경영정보학연구
제26권 제1호, 2024년 2월 http://dx.doi.org/10.14329/isr.2024.26.1.019

인공지능 학습용 데이터 품질에 대한 연구: 퍼지셋 질적비교분석

A Study on the Artificial Intelligence (AI) Training Data Quality: Fuzzy-set Qualitative Comparative Analysis (fsQCA) Approach

오 현 목 (Hyunmok Oh) 한국지능정보사회진흥원 수석연구원
이 서 연 (Seoyoun Lee) 북경이공대학교 관리경제학원 박사과정
장 영 훈 (Younghoon Chang) 노팅엄대학교 닝보캠퍼스 상학원 교수, 교신저자

요 약

본 연구는 한국의 인공지능 학습용 데이터 구축 사업과 데이터의 공공 개방에 관한 정책 수행 기관, 데이터 구축 기업, 그리고 이를 활용하는 다양한 기관의 데이터 품질에 대해 이해를 제고하고, 신뢰할 수 있는 인공지능 알고리즘 개발에 있어 가장 중요한 학습용 데이터 품질에 대한 이론적 토대를 만들기 위한 실증적 연구이다. 이를 위해, 데이터의 속성 요인, 데이터 구축환경 요인, 데이터 타입 관련 요인 등 인공지능 학습용 데이터 품질과 관련된 중요 선행요인을 도입하여 이론적 모형을 제안한다. 본 연구는 393명의 인공지능 학습용 데이터 구축 기업과 인공지능 서비스 개발 기업의 실무 담당자를 대상으로 설문조사를 실시하여 데이터를 수집하였다. 데이터 분석은 퍼지셋 질적비교분석 방법과 인공신경망 분석을 통해 이루어졌으며, 분석 결과를 통해 인공지능 학습용 데이터 관련 학술적 및 실무적 시사점을 도출했다.

키워드 : 인공지능 학습용 데이터, 데이터 품질, 퍼지셋 질적비교분석, 인공신경망 분석, 혼합 연구 방법

Ⅰ. 서 론

인공지능(Artificial Intelligence)의 시대가 도래하고 있다. 모바일 인터넷, 빅데이터, 사물인터넷 등 초연결 기반 기술들의 혁명으로 다양한 분야와 산업에서 인공지능과 디지털 기술을 활용한 다양한 서비스들과 제품들이 쏟아져 나오고 있다. 이러한 4차 산업혁명으로 불리는 데이터 분석과 정보기술 혁명은 산업 전반에서 디지털 기술을 통한 다양하고 새로운 비즈니스 모델과 서비스 산업으로의 전환 또한 촉발하고 있다. 미국과 중국이 벌이고 있는 기술 전쟁의 이면에도 이러한 인공지능 기술과 이를 지원하는 다양한 반도체 및 데이터 관련 기술의 선점 그리고 미래 기술력 확보가 자리 잡고 있다고도 하겠다. 이러한 새로운 기술혁신의 중심에는 인공지능 알고리즘을 개발하고 상업화하는 필수인력의 중요성이 대두되고 있다. 어찌 보면 이보다도 더 중요한 인공지능 알고리즘을

학습시키고 훈련할 수 있는 다양하고도 품질 좋은 데이터의 요구(Needs)가 또한 부상하고 있다. 인공지능처럼 자율적으로 알고리즘에 의해서 주어진 업무를 처리하는 시스템들은 초기에 중요한 업무 프로세스와 관련된 학습을 위한 정확하고도 무결한 고품질의 데이터가 필요하다. 만약 이러한 좋은 품질의 데이터가 확보되지 않고, 불완전하고 어느 한쪽에만 치우치거나 편향된 데이터로 알고리즘을 학습한다면 이를 통해 만들어진 인공지능 시스템은 다양한 사회적, 윤리적 문제를 일으키게 된다. 이러한 측면에서 인공지능 시스템 구축에 가장 근본이 되는 학습용 데이터의 품질에 대한 문제는 매우 중요하다고 하겠다.

국내의 경우는 미국이나 중국에 비해 인구수나 기업 수가 상대적으로 적기 때문에 인공지능 시스템을 학습하기 위한 공개 데이터나 전문적으로 제작된 데이터의 수가 부족하다. 이에 인공지능 시스템을 학습하기 위한 다양한 고품질의 인공지능 학습용 데이터가 매우 부족한 실정이다. 이러한 문제를 개선하기 위하여 한국 정부는 인공지능 학습용 데이터 수급의 전략적이고 체계적인 관리가 필요함을 인지하고 인공지능 학습용 데이터 구축 지원사업을 추진해 오고 있다. 이와 같은 인공지능 학습용 데이터 구축 지원사업에 수백 개의 다양한 기업들이 참여하고 있으며, 이를 통해 과거보다 매우 높은 양질의 학습용 데이터가 구축·개방되어 데이터를 필요로 하는 민간 기업, 학교, 연구기관 등이 더욱 발달한 인공지능 시스템을 개발하는 데 사용되고 있다.

그럼에도 불구하고, 현재 개방되어 활용되고 있는 다양한 학습용 데이터는 품질 제고 측면에서 반드시 해결해야 하는 한계점을 가지고 있다. 현재 인공지능 학습용 데이터를 주관하는 기관에서 데이터 품질에 대한 모니터링과 품질 기준들을 제안하고 있지만 이는 기존 일반적인 데이터에 대한 품질, 전반적인 정보 품질의 요인들을 기반으로 하여 만들어진 모형들이기 때문에 세부적인 인공지능 학습용 데이터의 품질을 측정하는 데에는 다소 부족함을 가지고 있다. 또한 인공지능 학습용 데이터에서의 품질이라는 개념을 교육 분야의 개념들로 비교해 보면 이러한 데이터의 품질은 교육 커리큘럼의 품질과 유사하다고 볼 수도 있다. 교육 기관이 얼마나 좋은 커리큘럼을 제공하느냐가 학생들의 학습 품질에 매우 중요한 역할을 하기 때문이다. 인공지능 학습용 데이터의 품질은 결국 인공지능 시스템이 인지적, 자율적 능력을 학습하기 위한 가장 중요한 학습 소스가 된다. 이러한 측면에서 본다면 단순한 데이터의 특성 이외에도 데이터를 제작하는 기관이나 회사의 기술적 측면과 관리적인 측면도 고려해야 더 정확한 데이터 품질을 측정할 수 있을 것으로 보인다.

이론적인 측면에서 보면 정보시스템 분야에서 기업이나 기관에 도입된 정보시스템의 성공 요인을 볼 때 가장 자주 사용되고 있는 정보시스템 성공 모형(Information Systems Success Model)이 있다(Rana *et al.*, 2015). 정보 품질과 시스템 품질이 정보시스템의 만족도에 중요한 선행변수임을 보여주는 이론이다. 정보시스템 성공 모형은 다양한 분야에서 현재까지 활용되고 있는데, 전자상거래 시스템, 온라인 교육, 전자정부 시스템 사용자들의 행동 관련 연구 등에서 빈번하게 활용되고 있다(Mohammadi, 2015; Rana *et al.*, 2015). 그러나 현재까지 연구된 정보시스템 성공 모형을 통해서는 새롭게 대두되고 있는 인공지능 시스템의 성공 요인을 보기 위해서는 다소 부족한 면이 있는데, 이는 바로 알고리즘을 학습하기 위한 데이터 품질 요인에 대한 고려가 없기 때문이다. 데이터를 통해 가공된 정보의 품질을 보는 정보 품질과 데이터 자체의 특성적인 면을 고려하는 데이터 품질은 높은 상호관계를 맺고 있기에 개념상으로 매우 유사하다. 이러한 점 때문에 Madnick *et al.*(2009)은 정보 품질과 데이터 품질은 상황에 따라 다르게 적용 사용이 가능하다고 보았다. 기존의 데이터 품질 관련 연구에서는 한 가지 측면으로 단순하게 측정되는 정보 품질에 초점을 맞춘 반면, 점차 다차원적으로 다양한 측면에서 데이터 품질에 관한 연구

가 수행되고 있다(Batini and Scannapieco, 2016). 그렇기에 본 연구의 맥락인 인공지능 학습용 데이터의 품질에 대해 깊이 있게 이해하기 위해서는 다차원적으로 구성된 선행요인들을 검증하고 제안하는 실증적 연구가 필요하다. 또한 향후 정보시스템 성공 모형의 인공지능 시스템에 대한 성공 요인을 알아보기 위해서도 정보 품질보다는 데이터의 품질을 사용하는 것이 매우 시의적절할 것으로 보인다(Cai and Zhu, 2015). 특별히 이러한 인공지능 학습용 데이터가 품질의 연구에서 데이터 자체가 가지고 있는 전문성과 윤리성에 대한 구체적인 패턴을 알아보는 부분까지 고려한 데이터 품질의 측정과 구체적인 요인들에 관한 연구는 전무하다.

본 연구에서는 국가 전략 차원에서 접근하고 있고 민간에 개방하고 있는 인공지능 학습용 데이터의 품질 요인들을 알아보고자 한다. 향후 수정 보완된 인공지능 시스템 기반 정보시스템 성공 모형의 새로운 데이터 품질 선행변수들을 제안하기 위하여 기존 데이터 품질 관련 연구들과 인공지능 시스템 관련 문헌 연구를 기반으로 인공지능 학습용 데이터 품질에 영향을 주는 데이터의 특성 요인과 데이터의 구축환경 요인들을 재정의하였다. 또한 인공지능 데이터를 구축하고 현업에서 인공지능 서비스를 개발하고 있는 5명의 전문가 인터뷰를 통해 본 연구가 검증하고자 하는 다양한 요인들을 질적 연구방법론을 통해 확인하였다. 이러한 전문가 인터뷰를 통해 검증된 요인들을 기반으로 연구모형을 수립하고 측정 항목을 개발하여 총 393명을 대상으로 온라인 설문조사를 시행하였다. 실제 인공지능 학습용 데이터를 구축하고 인공지능시스템을 개발하고 있는 실무 담당자의 설문 데이터를 수집하고, 수집된 데이터를 기반으로 데이터 특성 요인, 데이터 구축환경 요인, 전문성과 특수성 그리고 윤리와 프라이버시와 관련된 데이터 타입과 데이터 품질 간의 다양한 패턴들과 구성 요소들을 퍼지셋 질적 비교분석을 통해 알아보았다. 또한 사후 분석에서는 본 연구에서 제안하고 있는 총 7가지의 데이터 특성 요인과 데이터 구축 환경요인이 다른 데이터 타입에 따라 어떠한 변수가 가장 중요한지를 인공신경망 분석(Artificial Neural Network)을 통해서 분석하였다. 본 연구의 결과를 토대로 인공지능 학습용 데이터의 품질 측정 요인들의 중요 조합들을 제안하고, 향후 다양한 실증적 연구에서 활용할 수 있는 선행변수들과 잠재적인 조절 변수들을 제안하였으며, 이를 통해 인공지능 학습용 데이터 품질과 선행요인 관련 학술적, 실무적 시사점도 도출 제안하였다.

Ⅱ. 문헌 연구

2.1 인공지능 학습용 데이터 정책

인공지능 학습용 데이터 구축 사업은 '디지털 뉴딜' 핵심 프로젝트인 '데이터 댐'의 대표 사업으로 AI 스피커, 자율주행차, 정밀 의료 등 인공지능 서비스 개발의 필수적인 학습용 데이터를 대규모로 구축· 개방하는 사업이다. 양질의 인공지능 학습용 데이터를 대규모로 구축하여 중소기업, 스타트업 등 민간의 인공지능 기술개발 촉진 및 관련 산업을 육성하고, 코로나발 경제위기 극복을 위한 일자리 창출 등 민간 참여 기반의 인공지능 데이터 선순환 생태계를 조성하는데 목적이 있다. 한국지능정보사회진흥원(2020)에 따르면 사업을 통해 인공지능 학습용 데이터를 구축·개방하고자 하는 기업, 대학, 연구소, 공공기관, 협회, 지자체 등 민간과 공공 구분 없이 모두 수행기관으로 참여할 수 있다.

2019년까지 법률, 특허, 한국어 음성, 이상행동 CCTV 등 21종 4,650만 건, 2020년 170종 4억 8,000만 건, 2021년 190종 5억 6,000만 건 등 총 380종의 데이터가 AI 허브(http://aihub.or.kr)를 통해 개방되고 있으며, 2022년도에는 한국어 음성, 자연어, 의료, 교통물류, 농축 수산, 재난 안전 환경, 문화, 관광, 스포츠, 제조, 로봇공학(로보틱스), 교육, 금융 등 다양한 분야로 데이터 구축 대상을 확대하여 영상, 이미지, 음성, 텍스트 데이터 등 총 310종

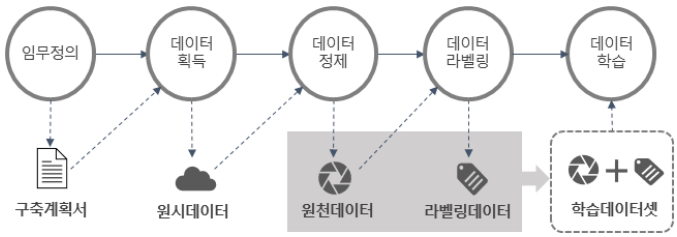
의 데이터를 추가 구축하여 개방하고 있다. 2022년 약 15억 건의 데이터 개방으로 AI 허브 이용자들은 총 691종, 약 26억 건의 데이터를 활용할 수 있게 되었다. 정부는 2025년까지 총 약 2조 5천억 원을 투자해 AI 학습용 데이터 1,300종을 구축하고 개방할 계획이다(과학기술정보통신부, 2022, 2023).

이러한 정부의 인공지능 학습용 구축 사업은 민간 또는 학술 연구기관들이 서비스 개발 및 연구를 위한 양질의 데이터를 확보하는 과정에서 경험하게 되는 데이터 수급의 한계를 극복시켜 줄 수 있는 중요한 사회적 해결책이 된다. 인공지능 학습용 데이터 구축 사업의 의의를 더욱 세부적으로 보면, 첫 번째로 정부 주도로 체계적으로 구축된 데이터들은 학습용 데이터의 질과 양을 향상시켜 알고리즘의 정확도와 효율성을 높인다. 두 번째는 언어 및 문화에 특화된 알고리즘을 개발하는데 필요한 데이터를 적시에 공급함으로써 한국어와 한국문화에 특화된 인공지능 알고리즘, 자연어처리 기술의 발전을 도모한다. 세 번째는 다양한 분야의 학습용 데이터를 정부 주도로 개발함으로써 의료, 금융, 제조, 서비스업 등 다양한 분야에서 인공지능 기술을 적용하여 광범위한 산업으로 빠르게 적용되고 응용될 수 있도록 하고 있다. 네 번째는 이러한 지원을 통하여 치열해지는 글로벌 경쟁에서 한국의 인공지능 기업들이 경쟁력을 확보할 수 있도록 한다. 다섯 번째는 새로운 데이터들을 지속해서 가공하여 공급함으로써 새로운 기술이나 알고리즘을 개발할 수 있는 아이디어를 제공하고, 향후 연구 및 개발에 새로운 방향들을 제시하고 있다. 마지막으로는 학계나 연구기관이 다양한 학습용 데이터를 비용 부담과 제약 없이 접근할 수 있게 됨으로써 연구 성과를 제고하고, 미래 인공지능 기술 분야의 연구 경쟁력을 확보하고 있다. 이렇게 한국의 인공지능학습용 데이터 구축 사업은 국가 전체의 인공지능 관련 기술 발전과 산업적 응용 그리고 글로벌 및 학술 연구의 경쟁력으로 이어지고 있다.

2.2 인공지능 학습용 데이터의 특성과 구축 과정

인공지능 기술 개발에 있어서 중요한 부분 중의 하나인 인공지능 학습용 데이터 확보는 먼저 어떠한 데이터들이 필요하고, 어떠한 특성과 품질을 가진 데이터들이 필요한지 그리고 어떻게 구축된 데이터가 관리되어야 하는지가 매우 중요하다. 인공지능학습용 데이터들은 우리가 일상생활에서 쉽게 구할 수 없는 데이터들이 다수를 차지하고 있다. 예를 들면 의학용 데이터 중, 난소 및 배아 데이터나 내시경 이미지 데이터 그리고 한국인 피부 상태 관련 데이터 등이 있고, 산업 관련 데이터들에는 제조시설 안전 데이터와 화학물질 유해성 예측 데이터와 같이 일반적인 기업들이나 개인들은 접근 자체를 할 수 없고, 자체적으로 구축하기도 매우 어렵다. 이러한 관점에서 정부 주도의 학습용 데이터 구축 사업은 기업들이나, 연구기관 그리고 교육 기관들에 학습용 데이터 구축에 드는 막대한 비용과 시간을 절약시켜 주고, 또한 인공지능 기술의 개발과 도입에 대한 장벽을 낮춰주는 역할도 하고 있다. 추가로, 인공지능학습용 데이터 구축 사업에서 확보한 데이터들은 상당 부분 전문성과 특수성을 요구하는 데이터들이 차지하고 있다. 이러한 방대한 데이터들은 구축 자체뿐만이 아니라 지속적인 업데이트와 관리도 매우 중요하다고 볼 수 있다.

더 나아가 인공지능 학습용 데이터 구축 과정에서 확보된 품질이 학습데이터 전체의 품질을 결정하기 때문에, 인공지능 학습용 데이터의 품질관리를 위해서는 인공지능 학습용 데이터의 구축 과정을 이해하는 것이 필요하다. 인공지능 학습용 데이터의 구축 과정은 다양한 유형의 데이터가 사전에 정의된 목적에 따라 구축되기 때문에 구축 과정이 세부적으로는 다를 수 있으나, 임무 정의, 데이터 획득, 데이터 정제, 데이터 라벨링, 그리고 데이터 학습의 순서로 진행된다. 아래의 <그림 1>은 인공지능 학습용 데이터의 일반적인 구축 과정



출처: 인공지능 학습용 데이터 품질관리 가이드라인 v1.0, 2021.

〈그림 1〉 인공지능 학습용 데이터의 구축과정

을 도식화한 것이다.

인공지능 학습용 데이터의 구축 과정과 이를 통해 도출되는 데이터는 다음과 같은 단계로 설명할 수 있다.

첫 번째, '임무정의' 단계에서는 인공지능이 기계 학습을 통해 해결하고자 하는 문제를 명확하게 정의한다. 또한 문제해결에 필요한 학습용 데이터를 구체적으로 정의하고 설계하는 활동을 수행한다. 두 번째, '데이터 획득' 단계에서는 인공지능의 기계 학습에 필요한 데이터를 현실 세계에서 직접 생성하거나, 이미 보유하고 있는 조직이나 시스템이 인공지능 학습에 필요한 데이터를 수집하는 시점에 포함될 수 있는 개인정보 및 저작권 등 법률적 제약이 없도록 '원시데이터'를 확보하는 활동을 수행한다. 여기서 '원시데이터'는 기계 학습을 목적으로 획득 단계에서 수집 또는 생성된 '음성', '이미지', '영상', '텍스트' 등의 데이터를 의미한다. 세 번째, '데이터 정제' 단계에서는 획득한 원시데이터를 기계 학습에 필요한 형식이나 크기로 맞추고, 데이터의 중복을 제거하며, 원시데이터 획득 시 포함된 개인정보를 비식별화하여 처리하는 등의 과정을 통해 '원천데이터'를 확보한다. 이 단계에서 확보된 '원천 데이터'는 원시데이터가 라벨링 작업에 투입하는 데 필요한 정제 작업이 수행된 데이터로 '라벨링 데이터'가 부여되지 않은 상태의 데이터를 의미한다. 네 번째, '데이터 라벨링' 단계에서는 인공지능이 기계 학습에 활용할 수 있도록 기능이나 목적에 부합하는 '라벨링 데이터'를 원천 데이터에 부착하는 활동을 수행한다. 여기서 '라벨링 데이터'란, 원천 데이터에 부여한 '참값(Ground Truth)', 파일 형식, 해상도 등의 데이터 속성과 설명, 주석 등이 포함된 '어노테이션(Annotation)'의 집합을 의미한다. 다섯 번째, '데이터 학습' 단계에서는 원천 데이터와 라벨링 데이터의 묶음인 '학습데이터 세트'을 이용하여 사전에 정의된 인공지능 알고리즘을 학습시키고, 학습된 인공지능 모델의 성능을 향상하거나 바로잡는 활동을 수행한다.

2.3 인공지능 데이터 품질 및 데이터 특성 조건

최근 사회의 모든 영역에서 인공지능과 기계학습에 대한 수요가 폭발적으로 많아지면서, 효과적으로 인공지능을 학습시키기 위한 다양한 방법들이 연구되고 있다. 전통적인 인공지능은 좋은 수학적 모델 기반 알고리즘을 연구하는 것이었으나, 최근의 인공지능은 데이터 기반의 귀납적 인공지능 모델이 대세를 이루고 있으며, 인공지능으로 성과를 내기 위해서는 양질의 데이터를 확보하는 것이 중요하다고 강조하고 있다(Bertossi and Geerts, 2020; Ng, 2021).

데이터 품질은 기계 학습 모델 성능에 영향을 미치며 데이터 과학자는 모델 교육 전에 데이터 정리에 상당한 시간을 소비한다. 기계 학습 애플리케이션의 품질은 훈련된 데이터의 품질만큼만 우수하며 데이터 정리는 고품질의 기계 학습 모델을 구축하는 초석이라고 할 수 있다(Li *et al.*, 2021).

특히, Ng(2021)의 사례 연구는 인공지능 시스템에서 데이터가 얼마나 중요한 역할을 수행하는지를 잘 보여준다. 사례에서 관련 기업들은 인공지능을 활용하여 철강, 태양광 패널에 결함을 찾는 문제를 개선하고자 하였다. 인공지능 시스템은 코드(모델, 알고리즘)와 데이터로 구성되는데, 모델 중심(Model-centric) 개선 방식으로 결함을 찾고 줄이고자 시도하였지만 시스템 개선 비율은 0~0.04%로 전혀 진척이 없었다. 반면에 데이터 중심(Data-centric) 개선 방식으로 데이터를 개선했더니 시스템 성능을 무려 16.9% 더 정확도가 올라가는 효과를 보였다. Ng(2021)은 이를 근거로 AI 모델 성능개선을 위해서는 데이터 품질을 향상하는 것이 무엇보다 중요하다고 강조하였다. 이후, 기계학습을 위한 데이터 품질 향상에 대하여 Data-centric AI 연구들이 주를 이루고 있다(Lee *et al.*, 2021).

일반적으로 데이터의 품질은 사용자가 운영, 의사결정, 계획을 하는 데 있어 의도된 활용에 적합한 경우에 품질이 우수하다고 이야기할 수 있다. 즉, 데이터가 어떤 목적으로 어떤 시스템상에서 활용되는지에 따라 데이터의 품질을 서로 다른 관점에서 바라볼 수 있는데 인공지능 학습용 데이터는 특정 문제해결을 위한 임무 정의와 기계 학습(Machine Learning) 시스템에 활용되는 데이터라는 특징을 고려하여 새로운 시각에서 데이터 품질을 해석하고 재정의해 볼 수 있다.

기존의 전통적인 데이터 품질특성을 인공지능 학습용 데이터 관점에서 살펴보면 다음과 같이 정리할 수 있다(한국지능정보사회진흥원, 2021, 2023).

- 정확성(Accuracy, 올바른 속성과 정확한 값을 갖는 데이터 특성): 라벨링 데이터의 구조와 형식이 올바르고 분류, 경계 박스 등의 정확한 라벨이 포함된 수준
- 완전성(Completeness, 필요한 모든 속성과 값을 갖는 데이터 특성): 데이터의 특성 정보가 활용 목적에 적합하도록 다양하게 구성되어 있고, 라벨이 누락 없이 포함된 수준
- 일관성(Consistency, 서로 간 모순점이 없고 일관된 데이터 특성): 데이터 라벨링 작업자, 작업방식, 작업시간과 관계없이 라벨링 기준이 전체 데이터 세트에 일관되게 적용된 수준
- 신뢰성(Reliability, 사용자가 믿을 수 있는 속성과 값을 갖는 데이터 특성): 데이터와 라벨 데이터에 대한 사용자의 신뢰 수준
- 현재성(Currentness, 사용 시점에 유효한 값을 가진 속성과 값을 갖는 데이터 특성): 현재 시점에 유효한 데이터와 라벨이 포함된 수준

위와 같이 전통적인 데이터 고유 품질특성을 인공지능 학습용 데이터 관점에서 재정의할 수 있다. 양질의 데이터 확보를 위해서는 품질 기준이 정확하게 확립되어야 한다. 현재 데이터 정제와 관련된 데이터 품질 요소인 일관성(Consistency), 완전성(Completeness) 등에 대해서는 상당 부분 정립된 상태지만 기존 산업과 기술에 맞추어 시도되었던 한계점을 지닌다. 새롭게 적용되는 인공지능 분야에서는 기존의 데이터 품질에서 고려하지 못했던 특화된 데이터 품질이 필요하다. 예를 들어, 최근 인공지능에서 중요한 문제로 대두된 인종 간, 성별 간 차별 문제는 인공지능 학습에 사용된 데이터의 편향성(Bias) 문제일 가능성이 크다. 또한, 원천 데이터에서 수정된 정도에 따라 다른 학습 결과가 나올 수 있으므로 어느 정도 수정이 되었는지, 원천 데이터가 무엇인지를 나타내는 것도 데이터 품질로 포함되어야 한다. 또한 표본 추출 여부와 비율, 믿을 만한 정보 출처인지를 제시하는 것도 인공지능 데이터 품질에서는 관심 사항이라고 할 수 있다(정원섭, 2020).

데이터 품질에 관해 다양한 선행연구들이 있지만, 연구자마다 각기 다른 관점으로 데이터 품질에 대한 정의에 차이를 보인다. 데이터 품질은 사용하는 사용자의 목적에 적합한지 사용자 관점에서 수준을 맞춰야 한다고 주장하였으며(Wang *et al.*, 1995), 사용자에 의해 데이터가 어떻게 인식되고 사용되는지에 따라 데이터 품질은 달라진다

(Miller, 1996). 한편, 다양한 선행연구를 기반으로 데이터 품질은 제공된 데이터가 얼마나 정해진 목적에 부합하여 좋은 결과를 만들어 내는지에 대한 정도로 보는 시각도 있다(Scannapieco *et al.*, 2004). 이러한 데이터 품질은 다양한 연구에서 활용되어 왔는데, 다양한 구성원들과의 협력적 정보시스템에서의 데이터의 품질 측면(Scannapieco *et al.*, 2004; 이용희, 2018), 빅데이터 분석을 위한 데이터의 품질(Cai and Zhu, 2015), 전사적 자원관리 시스템 도입 측면의 데이터 품질(Xu *et al.*, 2002), 공공데이터 품질(Park and Kim, 2015), 조직적 차원에서 데이터베이스관리 및 이해관계자 측면의 데이터 품질(Pipino *et al.*, 2002), 정보시스템의 데이터 품질관리 평가모델(Kim, 2020) 등에서 다양하게 사용되어 오고 있다(<Appendix 2> 참조).

데이터 품질 평가 시 고려해야 하는 데이터 품질 차원에 관한 연구들은 1990년대부터 활발히 전개되기 시작했다. Granick(1991)은 일관성, 수록범위, 최신성, 정확성, 이용자 용이성, 통합성, 사용자 지원성, 출력, 내용, 비용 대비 가치의 10가지 기준을 데이터 품질 요소로 제시하였고, Wang *et al.*(1995)은 사용자 관점에서 중요하다고 생각하는 데이터 품질 요소를 분석하여 본질적, 업무 환경적, 표현적, 접근성이라는 4개의 범주로 구분하여 15개 품질 요소로 상세화하여 제시하였다. Pipino *et al.*(2002)은 접근성, 데이터 양의 적절성, 완전성, 일관성 등 16개 차원을 제시하였다.

Wang and Strong(1996)은 데이터 사용자 관점에서 다양한 데이터 품질 속성을 분석하였는데, 사용자로부터 10개가 넘는 품질 항목을 취합해서 20개의 범주로 분류하고 이를 세부적으로 본질성(Intrinsic), 맥락성(Contextual), 대표성(Representational), 접근용이성(Accessibility)의 네 가지 차원으로 정의하였다. English(1999)는 데이터 품질은 데이터 정의 품질(Data Definition Quality), 내용 품질(Contents Quality), 표현 품질(Presentation Quality)의 세 가지 구성요소로 이루어져 있다고 정의했다. 정의 품질은 정보 제품으로서의 데이터에 대한 정보 제품 명세의 품질이며, 내용 품질은 데이터 값의 정확성을 말한다. 표현 품질은 최종 사용자에게 제공되었을 때의 정보의 품질로 정의하였다. Hoxmeier(1998)는 데이터 품질의 차원을 프로세스 사이클의 품질, 데이터의 품질, 데이터 모델의 품질(의미 구조의 품질), 행위 품질의 네 가지 영역으로 구분하였다. 이 연구는 기존의 데이터 값 및 서비스의 관점에 한정되던 품질 연구를, 데이터 베이스를 구현하는 개발・구현 프로세스 품질과 현행 및 미래의 문제 영역, 즉 사용자 요구사항을 적절히 모두 반영하였는가에 대한 행위 품질 관점으로 확대하였다.

데이터 품질 관련 선행연구에서 가장 많이 인용되는 데이터 품질관리 차원은 정확성, 시의적절성, 신뢰성, 관련성, 완전성 등의 5개의 차원이 인용 빈도가 가장 높았다(Wang *et al.*, 1995). 다른 선행연구에서는 완전성, 의미성, 명확성, 정확성으로 데이터 품질을 차원으로 분리하였고(Wand and Wang, 1996), 다른 연구에서는 고유성, 유용성, 접근성으로 데이터 품질의 카테고리를 분리하여 정의하였다(Haug *et al.*, 2009). 따라서 데이터 품질은 보는 사람의 관점이나 상황에 따라서 달라진다는 것을 알 수 있다.

먼저 이러한 데이터 품질 개념은 크게 가용성, 사용성, 신뢰성, 관련성, 제출 품질 등으로 다섯 가지로 나눠지고 있다(Cai and Zhu, 2015). 데이터 품질을 구성하는 첫 번째 요소는 가용성으로서, 데이터가 쉽게 접근이 가능한가, 그리고 데이터 자체가 얼마나 시의적절한가에 대한 특성을 의미한다(Cai and Zhu, 2015). 두 번째 사용성은 데이터 자체가 가지고 있는 신임성으로서 어떠한 기관에서 이러한 데이터가 제작, 배포되고 있는지에 대한 의미이다. 세 번째로 신뢰성은 데이터가 얼마나 일관성을 가지고 있는가, 무결한가, 완전하고, 정확한가를 보는 요소이다(Cai and Zhu, 2015). 네 번째로 관련성은 주어진 데이터가 활용하는 목적에 얼마나 잘 부합하는지에 대한 적합성을 의미하고, 마지막으로 제출 품질은 데이터 자체의 가독

성이 얼마나 높은가, 인공지능시스템을 학습할 때 얼마나 잘 학습시킬 수 있는가에 대한 요소이다. 기존의 연구에서는 이러한 데이터의 품질을 빅데이터 품질이나 공공데이터의 품질을 측정하는 목적으로 활용하였기 때문에 인공지능 학습용 데이터의 특징인 학습 가능성의 측면은 간과하고 있었으나, 본 연구에서는 이러한 인공지능 학습용 데이터의 특징을 고려하여 학습 가능성 측면을 제출 품질에 추가하였다.

2.4 데이터 구축환경 조건

기존 연구에서는 데이터 품질을 측정할 때 단순히 데이터의 특성 요인만을 고려한 연구가 대부분이다(Batini and Scannapieco, 2016). 그러나 인공지능 학습용 데이터는 데이터 생성 주기별 처리 과정이 기존 데이터들과는 프로세스가 다르다. 일반적인 데이터는 정보시스템에서 생성되고 관리하게 되지만, 인공지능 학습용 데이터는 데이터 생성, 수집 과정부터 프로세스가 다르게 적용된다. 예를 들어, 이미지 학습용 데이터는 이미지 데이터 수집, 생성, 가공 시 수많은 크라우드워커들이 참여하여 데이터를 수집하게 되고, 수집된 데이터는 각 이미지 데이터별로 어노테이션 라벨링 작업과정을 거쳐 학습용 데이터로 결과물이 도출된다. 따라서 데이터 수집 및 가공되는 프로세스 과정에서 데이터의 체계적인 관리가 수행되지 않는다면 데이터 품질이 낮아지게 된다. 따라서 데이터 품질을 향상하기 위해서는 데이터 품질관리가 체계적으로 기술적, 관리적 관점에서 관리되어야 한다(Roh *et al.*, 2019).

인공지능 학습용 데이터의 구축 시 생산되는 원시데이터와 라벨링 데이터의 품질관리가 학습데이터의 품질 측면에서 중요하다. 단, 구축하는 데이터 유형에 따라 텍스트, 음성, 이미지, 영상 등 다양한 유형의 데이터로 구분될 수 있고, 학습 목적에 따라서는 분류, 인식, 검색, 식별, 예측 등 매우 다양한 목적으로 다시 세분될 수 있어, 데이터 유형이나 학습 목적별로 품질 확보 방안을 상세화하여 제시하는 것이 필요하다.

기술적인 관점에서 살펴보면, 데이터 구축 시 라벨링 작업을 위한 플랫폼 또는 소프트웨어 도구의 지원이 필요하다. 각 데이터 객체별 바운딩 박스, 세그멘테이션, 폴리 라인 등 라벨링 작업이 쉽고 편리하게 사용될 수 있는 저작도구와 데이터 작업을 체계적으로 관리할 수 있는 플랫폼 지원이 무엇보다 중요하다. 관리적 관점에서 살펴보면, 데이터 구축 단계별 기준 가이드라인이 필요하다(Ercole *et al.*, 2020). 다양한 크라우드워커 참여로 데이터 구축 작업에 대한 구체적인 기준과 가이드라인 없이는 고품질의 데이터 확보를 기대하기가 어렵다. 데이터 수집단계의 품질관리를 위해서는 법제도 준수, 편향성 방지 등을 위한 기준을 제시하여 개인정보가 포함된 데이터를 수집하는 경우에는 반드시 수집에 대한 동의와 활용 및 제3자 제공 등에 대한 동의를 확보하도록 요구하여야 한다. 특히, 의료 데이터와 같은 경우는 의학연구 윤리심의위원회(Institutional Review Board, IRB)와 데이터 공개에 대한 해당 기관의 동의를 득해야 한다. 데이터 정제 단계의 품질 관리를 위해서는 정제 기준을 명확하게 정의하고, 데이터 구축 목적에 적절한 데이터를 선별하기 위한 명확한 정제 기준을 수립하여야 한다. 또한 기준 미달 또는 활용 불가능한 데이터를 효과적으로 제거할 수 있는 가이드라인을 제공하여야 한다. 가공(라벨링)단계의 품질 관리를 위해서는 라벨링을 위한 작업 매뉴얼을 확보하고, 해당 데이터의 특성을 고려해서 라벨링 작업 방법과 기준에 대한 가이드가 제공되어야 한다. 그리고 지속적인 데이터 모니터링 활동을 통한 품질 검수/피드백을 통하여 데이터 품질을 높이도록 해야 한다(Ercole *et al.*, 2020).

데이터 구축환경 요인들은 데이터를 구축하는 기술적인 측면과 데이터를 관리하는 측면의 요소들로 구분된다. 먼저 데이터 구축 환경에서의 기술적 요인들로 데이터 구축 플랫폼과 데이터 구축 도구, 그리고 데이터 구축을 위한 세부적인 기술이 얼마나 잘 갖춰져 있는가 등의 관리적 요인들

인 데이터 구축 프로세스, 데이터 구축 가이드라인과 데이터 구축 모니터링(평가)로 나눠볼 수 있다. 기존의 많은 정보시스템의 품질연구에서도 시스템들이 가지고 있는 측면(DeLone and McLean, 2004)을 보고 있다. 특별히 TOE(Technology-Organization-Environment) 이론은 조직의 의사결정에 영향을 미치는 영향 요인들을 식별하는 조직 차원의 이론으로(최유진, 양희태, 2023), 기술과 조직, 그리고 환경적인 측면으로 나누어서 정보시스템의 다양한 측면에 관한 연구가 이루어지고 있다(Pudjianto *et al.*, 2011). 따라서 본 연구의 인공지능 학습용 데이터 품질 측면에서도 이러한 기술적인 측면과 조직의 관리적인 측면의 요인들이 사용되는 것이 데이터 품질을 측정함에 있어서 정확도를 높이는 중요한 전략으로 보인다.

2.5 데이터 타입 조건

본 연구에서는 인공지능 학습용 데이터가 가진 특성을 구체적으로 알아보기 위해 두 가지의 데이터 타입 조건을 추가하였다. 먼저 첫 번째 조건 변수로는 설문에 참여하는 데이터 구축업체가 실제 제작하고 있는 데이터의 전문성이나 특수성을 고려하였다. 텍스트 데이터나 간단한 이미지 데이터 구축업체와 동영상이나 특수데이터(3D, 의료용) 등 전문성과 특수성 있는 데이터 구축기업과 난이도가 쉬운 일반적인 데이터를 구축하는 기업과의 데이터 품질 요인에서 어떠한 차이가 있는지를 보기 위함이다. 예를 들어 일반적인 이미지 라벨링 데이터의 경우는 크라우드워커들이 기본적인 교육 이수 후, 어려움 없이 데이터 구축 작업이 가능한 반면, 전문성이 요구되는 의료용 데이터의 경우는 IRB 승인을 거쳐 전문직 의사들이 직접 라벨링 작업에 참여하여야만 정확한 데이터를 확보할 수 있어 데이터 구축 시간과 비용이 많이 증가하게 된다.

두 번째로는 최근 쟁점이 되는 인공지능 윤리, 프라이버시에 민감한 데이터를 제작하는 기업과 이러한 이슈에서 자유로운 데이터 구축업체를 비교하여 데이터 품질 요인들에 대한 구성의 차이를 보이는지를 살펴보고자 한다. 개인의 프라이버시 이슈로 인하여 데이터 구축 시 개인정보 활용에 문제가 없도록 가명 정보 처리나 비식별조치를 추가로 진행함에 따라 추가 작업 진행되면서 시간이 소요되고 데이터 구축 비용이 증가하는 등 어려움이 발생하게 된다. 또한 인종, 성차별, 성적 비하 등 윤리적 이슈가 발생하지 않도록 데이터에 대한 추가 작업이 필요하여 일반적인 데이터에 비해 추가로 시간과 비용이 발생한다.

Ⅲ. 예비 조사: 전문가 인터뷰 결과

본 연구에서는 연구모형의 중요 변수들을 도출하기 위해 문헌연구와 더불어 전문가 인터뷰 또한 수행하였다. 인터뷰는 인공지능 학습용 데이터 구축과 실제 인공지능 서비스를 제공하고 있는 기업의 실무 임원 및 책임자들과 진행하였다. 총 5명의 전문가와 인터뷰를 진행하였으며, 인터뷰는 2023년 2월 7일부터 2월 9일까지 서면 인터뷰와 면대면 인터뷰로 진행하였다.

인터뷰 결과를 분석하기 위해 질적방법론 전문가 2인으로 구성된 코딩팀이 먼저 개방 코딩(Open Coding)을 통해 가장 중요한 키워드들을 도출해 내고, 이를 개념화하여, 실제 전문가들의 생각들을 구체화하였다. 2단계 축 코딩(Axial Coding)을 통해 전문가들이 제안하고 있는 중요한 개념들을 묶는 키워드를 찾아내어 본 연구가 제안하고 있는 요인들과 일치시키는 작업을 진행하였다. 아래 <표 1>은 본 연구의 코딩 결과물로서 전문가들이 제안하고 있는 중요한 요인들을 기존 연구에서 제안하고 있는 다양한 품질 선행요인들과 매칭시켜 보았다. 이러한 인터뷰 분석 결과를 통해서 보면 전문가 대부분이 데이터 특성 요인들과 데이터 구축 요인 그리고 데이터 형태 타입이 데이터의 품질에 매우 중요한 요인들임을 확인해 주고 있다. 이러한 전문가 인터뷰 결과를 기반으로 아래 <그림 2>의 개념적 모형을 개발하였다.

〈표 1〉 전문가 인터뷰 코딩 결과

구분	데이터 특성 요인		CEO (대표)	AI 프로젝트 책임자	데이터품질 담당	AI 프로젝트 책임자	AI 프로젝트 책임자
			자연어 처리 전문 인공지능 기업 A	자연어 처리 전문 인공지능 기업 B	자연어 처리 전문 인공지능 기업 B	이미지 프로세싱 전문 데이터 구축 기업 C	이미지 기반 인공지능 서비스 기업 D
데이터 특성요인	가용성 (Availability)	접근성 (Accessibility) 시의적절성 (Timeliness)	데이터 지속성	시기적절한 데이터 접근성	시기적절한 데이터 접근성		
	사용성 (Usability)	신임성 (Credibility)	응용 적용 가능	서비스 개발에 사용가능한 데이터, 저작권 문제해결	서비스 개발에 사용가능한 데이터, 저작권 문제해결		
	신뢰성 (Reliability)	정확성 (Accuracy) 일관성 (Consistency) 무결성 (Integrity) 완전성 (Completeness)	다양성, 구문 정확성, 의미 정확성	전문성 보장이 일관성으로 연결	전문성 보장이 일관성으로 연결	다양성, 구문 정확성, 의미 정확성	다양성, 구문 정확성, 의미 정확성
	관련성 (Relevance)	적합성 (Fitness)	현실과의 차이, 목적에 맞는 데이터 형식 필요	목적에 맞는 데이터 형식 필요	목적에 맞는 데이터 형식 필요		
	제출품질 (PresentationQuality)	가독성 (Readability) 학습가능성 (Trainability)	학습목적에 맞는 형식이 중요함			범용성 있는 데이터 구축필요	범용성 있는 데이터 구축필요
데이터 구축환경	기술적 측면 (Technology Perspective)	플랫폼 구축도구 구축 기술	데이터 모델의 변화, 구축관련 기술의 축적이 관리능력으로 연결	관리 저작도구 구축	관리 저작도구 구축	구축기술의 변화 시 즉시 적용	
	관리적 측면 (Management Perspective)	프로세스 가이드라인 모니터링	전문가 확보, 인력양성 프로그램. 세부 가이드라인 필요	데이터에 맞는 구체적 가이드라인, 평가기관의 긴밀한 모니터링	데이터에 맞는 구체적 가이드라인, 품질전담 조직 필요	시간준수 어려움, 평가기관의 긴밀한 모니터링	구체적인 가이드라인
데이터 형태	프라이버시, 윤리		윤리, 프라이버시 고려, 편향성 데이터 관리	윤리, 프라이버시 고려, 편향성 데이터 관리	윤리, 프라이버시 (비 식별 조치)	윤리, 프라이버시 (비 식별 조치)	윤리, 프라이버시 (비 식별 조치)
	전문성, 특수성		데이터의 전문성을 고려해야 일관성이 확보됨	데이터의 전문성을 고려해야 일관성이 확보됨	복합적 추론과정에 따른 전문성, 특수성 요구	멀티 모달 같은 복잡한 데이터로 전문성 요구	3D 데이터 구축 시 전문성 증가
기타	인증 및 교육 관련		품질인증제도, 등급제	품질인증제도, 등급제		품질교육	
	AI 구축 관련		AI 모델 정확도 개선, 시차문제 (개발과 출시)	AI 모델 정확도 개선, 시차문제 (개발과 출시)	AI 모델 정확도 개선, 시차문제 (개발과 출시)		

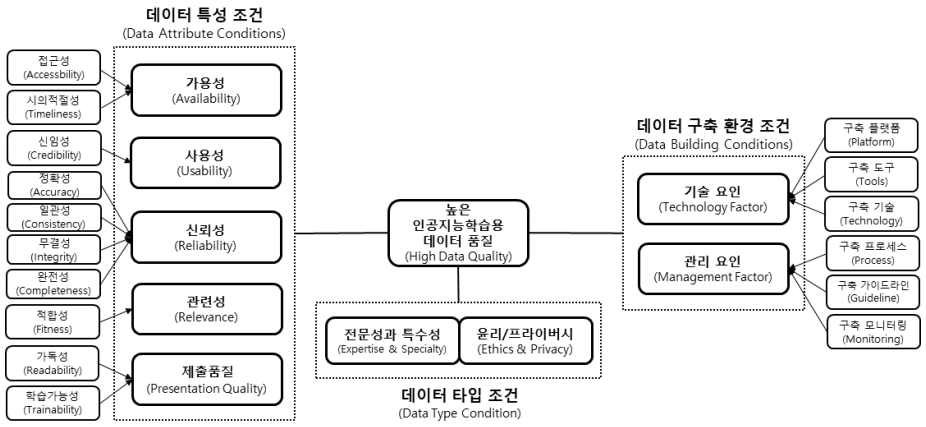
Ⅳ. 연구 방법

본 연구에서는 인공지능 학습용 데이터 품질 지표 개발을 위하여 기존 데이터 품질 및 빅데이터/인공지능 관련 데이터 품질 연구들의 문헌 연구를 통해 데이터 품질과 관련된 선행 변수들을 선정하였다. 선정된 변수들을 측정하고 분석하기 위하여 도출된 선행 변수의 설문 항목들을 기존 문헌들을 기반으로 수정하여 설계하였다. 제안하고 있는 변수들 간에 어떠한 인과적 구성을 만들어 내는지를 보기 위해서 퍼지셋 질적 비교 방법론을 수행하였으며, 마지막으로는 사후 분석으로 제안된 변수들 가운데 어떠한 변수가 학습용 데이터 품질에 가장 강하게 영향을 주는지를 예측하기 위해서 인공신경망 분석을 수행하였다.

퍼지셋 질적 분석과 인공신경망 분석을 위한 데이터 수집을 위해 실제 인공지능 학습용 데이터를 발굴하고 제작하고 있는 기업들의 임원과 실무 책임자를 대상으로 온라인 설문조사를 실시하여, 총 393개의 자료를 수집하였다. 수집된 데이터 세트는 fsQCA 2.5 소프트웨어를 이용하여 퍼지셋 질적비교분석 방법을 수행하였으며 통계 프로그램인 SPSS 18.0을 이용하여 인공신경망 분석(Artificial Neural Network)을 추가로 수행하였다. 아래의 <그림 2>는 본 연구의 개념적 모형으로 크게 데이터 특성 요인과 데이터 구축 환경요인으로 나누어 총 7개의 선행변수와 데이터 품질 간의 관계를 분석하였다.

4.1 연구방법론 및 측정 도구 개발

본 연구에서 제안하고자 하는 인공지능 학습용 데이터의 품질에 대한 선행요인은 크게 두 가지의 다른 품질 요인으로 나뉜다. 첫 번째는 데이터 자체가 가지고 있는 특성 요인으로서 기존의 데이터 품질과 정보 품질에서 다루어져 오던 변수들과 추가로 인공지능 학습용 데이터가 가지고 있는 특성을 재정의하여 이에 맞게 기존 연구의 변수를 수정한 요인들이고, 두 번째로는 데이터 구축 환경적 요인으로서 데이터를 수집, 처리, 제작하는 기업들이 가지고 있는 시스템적, 관리적 요인들로 나뉘어져 있다. 기존 연구에서는 이러한 데이터의 특성이나 구축환경 측면을 나누어서 분석해 왔는데, 본 연구에서 이러한 두 가지 측면의 변수들이 인공지능 학습용 데이터 품질에 어떻게 영향을 주는지를 통합적인 시각으로 접근하였다. 이에 이러한 다양한 요인들을 분석하기 위해 기존 사례나 변수 중심의 연구방법론이 아닌, 조합된 요인들과 결과의 상호 인과관계가 갖는 인과적 복잡성을 살펴보고, 주어진 조건에 따라 어떠한 다양한 구성



〈그림 2〉 개념적 모형

(Configuration)을 하는지를 분석해 볼 수 있는 퍼지셋 질적비교분석 방법론을 도입하여 데이터를 분석하였다(Ragin, 2006, 2009). 기존의 구조방정식이나 회귀분석처럼 하나의 예측 변수와 종속 변수의 관계를 중심으로 분석하는 것이 아닌, 다양한 변수들이 주어진 조건에 따라 다른 조합과 구성들을 만들어 내는 것을 검증하고, 실제 현장에서 접할 수 있는 인과적 복잡성과 데이터 품질의 역동성을 알아봄으로써 더욱 실제적인 데이터 품질 측정 방안을 제안할 수 있을 것이다. 추가로 퍼지셋 질적비교분석을 수행한 이유로는 해당 분석은 기존의 통계적 방법론처럼 기존 연구들에 의해 정해진 통계적 범위와 프로세스 안에서 분석하는 구조방정식이나 회귀분석과는 달리 분석자가 본인의 경험과 지식수준에 따라 데이터 안의 특정 사례들을 특징 하는 과정에서 퍼지셋 논리들이 어떠한 그룹에 소속(Membership)되어 있는지에 대한 소속 조건을 유연하게 조절할 수 있다. 이러한 프로세스를 통해 연구자는 질적 분석의 유연성과 정확성 및 정밀성을 조절하여 더욱 구체적인 분석 결과를 도출할 수 있다(Ragin, 2009). 또한 기존 질적 사례 연구나 인터뷰 연구와는 다르게 리커드 7점 척도를 활용하여 데이터를 수집하여 분석할 수 있으므로, 추출된 사례 간의 특징을 분석할 때 정량적인 접근이 가능하여 더 실증적인 정교한 분석을 수행할 수 있다(Ragin and Davey, 2014). 마지막으로는 퍼지셋 질적 비교분석은 사회과학의 집합 이론성(Set-theoretic)을 반영함으로 이를 활용하는 연구들이 이론들과 사례의 경험적인 자료의 통합적 정량 분석이 가능하게 한다(Fiss, 2007). 본 연구가 제안하고 있는 두 가지의 데이터 품질 선행요인들이 데이터가 가지고 있는 특수성/전문성이나 윤리 또는 프라이버시 관련 이슈의 낮고 높은 정도에 따라 어떠한 패턴을 보이는지 등의 복합적인 패턴 또한 분석할 수 있게 하는 방법론으로, 연구자들에게 좀 더 통합적인 이론적 시사점을 제안하도록 한다(Fiss, 2007, 2011; Ragin, 2009; Ragin *et al.*, 2006).

본 연구의 측정 항목들은 기존의 연구들이 제안하고 있는 측정 항목들을 기반으로 본 연구의 맥락에 맞게 수정하여 사용하였다. 설문에 참여하는 응답자는 인공지능 학습용 데이터를 제작, 관리하는 실무책임자와 임원으로 한정하고, 데이터 가공이 쉬운 데이터들과 복잡한 데이터를 처리하는 회사를 구분하였으며, 윤리적, 프라이버시 관련 이슈가 있는 데이터를 취급하는 기업들 나누어서 설문조사를 실시하였다(자세한 측정 항목은 <Appendix 1> 참조).

4.2 데이터 수집

본 연구의 설문 데이터 수집은 현재 한국지능정보사회진흥원(NIA, National Information Society Agency)에서 실시하고 있는 인공지능 학습용 데이터 구축 사업에 참여하고 있는 기업들의 임원들과 실무자들을 대상으로 이루어졌다. 2023년 4월 1일부터 2022년 4월 6일까지 실시하여, 총 405개를 수집하여 이중 불성실하게 응답한 12개의 데이터를 제거하고 총 393개의 데이터를 최종 분석에 사용하였다. <표 2>는 설문에 참여한 응답자들의 인구통계학적 정보이다.

〈표 2〉 응답자의 인구통계학적 정보

특성	항목정보	빈도	%
성별	남성	323	82.2%
	여성	70	17.8%
연령	18~29세	30	13.4%
	30~39세	97	25.8%
	40~49세	150	21.3%
	50~59세	98	12.1%
	60~69세	18	10.2%

〈표 2〉 응답자의 인구통계학적 정보(계속)

특성	항목정보	빈도	%
학력수준	고등학교 졸업 또는 동등한 자격	14	3.6%
	학사 학위	157	39.9%
	석사 학위	120	30.5%
	박사 학위	102	26.0%
회사 연 매출	10억원 미만	81	20.6%
	10억원 이상 50억원 미만	142	36.1%
	50억원 이상 100억원 미만	55	14.0%
	100억원 이상 500억원 미만	64	16.3%
	500억원 이상 1,500억원 미만	32	8.1%
	1,500억원 이상	19	4.8%
직급	과장 직급 또는 그 이하	110	28.0%
	차장 직급 또는 그와 동등한 직급	39	9.9%
	부장 직급 또는 그와 동등한 직급	82	20.9%
	상무 또는 그와 동등한 직급	70	17.8%
	전무 또는 그와 동등한 직급	23	5.9%
	전무 이상	69	17.6%
인공지능관련 기업 경력	6개월 미만	21	6.7%
	6개월 이상 1년 미만	42	13.4%
	1년 이상 2년 미만	28	8.9%
	2년 이상 3년 미만	71	22.6%
	4년 이상 5년 미만	53	16.9%
	5년 이상	24	7.6%
회사 규모	10명 미만	49	12.5%
	10명 이상 50명 미만	178	45.3%
	50명 이상 100명 미만	61	15.5%
	100명 이상 200명 미만	40	10.2%
	200명 이상 300명 미만	10	2.5%
	300명 이상	55	14.0%

4.3 눈금매기기

fsQCA에서는 기존의 선행연구나 연구자의 지식에 기반하여 결과 조건과 원인 조건에 해당하는 데이터들을 퍼지 점수(0-1)로 환산(Calibration)하는 과정으로 분석을 준비한다(Ragin, 2009). 본 연구에서는 결과 조건 변수인 인공지능 학습용 데이터 품질 변수와 7개의 원인 조건 변수를 기존 연구의 설문 항목들을 기반으로 본 연구의 맥락에 맞게 수정하여 측정하고, 측정된 값을 퍼지 점수로 환산하여 사용하였다. 또한 추가적인 원인 조건 변수인 데이터 형식(전문성/특수성과 윤리/프라이버시)을 실제 이러한 전문성이나 특수성을 요구하는지 그리고 윤리나 프라이버시 이슈가 있는지를 0, 1, 2로 측정하였고, 이 값들을 퍼지 점수로 환산하였다. 리커트 7점 척도를 활용하여 측정한 변수들은 최대값인 Full membership은 6으로, 중간값인 Cross-over point는 4로 그리고 최소값인 Non-full membership은 2로 하여 퍼지 점수를 환산하였다(이현애 등, 2019).

〈표 3〉 결과 및 원인조건 변수 눈금 매기기

구분			설명	퍼지점수 환산
결과 조건 변수		인공지능학습용 데이터 품질	동의 여부 (7점 척도)	항목별 동의 비중 최대값 6 중간값 4 최소값 2
원인 조건 변수	데이터 특성 요인	가용성 사용성 신뢰성 관련성 제출 품질	동의 여부 (7점 척도)	항목별 동의 비중 최대값 6 중간값 4 최소값 2
	데이터 구축 환경 요인	데이터 구축 기술 요인 데이터 구축 관리 요인	동의 여부 (7점 척도)	항목별 동의 비중 최대값 6 중간값 4 최소값 2
	데이터 타입	전문성/특수성 윤리/프라이버시	구축 중인 데이터 타입에 따른 분류	타입과 관련성 높음 2 타입과 관련성 낮음 1 타입과 관련성 없음 0

Ⅴ. 데이터 분석 및 결과

5.1 기초 데이터 분석

퍼지셋 질적 비교 분석을 하기 전 측정 모형의 신뢰성과 타당성 확인을 위해 7개의 변수의 세부 이차순위 변수들의 Cronbach's Alpha, Henseler's Rho (ρA), 복합신뢰도(CR), Average Variance Extract (AVE) 등과 각 변수의 요인 적재량(Loading) 값들을 Adanco 3.0 소프트웨어를 이용하여 확인하였다. 알파와 ρA 그리고 CR은 모두 기준치인 0.7 이상으로 본 연구에서 사용한 측정 변수들이 모두 내적 일관도와 신뢰성을 가지고 있는 것으로 보인다(Bagozzi and Yi, 1988; Bryman, 2016). 또한 요인 적재량도 모두 0.7 이상으로 Fornell and Lacker (1981) 집중 타당도를 확보하고 있으며(Gefen and Straub, 2005), AVE (평균 분산추출) 값도 0.5 이상으로 기존 연구들이 제안하고 있는 기준치를 상회하므로 집중 타당도 또한 확보한 것으로 보인다(Podsakoff *et al.*, 2003). 마지막으로 판별타당도는 모든 변수의 요인 적재량에 평균 분산 추출의 제곱근 값과 행과 열에 있는 다른 값들과 비교했을 때 상대적으로 큰 값을 가질 때 판별타당도를 확보했다고 보는데 <표 5>에 보면 모든 값이 행과 열에 있는 값들보다 큰 것으로 확인되어 판별타당도 역시 확보한 것으로 보았다(Fornell and Larcker, 1981). 이러한 결과를 통하여 본 연구에서 사용한 측정 모형이 퍼지셋 질적비교분석 진행에 아무런 문제가 없음을 확인하였고, 이에 분석을 시행하였다.

〈표 4〉 측정 조건 변수에 대한 기술통계와 신뢰도

변수 Construct	평균/표준편차 Mean/S.D.	요인적재량 Loadings	Cronbach's Alpha	rho_A	CR	AVE
접근성	5.856/1.151	0.85 0.78 0.87	0.787	0.839	0.869	0.690
정확성	5.351/1.789	0.93 0.93 0.89	0.904	0.905	0.940	0.839

〈표 4〉 측정 조건 변수에 대한 기술통계와 신뢰도(계속)

변수 Construct	평균/표준편차 Mean/S.D.	요인적재량 Loadings	Cronbach's Alpha	rho_A	CR	AVE
완전성	5.165/1.920	0.98 0.98	0.954	0.955	0.978	0.956
일관성	5.281/1.866	0.93 0.94 0.92	0.915	0.917	0.946	0.855
신임성	5.331/1.843	0.91 0.92 0.94	0.914	0.916	0.946	0.853
구축 가이드라인	5.313/1.868	0.92 0.93	0.830	0.831	0.922	0.855
구축 모니터링	5.072/1.040	0.96 0.96 0.96 0.92	0.964	0.965	0.974	0.904
구축 플랫폼	4.918/1.253	0.89 0.96 0.95	0.927	0.935	0.954	0.874
구축 프로세스	5.284/1.877	0.93 0.93 0.91 0.91	0.941	0.941	0.958	0.850
데이터 품질	5.189/1.851	0.90 0.91 0.89 0.90 0.89 0.92 0.88 0.91	0.966	0.966	0.971	0.806
구축 기술	5.053/1.227	0.96 0.98 0.97	0.967	0.970	0.978	0.937
구축 도구	5.928/1.281	0.93 0.97 0.97	0.953	0.956	0.970	0.914
적합성	5.333/1.801	0.96 0.94 0.94	0.941	0.941	0.962	0.894
무결성	5.332/1.804	0.90 0.95 0.94	0.917	0.922	0.948	0.858
가독성	5.234/1.858	0.94 0.95 0.93	0.932	0.932	0.957	0.880
시의적절성	4.925/1.108	0.81 0.90 0.92 0.93	0.914	0.917	0.940	0.797
학습가능성	5.304/1.847	0.97 0.97 0.95	0.960	0.960	0.974	0.926

〈표 5〉 측정 조건 변수에 대한 판별타당도

Construct	접근성	정확성	완전성	일관성	신임성	구축 가이드라인	구축 모니터링	구축 플랫폼	구축 프로세스	데이터품질	구축 기술	구축 도구	적합성	무결성	가독성	시의적절성	학습가능성
접근성	**0.83**																
정확성	0.59	**0.92**															
완전성	0.68	0.63	**0.98**														
일관성	0.66	0.72	0.73	**0.93**													
신임성	0.70	0.73	0.75	0.79	**0.92**												
구축가이드라인	0.59	0.71	0.65	0.77	0.76	**0.93**											
구축 모니터링	0.67	0.57	0.69	0.69	0.74	0.68	**0.95**										
구축 플랫폼	0.49	0.50	0.47	0.52	0.52	0.56	0.64	**0.94**									
구축 프로세스	0.77	0.69	0.74	0.76	0.83	0.75	0.79	0.61	**0.92**								
데이터품질	0.71	0.76	0.76	0.85	0.83	0.81	0.74	0.57	0.79	**0.90**							
구축 기술	0.55	0.48	0.51	0.53	0.55	0.53	0.69	0.80	0.62	0.57	**0.97**						
구축 도구	0.44	0.39	0.46	0.49	0.48	0.47	0.62	0.69	0.56	0.48	0.70	**0.96**					
적합성	0.70	0.74	0.75	0.82	0.84	0.78	0.70	0.51	0.80	0.85	0.54	0.48	**0.95**				
무결성	0.68	0.69	0.73	0.78	0.84	0.72	0.69	0.51	0.79	0.81	0.54	0.47	0.82	**0.93**			
가독성	0.66	0.67	0.74	0.77	0.80	0.72	0.75	0.52	0.80	0.81	0.52	0.48	0.83	0.79	**0.94**		
시의적절성	0.65	0.52	0.61	0.60	0.64	0.59	0.74	0.52	0.69	0.68	0.58	0.52	0.65	0.60	0.65	**0.89**	
학습가능성	0.70	0.72	0.78	0.82	0.85	0.74	0.73	0.52	0.83	0.82	0.58	0.50	0.85	0.82	0.79	0.64	**0.96**

주) 굵은 글씨는 해당 변수의 AVE의 제곱근 값.

5.2 퍼지셋 분석 결과

본 연구에서는 인공지능학습용 데이터의 높은 품질과 관련된 구체적인 조건 변수들의 조합을 알아보기 위해 데이터의 특성 조건과 데이터 구축환경 조건으로 주축으로 하여 데이터 타입 조건까지 도입하여 총합적인 관점에서 데이터 품질과 관련된 충분조건을 분석하였다. 퍼지셋 질적 비교분석 소프트웨어를 활용하여 분석한 결과 연구에서 제일 중요한 복합모형과 최소 간결 모형을 아래 <표 6>과 같이 도출되었다. 전체 모형에 대한 설명력(Coverage)은 96%와 일관성(Consistency)은 89%로 도출되었다. 퍼지셋 질적비교분석에서 설명력은 결과 조건인 높은 데이터 품질이 어느 정도 원인 조건들인 데이터 특성, 데이터 환경 그리고 데이터 타입에 의해 설명되는지를 보여주고, 일관성은 결과 조건인 높은 데이터 품질에 포함되는 항목들이 가진 퍼지 점수가 결과 조건의 부분집합에 어느 정도인지를 의미한다(Ragin, 2008). 기존 퍼지셋 질적비교분석에서는 설명력은 R^2와 비슷한 의미로 제안된 연구모형의 설명력과 유사하게 사용되고 있으며, 일관성은 주로 0.75 이상이 적절한 수준으로 받아들여지고 있다. 이러한 기존 연구를 토대로 보면 본 연구의 설명력과 일관성은 모두 적절한 것으로 보인다(Schneider and Wagemann, 2012).

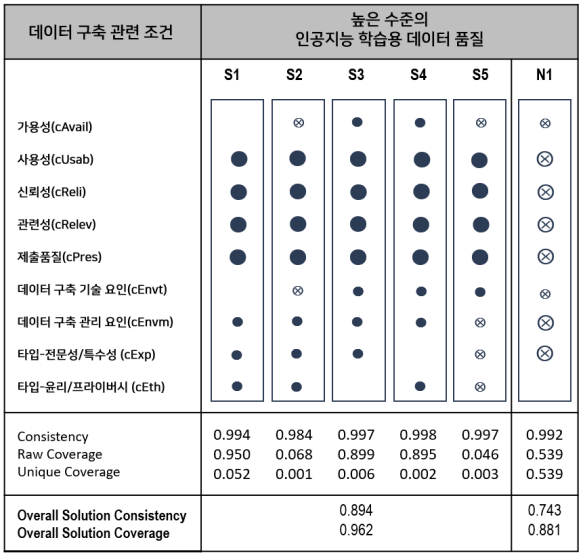
퍼지셋 질적비교분석 결과를 기반으로 <그림 3>과 같이 5개의 복합모형이 도출되었다. 5개의 복합모형은 크게 3개의 특징적인 조합으로 나누어 볼 수 있다.

〈표 6〉 조건 변수에 대한 퍼지셋 질적비교분석 결과

모형 구분	결합조건
복합모형 (Complex solution)	**cUsab*cReli*cRelev*cPres***cEnvm*cExp*cEth ~cAvail***cUsab*cReli*cRelev*cPres***~cEnvt*cExp*cEth cAvail***cUsab*cReli*cRelev*cPres***cEnvt*cEnvm*cExp cAvail***cUsab*cReli*cRelev*cPres***cEenvt*cEnvm*cEth ~cAvail***cUsab*cReli*cRelev*cPres***~cEnvt*cEnvm*~cExp*~ cEth
최소간결 모형 (Parsimonious solution	cUsab cUsab*cReli cReli*cRelev cUsab*cPres

주) cAvail: 가용성, cUsab: 사용성, cReli: 신뢰성, cRelev: 관련성, cPres: 제출품질, cEnvt: 기술요인, cEnvm: 관리요인, cExp: 전문성/특수성 데이터 타입, cEth: 윤리/프라이버시 데이터 타입

첫 번째 조합은 가장 기본이 되는 조합으로, S3 (cAvail * cUsab * cReli * cRelev * cPres * cEnvt * cEnvm * cExp), S4(cAvail * cUsab * cReli * cRelev * cPres * cEnvt * cEnvm * cEth)의 두 조합으로 나타났다. S3은 전문성/특수성이 있는 타입의 데이터를 가공하는 기업들, S4는 윤리/프라이버시와 관련 있는 데이터를 가공하는 기업들을 구분하여 분석한 결과를 보여준다. 이 두 가지 데이터 타입은 동일하게 신임성을 나타내는 사용성, 정확성, 일관성, 무결성, 완전성을 나타내는 신뢰성, 적합성을 나타내는 관련성, 가독성과 학습 가능성을 나타내는 제출 품질의 데이터 속성이 높은 데이터 품질에 가장 영향을 많이 주는 것으로 나타났다. 또한 접근성과 시의 적절성을 나타내는



주) ●: 조건의 존재, 핵심 요인, ⊗: 조건의 부재: 핵심 부재, •: 주변적 요인, ⊗: 주변적 부재, 공백은 상관없음을 나타낸다.

〈그림 3〉 높은 인공지능학습용 데이터 품질에 대한 퍼지셋 질적비교분석 결과

가용성과 플랫폼, 도구, 기술들과 관련된 데이터 기술 요인 그리고 프로세스, 가이드라인, 모니터링과 관련된 데이터 관리요인도 다소 낮지만 데이터 품질에 영향을 주는 것으로 분석되었다.

두 번째 조합은 데이터를 구축하는 기업이 전문성/특수성, 그리고 윤리/프라이버시와 관련 있는 데이터를 모두 취급하는 경우에서 나타나는 조합으로 S1(cUsab * cReli* cRelev * cPres * cEnvm * cExp * cEth), S2(~cAvail * cUsab * cReli * cRelev * cPres *~ cEnvt * cExp * cEth)로 서로 유사한 조합을 보인다. S1과 S2의 조합에서 공통적으로 가장 중요한 데이터의 특성 요인은 사용성, 신뢰성, 관련성, 제출 품질이며, 부가적인 특성으로는 데이터 관리요인이 중요한 것으로 나타났다. 특별히 가용성과 데이터 기술 요인은 S1에서는 크게 상관없는 특성으로 나왔으며, S2에서는 전혀 고려되지 않고 있는 것으로 나왔다. 이는 전문성/특수성이 필요하고, 윤리와 프라이버시에 민감한 데이터를 가공하는 기업이 높은 데이터 품질을 얻기 위해서는 가용성이나 데이터 구축 기술요인은 크게 중요하지 않으며, 더욱 중요한 것은 사용성, 신뢰성, 관련성, 제출 품질의 데이터 속성과 데이터 구축의 관리요인이 중요하게 고려되어야 함을 나타낸다.

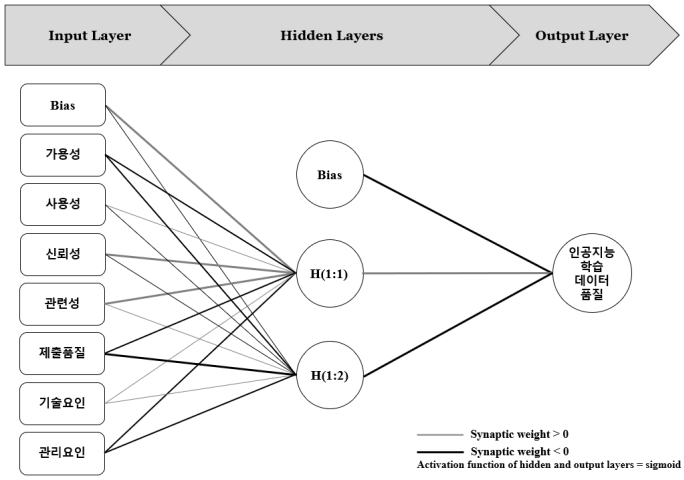
세 번째는 S5(~cAvail * cUsab * cReli * cRelev * cPres * cEnvt *~ cEnvm *~ cExp *~ cEth)의 조합으로, 전문성/특수성 그리고 윤리/프라이버시와 전혀 관련 없는 데이터를 구축하는 기업에 관한 결과를 보여준다. 이와 같은 기업들 역시 높은 데이터 품질을 얻기 위해서는 위의 네 가지의 조합들과 동일하게 사용성, 신뢰성, 관련성, 제출 품질이 중요하고, 추가로 데이터 구축 기술요인이 중요한 것으로 나타났다. 단 이들 기업에서는 상대적으로 가용성과 데이터 구축 관리요인은 크게 중요하지 않은 것으로 나타났다.

이러한 퍼지셋 질적비교분석 결과로 보면 기존 연구에서 중요하게 보고 있는 다섯 가지의 데이터 특성 중 네 가지의 속성이 데이터의 품질을 결정하는 데 중요한 조합변수로 보이며, 데이터 구축의 기술 요인과 관리 요인도 데이터의 타입에 따라 중요한 역할을 하는 것으로 보인다. 인공지능 학습용 데이터는 무엇보다 최근에 쟁점이 되는 책임있는 인공지능, 인공지능 편향(Bias) 등과 같은 윤리 및 프라이버시와 같은 민감한 문제들을 데이터 구축에 고려해야 하고, 이러한 데이터 타입에 따라서 데이터 품질을 결정하는 조건들과 조합들이 변화될 수 있음을 본 연구의 결과로 발견하였다.

5.3 사후분석-인공신경망 분석(ANN)

본 연구에서는 퍼지셋 질적비교분석 결과를 보조하기 위하여 최근에 많은 행동 연구에서 사용되고 있는 인공신경망 분석(Artificial Neural Network analysis, ANN) 방법을 도입하여 추가적인 사후 분석을 시행하였다. fsQCA 분석 결과로 확인된 서로 다른 타입의 데이터들을 분석하여 이들 간에 어떠한 차이를 보이는지를 비교 분석하고자 하였다. 인공신경망 분석은 비선형 관계와 선형 관계 모두에서 사용될 수 있으며(Chong, 2013; 이원국, 양희태, 2022) 특별히 독립변수와 종속변수 간에 선형 관계가 없는 상황에서도 어떠한 변수들이 종속변수에 영향을 주는지를 예측할 수 있다(Chong *et al.*, 2016). 또한 인공신경망 분석은 기존 구조방정식 등과 같은 방법론들이 요구되는 다변량의 가정도 필요 없으므로 가설을 설정하지 않고도 변수 간의 관계를 예측할 수 있는 장점이 있다(Hew *et al.*, 2019). 이러한 인공신경망 분석의 장점으로 인해 최근에는 PLS-SEM, fsQCA 등 다양한 방법론들과 함께 사용되고 있는데(Sharma M. et al., 2022, Jun-Jie Hew et al, 2023), 대표적으로 구조방정식 및 퍼지셋 질적비교분석 방법론들과도 함께 사용하여(Chan and Chong, 2012; Li et al., 2022), 모형의 예측이나, 주어진 조건 조합 중 가장 중요한 변수들을 예측하는 데 활용되고 있다(Chong, 2013). 본 연구에서도 퍼지셋 질적비교분석에 사용된 다섯 가지 데이터의 특성 변수들과 두 개의 데이터 환경 변수들이 어떻게 인공지능 학습용 데이터 품질에 영향을 주는지를 추가로 분석하고자 한다(Chong, 2013).

인공신경망 분석을 위해 총 다섯 개의 모형을 수립하였다. 먼저 가장 기본이 되는 (1). 전체표본을 기반한 모형과 (2), (3) 전문성/특수성이 있는 데이터를 다루는 기업들과 다루지 않는 기업들의 표본, (4), (5) 인공지능 윤리와 프라이버시와 관련된 데이터를 다루는 기업들과 그렇지 않은 기업들의 표본을 활용하여 각각 두 개의 모형을 만들었다. 전체 다섯 개의 모형의 예측 정확도를 측정하기 위해 Root Mean Squared Error(RMSE) 값을 분석하였다. RMSE 값 분석 결과 모든 다섯 개의 모형에서 최소 0.198에서 최대 0.469로 기존 연구에서 제안하고 있는 적정 값의 범위 안에 있으므로, 본 연구의 다섯 개의 모형은 예측 정확도를 확보한 것으로 보인다(Chong, 2013; Leong *et al.*, 2019). 전체표본에 대한 인공신경망 모형은 <그림 4>와 같으며, 각 모형의 RMSE 결과는 <표 7>과 같다.



주) N = 표본 개수 (Sample size), SSE = Sum square of error, RMSE = Root mean square of error.

〈그림 4〉 인공신경망 모형(전체표본)

〈표 7〉 5개 모형에 대한 Root Mean Squared Error(RMSE) 결과

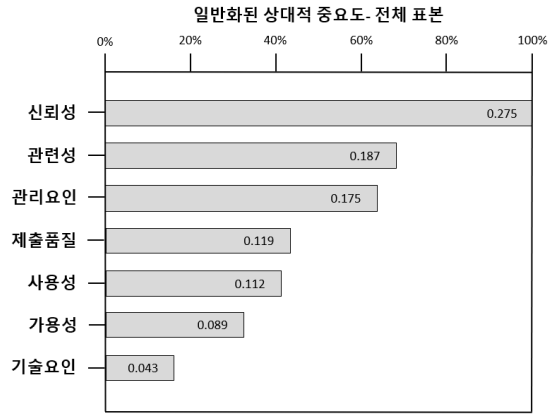
훈련(Training) 전체표본			검증(Testing) 전체표본			Total Samples
N1	SSE	RMSE	N2	SSE	RMSE	N1 + N2
267	31.555	0.344	126	5.253	0.204	393
285	32.416	0.337	108	7.230	0.259	393
268	20.948	0.280	125	8.830	0.266	393
266	30.001	0.336	127	11.046	0.295	393
270	21.983	0.285	123	10.695	0.295	393
282	23.584	0.289	111	16.809	0.389	393
272	23.856	0.296	121	7.938	0.256	393
276	20.190	0.270	117	8.567	0.271	393
275	22.671	0.287	118	8.890	0.274	393
285	33.896	0.345	108	8.240	0.276	393
Mean	26.110	0.307	Mean	9.350	0.279	
SD	5.241	0.030	SD	3.090	0.046	

〈표 7〉 5개 모형에 대한 Root Mean Squared Error(RMSE) 결과(계속)

훈련(Training) 전문성/특수성 관련 없는 표본			검증(Testing) 전문성/특수성 관련 없는 표본			Total Samples
N1	SSE	RMSE	N2	SSE	RMSE	N1 + N2
167	28.502	0.413	70	11.670	0.408	237
158	27.736	0.419	79	16.303	0.454	237
170	29.241	0.415	67	14.323	0.462	237
158	25.819	0.408	79	15.590	0.444	237
161	29.520	0.428	76	11.455	0.388	237
164	27.135	0.407	73	12.207	0.427	237
157	24.015	0.396	80	14.198	0.424	237
164	27.303	0.408	73	12.406	0.415	237
166	28.780	0.416	71	14.718	0.455	237
168	25.501	0.390	69	14.003	0.450	237
Mean	27.355	0.410	Mean	13.687	0.433	
SD	1.787	0.011	SD	1.672	0.024	
훈련(Training) 전문성/특수성 관련 있는 표본			검증(Testing) 전문성/특수성 관련 있는 표본			Total Samples
N1	SSE	RMSE	N2	SSE	RMSE	N1 + N2
106	4.713	0.211	49	3.722	0.276	155
107	6.250	0.242	48	4.545	0.308	155
108	5.106	0.217	47	4.709	0.317	155
103	7.278	0.266	52	9.814	0.434	155
105	4.136	0.198	50	4.566	0.302	155
103	4.374	0.206	52	7.452	0.379	155
110	5.144	0.216	45	2.843	0.251	155
105	9.774	0.305	50	5.129	0.320	155
122	12.426	0.319	33	0.883	0.164	155
101	13.016	0.359	54	5.262	0.312	155
Mean	7.222	0.254	Mean	4.893	0.306	
SD	3.347	0.056	SD	2.424	0.072	
훈련(Training) 윤리/프라이버시 관련 없는 표본			검증(Testing) 윤리/프라이버시 관련 없는 표본			Total Samples
N1	SSE	RMSE	N2	SSE	RMSE	N1 + N2
180	8.245	0.214	80	9.424	0.343	260
192	10.851	0.238	68	2.707	0.200	260
175	7.817	0.211	85	5.219	0.248	260
176	12.328	0.265	84	4.255	0.225	260
180	9.623	0.231	80	4.714	0.243	260
186	13.625	0.271	74	3.596	0.220	260
174	7.972	0.214	86	8.493	0.314	260
180	12.205	0.260	80	4.741	0.243	260
188	18.013	0.310	72	3.736	0.228	260
187	13.840	0.272	73	3.152	0.208	260
Mean	11.452	0.249	Mean	5.004	0.247	
SD	3.234	0.032	SD	2.230	0.046	

〈표 7〉 5개 모형에 대한 Root Mean Squared Error(RMSE) 결과(계속)

훈련(Training) 윤리/프라이버시 관련 있는 표본			검증(Testing) 윤리/프라이버시 관련 있는 표본			Total Samples
N1	SSE	RMSE	N2	SSE	RMSE	N1 + N2
96	17.826	0.431	37	3.973	0.328	133
94	14.780	0.397	39	1.957	0.224	133
89	9.926	0.334	44	12.677	0.437	133
85	6.655	0.280	48	7.309	0.390	133
93	12.480	0.366	40	5.760	0.379	133
86	14.725	0.414	47	4.013	0.292	133
95	16.660	0.419	38	3.174	0.289	133
95	8.210	0.294	38	8.322	0.468	133
79	9.632	0.349	54	11.881	0.469	133
94	10.698	0.337	39	13.076	0.579	133
Mean	12.159	0.362	Mean	7.214	0.386	
SD	3.732	0.052	SD	4.137	0.116	



〈그림 5〉 민감도 분석 결과(전체표본)

다섯 개의 모형에서 어떠한 외생 변수들이 내생 변수인 데이터 품질에 대해 정규화된 상대적 중요성이 있었는지를 확인하기 위해 <표 8> 및 <그림 5>, <그림 6>과 같이 민감도 분석을 실시하였다(Chong, 2013). 전체표본 모형에서는 신뢰성(100%)이 가장 중요한 변수로 나타났으며, 관련성(67.9%)과 관리 요인(63.5%)이 중요한 변수로 나타났다.

두 번째 전문성/특수성 관련 없는 표본 모형은 관련성(100%)이 가장 중요한 변수로 나타났고, 가용성(91.3%), 신뢰성(90.2%), 관리 요인(81.5%), 제출 품질(74.0%), 사용성(68.1%)이 다음으로 중요한 변수들로 나타났다. 특별히 두 번째 모형에서는 기술 요인(46.6%)을 제외하고는 대부분 변수가 60% 이상의 중요도를 나타냈다. 세 번째 모형인 전문성/특수성 관련 표본 모형은 신뢰성(100%)이 가장 중요한 변수로 나타났고, 다음으로는 관리 요인(57.4%)이 중요한 것으로 나타났다. 네 번째 모형인 윤리/프라이버시와 관련 없는 표본 모형은 신뢰성(100%)이 가장 중요한 변수로 나타났고, 관련성(95.9%), 관리 요인(91.7%), 제출 품질(83.7%)이 다음 중요한 변수로 나타났다. 마지막으로 윤리/프라이버시와 관련 있는 표본 모형에서는 사용성(100%)이 가장 중요한 변수로 나타났고, 다음으로는 신뢰성(77.7%), 가용성(72.5%), 관련성(57.5%), 제출 품질(56.1%), 관리 요인(54.5%)이 중요한 것으로 나타났다. 위의 다섯 가지의 모형을 기반으로 보면, 가장 중요한 변수로는 데이터의 정확성, 일관성, 무결성, 완전성을 의미하는 신뢰성이 가장 중요한 것으로 보이며, 전문성/특수성 측면의 데이터 타입을 구축하는 기업에서는 신뢰성과 관리 요인이 데이터 품질에 중요한 것으로 보인다. 또한 윤리/프라이버시와 관련 있는 데이

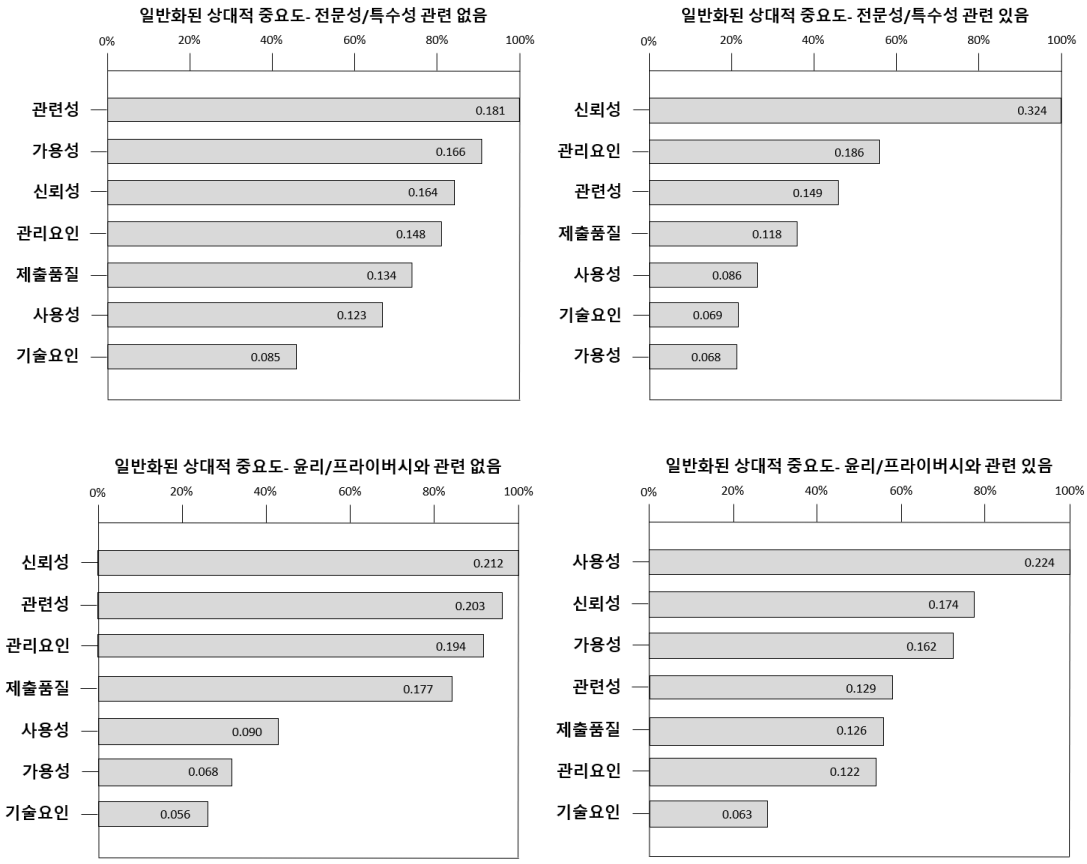
터를 구축하는 기업에서는 데이터의 신임성을 나타내는 사용성과 신뢰성, 그리고 접근성, 시의적절성을 나타내는 가용성을 다음으로 중요하게 보는 것으로 나타났다. 위의 결과는 퍼지셋 질적비교분석에서 나타나지 않은 다섯 가지 데이터 속성과 두 개의 데이터 구축환경 변수들이 데이터 품질을 예측하면서 얼마나 상대적으로 중요한 역할을 하고 있는지에 대한 시사점을 주고 있다.

〈표 8〉 민감도 분석 결과

전체표본(n=393)							
인공신경망	가용성	사용성	신뢰성	관련성	제출품질	기술요인	관리요인
ANN1	0.135	0.155	0.225	0.146	0.127	0.028	0.184
ANN2	0.093	0.036	0.274	0.192	0.223	0.066	0.116
ANN3	0.066	0.134	0.239	0.293	0.050	0.029	0.188
ANN4	0.047	0.164	0.348	0.093	0.069	0.085	0.194
ANN5	0.077	0.082	0.372	0.211	0.141	0.015	0.103
ANN6	0.128	0.121	0.196	0.184	0.090	0.017	0.264
ANN7	0.062	0.110	0.287	0.120	0.225	0.021	0.176
ANN8	0.123	0.069	0.179	0.208	0.142	0.109	0.168
ANN9	0.083	0.072	0.420	0.132	0.073	0.036	0.184
ANN10	0.077	0.179	0.210	0.288	0.048	0.027	0.170
평균 상대적 중요도	0.089	0.112	0.275	0.187	0.119	0.043	0.175
일반화된 상대적 중요도	32.4%	40.8%	100.0%	67.9%	43.2%	15.7%	63.5%
전문성/특수성 관련 없는 표본(n=237)							
인공신경망	가용성	사용성	신뢰성	관련성	제출품질	기술요인	관리요인
ANN1	0.158	0.144	0.164	0.197	0.090	0.053	0.195
ANN2	0.235	0.079	0.109	0.161	0.131	0.094	0.191
ANN3	0.145	0.081	0.176	0.173	0.194	0.060	0.170
ANN4	0.176	0.167	0.139	0.223	0.111	0.098	0.086
ANN5	0.130	0.177	0.187	0.157	0.162	0.062	0.124
ANN6	0.147	0.091	0.186	0.191	0.138	0.088	0.159
ANN7	0.200	0.144	0.207	0.186	0.085	0.105	0.072
ANN8	0.128	0.127	0.182	0.094	0.181	0.092	0.197
ANN9	0.167	0.117	0.129	0.221	0.173	0.083	0.111
ANN10	0.169	0.107	0.156	0.210	0.076	0.110	0.173
평균 상대적 중요도	0.166	0.123	0.164	0.181	0.134	0.085	0.148
일반화된 상대적 중요도	91.3%	68.1%	90.2%	100.0%	74.0%	46.6%	81.5%

〈표 8〉 민감도 분석 결과(계속)

전문성/특수성 관련 있는 표본(n=155)							
인공신경망	가용성	사용성	신뢰성	관련성	제출품질	기술요인	관리요인
ANN1	0.060	0.080	0.187	0.186	0.087	0.135	0.265
ANN2	0.080	0.074	0.375	0.049	0.184	0.067	0.171
ANN3	0.083	0.063	0.275	0.203	0.140	0.068	0.168
ANN4	0.080	0.085	0.294	0.156	0.094	0.130	0.162
ANN5	0.063	0.187	0.212	0.226	0.121	0.091	0.099
ANN6	0.141	0.099	0.286	0.164	0.137	0.038	0.134
ANN7	0.006	0.095	0.367	0.173	0.166	0.013	0.180
ANN8	0.026	0.054	0.471	0.047	0.093	0.046	0.263
ANN9	0.054	0.022	0.297	0.226	0.067	0.046	0.288
ANN10	0.087	0.105	0.478	0.058	0.089	0.054	0.129
평균 상대적 중요도	0.068	0.086	0.324	0.149	0.118	0.069	0.186
일반화된 상대적 중요도	21.0%	26.7%	100.0%	45.9%	36.4%	21.2%	57.4%
윤리/프라이버시와 관련 없는 표본(n=260)							
인공신경망	가용성	사용성	신뢰성	관련성	제출품질	기술요인	관리요인
ANN1	0.063	0.087	0.126	0.258	0.165	0.097	0.204
ANN2	0.039	0.110	0.278	0.122	0.242	0.055	0.155
ANN3	0.054	0.061	0.088	0.229	0.241	0.069	0.259
ANN4	0.064	0.051	0.251	0.272	0.091	0.064	0.208
ANN5	0.049	0.115	0.193	0.216	0.160	0.042	0.224
ANN6	0.078	0.082	0.183	0.224	0.166	0.050	0.217
ANN7	0.078	0.158	0.223	0.220	0.177	0.043	0.102
ANN8	0.076	0.076	0.236	0.106	0.225	0.062	0.220
ANN9	0.121	0.089	0.262	0.161	0.171	0.028	0.168
ANN10	0.059	0.066	0.280	0.226	0.136	0.045	0.187
평균 상대적 중요도	0.068	0.090	0.212	0.203	0.177	0.056	0.194
일반화된 상대적 중요도	32.1%	42.2%	100.0%	95.9%	83.7%	26.2%	91.7%
윤리/프라이버시와 관련 있는 표본(n=133)							
인공신경망	가용성	사용성	신뢰성	관련성	제출품질	기술요인	관리요인
ANN1	0.193	0.282	0.061	0.101	0.145	0.063	0.155
ANN2	0.094	0.304	0.175	0.093	0.099	0.028	0.206
ANN3	0.202	0.252	0.147	0.098	0.026	0.103	0.172
ANN4	0.153	0.224	0.250	0.157	0.110	0.064	0.043
ANN5	0.137	0.224	0.280	0.130	0.071	0.081	0.077
ANN6	0.173	0.194	0.103	0.157	0.272	0.055	0.046
ANN7	0.207	0.098	0.179	0.143	0.124	0.055	0.194
ANN8	0.068	0.260	0.233	0.216	0.104	0.016	0.104
ANN9	0.155	0.246	0.133	0.063	0.209	0.097	0.097
ANN10	0.241	0.160	0.179	0.131	0.097	0.065	0.127
평균 상대적 중요도	0.162	0.224	0.174	0.129	0.126	0.063	0.122
일반화된 상대적 중요도	72.5%	100.0%	77.7%	57.5%	56.1%	28.0%	54.5%



〈그림 6〉 민감도 분석 결과 - 데이터 타입에 따른 분리 분석

Ⅵ. 결론과 시사점

본 연구는 최근 국가적으로 중요한 기술로 주목받고 있는 인공지능기술 중에서도 알고리즘 개발에 가장 중요한 위치를 차지하고 있는 인공지능 학습용 데이터의 높은 품질 조건을 알아보고자 하였다. 이를 위해 본 연구는 기존 연구의 문헌 조사와 전문가 인터뷰를 통해 연구모형을 제안하고, 이를 퍼지셋 질적비교분석을 통해 다양한 조건들과 요인들이 어떠한 조합을 보이는지를 탐색하였다. 추가적으로 인공신경망 분석을 통해 주어진 데이터 타입에 따라 어떠한 요인들이 가장 큰 영향을 주는지를 실증적으로 본 연구이다.

첫 번째, 본 연구의 퍼지셋 질적비교분석 결과를 보면 사용성, 신뢰성, 관련성, 제출 품질이 결과 조건들에 가장 중요한 조건 변수로 나타났는데 이는 인공지능학습용 특성상, 단순한 데이터 분석을 통해 데이터에 담겨 있는 의미를 추출하는 것이 아니라 인공지능 서비스에 사용될 알고리즘을 학습하는 목적의 데이터이기 때문에 기존 데이터 품질과 관련된 대부분의 요인이 높은 데이터 품질을 결정하는 조건들에 중요하게 영향을 주는 것으로 나타났다. 이러한 관점에서 보면 기존 정보시스템이나 빅데이터 관점의 연구에서 강조하고 있는 데이터 품질 관련 변수들이 다른 인공지능 학습데이터의 데이터 타입 조건(특수성/전문성, 윤리/프라이버시)의 상황에서도 동일하게 중요한 역할을 하는 것을 볼 수 있었다. 추가로 기존 데이터 속성

변수에 추가한 구축 관련 환경 변수인 데이터 구축 기술요인과 데이터 구축 관리요인도 주어진 데이터의 타입 조건에 따라 다른 결과들을 보여줌으로써 인공지능 학습데이터 품질 향후 연구에 사용될 수 있는 중요한 변수임을 확인할 수 있었다. 기존의 빅데이터나 정보시스템의 데이터베이스 데이터들은 주로 마이닝을 통한 중요 정보 추출을 목적으로 하므로 데이터를 구축하는 환경과는 관계가 적다. 반면, 인공지능 학습용 데이터는 인공지능 시스템을 학습하는 목적으로 사용되는 데이터이기 때문에, 학습데이터의 구축부터 관리까지 시스템의 환경 측면 또한 반드시 데이터의 품질 측정에 함께 고려되어야 할 것으로 보인다.

마지막으로 추가로 시행한 인공신경망 분석에서는 대부분의 주어진 데이터 타입에 따른 분석에서 신뢰성이 가장 중요한 변수로 나왔는데 이는 학생들의 학습시키는 교과서 내용의 신뢰성이 실제 교육 기관에서 중요한 것과 같은 맥락의 결과라고 볼 수 있겠다. 미래에 지속해서 자율적으로 서비스를 제공하는 인공지능 시스템의 특성상 학습을 위한 데이터들의 가장 기본 되는 것은 바로 신뢰성이고, 그 신뢰성에 기반해서 주어진 데이터 타입에 따라서 관련성 사용성 등과 같은 변수들이 중요하게 데이터 품질에 영향을 주는 것으로 보인다.

본 연구의 학술적 시사점은 새롭게 대두되고 있는 인공지능 학습용 데이터 품질의 중요성과 이러한 품질을 구성하는 중요한 요인들이 어떠한 조합을 보이는 지를 퍼지셋 질적비교분석 방법론을 통해 알아보고, 인공지능학습용 데이터 품질에 대한 연구모형을 제안하였다는 데 있다. 기존의 데이터 품질 연구에서는 대부분 데이터의 속성에 집중하고 있어 데이터를 구축하는 환경이나 데이터를 구축하는 관리 환경 등에 대해서는 고려되지 않고 있었다. 대부분의 빅데이터나 데이터 마이닝을 위한 데이터들은 이미 생성된 데이터들을 기반으로 하므로 이러한 구축환경에 대한 충분히 고려되지 못한 면이 있다. 그러나 인공지능 학습용 데이터의 경우에는 기업이나 정부 기관들이 인공지능 시스템과 알고리즘의 품질을 보장하기 위해 계획적으로 구축하고 있으므로 이러한 측면에서 데이터를 구축하는 시스템과 관리 시스템의 측면도 고려되어야 한다. 또한 학습용 데이터에는 전문성이 필요하거나, 특수성이 있는 데이터와 윤리적, 프라이버시와 관련된 데이터가 있으므로 이러한 부분에 대한 고려도 필요하다. 이러한 관점에서 본 연구에서는 이러한 다양한 측면의 품질 선행요인들을 고려하여 연구모형을 제안하고 있다는데 그 첫 번째 가장 중요한 공헌이 있다고 하겠다. 두 번째, 본 연구는 퍼지셋 질적비교분석 방법론을 사용하여 데이터를 분석함으로써 다양한 요인들의 의미 있는 조합을 찾아내고, 이를 통해 연구모형을 검증하였다. 하지만 특정 요인들의 중요순위를 알아볼 수 없다는 방법론의 한계점을 극복하기 위해 최근 정보시스템 학계에서 활발하게 도입되고 있는 인공신경망 분석을 도입하여, 퍼지셋 질적비교분석 방법론과 혼용하여 사용함으로써, 더욱 다양한 결과를 도출하였다는데 두 번째 공헌이 있다.

본 연구의 실무적 시사점은 첫째, 인공지능 학습용 데이터 구축을 담당하는 기관과 정책 입안자들은 높은 데이터 품질을 제고하기 위해서는 단순히 데이터의 속성적인 측면으로 데이터 품질을 검증하고 판단하는 것이 아니라, 총합적인 관점에서 데이터를 구축하는 기업들이 사용하고 있는 데이터 구축 기술이나 데이터 구축 관리 시스템에 대해서도 구체적으로 검증할 필요가 있다는 것이다. 인공지능 학습용 데이터 구축은 단순히 몇 년만 하고 끝나는 사업이 아닌 장기적 프로젝트이기 때문에 국가의 미래 인공지능 기술의 경쟁력 확보를 위해 총합적인 접근방법과 관리가 필요하겠다. 둘째는 본 연구의 결과로도 확인된 바와 같이, 전문성/특수성, 윤리/프라이버시 등과 같이 데이터의 형태에 따라 나타나는 중요한 특성들을 분리하여 품질 평가를 해야 할 것으로 보인다. 일률적인 품질 기준으로는 이러한 데이터의 특수성이나 전문성이 요구되는 데이터들과 윤리 및 프라이버시 이

슈가 있는 데이터들의 품질을 평가하는 어려움이 있을 것이며, 형평성 또한 맞지 않는다. 이러한 관점에서 데이터 품질 평가 방법과 요인들의 구성에 다양성이 필요할 것으로 보인다. 데이터들을 구축하는 기업들도, 이러한 다양한 데이터 특성에 맞는 구축 기술과 관리기법들을 개발하여, 미래에 발생할 수 있는 다양한 데이터 품질 이슈들에 대해서 대비해야 할 것이다. 또한 데이터 속성에서 가장 중요한 요소들로 신뢰성과 관련성 그리고 사용성 등이 나타났음으로 기업은 데이터를 구축할 때 데이터의 신뢰성과 관련성 및 사용성을 어떻게 제고할지에 대한 부분도 심각하게 고려해야 할 것으로 보인다. 마지막으로는 본 연구에서 제안된 학습용 데이터 품질 모형을 기반으로 국내의 기업, 대학, 연구소, 공공기관, 협회, 지자체 등은 물론 민간 기업들까지도 인공지능 학습용 데이터 구축을 위한 역량을 제고하고, 지속적인 평가를 통해 더 많은 수의 질 좋은 인공지능 학습용 데이터를 구축할 수 있어야 할 것이다. 정부의 구축 사업에서도 이러한 품질 기준을 활용하여 선정 기준으로 사용한다면 적합한 기준을 충족하는 기업과 기관들이 참여함으로써 인공지능 산업의 긍정적인 선순환 구조를 창출할 수 있을 것으로 보인다.

모든 연구가 그렇듯이 본 연구에도 한계점을 가지고 있다. 첫 번째, 본 연구의 데이터 수집은 인공지능학습용 데이터를 가공하는 기업들의 실무자와 임원들을 대상으로 진행되어 실제 이를 활용하는 다양한 기업과 산업들의 관점을 포함하고 있지 않다. 이러한 한계점에 대하여, 향후 연구에서는 본 연구에서 제안하는 모형에 기반을 두어 인공지능학습용 데이터의 구체적인 특성들을 반영하여 데이터를 활용하는 다양한 산업과 다양한 데이터 형태를 만들고, 사용하는 기업에서 설문조사를 실시하여 이들 간의 비교 연구를 진행하여, 인공지능 학습용 데이터 품질에 대한 이해의 폭을 넓혀야 할 것이다. 구체적으로는 민감한 데이터를 다루거나, 편향성 및 프라이버시 이슈가 있는 산업들과 그렇지 않은 산업들을 비교해서 실제로 이러한 데이터의 특성이 학습용 데이터의 품질에는 어떠한 영향을 주고, 산업들 간에는 어떠한 차이가 있는지를 보는 것도 매우 중요할 것으로 보인다. 두 번째는 본 연구에서 제안하고 있는 데이터 타입에 대한 관점 외에도 다양한 조절 효과를 줄 수 있는 변수들이 존재한다. 예를 들어, 데이터 구축기업의 인공지능 데이터 및 서비스 구축 경험 연수나 실제 어떠한 기술들을 활용하고 있고, 이러한 기술들이 어떠한 우위를 가졌는지에 대한 부분에 대한 고려가 부족했다. 향후 연구에서는 이러한 다양한 조절 효과를 줄 수 있는 요인들을 발굴하고, 이를 이용하여 모형을 개발, 가설을 설정하여, 구조방정식이나 계층적 회귀분석을 통해서 실증적 연구를 수행해 보고자 한다. 이러한 한계점까지 고려하여 향후 연구가 진행된다면 인공지능학습용 데이터 품질과 관련된 심화한 학술적 공헌과 실무적 공헌이 증가할 것으로 보인다.

참 고 문 헌

[1] 과학기술정보통신부, “인공지능 학습용 데이터 개방, 2배(191→381종)로 늘어난다”, 2021 Available at https://www.msit.go.kr/bbs/view.do?sCode=user&mId=113&mPid=238&pageIndex=1&bbsSeqNo=94&nttSeqNo=3181904&searchOpt=ALL&searchTxt=%ED%95%99%EC%8A%B5%EC%9A%A9.

[2] 과학기술정보통신부, “인공지능 학습용 데이터, 역대 최대 규모로 개방한다”, 2023 Available at https://www.msit.go.kr/bbs/view.do?sCode=user&mId=113&mPid=238&pageIndex=1&bbsSeqNo=94&nttSeqNo=3183010&searchOpt=ALL&searchTxt=%ED%95%99%EC%8A%B5%EC%9A%A9.

[3] 김형섭, “데이터 품질관리 평가모델에 관한 연구”, *한국융합학회논문지*, 제11권, 제7호, 2020, pp. 217-222.

[4] 이용희, “빅데이터 품질향상 방안에 관한 고

찰", *한국IT정책경영학회논문지*, 제10권, 제5호, 2018, pp. 1007-1013.

[5] 이원국, 양희태, "퍼스널 모빌리티 사용의도에 관한 연구: SOR(Stimulus-Organism-Response) 모델을 중심으로", *경영정보학연구*, 제24권, 제2호, 2022, pp. 67-88.

[6] 이현애, 정희정, 함주연, 정남호, "퍼지셋 질적 비교 분석(fsQCA)을 활용한 관광지 거주민들의 삶의 질 저하에 영향을 미치는 요인 연구", *경영정보학연구*, 제21권, 제1호, 2019, pp. 113-133.

[7] 장경애, 김우제, 김자희, "고객의 요구사항에 기반한 데이터 품질 평가 속성 및 우선순위 도출", *정보처리학회논문지. 소프트웨어 및 데이터 공학*, 제4권, 제12호, 2015, pp. 549-560.

[8] 정원섭, "인공지능 알고리즘의 편향성과 공정성", *인간・환경・미래*, 제25권, 2020, pp. 55-73.

[9] 정원진, 박영태, "Data warehousing, contextual data quality, and problem-solving performance", *정보시스템연구*, 제14권, 제2호, 2005, pp. 237-256.

[10] 정혜정, "데이터 품질 평가에 관한 연구", *인터넷정보학회논문지*, 제8권, 제4호, 2007, pp. 119-128.

[11] 최유진, 양희태, "위드 코로나 시대의 원격근무 솔루션 지속 사용 의도에 관한 연구: TOE (Technology-Organization-Environment) 모델을 중심으로", *경영정보학연구*, 제25권, 제2호, 2023, pp. 163-180.

[12] 한국지능정보사회진흥원, "인공지능 학습용 데이터 품질관리 가이드라인", v1.0, 2021.

[13] 한국지능정보사회진흥원, "인공지능 학습용 데이터 품질관리 가이드라인", v3.0, 2023.

[14] Ardagna, D., C. Cappiello, W. Samá, and M. Vitali, "Context-aware data quality assessment for big data", *Future Generation Computer Systems*, Vol.89, 2018, pp. 548-562.

[15] Awa, H. O. and O. U. Ojiabo, "A model of adoption determinants of ERP within TOE framework", *Information Technology & People*, Vol. 29, No.4, 2016, pp. 901-930.

[16] Bagozzi, R. P. and Y. Yi, "On the evaluation of structural equation models", *Journal of the Academy of Marketing Science*, Vol.16, 1988, pp. 74-94.

[17] Batini, C. and M. Scannapieco, *Data and Information Quality*, Cham, Switzerland: Springer International Publishing. 2016.

[18] Bertossi, L. and F. Geerts, "Data quality and explainable AI", *Journal of Data and Information Quality (JDIQ)*, Vol.12, No.2, 2020, pp. 1-9.

[19] Bryman, A., *Social Research Methods*, Oxford university press, 2016.

[20] Cai, L. and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era", *Data Science Journal*, Vol.14, No.2, 2015, pp. 1-10.

[21] Chan, F. T. and A. Y. Chong, "A SEM-neural network approach for understanding determinants of interorganizational system standard adoption and performances", *Decision Support Systems*, Vol.54, No.1, 2012, pp. 621-630.

[22] Chong, A. Y. L., B. Li, E. W. Ngai, E. Ch'Ng, and F. Lee, "Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach", *International Journal of Operations & Production Management*, Vol.36, No.4, 2016, pp. 358-383.

[23] Chong, A. Y.-L., "Predicting m-commerce adoption determinants: A neural network approach", *Expert Systems with Applications*, Vol.40, No.2, 2013, pp. 523-530.

[24] Christine Patterso, *The Six Primary Dimensions For Data Quality Assessment*, DAMA UK

Working, 2017.

[25] David Loshin, *Dimensions of Data Quality*, In MK Series on Business Intelligence. The Practitioner's Guide to Data Quality Improvement, Morgan Kaufmann, 2011.

[26] David Plotkin, *Data Stewardship*, An Actionable Guide to Effective Data Management and Data Governance, Morgan Kaufmann, 2014.

[27] DeLone, W. H. and E. R. McLean, "Measuring e-commerce success: Applying the DeLone & McLean information systems success model", *International Journal of Electronic Commerce,* Vol.9, No.1, 2004, pp. 31-47.

[28] English, L. P., *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*, John Wiley & Sons, Inc, 1999.

[29] Ercole, A., V. Brinck, P. George, R. Hicks, J. Huijben, M. Jarrett, and L. Wilson, "Guidelines for data acquisition, quality and curation for observational research designs (DAQCORD)", *Journal of Clinical and Translational Science*, Vol.4, No.4, 2020, pp. 354-359.

[30] Fiss, P. C., "A set-theoretic approach to organizational configurations", *Academy of Management Review*, Vol.32, No.4, 2007, pp. 1180-1198.

[31] Fiss, P. C., "Building better causal theories: A fuzzy set approach to typologies in organization research", *Academy of Management Journal*, Vol.54, No.2, 2011, pp. 393-420.

[32] Fornell, C. and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error", *Journal of Marketing Research*, Vol.18, No.1, 1981, pp. 39-50.

[33] Gefen, D. and D. Straub, "A practical guide to factorial validity using PLS-Graph: Tutorial and annotated example", *Communications of the Association for Information Systems*, Vol.16, No.1, 2005, pp. 5.

[34] Granick, L., "Assuring the quality of information dissemination: Responsibilities of database producers", *Information Services & Use*, Vol.11, No.3, 1991, pp. 117-136.

[35] Haug, A., J. Stentoft Arlbjørn, and A. Pedersen, "A classification model of ERP system data quality", *Industrial Management & Data Systems*, Vol.109, No.8, 2009, pp. 1053-1068.

[36] Hew, J.-J., L.-Y. Leong, G. W.-H. Tan, K.-B. Ooi, and V.-H. Lee, "The age of mobile social commerce: An Artificial Neural Network analysis on its resistances", *Technological Forecasting and Social Change*, Vol.144, 2019, pp. 311-324.

[37] Hoxmeier, J. A., "Typology of database quality factors", *Software Quality Journal*, Vol.7, 1998, pp. 179-193.

[38] Hew, J., V. Lee, and L. Leong, "Why do mobile consumers resist mobile commerce applications? A hybrid fsQCA-ANN analysis", *Journal of Retailing and Consumer Services*, Vol.75, 2023, 103526, ISSN 0969-6989, https://doi.org/10.1016/j.jretconser.2023.103526.

[39] Kim, H.-S., "A study on the data quality management evaluation model", *Journal of the Korea Convergence Society*, Vol.11, No.7, 2020, pp. 217-222.

[40] Lee, Y., O. J. Kwon, H. Lee, J. Kim, K. Lee, and K.-E. Kim, "Augment & valuate: A data enhancement pipeline for data-centric AI", 2021, arXiv preprint arXiv:2112.03837.

[41] Leong, L.-Y., T.-S. Hew, K.-B. Ooi, V.-H. Lee, and J.-J. Hew, "A hybrid SEM-neural network analysis of social media addiction", *Expert Systems with Applications*, Vol.133, 2019, pp. 296-316.

[42] Li, F., E. C.-X. Aw, G. W.-H. Tan, T.-H. Cham, and K.-B. Ooi, "The Eureka moment in under-

standing luxury brand purchases! A non-linear fsQCA-ANN approach", *Journal of Retailing and Consumer Services*, Vol.68, 2022, 103039.

[43] Li, P., X. Rao, J. Blase, Y. Zhang, X. Chu, and C. Zhang, "CleanML: A study for evaluating the impact of data cleaning on ml classification tasks", *2021 IEEE 37th International Conference on Data Engineering* (ICDE), 2021.

[44] Madnick, S. E., R. Y. Wang, Y. W. Lee, and H. Zhu, "Overview and framework for data and information quality research", *Journal of Data and Information Quality (JDIQ)*, Vol.1, No.1, 2009, pp. 1-22.

[45] Miller, H., "The multiple dimensions of information quality", *Information Systems Management*, Vol.13, No.2, 1996, pp. 79-82.

[46] Mohammadi, H., "Investigating users' perspectives on e-learning: An integration of TAM and IS success model", *Computers in Human Behavior*, Vol.45, 2015, pp.359-374.

[47] Ng, A., "A Chat with Andrew on MLOps: From Model-centric to Data-centric AI", DeepLearningAI. 2021. Available at https://www.youtube.com/watch?v=06-AZXmwHjo.

[48] Nicolaou, A. I., M. Ibrahim, and E. Van Heck, "Information quality, trust, and risk perceptions in electronic data exchanges", *Decision Support Systems*, Vol.54, No.2, 2013, pp. 986-996.

[49] Park, G.-E. and C.-J. Kim, "Quality characteristics of public open data", *Journal of Digital Convergence*, Vol.13, No.10, 2015, pp. 135-146.

[50] Pipino, L. L., Y. W. Lee, and R. Y. Wang, "Data quality assessment", *Communications of the ACM*, Vol.45, No.4, 2002, pp. 211-218.

[51] Podsakoff, P. M., S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: A critical review of the literature and recommended remedies", *Journal of Applied Psychology*, Vol.88, No.5, 2003, pp. 879.

[52] Pudjianto, B., H. Zo, A. P. Ciganek, and J. J. Rho, "Determinants of e-government assimilation in Indonesia: An empirical investigation using a TOE framework", *Asia Pacific Journal of Information Systems*, Vol.21, No.1, 2011, pp. 49-80.

[53] Purwanto, S., "The effect of organizational governance on the performance and commitment of the lecturers", *Public Policy and Administration Research*, Vol.5, No.1, 2015, pp. 35-42.

[54] Ragin, C. C., *Redesigning social inquiry: Fuzzy Sets and Beyond*, University of Chicago Press. 2009.

[55] Ragin, C. C., "Set relations in social research: Evaluating their consistency and coverage", *Political Analysis*, Vol.14, No.3, 2006, pp. 291-310.

[56] Ragin, C. C., K. A. Drass, and S. Davey, "Fuzzy-set/qualitative comparative analysis 2.0", *Tucson, Arizona: Department of Sociology, University of Arizona*, Vol.23, No.6, 2006, pp. 1949-1955.

[57] Ragin, C. and S. Davey, *fs/QCA [Computer Programme], version 2.5. Irvine*, CA: University of California, 2014.

[58] Rana, N. P., Y. K. Dwivedi, M. D. Williams, and V. Weerakkody, "Investigating success of an e-government initiative: Validation of an integrated IS success model", *Information Systems Frontiers*, Vol.17, 2015, pp. 127-142.

[59] Roh, Y., G. Heo, and S. E. Whang, "A survey on data collection for machine learning: A big data-ai integration perspective", *IEEE Transactions on Knowledge and Data Engineering*, Vol.33, No.4, 2019, pp. 1328-1347.

[60] Scannapieco, M., A. Virgillito, C. Marchetti, M.

Mecella, and R. Baldoni, “The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems”, *Information Systems*, Vol.29, No.7, 2004, pp. 551-582.

[61] Schneider, C. Q. and C. Wagemann, *Set-theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*, Cambridge University Press, 2012.

[62] Sharma, M., S. Joshi, and S. Luthra, et al., “Impact of Digital Assistant Attributes on Millennials’ Purchasing Intentions: A Multi-Group Analysis using PLS-SEM, Artificial Neural Network and fsQCA”, *Information Systems Frontiers,* 2022, https://doi.org/10.1007/s10796-022-10339-5.

[63] Wand, Y. and R. Y. Wang, “Anchoring data quality dimensions in ontological foundations”, *Communications of the ACM*, Vol.39, No.11, 1996, pp. 86-95.

[64] Wang, R. Y. and D. M. Strong, “Beyond accuracy: What data quality means to data consumers”, *Journal of Management Information Systems,* Vol.12, No.4, 1996, pp. 5-33.

[65] Wang, R. Y., V. C. Storey, and C. P. Firth, “A framework for analysis of data quality research”, *IEEE Transactions on Knowledge and Data Engineering*, Vol.7, No.4, 1995, pp. 623-640.

[66] Xu, H., J. Horn Nord, N. Brown, and G. Daryl Nord, “Data quality issues in implementing an ERP”, *Industrial Management & Data Systems*, Vol.102, No.1, 2002, pp. 47-58.

〈Appendix 1〉 측정 변수 및 항목

변수	설문 항목	참고문헌
접근성 (Accessibility)	· 우리 회사는 AI 학습용 데이터를 구축하기 위한 적절한 접근 경로와 방법을 확보하고 있다. · 우리 회사에서는 AI 학습용 데이터 구축을 위한 다양한 데이터 소스를 회사 구성원이라면 누구나 접근/이용하도록 하고 있다. · 우리 회사에서 구축하고 있는 AI 학습용 데이터는 사용자에게 차별과 제약 없이 접근을 보장하고 있다.	Cai and Zhu (2015)
시의 적절성 (Timeliness)	· 우리 회사에서 구축하고 있는 AI 학습용 데이터는 계획된 기한에 적절하게 수집된다. · 우리 회사에서 구축하고 있는 AI 학습용 데이터는 정기적으로 업데이트된다. · 우리 회사에서 AI 학습용 데이터를 구축하기 위한 애플리케이션의 인터페이스와 데이터의 소스는 정기적으로 업데이트된다. · 우리 회사가 구축하여 제공하는 AI 학습용 데이터는 지속적인 업데이트(최신화)가 가능하다.	Cai and Zhu (2015)
신임성 (Credibility)	· 우리 회사에서 구축하고 있는 AI 학습용 데이터들은 검증된 출처에서 도입된다. · 우리 회사에서 구축하고 있는 AI 학습용 데이터는 전문가를 통해, 데이터의 정확성 등 상시적인 품질점검을 시행하고 있다. · 우리 회사에서 구축하고 있는 AI 학습용 데이터는 관련 기관의 품질 기준을 충족한 데이터를 활용하고 있다.	Cai and Zhu (2015)
정확성 (Accuracy)	· 우리 회사에서 구축하고 있는 AI 학습용 데이터는 정확성을 확보하고 있다. · 우리 회사에서 구축하고 있는 AI 학습용 데이터는 정보 출처의 실제 상태와 본연의 값을 적절하게 반영하고 있다. · 우리 회사에서 구축하고 있는 AI 학습용 데이터는 애매하거나 모호하지 않고 분명하고 명확한 데이터 구조로 표현되어 정의된다.	Cai and Zhu (2015)
일관성 (Consistency)	· 우리 회사에서 구축하고 있는 AI 학습용 데이터는 데이터 가공 후에도 원본 소스의 개념과 핵심적인 가치는 변하지 않는다. · 우리 회사에서 구축하고 있는 AI 학습용 데이터는 데이터 가공 후에도 원본 소스를 확인하거나 검증할 수 있는 추적성을 확보하고 있다. · 우리 회사에서 구축하고 있는 AI 학습용 데이터는 다른 출처의 학습용 데이터와 비교/검증할 수 있도록 일관성을 유지하고 있다.	Cai and Zhu (2015)
무결성 (Integrity)	· 우리 회사에서 수집하고 있는 AI 학습용 데이터는 형식이 명확하고 초상권, 저작권 등 사용에 문제가 없는 데이터로, 관련 기준을 충족한다. · 우리 회사에서 수집하고 있는 AI 학습용 데이터는 구조적으로 결함이 없다. · 우리 회사에서 수집하고 있는 AI 학습용 데이터는 내용상으로 결함이 없다.	Cai and Zhu (2015)
완전성 (Completeness)	· 우리 회사에서 수집하고 있는 AI 학습용 데이터는 일부 구성요소의 결함이 있어도 다른 구성요소의 데이터에 영향을 미치지 않는다. · 우리 회사에서 수집하고 있는 AI 학습용 데이터는 일부 구성요소의 결함이 있어도 다른 데이터의 정확성 및 완전성을 해치지 않는다.	Cai and Zhu (2015)
적합성 (Fitness)	· 우리 회사에서 구축하고 있는 AI 학습용 데이터는 사전에 정의된 학습 목적에 충분히 부합한다. · 우리 회사에서 구축하고 있는 AI 학습용 데이터는 사전에 정의된 주제 영역과 일치한다. · 우리 회사에서 구축하고 있는 AI 학습용 데이터는 사용자/플랫폼에 제공되는 인공지능 학습의 주제와 적절하게 조화를 이룬다.	Cai and Zhu (2015)

변수	설문 항목	참고문헌
가독성 (Readability)	· 우리 회사에서 구축(수집, 정제, 가공)하고 있는 AI 학습용 데이터는 사용자가 쉽게 이해할 수 있는 데이터 구조와 설명이 포함된 메타데이터를 제공하고 있다. · 우리 회사에서 구축(수집, 정제, 가공)하고 있는 AI 학습용 데이터는 인공지능 학습을 위한 목적과 요구사항을 충족하는지 쉽게 판단할 수 있다. · 우리 회사에서 구축(수집, 정제, 가공)하고 있는 AI 학습용 데이터는 인공지능 학습에 적합한 형식과 내용으로 구성되어 있다.	Cai and Zhu (2015)
학습가능성 (Trainability)	· 우리 회사에서 구축(수집, 정제, 가공)하고 있는 AI 학습용 데이터는 인공지능 학습에 적합한 형식과 내용으로 구성되어 있다. · 우리 회사에서 구축(수집, 정제, 가공)하고 있는 AI 학습용 데이터는 인공지능 학습 목적에 부합하는 특성을 확보하고 있다. · 우리 회사에서 구축(수집, 정제, 가공)하고 있는 AI 학습용 데이터는 인공지능 학습에 쉽게 적용 가능하고, 목표하는 특성을 잘 확보하고 있다.	Cai and Zhu (2015)
데이터 구축 플랫폼 (Data Building Platform)	· 우리 회사는 AI 학습용 데이터의 가공을 위한 자체적인 인프라를 확보하고 있다. · 우리 회사는 AI 학습용 데이터의 가공을 위한 자체적인 인프라 운영을 위한 전담 조직을 확보하여 운영하고 있다. · 우리 회사에서는 AI 학습용 데이터의 가공을 위한 자체적인 인프라의 관리 감독 체계를 구축하고 있다.	Awa and Ojiabo (2016)
데이터 구축 도구 (Data Building Tools)	· 우리 회사는 AI 학습용 데이터의 가공을 위한 자체적인 도구(소프트웨어 도구)를 확보하고 있다. · 우리 회사에서는 데이터 가공 도구를 관리 담당하는 직원을 고용/보유/교육하고 있다. · 우리 회사는 AI 학습용 데이터의 가공을 위한 자체적인 도구(소프트웨어 툴) 운영을 위한 전담 조직을 확보하여 운영하고 있다. · 우리 회사는 AI 학습용 데이터의 가공을 위한 자체적인(소프트웨어 툴) 운영을 위한 관리 감독 체계를 구축하고 있다.	Awa and Ojiabo (2016)
데이터 구축 기술 (Data Building Technology)	· 우리 회사는 AI 학습용 데이터의 가공을 위한 자체적인 기술을 확보하고 있다. · 우리 회사는 AI 학습용 데이터의 가공을 위한 자체적인 기술의 개발 및 유지 등을 위한 전담 조직을 확보하여 운영하고 있다. · 우리 회사는 AI 학습용 데이터의 가공을 위한 자체적인 기술의 관리 감독 체계를 구축하고 있다.	Awa and Ojiabo (2016)
데이터 구축 프로세스 (Data Building Process)	· 우리 회사의 AI 학습용 데이터 구축(수집, 정제, 가공)에 필요한 프로세스는 명확히 정의되어 있다. · 우리 회사는 AI 학습용 데이터 구축(수집, 정제, 가공)에 필요한 조직을 구성하고, 프로세스별 역할과 책임을 명확하게 정의하고 있다. · 우리 회사의 AI 학습용 데이터 구축(수집, 정제, 가공) 프로세스는 편향되지 않은 데이터 구축 프로세스를 확보하고 있다. · 우리 회사의 AI 학습용 데이터 구축(수집, 정제, 가공) 프로세스는 개인정보보호, 저작권 등 관련 법규 및 사회적 책무를 준수하고 있다.	Purwanto (2015)
데이터 구축 가이드라인 (Data Building Guideline)	· 우리 회사는 AI 학습용 데이터 구축(수집, 정제, 가공) 시 발주처가 요구하는 기준이나 가이드라인 등을 준수하고 있다. · 발주처가 요구하는 기준이나 가이드라인을 참조하여 우리 회사는 AI 학습용 데이터 구축(수집, 정제, 가공)에 부합하는 자체적인 품질관리 체계(조직, 절차, 품질관리 기준 등)를 확보하고 있다.	Purwanto (2015)

변수	설문 항목	참고문헌
데이터 구축 모니터링 (Data Building Monitoring)	· 우리 회사는 AI 학습용 데이터 구축(수집, 정제, 가공) 절차에 정의된 품질관리 활동의 준수 및 이행 상태를 상시로 점검하고 있다. · 우리 회사는 AI 학습용 데이터 구축(수집, 정제, 가공) 단계별 품질진단 점검 관리 활동을 정기적으로 수행하고 있다. · 우리 회사는 AI 학습용 데이터 구축(수집, 정제, 가공) 단계별 데이터의 품질진단 및 개선 활동을 수행하는 모니터링 프로세스를 운영하고 있다. · 우리 회사는 AI 학습용 데이터 구축(수집, 정제, 가공)에 참여하는 크라우드워커들의 작업 결과물을 점검하고 개선하기 위한 관리체계를 확보하고 있다.	Purwanto (2015)
인공지능 학습용 데이터 품질 (AI Training Data Quality)	· 우리 회사에서 수집하고 있는 AI 학습용 데이터는 인공지능 학습 목적에 부합되는 최신상태를 유지하고 있다. · 우리 회사에서 수집하고 있는 AI 학습용 데이터는 인공지능 학습을 위해 무결성을 유지하고 있다. · 우리 회사에서 수집하고 있는 AI 학습용 데이터는 인공지능 학습을 위해 편향적이지 않도록 보편성을 유지하고 있다. · 우리 회사가 구축(수집, 정제, 가공)한 AI 학습용 데이터는 인공지능 학습에 따른 성능 확보에 충분한 정확성을 확보하고 있다. · 우리 회사가 구축(수집, 정제, 가공)한 AI 학습용 데이터는 인공지능 학습에 충분한 데이터를 확보(제공)하고 있다. · 우리 회사가 구축(수집, 정제, 가공)한 AI 학습용 데이터는 사용자의 요구를 만족하는 최신의 상태를 유지하고 있다. · 우리 회사가 구축(수집, 정제, 가공)한 AI 학습용 데이터는 사용자의 요구사항을 충분히 반영하고 있다. · 우리 회사가 구축(수집, 정제, 가공)한 AI 학습용 데이터는 사용자의 요구를 만족하는 필수 데이터를 누락 없이 확보하고 있다.	Nicolaou *et al.* (2013)

〈Appendix 2〉 데이터/정보 품질에 대한 문헌 조사

저자 정보 (연도)	연구 맥락	데이터 품질 항목/차원 Or 독립변수/Mediator	데이터 품질 세부 항목/차원 Or 종속변수	Methods & Key Findings
Cai and Zhu (2015)	빅데이터 분석 및 활용을 위한 데이터 품질 기준	· Availability · Usability · Reliability · Relevance · Presentation Quality	· Accessibility · Timeliness · Authorization · Definition/ Documentation · Credibility · Meta Data · Accuracy · Integrity · Consistency · Completeness · Auditability · Fitness · Readability · Structure	· The hierarchical big data quality assessment framework 개발
Nicolaou *et al.* (2013)	정보 품질이 사용자의 사용 의도에 미치는	· Information Quality · Competence Trust	· Expected Transaction Performance	· 설문조사 데이터를 통한 실증적 분석

저자 정보 (연도)	연구 맥락	데이터 품질 항목/차원 Or 독립변수/Mediator	데이터 품질 세부 항목/차원 Or 종속변수	Methods & Key Findings
	영향 조사	· Perceived Performance Risk · Perceived Exchange Risk · Goodwill Trust	· Intent for continued Use	
Xu *et al.*(2002)	회사 내 전사적자원관리 시스템 도입과 구현 (SAP) 맥락에서 데이터 품질(DQ)이해	· Training, management support, organizational structure, change management, employee relations	· Training · Top management support · Communications · Manage change · Employee relations DQ Controls	· Case Study (2개 회사 조직 비교) · 데이터 품질 문제 프레임워크를 개발 및 적용을 통해 데이터 품질의 중요성 강조
정혜정(2007)	데이터 품질 평가 모델에 대한 연구	· 기능성 · 사용성 · 효율성 · 신뢰성 · 이식성 · 준수성	· 가변성(Exchangeability) · 정확성(Precision) · 접근제어성(Traceability) · 보안성(Security) · 표현일관성(Consistency) · 접근성(Accessibility) · 조작용이성(Ease of Handling) · 해석가능성 (Possibility of explain) · 현시성(Currentness) · 완전성(Completeness) · 이해성(Understandability) · 입출력이해가능성 (Understandable Input and Output) · 유용성(Availability) · 효율성(Efficiency) · 복구성(Recoverability) · 신뢰성(Reliability) · 변화성(Changeability) · 이식성(Portability) · 준수성(Regulatory Compliance)	· 데이터 품질 평가 항목을 제시 · 제시된 평가 항목의 평가 방안에 대한 정량적 평가 방법을 제시함
김형섭(2020)	데이터 품질 관리 평가 모델에 관한 연구	· 정확성(Accuracy) · 일관성 (Consistency) · 보안성(Security) · 완전성 (Completeness) · 준비성(Readiness)		· 우선순위 결정 방법(AHP기법) · 데이터 품질 관리에 영향을 미치는 요인 (전문가 관점) · 품질관리 요인의 중요도는 정확성-일관성-보안성-완전성-준비성의 순서로 분석됨
Ardagna *et al.* (2018)	스마트 시티	· Confidence/Cost/ Time(CCT) model	· time minimization, confidence	· 케이스 스터디로 수집된 데이터를

저자 정보 (연도)	연구 맥락	데이터 품질 항목/차원 Or 독립변수/Mediator	데이터 품질 세부 항목/차원 Or 종속변수	Methods & Key Findings
			maximization, and budget minimization	활용한 시나리오 기반 Experiments
정원진, 박영태 (2005)	데이터의 질이 의사결정 성과에 미치는 영향	· Contextual Data Quality · Information Quality · System Quality · Service Quality	· Intention to use · User Satisfaction · Net Benefit	· contextual data(상황적 데이터)의 질과 업무의 복잡성이 의사결정 성과에 영향을 미침 · contextual data의 Quality 향상 필요성 제시
장경애 등(2015)	고객데이터요구 사항에 맞춘 데이터 품질 평가	· 일관된 체계 · 정확한 데이터 · 효율적 환경 · 유연한 관리 · 지속적 개선	· 통제성 · 준거성 · 요구완전성 · 정확성 · 추적가능성	· RGT(Repertory Grid Technique) · AHP(Analytic Hierarchy Process)
이용희(2018)	빅데이터품질을 확보를 위한 데이터 품질 표준, 품질평가준거 및 모델 비교 분석	· 사업적 환경 · 개인정보와 보안 · 시간관련 요인	· 복잡성 · 접근성 · 명확성 · 연관성 · 정확성 · 타당성 · 응집 · 연결/지속성	· 빅데이터 품질을 향상시키기 위한 방안으로서 AI 알고리즘 모형 개선 및 실증적 테스트 모델 제시
Park and Kim (2015)	공공데이터 품질 관리와 표준화	· 공공성 Publicity · 활용성 Usability · 신뢰성 Reliability · 적합성 Suitability	· 접근성 Accessibility · 객관성 Objectivity · 공익성 Publicness · 요구반영성 Responsiveness · 대응성 Reactivity · 이해성 Understandability · 가공성 Processability · 연계성 Linkage · 이용편의성 Convenience · 표준준수성 Compliance · 보안성 Security · 보호성 Protection · 공유성 Share · 완전성 Completeness · 정확성 Accuracy · 유효성 Effectiveness · 적시성 Timeliness · 일관성 Consistency	· 전문가 설문/요인분석 · 공공개방데이터의 품질 향상과 활용 활성화를 위해 갖춰야 할 품질 특성 제시
Christine Patterso(2017)	데이터 품질에 대하여 업계 공통으로 사용할	· Completeness · Uniqueness	· Usability · Timing issues	· SIX CORE DATA QUALITY

저자 정보 (연도)	연구 맥락	데이터 품질 항목/차원 Or 독립변수/Mediator	데이터 품질 세부 항목/차원 Or 종속변수	Methods & Key Findings
	수 있는 핵심 함의 제시	· Timeliness · Validity · Accuracy · Consistency	· Flexibility · Confidence · Value	DIMENSIONS에 대한 정의 및 각 핵심 차원에 대한 품질 평가 방식 제시
Pipino *et al.* (2002)	조직적 차원에서 데이터 베이스 관리 및 사용의 이해관계자들이 동의할 수 있는 평가 항목에 대한 논의	Completeness Validity Uniqueness Internal/ External Consistency Anomaly detection	· Accessibility · Appropriate Amount of Data · Believability · Completeness · Concise Representation · Consistent Representation · Ease of Manipulation · Free- of Error · Interpretability · Objectivity · Relevancy · Reputation · Security · Timeliness · Understandability · Value-added	· 실질적으로 적용할 수 있는 데이터 품질에 대한 객관적/주관적 평가 매트릭스 개발 및 검증
David Loshin (2009/2011)	데이터 품질의 개선을 위한 실무자 가이드	· Uniqueness · Accuracy · Consistency · Completeness · Timeliness · Currency	· Accuracy · Lineage · Semantic · Structure · Completeness · Consistency · Currency · Timeliness · Reasonableness · Identifiability	· 데이터 품질에 대한 포괄적 관점을 제안하고, 데이터 품질 프로그램의 분석, 보고, 전략 도구 사용에 대한 심층적 검토
David Plotkin (2014)	조직 내 데이터 관리자(Data stewards)의 성공적인 역할과 거버넌스 수행 방법	· Completeness · Validity · Accessibility · Timeliness · Consistency · Accuracy		· 데이터 거버넌스의 차원에서 일관된 방식으로 사용자들에게 쉽게 접근할 수 있는 고품질의 데이터를 관리(Data stewardship)하는 실질적 방법 제안

Information Systems Review
Volume 26 Number 1
February 2024

A Study on the Artificial Intelligence (AI) Training Data Quality: Fuzzy-set Qualitative Comparative Analysis (fsQCA) Approach

Hyunmok Oh* · Seoyoun Lee** · Younghoon Chang***

Abstract

This study is empirical research to enhance understanding of AI (artificial intelligence) training data project in South Korea. It primarily focuses on the various concerns regarding data quality from policy-executing institutions, data construction companies, and organizations utilizing AI training data to develop the most reliable algorithm for society. For academic contribution, this study suggests a theoretical foundation and research model for understanding AI training data quality and its antecedents, as well as the unique data and ethical aspects of AI. For this purpose, this study proposes a research model with important antecedents related to AI training data quality, such as data attribute factors, data building environmental factors, and data type-related factors. The study collects 393 sample data from actual practitioners and personnel from companies building artificial intelligence training data and companies developing artificial intelligence services. Data analysis was conducted through Fuzzy Set Qualitative Comparative Analysis (fsQCA) and Artificial Neural Network analysis (ANN), presenting academic and practical implications related to the quality of AI training data.

Keywords: AI Training Data, Data Quality, AI Ethics, Fuzzy Set Qualitative Comparative Analysis (fsQCA), Artificial Neural Network Analysis (ANN)

* Principal Researcher, Ph.D. Candidate, National Information Society Agency, Daegu, Korea
** Ph.D. Candidate, School of Management and Economics, Beijing Institute of Technology, Beijing China
*** Corresponding Author, Associate Professor, Nottingham University Business School China, University of Nottingham Ningbo China, Ningbo, China

◐ 저 자 소 개 ◑

오 현 목 (ohm@nia.or.kr)

한국항공대학교에서 전산학을 공부하고, 한국과학기술원(KAIST)에서 IT 경영학 석사를 취득하였고, 성균관대학교 일반대학원 경영학과에서 경영학박사 과정을 수료하였다. 현재 한국지능정보사회진흥원(NIA)에 수석연구원으로 재직 중이며, 주요 관심 분야는 빅데이터 분석 및 인공지능 데이터 정책이다.

이 서 연 (seoyounlee@bit.edu.cn)

북경이공대학교 관리경제학원에서 경영과학 & 공학(정보시스템) 전공으로 경영학 박사 과정을 수료하였다. 이화여자대학교에서 방송영상학과 미디어학 복수전공으로 학사 학위를 졸업하고, 한국과학기술원(KAIST) 경영대학에서 정보미디어 MBA 석사 학위를 취득하였다. 대한항공 여객사업본부에서 정보시스템 관련 IT 기획 실무를 담당했다. Industrial Management & Data Systems, Service Business, Creativity & Innovation Management, 지식경영연구에 논문을 게재했으며, 한국경영정보학회, 한국지식경영학회, 인터넷전자상거래학회 학술대회에서 우수 논문상을 수상하고, 국내외 학회에서 다수의 논문을 발표하였다. 현재 중국 국가자연과학기금(NSFC)과 미국 국립과학재단(NRF) 지원의 우울증 환자를 위한 인공지능 챗봇 및 상담 서비스 개발 프로젝트의 연구원으로 참여 하고 있다. 관심 연구 분야는 메타버스, AI 윤리, 인공지능 디자인 & 관리, 서비스 로봇, 디지털 트랜스포메이션, IT 사용자 행동 등이다.

장 영 훈 (Younghoon.Chang@nottingham.edu.cn)

University of Nottingham Ningbo China, Nottingham University Business School China에서 Marketing & Analytics 부교수로 재직하고 있다. 한국과학기술원(KAIST) 경영대학에서 기술경영 (정보시스템) 전공으로 공학박사 학위를 취득하였고, 중국 북경이공대학교에서 경영공학 부교수(특별연구원)와 박사지도교수를 역임하였으며, 북경시 외국인 고급인재로 등재되어있다. 현재까지 50여 편의 논문을 SCI급 국제저널과 KCI 저널에 게재했으며, Industrial Management & Data Systems, Journal of Computer Information Systems와 APJIS에서 Associate Editor로 활동하고 있으며, 국내저널은 경영정보학연구와 지식경영연구의 편집위원으로도 활동하고 있다. 관심 연구 분야는 메타버스 생태계, 인공지능과 로보틱스 매니지먼트, 빅데이터 분석과 생태계, IT 디자인 & 관리, 디지털 트랜스포메이션, IT 사용자 행동 등이 있다.

논문접수일 : 2023년 10월 18일 **게재확정일 : 2024년 11월 23일**
1차 수정일 : 2023년 11월 18일