ADsP 31회 기출문제 답안

【 객관식 정답 】

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

【 주관식 정답 】

01	머신러닝 또는 기계학습
02	KMS(지식관리시스템)
03	ISP(정보전략계획)
04	IT 인프라
05	나이브 베이즈 분류
06	최단연결법(단일연결법)
07	스테밍(Stemming) 또는 어간 추출
08	차분
09	57.5%
10	배깅(Bagging)

01. 사물인터넷(Internet of Things)는 인터넷을 기반으로 모든 사물을 연결해 사람과 사물, 사물과 사물 간의 정보를 상호 소통하는 지능형 기술 및 서비스이며, 사물에서 생성되는 Data를 활용한 분석을 통해 마케팅 등에 활용할 수 있다.

02. 기능구조는 별도 분석조직이 없고 해당업무부서에서 분석 수행한다.

03. 통합된 데이터는 동일한 내용의 데이터가 중복되어 있지 않다는 것을 의미한다.

04. (a)는 가공하기 전의 순수한 수치나 기호인 데이터를 의미하며, (b)는 데이터의 가공 및 상관관계 이해를 통해 패턴 인식하고 의미를 부여하는 정보를 의미한다. (c)는 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물인 지식을 의미한다.

05. 재무관리분야의 분석 유즈 케이스에는 일별로 예정된 자금 지출과 입금을 추정하는 자금시재예측 등이 있다.

06. 다) 개인정보 사용자의 정보사용에 대한 무한책임의 한계로 개인정보 사용 동의제보다 책임체로 더욱 강화시켜야 한다.
라) 민주주의 국가의 형사 처벌과 같이 잠재적 위협이 아닌 명확하게 행동한 결과에 대해 책임을 묻기 때문에 빅데이터 사전 성향 분석을 실시한다면 책임 원칙을 훼손한다.

07. 데이터베이스는 종속성과 중복성을 배제한다. 데이터 종속성이란 응용프로그램별로 데이터를 별도 관리한다.

08. 빅데이터의 경우 데이터양의 급증으로 데이터의 생명 주기(수명주기) 관리방안을 수립하지 않으면 데이터 가용성 및 관리비용 증대되는 문제에 직면하게 될 수 있다.

09. 분산구조는 별도 분석전담조직, 분석조직 인력을 현업부서로 직접 배치한다.

10. 솔루션은 분석대상은 알고 있지만, 분석의 방법을 모를 때 나타나는 분석 주제의 유형이다.

11. 분석 유즈 케이스는 기업의 전사 또는 개별 업무별 주요 의사결정 포인트에 활용할 수 있는 분석의 후보들을 의미한다.

12. 반복적으로 위험분석을 수행하여 위험을 관리하며 순환적으로 개선하는 것은 나선형 모델이다.

13. 분석과제 발굴 방법론에서 상향식 접근 방식은 문제의 정의 자체가 어려운 경우, 데이터를 기반으로 문제의 재정의 및 해결방안을 탐색하고 이를 지속적으로 개선하는 방식이다.

14. (A) 단순히 대용량 데이터를 수집·축적하는 것보다는 어떤 목적으로 어떤 데이터를 어떻게 분석에 활용할 것인가가 더욱 중요하다.
    (B) 빅데이터의 경우 데이터양의 급증으로 데이터의 생명 주기 관리방안을 수립하지 않으면 데이터 가용성 및 관리비용 증대 문제에 직면할 수 있다.

15. 전략적 통찰력을 얻기 위해서는 내부뿐만 아니라 외부환경을 같이 분석해야 한다.

16. 분석 마스터플랜의 우선순위 고려요소는 전략적 중요도, 비즈니스 성과/ROI, 실행 용이성이 있으며, 적용 범위/방식의 고려요소는 업무 내재화 적용 수준, 분석데이터 적용 수준, 기술적용 수준이 있다.

17. 데이터 마트는 데이터웨어하우스와 사용자 사이에 위치한 것으로, 하나의 주제 또는 하나의 부서 중심의 데이터웨어하우스라고 할 수 있다.

18. 지지도는 전체 거래 중 항목 A와 항목 B를 동시에 포함하는 거래의 비율이다.

19. 향상도는 $\frac{P(B \mid A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)}$ 로 계산할 수 있다.

20. C의 지니지수는 $1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2 = 0.48$ 로 구할 수 있다.

21. MAPE는 MAE를 퍼센트로 변환한 것으로 실제값(A_t)과 예측값(F_t)으로 구할 수 있다.

$$MAPE = \left(\sum_{t=1}^{n} \frac{|A_t - F_t|}{A_t}\right) \div n \times 100 = 0.4/4 \times 100 = 10\%$$

22. 평균은 이상치에 민감한데 이를 보완하기 위해 군집의 가장 중심에 있는 값(메도이드(medoid))을 사용하여 군집을 찾는 방법이 k-medoids 군집화 방식이다.

23. 부스팅은 붓스트랩 표본을 구성하는 재표본 과정에서 분류가 잘못된 데이터에 더 큰 가중치를 주어 표본을 추출하는 기법/예측력이 약한 모형들을 결합하여 강한 예측모형을 만드는 방법이다.

24. 일반적으로 학습 모형의 유연성이 클수록 분산은 높고 편향은 낮다.

25. 데이터 마이닝은 대용량 데이터에서 의미있는 패턴을 파악하거나 예측하여 의사결정에 활용하는 방법이다.

26. Estimate값인 회귀계수는 Intercept의 회귀계수와의 차이를 의미하고 종속변수 wage의 평균과 차이는 아니다.

27. SOM은 비지도 학습에 해당하고 나머지 항목은 지도학습에 해당한다.

28. k-means clustering, Single linkage method, DBSCAN은 군집분석 방법이다.

29. Q1 - 1.5 * IQR ≤ 데이터 ≤ Q3 + 1.5* IQR 범위를 벗어난 데이터를 이상치라고 한다. 160은 중앙값(Q2)이고 백분율을 4등분 하였으므로 각각의 범위에는 25%의 데이터가 있다. IQR은 Q3-Q1으로 200-140=60이다.

30. 층화추출법은 이질적인 원소들로 구성된 모집단에서 각 계층을 고루 대표할 수 있도록 표본을 추출하는 방법.

31. 연관분석은 흔히 장바구니분석 또는 서열분석이라고 불리며, 상품의 구매, 서비스 등 일련의 거래 또는 사건들 간의 규칙을 발견하기 위해 적용한다.

32. 스피어만 상관계수는 서열척도인 두 변수들의 상관관계 측정방식으로 순위를 기준으로 상관관계를 측정하는 비모수적 방법이다.

33. 연결정도는 해당 노드에 직접 연결되어 있는 노드 또는 링크의 수이다.

34. 마할라노비스 거리는 통계적 개념이 포함된 거리이며 변수들의 산포를 고려하여 이를 표준화한 거리이다.

35. 분류분석은 레코드의 특정 속성의 값이 범주형으로 정해져 있으며 데이터의 실체가 어떤 그룹에 속하는지 예측하는데 사용되는 기법으로 사기방지모형, 이탈모형, 고객 세분화 모형 등을 개발할 때 활용한다.

36. matrix 함수는 행렬을 만드는 함수로 c(1,2,3,4,5,6)를 통해 행렬을 구성하는 원소를 입력한다.
그리고 ncol=2는 컬럼의 개수이며, byrow=T는 행방향(가로축)으로 채워넣는 인자이다.

37. 가, 다, 마는 지도학습이며, 나, 라는 비지도 학습이다.

38. 변수 k의 분산팽창요인 $\mathrm{VIF}_k = \frac{1}{1-R_k^2}$ 으로 결정계수에 영향을 받는다. 결정계수는 회귀모델에서 독립변수가 종속변수를 얼마나 잘 설명하는지를 나타내는 것으로 회귀식의 기울기와는 관계가 없다.

39. 순환변동은 경제적이나 자연적인 이유 없이 알려지지 않은 주기를 가지고 변화하는 자료를 의미한다.

40. 평균 고유값 방법은 고유값들의 평균을 구한 후 고유값이 평균값 이상이 되는 주성분을 제거하는 것이 아니라 설정하는 것이다.

【 단답형 】

단답형 01. 머신러닝은 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야이다.

단답형 02. 지식관리시스템은 기업의 환경이 물품을 주로 생산하던 산업사회에서 지적 재산의 중요성이 커지는 지식사회로 급격히 이동함에 따라, 기업 경영을 지식이라는 관점에서 새롭게 조명하는 접근방식이다.

단답형 03. 정보기술 또는 정보시스템을 전략적으로 활용하기 위하여 조직 내/외부 환경을 분석하여 기회나 문제점을 도출하고 사용자의 요구사항을 분석하여 시스템 구축 우선순위를 결정하는 등 중장기 마스터 플랜을 수립하는 절차이다.

단답형 04. 분석 준비도 중 IT인프라에는 운영시스템 데이터 통합, EAI, ETL 등 데이터 유통 체계, 분석 전용 서버 및 스토리지, 빅데이터 분석 환경, 통계 분석 환경, 비쥬얼 분석 환경이 있다.

단답형 05. 나이브 베이즈 분류는 특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기의 일종으로 1950년대 이후 광범위하게 연구되고 있다.

단답형 06. 최단연결법은 가장 가까운 데이터를 묶어서 군집을 형성한다.

단답형 07. 스테밍(Stemming)은 어형이 변형된 단어로부터 접사 등을 제거하고 그 단어의 어간을 분리해내는 것을 의미한다.

단답형 08. 차분은 시계열의 수준에서 나타내는 변화를 제거하여 시계열의 평균 변화를 일정하게 만드는 것을 돕는다.

단답형 09. 첫 번째 분산(Proportion of Variance)은 0.5748331로 나타났다.

단답형 10. 배깅(Bagging)은 원 데이터 집합으로부터 크기가 같은 표본을 여러 번 단순임의 복원추출하여 각 표본에 대해 분류기를 생성한 후 그 결과를 앙상블하는 방법이다.