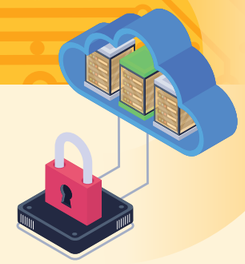
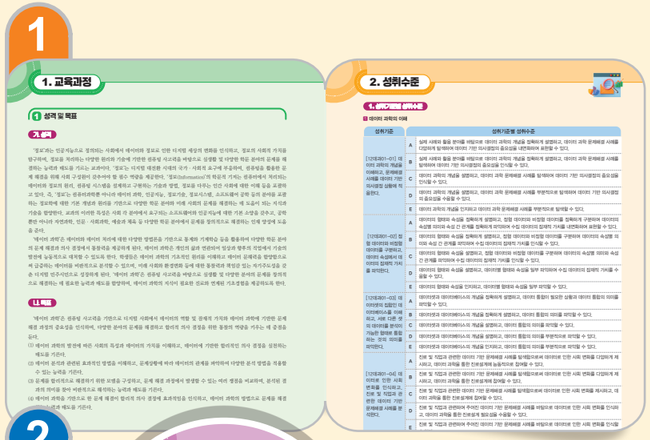
구성과 특징



교육과정 해설

2022 개정 교육과정의 총론과 성취기준을 제시하여 교육과정 편성에 참고하도록 하였습니다.

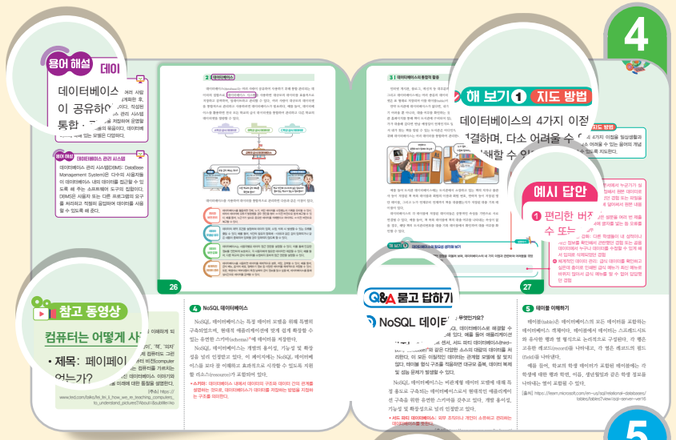


연간 지도 계획

64차시(4학점×16주)를 기준으로, 전체적인 구성과 차시를 표로 제시하여 연간 계획이 한눈에 들어오도록 하였습니다.

대단원 도입 · 차시 시작

- **단원 개관** | 대단원에서 배울 내용을 소개하고, 학습 목표와 방향을 명확하게 제시하였습니다.
- **단원 전개 계획** | 소단원명, 학습 주제, 지도 방법 등을 표로 제시하여 수업 계획에 도움이 되도록 하였습니다.
- **교수 · 학습 계획안** | 소단원마다 교수 · 학습 계획안을 제시하였습니다.



교수 참고 자료와 지도 방법, 문제 해설과 예시 답안

- **교수 참고 자료** | 교사들이 참고할 자료를 교과서 본문 내용과 유기적으로 연계하여 수록하였습니다.
- **지도 방법** | 실제 수업 지도 시 유의할 점이나 도움이 될 내용을 넣었습니다.
- **문제 해설과 예시 답안** | 해 보기, 탐구 활동 등의 문제에 대한 예시 답안과 해설을 수록하였습니다.

추가 활동지

교과서에 수록하지 못한 활동지를 추가로 제공하여 수업에 활용할 수 있도록 하였습니다.

시험 대비용 | 대단원 평가 문제

각 단원별로 단원을 정리할 수 있는 평가 문제와 상세한 해설을 두어 평가에 활용할 수 있도록 하였습니다.

차례

2022 개정 교육과정 해설·성취수준

01 | 교육과정

02 | 성취수준

03 | 연간 지도 계획

1 성격 및 목표

가. 성격

'정보'과는 인공지능으로 정의되는 사회에서 데이터와 정보로 인한 디지털 세상의 변화를 인식하고, 정보의 사회적 가치를 탐구하며, 정보를 처리하는 다양한 원리와 기술에 기반한 컴퓨팅 사고력을 바탕으로 실생활 및 다양한 학문 분야의 문제를 해결하는 능력과 태도를 기르는 교과이다. '정보'는 디지털 대전환 시대의 국가 · 사회적 요구에 부응하여, 컴퓨팅을 활용한 문제 해결을 위해 사회 구성원이 갖추어야 할 필수 역량을 제공한다. '정보(Informatics)'의 학문적 기저는 컴퓨터에서 처리되는 데이터와 정보의 원리, 컴퓨팅 시스템을 설계하고 구현하는 기술과 방법, 정보를 다루는 인간 사회에 대한 이해 등을 포괄하고 있다. 즉, '정보'는 컴퓨터과학뿐 아니라 데이터 과학, 인공지능, 정보기술, 정보시스템, 소프트웨어 공학 등의 분야를 포괄하는 정보학에 대한 기본 개념과 원리를 기반으로 다양한 학문 분야와 미래 사회의 문제를 해결하는 데 도움이 되는 지식과 기술을 함양한다. 교과의 이러한 특성은 사회 각 분야에서 요구되는 소프트웨어와 인공지능에 대한 기본 소양을 갖추고, 공학뿐만 아니라 자연과학, 인문 · 사회과학, 예술과 체육 등 다양한 학문 분야에서 문제를 창의적으로 해결하는 인재 양성에 도움을 준다.

'데이터 과학'은 데이터와 데이터 처리에 대한 다양한 방법론을 기반으로 통계와 기계학습 등을 활용하여 다양한 학문 분야의 문제 해결과 의사 결정에서 통찰력을 제공하게 된다. 데이터 과학은 개인의 삶과 연관되어 일상과 향후의 직업에서 기술의 발전에 능동적으로 대처할 수 있도록 한다. 학생들은 데이터 과학의 기초적인 원리를 이해하고 데이터 문해력을 함양함으로써 급증하는 데이터를 비판적으로 분석할 수 있으며, 미래 사회와 환경변화 등에 대한 통찰력과 책임감 있는 자기주도성을 갖춘 디지털 민주시민으로 성장하게 된다. '데이터 과학'은 컴퓨팅 사고력을 바탕으로 실생활 및 다양한 분야의 문제를 창의적으로 해결하는 데 필요한 능력과 태도를 함양하며, 데이터 과학의 지식이 필요한 진로와 연계된 기초경험을 제공하도록 한다.

나. 목표

'데이터 과학'은 컴퓨팅 사고력을 기반으로 디지털 사회에서 데이터의 역할 및 잠재적 가치와 데이터 과학에 기반한 문제 해결 과정의 중요성을 인식하며, 다양한 분야의 문제를 해결하고 합리적 의사 결정을 위한 통찰의 역량을 키우는 데 중점을 둔다.

(1) 데이터 과학의 발전에 따른 사회의 특성과 데이터의 가치를 이해하고, 데이터에 기반한 합리적인 의사 결정을 실천하는 태도를 기른다.

(2) 데이터 분석과 관련된 효과적인 방법을 이해하고, 문제상황에 따라 데이터의 관계를 파악하여 다양한 분석 방법을 적용할 수 있는 능력을 기른다.

(3) 문제를 합리적으로 해결하기 위한 모델을 구성하고, 문제 해결 과정에서 발생할 수 있는 여러 쟁점을 비교하며, 분석된 결과의 의미를 찾아 비판적으로 해석하는 능력과 태도를 기른다.

(4) 데이터 과학을 기반으로 한 문제 해결이 합리적 의사 결정에 효과적임을 인식하고, 데이터 과학의 방법으로 문제를 해결하는 능력과 태도를 기른다.

2 내용 체계 및 성취기준

가. 내용 체계

1 데이터 과학의 이해

핵심 아이디어		• 디지털 사회의 시민에게는 데이터에 기반한 합리적인 의사 결정이 필요하다. • 데이터 과학에 대한 이해는 데이터를 활용하여 복잡한 문제를 해결하는 데 도움을 준다.
내용 요소	지식 · 이해	• 데이터 과학의 개념 • 데이터의 형태와 속성 • 데이터셋과 데이터베이스
	과정 · 기능	• 데이터 과학의 문제 해결 사례 탐색하기 • 데이터의 형태와 속성 파악하기 • 데이터 통합의 의미 파악하기
	가치 · 태도	• 데이터 기반 의사 결정의 중요성 인식 • 데이터의 잠재적 가치 내면화 • 데이터 과학을 통한 진로설계 참여

2 데이터 준비와 분석

핵심 아이디어		• 데이터 분석을 위해서는 데이터를 수집, 전처리하는 과정이 필요하다. • 데이터 처리는 데이터를 분석에 효과적인 형태로 변환하며, 지식을 추출하는 데 도움을 준다.
내용 요소	지식 · 이해	• 데이터 전처리 • 데이터 분석 방법
	과정 · 기능	• 데이터 시각화하고 분석하기 • 이상치와 결측치 처리하고 정규화 활용하기 • 데이터 속성 간의 관계를 파악하고 통합하여 탐색하기 • 서로 다른 데이터 분석 방법 비교하기
	가치 · 태도	• 데이터가 편향되지 않도록 수집하는 자세 • 데이터의 불확실성과 오류 가능성 인식

3 데이터 모델링과 평가

핵심 아이디어		• 데이터 모델은 문제를 합리적으로 해결할 수 있도록 도움을 준다. • 데이터 기반의 합리적 의사 결정을 위해 데이터를 분석해서 새로운 지식을 추출하고, 의미를 해석한다.
내용 요소	지식 · 이해	• 데이터 모델의 개념 • 회귀 분석 • 군집 분석 • 연관 분석
	과정 · 기능	• 분석을 위한 도구 탐색하기 • 분석 결과 평가하기 • 분석 결과에 대한 의미 해석하기
	가치 · 태도	• 데이터에 대한 다양한 해석을 수용하는 태도 • 적절한 분석 방법을 선택하여 적용하는 자세

4 데이터 과학 프로젝트

핵심 아이디어	• 데이터 기반 문제 해결을 위해 데이터 과학의 기본 개념과 원리를 바탕으로 탐구 과정을 수행한다. • 데이터 과학으로 문제를 해결할 때, 통계적 방법이나 기계학습 등 다양한 방법을 활용한다.	
내용 요소	지식 · 이해	• 데이터 과학의 주제 • 탐색적 데이터 분석 • 결과의 의미 해석
	과정 · 기능	• 분야별 데이터 과학의 주제 조사하기 • 탐색적 데이터 분석으로 데이터 속 의미 파악하기 • 기계학습 방법으로 분석하기 • 결과를 활용하는 방법 탐색하기
	가치 · 태도	• 문제를 해결하기 위한 창의적인 방법을 고민하는 자세 • 복잡하고 어려운 문제를 끝까지 해결하기 위해 노력하는 태도 • 일반화 및 공유 과정에서 윤리 문제 등 사회적 영향 인식

나. 성취기준

1 데이터 과학의 이해

[12데과01-01] 데이터 과학의 개념을 이해하고, 문제 해결 사례를 데이터 기반 의사 결정 상황에 적용한다.
[12데과01-02] 정형 데이터와 비정형 데이터를 구분하고, 데이터 속성에서 데이터의 잠재적 가치를 파악한다.
[12데과01-03] 데이터셋의 집합인 데이터베이스를 이해하고, 서로 다른 데이터셋의 데이터를 분석이 가능한 형태로 통합하는 것의 의미를 파악한다.
[12데과01-04] 데이터로 인한 사회 변화를 인식하고, 진로 및 직업과 관련한 데이터 기반 문제 해결 사례를 분석한다.

(가) 성취기준 해설

- [12데과01-02] 데이터의 형태를 바탕으로 정형 데이터와 비정형 데이터를 구분하고, 정형 데이터에서 데이터 속성별 의미와 속성 간 관계를 파악하여 수집된 데이터가 분석 대상으로서 가치가 있는지 판단할 수 있어야 한다.
- [12데과01-03] 서로 다른 데이터셋 간 공통된 속성을 기준으로 데이터를 통합할 수 있음을 이해하고, 데이터셋 간의 관계를 바탕으로 데이터베이스의 개념을 설명할 수 있어야 한다. 대규모의 데이터를 여러 사람이 공유하기 위해서는 체계적인 시스템이 필요하다는 점을 바탕으로 데이터베이스의 필요성을 설명할 수 있어야 한다.

(나) 성취기준 적용 시 고려 사항

- 실제 활용 분야와 사례를 중심으로 데이터 과학이 어떻게 발전되어 왔는지 이해할 수 있도록 하며, 지속가능한 발전을 통한 미래 사회를 만들기 위한 데이터의 역할 및 중요성을 파악할 수 있도록 교수 · 학습을 구성한다.
- 우리 사회의 다양한 분야에서 사용될 수 있는 데이터의 잠재적 가치를 이해하고, 데이터와 데이터 분석이 활용된 문제 해결의 사례를 탐색할 수 있도록 교수 · 학습을 구성한다.
- 데이터 과학에 기반한 여러 가지 문제 해결 사례를 탐색하기 어려워하는 학습자의 경우, 교수자가 제시한 사례에서 사용된 데이터가 무엇인지, 문제 해결에 어떠한 역할을 하였는지를 기반으로 데이터의 잠재적 가치를 설명할 수 있는지를 평가하도록 한다.

2 데이터 준비와 분석

[12데과02-01] 데이터를 편향되지 않도록 수집하고, 수집된 데이터의 특성을 분석한다.
[12데과02-02] 이상치와 결측치 탐색 및 정규화를 통해 전처리하여 오류 가능성을 최소화하고, 데이터 분석을 위해 시각화한다.
[12데과02-03] 데이터를 분석하기 위해 데이터 속성 간의 관계를 파악하고 통합한다.
[12데과02-04] 동일한 데이터를 서로 다른 분석 방법을 적용하여 분석 결과를 비교한다.

(가) 성취기준 해설

- [12데과02-01] 데이터의 선택과 수집된 데이터를 활용하는 과정에서 발생할 수 있는 편향성을 최소화하고, 수집된 데이터의 출처, 규모, 데이터 속성별 자료형, 간단한 통계적 특성 등 데이터의 특성을 분석하고 파악할 수 있어야 한다.
- [12데과02-04] 동일한 데이터에 서로 다른 분석 방법을 적용하는 경우 분석 결과가 달라질 수 있음을 이해하고, 분석 방법과 연관지어 서로 다른 결과가 나온 이유 를 분석하고 비교할 수 있어야 한다.

(나) 성취기준 적용 시 고려 사항

- 데이터를 수집할 때 다양한 경로로 접근할 수 있는 공공 데이터 포털, 출처가 명확한 민간 데이터 포털을 통해 정확하고 신뢰할 수 있는 데이터를 수집하도록 활동을 구성하여 데이터 문해력을 기를 수 있도록 한다.
- 일상 속 데이터를 수집, 전처리, 분석하는 모든 과정에서 데이터 편향, 오류 가능성을 최소화 하기 위한 방법을 탐색하고, 관련 내용을 검증하는 과정을 교수 · 학습에 포함하도록 한다.
- 데이터 전처리에 어려움을 겪는 학습자의 경우, 전처리가 비교적 간단한 데이터를 제시하여 이상치와 결측치를 탐색하고, 데이터 특성에 적합한 분석 방법을 제시할 수 있는지 평가하도록 한다.

3 데이터 모델링과 평가

[12데과03-01] 데이터 모델 개념을 이해하고 데이터 분석에 활용할 수 있는 도구를 탐색한다.
[12데과03-02] 동일한 데이터를 통계적 회귀모델과 기계학습을 통한 회귀모델로 분석하여 결과 해석 내용을 비교한다.
[12데과03-03] 데이터의 속성에 대한 유사성을 측정하고 분석하여 군집을 형성하고, 군집 분석 결과의 의미를 해석한다.
[12데과03-04] 데이터 간의 관계를 분석하고 상호 연관성을 파악하여 결과의 의미를 해석한다.
[12데과03-05] 데이터 분석 방법에 따른 데이터 모델의 분석 결과를 비교하고 평가한다.
[12데과03-06] 다양한 분석 방법을 비교하고 평가하여 분석 목적에 가장 적합한 분석 방법을 적용한다.

(가) 성취기준 해설

- [12데과03-03] 데이터의 속성을 바탕으로 유사도가 높은 데이터끼리 묶어 다수의 군집으로 나누고 군집 내 유사성과 군집 간 상이성을 설명할 수 있어야 한다.
- [12데과03-04] 장바구니 분석의 사례를 통해 연관 분석의 원리를 이해하고, 데이터 속성 간 지지도, 신뢰도, 향상도를 측정하며 데이터 속성 간 연관 규칙을 찾아낼 수 있어야 한다.

• [12데과03-05] 데이터 특성과 분석 목적에 적합한 평가 방법(예측률, 정확도 등)을 선정하고, 이를 바탕으로 데이터 분석 결과를 해석하고 그 의미를 판단할 수 있어야 한다.

(나) 성취기준 적용 시 고려 사항

• 데이터 모델링 과정에서 다양한 데이터 분석 방법을 비교하여 활용하고, 데이터 해석 과정에서 등장하는 여러 견해를 상호 존중하고 비판적 시각으로 바라봄으로써 합리적으로 문제를 해결할 수 있도록 교수 · 학습을 구성하도록 한다.
• 데이터 분석을 위한 도구는 학습자의 디지털 역량을 사전에 파악하여 학습자의 인지적인 부담이 적은 방향으로 선정하도록 한다. 학습에 사용하는 기기나 운영 체제에 비교적 독립적인 소프트웨어나 프로그래밍 언어를 활용하여 다양한 학습 환경에서 학습자의 접근성을 보장하도록 한다.
• 최소 성취수준을 보장하기 위하여 교수자가 제시한 데이터 분석 과정을 바탕으로 데이터 모델의 개념을 제시하고, 여러 가지 분석 방법을 구별하여 설명할 수 있는지 평가한다. 또한, 비교적 명확하게 해석 결과가 도출되는 데이터셋을 제공하여 학습자가 최소한의 데이터 분석 과정을 체험하고 의미를 인식할 수 있는 학습 과정을 제공하도록 한다.

4 데이터 과학 프로젝트

[12데과04-01] 분야별 데이터 과학의 적용 사례를 조사하여 분석하고, 데이터로 해결 가능한 주제를 찾아 적합성을 판단한다.
[12데과04-02] 수집된 데이터를 탐색적으로 분석하여 데이터 속 의미를 파악하고, 문제 해결을 위한 창의적인 방법을 구상한다.
[12데과04-03] 데이터 분석을 진행할 때, 2개 이상의 방법을 사용하여 분석하고 결과를 비교한다.
[12데과04-04] 복잡하고 어려운 문제라도 끝까지 해결하기 위한 자세를 갖추고 분석하여, 분석 결과에 대한 의미를 해석한다.
[12데과04-05] 분석을 위한 목적부터 데이터 수집 및 분석에 이르는 전 과정을 성찰하고, 사회적 영향을 고려하여 분석 결과의 활용방안을 탐색한다.

(가) 성취기준 해설

• [12데과04-03] 동일한 데이터를 기반으로 서로 다른 데이터 전처리, 데이터 분석 방법을 적용한 결과를 비교하여 가장 적절한 데이터 모델링 방법을 설계할 수 있어야 한다.
• [12데과04-05] 데이터 분석을 통한 문제 해결 과정에서 파생되는 사회적인 영향, 윤리적인 문제를 성찰하고 데이터 모델을 수정한 후, 문제 해결 결과를 일반화하고 공유할 수 있어야 한다.

(나) 성취기준 적용 시 고려 사항

• 기후위기, 환경 문제, 에너지 문제 등의 주제를 가정, 학교, 지역 및 지구 차원에서 프로젝트 주제로 고려하여, 인류가 당면한 여러 가지 생태적 문제를 데이터 과학 프로젝트를 통해 심층적으로 탐구하도록 한다.
• 여러 가지 데이터 분석 방법 중 기계학습을 통한 분석 방법을 포함하여 분석 결과를 비교함으로써 디지털 · 인공지능 소양을 기를 수 있도록 교수 · 학습을 구성하도록 한다.
• 프로젝트 과정에서 온라인 공유 환경을 제공하여 온오프라인의 협업이 모두 일어날 수 있도록 하고, 협업 과정이 온라인 문서에 기록될 수 있도록 한다. 특히, 주제 선정, 데이터 수집 등에서 아이디어를 발산하고 수렴하는 과정이 기록되도록 하여 프로젝트 내 문제 해결 과정이 드러나는 데에 중점을 두도록 한다.
• 민주시민으로서 데이터 과학 프로젝트를 통해 생산한 정보의 사회적 영향 및 파급력에 대하여 논의할 수 있도록 문제기반 학습을 구성하고, 협력학습을 통해 학생들이 토론이나 토의 과정의 기회를 갖도록 한다.

3 교수·학습 및 평가

가. 교수·학습

1 교수 · 학습의 방향

(가) 실제적인 삶의 맥락에서 컴퓨팅을 통해 문제를 해결하도록 하는 학습 과제를 제시하여 학습자가 과제를 스스로 해결하는 과정에서 자연스럽게 컴퓨팅 사고력, 디지털 문화 소양, 인공지능 소양을 함양할 수 있도록 지도한다.
(나) 학습자의 흥미와 다양성을 고려하여 학습 소재, 학습 환경 및 학습 과정에 대한 선택의 기회를 제공하고, 교수-학습의 설계 과정에 학습자 참여 기회를 증진하는 등 학습자 맞춤형 교수 · 학습을 통해 역량 함양을 위한 깊이 있는 학습 지도 방안을 구성한다.
(다) 데이터 과학 과목의 지식 · 이해, 과정 · 기능을 활용하여 민주시민교육, 생태전환 교육 등 현 시대가 당면한 여러 사회문제와 더불어 지속가능발전 등의 범교과 주제를 교수 · 학습 과제로 제시하여 주도성 있는 문제 해결 경험을 제공한다.
(라) 디지털 교육 환경에 적응할 수 있도록 온오프라인 연계 수업, 다양한 디지털 도구의 활용 등을 통해 디지털 도구에 대한 인지적 부담은 최소화하고, 활용에 대한 경험은 높일 수 있도록 수업을 구성한다.
(마) 데이터 과학에 대한 이해를 통해 디지털 사회에서 데이터 기반의 사회 변화에 적극적으로 대응할 수 있는 태도와 능력을 함양할 수 있도록 교수 · 학습을 설계한다.
(바) 프로젝트형 실습은 협업을 통해 의사 소통 능력, 협력적 문제 해결력, 공유의 가치 인식 등을 함양하도록 한다.
(사) 특정 데이터 과학 기술이나 도구의 사용법 습득에 치중하지 않도록 유의하고, 문제 해결을 위한 데이터 과학 기술의 활용, 프로젝트 설계 및 수행을 통해 데이터 문해력과 인공지능 소양을 함양하는 데 중점을 둔다.

2 교수 · 학습 방법

(가) 교과 역량을 함양하기 위해 문제기반학습, 프로젝트 기반학습, 디자인기반학습, 짝 프로그래밍, 탐구학습 등 각 영역의 핵심 아이디어를 습득하는 데 적절한 교수 · 학습 방법을 선택하여 활용한다.
(나) 학습자 개인별로 학습하는 속도가 다양할 수 있음을 고려하고, 최소 성취수준을 보장할 수 있도록 학습관리시스템(LMS)을 활용하여 온라인 학습자료를 제작 및 제공함으로써 학습 격차를 최소화하도록 노력한다.
(다) 데이터를 수집하고 분석하는 과정에서 토의 · 토론을 통해 데이터 편향성, 윤리 문제 등 사회적 영향력을 판단하여 의사결정할 수 있는 과정을 포함한다. 데이터 모델링 과정에서는 토의 · 토론을 통해 다양한 데이터 분석 방법을 비교하여 선정하고 비판적 시각으로 결과를 해석할 수 있도록 안내한다.
(라) 학습자의 진로와 연계된 주제의 프로젝트를 선택하도록 하여, 학습자가 데이터 과학 기술의 활용과 자신의 미래를 연결하여 생각할 수 있도록 수업을 구성한다.
(마) 프로젝트 활동 과정에서 협업에 필요한 다양한 디지털 도구를 활용할 수 있으며, 학생들이 손쉽게 활용할 수 있는 디지털 도구를 도입하여 원격수업이나 협업 활동에서 디지털 도구 활용 방법을 익히는 데 인지적 부담을 최소화한다.

나. 평가

(1) 평가의 방향

(가) 평가 항목은 컴퓨팅 사고력, 디지털 문화 소양, 인공지능 소양의 하위 요소를 기반으로 구체화한다.

(나) 평가 내용은 지식 · 이해뿐 아니라, 과정 · 기능, 가치 · 태도의 측면 등을 다면적으로 반영하고 과정을 중시하는 평가를 통해 학생의 성장과 발달을 돕는 평가를 실현한다.

(다) 구체적인 평가 루브릭을 학생과 함께 구성하는 과정을 통해 학생이 자신의 학습 수준을 파악하고 스스로 학습을 성찰할 수 있는 기회를 제공하여, 적극적이고 능동적인 학습이 이루어지도록 한다.

(라) 작성한 프로그램의 정확성, 효율성과 더불어 프로그램 설계 과정의 논리성과 실습 과정을 통해 데이터 모델링의 과정을 이해하고 있는지에 중점을 두고 평가한다.

(마) 모둠별 탐구 활동의 성과물에 대한 평가뿐만 아니라 협업 및 발표, 토론 수행 등의 전 과정에서 합리적이고 객관적인 평가가 이루어질 수 있도록 평가기준과 구체적인 체크리스트를 마련하고, 이를 교사 평가뿐만 아니라 자기 평가, 동료 평가의 도구로 활용한다.

(2) 평가 방법

(가) 성취기준을 분석하고 재구성하여 지필평가에 국한하지 않고, 학생의 성장에 기여할 수 있는 평가 포트폴리오를 계획한다. 예를 들면, 관찰 평가, 서술형평가, 수행평가 등을 활용하거나, 자기 평가, 동료 평가 등과 같은 다면적 평가를 실행한다.

(나) 평가 내용이나 방법에 따라 다양한 디지털 도구(프로그램 자동 평가시스템(online judge 등), 학습관리시스템(LMS) 등)를 활용할 수 있으며, 평가 이전에 학생이 디지털 도구를 다룰 수 있도록 교육하여 평가의 불이익이 없도록 계획한다.

(다) 실습 과제를 평가할 경우, 산출물 평가와 더불어 과제 해결 과정을 꾸준히 관찰하여 학생의 학습 과정을 종합적으로 평가한다. 특히 프로젝트형 과제 수행 시 학습자의 수행 과정을 온오프라인 상에 누적하도록 하여 전반적인 과정을 종합적으로 평가하도록 한다.

(라) 협업 프로젝트를 평가할 때는 학습자별 역할을 구체적으로 기록하고, 동료 평가를 통해 모둠원에서 활동했던 비중을 논의하여 제시하도록 함으로써 최대한 공정성을 확보한다.

(마) 효율적인 평가를 위하여 다양한 디지털 도구를 활용할 수 있으나 학생이 디지털 도구 활용의 미숙으로 인해 평가에 불이익을 받지 않도록 디지털 도구의 사용법을 익히는 데 부담을 최소화하거나 충분히 익힐 기회를 제공한다.

2. 성취수준

1. 성취기준별 성취수준

1 데이터 과학의 이해

성취기준	성취기준별 성취수준	
[12데과01-01] 데이터 과학의 개념을 이해하고, 문제해결 사례를 데이터 기반 의사결정 상황에 적용한다.	A	실제 사례와 활용 분야를 바탕으로 데이터 과학의 개념을 정확하게 설명하고, 데이터 과학 문제해결 사례를 다양하게 탐색하여 데이터 기반 의사결정의 중요성을 내면화하여 표현할 수 있다.
	B	실제 사례와 활용 분야를 바탕으로 데이터 과학의 개념을 정확하게 설명하고, 데이터 과학 문제해결 사례를 탐색하여 데이터 기반 의사결정의 중요성을 인식할 수 있다.
	C	데이터 과학의 개념을 설명하고, 데이터 과학 문제해결 사례를 탐색하여 데이터 기반 의사결정의 중요성을 인식할 수 있다.
	D	데이터 과학의 개념을 설명하고, 데이터 과학 문제해결 사례를 부분적으로 탐색하여 데이터 기반 의사결정의 중요성을 수용할 수 있다.
	E	데이터 과학의 개념을 인지하고 데이터 과학 문제해결 사례를 부분적으로 탐색할 수 있다.
[12데과01-02] 정형 데이터와 비정형 데이터를 구분하고, 데이터 속성에서 데이터의 잠재적 가치를 파악한다.	A	데이터의 형태와 속성을 정확하게 설명하고, 정형 데이터와 비정형 데이터를 정확하게 구분하며 데이터의 속성별 의미와 속성 간 관계를 정확하게 파악하여 수집 데이터의 잠재적 가치를 내면화하여 표현할 수 있다.
	B	데이터의 형태와 속성을 정확하게 설명하고, 정형 데이터와 비정형 데이터를 구분하며 데이터의 속성별 의미와 속성 간 관계를 파악하여 수집 데이터의 잠재적 가치를 인식할 수 있다.
	C	데이터의 형태와 속성을 설명하고, 정형 데이터와 비정형 데이터를 구분하며 데이터의 속성별 의미와 속성 간 관계를 파악하여 수집 데이터의 잠재적 가치를 인식할 수 있다.
	D	데이터의 형태와 속성을 설명하고, 데이터별 형태와 속성을 일부 파악하여 수집 데이터의 잠재적 가치를 수용할 수 있다.
	E	데이터의 형태와 속성을 인지하고, 데이터별 형태와 속성을 일부 파악할 수 있다.
[12데과01-03] 데이터셋의 집합인 데이터베이스를 이해하고, 서로 다른 셋의 데이터를 분석이 가능한 형태로 통합하는 것의 의미를 파악한다.	A	데이터셋과 데이터베이스의 개념을 정확하게 설명하고, 데이터 통합이 필요한 상황과 데이터 통합의 의미를 파악할 수 있다.
	B	데이터셋과 데이터베이스의 개념을 정확하게 설명하고, 데이터 통합의 의미를 파악할 수 있다.
	C	데이터셋과 데이터베이스의 개념을 설명하고, 데이터 통합의 의미를 파악할 수 있다.
	D	데이터셋과 데이터베이스의 개념을 설명하고, 데이터 통합의 의미를 부분적으로 파악할 수 있다.
	E	데이터셋과 데이터베이스의 개념을 인지하고, 데이터 통합의 의미를 부분적으로 파악할 수 있다.
[12데과01-04] 데이터로 인한 사회 변화를 인식하고, 진로 및 직업과 관련한 데이터 기반 문제해결 사례를 분석한다.	A	진로 및 직업과 관련한 데이터 기반 문제해결 사례를 탐색함으로써 데이터로 인한 사회 변화를 다양하게 제시하고, 데이터 과학을 통한 진로설계에 능동적으로 참여할 수 있다.
	B	진로 및 직업과 관련한 데이터 기반 문제해결 사례를 탐색함으로써 데이터로 인한 사회 변화를 다양하게 제시하고, 데이터 과학을 통한 진로설계에 참여할 수 있다.
	C	진로 및 직업과 관련한 데이터 기반 문제해결 사례를 탐색함으로써 데이터로 인한 사회 변화를 제시하고, 데이터 과학을 통한 진로설계에 참여할 수 있다.
	D	진로 및 직업과 관련하여 주어진 데이터 기반 문제해결 사례를 바탕으로 데이터로 인한 사회 변화를 인식하고, 데이터 과학을 통한 진로설계 필요성을 수용할 수 있다.
	E	진로 및 직업과 관련하여 주어진 데이터 기반 문제해결 사례를 바탕으로 데이터로 인한 사회 변화를 인식할 수 있다.

2 데이터 준비와 분석

성취기준	성취기준별 성취수준	
[12데과02-01] 데이터를 편향되지 않도록 수집하고, 수집된 데이터의 특성을 분석한다.	A	데이터가 편향되지 않도록 수집하는 적극적인 자세를 갖추고, 수집된 데이터의 특성을 다양한 관점으로 분석할 수 있다.
	B	데이터가 편향되지 않도록 수집하는 자세를 갖추고, 수집된 데이터의 특성을 다양한 관점으로 분석할 수 있다.
	C	데이터가 편향되지 않도록 수집하는 자세를 갖추고, 수집된 데이터의 특성을 분석할 수 있다.
	D	데이터 편향 방지의 필요성을 수용하고, 수집된 데이터의 특성을 분석할 수 있다.
	E	데이터 편향 방지의 필요성을 수용하고, 수집된 데이터의 특성을 일부 분석할 수 있다.
[12데과02-02] 이상치와 결측치 탐색 및 정규화를 통해 전처리하여 오류 가능성을 최소화하고, 데이터 분석을 위해 시각화한다.	A	데이터 전처리의 필요성과 방법을 정확하게 제시하고, 데이터 특성에 맞는 시각화, 이상치와 결측치 탐색 및 처리, 정규화의 적절한 활용을 통해 데이터의 불확실성과 오류 가능성을 내면화하여 표현할 수 있다.
	B	데이터 전처리의 필요성과 방법을 정확하게 제시하고, 데이터 시각화, 이상치와 결측치 처리, 정규화의 활용을 통해 데이터의 불확실성과 오류 가능성을 인식할 수 있다.
	C	데이터 전처리의 필요성과 방법을 제시하고, 데이터 시각화, 이상치와 결측치 처리, 정규화의 활용을 통해 데이터의 불확실성과 오류 가능성을 인식할 수 있다.
	D	데이터 전처리의 필요성과 방법을 제시하고, 일부 데이터 시각화, 이상치 또는 결측치 처리를 통해 데이터의 불확실성과 오류 가능성을 수용할 수 있다.
	E	데이터 전처리의 필요성과 방법을 제시하고, 일부 데이터 시각화, 이상치 또는 결측치 처리를 할 수 있다.
[12데과02-03] 데이터를 분석하기 위해 데이터 속성 간의 관계를 파악하고 통합한다.	A	데이터 분석 목적에 적합한 데이터 속성을 구분하고, 데이터 속성 간의 관계를 파악하고 통합할 수 있다.
	B	데이터 분석 목적에 적합한 데이터 속성을 구분하고, 데이터 속성 간의 관계를 파악할 수 있다.
	C	데이터 속성을 구분하고, 데이터 속성 간의 관계를 파악할 수 있다.
	D	데이터 속성을 구분하고, 데이터 속성 간의 관계를 일부 파악할 수 있다.
	E	데이터 속성을 인지하고, 데이터 속성 간의 관계를 일부 파악할 수 있다.
[12데과02-04] 동일한 데이터를 서로 다른 분석 방법을 적용하여 분석 결과를 비교한다.	A	합리적인 데이터 분석 방법을 다양하게 제시할 수 있으며, 동일한 데이터에 서로 다른 데이터 분석 방법을 적용하고 분석 결과를 비교할 수 있다.
	B	합리적인 데이터 분석 방법을 다양하게 제시할 수 있으며, 동일한 데이터에 적용 가능한 서로 다른 데이터 분석 방법을 비교할 수 있다.
	C	데이터 분석 방법을 다양하게 제시할 수 있으며, 동일한 데이터에 적용 가능한 서로 다른 데이터 분석 방법을 비교할 수 있다.
	D	데이터 분석 방법을 다양하게 제시할 수 있으며, 동일한 데이터에 적용 가능한 서로 다른 데이터 분석 방법을 부분적으로 비교할 수 있다.
	E	동일한 데이터에 적용 가능한 서로 다른 데이터 분석 방법을 인지할 수 있다.

3 데이터 모델링과 평가

성취기준	성취기준별 성취수준	
[12데과03-01] 데이터 모델 개념을 이해하고 데이터 분석에 활용할 수 있는 도구를 탐색한다.	A	데이터 모델의 개념을 정확하게 설명하고, 데이터 분석을 위한 도구를 다양하게 탐색할 수 있다.
	B	데이터 모델의 개념을 정확하게 설명하고, 데이터 분석을 위한 도구를 탐색할 수 있다.
	C	데이터 모델의 개념을 설명하고, 데이터 분석을 위한 도구를 탐색할 수 있다.
	D	데이터 모델의 개념을 설명하고, 데이터 분석을 위한 도구를 부분적으로 탐색할 수 있다.
	E	데이터 모델의 개념을 인지하고, 데이터 분석을 위한 도구를 부분적으로 탐색할 수 있다.
[12데과03-02] 동일한 데이터를 통계적 회귀모델과 기계학습을 통한 회귀모델로 분석하여 결과 해석 내용을 비교한다.	A	회귀 분석의 개념을 정확하게 설명하고, 동일한 데이터를 통계적 회귀모델과 기계학습 기반 회귀모델로 분석한 후 분석 결과에 대한 의미를 비판적으로 해석할 수 있다.
	B	회귀 분석의 개념을 정확하게 설명하고, 동일한 데이터를 통계적 회귀모델과 기계학습 기반 회귀모델로 분석한 후 분석 결과에 대한 의미를 해석할 수 있다.
	C	회귀 분석의 개념을 설명하고, 동일한 데이터를 통계적 회귀모델과 기계학습 기반 회귀모델로 분석한 후 분석 결과에 대한 의미를 해석할 수 있다.
	D	회귀 분석의 개념을 설명하고, 동일한 데이터를 통계적 회귀모델과 기계학습 기반 회귀모델로 분석한 결과에 대한 의미를 부분적으로 해석할 수 있다.
	E	회귀 분석의 개념을 인지하고, 동일한 데이터를 통계적 회귀모델과 기계학습 기반 회귀모델로 분석한 결과에 대한 의미를 부분적으로 해석할 수 있다.
[12데과03-03] 데이터의 속성에 대한 유사성을 측정하고 분석하여 군집을 형성하고, 군집 분석 결과의 의미를 해석한다.	A	군집 분석의 개념을 정확하게 설명하고, 데이터의 여러 속성에 대한 유사성을 바탕으로 군집을 형성한 후 분석 결과에 대한 의미를 비판적으로 해석할 수 있다.
	B	군집 분석의 개념을 정확하게 설명하고, 데이터의 여러 속성에 대한 유사성을 바탕으로 군집을 형성한 후 분석 결과에 대한 의미를 해석할 수 있다.
	C	군집 분석의 개념을 설명하고, 데이터의 여러 속성에 대한 유사성을 바탕으로 군집을 형성한 후 분석 결과에 대한 의미를 해석할 수 있다.
	D	군집 분석의 개념을 설명하고, 데이터의 여러 속성에 대한 유사성을 바탕으로 군집을 형성한 후 분석 결과에 대한 의미를 부분적으로 해석할 수 있다.
	E	군집 분석의 개념을 인지하고, 데이터의 여러 속성에 대한 유사성을 바탕으로 군집을 형성한 후 분석 결과에 대한 의미를 부분적으로 해석할 수 있다.

[12데과03-04] 데이터 간의 관계를 분석하고 상호 연관성을 파악하여 결과의 의미를 해석한다.	A	연관 분석의 개념을 정확하게 설명하고, 데이터 간의 관계를 분석하여 상호 연관성을 파악한 후 분석 결과에 대한 의미를 비판적으로 해석할 수 있다.
	B	연관 분석의 개념을 정확하게 설명하고, 데이터 간의 관계를 분석하여 상호 연관성을 파악한 후 분석 결과에 대한 의미를 해석할 수 있다.
	C	연관 분석의 개념을 설명하고, 데이터 간의 관계를 분석하여 상호 연관성을 파악한 후 분석 결과에 대한 의미를 해석할 수 있다.
	D	연관 분석의 개념을 설명하고, 데이터 간의 관계를 분석하여 상호 연관성을 파악한 후 분석 결과에 대한 의미를 부분적으로 해석할 수 있다.
	E	연관 분석의 개념을 인지하고, 데이터 간의 관계를 분석하여 상호 연관성을 파악한 후 분석 결과에 대한 의미를 부분적으로 해석할 수 있다.
[12데과03-05] 데이터 분석 방법에 따른 데이터 모델의 분석 결과를 비교하고 평가한다.	A	여러 가지 데이터 분석 방법에 따른 데이터 모델의 분석 결과를 비교하고 평가하여 데이터에 대한 다양한 해석을 제시하고 비판적으로 수용할 수 있다.
	B	여러 가지 데이터 분석 방법에 따른 데이터 모델의 분석 결과를 비교하고 평가하여 데이터에 대한 다양한 해석을 비판적으로 수용할 수 있다.
	C	데이터 분석 방법에 따른 데이터 모델의 분석 결과를 평가하여 데이터에 대한 다양한 해석을 비판적으로 수용할 수 있다.
	D	데이터 분석 방법에 따른 데이터 모델의 분석 결과를 평가하여 데이터에 대한 해석을 수용할 수 있다.
	E	데이터 분석 방법에 따른 데이터 모델의 분석 결과를 부분적으로 평가하여 데이터에 대한 해석을 수용할 수 있다.
[12데과03-06] 다양한 분석 방법을 비교하고 평가하여 분석 목적에 가장 적합한 분석 방법을 적용한다.	A	다양한 분석 방법을 성능, 효율성 등 여러 기준으로 비교하고 평가하며, 적절한 분석 방법을 선택하여 적용하는 자세를 내면화하여 표현할 수 있다.
	B	다양한 분석 방법을 성능, 효율성 등 여러 기준으로 비교하고 평가하며, 적절한 분석 방법을 선택하여 적용하는 자세의 중요성을 인식할 수 있다.
	C	다양한 분석 방법을 비교하고 평가하며, 적절한 분석 방법을 선택하여 적용하는 자세의 중요성을 인식할 수 있다.
	D	다양한 분석 방법을 비교하고 평가하며, 적절한 분석 방법을 선택하여 적용하는 자세를 수용할 수 있다.
	E	다양한 분석 방법을 비교하고, 적절한 분석 방법을 선택하여 적용하는 자세를 수용할 수 있다.

4 데이터 과학 프로젝트

성취기준	성취기준별 성취수준	
[12데과04-01] 분야별 데이터 과학의 적용 사례를 조사하여 분석하고, 데이터로 해결 가능한 주제를 찾아 적합성을 판단한다.	A	분야별 데이터 과학의 주제를 다양하게 조사하며, 데이터로 해결 가능한 주제를 정확하게 설명하고 적합성을 판단할 수 있다.
	B	분야별 데이터 과학의 주제를 조사하며, 데이터로 해결 가능한 주제를 정확하게 설명하고 적합성을 판단할 수 있다.
	C	분야별 데이터 과학의 주제를 조사하며, 데이터로 해결 가능한 주제를 설명할 수 있다.
	D	데이터 과학의 주제를 조사하며, 데이터로 해결 가능한 주제를 설명할 수 있다.
	E	데이터 과학의 주제를 조사하고, 데이터로 해결 가능한 주제를 인지할 수 있다.
[12데과04-02] 수집된 데이터를 탐색적으로 분석하여 데이터 속 의미를 파악하고, 문제해결을 위한 창의적인 방법을 구상한다.	A	탐색적 데이터 분석을 정확하게 설명하고, 탐색적 데이터 분석으로 데이터 속 의미를 다각도로 파악함으로써 문제를 해결하기 위한 창의적인 방법을 고민하는 자세를 내면화하여 표현할 수 있다.
	B	탐색적 데이터 분석을 정확하게 설명하고, 탐색적 데이터 분석으로 데이터 속 의미를 파악함으로써 문제를 해결하기 위한 창의적인 방법을 고민하는 자세를 인식할 수 있다.
	C	탐색적 데이터 분석을 설명하고, 탐색적 데이터 분석으로 데이터 속 의미를 파악함으로써 문제를 해결하기 위한 창의적인 방법을 고민하는 자세를 인식할 수 있다.
	D	탐색적 데이터 분석을 설명하고, 탐색적 데이터 분석으로 데이터 속 의미를 부분적으로 파악함으로써 문제를 해결하기 위한 창의적인 방법을 고민하는 자세를 수용할 수 있다.
	E	탐색적 데이터 분석을 인지하고, 탐색적 데이터 분석으로 데이터 속 의미를 부분적으로 파악할 수 있다.
[12데과04-03] 데이터 분석을 진행할 때, 2개 이상의 방법을 사용하여 분석하고 결과를 비교한다.	A	동일한 데이터를 기반으로 서로 다른 데이터 분석 방법을 적용한 결과를 다각도로 비교하고, 가장 적절한 데이터 모델링 방법을 설계하는 자세를 내면화하여 표현할 수 있다.
	B	동일한 데이터를 기반으로 서로 다른 데이터 분석 방법을 적용한 결과를 다각도로 비교하고, 적절한 데이터 모델링 방법을 설계하는 자세를 인식할 수 있다.
	C	동일한 데이터를 기반으로 서로 다른 데이터 분석 방법을 적용한 결과를 비교하고, 적절한 데이터 모델링 방법을 설계하는 자세를 인식할 수 있다.
	D	동일한 데이터를 기반으로 서로 다른 데이터 분석 방법을 적용한 결과를 비교하고, 적절한 데이터 모델링 방법을 설계하는 자세를 수용할 수 있다.
	E	동일한 데이터를 기반으로 서로 다른 데이터 분석 방법을 적용한 결과를 부분적으로 비교할 수 있다.

[12데과04-04] 복잡하고 어려운 문제라도 끝까지 해결하기 위한 자세를 갖추고 분석하여, 분석 결과에 대한 의미를 해석한다.	A	분석 결과에 대한 의미를 비판적으로 해석하고, 복잡하고 어려운 문제를 끝까지 해결하기 위해 노력하는 태도를 내면화하여 표현할 수 있다.
	B	분석 결과에 대한 의미를 비판적으로 해석하고, 복잡하고 어려운 문제를 끝까지 해결하기 위해 노력하는 태도를 인식할 수 있다.
	C	분석 결과에 대한 의미를 해석하고, 복잡하고 어려운 문제를 끝까지 해결하기 위해 노력하는 태도를 인식할 수 있다.
	D	분석 결과에 대한 의미를 해석하고, 복잡하고 어려운 문제를 끝까지 해결하기 위해 노력하는 태도를 수용할 수 있다.
	E	분석 결과에 대한 의미를 일부 해석하고, 복잡하고 어려운 문제를 끝까지 해결하기 위해 노력하는 태도를 수용할 수 있다.
[12데과04-05] 분석을 위한 목적부터 데이터 수집 및 분석에 이르는 전 과정을 성찰하고, 사회적 영향을 고려하여 분석 결과의 활용방안을 탐색한다.	A	사회적 영향을 고려하여 분석 결과를 활용하는 방법을 다양하게 탐색하고, 일반화 및 공유 과정에서 윤리 문제 등 사회적 영향을 인식하는 자세를 내면화하여 표현할 수 있다.
	B	사회적 영향을 고려하여 분석 결과를 활용하는 방법을 다양하게 탐색하고, 일반화 및 공유 과정에서 윤리 문제 등 사회적 영향을 인식하는 자세를 수용할 수 있다.
	C	사회적 영향을 고려하여 분석 결과를 활용하는 방법을 탐색하고, 일반화 및 공유 과정에서 윤리 문제 등 사회적 영향을 인식하는 자세를 수용할 수 있다.
	D	사회적 영향을 고려하여 분석 결과를 활용하는 방법을 부분적으로 탐색하고, 일반화 및 공유 과정에서 윤리 문제 등 사회적 영향을 인식하는 자세를 수용할 수 있다.
	E	사회적 영향을 고려하여 분석 결과를 활용하는 방법을 부분적으로 탐색할 수 있다.

2. 영역별 성취수준

1 데이터 과학의 이해

성취기준	영역별 성취수준		
(1) 데이터 과학의 이해	A	지식 · 이해	데이터 과학의 개념, 데이터의 형태와 속성, 데이터셋과 데이터베이스의 개념을 정확하게 설명할 수 있다.
		과정 · 기능	데이터 과학 문제해결 사례를 다양하게 탐색하고, 데이터의 형태와 속성, 데이터 통합이 필요한 상황과 데이터 통합의 의미를 정확하게 파악하며, 진로 및 직업과 관련한 데이터 과학의 문제해결 사례를 탐색함으로써 데이터로 인한 사회 변화를 다양하게 제시할 수 있다.
		가치 · 태도	데이터 기반 의사결정의 중요성과 데이터의 잠재적 가치를 내면화하고, 데이터 과학을 통한 진로설계에 능동적으로 참여할 수 있다.
	B	지식 · 이해	데이터 과학의 개념을 정확하게 설명하고, 데이터의 형태와 속성, 데이터셋과 데이터베이스의 개념을 설명할 수 있다.
		과정 · 기능	데이터 과학 문제해결 사례를 다양하게 탐색하고, 데이터의 형태와 속성, 데이터 통합이 필요한 상황과 데이터 통합의 의미를 정확하게 파악하며, 진로 및 직업과 관련한 데이터 과학의 문제해결 사례를 탐색함으로써 데이터로 인한 사회 변화를 제시할 수 있다.
		가치 · 태도	데이터 기반 의사결정의 중요성과 데이터의 잠재적 가치를 내면화하고, 데이터 과학을 통한 진로설계에 참여할 수 있다.
	C	지식 · 이해	데이터 과학의 개념, 데이터의 형태와 속성, 데이터셋과 데이터베이스의 개념을 설명할 수 있다.
		과정 · 기능	데이터 과학 문제해결 사례를 탐색하고, 데이터의 형태와 속성, 데이터 통합이 필요한 상황과 데이터 통합의 의미를 파악하며, 진로 및 직업과 관련한 데이터 과학의 문제해결 사례를 탐색함으로써 데이터로 인한 사회 변화를 제시할 수 있다.
		가치 · 태도	데이터 기반 의사결정의 중요성과 데이터의 잠재적 가치를 인식하고, 데이터 과학을 통한 진로설계에 참여할 수 있다.
	D	지식 · 이해	데이터 과학의 개념을 설명하고, 데이터의 형태와 속성, 데이터셋과 데이터베이스의 개념을 인지할 수 있다.
		과정 · 기능	데이터 과학 문제해결 사례를 탐색하고, 데이터의 형태와 속성, 데이터 통합이 필요한 상황과 데이터 통합의 의미를 파악하며, 진로 및 직업과 관련한 데이터 과학의 문제해결 사례를 탐색함으로써 데이터로 인한 사회 변화를 인식할 수 있다.
		가치 · 태도	데이터 기반 의사결정의 중요성과 데이터의 잠재적 가치를 인식하고, 데이터 과학을 통한 진로설계의 필요성을 수용할 수 있다.
	E	지식 · 이해	데이터 과학의 개념, 데이터의 형태와 속성, 데이터셋과 데이터베이스의 개념을 인지할 수 있다.
		과정 · 기능	데이터 과학 문제해결 사례를 일부 탐색하고, 데이터의 형태와 속성, 데이터 통합이 필요한 상황과 데이터 통합의 의미를 부분적으로 파악하며, 진로 및 직업과 관련한 데이터 과학의 문제해결 사례를 탐색함으로써 데이터로 인한 사회 변화를 인식할 수 있다.
		가치 · 태도	데이터 기반 의사결정의 중요성, 데이터의 잠재적 가치, 데이터 과학을 통한 진로설계의 필요성을 수용할 수 있다.

2 데이터 준비와 분석

성취기준	영역별 성취수준		
(2) 데이터 준비와 분석	A	지식 · 이해	데이터 전처리의 필요성을 정확하게 설명하고 데이터 분석 목적에 맞는 데이터 속성을 분별하며, 여러 합리적인 데이터 분석 방법을 제시할 수 있다.
		과정 · 기능	데이터의 특성을 다양한 관점으로 분석하고 데이터 특성에 맞는 시각화, 이상치와 결측치 탐색 및 처리, 정규화를 하며, 데이터 속성 간의 관계를 파악 및 통합하고 서로 다른 데이터 분석 방법을 적용한 후 분석 결과를 비교할 수 있다.
		가치 · 태도	데이터가 편향되지 않도록 수집하는 적극적인 자세와 데이터의 불확실성과 오류 가능성을 내면화할 수 있다.
	B	지식 · 이해	데이터 전처리의 필요성을 정확하게 설명하고 데이터 속성을 분별하며, 여러 데이터 분석 방법을 제시할 수 있다.
		과정 · 기능	데이터의 특성을 다양한 관점으로 분석하고 데이터 특성에 맞는 시각화, 이상치와 결측치 탐색 및 처리, 정규화를 하며, 데이터 속성 간의 관계를 파악하고 서로 다른 데이터 분석 방법을 비교할 수 있다.
		가치 · 태도	데이터가 편향되지 않도록 수집하는 적극적인 자세를 갖추고 데이터의 불확실성과 오류 가능성을 인식할 수 있다.
	C	지식 · 이해	데이터 전처리의 필요성을 설명하고 데이터 속성을 분별하며, 여러 데이터 분석 방법을 제시할 수 있다.
		과정 · 기능	데이터의 특성을 분석하고 데이터 시각화, 이상치와 결측치 탐색 및 처리, 정규화를 하며, 데이터 속성 간의 관계를 파악하고 서로 다른 데이터 분석 방법을 비교할 수 있다.
		가치 · 태도	데이터가 편향 방지의 필요성, 데이터의 불확실성과 오류 가능성을 인식할 수 있다.
	D	지식 · 이해	데이터 전처리의 필요성을 설명하고, 데이터 속성과 여러 데이터 분석 방법을 인지할 수 있다.
		과정 · 기능	데이터의 특성을 분석하고 데이터 시각화, 이상치와 결측치 탐색 및 처리, 정규화를 하며, 데이터 속성 간의 관계를 일부 파악하며 서로 다른 데이터 분석 방법을 부분적으로 비교할 수 있다.
		가치 · 태도	데이터가 편향 방지의 필요성을 인식하고, 데이터의 불확실성과 오류 가능성을 수용할 수 있다.
	E	지식 · 이해	데이터 전처리의 필요성, 데이터 속성, 여러 데이터 분석 방법을 인지할 수 있다.
		과정 · 기능	데이터의 특성을 부분적으로 분석하고, 데이터 속성 간의 관계를 일부 파악하며 서로 다른 데이터 분석 방법을 부분적으로 비교할 수 있다.
		가치 · 태도	데이터가 편향 방지의 필요성, 데이터의 불확실성과 오류 가능성을 수용할 수 있다.

3 데이터 모델링과 평가

성취기준	영역별 성취수준		
(3) 데이터 모델링과 평가	A	지식 · 이해	데이터 모델의 개념, 회귀 분석, 군집 분석, 연관 분석을 정확하게 설명할 수 있다.
		과정 · 기능	데이터 분석을 위한 도구를 다양하게 탐색하고 데이터 모델의 분석 결과에 대한 의미를 비판적으로 해석하며, 여러 가지 데이터 분석 방법에 따른 분석 결과를 종합적으로 비교하고 평가할 수 있다.
		가치 · 태도	데이터에 대한 다양한 해석을 제시하고 비판적으로 수용하며, 적절한 분석 방법을 선택하여 적용하는 자세를 내면화할 수 있다.
	B	지식 · 이해	데이터 모델의 개념을 정확하게 설명하고, 회귀 분석, 군집 분석, 연관 분석을 설명할 수 있다.
		과정 · 기능	데이터 분석을 위한 도구를 다양하게 탐색하고 데이터 모델의 분석 결과에 대한 의미를 비판적으로 해석하며, 여러 가지 데이터 분석 방법에 따른 분석 결과를 비교하고 평가할 수 있다.
		가치 · 태도	데이터에 대한 다양한 해석을 제시하고 비판적으로 수용하며, 적절한 분석 방법을 선택하여 적용하는 자세의 중요성을 인식할 수 있다.
	C	지식 · 이해	데이터 모델의 개념, 회귀 분석, 군집 분석, 연관 분석을 설명할 수 있다.
		과정 · 기능	데이터 분석을 위한 도구를 탐색하고 데이터 모델의 분석 결과에 대한 의미를 해석하며, 데이터 분석 방법에 따른 분석 결과를 비교하고 평가할 수 있다.
		가치 · 태도	데이터에 대한 다양한 해석을 비판적으로 수용하며, 적절한 분석 방법을 선택하여 적용하는 자세의 중요성을 인식할 수 있다.
	D	지식 · 이해	데이터 모델의 개념을 설명하고, 회귀 분석, 군집 분석, 연관 분석을 인지할 수 있다.
		과정 · 기능	데이터 분석을 위한 도구를 탐색하고 데이터 모델의 분석 결과에 대한 의미를 일부 해석하며, 여러 가지 데이터 분석 방법에 따른 분석 결과를 부분적으로 비교하고 평가할 수 있다.
		가치 · 태도	데이터에 대한 다양한 해석, 적절한 분석 방법을 선택하여 적용하는 자세를 수용할 수 있다.
	E	지식 · 이해	데이터 모델의 개념, 회귀 분석, 군집 분석, 연관 분석을 인지할 수 있다.
		과정 · 기능	데이터 분석을 위한 도구를 부분적으로 탐색하고 데이터 모델의 분석 결과에 대한 의미를 일부 해석하며, 데이터 분석 방법에 따른 분석 결과를 부분적으로 비교하고 평가할 수 있다.
		가치 · 태도	데이터에 대한 해석, 적절한 분석 방법을 선택하여 적용하는 자세를 수용할 수 있다.

4 데이터 과학 프로젝트

성취기준	영역별 성취수준		
(4) 데이터 과학 프로젝트	A	지식 · 이해	데이터로 해결이 가능한 주제, 탐색적 데이터 분석을 정확하게 설명하고, 분석 결과에 대한 의미를 비판적으로 해석할 수 있다.
		과정 · 기능	분야별 데이터 과학의 주제를 다양하게 조사하고 탐색적 데이터 분석으로 데이터 속 의미를 다각도로 파악하며, 서로 다른 데이터 분석 방법을 적용한 결과를 비교하고 분석 결과를 활용하는 방법을 다양하게 탐색할 수 있다.
		가치 · 태도	문제를 해결하기 위한 창의적인 방법을 고민하는 자세, 가장 적절한 데이터 모델링 방법을 설계하는 자세, 복잡하고 어려운 문제를 끝까지 해결하기 위해 노력하는 태도, 일반화 및 공유 과정에서 윤리 문제 등 사회적 영향을 인식하는 자세를 내면화할 수 있다.
	B	지식 · 이해	데이터로 해결이 가능한 주제, 탐색적 데이터 분석을 정확하게 설명하고, 분석 결과에 대한 의미를 해석할 수 있다.
		과정 · 기능	분야별 데이터 과학의 주제를 다양하게 조사하고 탐색적 데이터 분석으로 데이터 속 의미를 다각도로 파악하며, 서로 다른 데이터 분석 방법을 적용한 결과를 비교하고 분석 결과를 활용하는 방법을 탐색할 수 있다.
		가치 · 태도	문제를 해결하기 위한 창의적인 방법을 고민하는 자세, 가장 적절한 데이터 모델링 방법을 설계하는 자세를 내면화하고, 복잡하고 어려운 문제를 끝까지 해결하기 위해 노력하는 태도, 일반화 및 공유 과정에서 윤리 문제 등 사회적 영향을 인식하는 자세를 인식할 수 있다.
	C	지식 · 이해	데이터로 해결이 가능한 주제, 탐색적 데이터 분석을 설명하고, 분석 결과에 대한 의미를 해석할 수 있다.
		과정 · 기능	분야별 데이터 과학의 주제를 조사하고 탐색적 데이터 분석으로 데이터 속 의미를 파악하며, 서로 다른 데이터 분석 방법을 적용한 결과를 비교하고 분석 결과를 활용하는 방법을 탐색할 수 있다.
		가치 · 태도	문제를 해결하기 위한 창의적인 방법을 고민하는 자세, 가장 적절한 데이터 모델링 방법을 설계하는 자세, 복잡하고 어려운 문제를 끝까지 해결하기 위해 노력하는 태도, 일반화 및 공유 과정에서 윤리 문제 등 사회적 영향을 인식하는 자세를 인식할 수 있다.
	D	지식 · 이해	데이터로 해결이 가능한 주제, 탐색적 데이터 분석을 설명하고, 분석 결과에 대한 의미를 부분적으로 해석할 수 있다.
		과정 · 기능	분야별 데이터 과학의 주제를 조사하고 탐색적 데이터 분석으로 데이터 속 의미를 파악하며, 서로 다른 데이터 분석 방법을 적용한 결과를 일부 비교하고 분석 결과를 활용하는 방법을 부분적으로 탐색할 수 있다.
		가치 · 태도	문제를 해결하기 위한 창의적인 방법을 고민하는 자세, 가장 적절한 데이터 모델링 방법을 설계하는 자세를 인식하고, 복잡하고 어려운 문제를 끝까지 해결하기 위해 노력하는 태도, 일반화 및 공유 과정에서 윤리 문제 등 사회적 영향을 인식하는 자세를 수용할 수 있다.
	E	지식 · 이해	데이터로 해결이 가능한 주제, 탐색적 데이터 분석을 인지하고, 분석 결과에 대한 의미를 부분적으로 해석할 수 있다.
		과정 · 기능	데이터 과학의 주제를 조사하고 탐색적 데이터 분석으로 데이터 속 의미를 부분적으로 파악하며, 서로 다른 데이터 분석 방법을 적용한 결과를 일부 비교하고 분석 결과를 활용하는 방법을 부분적으로 탐색할 수 있다.
		가치 · 태도	문제를 해결하기 위한 창의적인 방법을 고민하는 자세, 적절한 데이터 모델링 방법을 설계하는 자세, 복잡하고 어려운 문제를 끝까지 해결하기 위해 노력하는 태도를 수용할 수 있다.

3. 연간 지도 계획

● 본 연간 지도 계획은 64차시(4학점×16주)를 기준으로 작성되었으며, 학교 편성 계획에 따라 증감(±1학점×16주)이 가능합니다.

대단원명	소단원명	차시	시간	교과서 쪽수
Ⅰ 데이터 과학의 이해	01	데이터 과학과 문제 해결	2	10	10~17
	02	데이터의 형태와 속성	2		18~23
	03	데이터셋과 데이터베이스	3		24~31
	04	세상을 바꾸는 데이터 과학	2		32~41
		대단원 정리 및 평가 문제	1		42~43
Ⅱ 데이터 준비와 분석	01	데이터 수집 및 특성 분석	3	15	46~55
	02	데이터 전처리	3		56~67
	03	데이터 속성 간의 관계 파악	4		68~83
	04	동일한 데이터의 다양한 분석 · 비교	4		84~99
		대단원 정리 및 평가 문제	1		100~101
Ⅲ 데이터 모델링과 평가	01	데이터 모델과 모델링	4	21	104~117
	02	회귀 분석	7		118~143
	03	군집 분석	5		144~155
	04	연관 분석	4		156~165
		대단원 정리 및 평가 문제	1		166~167
Ⅳ 데이터 과학 프로젝트	01	데이터 과학 적용과 데이터 적합성	2	18	170~177
	02	[데이터 과학 프로젝트 ①] 슬기로운 의사 생활	6		178~195
	03	[데이터 과학 프로젝트 ②] 지혜로운 어부 생활	6		196~211
		대단원 프로젝트 대단원 정리 및 평가 문제	4		212~215
연간 시수			64		

데이터 과학의 이해

단원 개관

데이터의 폭발적인 증가와 데이터 처리 기술의 발전으로 인해 데이터 과학은 다양한 분야에서 더 나은 의사 결정을 도우며 우리의 삶과 사회의 변화를 이끌고 있다.

이 단원에서는 데이터 과학의 개념과 함께 데이터의 잠재적 가치를 살펴보고, 진로 및 직업과 관련한 데이터 기반의 문제 해결 사례를 분석하며 데이터의 역할 및 중요성을 파악하도록 한다.

단원 전개 계획

소단원명	차시	교과서 쪽수	학습 주제	지도 방법
01 데이터 과학과 문제 해결	2	10~17	1. 데이터와 데이터 과학 • 해 보기 ① 다양한 교과목과 데이터의 연계성 분석해 보기 • 해 보기 ② 하루 동안 생활 속에서 접하는 데이터 찾아보기 2. 데이터 과학의 문제 해결 과정	• 데이터 과학의 개념과 데이터 과학의 문제 해결 단계를 이해하고, 실생활의 문제 해결 사례를 데이터 기반의 의사 결정 상황에 적용할 수 있도록 지도한다.
02 데이터의 형태와 속성	2	18~23	1. 데이터의 다양한 형태 • 해 보기 ① 정형 데이터 찾아보기 2. 데이터의 속성 • 해 보기 ② 도서관 데이터의 다양한 속성이 가진 가치 살펴보기	• 생활 속 데이터에서 정형 데이터와 비정형 데이터를 구분하고, 데이터 속성에서 데이터의 잠재적 가치를 판단할 수 있도록 지도한다.
03 데이터셋과 데이터베이스	3	24~31	1. 데이터셋 2. 데이터베이스 3. 데이터베이스의 통합적 활용 • 해 보기 ① 데이터베이스의 필요성 생각해 보기 4. 데이터베이스의 생활 속 활용	• 데이터셋과 데이터베이스를 이해하고, 서로 다른 데이터셋을 통합하는 데이터베이스의 필요성을 설명할 수 있도록 지도한다.
04 세상을 바꾸는 데이터 과학	2	32~41	1. 데이터의 역사 • 해 보기 ① 관심 분야에서 데이터 과학이 접목된 사례 찾기 2. 지속 가능한 미래를 만드는 데이터 과학 • 해 보기 ② 지속 가능한 미래를 만들기 위한 사례 조사하기 3. 데이터 과학 속 데이터 윤리 • 해 보기 ③ 데이터 윤리 문제와 관련된 사례 조사하기	• 데이터로 인한 사회 변화를 이해하고, 나의 진로 및 관심 분야와 관련된 데이터 기반 문제 해결 사례를 분석할 수 있도록 지도한다.

성취기준

[12데과01-01] 데이터 과학의 개념을 이해하고, 문제 해결 사례를 데이터 기반 의사 결정 상황에 적용한다.

[12데과01-02] 정형 데이터와 비정형 데이터를 구분하고, 데이터 속성에서 데이터의 잠재적 가치를 파악한다.

[12데과01-03] 데이터셋의 집합인 데이터베이스를 이해하고, 서로 다른 데이터셋의 데이터를 분석이 가능한 형태로 통합하는 것의 의미를 파악한다.

[12데과01-04] 데이터로 인한 사회 변화를 인식하고, 진로 및 직업과 관련한 데이터 기반 문제 해결 사례를 분석한다.

1 데이터 과학과 문제 해결

수업 시간: 2시간 10~17쪽

단원명	Ⅰ. 데이터 과학의 이해 01. 데이터 과학과 문제 해결
학습 목표	• 데이터 과학의 개념과 데이터 과학의 문제 해결 단계를 설명할 수 있다. • 실생활의 문제 해결 사례를 데이터 기반의 의사 결정 상황에 적용할 수 있다.
수업 방법	강의, 토론, 발표
준비물	교사 교과서, 관련 교수 학습 자료 학생 필기도구

단계	교수 · 학습 방법	지도상의 유의점
도입	생각 열기 데이터가 우리 삶에 어떤 영향을 끼치고 있는지 친구들과 이야기를 나누어 보도록 한다.	• 데이터 과학이 적용된 다양한 문제 해결 사례를 바탕으로 이야기해 보게 하고, 흥미를 유발하도록 한다.
전개	**1. 데이터와 데이터 과학** • 데이터 과학의 개념과 데이터 기반 의사 결정의 중요성에 대해 탐구할 수 있도록 한다. • 데이터 과학의 특징을 이해할 수 있도록 한다.	• 데이터 과학의 개념과 특징을 다양한 관점에서 이해할 수 있게 한다.
	2. 데이터 과학의 문제 해결 과정 데이터 과학의 문제 해결 과정을 알고, 단계마다 수행해야 할 과업이 무엇인지 파악할 수 있도록 지도한다.	• 데이터 과학의 문제 해결 과정이 고정적인 것이 아니라 문제 상황에 따라 융통성 있게 변화시켜 적용할 수 있음을 이해하도록 지도한다.
정리	탐구 활동 데이터에 기반한 의사 결정 사례를 조사하고, 데이터 기반 의사 결정의 가치가 무엇인지 설명하고, 활동에 참고할 수 있는 사이트를 안내한다.	• 다양한 관점으로 데이터에 기반한 의사 결정 사례를 찾을 수 있도록 유도하고, 그 가치에 대한 의견을 자유롭게 나눌 수 있도록 지도한다. • 본 차시를 정리하고, 다음 차시를 예고한다.
평가	• 데이터 과학의 개념과 데이터 과학의 문제 해결 단계를 설명할 수 있는가? • 실생활의 문제 해결 사례를 데이터 기반의 의사 결정 상황에 적용할 수 있는가?	• 학생들이 자신의 생각을 정리할 수 있도록 지도하여 데이터 과학의 개념과 데이터 과학의 문제 해결 단계, 데이터에 기반한 의사 결정에 대해 이해할 수 있도록 돕는다.

1 데이터 과학과 문제 해결

제시 의도

데이터 과학은 의료, 축산, 마케팅, 안전 등 우리 생활과 밀접한 다양한 분야에서 많은 영향을 주고 있음을 다양한 실제 사례를 통해 학생들이 체감할 수 있도록 하였다. 특히 제시된 내용 외에도 학생들이 각자 관심 영역에서 데이터가 우리 삶에 어떤 영향을 끼치는지 생각해 볼 수 있는 시간을 갖도록 하였다.

지도 방법

- 교과서에 제시된 내용을 한 번 살펴보도록 하고, 이 중에서 어떤 사례가 가장 와닿는지 학생들의 생각을 표현하도록 안내한다.
- 이외에도 학생들이 알고 있는 내용 중 데이터가 생활 주변에 영향을 미치는 사례를 알고 있으면 소개해 보도록 한다.
- 교과서에 제시된 내용 외에 최신의 사례를 영상 등을 통해 제시한다.

예시 답안

❶ 스포츠 영역: 축구, 야구 등 다양한 스포츠에서 데이터를 분석하여 전략을 세우거나 선수를 선발하는 등 스포츠 분야에서도 데이터가 많이 활용된다.
❷ 자율 주행 영역: 주행 데이터를 수집하여 자율 주행 프로그램을 만드는 데 활용된다.

1 데이터와 데이터 과학

01 데이터 과학의 개념

데이터 과학은 데이터를 기반으로 실제 현상을 이해하고 분석하기 위해 통계, 데이터 분석, 기계학습 및 관련 방법을 통합하는 학문 분야다. 특히 데이터의 관찰, 측정, 데이터 수집 및 해석 등의 단계를 거치면서 지식과 '통찰'을 이끌어 내고, 과학적 접근 방법을 적용한다는 특징을 갖고 있다. 최근에는 기상청에서 날씨를 예측하거나, 스포츠 분야에서 데이터에 기반한 작전을 계획하는 데도 데이터 과학이 활용되고 있다.

통찰
사물이나 현상을 꿰뚫어 보는 능력 또는 특정 대상을 보편적인 시각 외에 다른 시각으로 볼 수 있는 능력

02 데이터 기반 의사 결정의 중요성

데이터를 수집 · 분석하여 의사 결정을 내리는 것을 '데이터 기반 의사 결정'이라고 한다. 축구 선수들의 데이터를 기반으로 선발 여부를 결정하는 것도 데이터 기반 의사 결정이다. 또한 우리 생활 속의 크고 작은 결정을 내릴 때도 데이터에 기반한 의사 결정을 하면 개인의 기억과 경험, 주관에 의해 의사 결정을 하는 방식에 비해 객관적이고 합리적인 선택을 할 수 있다.

예를 들어 학생들이 선택 과목을 결정할 때 이전 학기 성적, 과목에 대한 흥미도, 미래 산업의 발전 가능성 등 다양한 데이터를 고려하면 더 나은 선택을 할 수 있다. 또한 버스 노선을 결정할 때 승객들의 이동 패턴과 시간대별 승객 수 등의 데이터를 분석하면, 합리적인 버스 노선을 만들 수 있다.

▲ 경험, 주관에 의한 의사 결정과 데이터 기반 의사 결정의 사례

11

지도 방법

- 학생들이 관심을 가질 만한 주제의 최신 영상 등을 제공하며, 학생들의 흥미와 공감을 이끈다.
- 데이터에 기반한 버스 노선 결정 사례 등을 활용하여 데이터 기반 의사 결정 사례를 소개한다.
- 과목 선택에 있어서 데이터를 기반으로 의사 결정한다면 어떤 데이터에 기반하는 것이 좋을지, 학생들의 생각을 표현할 수 있는 기회를 준다.

용어 해설 통찰

데이터 과학에서 통찰(insight)이란 데이터 분석을 통해 얻은 깊은 이해와 통찰력을 의미한다. 이는 데이터를 통해 발견된 새로운 정보, 패턴, 관계성 또는 의미 있는 통계적 결과를 말한다.

1 데이터와 데이터 과학

데이터란 질적인 변수들 혹은 양적인 변수들의 가치 집합으로, 정보의 조합이라고 표현할 수 있다. 온라인 어원 사전(Online Etymology Dictionary)에 따르면 데이터(data)는 'datum'의 복수형으로, '주다(give)'라는 뜻의 라틴어 동사 'dare'에서 비롯되었다. 'datum'은 '주어진 것(thing given)'이라는 의미를 지니고 있다. 즉, 데이터라는 개념은 근본적으로 우리에게 주어진 것들이므로, 이를 찾아가는 방법을 탐구하는 것이 데이터 과학의 목표라고 할 수 있다.

빅데이터의 환경이 조성되면서 이러한 어원에 근거한 데이터의 의미가 더욱 현실적으로 다가오게 된다. 데이터 분석을 담당하는 사람들에게는 주어진 업무의 범주가 훨씬 확대되었으며, 더욱 과학적이고 합리적인 방법과 분별 능력이 요구된다고 할 수 있다. 이에 따라 데이터 과학의 중요성이 강조되고 있으며, 체계적이고 과학적인 접근법이 요구되고 있다.

[출처] 장영재, 유찬우, 『데이터 과학 개론』, 한국방송통신대학교(2022)

데이터 과학이라는 말이 널리 쓰이기 시작한 것은 1990년대 후반으로, 컴퓨터 과학자들과 컴퓨터를 이용한 큰 데이터셋 분석에 대한 논의를 할 때 통계학자들이 수학적 엄밀함을 강조하기 위해 쓰곤 했다. 1997년 C. F. 제프 우(C. F Jeff Wu)는 '통계학 = 데이터 과학?'이라는 대중 강연을 하면서 당시 통계학의 분명한 경향 몇 가지를 강조했는데, 거대 데이터 베이스의 크고 복잡한 데이터셋 활용, 컴퓨터 알고리즘과 모델이 점점 더 많이 쓰이는 현상 등이었다. 강연의 결론은 이제 통계학을 데이터 과학으로 바꿔 불러야 한다는 것이었다. 2001년 윌리엄 S. 클리브랜드(William S. Cleveland)는 데이터 과학을 대학교의 한 학과로 두는 실행 계획을 발표했다. 데이터 과학 학과가 수학과 컴퓨터 과학 부문의 파트너십으로 만들어져야 한다는 것이다. 또한 데이터 과학은 학제적인 노력의 산물이어야 하며 데이터 과학자는 여러 다른 주제의 전문가와 협력하는 법을 알아야 한다는 점도 강조했다. 같은 해에 레오 브레이먼(Leo Breiman)은 〈통계 모델링: 두 종류의 문화〉(Breiman, 2001)라는 논문을 발표했다. 브레이먼은 데이터 분석의 최우선 목표를 분석 대상 데이터가 어떻게 생겨났는지 설명하는 숨겨진 통계적 데이터 모델(예: 선형 회귀)을 찾아내는 것이라고 보는 전통적인 통계학적 접근법을 데이터 모델링 문화라고 보았다. 반면 데이터가 어떻게 생성되었는지 설명하기보다 컴퓨터 알고리즘을 이용해 예측 모델을 만드는 데 더 집중하는 문화를 알고리즘 모델링 문화라고 했다. 데이터를 설명하는 모델에 집중하는 통계적 접근과 정확한 예측에 집중하는 알고리즘적 접근이라는 구분과 차이는 통계학자와 기계학습 연구자 사이에도 그대로 적용된다. 둘 사이의 논쟁은 통계학계에서 여전히 진행 중이다(Shmueli, 2010). 일반적으로 오늘날 데이터 과학 프로젝트는 정확한 예측 모델을 만드는 기계학습 접근법에 더 가까우며 데이터를 설명하는 통계적 접근에는 관심이 덜하다. 즉, 데이터 과학은 통계학과 연관되어 주목을 받았고 여전히 많은 방법론과 모델을 통계학으로부터 빌려오고 있지만, 시간이 흐르면서 그와 구분되는 데이터 분석 방법을 개발해 오고 있는 셈이다.

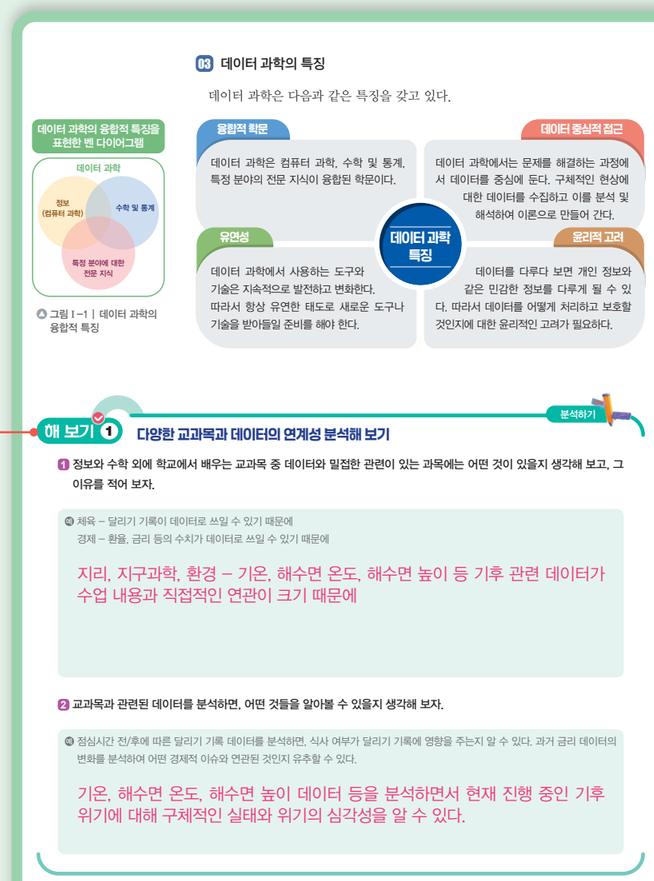
2001년부터는 데이터 과학의 개념이 통계학을 다시 정의하는 수준을 넘어섰다. 예를 들어, 지난 10년 동안 사람들이 활발하게 온라인 활동(온라인 구매, 소셜미디어 이용, 온라인 엔터테인먼트 등)을 하면서 생성되는 데이터의 양은 엄청나게 많아졌다. 데이터 과학자들은 이런 데이터를 수집하고 다

지도 방법

데이터 과학은 정보 과목 수업뿐 아니라 다양한 과목과 융합을 통해 이루어지는 학문이기 때문에 학생들이 관심 있는 과목과 연결 지을 수 있도록 지도한다.

해 보기 1 지도 방법

프로그래밍이나 데이터와 같은 정보 영역에 대한 전문성이 뛰어나지 않은 학생의 경우 [해 보기] 활동을 통해 자신이 데이터 과학 문제를 해결하는 데 중요한 역할을 담당할 수 있다는 사실을 알려 주도록 한다.
또한 다양한 학생들의 의견을 공유하며 데이터가 다양한 영역에서 활용 가능하다는 유용함을 느낄 수 있도록 지도한다.



03 데이터 과학의 특징

데이터 과학은 다음과 같은 특징을 갖고 있다.

데이터 과학의 융합적 특징을 표현한 벤 다이어그램

그림 Ⅰ-1 | 데이터 과학의 융합적 특징

융합적 학문
데이터 과학은 컴퓨터 과학, 수학 및 통계, 특정 분야의 전문 지식이 융합된 학문이다.

데이터 중심적 접근
데이터 과학에서는 문제를 해결하는 과정에서 데이터를 중심에 둔다. 구체적인 현상에 대한 데이터를 수집하고 이를 분석 및 해석하여 이론으로 만들어 간다.

데이터 과학 특징

유연성
데이터 과학에서 사용하는 도구와 기술은 지속적으로 발전하고 변화한다. 따라서 항상 유연한 태도로 새로운 도구나 기술을 받아들일 준비를 해야 한다.

윤리적 고려
데이터를 다루다 보면 개인 정보와 같은 민감한 정보를 다루게 될 수 있다. 따라서 데이터를 어떻게 처리하고 보호할 것인지에 대한 윤리적인 고려가 필요하다.

해 보기 1 다양한 교과목과 데이터의 연계성 분석해 보기

분석하기

1 정보와 수학 외에 학교에서 배우는 교과목 중 데이터와 밀접한 관련이 있는 과목에는 어떤 것이 있을지 생각해 보고, 그 이유를 적어 보자.

예 체육 – 달리기 기록이 데이터로 쓰일 수 있기 때문에
경제 – 환율, 금리 등의 수치가 데이터로 쓰일 수 있기 때문에

지리, 지구과학, 환경 – 기온, 해수면 온도, 해수면 높이 등 기후 관련 데이터가 수업 내용과 직접적인 연관이 크기 때문에

2 교과목과 관련된 데이터를 분석하면, 어떤 것들을 알아볼 수 있을지 생각해 보자.

예 점심시간 전/후에 따른 달리기 기록 데이터를 분석하면, 식사 여부가 달리기 기록에 영향을 주는지 알 수 있다. 과거 금리 데이터의 변화를 분석하여 어떤 경제적 이슈와 연관된 것인지 유추할 수 있다.

기온, 해수면 온도, 해수면 높이 데이터 등을 분석하면서 현재 진행 중인 기후 위기에 대해 구체적인 실태와 위기의 심각성을 알 수 있다.

해 보기 2 지도 방법

선생님의 예시 답안을 보여 주고, 이를 통해 생활 곳곳에 데이터가 숨어 있고, 우리가 이미 데이터에 기반하여 다양한 의사 결정을 하고 있음을 학생들이 체감할 수 있도록 지도한다.
특히, 선생님이 제시한 예시 답변 외에도 학생들의 삶과 밀접한 다양한 데이터가 있음을 질문을 통해 발굴하여 학생들이 공감할 수 있도록 한다.

해 보기 2 하루 동안 생활 속에서 접하는 데이터 찾아보기

분석하기

1 아침에 일어나서 밤에 잠들 때까지, 자신의 생활 속에서 어떤 데이터를 접하고 있는지 기록해 보자.

생활 속에서 접하는 데이터

날씨 애플리케이션 | 대중교통 전광판 | 헬스케어 애플리케이션

	언제	어디에서	무엇을	어떻게	왜
예	아침에 옷 입기 전	집에서	기온을	스마트폰으로	어떤 옷을 입을지 결정하려고
1	신발을 신기 전에	집에서	강수 확률을	스마트폰으로	우산을 가져갈지 말지 결정하려고
2	버스를 기다리며	버스 정류장에서	버스 도착 예정 시각을	버스 안내판으로	얼마나 기다려야 하는지 알아보려고
3	쉬는 시간에	교실에서	급식 메뉴를	스마트폰으로	메뉴가 궁금해서
4	운동 시간에	공원에서	달리기 거리를	스마트폰으로	운동량을 확인하기 위해서

2 생활 속에서 발생하는 데이터 중에는 어떤 목적을 위해 수집되는 경우도 있지만 그렇지 않은 경우도 있다. 만약 데이터를 추가로 수집한다면 어떤 문제를 해결할 수 있을지 생각해 보고, 친구들과 함께 이야기해 보자.

❶ 기존에는 수집되지 않았던 데이터 중에서 어떤 데이터를 추가로 수집할 수 있을까?

수업 시간 중 학생들의 시선 데이터

❷ ❶에서 말한 데이터를 수집할 수 있다면 어떤 문제를 해결할 수 있을까?

학생들이 얼마나 수업에 집중하는지 데이터를 기반으로 파악할 수 있어서 개별 지도가 가능해진다.

❸ 데이터를 수집하는 과정에서 윤리적인 문제가 발생할 가능성은 없을까?

학생들의 시선을 계속 파악하는 것이 민감한 개인 정보에 해당될 수 있고, 수업 외의 시선까지 수집한다면 인권 침해 요소도 될 수 있다.

루는 데이터 과학 프로젝트를 위해 웹에서 생성되는 종종 비구조화된 데이터를 수집, 병합, 정리하는 프로그래밍과 해킹 기술(악의적인 시스템 침투 기술이 아닌 네트워크를 파악하는 기술) 등을 개발할 필요가 생겼다. 또 빅데이터의 출현으로 하둡과 같은 빅데이터 기술도 다룰 줄 알아야 했다. 사실, 오늘날 데이터 과학자라는 개념은 워낙 넓어져서 그 역할과 요구되는 전문성, 기술 등을 어떻게 정의해야 하는지에 대해서도 논의가 한창 진행 중이다.

[출처] 존 켈러허, 브렌던 티어니, 『(더 나은 의사 결정을 위한 통찰의 도구) 데이터 과학』, 김영사(2019)

참고 동영상

스포츠 분야에서 빅데이터의 역할

- **제목**: 채널A 뉴스, 축구는 빅데이터로 특급 성장 중
- **영상 내용**: 독일 분데스리가 4부 리그에서 1부 리그로 성적이 급등한 축구팀이 있다. 바로 TSG 1899 호펜하임. 이 사례를 통해 스포츠 영역에서 빅데이터가 어떤 변화를 가져오는지 살펴본다.

|주소| https://youtu.be/fRjH0FwnKiY?si=gT1fmcwW9pnkRedW

2 데이터 과학의 융합적 성격

융합이란 다양한 요소나 요소들의 통합을 의미하는 용어이다. 예를 들어, 여러 가지 다른 분야나 영역의 지식, 기술, 자원 등을 결합하여 새로운 가치를 창출하거나 문제를 해결하는 것을 말한다.

데이터 과학의 융합적 성격은 다양한 분야와 학문의 지식과 기술을 결합하여 데이터를 분석하고 활용하는 것을 의미한다. 데이터 과학은 수학, 통계학, 컴퓨터 공학, 비즈니스, 사회과학 등 다양한 분야의 지식과 기술이 융합되어야만 효과적으로 데이터를 다룰 수 있다.

3 데이터 분석의 가치

데이터 기반 경제를 이끌어 나가는 핵심 능력은 데이터 분석이다. 검색 키워드를 분석한 결과가 높은 가치를 지니는 것과 같다. 데이터 분석을 통해서 얻는 통찰력이 데이터 자체의 가치를 넘어 새로운 가치와 서비스를 창출하는 것이다. 데이터 분석의 역사는 인류의 역사만큼이나 오래된 분야이지만 최근 데이터 이용량이 늘고 데이터 기반의 경제 시대가 되면서 데이터 분석이 새로운 차원으로 발전하고 있다.

소셜 네트워크 분석이 중요한 이유는 예전과 달리 제품이나 서비스에 대한 고객의 반응, 즉 VOC(voice of customer)를 파악하기가 매우 쉬워졌기 때문이다. 예전에는 우편엽서로 설문 조사를 하기도 하고, 전화로 물어보기도 했지만 소비자 불만과 피드백을 받기란 매우 어려웠다. 그러나 지금은 트위터, 블로그, 페이스북을 통해 자발적으로 직설적인 고객의 목소리가 쏟아져 나오고 있다. 이렇게 얻을 수 있는 고객 데이터를 분석하지 않는다는 것은 비즈니스를 포기한 것과 같다.

지도 방법

- 데이터 과학의 문제 해결 과정은 반드시 절차적으로 지켜야 할 절대적인 것이 아니라는 사실을 잘 전달하도록 한다.
- 아이스크림 생산 관련 이야기를 보면서 하나의 문제를 다양하게 정의할 수 있고, 이 과정에서 배경지식이 중요하게 작용함을 이해시킨다.

용어 해설 **문제 정의**

문제 정의 단계에서는 문제와 관계된 사람들과 협력하여 프로젝트의 목표 및 요구 사항과 함께, 해결하려는 문제 또는 답변해야 하는 질문을 명확하게 정의한다.

2 | 데이터 과학의 문제 해결 과정

그림 I-2 | 데이터 과학의 문제 해결 과정

맥락
사물이나 대상 등이 서로 연결되어 있는 관계

데이터 과학의 문제 해결 과정은 크게 문제 정의하기, 데이터 수집하기, 전처리 및 탐색하기, 모델링하기, 모델 평가하기, 모델 활용하기의 6단계로 구분할 수 있다. 그러나 6개의 단계를 순서대로 모두 거쳐야 하는 것은 아니다. 문제 상황에 따라 일부 단계는 생략될 수도 있고, 일부 단계는 반복될 수도 있으며, 경우에 따라 이전 단계를 다시 수행할 수도 있다. 따라서 문제 해결 과정의 전체적인 흐름을 이해하고, 각 상황에 따라 유연하게 문제 해결 과정을 설계할 수 있어야 한다.

01 문제 정의하기

문제 정의 단계는 문제와 관련된 다양한 맥락*을 바탕으로 문제를 어떻게 정의할지 결정하고, 전체적인 계획을 수립하는 단계다.

예를 들어, 아이스크림 회사에서 '올여름 아이스크림 생산량'을 결정해야 할 경우, 기존 아이스크림 판매 데이터를 바탕으로 아이스크림 판매에 영향을 끼치는 요인들을 분석하여 문제를 정의해야 한다. 이때 기온이나 장마와 같은 날씨 요인이 아이스크림 판매에 큰 영향을 끼친다고 생각하고 문제를 정의할 수도 있고, 아이스크림 소비를 주도하는 청소년 및 청년층의 기호에 맞는 신상품을 개발하는 것이 판매에 큰 영향을 끼친다고 생각하고 문제를 정의할 수도 있다.

02 데이터 수집하기

데이터 수집 단계에서는 정의된 문제와 직접 연관된 데이터를 수집해야 한다. 만약 데이터의 양이 부족하거나 누락된 데이터나 오류가 있는 데이터가 많다면 뛰어난 분석 방법을 적용한다고 하더라도 좋은 결과가 나오기 어렵다. 따라서 데이터 수집 단계는 데이터 과학의 문제 해결 과정에서 매우 중요한 단계다. 이때 수집된 데이터의 양과 질에 따라 문제를 다시 정의해야 할 수도 있기 때문에 문제 정의와 데이터 수집 단계는 매우 밀접한 관계가 있다.

모델링
지도는 실제 지형이나 도로, 건물 등의 위치를 토대로 만들지만, 현실 세계의 모습을 단순화하고 추상화하여 표현한 것이다. 이와 같이 어떤 대상을 단순화하고 추상화하여 일반적으로 표현하는 과정을 모델링이라고 하고, 모델링된 결과를 모델이라고 한다.

03 전처리 및 탐색하기

전처리 및 탐색 단계는 수집된 데이터에 비어 있는 값이나 정상적인 값에서 극단적으로 벗어난 값 등의 오류가 있는지 확인하고, 오류가 있을 경우 데이터를 분석하기 전에 적절한 방법으로 처리하는 단계다. 데이터에 이상이 없는지 확인하는 작업은 매우 중요한데, 데이터에 오류가 있을 경우 전체적인 문제 해결 과정에 심각한 장애를 초래하기 때문이다. 또한 데이터를 시각화하며 데이터의 다양한 특징을 탐색하는 것도 전처리 및 탐색 단계에서 해야 할 중요한 일이다.

04 모델링하기

모델링(modeling)* 단계는 데이터로부터 유용한 패턴을 추출하고, 그 패턴을 기반으로 모델을 구축하는 단계다. 모델은 데이터를 직접 분석하거나 자동화된 알고리즘을 통해 만들 수 있다. 예를 들어, 기온 및 강수량 데이터와 아이스크림 판매량 데이터 간의 패턴을 모델로 만들어서 아이스크림 판매량을 예측하는 모델을 만들 수 있을 것이다.

다양한 데이터 수집 방법

실제 데이터를 수집할 때는 파일 형태의 데이터를 다운로드하는 것뿐 아니라 서버에서 로그 데이터를 수집하거나 웹 문서에 있는 내용을 크롤링하거나 API를 활용하여 데이터를 수집하는 등 다양한 방법을 통해 수집할 수 있다.

지도 방법

- 문제가 해결되기 위해서는 양질의 데이터를 입력하는 것이 중요하다. 따라서 데이터 전처리 단계를 통해 데이터의 오류 등을 파악하는 것이 매우 중요함을 알려 주도록 한다.
- 데이터 전처리 및 탐색 단계를 통해 발견한 통찰을 바탕으로 문제를 다시 정의할 수도 있음을 알려 주도록 한다.

데이터 분석은 광고, 마케팅, 건강, 의료, 보험, 금융, 재난 관리, 범죄 예방 등 거의 모든 사회, 정치, 경제, 문화 분야에서 새로운 서비스를 만들어 낼 것이다. 그리고 데이터 분석은 점차 종합적인 형태로 발전할 것이다. 즉, 한 영역에서 나오는 데이터만 분석하는 것이 아니라 여러 소스에서 나오는 데이터를 연계하여 분석하는 것이 늘어날 것이다.

미국의 영수증 마케팅사(社)인 카탈리나는 1억 명 이상의 소비자 구매 이력을 상세히 분석하여 슈퍼마켓에서 소비자가 계산대에서 계산을 하고 영수증을 발행할 때 그 사람에게 현재 가장 적합한 할인 쿠폰을 영수증과 같이 출력해 준다. 카탈리나 마케팅에는 미국 대부분의 대형 슈퍼마켓들이 참여하고 있는데, 이렇게 서로 경쟁적인 슈퍼마켓들이 자사 고객의 민감한 쇼핑 정보를 카탈리나사(社)에게 제공하는 이유는 종합적인 소비자 분석을 하기 위해이다. 고객은 여러 슈퍼마켓을 돌아다니기 때문에 한 슈퍼마켓의 쇼핑 기록만 봐서는 정확한 고객 분석과 상품 추천을 할 수 없기 때문이다.

디지털 데이터는 복사본을 만들기도 쉬우며, 컴퓨터 프로그램으로 분석하기도 쉽다. 이제 데이터의 재생산과 재가공 그리고 재활용은 기하 급수적으로 늘어날 것이며, 여기서 가치를 추출하는 능력에서 기업과 기관의 실력 차이가 날 것이다.

예전에는 데이터 크기가 너무 커서 분석을 포기했던 경우도 이제는 컴퓨터와 정보 통신의 발달로 저렴하고 신속하게 분석하는 것이 가능해졌다. 데이터 분석 도구도 쉽고 다양하게 소개되고 있어 고도의 기술을 습득한 전문가가 아니어도 데이터 분석이 가능해졌고, 따라서 여러 가지 시각의 데이터 해석이 가능하게 되었다. 빅데이터의 중요한 특징은 많은 데이터의 존재가 아니라 데이터의 접근과 분석이 쉬워져서 다양한 시각의 분석이 가능해졌다는 것이다.

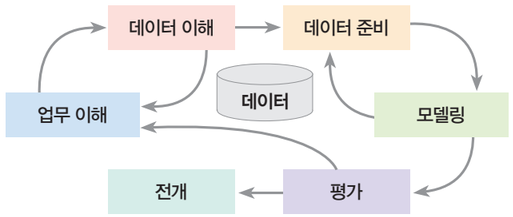
[출처] 김화종, 『데이터 사이언스 개론』, 홍릉과학출판사(2014)

4 CRISP-DM 방법론

전 세계에서 가장 많이 사용되는 데이터 마이닝(data mining) 표준 방법론으로서, 단계, 일반 과제, 세부 과제, 프로세스 실행 등의 4가지 레벨로 구성된 계층적 프로세스 모델이기도 하다.

CRISP-DM(Cross Industry Standard Process for Data Mining)의 절차는 6단계로 구성되어 있는데, 각 단계들은 순차적으로 진행되는 것이 아니라, 필요에 따라 단계 간의 반복 수행을 통해 분석의 품질을 향상시킨다.

[출처] https://www.2e.co.kr/news/articleView.html?idxno=301010



CRISP-DM 분석 절차

5 데이터 과학자에게 데이터 전처리 작업 비중은 얼마나 될까?

데이터 과학자는 주어진 시간의 60%를 데이터를 정리하고 구성하는 데 보낸다. 데이터셋 수집은 19%의 시간으로 두 번째로 이루어지며, 이는 데이터 과학자가 분석을 위해 데이터를 준비하고 관리하는 데 주어진 시간의 약 80%를 소비한다

지도 방법

모든 단계에 대한 설명이 끝난 후, 학생들을 모둠별로 편성하여 아이스크림 판매량 예측 모델 또는 새로운 모델을 만들어 가는 과정을 설명하도록 할 수 있다.
이때, 학생들이 새로운 주제를 만드는 것을 어려워할 경우, 아이스크림 판매량 예측 모델로 설명하도록 한다. 학생들이 새로운 아이디어를 담아 문제를 정의할 수 있는 역량이 충분하다면, 새로운 문제를 발견하여 각 단계에 맞는 스토리를 설명해 보도록 할 수 있다.

용어 해설 **모델 평가**

학습된 모델을 평가하여 모델이 얼마나 잘 작동하는지를 평가한다. 이를 통해 모델의 성능을 측정하고 필요에 따라 모델을 개선할 수 있다.

05 모델 평가하기

모델 평가 단계는 모델링 단계를 통해 만들어진 모델이 문제를 해결하는 데 적합한 모델인지 종합적인 관점에서 평가하는 단계다. 평가 단계를 통해 모델을 만들 때 고려하지 못한 중요한 요소가 있는지 최종적으로 점검하고, 모델을 문제 해결에 적용할지 여부를 결정한다. 예를 들어, 기존에 만들었던 모델이 작년의 아이스크림 판매량을 적절하게 예측하는지 확인하여 모델의 활용 여부를 결정할 수 있다.

06 모델 활용하기

모델 활용 단계는 이전 단계에서 만든 모델을 문제 해결 및 의사 결정에 적용하는 단계다. 또한 모델을 적용하는 과정에서 모델이 문제 해결에 적합한지를 지속적으로 점검하는 것도 모델 활용 단계에서 이루어진다. 예를 들어, 올여름 아이스크림 생산량을 결정하는 의사 결정에 아이스크림 판매량 예측 모델을 활용할 수 있다. 이렇게 데이터에 기반한 판매량을 예측 모델을 활용하면 적절한 양의 생산을 통해 더 효율적인 기업 활동을 할 수 있게 된다.

소단원 1분 요약

1. 데이터 과학은 데이터와 빅데이터의 기반 위에 실제 현상을 이해하고 분석하기 위해 통계, 데이터 분석, 기계학습 및 관련 방법을 종합적으로 활용하는 학문 분야다.
2. 데이터 과학의 문제 해결 과정은 크게 문제 정의하기, 데이터 수집하기, 전처리 및 탐색하기, 모델링하기, 모델 평가하기, 모델 활용하기의 6단계로 구분할 수 있다.

탐구 활동 데이터에 기반한 의사 결정 사례 조사하기

2020년에 발생한 코로나19 감염병 대응 과정은 데이터에 기반하여 문제를 해결한 대표적 사례다. 코로나19 등 감염병 방역 정책과 관련한 다음 기사를 읽고, 데이터 기반 의사 결정의 가치를 생각해 보자.

과학기술정보통신부(이하 과기정통부)와 질병관리청(이하 질병청), 정보통신산업진흥원은 코로나19 등 신종 호흡기계 감염병 대응 협력 상황을 점검하고, '한국형 신규 감염병 대응 시스템' 구축 등 추진 경과를 공유했다.
현재 우리나라 대표 출연 연구 기관, 인공지능 기업, 병원 등 15개 기관이 참여해 감염병 전파 매개 변수 분석, 공간 내 감염원 전파 양상 분석, 유행 예측 모델 및 방법론 개선, 온라인 기반 유행 양상 분석, 의료 자원 관리 등 5개 과제를 중심으로 해결책을 개발 중에 있다.
질병청 위기대응분석관은 "코로나19 팬데믹과 같은 긴급 상황 대응 시, 과학적 근거에 기반한 신속한 의사 결정이 무엇보다 중요하다."라며 "향후 질병청의 감염병 대응에 있어 과기정통부의 인공지능 기반 해법이 근거 중심 정책 수립에 큰 도움이 될 것으로 기대한다."라고 밝혔다.

국내 코로나19 발생 현황 대시보드

출처 인공지능 · 데이터에 기반한 감염병 방역 정책 수립, ZDNET(2021. 11. 24.), https://zdnet.co.kr/view/?no=20211124103726

1 코로나19 감염병을 해결하는 과정에서 데이터에 기반하여 의사 결정을 내렸던 사례를 찾아보고, 데이터 기반 의사 결정의 가치를 설명해 보자.

빅데이터에 기반한 코로나19 백신 예방 효과의 분석 체계를 표준화하여 예방 접종이 중증화 사망을 낮춘다는 과학적 분석 결과를 국민들에게 주기적으로 발표하였고, 이를 통해 예방 접종률을 향상시키는 등 방역 정책에 대한 신뢰도 제고에 크게 기여하였다.

2 자신의 관심 분야에서 데이터에 기반하여 문제를 해결한 사례를 찾아보고, 친구들에게 그 가치를 설명해 보자.

넷플릭스는 사용자 데이터를 철저히 분석하여 개인화된 콘텐츠 추천 시스템을 구축했다. 시청 기록, 평점, 검색 기록 시청 시간대 등의 데이터로 콘텐츠를 추천하여 고객들의 만족도를 향상시켰고, 고객들이 좋아하는 내용을 분석하여 효과적인 오리지널 콘텐츠 제작을 결정하는 등 데이터에 기반하여 문제를 해결하였다.

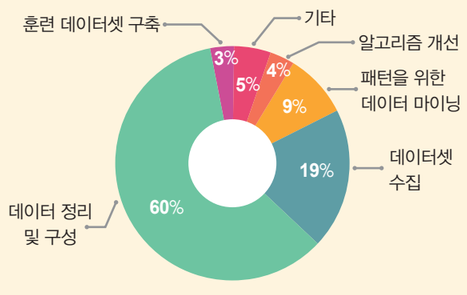
탐구 활동 지도 방법

데이터에 기반하여 문제를 해결하는 대표적인 분야가 방역 분야다. 우리나라에서 2020년에 발생한 코로나19 감염병 대응 과정을 살펴보며, 데이터에 기반한 의사 결정에 대해 이해하는 활동이다.
그러나 최신의 사례 중 더 적합한 사례가 있을 경우, 최신의 사례를 제시하고, 활동을 하는 것이 좋다.

는 것을 의미한다.

[출처] https://modulabs.co.kr/blog/data-preproccesing/

★데이터 과학자들은 시간을 어떻게 할당할까?★



6 모델 활용하기 단계 - 둘로 나누기

모델 활용하기 단계는 경우에 따라 모델 배포하기 단계와 결과 시각화하기 및 전달하기 단계로 나눌 수 있다.

❶ 모델 배포하기

평가가 끝난 최종 모델을 사람들이 사용할 수 있도록 배포한다.

❷ 결과 시각화하기 및 전달하기

다양한 시각화 도구를 사용하여 누구나 결과를 쉽게 이해할 수 있는 방식으로 시각화할 수 있다.

[출처] https://azure.microsoft.com/ko-kr/resources/cloud-computing-dictionary/what-is-data-science#processes-and-deliverables

7 데이터 기반 의사 결정을 통합하기 위한 6단계

모든 조직은 회사 전체에서 데이터 기반 의사 결정을 통합하기 위한 6단계를 수행하여 이점(利點)을 누릴 수 있다. 이러한 모범 사례를 채택하면 데이터 분석에서 파생된 전략을 구현하고 그 영향을 측정할 수 있다.

❶ 목표 정의

이 단계에는 조직의 목표를 명확하게 표현하는 것이 포함된다. 목표가 정의된 후 회사는 목표를 달성하기 위해 집중적이고 목적이 명확한 노력을 기울일 수 있다.

❷ 데이터 식별, 준비 및 수집

이 단계에서 조직은 명확한 목표를 설정하고 데이터 요구사항을 판별하며 데이터 소스를 평가 및 준비한 다음 체계적으로 데이터를 수집 및 검증한다.

❸ 구성 및 탐색

이 단계에서 데이터는 새로운 패턴, 추세 및 가치 있는 인사이트를 발견할 수 있도록 구조화된다. 데이터를 정리하면 정확성과 신뢰성이 보호된다. 데이터를 시각화하면 원시 데이터에서 즉시 명확하지 않은 패턴, 이상치 및 추세를 식별할 수 있다.

❹ 데이터 분석 수행

이 단계에서는 다양한 기술과 방법론을 사용하여 원시 데이터를 실행 가능한 인사이트로 변환해 비즈니스 전략에 영향을 미치는 패턴, 상관관계 및 추세를 파악한다. 데이터 분석을 수행함으로써 조직은 전략적 결정을 내리고 전반적인 성과를 개선할 수 있다.

❺ 결론 도출

이 단계에서는 주요 데이터 분석 결과를 검토하고 올바른

비즈니스 상황에 배치하여 실행 가능한 인사이트와 권장 사항을 구성한다. 실용적 인 결론은 비즈니스 성공을 이끄는 의미 있는 조치로 이어진다.

⑥ 구현 및 평가

이 마지막 단계는 인사이트를 검증하고 결과를 측정하는 데 중요하다. 데이터 기반 권장 사항을 기반으로 실행 계획을 개발하고 리소스(resource)*를 할당하며 이니셔티브(initiative)*의 진행 상황을 지속적으로 모니터링한다. 사전 정의된 핵심 성과 지표(KPI)*를 기준으로 결과를 측정하고 결과를 분석하며 피드백을 수집하고 해당 정보를 기반으로 지속적인 모니터링 및 반복적인 개선 주기를 통해 조정을 수행한다. 이 단계는 경험적 증거와 피드백을 기반으로 정보에 입각해 전략과 계획을 조정하기 위한 기초를 제공한다.

* **리소스(resource)**: 조직 내에서 특정 목표를 달성하기 위해 사용되는 모든 것을 의미한다. 이는 인력, 예산, 장비, 시설 등 다양한 형태로 나타날 수 있다.
* **이니셔티브(initiative)**: 조직이 특정 목표를 달성하기 위해 추진하는 계획이나 프로젝트를 의미한다. 이는 새로운 제품 출시, 프로세스 개선, 캠페인 등 다양한 형태로 나타날 수 있다.
* **핵심 성과 지표(KPI: Key Performance Indicator)**: 조직의 목표와 성과를 측정하기 위해 사용되는 주요 지표이다. 이는 조직의 목표와 전략에 따라 사전에 정의되며, 조직의 성과를 추적하고 개선할 수 있는 핵심적인 지표를 나타낸다.

[출처] IBM, https://www.ibm.com/kr-ko/think/topics/data-driven-decision-making

MEMO

2 데이터의 형태와 속성

수업 시간: 2시간 18~23쪽

단원명	Ⅰ. 데이터 과학의 이해 02. 데이터의 형태와 속성
학습 목표	• 생활 속 데이터에서 정형 데이터와 비정형 데이터를 구분할 수 있다. • 데이터 속성에서 데이터의 잠재적 가치를 판단할 수 있다.
수업 방법	강의, 토론, 발표, 실습
준비물	교사 교과서, 관련 교수 학습 자료, 컴퓨터 학생 필기도구

단계	교수 · 학습 방법	지도상의 유의점
도입	생각 열기 우리 주변의 문제를 해결하기 위해 수집할 수 있는 데이터의 형태를 생각해 보고, [생각 열기] 만화에 나오지 않은 어떤 데이터를 추가로 수집할 수 있을지 생각해 보도록 한다.	• 학생들이 스스로 생각할 수 있도록 지도하되, 어려움을 느낄 경우 키나 높이 뛰기 기록 등을 수집할 수 있다는 힌트를 제공할 수 있다. • 학생들이 이야기한 데이터가 어떤 형태의 데이터인지 생각해 보도록 지도한다.
전개	1. 데이터의 다양한 형태 • 정형 데이터와 비정형 데이터의 사례를 소개하며 각각의 특징을 설명한다. • 반정형 데이터도 최근 많이 사용되고 있음을 소개한다.	• 행과 열과 같은 데이터 관련 용어에 익숙해지도록 하고, 영어도 함께 친숙해질 수 있도록 지도한다.
	2. 데이터의 속성 실생활의 데이터를 중심으로 데이터에서 속성이란 무엇인지 이해할 수 있도록 안내하고, [해 보기] 활동을 통해 문제 상황에 따라 필요한 속성이 달라지며, 데이터 속성에서 잠재적 가치를 생각해 보도록 지도한다.	• 학생들이 관심사에 따라 최신의 흥미로운 데이터(음원 스트리밍 데이터 등)를 활용할 수도 있다.
정리	탐구 활동 공공데이터포털을 살펴보며 정형, 비정형, 반정형 데이터가 다양하게 제공됨을 경험하도록 하고, 수집한 데이터를 정형 데이터와 비정형 데이터를 중심으로 구분해 도록 한다. 또한 관심 분야의 데이터를 살펴보며 데이터의 잠재적 가치에 대해 함께 이야기를 나누도록 지도한다.	• 학생들의 역량이 충분한 경우 반정형 데이터에 대해서도 소개해 주는 것이 좋으나, 반정형 데이터는 다운로드하여 확인하기 어려우므로 교실 상황에 따라 생략할 수도 있다.
평가	• 생활 속 데이터를 정형 데이터와 비정형 데이터로 구분할 수 있는가? • 데이터 속성에서 데이터의 잠재적 가치를 판단할 수 있는가?	• 학생들이 실생활의 관심 데이터를 바탕으로 데이터의 유형을 구분하고, 데이터 속성에 따라 어떤 문제를 해결할 수 있을지 충분히 생각하여 자신의 생각을 표현하도록 한다.

2 데이터의 형태와 속성

제시 의도

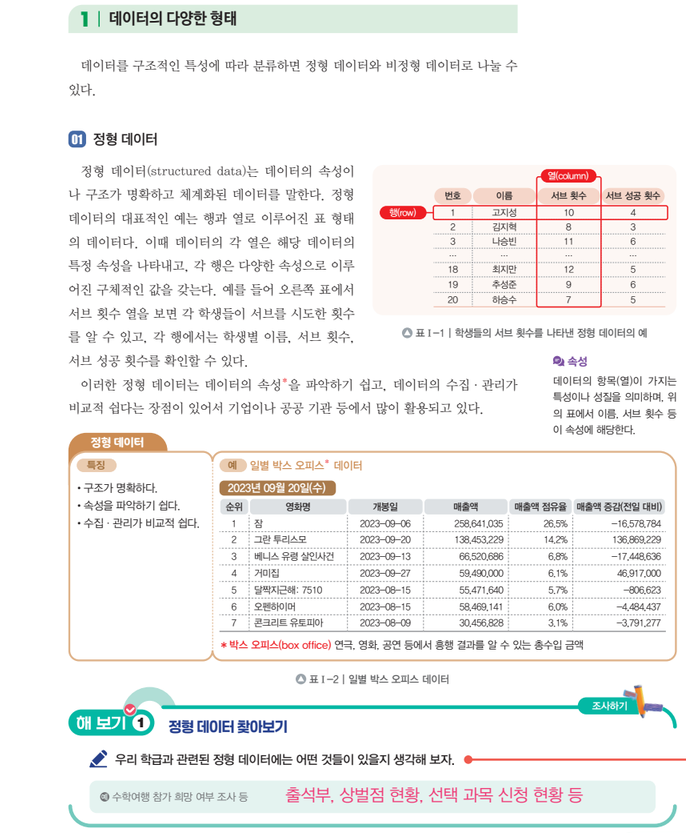
체육 시간과 같은 일상 속에서도 다양한 데이터가 존재하고, 데이터는 그 형태에 따라 표와 같은 형태의 정형 데이터와 이미지와 같은 형태의 비정형 데이터로 나뉜다는 사실을 알려 주기 위해 [생각 열기]를 제시하였다. 또한 어떤 속성의 데이터가 대표 선수 선발이라는 문제를 해결하는 데 필요할 것인지 생각해 보게 함으로써, 핵심 속성에 대해 이해할 수 있는 질문을 제시하였다.

지도 방법

- 제시된 스토리를 살펴보며, 표 형태의 데이터와 이미지 형태의 데이터의 장점과 단점을 생각해 보도록 한다.
- 대표 선수 선발을 위해 제시된 속성 외에 어떤 속성을 추가로 수집하면 좋을지 자유롭게 생각해 보도록 하고, 다양한 속성 중 어떤 속성이 가장 적합할지 함께 논의하는 시간을 가지며 핵심 속성을 추출하는 경험을 제공한다.

예시 답안

- **정형 데이터**: 키, 제자리 높이 점프 기록(cm) 등 데이터를 추가로 수집하면 대표 선수 선발에 도움이 될 수 있다.
- **비정형 데이터**: 서브하는 영상, 리시브하는 영상 등을 추가로 수집할 수 있다.



해 보기 1 지도 방법

정형 데이터가 생활 속에서 매우 많이 사용되고 있음을 학생들이 느낄 수 있도록 교실에서 쉽게 접할 수 있는 학급 학생들의 수학여행 참가 희망 여부, 출석부 등 정형 데이터의 사례를 소개해 준다. 또한 학생들이 스스로 정형 데이터의 사례를 생각해 볼 수 있도록 기회를 제공한다.

1 표와 이미지 데이터의 장점

표 형태의 데이터는 다양한 사실들을 일목요연하게 볼 수 있다는 장점이 있고, 이미지 형태의 데이터는 표로 표현하기 어려운 구체적인 자세나 표정 등의 정보를 알 수 있다는 장점이 있다.

2 정형 데이터와 비정형 데이터

정형 데이터는 특정한 구조를 갖고 있는 엑셀과 같은 프로그램에서 사용하는 표 형태의 데이터이고, 비정형 데이터는 특정한 구조를 갖고 있지 않은 소셜 미디어(social media)의 글이나 이미지, 영상 데이터이다.

3 속성, 피처, 레코드

정형 데이터는 행(column)과 열(row)을 기준으로 분석하는데, 이때 열을 속성(attribute)이나 피처(feature)라고 부르기도 하고, 행을 관찰값(observation)이나 레코드(record), 또는 인스턴스(instance)라고 부르기도 한다.

4 데이터의 유형 구분

❶ 정형 데이터, 비정형 데이터, 반정형 데이터

데이터 수집과 해석은 지식과 정보를 형성하는 데 필수적인 역할을 한다. 또한 실증적인 연구에 있어서도 데이터가 중요하고 핵심적인 역할을 하게 된다. 빅데이터 시대의 도래로 확장된 형태의 데이터도 흔히 접할 수 있는데, 음향 및 영상 신호, 이미지 구조 등이 그에 속한다고 할 수 있다. 이러한 환경 변화에 따라 데이터의 유형 구분에도 변화가 이루어졌다. 수치적인 계량에만 국한된 것이 아니고 시각이나 청각을 통해 인지할 수 있는 대상도 데이터의 범주에 포함되는 것이다. 이런 관점에서 데이터의 속성이나 구조가 상대적으로 명확하고 체계화된 데이터를 정형 데이터로 분류해 볼 수 있다. 즉, 데이터의 속성을 파악하는 데 제약이 비교적 적으며 데이터 수집이나 가공에 큰 어려움이 없고, 기존 데이터 분석 도구로 분석이 용이한 데이터를 정형 데이터라고 할 수 있다. 정형 데이터는 다시 범주형 데이터와 수치형 데이터로 구분해 볼 수 있다.

비정형 데이터는 정형 데이터에 비해서 형태가 매우 다양하며 데이터의 수집 과정이나 가공 · 정제 과정에 많은 노력이 필요한 데이터를 의미한다. 이러한 비정형 데이터는 전통적인 분석 도구를 이용하여 분석하기에 많은 제약이 따르므로 비정형 데이터를 분석하는 별도의 도구를 개발하여 분석하게 되는 경우가 많다. 필요에 따라서는 정형 데이터를 분석하는 도구를 이용하여 분석할 수도 있는데, 이 경우에는 비정형 데이터의 가공 또는 형태 변환 등의 다양한 기법을 사용하여 정형 데이터 형태에 최대한 가깝게 변환한 이후 전통적 분석 방법을 적용하게 된다. 대표적인 비정형 데이터의 형태로서 텍스트 데이터, 이미지 데이터, 음향 데이터 또는 이러한 형태가 융합된 데이터 등을 꼽을 수 있다.

정형 데이터와 비정형 데이터의 어느 한쪽으로 명확히 구분되지 않고 양쪽의 성격을 모두 가지고 있는 데이터도 존재하는데, 이를 반정형 데이터라고 부르기도 한다. 예를 들어 텍스트 데이터와 계량화된 데이터가 병합된 형태를 꼽을 수 있다. 반정형 데이터는 분석의 목적에 따라 변형하여 사용하거나 일부를 추출하여 사용할 수도 있다.

지도 방법

생활 속에 정형 데이터가 많긴 하지만, 이미지, 영상, 텍스트, 음성과 같은 데이터도 우리 일상에 깊숙하게 자리잡은 데이터들이다. 따라서 전체 데이터 중 비정형 데이터가 훨씬 더 많아지고 있고, 비정형 데이터를 분석하는 기법들이 발전함에 따라 비정형 데이터의 활용이 점점 더 쉽고 간편해지고 있음을 설명한다.
또한 정형 데이터와 비정형 데이터뿐 아니라 json과 같은 반정형 데이터들도 일상생활에서 점점 더 많이 활용되고 있는 흐름이기 때문에 각각의 데이터의 특징에 대해 이해할 수 있도록 지도한다.

02 비정형 데이터

비정형 데이터(unstructured data)는 구조를 명확하게 정의하기 어려운 데이터를 말한다. 비정형 데이터는 정형 데이터처럼 명확한 속성이 정의되어 있지 않지만, 데이터를 가공하여 속성을 추출할 수 있다. 예를 들어, 영화에 대한 사람들의 댓글 리뷰 데이터를 숫자로 변환하면, 이를 영화의 새로운 속성으로 사용할 수 있다. 대표적인 비정형 데이터로는 이미지, 텍스트, 음성 등 우리 생활에서 자주 접하는 데이터가 있다. 비정형 데이터는 구조에 제약을 받지 않고 자유롭게 표현할 수 있다는 장점이 있다.

알고 가기 반정형 데이터

정형 데이터와 비정형 데이터 외에도 반정형 데이터(semi-structured data) 형태의 데이터도 최근 데이터 과학에서 많이 활용되고 있다. 반정형 데이터는 일부분이 정형화된 데이터로, 구조가 존재하지만 구조를 바꿀 수도 있는 유연한 형태의 데이터를 말한다. 대표적인 반정형 데이터로는 XML(eXtensible Markup Language)과 JSON(JavaScript Object Notation)이 있다. 이러한 반정형 데이터는 데이터 구조를 쉽게 변경할 수 있고, 필요한 경우 표 형태로도 쉽게 변환할 수 있어 데이터 공유에 많이 활용된다.

공공데이터포털에서 활용되는 반정형 데이터(XML, JSON)

2 | 데이터의 속성

다음은 어느 지역의 도서관 관련 데이터를 표로 나타낸 것이다. 표와 같은 정형 데이터는 데이터의 속성을 파악하기 쉽기 때문에 별도의 설명이 없어도 다양한 사실을 알 수 있다. 예를 들어, '시작시각' 속성을 통해 도서관이 언제 시작하는지 알 수 있고, '휴관일' 속성을 통해 어떤 요일에 휴관하는지 알 수 있다. 또 '위도'와 '경도' 속성을 활용하면 도서관의 위치까지 알 수 있다. 이와 같이 행과 열로 이루어진 표 형태의 정형 데이터는 구조가 명확하여 데이터의 속성을 쉽게 파악할 수 있다.

도서관명	도서관 유형	휴관일	평일운영 시작시각	평일운영 종료시각	토요일 운영 시작시각	토요일 운영 종료시각	열람 좌석수	자료수 (도서)	대출가능 권수	대출가능 일수	위도	경도
A 도서관	공공 도서관	일	8:00	22:00	8:00	22:00	92	60080	5	14	36.34926753	127.330255
B 도서관	공공 도서관	월	8:00	22:00	8:00	22:00	212	90102	5	14	36.33771587	127.3374864
C 도서관	공공 도서관	월	8:00	22:00	8:00	22:00	98	48726	5	14	36.38210426	127.3204213
F 어린이 도서관	공공 도서관	금	9:00	18:00	9:00	18:00	150	65915	10	14	36.30504773	127.3684557
G 교육청 도서관	공공 도서관	둘째, 넷째주 월	7:30	22:00	7:30	22:00	1600	304305	10	15	34.80731478	126.3761686
H 정보도서관	공공 도서관	수	9:00	23:00	9:00	22:00	136	66137	5	14	37.486252	126.959668
J 주민자치 도서관	작은 도서관	토 · 일	9:00	18:00	0:00	0:00	20	3440	5	20	36.29561259	127.1412686
K 도서관	공공 도서관	월	8:00	21:00	9:00	18:00	115	68037	10	14	36.12620527	127.0973608
L 도서관	공공 도서관	월	9:00	18:00	9:00	18:00	158	55110	3	14	35.45273826	128.5305123

표 I-3 | 도서관 관련 데이터

한편 데이터의 속성은 문제 상황에 따라 그 중요성이 달라질 수 있다. 데이터를 활용해 우리 학교에서 가장 가까운 도서관이 어디인지 확인할 때는 위도와 경도 속성이 중요한 속성이 되고, 도서관에서 과제에 필요한 책을 언제 찾으러 갈 것인지 결정할 때는 휴관일과 시작시각이 중요한 속성이 될 수 있다. 또한 문제 해결에 필요한 새로운 속성이 추가되면 데이터의 가치가 더 높아질 수도 있기 때문에 데이터의 속성은 데이터의 가치를 결정하는 데 중요한 역할을 한다.

지도 방법

도서관 데이터를 보면서 각 행과 열을 살펴보되 각각의 속성을 살펴보며 해당 열의 데이터로 어떤 사실을 알 수 있는지 이해할 수 있도록 지도한다.
특히, 문제에 따라 필요한 속성이 달라짐을 주지시키도록 한다.

❷ 정형 데이터의 구분

실증적인 연구에 있어서 데이터는 매우 중요한 역할을 하게 된다. 연구를 위한 질문을 세우고 이를 증명하기 위해 데이터를 수집하고 분석하는데, 특히 자연 과학이나 사회 과학 등 다양한 연구에서 데이터는 어떤 형태의 측정을 통해 얻어지는 경우가 많다. 이러한 측정의 기준이 정량적(定量的)일 수도 있고, 정성적(定性的)일 수도 있다. 예를 들면, 신체 규격을 측정하는 것처럼 명확한 측정의 수단과 기준이 존재하기도 하는 반면 설문 조사 등을 통한 질적 연구와 같이 연구 대상의 특성치나 상태를 적절한 측도를 생성하여 측정하기도 한다. 정형 데이터 중 다음 세부적인 데이터의 유형은 측정 방법에 따라 데이터를 형태적 기준으로 분류한 체계이다.

(1) 범주형 데이터: 명목형 데이터 및 순서형 데이터

명목형 데이터란 순서가 없이 어떤 특성을 지니고 있는 관측치로 이루어진 데이터이다. 예를 들면, 거주 지역을 나타내는 우편번호 데이터라든지, 다양한 인종의 특징을 나타내는 인종 구분 데이터, 교과목을 구분하는 교과목 코드 데이터 등이 명목형 데이터라고 할 수 있다. 우편번호의 예를 조금 더 자세히 살펴보자. 우편번호는 비록 수치적 형태로 표현되어 있으나 수치 자체의 의미가 아니라 편의에 따른 지역 구분을 의미하는 것이므로 유의해야 할 필요가 있다.

인종 데이터는 우리나라 인구 동태적 데이터에는 잘 나타나지 않지만 미국 등과 같은 다인종 국가에서는 빈번하게 조사되고 발표되는 데이터로서 주로 백인, 흑인, 아시안, 하와이 원주민, 아메리칸 인디언 등의 범주로 나누게 된다. 우편번호와 마찬가지로 인종별 코드가 부여되어 수치적 형태로 표기되기도 한다. 교과목 코드나 질병 코드 등도 편의에 따라 수치 또는 문자적 형태로 나타낼 수도 있다.

순서형 데이터는 연속적인 수치형 데이터처럼 정교한 계량화는 어렵지만 순서는 지정할 수 있는 데이터이다. 즉, 수치적 정밀도는 제한적이지만 크고 작은 수량적 크기나 의미적인 순서 비교가 가능한 데이터라고 할 수 있다. 예를 들어 학력 데이터나 리커트 척도(Likert Scale)*와 같은 평가 데이터이다. 학력 데이터는 중학교 졸 이하, 고등학교 졸, 대학교 졸, 대학원 졸 등 범주를 구분하고 해당하는 범주를 선택한 관측치로 구성된다. 우편번호의 사례와는 정반대로 수치화되어 있지 않은 데이터이지만, 의미상 순서가 존재하는 데이터라고 할 수 있다. 설문지 작성에 종종 사용되는 리커트 척도는 응답자의 생각을 정확한 수치로 나타내기 어려우나 해당 질문에 동의하는 정도를 기입할 수 있도록 순서형 범주를 생성한 것이다. 특정 주제에 관한 질문에 '매우 동의함', '다소 동의함', '동의하지도 않고 반대하지도 않음', '다소 반대함', '매우 반대함' 등으로 응답지를 구성하는 사례를 생각해 볼 수 있다. 이를 5점 리커트 척도라고 부른다. 범주를 축소하거나 늘려서 3점 또는 7점 리커트 척도도 사용할 수 있겠으나 통상 5점 척도를 많이 사용한다. 이러한 순서형 데이터의 특징은 앞서 살펴본 것처럼 정확하고 정밀한 수치적 계량화가 불가능하지만 상대적인 크기를 비교함으로써 간접적이나마 수치적 비교를 할 수 있다는 것이다. 다만, 인접한 범주 간의 크기 비교가 정확한 간격으로 이루어질 수 없다는 제약이 있다. 예를 들면, '매우 동의함'과 '다소 동의함'의 간격이 '동의하지도 않고 반대하지도 않음'과 '다소 반대함'의 간격과 동일하다고 단언할 수는 없을 것이다. 따라서 어떤 주제를 정하고 질문을 구성할 때, 이러한 범주를 정하고 순서를 부여하여 평가하는 것이 적절하고 합리적인지를 면밀히 검토할 필요가 있다. 경우에 따라서는 리커트 척도를 이용하는 것이 측정 목적에 부합하지 않을 수 있기 때문이다.

지도 방법

[알고 가기]에 제시된 데이터형은 뒤에서 다루게 될 판다스(pandas)에서 주로 활용되는 데이터형으로, 일반적인 파이썬 프로그래밍 언어에서 사용되는 데이터형과는 조금 다름을 안내하도록 한다.

해 보기 ❶ 지도 방법

교과서 21쪽에 제시된 도서관 데이터를 활용할 수도 있지만, 실제 데이터를 다운로드하여 우리 지역의 도서관 데이터를 살펴보면서 수업을 하면 더 몰입감 있는 활동을 할 수 있다.

알고 가기 파이썬 언어의 다양한 데이터형

데이터의 각 속성에는 숫자, 문자 등 다양한 형태의 데이터가 저장된다. 이 중 파이썬(python) 프로그래밍 언어를 기반으로 데이터 분석에서 사용되는 대표적인 데이터형(data type)은 다음과 같다.

데이터형 이름	데이터의 형태	예시
object	텍스트(문자열)	'공공도서관', '작은 도서관'
int64	정수	92, 60080
float64	실수	36.34926753, 127.330255
bool	참(True) 또는 거짓(False)	True, False
datetime64	날짜와 시간	2025-03-01 13:30:03

해 보기 ❷ 도서관 데이터의 다양한 속성이 가진 가치 살펴보기

탐색하기

❶ 21쪽에 제시된 도서관 데이터를 사용하여 다음 문제 상황을 해결하고자 할 때, 문제 상황에 따라 필요한 속성이 어떻게 달라질지 생각해 보자.

문제 상황	우리 집이랑 가장 가까운 도서관은 어디일까?	빌려야 할 책이 많다면 어떤 도서관으로 가는 것이 좋을까?	다양한 책을 찾고 싶다면 어떤 도서관으로 가는 것이 좋을까?
필요한 속성	㉠ 위도, 경도	대출 가능 권수	자료 수

❷ 제시된 속성 외에 어떤 속성이 있으면 도서관 데이터의 가치를 높일 수 있을지 생각해 보자.

㉠ 내가 관심 있는 분야의 책이 몇 권씩 소장되어 있는지 기록된 속성

신착 도서 수, 청소년 열람실 도서 수

소단원 1분 요약

❶ 데이터를 구조적인 특성에 따라 분류하면 구조가 명확하게 정의된 정형 데이터와 구조를 명확하게 정의하기 어려운 비정형 데이터로 나눌 수 있다.
❷ 데이터의 속성은 데이터의 가치를 결정하는 데 중요한 역할을 하며, 문제 상황에 따라 그 중요성이 달라질 수 있다.

22

탐구 활동 '공공데이터포털'에서 다양한 형태의 데이터 수집하기

'공공데이터포털(https://www.data.go.kr)' 사이트는 공공 기관에서 제공하는 데이터를 쉽고 편리하게 사용할 수 있도록 정부에서 운영하는 곳이다. 이곳에서는 다양한 분야의 데이터가 정형, 비정형, 반정형 형태로 제공되고 있다.

공공데이터포털 화면

공공 데이터의 예

사고 유형별 교통사고 통계(도로교통공단)

독도의 사계절(외교부)

❶ 위에서 예시로 제시된 두 가지 데이터는 어떤 형태의 데이터인지 구분해 보자.

예시 데이터	정형 데이터	비정형 데이터
사고 유형별 교통사고 통계 데이터	○	
독도의 사계절 데이터		○

❷ '공공데이터포털' 사이트에서 관심 분야의 데이터를 내려받고, 어떤 형태의 데이터인지 구분해 보자.

내가 수집한 관심 분야의 데이터	정형 데이터	비정형 데이터
행정안전부_동물병원	○	
AI기반 생태통로 모니터링 야생동물 학습데이터		○

❸ 관심 분야의 데이터로 어떤 문제를 해결할 수 있을지 생각해 보고, 데이터의 잠재적 가치에 대해 이야기해 보자.

생태 통로에 어떤 동물들이 어떤 패턴으로 이동하는지 사람이 직접 관찰하지 않아도 확인할 수 있기 때문에 생태 연구에 도움을 받을 수 있다.

23

탐구 활동 지도 방법

- 표 형태의 구조가 정해진 데이터를 정형 데이터라고 하고, 이미지나 텍스트와 같이 구조가 정해지지 않은 데이터가 비정형 데이터임을 실제 데이터를 살펴보며 구분할 수 있도록 한다.
- 학생들이 처음부터 스스로 데이터를 탐색하는 것이 어렵기 때문에 선생님이 학급의 학생 중 한 명의 관심사를 물어보고, 그에 대한 데이터를 찾는 과정을 보여 줌으로써, 학생들이 자신의 관심사에 맞는 데이터를 찾는 과정을 이해할 수 있도록 지도한다.

예시 답안

- 행정안전부_동물병원 주소: https://www.data.go.kr/data/15045050/fileData.do
- AI기반 생태통로 모니터링 야생동물 학습데이터 주소: https://www.bigdata-transportation.kr/frn/prdt/detail?prdtId=PRDTNUM_000000020318

(2) 수치형 데이터: 이산형 데이터 및 연속형 데이터

수치형 데이터는 수치적 크기를 측정한 관측치로 이루어진 데이터이다. 수치적 크기는 객관적이고 정확한 측정 도구를 통해 얻어진 관측치로서 연구를 위한 실험에서 물리적인 도구를 이용하여 측정한 관측치 등이 이에 속한다. 수치형 데이터는 수치적 크기가 명확하게 구분되므로 순서형 데이터보다 더 많은 정보를 내포한 데이터라고 할 수 있다. 특히 각 관측치에 나타난 수치적 크기가 직접적으로 비교가 되고 인접한 관측치와의 격차는 그 자체로 명확한 의미를 지니고 있기 때문이다. 예를 들어, 리커트 척도에서 '매우 동의함'에 5점을 부여하고 '다소 동의함'에 4점을 부여하며 '동의하지도 않고 반대하지도 않음'에 3점을 부여한다고 할 때, 5점과 4점의 격차인 1점과 5점과 3점의 격차인 2점 사이에 정률적으로 2배 관계가 있다고 해석하기는 어렵다. 반면 지역 A에서 지역 B에 이르는 거리가 1.5km이고 지역 A에서 지역 C에 이르는 거리가 3km라고 한다면, 이 두 거리 사이에는 2배의 수치적 관계가 있다는 설명이 가능하다. 다만, 수치형 데이터를 해석함에 있어서도 유의할 점은 있다. 특히 비율 척도를 가지고 관측치를 해석하는 경우에 다소 부자연스러울 수 있다. 예를 들면, 거리의 사례에서는 비율 척도로 거리를 비교하는 데에는 문제가 없었으나, 이에 비해 온도라든지, 습도 등의 측도를 비교할 때 다소 어색한 부분이 발생할 수 있다. 온도나 습도의 경우에는 수치적 차이만을 가지고 2배만큼 따뜻하다거나 2배만큼 습하다고 해석하는 것이 의미가 없기 때문이다. 그러므로 수치형 데이터를 해석하고 비교함에 있어서 간격 또는 비율 척도를 목적에 맞도록 적절하게 활용해야 한다.

수치형 데이터의 형태를 조금 더 세부적으로 구분해 보면, 연속형 데이터와 이산형 데이터로 나누어 볼 수 있다. 양적인 크기가 있다는 점에서는 연속형 데이터와 이산형 데이터가 공통점이 있지만, 수치적 크기를 표현하고 비교하는 데에는 차이점이 있다. 이산형 데이터의 대표적인 예로는 가산(count)형 데이터를 꼽을 수 있다. 어떤 주차장의 주차 가능 대수, 각 가정별 구성원 수, 특정 학과의 신입생 수 또는 편입생 수 등 수량을 세어 나가는 데이터가 가산형 데이터이다. 수직선상에서는 0, 1, 2, 3 등 0과 자연수에 해당하는 지점에만 나타나는 관측치가 되겠다. 반면, 길이, 무게, 부피, 온도 등 측정의 기준이 연속적이라면 이러한 관측치가 모여 연속형 데이터를 이루게 된다. 수직선상에서도 실수의 범위 내 어느 곳에도 표현될 수 있는 데이터이다.

* **리커트 척도(Likert Scale):** 응답자가 진술문에 대해 동의나 불일치 정도를 표시하는 데 사용되는 심리 측정 척도이다. 이는 주관적인 태도나 의견을 측정하는 데 사용되며, 응답자는 일련의 문항에 대해 강하게 동의하지 않음에서 강하게 동의함까지 다양한 선택지 중에서 선택한다.

[출처] 장영재, 유찬우, 『데이터 과학 개론』, 방송통신대학교(2022)

5 JSON 이해하기

JSON은 데이터를 주고받을 때 쓰이는 형식으로, 사람이 읽을 수 있는 텍스트로 구성된다. 원래 자바스크립트(JavaScript)에서 시작되었지만, 모든 프로그래밍 언어에서 사용할 수 있다. 주로 인터넷에서 정보를 교환할 때 사용된다.

- **정의:** 64비트(8바이트) 크기의 부동 소수점형 데이터 타입
- **범위:** ±1.7E-308 ~ ±1.7E+308 (근삿값) 및 15~17자리의 유효 숫자
- **용도:** 실수(소수점이 있는 숫자) 연산을 정확하게 다루기 위해 사용되며, 일반적으로 과학적 계산 및 통계적 분석에서 사용된다.

☑ json 코드 예시

```
{
  "restaurant": "삼미 식당",
  "menu": [
    {
      "name": "불고기",
      "price": 12000,
      "category": "메인 요리"
    },
    {
      "name": "비빔밥",
      "price": 10000,
      "category": "메인 요리"
    },
    {
      "name": "식혜",
      "price": 4000,
      "category": "음료"
    }
  ]
}
```

☑ 도서관 데이터 수집 방법

[공공데이터포털] – [전국도서관표준데이터] – [CSV 또는 XLS 파일 다운로드]

☑ int64 이해하기

- 정의: 64비트(8바이트) 크기의 정수형 데이터 타입
- 범위: −2^63 ~ 2^63 − 1
- 용도: 매우 큰 정숫값을 다루기 위해 사용되며, 일반적으로 정수 연산이 필요할 때 사용된다.

☑ [탐구 활동]을 위한 데이터 찾기 – 방법 안내

공공데이터포털에서 학생들이 관심을 가질 만한 데이터를 찾기 위해서는 키워드 검색 방법도 좋지만, [데이터 찾기] – [국가 중점 데이터]를 살펴보거나 [데이터 찾기] – [이슈 및 추천 데이터]를 살펴보며 관심 있는 분야의 데이터를 찾는 것이 좋다.

단, 공공데이터포털의 데이터는 대부분 정형 데이터나 비정형 데이터 중심이기 때문에 비정형 데이터를 찾기 위해서는 AI허브(https://aihub.or.kr/)에 있는 다양한 데이터를 찾아보는 것도 좋은 방법이다. AI허브에는 이미지, 자연어 데이터를 포함한 다양한 비정형 데이터가 많이 있다.

MEMO

3 데이터셋과 데이터베이스

수업 시간: 3시간 24~31쪽

단원명	Ⅰ. 데이터 과학의 이해 03. 데이터셋과 데이터베이스
학습 목표	• 데이터셋과 데이터베이스를 설명할 수 있다. • 서로 다른 데이터셋을 통합하는 데이터베이스의 필요성을 설명할 수 있다.
수업 방법	강의, 토론, 발표, 실습
준비물	교사 교과서, 관련 교수 학습 자료, 컴퓨터 학생 필기도구

단계	교수 · 학습 방법	지도상의 유의점
도입	생각 열기 만화를 통해 서로 다른 데이터를 통합했을 때 어떤 효과가 나타날 수 있는지 생각해 보도록 안내한다.	• 학생에게 잔반을 줄인다면 어떻게하면 좋을지 의견을 이야기하도록 하고, 이를 데이터에 기반해서 해결하기 위해 어떤 데이터가 필요할지 생각해 보도록 지도한다.
전개	1. 데이터셋 데이터셋의 의미와 사례를 소개한다.	• 데이터셋의 중요성을 이해하는 데 도움이 되는 CIFAR-10 데이터셋에 대한 이야기를 들려줄 수 있다.
	2. 데이터베이스 데이터베이스의 개념과 이점을 안내한다.	• 데이터셋과 데이터베이스의 관계를 잘 이해할 수 있도록 지도한다.
	3. 데이터베이스의 통합적 활용 데이터베이스의 테이블에 대한 개념을 소개하고, 다양한 테이블이 서로 어떻게 연결되는지 이해하도록 한다.	• 너무 전문적인 내용을 설명하기보다는 다양한 테이블이 데이터베이스를 통해 서로 통합된다는 점을 이해시키는 데 중점을 두는 것이 좋다.
	4. 데이터베이스의 생활 속 활용 데이터베이스의 기본 기능을 살펴보며 생활 속에서 데이터베이스가 얼마나 많이 활용되는지 이해하도록 한다.	• 학생들에게 친숙한 서비스를 하며 어떤 데이터가 저장되고 관리되는지 생각해 보도록 지도한다.
정리	탐구 활동 인터넷 쇼핑몰을 만들기 위해 회원 관리 데이터베이스를 어떻게 만들 수 있을지 고민해 보고, 직접 장바구니 테이블을 설계해 보도록 한다.	• 쇼핑몰 사용자 테이블, 제품 테이블을 설계하는 방법을 차근차근 설명하며, 학생들이 주도적으로 장바구니 테이블을 설계해 보도록 지도한다.
평가	• 생활 속 데이터를 정형 데이터와 비정형 데이터로 구분할 수 있는가? • 데이터 속성에서 데이터의 잠재적 가치를 판단할 수 있는가?	• 학생들이 실생활의 관심 데이터를 바탕으로 데이터의 유형을 구분하고, 데이터 속성에 따라 어떤 문제를 해결할 수 있을지 충분히 생각하여 자신의 생각을 표현하도록 한다.

3 데이터셋과 데이터베이스

제시 의도

하나의 데이터셋은 특정한 목적을 바탕으로 수집된 것이다. 따라서 하나의 데이터셋만을 분석할 때도 의미 있는 결과를 도출할 수 있지만, 복잡한 문제를 해결하려고 할 때는 하나의 데이터셋이 아닌 서로 다른 데이터셋을 통합할 필요가 있다. [생각 열기]에서는 이런 개념을 실생활의 급식과 잔반이라는 주제로 제시하며 학생들이 데이터 통합의 필요성을 깨닫도록 하였다.

지도 방법

- 학생들에게 가장 많이 남기는 반찬이 무엇인지 적어 보도록 한다. 그리고 그 이유가 무엇인지 작성해 보도록 한다.
- 다양한 의견이 나오지만, 이것을 실제 데이터로 측정할 수 있는 방법이 있다면 문제를 해결하는 데 도움이 될 수 있음을 설명하며 함께 삽화를 보도록 한다.
- 이런 데이터를 통합하여 분석한다면 어떤 새로운 사실을 알아낼 수 있을지 함께 생각해 보도록 한다.

예시 답안

❶ 잔반이 많이 발생하는 메뉴가 무엇인지 알아낼 수 있다.
❷ 같은 메뉴라도 다른 메뉴와의 조합에 따라 잔반이 많아지기도 하고, 적어지기도 함을 알아낼 수 있다.

3 데이터셋과 데이터베이스

학습 목표
- 데이터셋과 데이터베이스를 설명할 수 있다.
- 서로 다른 데이터셋을 통합하는 데이터베이스의 필요성을 설명할 수 있다.

학습 요소
- 데이터셋, 데이터베이스

생각 열기

데이터 통합의 중요성

다음 사례를 통해 서로 다른 데이터들을 통합하는 것이 왜 중요한지 생각해 보자.

위 사례에서 서로 다른 데이터를 통합했을 때 기존 데이터에서는 알 수 없었던 어떤 새로운 사실을 알아낼 수 있었을까?

1 | 데이터셋

데이터셋(dataset)이란 하나의 주제에 대한 여러 데이터들이 모인 집합(set)을 말한다. 예를 들어 급식 메뉴에 대한 여러 데이터를 하나로 모은 표는 급식 데이터셋이라 할 수 있고, 생활 속의 사물이나 동물에 대한 이미지 데이터를 하나로 모은 것은 생활 이미지 데이터셋이라 할 수 있다.

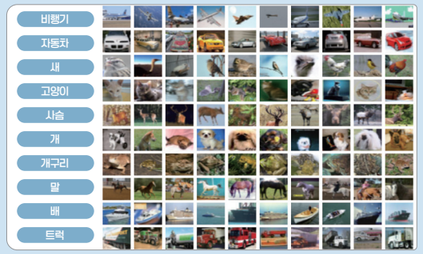
데이터와 데이터셋

〈급식 데이터셋〉 급식 메뉴에 대한 여러 데이터가 모여서 급식 데이터의 집합을 이룸

학교명	급식 일자	급식 인원 수	요리명	칼로리 정보
A중학교	20230927	513	오곡밥 알토란국 매운돈사태갈비찜 삼색나물 배추김치 전통식혜 송편	833.3 kcal
B중학교	20230927	170	다시마무채국 깍두기 요구르트 열대과일샐러드 옛날도시락비빔밥 스마일감자&계절	871.0 kcal
C고등학교	20230927	600	찰쌀밥 쇠고기토란국 매운돼지갈비찜 호박볶음 배추겉절이 미니약과&송편 육전&불고기산적&양파소스	781.4 kcal
D고등학교	20230927	1051	혼합잡곡밥 시금치된장국 매운돼지갈비찜 배추김치 송편 달빔라테 모듬전&초간장	1252.7 kcal
E고등학교	20230927	150	혼합잡곡밥 굴림고기만두국 계맛살오이냉채 스펌구이&계절 배추김치 레드자몽에이드 전무무침	1439.8 kcal

표 Ⅰ-4 | 급식 데이터셋

〈생활 이미지 데이터 및 데이터셋〉 CIFAR-10



CIFAR-10 데이터셋은 10개의 주제(비행기, 자동차 등)에 대한 이미지가 각각 6,000개로 구성된 이미지 데이터셋으로, 2009년에 알렉스 크리체프스키(Alex Krizhevsky), 제프리 힌튼(Geoffrey Hinton) 등이 수집 및 공개했다.

참고 도서

시빅 데이터의 모든 것

- **도서명**: 우리에게는 다른 데이터가 필요하다
- **도서 내용**: 시빅 데이터(Civic Data)의 개념과 활용법, 나아갈 방향에 이르기까지 시빅 데이터의 모든 것을 국내에 본격적으로 소개하는 최초의 책이다. 시빅 데이터란 '시민을 위한 데이터'를 의미한다. 복지뿐 아니라 행정 전반에서 시빅 데이터를 어떻게 활용하면 모두의 일상이 더 쉽고 편해지는지, 정부가 시빅 데이터를 어떻게 관리하면 한국의 민주주의가 더 성숙할 수 있는지를 조망한다. 공직자의 편의와 업무 중심으로 설계한 정책과 데이터는 복지 사각지대를 만들어 내는 것은 물론, 시민의 일상을 불편하고 짜증나게 만든다. 이 과정에서 생겨나는 시빅 데이터와 시민 간 공백은 약자들을 더욱 가난하고 아프게 만들고, 때로는 충분히 예측 가능한 사고조차 막지 못해 귀중한 목숨을 희생시킨다. 미국의 대표적 시빅 테크 단체인 '코드 포 아메리카' 소속 데이터 사이언티스트이자 존스홉킨스대 SNF 아고라 연구소 연구 위원이며 KDI 국제정책대학원 교수를 역임한 저자는, 이 책에서 10가지 키워드를 통해 시빅 데이터를 설명한다. 시빅 데이터의 발전사부터 한국과 미국의 현주소, 미국의 다양한 시빅 데이터 활용 사례, 한국이 고민해야 할 지점들을 조목조목 꼬집는다. 또한 '공공성'에 대한 인식 개선이 우리 사회에 어떤 긍정적 변화를 가져올 수 있는지 소개한다. 방대한 통계 자료와 사례를 바탕으로 쓴 이 책은 '공공성'과 '테크'를 둘러싼 여러 논쟁과 편견을 해소할 뿐 아니라, 사람이 중심이 되고 기술은 사람을 보조하는 사회를 만드는 데 영감을 주는 다양한 인사이트를 제공할 것이다.

|주소| 예스24, https://www.yes24.com/Product/Goods/122281427

1 ImageNet 데이터셋과 개인 정보 보호

ImageNet은 AI 분야 연구자라면 모르는 사람이 없을 정도로 AI 기술 발전에 큰 영향을 준 데이터셋이다. 많은 수의 이미지들과 그에 대한 메타데이터(metadata)로 구성된 이 데이터셋은 약 1,400만 개에 달하는 이미지들로 구성되어 있으며, 이들은 2만 개에 달하는 카테고리로 분류되어 있다. 수많은 카테고리들을 유의미하게 정의하기 위해서, 1980년에 만들어진 단어 계층 데이터셋인 WordNet을 활용하였으며, 이 카테고리별로 인터넷에서 이미지들을 크롤링하여 수집한 후, Amazon Mechanical Turk를 이용해 라벨링 작업이 진행되었다.

이렇게 수집된 데이터셋에서 1,000개의 카테고리, 100만 장의 이미지를 선별하여 ImageNet Competition이라는 이름의 유명한 이미지 관련 AI 경진대회가 지속적으로 개최되어 왔고, 이를 통해서 현재 잘 알려져 있는 AlexNet, VGG, GoogleNet, ResNet과 같은 네트워크 구조들이 유명세를 타게 된다. 다만, 크롤링에 의해서 수집되었기 때문에 ImageNet 데이터셋에는 수많은 사람들의 얼굴도 포함되어 있고, 그들의 동의를 별도로 얻지 않았다는 이슈가 제기되었다. 최근 AI 학습 데이터의 수집 적법성 이슈가 점차 중요하게 다루어지고 있는데, ImageNet 또한 예외는 아니어서, 이미지 내 포함된 얼굴 부분에 대해 흐림 처리를 하기로 했다는 소식이 전해지고 있다.

[출처] https://smilegate.ai/2021/04/01/imagenet-privacy/

2 데이터셋과 데이터베이스

데이터셋과 데이터베이스라는 용어를 혼동하는 경우도 많다. 데이터베이스와 데이터셋 모두 데이터의 구성과 관리를 설명하는 데 사용되는 관련성 높은 용어이지만, 다음과 같은 몇 가지 의미 있는 방식에서 차이가 있다.

데이터셋은 분석 및 모델링에 사용되는 데이터 모음이며, 일반적으로 구조화된 형식으로 구성된다. 구조화된 형식은 Excel 스프레드시트, CSV 파일, JSON 파일 또는 기타 형식일 수 있다. 데이터셋의 데이터는 여러 가지 방법으로 구성되는데, 고객 설문 조사, 실험 또는 기존 데이터베이스와 같은 다양한 소스에서 생성될 수 있다. 데이터셋은 기계학습 모델의 학습 및 테스트, 데이터 시각화, 연구 또는 통계 분석을 포함한 다양한 용도로 사용할 수 있다. 데이터셋은 공개적 또는 제한적으로 공유할 수 있다. 데이터셋은 일반적으로 데이터베이스에 비해 크기가 작다. 데이터베이스는 전자 방식으로 저장된 대량의 조직화된 데이터를 장기간 저장하고 관리하도록 설계되었으므로, 데이터에 쉽게 액세스하고 조직 및 업데이트할 수 있다. 즉, 데이터베이스는 여러 개의 데이터셋으로 저장된 데이터가 구성된 모음이다.

[출처] https://www.databricks.com/kr/glossary/what-is-dataset

3 공공 데이터를 활용한 우수 사례 알아보기

[공공데이터포털] – [데이터활용] – [공공데이터 우수사례]를 살펴보면 다양한 공공 데이터를 활용한 우수 사례를 살펴볼 수 있다.

[출처] https://www.data.go.kr/tcs/eds/ctm/selectContestDataList.do

용어 해설 **데이터베이스**

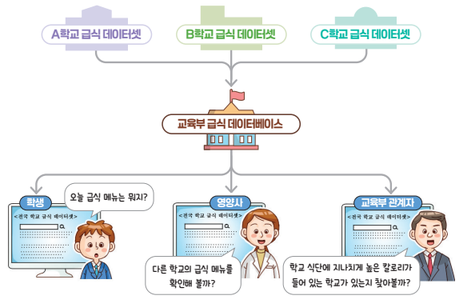
데이터베이스(DB: database)는 여러 사람이 공유하여 사용할 목적으로 체계화한 후, 통합 · 관리하는 데이터의 집합이다. 작성된 목록으로써 여러 데이터베이스 관리 시스템(DBMS)의 통합된 정보들을 저장하여 운영할 수 있는 공용 데이터들의 묶음이다. 데이터베이스에 속해 있는 모델은 다양하다.

용어 해설 **데이터베이스 관리 시스템**

데이터베이스 관리 시스템(DBMS: DataBase Management System)은 다수의 사용자들이 데이터베이스 내의 데이터를 접근할 수 있도록 해 주는 소프트웨어 도구의 집합이다. DBMS은 사용자 또는 다른 프로그램의 요구를 처리하고 적절히 응답하여 데이터를 사용할 수 있도록 해 준다.

2 | 데이터베이스

데이터베이스(database)는 여러 사람이 공유하여 사용하기 위해 통합 관리되는 데이터의 집합으로, 데이터베이스 시스템을 사용하면 대규모의 데이터를 효율적으로 저장하고 검색하며, 업데이트하고 관리할 수 있다. 여러 사람이 대규모의 데이터셋을 통합적으로 관리하고 사용하려면 데이터베이스가 필요하다. 예를 들어, 데이터베이스를 활용하면 전국 모든 학교의 급식 데이터셋을 통합하여 관리하고 다른 학교의 데이터셋을 열람할 수 있다.



데이터베이스를 사용하여 데이터를 통합적으로 관리하면 다음과 같은 이점이 있다.

편리한 버전 관리: 데이터베이스를 활용하면 언제, 누가, 어떤 데이터를 수정했는지 기록을 유지할 수 있다. 따라서 데이터에 오류가 발생했을 경우 원인을 찾아 그 이전 버전으로 쉽게 복구할 수 있다. 예를 들어, 누군가가 실수로 중요한 데이터를 삭제했다고 하더라도 그 이전 버전으로 복구할 수 있다.

데이터 무결성 유지: 데이터의 제약 조건을 설정하여 데이터 입력, 수정, 삭제 시 발생할 수 있는 오류를 줄일 수 있다. 예를 들어, 식단의 칼로리 항목에 −1000과 같은 값이 입력되거나 같은 내용이 중복되어 입력될 경우 입력되지 않도록 할 수 있다.

데이터 보안 강화: 데이터베이스는 사용자별로 데이터 접근 권한을 설정할 수 있다. 이를 통해 민감한 정보를 안전하게 보호하고, 각 사용자에게 필요한 데이터만 제공할 수 있다. 예를 들어, 다른 학교의 급식 데이터를 수정하지 못하게 접근 권한을 설정할 수 있다.

체계적인 데이터 관리: 데이터베이스를 사용하면 데이터를 체계적으로 분류, 저장, 검색할 수 있다. 예를 들어, 급식 메뉴, 음식의 재료, 알레르기 정보 등 다양한 데이터를 체계적으로 저장할 수 있다. 또한, 학생이나 학부모들이 특정 날짜의 급식 정보를 알고 싶을 때, 데이터베이스를 통해 실시간으로 데이터를 검색할 수 있다.

3 | 데이터베이스의 통합적 활용

인터넷 게시판, 블로그, 메신저 등 대부분의 서비스는 데이터베이스를 사용한다. 그리고 데이터베이스에는 여러 종류의 데이터셋이 들어가는데, 이때 각각의 데이터셋은 표 형태로 저장되며 이를 테이블(table)이라고 한다.

만약 도서관에 데이터베이스가 없다면, 내가 원하는 책이 도서관에 있는지 확인하기 어려울 뿐 아니라, 대출 여부를 확인하는 것은 더욱 어렵다. 하지만 우리는 도서관 홈페이지를 통해 책이 도서관에 구비되어 있는지 쉽게 확인할 수 있고, 만약 누군가가 대출해 갔다면 반납 예정일이 언제인지도 알 수 있다. 또한, 여러 도서관 중에서 내가 찾는 책을 빌릴 수 있는 도서관은 어디인지도 쉽게 검색할 수 있는데, 이를 위해 데이터베이스는 여러 데이터를 통합하여 관리한다.

예를 들어 도서관 데이터베이스에는 도서관에서 소장하고 있는 책의 저자나 출판사 등이 저장된 책 목록 테이블과 회원의 이름과 회원 번호, 연락처 등이 저장된 명단 테이블, 그리고 누가 언제부터 언제까지 책을 대출했는지가 저장된 대출 기록 테이블이 있다.

데이터베이스의 각 테이블에 저장된 데이터들은 공통적인 속성을 기반으로 서로 연결될 수 있다. 예를 들어, 책 목록 테이블에 책의 대출 여부를 나타내는 속성이 없을 경우, 해당 책의 도서관리번호를 대출 기록 테이블에서 확인하여 대출 여부를 확인할 수 있다.

해 보기 1 **데이터베이스의 필요성 생각해 보기** 설명하기

모둠 활동 또는 과제를 수행했던 경험을 떠올려 보며, 데이터베이스의 네 가지 이점과 관련하여 어려움을 겪었던 경험을 모둠원들과 이야기해 보자.

해 보기 1 **지도 방법**

데이터베이스의 4가지 이점을 일상생활과 연결하며, 다소 어려울 수 있는 용어의 개념을 이해할 수 있도록 지도한다.

예시 답안

1. 편리한 버전 관리: 공유 문서에서 누군가가 실수 또는 고의로 내용을 수정해서 원본 데이터로 되돌리는 데 어려움을 겪었던 경험 또는 파일을 잘못 저장해서 기존 파일에 덮어써서 원본 내용을 복구하지 못했던 경험
2. 데이터 무결성 유지: 같은 설문을 여러 번 제출하거나 전화번호 항목에 글자를 넣는 등 오류를 입력할 수 있는 경우
3. 데이터 보안 강화: 다른 학생들이 내 성적이나 개인 정보를 확인해서 곤란했던 경험 또는 공용 데이터에서 누구나 데이터를 수정할 수 있게 해서 임의로 삭제되었던 경험
4. 체계적인 데이터 관리: 급식 데이터를 확인하고 싶은데 종이로 인쇄된 급식 메뉴가 최신 메뉴로 바뀌지 않아서 급식 메뉴를 알 수 없어 답답했던 경험

참고 동영상

컴퓨터는 어떻게 사진을 이해할까?

- **제목**: 페이페이 리, 어떻게 컴퓨터가 사진을 이해하게 되었는가?
- **영상 내용**: 어린이가 사진을 볼 때, '고양이', '책', '의자'와 같이 단순한 것은 식별할 수 있다. 이제 컴퓨터도 그런 것을 할 수 있다. 그 다음은 뭘까? 컴퓨터 비전(computer vision) 분야의 전문가, 페이페이 리는 컴퓨터를 가르치는 데 사용한 1천 5백만 장의 사진 데이터베이스 이야기와 함께 기술의 현재와 다가올 미래에 대한 통찰을 설명한다.

|주소| https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures?About=&subtitle=ko

4 NoSQL 데이터베이스

NoSQL 데이터베이스는 특정 데이터 모델을 위해 특별히 구축되었으며, 현대적 애플리케이션에 맞게 쉽게 확장할 수 있는 유연한 스키마(schema)*에 데이터를 저장한다.

NoSQL 데이터베이스는 개발의 용이성, 기능성 및 확장성을 널리 인정받고 있다. 이 페이지에는 NoSQL 데이터베이스를 보다 잘 이해하고 효과적으로 시작할 수 있도록 지원할 리소스(resource)가 포함되어 있다.

* **스키마**: 데이터베이스 내에서 데이터의 구조와 데이터 간의 관계를 설명하는 것으로, 데이터베이스가 데이터를 저장하는 방법을 지정하는 구조를 의미한다.

Q&A 묻고 답하기

Q NoSQL 데이터베이스의 장점은 무엇인가요?

A 현대적 애플리케이션은 NoSQL 데이터베이스로 해결할 수 있는 몇 가지 문제에 직면해 있다. 예를 들어 애플리케이션은 소셜 미디어, 스마트 센서, 서드 파티 데이터베이스(third-party database)*와 같은 다양한 소스의 대량의 데이터를 처리한다. 이 모든 이질적인 데이터는 관계형 모델에 잘 맞지 않다. 테이블 형식 구조를 적용하면 대규모 중복, 데이터 복제 및 성능 문제가 발생할 수 있다.

NoSQL 데이터베이스는 비관계형 데이터 모델에 대해 특정 용도로 구축되는 데이터베이스로서 현대적인 애플리케이션 구축을 위한 유연한 스키마를 갖추고 있다. 개발 용이성, 기능성 및 확장성으로 널리 인정받고 있다.

* **서드 파티 데이터베이스**: 외부 조직이나 개인이 소유하고 관리하는 데이터베이스를 뜻한다.

[출처] https://aws.amazon.com/ko/nosql/

5 테이블 이해하기

테이블(table)은 데이터베이스의 모든 데이터를 포함하는 데이터베이스 객체이다. 테이블에서 데이터는 스프레드시트와 유사한 행과 열 형식으로 논리적으로 구성된다. 각 행은 고유한 레코드(record)를 나타내고, 각 열은 레코드의 필드(field)를 나타낸다.

예를 들어, 학교의 학생 데이터가 포함된 테이블에는 각 학생에 대한 행과 학번, 이름, 생년월일과 같은 학생 정보를 나타내는 열이 포함될 수 있다.

[출처] https://learn.microsoft.com/en-us/sql/relational-databases/tables/tables?view=sql-server-ver16

지도 방법

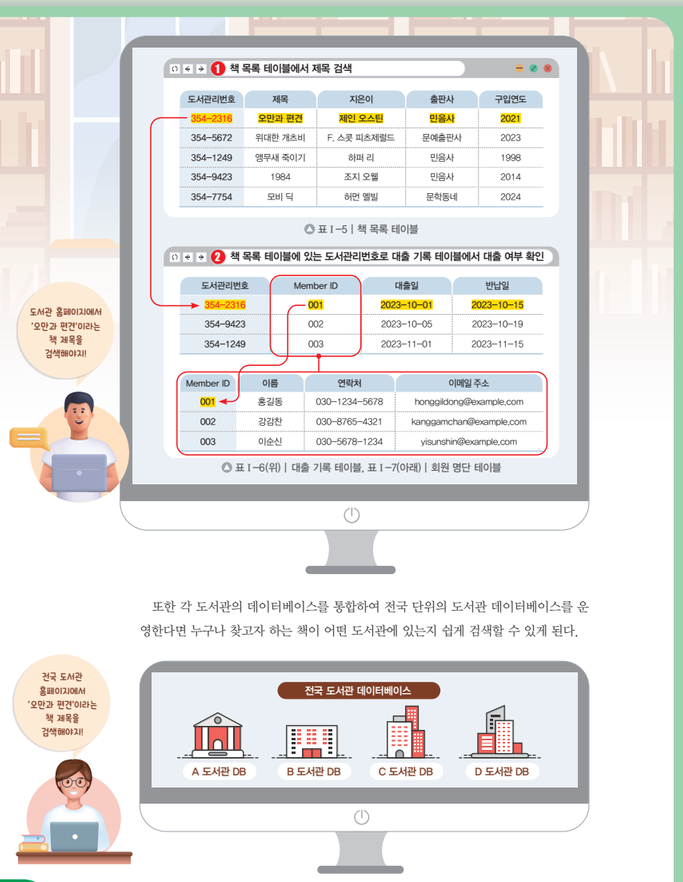
제시된 그림에서 제목과 회원 명단 사이에서 어떤 연결 고리가 있는지 흐름을 따라가면서 이해하도록 지도한다.

❶ 책 목록 테이블에서 사용자가 검색하기를 원하는 책 제목을 검색하고, 해당 책의 도서관 관리번호를 찾는다.

❷ 대출 기록 테이블에서 해당 도서관 관리번호의 책을 대출한 사람의 아이디를 찾는다.

❸ 회원 명단 테이블에서 대출한 사람의 아이디에 대한 연락처를 확인한다.

이를 통해 대출한 사람에게 연락을 하기 위해서는 서로 다른 테이블이 통합적으로 연결됨을 설명한다.



도서관리번호	제목	지은이	출판사	구입연도
354-2316	오만과 편견	제인 오스틴	민음사	2021
354-5672	위대한 개츠비	F. 스콧 피츠제럴드	문예출판사	2023
354-1249	앵무새 죽이기	하퍼 리	민음사	1998
354-9423	1984	조지 오웰	민음사	2014
354-7754	모비 딕	허먼 멜빌	문학동네	2024

표 Ⅰ-5 | 책 목록 테이블

도서관리번호	Member ID	대출일	반납일
354-2316	001	2023-10-01	2023-10-15
354-9423	002	2023-10-05	2023-10-19
354-1249	003	2023-11-01	2023-11-15

Member ID	이름	연락처	이메일 주소
001	홍길동	030-1234-5678	honggildong@example.com
002	강감찬	030-8765-4321	kanggamchan@example.com
003	이순신	030-5678-1234	yisunshin@example.com

표 Ⅰ-6(위) | 대출 기록 테이블, 표 Ⅰ-7(아래) | 회원 명단 테이블

또한 각 도서관의 데이터베이스를 통합하여 전국 단위의 도서관 데이터베이스를 운영한다면 누구나 찾고자 하는 책이 어떤 도서관에 있는지 쉽게 검색할 수 있게 된다.

4 | 데이터베이스의 생활 속 활용

데이터베이스는 기본적으로 추가 및 생성, 읽기, 갱신, 삭제와 같은 기능을 갖는다. 도서관의 사례를 통해 우리 생활에서 데이터베이스가 어떻게 활용되는지 간략히 살펴보자.

01 데이터 생성하기

데이터베이스에 새로운 데이터를 생성하거나 추가할 수 있다.

도서관에 새로운 책이 들어온 경우, 책 목록 테이블에 책에 대한 정보를 입력하면 데이터베이스에 새로운 책에 대한 데이터가 생성된다. 또한 새로운 테이블을 생성하거나 새로운 데이터베이스를 생성할 수도 있다.

도서관리번호	제목	지은이
354-2316	오만과 편견	제인 오스틴
354-5672	위대한 개츠비	F. 스콧 피츠제럴드
354-1249	앵무새 죽이기	하퍼 리
354-9423	1984	조지 오웰
354-7754	모비 딕	허먼 멜빌

새로운 책이 입고되면 해당 책에 대한 데이터를 생성한다.

표 Ⅰ-8 | 데이터 생성

02 데이터 읽기

데이터베이스에 저장된 데이터를 읽어올 수 있다.

도서관 데이터베이스에 접근하면 책의 제목을 기준으로 조회할 수도 있고, 지은이 이름, 도서관에 등록된 연도 등 다양한 기준으로 조회할 수도 있다.

용어 해설 CRUD

CRUD는 대부분의 컴퓨터 소프트웨어가 가지는 기본적인 데이터 처리 기능인 Create(생성), Read(읽기), Update(갱신), Delete(삭제)를 묶어서 일컫는 말이다. 사용자 인터페이스가 갖추어야 할 기능(정보의 참조/검색/갱신)을 가리키는 용어로도 사용된다.

6 전국 도서관 자료 통합 검색 서비스

전국 도서관 자료 통합 검색 서비스를 활용하여 데이터베이스의 통합이 주는 유익을 경험하도록 한다.

[출처] https://www.nl.go.kr/kolisnet/index.do

7 기본 키와 외래 키

기본 키(primary key)는 주 키 또는 프라이머리 키(primary key)라고 하며, 관계형 데이터베이스에서 레코드의 식별자로 이용하기에 가장 적합한 속성을 테이블마다 정해준 것을 말한다.

외래 키(foreign key)는 한 테이블에서 다른 테이블의 기본 키를 참조하는 속성을 말한다. 관계형 데이터베이스에서 테이블 간의 관계를 설정하는 데 사용되며, 외래 키는 참조된 테이블의 기본 키와 연결되어 두 테이블 간의 데이터 무결성을 유지하는 역할을 한다.

8 SQL 명령어 이해하기 1

❶ 데이터 생성 INSERT

구조화 질의어(SQL)에서, INSERT 문은 테이블에 한 개 이상의 행을 추가한다.

```
INSERT INTO 테이블 이름 (컬럼1, [컬럼2, ...]) values (값1, [값2, ...])
```

❷ 데이터 읽기 SELECT

SQL에서 SELECT 문은 하나 또는 그 이상의 테이블에서 데이터를 추출한다.

```
SELECT 컬럼1 [,컬럼2...] FROM 테이블 이름
```

9 스키마

스키마(schema)는 데이터베이스에서 데이터의 구조와 제약 조건을 정의하는 틀을 말한다. 테이블, 데이터 타입 등 데이터베이스 객체들의 논리적인 구성을 나타내며, 데이터베이스의 전체적인 설계와 관련된 정보가 포함된다. 스키마는 데이터베이스의 논리적 구조를 정의함으로써 데이터의 조직화와 관리를 용이하게 한다.

지도 방법

학생들이 도서관 데이터라는 맥락으로 데이터를 생성, 검색, 수정, 삭제하는 내용을 이해하도록 설명한다. 이를 바탕으로 학생들이 자주 사용하는 SNS와 같은 서비스에서 데이터가 어떻게 관리되는지 알려주면, 데이터베이스가 일상생활에 어떤 영향을 끼치는지 더 쉽게 이해할 수 있다.

03 데이터 갱신하기

데이터베이스에 저장된 데이터를 수정할 수도 있다.

만약 도서관에 책을 반납했을 경우 대출 기록 테이블의 데이터를 업데이트한다. 게시판의 경우 글의 내용을 수정할 때도 데이터 갱신이 이루어진다.

도서관리번호	Member ID	대출일	반납일
354-2316	[redacted]	[redacted]	
354-9423	002	2023-10-05	2023-10-19
354-1249	003	2023-11-01	2023-11-15

대출했던 도서를 반납하면 대출 현황 데이터베이스가 갱신된다.

▲ 표 I-9 | 데이터 갱신

04 데이터 삭제하기

데이터베이스에 저장된 데이터를 삭제할 수도 있다.

만약 도서관에서 오래된 책을 정리하여 버릴 때는 책 목록 테이블에서 해당 데이터를 삭제한다.

도서관리번호	제목	지은이	출판사	구입연도
354-2316	오만과 편견	제인 오스틴	민음사	2021
354-5672	위대한 개츠비	F. 스콧 피츠제럴드	문예출판사	2023
~~354-1249~~	~~앵무새 죽이기~~	~~하퍼 리~~	~~민음사~~	~~1998~~
354-9423	1984	조지 오웰	민음사	2014
354-7754	모비 딕	허먼 멜빌	문학동네	2024

오래된 도서를 처리하거나 도서가 분실된 경우 해당 데이터를 삭제한다.

▲ 표 I-10 | 데이터 삭제

소단원 1분 요약

❶ 하나의 주제에 대한 데이터들의 집합을 데이터셋이라고 하며, 대규모의 데이터셋들을 효율적으로 통합하고 관리하기 위해 데이터베이스를 사용한다.

❷ 생활 속에서 이용하는 블로그, SNS, 온라인 쇼핑몰 등 대규모 데이터를 여러 사람이 공유하는 서비스들은 데이터베이스를 활용하여 데이터를 추가, 조회, 수정, 삭제하는 등의 작업을 처리한다.

탐구 활동 인터넷 쇼핑몰 데이터베이스 설계하기

만약 인터넷 쇼핑몰을 만든다면, 쇼핑몰의 회원 데이터를 관리하기 위한 데이터베이스를 어떻게 설계하면 좋을지 모둠을 구성하여 논의해 보자.

1 [쇼핑몰 사용자 테이블]을 설계해 보자.

	속성명(영문)	설명	데이터형 (정수/실수/텍스트/이미지/영상 등)	조건
0	uid	회원들을 구분할 수 있는 사용자 아이디	텍스트	중복되면 안 됨. 8자 이상, 32자 이하
1	passwd	사용자 비밀번호	텍스트	숫자, 특수 문자, 영문자 조합 10글자 이상
2	name	사용자 이름	텍스트	6자 이상, 30자 이하
3	email	사용자 이메일 주소	텍스트	이메일 형식에 맞아야 함. 중복되면 안 됨.
4	point	쇼핑몰 포인트	정수	0 이상이어야 함.

2 [쇼핑몰 제품 테이블]을 설계해 보자.

	속성명(영문)	설명	데이터형 (정수/실수/텍스트/이미지/영상 등)	조건
0	pid	쇼핑몰에 등록된 물품에 대한 아이디	정수	중복되면 안 됨.
1	like	사람들이 '좋아요'를 누른 숫자	정수	0 이상이어야 함.
2	pname	상품명	텍스트	6자 이상, 100자 이하
3	price	상품 가격	정수	0 이상이어야 함.
4	image	제품 이미지	이미지	이미지 형식의 확장자

3 [장바구니 테이블]을 설계해 보자.

	속성명(영문)	설명	데이터형 (정수/실수/텍스트/이미지/영상 등)	조건
0	uid	회원들을 구분할 수 있는 사용자 아이디	텍스트	[쇼핑몰 사용자] 테이블에서 불러올 것
1	pid	쇼핑몰에 등록된 물품에 대한 아이디	정수	[쇼핑몰 제품] 테이블에서 불러올 것
2	history	이전에 구입한 횟수	정수	0 이상이어야 함.
3	quantity	장바구니에 담긴 물품의 수량	정수	0 이상이어야 함.
4	addedtime	장바구니에 담긴 날짜와 시간	시간	시간 형식이어야 함.

탐구 활동 지도 방법

- 학생들과 함께 인터넷 쇼핑몰 사이트를 살펴보며, 어떤 데이터들이 저장되어 있는지, 해당 데이터는 어떤 데이터형으로 되어 있는지, 어떤 제약 조건들이 있어야 하는지를 가볍게 알아본다.
- 이어서 [탐구 활동]을 실시한다.
- 위에 있는 2개의 테이블을 하나씩 살펴본다. 마지막 장바구니 테이블은 빈칸에 어떤 내용을 채우면 좋을지 생각해 보고, 학생들이 친구들과 함께 빈칸을 채우도록 지도한다.
- 이 과정에서 창의적인 아이디어를 제시한 학생이 있다면, 칭찬하여 다른 학생들이 또 다른 아이디어를 생각할 수 있는 마중물 역할을 하도록 한다.
- 특정 물품의 수량이 제한될 경우, 제품 테이블에서 물품 수량 속성을 추가하고, quantity 속성이 제품 테이블의 물품 수량 속성값보다 커서는 안 된다는 조건을 추가할 수 있다.

이와 같이 테이블 설계는 문제를 정의하는 사람에 따라 달라질 수 있음을 학생들이 이해하도록 안내한다.

10 SQL 명령어 이해하기 2

❶ 데이터 갱신 UPDATE

UPDATE 문은 구조화 질의어 중 하나로, 테이블에서 한 개 이상의 행을 바꾼다. 모든 행을 변경해야 되는 경우도 조건절을 사용하여 하위 집합을 선택할 수 있다.

```
UPDATE 테이블 이름 SET 컬럼1 = 값1 [, 컬럼2 = 값2 ...]
[WHERE 조건]
```

❷ 데이터 삭제 DELETE

구조화 질의어(SQL)에서, DELETE 문은 테이블에서 한 개 이상의 행을 삭제한다. 하위 집합은 삭제에 대한 조건을 정의할 수 있으며, 별도의 조건을 정의하지 않으면 모든 행이 삭제된다.

```
DELETE FROM 테이블 이름 [WHERE 조건]
```

11 쇼핑몰 데이터베이스에서 개인 정보를 보호하기 위한 조치

- 데이터 암호화 및 비밀번호 해싱: 고객의 민감한 정보를 저장할 때 암호화하고, 비밀번호는 해싱(hashing)*하여 보호한다. 이를 통해 데이터가 유출되더라도 안전하게 보호된다.

* **해싱:** 데이터를 고정된 길이의 해시 값으로 변환하는 과정을 말한다. 해시 함수는 입력된 데이터에 대해 동일한 해시 값을 반환하므로, 데이터의 무결성을 확인하거나 고유한 식별자를 생성하는 데 사용된다. 비밀번호 해싱은 사용자의 비밀번호를 해시 함수를 통해 변환하여 저장하는 보안 기법이다. 이렇게 저장된 해시 값은 원래의 비밀번호를 복원할 수 없으므로, 데이터베이스가 노출되더라도 해커가 비밀번호를 직접 알아낼 수 없다.

- 접근 권한 관리: 민감한 데이터에 접근할 수 있는 권한을 최소화하고, 권한을 엄격하게 관리하여 불필요한 접근을 방지한다.
- SQL 인젝션 방지 및 보안 프로토콜 사용: 데이터베이스 보안을 강화하기 위해 SQL 인젝션(injection)* 방지와 SSL/TLS 같은 보안 프로토콜(protocol)을 사용하여 안전한 데이터 전송과 처리 환경을 유지한다.

* **SQL 인젝션:** 데이터베이스 보안 취약점 중 하나로, 해커가 악의적인 SQL 코드를 입력하여 데이터베이스에 저장된 정보를 탈취하거나 수정, 삭제하는 공격 기법을 말한다. 인젝션은 주입이라는 뜻으로, SQL 명령어를 입력할 때 악의적인 코드를 함께 주입하여 데이터베이스를 조작하는 것을 의미한다.

12 벡터 데이터 베이스

인공지능(AI)은 급격한 발전을 거듭하며 우리의 삶에 많은 변화를 가져오고 있다. 특히 최근 딥러닝 기술의 비약적 성장으로 자연어 처리, 컴퓨터 비전, 음성 인식 등 다양한 분야에서 혁신적인 성과가 나타나고 있다. 이러한 AI 기술 발전의 중심에는 '벡터 데이터 베이스'라는 핵심 인프라가 자리 잡고 있다.

벡터 데이터 베이스는 비정형 데이터를 고차원의 밀집 벡터로 표현하고, 이를 효율적으로 저장하고 검색할 수 있게 해 준다. 텍스트나 이미지, 음성 등의 데이터를 의미론적으로 유사한 벡터 표현으로 변환해 벡터 공간상에서 가까운 위치에 배치하는 것이다. 이를 통해 의미 기반의 검색과 분석이 가능해진다.

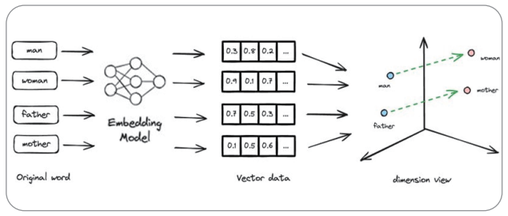
특히 최근 주목받고 있는 '생성형 AI' 분야에서 벡터 데이터 베이스는 필수 불가결한 역할을 한다. 대규모 언어 모델(LLM: Large Language Model)을 활용하는 생성형 AI 시스템에서는 RAG(Retrieval Augmented Generation) 아키텍처를 통해 관련 컨텍스트를 검색하고 답변을 생성하는데, 이 과정에서 벡터 데이터 베이스가 중추적인 역할을 수행하게 된다.

이처럼 벡터 데이터 베이스 기술은 AI 혁신의 근간이 되고 있다. 본 기고에서는 벡터 데이터 베이스의 원리와 구현 방식, 실제 활용 아키텍처들을 살펴보고, 이 분야의 최신 기술 동향과 전망에 대해서도 다뤄 보도록 한다.

☑ 벡터 데이터 베이스란?

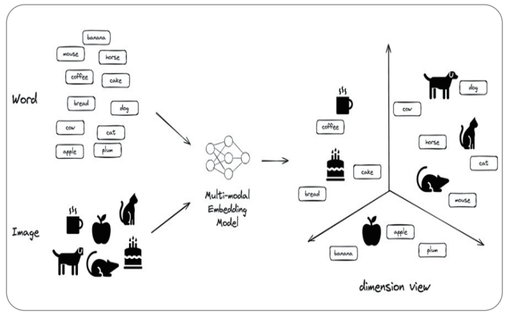
전통적인 데이터 베이스는 일반적으로 테이블 형태의 구조화된 데이터를 저장하고 관리한다. 이 데이터는 행과 열로 구성돼 있으며, 각 행은 특정 엔티티(예: 고객, 주문 등)를 나타내고 열은 해당 엔티티의 속성(예: 이름, 주소, 금액 등)을 나타낸다. 이러한 관계형 데이터 베이스는 데이터 무결성과 일관성을 유지하는 데 적합하다.

그러나 텍스트, 이미지, 오디오와 같은 비정형 데이터를 다루는 경우, 벡터 데이터 베이스가 더 효과적이다. 벡터 데이터 베이스에서는 텍스트를 고차원 벡터로 임베딩해 저장한다. 이를 위해 자연어 처리 기술을 활용한 임베딩 모델이 활용된다. 이러한 모델은 단어나 문장의 의미를 벡터 표현으로 변환해 벡터 공간상에 배치한다. 의미가 유사한 텍스트는 가까운 위치에, 상이한 텍스트는 먼 위치에 위치하게 된다. 예를 들면, 아래 그림에서와 같이 man은 woman이나 mother보다 father에 더 가까운 의미론적인 유사성을 가지고 있다. 이 의미론적 유사성은 차원 벡터 공간에서 벡터 거리가 가까운지를 계산해 판단한다.



▲ 텍스트 임베딩

벡터 데이터 베이스의 중요한 특징 중 하나는 멀티모달(multi-modal) 기능이다. 이를 통해 텍스트, 이미지, 오디오 등 다양한 유형의 데이터를 동일한 벡터 공간에 표현할 수 있다. 예를 들어, 이미지와 텍스트를 동시에 임베딩해 유사도를 비교할 수 있다. 이는 이미지 검색, 멀티모달 데이터 분석 등 다양한 응용 분야에 활용될 수 있다.



▲ 멀티모달 임베딩

요약하면 전통적 데이터 베이스는 구조화된 데이터를 효율적으로 관리하는 반면, 벡터 데이터 베이스는 비정형 데이터를 벡터 표현으로 변환해 의미 있는 패턴과 관계를 탐색할 수 있다. 특히 텍스트 임베딩과 멀티모달 기능은 벡터 데이터 베이스의 주요 강점이다.

☑ 벡터 데이터 베이스의 활용

벡터 데이터 베이스의 주요 활용 사례는 다음과 같다.

❶ **이미지 인식:** 이미지 임베딩을 사용한 유사 이미지 검색, 역방향 이미지 검색, 유사 제품 추천, 얼굴 인식 등

❷ **자연어 처리:** 단어, 문장, 문서 임베딩을 활용한 정보 검색, 문서 클러스터링, 텍스트 분류 등

❸ **추천 시스템:** 사용자 행동, 아이템 특징 임베딩을 사용한 개인화 추천, 크로스셀링(Cross-selling)*

❹ **이상 탐지:** 정상 행동 벡터 집합과 비교해 이상치 식별

❺ **생물 정보학:** 복잡한 생물 정보 데이터를 고차원 벡터로 표현, 패턴/유사성 발견

❻ **음성 인식:** 음성 샘플 벡터와 저장된 음성 프로필 비교

추가로 전자 상거래에서 개인화 추천과 크로스셀링, 콘텐츠 필터링, 지리 공간 분석, 네트워크 보안을 위한 이상 탐지, 지능형 검색 등에도 벡터 데이터 베이스가 활용된다.

*크로스셀링(Cross-selling): 한 상품을 구매한 고객에게 관련된 다른 상품을 추천하여 추가 구매를 유도하는 마케팅 전략이다. 추천 시스템에서 크로스셀링은 사용자가 구매한 상품이나 관심을 보인 상품과 연관된 다른 상품을 추천하여 추가적인 구매나 상호 작용을 유도하는 방식으로 사용된다.

☑ 생성형 AI 생태계에서의 벡터 데이터 베이스 활용

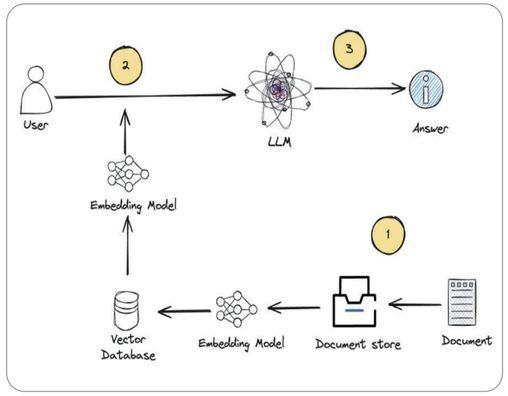
생성형 AI를 사용하는 환경에서는 벡터 데이터 베이스가 RAG 아키텍처에서 많이 활용된다. RAG의 아키텍처는 다음과 같은 특성을 갖고 있다.

❶ RAG는 방대한 양의 프라이빗 데이터 전체에서 검색하고, 최종 사용자가 질의한 내용과 가장 유사한 결과를 검색해 LLM에 컨텍스트로 전달한다.

❷ 검색된 관련 컨텍스트와 사용자 질의를 입력으로 받은 LLM은 이를 바탕으로 최종 답변을 생성한다. 단순 정보 검색이 아닌 LLM의 언어 이해/생성 능력을 활용해 높은 수준의 답변이 생성 가능하다.

❸ RAG는 단순 QA 외에도 요약, 분석, 태스크 완성 등 다양한 자연어 처리 과제에 응용 가능하다.

벡터 데이터 베이스는 RAG 아키텍처에서 중요한 역할을 한다. 이 아키텍처는 3단계로 구성돼 있다.



▲ RAG 아키텍처의 벡터 DB 활용(3단계 구성)

❶ 1단계 문서 임베딩

이 단계에서는 질문에 대한 답변으로 사용할 문서들을 LLM을 사용해 벡터로 임베딩한다. 각 문서는 고차원 벡터로 표현되며, 이 벡터들은 벡터 데이터 베이스에 저장된다.

❷ 2단계 질문 임베딩

사용자로부터 질문을 입력받으면, 동일한 LLM 모델을 사용해 질문 텍스트를 벡터로 임베딩한다. 이 질문 벡터는 1단계에서 저장된 문서 벡터들과 비교된다.

❸ 3단계 컨텍스트 검색 및 답변 생성

벡터 데이터 베이스에서는 질문 벡터와 가장 가까운(유사한) 문서 벡터들을 찾는다. 이를 위해 코사인 유사도나 유클리디안 거리 등의 유사도 측정 기법이 사용된다. 가장 관련성이 높은 문서들이 선택되면, 이 문서들을 컨텍스트로 활용해 LLM 모델이 최종 답변을 생성한다.

요약하면 RAG 아키텍처에서 벡터 데이터 베이스는 문서 임베딩, 질문 임베딩, 그리고 유사도 기반 컨텍스트 검색을 수행하는 핵심 역할을 한다. 이를 통해 대규모 문서 집합에서 사용자 질문에 관련된 컨텍스트를 효율적으로 찾아낼 수 있다.

☑ 다양한 유형의 벡터 데이터 베이스

벡터 DB는 크게 세 가지 유형으로 나눌 수 있다. 각각의 특징과 장단점을 이해하면 프로젝트 요구 사항에 맞는 최적의 솔루션을 선택할 수 있다.

❶ 벡터 라이브러리

벡터 라이브러리는 프로그래밍 언어에 내장해 사용할 수 있는 라이브러리나 오픈소스 라이브러리이다. 이러한 라이브러리는 간단한 벡터 연산을 수행하기에 용이하다. 하지만 대규모 데이터셋을 다루거나 복잡한 쿼리를 실행하기에는 한계가 있다. 벡터 라이브러리를 사용하는 솔루션은 다음과 같다.

(1) FAISS(Facebook AI Similarity Search): FAISS는 페이스북(Facebook)의 AI 팀에서 개발한 오픈소스 라이브러리로, 밀집 벡터의 효율적인 유사도 검색 및 클러스터링에 특화돼 있다. 대규모 벡터 검색 작업에 매우 적합하며 이미지 및 비디오 검색과 같은 AI 연구 분야에서 광범위하게 사용된다. FAISS는 고차원 데이터 처리에 탁월하지만 SQL이나 JSON과 같은 구조화된 데이터 유형을 직접 지원하지 않는다. 완전한 데이터 베이스라기보다는 주로 라이브러리 형태이며, 호스팅 또는 클라우드 서비스를 제공하지 않는다.

(2) ANNOY(Approximate Nearest Neighbors Oh Yeah): ANNOY 또한 오픈소스 프로젝트로, 고차원 공간에서 메모리 효율적이고 빠른 근사 최근접 이웃 검색을 위해 설계됐다. 스포티파이(Spotify)에서 개발했으며, 대략적인 결과로도 충분한 시나리오에서 일반적으로 사용된다. ANNOY는 라이브러리이며 데이터 베이스가 아니고 호스팅 서비스도 제공하지 않는다. 벡터 연산에 초점을 맞추고 있으며, SQL과 같은 구조화된 데이터 유형을 기본적으로 지원하지 않는다.

(3) SCANN(Scalable Nearest Neighbors): 구글 리서치(Google Research)에서 개발한 SCANN은 대규모 최근

접 이웃 검색에 특화된 오픈소스 라이브러리다. 고차원 공간에서 정확도와 효율성의 균형을 제공하며, 정밀한 벡터 검색 기능이 필요한 사용 사례를 위해 설계됐다. FAISS와 ANNOY와 마찬가지로, SCANN은 벡터 연산에 초점을 맞춘 라이브러리이며 구조화된 데이터 유형이나 호스팅 서비스를 기본적으로 지원하지 않는다.

(4) NMSLIB(Non-Metric Space Library): NMSLIB는 최근접 이웃 검색을 위한 효율적인 인덱싱 및 검색 알고리즘을 제공하는 범용 라이브러리다. 다양한 거리 함수를 지원하며 고차원 및 희소 데이터셋에서 좋은 성능을 보인다. 비트 벡터, 정수 벡터 및 부동 소수점 벡터를 모두 지원한다.

(5) 버치(Vearch, Vector Search): 버치는 벡타라(Vectara)에서 개발한 오픈소스 벡터 데이터 베이스다. 대규모 데이터에 대한 빠른 검색과 삽입을 지원하며 백업, 복제, 클러스터링 등의 기능을 제공한다. 파이썬(Python), 자바(Java) 및 고(Go) 클라이언트 라이브러리를 통해 간편하게 통합할 수 있다.

❷ 벡터 전용 데이터 베이스

벡터 전용 데이터 베이스는 고차원 벡터 데이터를 효율적으로 저장하고 검색하도록 설계됐다. 이러한 데이터 베이스는 근접 벡터 검색(Nearest Neighbor Search)에 최적화돼 있어 유사도 기반 검색이 가능하다. 또한 대규모 데이터셋을 처리할 수 있는 확장성을 갖추고 있다. 하지만 일반적인 관계형 데이터 베이스와 달리 벡터 데이터에 특화되어 있어 다른 유형의 데이터를 저장하기에는 적합하지 않다. 벡터 전용 데이터 베이스는 다음과 같다.

(1) 파인콘(Pinecone): 파인콘은 추천 시스템, AI 기반 검색과 같은 애플리케이션에서 확장 가능하고 높은 성능의 유사도 검색을 위해 설계된 벡터 데이터 베이스 서비스다. 완전 관리형 클라우드 서비스로, 벡터 검색 시스템의 배포와 확장을 단순화한다. 주로 벡터 데이터에 중점을 두고 있지만 다른 데이터 유형 및 시스템과의 통합을 지원할 수 있다.

(2) 위비에이트(Weaviate): 위비에이트는 확장 가능한 의미 검색을 위해 설계된 오픈소스 그래프 기반 벡터 데이터베이스다. 비정형 데이터를 포함한 다양한 데이터 유형을 지원하며 기계학습 모델과 통합해 데이터의 자동 벡터화가 가능하다. 위비에이트는 클라우드 및 자체 호스팅 배포 옵션을 모두 제공하며, 그래프 데이터 베이스 기능과 벡터 검색이 필요한 애플리케이션에 적합하다.

(3) 밀버스(Milvus): 밀버스는 대규모 고차원 벡터 데이터 처리에 최적화된 오픈소스 벡터 데이터 베이스다. 다양한 인덱스 유형과 메트릭을 지원해 효율적인 벡터 검색이 가능하며 다양한 데이터 유형과 통합할 수 있다. 밀버스는 온프레미스 또는 클라우드에서 배포할 수 있어 다양한 운영 환경에서 활용 가능하다.

(4) 크로마DB(ChromaDB): 크로마DB는 임베딩 벡터 데이터의 효율적인 저장 및 검색을 위한 오픈소스 벡터 데이터베이스다. 텍스트, 이미지, 오디오 등의 데이터를 벡터로 변환해 저장하고, 최근접 이웃 검색 알고리즘을 사용해 빠른 속도로 유사한 벡터를 찾을 수 있다. 크로마DB는 클러스터링을 지원해 확장성을 제공하며, 파이썬과 통합된 단순한 API를 갖추고 있다. 프로토타이핑이나 소규모 애플리케이션에서 높은 처리량과 낮은 대기 시간이 필요한 경우 크로마DB를 효과적으로 활용할 수 있다.

(5) 쿼드란트(Qdrant): 쿼드란트는 고차원 벡터 데이터를 지원하는 오픈소스 벡터 검색 엔진이다. 벡터 데이터의 효율적인 저장 및 검색을 위해 설계됐으며, 필터링과 전체 텍스트 검색 등의 기능을 제공한다. 쿼드란트는 클라우드 또는 온프레미스 환경에서 사용할 수 있어 효율적인 벡터 검색 기능이 필요한 다양한 애플리케이션에 활용 가능하다.

(6) 베스파(Vespa): 야후(Yahoo)에서 개발한 오픈소스 빅데이터 서빙 엔진인 베스파는 대규모 데이터셋의 저장, 검색 및 구성 기능을 제공한다. 정형 및 비정형 데이터를 포함한 다양한 데이터 유형을 지원하며, 실시간 계산 및 데이터 서빙이 필요한 애플리케이션에 적합하다. 베스파는 클라우드 및 자체 호스팅 환경에서 배포할 수 있다.

❸ 벡터도 지원하는 엔터프라이즈 데이터 베이스

일부 엔터프라이즈급 데이터 베이스는 벡터 데이터 타입을 지원한다. 이러한 데이터 베이스는 기존 데이터 타입의 데이터와 벡터 데이터를 모두 저장할 수 있어 통합 데이터 관리가 가능하다. 또한 기존 데이터 베이스 인프라와 연계할 수 있어 편리하다. 엔터프라이즈 데이터 베이스 환경에서 최적화돼 있는 벡터 데이터 베이스는 다음과 같다.

(1) 오픈서치(OpenSearch)/엘라스틱서치(Elasticsearch): 오픈서치와 엘라스틱서치는 강력한 전체 텍스트 검색 기능으로 널리 알려진 오픈소스 검색 및 분석 엔진이다. JSON 문서를 비롯한 다양한 데이터 유형을 지원하며 확장 가능한 검색 솔루션을 제공한다. 또한 클라우드 또는 온프레미스에 배포할 수 있으며, 벡터 검색 기능을 추가해 다양한 검색 및 분석 애플리케이션에 적합하게 설계됐다. AWS와 같은 클라우드 서비스에서는 관리형 서비스로 오픈서치를 지원하고 있다. 아마존 오픈서치 서비스(Amazon OpenSearch Service)는 오픈서치를 관리형 서비스로 제공하고 있고, 서버리스(severless) 형태의 관리형 서비스도 제공하고 있다.

(2) 몽고DB(Mongo, MongoDB): 몽고DB는 유연성과 사용 편의성으로 잘 알려진 인기 있는 오픈소스 문서 기반 데이터 베이스다. 주로 JSON 형식의 문서를 지원하며 다양한 데이터 유형을 처리할 수 있다. 몽고DB는 클라우드 기반 서비스인 몽고DB 아틀라스(MongoDB Atlas)와 온프레미스 배포 옵션을 모두 제공한다. 전통적으로 문서 저장에 중점을 뒀지만, 벡터 데이터 처리를 위한 기능을 점차 추가하고 있다.

(3) 레디스(Redis): 레디스는 오픈소스 인메모리 데이터 구조 스토어로, 데이터 베이스, 캐시, 메시지 브로커로 사용된다. 문자열, 해시, 리스트, 집합 등 다양한 데이터 유형을 지원한다. 레디스는 속도가 빠르기로 유명하며 캐싱, 세션 관리, 실시간 애플리케이션 등에 일반적으로 사용된다. 클라우드 호스팅 및 자체 호스팅 배포 옵션을 모두 제공한다.

(4) 포스트그레SQL(Postgres, PostgreSQL): 포스트그레SQL은 신뢰성, 기능 견고성, 성능으로 잘 알려진 강력한 오픈소스 객체-관계형 데이터 베이스 시스템이다. 구조화된 SQL 데이터와 JSON을 포함한 다양한 데이터 유형을 지원한다. 포스트그레SQL은 온프레미스나 클라우드에서 배포할 수 있으며, 소규모 프로젝트부터 대규모 엔터프라이즈 시스템까지 다양한 애플리케이션에서 널리 사용된다. pgvector 확장 기능을 사용하면 포스트그레SQL을 벡터 데이터 베이스로 활용할 수 있다. AWS와 같은 클라우드 환경에서도 사용이 가능하다. AWS에서는 아마존 오로라 포스트그레SQL(Amazon Aurora PostgreSQL)을 통해서 pgvector를 지원하고 있다.

(5) 아마존 켄드라(Amazon Kendra): 아마존 켄드라는 기업 환경에서 지식 검색을 위한 인텔리전트 검색 서비스다. 이 서비스는 내부적으로 벡터 데이터 베이스를 활용해 문서나 콘텐츠를 벡터로 표현하고, 자연어 쿼리에 대한 정확한 검색 결과를 제공한다. 기계학습 모델을 활용하여 텍스트 데이터의 의미를 파악하고, 관련성이 높은 결과를 반환한다. 켄드라는 다양한 데이터 소스로부터 정보를 인덱싱하고 통합할 수 있다.

(6) 날리지 베이스 포 아마존 베드록(Knowledge Bases for Amazon Bedrock, Bedrock Knowledge Base): 날리지 베이스 포 아마존 베드록은 대규모 기업 지식 베이스 구축 및 관리를 목적으로 특화된 종합 솔루션으로, 단순한 벡터 데이터 베이스 이상의 통합적인 기능을 제공한다. 특히 최신 자연어 처리(NLP: Natural Language Processing) 기술과 임베딩 모델을 내장하고 있어 텍스트 데이터를 의미 있는 벡터 표현으로 변환할 수 있고, 강력한 벡터 데이터 베이스와 벡터 유사도 검색 기능을 탑재하고 있다. 대규모 데이터에 대해서도 실시간으로 관련성이 높은 결과를 제공할 수 있다. AWS 클라우드 네이티브 서비스라는 장점으로 확장성, 가용성, 보안, 규정 준수 등의 요구 사항을 모두 충족한다. 온프레미스 환경에서는 구현하기 어려운 수준의 기능과 성능을 제공한다.

각 유형의 벡터 DB는 고유한 장단점을 갖고 있다. 프로젝트의 규모, 데이터 유형, 성능 요구사항 등을 종합적으로 고려해 가장 적합한 솔루션을 선택하는 것이 중요하다. 또한 하이브리드 아키텍처를 활용해 서로 다른 유형의 벡터 DB를 함께 사용하는 것도 가능하다.

☑ 벡터 데이터 베이스 구성 시, 고려 사항

벡터 데이터 베이스 안정적으로 구축하고 운영하려면 다음과 같은 사항들이 고려돼야 한다.

❶ 데이터 전처리 및 임베딩: 벡터 데이터 베이스에 데이터를 효과적으로 저장하고 검색하기 위해는 데이터 전처리와 임베딩 과정이 중요하다. 텍스트 데이터의 경우 적절한 토크나이저(Tokenizer)*를 선택하고 정제 과정을 거쳐야 한다. 또한 어떤 언어 모델을 사용해 임베딩할지, 미세 조정(fine-tuning)이 필요한지 등을 결정해야 한다.

❷ 성능 최적화: 벡터 검색의 성능은 응답 시간, 처리량 등 애플리케이션의 핵심 요구 사항과 직결된다. 인덱싱 전략 외에도 벡터 차원 축소, 캐싱, 분산 아키텍처 적용 등 다양한 기법을 통해 성능을 최적화할 수 있다.

❸ **모니터링 및 운영:** 실제 운영 환경에서는 벡터 데이터 베이스의 상태를 지속적으로 모니터링하고 관리해야 한다. 데이터 증가, 쿼리 패턴 변화 등에 따라 인덱싱을 재구성하거나 클러스터를 확장해야 할 수 있다. 또한 백업, 장애 대응 등 운영 관리 절차도 마련돼야 한다.

❹ **보안 및 프라이버시:** 벡터 데이터 베이스에는 민감한 데이터가 포함될 수 있으므로 보안과 프라이버시 보호 대책도 필수적이다. 데이터 암호화, 접근 제어, 암호화된 벡터 검색 등의 기술을 활용할 수 있다.

❺ **통합 및 상호 운용성:** 벡터 데이터 베이스를 기존 데이터 인프라와 통합하거나 다른 시스템과 연계해야 하는 경우가 많다. API, 데이터 파이프라인, 데이터 포맷 등의 측면에서 상호 운용성을 고려해야 한다.

이러한 요소들을 종합적으로 검토해 프로젝트 요구 사항에 가장 적합한 벡터 데이터 베이스 아키텍처를 설계하고 구축해야 한다. 운영 단계에서도 지속적인 모니터링과 최적화가 필요할 것이다.

* **토크나이저(Tokenizer):** 텍스트 데이터를 처리하고 분석하기 위해 텍스트를 기본적인 단위로 분할하는 도구를 말한다. 텍스트 데이터를 토큰으로 변환하는 과정을 토크나이제이션(tokenization)이라고 한다. 토큰은 일반적으로 단어, 구두점, 숫자 등을 포함할 수 있다.

☑ 결론

벡터 데이터 베이스는 AI 기술의 발전과 함께 그 중요성이 더욱 부각될 것으로 전망된다. 인공지능이 사람의 인지 능력을 모방하고 넘어서려면 데이터의 의미를 정확히 이해하고 처리할 수 있어야 한다. 벡터 데이터 베이스는 이러한 의미 기반 데이터 처리를 가능케 하는 핵심 인프라로 자리매김할 것이다.

머신러닝 모델이 고도화되고 멀티모달 AI의 활용이 확대되면서 벡터 데이터 베이스 기술도 지속적으로 진화할 것이다. 텍스트, 이미지, 음성, 센서 데이터 등 다양한 유형의 데이터를 벡터 표현으로 통합하고, 이를 기반으로 새로운 지식과 통찰력을 창출하는 것이 가능해질 것이다.

또한 벡터 데이터 베이스는 대규모 데이터 처리를 위한 분산 컴퓨팅, 메모리 및 스토리지 최적화 등의 기술과 긴밀히 연계돼 성능과 확장성을 높여갈 것으로 예상된다. 클라우드 네이티브 아키텍처와의 융합을 통해 벡터 데이터 베이스는 더욱 손쉬운 배포와 운영이 가능해질 것이다.

AI 기술의 급진전에 따라 벡터 데이터 베이스는 혁신적인 애플리케이션과 서비스를 견인하는 필수적인 데이터 인프라로 자리 잡게 될 것이다. 다양한 산업 분야에서 새로운 가치 창출과 AI 기반 의사 결정의 토대가 될 벡터 데이터 베이스 기술에 대한 연구와 투자가 지속될 것으로 기대된다.

[출처] http://www.itdaily.kr/news/articleView.html?idxno=225821

MEMO

4 세상을 바꾸는 데이터 과학

수업 시간: 2시간 32~41쪽

단원명	Ⅰ. 데이터 과학의 이해 04. 세상을 바꾸는 데이터 과학
학습 목표	• 데이터로 인한 사회 변화를 인식하고 설명할 수 있다. • 나의 진로 및 관심 분야와 관련된 데이터 기반 문제 해결 사례를 분석할 수 있다.
수업 방법	강의, 토론, 발표
준비물	교사 교과서, 관련 교수 학습 자료, 컴퓨터 학생 필기도구

단계	교수 · 학습 방법	지도상의 유의점
도입	생각 열기 만화를 통해 데이터가 우리 생활과 얼마나 밀접한 관련이 있으며, 지금까지 해결되지 않은 문제 중 데이터로 어떤 문제를 해결할 수 있을지 함께 생각해 보도록 한다.	• 와이파이가 안 되는 상황에서 답답한 이유는 데이터가 전송되지 않기 때문이라는 사실 등 데이터가 얼마나 생활 속 깊이 자리 잡고 있는지 깨닫도록 안내한다.
전개	1. 데이터의 역사 데이터는 디지털 시대 이전부터 중요하였으나 디지털 시대로 접어들며 더욱 중요해지고 있음을 안내하고, 의료, 제조, 교육, 경영, 스포츠 등 다양한 분야에서 데이터 과학이 접목되고 있음을 안내한다.	• 자신의 관심 분야에서 데이터 과학이 접목된 사례를 찾아보고, 함께 이야기를 나누도록 한다.
	2. 지속 가능한 미래를 만드는 데이터 과학 UN의 지속 가능 발전 목표를 소개하고, 이를 위해 데이터 과학이 어떻게 기여하고 있는지 알아본다. 이외에도 어떤 문제를 해결하는 데 활용될 수 있을지 조사하고, 발표하도록 한다.	• 지진, 산불, 감염병 예방 등 실생활과 밀접한 분야뿐 아니라 전문적인 영역에서의 사례도 자신의 진로 및 관심 분야와 연계하여 찾아보도록 지도한다.
	3. 데이터 과학 속 데이터 윤리 데이터 과학의 발전으로 인해 발생하는 윤리적 문제들을 안내하고, 데이터 과학으로 인한 윤리적 이슈에 대해 비판적으로 생각할 수 있는 관점을 갖도록 지도한다.	• 앞으로 점점 더 데이터가 중요해짐에 따라 데이터 윤리 문제에 대해 반드시 중요하게 다뤄야 함을 강조하도록 한다.
정리	탐구 활동 지금까지 학습한 내용을 바탕으로 데이터 과학과 자신의 진로에 대해 생각하고 함께 공유하도록 안내한다.	• 지식 충전소의 '데이터 과학과 사회 문제' 읽기 자료를 통해 데이터 과학으로 발생하는 사회 문제를 고민하고 해결하는 것 또한 중요하다는 사실을 깨닫도록 지도한다.
평가	• 데이터로 인해 사회가 어떻게 변화하고 있는지 설명할 수 있는가? • 자신의 진로 및 관심 분야의 데이터 기반한 문제 해결 사례를 소개할 수 있는가?	• 스포츠나 엔터테인먼트 등 학생들이 관심을 가질 만한 영역을 통해 동기를 유발하되, 결론적으로 자신의 진로와 연계지을 수 있도록 지도한다.

4 세상을 바꾸는 데이터 과학

제시 의도

데이터 과학에서는 데이터를 분석의 대상으로 인식하지만, 실제로 데이터는 우리 일상생활과 뗄 수 없는 것이 되었다. 특히 데이터로 인해 문제가 해결된 부분도 있지만, 데이터로 인해 새로운 문제가 발생하는 등 데이터에는 긍정적 측면과 부정적 측면이 모두 있음을 이해할 수 있는 사례를 제시하였다. 와이파이 연결이 안 되는 것이 문제가 되는 이유도 데이터를 주고받을 수 없기 때문이고, 추천 서비스가 내가 원하는 것을 잘 추천해 주는 것도 데이터에 기반했기 때문이다. 또한 미래의 도시를 설계하는 데 있어서도 가장 우선적으로 고려해야 할 점들이 데이터를 어떻게 활용할 것인가에 대한 문제이고, 데이터 윤리 문제 역시 데이터를 잘못 수집하거나 잘못 활용하기 때문에 발생하는 문제이다. 이를 통해 이미 우리는 데이터 없이는 살 수 없으며 데이터를 잘 다루는 것이 매우 중요하다는 것을 인식시키도록 한다.

예시 답안

- **자연재해 예측 및 대응:** 지진, 홍수, 허리케인 등의 자연재해 발생 가능성을 예측하고, 빠른 대응 체계 구축
- **교통 혼잡 해결:** 실시간 교통 데이터 분석을 통해 교통 혼잡 문제 해결 및 교통 시스템 최적화
- **농산물 수확량 예측:** 기후 데이터와 작물 데이터를 결합하여 수확량 예측 및 농업 계획 수립

4 세상을 바꾸는 데이터 과학

학습 목표
- 데이터로 인한 사회 변화를 인식하고 설명할 수 있다.
- 나의 진로 및 관심 분야와 관련된 데이터 기반 문제 해결 사례를 분석할 수 있다.

학습 요소
- 다양한 분야 속 데이터 과학, 미래 사회와 데이터 과학

생각 열기 데이터와 우리 생활

우리는 이미 데이터와 떼려야 뗄 수 없는 데이터 시대를 살고 있다. 데이터를 활용하여 많은 문제를 해결할 수 있지만, 아직 해결되지 않은 문제도 여전히 남아 있다. 다음 내용을 보고 질문에 답해 보자.

Q 지금까지 해결되지 않은 문제 중에서 앞으로 데이터로 해결할 수 있는 문제에는 어떤 것들이 있을까?

1 | 데이터의 역사

데이터의 역사는 날짜를 기록하기 위해 벽이나 막대기에 줄을 긋는 행위에서부터 시작되었다. 그 이후 글자와 숫자가 발명되면서 본격적으로 데이터가 기록되기 시작했다. 기원전 3,000년경 고대 메소포타미아 문명에서 만들어진 쐐기(설형) 문자는 상업 분야의 거래 내역을 기록하는 데 사용되었고, 이것이 오늘날 금융 데이터의 시초라고 할 수 있다. 최근에는 금융과 기술의 합성어인 '핀테크(fintech)'의 등장으로 다양한 금융 서비스에 혁신을 이루어 내고 있다.


최초 문자인 쐐기 문자가 새겨진 점토판(맥아와 보릿가루의 수령 내역을 적은 장부)

금융이란 의미의 'financial'과 기술이란 의미의 'technology'의 합성어

인구 데이터 역시 데이터의 역사와 함께 해 온 중요한 데이터다. 현재까지 알려진 가장 오래된 인구 데이터는 기원전 3,000년경 이집트에서 만들어졌다. 우리나라에서는 삼국 시대, 고려 시대, 조선 시대에 이르기까지 호구 조사란 명칭으로 인구 조사가 실시되었다. 최근에는 5년마다 전 국민의 일부를 대상으로 인구주택총조사가 이루어지는데, 이렇게 수집된 데이터는 국가 정책 수립에 중요하게 사용되며 누구나 다양하게 활용할 수 있도록 개방하고 있다.


조선 시대 전국의 호수와 인구수를 기록한 책

5년마다 전 국민의 일부를 대상으로 이루어지는 인구주택총조사

통계학(statistics)의 어원은 '국가'라는 의미가 담긴 이탈리아어 'statista'에서 유래되었다.

AI 시대, 데이터 과학의 지속적인 중요성

AI 시대에도 데이터 과학은 여전히 매우 중요하고 유효한 분야이다. 다음과 같은 이유로 데이터 과학의 역할이 더욱 중요해지고 있다.

1. 데이터 품질 관리와 전처리: AI 모델의 성능은 학습 데이터의 품질에 크게 좌우된다. 데이터 과학자들은 고품질 데이터를 준비하고 전처리하는 핵심 역할을 담당한다.
2. 인사이트(insight) 도출: AI가 대량의 데이터를 생성할 수 있지만, 그 데이터에서 의미 있는 인사이트를 추출하고 해석하는 것은 여전히 데이터 과학자의 몫이다.
3. 윤리적 고려 사항: AI 모델의 편향성과 공정성 문제를 해결하고 책임 있는 AI 구현을 위해서는 데이터 과학자의 전문성이 필요하다.

생각 열기 지도 방법

- 각각의 문제에 데이터가 어떤 관련이 있을지 학생들에게 질문하고, 학생들이 대답하기 어려운 경우에는 교사가 도움을 주도록 한다.
- 지금까지 데이터로 인해 많은 문제가 해결되었는데, 지금까지 해결되지 못한 문제 중에서 앞으로 데이터로 해결 가능한 문제에는 어떤 것들이 있을지 생각해 보도록 한다.
- 현대에는 데이터를 디지털 데이터에 한정하지만, 컴퓨터가 개발되기 전에도 다양한 데이터가 존재했고, 이러한 데이터 중 금융, 인구 데이터와 같은 데이터들은 인류 문명이 발전하는 데 매우 큰 영향을 끼쳤음을 안내한다.
- 과거의 데이터는 정확도가 떨어지고 관리가 어려운 등 여러 문제가 있었지만, 현대의 디지털 데이터를 활용하기 시작하면서 과거보다 훨씬 더 폭넓은 분야의 데이터가 수집되어 많은 문제들을 해결할 수 있게 되었음을 안내한다.

1 생명 과학과 데이터

데이터는 생명 과학 분야의 연구 방식을 근본적으로 변화시키고 있다. 대규모 데이터 분석을 통해 기존에 발견하기 어려웠던 패턴과 상관관계를 파악할 수 있게 되었다. 이는 유전체 분석 등 다양한 분야에서 새로운 통찰력을 제공하고 있다.

특히 데이터 기반 인공지능(AI)의 결합은 신약 개발 프로세스를 크게 가속화하고 있다. 방대한 양의 생물학적 데이터와 화합물 정보를 분석하여 잠재적인 약물 후보를 더 빠르고 효율적으로 식별할 수 있게 되었으며, 이는 개발 비용 절감과 시간 단축으로 이어져 제약 산업에 큰 변화를 가져오고 있다.

또한, 개인의 유전체 정보와 임상 데이터를 통합 분석함으로써, 개인 맞춤형 치료법 개발이 가능해졌다. 이는 환자 개인의 특성에 맞는 최적의 치료법을 제공하는 정밀 의학의 발전으로 이어지고 있으며, 빅데이터 분석을 통해 질병의 발생 패턴과 위험 요인을 더 정확히 파악할 수 있게 되었다. 이는 질병의 조기 진단과 예방에 큰 도움이 되고 있으며, 공중보건 정책 수립에도 중요한 역할을 하고 있다.

2 민주주의를 위한 데이터의 역할

1. **정책 결정의 투명성과 책임성 향상:** 데이터는 정책 결정 과정을 투명하게 하고 정부의 책임성을 강화한다.
2. **빈곤 감소와 경제 발전 지원:** 데이터는 빈곤 감소 전략과 경제 발전 정책 수립에 핵심적인 역할을 한다.
3. **자원 배분 최적화:** 데이터는 자원이 가장 필요한 곳을 파악하고 효율적인 배분을 가능하게 한다.
4. **정책 효과 모니터링:** 데이터는 다양한 정책의 진행 상황을 추적하고 그 영향을 평가하는 데 사용된다.

[참고 자료] The Role of Statistics in Promoting Good Governance in Nigeria's Democracy

3 생성형 AI 시대, 데이터 사이언티스트는 무슨 일을 할까?

생성형 AI는 의심할 여지 없이 툴, 프로세스, 결과물을 포함한 데이터 사이언티스트와 분석가의 업무 수행 방식에 변화를 일으키고 있다. 이에 대비하기 위해 데이터 사이언티스트는 지금 무엇을 해야 하는지 살펴본다.

최근까지 데이터 사이언티스트와 분석가의 주된 작업 결과물은 데이터 시각화, 머신러닝(Machine Learning) 모델, 대시보드, 보고서, 스토리텔링에 사용되는 분석 인사이트였다. 앞으로 데이터 사이언티스트는 생성형 AI의 기능을 활용해서 비정형 데이터 소스까지 포함하도록 분석 범위를 확장하고, 비즈니스팀이 데이터 기반 의사 결정으로 전환하도

지도 방법

- 다양한 분야 속 데이터 과학이 적용되는 사례를 소개한다. 데이터 과학 분야가 워낙 빠르게 변하는 분야이기 때문에 최신의 사례를 소개하는 기사나 영상을 최대한 활용하도록 한다.
- 교과서에 제시된 4가지 사례를 살펴보며, 데이터 과학이 우리의 삶에 얼마나 큰 영향을 끼치는지 알도록 한다. 또한 제시된 분야 외에도 자신의 관심 분야에서 어떻게 접목된 사례가 있는지 찾아보도록 한다.

예시 답안

❶ 경기력 분석 및 전략 최적화

- 축구: 축구에서는 빅데이터를 통해 경기 전략을 분석하고 최적화한다. 예를 들어, 선수들의 움직임, 패스 패턴, 슛 정확도 등을 분석하여 팀의 전술을 개선한다. 레스터 시티((Leicester City: 잉글랜드 레스터에 위치한 프로 축구 클럽)는 GPS가 장착된 웨어러블 기기(wearable device)와 다양한 각도에서 촬영된 경기 영상을 활용하여 선수들의 뛴 거리, 순간 속도 등을 분석하고 있다.
- 야구: 오클랜드 애슬레틱스(Oakland Athletics: 미국 메이저리그 야구 팀으로, 캘리포니아주 오클랜드를 연고지로 함)는 2002년부터 데이터 분석을 활용하여 저비용으로도 경쟁력 있는 팀을 구성하는 데 성공했다. 이를 통해 선수의 성과를 예측하고 최적의 전략을 수립하는 '머니볼' 접근법을 도입했다.

❷ 부상 방지 및 선수 관리

- 생체 센서: GPS 추적과 생체 센서를 활용하여 선수들의 움직임과 생리적 반응을 실시간으로 모니터링한다. 이를 통해 부상을 예방하고, 훈련 및 경기 중 성능을 최적화한다.
- 웨어러블 기기: 웨어러블 기기를 통해 선수들의 건강 상태를 모니터링하고, 뇌진탕 탐지 및 치료와 같은 부상 예방 조치를 취할 수 있다.

❸ 팬 경험 향상 및 상업적 활용

- 팬의 욕구 충족: 경기 분석을 통해 팬들이 원하는 정보를 제공하고, 팬 경험을 향상시킨다. 예를 들어, 실시간 경기 데이터와 통계를 제공하여 팬들이 경기를 더 깊이 이해할 수 있도록 돕는다.
- 소셜 빅데이터: 소셜 미디어 데이터를 분석하여 팬들의 관심사와 주요 이슈를 파악한다. 이를 통해 스포츠 이벤트의 마케팅 전략을 최적화하고, 팬들과의 소통을 강화한다.

록 돕고, AI 윤리와 거버넌스(Governance)*에 대해 자문하고, 갈수록 늘어나는 시민 데이터 사이언티스트를 위한 가드레일(Guardrail)을 마련해야 한다.

인포시스 코발트(Infosys Cobalt)의 부사장 아난트 아디야는 생성형 AI가 인사이트 도출 시간을 가속하고 기술 장벽을 낮추고 데이터 기반 의사 결정을 위해 역량을 확장할 수 있게 해 준다고 한다. 사람의 전문 지식은 여전히 중요하지만, 생성형 AI는 강력한 전력 승수(勝數) 역할을 하면서 인간의 역량을 강화하고 새로운 데이터 혁신 기회를 창출한다.

파운드리가 오픈텍스트(OpenText) 의뢰로 최근 실시한 AI와 분석에 대한 설문에서 응답자 75%는 데이터 시각화와 보고에 생성형 AI를 사용하는 것이 중요하다고 답했다. 그러나 데이터 아키텍처 및 분석 직무를 담당하는 응답자 중에서 생성형 AI가 '매우 중요하다'고 답한 비율은 27%에 그쳤다. AI에는 많은 비즈니스 기대가 따르고, 경영진은 데이터 사이언티스트와 분석가가 경쟁 우위를 확보하는 데 필요한 지식과 기술을 얻길 바란다. 데이터 사이언스팀은 현재의 목표를 점검하고 생성형 AI 활용을 위한 전략을 논의해야 한다.

젠팩트(Genpact)의 DTAI 부문 글로벌 책임자인 스리칸스 메논은 분석, 데이터 시각화, 머신러닝은 생성형 AI가 가진 역량을 통해 빠르게 발전하면서 더 직관적인 데이터 상호 작용, 자동화된 인사이트, 정교한 예측 모델을 실현하고 있다고 설명한다. 이와 같은 기술이 발전함에 따라 생성형 AI는 더 정확한 시각화를 생성하고 자연어 처리를 통해 복잡한 데이터 해석을 간소화하고 분석 보고서를 자동으로 생성함으로써 이런 분야를 강화한다.

AI 혁신과 새로운 비즈니스 동인에 대응해서 데이터 거버넌스와 소프트웨어 개발, 로우코드(Low-code)* 개발, 데브섹옵스(DevSecOps)*가 어떻게 발전하고 있을까? 여기서는 데이터 사이언티스트와 분석가의 역할과 책임, 그리고 이들이 사용하는 툴과 프로세스의 발전 동향에 기반해 데이터 사이언티스트와 분석가가 앞으로 갖춰야 할 5가지 역량을 알아본다.

❶ 매출과 성장을 위한 분석

데이터 사이언티스트는 마케팅팀을 위한 리드 생성, 영업팀을 위한 파이프라인 최적화, 재무팀을 위한 수익성 분석, HR팀을 위한 스킬 개발 등 항상 자신의 기술을 적용할 사용 사례 포트폴리오를 추구해 왔다. 생산성을 개선할 부분을 찾는 것도 중요하지만, 생성형 AI가 등장한 이후 기업이 AI를 활용해서 새로운 디지털 트랜스포메이션 기회를 추구함에 따라 데이터 사이언티스트는 매출 성장 영역에서 자신의 서비스에 대한 요구가 더 커질 것임을 예상해야 한다.

이노바 솔루션(Innova Solutions)의 CTO 스리다르 카지 페타는 단순한 생산성 향상 이상의 목표를 달성하기 위해는 이미 디지털 트랜스포메이션의 혜택을 얻고 있지만 여전히 사람의 분석에 의존하는 롱테일 매출을 가속하는 데 초점을 맞추는 것이 중요하다고 강조한다. 이제 AI로 이 영역을 강화해서 더 큰 매출 성장을 이룰 수 있다. 주요 영역으로는 롱테일 고객 요구 사항 분석을 통한 제품 및 서비스 조정, 가격 및 프로모션 최적화, 틈새 영역을 위한 타겟 마케팅 콘텐츠 제작, 전통적인 영업 전략 이상의 새로운 고객 세그먼트 파악 등이 있다.

컴퍼니 서치 인코퍼레이티드(Company Search Incorporated: CSI)의 COO인 폴 보인튼은 전략적 분석 사용 사례로서 생성형 AI는 시장 동향 분석, 제품 수요 예측, 공급망 효율성 최적화, 판매와 성장을 이끄는 파트너십 식별을 위한 사용자 인터페이스를 대폭 개선해 준다고 언급했다.

이처럼 증가하는 비즈니스 요구를 충족하려면 데이터 사이언티스트는 비즈니스에 대한 이해도를 높이고 매출 성장

지도 방법

지속 가능 발전 목표를 달성하기 위해 데이터 과학이 매우 중요한 역할을 차지하고 있음을 설명하도록 한다.

2 | 지속 가능한 미래를 만드는 데이터 과학

01 UN의 지속 가능 발전 목표와 데이터 과학

세계 연합(UN)은 전 세계 빈곤을 종식시키고, 지구를 보호하며, 2030년까지 모든 사람들이 평화와 번영을 누릴 수 있는 지속 가능한 미래를 만들기 위한 17개의 지속 가능 발전 목표(SDGs: Sustainable Development Goals)를 제시하였다. 이러한 지속 가능 발전 목표를 달성하는 데 데이터 과학은 핵심적인 역할을 한다. 데이터 과학을 활용하면 각 목표에 대한 정확한 모니터링과 평가가 가능하며, 이를 통해 국가와 국제 사회가 자원을 효율적으로 배분하고 전략을 수립할 수 있다. 보다 구체적으로 살펴보면 데이터 과학은 다음과 같은 다양한 기술과 접목되어 전 세계의 지속 가능한 발전에 도움을 주고 있다.

인공위성 기술

인공위성을 활용하면 환경 및 지리적 문제점과 위험 요소를 데이터에 기반하여 파악할 수 있다.

웨어러블 기술

스마트 워치와 같은 웨어러블 기기는 환자들의 건강 데이터를 수집하고 치료하는 데 활용된다.

핀테크 기술

핀테크 기술의 발달로 디지털 송금이 가능하여 사람들의 재산 접근성이 증대되었을 뿐만 아니라 보다 안전하게 돈을 관리할 수 있다.

교통 데이터 모니터링 기술

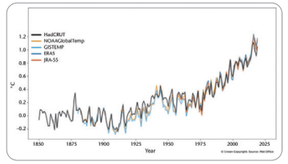
지속적인 교통량 모니터링 및 교통 데이터 수집을 통해 도시 교통 계획을 수립하고, 개인의 교통 카드 데이터를 활용하여 인구 이동에 대한 예측을 할 수 있다.

02 국가 지속 가능 발전 목표와 데이터 과학

우리나라에서도 국제 사회의 공동의 목표 달성에 기여하고 우리 사회가 처한 여러 문제들을 해결하기 위해 한국형 지속 가능 발전 목표인 국가 지속 가능 발전 목표(K-SDGs: Korean Sustainable Development Goals)를 수립했다. 여기에는 모두가 사람답게 살 수 있는 포용 사회 구현, 모든 세대가 누리는 깨끗한 환경 보전, 삶의 질을 향상시키는 경제 성장, 인권 보호와 남북 평화 구축, 지구촌 협력과 같은 5대 전략과 이를 실천하기 위한 17개 목표가 제시되었다.

국가 지속 가능 발전 목표

이러한 국가 지속 가능 발전 목표를 달성하는 데 데이터 과학은 다양한 영역에서 활용되고 있다. 예를 들어 데이터 분석을 통해 기후 위기를 진단하고 기후 위기에 대응하기 위한 구체적인 전략을 제시하고 있고, 데이터를 기반으로 작물을 재배하는 스마트팜을 통해 지속 가능한 식량 생산 체계를 구현하고 있다.



데이터 분석을 통한 기후 위기 진단: 대기질 및 폐기물 관리 등 도시가 갖는 부정적인 환경 영향을 검소시킨다(K-SDGs 11-6).

데이터 기반으로 작물을 재배하는 스마트팜: 지속 가능한 식량 생산 체계를 구축한다(K-SDSs 2-3).

해 보기 2 지속 가능한 미래를 만들기 위한 사례 조사하기 — 조사하기

본문에 제시된 사례 외에 인류의 지속 가능한 미래를 만들기 위해 데이터 과학이 활용되는 사례 또는 활용될 수 있는 분야에 대해 조사해 보자.

- 지진, 산불, 감염병, 해양 선박 사고 등 재난 관련 데이터로 사고 예방하기

예시 답안

어업 활동과 수산물 생산은 경제 성장과 생계, 영양, 생물 다양성, 생태계와 본질적으로 연관되어 있으며, 어업의 지속 가능성 향상을 위한 노력은 다양한 SDGs 달성에 기여한다.
지속 가능한 어업은 남획(濫獲: 짐승이나 물고기 따위를 마구 잡음)을 종식하고, 해양 생태계 보호 및 회복 등을 목표로 하는 『SDG 14. 해양 생태계 보호』 달성에 필수적이다.

[출처] https://www.msc.org/docs/default-source/kr-files/%EC%A7%80%EC%86%8D%EA%B0%80%EB%8A%A5%EC%96%B4%EC%97%85%EA%B3%BC-sdgs-%EB%B3%B4%EA%B3%A0%EC%84%9C.pdf

을 위한 새로운 데이터 집합을 발견하고 분석할 방법을 찾아야 한다.

② AI 생성 대시보드와 통합

데이터 사이언티스트는 전통적으로 새로운 데이터 집합에 대해 파악하거나 비즈니스 사용자가 데이터에 대한 질문에 답하는 데 도움을 주기 위한 빠르고 쉬운 방편으로 대시보드를 개발해 왔다. 데이터 시각화와 분석 플랫폼에는 지난 몇 년에 걸쳐 자연어 쿼리와 머신러닝 알고리즘이 추가됐지만, 데이터 사이언티스트는 생성형 AI가 이끄는 새로운 혁신의 물결을 예상하고 대비해야 한다.

IBM의 비즈니스 분석 제품 관리 부사장인 앨빈 프란시스는 향후 2년 동안 정적인 비즈니스 인텔리전스 대시보드에서 더 동적이고 개인화된 분석 경험으로 전환할 것으로 예상한다. 생성형 AI를 통해 사용자들이 분석에서 노이즈를 제거하고 실용적인 인사이트를 대화 형식으로 얻을 수 있게 되면서 기존 대시보드에 대한 의존도는 줄어들게 된다. 임시 대시보드를 생성할 필요가 없게 된 데이터 분석가와 데이터 사이언티스트는 조직의 지식을 의미 계층으로 문서화하고 전략적 분석을 통해 선순환을 구축하는 데 집중하게 될 것이다.

C데이터(CData)의 선임 기술 에반젤리스트인 제로드 존슨은 생성형 AI 플랫폼이 시각화 툴에 통합되면서 더 동적이고 인터랙티브한 데이터 표현이 실현되고 실시간 합성과 시나리오 분석이 가능해질 것이라고 한다. 이런 툴이 향후 몇 년 동안 발전해서 시각화의 직관성과 인사이트를 강화하고, 더 나아가 묻지 않은 질문에도 답하는 등 혁신적인 발견을 지원할 수 있게 될 것이라고 예상한다.

데이터 사이언티스트는 이 기간 동안 자신의 데이터 시각화 플랫폼에서 생성형 AI 기능을 사용하는 방법을 익혀야 한다. 시각화가 더 쉬워짐에 따라 데이터 사이언티스트는 고급 분석 기능을 사용해 새로운 유형의 인사이트를 제공할 준비를 해야 한다.

③ 시민 데이터 사이언티스트 지원

많은 전문가가 시민 데이터 사이언티스트를 대상으로 하는 기능이 증가하고 생성형 AI 기능을 갖춘 셀프서비스 비즈니스 인텔리전스 툴을 배우는 비즈니스 인력이 증가할 것으로 전망한다.

SAP 북미 사업부 AI 책임자인 제러드 코일은 생성형 AI가 데이터의 잠재력을 끌어내서 IT 전문가가 확장된 기능과 자동화된 워크플로를 통해 계획 및 분석 역량을 최적화할 수 있게 해준다고 한다. 이러한 발전을 통해 복잡한 작업이 간소화되고 기술 전문가가 아닌 사용자도 고급 툴에 더 쉽게 접근할 수 있게 된다. 향후 몇 년 동안 더 많은 일상적인 작업이 자동화되면서 팀은 더 전략적인 작업에 집중할 수 있게 되고 조직 전반적으로 더 효율적인 데이터 주도 의사 결정이 이뤄질 것이다.

이 성장은 데이터 시각화 툴을 통해 자연어 기능이 강화되고 ML 모델 적용이 자동화되는 데 따르는 결과이다. 이런 기능은 시민 데이터 사이언티스트의 작업을 간소화해서 전문 지식이 많지 않은 사용자도 보다 간편하게 데이터를 쿼리하고 이상값을 찾고 추세를 파악하고 대시보드를 만들고 유지할 수 있게 된다.

RR 도넬리(RR Donnelley)의 엔터프라이즈 AI 설계자 샤모딥 사카르는 생성형 AI 기반 애플리케이션과 플랫폼이 동적인 시각화, 데이터 스토리텔링 내러티브, 복잡한 데이터 인사이트에 대한 명확한 설명을 생성할 수 있다고 한다. 이를 통해 기술 전문가가 아닌 사용자도 데이터를 더 쉽게 이해할 수 있으므로 대규모 조직 전반에서 유능한 시민 데이터 분석가를 육성하는 데 도움이 될 것이라고 강조한다.

데이터, 분석, 시각화, 모델링 기술이 데이터 사이언스에

용어 해설 **데이터 윤리**

데이터 윤리는 시민과 기업의 권리를 넘어서는 사회적 가치를 담을 수 있는 기술의 윤리적 차원을 다룰 수 있어야 한다.
데이터를 책임감 있고 지속 가능하게 사용하는 것이며, 궁극적으로 사람과 사회에 대한 선한 일을 하는 것이다.
[출처] 인공지능(AI)의 학습용 데이터 윤리 가이드라인에 대한 연구. 한국교육학술정보원. 2023.

3 | 데이터 과학 속 데이터 윤리

데이터로 인해 세상의 많은 부분이 편리하게 변화하고 있지만, 데이터를 활용할 때 윤리적인 문제가 발생하기도 한다. 특히 데이터를 기반으로 한 인공지능의 발전에 따라 데이터 윤리는 더욱 중요해지고 있다.

01 데이터 편향성 문제

AI사에서 채용 심사 시스템을 구축할 때 남성 데이터를 중심으로 모델을 학습한 결과, 여성에게 불리하게 적용되었다는 사실이 밝혀져 결국 시스템을 폐기하였다.

수집된 데이터의 분포가 고르지 못하고, 특정한 지역, 연령, 성별 등에 치우쳐 있을 경우 데이터에 편향성이 있다고 한다. 실제 현실을 대표하지 못하고 특정한 방향으로 치우쳐 있는 데이터를 가지고 모델을 만들면, 특정 대상에게 유리하거나 불리한 모델이 만들어질 수 있다.

편향(偏向) 한쪽으로 치우침

데이터 편향성에 대한 자세한 설명은 47쪽을 참고한다.

이를 예방하기 위해서는 데이터 수집 단계에서부터 데이터가 편향되지 않았는지 검토하고, 데이터 분석의 전 과정에서 편향*을 최소화하기 위해 노력해야 한다.

02 데이터 속 혐오 표현 문제

데이터 편향성 문제가 수집된 데이터의 분포와 관련된 문제라면, 데이터의 질과 관련된 대표적인 문제로는 인공지능의 혐오 표현 문제가 있다. 과거 인공지능 챗봇이 잘못된 데이터를 학습하여 유태인 학살을 옹호하는 발언을 하여 문제가 된 것도 혐오 표현의 사례다. 이러한 문제가 발생하지 않도록 수집된 데이터를 필터링하는 과정도 필요하지만, 이는 결국 평소 사람들이 사용하는 데이터로 인해 발생한 문제이기 때문에 디지털 세상에서 바른 언어를 사용하는 습관을 가져야 한다.

알고 가기 혐오 표현을 줄이기 위한 노력 – 혐오 표현 데이터셋 공개

S사의 인공지능 센터는 악성 댓글과 혐오 표현에 대한 대규모 데이터셋을 공개하였다. 2019년 1월부터 2021년 7월 초까지 포털과 온라인 커뮤니티에서 수집한 55만 개의 데이터 중 1만 개를 선정하였다. 혐오 표현은 여성, 남성, 성소수자, 인종, 연령, 지역, 종교 등 8개의 범주로 분류하였다. 이러한 혐오 표현 데이터셋은 혐오 표현을 걸러내는 시스템을 구축하는 데 활용될 예정이다.

03 데이터 소유권 문제

학교 수업 중에 생성된 데이터는 학생의 것일까? 학교의 것일까? 서비스를 만든 기업의 것일까? 데이터가 중요한 자원이 됨에 따라 데이터의 소유권자를 누구로 볼 것인지 합리적으로 결정하는 것도 중요한 문제 중 하나다. 이러한 쟁점에 대해 충분히 논의하고 사회적 합의를 이끌어 내기 위해서는 사회 구성원들의 적극적인 참여와 관심이 필요하다.

데이터 소유권 문제

해 보기 3 데이터 윤리 문제와 관련된 사례 조사하기 — 조사하기

데이터 윤리 문제와 관련하여 직접 경험했거나 간접적으로 들었던 사례가 있다면 소개해 보자.

소단원 요약 1분

1 데이터는 과거부터 금융, 정치 등 다양한 분야에 활용되어 왔으나 최근 디지털화되면서 의료, 제조, 교육 등 모든 분야에서 사회 변화를 주도하고 있다.
2 데이터는 지속 가능한 발전을 만드는 데 활용될 수 있다. 하지만 데이터를 활용할 때 윤리적 문제가 발생할 수 있으므로, 이를 인식하고 비판적인 태도로 활용할 줄 알아야 한다.

해 보기 3 지도 방법

데이터 윤리와 관련된 최신 이슈를 살펴보며, 데이터 윤리가 우리 생활과 밀접하게 연관되어 있다는 사실을 깨닫도록 한다. 특히 데이터 윤리는 AI 윤리와 뗄 수 없는 관계라는 것을 인식하도록 한다.

예시 답안

1 이루다 AI 챗봇 사건
2020년 12월 출시된 AI 챗봇 '이루다'가 성소수자와 장애인에 대한 차별 발언, 개인 정보 유출 등의 문제로 한 달 만에 서비스가 중단되었다.

2 오픈AI ChatGPT 개인 정보 유출 사건
2023년 3월 오픈AI의 챗GPT 서비스에서 개인 정보 유출이 발생하여 한국 이용자 687명의 성명, 이메일, 결제지, 신용카드 번호 일부가 노출되었다.

서 비즈니스팀으로 이전되는 추세는 약 10년 전부터 일어나고 있지만, 생성형 AI는 이 전환을 가속하는 역할을 하게 될 가능성이 높다. 이것이 데이터 사이언티스트와 이들의 업무에 의미하는 바는 무엇일까?

EDB의 최고 제품 엔지니어링 책임자 조제프 드브라이스는 생성형 AI가 분석에 통합되면서 데이터 준비, 기본 분석과 같은 일상적인 작업은 점점 더 자동화되고 그에 따라 인사이트에 더 깊이 파고들 수 있는 시간이 늘어난다고 설명한다. 고급 AI 툴은 데이터 시각화와 스토리텔링을 더 직관적으로 만들어 주므로 데이터 사이언티스트가 복잡한 결과물을 기술 전문가가 아닌 동료에게 더 쉽게 전달할 수 있고, 동료는 자연어를 사용해 데이터를 탐색할 수 있다. 이는 데이터 팀과 다른 부서 간의 간극을 이어 더 협업적인 환경을 조성하는 데 도움이 될 것이다.

아스트로노머(Astronomer)의 CTO 줄리안 라니브는 데이터 사이언스팀이 생성형 AI 기능으로 인해 이해(利害) 관계자의 관심과 참여가 증가할 것을 예상해야 한다면서 '데이터를 다루고 데이터에서 인사이트를 추출하는 작업의 진입 장벽이 대폭 낮아지게 되므로 견고한 데이터 문화와 관행을 확립하는 것이 매우 중요하다'고 강조한다. 기술 전문가가 아닌 동료를 위해 데이터 엔지니어링 베스트 프랙티스와 잘 분류된 데이터 사전을 기반으로 적절한 데이터 플랫폼을 개발할 것을 권한다. 또 다른 역할은 최종 사용자를 위한 적절한 거버넌스와 가드레일에 대한 컨설팅이다.

4 비정형 데이터 집합 활용

비즈니스 사용자가 쉽게 데이터 행과 열을 분석할 수 있게 되면서 데이터 사이언티스트는 비정형 데이터 소스를 살펴보기 위한 기술과 분석 노력을 확장해야 한다. 마케팅, 영업, 고객 서비스 데이터 집합은 대부분 비정형이므로 이를 분석하면 성장과 경쟁 우위를 추구하는 비즈니스와 보조를 맞추는 데 도움이 된다.

m펄스(mPulse)의 CPO 사이드 아민자데는 생성형 AI가 고객 중심 기업이 대량의 자유 텍스트 대화를 종합하고 분석하는 방법에 혁신을 일으키고 있으며, 이런 고급 툴은 고객 의도와 요구를 대규모로 정확히 분류함으로써 더 풍부하고 실용적인 인사이트를 제공한다고 강조한다.

데이터 사이언티스트가 배워야 할 기술은 그래프 데이터베이스이다. 또 다른 기술인 지식 그래프는 도메인 인텔리전스로 LLM(Large Language Model) 모델을 증강하는 RAG를 개발하는 데 유용할 수 있다.

릴레이셔널AI(RelationalAI)의 연구 ML 부문 부사장 니콜라우스 바실로글로우는 데이터를 일반 SQL 테이블이 아닌 지식 그래프로 체계화하면 고급 분석 수행과 머신러닝 모델 실행 측면에서 큰 이점이 있다고 조언한다. 가장 빈번한 작업은 특징 공학(Feature Engineering)이며, LLM이 지식 그래프에 내장됨에 따라 데이터 사이언티스트는 더 유의미한 특징 생성을 기대할 수 있다.

쿠모 AI(Kumo AI)의 공동 창업자이자 엔지니어링 책임자인 헤마 라가반은 데이터 사이언티스트가 그래프 신경망(GNN: Graph Neural Networks)에 익숙해야 한다면서 'GNN이 여러 테이블을 살펴보고 예측 AI 작업에 필요한 신호를 찾는 기능을 통해 많은 특징 공학 워크플로우의 필요성을 없애 준다'고 강조한다. 그러면 데이터 사이언티스트는 영향과 예측을 연결할 수 있는 비즈니스의 기회를 식별하는 데 집중할 수 있다.

5 AI 에이전트 및 모델 활용

데이터 사이언티스트가 관심을 가져야 할 2가지 새로운 AI 기능은 업종별 AI 모델과 AI 에이전트이다.

최근 세일즈포스는 자동차, 금융 서비스, 의료, 제조, 소매를 포함한 15개 업종에 걸쳐 업종별 과제를 해결하는 맞

지도 방법

지금까지 살펴본 다양한 데이터 과학의 적용 분야를 바탕으로 자신의 진로 분야에서 데이터에 기반한 문제 해결 사례를 조사하고, 데이터 과학이 자신의 진로 분야에 끼치는 영향을 조사하는 활동이다.
내용을 조사하는 데 어려움을 겪는 학생들이 있을 경우, 최근 데이터 과학은 인공지능과 밀접한 연관이 있으므로 인공지능 관련 내용을 중심으로 탐구 방향을 제시하는 것도 좋은 방법이다.

예시 답안

❶ 넷플릭스(Netflix)는 수백만 명의 사용자가 접근하는 대규모 콘텐츠 라이브러리를 보유하고 있다. 이로 인해 사용자가 자신에게 맞는 콘텐츠를 찾기 어려운 문제가 발생한다.
▶**해결책:** 넷플릭스는 사용자 데이터(시청 기록, 평가, 검색 패턴 등)를 분석하여 개인화된 추천 시스템을 개발했다. 이를 통해 각 사용자에게 맞춤형 콘텐츠를 추천한다.

❷ 스포티파이(Spotify) 사용자들은 수백만 곡의 음악 중에서 자신이 좋아하는 음악을 발견하는 데 어려움을 겪는다.
▶**해결책:** 스포티파이는 사용자 청취 데이터, 좋아요 및 플레이리스트 정보를 분석하여 개인화된 음악 추천 알고리즘을 개발했다. 이를 통해 사용자는 자신의 취향에 맞는 새로운 음악을 쉽게 발견할 수 있다.

탐구 활동 데이터 과학과 나의 진로

데이터 과학과 함께 4차 산업 혁명 시대가 도래함에 따라 다양한 진로와 직업의 새로운 기회가 열리고 있다. 다음 세 가지 분야를 살펴보며, 나의 진로 분야에 데이터 과학이 어떤 영향을 끼칠지 분석해 보자.

분야 ❶ | 의료 분야의 진단 및 치료 개선

데이터 과학을 활용한 인공지능 및 빅데이터 기술은 의료 분야에서 질병 진단과 치료 개선에 도움을 준다. 이를 통해 의사들은 환자의 증상과 병력을 기반으로 한 정확한 진단을 내릴 수 있으며, 개별화된 치료 방안을 강구하는 데 도움을 받을 수 있다. 또한 데이터 과학은 새로운 약물 개발 및 임상 시험 과정에서도 큰 역할을 하며, 의료 분야에서의 연구와 발전을 돕는다.

분야 ❷ | 환경 보호와 지속 가능한 발전

데이터 과학은 환경 데이터를 분석하여 기후 변화, 환경 오염, 자원 소모 등에 대한 통찰을 제공한다. 이러한 정보를 바탕으로 정책 결정자들과 기업들은 지속 가능한 발전을 도모하고 환경 문제를 해결하기 위한 전략을 수립할 수 있다.

분야 ❸ | 스마트 도시 및 교통 관리

데이터 과학 기술은 도시의 교통 흐름, 에너지 소비, 안전 및 생활 편의를 개선하는 데 사용될 수 있다. 스마트 도시 기술을 통해 교통 정체를 줄이고, 에너지 효율을 높이며, 치안을 강화할 수 있다.

1 나의 진로 분야를 작성해 보자.

엔터테인먼트 분야

2 해당 진로 분야에서 데이터 기반 문제 해결 사례를 조사해 보자.

- 문제:
- 해결책:

3 데이터 과학이 해당 진로 분야에 끼치는 영향을 조사해 보자.

40

예시 답안

❶ **넷플릭스(Netflix) 영향**
- 시청 시간 증가: 개인화 추천 시스템은 사용자가 더 많은 시간을 보낼 수 있도록 한다. 사용자 만족도가 증가하고, 구독 취소율이 감소한다.
- 콘텐츠 소비 패턴 분석: 사용자 데이터 분석을 통해 어떤 유형의 콘텐츠가 인기가 있는지 파악하고, 이를 바탕으로 새로운 콘텐츠 제작에 반영한다.

❷ **스포티파이(Spotify) 영향**
- 사용자 참여 증가: 개인화된 추천이 사용자의 음악 청취 시간을 늘리고, 플랫폼에 대한 충성도를 높인다.
- 아티스트와의 연결 강화: 새로운 아티스트와 음악을 사용자에게 소개하여 음악 시장의 다양성을 증진한다.

춤 구성할 수 있는 사전 구축된 AI 기능 모음인 인더스트리 AI(Industries AI)를 발표했다. 한 의료 모델은 혜택 확인 기능을 제공하고 자동차 모델은 차량 텔레메트리 요약 기능을 제공한다.

AI 에이전트와 관련, 에이세라(Aisera) CEO 아비 마헤시 와리는 AI 에이전트가 추론, 계획, 의사 결정, 툴 사용에 관여해서 CRM, ERP 트랜잭션과 같은 작업을 자율적으로 처리함으로써 LLM을 강화한다고 설명한다. 이런 에이전트는 일반적으로 데이터 분석가가 수행하는 데이터 정제, 탐색을 위한 데이터 분석, 특징 공학, 예측 등의 데이터 작업을 간소화해 준다.

이 2가지 추세는 데이터 사이언스의 역할에 대한 2차적인 변화, 즉 데이터 랭글링에서 머신러닝 모델 개발, AI 에이전트 활용, 서드파티 모델 조사, 그리고 시민 데이터 사이언티스트와의 협업을 통한 AI, 머신러닝, 기타 데이터 사이언스 기능 적용으로의 변화를 보여 준다.

또한 데이터 사이언티스트는 AI 윤리와 이것이 기업의 AI 거버넌스에 어떻게 기여하는지에도 정통해야 한다. 플로우X.AI(FlowX.AI)의 AI 책임자 보그단 라두타는 생성형 AI가 분석에 더 깊이 개입함에 따라 데이터 사이언스팀은 새로운 기술을 습득하고 전략적 협업에 집중하고 AI 윤리를 우선하는 방식으로 적응해 나가야 한다고 언급했다. 젠팩트의 메논은 데이터 스토리텔링에서 생성형 AI를 사용하기 위해서는 윤리적인 사용, 투명성, 공정성을 보장하기 위한 책임감 있는 AI를 통해 생성된 콘텐츠의 정확성을 보장하고 편견을 줄이는 것과 같은 지속적인 과제를 해결해서 데이터 주도 의사 결정의 신뢰와 정확성을 강화해야 한다고 강조했다.

AI는 데이터 사이언티스트가 업무를 수행하는 방식과 이들이 집중하는 작업의 유형에 변화를 일으키고 있음은 의심할 여지가 없다. 따라서 진정한 기회는 이 기술을 통해 기업을 앞으로 이끌고 분석 기반 효과를 윤리적인 방식으로 제공하는 데 있다.

지식 충전소 데이터 과학과 사회 문제

데이터 과학으로 인해 많은 기회가 우리에게 찾아왔지만, 빛에 그림자가 있는 것처럼 그 이면에는 데이터 과학으로 인한 문제 역시 자리 잡고 있다. 다음 세 가지 사례를 살펴보자.

노동 시장의 다변화와 유연성

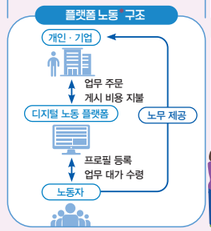
기술의 발달로 새로운 일자리가 창출되고 다양한 형태의 직업이 생겨나고 있지만, 그로 인해 배달 기사 등 디지털 플랫폼 노동자들의 처우 문제가 발생하고 있다. 이들은 종종 사회적 보장이 제한되며, 불안정한 고용 및 저임금 문제를 겪을 수 있다.

디지털 리터러시*에 따른 격차 문제

키오스크와 같은 무인 기기들 또는 모바일 앱을 활용해 주문을 하는 것이 일상화되면서 편리해진 부분이 있지만, 한편으로는 디지털 리터러시가 부족한 사람들을 위한 대책이 필요하다. 이는 특히 노년층, 저소득층 등 사회적 약자들에게 영향을 미치며, 학생들 역시 디지털 리터러시에 따라 배움의 격차가 발생할 수 있다.

초연결 사회와 개인 정보 문제

초연결 사회에는 언제 어디서나 연결될 수 있어 새로운 기회를 많이 만들 수 있고 빅데이터를 활용해 많은 문제들을 해결할 수 있지만, 한편으로 개인 정보 유출 및 사생활 침해, 초연결 사회에 대한 피로감을 호소하는 경우도 발생할 수 있다.



이와 같이 데이터 과학으로 인해 발생하는 사회적 문제를 인식하며 데이터 과학의 빛과 어두움을 모두 이해할 줄 아는 비판적 태도를 갖추어야 한다.

디지털 리터러시(digital literacy) 디지털 기술을 활용해 정보를 찾고, 이해하며, 활용하는 능력을 말한다. 즉, 디지털 기술을 사용하는 데 필요한 능력과 지식을 갖추는 것을 뜻한다.

플랫폼 노동 인터넷 기반의 디지털 플랫폼을 통해 제공되는 일시적이고 유연한 일자리를 말한다. 이러한 플랫폼은 앱 또는 웹 사이트 형태로 구성되어 있으며, 일자리 제공자와 일자리 수요자를 연결해 준다. 플랫폼 노동은 택배 배송, 음식 배달, 운송, 가사 도우미 등 다양한 서비스 분야에서 이루어질 수 있다.

41

지도 방법

데이터 과학의 발전에는 긍정적인 면만 존재하지 않는다. 데이터 과학의 발전으로 인해 생기는 사회적 부작용을 살펴봄으로써, 데이터 과학이 만능이거나 절대 선이 아니라는 비판적 시각을 갖는 것이 중요하다.
따라서 지식 충전소에 제시된 다양한 사례를 살펴보며, 데이터 과학으로 인한 어두운 문제들을 해결하기 위해 어떤 노력을 해야 하는지 함께 생각해 보는 시간을 갖도록 한다.

* **거버넌스(Governance):** 조직이나 시스템이 운영되는 방식과 의사 결정 과정을 의미한다. 일반적으로 조직 내에서 목표 달성을 위해 자원을 효율적으로 관리하고, 위험을 완화하며, 법규 준수와 투명성을 유지하는 것을 목표로 한다.
* **로우코드(Low-code):** 소프트웨어 개발에서 최소한의 코딩으로 애플리케이션을 개발할 수 있는 개발 플랫폼을 말한다. 로우코드 플랫폼은 시각적인 인터페이스와 미리 구축된 구성 요소를 제공하여 개발자가 코딩에 집중하는 대신 비즈니스 로직과 기능에 집중할 수 있도록 돕는다.
* **데브섹옵스(DevSecOps):** 개발(Development), 보안(Security), 운영(Operations)의 융합을 의미한다. 이는 소프트웨어 개발 생명 주기에서 보안을 통합하여 보안 취약점을 줄이고, 안전한 소프트웨어를 개발하고 배포하는 것을 목표로 한다. 데브섹옵스는 개발팀, 보안팀, 운영팀이 협력하여 보안 작업을 자동화하고, 지속적인 보안 테스트와 배포를 수행하는 것을 강조한다.

[출처] https://www.samsungsds.com/kr/insights/what-do-data-scientists-do.html

Q&A 묻고 답하기 • 교과서 36~37쪽 •

Q1 지속 가능 발전(Sustainable Develop ment)이란 무엇인가요?

A 1987년 세계환경개발위원회(WCED)에서 발표한 『우리 공동의 미래(Our Com mon Future)』 보고서에서 지속 가능 발전에 대한 개념이 공식적으로 정의되었다. 이에 따르면, 지속 가능 발전은 '미래 세대의 필요를 충족시킬 능력을 훼손하지 않으면서, 현 세대의 필요도 충족하는 발전'을 의미한다. 이후 지속 가능 발전 개념이 활발히 논의되었으며, 경제 발전 · 사회 통합 · 환경의 지속 가능성을 고려한 발전으로 그 의미가 확장되었다.

Q2 지속 가능 발전 목표(SDGs)란 무엇인가요?

A 2015년 9월 유엔 총회에서는 『세계의 변혁: 지속 가능 발전을 위한 2030 의제(Transforming Our World: The 2030 Agenda for Sustainable Development)』 결의문을 채택했다. 이 문서에는 전 세계가 인류의 지속 가능한 발전을 위해 2030년까지 공동 달성하기로 합의한 17개 목표와 후속 조치 방안이 담겨 있다.

Q3 SDGs 지표는 어떻게 개발되었나요?

A 각 지역 대표 통계청을 회원으로 하는 SDGs 지표 전문가 그룹(IAEG-SDGs: Inter Agency and Expert Groups on SDG Indicators)이 주도하여 지표를 개발했다. 이 그룹은 2015년 3월 유엔통계위원회에서 결성되었다. 지표 선정 원칙에 따라, 국제기구 및 다양한 이해 당사자 그룹과의 논의를 거쳐 2017년 3월에 232개 지표를 개발하였고, 이 지표는 같은 해 7월 유엔 총회에서 채택되었다. 2020년 종합 개편을 통해 36개 지표가 변경되었으며, 현재 231개 지표 체계로 운영 중이다.

Q4 SDGs 데이터는 어떻게 수집되나요?

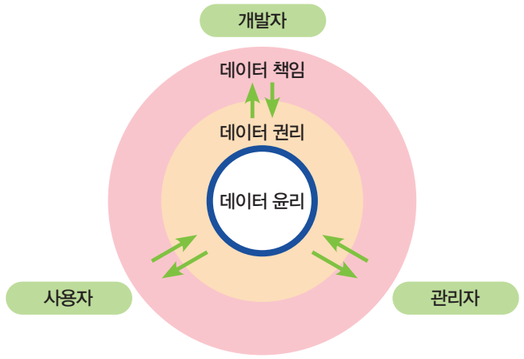
A 유엔은 지표별로 국제기구를 담당자로 지정하여 지표 방법론 개발 및 데이터 수집의 역할을 부여하였다. 또한, 국가별로는 SDGs 데이터 국가 책임 기관을 지정하였는데, 우리나라는 통계청 통계 개발원이 책임 기관이다. 지표 소관 국제기구에서 데이터 제공 요청 시, 통계청은 국내 28개 지표 관계 부처와 협력하여 데이터를 제공하고 있다.

Q5 국가에서 국제기구에 제공한 데이터는 어떻게 처리되나요?

A 국제기구는 국가로부터 제공받은 데이터를 국제 비교를 위해 유엔 SDGs 지표 개념 정의에 따라 보정한 후, 유엔 통계처에 제출한다. 국제기구 데이터와 국내 데이터 간에 차이가 여기에서 발생할 수 있다. 국제기구가 최종 제출한 데이터는 유엔 SDGs 데이터베이스(unstats.un.org/sdgs/indicators/database)에 수록된다.

[출처] https://kostat-sdg-kor.github.io/sdg-indicators/guidance/

데이터 윤리의 핵심 요소와 데이터 주체 • 교과서 38쪽 •



▲ 데이터 윤리의 핵심 요소와 데이터 주체

참고 자료 • 교과서 38쪽 •

생성형 AI와 데이터 편향성 문제

- **제목**: 데이터 편향이란 무엇이며 어떻게 해결할 수 있을까?
- **자료 내용**: 생성형 AI는 학습 데이터의 편향으로 인해 특정 인종이나 정치 성향을 반영하는 문제가 발생한다. 예를 들어 이미지 생성 AI인 Stable Diffusion에 'unprofessional한 사람의 이미지'를 그려달라고 요청하면 '고령의 흑인 남성'과 같은 특정 인종과 성별을 반영한 이미지를 그려낸다. 이를 해결하기 위해서는 데이터의 다양성 확보, 평가와 감독, 투명성 및 책임 강화가 필요하다.

▲ 이미지 생성형 AI의 편견

[그림 출처] Business Insider

|주소| https://blog-ko.superb-ai.com/generative-ai-and-the-data-bias-problem-what-is-data-bias-and-how-can-it-be-solved/

데이터 소유권 관련 주요 내용 • 교과서 39쪽 •

❶ 데이터의 특성과 소유권 개념의 적용

- 데이터는 무형의 자산으로, 기존의 물권이나 저작권 개념을 그대로 적용하기 어렵다.
- 데이터는 복제와 공유가 쉽고, 여러 사람이 동시에 사용할 수 있어 배타적 소유권 개념과 맞지 않다.

❷ 데이터 유형별 소유권 논의

- 개인 정보 데이터: 정보 주체인 개인의 권리와 데이터를 수집 · 처리하는 기업의 권리 사이의 균형이 중요하다.
- 생성 데이터: 기업이 생성하거나 수집한 데이터에 대한 소유권 문제가 있다.
- 공공 데이터: 정부나 공공 기관이 보유한 데이터의 소유권과 활용 방안에 대한 논의가 있다.

❸ 주요 쟁점

- 데이터 소유권 vs 개인 정보 보호: 데이터 활용과 개인 정보 보호 사이의 균형이 중요한 과제이다.
- 데이터 노동 개념: 개인이 생성하는 데이터를 노동으로 인정할지에 대한 논의가 있다.
- 의료 정보 소유권: 환자의 개인 정보와 의료인의 전문 지식이 결합된 의료 정보의 소유권 문제가 있다.

참고 자료 • 교과서 39쪽 •

데이터 소유권 문제

- **제목**: 데이터 소유권과 데이터 노동 문제
- **자료 내용**: 개인이 서비스 이용 시 제공하는 개인정보와 생성된 데이터의 소유권과 노동에 대한 문제가 제기된다. 데이터를 생성하는 행위가 노동으로 간주될 수 있는지에 대한 논의가 필요하다.

|주소| https://www.etnews.com/20211011000062

참고 동영상 • 교과서 41쪽 •

플랫폼 노동이란?

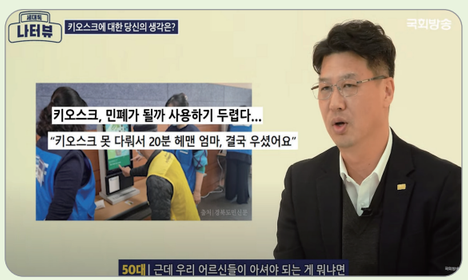
- **제목**: [KDI 경제정보센터] e-경제정보리뷰: 플랫폼 노동 개념편
- **영상 내용**: 거스를 수 없는 시대의 흐름, 플랫폼 노동이란 무엇인지 알아본다.

|주소| https://youtu.be/EczulafZkzg?si=TzlhfZxwkq3-ozDg

참고 동영상 • 교과서 41쪽 •

디지털 격차에 대한 각 세대별 해결 방안은?

- **제목**: #디지털 시대 심해지는 #디지털 격차 각 세대의 생각은?
- **영상 내용**: 디지털 보편화 시대, 갈수록 커져 가는 디지털 격차 해소 방안은 무엇인지 알아본다.

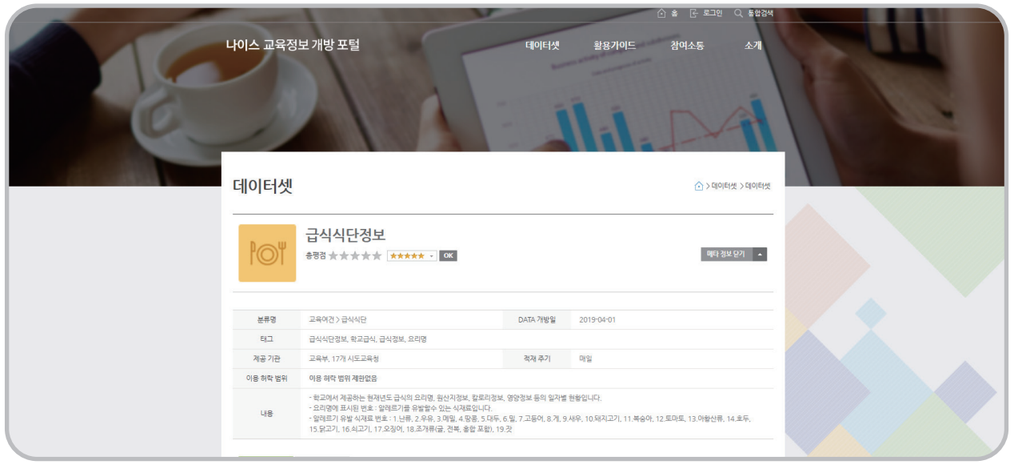


|주소| https://youtu.be/r2cq2R8mlqg?si=fJ0YZxdEaM4F2PUD

추가 활동지 1

우리 학교 급식 식단 데이터셋 살펴보기

‘나이스 교육정보 개방 포털(open.neis.go.kr)’ 홈페이지에서 [데이터셋] – [급식식단정보] – [Sheet]를 선택하고, 우리 학교의 올해 급식 데이터셋을 다운로드해 보자.



1 데이터셋에 어떤 속성이 있는지 살펴보자.

예시 답안 학교명, 식사명, 급식일자, 급식인원수, 요리명, 원산지정보, 칼로리정보, 영양정보, 수정일자

2 이 데이터로 어떤 분석을 할 수 있을지 생각해 보자.

예시 답안 우리 학교에서 가장 많이 나오는 메뉴, 요일별 칼로리의 차이, 우리 학교와 인근 학교의 칼로리 차이 등을 분석할 수 있다.

3 데이터를 효율적으로 분석하기 위해 어떤 전처리가 필요할지 생각해 보자.

예시 답안
- 요리명에 〈br/〉와 같은 특수 문자들이 포함되어 있어서 분석하기 어려울 것 같다. 제거해야 한다.
- Kcal라는 글자가 있어서 분석하기 어려울 것 같다. 숫자만 남겨두어야 한다.

지도 방법

우리 학교의 올해 데이터셋을 받도록 안내하고, 학생들의 역량이 충분한 경우에는 다른 기간의 데이터셋과 다른 학교와의 데이터셋을 비교해 보고, 어떤 사실을 알아냈는지 발표시킬 수 있다.

해설

학생들이 가장 관심이 많은 데이터인 급식 데이터셋이 이렇게 공개되어 있음을 학생들에게 알려 주고, 데이터셋에서 의미 있는 결과를 도출할 수 있음을 경험시키기 위한 활동이다.
급식식단정보 외에도 학교기본정보 등 다양한 학교 관련 데이터를 경험할 수 있도록 안내하는 것도 좋은 방법이다.

추가 활동지 2

앞으로 우리나라의 인구 구조는 어떻게 변할까?

통계청에서 운영하는 ’인구로 보는 대한민국(kosis.kr/visual/populationKorea)‘ 홈페이지에서 우리나라의 인구 데이터를 살펴보자.

1 현재 우리나라에 가장 많은 인구가 있는 연령은 몇 살인지 살펴보자.

예시 답안 (2024년 기준) 남: 53세, 여: 63세

2 이 데이터로 어떤 분석을 할 수 있을지 생각해 보자.

예시 답안 (2024년, 만 17세 기준) 남: 233,503명, 여: 219,411명

3 데이터를 효율적으로 분석하기 위해 어떤 전처리가 필요할지 생각해 보자.

예시 답안
- (2024년 기준) 7,824명인 청소년 인구가 2050년에는 4291명으로 감소할 것으로 보이는데, 전체적으로 교육의 구조가 변화될 것이라고 예상된다.
- 독거노인 가구가 2,113,924명에서 2050년까지 4,670,693명으로 2배 이상 늘어나는 것으로 보이는데, 이로 인한 사회적 문제를 미리 예상하고 정책을 세워야겠다는 생각이 들었다.

지도 방법

인구로 보는 대한민국 사이트를 선생님이 먼저 살펴보신 후, 학생들이 흥미로워할 만한 2~3가지의 소재를 살펴보고 수업 시간에 학생들에게 안내한다. 이후 모둠별로 함께 다양한 데이터를 둘러보는 시간을 주고, 이 시간 동안 서로가 발견한 사실들을 서로 공유하도록 한다.

해설

데이터로 미래를 예측한다는 것이 앞으로 배울 다양한 프로그래밍을 통해서도 가능하지만, 시각화만으로도 미래에 대한 가설을 세울 수 있음을 설명하기 위한 활동이다. 특히, 현재까지의 데이터를 바탕으로 미래를 예측하고 합리적인 의사 결정을 하는 것이 데이터 과학의 목적이라는 사실을 별도의 복잡한 기술 없이 이해시킬 수 있다.

I 시험 대비용 | 대단원 평가 문제

선택형

01 데이터 과학을 정의한 설명으로 가장 적절한 것은?

① 데이터를 무작위로 수집하는 학문이다.
❷ 데이터를 기반으로 현상을 이해하고 분석하는 학문이다.
③ 데이터를 보호하고 보관하는 학문이다.
④ 데이터를 그래픽으로 시각화하는 학문이다.
⑤ 데이터를 복구하고 삭제하는 학문이다.

해설 데이터 과학은 데이터를 기반으로 현상을 분석하고 이해하기 위한 통계, 데이터 분석, 기계학습 등을 통합하는 학문이다.

02 데이터 기반 의사 결정의 장점으로 옳지 <u>않은</u> 것은?

① 주관적 결정을 피할 수 있다.
② 다양한 데이터를 바탕으로 합리적 선택을 할 수 있다.
❸ 모든 결정을 수동으로 처리할 수 있다.
④ 과거 데이터를 통해 미래를 예측할 수 있다.
⑤ 데이터 분석을 통해 최적의 결정을 내릴 수 있다.

해설 데이터 기반 의사 결정은 주관적 판단을 배제하고, 데이터를 통해 더 객관적이고 합리적인 결정을 내리게 한다.

03 데이터 과학의 문제 해결 과정에 포함되지 <u>않은</u> 단계는?

① 문제 정의하기
② 데이터 수집하기
❸ 수학적 증명하기
④ 모델링하기
⑤ 모델 평가하기

해설 데이터 과학의 문제 해결 과정은 문제 정의, 데이터 수집, 전처리 및 탐색, 모델링, 모델 평가, 모델 활용 단계로 구성된다.

04 데이터 과학에서 데이터를 전처리하는 이유로 가장 적절한 것은?

① 데이터를 무작위로 삭제하기 위함이다.
❷ 데이터를 분석하기 전에 정제하고 오류를 수정하기 위함이다.
③ 데이터를 보호하고 저장하기 위함이다.
④ 데이터를 아름답게 시각화하기 위함이다.
⑤ 데이터를 복구하여 다시 사용하기 위함이다.

해설 전처리는 데이터를 분석하기 전에 오류나 결측값을 수정하고 정제하는 과정이다.

05 모델 평가하기 단계의 주요 역할을 가장 잘 설명한 것은?

① 데이터를 수집하는 단계이다.
② 수집한 데이터를 저장하는 단계이다.
❸ 모델이 문제를 해결하는 데 적합한지 평가하는 단계이다.
④ 데이터를 삭제하는 단계이다.
⑤ 데이터를 시각적으로 표현하는 단계이다.

해설 모델 평가하기 단계는 모델이 문제 해결에 적합한지 종합적으로 평가하는 단계이다.

06 데이터 과학이 사회 문제 해결에 기여할 수 있는 방식으로 알맞지 <u>않은</u> 것은?

① 데이터를 통해 기후 변화를 예측한다.
② 데이터를 분석하여 교통 혼잡 문제를 해결한다.
③ 데이터를 활용해 질병 예측을 가능하게 한다.
❹ 데이터를 삭제하여 과거 문제를 해결하지 못하게 한다.
⑤ 데이터를 기반으로 정책을 수립하는 데 도움을 준다.

해설 데이터 과학은 데이터를 통해 사회 문제를 분석하고 해결 방안을 제시하는 데 기여한다.

07 데이터 과학에서 모델링의 의미를 가장 잘 설명한 것은?

① 데이터를 무작위로 수집하는 과정이다.
② 데이터를 시각화하는 과정이다.
❸ 데이터를 분석하여 패턴을 추출하고, 모델을 구축하는 과정이다.
④ 데이터를 삭제하는 과정이다.
⑤ 데이터를 저장하고 보호하는 과정이다.

해설 모델링은 데이터를 분석하여 유용한 패턴을 추출하고, 이를 기반으로 모델을 구축하는 과정이다.

08 다음 중 정형 데이터의 특징으로 가장 적절한 것은?

① 구조가 명확하지 않다.
② 이미지 데이터를 저장할 수 있다.
❸ 주로 행과 열로 구성된 표 형태이다.
④ 수집과 관리가 어렵다.
⑤ 속성이 불분명하다.

해설 정형 데이터는 명확한 구조를 가지며, 주로 행과 열로 구성된 표 형태의 데이터를 의미한다.

09 반정형 데이터의 예로 알맞은 것은?

① 이미지 파일
❷ XML 파일
③ Excel 파일
④ 관계형 데이터베이스
⑤ 텍스트 파일

해설 반정형 데이터는 구조가 일부 존재하면서도 유연한 형태를 가진 데이터로, XML이나 JSON 파일이 그 예이다.

10 정형 데이터의 예로 가장 적절한 것은?

① 소셜 미디어 게시글
② 영화 리뷰
❸ 엑셀로 작성된 판매 기록
④ 음성 파일
⑤ 비디오 클립

해설 정형 데이터는 명확한 구조를 가지고 있으며, 엑셀로 작성된 판매 기록이 그 예에 해당한다.

11 데이터셋을 정의한 설명으로 가장 적절한 것은?

① 데이터를 무작위로 저장한 파일
❷ 하나의 주제에 대한 여러 데이터의 집합
③ 데이터를 무작위로 삭제한 후 복구한 파일
④ 소규모 데이터의 집합
⑤ 데이터의 무결성을 보장하는 방법

해설 데이터셋은 하나의 주제에 대한 여러 데이터가 모인 집합을 의미한다.

12 데이터베이스를 사용하면 얻을 수 있는 이점으로 가장 적절하지 <u>않은</u> 것은?

① 데이터의 무결성을 유지할 수 있다.
② 데이터의 접근 권한을 설정할 수 있다.
③ 대규모 데이터를 체계적으로 관리할 수 있다.
④ 데이터를 통합하여 쉽게 조회할 수 있다.
❺ 데이터를 쉽게 변경할 수 없다.

해설 데이터베이스는 데이터를 체계적으로 관리하며, 필요할 때 데이터를 쉽게 갱신하거나 수정할 수 있다.

13 데이터베이스에서 테이블의 의미를 가장 잘 설명한 것은?

① 데이터를 구조화하지 않은 목록
❷ 특정 주제에 대한 데이터를 저장하는 구조화된 표
③ 데이터를 삭제한 후 남은 기록
④ 데이터를 실시간으로 복사하는 도구
⑤ 데이터를 무작위로 선택하는 방법

해설 데이터베이스에서 테이블은 특정 주제에 대한 데이터를 저장하는 구조화된 표 형태를 의미한다.

서술형

14 데이터 과학에서 '데이터 전처리'의 주요 역할을 설명하시오.

예시 답안 데이터 전처리는 데이터를 분석하기 전에 오류를 수정하고 결측값을 처리하는 단계이다.

15 데이터 기반 의사 결정의 장점 중 한 가지를 설명하시오.

예시 답안 데이터 기반 의사 결정은 주관적인 판단을 배제하고, 데이터를 기반으로 객관적이고 합리적인 결정을 내릴 수 있다.

16 데이터 과학의 융합적 특징을 설명하시오.

예시 답안 데이터 과학은 컴퓨터 과학, 수학, 통계, 특정 분야의 전문 지식이 융합된 학문이다.

17 데이터 과학이 의료 분야에서 어떻게 활용될 수 있는지 설명하시오.

예시 답안 데이터 과학은 환자의 의료 데이터를 분석하여 질병을 예측하고 맞춤형 치료 방안을 제시하는 데 사용될 수 있다. 예를 들어, 유전자 데이터를 분석하여 특정 질병의 발병 가능성을 예측하고 예방할 수 있다.

18 데이터 과학이 환경 문제 해결에 기여할 수 있는 방법을 설명하시오.

예시 답안 데이터 과학은 환경 데이터를 분석하여 기후 변화나 대기 오염과 같은 문제를 파악하고, 이를 바탕으로 정책을 수립하거나 문제 해결 방안을 제시하는 데 기여할 수 있다.

19 데이터 과학의 윤리적 고려 사항에 대해 설명하시오.

예시 답안 데이터 과학에서 개인 정보와 같은 민감한 데이터를 다룰 때는 이를 보호하고 윤리적으로 처리하는 것이 중요하다. 데이터의 잘못된 사용은 개인의 권리를 침해할 수 있기 때문에, 데이터 보호 방안이 반드시 고려되어야 한다.

20 정형 데이터의 주요 특징을 설명하시오.

예시 답안 정형 데이터는 명확한 구조를 가지고 있으며, 주로 행과 열로 구성된 표 형태로 이루어져 있어 데이터의 속성을 쉽게 파악할 수 있다.

21 정형 데이터와 비정형 데이터의 차이점을 설명하시오.

예시 답안 정형 데이터는 명확한 구조를 가지고 있으며, 주로 표 형태로 표현되는 데이터를 의미한다. 반면, 비정형 데이터는 명확한 구조가 없어 자유로운 형태로 표현되며, 이미지, 텍스트, 음성 등이 이에 해당한다.

22 데이터의 속성에 따른 문제 해결 과정의 중요성을 구체적으로 설명하시오.

예시 답안 데이터의 속성은 문제 해결 과정에서 매우 중요한 역할을 한다. 예를 들어, 도서관의 위치를 확인하는 문제에서는 위도와 경도 속성이 중요하게 작용하며, 도서관 운영 시간을 확인하는 문제에서는 휴관일과 시작 시각이 중요한 속성이 된다.

23 데이터셋이란 무엇인지 설명하시오.

예시 답안 하나의 주제에 대한 여러 데이터가 모인 집합을 말한다.

24 데이터셋과 데이터베이스의 차이점을 설명하시오.

예시 답안 데이터셋은 하나의 주제에 대한 데이터 집합이며, 데이터베이스는 이러한 데이터셋을 통합적으로 관리하고 여러 사용자가 공유할 수 있도록 지원하는 시스템이다.

25 UN의 지속 가능 발전 목표(SDGs)와 관련하여 데이터 과학이 기여할 수 있는 두 가지 방법을 설명하시오.

예시 답안 데이터 과학은 기후 변화와 같은 환경 문제를 모니터링하고 평가하며, 도시 교통 계획을 세우고 자원을 효율적으로 배분하는 데 기여할 수 있다.

26 데이터 편향성 문제가 사회에 미치는 영향을 설명하고, 이를 해결하기 위한 방안을 제시하시오.

예시 답안 데이터 편향성 문제는 특정 성별, 연령, 지역 등에 치우친 데이터를 바탕으로 한 모델이 특정 집단에 불리하게 작용할 수 있다. 이를 해결하기 위해서는 데이터 수집 시 다양한 표본을 포함하고, 데이터 분석 과정에서 편향성을 검토하여 수정하는 노력이 필요하다.

27 데이터 소유권 문제와 관련된 윤리적 쟁점에 대해 설명하시오.

예시 답안 데이터 소유권 문제는 데이터가 중요한 자원으로 인식되면서, 데이터를 생성한 주체와 이를 활용하는 기업 간에 소유권을 누구에게 부여할지에 대한 논의가 필요해졌다. 이 문제를 해결하기 위해서는 사회적 합의를 통해 데이터 소유권에 대한 기준을 명확히 설정해야 한다.

memo