3과목.빅데이터 모델링

(Ch_01. 분석 모형 설계 - SEC 01. 분석 절차 수립, SEC 02. 분석 환경 구축)

빅데이터 분석 기사(3과목. 빅데이터 모델링)

CHAPTER 1. 분석 모형 설계

CHAPTER 2. 분석기법 적용

분석 모형 설계

분석 모형 설계 챕터는 총 2개의 작은 섹션으로 구성된다.

1. 분석 절차 수립
2. 분석 환경 구축

3. 분석 모형 설계 – 분석 절차 수립

01 분석 모형 선정

- 분석 모형 선정은 분석 목적과 수집된 데이터의 변수들을 고려하여 적합한 빅데이터 분석 모형을 선정하는 것이다.
- 통계, 데이터 마이닝, 머신러닝 기반 분석 모형 선정 방법을 고려하여 적절한 데이터 분석 모델을 선정한다.

1) 통계기반 분석 모형 선정

- 통계분석은 수치화된 자료를 분석하여 사회현상을 예측하고 이해하고자 할 때 사용된다.
- 통계기반 분석 모형에는 기술 통계, 추론 통계, 상관 분석, 회귀 분석, 인과관계 분석, 분산 분석, 주성분 분석 등이 있다.

① 기술 통계(Descriptive Statistic)

▶ 데이터의 특징을 파악하기 위해 평균, 분산, 표준편차 등의 기초통계량을 구하거나 시각화 도구인 그래프를 활용하는 분석 방법이다.

② 추론 통계(Inferential Statistic)

▶ 모집단에서 추출된 표본으로부터 모수와 관련된 통계량들의 값을 계산하고, 이것을 이용하여 모집단의 특성을 알아내는 방법이다.

3. 분석 모형 설계 – 분석 절차 수립

1) 통계기반 분석 모형 선정

③ 상관분석(Correlation Analysis)

▶ 두 개 이상의 변수 사이에 존재하는 상호 연관성을 분석하는 방법으로 상관계수(r)를 이용하여 상관관계를 분석한다.

④ 회귀 분석(Regression Analysis)

▶ 하나 이상의 독립변수(X)가 종속변수(Y)에 끼치는 영향을 수치적으로 추정하는 통계 방법이다.

예) 흡연량에 따른 폐암 발병률 연구

⑤ 인과관계 분석(Causality Analysis)

▶ 독립변수(X)와 종속변수(Y) 간의 인과관계를 분석하는 방법이다.

⑥ 분산 분석(ANOVA, Analysis of Variance)

▶ 서로 다른 집단의 평균에서 분산값(총 평균과 각 집단 간의 평균 차이에 의해 생긴 분산)을 비교하여 집단 간의 통계학적 차이를 확인하는 방법이다.

⑦ 주성분 분석(PCA, Principal Component Analysis)

▶ 기존 데이터의 분포를 최대한 보존하면서 고차원 공간의 데이터들을 저차원 공간으로 변환하는 방법이다.

3. 분석 모형 설계 – 분석 절차 수립

2) 데이터 마이닝 기반 분석 모형 선정

● 데이터 마이닝(Data Mining)이란 많은 양의 데이터 속에서 데이터의 패턴, 규칙 등을 탐색하고, 통계기법을 활용하여 분석한 뒤, 이러한 분석을 기반으로 가치 있는 정보를 추출하는 과정을 의미한다.

● 데이터 마이닝 기반 분석 모델에는 분류 모델, 예측 모델, 군집화 모델, 연관규칙 모델이 있다.

① 분류 모델(Classification)

▶ 다수의 속성을 갖는 객체들을 사전에 정해진 그룹 중 하나로 분류하는 기법이다.

예) 통계적 기법, 트리 기반 기법, 최적화 기법, 기계학습 모델

② 예측 모델(Prediction)

▶ 과거 데이터로부터 데이터의 특성을 분석하여 다른 데이터의 결과값을 예측하는 기법이다.

예) 회귀 분석, 의사결정나무, 시계열 분석, 인공신경망

의사결정나무(Decision Tree) : 데이터를 분석하여 이들 사이에 존재하는 패턴을 예측 가능한 규칙들의 조합으로 나타내는 모형으로 그 모양이 나무와 비슷하여 의사결정 나무라고 부른다.
인공신경망(Artificial Neural Network, ANN) : 사람 두뇌의 신경세포인 뉴런이 전기신호를 전달하는 모습을 모방한 기계 학습 모델이다. 인공신경망은 활성화 함수를 사용하고, 가중치를 알아내는 것이 목적이다.

3. 분석 모형 설계 – 분석 절차 수립

2) 데이터 마이닝 기반 분석 모형 선정

③ 군집화 모델(Clustering)

▶ 관측된 여러 개의 변숫값에서 유사한 성격을 갖는 몇 개의 군집으로 그룹화하여 그룹들 사이의 관계를 분석하는 다변량 분석기법이다.

예) 계층적 방법 : 병합적 방법, 분할적 방법

비계층적 방법: K-평균군집

④ 연관규칙 모델(Association Rule)

▶ 주어지는 데이터에서 동시에 발생하는 사건 혹은 항목 간의 규칙을 수치화하는 기법으로 장바구니 분석이라고도 하며, 주로 마케팅 분야에서 활용된다.

예) 우유를 구입한 고객이 식빵을 함께 구입한 경우

K-평균 군집 분석(K-means clustering) : 주어진 데이터를 K개의 군집으로 묶는 알고리즘으로 군집수를 K개큼 초깃값으로 지정하고, 각 객체를 가까운 초깃값에 할당하여 군집을 형성하는 방법이다. 각 군집의 평균을 재계산하여 초깃값을 갱신하는 과정을 반복하여 K개의 최종 군집을 형성한다.
다변량 분석(Multivariate analysis) : 여러 현상이나 사건에 대한 측정치를 개별적으로 분석하지 않고 동시에 한번에 분석 하는 통계적 기법을 말한다.

3. 분석 모형 설계 – 분석 절차 수립

개념 체크

01 분석 모형에 대한 설명 중 틀린 것은?

① 분석 모형 선정은 분석 목적과 수집된 데이터의 변수들 을 고려하여 적합한 빅데이터 분석 모형을 선정하는 것이다.

② 통계 분석은 수치화된 자료를 분석하여 사회현상을 예측하고 이해하고자 할 때 사용된다.

③ 분석 모형 선정 방법에는 통계기반, 데이터 마이닝 기반, 머신러닝 기반, 분할 기반이 있다.

④ 통계기반 분석 모형에는 기술통계, 추론통계, 상관분석, 회귀 분석, 분산 분석, 주성분 분석 등이 있다.

분석 모형 선정 방법에는 통계기반, 데이터 마이닝 기반, 머신러닝 기반 3가지가 있다.

분석 모형 선정

● 분석 모형 선정은 분석 목적과 수집된 데이터의 변수들 을 고려하여 적합한 빅데이터 분석 모형을 선정하는 것이다.

● 통계, 데이터 마이닝, 머신러닝 기반 분석 모형 선정 방법을 고려하여 적절한 데이터 분석 모델을 선정한다.

02 다음과 같은 연구를 위해 사용할 수 있는 분석 방법은?

- 흡연량에 따른 폐암 발병률 연구
- 처벌정책이 범죄율에 미치는 영향 연구
- 주택당 방수에 따른 보스턴 집값 예측 연구

① 회귀 분석

② 추론 통계

③ 요인 분석

④ 상관 분석

하나 또는 그 이상의 독립변수(X)가 종속변수(Y)에 끼치는 영향을 확인할 수 있는 회귀 분석을 통해 분석할 수 있다.

추론통계(Inferential Statistic)

▶ 모집단에서 추출된 표본으로부터 모수와 관련된 통계량들의 값을 계산하고, 이것을 이용하여 모집단의 특성을 알아내는 방법이다.

상관분석(Correlation Analysis)

▶ 두 개 이상의 변수 사이에 존재하는 상호 연관성을 분석하는 방법으로 상관계수(r)를 이용하여 상관관계를 분석 하는 방법이다.

3. 분석 모형 설계 – 분석 절차 수립

03 다음 중 통계기반 분석 모형에 속하지 않는 것은?

① 기술 통계

② 비율 통계

③ 추론 통계

④ 상관분석

통계기반 분석 모형에는 기술통계, 추론통계, 상관분석, 회귀분석, 인과관계 분석, 분산분석, 주성분 분석 등이 있다.

04 다음 중 설명이 틀린 것은?

① 기술 통계는 데이터의 특징을 파악하기 위해 평균, 분산, 표준편차 등의 기초통계량을 구하거나 시각화 도구인 그래프를 활용하는 분석 방법이다.

② 추론 통계는 모집단에서 추출된 표본으로부터 모수와 관련된 통계량들의 값을 계산하고, 이것을 이용하여 모집단의 특성을 알아내는 방법이다.

③ 상관 분석은 두 개 이상의 변수 사이에 존재하는 상호 연관성을 분석하는 방법이다.

④ 회귀 분석은 기존 데이터의 분포를 최대한 보존하면서 고차원 공간의 데이터들을 저차원 공간으로 변환하는 방법이다.

기존 데이터의 분포를 최대한 보존하면서 고차원 공간의 데이터들을 저차원 공간으로 변환하는 방법은 차원 축소 방법 중 하나인 주성분 분석(PCA)에 대한 설명이다.

회귀 분선은 하나 이상의 독립변수(X)가 종속변수(Y)에 끼치는 영향을 추정하는 통계 방법이다.

3. 분석 모형 설계 – 분석 절차 수립

05 다음 설명에 해당하는 것은?

많은 양의 데이터 속에서 데이터의 패턴, 규칙 등을 탐색하고, 통계기법을 활용하여 분석한 뒤, 이러한 분석을 기반으로 가치 있는 정보를 추출하는 과정을 의미한다.

① 데이터 분석 ② 데이터 마이닝

③ 데이터 모델링 ④ 데이터 추출

06 다음 설명에 해당하는 것은?

- 주어지는 데이터에서 동시에 발생하는 사건 혹은 항목 간의 규칙을 수치화하는 기법으로 '장바구니 분석'이라고도 하며, 주로 마케팅 분야에서 활용된다.
- 예를 들어 우유를 구입한 고객이 식빵을 함께 구입한 경우를 들 수 있다.

① 군집화 모델 ② 예측 모델

③ 분류 모델 ④ 연관규칙 모델

군집화 모델(Clustering)

▶ 관측된 여러 개의 변숫값에서 유사한 성격을 갖는 몇 개의 군집으로 그룹화하여 그룹들 사이에 관계를 분석하는 다변량 분석기법이다.

예측 모델(Prediction)

▶ 과거 데이터로부터 데이터의 특성을 분석하여 다른 데이터의 결과값을 예측하는 기법이다.

분류 모델(Classification)

▶ 다수의 속성을 갖는 객체들을 사전에 정해진 그룹 중

3. 분석 모형 설계 – 분석 절차 수립

3) 머신러닝 기반 분석 모형 선정

- 머신러닝(Machine Learning)이란 컴퓨터가 스스로 데이터를 분석하고 학습하여 인공지능 성능을 향상시키는 기술이다.
- 머신러닝 처리 단계는 표현(Representaion), 평가(Evaluation), 최적화(Optimization), 일반화(Generalization) 순이다.
- 머신러닝 학습 방법은 지도 학습, 비지도 학습, 강화 학습, 준지도 학습, 전이 학습으로 나뉜다.

① 지도 학습(Supervised Learning)

- ▶ 정답인 레이블(Label)이 포함된 학습 데이터를 통해 컴퓨터를 학습시키는 방법이다.
- ▶ 인식, 분류, 진단, 예측 등의 문제에 적합하다.

② 비지도 학습(Unsupervised Learning)

- ▶ 정답인 레이블(Label)이 없는 상태에서 컴퓨터를 학습시키는 방법이다.
- ▶ 현상에 대한 설명, 특징 도출, 패턴 도출 등에 적합하다.

③ 강화 학습(Reinforcement Learning)

- ▶ 컴퓨터가 선택 가능한 행동(Action) 중 보상(Reward)을 최대화하는 행동을 선택하도록 하는 학습 방법이다.

3. 분석 모형 설계 – 분석 절차 수립

3) 머신러닝 기반 분석 모형 선정

④ 준지도 학습(Semi-Supervised Learning)

▶ 정답이 포함된 데이터와 정답이 없는 데이터를 모두 훈련에 사용하는 학습 방법이다.

⑤ 전이 학습(Transfer Learning)

▶ 학습된 모형을 기반으로 최종 출력층을 바꾸어 재학습하는 방법이다.

▶ 하나의 작업을 위해 훈련된 모델을 유사 작업 수행 모델의 시작점으로 활용하는 딥러닝 학습 방법이다.

예) 고양이를 인식하기 위해 학습하는 동안 얻은 지식을 호랑이를 인식하려고 할 때 적용한다.

학습 방법에 따른 분석 모형	
지도 학습 분석 모형	비지도 학습 분석 모형
• 회귀 분석 • 로지스틱 회귀 분석 • 나이브 베이즈 • KNN(K-최근접 이웃 알고리즘) • 의사결정나무 • 인공신경망 • 서포트 벡터 머신(SVM) • 랜덤 포레스트 / 감성 분석	• 군집화 (k-means, SOM, 계층군집 등) • 차원 축소(주성분 분석, 선형판별 분석 등) • 연관 분석 • 자율학습 인공신경망

딥러닝(Deep Learning) : 대용량 데이터를 처리하기 위해 인공신경망을 기반으로 구현되는 기계학습 알고리즘이다.

3. 분석 모형 설계 – 분석 절차 수립

4) 독립변수와 종속변수의 데이터 유형에 따른 분석기법

; 독립변수와 종속변수의 데이터 유형에 따라 다양한 분석기법을 활용할 수 있다.

독립변수와 종속변수에 따른 분석기법

		종속변수(Y)		
		연속형 변수	범주형 변수	없음
독립변수(X)	연속형 변수	• 회귀 분석 • 인공신경망 모델 • KNN • 의사결정나무(회귀)	• 로지스틱 회귀 분석 • 판별 분석 • KNN • 의사결정나무(분류)	• 주성분 분석 • 군집 분석
	범주형 변수	• 회귀 분석 • 인공신경망 모델 • 의사결정나무(회귀)	• 로지스틱 회귀 분석 • 인공신경망 모델 • 의사결정나무(분류)	• 연관 분석 • 판별 분석

3. 분석 모형 설계 – 분석 절차 수립

개념 체크

01 다음 설명에 해당하는 것은?

컴퓨터가 스스로 데이터를 분석하고 학습하여 인공지능 성능을 향상시키는 기술이다.

① 딥러닝

② 데이터 마이닝

③ 머신러닝

④ 데이터 분석

딥러닝(Deep Learning) : 대용량 데이터를 처리하기 위해 인공신경망을 기반으로 구현되는 기계학습 알고리즘이다.

데이터 마이닝(Data Mining) : 많은 양의 데이터 속에서 데이터의 패턴, 규칙 등을 탐색하고, 통계기법을 활용하여 분석한 뒤, 이러한 분석을 기반으로 가치 있는 정보를 추출하는 과정을 의미한다.

02 머신러닝 기반 분석 모형 선정에 대한 설명 중 틀린 것은?

① 머신러닝 기반 분석 모형 선정은 분석 모형 선정 방법 중 하나이다.

② 머신러닝 학습 방법은 지도 학습, 비지도 학습, 강화 학습, 반지도 학습으로 나뉜다.

③ 지도 학습은 정답이 주어진 상태에서 데이터를 학습하는 방법이다.

④ 비지도 학습은 정답이 주어지지 않은 상태에서 데이터를 학습하는 방법이다.

머신러닝 학습 방법에는 지도 학습, 비지도 학습, 강화 학습, 준지도 학습, 전이 학습이 있다.

지도 학습

▶ 정답인 레이블이 포함된 학습 데이터를 통해 컴퓨터를 학습시키는 방법이다.

▶ 인식, 분류, 진단, 예측 등의 문제에 적합하다.

비지도 학습

▶ 정답인 레이블이 없는 상태에서 컴퓨터를 학습시키는 방법이다.

3. 분석 모형 설계 – 분석 절차 수립

03 다음 설명에 해당하는 학습 방법은?

- 2016년 이세돌 프로와 대국했던 알파고가 사용했던 학습 방법이다.
- 컴퓨터가 선택 가능한 행동(Action) 중 보상(Reward)을 최대화하는 행동을 선택하도록 하는 학습 방법이다.

① 강화 학습　② 지도 학습
③ 비지도 학습　④ 준지도 학습

04 다음 중 성질이 다른 모형은?

① SVM　② KNN
③ 회귀 분석　④ SOM

지도 학습 분석 모형

▶ 회귀분석, 로지스틱 회귀분석, 나이브 베이즈, KNN, 의사결정나무, 인공신경망, SVM(서포트 벡터 머신), 랜덤 포레스트 / 감성 분석

비지도 학습 분석 모형

▶ 군집화(K-means, SOM, 계층군집)
▶ 차원 축소(주성분 분석, 선형판별 분석 등)
▶ 연관 분석
▶ 자율학습 인공 신경망

3. 분석 모형 설계 – 분석 절차 수립

02 분석 모형 정의

● **분석 모형 정의는 분석 모형을 선정하고 모형(Model)에 적합한 변수를 선택하여 모형의 사양(Specification) 을 정의하는 기법이다.**

● **선택된 모형에 적합한 변수를 사용하기 위해 매개변수와 초매개변수를 선정한다.**

매개변수와 초매개변수

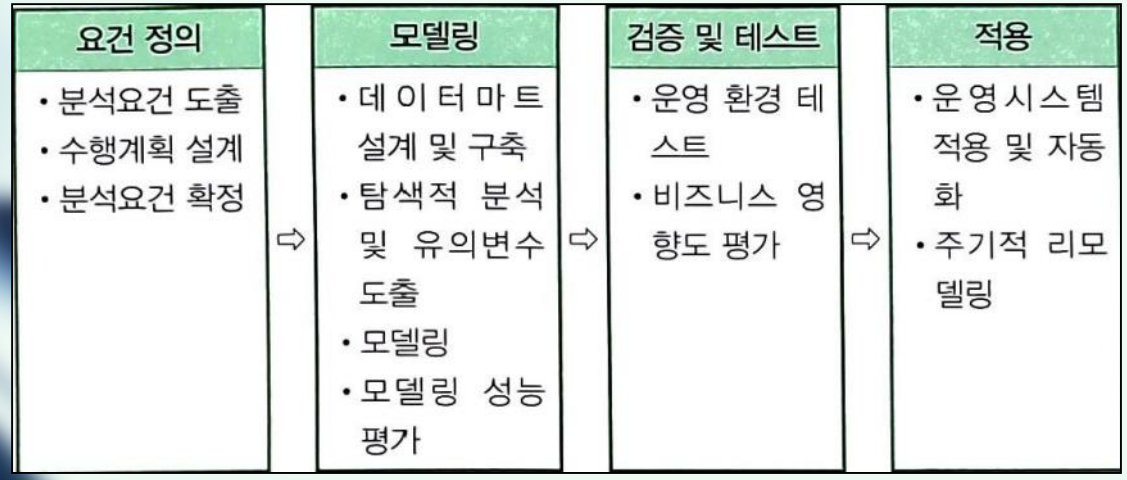
변수 명칭	설명
매개변수 (Parameter)	• 모델 내부에서 획인 가능한 변수로 데이터를 통해 자동으로 산출된 값이며, 수작업으로 측정되지 않는다. • 매개변수가 모델의 성능에 영향을 미친다. 예 인공신경망의 가중치, SVM에서 SV, 선형회귀에서 결정계수
초매개변수 (Hyper Parameter)	• 모델 외부 요소로 사용자가 직접 수작업으로 설정해주는 값이다. • 학습 과정과 학습 결과에 영향을 미친다. 예 학습률, 의사결정나무 깊이(Depth), 신경망에서 은닉층(Hidden Layer)의 개수, SVM에서 코스트 값인 C, KNN에서 K개수

서포트 벡터(Support Vector, SV) : 데이터 중 결정경계와 가장 가까이에 있는 데이터
결정경계(Decision Boundary) : 데이터 분류의 기준 경계선
결정계수(R^2, Coefficient of Determination) : 전체 데이터를 회귀모형이 얼마나 잘 설명하고 있는지를 보여주는 지표이다.
회귀 모형 : 어떤 관계가 있을지에 대한 여러 가지 가설들을 말한다.
은닉층(또는 숨겨진 층, hidden layer) : 인공신경망에서 입력층과 출력층 사이에 위치하는 층이다.
KNN : K - 최근접 이웃법으로 분류 문제에 사용하는 알고리즘이다.

3. 분석 모형 설계 – 분석 절차 수립

03 분석 모형 구축 절차

● **분석 모형 구축 절차는 요건 정의, 모델링, 검증 및 테스트, 적용 순으로 진행된다.**



① 요건 정의 : 기획 단계에서 분석요건을 도출하고, 수행계획을 설계하며, 분석요건을 확정시키는 단계 .

② 모델링 : 정의된 요건에 근거하여 분석 작업을 수행하는 단계이며, 데이터 탐색과 분석을 통해 모델링 작업을 하고, 모델링 성능평가를 통해 최종 모델을 선정한다.

③ 검증 및 테스트: 분석 모델을 가상 운영 환경에서 테스트하는 단계

④ 적용 : 분석 결과를 실제 운영 환경에 적용하는 단계

3. 분석 모형 설계 – 분석 절차 수립

개념 체크

01 다음 설명에 해당하는 것은?

- 모델 내부에서 확인 가능한 변수로 데이터를 통해 자동으로 산출된 값이며, 수작업으로 측정되지 않는다.
- 예시로 인공신경망의 가중치, 선형회귀에서 결정계수가 있다.

① 반응변수

② 초매개변수

③ 매개변수

④ 독립변수

매개변수와 초매개변수

매개변수(Parameter)

● 모델 내부에서 확인 가능한 변수로 데이터를 통해 자동으로 산출된 값이며, 수작업으로 측정되지 않는다.

● 예시로 인공신경망의 가중치, SVM에서 SV, 선형회귀에서 결정계수가 있다.

초매개변수(Hyper Parameter)

02 다음 중 성질이 다른 변수는?

① SVM에서 SV

② 신경망에서 은닉층(Hidden Layer)의 개수

③ SVM에서 코스트 값인 C

④ KNN에서 K개수

3. 분석 모형 설계 – 분석 절차 수립

03 다음 중 분석 모형 구축 절차로 옳은 것은?

① 모델링 → 요건 정의 → 검증 및 테스트 → 적용

② 모델링 → 적용 → 검증 및 테스트 → 요건정의

③ 요건 정의 → 검증 및 테스트 → 모델링 → 적용

④ 요건 정의 → 모델링 → 검증 및 테스트 → 적용

요모검적 으로 외우자.

요건 정의 : 분석요건 도출, 수행계획 설계, 분석요건 확정

모델링 : 데이터마트 설계 및 구축, 탐색적 분석 및 유의 변수 도출, 모델링, 모델링 성능 평가

검증 및 테스트 : 가상 운영 환경 테스트, 비즈니스 영향도 평가

적용 : 실제 운영 환경에 적용 및 자동화, 주기적인 리모델링

04 분석 모형 구축 절차 중 요건 정의에 속하지 않는 업무 는?

① 수행계획 설계

② 분석요건 도출

③ 모델링

④ 분석요건 확정

모델링은 모델링 단계에서 진행한다.

3. 분석 모형 설계 – 분석 환경 구축

01 분석 도구 선정

● **대표적인 데이터 분석 도구로는 R, 파이썬(Python)이 있다. R과 파이썬은 모두 오픈소스로 무료로 사용이 가능하다.**

1) R

- R은 통계 프로그래밍 언어인 S언어를 기반으로 만들어진 오픈소스 프로그래밍 언어이다.
- R은 데이터 분석에 특화된 언어로 강력한 시각화 기능을 제공한다.
- 또한, 핵심 패키지 이외에 15,000개 이상의 패키지와 테스트 데이터를 다운받아 사용할 수 있다.
- R의 대표적인 통합 개발 환경(IDE)은 RStudio이다.
- Microsoft Windows, Mac OS, Linux 등 다양한 OS를 지원한다.

3. 분석 모형 설계 – 분석 환경 구축

2) 파이썬(Python)

- 파이썬은 C언어 기반의 오픈소스 프로그래밍 언어이다.
- R과 달리 특정 영역에 특화된 언어가 아닌 범용으로 사용 가능한 언어이다.
- 파이썬 역시 다양한 시각화 라이브러리를 지원하지만 R에 비해서는 선택의 폭이 좁다.
- 파이썬은 TensorFlow, Keras 등 인공지능 패키지 분석에 용이하다.
- 파이썬의 대표적인 통합 개발 환경(IDE)은 주피터 노트북(Jupyter Notebook), 파이참(Pycharm) 등이 있다.
- Microsoft Windows, Mac OS, Linux 등 다양한 OS를 지원한다.

텐서플로(TensorFlow) : 구글(Google)에서 만든, 딥러닝 프로그램을 쉽게 구현할 수 있도록 다양한 기능을 제공해주는 라이브러리다. 텐서플로 자체는 기본적으로 C++로 구현되어 있다.
Keras(케라스) : 파이썬으로 구현된 쉽고 간결한 딥러닝 라이브러리이다.

3. 분석 모형 설계 – 분석 환경 구축

02 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.

구분	설명
학습 데이터 (Training Data)	알고리즘을 학습하기 위한 데이터이다.
검증 데이터 (Validation Data)	• 학습된 모델의 성능을 검증하고, 모델을 선택하기 위한 데이터이다. • 초매개변수의 조정을 위해 필요한 데이터를 일반적으로 검증 데이터(validation data)라고 한다.
평가 데이터 (Test Data)	• 최종 모델의 성능을 평가하기 위한 데이터이다. • 주의! 학습 과정에서 사용되지 않음

- 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.

3. 분석 모형 설계 – 분석 환경 구축

개념 체크

01 다음 중 데이터 분석 도구에 대한 설명으로 옳지 않은 것은?

① 대표적인 데이터 분석 언어에는 R과 파이썬(Python)이 있다.

② 파이썬(Python)은 강력한 시각화 도구를 지원한다는 특징이 있다.

③ R과 파이썬(Python)은 모두 오픈소스로 무료로 사용 가능하다.

④ R과 파이썬(Python)은 모두 Windows, Linux 등 다양한 OS에서 사용이 가능하다.

강력한 시각화 도구를 지원하는 것은 R에 대한 설명이다.

R

● R은 통계 프로그래밍 언어인 S언어를 기반으로 만들어진 오픈소스 프로그래밍 언어이다.

● R은 데이터 분석에 특화된 언어로 강력한 시각화 기능을 제공한다.

● 또한, 핵심 패키지 이외에 15,000개 이상의 패키지와 테스트 데이터를 다운받아 사용할 수 있다.

02 다음 중 데이터 분할에 대한 설명으로 옳지 않은 것은?

① 데이터는 분석되기 전 목적에 맞게 분할되어야 하는데, 이는 분석 모형의 과적합을 방지하고, 일반화 성능을 향상시키기 위함이다.

② 데이터가 충분하지 않은 경우에도 학습, 검증, 평가 데이터로 분할하여 분석한다.

③ 일반적으로 데이터는 학습 데이터, 검증 데이터, 평가 데이터로 나뉜다.

④ 보통의 경우 학습 데이터와 검증 데이터를 60~80%로 사용하고, 평가 데이터를 20~40%로 사용한다.

데이터가 충분하지 않은 경우에는 학습 데이터와 평가 데이터로만 분할하여 분석하기도 한다.

3. 분석 모형 설계 – 분석 환경 구축

03 다음 중 분석도구에 대한 설명으로 옳지 않은 것은?

① 대표적인 분석도구로는 R과 Python이 있다.

② R은 데이터 분석에 특화된 언어로 강력한 시각화 기능을 제공한다.

③ R은 TensorFlow, Keras 등 인공지능 패키지 분석에 용이하다.

④ Python은 C언어 기반의 오픈소스 프로그래밍 언어이다.

Python은 TensorFlow, Keras 등 인공지능 패키지 분석에 용이하다.

파이썬(Python)

- 파이썬은 C언어 기반의 오픈 소스 프로그래밍 언어이다.
- R과 달리 특정 영역에 특화된 언어가 아닌 범용으로 사용 가능한 언어이다.
- 파이썬 역시 다양한 시각화 라이브러리를 지원하지만 R에 비해서는 선택의 폭이 좁다.
- Python은 TensorFlow, Keras 등 인공지능 패키지 분석에 용이하다.
- 파이썬의 대표적인 통합 개발 환경은 주피터 노트북과

3. 분석 모형 설계 예상 문제

예상문제

01 다음 설명에 해당하는 것은?

많은 양의 데이터 속에서 데이터의 패턴, 규칙 등을 탐색하고, 통계기법을 활용하여 분석한 뒤, 이러한 분석을 기반으로 가치 있는 정보를 추출하는 과정을 의미한다.

① 데이터 분석

② 데이터 마이닝

③ 데이터 모델링

④ 데이터 추출

02 다음 중 데이터 마이닝 기반 분석 모형이 아닌 것은?

① 비교 모델

② 분류 모델

③ 예측 모델

④ 군집화 모델

데이터 마이닝 기반 분석 모델에는 분류, 예측, 군집화, 연관 규칙 모델이 있다.

03 다음 설명에 해당하는 학습 방법은?

- 정답이 포함된 데이터와 정답이 없는 데이터를 모두 훈련에 사용하는 학습 방법으로 이 방법을 사용한다
- 대표적인 예로 사용자가 올린 사진을 자동으로 정리해주는 구글 포토가 있다.

① 준지도 학습 ② 지도 학습

③ 강화 학습 ④ 비지도 학습

지도 학습(Supervised Learning)

▶ 정답인 레이블이 포함된 학습 데이터를 통해 컴퓨터를 학습시키는 방법이다.

▶ 인식, 분류, 진단, 예측 등의 문제에 적합하다.

비지도 학습(Unsupervised Learning)

▶ 정답인 레이블이 없는 상태에서 컴퓨터를 학습시키는 방법이다.

▶ 현상에 대한 설명, 특징 도출, 패턴 도출 등에 적합하다.

강화 학습(Reinforcement Learning)

▶ 컴퓨터가 선택 가능한 행동(Action) 중 보상(Reward)을 최대화하는 행동을 선택하도록 하는 학습 방법이다.

3. 분석 모형 설계 예상 문제

05 다음 중 독립변수와 종속변수가 모두 연속형 변수일 때 활용 가능한 분석기법이 아닌 것은?

① 회귀 분석 ② 판별 분석

③ 인공신경망 모델 ④ KNN

독립변수와 종속변수에 따른 분석기법

독립변수(X)가 연속형, 종속변수(Y)가 연속형 : 회귀 분석, 인공신경망 모델, KNN, 의사결정나무(회귀)

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

독립변수(X)가 연속형, 종속변수(Y)가 없음 : 주성분 분석, 군집 분석

독립변수(X)가 범주형, 종속변수(Y)가 연속형 : 회귀 분석, 인공신경망 모델, 의사결정나무(회귀)

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

독립변수(X)가 범주형, 종속변수(Y)가 없음 : 연관 분석, 판별 분석

06 다음 설명 중 틀린 것은?

07 다음 통계기반 분석 모형 선정에 대한 설명 중 틀린 것은?

① 통계기반 분석 모형에는 기술 통계, 추론통계, 상관분석, 회귀분석, 분산 분석, 주성분 분석 등이 있다.

② 기술 통계는 데이터의 특징을 파악하기 위해 평균, 분산, 표준편차 등의 기초통계량을 구하거나 시각화 도구인 그래프를 활용하는 분석 방법이다.

③ 추론 통계는 두 개 이상의 변수 사이에 존재하는 상호 연관성을 분석하는 방법이다.

④ 회귀 분석은 하나 이상의 독립변수(X)가 종속변수(Y)에 끼치는 영향을 추정하는 통계 방법이다.

추론통계는 모집단에서 추출된 표본으로부터 모수와 관련된 통계량들의 값을 계산하고, 이것을 이용하여 모집단의 특성을 알아내는 방법이다. 두 개 이상의 변수 사이에 존재하는 상호 연관성을 분석하는 방법은 상관분석이다.

08 다음 중 기존 데이터의 분포를 최대한 보존하면서 고차원 공간의 데이터들을 저차원 공간으로 변환하는 분석 방법은?

① 상관분석 ② 회귀 분석

③ 주성분 분석 ④ 분산 분석

3. 분석 모형 설계 예상 문제

09 다음 중 데이터 마이닝 기반 분석 모형 선정에 대한 설명으로 옳지 않은 것은?

① 데이터 마이닝 기반 분석 모델에는 분류, 예측, 군집화, 연관규칙이 있다.

② 분류 모델은 다수의 속성을 갖는 객체들을 사전에 정해진 그룹 중 하나로 분류하는 기법이다.

③ 예측 모델의 예시에는 통계적 기법, 트리 기반 기법, 최적화 기법, 기계학습 모델이 있다.

④ 군집화 모델은 관측된 여러 개의 변숫값에서 유사한 성격을 갖는 몇 개의 군집으로 그룹화하여 그룹들 사이의 관계를 분석하는 다변량 분석기법이다.

분류 모델의 예시 : 통계적 기법, 트리 기반 기법, 최적화 기법, 기계학습 모델

예측 모델의 예시 : 회귀 분석, 의사결정나무, 시계열 분석, 인공신경망

군집화 모델의 예시

계층적 방법 : 병합적 방법, 분할적 방법

비계층적 방법 : K-평균군집

11 다음 중 분석 모형 정의에 대한 설명으로 옳지 않은 것은?

① 분석 모형 정의는 분석 모형을 선정하고 모형에 적합한 변수를 선택하여 모형의 사양을 정의하는 기법이다.

② 선택된 모형에 적합한 변수를 사용하기 위해 매개변수와 초매개변수를 선정한다.

③ 매개변수의 예로는 학습률, 의사결정나무 깊이(Depth) 등이 있다.

④ 초매개변수는 모델 외부 요소로 사용자가 직접 수작업으로 설정해주는 값이다.

학습률, 의사결정나무 깊이(Depth)는 초매개변수의 예이다.

● 분석 모형 정의는 분석 모형을 선정하고 모형에 적합한 변수를 선택하여 모형의 사양을 정의하는 기법이다.

● 선택된 모형에 적합한 변수를 사용하기 위해 매개변수와 초매개변수를 선정한다.

● **매개변수(Parameter)**

▶ 모델 내부에서 확인 가능한 변수로 데이터를 통해 자동으로 산출된 값이며, 수작업으로는 측정되지 않는다.

3. 분석 모형 설계 예상 문제

13 다음 중 분석 도구에 대한 설명으로 옳지 않은 것은?

① 빅데이터 분석을 위해 사용되는 대표적인 언어에는 R, Python이 있다.

② R은 데이터 분석에 특화된 언어로 강력한 시각화 기능을 제공한다.

③ Python은 TensorFlow, Keras 등 인공지능 패키지 분석에 용이하다.

④ R은 C언어를 기반으로 만들어진 오픈 소스 프로그래밍 언어이다.

R은 통계 프로그래밍 언어인 S언어를 기반으로 만들어진 오픈 소스 프로그래밍 언어이다.

R

● R은 통계 프로그래밍 언어인 S언어를 기반으로 만들어진 오픈 소스 프로그래밍 언어이다.

● R은 데이터 분석에 특화된 언어로 강력한 시각화 기능을 제공한다.

● 또한, 핵심 패키지 이외에 15,000개 이상의 패키지와 테스트 데이터를 다운받아 사용할 수 있다.

15 다음 중 정답인 레이블이 포함된 학습 데이터를 통해 컴퓨터를 학습시키는 방법은?

① 비지도 학습 ② 지도 학습

③ 강화 학습 ④ 준지도 학습

16 다음 중 C언어를 기반으로 만들어진 데이터 분석과 머신러닝 프로그래밍이 가능한 오픈소스 언어는?

① Swift ② Java

③ R ④ 파이썬

3. 분석 모형 설계 예상 문제

17 다음과 같은 형태의 데이터 분석 언어는?

```
import pandas as pd
import numpy as np

df=pd.read_csv("airquality.csv")
df_mean=df['Ozone'].mean()
print(df_mean)
```

① Python ② Java

③ R ④ PHP

import 예약어 같은 내용이 나오면 python 코드라고 생각하자. R같은 경우는 install.package(), library(패키지명) 코드가 나오면 R코드라고 생각하시면 된다.

18 다음과 같은 분석 모형 구축 절차 단계에 해당하는 것은?

•데이터 마트를 설계하고 구축한다.

•탐색적 분석을 하고, 유의변수를 도출한다.

① 모델링 ② 요건 정의

③ 검증 및 테스트 ④ 적용

19 다음 중 지도 학습 모형으로 알맞게 짝지어진 것은?

㉠ K-means	㉡ 회귀 분석
㉢ SOM	㉣ PCA
㉤ 로지스틱 회귀 분석	㉥ SVM
㉦ FDA	㉧ KNN

① ㉠, ㉡, ㉣, ㉦ ② ㉡, ㉤, ㉥, ㉧

③ ㉡, ㉢, ㉧ ④ ㉤, ㉥, ㉦

지도 학습 분석 모형의 종류

① 회귀 분석

② 로지스틱 회귀 분석

③ 나이브 베이즈

④ KNN(K-최근접 이웃 알고리즘)

⑤ 의사결정나무

⑥ 인공신경망

⑦ 서포트 벡터 머신(SVM)

⑧ 랜덤 포레스트

⑨ 감성 분석

비지도 학습 분석 모형의 종류

① 군집화(K-means, SOM, 계층군집)

감사합니다.