빅데이터 분석기사 필기 기출 문제(6회, 2023년 4월 8일)

1과목 빅데이터 분석 기획

1. 맵리듀스 디자인 패턴 중 다른 데이터와 연결하여 분석하는 패턴은?

① 요약 패턴 ② 조인 패턴

③ 필터링 패턴 ④ 메타 패턴

맵리듀스 디자인 패턴에는 요약 패턴, 필터링 패턴, 데이터 조직화 패턴, 조인 패턴, 메타 패턴, 입출력 패턴이 있고, 여기서 다른 데이터와 연결하여 분석하는 패턴은 조인 패턴이다.

맵리듀스(MapReduce)

● 구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004년에 발표한 소프트웨어 프레임워크이다.

● 맵리듀스는 맵(Map) 작업과 리듀스(Reduce) 작업의 결합이다.

● 맵(Map) 작업은 여러 데이터를 Key-Value의 형태로 연관성 있는 데이터로 분류하여 묶는 작업이다.

● 리듀스(Reduce) 작업은 맵 작업한 데이터 중 중복 데이터를 제거하고 원하는 데이터를 추출하는 작업이다.

● 맵리듀스 과정 : Input -> Splitting -> Shuffling -> Reducing -> Final Result

2. 다음 중 데이터 탐색에 대한 설명으로 옳지 않은 것은?

① 데이터 탐색은 수집한 데이터를 분석하기 전에 통계적인 방법을 이용하여 다양한 각도에서 데이터의 특징을 파악하는 분석 방법이다.

② 탐색적 데이터 분석의 특징으로는 저항성, 잔차해석, 자료 재표현, 현시성이 있다.

③ 범주형↔범주형 데이터의 시각화는 막대형 그래프를 사용한다.

④ 데이터 탐색은 모형 해석 시에 필요하다.

데이터 탐색은 수집한 데이터를 분석하기 전에 통계적인 방법을 이용하여 다양한 각도에서 데이터의 특징을 파악하는 분석 방법이다.

데이터의 특성을 파악하기 위해 시각화하여 분석하기도 하며, 데이터 탐색 도구에는 도표, 그래프, 요약 통계 등이 있다.

탐색적 데이터 분석이란 노키아 벨 연구소의 수학자 존 튜키가 개발한 개념으로, 데이터를 분석하고 결과를 내는 과정에서 지속적으로 해당 데이터에 대한 '탐색과 이해'를 기본으로 가져야 한다는 것을 의미한다.

탐색적 데이터 분석의 4가지 특징 저항성, 잔차해석, 자료 재표현, 현시성이 있다.

① 저항성(Resistance) : 오류의 영향을 적게 받는 성질로 저항성이 큰 데이터를 사용한다.

② 잔차해석(Redidual) : 잔차는 관찰 값들이 주 경향으로부터 벗어난 정도이며, 이를 해석하며 데이터의 특징을 파악한다.

③ 자료 재표현(Re-expression) : 데이터 분석 및 해석의 용이성을 위해 변수를 적당한 척도로 바꾸는 것이다.

④ 현시성(Graphic Representation) : 데이터 시각화라고도 할 수 있으며, 분석 결과를 쉽게 이해할 수 있도록 데이터를 시각적으로 표현하는 것이다.

모형 해석은 데이터 탐색을 통해 모델링한 후에 진행한다.

3. 다음 중 외부 공공데이터 이용의 장점은?

① 데이터 제공자와 상호협약에 의한 의사소통이 가능하다.

② 제공되는 데이터의 범위가 넓다.

③ 주로 정형 데이터 형태로 수집이 용이하다.

④ 개인정보보호에 관한 문제점을 사전에 점검할 수 있다.

외부 데이터는 수집하려는 데이터가 외부 저장소에 저장된 것으로 주로 수집이 어려운 비정형 데이터이나, 제공되는 데이터의 범위가 넓다는 특징이 있다.

①, ③, ④은 내부 데이터에 대한 설명이다.

저장 위치에 따른 데이터 구분

1. 내부 데이터

● 부서 간 업무 협조와 개인정보보호 및 정보보안 관련된 문제점을 사전에 점검하여 수집

● 데이터 제공자와 상호 협약에 의한 의사소통이 가능

● 주로 수집이 용이한 정형 데이터의 형태

2. 외부 데이터

● 시스템 간 다양한 인터페이스 및 법적인 문제점을 고려하여 상세한 데이터 수집 계획 수립

● 데이터 제공자와 협약된 관계가 아닐 경우 상호 의사소통 불가능

● 주로 수집이 어려운 비정형 데이터의 형태

4. 다음 중 빅데이터 시대 위기 요인이 아닌 것은?

① 데이터 오용

② 책임 원칙 훼손

③ M2M시대 본격화

④ 사생활 침해

빅데이터 시대의 위기 요인으로는 사생활 침해, 책임 원칙 훼손, 데이터 오용이 있다.

M2M(Machine To Machine)은 네트워크를 통한 사물간의 통신을 의미한다.

5. 다음 중 탐색적 데이터 분석에 대한 설명으로 옳은 것은?
① 탐색적 데이터 분석으로 데이터를 시각화할 수는 없다.
② 변숫값과 자료구조 간의 관계를 알 수 있다.
③ 범주형 데이터의 시각화는 주로 박스플롯을 사용한다.
④ 수치형 데이터의 시각화는 주로 막대형 그래프를 사용한다.
탐색적 데이터 분석은 수집된 데이터를 다양한 방법을 활용하여 분석하여 데이터의 특징을 정확하게 파악하는 것으로 개별 변수 탐색, 다차원 데이터 탐색 등으로 데이터의 구조를 파악할 수 있다. 탐색적 데이터 분석으로 데이터를 시각화 할 수 있다. (현시성)
개별 변수 탐색 방법은 변수가 범주형과 수치형인 경우로 나뉜다.
1. 범주형 데이터(질적 데이터)
● 명목형 변수와 순서형 변수에 대한 데이터 탐색 방법
● 빈도수, 최빈값, 비율, 백분율 등을 활용하여 데이터 분포의 특징을 중심성, 변동성 측정에서 파악
● 시각화는 막대형 그래프(Bar Plot)를 주로 사용
2. 수치형 데이터(양적 데이터)
● 이산형 변수와 연속형 변수에 대한 데이터 탐색 방법
● 평균, 분산, 표준편차, 첨도, 왜도 등을 이용하여 데이터 분포의 특징을 정규성 측면에서 파악
● 시각화는 박스플롯(Box-Plot) 또는 히스토그램을 주로 사용

6. 다음 중 데이터 전처리 과정에 해당하는 분석 과정은?
① 데이터 시각화 ② 모델링
③ 적합도 검정 ④ 데이터 축소
데이터 전처리 과정에서는 데이터를 정제하고, 분석 변수를 처리한다. 이 과정에서 **결측값, 이상값을 처리하며, 차원을 축소하고 변수를 변환한다.** 데이터 시각화 및 적합도 검정은 빅데이터 결과 해석 과정에서 수행되고, 모델링은 빅데이터 모델링 과정에서 수행된다.

7. 다음 중 데이터 사이언스에 대한 설명으로 옳은 것은?
① 인문, 사회, 공학 등 전반적인 영역에 골고루 퍼져 있다.
② 데이터 사이언스에는 딥러닝 기술이 활용되지 않는다.
③ 데이터 사이언스를 위해 활용되는 데이터는 주로 소규모 데이터이다.
④ 데이터 사이언스에 필요한 기술에 비즈니스 관련 기술은 포함되지 않는다.
데이터 사이언스는 분석 방법, 도메인 전문성 및 기술의 융합을 통해 데이터에서 패턴을 찾고, 추출하고, 표면화하는 다학문적인 접근 방식이다. 데이터 사이언스에 활용되는 기술은 대부분 대규모 데이터이고, 딥러닝 기술이 활용된다. 데이터 사이언스에 필요한 기술에는 비즈니스 기술, 분석 기술, IT 기술이 있다.

8. 다음 중 분석 준비도의 척도가 아닌 것은?
① 분석 문화 ② 분석 업무
③ 분석 결과 활용 ④ 분석 인력
분석 준비도(Readiness)
● 데이터를 분석하여 업무 및 의사결정에 활용하기 위해 준비가 어느 정도 되어 있는지를 점검하는 체계이다.
● 조직 내 데이터 분석 업무 도입을 목적으로 현재 수준을 파악하기 위한 진단 방법이다.
● 총 6가지 영역을 대상으로 현재 수준을 파악한다.
● **분석 업무, 인력 및 조직, 분석 비법, 분석 데이터, 분석 문화, IT 인프라가 있다.**
● 진단 영역별 세부 항목에 대한 수준까지 파악을 해야 한다.
● 각 진단 결과 전체 요건 중 일정 수준 이상 충족 시 데이터 분석 업무 도입을 한다.
● 만일 일정 수준 이상 충족되지 못하면 데이터 분석 환경을 먼저 조성한다.
분석 결과 활용은 분석 성숙도 진단 척도이다. 분석 성숙도 진단 단계에는 도입, 활용, 확산, 최적화가 있다.

9. 다음 중 연속형 변수가 아닌 것은?
① 형광등 수명 ② 혈액형
③ 키 ④ 나이
혈액형은 범주형 변수로써 이름으로 범주를 구분한다. A형, B형, AB형, O형과 같이 범주 형태로 데이터를 분류할 수 있다. 성별(남자/여자), 지역(서울/부산/광주/대구) 등이 있다.
연속형 변수는 연속적인 수로 수량화가 가능한 자료를 의미한다.

10. 빅데이터를 정형, 비정형, 반정형으로 나눌 경우 빅데이터의 어떠한 특성을 기준으로 나눈 것인가?
① 저장 위치 ② 변수 개수
③ 수집 방법 ④ 다양성
데이터는 저장 위치에 따라 내부 데이터와 외부 데이터로 나눌 수 있고, 데이터의 형태에 따라 정형, 비정형, 반정형 데이터로 나뉜다.

1. 정형 데이터

● 정해진 형식과 구조에 따라 저장된 데이터

● 예) Excel, 스프레드시트, 관계형 데이터베이스 테이블

2. 반정형 데이터

● 데이터 구조 정보를 데이터와 함께 제공하는 형식의 데이터

● 예) JSON, XML, HTML 등

JSON(JavaScript Object Notation) : 사람이 읽을 수 있는 데이터 교환용으로 설계된 경량 텍스트 기반 개방형 표준 포맷으로 '키(Key) - 값(Value)' 으로 구성된다.

XML(Extensible Markup Language) : 데이터를 정의하는 규칙을 제공하는 마크업 언어

HTML(Hyper Text Markup Lanuage) : 웹 페이지의 표시를 위해 개발된 마크업 언어

3. 비정형 데이터

● 정의된 구조가 없는 형태의 정형화되지 않은 데이터

● 예) 동영상 파일, 오디오 파일, 사진, 보고서, 이메일 등등

11. 다음 중 데이터셋의 noise를 제거하거나 최소화하기 위한 알고리즘은?

① 일반화(generalization)

② 집계(aggregation)

③ 평활(smoothing)

④ 속성 생성(feature construction)

데이터셋의 noise를 제거하거나 최소화하기 위한 알고리즘은 데이터 변환 기술 중 하나인 평활화(Smoothing)이다.

데이터 변환 기술

1. 평활화(Smoothing) : 데이터의 노이즈를 구간과 군집화 등으로 다듬는 기법

2. 집계(Aggregation) : 다양한 차원으로 데이터를 요약하는 기법

3. 일반화(Generalization) : 특정 구간으로 값을 스케일링 하는 기법

4. 정규화(Normalization) : 데이터를 정해진 구간(0~1)으로 전환하는 기법

5. 속성 생성(Feature Construction) : 여러 데이터를 대표할 수 있는 새로운 속성값을 생성하는 기법

12. 다음 중 데이터 분석 조직 구조에 대한 설명으로 옳지 않은 것은?

① 빅데이터 조직 구조 유형에는 집중 구조, 기능 구조, 분산 구조가 있다.

② 집중 구조는 별도의 분석 조직이 존재하고, 협업 부서와 기능이 겹치지 않는다.

③ 기능 구조는 전사적 핵심 분석이 어려우며, 과거에 국한된 분석 수행 가능성이 높다.

④ 분산 구조는 업무 과다, 이원화 가능성이 존재할 수 있기 때문에 부서 분석 업무와 역할 분담이 명확해야 한다.

빅데이터 조직 구조 유형에는 집중 구조, 기능 구조, 분산 구조가 있다.

집중 구조는 전사의 분석 업무를 별도의 분석 전담 조직에서 담당하는 구조이다.

집중 구조는 일반 업무 부서의 분석 업무와 중복 혹은 이원화될 가능성이 높다.

13. 다음 중 데이터 거버넌스의 3요소가 아닌 것은?

① 원칙 ② 조직

③ 시스템 ④ 프로세스

데이터 거버넌스의 3요소는 원칙, 조직, 프로세스이다.

데이터 거버넌스란 기업에서 사용하는 데이터의 가용성, 유용성, 통합성, 보안성을 관리하기 위한 정책과 프로세스를 다루며 프라이버시 보안성, 데이터 품질, 관리 규정 준수를 강조하는 모델을 의미한다.

14. 다음 중 네트워크를 기반으로 파일의 수집 및 공유가 가능한 시스템은?

① 관계형 데이터베이스

② NoSQL

③ HBase

④ 분산 파일 시스템

네트워크를 기반으로 파일의 수집 및 공유가 가능한 시스템은 분산 파일 시스템(DFS)이다. 정형 데이터는 관계형 데이터베이스(RDBMS), 반정형 데이터는 비관계형 데이터베이스(NoSQL), 비정형 데이터는 분산 파일 시스템(DFS)에 저장된다.

HBase는 HDFS(Hadoop Distributed File System)의 분산 컬럼 기반 데이터베이스이다. 실시간 랜덤 조회 및 업데이트를 할 수 있으며, 각각의 프로세스는 개인의 데이터를 비동기적으로 업데이트 할 수 있다.

15. 다음 중 데이터 분석 수행을 위한 현황 파악 및 분석을 통한 문제를 정의하는 단계는?
① 분석 목표 수립
② 프로젝트 계획 수립
③ 보유 데이터 자산 확인
④ 도메인 이슈 도출
데이터 분석 영역에서 데이터 분석 수행을 위한 현황 파악 및 분석을 통한 문제를 정의하는 단계는 도메인 이슈 도출 단계이다.

데이터 분석 영역

● 저장된 데이터를 추출하여 분석 목적과 방법에 맞게 가공하여 데이터 분석을 수행하고 결과를 표현하는 영역이다.
● 국가직무표준능력(NCS) 데이터 분석은 다음과 같이 구분된다.

1. 도메인 이슈 도출

● 분석 대상 과제 현황 파악 및 개선 과제 정의
● 문제의 주요 이슈별로 개선 방향 도출, 개선 방안 수립, 빅데이터 요건 정의서 작성

2. 분석 목표 수립

● 빅데이터 요건 정의서 기반 현실적인 분석 목표를 수립
● 데이터 관련 정보, 분석 타당성 검토, 성과 측정 방법 등을 포함한 분석 목표 정의서 작성

3. 프로젝트 계획 수립

● 사전에 책정된 자원과 예산, 기간 등을 고려하여 분석 프로젝트 계획 수립
● 분석목표정의서, 프로젝트 소요 비용, 배분 계획을 바탕으로 작업분할구조도 작성

4. 보유 데이터 자산 확인

● 분석 목표와 프로젝트 계획을 기반으로 현재 보유 중인 데이터의 품질이나 규모, 유형 등을 확인하고 법률적 이슈 혹은 제약사항 검토

16. 다음 중 분석 마스터 플랜에 대한 설명으로 옳은 것은?
① 전략적 중요도, 비즈니스 성과 및 ROI, 분석 과제의 실행 용이성을 고려하여 분석 구현 로드맵을 수립한다.
② 업무 내재화 적용 수준, 분석 데이터 적용수준, 기술 적용 수준을 고려하여 우선순위를 설정한다.
③ ISP는 정보기술 및 정보 시스템을 전략적으로 활용하기 위해 중장기 마스터플랜을 수립하는 절차이다.
④ 과제 우선순위 평가기준의 시급성에는 분석 수준, 분석 적용 비용이 포함된다.

분석 마스터 플랜

● 지속적으로 분석이 주는 가치를 체계적으로 관리하고 분석 역량을 내재화하려면 단기적인 과제 수행뿐만 아니라 중/장기적 관점의 마스터 플랜 수립이 필요하다.
● 분석 마스터 플랜 과정에서는 전략적 중요도, 비즈니스 성과 및 ROI, 분석 과제의 실행 용이성을 고려하여 우선순위를 설정한다.
● 분석 마스터 플랜 과정에서는 업무 내재화 적용 수준, 분석 데이터 적용 수준, 기술 적용 수준을 고려하여 분석 구현 로드맵을 수립한다.
● 정보 전략 계획(ISP; Information Strategy Planning)은 정보기술 및 정보 시스템을 전략적으로 활용하기 위해 중장기 마스터 플랜을 수립하는 절차이다.

17. 다음 중 기업의 분석 수준 진단에 대한 설명으로 옳지 않은 것은?
① 확산형은 기업에 필요한 분석 구성 요소를 갖추고 있고, 높은 성숙도를 갖는 유형이다.
② 정착형은 조직 및 인력, 분석 업무, 분석 기법이 내부에 오픈되어 있다.
③ 도입형은 기업에서 활용하는 분석 업무 및 기법은 부족하지만 준비도가 높아 바로 도입할 수 있는 유형이다.
④ 준비형은 기업에 필요한 구성 요소 등이 준비되지 않아 사전 준비가 필요한 유형이다.

기업의 분석 수준 진단

1. 준비형
● 낮은 준비도, 낮은 성숙도, 사전 준비가 필요
2. 정착형
● 낮은 준비도, 높은 성숙도, 분석의 정착이 필요
3. 도입형
● 높은 준비도, 낮은 성숙도, 데이터 분석 도입 가능
4. 확산형
● 높은 준비도, 높은 성숙도, 지속적 확산이 가능

18. 다음 설명하는 파생변수 생성 방법에 해당하는 것은?
타이타닉 생존자 데이터에서 형제, 부모 데이터를 가족 데이터로 결합
① 단위 변환
② 표현방식 변환
③ 요약 통계량 변환
④ 변수 결합

변수를 결합하여 새로운 변수를 정의하는 방법은 변수 결합 방법이다.

파생변수 생성 방법

1. 단위 변환

● 주어진 변수의 단위를 새로운 단위로 변환하여 데이터를 표현하는 방법

2. 표현방식 변환

● 표현방식을 단순화하는 변환 방법

● 예) 미혼, 기혼 데이터를 0, 1로 변환

3. 요약 통계량 변환

● 요약 통계량을 활용하는 변환 방법

● 예) 성별에 따른 A제품 재구매율 확인

4. 정보 추출

● 하나의 정보에서 새로운 정보를 추출하는 방법

● 예) 차량 번호판에서 개인소유 혹은 렌터카 여부 확인

5. 변수 결합

● 변수를 결합하여 새로운 변수를 정의하는 방법은 변수 결합 방법이다.

6 조건문 이용

● 조건문을 활용하여 파생변수를 생성하는 방법

● 예) 백화점 구매 데이터에서 고객의 1년 구매액이 8,000만원 이상인 경우 A그룹, 아닌 경우 B 그룹의 변수로 분류

19. 다음 중 데이터 정제 방법이 아닌 것은?

① 삭제 ② 표준화

③ 대체 ④ 예측값 삽입

데이터 정제(Data Cleansing)

● 데이터 정제는 데이터를 깨끗하게 다듬어서 데이터의 신뢰도를 높이는 작업

● **데이터 정제 방법에는 삭제, 대체(최빈값, 중앙값, 평균값 활용), 예측값 삽입**

20. 다음 중 개인정보 비식별화 조치에 대한 설명으로 옳지 않은 것은?

① 가명처리는 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보 없이는 특정 개인을 알아볼 수 없도록 처리하는 방법이다.

② 데이터 범주화는 특정 정보를 해당 그룹의 대푯값으로 변환하거나 구간값으로 변환하여 특정 개인을 식별할 수 없도록 하는 방법이다.

③ 총계처리는 통계값을 적용하여 특정 개인을 식별할 수 없도록 하는 방법이다.

④ 데이터 마스킹은 민감 데이터 부분을 국소적으로 삭제하는 것이다.

데이터 마스킹은 민감 정보 일부를 * 와 같은 기호로 표기하는 방법이다.

비식별화 조치방법

1. 가명처리

● 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보 없이는 특정 개인을 알아볼 수 없도록 처리하는 방법이다.

세부 기술 : 휴리스틱 가명화, 암호화, 교환 방법

예) 신은혁, 17세, 대구 거주, A 고등학교 재학

김은요, 10대, 대구 거주, B 고등학교 재학

2. 총계처리

● 통계값을 적용하여 특정 개인을 식별할 수 없도록 하는 방법

세부 기술 : 총계 처리, 부분 총계, 라운딩, 재배열

예) 신은혁, 173cm, 김은혁, 170cm,....

학생 키 합 : 343cm, 평균 키 : 171.5cm ...

3. 데이터 삭제

● 민감 데이터 일부 혹은 전체를 삭제하여 개인을 식별할 수 없도록 하는 방법

세부 기술 : 식별자 삭제, 식별자 부분 삭제, 레코드 삭제, 식별 요소 전부 삭제

예) 주민등록번호 : 800111-1234567 -> 80년대생 남자

4. 데이터 범주화

● 특정 정보를 해당 그룹의 대푯값으로 변환하거나 구간값으로 변환하여 특정 개인을 식별할 수 없도록 하는 방법이다.

세부 기술 : 감추기, 랜덤 라운딩, 범위 방법, 제어 라운딩

예) 신은혁, 17세 -> 신씨, 10~20세

5. 데이터 마스킹

● 데이터 마스킹은 민감 정보 일부를 * 와 같은 기호로 표기하는 방법이다.

세부 기술 : 임의 잡음 추가, 공백화 대체

예) 신은혁, 17세, 대구 거주, A 고등학교 재학

신**, 17세, 대구 거주, **고등학교 재학

2과목 빅데이터 탐색

21. 다음 중 주성분 분석(PCA)에 대한 설명으로 옳지 않은 것은?

① 비정방 행렬을 음상관 행렬의 곱으로 바꾼다.

② 가장 보편적으로 사용되는 차원 축소 기법 중 하나다.

③ 원본 데이터를 최대한 보존하면서 고차원 공간의 데이터를 저차원 공간 데이터로 변환하는 기법이다.

④ 기존 변수들을 조합하여 서로 연관성이 없는 새로운 변수를 생성한다.

주성분 분석(PCA; Principal Component Analysis)

● 가장 보편적으로 사용되는 차원 축소 기법 중 하나로 원본 데이터를 최대한 보존하면서 고차원 공간의 데이터를 저차원 공간 데이터로 변환하는 기법이다.

● 기존 변수들을 조합하여 서로 연관성이 없는 새로운 변수(주성분 PC(Principal Component))를 생성한다.

● 행과 열의 크기가 같은 정방행렬에서만 사용한다.

● 첫 번째 주성분(PC1)은 원 데이터의 분포를 가장 많이 보존하고, 두 번째 주성분(PC2)이 그 다음으로 원 데이터의 분포를 많이 보존한다.

22. 다음 설명하는 결측값 대치법에 해당하는 것은?

> 단순대치법을 한 번 하지 않고, n번 대치를 통해 n개의 완전한 자료를 만들어 분석하는 방법으로, 대치 → 분석 → 결합의 3단계로 구성된다.

① 핫-덱 대체　　② 콜드덱 대체
③ 다중대치법　　④ 혼합방법

위의 구문의 설명은 다중 대치법에 대한 설명이다. 나머지는 단순확률대치법에 속한다.

단순확률대치법(Single Stochastic Imputation) : 적절할 확률값을 부여한 후 이를 결측값으로 대치하는 방법이다.

● **핫-덱(Hot-Deck) 대체** : 진행 중인 연구 내에서 비슷한 성향의 자료로 결측값을 대체하는 방법

● **콜드덱(Cold-Deck) 대체** : 진행 중 연구 내부가 아닌 외부 출처 또는 이전에 비슷한 연구에서 대체 값을 가져오는 방법

● **혼합방법** : 다양한 방법을 혼합하는 방법

23. 다음 중 표현하고 싶은 데이터를 1값으로, 그렇지 않은 데이터를 0값으로 표현하는 인코딩 방식은?

① 레이블 인코딩
② 대상 인코딩
③ 카운트 인코딩
④ 원-핫 인코딩

인코딩 : 데이터의 형태나 형식을 변환하는 처리 방법으로 데이터 분석에서는 문자열 데이터를 숫자형 데이터로 변환하는 기술을 의미한다.\

인코딩의 종류

● **원-핫 인코딩(One-Hot Encoding)**

원-핫 인코딩은 표현하고자 하는 데이터를 1값으로, 그렇치 않은 데이터를 0값으로 표현하는 방식이다.

● **레이블 인코딩**

범주형 변수의 문자열 데이터를 수치형으로 변환하는 방식이다.

● **카운트 인코딩**

각 범주의 데이터 개수를 총합하여 그 개수의 수치값을 인코딩하는 방식이다.

● **대상 인코딩**

범주형 데이터의 값들을 목표하는 데이터 값으로 바꿔주는 방식이다.

대상 인코딩은 원-핫 인코딩에서 변수의 값이 많아지는 문제를 해결해 준다.

24. 다음 중 데이터 일관성 유지를 위한 방법이 아닌 것은?

① 삭제　　② 변환
③ 파싱　　④ 보강

데이터 일관성 유지를 위한 정제 기법

● **변환(Transform)** : 다양한 형태로 표현된 데이터를 일관된 형태로 변환하는 작업

● **파싱(Parsing)** : 데이터를 유의미한 최소 단위로 분할하는 작업

● **보강(Enhancement)** : 변환, 파싱, 표준화 등을 통한 추가적인 정보를 반영하는 작업

삭제는 이상값 처리 방법이다.

25. 다음 중 이상값 처리에 대한 설명으로 옳지 않은 것은?

① 이상값 처리 방법에는 삭제, 대체, 변환이 있다.
② 평균값으로 이상값을 대체해도 데이터 변환 시에 신뢰도 문제가 발생하지 않는다.
③ ESD는 평균(μ)으로부터 3시그마(σ, 표준 편차) 떨어진 값을 이상치로 인식하는 방법으로, 양쪽 0.15%에 해당하는 값을 이상치로 인식한다.
④ 머신러닝 기법을 활용하여 이상값을 검출할 수 있다.

평균값은 이상값 가장 많이 영향을 받기 때문에 이상값에 영향을 받지 않는 중앙값으로 이상값을 대체한다.

이상값(Outlier) : 일반적인 데이터 범위를 많이 벗어난 아주 작은 값 또는 큰 값을 의미한다.

이상값 처리 방법

● 삭제(Deleting Observation) : 이상값으로 확인된 값을 삭제하는 방법이다.

● 대체(Imputation) : 이상값을 평균 또는 중위수로 대체하는 방법이다.

● 변환(Transformation) : 극단적인 값으로 인해 발생된 이상값의 경우 데이터에 자연로그를 취해서 값을 감소시키는 방법이다.

ESD(Extreme Studentized Deviate Test)
평균으로부터 3표준편차 떨어진 값을 이상치로 인식하는 방법으로 3표준편차에 해당하는 값은 99.7%이다.

26. 다음과 같은 표본집단 데이터의 평균값과 분산은 얼마인가?

2, 4, 6, 8, 10

	평균	분산
①	5	10
②	5	8
③	6	10
④	6	8

평균값은 6이다.

분산의 공식 : $s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ 된다.

공식에 대입하면 $\frac{(2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2}{4}$ = 10이 된다.

27. 다음 중 데이터 정제(Data Cleansing)에 대한 설명으로 옳지 않은 것은?
① 데이터 정제는 원본 데이터를 다듬어서 데이터의 신뢰도를 높이는 작업이다.
② 데이터 정제의 목적은 데이터를 이해하기 쉽게 표현하는 것이다.
③ 데이터 정제 과정은 데이터 오류 원인 분석 -> 데이터 정제 대상 선정 -> 데이터 정제 방법 결정 순이다.
④ 데이터 정제 방법에는 삭제, 대체, 예측값 삽입이 있다.
데이터 정제의 목적은 데이터를 정제하여 데이터의 신뢰도를 높이는 것이다.
데이터를 이해하기 쉽게 표현하는 것은 데이터 시각화의 목적에 해당한다.

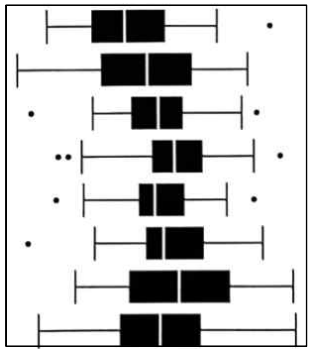
28. 다음 중 산포도 통계량에 대한 설명으로 옳지 않은 것은?
① 산포도 통계량은 데이터의 흩어진 정도를 나타내는 통계량이다.
② IQR은 사분위수 범위로 $Q_3 - Q_1$ 와 같이 연산된다.
③ 사분편차는 IQR의 절반값이다.
④ 변동계수는 분산을 평균으로 나눈 값이다.
변동계수는 표준편차를 평균으로 나눈 값이다.
산포도 통계량(데이터 흩어진 정도)
● 분산(Variance) : 평균으로부터 얼마나 떨어져 있는지 나타내는 값
● 표준편차(Standard Deviation) : 분산에 양의 제곱근을 취한 값. 자료의 산포도를 나타내는 수치이다.
● 범위(Range) : 데이터 값 중에서 최댓값과 최솟값의 차이이다.
● **IQR(InterQuatile Range, 사분위수 범위)** : 3사분위수와 1사분위수의 차이 값이다. 사분위수 범위는 Q3 - Q1이다.
● 사분편차(Quartile Deviation)
IQR의 절반값이다.

29. 다음 설명에 해당하는 확률 분포는?

• 단위시간 또는 영역에서 어떤 사건의 발생횟수를 나타내는 확률 분포이다. • 수식은 $P=\frac{\lambda^n e^{-\lambda}}{n!}$ (λ : 평균, n : 발생횟수)와 같이 표현된다.

① 베르누이 분포
② 푸아송 분포
③ 이항분포
④ 연속확률분포
위의 내용은 푸아송 분포에 대한 내용이다.
이산확률분포의 종류에는 푸아송 분포, 베르누이 분포, 이항 분포가 있다.
베이누이 분포
● 특정 실험의 결과가 성공 또는 실패로 두 가지 중 하나의 결과를 얻는 확률 분포
이항 분포
● n번 시행 중에 각 시행의 확률이 P일 때, k번 성공할 확률 분포를 의미한다.
● 공식 : $P = \binom{n}{k} P^k (1-P)^{n-k}$ (n : 시행 횟수, P : 특정 사건이 성공할 확률, k : 성공 횟수)
연속확률분포
● 연속확률변수 X가 갖는 확률 분포를 나타낸다.
● 연속확률분포의 종류에는 정규분포, 표준정규분포(Z-분포), T-분포, 카이제곱 분포, F-분포, 지수 분포, 감마 분포가 있다.

30. 다음과 같은 형태의 차트 이름은?



① Catogram
② Box-plot
③ Histogram
④ Heat Map

위의 그림은 상자수염그림(Box-Plot)을 나타낸다.

카토그램(Catogram)

● 특정한 데이터 수치의 변화에 따라서 지도의 면적이 왜곡되는 그래프이다.
● 주로 의석 수나 선거인단 수, 인구 등의 데이터를 표현한다.

히스토그램(Histogram)

● 데이터의 분포를 서로 붙어 있는 직사각형 형태로 시각화하여 표현하는 그래프이다.
● 가로축은 반드시 수량을 나타내고, 막대의 너비는 항상 일정해야 한다.

히트맵(Heat Map)

● 색상으로 표현할 수 있는 다양한 정보를 일정한 이미지 위에 열분포 형태로 표현한 그래프이다.
● 각 칸별 색상은 데이터 값의 크기를 나타내고, 색상이 짙을수록 데이터 값이 큰 것을 의미한다.
● 웹 사이트 방문자 분석 및 다양한 분야에서 사용된다.

박스플롯(Box-Plot, 상자수염그림, 상자그림)

● 많은 데이터를 그림을 이용하여 집합의 범위와 중위수를 빠르게 확인할 수 있다.
● 통계적으로 이상값이 있는지 빠르게 확인 가능하다.

31. 시간 시각화 자료 중 일정 기간 동안 측정된 데이터들의 경향성을 보여주는 직선 또는 곡선은?

① 누적막대그래프　　② 추세선
③ 점그래프　　④ 계단그래프

누적막대그래프(Stacked Bar Chart)

● 하나의 막대로 데이터의 여러 범주별 비율을 확인할 수 있는 그래프이다.

점그래프(Dot Plot)

● x축에 따른 y축의 값을 점으로 표시한 그래프로, x축은 시간, y축은 데이터인 경우 시간의 흐름에 따른 데이터의 변화를 확인할 수 있다.

계단식그래프(Step Line Graph)

● 각 범주별 측정된 데이터를 선분으로 연결하는 것이 아니라 x축과 평행한 일정한 선을 유지하고 값이 급격히 변하는 지점을 이전 데이터와 계단식으로 이어 표현해 주는 그래프이다.

32. 다음 중 표본분포에 대한 설명으로 옳지 않은 것은?

① 중심 극한 정리는 데이터의 크기가 작아지면 데이터의 표본분포는 최종적으로 정규분포의 형태를 따른다는 것이다.
② 표본분포는 모집단에서 추출한 일정한 크기의 표본에 대한 분포 상태를 의미한다.
③ 모수는 모집단 분포 특성을 규정짓는 척도로 관심의 대상이 되는 모집단의 대푯값이다.
④ 큰 수의 법칙은 데이터를 많이 선택할수록 표본평균의 분산은 0에 가까워진다는 것이다.

중심 극한 정리는 데이터의 크기가 커지면 데이터의 표본분포는 최종적으로 정규분포의 형태를 따른다는 것이다.

표본분포(Sample Distribution)

● 표본분포는 모집단에서 추출한 일정한 크기의 표본에 대한 분포 상태를 의미한다.

표본분포 용어

● 모집단(Population) : 정보를 얻고자 하는 관심 대상의 전체 집합이다.
● 모수(Parameter) : 모집단 분포 특성을 규정짓는 척도로 관심의 대상이 되는 모집단의 대푯값이다.
● 통계량(Statistic) : 표본의 몇몇 특징을 수치화 한 값(평균, 표준오차)
통계량을 통해 모수를 추정하고, 무작위로 추출할 경우 각 표본에 따라 달라지는 확률변수이다.
● 추정량(Estimator) : 모수의 추정을 위해 구해진 통계량이다.

큰 수의 법칙(Law Large Number)

● 데이터를 많이 선택할수록(n이 커질수록) 표본평균의 분산은 0에 가까워진다.

33. 다음 중 클래스 불균형에 대한 설명으로 옳지 않은 것은?
① 불균형 클래스 처리를 위해서 다수 클래스의 데이터 중 일부만 선택하여 사용하는 것을 과소표집이라고 한다.
② 가중치 균형(weight balancing)으로는 불균형 클래스를 처리할 수 없다.
③ 임곗값은 학습 단계에서는 변화 없이 학습하고, 테스트 단계에서 이동한다.
④ 과대표집 기법으로는 SMOTE, ADASYN 등이 있다.

불균형 데이터 처리 방법에는 과소표집, 과대표집, 임곗값 이동, 앙상블 기법, 가중치 균형이 있다. 가중치 균형은 학습 데이터셋의 각 데이터에서 손실(loss)을 계산할 때 특정 클래스의 데이터에 더 큰 손실(loss) 값을 갖도록 하는 방법이다.
불균형 데이터 처리 시 **임곗값 이동**은 데이터 많은 쪽으로 이동시키는 방법으로 학습 단계에서는 변화 없이 학습하고, 테스트 단계에서 임곗값 이동한다.

과대표집(Over-Sampling)
● 소수 클래스의 데이터를 복제 또는 생성하여 데이터 비율을 맞추는 방법이다.
● 과적합 가능성이 존재한다.
● 알고리즘 성능은 높지만, 검증 성능은 나빠질 수 있다.
[기법] 랜덤 과대표집, SMOTE, Borderline-SMOTE, ADASYN(Adaptive Synthetic ampling)

SMOTE(Synthetic Minority Over-Sampling TEchnique)
● 소수 클래스에서 중심이 되는 데이터와 주변 데이터 사이에 가상의 직선을 만들고 그 위에 데이터를 추가하는 방법

ADASYN(Adaptive Synthetic ampling)
● 소수 클래스 데이터와 그 데이터에서 가장 가까운 k개의 소수 클래스 데이터 중 무작위로 선택된 데이터 사이의 직선상에 가상의 소수 클래스 데이터를 만드는 방법이다.

34. 다음 중 기초 통계량에 대한 설명으로 옳지 않은 것은?
① 표준편차는 분산에 양의 제곱근을 취한 값이다.
② 사분편차는 사분위수 범위(IQR)의 절반값이다.
③ 첨도는 데이터 분포의 뾰족한 정도를 나타내는 통계량이다.
④ 사분위수는 3분위수에서 1사분위수를 뺀 값이다.

사분위수(Quartile)
● 사분위수는 모든 데이터를 순서대로 배열했을 때, 4등분한 지점에 있는 값을 의미한다. 3사분위수에서 1사분위수를 뺀 값은 IQR이다.

35. 다음 중 파생변수 사용 예시로 옳지 않은 것은?
① 크루즈 탑승자 명단에서 형제, 부모 데이터를 가족 데이터로 변환하여 사용한다.
② A, B, O, AB 혈액형 데이터를 0, 1, 2, 3으로 변환하여 사용한다.
③ 화장품 업체의 분기별 매출 자료를 총 매출액으로 사용한다.
④ 차량 번호판에서 개인소유 혹은 렌터카 여부를 확인하여 사용한다.

36. 측정된 데이터들을 x축과 y축을 기반으로 점으로 표시한 그래프로, 측정된 데이터의 분포를 통해 변수간의 관계 파악이 가능한 그래프는?
① 점그래프 ② 산점도
③ 버블차트 ④ 네트워크그래프

버블차트(Bubble Chart)
● 산점도는 데이터를 점으로 표현하지만, 버블차트는 데이터의 크기를 추가적으로 고려하여 표현한 그래프이다.

네트워크그래프(Network Graph)
● 서로 연관된 개체들 간의 관계를 표현하는 그래프이다.
● 각 개체들은 선으로 연결되며, 연될된 선의 빈도수 등을 통해 개체 간의 관계를 파악할 수 있다.

37. 다음 중 차원 축소에 대한 설명으로 옳은 것은?
① 데이터가 많고 고차원일수록 모델의 정확도가 높다.
② 선형판별 분석은 다변량의 신호를 통계적으로 독립적인 하부 성분으로 분리하여 차원을 축소하는 기법이다.
③ 차원 축소는 분석에 활용되는 데이터의 변수 정보는 최대한 유지하면서 데이터셋 변수의 개수를 줄이는 데이터 분석 기법이다.
④ 주성분 분석(PCA)은 행과 열의 크기가 다른 임의의 M*N 차원의 행렬에서 특이값을 추출하여 효율적으로 차원을 축소하는 기법이다.

차원 축소(Dimensionality Reduction)의 개념
● 막연히 데이터의 개수가 많다고 하여 정확한 분석 결과을 얻을 수 있는 것은 아니다. 원활한 데이터 분석 작업을 위해서 차원축소 기법을 사용한다.
● 차원축소는 분석에 활용되는 데이터의 변수 정보는 최대한 유지하면서 데이터 셋 변수의 개수를 줄이는 데이터 분석 기법이다.

선형판별 분석(LDA; Linear Discriminant Analysis)
● 데이터를 특정한 직선(축)에 사영(projection)하여 두 범주를 잘 구분할 수 있는 직선을 찾는 기법이다.

독립성분 분석(ICA; Independent Component Analysis)

● 데이터를 가장 잘 설명할수 있는 축을 찾는 주성분 분석(PCA)과 다르게 가장 독립적인 축을 찾는 기법이다.

● 다변량의 신호를 통계적으로 독립적인 하부 성분으로 분리하여 차원을 축소하는 기법이다.

● 비정규 분포를 따르는 데이터들의 관계를 독립적으로 변환시키는 방법이다.

38. 세 학생의 중간고사 성적이 각각 60, 70, 80점이었다. 최소-최대 정규화를 했을 때, 세 학생의 성적의 합은 얼마인가?

① 1.5 ② 1

③ 0.5 ④ 2

최소-최대 정규화는 데이터의 최솟값을 0으로, 최댓값을 1로 설정하고, 중간값을 공식에 의하여 연산한다.

중간값 공식 : $X = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$

(X_i : 정규화 대상 i번째 데이터, $X_{\max}$: 정규화 대상 최대 데이터, $X_{\min}$: 정규화 대상 최소 데이터)

위의 공식에 의거하여 데이터의 최솟값인 60점은 0이 되고, 최댓값인 80점은 1이 되면 중간값인 70은 $\frac{70-60}{80-60} = \frac{10}{20} = 0.5$ 가 되어서 세 학생의 성적의 합은 1.5가 된다.

39. 다음 중 다중회귀 분석의 가정이 아닌 것은?

① 잔차와 독립변수의 독립성

② 잔차와 종속변수의 선형성

③ 잔차의 분산이 독립변수와 무관한 등분산성

④ 잔차항의 정규성

회귀 분석 가정 중 선형성은 잔차와 관련이 없다.

잔차(residual) : 표본(Sample)으로 추정한 회귀식과 실제 관측값의 차이로 각각의 자료가 직선에 얼마나 잘 맞는지 확인하는 도구

회귀 분석은 선형성, 독립성, 등분산성, 정상성(정규성)의 4가지 가정을 만족해야 한다.

● **선형성**

- 독립변수 변화에 따라 종속변수도 선형적인 일정 크기로 변화한다.
- 산점도를 통해 선형성 확인이 가능하다.

● **독립성**

- 잔차와 독립변수의 값이 서로 독립적이야 한다.
- 더빈-왓슨 검정을 통해 통계량 확인이 가능하다.

● **등분산성**

- 잔차의 분산이 독립변수와 무관하게 일정해야 한다.
- 잔차가 고르게 분포되어 있어야 한다.

● **정상성(정규성)**

- 잔차항이 평균 0인 정규분포 형태를 이뤄야 한다.
- 샤피로-윌크 검정, 콜모고로프-스미르노프 검정을 통해 통계량 확인이 가능하다.
- Q-Q plot에서 잔차가 오른쪽으로 치우친 직선 형태의 경우 정규성을 띈다고 할 수 있다.

40. 다음 설명에 해당되는 시스템은?

> • 대규모 데이터를 저장하기 위한 데이터베이스 관리 시스템이다.
> • 고정된 테이블 스키마가 없고, 조인(JOIN) 연산을 사용할 수 없다.
> • 수평적 확장이 가능하다.
> • 활용 예시로는 HBase, Cassandra, MongoDB 등이 있다.

① RDBMS

② MySQL

③ DFS

④ NoSQL

위에서 설명하고 있는 것은 비관계형 데이터베이스(NoSQL)에 대한 내용이다.

RDBMS는 정형 데이터를 저장하는 관계형 데이터베이스를 의미하고, DFS는 비정형 데이터를 저장하는 분산 파일 시스템을 의미한다. MySQL은 RDBMS의 한 종류의 프로그램이다.

3과목 빅데이터 모델링

41. 다음 중 Causality Analysis에 대한 설명으로 옳은 것은?

① 하나 이상의 독립변수가 종속변수에 끼치는 영향을 추정하는 통계 방법이다.

② 두 개 이상의 변수 사이에 존재하는 상호 연관성을 분석하는 방법이다.

③ 독립변수와 종속변수 간의 인과관계를 분석하는 방법이다.

④ 서로 다른 집단의 평균에서 분산값을 비교하여 집단 간의 통계학적 차이를 확인하는 방법이다.

Causality Analysis는 **인과관계 분석**으로 독립변수와 종속변수 간의 인관관계를 분석하는 방법이다. ①은 회귀 분석, ②은 상관분석, ④은 분산 분석에 대한 설명이다.

42. 다음 중 다중공선성을 진단하기 위한 지표는?
① 회귀계수(Regression Coefficient)
② 분산팽창지수(Variance Inflation Factor)
③ 자카드계수(Jaccard)
④ 순위상관계수(Rank Correlation Coeffecient)

다중공선성(Multicollinearity) : 다중공선성은 설명변수들 사이에 선형관계가 존재하게 되면 회귀 계수의 정확한 추정이 어려워지는 것을 의미하며, 이 경우 문제가 있는 변수를 제거하거나 주성분 회귀 모형을 적용하여 문제를 해결할 수 있다. 다중공선성을 진단하기 위해서 **분산팽창지수(VIF)**를 활용한다.
분산팽창지수는 결정계수를 활용하여 독립변수들 간의 상호 연관성을 수치로 분석한다. 분산팽창지수가 10 이상인 경우 해당 독립변수는 독립적인 변수로 역할을 하기 어렵다고 판단한다.

회귀계수(Regression Coefficient)
● 회귀계수는 독립변수가 한 단위 변환함에 따라 종속변수에 미치는 영향력의 크기를 의미하며, 최소제곱법을 사용한다.
● 최소제곱법은 구하려는 값과 실제 값의 오차를 제곱한 합이 최소가 되는 해를 구하는 방법이다.

자카드계수(Jaccard)
● 두 집합 사이의 유사도를 측정하는 방법
● 0과 1사이의 값을 가지며 두 집단이 동일하면 1값을, 공통 원소가 하나도 없는 경우 0값을 가진다.

순위상관계수(Rank Correlation Coeffecient)
● 값에 순위를 매겨 그 순위에 대해 상관계수를 구하는 방법

43. 교차 검증 방법 중 N개 데이터 중 1개만 평가 데이터로 사용하고, 나머지 N-1개는 훈련 데이터로 사용하는 과정을 N번 반복하는 검증 방법은?
① K-fold 교차 검증
② Hold-out 교차 검증
③ LOOCV
④ LpOCV

교차 검증 방법
K-fold 교차 검증
● 학습 데이터를 K개의 그룹(fold)으로 나누어 (K-1)개는 학습에, 나머지 하나는 검증에 사용하는 방법이다.
● 방법 : 테스트 데이터를 제외한 데이터를 무작위로 중복되지 않는 K개의 데이터로 분할 -> K-1 개의 데이터를 학습 데이터로 사용하고, 나머지 1개는 데이터를 검증 데이터로 사용 -> 검증 데이터를 바꾸며 K번 반복해서 분할 데이터가 한 번씩 검증 데이터로 사용된다.
● LOOCV보다 연산량이 작고, 중간 정도의 편향과 분산을 가진다.

Hold-out 교차 검증
● 데이터를 무작위로 7:3 또는 8:2 비율로 학습 데이터와 검증 데이터로 나누는 방법이다.
● 가장 보편적으로 랜덤 추출을 통해 데이터를 분할하는 방법으로 학습, 검증 데이터가 60~80%이고, 테스트 데이터가 20~40%이다.

LpOCV
● 데이터 중 p개의 관측치를 검증 데이터로 사용하고, 나머지는 학습 데이터로 사용하는 방법이다.

44. 다음 중 인공신경망에 대한 설명으로 옳지 않은 것은?
① 머신러닝은 딥러닝의 일부이다.
② 인공신경망은 활성화 함수를 사용하고, 가중치를 알아내는 것이 목적이다.
③ 인공신경망의 활성화 함수는 입력 신호의 총합을 출력 신호로 변환하는 함수이다.
④ 퍼셉트론은 XOR 선형 분리 불가 문제가 발생하여 이를 보완하기 위해 다중 퍼셉트론이 개발되었다.

딥러닝은 머신러닝의 일부이고, 머신러닝은 인공지능의 일부이다.

인공신경망(Artificial Neural Network, ANN)
● 사람 두뇌의 신경세포인 뉴런이 전기신호를 전달하는 모습을 모방한 기계학습 모델이다.
● 인공신경망은 활성화 함수를 사용하고, 가중치를 알아내는 것이 목적이다.

인공신경망의 구조
● 퍼셉트론(Perceptron) : 신경망의 뉴런 모델을 모방하여 입력층과 출력층으로 구성된 최초의 인공신경망 모델이다.
● 다중 퍼셉트론 : 퍼셉트론은 XOR 선형 분리 불가 문제가 발생하여, 이를 보완하기 위해서 다중 퍼셉트론이 개발이 되었다. 다중 퍼셉트론은 입력층, 은닉층, 출력층으로 이루어져 있고, 활성화 함수로 시그모이드 함수를 사용한다.

45. 다음과 같은 분할표에서 흡연 여부에 따른 폐암 발생률에 대한 오즈비는 얼마인가?

구분	폐암 발생	폐암 미발생	합계
흡연	6	5	11
비흡연	2	10	12
합계	8	15	23

① 8 ② 4 ③ 10 ④ 6

오즈비(Odds Ratio)는 특정 사건이 발생할 확률(p)과 그 사건이 발생하지 않을 확률(1-p)의 비를 의미한다.

오즈비의 공식 : $\frac{A\text{집단 사건 발생 확률}}{B\text{집단 사건 발생 확률}} = \frac{ad}{bc}$ 와 같이 연산된다.

따라서 흡연 여부에 따른 폐암 발생률에 대한 오즈비는 $\frac{6*10}{5*2} = \frac{60}{10} = 6$이 된다.

46. 다음 중 분석 모형 구축 절차로 옳은 것은?

① 비즈니스 영향도 평가 → 유의변수 도출 → 분석요건 확정 → 운영시스템 적용

② 유의변수 도출 → 비즈니스 영향도 평가 → 분석요건 확정 → 운영시스템 적용

③ 분석요건 확정 → 유의변수 도출 → 비즈니스 영향도 평가 → 운영시스템 적용

④ 비즈니스 영향도 평가 → 분석요건 확정 → 운영시스템 적용 → 유의변수 도출

분석 모형 구축 절차는 **분석요건 확정 → 유의변수 도출 → 비즈니스 영향도 평가 → 운영시스템 적용**이다. 분석요건 확정은 요건 정의에 해당하고, 유의변수 도출은 모델링에 해당된다. 비즈니스 평가는 검증 및 테스트에 해당되고, 운영 시스템 적용은 적용에 해당된다.

47. 다음 중 시계열 데이터의 장기의존성 문제에 대한 LSTM기법을 보완한 방법은?

① SMOTE ② LOF

③ SEMMA ④ GRU

시계열 데이터의 장기의존성 문제에 대한 해결책인 LSTM(장단기 메모리기법)의 복잡한 연산 구조를 보완한 방법은 게이트 순환 유닛(GRU; Gated Recurrent Unit)이다. GRU는 LSTM의 장기 의존성 문제에 대한 해결책은 유지하면서 은닉 상태를 업데이트 하는 계산량을 줄였다.

SMOTE는 과대표집 기법 중 하나이다.

LOF(Local Outlier Factor)는 전체 데이터 분포에서 지역적인 밀집도(Density)를 고려하여 이상값을 확인하는 방법이다.

SEMMA는 분석 솔루션 업체 SAS사가 주도한 통계중심의 분석 방법론이다.

48. 다음 중 앙상블 분석에 대한 설명으로 옳지 않은 것은?

① 앙상블 분석 방법에는 배깅, 부스팅, 랜덤 포레스트, 보팅, 스태킹이 있다.

② 배깅(Bagging)은 데이터 사이즈가 크거나 결측값이 없는 경우에 사용하기 유리하다.

③ 부스팅(Boosting)의 알고리즘에는 AdaBoost, GBM, XGBoost 이 있다.

④ 간접투표(Soft Voting)는 각 모형의 클래스 확률값을 평균 내어 확률이 가장 높은 클래스를 최종 결과로 예측하는 방법이다.

배깅(Bagging)은 사이즈가 작거나 결측값이 있는 경우에 사용하기 유리하고, 성능 향상에 효과적인 특징이 있다.

앙상블 분석 방법에는 배깅, 부스팅, 랜덤 포레스트, 보팅, 스태킹이 있다.

배깅(Bagging)

- 부트스트랩(Bootstrap) 샘플링으로 추출한 여러 개의 표본에 각각 모형을 병렬적으로 학습하고, 추출된 결과를 집계하는 기법이다.
- 배깅(Bagging)은 사이즈가 작거나 결측값이 있는 경우에 사용하기 유리하고, 성능 향상에 효과적인 특징이 있다.

부스팅(Boosting)

- 예측력이 약한 모형들을 결합하여 예측력이 강한 모형을 만드는 알고리즘으로 분류가 잘못된 데이터에 가중치를 적용하여 표본을 추출하는 기법이다.
- 대용량 데이터 분석에 유리하고, 높은 계산 복잡도를 가진다.
- 알고리즘 : AdaBoost, GBM, XGBoost

보팅(Voting)

- 여러 개의 분석 모형 결과를 조합하는 방법이다.
- 직접투표와 간접투표가 있다.

직접투표(Hard Voting) : 많이 선택된 클래스를 최종 결과로 예측한다.

간접투표(Soft Voting) : 각 모형의 클래스 확률값을 평균 내어 확률이 가장 높은 클래스를 최종 결과로 예측하는 방법이다.

49. 다음 중 기계학습과 통계분석에 대한 설명으로 옳지 않은 것은?

① 기계학습은 다양한 알고리즘을 활용한 학습 방법을 의미한다.

② 통계분석은 다양한 통계량을 활용한 분석방법으로 분석 결과를 시각화하여 표현할 수 있다.

③ 기계학습은 통계분석과 다르게 결과물에 대한 수식을 도출할 수 없다.

④ 기계학습을 위한 알고리즘 선정은 분석 대상에 따라 다르게 설정된다.

기계학습 역시 결과물에 대한 수식을 도출할 수 있다.

50. 다음 중 데이터 분할(split) 방법에 대한 설명으로 옳지 않은 것은?
① 데이터가 충분하지 않은 경우에는 학습 데이터와 검증 데이터로만 분할하여 분석하기도 한다.
② 훈련 데이터셋으로 학습한다.
③ 검증 데이터는 하이퍼파라미터의 성능을 평가하는 데 사용된다.
④ 테스트 데이터셋으로 성능을 확인한다.
데이터가 충분하지 않은 경우에는 학습 데이터와 평가 데이터로만 분할하여 분석하기도 한다.
데이터 분할(Data Split)
● 데이터는 분석되기 전 목적에 맞게 분할되어야 하는데, 이는 분석 모형의 과적합을 방지하고, 일반화 성능을 향상시키기 위함이다.
● 일반적으로 데이터는 학습(훈련) 데이터, 검증 데이터, 평가(테스트) 데이터로 나뉜다.
● 보통의 경우 학습 데이터와 검증 데이터를 60~80%로 사용하고, 평가 데이터를 20~40%로 사용하지만 절대적인 수치는 아니다.
● 데이터가 충분하지 않은 경우에는 학습 데이터와 평가 데이터로만 분할하여 분석하기도 한다.

51. 다음 중 과적합 방지 방법이 아닌 것은?
① 데이터 삭제 ② LASSO
③ 데이터 증강 ④ Drop Out
과적합 방지 방법에는 데이터 증강, 모델의 복잡도 감소, 가중치 규제 적용, 드롭 아웃이 있다.
● 데이터 증강(Data Augmentation)
- 데이터의 개수가 적을 경우 지나치게 세세한 학습이 진행될 수 있기 때문에 과적합을 유발할 수 있어 데이터를 증강시켜 데이터 분석을 위한 충분한 데이터셋을 확보해야 한다.
- 데이터의 양이 적을 경우 데이터 변형, 데이터 표집 등의 방법을 활용하여 데이터의 수를 늘릴 수 있다.
● 모델의 복잡도 감소
- 모델의 복잡도가 높은 경우 데이터 과대적합의 위험이 있다.
- 이 경우 모델의 복잡도와 관련되는 인공신경망은 은닉층 수 감소, 매개변수의 수 조절 등의 방법으로 모델의 복잡도를 감소시킬 수 있다.
● 가중치 규제
- 가중치 규제(Weight Regularization)란 가중치의 값을 제한하여 모델의 복잡도를 간단하게 만드는 것을 의미한다.
- 가중치 규제의 종류에는 라쏘(LASSO, L1 노름 규제), 릿지(Ridge, L2 노름 규재), 엘라스틱 넷이 있다.
● 드롭 아웃
- 드롭 아웃은 학습 과정에서 신경망 일부를 사용하지 않는 방법이다.
- 드롭 아웃은 서로 연결된 연결망에서 0~1 사이의 확률로 뉴런을 제거하는 방법이다.
- 드롭 아웃은 신경망 학습 시에만 사용하고, 예측 시에는 사용하지 않는다.
- 드롭 아웃의 유형에는 초기, 공간적, 시간적 드롭 아웃이 있다.

52. 다음 중 랜덤 포레스트에 대한 설명으로 옳지 않은 것은?
① 랜덤 포레스트는 의사결정나무 기반 앙상블 알고리즘이다.
② 이상치의 영향을 적게 받는다.
③ 분류기를 여러 개 사용할수록 예측편향이 줄어든다.
④ 랜덤 포레스트 모형에서는 모든 변수(Feature)를 학습시킨다.
랜덤 포레스트(Random Forest)
● 의사결정나무 기반 앙상블 알고리즘으로 모든 속성(Feature)들에서 임의로 일부를 선택하고, 그 중에서 정보 획득량이 가장 높은 것을 기준으로 데이터를 분할한다.
● 분류기를 여러 개 사용할수록 성능이 좋아지고, 예측편향을 줄이고, 과대적합을 피할 수 있으며, 이상치에 영향을 적게 받는다.

53. 다음 중 변수의 성질이 다른 하나는?
① 결과변수
② 회귀변수
③ 실험변수
④ 통제변수
독립변수(X)와 같은 표현 : 설명변수, 원인변수, 예측변수, 실험변수, 회귀변수, 통제변수, 조작변수, 노출변수
종속변수(Y)와 같은 표현 : 반응변수, 결과변수, 목표변수, 준거변수

54. 다음 중 종속변수가 범주형일 때 사용되는 분석 기법이 아닌 것은?
① 판별 분석
② KNN
③ 다중선형 회귀 분석
④ 로지스틱 회귀 분석

다중선형 회귀 분석을 포함하는 회귀 분석은 독립변수가 연속형 혹은 범주형이고, 종속변수가 연속형일 때 사용된다.

독립변수와 종속변수에 따른 분석기법

독립변수(연속형), 종속변수(연속형) : 회귀분석, 인공신경망 모델, KNN, 의사결정나무(회귀)

독립변수(연속형), 종속변수(범주형) : 로지스틱 회귀 분석, 판별 분석, KNN, 의사결정나무(분류)

독립변수(연속형), 종속변수 없음 : 주성분 분석, 군집 분석

독립변수(범주형), 종속변수(연속형) : 회귀분석, 인공신경망 모델, 의사결정나무(회귀)

독립변수(범주형), 종속변수(범주형) : 로지스틱 회귀 분석, 인공신경망 모델, 의사결정나무(분류)

독립변수(범주형), 종속변수 없음 : 연관 분석, 판별 분석

55. 다음 중 다중선형 회귀 모형의 평가 지표는?

① ROC 곡선

② 결정계수(R^2)

③ 정밀도

④ 재현율

회귀모형 평가지표 : 평균절대오차(MAE), 평균제곱오차(MSE), 평균제곱근오차(RMSE), 평균절대백분율오차(MAPE), 결정계수(R^2)가 있다.

ROC곡선, 정밀도, 재현율은 분류모형 평가 지표에 해당된다.

회귀 모형 평가 지표

평균절대오차(MAE, Mean Absolute Error)

● 모델의 실젯값과 예측값 차이에 절댓값을 취하여 평균한 값

● 수식 : $\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}|$

평균제곱오차(MSE; Mean Squared Error)

● 모델의 실젯값과 예측값 차이를 제곱하여 평균한 값

● 수식 : $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y_i})^2$

평균제곱근오차(RMSE; Root Mean Squared Error)

● 평균제곱오차(MSE) 제곱근을 씌운 값

● MSE는 값을 커지는 경향이 있으므로 제곱근을 씌운 RMSE를 실무에서 일반적으로 사용한다.

● 수식 : $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$

평균절대백분율오차(MAPE; Mean Absolute Percentage Error)

● 평균절대오차(MAE)를 퍼센트로 변환한 값

● 다른 변수 사이의 오차를 비교할 수 있다.

● 수식 : $\frac{100}{n}*\sum_{i=1}^{n}|\frac{y_i-\hat{y}}{y_i}|$

결정계수(Coefficient of Determination, R^2)

● 결정계수는 선형회귀 모형의 성능 검증 지표로 많이 사용되고, 회귀 모형의 예측값이 실젯값과 얼마나 유사한지를 나타내는 지표이다.

● 결정계수는 0~1의 범위를 갖고, 결정계수 값이 1에 가까울수록 모형의 설명력이 높다고 할 수 있다.

● 결정계수의 수식 : $R^2=\frac{\text{회귀제곱합}}{\text{전체제곱합}}=\frac{SSR}{SST}=\frac{SSR}{SSR+SSE}=(1-\frac{SSE}{SST})$

56. 다음 중 시계열 데이터의 공분산 기법은?

① 연관 분석

② 계절성 분석

③ 추세 분석

④ 자기상관 분석

시계열 데이터 공분산 기법

● 시계열 데이터의 공분산 기법으로는 자기상관(autocorrelation)이 있다.

● 상관계수가 두 변수 사이의 선형 관계의 크기를 측정하는 것과 같이 자기상관은 시계열 데이터의 시차값(logged values) 사이의 선형 관계를 측정한다.

● 자기상관계수는 동일한 변수의 서로 다른 시간 차이(time lag)를 두고 관계를 분석하는 것이다.

● 자기상관함수(ACF)는 임의의 어떤 신호와 그 신호를 임의의 시간(t)만큼 지연시킨 신호(t+t) 사이의 상관관계를 파악할 수 있는 함수이다.

● 데이터에 추세가 존재할 때 자기상관함수는 양의 값을 갖는 경향을 보이고, 이러한 자기상관함수 값은 시차가 증가함에 따라 서서히 감소한다.

연관 분석, 계절성 분석, 추세 분석, 불규칙 요인은 시계열 분해 구성 요소에 속한다.

57. 다음 중 시계열 데이터 예측 방법에 대한 설명으로 옳지 않은 것은?

① 시계열 데이터 예측 방법은 확률적 방법과 고전적 방법으로 나뉜다.

② 지수평활법은 과거 값에 가중치를 두고, 최근 값에 적은 비중을 두는 방법이다.

③ 이동평균법은 일정 기간의 관측치를 이용하여 평균을 구하고, 이를 이용해 예측하는 방법이다.

④ 확률적 방법은 주파수 영역과 시간 영역으로 나뉜다.

시계열 데이터 예측 방법

● 시계열 데이터의 예측 방법은 확률적 방법과 고전적 방법으로 나뉜다.

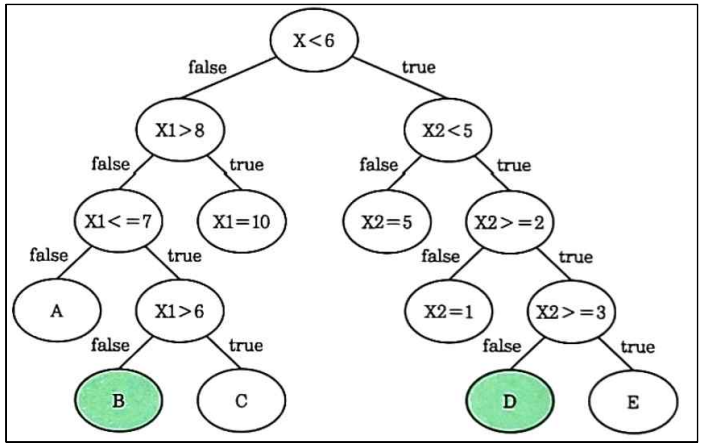
● 확률적 방법은 주파수 영역과 시간 영역으로 나뉘고, 시간 영역에는 자기회귀 모형, 이동평균 모형, 자기회귀이동평균 모형이 있다.

● 고전적 방법은 분해분석법과 평활법으로 나뉘고, 평활법에는 이동평균법(MA)과 지수평활법이 있다.

● 이동평균법은 일정 기간의 관측치를 이용하여 평균을 구하고, 이를 이용해 예측하는 방법으로 장기적인 추세를 쉽게 파악할 수 있다.

● 지수평활법은 일정기간의 평균을 활용하는 이동평균법과는 다르게 모든 시계열 데이터를 사용하여 평균을 구하고, 시간의 흐름에 따라 최근 시계열 데이터에 더 많은 가중치를 부여하여 미래를 예측하는 방법이다.

58. 다음과 같은 의사결정나무에서 B에 해당하는 X1 값과 D에 해당하는 X2값은?



	B(X1)	D(X2)
①	6	1
②	8	0
③	6	2
④	7	2

그림의 의사결정나무의 값을 연산하면 B(X1)은 최종적으로 $6 \leq X1 < 7$의 범위를 갖게 되므로 6이 되고, D(X2)는 최종적으로 $2 \leq X2 < 3$의 범위를 갖게 되어 2가 된다.

59. 다음 중 ReLU 함수의 뉴런이 죽는 현상(Dying ReLU)을 해결한 활성화 함수는?

① Sigmoid ② tanh

③ Leaky ReLU ④ Softmax

ReLU 함수의 뉴런이 죽는 현상(Dying ReLU)을 해결한 활성화 함수는 Leaky ReLU 함수이다.

활성화 함수

● 활성화 함수(Activation Function)는 입력 신호의 총합을 출력 신호로 변화해주는 함수이다.

① 시그모이드(Sigmoid) 함수

● 로지스틱 회귀 함수의 로짓 변환을 한 형태이다.

● 기울기 소실 문제의 원인이 된다.

② tanh(Hyperbolic Tangent Function)

● 시그모이드 함수의 확장된 형태이다.

● 시그모이드보다 학습 속도가 빠르다.

③ ReLU 함수

● 양수 입력 시 어떠한 값의 변형 없이 입력값 그대로 출력하고, 음수 입력 시 항상 0값을 리턴하는 함수이다.

● 시그모이드 함수의 기울기 소실 문제를 해결한다.

● 상대적으로 가중치 업데이트 속도가 빠르다.

④ Leaky ReLU

● 임계치보다 작을 때 0을 곱하는 ReLU와 달리 0.01을 곱한다.

● ReLU 함수의 뉴런이 죽는 현상(Dying ReLU)을 해결한다.

⑤ 소프트맥스 함수(Softmax Function)

● 세 개 이상으로 분류하는 다중 클래스 분류 모델을 만들 때 사용된다.

● 분류될 클래스의 개수가 n인 경우 n차원의 벡터를 입력받아 각 클래스에 속할 확률을 추정한다.

● 입력받은 값을 출력할 때 0~1 사이의 값으로 모두 정규화하며, 출력값의 총합은 항상 1이 된다.

60. 다음과 같은 분석 방법에 해당하는 것은?

• 독립변수가 종속변수에 얼마나 부정적인(−) 혹은 긍정적인(+) 영향을 주는지 확인하는 분석 방법으로 주로 의료통계 분야에서 많이 사용된다. • 종속변수(Y)가 이진 형태(남성 또는 여성, 성공 또는 실패, 증가 또는 감소)여야 하고, 독립변수(X)는 연속형 또는 범주형일 수 있다.

① 비선형 회귀 분석
② 다중선형 회귀 분석
③ 로지스틱 회귀 분석
④ 이항 로지스틱 회귀 분석
다중선형 회귀 분석 : 독립변수가 k개이고, 종속변수와의 관계가 선형인 경우(1차 함수)
로지스틱 회귀 분석
● 독립변수가 수치형이고, 반응변수(종속변수)가 범주형일 때 사용되는 분석 모형이다.
● 어떤 사건이 발생할지에 대한 직접적인 예측이 아닌 그 사건이 발생할 확률을 예측하는 방법이다.

4과목 빅데이터 결과 해석

61. 다음 중 K-fold 교차 검증 학습 과정에 대한 설명으로 옳지 않은 것은?
① 데이터 학습과 검증 과정에서 테스트 데이터는 사용되지 않는다.
② K-1개의 검증 데이터를 만들고, 1개의 훈련 데이터를 만들어서 학습한다.
③ 데이터를 학습, 검증, 테스트 데이터로 나누어 교차 검증하는 방법이다.
④ 검증 데이터를 계속 바꾸어 사용하기 때문에 분할된 데이터는 한 번씩 검증 데이터로 사용된다.
(K-1)개의 학습에 사용하고, 나머지 1개는 검증에 사용하는 방법이다.

62. 다음 중 시간 시각화에 대한 설명으로 옳지 않은 것은?
① 시간 시각화는 시간의 흐름에 따른 데이터의 변화를 나타낸 것을 의미한다.
② 추세선은 일정 기간 동안 측정된 데이터들의 경향성을 보여주는 직선 또는 곡선이다.
③ 일반적으로 y축은 시간을, x축은 데이터 값을 나타낸다.
④ 점그래프의 점들을 선으로 연결하면 선그래프로 표현할 수 있다.
시간 시각화 자료는 일반적으로 x축은 시간을, y축은 데이터 값(value)을 나타내고, 시계열 데이터를 통한 데이터의 경향성과 흐름을 파악하는 것이 목적이다.

63. 다음 설명에 해당하는 분석 방법은?

• 비계층적 군집분석 방법 중 하나로, 군집의 수를 지정하지 않아도 된다. • 밀도를 기반으로 군집을 이루기 때문에 기하학적인 모양의 군집도 찾을 수 있고, 이상값을 검출할 수 있다.

① K-means clustering
② DBSCAN
③ SOM
④ SVM
비계층적 군집 분석의 종류
K-평균 군집 분석(K-means clustering), 밀도 기반 군집 분석(DBSCAN; Density-based Spatial Clustering of application with noise), 자기 조직화 지도(SOM)가 있다.
K-평균 군집 분석(K-means clustering)
● 주어진 데이터를 K개의 군집으로 묶는 알고리즘으로 군집 수를 K개 만큼 초깃값으로 지정하고, 각 객체를 가까운 초깃값에 할당하여 군집을 형성하는 방뻡이다.
각 군집의 평균을 재계산하여 초깃값을 갱신하는 과정을 반복하여 K개의 최종 군집을 형성한다.
● K-평균 군집 분석은 이상값에 민감하게 반응하는 단점이 있고, 이를 보완하는 방법으로는 K-중앙값 군집 사용 및 이상값 제거가 있다.
자기 조직화 지도(SOM; Self-Organizing Maps, 코호넨 네트워크)
● 대뇌 피질과 시각 피질의 학습 과정을 기반으로 모델화한 인공신경망으로 자율학습 방법에 의 클러스터링 방법을 적용한 알고리즘이다.
서포트 벡터 머신(SVM; Support Vector Machine)
● 두 집단의 데이터를 분리해주는 가장 적합한 경계선(결정 경계)을 찾아주는 지도학습 기반의 이진 선형 분류기이다.
● 마진(margin, 여유 공간)을 최대화하는 것을 목표로 한다.

64. 다음 중 파라미터 최적화 방법으로 옳지 않은 것은?
① 손실함수 최소화
② AdaGrad
③ 확률적 경사하강법(SGD)
④ 베이지안 최적화(Bayseian Optimization)
베이지안 최적화는 초매개변수 최적화 방법이다. 초매개변수 최적화 방법으로는 매뉴얼 탐색, 그리드 탐색, 랜덤 탐색, 베이지안 최적화가 있다.
매개변수(parameter) 최적화 기법에는 경사하강법(배치 경사하강법, 확률적 경사하강법, 미니 배치 경사하강법), 모멘텀, 네스테로프 모멘텀, AdaGrad(Adaptive Gradient), RMSProp, Adam 이 있다.

65. 다음 중 ROC 곡선에 대한 설명으로 옳지 않은 것은?
① ROC 곡선의 x축은 1-Specificity이고, y축은 Sensitivity이다.
② ROC 곡선은 항상 0.5 이상의 값을 갖는다.
③ ROC 곡선은 가능한 모든 임곗값에 대한 참 긍정률과 거짓 긍정률을 확인한다.
④ ROC 곡선은 회귀 모형 평가 지표이다.
ROC 곡선은 분류 모형 평가 지표이다.

66. 다음 중 혼동행렬(Confusion Matrix)에 대한 설명으로 옳지 않은 것은?
① TPR은 $\frac{TP}{TP+FN}$와 같이 연산된다.
② F1-Score는 정밀도와 재현율의 기하평균이다.
③ Specificity는 실제 '부정' 범주 중 '부정'의 비율이다.
④ Precision은 $\frac{TP}{TP+FP}$ 와 같이 연산된다.
F-Measure(F1-Score) : 0~1 사이의 범위를 가짐
공식 : $2*\frac{Precision * Recall}{Precision + Recall}$

67. 다음 중 특정 사건 혹은 주제에 대한 정보를 이야기 들려주듯이 표현하는 인포그래픽 종류는?
① 비교분석형
② 만화형
③ 스토리텔링형
④ 타임라인형

68. 다음 중 역사적 사건이나 특정 주제와 관련된 히스토리를 시간 순서 형식으로 표현한 것으로 기업의 발전과정을 표현할 때 사용되는 인포그래픽 유형은?
① 통계
② 프로세스
③ 도표
④ 타임라인

69. 다음 중 스타차트에 대한 설명으로 옳지 않은 것은?
① 스타차트의 중요도는 별의 개수로 확인할 수 있다.
② 스타차트는 비교 시각화 유형에 속한다.
③ 스타차트의 축은 3개 이상이다.
④ 스타차트로 데이터의 이상값을 확인할 수 있다.
스타차트의 중요도는 별의 개수가 아닌 중심점에서 축까지 이어진 변수들의 값을 연결한 영역을 통해 확인할 수 있다.

70. 다음과 같은 실젯값과 예측값 데이터가 있을 때 평균제곱오차(RMSE)는?

실젯값	10	20	15	8
예측값	8	18	13	6

① 1 ② 2
③ 3 ④ 4
RMSE는 모델의 실젯값과 예측값 차이를 제곱하여 평균한 값에 제곱근을 씌운 값
공식 : $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$
$\sqrt{\frac{(10-8)^2+(20-18)^2+(15-13)^2+(8-6)^2}{4}}=\sqrt{\frac{16}{4}}=2$ 가 된다.

71. 다음 중 ( ) 안에 알맞은 것은?

> • (㉠)은 학습 알고리즘에서 잘못된 가정으로 인한 오류를 의미하고, (㉡)은 학습 데이터의 내재된 작은 변동으로 발생하는 오차를 의미한다.
> • 이상적인 분석 모형은 낮은 (㉠)과 낮은 (㉡)으로 설정되어야 한다.

	㉠	㉡
①	오차	편향
②	잔차	분산
③	분산	편향
④	편향	분산

72. 다음과 같은 혼동행렬에서 정밀도는 얼마인가?

		예측 범줏값	
		Predicted Positive	Predicted Negative
실제 범줏값	Actual Positive	50	150
	Actual Negative	60	140

① 0.54 ② 0.45
③ 0.25 ④ 0.75

혼동행렬에서의 정밀도는 예측 '긍정' 범주 중 '긍정'의 비율을 의미한다.

정밀도(Precision) 공식 : $\frac{TP}{TP+FP} = \frac{50}{50+60} = \frac{50}{110} = \frac{5}{11} = 0.4545...$

정답은 0.45가 된다.

73. 다음 중 비즈니스 기여도 평가 기법에 대한 설명으로 옳지 않은 것은?

① 순 현재 가치(NPV)는 투자로부터 유입되는 미래 현금의 현재 가치와 해당 투자를 위해 투입된 비용의 차액으로 미래 시점의 순이익 규모이다.

② 투자대비효과(ROI)는 $\frac{\text{순이익}}{\text{투자비용}} * 100$ 으로 계산된다.

③ 투자회수기간(PP)은 누적투자금액과 매출금액의 합이 같아지는 기간으로 투자에 소요되는 모든 비용을 회수하는 데 걸리는 기간으로 보통 월(month) 단위로 기록한다.

④ 내부수익률(IRR)은 순 현재 가치를 '0'으로 만드는 할인율이다.

투자회수기간(PP)은 누적투자금액과 매출금액의 합이 같아지는 기간으로 투자에 소요되는 모든 비용을 회수하는 데 걸리는 기간으로 보통 연(year) 단위로 기록한다.

74. 다음 중 시간 시각화 유형에 속하지 않는 그래프는?

① 선그래프 ② 히스토그램
③ 계단식그래프 ④ 막대그래프

히스토그램은 관계시각화 유형에 속한다.

시간 시각화 유형 : 막대그래프, 누적막대그래프, 점그래프, 선그래프, 영역차트, 계단식 그래프, 추세선

공간 시각화 유형 : 등치지역도, 등치선도, 도트맵, 버블맵, 카토그램

분포 시각화 유형 : 파이차트, 도넛차트, 트리맵, 누적영역그래프

관계 시각화 유형 : 산점도, 산점도 행렬, 버블차트, 히스토그램, 네트워크그래프

비교 시각화 유형 : 플로팅 바 차트, 히트맵, 체르노프페이스, 스타차트, 평행좌표 그래프

75. 정밀도가 80%이고, 재현율이 90%일 때 F1-Score는 얼마인가?

① 80.2% ② 83.1% ③ 84.7% ④ 85.3%

F1-Score의 공식 : $2 * \frac{\text{정밀도}(Precision) * \text{재현율}(Recall)}{\text{정밀도}(Precision) + \text{재현율}(Recall)} = 2 * \frac{0.8*0.9}{0.8+0.9} = 2 * \frac{0.72}{1.7} = 2 * 0.423... = 0.847 = 84.7\%$

76. 다음 중 실젯값과 가장 오차가 작은 가설 함수를 도출하기 위해 사용되는 함수는?

① 손실 함수 ② 비용 함수
③ 활성화 함수 ④ 확률밀도함수

77. 다음 중 교차 검증에 대한 설명으로 옳지 않은 것은?

① Hold-Out 교차 검증은 가장 보편적으로 랜덤추출을 통해 데이터를 분할하는 방법으로 학습 데이터와 검증 데이터가 20~40%이고, 테스트 데이터가 60~80% 이다.

② Bootstrap은 주어진 자료에서 단순 랜덤 복원추출 방법을 활용해 동일한 크기의 표본을 여러 개 생성하는 방법이다.

③ LOOCV는 N개 데이터 중 1개만 평가 데이터로 사용하고, 나머지 N-1개는 훈련 데이터로 사용하는 과정을 N번 반복하는 방법이다.

④ K-fold 교차 검증은 데이터를 K개의 fold로 나누어 (K-1)개는 학습에, 나머지 하나는 검증에 사용하는 방법이다.

Hold-Out 교차 검증은 가장 보편적으로 랜덤추출을 통해 데이터를 분할하는 방법으로 학습 데이터와 검증 데이터가 60~80%이고, 테스트 데이터가 20~40% 이다.

78. CNN에서 원본 이미지가 3×3, stride가 2, 필터가 5×5, padding의 크기가 2일 때 Feature Map은 얼마인가?

① (4, 4)　② (3, 3)

③ (1, 1)　④ (2, 2)

CNN Feature Map 계산

● 스트라이드(지정된 간격으로 필터를 순회하는 간격)가 적용되었을 때, 원본 이미지의 크기가 n * n, 스트라이드가 s, 패딩이 p, 필터가 f * f 일 때, 피쳐맵의 크기는 아래 공식과 같이 계산된다.

● 공식 : $Feature\ Map = (\frac{n+sp-f}{s}+1, \frac{n+sp-f}{s}+1) =$

$(\frac{3+4-5}{2}+1, \frac{3+4-5}{2}+1) = (2,2)$ 가 된다.

79. 다음 설명에 해당하는 오류는?

> 분석 모형을 만들 때 주어진 데이터의 특성이 지나치게 반영되어 발생하는 오류를 의미하고, 이를 과대적합(Over-Fitting)되었다고 표현한다.

① 분석 오류　② 가정 오류

③ 일반화 오류　④ 학습 오류

80. 다음 중 드롭 아웃(DropOut)에 대한 설명으로 옳지 않은 것은?

① 드롭 아웃은 학습과정에서 신경망의 일부를 사용하지 않는 기법이다.

② 제거되는 신경망의 종류와 개수는 랜덤하게 드롭 아웃 확률에 의해 결정된다.

③ 드롭 아웃은 서로 연결된 연결망에서 0~1사이의 확률(Drop Out Rate)로 뉴런을 제거하는 방법이다.

④ 드롭 아웃은 신경망 예측 시에 사용하고, 학습 시에는 사용하지 않는다.

드롭 아웃은 신경망 학습 시에 사용하고, 예측 시에는 사용하지 않는다.