



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

교육학 석사 학위논문

과학 데이터 기반  
인공지능(AI) · 고등학교 과학 융합  
교육 프로그램의 개발 및 적용  
- pH 예측을 중심으로 -

2023년 8월

서울대학교 대학원  
AI 융합교육학과 AI 융합교육 전공  
노 동 규

과학 데이터 기반  
인공지능(AI) · 고등학교 과학 융합  
교육 프로그램의 개발 및 적용  
- pH 예측을 중심으로 -

지도교수 정 대 홍

이 논문을 교육학 석사 학위논문으로 제출함  
2023년 8월

서울대학교 대학원  
AI 융합교육학과 AI 융합교육전공  
노 동 규

노동규의 석사 학위논문을 인준함  
2023년 8월

위 원 장           유 준 희           (인)

부위원장           조 정 효           (인)

위     원           정 대 홍           (인)

## 국문 초록(16pt)

4차 산업혁명 시대의 핵심 ICT 기술은 빅데이터(Big data), 인공지능(AI), 클라우드(Cloud)이다. 이중 인공지능 기술은 다양한 영역에서 변화를 가져오고 있으며, 교육 분야에도 중대한 영향을 미칠 것으로 예상되고 있다. 이에 4차 산업혁명 시대의 교육계의 화두는 인공지능 융합 교육이다.

인공지능 융합 교육의 선행 연구들은 2020년부터 급증하고 있는 상황이고, 학교 급간 별로는 초등학생이 가장 높은 비율을 차지했으며, 그중 고등학생을 대상으로 하는 선행 연구의 비율이 가장 낮았다. 따라서 고등학생을 대상으로 하는 인공지능 융합 교육 연구가 필요한 상황이다. 또한 대부분의 인공지능 융합 교육 실습 활동에서는 지도학습을 이용한 분류 모델만 다루고 있으므로 인공지능의 또 다른 영역인 회귀나 비지도 학습을 이용한 교육 프로그램의 개발이 필요한 상황이다.

2022개정 교육과정에서는 인공지능을 활용한 예측과 과학탐구, 사회문제 해결을 위한 인공지능 과학탐구(가칭)과 같은 융합선택 과목의 개발의 필요성이 있으나, 전문성을 갖춘 교원 수의 부족과 선행 연구된 교육 프로그램의 수가 극히 부족하여 융합선택과목에서 인공지능과 관련된 융합과목은 신설되지 못하고 있다.

제 4차 과학교육 종합 계획에서는 빅데이터를 이용한 과학탐구를 통해 과학적 문제 해결 과정에서 인공지능과 빅데이터를 활용한 디지털 도구를 활용하는 방안을 추진하고 있고, 전 세계적으로 인공지능을 위한 데이터셋을 제공하는 플랫폼이 많아지고 있어, 이러한 데이터셋은 현실 세계의 문제들을 해결하는 데 좋은 수업 소재가 될 가능성이 있다.

이에 본 연구에서는 공개된 과학 데이터셋을 기반으로 인공지능

의 기술 중 지도학습의 회귀(Regression) 모델 알고리즘을 적용하여 pH 예측을 목적으로 하는 인공지능 모델을 만드는 과정을 포함하는 과학 데이터 기반 인공지능·고등학교 과학 융합 프로그램을 개발하였다. 이 과정에서는 지식정보처리역량 함양을 위한 데이터 기반 과학 데이터 분석 탐구 모형을 적용하고, 각 과정에서 학생들의 데이터리터러시가 길러질 수 있도록 설계하였으며 총 6차시의 수업으로 개발하였다.

개발된 pH 예측을 위한 과학 데이터 기반 인공지능(AI)·고등학교 과학 융합 프로그램은 3인의 전문가와 3인의 현장교사에게 자문을 받아 내적 타당화 과정을 거쳐 타당성을 높였으며, 학생 27명을 대상으로 수업을 진행하여 사전 사후 설문을 대응표본  $t$ -검정으로 분석한 결과 데이터리터러시의 향상( $p < 0.01$ )을 볼 수 있었다.

본 연구에서는 과학 데이터기반 인공지능 모델을 만들고 이를 기반으로 학생들의 데이터리터러시 향상을 위한 과학 데이터기반 인공지능·고등학교 과학 융합 프로그램을 개발 및 적용한 데에 의미가 있다. 또한 과학 데이터 기반 분석 탐구모형(ESDA)에 맞춰 데이터 리터러시의 각 요소를 향상 시킬수 있는 프로그램이고, 클라우드 기반의 활동지가 제공됨으로써 교사 학생들 누구나 쉽게 접근이 가능한 인공지능 융합 교육 프로그램을 제시하여 다양한 과학 데이터를 이용하여 새로운 교육 프로그램의 개발에도 도움이 될 방식을 제안했다는 데 의미가 있다.

**주요어** : 인공지능 융합 교육 프로그램, ESDA 모형 적용 AI 융합 교육 프로그램, 과학 데이터 기반 교육 프로그램, 고등학교 과학 융합 프로그램, pH 예측 교육 프로그램

**학 번** : 2021-23621

# 목 차

I. 서론 .....	1
1. 연구의 필요성과 목적 .....	1
2. 연구 문제 .....	5
II. 이론적 배경 .....	6
1. 데이터 리터러시 .....	6
2. 탐색적 과학 데이터 분석 과학 탐구 모형 .....	15
3. 인공지능 · 과학 교과 융합 프로그램 .....	21
4. 지도 학습(Supervised Learning) .....	22
5. 머신러닝 플랫폼과 프로그래밍 언어 .....	24
가. Python .....	24
나. 싸이킷런(Scikit-learn) 라이브러리 .....	24
다. Colab .....	24
라. Lucifer-ML 패키지 .....	25
6. 교육과정과 데이터셋 .....	26
III. 연구 방법 .....	29
1. 연구 절차 .....	29
2. 연구 대상 .....	29
3. 연구 도구 .....	30
4. 자료 수집 .....	31
가. 선행 문헌 검토 .....	31
나. 내적 타당화 .....	32
다. 외적 타당화 .....	33

5. 자료 분석 .....	35
가. 전문가 타당화 자료 .....	35
나. 데이터 리터러시 향상도 측정 분석 .....	35
다. 외적 타당화 .....	36
IV. 연구 결과 .....	37
1. 과학 데이터기반 인공지능(AI) · 고등학교 과학 융합 교육 프로그램의 개발 .....	
가. 예측 모델의 개발 .....	37
나. 교육 프로그램의 개발 .....	40
2. 프로그램의 적용 효과 .....	45
가. 데이터 리터러시 사전, 사후 검사(외적 타당화) .....	45
나. 학습자 반응 .....	46
V. 결론 및 제언 .....	48
1. 결론 .....	48
2. 제언 .....	49
참고문헌 .....	51
부    록 .....	55
Abstract .....	69

## 표 목 차

[표 II-1] Definition of Data literacy by Researcher .....	7
[표 II-2] 데이터 리터러시의 정의와 구성 요소 .....	10
[표 II-3] 데이터 리터러시 구성요소 .....	10
[표 II-4] Matrix of different facets of data literacy and the features of understanding and competencies associated with different levels of data literacy for them .....	12
[표 II-5] 데이터 리터러시 향상을 위한 5단계 .....	13
[표 II-6] ESDA 탐구 모형 설계 원리 .....	16
[표 II-7] 인공지능(AI)· 과학 융합 프로그램 .....	21
[표 II-8] 지도학습에 주로 활용되는 알고리즘 .....	23
[표 II-9] Lucifer-ML 패키지의 데이터 처리 절차 .....	25
[표 II-10] 고등학교 AI 교육 주요 성취 기준 .....	26
[표 II-11] AI 교육용 데이터셋 제공 플랫폼과 라이브러리 ..	27
[표 II-12] pH recognition 데이터셋에 적용 가능한 2022개정 고등학교 과학과 성취기준 .....	28
[표 III-1] 연구 대상 .....	30
[표 III-2] 프로그램 구성 내용에 전문가 타당화 평가 문항 .....	31
[표 III-3] 전문가 타당화에 참여한 전문가 프로필 .....	32
[표 III-4] 데이터 리터러시 역량 측정 도구 설문지 문항 1차 ·	33
[표 III-5] 데이터 리터러시 역량 측정 도구 설문지 문항 2차 ·	34
[표 IV-1] ESDA 모형이 적용된 과학 데이터 기반 인공지능(AI) · 과학 융합 교육 프로그램 초안 .....	41
[표 IV-2] 전문가 타당화 검사 결과(1차) .....	42
[표 IV-3] 전문가 검토 의견과 수정 사항 .....	42



[표 IV-4] ESDA 모형이 적용된 과학 데이터 기반 인공지능(AI) · 고등학교 과학 융합 교육 프로그램 .....	43
[표 IV-5] 전문가 타당화 검사 결과(2차) .....	45
[표 IV-6] 대응표본 t-검증 검사 결과 .....	46
[표 IV-7] 학습자 의견 - 좋았던 점 .....	46

## 그 립 목 차

[그림 II-1] 데이터 리터러시 연구의 word2vec 분석 결과 .....	9
[그림 II-2] ESDA 탐구 모형 .....	15
[그림 III-1] 프로그램 개발 절차 .....	29
[그림 IV-1] pH-prediction 데이터셋 정보 .....	37
[그림 IV-2] pH-recognition 데이터셋으로 적합한 회귀(Regression) 모형을 찾는 Lucifer-ML 실행 코드 .....	38
[그림 IV-3] Lucifer-ML 분류 모델의 알고리즘 성능 결과 .....	39
[그림 IV-4] Scikit-Learn 라이브러리의 Catboost Regressor 알고리즘을 이용한 머신러닝 pH 예측 모델의 개발 코드 .....	40
[그림 IV-5] Colab용 텍스트 코딩 활동지 QR코드 .....	39

# I. 서론

## 1. 연구의 필요성과 목적

4차 산업혁명은 스위스 다보스에서 열린 세계경제포럼(WEF, 2016)에서 처음 언급한 개념으로 디지털 혁명 또는 지식정보혁명으로 정의되는 3차 산업혁명을 기반으로 수학, 물리학, 생물학 등의 기초과학과 정보통신기술(ICT) 융합으로 이루어지는 지식혁명이다. 4차 산업혁명의 핵심 ICT 기술은 빅데이터(Big Data), 인공지능(AI), 클라우드(Cloud)이다. 이 중 인공지능 기술은 영상이나 음성 등의 디지털 신호를 분석하여 물체를 인식하고, 말과 문자로 사람과 소통하며, 미래에 일어날 일을 예측하거나 의사 결정을 내리고, 데이터로부터 새로운 정보와 지식을 학습하는 기술을 일컫는다(김진형, 2020).

인공지능 기술은 영상이나 음성을 전 세계의 정치·경제·사회·문화 등의 다양한 영역에서 변화를 가져오고 있으며, 교육 분야에도 중대한 영향을 미칠 것으로 예상되고 있다(한송이, 2022). 4차 산업혁명 시대의 교육계의 화두는 인공지능 융합 교육이다. 인공지능 융합 교육을 위해서 인공지능의 기초와 원리를 직접적으로 교육해야 하는 교과인 정보 교과가 2009 개정 교육 과정에서는 고등학교에서 생활·교양 영역, 기술·가정 교과(군)의 심화 선택 과목으로 포함되어 있었으나 2015 개정 교육과정에서 일반 선택 과목으로 전환되는 등의 인공지능과 관련된 교과목의 선택을 강화하는 방향으로 정책이 추진되었고, 전문 교과 교과목으로 <정보 과학>, 진로 선택 교과목으로 <인공지능 기초>, <인공지능 수학>이 신설되어 현장에서 적용하게 되었다.

이러한 인공지능 과목들의 바탕이 되는 인공지능 교육에 관련된 선행

연구들은 2017년도부터 시작되기 시작하여 2018년도에 잠시 주춤하였으나, 2019년도에 다시 활성화되기 시작해 2020년에 관련 연구 수가 급증하고 있는 상황이고, 학교급 별로는 초등학생이 48.5%로 가장 높은 비율을 차지했으며, 대학생과 중학생(18.2%), 일반(12.1%), 고등학생(3.0%)으로 고등학생을 대상으로 하는 교육 연구가 필요한 상황이다. 또한 대부분의 인공지능 교육 실습 활동에서는 지도학습(Supervised Learning)을 이용한 분류(Classification) 모델만 다루고 있으므로 인공지능의 또 다른 영역인 회귀(Regression)나 비지도학습(Unsupervised learning)을 이용한 교육 프로그램의 개발이 필요한 상황이다(한정윤, 허선영, 2021).

2019년 한국과학창의재단과 4차 산업혁명과 미래교육포럼과 서울시 교육청이 공동 주최한 AI 융합 교육 컨퍼런스에서는 한국 교육학회, 정보과학교육연합회, 한국과학기술단체총연합회, 한국과학기술한림원 등이 인공지능 융합 교육의 시작을 알리는 공동 선언문을 작성하여 배포하였다. 인공지능은 기계, 인간, 환경을 지능적으로 만드는 방법론이고, 인공지능의 발전은 우리 삶의 모습과 사회 전 부문의 운영방식을 획기적으로 변화시킬 수 있다고 하였다. 따라서 인공지능에 친숙해지는 것은 인공지능을 연구하고 개발하는 소수의 사람에게만 요구되는 것이 아니라 모두에게 필요한 일이 되어 있다고 하였다. 인공지능 시대를 살아갈 모든 학생들이 미래 사회에 적응할 수 있게 하기 위해서 인공지능은 교육 현장에서 효과적으로 활용될 수 있어야 하고, 수학, 정보, 과학교육의 재구조화가 필요하다고 하였다. 또한 모든 학생이 손에 인공지능을 익히기 위해서 인공지능의 구구단에 해당하는 코딩능력, 데이터에 기반한 의사결정 알고리즘 설계 및 활용 학습의 기회를 가져야 하며, 국가 교육과정에 반영해야 한다고 하였다(한국과학창의재단, 2019).

2022 개정 교육과정 시안에서는 융합선택 과목 체계가 신설되었으나 융합선택 과목으로는 수학과에서는 <수학과 문화>, <실용 통계>, <수

학과제 탐구>, 정보과에서는 <소프트웨어와 생활>, 과학과에서는 <과학의 역사와 문화>, <융합과학 탐구>, <기후변화와 환경생태>의 과목이 신설될 예정이지만 융합선택 과목에 인공지능이 융합되는 과목은 제시되지 못하고 있다. STEAM 교육을 필두로 다양한 융합 교육이 교육계 전반으로 이어지고 있는 상황에서 교사나 학교에서는 학교에서 활용할 수 있는 다양한 연계·융합 프로그램과 교수·학습 자료를 요구하고 있는 반면에, 교육부나 한국과학창의재단 등에서 개발하고 있는 프로그램이나 교수·학습 자료는 활용이 어려운 상황이다(권점례 외, 2019). 또한 인공지능을 활용한 예측과 과학탐구, 사회문제해결을 위한 인공지능과학탐구(가칭)와 같은 융합선택과목 개발의 필요성이 있으나, 전문성을 갖춘 교원이 충분하게 양성되지 못하고 있고, 선행 연구된 프로그램의 수가 극히 부족하여 융합선택과목에서 인공지능과 관련된 융합과목은 신설되지 못하고 있다(곽영순, 2021).

교육부와 한국과학창의재단에서 제시하고 있는 인공지능 원리 교육의 정의 중에는 ‘데이터에 대한 이해’가 포함되어 있으며, 인공지능 융합 교육의 정의에는 ‘인공지능에 대한 원리와 핵심 개념 이해를 기반으로 융합 태도를 함양하는 교육’이 포함되어 있다. 따라서 인공지능 융합 교육을 위해서는 데이터 이해, 수집, 분석, 처리 등 데이터 리터러시 교육이 중요하다(김준영, 한선관, 2022).

제 4차 과학교육 종합계획(2020~2024)에서는 빅데이터를 이용한 과학탐구를 통해 과학적 문제 해결 과정에서 인공지능과 빅데이터를 활용한 디지털 도구를 활용하는 방안을 추진하고 있다(교육부, 2020). 인공지능에 대한 중요성이 강조되면서 전세계적으로 인공지능을 위한 데이터셋을 제공하는 플랫폼이 많아지고 있으며, 이러한 데이터셋은 현실 세계의 문제들을 해결하는 데 좋은 수업 소재가 될 가능성이 있다(김슬기 외, 2021).

이에 본 연구에서는 만능 지시약의 색 변화 이미지로부터 데이터를 얻는 과정을 수행하고, 공개된 데이터셋으로부터 인공지능 기술 중 지도 학습의 회귀 모델을 만들어 pH 예측을 목적으로 하는 인공지능·고등학교 과학 융합 프로그램을 개발하여 학생들의 데이터 리터러시 역량에 미치는 효과를 탐색해 보고자 한다. 개발된 교육 프로그램은 학교 현장의 컴퓨팅 환경을 고려하여 프로그램을 설치하지 않고도 python 스크립트를 입력하여 데이터를 처리할 수 있고 인공지능 모델을 구동시킬 수 있는 구글 Colaboratory(이하 Colab)과 같은 클라우드 컴퓨팅 기술을 이용하였다. 이를 통해 학교 현장에서 인공지능 융합 교육을 위한 빅데이터 기반의 교육 프로그램들이 개발되어 인공지능과 관련된 융합과목의 신설로 이어지는데 기여하고자 하였다.

## 2. 연구 문제

본 연구의 구체적인 연구 문제는 다음과 같다.

첫째. 과학 데이터 기반 인공지능(AI)·고등학교 과학 융합 프로그램은 어떻게 구성되는가?

둘째. 과학 데이터 기반 인공지능(AI)·고등학교 과학 융합 프로그램은 데이터 리터러시 향상에 효과적인가?

## II. 이론적 배경

### 1. 데이터 리터러시와 교육

#### 가. 데이터 리터러시(Data Literacy)

데이터 리터러시는 관찰, 측정을 통해 수집된 데이터(data)와 읽고 쓰고 해석할 수 있는 문해 능력을 뜻하는(literacy)가 합쳐진 단어로 최근 강조되고 있는 AI 기술과 빅데이터에 대한 역량의 함양이 강조되면서 주목받고 있다(배화순, 2019). 컴퓨터 교육에서 데이터 리터러시는 데이터를 수집, 분석, 표현하는 방법을 이해하고, 정확하고 효과적인 실제 데이터를 활용하여 문제를 해결할 수 있는 능력(이승철 외, 2019) 또는 데이터 수집/준비, 데이터 분석, 데이터 평가, 데이터 시각화 및 표현, 데이터 기반 의사 소통(송유경, 2021)으로 나뉘기도 하며, 데이터 이해, 데이터 수집, 데이터 해석 및 평가, 데이터 관리, 데이터 활용(Prado, Marzal, 2013)으로 구성하여 수업 모형이나 프로그램에서 길러야 할 역량으로 제시하고 있다.

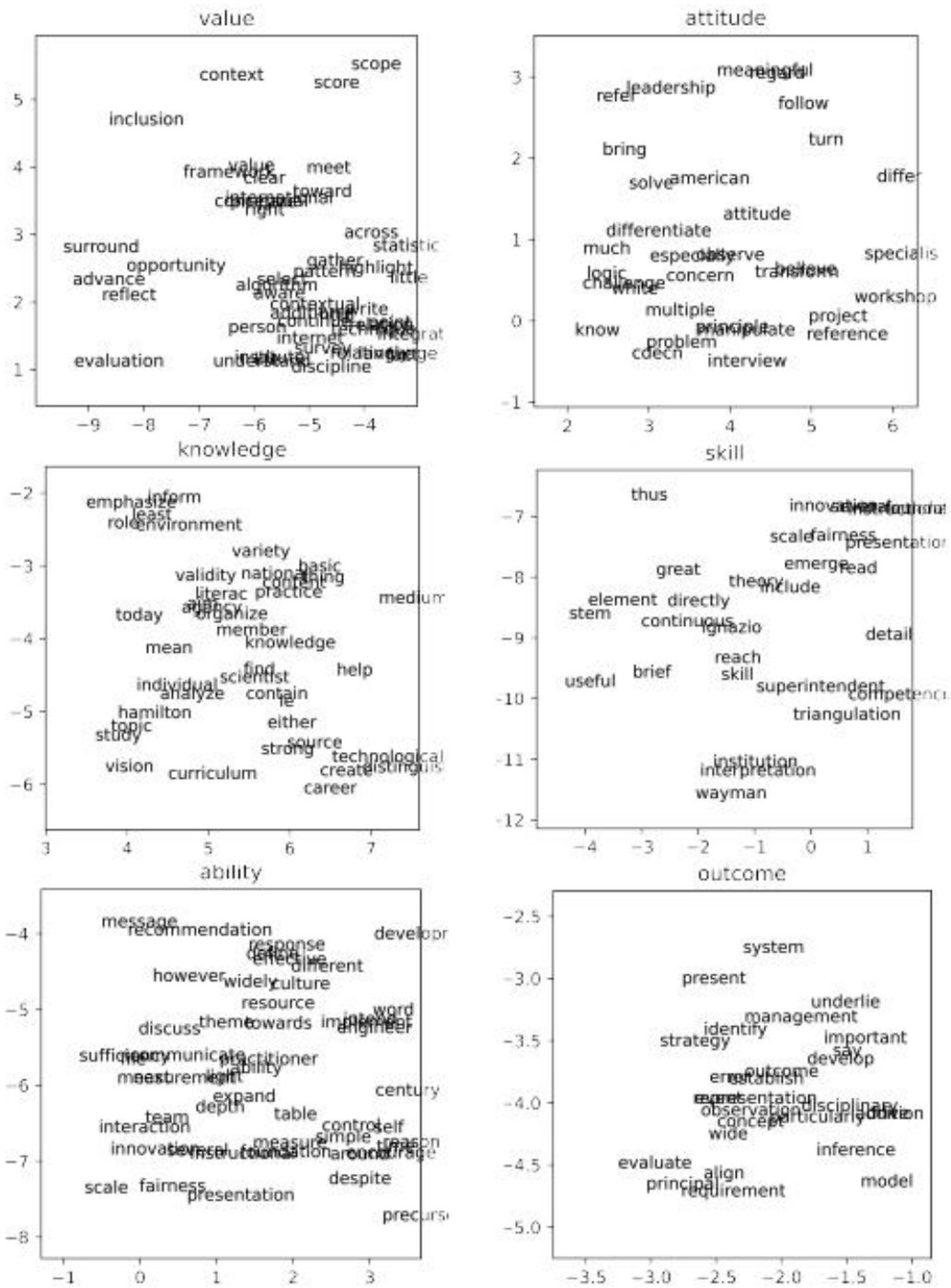
해외의 연구 사례에서는 데이터 리터러시에 대한 논의가 일찍 시작되어 다양한 정의들을 확인할 수 있다. 주요 연구자들의 데이터 리터러시의 정의는 <표 II-1>과 같이 주로 해석, 활용, 이해, 의사소통에 초점을 맞춰서 정의하고 있지만, 구체적인 정의나 개념의 접근 방법 및 범위 등이 연구자마다 다양하다(김슬기, 김태영, 2021).

<표 II-1> Definition of Data literacy by Researcher

Researcher	Definition of Data literacy
Shields, M. (2005)	정보 리터러시와 통계 리터러시와 상호 보완적인 관계에 있는 데이터에 접근하고, 조작하고, 요약하는 능력
Qin, J. et. al (2010)	복잡하고 다양한 교무와 수준의 데이터를 측정하고 추론하는 능력
Carlson, J. et. al (2011)	그래프와 차트를 올바르게 읽고, 데이터에서 올바른 결론을 도출하여 데이터를 이해하는 능력
Mandinach, E. B. et. al (2013)	의사 결정에서 필요한 데이터를 이해하고 효과적으로 사용할 수 있는 능력
Prado, J. C. et. al (2013)	검색, 비판적 평가, 관리, 분석 및 종합을 포함하여 정보 리터러시의 하위 구성 요소로서 정보 요구를 정확하게 정의할 수 있는 능력
Deahl, E. (2014)	정량적 및 정성적 데이터를 사용하여 주장을 이해하고, 찾고, 수집하고, 해석하고, 시각화하고, 지원하는 능력
Bhargava, R. et. al (2015)	데이터를 읽고, 작업하고, 분석하고, 논증할 수 있는 능력
D'Ignazio, C. et. al (2015)	데이터를 읽고, 분석하고, 활용하여 논증하고, 나아가 빅데이터를 식별, 이해, 평가할 수 있는 능력
Frank, M. et. al (2016)	데이터를 이해하고 효과적으로 사용하는 능력
김슬기 외 (2021)	문제를 해결하기 위해서 데이터를 수집하고 분석 및 활용하여 정보로 처리하는 지식 구성과 의사소통의 기초 능력
송유경 (2021)	데이터로부터 의미 있는 정보를 추출해내고 실생활의 다양한 문제를 해결하기 위해 데이터를 활용하며 적절한 도구를 활용해 데이터를 분석하고 결론을 도출해 내는 능력뿐만 아니라, 더 나아가 데이터를 활용하여 타인과 효과적인 의사소통을 하는 능력



김슬기와 김태영(2021)은 데이터 리터러시 관련 주요 연구의 정의와 구성 요소에 활용된 단어 빈도 분석과 word2vec 딥러닝 자연어 처리 방법을 통해 [그림 II-1]와 같은 분석을 하였고, 데이터 리터러시를 ‘문제를 해결하기 위해서 데이터를 수집하고 분석 및 활용하여 정보로 처리하는 지식 구성과 의사소통의 기초 능력’으로 정의하고 데이터의 수집, 분석, 활용에 주목하고, 데이터를 다방면의 문제 해결에 사용할 수 있는 기능과 현재의 데이터로부터 미래를 예측하고 새로운 지식을 구성할 수 있으며, 구성된 지식을 바탕으로 의사 결정과 소통을 할 수 있는 능력으로 정리하였고, 구체적인 구성 요소를 지식, 기능, 가치와 태도의 측면으로 하여 <표 II-2>와 같이 제안하였다.



[그림 II-1] 데이터 리터러시 연구의 word2vec 분석 결과(김슬기, 2021)

<표 II-2> 데이터 리터러시의 정의와 구성 요소(김슬기, 2021)

Definition		
Basic ability of knowledge construction and communication to collect, analyze, and use data and process it as information for problem solving		
Knowledge	Skills	Values and Attitudes
<ul style="list-style-type: none"> <li>· Concept of Data</li> <li>· Role of Data</li> <li>· Type of Data</li> <li>· Structure of Data</li> <li>· Source of Data</li> <li>· Analysis Method of Data</li> <li>· Data Visualization</li> <li>· Data Pre-processing</li> </ul>	<ul style="list-style-type: none"> <li>· Collect</li> <li>· Interpret</li> <li>· Reason</li> <li>· Summarize</li> <li>· Presentation.</li> <li>· Manage</li> <li>· Process</li> <li>· Communicate</li> </ul>	<ul style="list-style-type: none"> <li>· Relationship</li> <li>· integration</li> <li>· Continue</li> <li>· Human-centered</li> <li>· Fairness</li> </ul>

또한 송유경(2021)의 연구에서는 데이터 리터러시의 구성 요소를 크게 ‘데이터의 통계적 분석’과 ‘데이터를 활용한 의사 소통’으로 나누고 데이터의 통계적 분석 부분을 구성하는 역량은 ‘데이터 이해’, ‘데이터 수집/준비’, ‘데이터 분석’, ‘데이터 평가’로 구분하였고, 데이터를 활용한 의사소통은 ‘데이터 시각화 및 표현’, ‘데이터 기반 의사 소통’으로 <표 II-3>과 같이 나누고 있다.

<표 II-3> 데이터 리터러시 구성요소(송유경, 2021)

주요 영역	세부 항목	관련 선행연구
데이터의 통계적 분석	데이터 이해	이승철과 김태영(2019)
	데이터 수집/준비	배화순(2019)
	데이터 분석	Borner 외(2016), Carlson 외(2011), Gray 외(2018)
	데이터 평가	Prado와 Marzal(2013)
데이터를 활용한 의사소통	데이터 시각화 및 표현	한상우(2018), Borner 외(2016)
	데이터 기반 의사소통	한상우(2018)

최근의 국내 연구들은 데이터 리터러시에 국외 연구자들의 선행 연구에 문제 해결을 위한 부분을 포함시키면서 데이터 리터러시 향상을 위한 교육에서 문제 해결 능력을 향상시키기 위한 의사 소통이나 협업 능력을 데이터 리터러시 역량의 구성 요소로 요구하는 추세이다. 이는 융합 교육으로 제안되는 STEAM 교육프로그램에서 우리 주변의 문제를 해결하는 아이디어를 산출하는 교수 학습 방법이 포함되어 있는(홍석영 외, 2020)의 영향이 있다. 이러한 데이터 리터러시의 확장된 정의를 실현하기 위해서는 문제의 발견이나 인식 단계 등이 필요하나 이는 고차원적인 학업 능력이므로 교육 프로그램이 적용되는 프로그램에 이러한 단계를 포함시키거나, 학습자의 레벨에 따라 필요한 데이터 리터러시의 역량이 다르므로 이를 학습자의 단계별에 맞춰 프로그램의 구성 요소를 달리할 필요성이 있다.

Gibson et al(2018)은 생명과학을 위한 데이터 리터러시의 향상을 위해 데이터 리터러시의 세부 항목을 수집과 기록(Collection and recording), 계산(Calculation), 분석 및 해석(Analysis and interpretation), 소통(Communication)으로 나누고 각 데이터 리터러시의 다양한 측면과 이와 관련된 이해 및 역량의 특징에 대하여 <표 II-4>와 같이 제안하였다. 이 중 고급(advanced) 데이터 리터러시 측면에서 계산의 단계에서는 생명 시스템의 연구에서 적절한 수학적 틀이 선택되는가를 알고, 적절한 생명과학적 모델이 데이터로부터 만들어지는지를 이해하는 단계가 포함되어 있다. 이와 같이 과학 교육에서도 데이터를 통해서 모델을 만들고 이를 이해하는 과정이 필요한 것으로부터 데이터 리터러시 역량을 기르기 위해서는 인공지능에서 데이터 학습을 통해서 모델링을 하는 단계를 고급 데이터 리터러시 프로그램의 적용 과정에서 포함시켜야 할 필요성이 있다.

<表 II-4> Matrix of different facets of data literacy and the features of understanding and competencies associated with different levels of data literacy for them (Gibson et al(2018))

Facets of data literacy	Basic data literacy	Intermediate data literacy	Advanced data literacy
Collection and recording	<ul style="list-style-type: none"> <li>• Know how to use instrumentation and technology to collect and store data</li> <li>• Able to collect and record data accurately using technology</li> </ul>	<ul style="list-style-type: none"> <li>• Able to identify appropriate data to collect relative to a biological question and hypothesis</li> <li>• Know how to enter data into a spreadsheet or database</li> </ul>	<ul style="list-style-type: none"> <li>• Understand how to incorporate rigorous data collection and sampling methods into experimental design</li> <li>• Know how to store, manage, manipulate, or query a database</li> </ul>
Calculation	<ul style="list-style-type: none"> <li>• Know how to conduct mathematical calculations</li> <li>• Able to use spreadsheets and software to conduct calculations</li> </ul>	<ul style="list-style-type: none"> <li>• Understand the relationship between calculations and a biological question</li> <li>• Know how to apply mathematical tools and technology to conduct calculations with biological data</li> </ul>	<ul style="list-style-type: none"> <li>• Know how to choose appropriate mathematical tools for studies of biological systems</li> <li>• Understand how data are used to develop quantitative biological models</li> </ul>
Analysis and interpretation	<ul style="list-style-type: none"> <li>• Know how to describe data with statistics</li> <li>• Able to describe patterns in data</li> </ul>	<ul style="list-style-type: none"> <li>• Know how to analyze and interpret data using statistical tests</li> <li>• Be able to interpret results of statistical test relative to a biological question or hypothesis</li> </ul>	<ul style="list-style-type: none"> <li>• Understand how to incorporate data analysis and statistical methods into experimental design</li> <li>• Understand assumptions in analyses</li> <li>• Able to compare results among analyses</li> </ul>
Communication	<ul style="list-style-type: none"> <li>• Know how to use technology to construct tables and figures.</li> <li>• Able to describe graphical and tabular presentations of data</li> </ul>	<ul style="list-style-type: none"> <li>• Able to explain the relationships among data presented in figures and tables</li> <li>• Understand how to use data and analyses to argue from evidence</li> </ul>	<ul style="list-style-type: none"> <li>• Able to evaluate the strengths and limitations of their data and analyses</li> <li>• Understand the relationship between their data and other biological, scientific or societal issues</li> </ul>

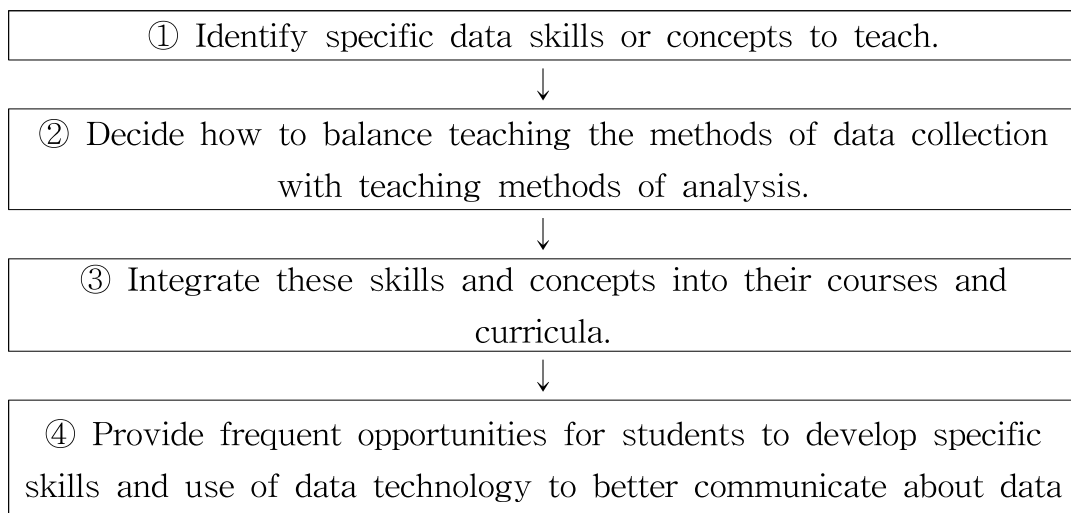
이와 같이 고급 데이터 리터러시 역량의 향상을 위해서는 모델링이 완료된 인공지능을 사용하는 것은 인공지능을 체험하는 수준에 그칠 수 있고, 초등학교 인공지능 융합 교육에서 많이 사용되고 있으므로 데이터를 이용한 인공지능 모델을 직접 만들어보고 이를 통해 데이터 리터러시의 향상을 볼 수 있는 고등학생을 위한 교육 프로그램의 개발이 필요하다.

인공지능 융합 교육을 위해서는 데이터 리터러시의 구성 요소들에 해당하는 역량의 향상이 기본이 되어야 하므로 교수학습에 활용되는 데이터셋부터 교육 내용, 평가 자료를 포함하는 패키지 형태의 프로그램이 필요하다(김슬기, 2021). 이에 본 연구에서는 고등학교 과학 교과인 물리학과 화학에서 활용할 수 있는 데이터셋을 이용하여 패키지 형태의 AI 고등학교 과학 융합 프로그램을 개발하고 학생들에게 실제로 적용하여 데이터 리터러시의 향상을 평가해 보고자 한다.

#### 나. 과학 교육에서 데이터 리터러시

과학 교육에서 데이터 리터러시를 기르기 위해 교수자는 <표 II-5>와 같이 5가지 단계를 통할 수 있다(Gibson et al, 2018).

#### <표 II-5> 데이터 리터러시 향상을 위한 5단계(Gibson et al, 2018)



in science.



⑤ Provide multiple opportunities for students to use technology to produce graphics that effectively communicate the information in the data.

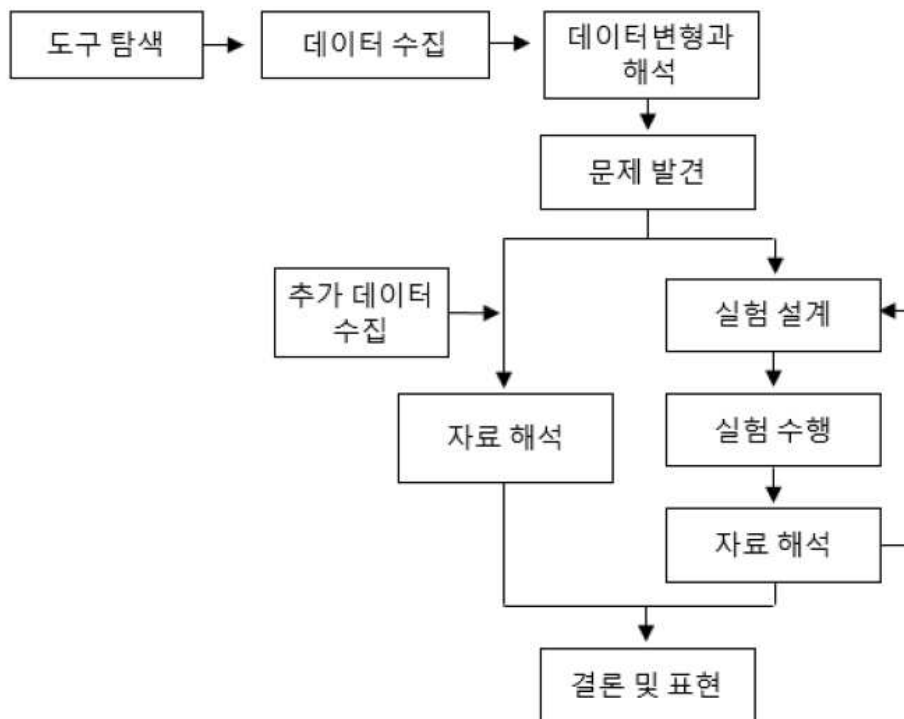
첫 번째 단계는 가르칠 특정 데이터 기술 및 개념을 파악하는 단계이다. 이 단계에서는 교수자가 학생들에게 안내할 데이터의 특성을 파악해야 하는 단계이다. 두 번째 단계는 데이터를 수집하는 방법과 분석 방법의 균형을 결정하는 단계이다. 이 단계에서는 학생들에게 컴퓨터를 활용하여 과학 데이터를 수집하거나 분석하는 방법을 어느 수준까지 할 것인가를 교수자가 결정해야 하는 단계이다. 세 번째 단계에서는 기술과 개념을 코스와 커리큘럼에 통합하는 단계이다. 데이터의 수집과 분석 기술을 어떤 교육과정에 통합하여 사용할지 교수자가 결정해야 하는 단계이다. 네 번째 단계는 과학에서 데이터에 대해 더 잘 소통할 수 있도록 데이터를 활용하는 기술을 개발할 수 있는 기회를 제공하는 단계이다. 학생들이 직접 다양한 과학 데이터를 수집하고 이를 활용하는 기술을 사용해보면서 데이터 리터러시의 향상이 이루어질 수 있는 단계이다. 다섯 번째 단계는 학생들이 기술을 사용하여 데이터의 정보를 효과적으로 전달할 수 있는 그래픽을 제공할 수 있는 다양한 기회를 제공해야 하는 단계이다. 정보의 효과적인 전달은 데이터를 이용한 의사 소통 능력의 향상에 매우 중요하므로 이러한 단계를 교수자는 기획하여 학생들의 역량 향상을 고려해야 하는 단계이다.

따라서 교사가 직접 교육과정에 활용할 수 있는 과학 데이터를 통해 인공지능 모델을 만들고, 데이터를 시각화 하고, 문제 해결이나 의사 소통 기회를 제공함으로써 학생들에게 데이터 리터러시의 향상이 이루어지게 할 수 있는 교육 프로그램의 필요성이 있다.

## 2. 탐색적 과학 데이터 분석 과학 탐구 모형(ESDA 탐구 모형)

과학 교과에서는 교사와 학생이 함께 활동하는 구성주의적 탐구를 강조하고(Arends, 1998), 이러한 탐구의 과정에서 데이터를 활용하여 과학 데이터 소양(Science Data Literacy)을 기를 수 있는 데이터 수집, 선별, 처리 활용, 평가를 할 수 있는 역량 교육 등이 강조되고 있다(Qin, J외, 2010).

이러한 과학 교육의 방향에 데이터를 기반으로 중고등학생 대상으로 하는 교육에 활용할 수 있는 과학 탐구 모형과 수업 전략이 제시된 지식 정보처리역량 함양을 위한 데이터 기반의 과학 탐구 모형(Exploratory Scientific Data Analysis Inquiry Model, 이하 ESDA)이 개발(손미현, 2020)되어 과학 교육에서 지식정보 처리 역량을 기르는 노력이 이어지고 있다([그림 II-2]).



[그림 II-2] ESDA 탐구 모형(손미현, 2020)



ESDA 탐구모형에서는 도구 탐색의 원리, 실생활 데이터 수집의 원리, 데이터 변형의 원리, 데이터 해석의 원리, 문제 구체화의 원리, 문제 해결의 원리, 표현과 공유의 원리와 같이 7개의 교수 전략이 포함되어 있는데 이는 <표 II-6>과 같다.

<표 II-6> ESDA 탐구 모형 설계 원리(손미현, 2020)

도구 탐색의 원리	
교수전략	<ul style="list-style-type: none"> <li>• 도구에서 측정할 수 있는 변인(센서)의 수를 중학생은 2-3개, 고등학생 이상은 그보다 복잡한 형태의 데이터를 활용할 수 있게 한다.</li> </ul>
	<ul style="list-style-type: none"> <li>• 익숙하고 간단한 형태로 데이터가 산출되는 도구를 제시한다.</li> </ul>
	<p>(예) 엑셀이나 텍스트 파일 형태</p>
	<ul style="list-style-type: none"> <li>• 교사가 도구를 제시할 경우는 도구의 기본 사용법은 교사가 지도하고, 학생이 도구를 선택한 경우는 학생 스스로 인터넷 검색, 사용설명서 등을 통해 익히도록 한다.</li> <li>• 필요한 도구를 제작할 경우는 코딩, 3D 프린터 이용법 등의 제작 방법을 가르친다.</li> </ul>
환경구성	<ul style="list-style-type: none"> <li>• 학생들이 원활하게 정보를 검색하며 탐구를 진행할 수 있는지 컴퓨터와 인터넷 여부를 확인한다.</li> </ul>
실생활 데이터 수집의 원리	
교수전략	<ul style="list-style-type: none"> <li>• 학생들이 측정하고 싶은 장소를 선택하게 한다.</li> </ul>
	<ul style="list-style-type: none"> <li>• 수집되는 데이터의 측정 간격이나 측정 장소의 위치, 환경 등을 자세히 기록하게 한다.</li> </ul>
	<ul style="list-style-type: none"> <li>• 교사는 주제에 관련된 기초 과학지식을 학생들이 익힐 수 있도록 지도한다.</li> <li>• 데이터베이스의 사용법을 익히게 한다.</li> </ul>
	<ul style="list-style-type: none"> <li>• 센서를 설치하는 곳에 인터넷 여부를 확인한다.</li> </ul>
환경구성	<ul style="list-style-type: none"> <li>• 데이터베이스를 사용하기 위한 사전 준비사항 등을 미리 확인한다.</li> </ul>
데이터 변형의 원리	

교수전략	<ul style="list-style-type: none"> <li>수집된 데이터가 정확한지, 연속적으로 누적되어 있는지 확인하도록 한다.</li> <li>데이터를 그래프 또는 도표로 변형할 수 있는 스프레드시트 프로그램 또는 데이터 모델링 프로그램에서 필요한 기능을 위주로 사용법을 가르친다.</li> </ul> <p>(예) 엑셀, 지오지브라</p>
환경구성	<ul style="list-style-type: none"> <li>다양한 그래프의 쓰임새와 예시를 제시한다.</li> <li>많은 양의 데이터를 변형하고 분석할 수 있을 만큼 성능이 좋은 컴퓨터를 준비한다.</li> <li>컴퓨터는 1~2인당 1대씩 준비하여 학생들이 직접 실습할 수 있도록 한다.</li> </ul>
<b>데이터 해석의 원리</b>	
교수전략	<ul style="list-style-type: none"> <li>도표나 그래프를 이용하여 상관 관계에 대한 내용을 설명할 수 있게 한다.</li> <li>변인 사이의 관계를 서술할 때 기존 학습된 지식, 인터넷으로 검색한 과학지식 등과 연관 지어 설명할 수 있게 한다.</li> <li>정보원에 대한 교육, 정보 검색의 원리 교육, 정보 검색의 전략 등을 미리 학습시킨다.</li> <li>실제 데이터를 이용한 통계, 그래프 등을 이용하여 데이터에서 의미를 찾아내는 활동을 연습하게 한다.</li> </ul> <p>(예) trends.google.co.kr, www.gapminder.org/tools/ 등의 사이트를 이용하여 데이터 해석 및 추론을 연습할 수 있도록 한다.</p>
환경구성	<ul style="list-style-type: none"> <li>학생들이 원활하게 정보를 검색하며 탐구를 진행할 수 있는지 컴퓨터와 인터넷 여부를 확인한다.</li> </ul>
<b>문제 구체화의 원리</b>	
교수전략	<ul style="list-style-type: none"> <li>데이터 해석 과정에서 발생한 질문 중 인터넷 검색, 자료 조사 등을 통해 답을 찾을 수 있는 문항은 문제 선정에서 제외한다.</li> <li>실험이나 활동을 통해 문제를 해결할 수 있도록 구체적으로 문제를 서술하게 한다.</li> <li>데이터를 해석한 내용과 적힌 질문을 구체적으로 서술하게 한다.</li> </ul>

	(예) 정확하게 어떤 점이 궁금한가? 왜 궁금증이 생겼는가?
환경구성	<ul style="list-style-type: none"> <li>• 학생들이 원활하게 정보를 검색하며 탐구를 진행할 수 있는지 컴퓨터와 인터넷 여부를 확인한다.</li> </ul>
<b>문제 해결의 원리</b>	
	<ul style="list-style-type: none"> <li>• 데이터 분석 과정에서 찾아낸 결과가 다른 상황에서도 적합한 지 추가적인 데이터를 수집하게 한다.</li> </ul>
교수전략	<ul style="list-style-type: none"> <li>• 데이터를 수집할 때는 탐구 문제를 고려하여 측정 간격과 기간을 정할 수 있도록 지도한다.</li> <li>• 통제 변인과 조작 변인을 고려하여 데이터를 수집한다.</li> <li>• 수집한 데이터의 결과와 과학적 지식을 연계하여 해석할 수 있도록 한다.</li> <li>• 충분한 논의시간을 확보하여 학생들이 깊이 있는 논의를 진행할 수 있게 한다.</li> </ul>
환경구성	<ul style="list-style-type: none"> <li>• 학생들이 원활하게 정보를 검색하며 탐구를 진행할 수 있는지 컴퓨터와 인터넷 여부를 확인한다.</li> <li>• 데이터 수집 단계의 환경 조건과 동일</li> </ul>
<b>표현과 공유의 원리</b>	
	<ul style="list-style-type: none"> <li>• 인터넷 검색 또는 탐구 결과 새롭게 알게 된 사실이나 지식을 웹문서로 기록하고, 쉽게 공유하게 한다.</li> <li>• 표현 수단이 되는 웹문서나 소프트웨어는 정해진 프레임이 있어 사용 방법이 간편하고, 다양한 형태의 파일을 포함할 수 있는 것을 활용한다.</li> </ul>
교수전략	<p>(예)sites.google.co.kr 이나 망고보드, x-mind zen 등의 소프트웨어를 이용한다.</p> <ul style="list-style-type: none"> <li>• 효과적으로 정보를 전달할 수 있도록 다양한 형태의 자료를 활용할 수 있도록 안내한다.</li> <li>• 효과적인 표현법의 전략을 익히고, 모방을 통해 연습하게 한다.</li> </ul>
환경구성	<ul style="list-style-type: none"> <li>• 표현 수단이 되는 웹문서나 소프트웨어를 미리 설치하거나 계정을 만들도록 한다.</li> <li>• 사용하는 컴퓨터의 사양에서 활용 가능한 형태의 소프트웨어인지 확인하도록 한다.</li> </ul>

도구 탐색의 원리에서 고등학생 이상은 3개 이상의 복잡한 형태의 데

이터를 활용할 수 있어야 하고, 익숙하고 간단한 형태로 데이터가 산출되는 도구를 사용하며, 교사가 직접 도구 사용을 지도해야 하며, 필요한 도구를 제작할 경우 코딩을 직접 가르쳐야 할 필요성이 있으므로 본 연구에서는 3개 이상의 과학 데이터를 이용하여, 학생들에게 파이썬 프로그래밍 언어를 활용할 수 있도록 구글 Colab 온라인 활동지를 통하여 텍스트 코딩을 도구로 활용하면서 사용법을 가르치는 과정을 포함하는 프로그램을 개발하고자 한다.

실생활 데이터 수집의 원리에서는 학생들이 측정하고 싶은 장소나 환경을 선택하게 하고, 측정 관련 데이터들을 자세하게 기록하게 해야 하므로 본 연구에서는 학생들이 측정하고 싶은 데이터의 측정 대상들을 정하고 이를 공유하면서, 수집된 데이터를 정리하여 공유하는 스프레드시트 문서를 만들어 제공하고자 한다.

데이터 변형의 원리에서는 수집된 데이터의 정확도와 누적 여부를 확인하고, 데이터 모델링 프로그램에서 필요한 기능을 위주로 가르쳐야 하는 교수 전략이 포함되어 있으므로 이에 본 연구에서는 수집된 pH 데이터를 정리하고 누적적으로 기록하게 하며, 알려진 만능 지시약의 pH를 기반으로 데이터의 정확성을 확인하는 과정을 내용으로 포함하고자 한다.

데이터 해석의 원리에서는 도표나 그래프를 이용하여 상관관계에 있는 내용을 설명할 수 있게 하여야 하므로 파이썬 라이브러리를 통하여 실제로 수집된 과학 데이터들의 상관 관계를 한눈에 시각화해주는 내용을 코딩 프로그램에 포함시켜 활용하고자 한다. 다만 본 연구에서 개발된 프로그램은 기존 연구나 분석이 없는 내용이므로 인터넷으로 검색한 지식을 활용할 수는 없으므로 상호간의 토의를 통하여 이를 보완하고자 한다.

문제 구체화의 원리에서는 실험이나 활동을 통해 문제를 구체적으로

서술하게 해야 한다. 따라서 본 연구에서는 준비된 다양한 음료들에 넣은 만능 지시약의 색 변화가 유사한 경우에는 구체적으로 pH를 구할 수 있는 방법에 대한 탐구가 필요하다는 문제 구체화 과정을 거치게 하는 내용을 포함하고자 한다.

문제 해결의 원리에서는 데이터 분석 과정에서 찾아낸 결과를 다른 상황에서도 적합한지 데이터를 추가적으로 수집하는 과정과 수집한 데이터의 결과와 과학적 지식을 연계하는 해석 과정을 포함한다. 이에 본 연구에서는 머신러닝으로 구현한 모델을 새로 수집된 데이터를 해석하는데 사용하여 모델이 예측한 결과와 기존 지식과의 연결을 이어가는 과정을 포함하고자 한다.

표현과 공유의 원리에서는 탐구 결과 새롭게 알게된 사실이나 지식을 웹문서로 기록하여 쉽게 공유하게 한다. 이에 본 연구에서는 학생들이 머신러닝으로 구현된 인공지능 모델을 토대로 얻은 결과를 활동지에 정리하고 이를 통해 확인할 수 있는 내용을 과학적 지식을 활용하여 서술할 수 있도록 활동지를 제공하고자 한다.

과학 탐구에서 데이터와 정보를 분리하는 것보다 데이터를 처리하는 역량과 정보를 처리하는 역량을 문제 해결을 위해서 융합적으로 활용해야 할 필요성이 있고, ESDA 탐구 모형은 융합교육의 하나의 유형으로 현장에서 진행할 구체적인 교수학습 방안에 대한 연구 필요성이 있다(손미현, 2020). 따라서 데이터 리터러시 역량의 향상을 위해 과학 데이터를 기반으로 ESDA 탐구 모형의 설계 원리를 적용한 인공지능 고등학교 과학 융합 프로그램을 개발이 이어져야 할 필요성이 있다.

### 3. 인공지능(AI) · 과학 교과 융합 프로그램

본 연구를 위해 인공지능(AI) · 과학 융합 프로그램에 관한 선행 연구

를 연구년도, 연구주제, 이용 플랫폼, 학교급으로 분석하면 <표 II-7>과 같다.

<표 II-7> 인공지능(AI)·과학 융합 프로그램

연구자	연구 년도	연구주제	플랫폼	학교 급
신원섭	2020	초등 생물 분류 학습에서 인공지능 융합 교육의 적용 사례 연구	ML4K, scratch	초
손원성	2020	인공지능(AI) 교육 플랫폼을 활용한 SW교육 수업안 개발 : 초등학교 고학년을 중심으로	AI Oceans, 엔트리	초
박민솔 외	2020	미래 지능형 과학실 활용을 위한 ‘화학 원소 기호 이미지 기계학습 AI·SW교육 프로그램’ 제안	Teachable machine, Visual Studio Code	고
조영생 외	2021	AI 분류 모델을 활용한 초등 AI 과학 융합 교육프로그램 개발	Teachable machine	초
이준행 외	2021	인공지능 융합교육을 위한 데이터 기반 교육자료 개발 : 감쇠진동을 중심으로	Python, Pytorch, Google Colab	고
이소율 외	2021	머신러닝 교육 플랫폼 활용 ‘분자 구조의 이해’를 위한 융합교육 프로그램 개발	Avogadro ML4K	고
조연수	2022	인공지능을 융합한 과학 수업이 중학생의 인공지능에 대한 태도 및 데이터 리터러시 역량에 미치는 효과 : 중학교 과학 ‘별과 우주’ 단원을 중심으로	Teachable machine, AI4Mars	중
허희정 외	2022	인공지능을 주제로 한 과학탐구실험 교과 내 ‘첨단과학탐구’ 단원 수업 프로그램의 개발 및 적용	Autodraw, Quickdraw, Teachable machine, 엔트리	고
문샛별 외	2022	인공지능 융합 화학수업에 대한 고등학생의 인식과 만족도 분석	Ornage	고
이소율 외	2021	고등학생을 위한 머신러닝 교육 플랫폼 활용 과학 융합교육 콘텐츠 개발	엔트리	고

<표 II-7>에서 볼 수 있듯이 대부분의 인공지능(AI)·과학 융합 프로그램은 머신러닝 플랫폼을 기반으로 하고 있으며, 교과 내용을 인공지능 플랫폼을 활용하여 새로운 방식으로 접근해 보고자 하는 프로그램들이다. 적용된 플랫폼의 형태는 연구년도가 지나갈수록 다양해지고 있으며 블록 코딩의 인공지능 활용에서 텍스트 코딩을 이용할 수 있는 방식으로 변화되고 있다. 기 개발된 프로그램들은 초등학교 급에서 적용되는 프로그램들이 많았으나, 최근에는 중고등학생들이 이용할 수 있는 다양한 프로그램의 개발이 이루어지고 있다. 하지만 대부분의 프로그램들이 블록코딩으로 모델을 만드는 방식이 많으며, 교과 융합적인 요소보다는 여러 과학 과목 중 한 과목에 인공지능 융합 교육 프로그램이 개발되고 있다. 이에 본 연구에서는 교과 내용 중 고등학교 물리학, 화학 과목에 적용할 수 있고, Colab 플랫폼에서 텍스트 코딩을 활용하여 과학 데이터 기반 인공지능(AI)·고등학교 과학 융합 프로그램을 개발하고자 한다.

#### 4. 지도 학습(Supervised Learning)

인공지능에서 머신러닝(Machine learning)은 성능이 향상되는 컴퓨터 알고리즘에 관한 연구를 총칭하는 것으로, 개발자가 프로그래밍하지 않아도 학습 알고리즘이 프로그램을 만드는 것이라고 할 수 있다.(김진형, 2020) 머신러닝의 알고리즘은 지도 학습, 비지도 학습, 강화 학습으로 나눌 수 있고, 그 중 지도 학습은 정답이 있는 훈련 데이터셋을 이용하여 학습시키는 것이다. 입력 값으로부터 출력 데이터를 나타낼 수 있는 과제를 수행하는 것으로, 대표적인 지도 학습의 과제는 분류(Classification)와 회귀(Regression)가 있다. 분류는 주어진 훈련 데이터를 정해진 라벨(Label)에 따라 학습시키고 예측하게 한 뒤에 예측된 결

과가 정해진 라벨에 정확하게 찾아갈 수 있도록 지도하는 것이다. 회귀는 훈련 데이터의 특징(Feature)를 토대로 라벨을 잘 찾아갈 수 있는 함수를 만들어가는 학습이다. 지도 학습에서 주로 활용되는 알고리즘은 <표 II-8>와 같다.

<표 II-8> 지도학습에 주로 활용되는 알고리즘

K-최근접 이웃(k-Nearest Neighbors)
선형 회귀(Linear Regression)
로지스틱 회귀(Logistic Regression)
서포트 벡터 머신(SVM, Support Vector Machine)
결정 트리(Decision Tree)
랜덤 포레스트(Random Forest)
신경망(Neural Network)

본 연구에서는 공개된 과학 관련 내용인 만능 지시약의 색 변화를 RGG data와 pH로 라벨링한 빅데이터 세트(Big data set)으로 만든 머신러닝 인공지능 모델의 성능을 비교할 수 있는 Lucifer-ML Python 라이브러리를 Colab에 설치하여 Lucifer-ML 패키지에 내장되어 있는 분류 모델의 성능을 비교해보고 가장 높은 알고리즘 모델을 바탕으로 인공지능 모델을 구현하고 실험적으로 얻은 이미지 데이터로부터 pH를 예측하는 인공지능 모델을 개발하여 교육 프로그램에 적용해 보고자 한다.



## 5. 머신러닝 플랫폼과 프로그래밍 언어

### 가. Python

Python은 데이터 과학 분야에서 가장 많이 사용되고 있는 프로그래밍 언어로 오늘날 수집되는 대용량 데이터를 처리하기에 계산 효율성과 사용자 편의성을 갖고 있다. 이에 지난 10년간 과학 계산(Scientific computing) 커뮤니티에서 가장 많이 활용되는 프로그래밍 언어이다.(Raschka et al., 2020). Python은 Numpy와 Scipy 등의 계산을 위한 강력한 라이브러리를 불러와서 데이터 처리를 할 수 있으므로, 오늘날의 많은 인공지능 툴들은 Python을 기반으로 개발되어 있다.

### 나. 싸이킷런(Scikit-learn) 라이브러리

싸이킷런(Scikit-learn)은 2007년에 개발된 Python으로 구현된 가장 유명한 기계학습 오픈 소스 라이브러리이다. 싸이킷런은 여러 라이브러리와 호환성이 좋으며, 다양한 분류, 회귀, 클러스터링, 차원 축소처럼 기계학습에 자주 사용되는 알고리즘을 지원하고 있고, 머신러닝 결과를 검증하는 기능도 갖추고 있다. 또한 Python 생태계를 기반으로 하기 때문에 기존의 통계 데이터 분석 범위를 벗어난 애플리케이션에도 쉽게 통합할 수 있어 다양한 분야에서 활용되고 있다(Pedregosa et al, 2011).

### 다. Colab

Google사에서 개발한 웹브라우저 기반의 대화형 클라우드 컴퓨터 환경이다. 현재 교육 자료 개발에 많이 활용되는 Jupyter notebook을 기반으로 개발된 Colab은 웹브라우저 상에서 프로그래밍 언어인 코드(Code)와 문서를 동시에 편집할 수 있어 Python의 작업 환경을 편리한 형태로 제공하고 있다(Nelson & Kinder, 2018). 웹브라우저 형태로 작업

환경이 구현되므로 디지털 도구의 OS에 크게 구애받지 않고, 문서의 공유와 공동작업이 가능한 장점이 있으므로 수업 자료의 개발과 배포에 유용한 점이 있다. 또한 빅데이터를 이용할 때에도 머신러닝의 성능이 뛰어나므로 활용 가능성이 높다(이준행 외, 2021). 이에 본 연구에서는 Colab 기반의 텍스트 코딩 온라인 활동지에 텍스트 코딩과 마크다운 문서를 포함하여 학생들에게 온라인 활동지로 공유하고자 한다. 온라인 활동은 Google사에서 개발한 Google Classroom 학습 플랫폼을 기반으로 하고 최고 관리자 계정에서 Colab 환경을 학생들에게 허용하여 모든 학생이 파이썬 기반의 언어를 온라인에서 실행하고 변형하는 과정을 거치게 하고자 한다. 또한 학생들이 의견을 공유하거나 하는 과정에서는 padlet과 같은 온라인 의견 공유 프로그램도 활용하여 데이터의 저장 및 공유가 원활하게 이루어질 수 있도록 하고자 한다.

#### 라. Lucifer-ML 패키지

2021년에 개발된 머신러닝 패키지로 여러 가지 머신러닝 알고리즘을 한꺼번에 분석해 줄 수 있는 뛰어난 성능을 가진 패키지이다. Dark-Lucifer의 닉네임을 가진 깃허브(<https://github.com/d4rk-lucif3r/LuciferML>)에 공개되어 있으며 <표 II-9>와 같은 절차를 거쳐서 data 분석 중 분류(Classification)와 회귀(Regression)모델의 성능을 한번에 확인할 수 있는 유용한 도구이다.

<표 II-9> Lucifer-ML 패키지의 데이터 처리 절차(Riyantoko et, 2021)

				
Lucifer-ML Python 라이브러리 설치	사전 데이터 분석	데이터 분석	데이터 왜곡 (Skewness)수정	모델별 분석 결과 제공

Lucifer-ML 패키지의 머신러닝 알고리즘 분석 결과를 토대로 학생들이 가장 최적화된 머신러닝 알고리즘을 정하고 데이터를 기반으로 인공지능 모델을 만드는 과정이 인공지능 융합교육 프로그램에 포함된다면 데이터 리터러시 향상에 도움이 될 것이다.

## 6. 데이터셋과 교육과정 성취 기준의 융합

교육부는 '2022 개정 교육과정 총론 주요사항'에서 개정 중점 사항 중 하나로 '미래 사회가 요구하는 역량 함양이 가능한 교육과정'과 AI·SW 교육을 비롯한 디지털 기초 소양 강화를 강조하였다. 미래 사회의 역량을 증진시켜주는 2022 개정 교육과정과 AI 교육의 성공을 위해서는 적절한 교육 프로그램의 개발이 필요하다(김슬기 외, 2023). AI 교육 내용을 성취 기준에서 다루고 있는 고등학교 과목은 '정보', '데이터과학', '인공지능 기초' 과목에서 중점적으로 다루고 있으며, 관련 성취 기준은 <표 II-10>와 같다.

<표 II-10> 고등학교 AI 교육 주요 성취 기준

과목	성취 기준
정보	[12정04-02] 기계학습의 개념을 이해하고, 지도학습과 비지도학습의 차이를 비교·분석한다.
데이터과학	[12데과03-02] 동일한 데이터를 통계적 회귀모델과 기계학습을 통한 회귀 모델로 분석하여 결과 해석 내용을 비교한다
인공지능 기초	[12인기02-01] 기계학습을 적용할 문제를 정의하고, 문제해결에 필요한 데이터를 선정하여 수집한다. [12인기02-06] 딥러닝을 활용하여 실생활 및 다양한 학문 분야의 문제를 해결하고, 성능을 평가한다.

AI 교육의 발전을 위해서는 양질의 데이터셋이 발굴 및 제공 되어야 할 필요성이 있으므로 대한민국 정부에서도 '인공지능 국가 전략', '디지털 뉴딜' 등의 정책을 통해 노력하고 있다.

국내 뿐만 아니라 해외에서도 다양한 연구자와 기업, 연구 기관등이 AI를 위한 데이터셋의 중요성에 대하여 공감하고 데이터셋을 제공하기 위한 플랫폼이나 라이브러리를 개발하여 활용하도록 하고 있다(김슬기 외, 2023). 국내외의 주요 데이터셋 제공 플랫폼과 라이브러리는 <표 II-11>와 같다.

<표 II-11> AI 교육용 데이터셋 제공 플랫폼과 라이브러리(김슬기 외, 2023)

플랫폼 및 라이브러리	특징
기상 자료 개방 포털 ( <a href="https://data.kma.go.kr/">https://data.kma.go.kr/</a> )	오랜 기간 수집된 기상자료 관련 특화된 다양한 데이터를 상세한 설명과 함께 제공
공공 데이터 포털 ( <a href="https://www.data.go.kr/">https://www.data.go.kr/</a> )	국내 공공기관이 생성 또는 취득한 데이터를 제공하는 다양한 플랫폼을 연계해주며 데이터를 손쉽게 검색하고 다운 받을 수 있는 링크 제공
Kaggle ( <a href="https://www.kaggle.com/">https://www.kaggle.com/</a> )	전 세계의 사용자들이 수 많은 데이터셋과 데이터셋을 활용한 분석 및 AI 모델링 결과를 공유하는 커뮤니티 성격의 플랫폼
UCI ML Respository ( <a href="https://archive.ics.uci.edu/">https://archive.ics.uci.edu/</a> )	데이터의 형태, 특성의 수, AI 알고리즘 등 다양한 기준으로 유목화 된 600여 종 이상의 정제된 데이터셋 제공
Pydataset ( <a href="https://pypi.org/">https://pypi.org/</a> )	파이썬 개발 환경에서 데이터셋을 쉽게 활용하기 위한 목적으로 개발되어 756종의 데이터셋 활용 가능, 파이썬 설치 라이브러리

김슬기 외(2022)는 AI 교육 데이터 셋중 국내 공공 데이터셋 플랫폼은 현실의 데이터셋을 제공하여 학생의 삶과 연관성이 높은 소재이지만, 정확도가 상대적으로 높은 AI 모델의 생성에는 적합하지 않고, 외국의 데이터셋 플랫폼 및 라이브러리는 정확도가 높은 AI 구성 및 결과를 제공하지만, 학생들의 생활 환경과는 관련성이 떨어지는 소재들이 많음을 제시하고 있다.

과학 교육에서는 정확도가 높은 데이터셋을 활용하여 성능이 높은 AI 모델을 만들어 제공해야 할 필요성이 있으므로 본 연구에서는 Kaggle 플랫폼에 제공되는 pH recognition 데이터셋을 활용하여 고등학교 과학 데이터셋을 이용한 정확도가 높은 인공지능 모델을 만들어 학생들에게 인공지능 모델을 만들어보는 경험을 교육 프로그램에서 제공하고자 한다.

pH recognition 데이터셋은 pH에 따른 만능 지시약의 색변화를 RGB 데이터로 정리해 놓은 것으로 이와 관련된 성취 기준으로는 2022 개정 고등학교 과학과 교육과정의 물리학과 화학에서 <표 II-12>와 같이 찾아볼 수 있다. 이에 본 연구에서는 적절한 데이터셋과 교육과정을 융합한 인공지능 고등학교 과학 융합 프로그램을 개발하고자 한다.

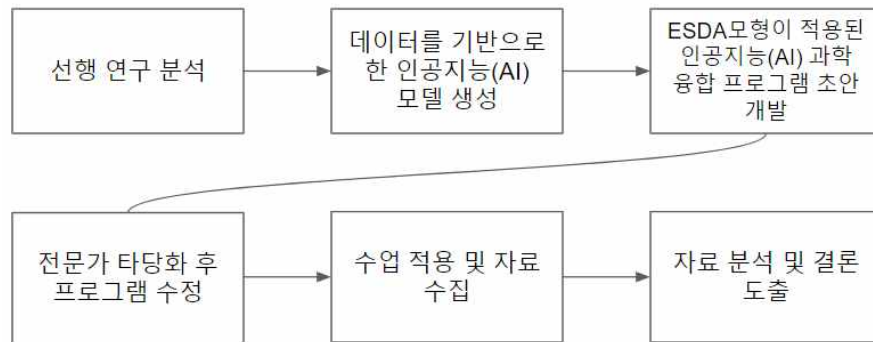
<표 II-12> pH recognition 데이터셋에 적용 가능한 2022개정 고등학교 과학과 성취기준

과목	성취 기준
물리학	[12물리03-01] 빛의 중첩과 간섭을 통해 빛의 파동성을 알고, 이를 이용한 기술과 현상을 예를 들어 설명할 수 있다. [12물리03-03] 빛과 물질의 이중성이 전자 현미경과 영상 정보 저장 등 다양한 분야에 활용됨을 설명할 수 있다.
화학	[12화학04-01] 물의 자동 이온화와 물의 이온화 상수를 이해하고, 수소 이온의 농도를 pH로 표현할 수 있다.

### Ⅲ. 연구 방법

#### 1. 연구 절차

데이터나 경험을 통해서 스스로의 능력을 향상 시키는 방법인 머신러닝은 인공지능의 중요한 연구 영역이고(김진형, 2020), 변형에 따른 무수한 변칙까지도 데이터를 이용해 모두 찾아낼 수 있으며 결과도 추론할 수 있게 한다(박상길, 2022). 따라서 인공지능(AI) 융합 프로그램에서는 인공지능(AI) 모델을 생성해야 할 필요성이 있다. 본 연구에서는 머신러닝의 회귀(Regression) 알고리즘을 이용하여 pH 값을 예측할 수 있는 모델을 공개된 데이터를 기반으로 개발하고, 이 과정에서 만들어진 인공지능 모델을 활용할 수 있게 ESDA 모형에 적용하여 수업 프로그램을 개발하고자 한다. 프로그램의 개발 절차는 [그림 Ⅲ-1]과 같다.



[그림 Ⅲ-1] 프로그램 개발 절차

#### 2. 연구 대상

본 연구는 2022년 2학기에 서울 소재 일반계 고등학교 1, 2학년 학생 총 27명을 대상으로 실시하였다(<표 Ⅲ-1>). 수업은 2학년 학생들의 경우 공동 교육과정의 과학 거점학교 화학 실험 수강 학생을 대상(14명)

으로 진행하였고, 1학년 학생들은 과학 탐구 실험 교과 시간을 활용하여 인공지능을 활용한 과학 수업에 관심이 있는 학생들(14명)을 대상으로 진행하였다.

<표 III-1> 연구 대상

학교	수업 형태	학년	남학생	여학생	계
과학 거점학교	공동 교육과정 화학 실험 과목	2	6	5	11
I 고등학교	과학탐구실험 과목	1	16		16
계			22	5	27

### 3. 연구 도구

본 연구에서 사용된 연구 도구는 내적 타당화를 위해 4점 척도( 4: 매우 그렇다, 3: 그렇다, 2: 그렇지 않다, 1: 매우 그렇지 않다) 척도의 전문가 타당화 검사지, 학습자 대상 외적 타당화를 위한 설문지, 면담 질문지, 수업 플랫폼인 구글 클래스룸(Google Classroom)과 패들릿(Padlet), 구글 문서(Google docs), 구글 스프레드시트(Google spreadsheet) 등이다. 또한 현장 적용 후 프로그램의 데이터 리터러시 향상 여부를 판단하기 위한 학습자 대상 수업 데이터 리터러시 검사지(구글 설문 활용)를 기존 연구자의 문항을 토대로 수정 보완하여 4점 척도 4: 매우 그렇다, 3: 그렇다, 2: 그렇지 않다, 1: 매우 그렇지 않다)로 사용하였다.

전문가 타당화에 사용한 검사지는 나일주와 정현미(2001)의 평가 문항을 바탕으로 문항을 재구성하였고, 김근재(2019)에서 사용한 전문가 타당화 검사지를 수정 및 보완하여 <표 III-2>와 같이 개발하여 사용하였다.

<표 III-2> 프로그램 구성 내용에 전문가 타당화 평가 문항

항목	프로그램 평가 문항
타당성	본 프로그램은 데이터 리터러시 향상을 위한 인공지능(AI) · 고등학교 과학 융합 프로그램으로 타당하다.
설명력	본 프로그램은 데이터 리터러시 향상을 위한 인공지능(AI) · 고등학교 과학 융합 프로그램의 각 단계를 잘 설명하고 있다.
유용성	본 프로그램은 데이터 리터러시 향상을 위한 인공지능(AI) · 고등학교 과학 융합 프로그램으로 유용하게 활용될 수 있다.
보편성	본 프로그램은 데이터 리터러시 향상을 위한 인공지능(AI) · 고등학교 과학 융합 프로그램으로 보편적인 고등학생들을 위해 적용될 수 있다.
이해도	본 프로그램은 데이터 리터러시 향상을 위한 인공지능(AI) · 고등학교 과학 융합 프로그램으로 적용시 필요한 요소들이 이해하기 쉽게 표현되어 있다.

학생들을 대상으로 프로그램 전과 후의 데이터리터러시의 향상을 보기 위해서 외적 타당화 검사지는 기 개발된 송유경(2021)의 데이터 리터러시 검사지와 조연수(2022)의 데이터 리터러시 역량 측정 도구를 프로그램에 적합하게 세부 항목을 변경하여 구글 설문으로 사용하였다.

## 4. 자료 수집

### 가. 선행 문헌 검토

인공지능(AI) 고등학교 과학 융합 프로그램의 개발을 위하여 과학 교육 및 머신 러닝과 관련된 선행 문헌들을 검토하여 인공지능(AI)과 과학 융합 프로그램들을 분석하는 과정을 거쳤다. 관련 문헌을 수집하고 정하기 위하여 구글 학술검색(<https://scholar.google.co.kr/>)와 디비피아(<https://www.dbpia.co.kr/>), KISS 한국 학술 정보



(<https://kiss.kstudy.com/>), 스콜라(<https://scholar.kyobobook.co.kr/>) 등의 사이트를 활용하였다.

선행문헌 검토를 위해 ‘인공지능과 과학’, ‘머신러닝과 과학’, ‘머신러닝과 과학’, ‘머신러닝’, ‘교과 융합 프로그램’ 등의 키워드를 사용하였다.

## 나. 내적 타당화

선행문헌 분석과 인공지능(AI) 과학 융합 프로그램의 타당화를 위하여 두 차례의 타당화를 받았다. 1, 2차 타당화에서는 프로그램의 타당성을 각 영역별로 4점 척도( 4: 매우 그렇다, 3: 그렇다, 2: 그렇지 않다, 1: 매우 그렇지 않다)으로 평가할 것을 요청하였다. 1, 2차 타당화에는 화학교육, 물리교육, 과학교육 전문가 6인이 참여하였다. 1차 타당화에서 나온 의견을 바탕으로 프로그램의 내용과 학생들에게 주어질 설문의 내용을 수정하고 2차 타당화를 거쳤다.

전문가 타당화를 통해 인공지능(AI) 과학 융합 프로그램 초안의 타당성, 설명력, 유용성, 이해도, 보편성과 데이터 리터러시 역량 측정 도구의 타당도를 검토하였다. 전문가 타당화에 참여한 전문가들의 프로필은 다음 <표 III-3>과 같다.

<표 III-3> 전문가 타당화에 참여한 전문가 프로필

전문가	직업	경력	최종학력	전공분야
A	연구원	21	박사	화학교육
B	교수	18	박사	화학교육
C	연구원	5	박사	과학교육
D	교사	11	석사	물리교육
E	교사	16	석사	화학교육
F	교사	9	학사	화학교육

## 다. 외적 타당화

본 연구는 고등학생을 위한 인공지능(AI)·과학 융합 프로그램을 개발하여 고등학생의 데이터 리터러시 향상에 효과적인가를 보고자한다. 전문가 타당화에서 외적 타당화 질문에 대한 의견을 구했을 때, 기존 검사지를 활용하거나 외적 타당화 질문을 구체적으로 표현해야 할 필요성에 대하여 언급되었다. 또한 프로그램에 본 프로그램은 인공지능 모델의 생성 과정이 포함되어 있다. 따라서 외적 타당화 검사지는 기 개발된 송유경(2021)의 데이터 리터러시 검사지와 조연수(2022)의 데이터 리터러시 역량 측정 도구를 프로그램에 적합하게 세부 항목을 변경하고, 머신러닝 모델 생성의 부분을 추가하였고, 의사 소통 부분에 대한 강조보다는 과학 지식에 대한 창의적 해석을 강조하여 설문지를 재구성하여 <표 III-4>과 같이 제공하였고, 각 영역별로 4점 척도( 4: 매우 그렇다, 3: 그렇다, 2: 그렇지 않다, 1: 매우 그렇지 않다)로 제공하였다.

<표 III-4> 데이터 리터러시 역량 측정 도구 설문지 문항 1차

데이터 리터러시	문항	연구자
데이터 이해	1. 나는 데이터에 다양한 종류가 있음을 이해하고 있다.	조연수(2022)
	2. 나는 데이터에 각 행과 열이 의미하는 바를 이해할 수 있다.	송유경(2021)
데이터 수집	3. 나는 필요한 데이터가 없을 때 새로운 방법을 사용하여 데이터를 얻어낼 수 있다.	조연수(2022)
	4. 나는 데이터의 출처를 평가하고 선택할 수 있다.	조연수(2022)
데이터 관리	5. 나는 데이터를 적절한 기준에 따라 나누어 저장할 수 있다.	조연수(2022)
	6. 나는 데이터 분석을 위한 준비를 하고 적절한 도구를 사용하여 분석할 수 있다.	조연수(2022)

데이터 해석 및 평가	7. 나는 서로 다른 두 종류의 데이터가 어떻게 해석되는지 이해하고 있다.	조연수(2022)
	8. 나는 데이터 해석을 위해 모델을 설계할 수 있다.	송유경(2021) 세부 항목 변경
데이터 활용	9. 나는 데이터로부터 새로운 통찰을 발견하고 데이터를 창의적으로 해석할 수 있다.	송유경(2021) 세부 항목 변경
	10. 나는 데이터 해석을 위해 만든 모델을 활용할 수 있다.	신규 개발
데이터 시각화	11. 나는 표, 차트, 그래프 등 여러 시각화 방식의 특징을 이해하고 있다.	송유경(2021)
	12. 나는 필요에 따라 표, 차트, 그래프 등 여러 시각화 방식 중 적절한 것을 선택하여 데이터를 시각화 할 수 있다.	송유경(2021)

개발된 초기 문항에 대하여 타당도를 검증하기 위하여 <표 III-3>의 전문가들에게 각 검사 문항이 데이터 리터러시를 제대로 측정하고 있는지 의견을 구하였다. 전문가의 의견 중에는 학생들에게 적절한 답을 위하여 문항을 구체적으로 서술해야할 필요성이 제기되었으므로 이를 반영하여 수정한 12개의 문항을 <표 III-5>와 같이 사용하여 검사를 실시하였다.

<표 III-5> 데이터 리터러시 역량 측정 도구 설문지 문항 2차

데이터 리터러시	문항	연구자
데이터 이해	1. 나는 데이터에 다양한 종류(숫자, 이미지, 터치 등)가 있음을 이해하고 있다.	조연수(2022)
	2. 나는 데이터에 각 행과 열이 의미하는 바를 이해할 수 있다.	송유경(2021)
데이터 수집	3. 나는 필요한 데이터가 없을 때 새로운 방법을 사용하여 데이터를 얻어낼 수 있다.	조연수(2022)

	4. 나는 데이터의 출처를 평가하고 적절한 데이터를 선택할 수 있다.	조연수(2022) 변형
데이터 관리	5. 나는 데이터를 적절한 기준에 따라 나누어 저장 또는 보관할 수 있다.	조연수(2022) 변형
	6. 나는 데이터 분석을 위한 준비를 하고 적절한 도구를 사용하여 분석할 수 있다.	조연수(2022)
데이터 해석 및 평가	7. 나는 서로 다른 두 종류의 데이터가 어떻게 해석되는지 이해하고 있다.	조연수(2022) 변형
	8. 나는 데이터 해석 및 적용을 위해 인공지능 모델(분류, 회귀)의 성능을 확인할 수 있다.	송유경(2021) 변형
데이터 활용	9. 나는 데이터로부터 새로운 사실이나 의견을 발견하고 데이터를 창의적으로 해석할 수 있다.	송유경(2021) 변형
	10. 나는 데이터 해석을 위해 만든 인공지능 모델(분류, 회귀)을 활용할 수 있다.	신규 개발
데이터 시각화	11. 나는 표, 차트, 그래프 등 여러 시각화 방식의 특징을 이해하고 있다.	송유경(2021)
	12. 나는 필요에 따라 표, 차트, 그래프 등 여러 시각화 방식 중 적절한 것을 선택하여 데이터를 시각화 할 수 있다.	송유경(2021)

## 5. 자료 분석

### 가. 전문가 타당화 자료

내적 타당화는 전문가 타당화를 진행하였고 해당 프로그램의 타당도와 데이터 리터러시 역량 측정 질문의 타당도를 측정하는 문항으로 구성하였다. 전문가가 질문지에 응답한 문항에 대한 타당성과 신뢰성을 확보하기 위하여 Rubio와 연구자들(2003)이 제안한 내용 타당도 지수

(Content Validity Index: CVI)와 평가자 간 내용 일치도 지수 (Inter-Rater Agreement : IRA)를 구하였다. 1차 타당화 결과 프로그램은 타당성, 설명력, 유용성 부분에서는 CVI가 1.00으로 높게 나왔으나 이해도와 보편성 부분에서 0.80이하의 결과가 나왔으므로 프로그램의 이해도와 보편성을 높이기 위하여 코딩 온라인 활동지에 기초 부분을 좀 더 자세하게 구성하여 2차 타당화를 거쳤다.

#### 나. 데이터 리터러시 향상도 측정 분석

본 연구는 인공지능(AI) 과학 융합 프로그램을 개발하고 데이터 리터러시의 향상을 살펴보는 것을 목적으로 한다. 이를 위해 <표 III-3>의 전문가들의 타당화를 거쳐 완성된 데이터 리터러시 역량을 수업 프로그램 실시 전과 후에 학습자 27명에게 측정한 뒤에 대응표본 t 검정을 실시하여 데이터 리터러시의 사전, 사후 평균과 표준편차, 평균 차이를 구하고 평균 차이의 유의미성을 검증하고자 하였다(Creswell, 2014).

#### 다. 외적 타당화

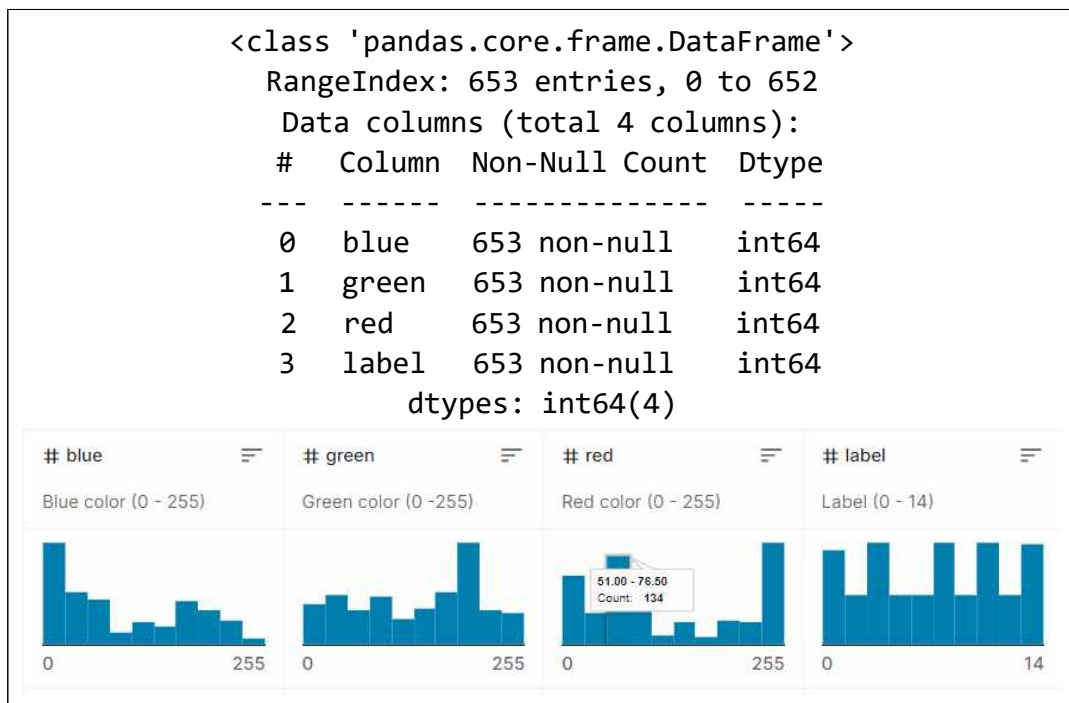
본 연구는 인공지능(AI) 고등학교 과학 프로그램을 개발하고 학생들에게 적용하여 데이터 리터러시 역량의 향상을 살펴보는 것을 목적으로 한다. 이를 위해 본 연구에서 개발된 프로그램이 적용된 수업을 고등학교 현장에 실시하였고, 총 27명의 학생이 참여하였다. 현장 적용 후 학습자를 대상으로 외적 타당화 설문지를 구글 설문으로 입력하게 하고 개별 내용을 적게 하여 학생들의 수업 전 후의 의견을 파악하였다.

## IV. 연구결과

### 1. 과학 데이터기반 인공지능(AI) · 고등학교 과학 융합수업 프로그램의 개발

#### 가. 예측 모델의 개발

인공지능(AI)의 한 분야인 머신러닝을 위해서는 data를 학습하는 과정이 필요하다. 먼저 pH 예측 모델을 만들기 위해서 데이터셋을 공유하는 Kaggle 웹사이트에 공유되어 있는 만능 지시약의 pH에 따른 색 변화에서 RGB data를 분류하여 저장해 놓은 과학 데이터인 pH-prediction 데이터셋([그림 IV-1])을 준비한다.



[그림 IV-1] pH-prediction 데이터셋 정보

pH-prediction 데이터셋은 0~14의 pH 데이터를 Label로 하여 만능 지시약(universal indicator)의 색변화를 빛의 3원색인 RGB 데이터로 정리해 놓은 653개의 행으로 구성된 데이터셋이다. 데이터셋 내부는 숫자 데이터(int64)로 구성되어 있고, 누락된 데이터가 없는(non-null) 데이터 형태이고, Label이 주어진 데이터셋이므로 머신 러닝의 지도학습에 적합한 데이터셋임을 알 수 있다. pH-recognition 데이터셋을 기반으로 pH를 예측하기 위해 python 라이브러리인 Lucifer-ML을 이용하는 코드를 [그림 IV-2]와 같이 작성하면 [그림 IV-3]의 결과를 출력해 내므로 가장 적합한 분류 모델로 Catboost Regressor가 나타남을 확인할 수 있다.

```

✓ [1] !pip install lucifer-ml #lucifer-ML 설치

✓ [2] import pandas as pd #pandas library를 불러옵니다. pd로 지칭합니다.
1초

✓ [3] from luciferml.supervised.regression import Regression
17초
dataset = pd.read_csv('/content/ph-data.csv')
X = dataset.iloc[:, :-1]
y = dataset.iloc[:, -1]

``multivariate`` option is an experimental feature. The interface can change in the future.

5초
regressor = Regression(
    predictor=["all"],
    cv_folds=2,
    # tune=True,
    optuna_n_trials=2,
)
regressor.fit(X, y)

```

[그림 IV-2] pH-recognition 데이터셋으로 적합한 회귀(Regression)모델을 찾는 Lucifer-ML 실행 코드

	Name	R2 Score	Mean Absolute Error	Root Mean Squared Error	KFold Accuracy
0	Linear Regression	71.617015	1.813891	2.272352	69.562803
1	Stochastic Gradient Descent Regressor	71.591600	1.816997	2.273369	69.583596
2	Kernel Ridge Regressor	-208.611412	7.142015	7.492945	-217.230725
3	Elastic Net Regressor	63.992122	2.005355	2.559442	61.063814
4	Bayesian Ridge Regressor	71.651771	1.812119	2.270960	69.572223
5	Support Vector Regressor	94.425714	0.656434	1.007027	93.519444
6	K-Neighbors Regressor	95.268589	0.462595	0.927773	95.513292
7	Decision Trees Regressor	85.689013	0.604962	1.613546	93.169354
8	Random Forest Regressor	92.156887	0.512387	1.194513	95.761413
9	Gradient Boost Regressor	91.966819	0.554590	1.208900	95.404427
10	AdaBoost Regressor	86.619288	0.889896	1.560221	92.783733
11	Bagging Regressor	89.246401	0.576719	1.398696	95.729442
12	Extra Trees Regressor	96.231656	0.395085	0.827984	95.661016
13	LightGBM Regressor	92.765859	0.530159	1.147202	95.309390
14	XGBoost Regressor	90.233054	0.523996	1.332987	94.960429
15	Catboost Regressor	95.539722	0.486899	0.900798	95.792678
16	Multi-Layer Perceptron Regressor	83.489035	1.167257	1.733137	81.962407

[그림 IV-3] Lucifer-ML 분류 모델의 알고리즘 성능 결과

이 성능 출력 결과를 바탕으로 Catboost Regressor 알고리즘을 Scikit-Learn 라이브러리를 사용하여 [그림 IV-4]와 코드를 작성하여 RGB 데이터를 이용하여 pH를 예측하는 모델을 머신러닝으로 개발하였다.



```
[17] from sklearn import ensemble, model_selection
      from sklearn.metrics import mean_absolute_error, r2_score

# y값에 label(pH값), X값에(blue, green, red)를 학습 데이터로 저장
y = dataset.label
X = dataset[['blue', 'green', 'red']]

[19] # 학습 set과 목표 set로 data를 분리
      X_train, X_test, y_train, y_test = model_selection.train_test_split(X,y,
                                                                           test_size=0.25, random_state=42)

[20] from catboost import CatBoostRegressor, cv, Pool

[21] #Initiate a CatBoost Regressor model and train it
      # CatBoost Regressor 모델을 사용하여 학습 시킴
      CB_model = CatBoostRegressor(random_state=12)
      CB_model.fit(X_train, y_train)
```

[그림 IV-4] Scikit-Learn 라이브러리의 Catboost Regressor 알고리즘을 이용한 머신러닝 pH 예측 모델의 개발 코드

## 나. 교육 프로그램의 개발

프로그램의 각 단계마다 데이터 리터러시 구성 요소인 데이터 관리, 데이터 수집, 데이터 이해, 데이터 해석 및 평가, 데이터 활용, 데이터 시각화(송유경, 2021; 조연수, 2022)와 ESDA 과학 교육 프로그램 모형의 설계 원리가 반영되도록 프로그램 초안을 <표 IV-2>과 같이 설계하였다.

<표 IV-1> ESDA 모형이 적용된 과학 데이터 기반 인공지능(AI)·과학 융합 교육 프로그램 초안

단계	학습자 활동	데이터 리터러시
(1) 도구 탐색	-교사가 탐구 소재(pH 예측)을 선정함. -colab에서 python 텍스트 코딩의 도구 사용법을 익힘	데이터 관리
(2) 데이터 수집	만능지시약 실험으로 얻은 여러 가지 시약의 이미지로부터 colab에서 python의 PIL 라이브러리를 이용하여 RGB데이터를 수집함	데이터 수집 데이터 이해
(3) 데이터 변형과 해석	만능지시약 실험 결과 pH와 이미지의 RGB데이터를 정리함	데이터 관리
(4) 문제 발견과 수립	&만능지시약의 색 변화를 학습하면 pH를 예측할 수 있을 것이다 & 라는 문제를 수립	데이터 해석 및 평가
(5) 귀납적 탐구	Kaggle의 pH 예측 big data를 기반으로 머신러닝 회귀 모델을 만듦. Kaggle의 pH 예측 big data를 기반 데이터 시각화 코드 구현	데이터 활용 데이터 시각화
(6) 결론 및 표현	머신러닝 모델로부터 pH를 예측해보고 각 물질들의 [H3O+]를 구하는 과정을 수행함.	데이터 해석 및 평가

프로그램의 구성에 대한 전문가 타당화 1차 검사 결과는 <표 IV-2>과 같다. 1차 전문가 타당화 결과, 타당성(평균 3.67), 설명력(평균 3.17), 유용성(평균 3.50), 보편성(평균 3.00)은 비교적 타당한 것으로 나왔으나 이해도(평균 2.83) 측면에서는 다소 미흡한 것으로 나타났고, 평가자간 신뢰도인 IRA는 0.67로 다소 낮게 나왔다.

<표 IV-2> 전문가 타당화 검사 결과(1차)

영역	전문가						평균	CVI	IRA
	A	B	C	D	E	F			
타당성	3	4	4	3	4	4	3.67	1.00	0.67
설명력	3	3	3	3	4	3	3.17	1.00	
유용성	3	3	4	3	4	4	3.50	1.00	
이해도	2	2	3	3	4	3	2.83	0.67	
보편성	2	2	4	3	4	3	3.00	0.67	

1차 전문가 타당화에서 입수된 전문가들의 의견과 이를 반영하여 프로그램의 구성 요소와 활동을 수정한 사항은 <표 IV-3>과 같다.

<표 IV-3> 전문가 검토 의견과 수정 사항

항목	전문가 검토 의견	수정 사항
설명력	탐구 모형이 다소 추상적으로 반영된 것으로 보이므로 구체적인 교수 전략이 프로그램에 반영될 수 있어야 함.	프로그램이 구성된 활동지를 구글 스프레드시트, 패들릿 등으로 구성하여 프로그램에서 활동하는 내용을 구체화함.
이해도	프로그램에 대한 기초 지식이 없는 학생들은 핵심적인 실험 내용과 pH 예측 과정 보다 프로그램 자체에 대한 이해에 시간을 할애할 수 있음. 프로그램에 대한 구체적인 흐름과 활동을 제시해야 함.	프로그램의 활동지를 Colab을 기반으로 제공하고 전체 차시를 모두 공유 문서로 작성함.
	colab에 제시된 내용만으로 교사의 안내가 충분히 전달될지 충분한 안내가 있어야 할 것으로 보임.	교사의 안내가 충분히 전달될 수 있도록 차시별 필요한 내용을 줄이고 마크다운을 강조하는 방향으

		로 구성함.
보편성	프로그램 자체에 대한 거부감이 있는 경우 학습 단계까지 이르지 못할 수 있음이 우려됨.	사전에 학생 대상 교육 프로그램의 적용 대상을 자발적인 참여 의사를 나타낸 학생을 대상으로 진행함.
	Python의 학습이 어느정도 수준으로 되어야 하는지, 기초 지식이 없는 학생들에 대한 지도를 위한 지도 계획이 필요함.	Colab에서 python을 활용한 기초 내용을 포함하는 방향으로 프로그램 내부의 구성을 수정함.

프로그램 초안에 대한 피드백을 바탕으로 프로그램의 구성 요소들을 구체화하여 하위 내용들을 개선하고, 다양한 에듀테크 도구들(패들릿, 구글 문서, 스프레드시트)을 활용하여 개선한 프로그램은 <표 IV-4>와 같다.

<표 IV-4> ESDA 모형이 적용된 과학 데이터 기반 인공지능(AI)·고등학교 과학 융합 교육 프로그램

단계	학습자 활동	데이터 리터러시
(1) 도구 탐색	<ul style="list-style-type: none"> <li>- Colab에서 python 텍스트 코딩의 기초 문법을 익히게 함.</li> <li>- pandas 라이브러리를 import하여 데이터를 저장하고 출력하는 방법을 익히게 함.</li> <li>- 교사가 탐구 소재(pH 예측)를 다룰 것임을 알림.</li> </ul>	데이터 관리
(2) 데이터 수집	<ul style="list-style-type: none"> <li>- 여러 가지 용액(10가지 이상)중 개인별로 1가지씩 선택하고, 이를 흡판에 넣고 만능지시약 실험으로 색 변화를 관찰함.</li> <li>- 개인별 디지털 도구(태블릿, 스마트폰)등으로 색 변화를 촬영하고 이미지를 패들릿(padlet)링크로 공유함.</li> </ul>	데이터 수집 데이터 이해

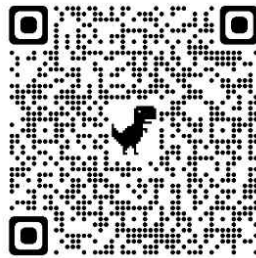
	- 만능 지시약 색변화의 사진을 구글 드라이브에 저장하고 colab에서 python의 PIL 라이브러리를 이용하여 RGB데이터를 수집함	
(3) 데이터 변형과 해석	- 용액의 종류와 눈으로 본 pH 예측값 및 python 이미지 데이터 수집을 통해 얻어진 RGB 데이터를 입력하고, 기종별 편차를 확인하기 위해서 촬영 기종등을 정리하게 함. - 각 데이터들의 차이 유무와 용액의 pH 변화가 올바른지 토의하여 해석하는 과정을 거침.	데이터 관리
(4) 문제 발견과 수립	- 색변화 data를 기반으로 해석한 내용을 발표하고 공유함. - 비슷한 색으로 변하는 용액들이 있으므로 만능 지시약의 색 변화를 학습하면 pH를 예측할 수 있을 것이다 라는 문제를 수립.	데이터 해석 및 평가
(5) 귀납적 탐구	- pH-recognition 데이터 셋의 데이터가 어떤 분포를 보이고 있는지를 시각화하는 과정을 수행하고 출력함. - pH 예측이 가능한 AI 모델을 머신러닝 회귀(regression) 알고리즘을 사용하여 구현함.	데이터 시각화 데이터 활용
(6) 결론 및 표현	- 자신이 데이터를 수집한 용액의 RGB 값을 대입하여 머신러닝 모델로부터 pH를 예측해보고 각 물질들의 $[H_3O^+]$ 를 구하는 과정을 모듈별로 수행함.	데이터 해석 및 평가

2차 전문가 타당화는 1차 타당화의 전문가들에게 다시 의뢰하였고, <표 IV-5>과 같이 모든 항목에서 3이상의 타당성을 보였다. CVI와 IRA는 1로 나타나 전문가 타당화 결과의 신뢰도가 향상되었고 검증되었다.

<표 IV-5> 전문가 타당화 검사 결과(2차)

영역	전문가						평균	CVI	IRA
	A	B	C	D	E	F			
타당성	3	4	4	3	4	4	3.67	1.00	1.00
설명력	3	3	3	3	4	3	3.17	1.00	
유용성	3	3	4	3	4	4	3.50	1.00	
이해도	3	3	3	3	4	3	3.17	1.00	
보편성	3	3	4	3	4	3	3.33	1.00	

본 연구에서는 클라우드 기반의 컴퓨터인 Colab을 이용하여 학생들에게 코딩 활동지를 제공하였고, 학생들은 이 활동지를 수정, 변형, 실행하면서 인공지능(AI)·고등학교 과학 융합 프로그램에 참여할 수 있었다. 각각의 차시별로 총 6차시의 수업을 진행하였고, 클라우드 컴퓨팅의 특성상 기존 차시들에서 활동한 내용이 저장되지 않으므로 지난 차시의 활동들을 사전에 컴퓨팅 환경에서 실행시키는 과정이 복잡 형태로 운영되게 하였다. 다음 QR코드([그림 IV-5])를 스캔하여 확인할 수 있다.



[그림 IV-5] Colab용 텍스트 코딩 활동지 QR코드

## 2. 프로그램의 적용 효과

### 가. 데이터 리터러시 사전, 사후 검사(외적 타당화)

데이터 리터러시의 향상을 확인해보기 위하여 본 연구에서는 수업 대

상 학생 27명의 대응표본 t-검정을 실시하여 데이터 리터러시의 사전, 사후 평균과 표준편차, 평균 차이를 구하고 평균의 차이가 통계적으로 유의미한지 검토해보았다(Creswell, 2014). 수업에 참여한 27명의 데이터 리터러시 사전, 사후 검사 결과 평균이 4점 만점에 3.19점에서 3.74점으로 상승하였다. 또한 대응표본 t-검정의 검사 결과(<표 IV-6>)에서  $p < 0.01$ 의 결과를 보였으므로 개발된 프로그램은 학생들의 데이터 리터러시 향상에 효과적임을 확인 할 수 있었다.

<표 IV-6> 대응표본 t-검증 검사 결과

대응표본 검정								
대응차						t	자유도	유의확률 (양측)
사후검사 - 사전검사	평균	표준화 편차	표준오차 평균	차이의 95% 신뢰구간				
				하한	상한			
	0.605	0.969	0.186	0.222	0.989	3.245	26	0.003

\*\*  $p < 0.01$

#### 나. 학습자 반응

데이터 리터러시 검사 사후 설문지의 마지막 문항으로 프로그램에 참여한 학생들에게 수업 후에 느낀점과 장점 단점을 묻는 설문을 입력하게 하였다. 데이터를 다루는 능력이 향상되었음을 나타내는 의견, 인공지능을 이용하면 더 정확한 pH를 예측할 수도 있겠다는 의견, 교사의 지도로 인하여 수업에 흥미와 참여도가 높아질수 있었다는 의견이 다수 있었다. 구체적인 학습자의 긍정 의견을 정리하면 다음의 <표 IV-7>과 같다.

<표 IV-7> 학습자 의견 - 좋았던 점

구분	내용
<p>데이터 리터러시 향상</p>	<ul style="list-style-type: none"> <li>- 파이썬이 누가 하는 거 옆에서 보면 정말 어렵고 전문가나 다룰 수 있을 것 같았는데 막상 해보니까 그렇게 어렵지만은 않아 보인다. 혼자 해보라고 하면 아직은 절대 못 하겠지만 적어도 파이썬 문자? 코드? 에 대해서는 조금 더 친근하게 느끼게 된 것 같다. 몇 개는 직접 입력해 보니까 재미있고 신기했고, 직접 못 하고 그냥 실행만 해본 거는 나중에 시간 많을 때 한 번 배워보고 싶다. 변화가 딱히 눈에 보이진 않았는데 쉽게 다운로드 받고 실행하고 할 수 있다는 게 신기하다.</li> <li>- pH 실험을 하고 회귀 모델을 만들어 데이터를 시각화하는 수업이었다. 수업을 하면서 가장 좋았던 점 중 하나는, 처음 지시약의 RGB 값을 분석하였을 때에 눈으로 본 것과 너무 편차가 커서 당황하였다. 그러나, 선생님께서 실패의 원인 (너무 적었던 만능 지시약의 양)을 정확히 지적해주셨고, 문제점을 고쳐 다시 시도해본 결과, 이번에는 예측 범위와 편차가 크지 않게 되었다! 이처럼 오늘 수업을 하면서 물론 ai와 여러 과학 지식에 관련된 것들도 많이 얻어가지만, 무엇보다 한 번 실패하였을 때에 포기하지 않고 문제점을 고치려 노력하는 태도를 배울 수 있어서 좋은 수업이었던 것 같다.</li> </ul>
<p>인공지능 관련</p>	<ul style="list-style-type: none"> <li>- 인공지능을 만드는 원리는 어렵지만 만드는 명령어를 만드는 것 자체는 익히기만 한다면 생각보다 많이 어렵지 않음을 알 수 있었다. 평소에 이런 실험을 하게 되면 pH가 이정도 될 것 같다고 예측을 하고, 비슷한 색을 띠다면 pH가 비슷하구나 까지만 예측할 수 있는데 오늘 배운 것을 활용하면 더 정확하게 구현할 수 있음을 깨달았다. 머신 러닝의 원리도 x의 값에 따라 y의 결과가 나오는 단순한 원리인 것이 신기했다.</li> <li>- pH를 모델을 통해 분석할 수 있다는 것이 새로웠습니다. 데이터 분석에 파이썬이 어떻게 사용되는지 알 수 있었던 유익한 시간이었습니다!</li> </ul>



	- 장점: 회귀모델과 분류모델을 이용해 데이터를 분석해 코딩 사용법을 알게 되었다.
흥미	- 어려워 보이는 수업이었지만 선생님이 다 준비해주신 자료로 수업을 해서 생각보다 어렵지 않았다. 선생님이 말씀하신 내용중 코딩은 컴퓨터와 나의 대화라고 말씀하신 내용이 기억에 남는다. 어렵게 생각만 들던 코딩을 일부분이지만 직접 실행해 RGB값을 찾아내는 과정이 신기했다.

한편, 학습자 의견 중 아쉬운 점 혹은 어려웠던 점에 대해 정리한 결과 과정이 어렵고, 지루하다는 의견과 pH와 색상과의 연관성의 원리는 알 수 없었다는 의견이 있었다. 대부분의 의견은 긍정적으로 학습자들이 교육 프로그램을 수행하면서 새롭게 접한 내용들이 흥미를 이끌고 데이터를 다루고 인공지능 모델을 만드는 과정에 대한 이해와 적용을 높이는 결과를 나타낸 것을 알 수 있다.

## V. 결론 및 제언

### 가. 결론

기존의 인공 지능(AI)·과학 융합 교육 프로그램은 주로 초등학교 급에서 적용할 수 있는 간단한 블록 코딩 플랫폼을 활용한 경우가 많았다. 또한 교과 간의 융합적인 모습보다는 한 과목에서 국한된 머신러닝 알고리즘을 적용한 경우가 더러 있었다.

본 연구에서는 공개된 데이터셋을 이용하여 가장 적합한 인공지능(AI) 모델을 구현하고 선택하는 과정을 포함하여 교수 학습 자료를 만들 때의 교사들이 참고할 만한 아이디어를 제공한 점, 이미지에서 데이터를 추출하고 데이터를 다루는 과정 및 화학 실험을 포함하는 탐구 활동을

통해서 인공지능 모델과 학습자 자신이 생각하는 pH를 비교해보는 과정을 포함시켜서 고등학교 과학 과목인 물리학과 화학 과목간의 융합과 탐구 활동을 연계시킨 점, 인공지능(AI) 교육에서 필수적인 데이터 리터러시 향상을 위한 프로그램을 개발하여 학생들의 데이터 리터러시 향상을 확인할 수 있었던 점에서 의의를 지닌다. 본 연구의 성과는 다음과 같다.

첫째, 과학 데이터 기반 인공지능(AI)·고등학교 과학 융합 교육 프로그램은 공개된 데이터셋으로부터 pH를 예측하는 인공지능 모델을 개발하고, 이를 탐색적 과학 데이터 분석 탐구 모형(ESDA 탐구 모형)에 적용하여 구성할 수 있었다.

둘째, 본 연구에서 개발된 과학 데이터 기반 인공지능(AI)·고등학교 과학 융합 교육 프로그램이 ‘학생들의 데이터 리터러시 향상에 긍정적인 영향을 줄 것이다’라는 가설을 토대로 수업을 진행한 결과 본 연구에서 개발된 교육 프로그램은 데이터 리터러시 향상에 효과적인 것으로 나타났다.

## 나. 제언

본 연구의 한계점과 추후 연구에 대한 제언은 다음과 같다. 첫째, 본 연구에서 개발된 프로그램은 물리학, 화학, 정보 과목의 내용 요소가 포함되어 있으므로 교과 내용에서 적용하기는 다소 어려운 부분이 있고, 고등학교의 수업량 유연화 부분에서 활용될 수 있을 것이다. 고등학교 학년 급이 올라갈수록 입시 부담에 대한 부분이 있으므로 고등학교 1, 2학년 급에 적용하는 것이 적절하다고 여겨진다. 둘째, 데이터 리터러시 향상을 보는 데에 집중되어 있으나 인공지능 리터러시 향상을 확인해 볼 수 있는 도구의 개발이 이루어진다면, 인공지능 융합 프로그램의 결과를

확인해 볼 수 있을 것으로 판단된다. 추후 인공지능 리터러시에 대한 보편화된 검사 도구가 개발된다면 본 연구에서 개발된 교육 프로그램을 적용 전과 후의 비교를 통해서 효과성을 확인해 볼 수 있을 것이다. 셋째, 실험 데이터에서 정확성을 요구하는 결과들에는 본 프로그램이 적용되기 힘들 것이다. 이미지 데이터는 기기의 성능 때문에 서로 다른 결과를 보일 수 있으므로, 데이터를 다루고 인공지능 모델을 만들어 예측하게 하는 수준으로 학생들에게 다루어야 할 것이지만, 추후 이 모델들을 활용하여 흡광도 비교 분석, 중화 적정 분석 등의 데이터를 기반으로 한 프로그램들이 개발되길 기대해 본다. 마지막으로 본 연구에서 개발된 인공지능 알고리즘을 기반으로 한 앱을 개발하는 과정까지 이루어지게 하는 교육 프로그램으로 확장한다면 과학 데이터를 기반으로 최적화된 알고리즘이 적용된 인공지능 앱 메이킹 교육까지 할 수 있는 좋은 계기가 될 수 있을 것이다.

## 참 고 문 헌

- 곽영순. (2021). 고교학점제와 2022 개정 교육과정에 대비한 과학과 융합선택과목 재구조화 방안 탐색. **대한지구과학교육학회지**, 14(2), 112-122.
- 권점례, 이광우, 신호재, & 김종윤. (2018). 2015 개정 교육과정에 따른 초, 중학교 교과 간 연계·융합 교육 적용 방안 연구. **한국교육학회 학술대회논문집**, 1-13.
- 김근재 (2019). 피지컬 컴퓨팅 도구를 활용한 메이커 교육 수업 모형 개발. **석사학위논문, 서울대학교**.
- 김대엽, 김영배. (2019). 4차 산업혁명 시대의 핵심 ICT 기술: 빅데이터, 인공지능, 클라우드 기술 동향. **정보처리학회지**, 26(1), 7-17.
- 김슬기, 전용주, 이현아, 김영애, & 김태영. (2021). 초·중등 AI 교육을 위한 데이터셋 분석 및 활용 제안. **한국컴퓨터교육학회 학술발표대회논문집**, 25(1 (A)), 55-58.
- 김슬기, 김태영.(2021).초·중등 AI 교육을 위한 데이터 리터러시 정의 및 구성 요소 연구. **정보교육학회논문지**, 25(5), 691-704.
- 김슬기, 신한나, 김태영.(2023).2022 개정 교육과정을 중심으로 한 교육용 데이터셋 활용 초·중학교 AI 교육 프로그램 개발.**한국컴퓨터교육학회 학술발표대회논문집**,27(1),133-136.
- 김준영, 한선관. (2022). 데이터 리터러시 신장을 위한 데이터과학 프로그램 개발 및 적용. **인공지능연구 논문지**, 3(2), 23-32.
- 김진형(2020), KAIST 김진형 교수에게 듣는 AI 최강의 수업. **매일경제신문사**.
- 김태령(Kim, Tae-Ryeong);한선관(Han, Sun-Gwan). (2020). 인공지능교육에 관한 초중등교사의 인식에 관한 연구. **교육논총**, 40(3), 181-204.
- 나일주 정현미 , (2001). 웹기반 가상교육 프로그램 설계를 위한 활동모형개발. **교육공학연구**, 17(2), 27-52.
- 대한민국 정부. (2019). 인공지능 국가 전략.
- 문셋별, 이원태, 허희욱.(2022).인공지능 융합 화학수업에 대한 고등학생의 인식과 만족도 분석.**한국컴퓨터교육학회 학술발표대회논문집**,26(1),113-116.
- 박상길, & 정진호. (2022). (비전공자도 이해할 수 있는) AI 지식 / 박상길

- 지음 ; 정진호 그림.
- 배화순. (2019). 데이터 리터러시의 사회과 교육적 함의. **시민교육연구**, 51(1), 95-120.
- 손미현. (2020). 지식정보처리역량 함양을 위한 데이터 기반 과학탐구 모형 개발, **박사학위논문, 서울대학교**
- 이경화 외. (2022). 2022 개정 수학과 교육과정 시안 개발 연구 토론회 자료집(이경화 외, 2022), **한국교육과정 평가원**
- 이소율, 이영준.(2021).고등학생을 위한 머신러닝 교육 플랫폼 활용 과학 융합교육 콘텐츠 개발.**한국컴퓨터교육학회 학술발표대회논문집,25(2(A)),135-136.**
- 이승철, 김태영. (2019). 컴퓨터 교육 분야에서 데이터 리터러시의 개념과 구성요소 탐색. **한국컴퓨터교육학회 학술발표대회논문집, 23(2), 33-36.**
- 이준행, 조정효, 채승철. (2021). 인공지능 융합교육을 위한 데이터 기반 교육자료 개발: 감쇠진동을 중심으로. **현장과학교육, 15(2), 121-136.**
- 조연수. (2022). 인공지능을 융합한 과학 수업이 중학생들의 인공지능에 대한 태도 및 데이터 리터러시 역량에 미치는 효과. **석사학위논문, 이화여자대학교**
- 초중등 인공지능(AI)교육 학교 적용 방안 연구보고서(임다미 외, 2021), **한국교육과정 평가원**
- 한송이(Songlee Han); 김태중(Taejong Kim). (2022). 국내 인공지능 교육 연구 동향 분석. **학습자중심교과교육연구, 22(13), 281-294. 10.22251/jlcci.2022.22.13.281.**
- 한정윤 and 허선영. (2021). 국내 AI 교육 프로그램 연구동향 분석:주제범위 문헌고찰 방법론을 적용하여. **정보교육학회논문지, 25(6), 879-890.**
- 홍석영, 한신, 김형범. (2020). 데이터 기반 STEAM 교육을 통한 문제 해결 과정 분석 : 대기대순환과 표층 해류 내용을 중심으로. **대한지구과학교육학회지, 13(3), 330-343.**
- 2022 개정 정보과 교육과정 시안 개발 연구 토론회 자료집(김자미 외, 2022), **한국교육과정 평가원**
- 2019 글로벌 소프트웨어 교육 컨퍼런스 자료집. **한국과학창의재단**
- 2022 개정 과학과 교육과정 시안 개발 연구 추진 과정과 내용체계 개선 방안 토론회 발제 자료집(신영준 외, 2022), **한국교육과정 평가원**
- Arends, R., & Castle, S. (1998). Learning to teach (Vol. 4). New York:

*McGraw-Hill.*

- Bae, H. (2019). Educational Implications of Data Literacy in Social Studies. *Theory and Research in Citizenship Education, 51(1), 95-120.*
- Bhargava, R., & D'Ignazio, C. (2015). Designing tools and activities for data literacy learners. *In Workshop on Data Literacy, Webscience.*
- Calzada Prado, J., & Marzal, M. (2013). Incorporating Data Literacy into Information Literacy Programs: *Core Competencies and Contents. Libri (København), 63(2), 123-134.*
- Creswell, J. W. (2014). 연구방법: 질적 · 양적 및 혼합적 연구의 설계(정종진 외 역). 서울: 시그마프레스.
- Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining data information literacy needs: A study of students and research faculty. portal: *Libraries and the Academy, 11(2), 629-657.*
- Dark-Lucifer. "LuciferML a semi-supervised machine learning Library by dark-lucifer". Diakses pada tanggal: 18 Agustus 2021. Diakses dari: <https://github.com/d4rk-lucif3r/LuciferML>
- Deahl, E. (2014). Better the Data You Know: Developing Youth Data Literacy in Schools and Informal Learning Environments. Master thesis. *MIT University.*
- D'Ignazio, C., & Bhargava, R. (2015). Approaches to building big data literacy. *In Proceedings of the Bloomberg data for good exchange conference.*
- Frank, M., Walker, J., Attard, J., & Tygel, A. (2016). Data Literacy-What is it and how can we make it happen?. *The Journal of Community Informatics, 12(3).*
- Gibson, P., & Mourad, T. (2018). The growing importance of data literacy in life science education. *American journal of botany, 105(12).*
- Jeng, W., & D'Ignazio, J. (2010). The central role of metadata in a science data literacy course. *Journal of Library Metadata, 10(2-3), 188-204.*
- Mandinach, E. B., & Gummer, E. S. (2012). Navigating the Landscape of Data Literacy: It Is Complex. *WestEd.*
- Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational Researcher. 42(1), 30-37.*
- Nelson, J. A., Kinder, A., Johnson, A. S., Hall, H. I., Hu, X., Sweet, D., ... &

- Harris, J. (2018). Differences in selected HIV care continuum outcomes among people residing in rural, urban, and metropolitan areas—28 US jurisdictions. *The Journal of Rural Health, 34(1), 63-70.*
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.*
- Prado, J. C., & Marzal, M. Á. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri, 63(2), 123-134.*
- Qin, J., & D'Ignazio, J. (2010). *Lessons learned from a two-year experience in science data literacy education.*
- Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in python: *Main developments and technology trends in data science, machine learning, and artificial intelligence. Information, 11(4), 193.*
- Riyantoko, P. A., Fahrudin, T. M., & Hindrayani, K. M. (2021). Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan Menggunakan Lucifer Machine Learning. *SENADA, 1(01), 12-18.*
- Shields, M. (2005). Information literacy, statistical literacy, data literacy. *IASSIST quarterly, 28(2-3), 6-6.*

[부록 1] 전문가 타당화 설문지

**‘인공지능(AI)·고등학교 과학 융합 프로그램의 개발 및 적용’**  
**연구의 전문가 타당화 설문지 (1차, 2차)**

안녕하십니까?


저는 서울대학교 사범대학 AI융합교육학과 석사과정 3학기에 재학 중인 노동규입니다. 본 설문지는 ‘인공지능(AI)·고등학교 과학 융합 프로그램의 개발 및 적용’ 연구에서 개발된 초기 수업 프로그램과 데이터 리터러시 역량 측정 도구에 대한 전문가 타당화 설문지입니다. 전문가로서 선생님의 검토 의견은 보다 나은 프로그램을 개발하는 데에 큰 도움이 될 것입니다.

본 설문지는 1. 전문가 인적사항, 2. 연구의 소개, 3. 프로그램의 타당도 검토, 4. 데이터 리터러시 역량 측정 도구 타당도 검토로 구성되어 있습니다.

질문에 응답하실 때 이해가 되지 않거나 추가적으로 설명이 필요한 부분은 연구자에게 질문하실 수 있습니다. 본 설문지의 응답 예상 소요 시간은 약 20분입니다.

전문가 인적사항에 작성해주시는 성함은 자료 식별용으로만 사용되며 논문에는 언급되지 않을 것입니다. 다만 전공 분야, 최종 학력, 소속, 경력 등의 정보는 논문에 언급될 예정입니다. 바쁘신 와중에 소중한 시간을 내어주셔서 진심으로 감사드립니다.

서울대학교 대학원 AI융합교육학과 석사과정  
노동규 올림

	연구 담당자 :
	연락처:



## 1. 전문가 인적사항

- 전공 분야 :
- 최종 학력 :
- 소속 :
- 경력 :

## 2. 연구의 소개

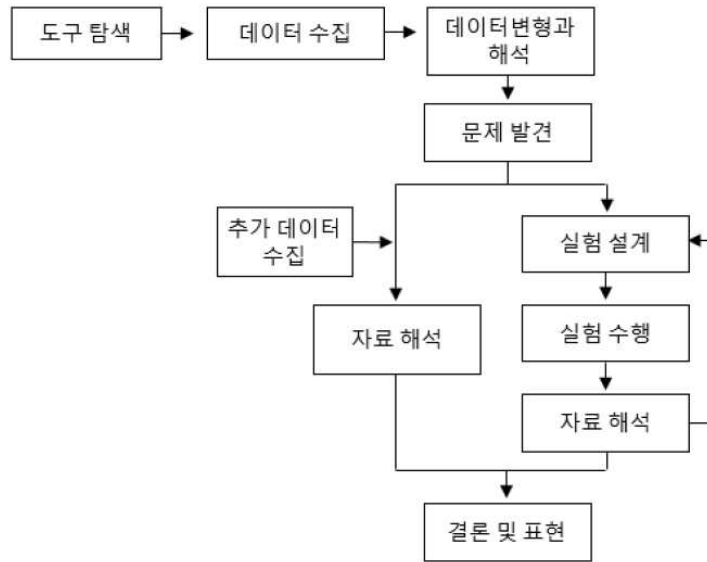
인공지능 융합 교육 컨퍼런스(KOSAF, 2019)에서는 인공지능 시대를 살아갈 모든 학생들에게 미래 사회에 적응할 수 있게 하기 위해서 인공 지능은 교육 현장에서 효과적으로 활용될 수 있어야 하고, 수학, 정보, 과학의 재구조화가 필요합니다.

인공지능을 활용한 예측과 과학탐구, 사회문제해결을 위한 인공지능과학탐구(가칭)와 같은 융합선택과목 개발의 필요성이 있으나, 전문성을 갖춘 교원이 충분하게 양성되지 못하고 있고, 선행 연구된 프로그램의 수가 극히 부족하여 융합선택 과목에서 인공지능과 관련된 융합과목은 신설되지 못하고 있습니다.(곽영순, 2021).

이에 본 연구에서는 만능 지시약의 색 변화 이미지로부터 데이터를 얻는 과정을 수행하고, 공개된 데이터셋으로부터 인공지능 기술 중 지도학습의 회귀 모델을 만들어 pH 예측을 목적으로 하는 인공지능·고등학교 과학 융합 프로그램을 개발하여 학생들의 데이터 리터러시 역량에 미치는 효과를 탐색해 보고자 합니다.

## 3. 탐색적 과학 데이터 분석 탐구 모형(ESDA 탐구모형)

본 연구의 프로그램이 적용된 과학 데이터 분석 탐구 모형(ESDA 탐구 모형)은 다음과 같습니다.



<그림 1> ESDA 탐구 모형(손미현, 2020)

또한 각 단계별 수업 전략은 <표 1>과 같습니다.

<표 1> ESDA 탐구 모형 설계 원리(손미현, 2020)

도구 탐색의 원리	
	<ul style="list-style-type: none"> <li>• 도구에서 측정할 수 있는 변인(센서)의 수를 중학생은 2-3개, 고등학생 이상은 그보다 복잡한 형태의 데이터를 활용할 수 있게 한다.</li> <li>• 익숙하고 간단한 형태로 데이터가 산출되는 도구를 제시한다.</li> </ul>
교수전략	(예) 엑셀이나 텍스트 파일 형태 <ul style="list-style-type: none"> <li>• 교사가 도구를 제시할 경우는 도구의 기본 사용법은 교사가 지도하고, 학생이 도구를 선택한 경우는 학생 스스로 인터넷 검색, 사용설명서 등을 통해 익히도록 한다.</li> <li>• 필요한 도구를 제작할 경우는 코딩, 3D 프린터 이용법 등의 제작 방법을 가르친다.</li> </ul>
환경구성	<ul style="list-style-type: none"> <li>• 학생들이 원활하게 정보를 검색하며 탐구를 진행할 수 있는지 컴퓨터와 인터넷 여부를 확인한다.</li> </ul>
실생활 데이터 수집의 원리	

교수전략	<ul style="list-style-type: none"> <li>• 학생들이 측정하고 싶은 장소를 선택하게 한다.</li> <li>• 수집되는 데이터의 측정 간격이나 측정 장소의 위치, 환경 등을 자세하게 기록하게 한다.</li> <li>• 교사는 주제에 관련된 기초 과학지식을 학생들이 익힐 수 있도록 지도한다.</li> <li>• 데이터베이스의 사용법을 익히게 한다.</li> </ul>
환경구성	<ul style="list-style-type: none"> <li>• 센서를 설치하는 곳에 인터넷 여부를 확인한다.</li> <li>• 데이터베이스를 사용하기 위한 사전 준비사항 등을 미리 확인한다.</li> </ul>
<b>데이터 변형의 원리</b>	
교수전략	<ul style="list-style-type: none"> <li>• 수집된 데이터가 정확한지, 연속적으로 누적되어 있는지 확인하도록 한다.</li> <li>• 데이터를 그래프 또는 도표로 변형할 수 있는 스프레드시트 프로그램 또는 데이터 모델링 프로그램에서 필요한 기능을 위주로 사용법을 가르친다.</li> </ul> <p>(예) 엑셀, 지오지브라</p>
환경구성	<ul style="list-style-type: none"> <li>• 다양한 그래프의 쓰임새와 예시를 제시한다.</li> <li>• 많은 양의 데이터를 변형하고 분석할 수 있을 만큼 성능이 좋은 컴퓨터를 준비한다.</li> <li>• 컴퓨터는 1~2인당 1대씩 준비하여 학생들이 직접 실습할 수 있도록 한다.</li> </ul>
<b>데이터 해석의 원리</b>	
교수전략	<ul style="list-style-type: none"> <li>• 도표나 그래프를 이용하여 상관관계에 대한 내용을 설명할 수 있게 한다.</li> <li>• 변인 사이의 관계를 서술할 때 기존 학습된 지식, 인터넷으로 검색한 과학지식 등과 연관 지어 설명할 수 있게 한다.</li> <li>• 정보원에 대한 교육, 정보 검색의 원리 교육, 정보 검색의 전략 등을 미리 학습시킨다.</li> <li>• 실제 데이터를 이용한 통계, 그래프 등을 이용하여 데이터에서 의미를 찾아내는 활동을 연습하게 한다.</li> </ul> <p>(예) trends.google.co.kr, www.gapminder.org/tools/ 등의 사이트를 이용하여 데이터 해석 및 추론을 연습할 수 있도록 한다.</p>
환경구성	<ul style="list-style-type: none"> <li>• 학생들이 원활하게 정보를 검색하며 탐구를 진행할 수 있는지 컴퓨터와 인터넷 여부를 확인한다.</li> </ul>

### 문제 구체화의 원리

교수전략	<ul style="list-style-type: none"><li>• 데이터 해석 과정에서 발생한 질문 중 인터넷 검색, 자료 조사 등을 통해 답을 찾을 수 있는 문항은 문제 선정에서 제외한다.</li><li>• 실험이나 활동을 통해 문제를 해결할 수 있도록 구체적으로 문제를 서술하게 한다.</li><li>• 데이터를 해석한 내용과 적힌 질문을 구체적으로 서술하게 한다.</li></ul>
환경구성	<p>(예) 정확하게 어떤 점이 궁금한가? 왜 궁금증이 생겼는가?</p> <ul style="list-style-type: none"><li>• 학생들이 원활하게 정보를 검색하며 탐구를 진행할 수 있는지 컴퓨터와 인터넷 여부를 확인한다.</li></ul>

### 문제 해결의 원리

교수전략	<ul style="list-style-type: none"><li>• 데이터 분석 과정에서 찾아낸 결과가 다른 상황에서도 적합한지 추가적인 데이터를 수집하게 한다.</li><li>• 데이터를 수집할 때는 탐구 문제를 고려하여 측정 간격과 기간을 정할 수 있도록 지도한다.</li><li>• 통제 변인과 조작 변인을 고려하여 데이터를 수집한다.</li><li>• 수집한 데이터의 결과와 과학적 지식을 연계하여 해석할 수 있도록 한다.</li><li>• 충분한 논의시간을 확보하여 학생들이 깊이 있는 논의를 진행할 수 있게 한다.</li></ul>
환경구성	<ul style="list-style-type: none"><li>• 학생들이 원활하게 정보를 검색하며 탐구를 진행할 수 있는지 컴퓨터와 인터넷 여부를 확인한다.</li><li>• 데이터 수집 단계의 환경 조건과 동일</li></ul>

### 표현과 공유의 원리

교수전략	<ul style="list-style-type: none"><li>• 인터넷 검색 또는 탐구 결과 새롭게 알게 된 사실이나 지식을 웹문서로 기록하고, 쉽게 공유하게 한다.</li><li>• 표현 수단이 되는 웹문서나 소프트웨어는 정해진 프레임이 있어 사용 방법이 간편하고, 다양한 형태의 파일을 포함할 수 있는 것을 활용한다.</li></ul> <p>(예)sites.google.co.kr 이나 망고보드, x-mind zen 등의 소프트웨어를 이용한다.</p> <ul style="list-style-type: none"><li>• 효과적으로 정보를 전달할 수 있도록 다양한 형태의 자료를 활용할 수 있도록 안내한다.</li><li>• 효과적인 표현법의 전략을 익히고, 모방을 통해 연습하게 한다.</li></ul>
환경구성	<ul style="list-style-type: none"><li>• 표현 수단이 되는 웹문서나 소프트웨어를 미리 설치하거나 계정을 만들도록 한다.</li></ul>

- 사용하는 컴퓨터의 사양에서 활용 가능한 형태의 소프트웨어인지 확인하도록 한다.

본 연구에서 개발된 인공지능(AI) · 고등학교 과학 융합 프로그램 초안은 <표 2>와 같고, 주 활동지는 Google Colab 텍스트 코딩 페이지로 이루어져 있습니다. 학생들은 이 코드를 실행하면서 수업에 참여하게 됩니다.

<표 2> ESDA 탐구 모형이 적용된 인공지능(AI) · 과학 융합 프로그램 초안

단계	학습자 활동	데이터 리터러시
(1) 도구 탐색	-교사가 탐구 소재(pH 예측)을 선정함. -colab에서 python 텍스트 코딩의 도구 사용법을 익힘	데이터 관리
(2) 데이터 수집	- 여러가지 물질들의 만능 지시약 실험으로 얻은 여러가지 시약의 이미지 data를 수집하여 padlet에 공유함. - 저장된 이미지로부터 colab에서 python의 PIL 라이브러리를 이용하여 RGB데이터를 수집하고 이미지가 데이터가 될 수 있음을 이해함.	데이터 수집 데이터 이해
(3) 데이터 변형과 해석	- 만능지시약 실험 결과 pH와 이미지의 RGB데이터를 스프레드 시트에 정리함	데이터 관리
(4) 문제 발견과 수립	- 만능지시약의 색 변화를 학습하면 pH를 예측할 수 있을 것이다& 라는 문제를 수립	데이터 해석 및 평가
(5) 귀납적 탐구	- Kaggle의 pH 예측 big data를 기반으로 머신러닝 회귀 모델을 만들음. - Kaggle의 pH 예측 big data를 기반데이터 시각화 코드 구현	데이터 활용 데이터 시각화
(6) 결론 및 표현	- 머신러닝 모델로부터 pH를 예측해보고 각 물질들의 [H3O+]를 구하는 과정을 수행함.	데이터 해석 및 평가



<그림 2> [Colab용 텍스트 코딩 활동지 QR코드](https://bit.ly/3Wry8Xc) (제작중) <https://bit.ly/3Wry8Xc>

#### 4. 탐색적 과학 데이터 분석 탐구 모형(ESDA 탐구모형)이 적용된 인공지능 고등학교 과학 융합 프로그램 개발 타당도 검토

##### 1) 프로그램 전반에 대한 타당도

- 인공지능(AI) · 고등학교 과학 융합 프로그램에 대한 타당성을 묻는 문항입니다. 질문을 읽고 해당하는 곳에 √(체크) 표시하여 주시기 바랍니다.

구분		전혀 그렇지 않다	그렇지 않다	그렇다	매우 그렇다
타당성	본 프로그램은 ESDA 탐구 모형의 설계 원리가 적용된 데이터 리터러시 향상을 위한 인공지능(AI) · 고등학교 과학 융합 프로그램으로 타당하다.	①	②	③	④
설명력	본 프로그램은 데이터 리터러시 향상을 위한 각 단계를 잘 설명하고 있다.	①	②	③	④
유용성	본 프로그램은 데이터 리터러시 향상을 위한 인공지능(AI) · 고등학교 과학 융합 프로그램으로 유용하게 활용될 수 있다.	①	②	③	④
이해도	본 프로그램은 데이터 리터러시 향상을 위한 인공지능(AI) · 고등학교 과학 융합 프로그램으로 적용시 필요한 요소들이 이해하기 쉽게 표현되어 있다.	①	②	③	④
보편성	본 프로그램은 데이터 리터러시 향상을 위한 인공지능(AI) · 고등학교 과학 융합 프로그램으로 보편적인 고등학생들을 위해 적용될 수 있다.	①	②	③	④
기타의견					

2) 데이터 리터러시 역량 측정 도구 타당도

- 관련 선행 연구와 수업 프로그램에서 얻고자 하는 역량을 추가하여 데이터 리터러시 역량 측정 설문 문항을 도출하였습니다.
- 다음의 데이터 리터러시 역량 측정 도구가 타당한지 판단하여 해당하는 곳에  $\surd$ (체크) 표시하여 주시기 바랍니다.

데이터 리터러시	문항	타당도			
		전혀 그렇지 않다	그렇지 않다	그렇다	매우 그렇다
데이터 이해	1. 나는 데이터에 다양한 종류가 있음을 이해하고 있다(조연수, 2022).	①	②	③	④
	2. 나는 데이터에 각 행과 열이 의미하는 바를 이해할 수 있다(송유경, 2021).	①	②	③	④
	의견 :				
데이터 수집	3. 나는 필요한 데이터가 없을 때 새로운 방법을 사용하여 데이터를 얻어낼 수 있다(조연수, 2022).	①	②	③	④
	4. 나는 데이터의 출처를 평가하고 선택할 수 있다(조연수, 2022).	①	②	③	④
	의견 :				
데이터 관리	5. 나는 데이터를 적절한 기준에 따라 나누어 저장할 수 있다(조연수, 2022).	①	②	③	④
	6. 나는 데이터 분석을 위한 준비를 하고 적절한 도구를 사용하여 분석할 수 있다(조연수, 2022).	①	②	③	④
	의견 :				
데이터 해석 및 평가	7. 나는 서로 다른 두 종류의 데이터가 어떻게 해석되는지 이해하고 있다(조연수, 2022).	①	②	③	④

	8. 나는 데이터 해석을 위해 모델을 설계할 수 있다(송유경. 2021 세부 항목 변경).	①	②	③	④
	의견 :				
데이터 활용	9. 나는 데이터로부터 새로운 통찰을 발견하고 데이터를 창의적으로 해석할 수 있다(송유경. 2021 세부 항목 변경).	①	②	③	④
	10. 나는 데이터 해석을 위해 만든 모델을 활용할 수 있다(신규 개발).	①	②	③	④
	의견 :				
데이터 시각화	11. 나는 표, 차트, 그래프 등 여러 시각화 방식의 특징을 이해하고 있다(송유경. 2021).	①	②	③	④
	12. 나는 필요에 따라 표, 차트, 그래프 등 여러 시각화 방식 중 적절한 것을 선택하여 데이터를 시각화 할 수 있다(송유경. 2021).	①	②	③	④
	의견 :				

- 귀중한 시간 내주셔서 감사합니다 -



단계	학습자 활동	데이터 리터러시
(1) 도구 탐색	<ul style="list-style-type: none"> <li>- Colab에서 python 텍스트 코딩의 기초 문법을 익히게 함.</li> <li>- pandas 라이브러리를 import하여 데이터를 저장하고 출력하는 방법을 익히게 함.</li> <li>-교사가 탐구 소재(pH 예측)를 다룰 것임을 알림.</li> </ul>	데 이 터 관리
(2) 데이터 수집	<ul style="list-style-type: none"> <li>- 여러 가지 용액(10가지 이상)중 개인별로 1가지씩 선택하고, 이를 홈판에 넣고 만능 지시약 실험으로 색 변화를 관찰함.</li> <li>- 개인별 디지털 도구(태블릿, 스마트폰)등으로 색 변화를 촬영하고 이미지를 패들릿(padlet)링크로 공유함.</li> <li>- 만능 지시약 색변화의 사진을 구글 드라이브에 저장하고 colab에서 python의 PIL 라이브러리를 이용하여 RGB데이터를 수집함</li> </ul>	데 이 터 수집 데 이 터 이해
(3) 데이터 변형과 해석	<ul style="list-style-type: none"> <li>- 용액의 종류와 눈으로 본 pH 예측값 및 python 이미지 데이터 수집을 통해 얻어진 RGB 데이터를 입력하고, 기종별 편차를 확인하기 위해서 촬영 기종등을 정리하게 함.</li> <li>- 각 데이터들의 차이 유무와 용액의 pH 변화가 올바른지 토의하여 해석하는 과정을 거침.</li> </ul>	데 이 터 관리
(4) 문제 발견과 수립	<ul style="list-style-type: none"> <li>- 색변화 data를 기반으로 해석한 내용을 발표하고 공유함.</li> <li>- 비슷한 색으로 변하는 용액들이 있으므로 만능 지시약의 색 변화를 학습하면 pH를 예측할 수 있을 것이다 라는 문제를 수립.</li> </ul>	데 이 터 해석 및 평가
(5) 귀납적 탐구	<ul style="list-style-type: none"> <li>- pH-recognition 데이터 셋의 데이터가 어떤 분포를 보이고 있는지를 시각화하는 과정을 수행하고 출력함.</li> <li>- pH 예측이 가능한 AI 모델을 머신러닝 회귀(regression) 알고리즘을 사용하여 구현함.</li> </ul>	데 이 터 시각화 데 이 터 활용

(6) 결론 및 표현	- 자신이 데이터를 수집한 용액의 RGB 값을 대입하여 머신러닝 모델로부터 pH를 예측해보고 각 물질들의 $[H_3O^+]$ 를 구하는 과정을 모둠별로 수행함.	데 이 터 해 석 및 평 가
-------------	---	-----------------------

데이터 리터러시	문항	타당도			
		전혀 그렇지 않다	그렇지 않다	그렇다	매우 그렇다
데이터 이해	1. 나는 데이터에 각 행과 열이 의미하는 바를 이해할 수 있다.(조연수, 2022).	①	②	③	④
	2. 나는 데이터에 각 행과 열이 의미하는 바를 이해할 수 있다.(송유경, 2021).	①	②	③	④
	의견 :				
데이터 수집	3. 나는 필요한 데이터가 없을 때 새로운 방법을 사용하여 데이터를 얻어낼 수 있다(조연수, 2022).	①	②	③	④
	4. 나는 데이터의 출처를 평가하고 적절한 데이터를 선택할 수 있다(조연수, 2022).	①	②	③	④
	의견 :				
데이터 관리	5. 나는 데이터를 적절한 기준에 따라 나누어 저장 또는 보관할 수 있다(조연수, 2022).	①	②	③	④
	6. 나는 데이터 분석을 위한 준비를 하고 적절한 도구를 사용하여 분석할 수 있다(조연수, 2022).	①	②	③	④
	의견 :				
데이터 해석 및 평가	7. 나는 서로 다른 두 종류의 데이터가 어떻게 해석되는지 이해하고 있다(조연수, 2022).	①	②	③	④
	8. 나는 데이터 해석 및 적용을 위해 인공지능 모델(분류, 회귀)의 성능을 확인할 수 있다(송유경, 2021 세부 항목 변경).	①	②	③	④
	의견 :				
데이터 활용	9. 나는 데이터로부터 새로운 사실이나 의	①	②	③	④

	견을 발견하고 데이터를 창의적으로 해석할 수 있다(송유경. 2021 세부 항목 변경).				
	10. 나는 데이터 해석을 위해 만든 모델을 활용할 수 있다(신규 개발).	①	②	③	④
	의견 :				
데이터 시각화	11. 나는 표, 차트, 그래프 등 여러 시각화 방식의 특징을 이해하고 있다(송유경. 2021).	①	②	③	④
	12. 나는 필요에 따라 표, 차트, 그래프 등 여러 시각화 방식 중 적절한 것을 선택하여 데이터를 시각화 할 수 있다(송유경. 2021).	①	②	③	④
	의견 :				

[부록 2] 데이터 리터러시 사전, 사후 검사 결과

- 일시 : 2022년 11월 19일 및 2023년 2월 7일
- 방법 : 구글 설문조사
- 대상 : 과학거점학교 2학년 학생 및 I 고등학교 1학년 학생 총 26명

영역	문항 번호	사전 검사		사후 검사	
		평균	표준편차	평균	표준편차
데이터 이해	1	3.43	0.89	3.96	0.87
	2	3.46	0.89	4.00	1.27
데이터 수집	3	2.86	0.92	3.67	1.03
	4	3.50	0.75	3.93	0.75
데이터 관리	5	3.29	0.70	3.80	0.83
	6	3.07	0.83	4.11	0.92
데이터 해석 및 평가	7	3.14	0.93	3.67	0.81
	8	2.64	1.11	3.89	1.08
데이터 활용	9	2.93	1.04	3.74	0.89
	10	2.79	0.94	3.33	1.25
데이터 시각화	11	3.75	0.91	3.54	0.70
	12	3.43	0.88	3.44	0.79
합계		3.19	0.64	3.74	0.27

## [부록 3] 과학 데이터 기반 인공지능(AI) 고등학교 과학 융합 프로그램 Colab 활동지

### < 1차시 : 도구 탐색 >

#### 1차시. 파이썬 기초 문법 익히기 <도구 탐색>

구글 Colab : 주피터 노트북(웹 브라우저에서 파이썬 코드 실행이 가능한 툴, 설치 필요)의 구글 클라우드 버전 설치 없이 구글 클라우드 웹 사이트에 접속해서 무료로 사용 가능합니다.

지원 되는 웹 브라우저는 구글 크롬, 파이어 폭스, 사파리 등이며 인터넷 익스플로러는 현재 지원 되지 않습니다. 사용하는 웹 브라우저가 지원되지 않는 웹 브라우저라면 구글 크롬 브라우저를 설치하는 것을 권장합니다.

print("Hello World") 를 코드에 적으면 어떻게 출력될까요?

```
▶ ### 아래에 코드를 작성합니다. ###
```

print(Hello World) 를 코드에 적으면 어떻게 출력될까요?

```
▶ ### 아래에 코드를 작성합니다. ###
```

Python에서는 숫자와 문자를 구분합니다.

다음 코드를 보고 무엇이 출력될지 이야기해봅시다.

```
a = 3+1  
b = a+1  
c = b+1  
print(c)
```

```
[ ] ### 아래에 코드를 작성합니다. ###
```

다음 코드를 실행해보세요. 무엇을 하는 것 같나요?

```
[ ] from google.colab import drive  
drive.mount('/content/drive')
```

Drive already mounted at /content/drive: to attempt to forcibly remount, call drive.mount("/content/drive", force\_remount=True).

csv파일을 불러와서 dataframe으로 저장해 보겠습니다.

다음 링크의 data를 불러와서 압축을 풀어보세요.

- 캐글 데이터 pH-recognition : <https://www.kaggle.com/datasets/robjan/ph-recognition>

#### Data Explorer

Version 1 (8.81 kB)

▣ ph-data.csv

```
[ ] import pandas as pd #pandas library를 불러옵니다. pd로 지칭합니다.

[ ] df = pd.read_csv('/content/ph-data.csv') # ' ' 사이에 경로를 저장하여 df라는 명칭으로 저장시킵니다. #다음표 안의 주소는 바뀌어야합니다.
```

다운 받은 data는 df에 데이터프레임으로 저장되었습니다.

```
[ ] df.info() #df에 저장된 데이터의 개요를 출력합니다.

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 653 entries, 0 to 652
Data columns (total 4 columns):
#   Column  Non-Null Count  Dtype
---  -
0    blue    653 non-null    int64
1    green   653 non-null    int64
2    red     653 non-null    int64
3    label   653 non-null    int64
dtypes: int64(4)
memory usage: 20.5 KB
```

```
df.head() #df에 저장된 데이터의 첫부분 5줄을 출력합니다.
```

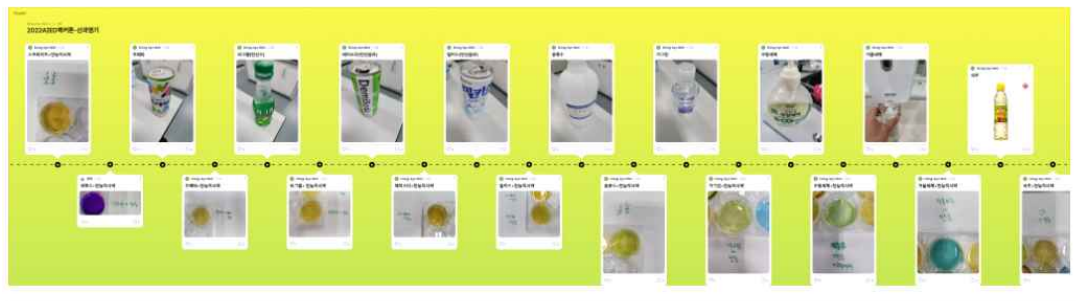
	blue	green	red	label
0	36	27	231	0
1	36	84	250	1
2	37	164	255	2
3	22	205	255	3
4	38	223	221	4

## < 2차시 : 데이터 수집 >

2차시. 만능지시약 실험으로 얻은 여러가지 시약의 이미지로부터 colab에서 python의 PIL 라이브러리를 이용하여 RGB데이터를 수집해 보겠습니다. <데이터 수집>

- (1) pH를 측정하고자 하는 용액에 만능 지시약을 2~3방울 떨어뜨린 후 색 변화를 관찰한다.
- (2) 측정하고자 하는 용액의 사진과 수용액의 색변화를 사진으로 촬영하여 이미지를 다음 패들릿 링크에 저장합니다.

[용액과 만능지시약의 색변화 공유 링크](#)



(3) 만능 지시약을 떨어뜨린 용액 사진의 중심 부분을 캡처하여 구글 드라이브에 저장합니다. 가급적 한 색깔로 보이는 부분이어야 합니다.

(4) PIL 라이브러리를 이용하여 RGB 데이터를 수집합니다. PIL 라이브러리는 이미지에서 R, G, B data를 추출합니다. (이미지가 데이터화 되는 것입니다.)

```

from PIL import Image
import os
import cv2

[ ] im2 = Image.open('/content/drive/MyDrive/22 AIED HAC/석회수만능색.PNG') #다음표 안의 주소는 바꿔야합니다.
rgb2_im = im2.convert('RGB')
r, g, b = rgb2_im.getpixel((1, 1))

print(b, g, r) #이미지에서 B,G,R data를 추출

99 2 46

```

### < 3차시 : 데이터의 변형 및 해석 >

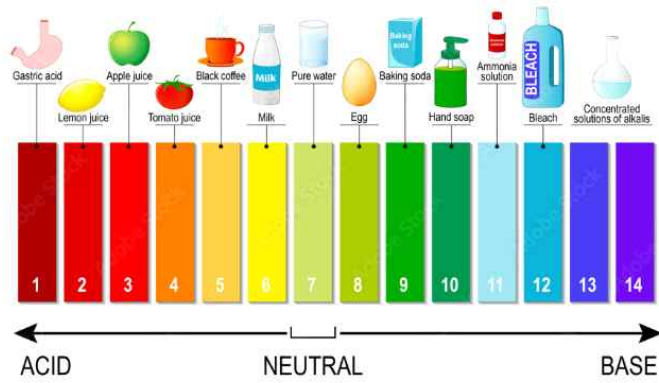
3차시. 만능 지시약의 실험으로 얻은 여러가지 시약의 실험 결과로부터 pH와 이미지의 RGB data를 저장합니다. <데이터의 변형 및 해석>

연번	용액의 종류	눈으로 본pH	수집된 data의 정리			촬영 기종
			R 값	G 값	B 값	
1	석회수	13	47	3	101	
2	술의 눈	8	111	103	67	
3	포카리	6	191	184	156	
4	토레타	6	134	132	103	
5	베이킹 소다	11	103	125	138	
6	주방 세제	8	179	179	140	
7	밀키스	7	166	167	172	
8	비누	9	114	128	84	
9	샤이다	6	163	160	134	
10	비누	9	108	130	77	
11	베이킹 소다	11	102	125	138	
12	염산	2	126	112	144	아이폰13프로
13	갈아만든 배	3	202	202	197	노트 9
14	밀키스	7	166	167	172	iphone 12 pro
15	세제	9	114	128	84	아이폰12 프로
16	가그린	5	123	106	63	iphone 12 pro
17	켈치스	3.5	155	95	63	아이폰12 프로
18	아이시스8.0	8	76	108	81	lphone 12 pro
19	비눗물	7	101	107	82	lphone 12 pro
20	아이시스8.0	8	137	152	133	갤럭시s21
21	씨그램	4	152	146	57	갤럭시s21
22	아세트산나트륨	7	96	116	93	아이폰 13 프로
23	베이킹소다	12	139	124	74	아이폰 13 프로
24	포카리스웨트	4	189	136	113	구글 픽셀 6 pro
25	데미소다	4	140	56	11	구글 픽셀 6 pro
26	토레타	4	200	123	76	구글 픽셀 6 pro
27	소금물	9	43	72	48	구글 픽셀 6 pro



## < 4차시 : 문제의 발견과 수립 >

4차시. 만능 지시약의 색변화 data를 기반으로 이야기를 나누어 봅시다. 비슷한 경우에는 비교할 수 있을까요? <문제의 발견과 수립>



## < 5차시 귀납적 탐구 >

<귀납적 탐구>

```
[ ] import matplotlib.pyplot as plt
    from mpl_toolkits.mplot3d import Axes3D # 3차원으로 그래프 그리는 라이브러리를 불러옵니다.
    import numpy as np
```

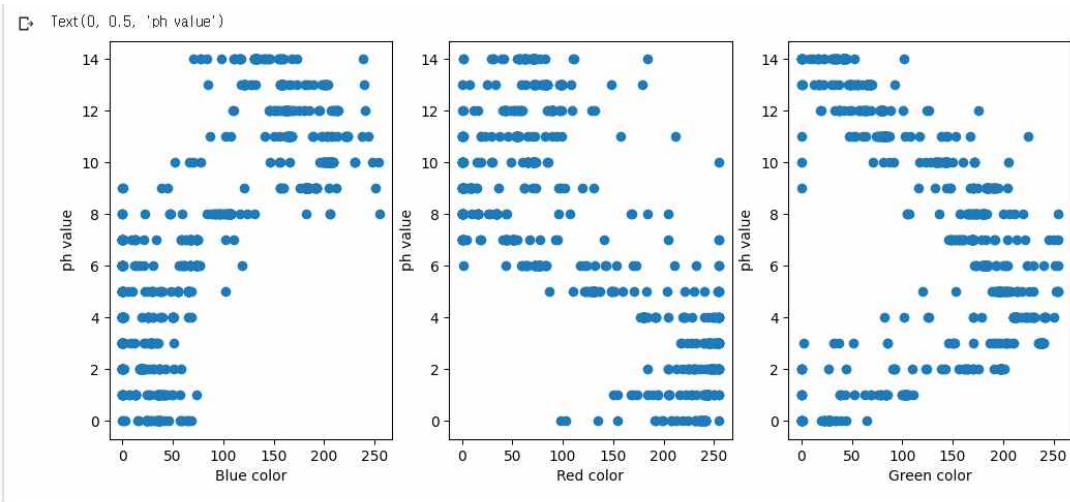
```
[ ] dataset = pd.read_csv('/content/ph-data.csv')
    X = dataset.iloc[:, :-1]
    y = dataset.iloc[:, -1]
```

(5-2) pH data를 기반으로 각 RGB값을 3차원 그래프로 시각화 해 보겠습니다.

```
[ ] ## pH-data 파일의 B, G, R 값을 시각화
    plt.figure(figsize=(12,5))
    #plotting blue spectrum with ph
    plt.subplot(1,3,1)
    plt.scatter(X.blue,y)
    plt.xlabel('Blue color')
    plt.ylabel('ph value')

    #plotting red spectrum with ph
    plt.subplot(1,3,2)
    plt.scatter(X.red,y)
    plt.xlabel('Red color')
    plt.ylabel('ph value')

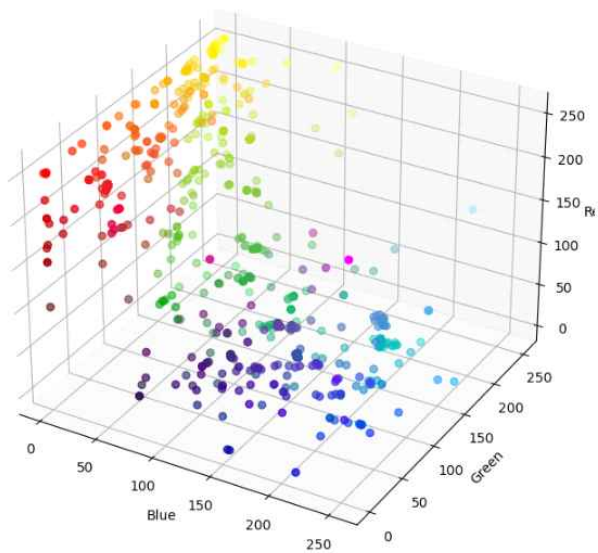
    #plotting green spectrum with ph
    plt.subplot(1,3,3)
    plt.scatter(X.green,y)
    plt.xlabel('Green color')
    plt.ylabel('ph value')
```



```
colors = np.array([df.red, df.green, df.blue]).T
```

```
fig = plt.figure(figsize=(8, 8))
ax = fig.add_subplot(111, projection='3d')
x = df.blue
y = df.green
z = df.red
ax.scatter(x, y, z, c=colors/255.0, s=30)
ax.set_title("Color distribution")
ax.set_xlabel("Blue")
ax.set_ylabel("Green")
ax.set_zlabel("Red")
plt.show()
```

Color distribution



링크 텍스트(5-1) pH-data를 이용하여 pH값을 예측을 학습 모델들을 분류 알고리즘 중에 가장 성능이 뛰어난 시 모델을 찾아봅니다.

- 캐글 데이터 pH-recognition : <https://www.kaggle.com/datasets/robian/ph-recognition>

```
[ ] !pip install lucifer-ml #lucifer-ML 설치
```

```
[ ] from luciferml.supervised.regression import Regression
```

```
dataset = pd.read_csv('/content/ph-data.csv')
X = dataset.iloc[:, :-1]
y = dataset.iloc[:, -1]
```

```
[ ] regressor = Regression(
    predictor=["all"],
    cv_folds=2,
    # tune=True,
    optuna_n_trials=2,
)
regressor.fit(X, y)
```

Results Below

	Name	R2 Score	Mean Absolute Error	Root Mean Squared Error	KFold Accuracy	Model
0	Linear Regression	71.617015	1.813891	2.272352	69.562803	LinearRegression()
1	Stochastic Gradient Descent Regressor	71.595053	1.816497	2.273231	69.502463	SGDRegressor()
2	Kernel Ridge Regressor	-208.611412	7.142015	7.492945	-217.230725	KernelRidge()
3	Elastic Net Regressor	63.992122	2.005355	2.559442	61.063814	ElasticNet()
4	Bayesian Ridge Regressor	71.651771	1.812119	2.270960	69.572223	BayesianRidge()
5	Support Vector Regressor	94.425714	0.656434	1.007027	93.519444	SVR()
6	K-Neighbors Regressor	95.268589	0.462595	0.927773	95.513292	KNeighborsRegressor()
7	Decision Trees Regressor	85.604437	0.621183	1.618307	93.925635	DecisionTreeRegressor()
8	Random Forest Regressor	92.021006	0.517095	1.204816	95.597882	(DecisionTreeRegressor(max_features=1.0, rando...
9	Gradient Boost Regressor	91.974906	0.553783	1.208291	95.446798	((DecisionTreeRegressor(criterion='friedman_ms...
10	AdaBoost Regressor	87.292631	0.970283	1.520458	94.277068	(DecisionTreeRegressor(max_depth=3, random_sta...
11	Bagging Regressor	89.026580	0.546904	1.412920	95.227493	(DecisionTreeRegressor(random_state=1616662711...
12	Extra Trees Regressor	96.021097	0.411734	0.850801	95.767464	(ExtraTreeRegressor(random_state=921055523), E...
13	LightGBM Regressor	92.765859	0.530159	1.147202	95.309390	LGMRRegressor()
14	XGBoost Regressor	90.233054	0.523996	1.332987	94.960429	XGBRegressor(base_score=None, booster=None, ca...
15	Catboost Regressor	95.539722	0.486899	0.900798	95.792678	<catboost.core.CatBoostRegressor object at 0x7...
16	Multi-Layer Perceptron Regressor	83.979030	1.188182	1.707226	81.321958	MLPRegressor()

Completed LuciferML Run [ ✓ ]

Saved Best Model to lucifer\_ml\_info/best/regression/models/Catboost\_Regressor\_1665770459.pkl and its scaler to lucifer\_ml\_info/best/regression/scalers/Catboost\_Regr

Time Elapsed : 9.67 seconds

```
print(regressor.best_regressor.name)
print(regressor.best_regressor.r2_score)
print(regressor.best_regressor.rmse)
print(regressor.best_regressor.mae)
print(regressor.best_regressor.kfold_acc)
```

```
Catboost Regressor
95.53972177189756
0.9007982214013077
0.486899259444034
95.79267631670652
```

출력 결과 Catboost Regressor 가 가장 성능이 뛰어난을 알 수 있습니다.

아래 코드는 Catboost Regressor 알고리즘으로 pH를 예측하는 모델을 구현합니다.

```

❶ from sklearn import ensemble, model_selection #알고리즘 모델이 포함된 라이브러리를 가져옵니다.
from sklearn.metrics import mean_absolute_error, r2_score # MAE 성능 지표를 계산하는 라이브러리를 가져옵니다.
from catboost import CatBoostRegressor, cv, Pool # Catboost Regressor 알고리즘 라이브러리를 가져옵니다.

```

```

[] # y값에 label(pH값), X값에(blue, green, red)를 학습 데이터로 저장
y = dataset.label
X = dataset[['blue', 'green', 'red']]

```

```

[] # 학습 set과 목표 set로 data를 분리
X_train, X_test, y_train, y_test = model_selection.train_test_split(X, y,
                                                                    test_size=0.25, random_state=42)

```

```

[] # X값 학습 데이터 수
X_train.shape

(489, 3)

```

```

[] #Initiate a CatBoost Regressor model and train it
# CatBoost Regressor 모델을 사용하여 학습 시킴
CB_model = CatBoostRegressor(random_state=12)
CB_model.fit(X_train, y_train)

```

```

❷ #학습 모델의 성능을 확인
#Prediction on the test data
y_pred = CB_model.predict(X_test)
#Calculation of Mean Absolute Error
mae = mean_absolute_error(y_test, y_pred)
#Calculation of R Squared value
r2_val = r2_score(y_test, y_pred)
print('Mean Absolute Error of the model is: ', mae)
print('R Squared value is: ', r2_val)

```

```

□ Mean Absolute Error of the model is: 0.5080556814351926
R Squared value is: 0.958978920071152

```

CatBoost Regressor 모델을 사용하여 pH 예측 모델이 만들어 졌습니다.

## < 6차시 : 결론 및 표현 >

6차시. 머신러닝 모델로부터 pH 예측하고 [H3O+]를 구해보고 비교하기

<결론 및 표현>

(6-1) 용액의 만능지시약과의 색 변화 data를 저장 후에 예측 모델에 넣어 pH를 예측합니다.

```

[] CaOH2 = pd.DataFrame({'blue':[99], 'green':[2], 'red':[46]}) #자신이 고른 용액의 데이터를 넣습니다.

```

```

[] CB_model.predict(CaOH2) # 석회수 pH
array([13.34113198])

```

이미지 사진으로부터 예측된 석회수의 pH=13.34 입니다.

(6-2) pH = -log[H3O+] 이므로 예측된 모델로부터 여러가지 수용액의 [H3O+]를 구하고 산의 세기를 비교해 봅시다.

[활동지 링크](#)

(6-2) pH 비교 활동지

학번 : 10808

이름 :

(6-1) 만들어진 AI 모델의 성능 지표를 적어보세요.

```
Mean Absolute Error of the model is: 0.5120835253838303
R Squared value is: 0.9324838563149832
```

(6-2)  $\text{pH} = -\log[\text{H}_3\text{O}^+]$  이므로 예측된 모델로부터 유사한 색을 보이는 물질들의  $[\text{H}_3\text{O}^+]$ 을 구하고  $[\text{H}_3\text{O}^+]$ 의 크기를 비교해봅시다.

수용액	비누	포카리	밀키스	비누	베이킹소다	사이다	토레타	솔의눈		
예측된 pH	13.84	9.54	7	10.5	10.71	8.5	6	4.22		
$[\text{H}_3\text{O}^+]$	$10^{-13.84}$	$10^{-9.54}$	$10^{-7}$	$10^{-10.5}$	$10^{-10.71}$	$10^{-8.5}$	$10^{-6}$	$10^{-4.22}$	$10^{-4}$	

여러가지 수용액의  $[\text{H}_3\text{O}^+]$ 비교

## Abstract

# Development and application of artificial intelligence (AI) and high school science integrated education programs based on scientific data

Noh Dong kyu

Artificial Intelligence Integrated Education

The Graduate School

Seoul National University

The core ICT technologies of the Fourth Industrial Revolution are big data, artificial intelligence (AI), and cloud. Among them, AI technology is bringing changes in various fields and is expected to have a significant impact on the education sector. Therefore, AI integrated education is a hot topic in education in the era of the Fourth Industrial Revolution.

Prior research on AI integrated education has been rapidly increasing since 2020, and by school level, elementary school students accounted

for the highest proportion, while the proportion of prior research targeting high school students was the lowest. Therefore, there is a need for AI integrated education research targeting high school students. In addition, most of the AI integrated education practice activities only deal with classification models using supervised learning, so it is necessary to develop educational programs using regression or unsupervised learning, another area of AI.

There is a need to develop integrated elective courses such as prediction and scientific research using artificial intelligence, and artificial intelligence scientific research for solving social problems (tentative title), but the number of teachers with expertise and the number of previously researched educational programs are extremely low, so integrated courses related to artificial intelligence in integrated elective courses have not been established.

The 4th General Conference on Science Education is promoting the use of digital tools using AI and big data in the process of solving scientific problems through scientific exploration using big data, and there are many platforms providing datasets for AI around the world, and these datasets have the potential to be good teaching materials for solving real-world problems.

In this study, we developed a scientific data-based AI and high school science integrated program that includes the process of creating an AI model for the purpose of pH prediction by applying the regression model algorithm of supervised learning among the techniques of AI based on publicly available scientific datasets. In this course, a data-based scientific data analysis exploration model is applied to foster knowledge information processing capabilities, and students' data literacy is designed to be developed in each course, and a total of six classes were developed.

The developed scientific data-based artificial intelligence (AI) and high school science integrated program was consulted by three experts and three field teachers to enhance its validity through an internal validation process, and the pre- and post-surveys were analyzed using a paired sample t-test, showing an improvement in data literacy( $p < 0.01^{**}$ ).

This study is meaningful for developing and applying a scientific data-based AI model and a scientific data-based AI and high school science integrated program to improve students' data literacy. In addition, it is a program that can improve each element of data literacy according to the scientific data-based analysis exploration model (ESDA), and a cloud-based activity sheet is provided, so it is meaningful that we proposed an AI integrated education program that is easily accessible to teachers and students, and suggested a method that will help develop new programs using various scientific data.

**keywords : AI integrated Education Program, AI integrated Education Program with ESDA Model, Science Data-Driven Education Program, High School Science integrated Program, pH Prediction Education Program**

***Student Number : 2021-23621***