2022년 35회 ADsP 기출복원 (인쇄용)

총점 50/50

✓ **R1. 사용자와 데이터베이스 사이에서 사용자 요구에 따라 정보처리 데이터베이스 관리를 하는 소프트웨어를 무엇이라고 하는가?** 1/1

- (●) DBMS ✓
- ( ) Data Dictionary
- ( ) SQL
- ( ) ERD

의견 보내기

DBMS : 사용자와 데이터베이스 사이에서 사용자의 요구에 따라 정보를 처리해주고 데이터베이스를 관리해주는 소프트웨어
Data Dictionary : 자료에 관한 정보를 모아 두는 저장소. 자료 사전이라고도 한다. 자료의 이름, 표현 방식, 자료의 의미와 사용 방식, 그리고 다른 자료와의 관계를 저장한다.
SQL : Structured Query Language, 데이터베이스를 구축하고 활용하기 위해 사용하는 언어♂RDBMS의 데이터를 관리하기 위해 설계된 특수 목적의 프로그래밍 언어
ERD : Entity Relationship Diagram, 실체와 이들의 관계를 도형으로 표현한 것, 실체의 상관관계 다이어그램은 사용자와 어플리케이션 개발자 간의 자료를 공통적으로 이해할 수 있도록 하는 유용한 매체가 됨

✓ **2. 데이터 사이언티스트의 필요 역량으로 적절하기 않은 것은?** 1/1

- ○ 하드 스킬과 소프트 스킬 능력
- ○ 데이터 처리 기술
- ◉ 네트워크 최적화 능력 ✓
- ○ 고객들에 대한 공감 능력

의견 보내기

1-23. 데이터 사이언티스트의 역량
- 데이터 관리, 분석 모델링, 비즈니스 분석, 소프트 스킬
- 데이터 해커, 애널리스트, 커뮤니케이션, 신뢰받는 어드바이저 등의 조합
- 하드 스킬과 소프트 스킬 능력을 동시에 갖추어야 함
-데이터 처리 기술 이외에 사고방식, 비즈니스 이슈에 대한 감각, 고객들에 대한 공감 능력이 필요함

✓ **3. 사생활 침해 방지 기술에 해당하는 것으로 개인 식별 정보를 삭제하거나 알아볼 수 없는 형태로 변환하는 포괄적 기술로 올바른 것은?** 1/1

- ◉ 익명화 ✓
- ○ 데이터 마스킹
- ○ 가명
- ○ 데이터값 삭제

의견 보내기

1-16, 1-17. 빅데이터 위기 요인과 통제방안
- 익명화 : 발생하는 사생활 침해를 막기 위해 데이터에 포함된 개인 식별 정보를 삭제하거나 알아볼 수 없는 형태로 변환하는 포괄적 기술(다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없는 정보)
- 데이터 마스킹 : 다양한 유형의 데이터 관리 시스템에 저장된 정보를 보호하는 데 사용되는 프로세스 (카드 뒤 4자리 숨기기, 주민번호 뒤 6자리 숨기기)
- 가명 : 개인식별 정보를 삭제, 알아볼 수 없는 형태로 변환 (개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가정보 없이는 특정 개인을 알아볼 수 없도록 처리하는 것)
- 데이터값 삭제 : 데이터 셋의 값 중 필요 없는 값 또는 개인 식별에 중요한 값 삭제

✓ **4. 데이터 분석 알고리즘으로 부당한 피해를 보는 사람을 방지하기 위해서 생겨난 직업으로 데이터 분석 알고리즘으로 인해 피해를 입은 사람을 구제하는 전문가를 무엇이라 하는가?** 1/1

- ○ 데이터 엔지니어
- ◉ 알고리즈미스트 ✓
- ○ 데이터 사이언티스트
- ○ 데이터 분석가

의견 보내기

1-16. 빅데이터 위기 요인과 통제방안, 1-23. 데이터 분석가, 사이언티스트, 엔지니어의 역량
데이터 분석가, 데이터 엔지니어, 데이터 사이언티스트는 데이터 분석 알고리즘을 생성하는 것과 관련된 전문가들이다.

✓ **5. 빅데이터의 영향에 대해 올바르지 않은 것은 무엇인가?** 1/1

- ○ 산업 전체의 생산성이 향상되었다
- ○ 맞춤형 서비스를 저렴한 비용으로 이용할 수 있게 되었다
- ○ 사회 변화를 추정, 각종 재해 관련 정보를 추출할 수 있게 되었다
- ◉ 사물인터넷(IoT)의 발달로 인해 사람이 최대로 개입하게 되었다 ✓

의견 보내기

1-14. 빅데이터의 영향
기업 : 비즈니스 모델 혁신, 신사업 발굴, 경쟁력 확보, 산업 전체의 생산성 향상, GDP 상승 효과
정부 : 환경탐색, 상황분석, 미래 대응, 사회 변화를 추정, 각종 재해 관련 정보를 추출할 수 있음
개인 : 맞춤형 서비스를 저렴한 비용으로 이용
사물 인터넷의 발달로 인해 사람의 개입이 최소화 됨

✓ **6. 빅데이터의 기술 활용에 관련된 설명으로 거리가 먼 것은?** 1/1

- ○ 기업은 원가절감, 제품차별화, 기업활동의 투명성 제공 등에 활용한다
- ○ 미래 사회 도래에 대비한 법제도 및 거버넌스 시스템 정비 방향, 미래 성장 전략 등에 대한 정보 제공한다
- ◉ 정부의 이익을 위해 개인의 정보를 활용한다 ✓
- ○ 적시에 필요한 정보를 얻어 다양한 형태로 기회 비용을 절약할 수 있다

의견 보내기

1-14. 빅데이터의 영향
- 빅데이터 활용에 있어 '정부의 이익을 위해 개인의 정보를 활용하지 않는다'.

✓ **7. 다음 중 빅데이터 위기요인과 통제방안에 대한 내용과 관련이 없는 것은?** 1/1

- ○ 사생활 침해
- ○ 데이터 오용
- ◉ 데이터 변화 관리 ✓
- ○ 책임 원칙의 훼손

의견 보내기

1-16. 빅데이터 위기 요인과 통제방안
- 빅데이터 위기요인의 종류에는 사생활 침해, 책임 원칙의 훼손, 데이터의 오용이 있다

✓ **8. 다음 중 데이터베이스와의 통신을 위해 고안된 언어는 무엇인가?** 1/1

- Python
- Java
- R
- SQL ✓

의견 보내기

1-15-a2. SQL(Structured Query Language)
데이터베이스를 구축하고 활용하기 위해 사용하는 언어
데이터베이스와 통신을 위해 고안된 언어
RDBMS의 데이터를 관리하기 위해 설계된 특수 목적의 프로그래밍 언어

✓ **9. 문자, 기호, 음성, 화상, 영상 등 상호 관련된 다수의 콘텐츠를 정보처리 및 정보통신 기기에 의해 체계적으로 수집, 축적하여 다양한 용도와 방법으로 이용할 수 있도록 정리한 정보의 집합체는?** 1/1

데이터베이스 ✓

의견 보내기

1-05. 데이터베이스
- 초기에는 텍스트, 숫자 형태의 데이터를 있는 그대로 저장하는 장치 (정형 데이터)
- 정보기술 발달 후 저장하는 데이터가 이미지, 동영상을 포함한 멀티미디어로 확대됨 (비정형 데이터)
- 이후, 단순한 데이터 저장에서 정보를 저장하는 지식베이스로 진화
- 단순한 저장소의 개념을 넘어 첨단 정보기술을 바탕으로 원하는 데이터를 저장 검색할 수 있는 복합체

✓ **10. 다음 설명에 해당하는 빅데이터 활용 테크닉은 무엇인가?**
“최대의 시청률을 얻으려면 어떤 프로그램을 어떤 시간대에 방송해야 하는가?와 같은
최적화의 메커니즘을 찾아가는 방법이다 " 1/1

유전 알고리즘 ✓

✓ **1. 데이터 분석을 위한 조직 구조 중 분석 조직 인력들을 현업부서로 직접 배치하여 신속한 업무 수행이 가능한 구조는 무엇인가 ?** 1/1

- (●) 분산 조직 구조 ✓
- ( ) 집중형 조직 구조
- ( ) 기능 중심 조직 구조
- ( ) 혼합형 조직 구조

의견 보내기

2-29. 데이터 분석을 위한 조직 구조
- 집중형 조직 구조 : 조직내 별도 독립적인 분석 전담 조직 구성
- 기능 중심 조직 구조 : 일반적인 분석 수행구조 별도 분석 조직을 구성하지 않고 각 해당 업무부서에서 직접 분석
- 분산 조직 구조 : 분석 조직의 인력들이 현업부서에 배치되어 업무를 수행함

✓ **2. 데이터 거버넌스 체계 단계 중 메타데이터와 데이터 사전(Data Dictionary)의 관리 원칙 수립과 관련된 단계는 무엇인가?** 1/1

- ( ) 데이터 표준화
- (●) 데이터 관리체계 ✓
- ( ) 데이터 저장소관리
- ( ) 표준화 활동

의견 보내기

2-28. 데이터 거버넌스 체계 수립
- 데이터 표준화 단계 : 데이터 표준용어 설정, 명명규칙 수립, 메타 데이터 구축, 데이터 사전 구축
- 데이터 관리체계 : 메타데이터와 데이터 사전(Data Dictionary)의 관리 원칙 수립
- 데이터 저장소관리 : 메타데이터 및 표준 데이터를 관리하기 위한 전사 차원의 저장소를 구성
- 표준화 활동 : 데이터 거버넌스 체계 구축 후, 표준 준수 여부를 주기적으로 점검, 모니터링

✓ **3. 분석 마스터 플랜의 과제 우선순위 결정과 관련된 내용으로 적절하지 않은 것은?** 1/1

- 난이도 판단기준은 데이터 획득/저장/가공 비용 및 분석 적용 비용, 분석 수준 등이 있다
- 시급성의 판단기준은 전략적 중요도가 핵심이다
- (●) Value(가치)는 투자비용 요소이다 ✓
- ROI 관점에서의 분석 과제 우선순위 평가 기준은 시급성과 난이도가 있다

의견 보내기

2-20. 분석 마스터 플랜의 분석 과제 우선순위 선정 기법
- 분석 과제 우선순위 평가기준에 시급성, 난이도가 있음
- 시급성 : 판단 기준에 전략적 중요도, 목표가치, Value(비즈니스 효과, Return)와 관련 있음
- 난이도 : 판단 기준에 데이터 획득/저장/가공 비용 및 분석 적용 비용, 분석 수준 등이 있으며, Volume, Variety, Velocity의 투자비용 요소(Investment)와 관련 있음

✓ **4. 다음 중 데이터 거버넌스의 구성요소가 아닌 것은?** 1/1

- 원칙(Principle)
- 조직(Organization)
- (●) 분석 방법(Method) ✓
- 프로세스(Process)

의견 보내기

2-23. 데이터 거버넌스 구성 요소 : 원칙, 조직, 프로세스
- 원칙 : 데이터를 유지 관리하기 위한 지침과 가이드 및 보안, 품질 기준, 변경 관리
- 조직 : 데이터를 관리할 조직의 역할과 책임 및 데이터 관리자, 데이터 아키텍트
- 프로세스 : 데이터 관리를 위한 활동과 체계 및 작업 절차, 모니터링 활동

✓ **5. 분석 과제 도출 방법 중 상향식 접근 방식의 절차로 알맞은 것은?** 1/1

- (●) 프로세스 분류 -> 프로세스 흐름분석 -> 분석요건 식별 -> 분석요건 정의 ✓
- ( ) 프로세스 흐름 분석 -> 분석요건 식별 -> 분석 요건 정의 -> 프로세스 분류
- ( ) 프로세스 흐름 분석 -> 분석요건 식별 -> 프로세스 분류 -> 분석 요건 정의
- ( ) 프로세스 분류 -> 분석요건 식별 -> 분석 요건 정의 -> 프로세스 흐름 분석

의견 보내기

2-13. 분석 과제 도출 방법 – 상향식 접근 방식 - 문제의 정의 자체가 어려운 경우 사용하는 방식
- 상향식 접근 방식의 절차: 프로세스 분류 -> 프로세스 흐름 분석 -> 분석요건 식별 -> 분석 요건 정의

✓ **6. 다음 중 분석과제의 우선순위 선정 시 난이도와 시급성을 모두 고려하였을 때 우선적으로 추진해야하는 분석 과제는 무엇인가?** 1/1

- ( ) 난이도 – 어려움, 시급성 - 미래
- ( ) 난이도 – 쉬움, 시급성 - 미래
- ( ) 난이도 – 어려움, 시급성 - 현재
- (●) 난이도 – 쉬움, 시급성 - 현재 ✓

의견 보내기

2-20. 분석 과제 우선순위 선정 기법
- 3사 분면: 난이도 쉬움, 시급성 현재에 해당하는 것으로 일반적으로 가장 먼저 하는 것
- 우선순위를 '시급성'에 둔다면 Ⅲ - Ⅳ - Ⅱ순서 진행
- 우선순위를 '난이도'에 둔다면 Ⅲ - Ⅰ - Ⅱ순서 진행
- 시급성이 높고(현재) 난이도가 높은(Difficult) 영역(1사분면)은 경영진 또는 실무 담당자의 의사결정에 따라 적용 우선순위를 조정할 수 있음

✓ **7. 다음 중 기업의 분석 도입의 수준을 파악하기 위한 분석 준비도와 관계가 적은 항목은 무엇인가?** 1/1

- ○ 분석 인력 및 조직
- ○ 분석 기법
- ◉ 목표와 정책 ✓
- ○ 분석 데이터

의견 보내기

2-24. 데이터 분석 수준 진단
분석 준비도 :기업의 데이터 분석 도입의 수준을 파악하기 위한 진단방법
- 분석 업무 파악, 인력 및 조직, 분석 기법, 분석 데이터, 분석 문화, IT 인프라(=분석 인프라)의 6가지 영역을 대상으로 현 수준을 파악함

분석 성숙도 : 시스템 개발 업무능력과 조직의 성숙도 파악을 위해 CMMI 모델을 기반으로 분석 성숙도를 평가함
- 비즈니스 부문, 조직/ 역량 부문, IT 부문을 대상으로 성숙도 수준에 따라 도입, 활용, 확산, 최적화 단계로 구분해 살펴 볼 수 있음

✓ **8. 다음 중 빅데이터 분석 방법론의 분석 기획 단계에서 프로젝트 위험 계획 수립 시 위험에 대한 대응 방법의 종류에 포함되지 않는 것은?** 1/1

- ○ 회피(avoid)
- ○ 전이(transfer)
- ○ 완화(mitigate)
- ◉ 관리(management) ✓

의견 보내기

2-10-1. 프로젝트 위험 계획 수립
위험에 대한 대응 방법 : 회피(Avoid), 전이(Transfer), 완화(Mitigate), 수용(Accept)
회피 : 계획 변경 등 원인 제거 (기간 연장, 범위 축소)
전이 : 보험, 사후 보증
완화 : 용인가능 임계치까지 절감노력
수용 : 적극적 수용(긴급 대책), 소극적 수용(아무 조치 안함), Fallback plan(위험의 영향이 클 경우

✓ **9. 문제가 주어지고 해답을 찾기 위해 각 과정이 체계적이고 단계화 되어 수행하는 분석 과제 도출 방식은 무엇인가?** 1/1

하향식 접근 방법 ✓

의견 보내기

2-11. 분석 과제 도출 방법
하향식 접근 방법: 문제가 확실할 때 사용하는 방법으로 문제가 주어지고 해법을 찾기 위해 사용함
상향식 접근 방법: 문제의 정의 자체가 어려운 경우 사용함

✓ **10. 다음이 설명하는 분석 조직 구조는 무엇인가?** 1/1
-조직내 별도 독립적인 분석 전담 조직 구성, 분석 전담 조직에서 회사의 모든 분석 업무를 담당
-전사분석 과제의 전략적 중요도에 따라 우선 순위를 정해 추진
-일부 협업 부서와 분석 업무가 중복 또는 이원화될 가능성이 있음

집중형 조직 구조 ✓

✓ **1. 다음 중 회귀분석의 결정계수에 관한 설명으로 적절하지 않은 것은 무엇인가?** 1/1

- ○ 결정계수는 회귀제곱합(SSR) / 총제곱합(SST) 로 구할 수 있다
- ◉ 종속변수와 독립변수 사이의 표본 상관계수와 값이 같다 ✓
- ○ 결정계수가 커질수록 회귀방정식의 설명력이 높다고 할 수 있다
- ○ 일반적으로 결정계수는 0 ~ 1의 값을 갖는다

의견 보내기

3-65. 회귀모형 해석 – 결정계수
- 회귀식의 적합도를 재는 척도
- 결정계수(R2) = 회귀제곱합(SSR) / 총제곱합(SST), 1-(SSE/SST)
- 결정계수는 0~1 사이의 범위를 갖음
- 전체 분산 중 모델에 의해 설명되는 분산의 양
- 결정계수가 커질수록 회귀방정식의 설명력이 높아짐

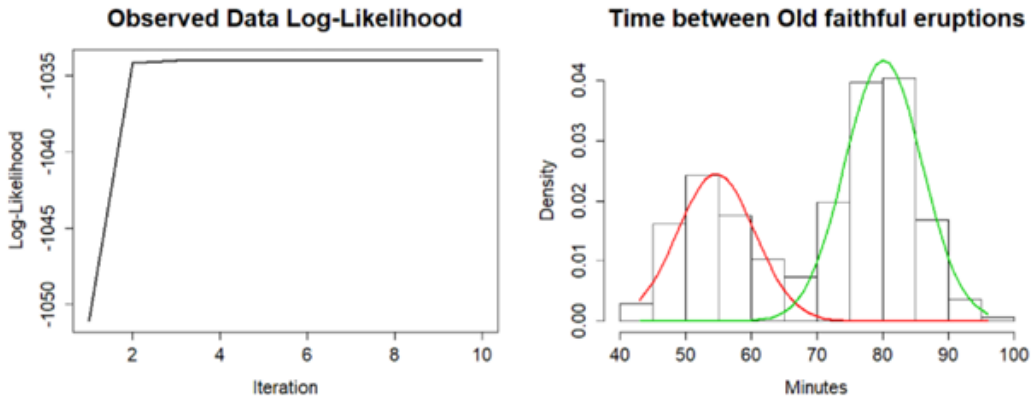
✓ **2. 다음 중 목표변수가 연속형인 회귀나무의 분류 기준값을 선택하는 기준으로 구성된 것은 무엇인가?** 1/1

- ○ 지니 지수(gini), 엔트로피 지수(entropy)
- ○ 카이 제곱 통계량, 분산 감소량
- ◉ F 통계량, 분산 감소량 ✓
- ○ 엔트로피 지수(entropy), 카이 제곱 통계량

의견 보내기

3-81. 의사결정나무를 위한 알고리즘
이산형 목표변수(분류나무) : 지니지수, 엔트로피 지수, 카이제곱 통계량
연속형 목표변수(회귀나무) : 분산 감소량, ANOVA F-통계량

✓ **3. EM알고리즘을 사용한 혼합분포 모형의 결과 해석에 대한 설명으로 적절한 것은 무엇인가?** 1/1



- ◉ 반복횟수 2회 만에 로그가능도 함수가 최대가 됨을 알 수 있다 ✓
- ○ 로그 가능도 함수의 최대값은 -1050이다
- ○ 결과적으로 3개의 정규분포가 혼합된 것을 알 수 있다
- ○ 모수 추정을 위해 8회 이상의 반복이 필요함을 알 수 있다

의견 보내기

로그 가능도 함수의 최대값은 -1035 이다
2개의 정규분포가 혼합된 것을 알 수 있음
모수 추정을 위해 2회의 반복이 필요함을 알 수 있다

✓ **4. Lasso 회귀 모형의 정의로 옳지 않은 것은?** 1/1

- ○ 모형에 포함된 회귀계수의 절댓값이 클수록 penalty를 부여하는 방식이다.
- ○ 람다값(lambda)으로 penalty의 정도를 조정한다.
- ○ 자동적으로 변수 선택을 하는 효과가 있다.
- ◉ L2 norm을 사용하여 penalty를 부여한다. ✓

의견 보내기

3-70. Lasso(라쏘) 회귀 모형
- 변수 선택이 가능하며, 변수간 상관관계가 높으면 성능이 떨어짐
- L1 norm을 패널티로 가진 선형 회귀 방법, 회귀계수의 절댓값이 클수록 패널티 부여
- MSE가 최소가 되게 하는 w, b를 찾는 동시에 w의 절대값들의 합이 최소가 되게 해야함
- W의 모든 원소가 0이 되거나 0에 가깝게 되게 해야 함 => 불필요 특성 제거
- 어떤 특성은 모델을 만들 때 사용되지 않게 됨

✓ **5. SOM에 대한 설명으로 옳지 않은 것은?** 1/1

- ○ SOM은 비지도 학습이다
- ○ SOM은 차원축소와 군집화를 동시에 수행하는 기법이다
- ◉ 입력층과 출력층이 부분연결 되어 있다 ✓
- ○ 출력 뉴런들은 승자 뉴런이 되기 위해 경쟁하고 오직 승자만이 학습한다.

의견 보내기

3-97. SOM(Self Organizing Maps, 자기조직화지도)
- 인공신경망의 한 종류로, 차원축소와 군집화를 동시에 수행하는 기법
- 비지도 학습(Unsupervised Learning)의 한 가지 방법
- 고차원으로 표현된 데이터를 저차원으로 변환해서 보는데 유용함
- 입력층과 2차원의 격자 형태의 경쟁층(=출력층) 으로 이루어져 있음(2개의 층으로 구성)
- 입력층과 출력층은 완전연결 되어 있다

✓ **6. 다음 오분류표를 사용하여 특이도(Specificity)를 구한 결과는 무엇인가?** 1/1

오분류표		예측값		합계
		TRUE	FALSE	
실제값	TRUE	300	300	600
	FALSE	450	150	600
합계		750	450	1200

- ○ 0.375
- ○ 0.75
- ◉ 0.25 ✓
- ○ 0.5

의견 보내기

3-91. 오분류표를 활용한 평가지표 – 특이도(Specificity)
실제로 N(=FALSE) 인 것들 중 예측이 N으로 된 경우의 비율
TN / (TN + FP) = 150/ (150 + 450) = 150/600 = 0.25

✓ **7. 다음 중 입력신호를 받아 출력신호로 연결하기 위한 활성화 함수로 로지스틱 회귀 모델에서도 사용하는 함수는 무엇인가?** 1/1

- ◉ sigmoid ✓
- ○ ReLU
- ○ tanh
- ○ log

의견 보내기

3-81. 로지스틱 회귀분석 – sigmoid 함수
- Logistic 함수라 불리기도 하며, log_odds 값을 연속형 0~1 사이의 값으로 바꾸는 함수
- 비선형 값을 얻기 위해 사용
- 인공신경망에서는 sigmoid 함수가 활성화 함수로 사용됨

3-87. 신경망 활성화 함수(activation function)
- 입력 신호를 받아 출력 신호로 연결하는 함수로 sigmoid, ReLU, tanh 등 다양한 함수가 있음
- [참고] ReLU : 음수를 0으로 변환하는 함수, tanh : -1~ 1 범위의 연속값으로 변환하는 함수

✓ **8. 로지스틱 회귀에 대한 특징으로 적절한 것은?** 1/1

- ○ 모형 검정에는 F 검정이 사용된다
- ◉ 종속변수(=반응변수)가 범주형인 경우 적용되는 회귀분석 모형이다 ✓
- ○ softmax 함수를 사용하여 종속변수를 전체 실수 범위로 확장하여 분석한다
- ○ 모형 탐색 방법에는 최소자승법(최소제곱법)이 있다

의견 보내기

3-81. 로지스틱 회귀분석
- 종속변수가 범주형인 경우 적용되는 회귀분석 모형
- 모형 탐색 방법으로 최대우도법(MLE), 가중최소자승법이 있다
- 모형 검정에는 카이제곱검정을 사용한다
- Sigmoid 함수를 사용하여 종속변수를 전체 실수 범위로 확장하여 분석한다

✓ **9. 선형회귀 모델의 통계적 유의성 검증을 위해 사용하는 것은?** 1/1

- ○ 회귀계수의 t 통계량
- ○ 결정 계수
- ○ 잔차 통계량
- ◉ F 통계량 ✓

의견 보내기

3-65. 회귀모형 해석
- 모형이 통계적으로 유의미한가? F 통계량, 유의확률(p-value)로 확인
- 회귀계수들이 유의미한가? 회귀계수의 t값, 유의확률(p-value)로 확인
- 모형이 얼마나 설명력을 갖는가? 결정계수(R2) 확인
- 모형이 데이터를 잘 적합하고 있는가? 잔차 통계량 확인, 회귀진단 진행(선형성~ 정상성)

✓ **10. 의사결정 나무에 대한 설명 중 적절하지 않은 것은?** 1/1

- (●) 비지도 학습으로 상향식 접근 방법을 이용한다. ✓
- ( ) 구조가 단순하여 해석이 용이하다.
- ( ) 목표변수가 이산형인 경우 분류나무, 목표변수가 연속형인 경우 회귀나무가 있다.
- ( ) 부모마디보다 자식마디의 순수도가 증가하도록 분류나무를 형성해 나간다

의견 보내기

3-82. 의사결정 나무
- 지도학습, 비모수적 모형
- 구조가 단순하여 해석이 용이
- 목표변수가 이산형인 분류나무, 연속형인 회귀나무가 있음
- 부모마디보다 자식마디의 순수도가 증가하도록 분류나무를 형성해 나감

✓ **11. 다음 중 군집의 수를 미리 지정하지 않으며 탐색적 기법에 적합한 군집 방법은 무엇인가?** 1/1

- (●) 계층적 군집 ✓
- ( ) 비계층적 군집
- ( ) K-means 군집
- ( ) 혼합분포 군집

의견 보내기

3-94. 계층적 군집
- 가장 유사한 개체를 묶어 나가는 과정을 반복하여 원하는 개수의 군집을 형성하는 방법
- 사전에 군집 수 k를 설정할 필요가 없는 탐색적 모형
- 비계층적 군집 중 K-means, 혼합분포 군집은 K개의 군집을 지정해 주어야 하며, DBSCAN은 군집의 수를 미리 지정하지 않아도 됨

✓ **12. 확률변수 x의 기댓값은 무엇인가?** 1/1

X	1	2	3
f(X)	0.5	0.3	0.2

- ○ 0.5
- ◉ 1.7 ✓
- ○ 6
- ○ 2

의견 보내기

3-53. 기댓값
이산형 확률변수 x의 기댓값 : $\sum x \cdot f(x)$
$= 1*1/2 + 2*3/10 + 3*1/5 = (5+6+6)/10 = 17/10 = 1.7$

✓ **13. 다음 두 좌표(A, B) 간의 맨해튼 거리(Manhattan distance)를 구하시오.** 1/1

	A	B
키	175	180
몸무게	70	65

- ◉ 10 ✓
- ○ 50
- ○ $\sqrt{10}$
- ○ $\sqrt{50}$

의견 보내기

3-94. 계층적 군집의 거리
맨해튼 거리, 유클리드 거리
=(5+5)=10, =$\sqrt{(25+25)}=\sqrt{50}$
식은 교재를 참조하세요!

✓ **14. 혼합분포 군집의 특징으로 적절하 않은 것은 무엇인가?** 1/1

- ○ 군집을 몇 개의 모수로 표현할 수 있으며, 확률분포를 도입하여 군집 수행한다
- ○ EM 알고리즘을 이용한 모수 추정에서 데이터가 커지면 수렴에 시간이 더 많이 걸릴 수 있다.
- ◉ 군집의 크기가 작을 수록 추정이 쉽고, 정밀한 추정이 가능하다 ✓
- ○ 복잡한 형태를 가진 분포의 경우 여러 분포를 확률적으로 선형 결합한 혼합분포로 설명할 수 있다

의견 보내기

3-95. 비계층적 군집 (혼합분포)
- 군집의 크기가 작으면 추정의 정도가 떨어진다
- 복잡한 형태를 가진 분포의 경우 여러 분포를 확률적으로 선형 결합한 혼합분포로 설명할 수 있다
- 모수와 가중치 추정에 EM 알고리즘이 사용됨(Expectation Maximization)

✓ **15. 다음 시계열 분석에 대한 설명 중 옳지 않은 것은 무엇인가?** 1/1

- ◉ 데이터가 추세를 가지면 변환(자연로그)를 사용하여 정상시계열로 만든다 ✓
- ○ 정상 시계열인 경우 평균값 주변에서의 변동은 대체로 일정한 폭을 갖는다
- ○ 시계열 데이터는 대부분 비정상 시계열이기 때문에 정상 시계열로 만든 후 분석을 수행한다
- ○ 시계열 그래프를 보면서 이상치와 정상성 여부를 확인할 수 있다

의견 보내기

3-75. 시계열 자료(time series), 정상 시계열 전환
- 비정상시계열 자료는 정상성을 만족하도록 데이터를 정상시계열로 만든 후 시계열 분석을 수행한다
- 데이터가 추세를 가지면(평균이 일정하지 않으면) 차분을 사용하여 정상시계열로 만든다
- 데이터의 분산이 일정하지 않은 경우 자연로그(변환)을 사용하여 정상시계열로 만든다
- 시계열 그래프를 보면서 이상치와 정상성 여부를 확인할 수 있다 (P445 그래프 확인)

✓ **16. 다음 중 군집분석에 대한 설명으로 옳지 않은 것은 무엇인가?** 1/1

- ○ 유사성을 이용하여 몇 개의 집단으로 그룹화하는 분석이다
- ○ 집단별 특성이 유사할 경우 안정성이 높다
- ○ 군집 분석은 이상치 자료에 민감한 특성이 있다
- ◉ 안정성 검토 방법으로 지도학습과 동일한 교차타당성(Cross Validation)을 사용한다 ✓

의견 보내기

3-96. 군집분석
- 군집분석에 있어 군집 타당성 검증을 위해 논리성과 안정성 모두가 중요한 부분이다.
- 안정성은 일부 입력값이 변경되었을 때 군집의 변화가 유의하게 변하는지에 대한 개념이다
- 집단별 특성이 유사할 경우 안정성이 높다
- 군집화 평가에는 실루엣 계수와 Dunn Index 가 사용된다
- 유사성을 이용하여 몇 개의 집단으로 그룹화하는 분석이다
- 이상치 자료에 민감한 특성이 있다

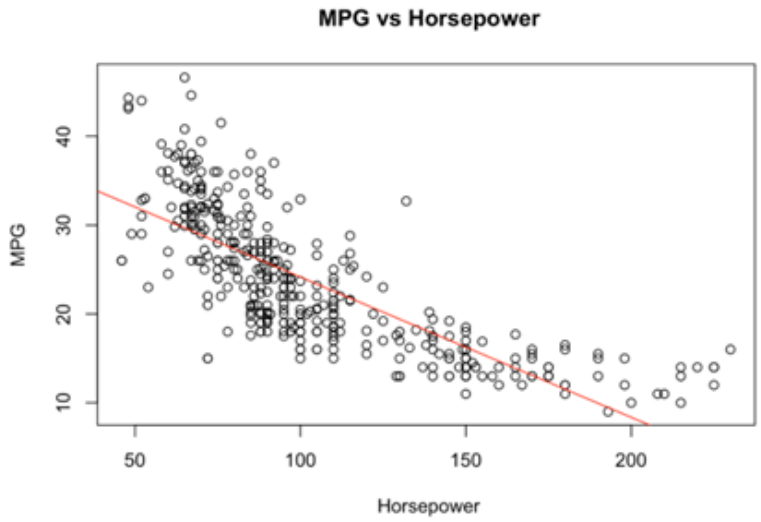
✓ **17. 다음 연관규칙 관련 식 중 A --> B 일 때의 지지도(Support)에 대한 올바른 식은 무엇인가?** 1/1

- ○ A와 B가 동시에 포함된 거래 수 / B가 포함된 거래 수
- ○ A와 B가 동시에 포함된 거래 수 / A가 포함된 거래 수
- ◉ A와 B가 동시에 포함된 거래 수 / 전체 거래 수 ✓
- ○ 품목B를 구매한 고객 대비 품목 A를 구매한 후 품목 B를 구매하는 고객에 대한 확률

의견 보내기

3-99. 연관규칙 측정지표 (A --> B 일 때)
- 신뢰도 : 상품 A를 구매했을 때 상품 B를 구매할 확률이 어느 정도 되는지를 확인
- 신뢰도 = P(B|A) = P(A∩B) / P(A) : A와 B가 동시에 포함된 거래 수 / A가 포함된 거래 수
- 지지도 : 전체 거래항목 중 상품 A와 상품 B를 동시에 포함하여 거래하는 비율
- 지지도 = P(A∩B) : A와 B가 동시에 포함된 거래 수 / 전체 거래 수
- 향상도 : A가 주어지지 않았을 때 B의 확률 대비 A가 주어졌을 때 B의 확률 증가 비율
- 품목B를 구매한 고객 대비 품목 A를 구매한 후 품목 B를 구매하는 고객에 대한 확률
*- 향상도 = P(B|A)/P(B) = P(A∩B) / (P(A)*P(B))*

✓ **18. 아래 산점도는 차량 392대의 연비(mpg)와 마력(horsepower)에 관한 그래프이다. 이와 관한 설명으로 가장 적절하지 않은 것은 무엇인가?** 1/1



- 연비와 마력은 음의 상관관계이다
- 연비-마력의 상관관계는 피어슨 상관계수로 분석이 가능하지 않다 ✓
- 연비와 마력 간의 영향력으로 단순 선형회귀모형 추정이 가능하다
- 마력이 증가할 때 연비가 감소하는 경향이 있다

의견 보내기

3-72. 상관 분석
- 양의 상관은 좌하단에서 우상단으로 기울기를 갖고, 음의 상관은 좌상단에서 우하단의 기울기를 갖음
- 연비, 마력은 비율척도이므로 피어슨 상관계수로 분석이 가능함
- 선형성을 갖고 있고, 종속변수가 연속형이고, 1개의 독립 변수이므로 단순 선형회귀모형 추정 가능
- 음의 상관은 한 가지 값이 증가할 때 다른 값은 감소하는 경향을 보임

✓ **19. 다음 이산형 확률분포의 확률변수 x에 대한 설명 중 적절한 것은 무엇인가?** 1/1

x	1	2	3
f(x)	1/6	1/2	1/3

- ○ 확률변수 x의 확률의 합은 1보다 작거나 클 수 있다
- ○ 확률변수 x가 0이거나 4일 확률은 0이 아니다
- ◉ 확률변수 x에 대한 기댓값은 $\mathbf{13/6}$ 이다 ✓
- ○ 확률변수 x가 1이거나 2일 확률은 5/6 이다

의견 보내기

3-53. 기댓값
이산형 확률변수 x의 기댓값 : $\sum x \cdot f(x)$
$= 1*1/6+2*3/6+3*2/6=(1+6+6)/6=13/6$

3-51. 확률분포
- 확률변수 x의 확률의 합은 1이다
- 확률변수 x가 0이거나 4일 확률은 0이다
- 확률변수 x가 1이거나 2일 확률은 2/3 이다

✓ **20. 다음 회귀 분석 결과에 대한 해석 중 옳지 않은 설명은 무엇인가?** 1/1

```
Call:
glm(formula = default ~ ., family = "binomial", data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

- ○ 로지스틱 회귀 모형을 사용한 분석 결과이다
- ◉ income은 default를 설명하는데 통계적으로 유의미한 변수이다. ✓
- ○ studentYes의 값이 Yes 일 때, 채무불이행(default) 될 확률이 낮다.
- ○ balance는 default를 설명하는데 통계적으로 유의미한 변수이다

의견 보내기

balance는 default를 설명하는데 통계적으로 유의미한 변수이다

✓ **21. 다음 중 데이터 마이닝 프로세스 순서로 올바르게 나열한 것은 무엇인가?** 1/1

(가) 목적 정의

(나) 데이터 준비

(다) 데이터 가공

(라) 데이터 마이닝 기법의 적용

(마) 검증

- ( ) (가)-(마)-(다)-(나)-(라)
- (•) (가)-(나)-(다)-(라)-(마) ✓
- ( ) (가)-(나)-(라)-(마)-(다)
- ( ) (나)-(가)-(다)-(라)-(마)

의견 보내기

3-79. 데이터 마이닝 5단계
목적 정의-> 데이터 준비-> 데이터 가공-> 데이터 마이닝 기법의 적용-> 검증
- 목적 정의: 데이터 마이닝 도입 목적을 명확하게 함
- 데이터 준비: 데이터 정제를 통해 데이터의 품질 확보까지 포함
- 데이터 가공: 목적 변수를 정의하고, 필요한 데이터를 데이터 마이닝 SW에 적용할 수 있게 가공 및 준비하는 단계
- 데이터 마이닝 기법 적용 : 모델을 목적에 맞게 선택하고 소프트웨어를 사용하는 데 필요한 값 지정
- 검증 : 결과에 대한 검증 시행

✓ **22. 다음 중 과대적합 방지를 위한 방법이 아닌 것은 무엇인가?** 1/1

- ○ 배깅(bagging)
- ○ 홀드 아웃(Hold Out)
- ◉ 의사결정 나무 ✓
- ○ Lasso, Ridge 모델

의견 보내기

3-68. 과대적합(Overfitting), 3-83. 앙상블(Ensemble) 모형
- 학습 데이터에 너무 잘 맞게 학습되어 학습 데이터에 대한 성능은 매우 높지만 평가 데이터에 대한 성능은 낮음
- 규제 모델, 앙상블 등의 방법으로 과대적합을 해결하거나 피할 수 있음
- 과대적합인 경우 평가 데이터(test data)의 작은 변화에도 민감하게 반응 함

✓ **23. 다음 중 분류 모형에 대한 설명으로 적절한 것은 무엇인가?** 1/1

- ○ 레코드 자체가 가진 다른 레코드와의 유사성에 의해 그룹화되고 이질성에 의해 세분화 된다
- ○ 카탈로그 배열, 교차판매 등의 마케팅 계획에 사용되는 데이터마이닝 기법이다
- ◉ 새롭게 나타난 현상을 검토하여 기존의 분류, 정의된 집합에 배정하는 것으로 현상 이해를 위해 데이터를 범주, 등급 등으로 나눈다 ✓
- ○ 데이터가 가진 특징 및 의미를 단순하게 설명하는 것이다

의견 보내기

3-80. 대표적 데이터 마이닝 기법
- 분류 : 새롭게 나타난 현상을 검토하여 기존의 분류, 정의된 집합에 배정하는 것으로 현상 이해를 위해 데이터를 범주, 등급 등으로 나눈다
- 군집 : 레코드 자체가 가진 다른 레코드와의 유사성에 의해 그룹화되고 이질성에 의해 세분화 된다
- 연관분석 : 카탈로그 배열, 교차판매 등의 마케팅 계획에 사용되는 데이터마이닝 기법이다
- 기술(description): 데이터가 가진 특징 및 의미를 단순하게 설명하는 것이다

✓ **24. 다음 중 연관분석에 대한 특징으로 적절한 것은 무엇인가?** 1/1

- (●) 조건반응(if ~ then)으로 표현되는 연관 분석의 결과를 이해하기 쉽다 ✓
- ( ) 강력한 목적성 분석 기법에 해당한다
- ( ) 세분화된 품목을 가지고 연관규칙을 찾아야만 의미 있는 분석 결과가 도출된다
- ( ) 분석 품목 수가 증가하더라도 분석 계산이 많이 증가하지는 않는다

의견 보내기

3-98. 연관분석(Association Analysis)
- 조건반응(if-then)으로 표현되는 연관 분석의 결과를 이해하기 쉬움
- 강력한 비목적성 분석 기법이며, 분석 계산이 간편함
- 분석 품목 수가 증가하면 분석 계산이 기하급수적으로 증가함
- 너무 세분화된 품목을 가지고 연관규칙을 찾으려면 의미 없는 분석 결과가 도출됨
- 상대적 거래량이 적으면 규칙 발견 시 제외되기 쉬움

✓ **25. 다음이 설명하는 이산형 확률분포는 무엇인가?** 1/1
"단위 시간이나 단위 공간에서 어떤 사건이 몇 번 발생할 것인지를 표현하는 분포로 특정 기간 동안 사건 발생의 확률을 구할 때 사용된다"

포아송 분포 ✓

의견 보내기

3-52. 이산형 확률분포 – 포아송분포
- 단위 시간이나 단위 공간에서 어떤 사건이 몇 번 발생할 것인지를 표현하는 분포
- 특정 기간 동안 사건(events) 발생의 확률을 구할 때 쓰임

✓ **26. P(A)=0.3, P(B)=0.4 일 때, 사건 A와 사건 B가 독립사건일 경우 P(B|A)는?** 1/1

0.4 ✓

의견 보내기

3-48. 독립사건
- 두 사건 A, B가 독립이면 P(B|A)=P(B), P(A|B) = P(A) 이다

✓ **27. 다음 빈 칸에 들어갈 알맞은 용어는 무엇인가?** 1/1

( ) 두 군집 사이의 거리를 군집에서 하나씩 관측 값을 뽑았을 때 나타날 수 있는 거리의 최솟값을 측정하는 계층적 군집의 거리 기반 측정 방법이다. 사슬 모양으로 생길 수 있으며 고립된 군집을 찾는데 중점을 두는 방식이다.

최단연결법 ✓

✓ **28. 다음이 설명하는 데이터마이닝의 모형평가 방법은 무엇인가?** 1/1

원천 데이터를 랜덤하게 두 분류로 분리하여 교차검정을 실시하는 방법으로 하나는 모형 학습 및 구축을 위한 훈련용 자료로 다른 하나는 성과평가를 위한 검증용 자료로 사용하는 방법이다.

홀드 아웃 ✓

✓ **29. 설명변수 선택 방법 중에서 독립변수 후보를 모두 포함한 모형에서 출발해 제곱합의 기준으로 가장 적은 영향을 주는 변수부터 하나씩 제거하면서 더 이상 유의하지 않은 변수가 없을 때까지 설명변수를 제거하는 모형은 무엇인가?** 1/1

후진제거법 ✓

의견 보내기

3-67. 설명 변수 선택 방법 – 후진제거법
모든 가능한 조합, 전진선택법, 후진제거법, 단계별 선택법 등의 설명 변수 선택 방법이 있음

✓ **30. 차원 축소 기법 중, 객체들 사이의 유사성, 비유사성을 2차원 혹은 3차원 공간상에 점으로 표현하여 개체 사이의 군집을 시각적으로 표현하는 기법은 무엇인가?** 1/1

다차원 척도법 ✓

의견 보내기

3-73. 차원 축소 기법 – 다차원 척도법
객체 간 근접성을 시각화 하는 기법으로, 개체들 사이의 유사성, 비유사성을 2차원 혹은 3차원 공간상에 점으로 표현하여 개체 사이의 군집을 시각적으로 표현하는 방법

Google 설문지