Journal of The Korean Association
of Information Education Vol. 27, No. 3,
June 2023, pp. 255-267

http://dx.doi.org/10.14352/jkaie.2023.27.3.255

초·중등 인공지능 교육을 위한 PISA 수학 맥락 중심의 데이터셋 개발

김슬기* · 김귀훈* · 김태영*
한국교원대학교*

요약

AI의 발전은 역사적으로 데이터셋과 깊은 관계가 있다. 최근 AI를 구성하는 데이터셋의 중요성이 강조됨에 따라 관련 연구가 많이 이루어지고 있지만 AI 교육 측면에서 데이터셋 관련 연구는 상대적으로 부족한 실정이다. 이에 본 연구는 학생들에게 유의미한 AI 교육용 데이터셋을 제공하기 위해 교수·학습 환경에서 맥락의 중요성을 확인하고, 컴퓨팅 사고력을 포함하는 PISA 2022 수학의 맥락을 중심으로 AI 교육에서 학생들이 선호하는 데이터셋의 맥락과 형태 및 교육용 도구를 확인하였다. 이를 바탕으로 AI 교육에 적합한 데이터셋 주제를 탐색하고 다양한 데이터셋을 합성, 수집, 수정 및 개발하였다. 또한 데이터셋의 적합성을 검증하기 위해 전문가 검토를 실시하고 그 결과를 반영하여 25종의 AI 교육용 데이터셋을 도출 및 배포하였다. 본 연구의 결과가 다양한 관점의 AI 교육을 위한 데이터셋 관련 연구에 기반이 되어 학생들의 AI 소양을 기르는데 도움이 될 수 있기를 기대한다.

키워드 : SW · AI교육, 데이터과학, 빅데이터, 데이터셋, 합성데이터셋

Development of PISA Mathematical Context-oriented Dataset for K-12 Artificial Intelligence Education

SeulKi Kim* · Kwihoon Kim* · Taeyoung Kim*
Korea National University of Education*

Abstract

The development of AI has historically been strongly tied to datasets. Recently, as the importance of datasets in AI has been emphasized, there has been a lot of related research, but there is a relative lack of research on datasets in the context of AI education. In order to provide students with meaningful datasets for AI education, this study identified the importance of context in the teaching-learning environment and identified the context, form, and educational tools of datasets preferred by students in AI education, focusing on the context of PISA 2022 mathematics, which includes computational thinking skills. Based on this, we explored dataset topics suitable for AI education and synthesized, collected, modified, and developed various datasets. We also conducted an expert review to verify the suitability of the datasets, and based on the results, we derived and distributed 25 datasets for AI education. We hope that the results of this study will serve as a basis for research on datasets for AI education from various perspectives and help students develop their AI competency.

Keywords : SW · AI Education, Data Science, Big Data, Dataset, Synthetic Dataset

교신저자 : 김태영(한국교원대학교)
논문투고 : 2023-03-23
논문심사 : 2023-04-06
심사완료 : 2023-05-15

1. 서론

전 세계는 인공지능(Artificial Intelligence, 이하 AI) 기술을 중심으로 한 디지털 기술의 발달로 급격한 사회적 변화를 겪고 있다. AI를 통해 창출되는 막대한 부가가치로 인해 기존 산업 및 일자리의 근본적인 구조가 변화하고 있으며 AI 역량이 국가와 기업의 핵심 경쟁력으로 강조되고 있다[14]. 교육은 미래 지향적인 것으로 미래를 살아갈 학생들의 역량을 길러주는 목적을 가지고 있다. 이에 최근 미래 사회를 살아갈 학생들이 가져야 할 핵심 역량으로 AI 역량과 디지털 소양을 길러줄 수 있는 교육의 사회적 요구가 커지고 있다[13].

이러한 사회적 요구에 따라 세계 각국은 다양한 정책을 통해 AI 교육을 지원하고 있다. 대한민국도 '인공지능 국가 전략'을 통해 일반 국민부터 AI 고급 인재에 이르는 전 국민 AI 교육을 위한 기반을 마련하였으며, '전 국민 AI·SW교육 확산 방안'을 통해 상세한 교육의 가이드라인을 제시하였다. 또한, '인공지능시대 교육정책 방향과 핵심과제' 발표를 통해 '인간', '시대', '기술'의 키워드를 중심으로 한 핵심 추진과제를 제시하였으며, 시대에 부합하는 핵심 역량을 기르기 위한 교육으로 AI 교육을 강조하였다[11,13,14].

교육부(2021, 2022)는 2022 개정 교육과정의 개정 중점 사항으로 '미래사회가 요구하는 역량 함양이 가능한 교육과정'을 제시하였으며 AI·소프트웨어 교육을 비롯한 디지털 기초 소양의 강화를 강조하였다. 그리고 2022 개정 교육과정을 통해 초등의 실과부터 중학교의 정보, 고등학교의 인공지능 기초 등의 교과를 중심으로 AI 교육의 내용체계와 성취기준을 제시하였다[18,19].

AI 기술은 본격적으로 명명되고 연구되기 시작한 1956년을 시작으로 기술의 한계로 인한 암흑기와 이를 극복한 황금기를 거쳐 발전해왔다. 최근 알파고를 기점으로 사회 전반의 다양한 분야에서 활용되고 있는 AI 기술인 딥러닝은 AI의 새로운 황금기를 불러왔으며 그 중심에는 폭발적으로 증가하고 수집과 활용이 용이해진 데이터와 수많은 데이터를 연산할 수 있는 컴퓨팅 하드웨어의 발전이 자리 잡고 있다[21].

최근에는 AI를 구성하는데 주재료로 활용되는 데이터셋의 중요성이 강조되고 있으며, 대한민국 정부(2020)는 '데이터 댐' 구축 및 활용과 같은 정책을 통해 양질의 데이터셋을 AI 산업 현장에서 활용할 수 있도록 지원하고 있다[12]. AI 교육의 측면에서도 데이터셋은 학생들이 다루게 되는 학습 문제 상황과 관련성이 높으며, 컴퓨팅 실습 활동이 이루어지는 교수·학습 환경에서 학생과 교사에 의해 직접적으로 활용되기 때문에 AI 교육 활동에 필수 요소로 강조되고 있다[9].

하지만, AI 교육 관련 연구 혹은 교육 현장에서 주로 사용되는 데이터셋은 외국에서 개발되어 많은 활용 사례를 가진 데이터셋으로 학생들의 삶과 관련성이 부족하다는 한계를 가지고 있다[8]. 이에 본 연구는 학생들의 삶과 가까운 맥락의 데이터셋을 연구하고 탐색, 수집 및 개발하여 AI 교육의 교수학습 환경에서 활용될 수 있는 유의미한 데이터셋을 제공하는데 도움을 주고자 하였다.

2. 연구 배경

2.1. AI와 데이터셋

AI를 구성하고 활용함에 있어 다양한 알고리즘이나 기술에 대한 이해와 함께 강조되는 것이 AI 학습의 주재료가 되는 데이터셋이다.

대한민국 정부(2020)는 AI 산업 활성화를 위한 정책으로 'D.N.A 생태계 강화'를 슬로건으로 제시하였다. 특히 AI를 위한 '데이터 댐' 구축과 데이터셋의 개방 및 공유의 목표를 제시하였으며, 2021년까지 190종, 4.8억 건 이상의 데이터를 개방하고 활용될 수 있도록 지원하였다[12]. 산학연이 함께하는 '제2회 인공지능 최고위전략대화'를 통해서는 데이터 댐의 고도화를 위해 구축된 데이터셋의 활용성을 강조하고 데이터의 양 뿐만 아니라 질 관리를 위한 선제적 품질 관리 과정을 도입하였다[15].

앞서 살펴본 국가 중심의 다양한 데이터 육성 정책을 통해 AI 산업계를 중심으로 데이터셋의 중요성을 명확하게 인식하고 산업 현장에서 활용할 수 있는 다양한 학습용 데이터셋을 제공하기 위해 많은 노력을 기울이고 있음을 확인할 수 있다.

2.2. AI 교육 관련 연구 동향

최근 발표된 2022 개정 교육과정 속 AI 교육은 초등학교 실과의 관련 영역과 중·고등학교의 정보과를 중심으로 구성되었으며, 교육의 목표를 '인공지능으로 인한

세상의 변화를 이해하고 인공지능을 통해 해결 가능한 문제를 탐색하고 해결하기 위한 능력과 태도를 갖추는 것'으로 제시되었다[19].

특히 정보과의 교과 역량으로 (Fig. 1)과 같이 '컴퓨팅 사고력', '디지털 문화 소양', '인공지능 소양'을 기를 수 있도록 구성하였으며, '인공지능 소양'은 '인공지능 문제해결력', '데이터 문해력', '인공지능 윤리의식'의 하위 역량을 포함하도록 제시하였다[19].

(Fig. 1) 2022 Revised Informatics Curriculum Design Overview

AI 교육의 중요성 및 사회적 요구가 높아짐에 따라 양질의 AI 교육을 학생들에게 제공하기 위한 많은 교육 연구들이 진행되었으며, AI 교육 연구 동향 분석 결과를 통해서 최근 교육 연구의 주된 관점을 확인할 수 있다.

한지윤 외(2021)는 2017년에서 2020년까지의 AI 교육 관련 연구 동향 분석을 통해 2019년부터 그 빈도가 크게 증가했으며, '머신러닝', '알고리즘', '초등' 등의 중심 키워드를 도출하였다. 도출된 키워드를 중심으로 네트워크 분석을 실시한 결과, 주로 AI 개념이나 원리를 적용한 교수·학습과 관련된 연구가 많이 이루어졌음을 확인하고 시사점을 도출하였다[6].

이다겸 외(2021)는 AI의 리터러시 측면을 강조하고 2019년부터 2021년까지 AI 역량을 길러 줄 수 있는 교육 연구의 동향을 분석하였다. 이를 통해 현재는 이론 연구가 주로 이루어지는 단계이며, 세부적으로 문헌분석과 프로그램 개발, 프로그램 적용 등의 연구가 주로 이루어짐을 확인하였다[16].

앞선 동향 분석 연구들을 통해 AI 교육 연구는 주로 교육 프로그램을 개발, 적용하고 효과성을 확인하는 관점에서 활발하게 이루어졌으며 상대적으로 AI 교육을 위한 중요 소재인 데이터셋에 초점을 맞춘 교육 연구는 부족하다는 한계를 확인할 수 있다.

2.3. AI 교육용 데이터셋의 맥락 관련 선행 연구

교육부(2021)는 2022 개정 교육과정 총론 주요 사항에서 개정 추진 배경의 한 축으로 '새로운 교육환경 변화에 적합한 역량 함양 교육 필요'를 제시하였으며, 단편적인 지식의 습득으로 끝나는 것이 아니라 학습한 내용을 삶의 맥락에서 적용하고 복잡한 문제를 해결하는 역량이 중요함을 언급하였다. 또한, 실생활 맥락과 연계한 교수·학습 및 평가를 통해 학생의 자발적·능동적 참여를 강화할 수 있는 교육 환경 구성을 강조하고 있다.

이를 바탕으로 2022 개정 실과 및 정보과 교육과정의 세부 내용 중 '교수·학습 및 평가'에서는 <Table 1>과 같이 학생들의 삶의 맥락을 강조하여 문제 해결 경험을 제공해야 함을 안내하였다[19].

<Table 1> Teaching and Learning, Evaluation Guide in the Curriculum

Subject	The Direction of Teaching and Learning
Practical Arts	In the context of life, a learning task that can be solved through Computational Thinking is presented and a learning method centered on play experience is applied to naturally cultivate artificial intelligence in the process of solving the task.
Informatics Data Science Basic Artificial Intelligence	It presents a learning task that allows learners to solve problems through computing in the context of real life to naturally cultivate Computational Thinking, digital culture, and artificial intelligence in the process of solving tasks on their own.

AI 교육이 이루어지는 교수·학습 환경에서 학생들에게 제시되는 학습 과제는 데이터셋의 주제와 밀접한 관련을 가지게 되며, 이러한 과제의 맥락 혹은 데이터셋의 맥락은 학습 과정의 전반에 직접적으로 영향을 미치게 된다[10]. 이에, 학습 혹은 교육의 맥락과 관련된 선행 연구를 살펴보면 다음과 같다.

Brown 외(1989)는 상황학습 이론에 따라 지식은 맥락과 독립적일 수 없으며, 상황 속에서 존재하기 때문에 새로운 지식을 제공할 때에는 맥락을 함께 제공해야 함을 강조하고 맥락 속 지식 형성의 중요성을 주장하였다[3].

Jonassen(2000)은 현대 학습 이론은 학습이 의미있는 활동의 맥락에서만 일어난다고 주장하며, 교육 설계 과정의 일부로 활동과 맥락을 분석하는 것의 중요성을 강조하였다. 특히 학생들이 실제로 접하게 되는 구조화된 문제와 현실 세계 문제의 차이점을 통해 문제를 해결하는 학습 과정에 있어 사회적, 문화적, 지적 가치가 중요함을 주장하였다[7].

범아영 외(2012)는 교수·학습 환경에서의 맥락 문제와 비맥락 문제의 차이점을 비교하였다. 학습자가 경험하게 될 문제 상황은 맥락을 포함해야 하며 맥락을 포함한 문제 상황을 해결한 학습자의 성취도가 높았고 해결 방식이 더 다양하게 나타남을 확인하였다[2].

OECD(2022)는 문제가 제시되는 상황 또는 문제가 배경으로 하는 세계를 '맥락'으로 표현하였다. 또한 문제를 해결하기 위한 전략과 표현의 선택은 맥락을 중심으로 이루어져야 함을 언급하고 미래 사회 시민으로서의 역량 평가를 위한 PISA(Program for International Student Assessment)에서 다양한 맥락을 사용하여 학생들의 소양을 측정할 수 있도록 하였다[20]. PISA는 시행 년도 마다 '읽기', '수학', '과학' 영역 중 하나의 영역을 주 영역으로 설정하며, PISA 2022는 수학 영역을 주 영역으로 하여 '수학적 추론', '컴퓨팅 사고력' 등을 평가하는 문항으로 구성되어 있다. 수학 영역의 맥락은 '개인적', '직업적', '사회적', '과학적' 4가지 범주로 구분되며 상세한 사항은 <Table 2>와 같다[20].

<Table 2> Context Categories in PISA 2022 Math

Context	
Personal	Personal context category focus on activities of one's self, one's family, or one's peer group
Occupational	Occupational context category are centered on the world of work
Societal	Societal context category focus on one's community (whether local, national, or global)
Scientific	Scientific category relate to the application of mathematics to the natural world and issues and topics related to science and technology

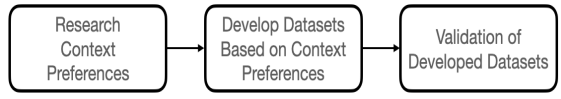
김슬기 외(2022)는 인공지능 기초 7종의 교과서 속 컴퓨팅 실습 활동에 활용되는 데이터셋을 분석하였다. 그 결과, 교과서에서 AI 모델링 등의 컴퓨팅 실습 활동에 주로 활용되는 정형 데이터는 붓꽃, 타이타닉 생존자, 손글씨 등 외국에서 생성된 데이터셋을 활용하고 있어 학생들의 삶의 맥락과 다소 거리가 있음을 확인하였다[8].

신승기(2022)는 AI 교육용 데이터셋 발굴을 위해 공공데이터 포털의 다양한 분류체계를 확인하고 학생들의 삶에 가까운 데이터셋 도출을 위해 '교육'을 키워드로 하여 LDA 기반 토픽 모델링을 수행하였다. 연구 결과로 공공데이터 포털을 통해 AI 교육을 위한 실제적인 데이터를 수집할 때 고려할 수 있는 데이터의 유형을 상세히 안내하였으며, 공공데이터 포털은 주로 프로그램 및 교육 지원의 환경적 측면에서 현황을 제공하며 교육의 실제적인 내용은 소수의 사례만 제공한다는 한계점을 확인하였다[23].

앞선 선행 연구를 통해 학생들의 AI 소양을 길러 주기 위해서는 삶의 맥락과 가까운 데이터셋이 필요함을 확인하였으며 현재 AI 교육에 주로 활용되고 있는 데이터셋 맥락의 한계와 공공데이터 포털을 통한 교육용 데이터셋 수집의 한계를 확인하였다. 이에 본 연구는 AI 교육을 위한 데이터셋 맥락의 학생 선호도를 조사하고 다양한 플랫폼에서 교수·학습 환경에 적합하며 학생들의 삶의 맥락과 가까운 데이터셋 주제 도출을 통해 AI 교육용 데이터셋을 제시하고자 하였다.

3. 연구 방법

본 연구는 학생들이 선호하는 맥락 중심의 AI 교육용 데이터셋을 개발하기 위해 (Fig. 2)와 같은 과정으로 연구를 진행하였다. PISA 2022 수학 맥락의 범주를 중심으로 학생들의 선호도를 조사하고, 공공데이터 및 다양한 데이터셋을 탐색하여 발굴 및 개발한다. 또한 개발된 데이터셋을 전문가 검토를 통해 타당도를 검증하고자 하였다.



(Fig. 2) The Process for Context-oriented Dataset Development

3.1. 연구 대상

학생들에게 유의미한 데이터셋의 맥락과 선호도를 도출하기 위해 AI 선도학교 혹은 AI 교육과 정보교육에 대한 경험이 있는 초등학생 76명, 중학생 89명, 고등학

생 27명, 총 192명의 대상을 선정하고 교사의 지도하에 온라인 설문에 답변할 수 있도록 하였다. 연구 대상에 대한 상세한 정보는 <Table 3>과 같다.

<Table 3> Research Participants Details

School	Male	Female	N	Sum
Elementary	32	44	76	192
Middle	47	42	89	
High	14	13	27	

3.2. 연구 도구

맥락에 대한 선호도를 조사하기 위한 설문은 PISA 2022의 수학 맥락 4가지 범주에 대한 선호도를 중심으로 '전혀 관심없음' 1점, '관심도가 매우 높음'을 5점으로 하는 5점 척도 문항과 AI로 해결하고 싶은 문제의 주제에 대한 개방형 문항, AI 수업에 활용되는 데이터셋의 형태로 선호되는 것을 2가지 선택하는 문항 및 AI 프로그래밍을 위해 활용할 수 있는 프로그래밍 도구를 모두 선택하는 문항으로 구성하였다. 구성된 문항을 컴퓨터 교육 전공 교수 1인과 동일한 전공 박사 3인이 검토하였으며, 맥락에 대한 설명을 번역하는 과정에서 사용된 단어들의 이해도를 높이고자 학생들이 이해하기 쉬운 문장으로 수정하여 최종안을 완성하였다.

4. 연구결과

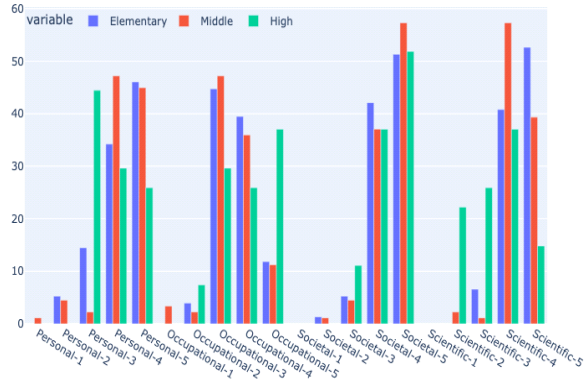
4.1. AI 교육용 데이터셋 맥락 선호도 설문 결과

AI 교육을 위한 데이터셋 맥락 선호도의 응답 결과는 <Table 4>와 같다. 모든 학교 급에서 사회적 맥락에 대한 선호도 평균이 가장 높게 나타났으며 전체 집단의 결과 또한 사회적 맥락의 평균이 가장 높고 표준편차가 가장 작게 나타났다.

<Table 4> Dataset Preference Response Results

Context	Personal		Occupational		Societal		Scientific	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Elementary	4.21	0.88	3.59	0.75	4.43	0.66	4.46	0.62
Middle	4.30	0.81	3.49	0.85	4.50	0.64	4.33	0.62
High	3.81	0.83	3.92	0.99	4.40	0.69	3.44	1.01
Whole Group	4.19	0.85	3.59	0.84	4.46	0.65	4.26	0.76

세부적으로 결과를 살펴보면 초등학생과 중학생의 경우 직업적 맥락을 제외한 나머지 3가지 맥락에 대해 비교적 고른 선호도를 보였다. 고등학생의 경우 사회적 맥락의 선호도가 가장 높았고 두번째로 선호도가 높았던 직업적 맥락의 응답이 전체 평균 대비 높게 나타났다. 상세한 분석을 위해 각 학교 급에 따른 맥락 선호도 응답의 비율을 (Fig. 3)과 같이 시각화하였다.



(Fig. 3) Visualization of Preferred Context Response Rates

긍정을 의미하는 4와 5의 응답 비율을 중심으로 분석하면 사회적 맥락의 선호도는 모든 학교 급에서 가장 높게 나타났으며, 직업적 맥락의 경우 고등학생의 긍정 응답 비율이 초등학생과 중학생에 비해 매우 높게 나타남을 확인할 수 있었다. 또한, 초등학생과 중학생의 과학적 맥락에 대한 선호도가 고등학생에 비해 높게 나타났으며 개인적 맥락의 응답 결과는 초등학생과 중학생의 긍정 응답 비율이 고등학생보다 높게 나타났다. 특히 중학생의 개인적 맥락에 대한 선호도가 더 높은 것을 확인할 수 있었다.

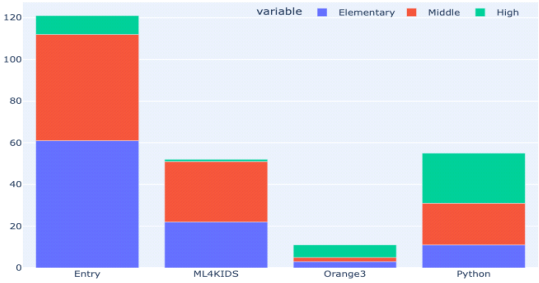
위의 결과를 바탕으로 AI 교육용 데이터셋의 주제는 모든 학교 급에서 높은 선호도를 보인 '사회적 맥락'을 중심으로 탐색하며, 학생들의 다양한 선호도 충족과 문제 해결 경험을 제공하기 위해, 초등학생과 중학생을 위한 '과학적 맥락', '개인적 맥락'의 데이터셋을 포함하며 고등학생을 위한 '직업적 맥락'의 데이터셋을 포함하도록 설정하였다.

두 번째 설문 결과인 AI 교육 활동 시 선호하는 데이터셋의 형태에 대한 응답을 분석하기 위해 각 학교 급별 응답의 빈도를 (Fig. 4)와 같이 시각화하였다.

(Fig. 4) Visualization of Preferred Data type Response

데이터셋의 형태는 수와 문자로 구성된 정형 데이터와 비정형 데이터 중 하나인 이미지 데이터를 선호하는 것으로 나타났다. 특히 이미지 데이터의 경우 초등학생과 중학생에 대한 선호도가 다소 높게 나타났으며, 음성 데이터의 경우 중학생, 영상 데이터의 경우 초등학생의 선호도가 높게 나타났다.

세번째 설문 결과인 선호하는 교육용 프로그래밍 도구에 대한 응답을 분석하기 위해 빈도를 시각화한 결과는 (Fig. 5)와 같다.



(Fig. 5) Visualization Preferred Programming Tool Response

프로그래밍 도구에 대한 선호도는 엔트리가 가장 높게 나타났으며 파이썬과 ML4KIDS가 비슷하게 나타났다. 학교 급별로 특징을 살펴보면 엔트리의 경우 초등과 중학생의 선호도가 높게 나타났으며, 고등학생의 경우 파이썬에 대한 선호도가 높은 것으로 나타났다. Orange3의 경우 전체적인 선호도가 가장 낮았으며 고등학교의 Orange3 선호도는 비교적 높게 나타났다.

위의 응답 결과를 바탕으로 AI 교육을 위한 데이터셋의 형태는 모든 학교급에서 가장 선호도가 높았던 형태인 수와 문자로 이루어져 있으며 엔트리와 파이썬에서 바로 활용 가능한 정형 데이터로 설정하였다.

4.2. AI 교육용 데이터셋 탐색 및 개발

국내외 데이터셋 제공 플랫폼에서 학생들이 선호하는 맥락의 주제이며, 2차 저작물 생산 및 교육 목적으로 활용 가능한 저작권을 명시하는 데이터셋 및 관련 주제를 탐색하고 선별하였다. 선별된 데이터셋의 주제와 출처 및 크기를 정리하면 <Table 5>와 같다.

<Table 5> Source and Basic Information about the Dataset

Datasets	Source	Rows*Columns
Age and Culture	www.culture.go.kr	44 * 314
Amongus	www.kaggle.com	2,090 * 13
Bakery Sales	www.kaggle.com	2,419 * 27
Baseball	koreabaseball.com	1369 * 28
Body Performance	www.culture.go.kr	13,393 * 14
Job and Crime	www.data.go.kr	164 * 47
Performance Review	www.culture.go.kr	458 * 14
Date and Delivery	bdp.kt.co.kr	10,000 * 4
Disease Prediction	www.data.go.kr	46,545 * 8
Earthquake	devweather.kma.go.kr	2,429 * 4
Premier League Soccer	www.kaggle.com	11,113 * 23
Future Incomes	www.kaggle.com	92,858 * 14
Medical Examination	www.kaggle.com	1,000 * 64
Super Heroes Info	www.kaggle.com	734 * 10
Super Heroes ability	www.kaggle.com	667 * 1681
Cancer	www.kaggle.com	7,288 * 26
Life quality	www.culture.go.kr	3,066 * 9
League of Legend	www.kaggle.com	26,904 * 40
Lotto	www.kaggle.com	1,003 * 9
MBTI	www.kaggle.com	10,0000 * 75
Energy	www.kaggle.com	299 * 14
Mobile Phone	www.kaggle.com	2,000 * 21
Mosquito	news.seoul.go.kr	1,090 * 13
Noodle	www.kaggle.com	109 * 4
Drama	www.kaggle.com	100 * 14
Notional Temperature	data.kma.go.kr	18,117 * 4
Pokemon	www.kaggle.com	801 * 44
Sensory Temperature	data.kma.go.kr	517 * 4
Travel Plan	www.culture.go.kr	1,500 * 9
T-Shirts	Collected by Researcher	30 * 3
Twitch	www.kaggle.com	1,000 * 11
Video Games	www.kaggle.com	6,822 * 16
Weather and Delivery	bdp.kt.co.kr	34,640 * 27
Weather and Park	www.culture.go.kr	85,157 * 21

또한 2022 개정 실과 및 정보 교육과정에서 제시된 교과역량인 '인공지능 소양'과 하위 역량인 '인공지능 문제 해결력', '데이터 문해력'에 초점을 맞춰 '활용 예시'를 함께 제시하고자 하였다. 데이터셋 내에 대표적인 종속변수가 있는 경우, 지도학습(Supervised Learning)으로 설정하며, 종속변수가 수치형 및 연속형인 경우 예측(Prediction, Ⓟ), 이산형 및 범주형일 경우 이진 분류(Binary Classification, Ⓑ) 혹은 다중 분류(Multiple Classification, Ⓜ)로 설정하였다. 그리고 종속 변수가 데이터셋 내에서 명시적으로 드러나지 않는 경우, 데이터셋의 특성에 따라 비지도학습(Unsupervised Learning, Ⓤ) 혹은 데이터 분석(Data Analysis, Ⓓ)으로 설정하였다. 마지막으로 데이터셋의 주제, 맥락 및 형태 등을 종합적으로 고려하여 초(E), 중(M), 고등학생(H)의 교육 대상을 설정하였다. 각각의 데이터셋에 설정된 맥락과 대상, 활용 예시에 대한 분석 결과는 <Table 6>과 같다.

<Table 6> Results of Analyzing Datasets for AI Education

Datasets	Context	Target	Usage Example
Age and Culture	Societal Occupational	E, M, H	Ⓓ
Amongus	Personal	E, M, H	Ⓑ
Bakery Sales	Occupational	E, M, H	Ⓟ
Baseball	Scientific Occupational	E, M, H	Ⓑ
Body Performance	Scientific	E, M, H	Ⓟ
Job and Crime	Societal	E, M, H	Ⓓ
Performance Review	Societal	M	Ⓓ
Date and Delivery	Occupational	M	Ⓓ
Disease Prediction	Societal	E, M, H	Ⓤ
Earthquake	Societal Scientific	E, M	Ⓓ
Premier League Soccer	Personal Occupational	E, M, H	Ⓑ
Future Incomes	Societal	H	Ⓟ
Medical Examination	Societal Scientific	M, H	Ⓤ
Super Heroes Info	Personal	E, M, H	Ⓓ
Super Heroes ability	Personal	M, H	Ⓓ
Cancer	Scientific	M	Ⓓ
Life quality	Societal	M, H	Ⓜ
League of Legend	Personal	M, H	Ⓑ
Lotto	Personal	E, M	Ⓓ
MBTI	Societal	M, H	Ⓜ
Energy	Societal	E, M, H	Ⓓ
Mobile Phone	Occupational	E, M, H	Ⓜ
Mosquito	Societal Scientific	E, M	Ⓟ
Noodle	Societal	E	Ⓓ
Drama	Societal	M, H	Ⓤ
Notional Temperature	Societal	E, M, H	Ⓓ
Pokemon	Personal	E, M, H	Ⓓ
Sensory Temperature	Societal Scientific	E, M	Ⓟ
Travel Plan	Societal	M, H	Ⓜ
T-Shirts	Societal	E, M	Ⓜ
Twitch	Personal Occupational	M, H	Ⓓ
Video Games	Personal Occupational	E, M, H	Ⓓ
Weather and Delivery	Societal Occupational	M, H	Ⓜ
Weather and Park	Societal Scientific	M, H	Ⓟ

1차적으로 도출된 AI 교육용 데이터셋의 활용도를 높이기 위해 데이터셋의 일부 구성 요소를 다음과 같이 재구성하였다. 먼저 모든 데이터셋 내의 영문으로 이루어진 행과 열 및 개별 데이터를 초, 중, 고등학교 학생들이 이해 가능한 표현으로 수정 및 번역하였다. 그리고 데이터셋 출처에서 여러 개의 파일로 나누어서 제공하는 경우 하나의 파일로 병합하였으며(Merge), 앞서 확인한 활용 예시에 적합하게 열의 순서를 변경하거나 사용도가 상대적으로 낮을 것으로 예상되는 열을 삭제하였다(Column Reconfiguration). 그리고 결측치나 적합하지 않은 정보를 포함하는 행을 수정하거나 삭제하였다(Row Reconfiguration). 마지막으로 데이터의 수가 상대적으로 작거나 정보 식별 및 개인정보 문제가 있을 것으로 예상되는 데이터셋의 경우 원데이터셋의 통계적 특성을 반영한 합성데이터셋 생성 방법을 이용하여 재생산하였다(Synthesis)[8]. 각 데이터셋의 재구성된 항목을 정리하면 <Table 7>과 같다.

<Table 7> Datasets Reconfiguration Components

Datasets	Reconfiguration Components			
	Synthesis	Merge	Row Reconfiguration	Column Reconfiguration
Age and Culture		○		○
Amongus		○		○
Bakery Sales				○
Baseball	○			
Body Performance		○		○
Performance Review		○		
Disease Prediction		○		○
Earthquake		○		○
Premier League Soccer			○	○
Medical Examination		○		
Life quality		○		○
League of Legend				○
Lotto				○
MBTI				○
Energy		○		
Pokemon				○
Sensory Temperature				○
Travel Plan		○		○
T-Shirts	○			
Twitch			○	○
Video Games			○	○
Weather and Delivery		○		○
Weather and Park		○		

4.3. AI 교육용 데이터셋 타당도 평가

개발된 데이터셋의 타당도를 검증하기 위해 컴퓨터 공학 전공 박사 학위 소지자 2인, 컴퓨터 교육학 박사 학위 소지자 1인의 연구자와 컴퓨터 교육 관련 석사 학위를 가지며 교육 경력 10년 이상의 초등학교 교사 2인, 중학교 교사 3인, 고등학교 교사 6인의 현장 전문가가 포함된 14인의 전문가 집단을 구성하였다.

AI 교육용 데이터셋의 타당도를 검증하기 위한 전문가 설문은 AI 교육을 위한 데이터셋이 갖추어야 할 요건을 '학생 발달 단계 측면', '공학적 측면', '일반적인 교수학습 자료 측면'의 요건으로 나누어 각각의 기준을 제시한 김슬기 외(2021)의 '초·중등 인공지능 교육을 위한 데이터셋 기준'을 활용하였다.

개발된 AI 교육용 데이터셋의 맥락 및 AI 교육 목적에서 데이터셋의 완성도를 평가하기 위해 전문가 집단에게 데이터셋의 전체적인 구성과 사용 예시를 함께 안내하여 개별 데이터셋에 대한 이해도를 높였으며, 각 요건에 따른 적합도를 5점 리커트 척도로 '전혀 적합하지 않음' 1점부터 '매우 적합함' 5점으로 응답하도록 구성하였다. 3점 이하의 긍정이 아닌 응답에 대해서는 이유를 함께 기재 요청하고 개방형 질문을 통해 추가적인 의견을 요청하였다[10].

또한, 타당도를 정량적으로 평가하기 위해 전문가 응답 결과를 활용하여 CVR(Content Validity Ratio) 값을 도출하였으며, 14명의 전문가 수에 대한 최소 CVR 값의 기준인 0.571 미만 응답의 데이터셋과 항목을 중심으로 추가 분석 및 수정하였다[1]. 1차 설문결과를 통해 각 요건의 CVR 값을 도출한 결과는 <Table 8>과 같다.

<Table 8> Results of CVR Values for Datasets Validity Assessment-1

Datasets	Teaching and Learning Material	Student Development Stage	Engineering
Age and Culture	0.571	0.429	1.000
Amongus	0.286	0.714	0.857
Bakery Sales	0.571	1.000	0.143
Baseball	0.571	0.571	1.000
Body Performance	0.571	1.000	1.000
Job and Crime	-0.571	-0.286	0.143
Performance Review	1.000	1.000	0.857
Date and Delivery	1.000	1.000	0.571
Disease Prediction	1.000	1.000	1.000
Earthquake	1.000	1.000	0.571
Premier League Soccer	1.000	0.714	1.000
Future Incomes	0.143	0.714	0.571
Medical Examination	0.571	0.571	1.000
Super Heroes Info	0.714	0.714	1.000
Super Heroes Ability	-0.286	0.429	-0.143
Cancer	-0.286	1.000	0.714
Life Quality	-0.143	0.714	0.571
League of Legend	0.714	0.571	0.714
Lotto	-0.571	0.571	0.571
MBTI	1.000	0.571	1.000
Energy	0.571	0.571	0.571

Mobile Phone	1.000	1.000	1.000
Mosquito	1.000	1.000	1.000
Noodle	1.000	1.000	1.000
Drama	-0.286	0.571	0.143
Notional Temperature	1.000	1.000	1.000
Pokemon	0.714	1.000	0.571
Sensory Temperature	1.000	1.000	1.000
Travel Plan	0.714	0.714	1.000
T-Shirts	1.000	1.000	1.000
Twitch	0.714	0.714	1.000
Video Games	0.143	0.714	0.286
Weather and Delivery	1.000	1.000	1.000
Weather and Park	1.000	1.000	0.571

‘교수·학습 자료 측면’에서 타당도를 먼저 분석하면 1차적으로 도출된 데이터셋 중 ‘어몽어스’, ‘직업별 범죄 현황’, ‘미래 수입 예측’, ‘슈퍼 히어로 능력’, ‘암 발병 현황’, ‘삶의 질 인식’, ‘로또 번호’, ‘한국 드라마 평가’, ‘비디오 게임’의 데이터셋이 기준을 만족하지 못한 것으로 나타났다. 상세 의견으로는 제시된 대상의 학생에게 적합하지 않거나 활용 예시에 적합하지 않다는 의견과 함께 데이터셋의 주제 및 내용이 학생에게 정서적으로 적절하지 않다는 의견 등이 제시되었다. 이에 ‘어몽어스’ 데이터셋의 대상을 중학생과 고등학생으로 한정하였으며 나머지 데이터셋을 모두 제외하였다.

‘학생 발달 단계 측면’에서의 타당도 검토 결과는 ‘연령별 인기 문화생활’ 데이터셋이 만족하지 못하는 것으로 나타났다. 관련 의견으로 불필요한 특성이 많아 학생들에게 적절하지 않다는 의견이 있어 추가적으로 일부 특성을 삭제하여 재구성하였다.

‘공학적 측면’에서의 검토 결과를 살펴보면, ‘빵 판매 현황’ 데이터셋이 기준을 만족하지 못한 것으로 나타났다. 해당 데이터셋의 경우 기계학습을 통해 예측하기에 데이터셋의 구성이 적절하지 않다는 의견이 제시되어 해당 데이터셋을 제외하였다.

기타 의견으로는 데이터셋의 특성을 학생들이 이해하기 쉬운 언어로 수정이 필요하다는 의견과 기온 정보를 포함하는 데이터셋의 경우 예외적으로 초등에서도 음수 데이터를 활용할 수 있도록 구성해야 한다는 의견이 제시되었다. 이를 반영하여 ‘어몽어스’, ‘야구 경기 결과’, ‘프리미어리그 축구 결과’, ‘리그 오브 레전드’, ‘트위치 스트리머’ ‘체감 온도’, ‘전국 기온 변화’, ‘모기 발생’ 데이터셋의 열과 행 및 개별 데이터를 재구성하였다.

데이터셋의 타당도를 평가한 3가지 요건 중 1가지 이상의 항목에서 정량적 평가 기준인 CVR 값 0.571 미만의 값을 가진 데이터셋과 전문가의 의견을 반영하여 수정한 데이터셋을 중심으로 2차 타당도 검토를 진행하였다. 최종 목록에서 제외하였거나 추가적으로 수정한 데이터셋의 2차 전문가 검토 결과는 <Table 9>와 같으며, 모든 데이터셋이 타당도를 만족하는 것으로 나타났다.

<Table 9> Results of CVR Values for Datasets Validity Assessment-2

Datasets	Teaching and Learning Material	Student Development Stage	Engineering
Age and Culture	1.000	1.000	1.000
Amongus	0.714	1.000	1.000
Baseball	1.000	1.000	1.000
Job and Crime		Excluded	
Premier League Soccer	0.714	1.000	1.000
Future Incomes		Excluded	
Super Heroes Ability		Excluded	
Cancer		Excluded	
Life Quality		Excluded	
League of Legend	0.571	1.000	1.000
Lotto		Excluded	
Drama		Excluded	
Notional Temperature	1.000	1.000	1.000
Sensory Temperature	1.000	1.000	1.000
Twitch	0.571	0.857	0.857
Video Games		Excluded	

4.4. 맥락형 데이터셋 구축

본 연구의 최종 결과인 AI 교육용 데이터셋을 맥락과 대상, 활용 예시로 정리한 결과는 <Table 10>과 같다.

중복 적용된 맥락을 포함하여 ‘사회적 맥락’ 15종, ‘과학적 맥락’ 7종, ‘개인적 맥락’ 6종, ‘직업적 맥락’ 7종의 데이터셋을 개발, 수집 및 재구성하였으며 데이터셋의 주제와 형태를 고려하여 활용 가능한 초, 중, 고등학생의 대상을 제시하였다. 또한 2022 개정교육과정의 AI

교육에 적합한 활용 예시를 기준으로 지도학습 13종(예측 4종, 이진 분류 4종, 다중 분류 5종), 비지도학습 3종, 데이터 분석 9종으로 세분화하여 구성하였다.

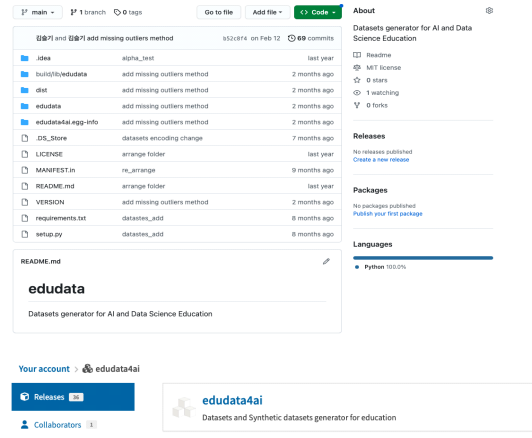
<Table 10> Results of AI Educational Dataset Development

Datasets	Context	Target	Usage Example
Age and Culture	Societal Occupational	E, M, H	Ⓓ
Amongus	Personal	M, H	Ⓑ
Baseball	Scientific Occupational	E, M, H	Ⓑ
Body Performance	Scientific	E, M, H	Ⓟ
Performance Review	Societal	M	Ⓓ
Date and Delivery	Occupational	M	Ⓓ
Disease Prediction	Societal	E, M, H	Ⓤ
Earthquake	Societal Scientific	E, M	Ⓤ
Premier League Soccer	Personal Occupational	E, M, H	Ⓑ
Medical Examination	Societal Scientific	M, H	Ⓤ
Super Heroes Info	Personal	E, M, H	Ⓓ
League of Legend	Personal	M, H	Ⓑ
MBTI	Societal	M, H	Ⓜ
Energy	Societal	E, M, H	Ⓓ
Mobile Phone	Occupational	E, M, H	Ⓜ
Mosquito	Societal Scientific	E, M	Ⓟ
Noodle	Societal	E	Ⓓ
Notional Temperature	Societal	E, M, H	Ⓓ
Pokemon	Personal	E, M, H	Ⓓ
Sensory Temperature	Societal Scientific	E, M	Ⓟ
Travel Plan	Societal	M, H	Ⓜ
T-Shirts	Societal	E, M	Ⓜ
Twitch	Personal Occupational	M, H	Ⓓ
Weather and Delivery	Societal Occupational	M, H	Ⓜ
Weather and Park	Societal Scientific	M, H	Ⓟ

본 연구를 통해 구성된 맥락형 데이터셋의 초등학교 및 중학교 AI 교육에 대한 활용성을 높이고자 국내 AI 교육을 위한 플랫폼으로 선호도가 높으며 AI 교육에 활용하기에 적합한 엔트리 플랫폼을 주요 대상으로 선정하였다[17]. 그리고 엔트리 플랫폼의 '데이터 과학' 메뉴에 (Fig. 6)과 같이 일부 데이터셋을 활용 예시별로 선별하여 탑재하였다[4].

(Fig. 6) Datasets on Entry Platform(Data Science-Datasets)

고등학교 AI 교육에 대한 활용성을 높이고자 고등학생이 선호하는 프로그래밍 도구인 파이썬을 주요 대상으로 설정하고 파이썬 패키지 관리 시스템인 PyPi와 소스코드 공유 플랫폼인 Github에 업로드 하여 (Fig. 7)과 같이 데이터셋 및 소스코드를 다운로드 받을 수 있도록 하였으며 pip 명령어를 활용하여 실습환경에서 설치할 수 있도록 배포하였다[5, 22].



(Fig. 7) Datasets on PyPi(edudata4ai) and Github(edudata)

5. 결론 및 제언

AI 기술의 발달로 인한 급격한 사회 변화에 적응하며, 미래를 살아갈 학생들의 기초 역량을 기르기 위한 교육으로 AI 교육에 대한 수요가 증가하고 있다. 이에 다양한 AI 교육 연구들이 유의미한 결과들을 도출하고 있지만 AI 교육에 필수 요소이며 교수·학습 환경에서 교사와 학생이 직접적으로 활용하는 데이터셋 관련 연구는 상대적으로 부족한 실정이다.

이러한 연구의 필요성에서 출발하여 본 연구는 학생들에게 유의미한 AI 교육을 제공하기 위해 데이터셋에 초점을 맞추어 2022 개정 교육과정과 선행연구를 분석하였다. 이를 통해 AI 교육에 활용되는 데이터셋 맥락의 중요성을 도출하였으며 현재 활용되는 데이터셋과 공공데이터 플랫폼에서 수집할 수 있는 데이터셋의 한계를 확인하였다.

학생들에게 유의미한 AI 교육용 데이터셋을 탐색하기 위해 PISA 2022 수학의 맥락에 대한 학생들의 선호도를 조사하고 그 결과를 다각도로 분석하여 시사점을 도출하였다. 또한 국내 외 데이터셋 제공 플랫폼에서 AI 교육에 적합한 데이터셋 주제를 탐색하고 이를 활용하여 다양한 데이터셋을 개발 및 재구성하였다.

1차적으로 도출된 AI 교육용 데이터셋의 적합성을 확인하기 위해 전문가 검토를 진행하고 전문가 의견을 반영하여 데이터셋을 수정 및 보완하였다. 개발된 25종의 데이터셋은 '사회적 맥락', '과학적 맥락', '개인적 맥락', '직업적 맥락'을 가지며 2022 개정 교육과정에서 제시된 AI 교육 환경에서 활용하기에 적합한 요소로 구성되었다. 또한 학생들의 선호도가 높은 엔트리의 데이터 과학 메뉴와 파이썬의 패키지 관리 시스템에 탑재하고 배포하여 교육적 환경에서 접근성을 높이고자 하였다.

본 연구를 통해 도출된 데이터셋은 교사와 학생 모두 교수·학습 환경에서 손쉽게 활용 가능하며 데이터셋의 주제가 학생들의 삶의 맥락과 가깝고 교육에 적합한 형태로 구성되어 있어 AI 역량을 효과적으로 증진시키는데 도움을 줄 수 있을 것으로 예상된다.

AI 교육용 데이터셋은 학생들에게 양질의 AI 교육을 제공하기 위한 필수요소 중 하나이다. 추후 본 연구의 결과를 발전시켜 교사와 학생의 접근성을 더욱 높인 AI 교육용 데이터셋 제공 시스템을 개발하고 개별 데이터셋에 대한 교수·학습 적용 효과성을 다방면으로 분석하는 연구가 이루어진다면 AI 교육이 현장에 안착하는데 더 많은 도움을 줄 수 있을 것이다.

본 연구의 결과가 다양한 관점에서의 AI 교육용 데이터셋 관련 연구에 밑거름이 되어 양질의 AI 교육을 현장에 보급하는데 도움을 줄 수 있기를 기대한다.

참고문헌

[1] Ayre, C., & Scally, A. J. (2014), Critical values for lawshe's content validity ratio: Revisiting the original methods of calculation, Measurement and Evaluation in Counseling and Development, 47(1), 79 - 86.

[2] Beom, A., & Lee, D. (2012), Analysis on the problem-solving methods of students on contextual and noncontextual problems of fractional computation and comparing quantities, Education of Primary School Mathematics, 15(3), 219-233

[3] Brown, J. S., Collins, A., & Duguid, P. (1989), Situated cognition and the culture of learning, Educational Researcher, 18(1), 32-42.

[4] Entry. (2023, February 5), Make-DataScience, https://playentry.org

[5] Github. (2023, February 5), Edudata, https:/ /github.com/tmfrlska/edudata

[6] Han, J., & Shin, Y. (2020), Analysis of research trends in artificial intelligence education: keyword network analysis, Korean Association of Artificial Intelligence Education, 1(2), 20-33.

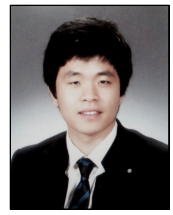
[7] Jonassen, D. (1997), Instructional design models for well-structured and ill-structured problem-solving learning outcomes, Educational Technology Research and Development, 45(1), 65-94.

[8] Kim, S., Jeon, Y., & Kim, T. (2022), Research on the development and utility analysis of K-12 artificial intelligence educational datasets using synthetic datasets generation method, The Journal of Korean Association of Computer Education, 25(3), 9-21.

[9] Kim, S., Kim, G., & Kim, T. (2022), Exploring public datasets topics for K-12 artificial intelligence edu-

cation using datasets. Proceedings of Journal of the Korea Computer Education Association, Korea, 26(1), 89-92.
[10] Kim, S., & Kim, T. (2022), A study on educational dataset standards for K-12 artificial intelligence education, The Journal of Korean Association of Computer Education, 25(1), 29-40.
[11] Korean Government (2019), AI National Strategy, https://www.koea.kr/news/pressReleaseView.do?newsId=156366736
[12] Korean Government (2020), Digital newdeal, https://digital.go.kr/front/promotion/policyList.do
[13] Korean Government. (2020), A plan to spread AI SW education for people across the country, https://www.4th-ir.go.kr/article/download/710
[14] Korean Government (2020), Education policy directions and key challenges in the age of AI, https://www.korea.kr/archive/expDoc15View.do?docId=39237
[15] Korean Government (2022). Policy on the advancement of data construction for artificial intelligence learning, https://www.msit.go.kr/bbs/view.do?sCode=user&bbsSeqNo=94&nttSeqNo=3181298
[16] Lee, D., Kim, S., Lee, Y. (2021), The analysis on research trends for artificial intelligence literacy education in korea, Proceedings of The Journal of Korea Computer Education Association, Korea, 25(2(A)), 25-27
[17] Lee, S., & Kim, T. (2020), Analysis and suggestions of web-based online platform for artificial intelligence education, The Journal of Korean Association of Computer Education, Korea, 24(2(A)), 77-80.
[18] National curriculum information center. (2022), 2022 revised curriculum main points
[19] National curriculum information center (2022), 2022 revised curriculum
[20] OECD (2022), PISA 2022 mathematics framework, https://pisa2022-maths.oecd.org/
[21] Park. H. S. (2020), Machine learning and deep learning that you study alone, Seoul, Hanbit Media.
[22] Pypi (2023, February 5), edudata4ai, https://pypi.org/project/edudata4ai
[23] Shin, S. (2022), A study on educational data mining for public data portal through topic modeling method with latent dirichlet allocation. Journal of The Korean Association of Information Education, 26(5), 439-448.

저자소개



김 슬 기

2008년 경인교육대학교 초등교육학과(교육학학사)
2016년 경인교육대학교 융합교육학과(교육학석사)
2023년 한국교원대학교 컴퓨터교육과(교육학박사)
관심분야: 정보(SW·AI) 교육, 컴퓨팅사고력, 데이터과학, 데이터리터러시
e-mail: tmfrlska85@gmail.com

김 귀 훈

1998년 KAIST 공학사
2000년 KAIST 공학석사
2019년 KAIST Ph.D.
2000년~2005년 LG데이콤 주임연구원
2005년~2020년 ETRI 책임연구원
2020년~현재 한국교원대학교 교수
2006년~현재 ITU-T SG11 Rapporteur, Editor
관심분야: 인공지능융합교육, 지능형에지컴퓨팅, 강화학습
e-mail: kimkh@knue.ac.kr

김 태 영

1985년 한양대학교
산업공학과(공학사)

1990년 Texas A&M University
Computer Science(M.S.)

1994년 Texas A&M University
Computer Science(Ph.D)

1994년~현재 한국교원대학교
컴퓨터교육과 교수

관심분야: 컴퓨터교육, 데이터베이스, 프로그래밍

e-mail: tykim@knue.ac.kr

부 록

<Table 11> AI 교육용 데이터셋 개발 결과

데이터셋	맥락	대상	활용예시
연령별 인기 문화 생활	사회적 직업적	초, 중, 고	데이터 분석
어몽 어스	개인적	중, 고	지도학습(이진분류)
야구 경기 결과	과학적 직업적	초, 중, 고	지도학습(이진분류)
국민 체력 100 측정	과학적	초, 중, 고	지도학습(예측)
공연 분야별 리뷰	사회적	중	데이터 분석
요일과 배달 음식	직업적	중	데이터 분석
요일과 배달 음식	사회적	초, 중, 고	비지도학습
지진 발생 정보	사회적 과학적	초, 중	비지도학습
프리미어리그 축구 결과	개인적 직업적	초, 중, 고	지도학습(이진분류)
건강 검진 결과	사회적 과학적	중, 고	비지도학습
슈퍼히어로 정보	개인적	초, 중, 고	데이터 분석
리그 오브 레전드	개인적	중, 고	지도학습(이진분류)
MBTI	사회적	중, 고	지도학습(다중분류)
에너지 사용	사회적	초, 중, 고	데이터 분석
휴대폰 가격	직업적	초, 중, 고	지도학습(다중분류)
모기 발생	사회적 과학적	초, 중	지도학습(예측)
라면 가격	사회적	초	데이터 분석
전국의 기온 변화	사회적	초, 중, 고	데이터 분석
포켓몬스터	개인적	초, 중, 고	데이터 분석
체감 온도	사회적 과학적	초, 중	지도학습(예측)
여행 계획	사회적	중, 고	지도학습(다중분류)
티셔츠 사이즈	사회적	초, 중	지도학습(다중분류)
트위치 스트리머	개인적 직업적	중, 고	데이터 분석
날씨별 배달 품목	사회적 직업적	중, 고	지도학습(다중분류)
날씨에 따른 공원 이용	사회적 과학적	중, 고	지도학습(예측)

<Table 12> AI 교육용 데이터셋 구성 예시(야구 경기 결과)

	상대팀	홈/어웨이	타점	득점	…	승패
1	K	Away	3	4		Win
2	D	Home	9	9		Win
3	I	Away	5	5	…	Win
4	I	Home	3	3		Lose
5	H	Away	7	8		Win
6	D	Away	11	12		Win
			…			
515	C	Home	6	7		Win
516	E	Away	6	8	…	Win
517	K	Home	4	4		Win