트(Random Forest)는 의사결정 나무 여러 개로 만들어진 모델이다.

단답형 06. 부스팅(Boosting)은 예측력이 약한 모형들을 결합하여 강한 예측모형을 만드는 방법으로 GBM, XgBoost, LightGBM 등이 있다.

단답형 07. 출력값에 대한 입력값의 기울기(미분값)을 출력층 layer에서부터 계산하여 거꾸로 전파시키는 것이다.

단답형 08. 실루엣(Shilouette) 계수는 군집 모형 평가 기준 중 군집의 밀집 정도를 계산하는 방법으로 군집 내의 거리와 군집간의 거리를 기준으로 군집 분할의 성과를 평가하는 것이다.

단답형 09. 점추정이란 추정하고자 하는 하나의 모수에 대하여 모집단에서 임의로 추출된 n개 표본의 확률변수로 하나의 통계량을 만들고 주어진 표본으로부터 그 값을 계산하여 하나의 수치를 제시하려고 하는 것이다.

단답형 10. 지니지수는 노드의 불순도를 나타내는 값으로 지니지수의 값이 클수록 이질적이며 순수도가 낮다.