ADsP 단답형 정리

15-단1

데이터 사이언티스트의 요구 역량 두가지, 이론적 지식이나 분석 기술에 대한 숙련은 ㅇㅇ skill이고 통찰력 있는 분석이나 설득력은 ㅁㅁ skill이다.

**답 : 하드, 소프트**

15-단2

SQL 언어에서 A와 B 사이에 있는 값을 구하는 함수는?

**답 : BETWEEN**

15-단3

구축된 시스템의 검증을 위해 단위 테스트, 통합 테스트, 시스템 테스트 등을 실시한다. ㅇㅇ테스트는 품질 관리 차원에서 진행함으로써 적용된 시스템의 객관성과 완전성을 확보한다.

**답 : 시스템**

15-단4

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 ㅇㅇ도 수행되어야 효율적인 활용이 가능하다.

**답 : 사전 영향 평가**

15-단5

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

**답 : 후진제거법**

15-단6

상자그림 결과값에서 3/4분위수보다 큰 자료는 몇 % 인가?

**답 : 25%**

15-단7

민감도와 1-특이도를 활용하여 모형을 평가하는 그림은 무엇인가?

**답 : ROC Curve**

15-단9

데이터마이닝 성과분석 중 오분류에 대한 추정치에서 TN/(FP+TN)은 뭐라고 하는가?

**답 : Specificity**

16-단1

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

**단 : 블록체인**

16-단2

수치로 명확하게 표현되는 데이터

**답 : 정량적 데이터**

16-단3

CRISP-DM에서 잡음, 이상치, 결측치 식별 및 제거와 데이터셋을 선택하는 단계는?

**답 : 데이터 준비**

16-단4

데이터 분석 기획에서 데이터 분석가에게 요구되는 기술은?

**답: 소프트스킬**

16-단5

로지스틱 회귀모형에서 단위가 증가할 때마다 성공의 ㅇㅇ이 몇 배 증가하는가?

**답 : 오즈(odds)**

16-단6

주성분 분석에서 3차원으로 축소할 경우 잃게 되는 정보량은?

**답 : 1 - 0.7966 = 0.20**

16-단7

베이즈 정리를 활용한 분류방법 알고리즘은?

**답 : 나이브 베이지안 분류**

16-단8

오분류표의 F1값은?

**답 : F1 = 2\*(재현율\*정확도)/(재현율+정확도)**

16-단9

시점에 상관없이 시계열의 특성이 일정하다는 용어는?

**답 : 정상 시계열**

16-단10

군집분석의 품질 평가 지표로 응집도와 분리도를 계산하는 지표는?

**답 : 실루엣 계수**

17-단1

기업이 외부 공급업체 또는 제휴업체와 통합된 정보시스템으로 연계하여 시간과 비용을 최적화시키는 데이터베이스 시스템은?

**답 : SCM**

17-단2

데이터 가공 및 상관관계를 통한 패턴을 인식하고 그 의미를 부여하는 것은?

**답 : 정보**

17-단3

소프트웨어와 시스템공학의 역량 성숙도를 측정하기 위한 모델은?

**답 : 능력 성숙도 통합모델 (CMMI)**

17-단4

동일한 사실에도 판단이 달라지는 현상은?

**답 : 프레이밍 효과**

17-단5

유사한 클러스터끼리 합쳐가면서 원하는 클러스터 개수가 될 때까지 진행하는 방식은?

**답 : 병합적 군집 (상향적 군집)**

17-단6

(관찰된 일치율 – 우연에 의한 일치율) / (1 – 우연한 일치율)은?

**답 : 카파(kappa)**

17-단7

민감도와 특이도가 어떤 관계를 갖고 변하는지 이차원 평면상에 표현한 것은?

**답 : ROC**

17-단8

시계열에서 분리해서 분석하는 방법은?

**답 : 분해 시계열**

17-단9

SOM의 경쟁층의 프로토타입 벡터와의 거리를 계산하고 가장 가까운 프로토타입 벡터는?

**답 : BMU(Best-Matching-Neighbors)**

17-단10

목표치가 다범주인 경우 각 범주에 속할 사후 확률을 제공하는 것은?

**답 : 소프트맥스 함수**

18-단1

**답 : 인과관계, 상관관계**

18-단2

**답 : 플랫폼**

18-단3

**답 : 하향식 접근방식**

18-단4

**답 : 모델링 과정**

18-단5

오분류표의 정확도 표기

**답 : (a+b) / (a+b+c+d)**

18-단6

맨하탄 거리 구하기

18-단7

분류가 잘못된 데이터에 큰 가중치를 부여하는 기법

**답 : 부스팅**

18-단8

현시점에서 전시점 자료를 빼는 것

**답 : 차분**

18-단9

회귀모형의 계수 추정 방식

**답 : 최소제곱법**

18-단10

군집의 개수 구하기

19-단1 **데이터사이언스**

19-단2 **유전자 알고리즘**

19-단3 **나선형 모델**

19-단4 **정보전략계획 (ISP)**

19-단5

모든 변수 적용 후 하나씩 변수 제거 방법

**답 : 후진제거법**

19-단6

의사결정나무에서 MSE 등을 고려한 과적합 문제 해결 방법은?

**답 : 가지치기**

19-단7

랜덤모델과 비교하여 모델 성능 평가는?

**답 : 향상도곡선 (lift curve)**

19-단8

A와 B의 지지도 구하기

19-단9

시계열 모형 중 백색 잡음의 선형 가중합 모형

**답 : MA 모형**

19-단10

선형회귀모형에서 결정계수 구하기

**공식) SSR / SST**

20-단1

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

**답 : BI (Business Intelligence)**

20-단2

데이터 사이언티스트들은 주로 데이터 처리나 분석 기술과 관련된 ㅇㅇ만을 요구받고 있는 것처럼 보인다. 하지만 이러한 ㅇㅇ은 훌륭한 데이터 사이언티스트가 갖춰야 하는 능력의 절반에 불과하다. 나머지 절반은 통찰력 있는 분석, 설득력 있는 전달, 협력 등 ㅁㅁ이다.

**답 : 하드스킬, 소프트스킬**

20-단3

동일한 사안이라고 해도 제시되는 방법에 따라 그에 관 한 해석이나 의사결정이 달라지는 왜곡현상은?

**답 : 프레이밍 효과**

20-단4

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

**답 : 사전영향평가**

20-단5

다음 값은 얼마인가?

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

**답 : 0.4**

20-단6

두 군집사이의 거리를 군집에서 하나씩 관측 값을 뽑았을 때 나타날 수 있는 거리의 최소값으로 측정하며, 고립 된 군집을 찾는데 중점을 둔 방법

**답 : 최단연결법**

20-단7

회귀모형에 대한 가정에 대한 설명이다. 빈칸에 알맞은 용어는?

∙ 잔차와 독립변수의 값이 관련되어 있지 않다. - 독립성

∙ 잔차들끼리 상관이 없어야 한다. - 비상관성

∙ ( )이 정규분포를 이뤄야 한다. – 정상성

**답 : 잔차항**

20-단8

다음을 무엇이라고 하는가?

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

**답 : 층화추출**

20-단9

다음을 무엇이라고 하는가?

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

**답 : 배깅**

20-단10

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

**답 : 홀드아웃**

21-단1 **데이터 사이언스**

21-단2 **분류분석, 유형분석**

21-단4 **디자인 씽킹 (Design Thinking)**

21-단6 **다차원 척도법**

21-단7 **와드연결법**

21-단8 **정상성**

21-단9 **정지규칙**

21-단10 **향상도**

22-단1 **SCM**

22-단2 **유전자 알고리즘**

22-단4 **ISP (Information Strategy Planning)**

22-단6 **배깅**

22-단7 **홀드아웃**

22-단8 **ESD (Extream Studentized Deviation)**

22-단9 **포아송분포**

22-단10 **ROC**

26-단1

이것은 비즈니스 측면에서 일반적으로 ‘공동 활용의 목적으로 구축된 유무형의 구조물’을 의미한다. 각종 사용자 데이터나 M2M센서 등에서 수집된 데이터를 가공, 처리, 저장해두고 이 데이터에 접근할 수 있도록 API를 공개한다. 그러면 다양한 서드파티 사업자들이 비즈니스에 필요한 정보를 추출해 활용하게 되고 빅데이터는 그 자체로 이 역할을 수행하게 된다.

**답 : 플랫폼**

26-단2

데이터의 가공 및 상관관계간 이해를 통해 패턴을 인식하고 그 의미를 부여한 데이터로 지식을 추출하기 위한 것이다.

**답 : 정보**

26-단3

데이터 표준화는 데이터표준 용어 설정, 명명규칙 수립, ( ) 구축, 데이터 사전 구축 등의 업무로 구성된다. 데이터 표준 용어는 표준 단어 사전, 표준 도메인사전, 표준 코드 등으로 구성되며 사전 간 상호 검증이 가능하도록 점검 프로세스를 포함해야 한다.

데이터의 데이터, 데이터에 관한 구조화된 데이터로 다른 데이터를 설명해주는 데이터

**답 : 메타데이터**

26-단4

ㅇㅇ란 전사차원의 모든 데이터에 대하여 정책 및 지침, 표준화, 운용조직 및 책임 등의 표준화된 관리체계를 수립하고 운영을 위한 프레임워크 및 저장소를 구축하는 것을 말한다. 특히 마스터 데이터, 메타 데이터, 데이터 사전은 ㅇㅇ의 중요한 관리 대상이다.

**답 : 데이터 거버넌스**

26-단5

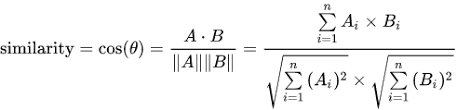
ㅇㅇ은 배깅에 랜덤과정을 추가한 방법이다. 원 자료로부터 부트스트랩 샘플을 추출하고, 각 부트스트랩 샘플에 대해 트리를 형성해 나가는 과정은 배깅과 유사하나, 각 노드마다 모두 예측변수 안에서 최적의 분할을 선택하는 방법 대신 예측변수를 임의로 추출하고 추출된 변수 내에서 최적의 분할을 만들어 나가는 방법을 사용한다.

의사결정나무의 특징인 분산이 크다는 점을 고려하여 배깅과 부스팅보다 더 많은 무작위성을 주어 약한 학습기들을 생성한 후 이를 선형 결합하여 최종 학습기를 만드는 방법이다.

**답 : 랜덤 포레스트**

26-단6

두 객체 간의 거리에 기반하여 군집을 형성해가는 계층적 군집방법에서 사용되는 측도 중 두 개체의 벡터 내적을 기반하여 아래의 수식으로 계산할 수 있는 유사성 측도는 무엇인가?



답 : 코사인 유사도

26-단7

코호넨에 의해 제시되었으며, 비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도의 형태로 형상화하는 클러스터링 방법은 무엇인가?

**답 : SOM**

26-단8

레이더 이미지 분석의 성과를 측정하기 위해 개발된 이 그래프는 두 분류 분석 모형을 비교 분석 결과를 가시화할 수 있다는 점에서 유용한 평가도구이다. x축에는 **FP Ratio(1-x특이도)**를 나타내며 y축에는 **민감도**를 나타내 두 평가 값의 관계로 모형을 평가한다. 모형의 성과를 평가하는 기준은 그래프의 밑 부분 면적이 넓을수록 좋은 모형으로 평가한다.

**답 : ROC Curve**

26-단9

정분류율 공식

답 : (TP+FP)/(TP+TN+FP+FN)

26-단10

시계열 모델 중 자기 자신의 과거 값을 사용하여 설명하는 모형임

백색 잡음의 현재 값과 자리 자신의 과거 값의 선형 가중합으로 이루어진 정상확률 모형

모형에 사용하는 시계열 자료의 시점에 따라 1차, 2차, …, p차 등을 사용하나 정상시계열 모형에서는 주로 1, 2차를 사용함

**답 : AR모형**

27-단1

ㅇㅇ은 어떤 현상에 대하여 현상을 발생시킨 원인과 그 결과 사이의 관계를 말하고, ㅁㅁ은 어떤 두 현상이 관계가 있음을 말하지만 어느 쪽이 원인인지 알 수 없다.

**답 : 인과관계, 상관관계**

27-단2

페이스북은 2006년 F8 행사를 기점으로 자신들의 소셜 그래프 자산을 외부 개발자들에게 공개하고 서드 파티 개발자들이 페이스북 위에서 작동하는 앱을 만들기 시작하면서 ㅇㅇ 역할을 하기 시작했다.

하둡은 대규모 분산 병렬 처리의 업계 표준으로 맵리듀스 시스템과 분산 파일 시스템인 HDFS로 구성된 ㅇㅇ 기술이며, 선형적인 성능과 용량 확장성, 고장 감내성을 가지고 있다. 아마존은 S3와 BC2 환경을 제공함으로써 ㅇㅇ을 위한 클라우드 서비스를 최초로 실현하였다.

**답 : 플랫폼**

27-단3

문제 탐색을 통해서 식별된 비즈니스 문제를 변환하는 단계로써, 문제 탐색 단계가 무엇을 어떤 목적으로 수행해야 하는가에 대한 관점이었다면, ㅇㅇ 단계는 이를 달성하기 위해서 필요한 데이터 및 기법을 도출하기 위한 데이터 분석의 문제로의 변환을 수행하게 된다.

**답 : 문제 정의**

27-단4

분석용 데이터를 이용한 가설 설정을 통해 통계모델을 만들거나 기계학습을 이용한 데이터의 분류, 예측, 군집 등의 기능을 수행하는 모델을 만드는 과정

**답 : 모델링**

27-단6

원 자료로부터 부트스트랩 샘플을 추출하고, 각 부트스트랩 샘플에 대해 트리를 형성해 나가는 과정은 유사하나, 각 노드마다 모든 예측변수 안에서 최적의 분할을 선택하는 대신 예측변수들을 임의로 추출하고, 추출된 변수 내에서 최적의 분할을 만들어 나가는 방법을 사용하는 기법은?

**답 : 랜덤 포레스트**

27-단8

아래의 식으로 계산되는 함수로 신경망 모형에서 표준화 지수 함수로 불리며, 출력값 z가 여러 개로 주어지고, 목표치가 다범주인 경우 각 범주에 속할 사후 확률을 제공하여 출력 노드에 주로 사용되는 함수는?

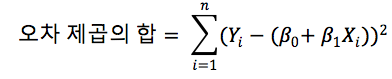
텍스트이(가) 표시된 사진

자동 생성된 설명

**답 : 소프트맥스 함수**

27-단9

회귀모형의 계수를 추정하는 방법으로써 잔차제곱합(SSR)을 최소화하는 계수를 찾는 방법을 무엇이라고 하는가?



**답 : 최소제곱법**

27-단10

로지스틱 회귀분석에서 어떠한 일이 일어날 확률을 일어나지 않을 확률을 나누어 log를 취하고 이를 0~1의 값이 아닌 [ 범위에서 선형함수를 시그모이드 함수로 변환하는 방법은 무엇인가?

**답 : 로짓 변환**

28-단1

OLAP : 다차원의 데이터를 대화식으로 분석하기 위한 소프트웨어

ㅇㅇ : 데이터 기반 의사결정을 지원하기 위한 리포트 중심의 도구

Analytics : 의사결정을 위한 통계적이고 수학적인 분석에 초점을 둔 기법

**답 : BI (Business Intelligence)**

28-단2 데이터 사이언티스트가 갖춰야 할 역량은 빅데이터 처리 및 분석에 필요한 이론적 지식과 기술적 숙련과 관련된 능력인 ㅇㅇ와 데이터 속에 숨겨진 가치를 발견하고 새로운 발전 기회를 만들어 내기 위한 능력인 ㅁㅁ로 나누어진다.

**답 : 하드스킬, 소프트스킬**

28-단3

비즈니스 모델 캔버스는 9가지 블록을 단순화하여 ㅇㅇ, ㅁㅁ, 고객단위로 문제를 발굴하고 이를 관리하는 규제와 감사, ㅎㅎ 영역으로 나눠 분석 기회를 도출한다.

**답 : 업무, 제품, 지원 인프라**

28-단4

소프트웨어와 시스템공학의 역량 성숙도를 측정하기 위한 모델로 소프트웨어 품질보증과 시스템 엔지니어링 분야의 품질보증 기술을 통합하여 개발된 평가모델로 1~5단계로 구성된 성숙도 모델이다.

**답 : 능력 성숙도 통합 모델 (CMMI)**

28-단5

결정계수 구하기

공식) **R 제곱 = SSR / SST, SST = SSR + SSE**

28-단6

시계열 분석에서 일정한 주기를 가지고 반복적으로 같은 패턴을 보이는 변화를 나타내는 요인은?

**답 : 계절 요인**

28-단7

연관규칙 분석에서 품목 A의 거래수가 50, 품목 B의 거래수가 30, 품목 A,B가 동시에 포함된 거래수가 20, 전체 거래수가 100일 때 품목 A, B의 지지도를 계산하시오.

답 : 20/100 = 0.2

\*지지도는 전체 거래 중 A, B를 동시에 포함하는 거래 비율

28-단8

의사결정 나무에서 더 이상 분기가 되지 않고 현재의 마디가 끝마디가 되도록 하는 규칙을 나타내는 용어는 무엇인가?

**답 : 정지규칙**

28-단9

상당히 이질적인 원소들로 구성된 모집단에서 각 계층을 고루 대표할 수 있도록 표본을 추출하는 방법이다. 이질적인 모집단의 원소들로 서로 유사한 것끼리 몇 개의 층을 나눈 후, 각 계층에서 표본을 랜덤하게 추출한다.

**답 : 층화추출**

28-단10

의사결정나무와 같이 선형성, 정규성, 등분산성 등의 가정을 필요로 하지 않는 모형은?

**답 : 비모수 모형**

29-단1

ㅇㅇ은 하둡 분산 파일시스템(HDFS)에 저장된 대용량의 데이터들을 대상으로 SQL을 이용하여 사용자의 질의를 실시간으로 처리하는 기술이다. 이것의 대표적인 예로는 Apache Hive, Apache Tajo, Cloudera의 Impala, Facebook의 Presto, Pivotal HD의 HAWQ, Apache Drill 등이 있다.

**답 : 맵리듀스**

29-단2

ㅇㅇ는 거래정보를 하나의 덩어리로 보고 이를 차례로 연결한 거래장부다. 기존 금융회사의 경우 중앙 집중형 서버에 거래 기록을 보관하는 반면, ㅇㅇ는 거래에 참여하는 모든 사용자에게 거래 내역을 보내 주며 거래 때마다 이를 대조해 데이터 위조를 막는 방식을 사용한다.

**답 : 블록체인**

29-단3

문제의 정의 자체가 어려운 경우 데이터를 기반으로 문제의 재정의 및 해결방안을 탐색하고 이를 지속적으로 개선하는 분석과제 발굴 방식을 무엇이라고 하는가?

**답 : 상향식 접근 방식**

29-단4

기업 및 공공기관에서는 시스템의 중장기 로드맵을 정의하기 위한 ㅇㅇ을 수행한다. ㅇㅇ은 정보기술 또는 정보시스템을 전략적으로 활용하기 위하여 조직 내,외부 환경을 분석하여 기회나 문제점을 도출하고 사용자의 요구사항을 분석하여 시스템 구축 우선순위를 결정하는 등 중장기 마스터플랜을 수립하는 절차이다.

**답 : ISP (Information Strategy Planning)**

29-단5

데이터마이닝을 적용한 후 그 결과의 신빙성을 검증하기 위해 데이터를 분할하는데 구축된 모델의 과잉 또는 과소 맞춤 등에 대한 미세조정 절차를 위해 사용되는 데이터는?

**답 : 검증데이터**

29-단6

P(A)=0.3, P(B)=0.4이다. 두 사건 A와 B가 독립일 경우 P(B|A)는 얼마인가?

답 : 0.4

해설) 사건 A와 B가 독립일 경우,

29-단7

실험 결과가 단지 성공과 실패만 있다고 가정하자. 성공일 경우 확률변수는 1의 값을 가지며, 실패일 경우 확률변수는 0의 값을 가진다. 이때 성공일 확률은 0.3이다. 이 경우 기댓값은 얼마인가?

답 : 0.3

해설) **베르누이 확률분포**,

베르누이 확률분포 기댓값 공식 E(X)=p\*1+(1-p)\*0

29-단9

원 데이터 집합으로부터 크기가 같은 표본을 여러 번 단순 임의 복원 추출하여 각 표본에 대해 분류기를 생성한 후 그 결과를 앙상블하는 방법을 무엇이라 하는가?

**답 : 배깅**

29-단10

어느 마트에서 A와 B를 판매하고 있다. A와 B의 지지도는 0.3이고 신뢰도가 0.6이다. A와 B의 판매 수량이 동일할 때, 향상도를 구하시오.

**답 : 1.2**

계산과정) 신뢰도 = 지지도 / P(A)

0.6 = 0.3 / P(A)

P(A) = P(B) = 0.5

향상도 = 신뢰도 / P(B)

= 0.6 / 0.5

= 1.2

30-단1

데이터의 가공 및 상관관계 간 이해를 통해 패턴을 인식하고 의미를 부여한 데이터

지식을 도출할 때 사용하는 데이터

**답 : 정보**

30-단2

기업내부 데이터베이스 중 기업 전체가 경영자원을 효과적으로 이용하기 위해 통합적으로 관리하고 경영의 효율화를 기하기 위한 수단으로 정보의 통합을 위해 기업의 모든 자원을 최적으로 관리하기 위한 기업 경영 정보시스템

**답 : ERP**

30-단3

데이터 거버넌스란 전사차원의 모든 데이터에 대해 정책 및 지침, 표준화, 운용조직 및 책임 등의 표준화된 관리체계를 수립하고 운영을 위한 프레임워크 및 저장소를 구축하는 것을 말한다. 특히 ㅇㅇ, ㅁㅁ, ㅎㅎ는 데이터 거버넌스의 중요한 관리 대상이다.

**답 : 마스터데이터, 메타데이터, 데이터 사전**

30-단4

ㅇㅇ는 전략적 중요도가 핵심이며, 이는 현재의 관점에서 전략적 가치를 둘 것인지, 미래의 중장기적 관점에 전략적인 가치를 둘 것인지를 고려하고, 분석 과제의 목표가치(KPI)를 함께 고려하여 ㅇㅇ의 여부를 판단할 수 있다.

**답 : 시급성**

30-단6

다수 모델의 예측을 관리하고 조합하는 기술을 메타 학습이라 한다. 여러 분류기들의 예측을 조합함으로써 분류 정확성을 향상시키는 기법

**답 : 앙상블 기법**

30-단8

이산형 확률분포 중 주어진 시간 또는 영역에서 어떤 사건의 발생 횟수를 나타내는 확률 분포는 무엇인가?

**답 : 포아송 분포**

30-단10

데이터의 패턴을 발견하고 데이터 모델의 매개 변수를 자동으로 학습함

자체 알고리즘을 사용하여 시간이 경과함에 따라서 경험을 축적하면서 작업 성능이 향상됨

**답 : 머신러닝**

31-단1

인공지능의 한 분야로, 컴퓨터가 스스로 많은 데이터를 분석해서 패턴과 규칙을 찾아내고, 학습된 패턴과 규칙을 활용하여 분류나 예측을 하는 것

**답 : 머신러닝**

31-단2

조직 내 구성원들이 축적하고 있는 노하우 등 암묵적 지식을 형식지로 표출화 될 수 있도록 지원하는 등 조직의 경쟁력 향상을 위해 지식자원을 체계화하고 원활하게 공유가 될 수 있도록 지원하는 시스템

**답 : 지식관리시스템 (KMS)**

31-단3

기업 및 공공기관에서 시스템의 **중장기 로드맵**을 정의하기 위해 수행하는 것

정보기술 또는 정보시스템을 전략적으로 활용하기 위해 조직 내,외부 환경을 분석하여 기회나 문제점을 도출하고 사용자의 요구사항을 분석하여 시스템 구축 우선순위를 결정하는 등 중장기 마스터플랜을 수립하는 절차

**답 : ISP (정보전략계획)**

31-단4

데이터 분석 도입의 수준을 파악하기 위한 분석 준비도의 6가지 구성요소 중 하나로서 운영시스템 데이터 통합, 빅데이터 분석 환경, 통계분석 환경 등을 진단하는 구성요소는 무엇인가

**답 : IT 인프라**

31-단5

베이즈 정리와 특징에 대한 조건부 독립을 가설로 하는 알고리즘으로 클래스에 대한 사전 정보와 데이터로부터 추출된 정보를 결합하고 **베이즈 정리**를 이용하여 특정 데이터가 특정 클래스에 속하는지를 분류하는 알고리즘은 무엇인가

특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기

**답 : 나이브 베이즈 분류**

31-단6

계층적 군집분석에서 두 군집을 병합하는 방법 중, 군집과 군집, 또는 데이터와의 거리 계산 시, **최단거리**를 계산하여 거리가 가까운 데이터, 군집을 새로운 군집으로 형성하는 방법

**답 : 최단연결법**

31-단7

텍스트 마이닝에서 어근에 차이가 있더라도 관련이 있는 단어들을 동일한 어간으로 맵핑 될 수 있도록 정해진 규칙에 따라 단어에서 어간을 분리하여 공통 어간을 가지는 단어를 묶는 작업

**답 : 어간 추출 (Stemming)**

31-단8

시계열 분석을 위해서는 정상성을 만족해야 한다. 따라서 주어진 자료가 정상성을 만족하는지 판단하는 과정이 필요하다. 자료가 추세를 보이는 경우에는 현 시점의 자료 값에서 전 시점의 자료를 빼는 방법을 통해 비정상시계열을 정상시계열로 바꿔 준다.

**답 : 차분**

31-단10

데이터 집합으로부터 크기가 같은 표본을 여러 번 단순 임의 복원 추출하여 각 표본에 대해 분류기를 생성한 후 그 결과를 앙상블 하는 방법

**답 : 배깅**