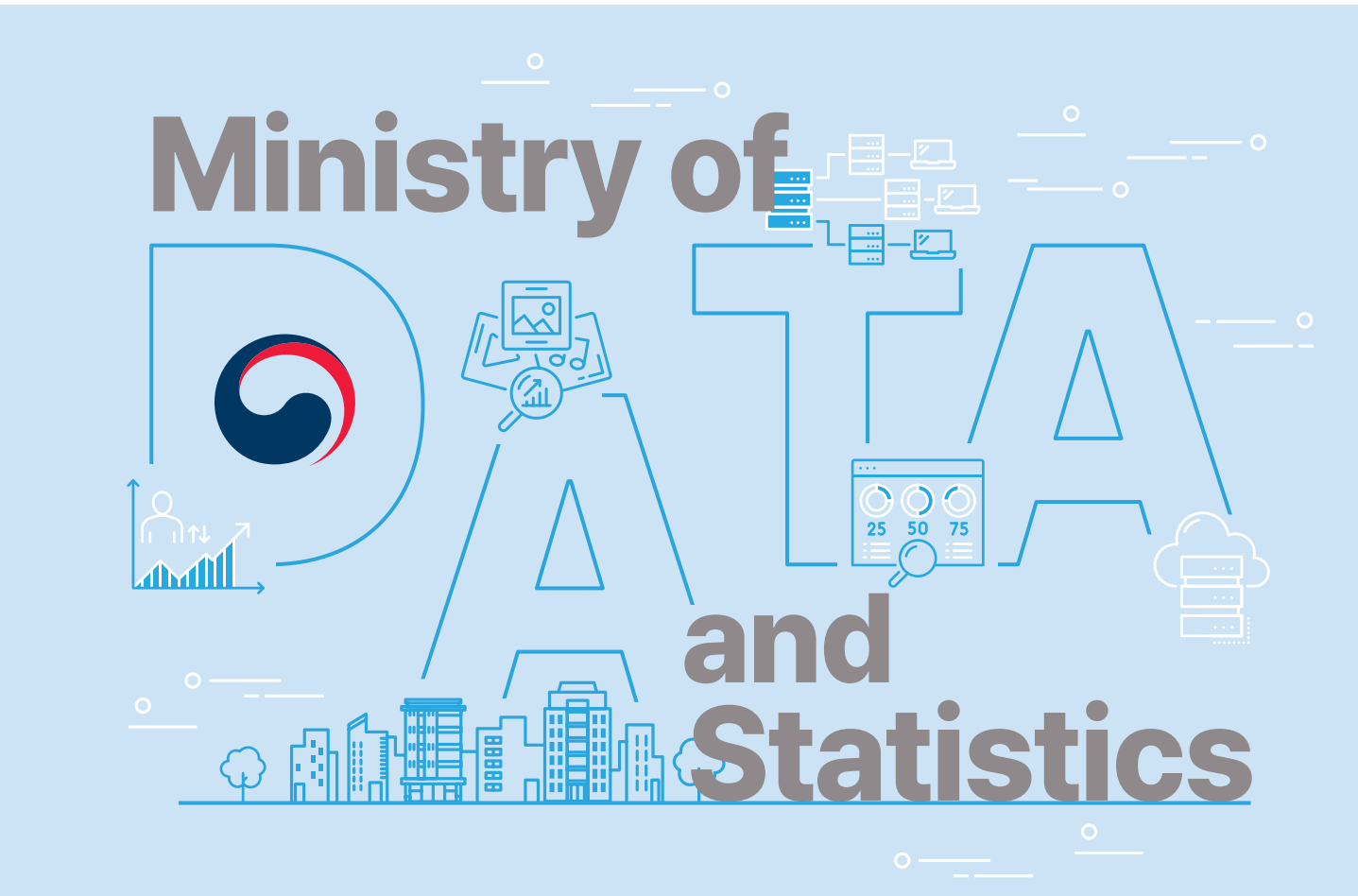
통계의 창

WINDOW OF STATISTICS

2026 SUMMER
Vol. 37



통계청이 국가데이터처로 다시 출발합니다.

'국민이 믿을 수 있고, 쉽게 쓸 수 있는 데이터 제공'을 최우선 가치로

CONTENTS

통계의 창
2026 SUMMER
Vol. 37

발행일 2026년 5월 29일
발행인 이인수
발행처 국가데이터인재개발원
기 획 차진숙, 김록환, 박상미
주 소 대전광역시 서구 한밭대로 713(월평동) 통계센터
국가데이터인재개발원
전 화 042-366-6151, 6155
팩 스 042-366-6498
이메일 knowan@korea.kr, tkdal1213@korea.kr
디자인 및 인쇄 씨앤피(042-825-0701)

ISSN 2005-1379

통계 ISSUE

6 **국가데이터처로의 새출발, 국가데이터관리본부와 함께**
공미숙 | 국가데이터처 국가데이터관리본부 본부장

통계 이야기

10 **AI·디지털 학습관인 채움관을 개관하며**
김록환 | 국가데이터처 국가데이터인재개발원 교육기획과 사무관

14 **통계데이터센터와 우리 생활의 데이터**
심재호 | 국가데이터처 국가데이터기획협력과 주무관

20 **숫자로 읽는 우리의 일상, 소비자물가지수**
김유미 | 국가데이터처 물가동향과 과장

24 **AI는 어떻게 판단할까-1 : 세상을 읽는 데이터, 데이터를 읽는 통계**
이동준 | 이화여자고등학교 교사

30 **기업이 원하는 AI 인재, 교육은 준비되어 있는가**
– 국내 AI 교육의 현주소와 데이터 활용 사례
진회승 | 소프트웨어정책연구소 책임연구원

38 **AI가 학습하는 데이터는 누구의 것인가**
이청호 | 상명대학교 교수

44 **인공지능 공정성의 명과 암**
김효은 | 국립한밭대학교 인문교양학부 교수

50 **통계로 바라보는 세상 이야기**
신동헌 | 도서출판 지일북스 대표

54 **생성형 AI를 마켓 리서치에 활용하기**
구자룡 | 밸류바인 대표

국가데이터처 시대의 서막,
데이터로 혁신하는 대한민국의 내일

초록이 짙어지는 활기찬 여름, 독자 여러분의 일상에도 기분 좋은 생동감이 가득하시길 바랍니다.
우리 사회를 둘러싼 데이터와 기술의 지형은 그 어느 때보다 빠르고 강렬하게 변화하고 있습니다.
과거의 데이터가 지난 발자취를 기록하는 '일기장'이었다면,
이제의 데이터는 내일을 예측하고 새로운 가치를 창출하는 '나침반'이 되었습니다.
통계·데이터 전문 교양지 『통계의 창』 2026년 여름호(Vol. 37)는
이 거대한 변화의 물결 한가운데서, 대한민국 데이터 생태계의 거대한 전환점이 될
국가 정책의 지향점을 심도 있게 공유하고자 합니다.

대한민국 데이터 혁신의 중심, '국가데이터처'로의 위대한 출범

이번 여름호의 문을 여는 가장 묵직한 화두는 단연 〈국가데이터처로의 새출발, 국가데이터관리본부와 함께〉입니다. 1948년 공보처 통계국으로 출발하여 1990년 통계청 개청을 거친 우리 조직은, 지난 2025년 9월 국무총리 소속의 '국가데이터처'로 격상되는 역사적인 승격을 맞이했습니다. 이는 단순히 기관의 위상이 높아진 것을 넘어, 분야별·부처별로 분절되어 있던 데이터 거버넌스의 한계를 극복하고 범정부 차원의 데이터 연계·활용을 총괄 조정하겠다는 국가적 결단입니다.

이 거대한 도약의 실질적인 실행축으로 출범한 '국가데이터관리본부'의 세 가지 핵심 과제는 대한민국의 내일을 바꿀 강력한 이정표가 될 것입니다.

첫째, 공공과 민간의 데이터 연계를 위한 법적 근거인 「국가데이터기본법」 제정을 추진하고 정책 컨트롤 타워인 '국가데이터위원회'를 설치하여 확고한 제도적 인프라를 구축합니다.

둘째, 정형화된 공식 통계를 AI가 올바르게 해석할 수 있도록 온톨로지(ontology) 기반의 'AI 친화적(AI-ready) 통계 메타데이터'를 구축하여, 잘못된 수치를 인용하는 인공지능의 환각(hallucination) 현상을 근본적으로 해결해 나갈 것입니다.

셋째, 여러 부처의 자료를 안전하게 결합한 융합데이터(소득이동DB, 자살DB, 사회보장DB 등)를 개발하고 통계데이터센터 내에 AI를 도입하여 데이터 이용 편의성을 극대화하는 성과(Value-up)를 증명해 보이겠습니다.

AI 시대를 바라보는 입체적이고 비판적인 시선

국가데이터처가 주도하는 지식 생태계 혁신에 발맞추어, 교육과 실무 현장 또한 빠르게 체질을 개선하고 있습니다. 미래 인재 양성을 위해 새롭게 문을 연 첨단 인프라를 조명한 〈AI·디지털 학습관인 채움관을 개관하며〉와, 과거의 '발품' 중심 상권 분석에서 벗어나 과학적 데이터 허브로 진화한 〈통계데이터센터와 우리 생활의 데이터〉 기사를 통해 삶의 질을 바꾸는 데이터 공간의 역동성을 전합니다.

본격적인 기술 담론에서는 마법처럼 보이는 인공지능의 판단 과정을 정교한 통계적 확률로 해부한 〈AI는 어떻게 판단할까 1: 세상을 읽는 데이터, 데이터를 읽는 통계〉와, 단순 코딩 기술을 넘어 비즈니스 문해력을 갖춘 진짜 인재 양성을 고민한 〈기업이 원하는 AI 인재, 교육은 준비되어 있는가〉를 수록했습니다.

특히 이번 호에서는 기술의 화려함 뒤에 숨은 윤리적 난제들을 날카롭게 직시합니다. 평범한 개인들의 삶의 흔적이 담긴 데이터의 주권과 진정한 동의의 가치를 성찰한 〈AI가 학습하는 데이터는 누구의 것인가〉와, 미국의 재범 예측 알고리즘(COMPAS) 사례를 통해 수학적으로 양립하기 어려운 공정성 기준의 모순을 짚어낸 〈인공지능 공정성의 명과 암〉은 신뢰 기반의 AI 사회를 위한 깊은 이정표가 될 것입니다. 아울러 실무자들을 위해 가상 고객 페르소나를 활용한 최신 실전 가이드 〈생성형 AI를 마켓 리서치에 활용하기〉도 알차게 준비했습니다.

차가운 숫자 속, 국민의 따뜻한 일상을 품다

거대한 담론의 최종 목적지는 결국 국민의 일상입니다. 국가데이터처 물가동향과에서 준비한 〈숫자로 읽는 우리의 일상, 소비자물가지수〉에서는 5년 주기로 찾아온 '2025년 기준 대개편' 소식을 전하며, 밀키트나 전기차 연료 등 변화된 라이프스타일을 통계에 정확히 반영하여 공식 물가와 체감 물가의 차이를 줄이려는 치열한 노력을 소개합니다. 마지막으로 최근의 러닝 열풍, 디저트 소비 트렌드(두쫀쿠), 소도시 여행과 로컬힙 문화, 맞벌이 가구의 일상까지 우리 시대의 생생한 자화상을 따뜻한 텍스트로 엮어낸 〈통계로 바라보는 세상 이야기〉가 여름날의 풍요로운 지적 읽을거리를 완성해 줄 것입니다.

국가데이터처라는 새로운 이름으로 돛을 올린 만큼, 『통계의 창』 역시 국민과 데이터를 가장 투명하고 안전하게 잇는 깨끗한 창(窓)이 되겠습니다. 이번 여름호가 독자 여러분의 '데이터 리터러시'를 한 뼘 더 성장시키는 유익한 길잡이가 되기를 바랍니다.

독자 여러분 모두, 데이터처럼 명쾌하고 시원한 여름 보내시길 기원합니다.

-국가데이터처 국가데이터인재개발원 교육기획과 올림-

ISSUE

국가데이터처로의 새출발, 국가데이터관리본부와 함께

공미숙 | 국가데이터처 국가데이터관리본부 본부장

1 들어가며

"국가 통계의 총괄·조정 및 통계·데이터 관리 기능을 강화하기 위해, 통계청을 국무총리 소속 '국가데이터처'로 승격하겠습니다."

2025년 9월 7일, 새 정부의 정부조직개편 방안이 발표되었다. 그 중에는 기획재정부 산하의 외청 기관이었던 통계청을 국무총리 소속 처(處)급 기관으로 승격시키는 내용도 포함되었다. 1990년 경제기획원 조사통계국으로부터 통계청으로 승격된 이후, 35년 만의 중요한 전환이다.

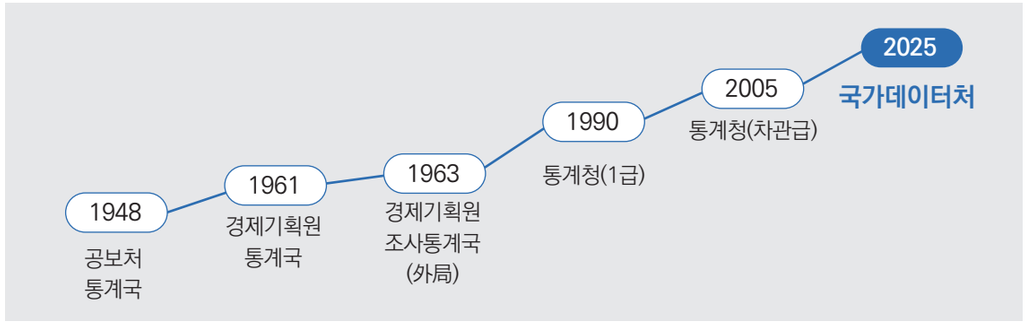
1948년 공보처 통계국으로 출발하여 1990년 1급 기관으로 개청한 통계청은 2005년 차관급으로 승격되었다. 1966년 경제기획원 통계국 시절에는 국내 최초로 컴퓨터를 도입하였고, 이후 현재까지 1,378종에 이르는 국가통계의 총괄·조정, 국가통계포털(KOSIS)과 통계데이터센터(SDC) 등을 통한 통계데이터 활용 확대 등 국가의 데이터 혁신을 위해 끊임없이 노력해 왔다.

통계청이 왜 국가데이터처가 되었는지, 어떤 역할을 새롭게 맡게 되었고 앞으로 무엇을 추진하고자 하는지 이하에서 살펴본다.

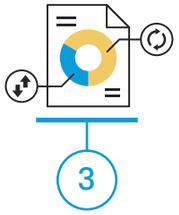
2 왜 국가데이터처인가?

인공지능(AI)이 일상과 산업 전반에 빠르게 확산되면서, 국가 경쟁력을 좌우하게 될 정도로 중요해진 시대가 되었다. 인공지능이 더 나은 지식에 기반하여 사용자의 요구에 정확하고 신속하게 답하기 위해서는 무엇보다 고품질의 학습데이터를 확보하는 것이 중요하다. 즉, AI 경쟁력의 기반은 데이터의 품질과 활용 가능성에 있다.

AI를 비롯한 다양한 데이터 수요에 대응하고, 고품질의 데이터 생산을 위해서는 데이터 보유기관과의 상호 협력이 필수적이다. 하지만 산업, 공공, 농업, 보건 등 분야별·부처별로 분절된 데이터 거버넌스 하에서 범정부적 협업을 도모하기에는 한계가 존재한다. 청년 일자리, 지역 소멸 등 사회문제를 해결하기 위해서는 더 복합적이고 다각적인 데이터 분석을 필요로 하는데, 이는 다양한 데이터의 연계·결합을 통해서만 가능하다. 이러한 배경에서 국가데이터처의 승격은 인공지능 시대에 통계 및 공공·민간 데이터를 아우르는 범정부 데이터 거버넌스를 확립하여 각종 통계와 데이터 연계·활용 기능을 강화하기 위해 이루어졌다.



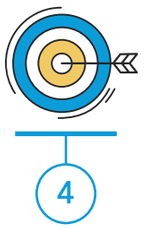
〈국가데이터처 위상 변화〉



3 무엇이 달라지는가?

국가데이터처는 경제·사회·환경 등 전 분야 통계를 관장하는 국무총리 소속 '처'로서, 공공·민간데이터 연계·활용을 강화하고, 범정부 총괄·조정 역할을 수행함으로써 데이터 시너지 효과를 기대할 수 있다.

통계청 시절의 핵심 역할이 국가통계의 생산·품질관리·조정에 있었다면, 국가데이터처의 역할은 이를 토대로 공공과 민간에 흩어져 있는 데이터를 연계하고, 국가 차원의 활용 체계를 설계하며, AI 시대에 필요한 고품질 데이터 기반을 조성하는 데까지 확장된다. 이를 반영하여 「데이터로 통하는 대한민국, 국민과 함께하는 국가데이터처」라는 비전과 「국민이 신뢰하는 데이터, 국민에게 유용한 데이터 서비스」라는 미션을 새롭게 설정하고, 데이터를 수집하고 관리하는 역할을 넘어 데이터 혁신을 주도하는 중심 기관으로 거듭나고자 한다.



4 국가데이터관리본부의 역할은?

국가데이터처가 '데이터 총괄·조정'이라는 새로운 임무를 부여받으면서, 이를 실질적으로 추진할 전담 조직으로 '국가데이터관리본부'가 출범[1]하였다. 국가데이터관리본부는 범정부 데이터 총괄·조정 및 통계와 공공·민간 데이터 간 연계·협력 지원을 위해 만들어졌기에, 국가데이터처 승격의 실행축으로서 다음의 세 가지 과제를 추진한다.

첫째, 국가데이터 제도 기반을 마련한다. 먼저, 범정부 데이터 거버넌스 구축과 민·관 데이터 연계·활용을 위한 법적 근거로서 「국가데이터기본법」 제정을 추진한다. 국가 차원에서 관리·연계 및 활용이 필요한 인구·고용·보건 등 다양한 분야의 데이터를 '국가데이터'로 지정하고, 데이터 관련 부처가 참여하는 '국가데이터위원회'를 설치하여 공공·민간 데이터 정책을 범정부 차원의 최우선 과제로 다루고자 한다. 국가데이터 지정이 개별 기관이 보유한 데이터를 단순히 한 곳에 모으기 위한 것은 아니다. 국가적 활용가치가 큰 데이터를 체계적으로 지정·관리하고, 품질을 높이며, 필요한 경우 안전하게 연계해 정책과 연구, 국민 서비스에 활용할 수 있도록 하는 기반을 마련하기 위한 것이다.

둘째, AI 친화적(AI-ready) 통계 메타데이터(metadata) 구축을 추진한다. 현재 서비스되고 있는 대부분의 인공지능 모델들은 공식 통계를 찾거나 활용하는 데 어려움을 보인다. 공식 통계가 서비스되는 국가통계포털(KOSIS) 내의 통계표를 읽어내지 못하기 때문이다. 정확한 통계를 인용하지 못하고, 잘못된 수치나 오래된 통계를 그럴듯하게 설명하는 인공지능 '환각(hallucination)'이 나타난다. 이에 온톨로지(ontology) 기반으로 AI가 통계 데이터 구조와 의미를 해석하고 추론할 수 있도록, 정형화된 형식과 연결구조를 갖춘 메타데이터를 구축하고자 한다. 사람이 통계표의 제목, 주석, 분류체계, 작성 기준을 함께 읽고 의미를 파악하듯이, AI도 통계를 올바르게 활용하려면 수치뿐 아니라 그 수치가 어떤 기준으로 작성되었고 어떤 개념과 연결되는지를 이해해야 한다. AI 친화적 통계 메타데이터는 바로 이러한 '통계의 의미 정보'를 기계가 읽고 해석할 수 있는 형태로 정비하는 작업이다.

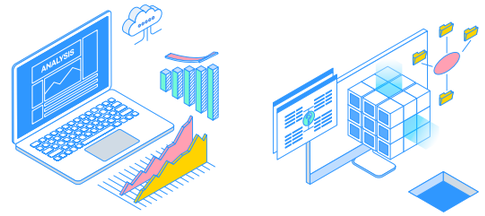
셋째, 국민이 체감하는 데이터 성과를 만드는 데이터 가치 제고(value-up)를 추진한다. 국가데이터처가 운영하는 통계데이터센터(Statistical Data Center)[2] 내에 AI를 도입하여 데이터 간 연계 가능성 진단부터 분석 및 반출까지의 과정을 자동화함으로써 데이터 이용의 편리성을 높이고자 한다. 또한 단일 영역의 데이터만으로 접근하기 어려운 복잡한 사회적 문제의 해결을 위해 소득이동DB, 자살DB, 사회보장DB 등과 같이 여러 부처의 다양한 자료를 결합한 융합데이터를 개발해 데이터의 가치를 높여 나갈 예정이다.

1) 국가데이터관리본부 외에 국가데이터기획협력관, 국가데이터기획협력과, 인공지능통계혁신과도 신설되었다.

5 나가며

데이터에 대한 관심은 하루 이틀의 일은 아니다. 알파고와 빅데이터가 휩쓸고 간 이른바 '4차 산업혁명' 시기를 거쳐, 가명정보 이용을 활성화하기 위해 데이터 3법[3]이 개정됐고, 데이터를 한데 모아 활용하려는 '데이터 댐'이 등장하기도 하였다. 하지만 데이터를 한 곳으로 모으는 것도 쉽지 않을 뿐만 아니라, 모아놓는 것만으로는 데이터의 활용 확대까지 나아가지 못했다. 데이터의 부처 간 칸막이를 해소하기 위한 '디지털플랫폼 정부' 구현 노력도 한계가 있었다. 과거에 비하면 데이터의 양적 성장은 분명하지만, 품질 높은 데이터를 제대로 활용하는 일은 아직 갈 길이 먼 상황이다.

국가데이터처는 1,378종의 국가통계를 총괄·조정하고, 67종의 통계를 생산하는 과정에서 데이터를 직접 다뤄 본 경험이 가장 풍부한 기관이다. 국가통계는 가장 정제된 데이터로서 사회문제 해결에 널리 활용되고 있다. 이러한 경험과 지식을 바탕으로 국가데이터처는 통계청 시절부터 데이터 활용의 중요성을 간파했고, 이에 통계데이터가



다양한 데이터와 연계·결합하여 활용될 수 있는 기반인 '통계등록부[4]'를 구축했다. 그리고 통계등록부와 여러 기관에 흩어져 있던 11종의 공적·사적 연금데이터를 연계·결합하여 세계 최초의 '포괄적 연금통계'를 개발하고, 인구동태코호트DB에 취업활동, 아동가구, 청년 통계등록부를 연계하여 개인의 혼인·출산 요인을 확인하는 '인구동태패널통계'를 개발하는 등 다양한 데이터 연계로 새로운 가치를 창출하는 데에 힘써왔다.

국가데이터처가 지향하는 바는 명확하다. 신뢰할 수 있는 데이터가 필요한 곳에 안전하게 연결되고, 국민의 삶을 개선하는 정책과 서비스로 이어지는 사회를 만드는 것이다. 데이터의 가치는 축적 그 자체에 있지 않다. 데이터가 정확하게 해석되고, 적절하게 결합되며, 국민이 체감할 수 있는 성과로 전환될 때 비로소 공익적 가치가 실현된다.

국가데이터관리본부는 국가데이터처의 이러한 전환을 실질적인 성과로 연결하는 실행 조직이다. 앞으로 「국가데이터기본법」 제정, AI 친화적 통계 메타데이터 구축, 통계·공공·민간데이터의 연계·융합 확대를 통해 범정부 데이터 협력의 기반을 다져 나갈 것이다. 이를 통해 국민이 데이터를 신뢰하고, 정부와 사회가 데이터를 더 쉽고 안전하게 활용할 수 있는 국가데이터 생태계를 만들어 가고자 한다.

2) 이용자가 국가통계 마이크로데이터, 행정통계자료 및 민간자료를 편리하게 이용하고, 연계·융합이 가능하도록 구축된 오프라인 데이터 플랫폼

3) 「개인정보 보호법」, 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」, 「신용정보의 이용 및 보호에 관한 법률」

4) 통계작성 및 데이터 간 연계·결합을 지원하기 위해 각종 조사자료와 행정자료를 결합하여 국내 모든 인구, 사업체의 특성을 수록한 자료로, 정책 대상·주제별로 필요한 자료를 연계·분석하는 통계데이터 허브(hub) 역할

AI·디지털 학습관인 채움관을 개관하며

김록환 | 국가데이터처 국가데이터인재개발원 교육기획과 사무관

1 들어가기

인공지능(AI) 강국 실현을 위한 산·학·연의 정책과 성과 등이 각종 미디어를 통해 숨 가쁘게 전해지고 있다. 교육도 선도적인 분야 중 하나로 교육부는 지난해 「모두를 위한 인공지능 인재양성 방안(AI for All)」을 발표하고, 국민 누구나 인공지능을 쉽게 활용할 수 있도록 전 생애주기에 걸친 보편적 인공지능 교육 확대와 인공지능 세계 3강 도약을 견인하는 다층적 인공지능 인재양성을 추진하고 있다. 과학기술정보통신부는 인공지능기본법을 제정하고 전문인재 양성을 위해 인공지능 혁신대학원과 「코디세이(Codyssey)」통합 과정을 신설하는 등 정부부처들도 다양한 정책을 시행하고 있다. 가시적인 성과들도 나오고 있는데, 미국 스탠퍼드대에서 발표한 「인공지능 지수(AI Index) 2026」에서 '25년 출시 주목할 만한 인공지능 모델 수' 3위, '인공지능 도입률 상승폭' 1위 등이 주목할 만하다.

2 채움관 알고 활용하기

한편 국가데이터처는 통계와 데이터 분야의 오랜 경험과 전문인력, 대규모의 국가통계DB 및 관련 암호기술 등을 기반으로 데이터시대 기관 승격과 더불어 'AI 기반 사회 구현을 위한 데이터 활용·관리체계 확립'을 위한 범정부 데이터 거버넌스를 구축하고 있다. 또한, 국가데이터인재개발원은 AI 기반 데이터 혁신 전문인재 양성 및 대국민 통계·데이터 리터러시 제고를 목표로 교육과정을 시대에 맞게 전환하고, 교육시설 인프라를 강화하는 등 국가 데이터·통계 전문교육기관으로 역할을 선도적으로 수행해 나가고 있다. 이러한 시점에 2026년 4월 개관한 채움관은 AI 활용 능력을 갖춘 데이터 전문인재 양성을 위한 공공 교육시설 인프라 제공에 큰 의미가 있다고 본다.

채움관은 증가하는 통계 관련 교육 수요에 대응하고, 교육생을 위한 자기주도형 및 체험형 교육시설을 갖추고자 2020년 '통계교육원 다목적교육관 신축'사업으로 시작하여, 2025년 9월 준공까지 한국자산관리공사에서 위탁개발한 교육연구시설로 건축물에너지효율 최고등급(1+++) 및 BF인증을 받은 공공건축물이다.

채움관은 국가데이터인재개발원 본관동인 통계센터 건물 좌측에 위치하며 연면적 4,397㎡의 3층 건물로 우수한 접근성과 주변에 다양한 편의시설 및 상권이 형성되어 있다. 채움관의 주요시설로는 AI컴퓨팅실과 하이브리드 강의실, 모듈형 강의실, 일반 강의실, 다목적강당 등이 있다.

〈그림1〉 채움관 전경 및 교육시설 현황

특히, AI컴퓨팅실(AI Computing Lab)은 2026년 도입한 최신 AI컴퓨터로 RTX 4080 SUPER GPU와 CPU i9 등 사양을 갖추고 있다. 이를 활용해 빠른 모델 학습 속도로 LLM 등 대용량 데이터셋 딥러닝 교육이나 멀티태스킹 실습이 가능한 교육환경을 제공한다.

하이브리드 강의실(Hybrid Room)은 40석 규모로 영상회의시스템과 전자칠판 등 기기가 갖추어진 계단형 강의실로 격자형 좌석 배치와 편안한 환경을 제공하여 강사와 교육생 또는 교육생 간 자유로운 토론과 활발한 의사소통이 필요한 교육에 최적화된 강의실이다.

또한 모듈형 강의실(Flexible Room)은 40석 규모로 교육 특성에 맞춰 강의시설 재배치와 구성이 용이한 맞춤형 강의실로 토론 및 실습, 협업이 필요한 교육에 유용하다.

다목적강당은 200석 규모지만 별도의 좌석은 설치되지 않은 광장형 구조로 각종 교육관련 행사나 체육활동이 가능한 공간이다. 이외 공용시설로 교육생의 편안한 교류와 소통이 가능한 야외 테라스 라운지를 비롯한 휴게공간이 마련되어 있다.

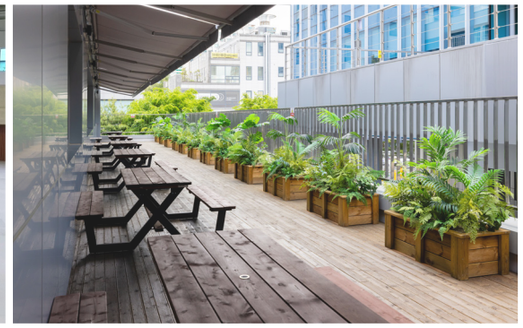
국가데이터인재개발원은 타 교육기관에서 제공하는 AI 및 디지털 인재 양성 프로그램과는 차별화된 교육과정을 설계하고, AI를 활용한 통계 및 데이터 실습에 최적화된 교육 환경을 제공한다. 수요자 맞춤형 특화교육으로 교육대상 또한 다양하여 학생을 위한 실용통계 및 수학


〈그림2〉 AI컴퓨팅실


〈그림3〉 하이브리드 강의실과 모듈형 강의실

과정(AI수학·통계과정, 통계캠프, 통계진로체험 등), 교사 대상 통계교육 설계 과정(AI시대의 통계교육 교사 연수 등), 데이터 생산 및 서비스 업무 담당자를 위한 AI 활용 데이터 서비스 혁신 과정, 증거기반 정책 수립에 활용하기 위한 데이터 분석 및 시각화, 그리고 국가데이터처 직원이나 통계작성기관 담당자 중심의 LLM 기반 통계서비스 혁신이나 데이터마이닝·딥러닝 기반 통계작성 고도화 과정, 일반 교양 함양을 위한 통계리터러시 향상 과정 등을 제공할 예정이다.

특히 하루가 다르게 발전하는 인공지능 및 데이터 기술을 교육과정에 빠르게 반영하기 위해 AI 전공 대학교수 등 학계 전문가와 협력하여 전문 강사진을 구성하고, 최신 교육시설인 채움관을 활용해 교육효과 극대화를 목표로 하고 있다.

〈그림4〉 다목적강당과 야외 테라스 라운지

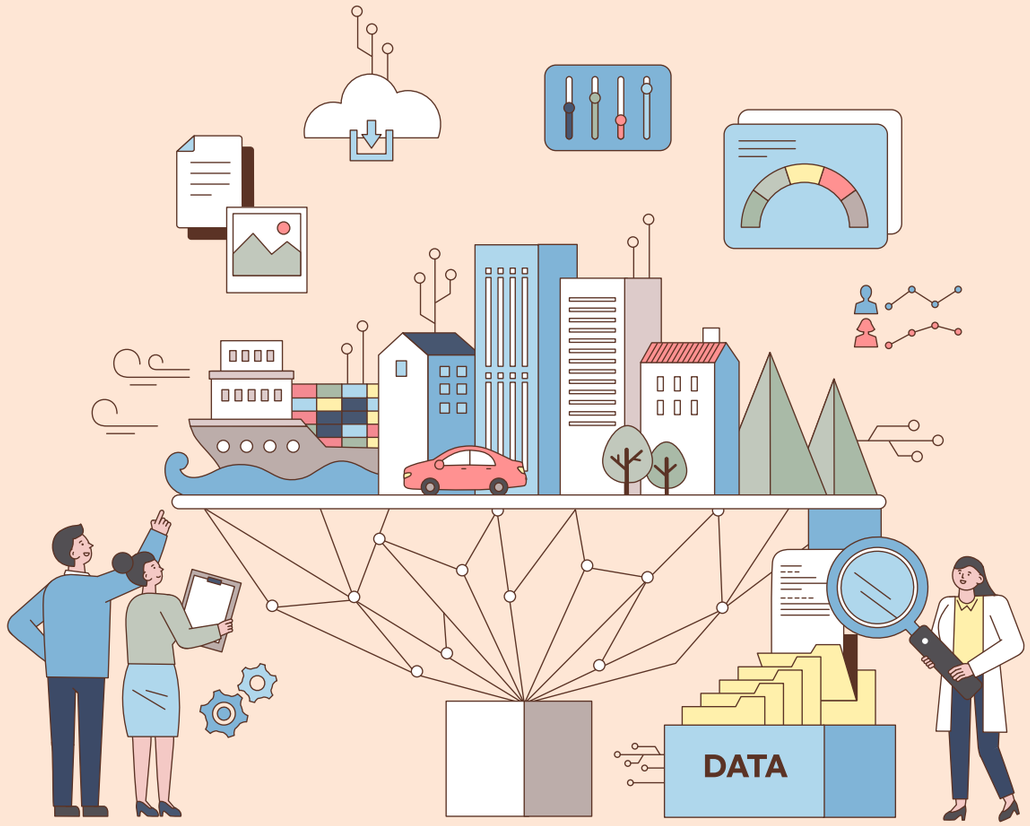
3 나가기

국가데이터인재개발원 채움관은 '누구도 뒤처지지 않는(Leave no one behind)' AI·데이터 교육 기회를 제공하여 지역 격차를 완화하고, 민간 교육기관 대비 합리적인 교육을 제공할 수 있는 기반 시설로 정식 개관함으로써, 통계와 데이터 분야의 오랜 교육 기획 및 운영 경험과 전문적인 강사진뿐만 아니라 첨단 시설 인프라까지 갖추게 된 셈이다. 향후 국가데이터인재개발원은 채움관 개관에 따라 AI·데이터 중심 교육과정 전환을 목표로 전문인재 양성, 실무중심의 이러닝 콘텐츠 개발, 그리고 AI·디지털 교육 확대 기반 강화 등을 위해 채움관을 적극 활용할 계획이다.

인공지능 기술 발전은 생활방식과 일하는 방식, 의사결정 방식 등 대부분의 사람들에게 기본적인 변화를 요구하고 있다. 고도로 발달된 기술일수록 사용자의 능력에 따라 그 기술의 효용성은 크게 달라진다. 전문교육의 중요성이 점증하는 이유 중 하나가 바로 이 때문이다. AI·데이터 시대 첨단 컴퓨터뿐 아니라 다양한 교육시설과 교육 콘텐츠로 꽉 채워진 채움관은 데이터 생태계를 채울 인재 양성을 목표로 전문지식을 채워주는 학습공간이 될 것으로 기대한다.

통계데이터센터와 우리 생활의 데이터

심재호 | 국가데이터처 국가데이터기획협력과 주무관



| 1 | 발품에서 데이터로 : 통계데이터센터가 이끈 생활의 변화

16년 전 프랜차이즈 유통업을 운영하는 회사에 입사하면서 나의 첫 직장생활이 시작되었다. 신입사원들은 OJT(On the Job Training) 과정을 거쳤는데, 이는 다양한 분야의 업무를 담당하는 선배들을 따라다니면서 사무실이나 현장에서 어떻게 업무가 이루어지는지 직접 배우라는 취지에서 도입된 제도였다.

OJT 과정에서 가장 인상 깊었던 시간은 “개발업무”를 담당하는 선배를 따라 3일간 현장에 동행했을 때였다. “개발업무”란 점포가 입점하기에 적절한 입지의 물건을 발견하고 임대차계약을 맺은 뒤, 해당 점포를 운영할 점주를 찾아 연결해 주기까지의 과정을 맡는 업무였다. 이는 프랜차이즈 사업의 확장과 성공을 위해 가장 중요한 업무라고 할 수 있는데 당시 나와 동행했던 L 선배는, 전년도 개발 실적 전국 1위를 차지한 우수사원이었다.

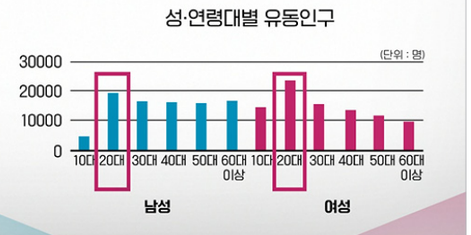
L 선배는 상점이 입점하기에 괜찮은 입지를 찾았다고 판단되면, 한구석에 주차를 한 뒤 마치 영화 속에서나 볼 수 있는 잠복경찰처럼 차 안에만 머물며 그 거리를 지나다니는 사람들의 수를 세기 시작했다. 그 과정은 아침 일찍부터 저녁 늦게까지 계속됐는데, 단순히 유동인구만 체크한 게 아니라 남성과 여성을 나누어 기록했으며, 연령대도 꼼꼼히 구분하여 정리했다. L 선배는 이렇게 최소 3, 4일간 해당 지역의 유동인구와 그 유동인구의 성비 및 연령대를 파악했고, 이를 토대로 상점이 입점할 만한 곳인지 판단을 내렸다. 이러한 꼼꼼함이 그를 우수사원으로 만들어 준 것이다.

그 후로 창업을 하려는 주변의 지인들이 나에게 창업 관련 문의를 해오곤 했는데, 나는 위에서 언급한 OJT 때의 간접경험과 영업 현장에서 직접 뛰던 경험을 바탕으로 다음과 같이 조언하곤 했다.

유동인구 데이터 분석

(단위 : 명)

성별	연령	유동인구
남성	10대	3736.46
	20대	19961.43
	30대	17472.47
	40대	16908.66
	50대	16623.28
	60대 이상	17503.79
여성	10대	5290.84
	20대	24497.19
	30대	15872.93
	40대	13788.31
	50대	11802.38
	60대 이상	10117.65



〈유동인구데이터를 활용한 성별·연령대별 유동인구 가상분석 예시〉

"창업을 위해서는 발품도 좀 팔아야 하고, 적어도 2~3일 정도는 창업을 희망하는 입지에서 계속 관찰하며 유동인구와 성비, 연령별 비율 등을 꼼꼼히 체크해야 한다."

하지만 이러한 조언은 옛날이야기가 되었다. 15년이란 시간이 흘러 국가데이터처의 통계데이터센터팀에 근무하게 된 이후로, 나는 같은 질문을 받았을 때 다음과 같이 대답한다.

"굳이 발품을 팔거나 많은 시간을 쓸 필요 없이, 국가데이터처 통계데이터센터를 이용하면 자신이 창업하고 싶은 위치의 유동인구를 쉽게 파악할 수 있고, 해당 입지에서 발생하는 업종별 매출 현황과 업종별 사업체 수까지도 파악할 수 있다."

과거에는 많은 발품과 시간을 들여야만 알 수 있던 사실들을 이제는 국가데이터처 통계데이터센터에서 몇 시간 만에 파악할 수 있는 시대가 왔기 때문이다. 즉, 대한민국 국민이라면 누구나 통계데이터센터에서 보유 중인 소지역 단위 유동인구 자료를 활용하여 자신이 관심 있는 지역의 요일별, 시간대별, 성별, 연령별 유동인구를 확인할 수 있다. 그리고 통계데이터센터에서는 카드매출 정보 또한 제공하는데 이를 활용하면 업종별 매출 현황까지도 파악할 수 있다. 이게 전부가 아니다. 기업통계등록부라는 자료를 활용하면 어떤 지역 내에 특정 업종을 영위하고 있는 기존 사업체 수가 몇 개나 되는지, 1년 사이에 폐업한 업체 수는 몇 개나 되는지까지도 파악할 수 있어, 사실상 창업을 위한 모든 정보를 통계데이터센터에서 확인할 수 있는 것이다.

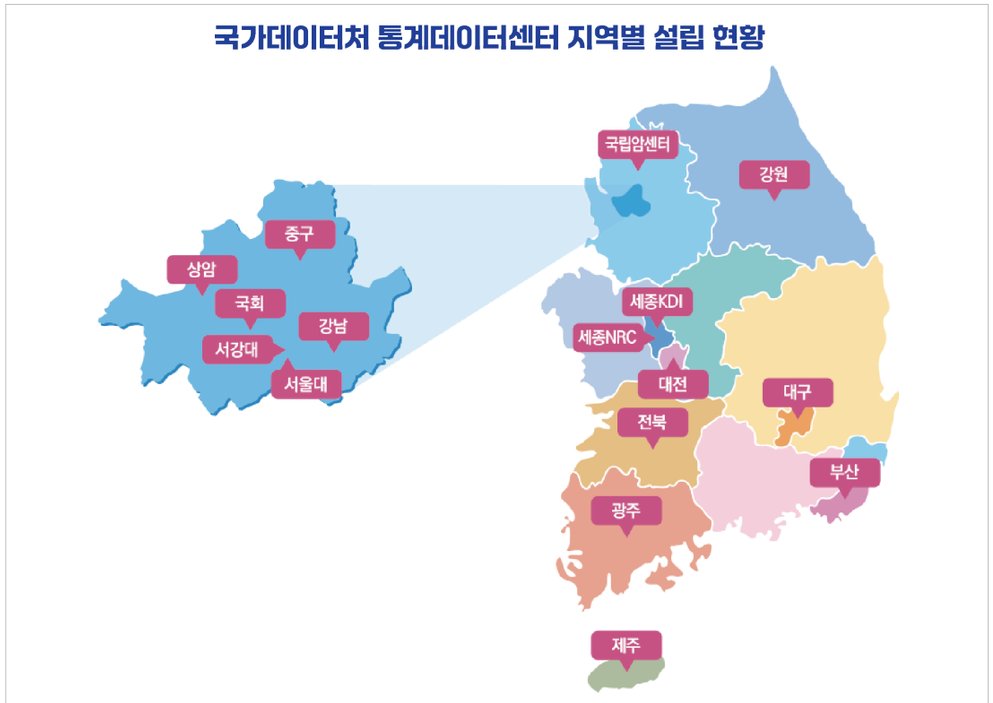
발품을 팔아 수십 시간 노력과 수고를 들이던 과거에 비하면 얼마나 큰 발전인가? 이 사례는 데이터가 우리의 생활과 의사결정에 얼마나 큰 영향을 주는지 보여주는 대표적인 사례라고 할 수 있다. 그리고 이러한 변화의 중심에는 다양한 분야의 데이터를 국민들이 편리하게 이용할 수 있게끔 제공하는 통계데이터센터가 있다.

| 2 | 통계데이터센터가 만드는 생활 속 새로운 의사결정방식

2025년 국가데이터처 통계데이터센터의 연간 이용자 수가 최초로 1만 명을 넘겼다. 이는 2020년 연간 이용자 수 1,630명에 비해 500% 이상 증가한 수치로, 그만큼 점점 통계데이터센터를 이용하는 사람 수가 증가하고 있다는 뜻이다. 또한 2025년 9월에는 강원대학교에 통계데이터 강원센터가 설립되면서 수도권에 위치한 7개의 센터를 포함, 전국에 16개의 센터가 운영되고 있다. 하지만 아직도 통계데이터센터가 제공할 수 있는 가치와 가능성은 충분히 알려지지 않았다.

이는 아직도 사람들이 데이터를 "전문가의 영역" 혹은 "어려운 분석 도구" 정도로만 인식하는 경향이 있기 때문이다. 하지만 내가 통계데이터센터에서 근무하며 느낀 가장 큰 변화는, 데이터가 더 이상 특정 집단의 전유물이 아니라 누구나 자신의 삶에 직접 활용할 수 있는 실질적인 도구가 되었다는 점이다.

과거에는 사업을 시작하기 위해서는 발로 뛰어야 했고, 정책을 설계하기 위해서는 오랜 경험과 직관에 의존해야 했으며, 개인의 중요한 의사결정 또한 제한된 정보 속에서 이루어질 수밖에 없었다. 그러나 지금은 다르다. 우리는 데이터를 통해 보이지 않던 패턴을 발견하고, 경험이 아닌 근거를 기반으로

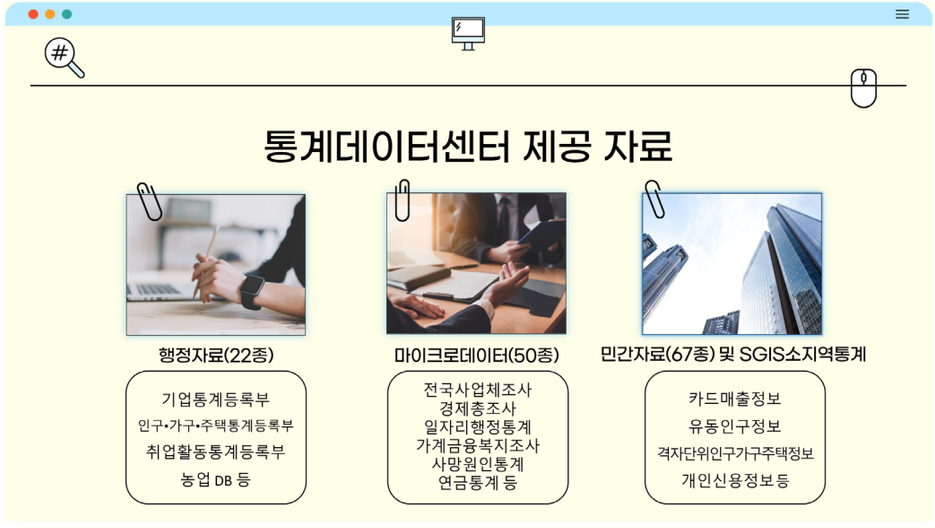


선택할 수 있는 시대에 살고 있다. 그리고 그 근거가 될 다양한 분야의 데이터를 제공해주는 공간이 바로 통계데이터센터이다.

통계데이터센터에서는 기업통계등록부, 인구·가구·통계등록부, 취업활동통계등록부와 같은 행정자료 및 통계조사과정에서 획득한 다양한 분야의 마이크로데이터, 통신사 유동인구정보, 카드사 카드매출정보 그리고 지리 정보 기반 소지역 통계 등도 제공하고 있는데, 이를 활용할 경우 일상생활이나 정책 수립에 유용한 많은 정보들을 얻을 수 있다.

예를 들어, 학령기 자녀를 둔 부모가 거주지를 선택할 때도 통계데이터센터의 데이터를 활용할 수 있다. 특정 지역의 연령별 인구 구조, 가구 형태, 주택 유형 등을 열람·분석하면 해당 지역의 교육 환경이나 생활 인프라 수준을 보다 객관적으로 판단할 수 있다. 또한 특정한 형태와 분위기의 거주 공간을 원하는 사람이 적합한 거주지를 찾을 때에도, 지역별 인구 구조나 가구 형태, 주택 형태 등에 대한 분석을 통해 본인이 원하는 형태의 거주지가 많은 지역을 찾을 수 있다. 결국 통계데이터센터에서 제공하는 정보들이 단순한 참고 자료를 넘어 삶의 방향을 결정하는 중요한 기준이 될 수도 있는 것이다.

공공기관에서도 인구·가구·주택통계등록부를 활용하여 담당 지역의 인구 및 가구 특성을 파악하고 그에 적합한 복지 정책이나 주거 및 도시 정책을 계획할 수 있고, 정책 효과를 판단할 때도 마이크로



〈 통계데이터센터 제공자료(26.3.31.기준) 〉

데이터 및 각종 통계등록부 등을 활용할 수 있다. 이런 과정을 통해 수립된 효율적인 정책은 개인의 삶에도 긍정적 영향을 미친다는 점에서 통계데이터센터에서 제공하는 모든 자료들이 개개인의 삶과 연결되어 있다고 할 수 있다.

| 3 | 데이터기반의 사회와 통계데이터센터의 역할

또 하나 주목해야 할 점은, 데이터의 활용이 단순히 결과를 분석하는 수준을 넘어 미래를 예측하는 도구로 확장되고 있다는 것이다. 현재 통계데이터센터에서는 AI의 도입을 준비하고 있는데, 완료될 경우 단순히 과거 데이터를 분석하는 것에 머물지 않고 누구나 인공지능의 도움을 통해 미래를 예측하고 구상해 볼 수도 있을 것이다. AI의 도움을 받을 경우 우리는 과거의 데이터들을 바탕으로 미래의 인구구조 및 산업구조가 어떻게 변화할지도 예측할 수 있는데, 이는 국민들의 의사결정 및 삶의 질에도 큰 영향을 미칠 것으로 예상된다.

이러한 변화 속에서 통계데이터센터는 단순한 데이터 제공 기관을 넘어, 데이터 기반 사회로 나아가기 위한 핵심 인프라로서의 역할을 수행하고 있다. 또한 다양한 데이터를 통합적으로 제공하고 자료들 간의 연계를 통해, 이용자들이 보다 깊이 있는 분석을 수행할 수 있도록 지원하고 있다.

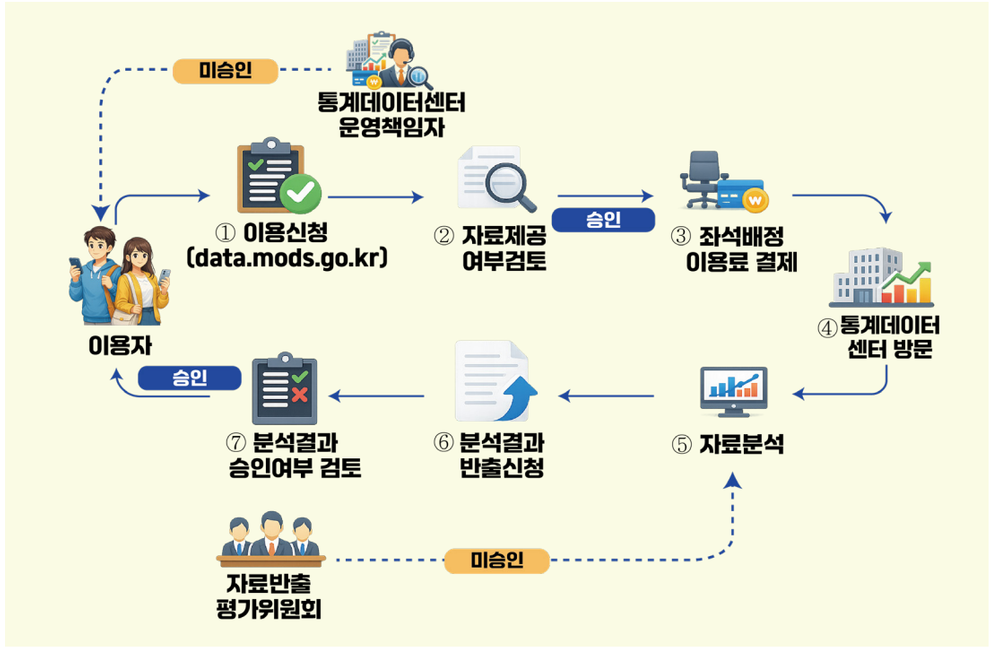
한편 안전한 분석 환경을 제공하여 개인정보 보호와 데이터 활용이라는 두 가지 가치를 동시에 실현하고 있는 공간이라는 점에서도 통계데이터센터의 의의가 크다고 할 수 있다. 현재 국민 누구나 통계데이터센터 홈페이지(data.mods.go.kr)를 통해 회원가입 후 센터 이용신청 예약이 가능하며, 어떤 자료를 이용할 것인지 미리 정하여 신청하면 예약일에 해당 자료를 열람·활용할 수 있다.

여전히 많은 사람들에게 데이터는 “어렵고 복잡한 것”으로 인식되기도 한다. 이는 데이터 자체의 문제가 아니라, 데이터를 활용하는 방법에 대한 경험이 부족하기 때문이라고 생각한다. 실제로 통계데이터센터를 방문하거나 서비스를 이용해 본 사람들은, 생각보다 쉽게 데이터를 활용할 수 있다는 점에 놀라는 경우가 많다. 즉, 데이터 활용의 장벽은 기술적인 문제가 아니라 접근과 경험의 문제인 것이다. 앞으로 중요한 것은 더 많은 사람들이 데이터를 일상 속에서 자연스럽게 활용할 수 있도록 하는 것이다.

이를 위해서는 다양한 데이터 제공뿐만 아니라 데이터를 좀 더 쉽게 활용할 수 있도록 하는 환경조성 및 상세한 교육과정, 그리고 통계데이터센터에 대한 홍보가 필요하다. 데이터는 그 자체로 가치를 가지는 것이 아니라, 어떻게 활용되느냐에 따라 그 가치가 결정되기 때문이다.

나는 통계데이터센터에서 근무하며, 데이터가 단순한 숫자의 집합이 아니라 사람들의 삶과 직접적으로 연결된다는 사실을 체감하고 있다. 과거에 L 선배가 차 안에서 손으로 직접 세던 유동인구는 이제 데이터로 축적되어 누구나 활용할 수 있는 자원이 되었고, 그 데이터는 누군가의 창업을 돕고, 누군가의 정책을 만들며, 또 다른 누군가의 삶의 방향을 결정하는 데 기여하고 있다. 그리고 통계데이터센터는 데이터를 누구나 이해하고 활용할 수 있도록 돕는 역할을 하고 있다.

앞으로 더 많은 사람들이 데이터를 통해 자신의 삶을 설계하고, 더 나은 선택을 할 수 있게 되기를 기대한다. 그리고 그 중심에는 언제나 우리 생활과 가장 가까운 곳에서 데이터를 제공하는 통계데이터센터가 자리하고 있을 것이다.

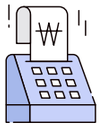


〈통계데이터센터 이용절차〉

숫자로 읽는 우리의 일상, 소비자물가지수

김유미 | 국가데이터처 물가동향과 과장

오늘도 점심 메뉴를 고르며 “내 월급 빼고 다 올랐네” 하고
한숨 섞인 농담을 던지진 않았는가?
워킹맘에게는 아이들 간식인 사과 한 알의 가격이,
직장인에게는 식후 커피 한 잔의 가격이 곧 ‘경제’ 그 자체일 것이다.
국가데이터처는 바로 그 ‘한숨’의 무게를 데이터로 기록하여,
우리 삶의 현주소를 숫자로 짚어내는 곳이다.



소비자물가지수, 그것이 알고 싶다

➜ 소비자물가지수의 의의

소비자물가지수(CPI: Consumer Price Index)는 ‘가구가 일상생활을 영위하기 위해 구입하는 상품 및 서비스의 가격 변동을 종합적으로 측정하기 위해 지수화한 통계’이다. 단순히 물건값이 얼마인지 합산하는 것을 넘어, 국민연금 수령액과 임금을 조정하기 위한 수단이며 디플레이터(deflator)이자 금리를 결정하는 통화정책의 핵심 도구이다. 우리나라의 소비자물가조사는 1936년에 처음 실시되어 1965년에는 전국 단위의 지수를 작성하였고, 1990년부터 국가데이터처가 작성하여 공표하고 있다. 또한 기본분류지수 외에 생활물가지수, 신선식품지수 등 특수분류지수를 작성하여 공식물가와 체감물가의 괴리에 대한 국민의 이해를 돕기 위해 노력하고 있다.

➜ 조사대상 및 방법

소비자물가는 우리가 소비하는 ‘모든’ 물건의 가격을 조사하는 것일까? 그렇지 않다. 가계 소비지출에서 차지하는 비중이 큰(가계동향조사 월평균 소비지출액 1/10,000 이상) 458개(2020년 기준)의 대표 품목을 선정해 매월 전국 40여 개 주요 도시의 26,000여 개 대상처를 직접 방문하여 가격을 수집한다.
이때, ‘조사규격’이라는 엄격한 잣대를 적용하는데, 시장점유율(소비량)이 높고 지속적으로 가격을 조사할 수 있는 상품 및 서비스와 그에 대한 거래단위를 정한다. 예를 들어 라면이라면, '특정 브랜드의 5개입 멀티팩'처럼 구체적인 단위를 정해 오직 가격의 순수한 변화만을 추적한다. 만약 단일 제품으로 해당 품목의 가격변동을 대표할 수 없다면, 2개 이상의 복수 규격을 지정하여 가격변동의 대표성을 높이고, 조사의 정확성을 제고한다.

➜ 지수 산식

지수는 라스파이레스(Laspeyres) 산식을 통해 계산한다. 기준년도의 소비 형태를 고정해 두고, 가격 변화만을 추적하는 방식이다.

$$\text{품목 지수} = \frac{P^t}{P^o} \times 100,$$

$$\text{상위 지수(총 지수 등)} = \left(\frac{\sum_{i=1}^{n} P_i^t Q_i^0}{\sum_{i=1}^{n} P_i^0 Q_i^0}\right) \times 100 = \sum_{i=1}^{n} W_i^0 \left(\frac{P_i^t}{P_i^0}\right) \times 100, \quad W_i^o = \frac{P_i^0 Q_i^0}{\sum_{i=1}^{n} P_i^0 Q_i^0}$$

* t : 비교시점, 0 : 기준시점

어려워 보이지만 원리는 간단하다. “5년 전(기준시점)에 사던 양만큼을 지금 가격으로 사면 돈이 얼마나 더 들까?”를 계산하는 것이다. 그런데 만약 장바구니 속 품목 자체가 구식이 된다면 통계는 현실과 따로 놀기 시작한다. 이것이 바로 5년마다 ‘장바구니(기준)’를 통째로 바꾸는 이유이다.

➜ 지수를 보는 방법

개개의 상품 또는 서비스의 가격을 과거에 ○○○원 하던 것이 현재 △△△원 한다고 표현할 수 있는 것처럼, 물가지수도 '기준시점(2020년)을 100으로 할 때 2025년 연평균 지수는 116.61이다'라고 표현한다. 이것은 기준년도와 동일한 품질의 상품 또는 서비스를 동일한 양만큼 소비한다고 가정할 때 예상되는 총 비용이 기준년도에 비해 약 16.61% 증가했음을 의미한다.

물가수준의 변동은 전월대비, 전년동월대비 물가변동률이 많이 이용되며 전년동월대비 변동률은 다음과 같이 계산된다.

$$\text{전년동월대비 변동률}(\%) = \frac{\text{금월의 물가수준} - \text{전년동월의 물가수준}}{\text{전년동월의 물가수준}} \times 100$$

$$= \frac{\text{금월지수} - \text{전년동월지수}}{\text{전년동월지수}} \times 100$$

또한, 개별 품목의 변동이 총지수의 변동률에 기여하는 정도를 '기여도'라고 하는데, 단위는 퍼센트 포인트(%p)이며, 계산식은 아래와 같다.

$$\text{기여도} = \frac{(\text{비교시점 품목지수} - \text{기준시점 품목지수})}{\text{기준시점 총지수}} \times \frac{\text{품목 가중치}}{\text{전체 가중치}} \times 100$$

2025년 기준 대개편: "우리의 장바구니가 바뀝니다"

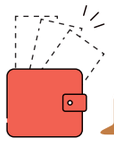
올해는 바로 5년 주기의 개편이 진행되는 해이다. 현재 지수의 기준년도는 2020년, 어떤 품목이 지수에 더 큰 영향을 주는지를 나타내는 가중치의 기준년도는 2022년인데 모두 2025년 기준으로 변경하는 것이다. 이것은 단순한 수정이 아니라 '현실 소비를 다시 반영하는 조사 재설계'이다. 현재 2020년에 멈춰있는 기준시점을 2025년으로 옮겨 통계의 '유통기한'을 갱신하는 작업인데, 이 모든 결과는 올해 12월에 발표될 예정이다.

품목의 세대교체

사람들의 손길이 뜸해진 품목들은 명예로운 은퇴를 맞이하고, 새로운 시대적 흐름이 반영된다. 이번 개편에서는 밀키트, 전기차연료 등 우리 일상에 깊숙이 들어온 항목들이 새롭게 이름을 올리거나 비중이 커질 것으로 보인다.

가중치의 재설계

"요즘 사람들은 어디에 돈을 더 많이 쓰는가?"도 다시 계산한다. 외식 비중이 늘었다면 외식 물가의 영향력을 키우고, 내구재 소비가 줄었다면 그 비중을 낮추어 통계가 실제 지갑 사정과 비슷하게 움직이도록 체질을 개선한다.



"지수는 낮다는데 왜 내 지갑은 텅텅?": 체감물가와의 전쟁

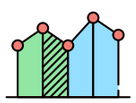
"물가 상승률이 2%라는데, 왜 마트만 가면 손이 떨릴까?" 공식물가와 체감물가의 차이 이유를 좀 더 구체적으로 살펴보면, 크게 두 가지 때문이다. 첫 번째는 개인별 소비의 차이다. 자취하는 대학생은 식비에 민감하고, 통학하는 학생은 지하철, 버스요금과 같은 교통비 변화에 더 민감하다. 두 번째는 자주 사는 품목의 영향이다. 우리는 자주 사는 물건의 가격 상승에 더 크게 반응한다. 예를 들어 커피 가격이 오르면 사람들은 실제 물가 상승률보다 더 크게 체감하게 된다. 결국 소비자물가지수는 '평균의 이야기'이고, 체감물가는 '개인의 이야기'이기 때문에 차이가 발생한다. 이를 더 잘 설명하기 위해 국가데이터처에서는 두 가지 지수를 별도로 관리한다.

생활물가지수

쌀, 삼겹살, 소주, 휘발유 등 구입 빈도가 높고 지출 비중이 커서 안 살 수 없는 품목 144개를 따로 모은 것으로, 일명 '장바구니 물가'라고 하며, 우리의 생존 체감도와 직결된다.

신선식품지수

상추, 배추, 사과처럼 기상 여건에 따라 가격이 롤러코스터를 타는 품목 55개로 작성하며 "금(金)사과" 논란처럼 일시적인 수급 불안에 의한 착시 현상을 걸러내고 먹거리 안정 대책을 세우는 기초자료로 활용한다.



물가는 숫자가 아니라 우리의 '오늘'이다.

소비자물가지수는 차가운 숫자로 보이지만, 사실 우리가 오늘 무엇을 먹고 어떤 시간을 보냈는지에 대한 가장 뜨거운 기록이다. 이번 2025년 기준 개편은 바로 우리들의 변해가는 삶을 가장 정확하게 담아내기 위한 국가데이터처의 노력이다.

오는 12월, 새로운 장바구니 목록이 공개될 때 "아, 내 삶의 변화가 이렇게 국가 통계에 반영됐구나!" 하고 반갑게 살펴봐 주시길 부탁드린다. 통계가 현실을 더 정확히 비출 때, 우리의 내일을 위한 더 나은 대책도 시작될 수 있다.

AI는 어떻게 판단할까-1
: 세상을 읽는 데이터, 데이터를 읽는 통계

이동준 | 이화여자고등학교 교사

세상을 읽는 데이터, 데이터를 읽는 통계

우리는 수많은 데이터 속에서 살아갑니다. 음식점 하나를 고르더라도 리뷰를 살펴보고, SNS에 올라온 사진 한 장이 새로운 맛집의 유행을 만들어내기도 합니다. 인스타그램에서는 하루에 9,500만 건[1] 이상의 사진과 영상이 공유되고 있으며, 이러한 데이터는 단순한 기록을 넘어 사회의 흐름을 보여주는 창이 되었습니다. 최근 우리나라에서 유행한 두쫀쿠나 버터떡의 인증샷이 SNS에 넘쳐났던 것처럼, 사람들이 일상에서 만들어내는 데이터는 먹거리 트렌드부터 세대의 심리까지 다양한 해석의 근거가 됩니다.

그렇다면 이 넘쳐나는 데이터를 어떻게 읽어야 할까요? 수천만 장의 사진을 하나씩 들여다본다고 해서 트렌드가 보이는 것은 아닙니다. 데이터 속에 숨겨진 패턴을 찾아내고 의미 있는 정보를 추출하는 과정이 필요한데, 이때 통계가 핵심적인 역할을 담당합니다. 보통 통계의 역할이 평균을 구하거나 히스토그램과 같은 도구로 자료를 이해할 수 있도록 도와주는 것으로 생각하는 경우가 많지요. 하지만 통계는 방대한 데이터의 특징을 추출하고, 데이터 안에 담긴 수학적 규칙을 발견함으로써 현상을 이해하고, 판단의 근거를 만드는 다양한 역할을 담당하고 있는 강력한 도구입니다.

특히 인공지능의 영역에서 통계의 역할은 더욱 두드러집니다. 인공지능이 학습할 데이터를 준비하는 과정에서, 데이터로부터 규칙을 찾아 모델을 만들어가는 과정에서, 그리고 학습한 패턴을 바탕으로 새로운 결과물을 생성해내는 과정에서 통계는 빠짐없이 등장합니다. 먼저 이 글에서는 인공지능의 학습 재료가 되는 데이터를 준비하는 과정에서 통계의 역할을 중심으로 살펴보려고 합니다.

인공지능과 데이터

그렇다면 인공지능은 데이터를 어떻게 활용할까요? 우리가 일상에서 접하는 인공지능 서비스의 대부분은 기계학습(머신러닝)이라는 방법으로 만들어집니다. 기계학습은 수많은 데이터를 학습한 다음 어떤 값을 예측하거나, 특정한 기준에 따라 값을 분류하기도 합니다. 때로는 사용자의 사용 패턴이 담긴 데이터를 통해 콘텐츠를 추천하기도 하고, 사용자의 요청에 따라 대화를 주고받기도 하는데 이러한 인공지능 모델을 구현하기 위해서는 많은 양의 데이터를 학습하는 과정이 반드시 필요합니다.

카페를 운영하는 사장님의 이야기를 예로 들어보겠습니다. 카페 사장님은 매일의 매출을 기록하면서 매출에 영향을 주는 요인들은 무엇이 있으며, 이를 바탕으로 매출을 예측할 수 있을지 궁금해졌습니다. 그래서 날짜, 요일, 날씨, 기온을 비롯해 주변 행사 여부 등과 같은 조건들을 함께 기록하기 시작했습니다. 아래 표는 카페 사장님이 일주일간 기록한 예시 데이터입니다.[2]

1) https://keywordseverywhere.com/blog/instagram-stats/?utm_source=chatgpt.com
2) 데이터를 일반화하기에는 충분하지 않지만 개념을 익히는 용도로 활용할 예정입니다.

날짜	요일	날씨	기온(℃)	습도	방문객 수	주변 행사	일 매출(만 원)
6/1	월	맑음	27	60	?	×	52
6/2	화	맑음	25	52	90	×	56
6/3	수	흐림	22	60	68	×	40
6/4	목	비	18	70	60	×	38
6/5	금	맑음	26	30	85	축제	42
6/6	토	맑음	27	50	530	축제	58
6/7	일	흐림	24	20	90	축제	50
6/8	월	맑음	26	70	84	×	55
6/9	화	비	17	85	55	×	35
6/10	수	맑음	24	70	92	×	57

데이터를 잘 들여다보면 카페의 일상을 읽어낼 수 있습니다. 주중과 주말은 사람들의 일상이 달라지므로 카페의 방문이나 매출에 변화가 있을 수도 있고, 날씨와 기온에 많은 영향을 받기도 합니다. 주변에서 축제와 같은 행사가 열린 금요일과 토요일에는 매출에 얼마나 영향을 줄까요? 이 표에서 요일, 날씨, 기온, 방문객 수 등 데이터를 설명하는 각 열에 해당하는 항목을 '속성'이라고 부르고, 카페 사장님이 궁극적으로 예측하고 싶은 속성인 일 매출을 '레이블'이라고 부릅니다. 어떤 속성들이 레이블에 얼마나 영향을 주는지 우리는 통계를 통해 살펴볼 수 있지요.

데이터의 정돈
학습할 수 있는 데이터로 만들기

그런데 앞의 데이터를 다시 한번 살펴보면 몇 가지 문제가 눈에 들어옵니다. 먼저 6월 1일 월요일의 방문객 수가 비어 있습니다. 기록이 누락된 것일 수도 있고, 단순한 실수일 수도 있는데, 이렇게 비어 있는 값을 결측값(missing value)이라고 합니다. 반면 6월 6일 토요일의 방문객 수는 530명으로, 다른 날과 비교해 지나치게 높은 값을 보이고 있죠? 매출이 50만 원인 것에 비해 방문객이 530명이라는 것은 입력 오류일 가능성이 높습니다. 이렇게 전체 흐름에서 크게 벗어나는 값을 이상값(outlier)이라고 합니다.

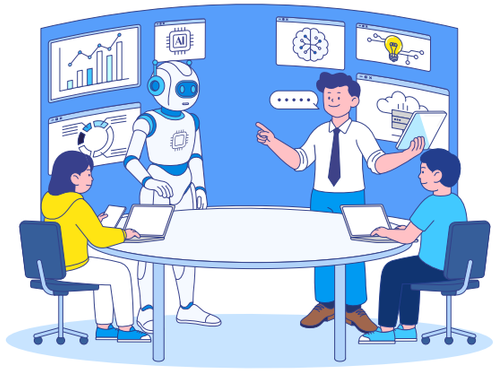
결측값과 이상값이 포함된 데이터를 그대로 인공지능의 학습에 사용하면 어떻게 될까요? 인공지능은 데이터에서 패턴을 찾아 학습하기 때문에, 만약 데이터에 이상한 값이 포함되어 있으면 엉뚱한 패턴을 학습하게 됩니다. 가령 530명이라는 이상값을 그대로 두면 인공지능은 "토요일에는 방문객이 500명이 넘을 수 있다"는 엉뚱한 규칙을 만들어낼 수도 있지요. 이때 통계적 방법을 활용하여 이상값, 결측값을 처리해야 합니다. 전체적인 경향성을 해치지 않는 정도라면 해당 데이터를 삭제하거나 통계를 활용해 평균이나 중앙값, 혹은 가장 유사한 조건의 값으로 대체합니다. 예를 들어 6월 1일은 맑고 기온이 27도이므로 비슷한 6월 7일의 데이터를 참고할 수 있겠죠. 또한 6월 6일은 다음 날인 7일의 데이터와 기온을 참고해 100명 정도로 데이터를 고칠 수 있습니다. 이때 정확한 값을 맞히기는 어렵기 때문에, 전체적인 경향성을 해치지 않는 범위에서 이상값, 결측치를 처리하는 것이 중요합니다.

이러한 과정을 데이터 전처리라고 합니다. 실제 인공지능 프로젝트에서 데이터 전처리가 차지하는 비중은 전체 작업 시간의 60%에서 80%에 달한다고 알려져 있습니다. 'Garbage In, Garbage Out', 즉 쓰레기를 넣으면 쓰레기가 나온다는 말처럼 데이터의 질이 인공지능의 성능을 좌우할 만큼 중요하기 때문입니다.

핵심 속성 추출
무엇이 매출을 좌우하는가

날짜	요일	날씨	기온(℃)	습도	방문객 수	주변 행사	일 매출(만 원)
6/1	월	맑음	27	60	88	0	52
6/2	화	맑음	25	52	82	0	56
6/3	수	흐림	22	60	68	0	40
6/4	목	비	18	70	60	0	38
6/5	금	맑음	26	30	85	1	42
6/6	토	맑음	27	50	100	1	58
6/7	일	흐림	24	20	90	1	50
6/8	월	맑음	26	70	84	0	55
6/9	화	비	17	85	55	0	35
6/10	수	맑음	24	70	92	0	57

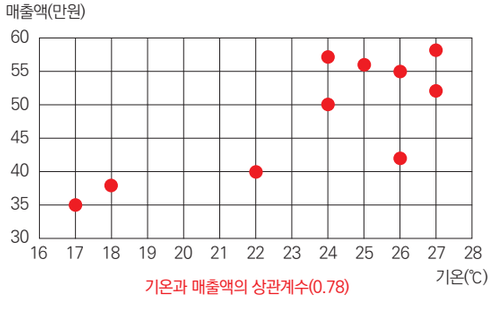
위의 표는 결측값과 이상값을 처리한 후의 데이터입니다. 데이터가 정돈되었다면, 이제 어떤 속성이 매출에 실제로 영향을 미치는지 살펴볼 차례입니다. 데이터에 포함된 모든 속성을 인공지능의 학습에 그대로 사용하면 좋을 것 같지만, 실제로는 그렇지 않습니다. 매출과 큰 연관성이 없는 속성까지 포함하면 오히려 예측을 왜곡하거나, 불필요한 속성이 억지로 반영되어 엉뚱한 모델이 만들어지기도 합니다. 그래서 레이블인 매출과 각 속성 사이에 어떠한 상관관계가 얼마나 있는지 살펴봄으로써, 학습에 필요한 속성이 무엇인지 판단할 필요가 있습니다.



두 속성 사이의 관계를 수치로 나타내는 대표적인 방법이 상관계수입니다. 상관계수는 -1에서 1 사이의 값을 가지며, 1에 가까울수록 한 속성이 증가할 때 다른 속성도 함께 증가하는 강한 관계를, -1에 가까울수록 한 속성이 증가할 때 다른 속성이 감소하는 강한 관계를, 0에 가까울수록 두 속성 사이에 뚜렷한 관계가 없음을 의미합니다. 이러한 속성 간의 상관관계는 시각화 자료를 통 해 쉽게 파악할 수 있습니다. 아래는 전처리를 거친 데이터 의 기온과 습도에 따른 매출액의 산점도입니다. 전처리 후 데이터에서

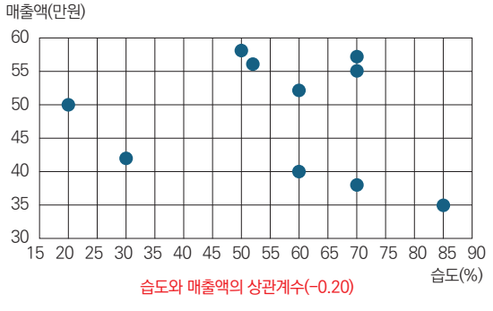
기온과 매출의 상관계수는 0.78로, 더울수록 매출이 올라가는 경향성을 보입니다. 그러나 습도의 경우 상관계수가 -0.20으로 뚜렷한 상관관계를 보이지 않지요. 따라서 매출액을 예측할 때는 습도보다는 기온이 더 유용한 요인이 됨을 알 수 있습니다.

기온에 따른 일 매출 산점도



기온과 매출액의 상관계수(0.78)

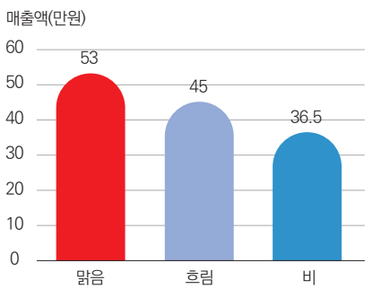
습도에 따른 매출액 산점도



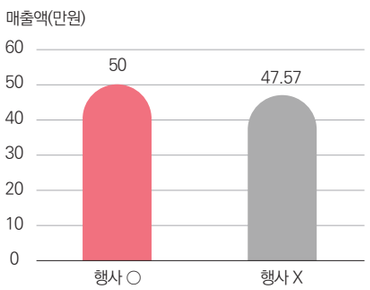
습도와 매출액의 상관계수(-0.20)

날씨나 주변 행사 등과 같은 범주형 데이터로 이루어진 요인들도 매출에 영향을 줄 수 있습니다. 날씨가 맑을수록 매출이 높은 경향을 보이며, 비가 올수록 매출이 낮은 경향을 보입니다. 그 이유로는 비가 올 때 외출하는 인원이 줄어들어 카페를 찾는 인원이 줄어들 수 있기 때문입니다. 반면 위의 기간 중 생겼던 행사 유무는 카페 매출에 큰 영향이 없어 보입니다. 행사에서 자체적으로 음료를 판매할 수도 있고, 행사의 주요 대상이나 연령대, 인원 등에 따라 영향이 없었던 것으로 원인 해석이 가능합니다.

날씨에 따른 평균 매출



행사 유무에 따른 평균 매출



이처럼 상관계수나 다양한 시각화 도구를 활용하면, 수많은 속성 중에서 매출 예측에 실제로 도움이 되는 속성을 찾아낼 수 있는데, 이 과정을 핵심 속성 추출이라고 합니다. 카페 사장님의 데이터에서는 요일, 날씨, 기온 등이 매출을 예측하는 데 핵심적인 속성이 되고, 날짜, 습도, 주변 행사 등은 상대적으로 덜 중요한 속성이 되는 셈입니다. 인공지능 학습에서 핵심 속성을 잘 선별하는 것은 마치 시험 공부를 할 때 핵심 내용만 정리하는 것과 비슷합니다. 교과서 전체를 무작정 외우는 것보다 중요한 부분을 골라 집중하는 편이 훨씬 효과적인 것과 같은 원리입니다.

다음 이야기
데이터에서 모델로

이렇게 데이터를 정제하고 핵심 속성을 선별하는 과정을 거치면, 카페 사장님의 데이터는 비로소 인공지능이 학습할 수 있는 상태가 됩니다. 이제 다음 단계는 이 데이터를 바탕으로 '모델'을 만드는 일입니다. 모델이란 데이터 속에 담긴 패턴을 수학적으로 표현한 것으로, 예를 들어 '기온이 27도이고 날씨가 맑음이면 매출액은 이 정도가 된다.'와 같은 규칙을 수식으로 나타낸 것이라고 할 수 있습니다.
통계에서 오래전부터 사용해온 회귀 모델이나 분포 모델이 바로 이러한 예이며, 인공지능은 이 원리를 더욱 정교하게 발전시킨 것입니다. 다음 글에서는 인공지능이 데이터로부터 어떻게 모델을 만들고, 오차를 줄여가며 스스로 판단 기준을 찾아가는지, 그 과정에서 수학과 통계가 어떤 역할을 하는지 살펴보겠습니다.

기업이 원하는 AI 인재, 교육은 준비되어 있는가
- 국내 AI 교육의 현주소와 데이터 활용 사례

진회승 | 소프트웨어정책연구소 책임연구원

1 기업이 필요로 하는 AI 전문가

생성형 AI 기술이 등장한 이후, AI는 금융·자동차·의료·법률 등 개별 산업을 넘어 경제·사회 구조 전반에 걸쳐 광범위한 변화를 일으키고 있다. 이러한 변화에 따라 AI 인재 수요의 양상도 달라지고 있다. 기업은 AI 모델을 개발하는 기술 전문가보다 AI를 실제 업무에 적용하는 실무 인재를 더 필요로 하며, 신입사원 채용 시에도 학벌이나 AI 기술 자체보다 실무 프로젝트 경험을 중요한 기준으로 삼고 있다. 구체적으로 데이터 분석가, AI 모델 기반 소프트웨어 개발자, AI 모델 적용 전문가에 대한 수요가 증가[1]하고 있으며, 이러한 흐름에 맞추어 AI 교육과정에도 실무 프로젝트 수행, 데이터 분석 및 처리, AI 모델 활용 등의 내용을 포함할 것이 요구되고 있다.

이러한 산업계의 수요에 대응하기 위해 정부와 교육계는 초·중등 및 대학의 재학생, 구직자, 재직자 등 다양한 학습자를 대상으로 AI 인재 양성 체계를 빠르게 정비하고 있다. 여기서 AI 인재는 AI 산업에 종사하는 AI 핵심인재·활용인재와, 제조·금융·의료 등 타 산업에 AI를 접목하여 활용하는 AI 융합인재로 구분된다. AI 기술의 여러 구성 요소 중에서도 데이터의 수집·정제·분석은 실무에서 중요한 비중을 차지하며, 이는 통계 분야의 전통적 업무와 본질을 공유한다. 따라서 AI 인재는 유형을 막론하고 데이터를 다루는 역량을 일정 수준 이상 갖추어야 한다. 산업계는 AI 인재의 부족 현상을 단순한 양적 문제보다 질적 문제로 인식하고 있어, 현행 교육 체계가 기업이 실제로 요구하는 실무 역량을 길러내고 있는지 점검이 필요한 시점이다. 이에 본고에서는 먼저 초·중등교육, 고등교육, 직업교육의 각 단계별 AI 교육 현황을 점검하고, 이를 통해 도출되는 핵심 과제를 바탕으로 교육 현장에서 AI와 데이터가 실제로 어떻게 활용되고 있는지 그 구체적 사례와 한계를 검토하고자 한다.

2 AI 교육 체계 점검

2.1. 초·중등 교육

초·중등 교육에서 AI 교육은 모든 학생을 위한 AI 리터러시 함양과 잠재적 AI 핵심 인재의 조기 발굴이라는 두 가지 방향을 축으로 한다. 교육부는 디지털·AI 시대에 필요한 미래 역량을 기르기 위해 2022년 교육과정을 개편하고, 모든 교과에서 AI 등 신기술 분야의 학습을 내실화하였다.[2]

먼저 AI 리터러시 교육은 AI 기술을 제대로 이해하고 윤리적으로 활용하며, 일상과 학습의 도구로 사용할 수 있는 역량을 기르는 것을 목표로 한다. AI 기술을 맹목적으로 수용하는 것이 아니라 제대로 활용하기 위해서는 AI의 작동 원리에 대한 기초 지식이 전제되어야 한다. 환각 현상, 학습 데이터의 편향, 딥페이크 등 AI가 야기할 수 있는 문제는 모두 AI의 구조적 특성에서 비롯되므로, 그 원리를 이해해야 AI가 생성하는 정보를 비판적으로 수용하고 올바르게 활용할 수 있기 때문이다. 보편적 AI 소양 교육과 병행하여, 정부는 'AI·디지털 영재교육원'[3] 운영, AI 특화 교육과정 편성, 학생 동아리 지원 등을 통해 잠재적 AI 인재를 조기에 발굴하고 심화 학습 경로를 제공하는 제도적 기반도 마련하고 있다.

1) 2026 기업 AI 인재 수요 트렌드 리포트, 2026, 한국표준협회
2) 교육부(2021.11.24.). 「2022 개정 교육과정 총론 주요사항」

2.2. 고등교육

대학과 대학원의 AI 교육은 산업 현장이 요구하는 AI 실무 인재 양성과 국가 경쟁력 확보를 위한 AI 최고급 인재 양성이라는 두 가지 목표를 동시에 추구한다.

먼저, 대학은 AI 전공자 대상 심화 교육과 전교생 대상 AI 융합 교육을 병행하고 있다. AI 기술이 고도화될수록 컴퓨터 과학의 근본 원리에 대한 이해가 중요해지므로, 이산수학, 컴퓨터 프로그래밍, 자료구조, 알고리즘 등을 전공 필수 과목으로 편성하여 기초 핵심 역량을 강화하고 있다. 이와 함께 실무 중심 커리큘럼과 산학협력 프로젝트, 인턴십 등을 통해 산업 현장에 즉시 투입 가능한 실무 전문가를 길러내고 있다.

나아가 AI 최고 인재 양성을 위해서는 대학원 수준의 집중적 지원이 필수적이다. 이에 따라 'AI 대학원'과 'AI융합혁신대학원'을 신설[4]하여 석·박사급 핵심 인재 양성을 뒷받침하고 있다. 'AI 대학원'은 핵심 기술 연구에 중점을 둔 엘리트 연구 중심 대학원으로서, 특화된 커리큘럼과 전임 교원 확충, 연구 인프라를 바탕으로 산학협력을 선도하는 거점 역할을 수행하며, 실제 'KAIST-Google 파트너십'이나 '서울대-네이버 초대규모 AI 연구센터' 등을 운영하고 있다. 'AI융합혁신대학원'은 AI 핵심 기술을 제조·금융·바이오 등 타 산업 분야의 데이터와 결합함으로써 산업계가 요구하는 AI 융합 전문가를 양성하는 역할을 담당한다.

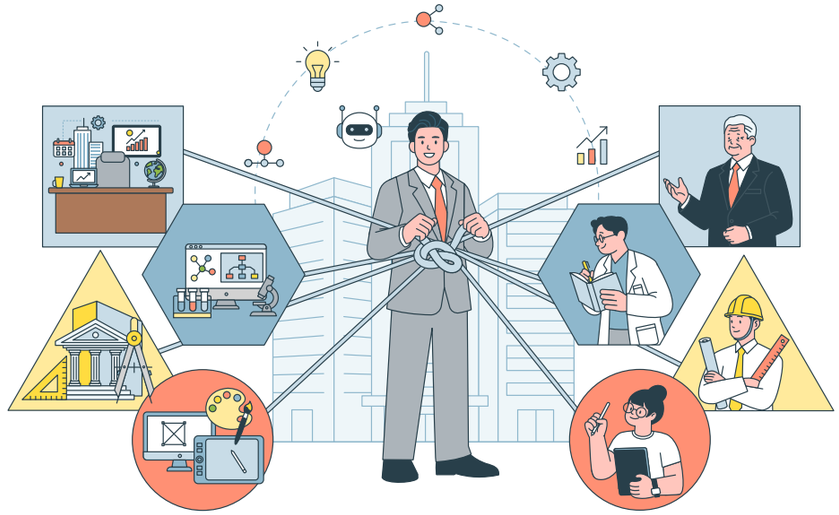
2.3. 직업교육

정부 주도 직업 교육은 단기간에 실무 투입이 가능한 AI·SW 인력을 대규모로 공급하여 인력 저변을 확대해 왔으며, 특히 비전공자에게 이 분야로 진입할 수 있는 효과적인 경로를 제공해 왔다. 이러한 성과가 가능했던 것은 실제 데이터를 활용한 프로젝트 기반 교육 방식과, 이를 채용으로 연결하는 민관협력 구조라는 두 가지 축 덕분이다.

교육 방식 측면에서는 강의식 수업에서 벗어나, 학습을 위한 데이터를 직접 수집·분석·활용하는 자기주도 프로젝트와 동료 학습을 교육의 중심에 두고 있다. 학습자는 정제된 예시 데이터가 아닌 실제 데이터를 다루는 과정에서 기업이 요구하는 문제 해결력과 실전 역량을 체득하게 되며, 이를 현장 전문가의 멘토링이 뒷받침한다. 채용 연계 측면에서는 기업 수요에 기반하여 교육과정을 설계하고, 수료생에게 수요 기업 추천, 면접 기회 제공, 인턴십 연계 등 실질적인 취업 지원을 병행함으로써 교육과 채용을 유기적으로 잇는 민관협력 모델을 구축하였고, 이를 통해 청년 고용 활성화에도 기여하고 있다. 과학기술정보통신부의 '채용 연계형 SW 전문 인재양성'[5]과 고용노동부의 'K-디지털 트레이닝(KDT) 사업'이 이러한 모델의 대표적인 사례에 해당한다.

나아가 정부 주도 교육은 AI·SW 산업 내 인력 양성에 그치지 않고, 산업별 도메인 데이터를 AI와 결합할 수 있는 융합인재 양성에도 주력하고 있다. 융합인재 양성은 AI 기술을 보유한 인재가 특정 산업의 도메인 지식과 데이터를 확보하는 '기술 주도형'과, 산업 현장에서 축적된 데이터를 기반으로 실제 문제를 해결하는 '현장 문제해결형'의 두 축으로 추진되고 있으며, 학위과정, 비학위 과정, 공동 연구, 수요·공급 매칭 등 다양한 형태로 운영되고 있다. 기술 주도형 인재 양성은 AI 산업 내에서의 융합인재 양성을 목표로 하며, '이노베이션아카데미', '인공지능산업융합기술개발' 등이 이에 해당한다.

3) 교육부(2024.12.). 「제2차 정보교육 종합계획(안) (2025년 ~ 2029년)」에서 영재 발굴, 재능, 진로 등 유형별 맞춤 학습, 우수 성과물 관리 등의 AI·디지털 영재교육원 운영(안)을 발표
4) 인공지능대학원협의회, https://aigs.kr/default/company/company_01.php?topmenu=1&left=1
5) 채용연계형 SW전문인재 양성사업은 멤버십기업 맞춤형 교육과정을 설계하고 교육생 운영, 프로젝트 운영, 멘토 및 현업 특강 등을 통해 현장에 즉시 투입 가능한 중·고급 인재를 양성하는 사업



현장 문제해결형 사업은 AI 활용 산업에서 도메인 데이터 기반의 AI 융합인재를 양성하는 데 초점을 두며, 'ICT 글로벌 전문 융합 인재양성', '의료인공지능 특화 융합인재 양성사업', '산업맞춤형 혁신바우처 지원' 등이 대표적이다.

한편, 기업들은 자사의 기업 문화와 기업이 요구하는 기술 스택에 최적화된 인재를 선제적으로 확보하기 위해 기업주도 교육 프로그램을 직접 운영하고 있으며, 민간 AI 부트캠프 등은 단기 집중 교육을 통해 재직자들의 AI 활용 능력 향상을 지원하고 있다. 기업주도 교육은 취업 준비생에게 6개월에서 1년간 집중 코딩 교육과 실전 프로젝트를 경험하게 하여 취업률 등에서 높은 성과를 나타내고 있으며, '삼성청년SW아카데미', '네이버 부스트캠프', 'KT 에이블스쿨' 등이 대표적이다.

민간 교육기관은 대학의 경직된 커리큘럼과 달리 신기술 교육과정을 빠르게 개설하고, 시장 수요가 높은 분야의 교육을 민첩하게 제공한다는 강점을 지닌다. 현직 개발자와 엔지니어를 강사진으로 확보하여 실무 중심으로 강의를 구성하고, 교육과정 또한 짧은 주기로 갱신하고 있다.

2.4. 대안교육[6]

AI 기술이 급속도로 발전하는 오늘날, 기존의 정규교육이나 정부·기업 주도의 교육만으로는 최신 기술에 대한 학습 수요와 산업계의 인재 양성 요구를 충분히 충족하기 어렵다. 이에 따라 대안적 교육의 중요성이 점차 커지고 있다. 현행 AI 교육 체계는 정규교육과 정부·기업이 주도하는 직업교육 등의 경로를 구축하고 있으나, 빠르게 변화하는 AI 기술에 대응하기에는 구조적 한계가 존재한다. 정규교육 체계는 깊이 있는 이론을 제공하지만 커리큘럼 개편, 교수자 양성, 인프라 확보 등에 오랜 시간이 소요되며, 교육에 적용까지의 기간이 길다. 또한 부정행위, 표절, 개인정보 및 데이터 보안, 지적재산권

6) 대안교육기관에 관한 법률(2025.7.22.)에서는 대안교육이란 개인적 특성과 필요에 맞는 다양한 교육내용 및 교육방법을 통하여 개개인의 소질과 적성 개발을 목적으로 하는 학습자 중심의 교육이라고 정의

침해 등 해결해야 할 쟁점이 산재해 있어 교육 현장에서 AI 활용에 대한 지침이 아직 정립되지 못하고 있으며, 이에 따라 AI 교육의 확산 또한 지체되고 있다. 나아가, 공급자 주도 교육시스템은 AI 기술 시대에 빠르게 변화하는 다양한 전문 분야와 직무에 대해 맞춤형 진로 경로를 포괄하기 어렵다는 근본적 한계를 안고 있다.

이러한 한계로 인해 정규 교육기관과 기업이 교육을 기획하여 일괄적으로 제공하는 방식을 넘어, AI 역량 확보 방안으로서 개인 주도적인 교육이 확대되고 있다. AI 시대에는 자신에게 맞는 학습 경로를 스스로 설계하고 조정할 수 있는 개인 주도 학습 역량이 중요시되고 있으며, 실제로 AI·SW 분야는 전통적인 학사·석사 학위 없이 전문성을 획득하여 실무에 투입되고 성공적인 성과를 내는 사례가 가장 많은 분야이다. 대안적 교육 경로는 비전공자, 경력전환자, 재직자 등 다양한 학습자에게 시간과 내용 면에서 더 유연한 선택지를 제공한다는 점에서 그 의의가 크다.

개인 주도의 학습 경로는 소극적으로는 부트캠프 등을 선택하여 활용하는 방식이 있고, 적극적으로는 커뮤니티 기반 학습, 동영상, 논문 등을 스스로 선별하여 학습하는 방식이 있다. 커뮤니티 기반 학습은 빠르게 변화하는 기술을 적용하여 문제 해결을 가속화하고 기술 트렌드를 습득하는 것은 물론, 학습자의 동기 부여와 자기효능감 강화에도 매우 중요한 활동이다. '페이스북 AGI KR', 대덕연구단지를 기반으로 시작한 'AIFrenz' 등이 대표적인 사례이다. 또한 오픈소스와 공개데이터 등을 활용하여 논문의 알고리즘을 직접 구현해 보는 자기주도적 학습 역시 주요한 대안 교육 사례로 주목받고 있다.

2.5. AI 교육체계 시사점

각 교육 단계의 AI 교육 현황을 종합하면, 단계별로 고유한 과제와 방향성이 존재한다. 본고에서는 이 중에서 기업이 요구하는 실무 인재 양성이라는 관점에 초점을 맞추어, 각 단계에서 공통적으로 요구되는 조건을 도출하고자 한다. 다만, 각 단계는 그 고유한 목표를 가지고 있으므로, 본고는 이를 일률적으로 실무 인재 양성의 틀로 환원하기보다, 각 단계가 지향하는 본연의 목표를 존중하면서 그 안에서 데이터 활용이라는 공통 축이 어떻게 작동하는지를 살펴본다.

초·중등 교육은 실무 인재 양성보다는 AI 리터러시 함양에 중점을 두고 있다. 2022 개정 교육과정은 이를 명확히 제시하고 있으며, AI 리터러시는 AI에 대한 지식을 암기하는 것이 아니라 AI의 작동 원리를 이해하고 그 결과를 비판적으로 수용할 수 있는 능력이라는 점에서, 학생들이 데이터를 직접 다루는 경험 없이는 길러지기 어렵다. AI의 환각이나 편향 문제는 교과서적 설명만으로는 충분히 전달되지 않는다. 학습자가 데이터를 직접 가공·분석하고 이를 활용해 AI를 다뤄보는 경험을 거쳐야 비로소 AI 기술을 비판적으로 이해할 수 있다. 이를 위해 충분한 교육 시수 확보, AI 전문 교원 양성, 실습 인프라 확대 등 운영 측면의 보완이 시급하다.

고등교육은 자본과 인프라, 특히 산업 현장 데이터에 대한 접근성의 격차로 인해 실무형 AI 인재 양성 기능이 점차 기업으로 이전되고 있다는 구조적 한계에 직면해 있다. 대학의 산학협력 프로젝트와 인턴십만으로는 기업이 보유한 방대한 현장 데이터와 실전 경험을 충분히 제공하기 어렵고, 그 결과 이론과 실무 사이의 괴리가 심화되고 있다. 대학이 실무 인재 양성 기능을 지속하려면 현장 데이터에 대한 접근성 확보가 관건이며, 현재의 산학협력 수준을 넘어 기업 데이터 공동 활용 체계를 제도화하고 학생들이 교육 과정 안에서 산업별 현장 데이터를 직접 다루며 실제 문제를 해결해 보는 환경을 마련해야 한다. 특히 AI 융합혁신대학원이 지향하는 도메인 데이터 기반 문제 해결 교육이 실효성을 거두려면, 특정 산업과의 장기적 데이터 협력 관계 구축이 뒷받침되어야 한다.

직업교육은 기업 현장으로의 실무 투입을 직접적 목표로 한다는 점에서, 실무 데이터의 확보가 교육 품질을 좌우하는 결정적 요인이다. 정부 주도 교육은 단기간에 실무 투입이 가능한 인력의 저변을 확대하는 데 기여해 왔으며, 기업 주도 교육과 민간 부트캠프가 높은 성과를 보이는 이유 또한 교육과정을 현장 데이터와 실전 프로젝트 중심으로 설계하고 있기 때문이다. 이는 정부·기업·교육기관 간의 유기적 협력을 통해 교육 내용과 기업이 요구하는 역량 간의 격차를 줄이는 것이 중요함을 시사한다.

대안교육은 정규교육이 포괄하기 어려운 다양한 진로와 빠르게 변화하는 기술 트렌드에 유연하게 대응한다는 점에서 고유한 의의를 지닌다. 이를 활성화하기 위해서는 개인 주도의 비형식 학습을 제도적으로 인정하고, 커뮤니티 기반의 능동적 학습 생태계를 지원할 필요가 있다. 공급자 위주의 교육 체계에서 벗어나 수요자 중심의 지원을 확대하되, 이 역시 학습자가 실제 데이터를 기반으로 문제를 해결하는 경험을 축적할 수 있는 환경이 전제되어야 한다.

결국, AI 교육체계의 각 단계는 고유한 목표를 지니지만, 어떤 단계에서도 AI 기술을 제대로 익히고 이를 실무에 적용하기 위해서는 학습자가 실제 데이터를 다루는 경험이 필수적이다. 특히 기업이 요구하는 실무 역량을 길러내기 위한 핵심 조건은 학습자가 실제 데이터를 다루며 AI를 적용하는 프로젝트형 교육 경험에 있으며, 이에 대한 수요는 갈수록 높아지고 있다. 그렇다면 이러한 데이터 기반 AI 교육이 교육 현장에서 실제로 어떤 형태로 구현되고 있으며, 어떤 성과를 내고 있고, 또 어떤 한계에 부딪히고 있는가? 이하에서는 교육 현장에서의 AI·데이터 활용 사례를 구체적으로 살펴봄으로써 이 질문에 답하고자 한다.

3 교육 현장에서의 AI, 데이터 활용 사례

3.1. 정부·지자체 주도 사례 : 청년취업사관학교(SeSAC)[7] 빅테크 전담 캠퍼스

서울시의 대표적인 청년 AI 인재 양성기관인 청년취업사관학교(SeSAC, Seoul Software Academy)는 그 이름처럼 AI 개발자를 꿈꾸는 청년들의 가능성을 발굴하고, 체계적인 교육과 산업 데이터를 활용한 실무 프로젝트를 통해 산업 현장에 투입 가능한 인재로 길러내는 교육 프로그램이다.

이 프로그램의 핵심은 산업 도메인별 실무 데이터를 교육의 중심에 배치한 프로젝트형 커리큘럼에 있다. 동작 캠퍼스에서는 오라클과 협력하여 의료 데이터의 구조와 특성을 이해하고 이를 기반으로 AI 서비스를 설계하는 의료·바이오 전문가를 양성하며, 서대문 캠퍼스에서는 엔비디아와 협력하여 산업 현장의 대규모 데이터를 활용한 생성형 AI 및 딥러닝 모델 개발 역량을 갖춘 전문가를 양성한다. 각 과정은 4~5개월간 기초 이론부터 실무 프로젝트까지 체계적인 커리큘럼으로 구성되며, 글로벌 빅테크의 현업자 및 인증 강사가 직접 교육에 참여한다. 이러한 기업 연계 교육의 실효성은 성과로도 확인되고 있는데, 2025년 기준 빅테크 특화캠퍼스의 취업률은 86%에 이른다.[8]

SeSAC 사례는 정부·지자체가 글로벌 기업과의 전략적 파트너십을 통해 교육 내용과 기업 수요 간의 격차를 줄이고, 실무 데이터 기반 프로젝트 교육과 채용 연계를 결합한 모델이라는 점에서 의의가 있다. 다만, 대규모 확산 과정에서 캠퍼스별 교육 품질의 균질성 확보, 빅테크 이외 산업 분야로의 확대, 그리고 지방 소재 청년의 접근성 제고 등의 과제가 남아 있다.

7) 청년취업사관학교(SeSAC) 홈페이지, https://sesac.seoul.kr/sesac/main/main.do
8) 서울특별시(2026.4.13.), 엔비디아·MS와 AI 인재 키운다! 빅테크 캠퍼스 교육생 모집, https://mediahub.seoul.go.kr/archives/2017765

3.2. 기업 주도 사례 : 카카오테크 부트캠프 AI 실무 개발 과정[9)]

오늘날 인공지능 기술은 단순히 모델을 만드는 데 그치지 않고, 사용자에게 서비스로 제공되면서 지속적으로 개선·운영되는 형태로 발전하고 있다. 이러한 AI 서비스가 제대로 작동하려면 알고리즘, 모델 설계, 컴퓨팅 환경, 운영 관리 등 여러 요소가 함께 맞물려야 하며, 그중에서도 데이터를 수집하고, 정리하고, 분석하는 작업이 상당한 비중을 차지한다. 아무리 뛰어난 알고리즘이라도 이를 이해하는 것만으로는 실제 업무의 요구를 충족시키기 어려우며, 실무자는 데이터를 직접 다룰 수 있는 능력을 함께 갖추어야 한다. 최근 확산되고 있는 대규모 언어 모델(LLM)이나 AI 에이전트처럼 여러 AI 모델이 복합적으로 연결된 기술이 등장하면서, 다양한 데이터를 기반으로 여러 AI 도구를 적절히 조합하여 안전하게 서비스에 연결하는 역량의 중요성은 더욱 커지고 있다.

카카오테크 부트캠프는 이러한 현업의 요구를 교육에 반영한 대표적인 사례이다. 이 과정은 단순한 AI 모델 이해에 그치지 않고, 데이터 수집부터 모델 학습, API 배포, 프런트엔드 연동까지 AI 서비스 개발의 전 과정을 실전 프로젝트를 통해 경험하도록 설계되어 있다. 컴퓨터 비전, 음성 등 산업 현장의 실제 데이터를 활용한 프로젝트를 다수 수행한다. 수료생들은 체계적 커리큘럼과 실습 중심 환경을 통해 실력과 협업 능력을 동시에 향상시킬 수 있었다는 평가를 내놓고 있다.

이 사례는 기업이 자사의 기술 스택과 실무 환경에 최적화된 교육을 직접 설계·운영함으로써 이론과 현장 사이의 괴리를 최소화하고 있다는 점에서 주목할 만하다. 기업 주도 교육이 높은 성과를 거두는 근본적인 이유 역시 현장 데이터와 이를 활용한 실전 프로젝트가 교육과정 전반의 토대를 이루고 있기 때문이며, 이는 앞서 도출한 핵심 과제, 즉 실제 데이터 기반 프로젝트형 교육의 확대를 기업이 자체적으로 실현하고 있는 사례라 할 수 있다.

3.3. 시사점

두 사례를 종합하면, 교육 현장에서 AI·데이터 활용이 실질적인 성과를 내고 있는 곳에는 공통적인 특징이 있다. 첫째, 교육 과정이 기업 현업의 업무 맥락과 최대한 가깝게 설계되어 있다. 완전히 교육용으로 정제된 환경이 아니라, 기업이 실제로 해결하고자 하는 문제를 소재로 삼고, 실제 업무 환경을 참고한 데이터를 활용함으로써 학습자는 데이터에 누락이나 오류가 있는 상황, 분석에 앞서 데이터를 정리하고 가공하는 과정의 복잡함 등을 어느 정도 체험할 수 있다. 둘째, 현업 전문가의 멘토링과 기업·교육기관 간 협력이 교육의 품질을 뒷받침하고 있다. 셋째, 교육을 마친 뒤 곧바로 취업이나 실무 투입으로 연결되는 구조를 갖추고 있어 학습자의 동기를 유지하고 교육의 실효성을 높이고 있다.

그러나 과제도 존재한다. 가장 근본적인 한계는 교육 현장에서 활용할 수 있는 산업 데이터의 범위가 여전히 제한적이라는 점이다. 기업이 실제로 다루는 현장 데이터는 개인정보 보호, 데이터 보안, 지식재산권 등 법적·제도적 제약에 묶여 있어, 교육에서는 공개 데이터셋이나 가공된 샘플 데이터에 의존하는 경우가 많다. 그 결과 학습자가 진짜 현장에서 발생하는 데이터의 복잡성과 규모를 충분히 경험하지 못한 채 교육을 마치게 될 우려가 있다.

8) 서울특별시(2026.4.13.), 엔비디아·MS와 AI 인재 키운다! 빅테크 캠퍼스 교육생 모집, https://mediahub.seoul.go.kr/archives/2017765
9) 카카오테크 부트캠프 공식 사이트, https://kakaotechbootcamp.com/

4 결 론

본고는 기업이 필요로 하는 AI 인재상을 출발점으로 삼아, 초·중등교육, 고등교육, 직업교육, 대안교육의 각 단계별 AI 교육 현황을 점검하고, 이를 통해 실무 인재 양성의 핵심 조건으로 도출된 데이터 기반 프로젝트형 교육이 현장에서 어떻게 구현되고 있는지를 구체적 사례를 통해 살펴보았다.

교육체계 점검 결과, 산업계는 AI 인재 부족을 양적 문제보다 질적 문제로 인식하고 있었으며, 모든 교육 단계에서 실제 현장 데이터를 활용한 프로젝트형 교육의 확대가 공통 과제로 확인되었다. 교육 현장에서는 대학, 정부·지자체, 기업, 민간 교육기관 등 다양한 주체가 각자의 방식으로 데이터 기반 AI 교육을 추진하고 있으며, 산업 현장의 수요와 교육 내용을 연계하려는 노력이 점차 확대되는 추세이다.

그러나 AI 교육이 기업이 요구하는 실무 역량을 효과적으로 길러내기 위해서는 여전히 많은 과제가 남아 있다. 교육용 현장 데이터의 확보와 공유를 위한 법적·제도적 기반을 정비하고, 교육 성과를 체계적으로 측정·환류하며, 정부·기업·교육기관 간 유기적 협력을 강화해 나갈 필요가 있다. AI가 국가 경쟁력의 핵심이고 AI 인재가 산업의 성패를 좌우하고 있다. 이러한 현실에서 실무 인재 양성의 방향은 데이터 기반 AI 교육에 있으며, 이를 지속적으로 확대하기 위해서는 앞서 제시한 과제들을 적극적으로 풀어가야 할 것이다.

AI가 학습하는 데이터는 누구의 것인가

이청호 | 상명대학교 교수

최근 생성형 AI 서비스가 일상 깊숙이 들어오면서 학생들의 보고서, 직장인의 업무 메일, 공공기관의 민원 응대까지 AI를 거치지 않는 분야가 거의 없을 정도가 되었다. 그런데 이렇게 AI가 만들어 내는 풍요로운 결과물 이면에는 다음과 같이 쉽사리 정리되지 않는 질문이 하나 자리하고 있다. "AI가 학습한 방대한 데이터는 도대체 누구의 것인가?" AI 서비스가 일상화되면서 덩달아 수많은 데이터의 활용이 일상화되었고 이 질문에 대한 대답은 점차 더 대답하기 힘들어진다. 본 글에서는 이 질문과 관련하여 소위 '데이터 윤리(data ethics)'에 답하기 위해 점검해 보아야 할 몇 가지 지점들을 언급하고자 한다.

참고로 필자가 데이터 윤리 관련 강의나 논의에서 가장 자주 듣는 질문 중 하나는 "이미 다 공개된 데이터인데 AI가 학습에 자유롭게 사용해도 되는 것 아닌가요?"라는 질문이다. 이러한 질문은 데이터 윤리에 대한 우리의 일반적인 상식이 얼마나 올바른 기준과 거리가 있는지를 잘 드러낸다. 많은 이들은 공개되어 있다는 사실과 누군가가 마음대로 가져가서 사용해도 된다는 것이 분명히 다른 문제임을 제대로 인지하지 못한다. 흔히 말하는 것처럼, 거리에 사람이 많이 다닌다고 해서 누군가의 사진을 찍어 상업적으로 이용하는 것은 초상권을 침해하는 것이라는 것과 동일한 논리이다. 데이터가 '눈에 보인다'는 사실과 '가져갈 권리가 있다'는 사실은 결코 동일한 의미가 아니다.

AI는 인간이 남긴 흔적을 먹고 자란다

AI는 인간이 수십 년에 걸쳐 만들고 사용해 온 디지털화된 자료들, 텍스트, 그림, 영상, 음성, 검색 기록, 구매 흔적 등을 자양분 삼아 학습한다. 최신 대형 언어모델(LLM) 하나가 학습하는 텍스트의 양만 해도 적게 잡아 수조 단어에 이르는 것으로 여겨진다. 인간 한 사람이 평생 동안 다 읽지 못할 엄청난 분량의 정보가 모델 하나에 의해 빨려 들어가고 있는 셈이다.

어제 SNS에 올린 사진, 무심코 작성한 블로그 글 하나, 친구에게 보낸 이메일의 한 문장, 동네 식당에 남긴 별점 후기까지도 어느새 AI 모델에 흘러 들어가 그들의 학습 재료가 되었을지도 모른다. 데이터는 평범한 개인들이 모두 끊임없이 만들어 내고 흘려보내는 일종의 자원이며, 이러한 흐름은 너무 빠르고 너무 거대하여 자기가 '생성'한 데이터에 대한 '주권' 개념이 사실상 그 효력을 제대로 발휘하지 못하고 있다.

문제는 이러한 흐름이 한 번 몰아닥치게 되면 그 흐름을 거슬러 되돌릴 수 없다는 점이다. 종이에 적힌 글은 찢어 버리거나 태워버릴 수 있지만, 한 번 학습된 데이터는 모델의 어딘가에 '영원히' 새겨진다. 더 심각한 점은 한 사람이 적은 글 한 조각이 단순하고 무미건조한 데이터만은 아니라는 점이다. 그러한 글에는 적은 사람의 사고방식과 가치관이 스며들어 있다. 마찬가지로 사진 한 장에는 찍은 사람의 시선이 올곧이 담겨 있다.

데이터는 정량화된 정보의 영혼 없는 묶음이 아니라 한 사람의 삶과 인격의 흔적이 깊이 새겨진 의미들의 기본 단위이다. 그러므로 '데이터 소유권(주권)'에 대한 물음은 결국 한 사람의 '삶의 의미'를 묻는 것과 동일하다.

데이터 주권을 둘러싼 세계적 경향

데이터 주권을 둘러싼 국제 사회의 대응은 결코 한 방향으로 수렴되지는 않는다. 오히려 세계 각국은 각자의 정한 방향에 따라 서로 다른 길을 걷고 있다는 점을 주목할 필요가 있다.

데이터 주권에 대해 가장 강력한 입장을 옹호하는 주체는 유럽연합(EU)이다. 2024년 발효되어 2026년부터 본격 시행되는 EU의 AI법(AI Act)은 AI 모델 개발자에게 학습 데이터의 출처를 투명하게 공개하도록 '강제'한다. 창작자는 자신의 저작물이 AI 학습에 쓰이지 않도록 거부할 권리(opt-out)를 보장받으며, 기업은 이를 확인하고 준수했다는 증거를 기록으로 남겨야 한다. EU의 AI법의 이러한 노선은 GDPR의 '시민 중심 데이터 보호'의 전통이 AI 시대에도 그대로 연장된 것으로 볼 수 있다. 일부 기업들은 이러한 규제가 혁신의 발목을 잡을 수 있다고 주장하지만, EU는 '신뢰할 수 있는 AI'야말로 장기적인 관점에서 더 바람직한 가치를 창출할 것이라는 신념을 굽히지 않고 있다.

반면 미국은 2025년 초에 AI 정책 기조를 크게 전환했다. 당시 발의된 행정명령은 기존의 다소 폐쇄적인 윤리적 가이드라인보다 훨씬 야심찬 행보를 보인다. 미국은 AI 시장에서 선두를 유지하는 혁신과 더불어 국가 이익과 안보를 우선시하는 방향으로 선회하게 된다. 연방정부는 주(州) 정부가 과도한 AI 규제를 부과하지 못하도록 제한하고, AI 기술에 관한 기업의 자율성을 폭넓게 보장함으로써 기술 주도권을 유지하겠다는 셈법으로 보인다. 일종의 '느슨한 규제를 발판으로 하는 발전' 전략이라 할 수 있다. 미국의 이러한 노선은 글로벌 차원의 AI 경쟁에서

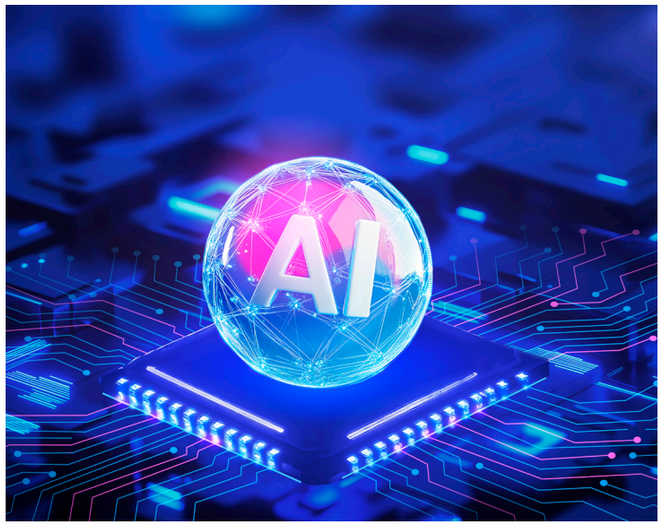
우위를 잃지 않으려는 전략적 선택이지만, 동시에 개인 데이터 보호에 악영향을 끼칠 수 있다는 적신호로 보인다. 실제로 뉴욕타임스가 자사의 기사를 무단으로 학습에 활용했다며 빅테크 기업을 상대로 제기한 소송 등의 사례를 본다면, 이러한 미국식 자율주의 노선이 종국에 맞이하게 될 파국으로 치닫는 균열의 시작을 단적으로 보여 준다.

일본은 또 다른 독자적인 방향을 택했다. 일본의 저작권법은 단순히 '향유'하기 위한 목적으로만 활용하는 것뿐 아니라, AI 학습을 위해 저작물을 이용하는 것을 폭넓게 허용하는 소위 '실용적 유연성(pragmatic flexibility)' 노선을 택하였다. 다만 동시에 일본 문화청은 2024년 발표한 가이드라인을 통해 AI가 만들어 낸 출력물이 기존 저작물과 유사할 경우 침해로 간주하는 기준을 명확히 함으로써 창작자의 권리도 균형 있게 보호하려 하고 있다.

이러한 일본의 접근은 'AI에 의한 학습은 자유롭게, 그러나 AI의 결과물은 엄격하게' 제어하려는 방향으로 여겨진다.

이렇듯 동일한 문제를 두고 세계 각국은 '강한 규제 중심', '시장 우선주의', '실용적 절충'의 세 방향으로 접근하고 있다. 그렇다면 우리나라는 어디쯤에 위치해 있을까. 한국은 EU 수준의 강력한 미국 정도의 자유로운 시장중심 전략, 일본의 실용적 절충의 측면 모두를 고려하고 있는 상황이었다. 2020년 발표된 '사람이 중심이 되는 인공지능 윤리기준'과 2022년의 '교육분야 인공지능 윤리원칙'이 우리나라의 AI 정책과 관련한 큰 방향을 제시하고는 있으나, 학습 데이터의 출처 공개나 창작자의 거부권에 관한 구체적 제도에 대해서는 정비를 제대로 하지 못한 상태였다.

그러다가 최근 2026년 1월 22일자로 "인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(AI 기본법)"을 전면적으로 시행하게 되었다. 대통령을 위원장으로 하는 국가인공지능위원회가 컨트롤타워 역할을 맡고, 가짜 뉴스 방지를 위한 생성형 AI 콘텐츠의 워터마크 표시가 의무화된 것은 적지 않은 의미를 지닌다. 다만 AI 기본법이 '금지' 대신 '사후 규제'와 '자율성'에 방점을 둔 최소 규제 원칙을



채택하고 있다는 점에서, 학습 데이터의 출처 공개나 창작자의 거부권 같은 핵심 쟁점은 여전히 논의의 영역에 남아 있는 것으로 보인다.

우리나라의 개인정보 영역에 대한 정비는 비교적 빠르게 이루어지는 추세다. 2026년 3월 개정된 개인정보 보호법(PIPA)은 중대한 정보 유출 시 전체 매출액의 최대 10%까지 과징금을 부과할 수 있도록 했으며, 채용이나 대출 심사에 사용되는 AI에 의한 '자동화된 결정'에 대해 시민이 거부하거나 설명을 요구할 수 있는 권리도 2026년 9월부터 본격 적용될 예정이다.

반면 학습 데이터와 직접 관련되는 저작권은 여전히 진통을 겪고 있다. 정부는 'TDM(텍스트·데이터 마이닝) 면책'과 함께 '선사용·후보상' 모델을 검토하는 한편, 문화체육관광부는 "저작권은 오직 인간만이 가질 수 있다"는 원칙을 재확인하며 인간의 창작적 기여가 약 30% 이상 반영된 경우에만 해당 부분에 대한 권리를 인정한다는 가이드라인을 발표한 바 있다. 그러나 신문협회 등 창작자 단체는 TDM 면책이 원본 시장을 대체할 수 있는 위험이 있다며 강하게 반발하고 있어, 산업 진흥과 창작자 보호 사이의 합의는 아직도 진행 중이라 할 수 있다.

특히 한국은 OECD 국가 중에서도 데이터 생산량과 디지털 활용도가 매우 높은 나라에 속한다. 그만큼 제도 정비가 미치는 파급력도 크다. AI 기본법이라는 큰 방향은 마련되었지만, 이러한 방향과 학습 데이터 윤리, 창작자 권리, 시민의 동의 등과 어떻게 조화시킬 것인가는 여전히 숙제로 남아 있다.

'선활용, 후규율'같은 식의 접근을 단기적으로는 빠른 성과를 낼 수 있겠지만, 장기적으로 고려해 볼 때 신뢰할 만하고 AI 기술사회로 자연스럽게 진입할 것인가에 대해 선뜻 긍정적인 대답을 내리기 힘든 것으로 보인다.

데이터 활용을 바라보는 세 가지 입장

그렇다면 우리나라가 데이터 주권과 관련하여 어느 노선을 취할 것인가에 대한 모종의 지침은 데이터 활용을 바라보는 다음과 같은 세 방향의 윤리 이론적 차원을 살펴봄으로써 얻을 수 있을 것으로 생각된다.

먼저 데이터의 소유와 활용에 대해 비판적 입장이 존재한다. 의무론적 윤리관에 가까운 이 시선은 "개인은 자신의 데이터에 대해 절대적인 소유권을 가진다"는 주장을 견지한다. 데이터는 곧 인간 존엄성의 연장이며, 데이터를 함부로 다루는 것은 인격을 침해하는 것과 다르지 않다는 입장이다.

이러한 입장에서는 데이터의 수집과 활용은 최소한에 그쳐야 하며, 상업적 이익보다 프라이버시 보호가 언제나 우선해야 한다고 본다. 자신의 사진이 본인의 동의 없이 얼굴 인식 AI 기술의 학습에 활용되었다는 사실에 분노하는 시민의 감정은 이러한 입장과 우리의 상식적인 도덕과의 모종의 일치를 보여준다.

또한 공격적 입장도 존재하며 이는 공리주의적 시각에 뿌리를 두어 "좋은 사회란 모든 이의 데이터를 공공선을 위해 최대한 활용하는 사회"라는 믿음에 기반한다. 다소 과감해 보이지만, 의료 데이터를 활용한 신약 개발, 교통 데이터에 기반한 도시 설계, 재난 대응을 위한 빅데이터 분석 등을 떠올리면 이러한 입장이 강조하는 '사회적 차원의 이익'도 결코 가볍게 다루지 않아야 한다는 점을 일깨워준다.

코로나 팬데믹 시기에 우리가 경험한 시민들의 동선 추적과 확진자 데이터의 신속한 공유는 데이터를 공공선에 적극 활용했을 때 어떤 효용이 발생하는지를 보여주는 사례라 할 수 있다. 다만 이러한 입장에서 주장하는 공익 추구는 개인의 자기결정권을 일정 부분 양보할 경우에 가능하다는 부담을 안고 있다.

마지막으로 진보적 입장으로 불리는 현실적인 절충안이 존재하며, "개인의 동의가 전제된다면, 데이터를 적극 활용하여 더 많은 혜택을 누리는 것이 바람직하다"는 주장과 궤를 같이 한다. 이는 데이터의 상업적 가치를 긍정하면서도 개인의 자기결정권을 존중하려는 균형 잡힌 태도이며, 오늘날 다수의 기업과 기관이 채택하고 있는 입장이기도 하다. 정보제공자의 동의에 기반할 경우, 의료 분야의 임상 데이터 공유, 금융권의 마이데이터 사업, 공공기관의 데이터 개방 정책 등이 모두 이러한 진보적 입장의 변주들로 가능하다.

세 입장 중 어느 하나가 정답일 수는 없다. 다만 극단적인 입장보다는 '동의'에 기반한 진보적 입장이 절차적 정당성을 확보하면서도 결국 가장 큰 공익을 가져올 것이라는 것이 필자의 판단이다. 그러나 우리에게 남은 과제는 어느 입장을 취사선택하는 것의 문제가 아니라, 어느 입장을 취하더라도 '동의'라는 최소한의 데이터 주권을 인정하는 절차를 확보하고 실행해야 한다는 점이다.

가이드라인 분석에서 드러난 '동의'의 빈자리

필자가 국내외에서 발표된 데이터 윤리 가이드라인 37건을 분석한 결과, 윤리 원칙들 사이의 강조 정도는 상당히 다양한 것으로 나타났다. 가장 압도적으로 지지를 받는 가치는 '투명성'이었으며, '공정성'과 '보안'이 그 뒤를 이었다. 이는 데이터를 어떻게 안전하게 보관하고 처리 과정을 얼마나 투명하게 공개할 것인가에 대해 사회적 합의가 비교적 단단하게 형성되어 있음을 시사한다. 데이터의 '관리'에 있어서는 대부분의 데이터 윤리 가이드라인이 수렴하는 윤리적 기준점을 마련했다는 의미로 파악된다.

그런데 흥미롭게도 이러한 '동의'야말로 현재 전 세계의 데이터 윤리가 그 중요성에 비해 소홀히 다루고 있는 것으로 파악된다. 정작 데이터 주권의 핵심인 '동의'와 관련된 항목에 대한 언급과 강조가 상대적으로 적게 나타났다. 이와 더불어 사용 목적에 부합하는지를 따지는 '합목적성'이나 체계적인 관리를 위한 '거버넌스' 역시 상대적으로 그 존재감이 약했다. 결국 많은 지침이 "데이터를 어떻게 안전하게 지키고 투명성을 보장할 수 있는가"에는 집중하지만, 데이터 주권에 관한 근본적인 물음에는 다소 침묵하고 있는 셈이다. 이러한 불균형은 우리가 '관리의 윤리'에 치중하느라 '동의의 윤리'가 지닌 무게를 간과하고 있음을 보여준다.

이는 단순한 우연이 아니라는 것이 필자의 생각이다. 동의를 받는 절차는 시간과 비용이 든다. 모든 사용자에게 일일이 허락을 구하는 일은 데이터 활용의 효율성을 떨어뜨린다. 따라서 효율성의 논리가 우선시될수록 '동의'의 자리는 좁아지기 마련이다. 효율을 위해 자기결정권을 양보하는 사회는 결국 개인의 데이터를 자원으로, 그리고 그 주체의 권리 또한 제대로 존중하지 않는 사회가 된다.

데이터 윤리들이 투명성을 강조하고 있다는 점은 바람직하지만 '동의'에 대한 중요성이 상대적으로 강조되지 않는 상황은 관심을 기울여야 할 부분으로 생각된다.

그런데 더욱 심각한 문제는 우리가 이미 데이터 활용에 '동의'했다는 착각이다. 데이터에 대한 자기결정권을 양도한다는 내용이 길고 복잡한 약관에 자그마하게 표현되어 있고, 이를 주지하지 않은 상태로 무심코 누른 동의가 진정한 의미의 동의인지 자문해 볼 일이다. 인터넷상의 사이트나 서비스에 가입하는 절차에서 '필수 동의' 항목을 거부하면 서비스 자체를 사용할 수 없는 구조는, 그 내용을 인지했다 하더라도 동의라기보다는 강요에 가까운 것으로 보인다. 진정한 동의는 거절할 자유를 전제로 하는 것이지만, 이러한 디지털 환경에서는 거절할 자유는 사실상 서비스를 포기해야만 실현될 수 있다.

그뿐만 아니라 '동의 피로(consent fatigue)'라는 현상도 짚어볼 필요가 있다. 우리는 하루에도 수십 번씩 가입할 때, 혹은 새로운 앱을 설치할 때마다 동의 여부를 묻거나 권한을 요청하는 팝업을 마주한다. 이처럼 너무 많이 동의에 대한 물음이 존재하는 상황에서 정작 우리는 더 이상 묻지 않았으면 좋겠다는 마음이 강해지게 되고 무조건 동의한다는 선택을 하게 된다. 이러한 형식적인 동의로 부터 탈출하여 진정한 의미의 동의를 실현할 수 있도록 동의 절차를 설계해야 할 필요가 있다.

사람 중심의 데이터 거버넌스가 필요하다

그렇다면 우리는 어디로 가야 할까. 식상한 결론이긴 하지만, 추상적인 원칙의 나열만으로는 부족하다. 이제는 구체적인 윤리 프레임워크가 실제로 작동해야 한다. 필자는 다음 네 가지 정도가 시급하다고 생각한다.

먼저 기관과 기업에 '데이터 윤리 위원회'를 구성할 필요가 있다. 위원회는 기술 전문가만이 아니라 윤리 전문가, 법률가, 그리고 시민 참여자를 포함해야 한다. 다양한 시선이 교차할 때 비로소 편향과 이해 충돌을 줄일 수 있기 때문이다. 또한 데이터의 전 생애주기에 걸쳐 윤리적 점검표(체크리스트)가 작동해야 하며, 조직의 윤리적 성숙도를 정기적으로 평가하는 체계 또한 마련되어야 한다. 위원회가 사후 감사 기구가 아니라 사전 설계 기구로 기능할 때 비로소 그 효력이 발현된다는 점도 강조하고 싶다. 어느 프로젝트가 끝난 뒤에야 데이터 활용의 윤리성을 묻는 것보다는, 시작하기 전에 묻는 것이 훨씬 적은 비용으로 더 큰 효과를 야기할 수 있다.

또한 동의 절차의 형식화를 극복할 필요가 있다. 끝없이 길고 복잡한 약관에 무심코 클릭하는 동의가 진정한 동의가 아니라는 사실은 누구나 알고 있다. 사용자가 자신의 데이터가 어디로 어떻게 흘러가는지, 어떤 모델을 학습시키는 데 쓰이는지를 '이해하고 결정할 수 있는 동의'를 할 수 있는 환경을 만드는 일이 필요하다. 짧고 명확한 설명, 단계별 동의, 그리고 언제든 철회할 수 있는 권리를 언급하는 과정이 포함될 때 동의는 비로소 신뢰라는 본래의 의미를 회복할 수 있을 것이다.

다음으로 사회적 합의의 문제가 있다. 우리 사회가 데이터 주권에 대한 공통된 언어를 가져야 한다는 의미다. 유럽처럼 강력한 규제가 답일지, 일본처럼 실용적 유연성이 답일지, 아니면 한국 고유의 길을 모색해야 할지에 대한 사회적 합의가 필요하다. 이는 정부와 기업뿐만 아니라 시민의 참여 없이는 불가능한 일이며, 데이터 윤리에 관한 논의가 더 이상 전문가만의 영역이 될 수 없다는 점도 함께 고려해야 한다. 시민이 자신의 데이터가 어디에 어떻게 사용되는지를 묻고 답을 요구하는 문화가 자리 잡을 때 비로소 살아 있는 거버넌스가 작동하게 될 것이다.

마지막은 교육의 문제다. 디지털 리터러시가 이제는 '문해력'의 일부가 되었듯, 데이터 윤리 또한 시민 교육의 한 축이 되어야 한다는 생각이다. 자신의 데이터가 어떻게 흐르는지, 어떻게 활용되며, 어떤 권리가 자신에게 있는지를 아는 시민이 많아질수록 우리 사회의 데이터 주권은 단단해질 것이다. 가르치지 않은 권리는 행사되지 않기 때문이다.

"AI가 학습한 데이터는 누구의 것인가?"라는 질문에 대한 답은 결국 우리가 어떤 사회를 원하는가에 대한 답이기도 하다. 데이터가 사람을 위한 것일 때 그 데이터를 학습한 AI도 비로소 사람을 위한 도구가 된다. 거꾸로 데이터가 사람을 도구화하는 자원이 될 때 AI 또한 사람을 향한 통제의 수단으로 변질될 수 있다. 기술의 화려함에 눈이 멀기보다는 종종 기술 발전의 자양분이 된 '사람의 흔적'을 잊어버리지 않는 것, 즉 데이터의 주인이 결국 '사람'임을 잊지 않는 것이 투명성과 공정성을 양보하지 않으면서도 동의라는 절차의 무게를 가벼이 여기지 않는 길로 생각된다. 이처럼 데이터 주권을 확립하기 위한 작은 시도와 노력들이 데이터가 무수히 많은 방식으로 활용되는 사회에서 반드시 뒷받침되어야 할 디딤돌임을 잊지 말아야 할 것이다.

인공지능 공정성의 명과 암

김효은 | 국립한밭대학교 인문교양학부 교수

1

AI 책임의 시대와 '고영향AI'의 공정성

2026년 1월, 한국에서 「인공지능 산업 육성 및 신뢰 기반 조성 등에 관한 법률(인공지능기본법)」이 본격 시행되면서 우리나라도 본격적인 'AI 책임의 시대'에 접어들었다. 이 법안은 인공지능을 산업적으로 육성하는 동시에, 인간의 생명이나 권익에 중대한 영향을 미치는 분야를 '고영향 인공지능'으로 분류하여 엄격한 안전성과 투명성을 요구한다. 인공지능이 인간의 주관적 편견을 배제하고 객관적 의사결정을 내릴 것이라는 기대는 이제 법적 권리와 의무의 영역으로 들어왔다. 그러나 재범 예측 알고리즘처럼 한 개인의 자유를 구속하는 결정에 참조되는 기술의 경우, '신뢰성'과 '공정성'을 기술적으로 어떻게 구현할 것인가는 중요한 문제이다.

재범 예측 알고리즘은 미국의 몇몇 주에서 형사 피고인의 범죄가능성을 평가하여 재판 전, 가석방, 선고 결정에서 참조자료 중 하나로 2000년부터 사용되어왔다. 널리 사용되는 범죄 위험 평가 도구 중 하나인 '교정 대상 범죄자대체 제재를 위한 관리 프로파일링'(Correctional Offender Management Profiling for Alternative Sanctions; COMPAS)는 Northpointe사(현재 'equivant')에서 1998년 개발 이후 100만 명이 넘는 범죄자를 평가하는 데 사용되었다. COMPAS는 개인에 대한 137가지 특징과 개인의 과거 범죄 기록을 바탕으로 평가 후 2년 이내에 피고인이 경범죄 또는 중범죄를 저지를 위험을 예측한다.

이 소프트웨어 도구가 이슈가 된 것은 비영리 탐사 매체인 ProPublica가 탐사보도 "Machine Bias"와 방법론 보고서 "How We Analyzed the COMPAS Recidivism Algorithm"에서 미국에서 비특권층으로 여겨지는 흑인종에 불리하게 설계가 되었다고 분석을 제시했기 때문이다.

인공지능이 사회적 차별을 고정하는 데 사용된다는 비판 자체를 흥미롭게 생각할 수 있다. 그러나 중요한 점은 데이터 혹은 알고리즘 편향을 어떤 근거로, 어떤 유형의 통계적 계산 기준으로 평가하는 것이 적절한가이다. 특정 인종에 알고리즘이 편향되었다는 등의 평가는 특정한 맥락에서의 공정성 기준을 전제로 한다. '공정성'이라는 단어는 더 이상 인문사회 분야에서 주로 사용되는 이론적 개념이 아니다. 인공지능 분야에서는 '알고리즘 공정성' 혹은 '통계적 공정성'이 편향된 알고리즘 및 데이터를 보정하는 기술의 한 분야로 자리 잡았다. 그러나 편향 기술이 인공지능의 편향성을 해결해주지 못한다. 이는 현재 기술단계가 미성숙해서가 아니라 근본적인 한계가 있기 때문이다.

2

COMPAS 논쟁: 통계적 공정성 기준의 대립

프로퍼블리카 매체가 COMPAS 도구의 편향성을 비판한 근거는 실제 데이터이다. 재소자의 보석 전 석방 결정에 사용하고 있었던 플로리다주 브로워드 카운티를 조사 대상으로 선택하였고, 2013~2014년에 체포된 7,000여 명의 COMPAS 위험 점수를 확보하여, 위험점수와 실제 2년간 재범 여부를 대조 분석하여 평가하였다. 분석 결과는 위험점수는 높으나 실제 재범을 하지 않은 비율(FPR; 위양성률)은 흑인(44.9%)이 백인(23.5%)보다 약 2배 높게 나타났고, 위험점수는 낮으나 실제 재범을 한 비율(FNR: 위음성률)은 거꾸로 백인이 흑인보다 약 1.7배 높은 것으로 분석되었다.(표 1) 이는 피고인 개인의 입장에서 "내가 재범하지 않을 사람인데도 고위험군으로 잘못 분류될 확률이 인종에 따라 달라지는가?"라는 질문에 답하기에 적절한 기준이다. 알고리즘이 객관적이고 중립적일 것이라는 일반적인 믿음과 달리, 수집된 데이터 자체가 과거의 차별적 관행이나 사회적 편견을 내포하고 있을 경우 알고리즘이 이를 학습하여 편향된 결과를 재생산할 수 있음을 보여준다. 이는 기술이 사회적 불평등을 과학적 근거라는 미명하에 고착화할 위험이 있다는 점을 시사한다. 다만, 인종 간 오류율을 유사하게 맞추려다 보면, 억울하게 고위험군으로 잘못 분류될 사람들의 비율을 줄일 수는 있지만, 재범 위험이 높은 집단의 예측 점수를 낮추거나 그 반대의 상황이 발생할 수 있다. 이는 결과적으로 전체적인 예측 정확도를 떨어뜨릴 수 있다는 단점이 있다.

표 1. ProPublica의 위양성률(FPR) / 위음성률(FNR) 수치

기준	흑인 피고인	백인 피고인	차이 (흑인-백인)	해석
▶ FPR (False Positive Rate): 실제 비재범자 중 고위험으로 잘못 분류된 비율 = FP ÷ (FP + TN)				
FPR: 실제 비재범자 중 고위험으로 오분류된 비율	44.9%	23.5%	+21.4%p	흑인이 약 2배 높음 → 위양성 불균형
▶ FNR (False Negative Rate): 실제 재범자 중 저위험으로 잘못 분류된 비율 = FN ÷ (FN + TP)				
FNR: 실제 재범자 중 저위험으로 오분류된 비율	28.0%	47.7%	-19.7%p	백인이 약 1.7배 높음 → 위음성 불균형

〈출처〉 Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. ProPublica, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS를 만든 Northpointe사는 프로퍼블리카의 분석을 즉시 반박하였다. 위양성률과 위음성률(FPR과 FNR)은 기저율을 고려하지 않은 기준이기에 적절한 공정성 기준이 되지 못한다는 것이다. 흑인 피고인의 위양성률이 높은 부분에 대해서는, 인종을 입력변수로 사용해서가 아니라 역사적으로 흑인 재범자 수가 더 많았기 때문에 실제 기저 재범률이 더 높게 된 상황에 기인한다고 설명했다. 노스포인트사는 이러한 기저율을 반영한 기준인 '예측적 균등성'이 더 적절한 기준(Dieterich et.al., 2016)이라고 주장했다. '예측적 균등성(predictive parity)'은 긍정적 결과의 정확도(Positive Predictive Value: 양성예측치) 즉 고위험으로 예측한 사람 중 실제 재범자 비율이 모든 집단에 동일하게 나타난다는 의미이다. 이 공정성 기준에 따라 계산한 수치는 흑인과 백인 간 차이가 거의 없거나 매우 적다(표 2). 이 기준은 판사 입장에서 점수가 높게 나온 이 사람이 실제로 재범할 확률이 인종과 상관없이 동일한지에 대해 실용적으로 답할 수 있는 기준이다.

반면, 전체 데이터의 차원에서는 과거의 차별적 수사 관행이나 사회구조적 요인으로 인해 특정 집단의 재범률(기저율)이 높게 측정되었다면, 알고리즘은 이를 자연스러운 사실로 수용하기에 구조적 차별을 영구화할 위험이 있다.

표 2. Northpointe의 예측적 균등성(PPV) 수치

기준	흑인 피고인	백인 피고인	차이 (흑인-백인)	해석
▶ PPV (Positive Predictive Value): 고위험 분류 중 실제 재범자 비율 = TP ÷ (TP + FP)				
PPV: 고위험으로 분류된 피고인 중 실제 재범한 비율	63%	59%	+4%p	거의 동일
NPV: 저위험으로 분류된 피고인 중 실제 재범하지 않은 비율	65%	71%	-6%p	소폭 차이

〈출처〉 Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Northpointe Inc.

위에서 살펴본 두 공정성 기준의 면면을 비교해보면, 공정성 기준을 선택할 것인가는 단순한 기술적 문제가 아니라, 우리 사회가 '효율적인 치안 유지'와 '사회적 약자에 대한 보호' 중 어느 가치에 더 무게를 둘 것인가를 결정하는 정치적·윤리적 선택의 영역임을 보여준다.

3

대리 차별과 양립 불가능성: 기술적 보정을 넘어선 가치 판단

COMPAS를 만든 회사의 반박이 사실인지 여부는 모델 내부 구조를 영업비밀로 하였기에 직접 알 수는 없다. 그러면 인종이 주요 입력변수였는지의 여부는 어떻게 알 수 있을까? 인종 변수의 값을 흑인에서 백인으로, 혹은 백인에서 흑인으로 변경하였을 때 최종 재범 위험 값이 어떻게 변화하는지를 통해 아주 정확하지는 않지만 조금이나마 추측해볼 수 있을 것이다. 또, 인종정보가 명시될 때와 아닐 때를 비교할 수 있다. Dressel 등의 연구자들(2018)은 COMPAS와 인간의 예측 모두 인종 정보가 명시되든, 되지 않든 흑인 피고인에 대한 더 높은 오탐률뿐만 아니라 예측 정확도에는 유의미한 변화가 없었음을 실험 및 분석하였다.

이는 인종을 직접 변수로 사용하지 않더라도 다른 데이터(전과 등)들이 인종과 상관관계를 가지기 때문이다. 예컨대, '인종'을 입력변수로 사용하지 않더라도 이와 밀접한 관련이 있는 교육수준, 재정적 어려움의 경험 여부, 가족·지인의 범죄 이력, 부모·형제·친구의 체포·수감 경험, 거주 지역의 범죄 환경 인식, 청소년기 행동 문제, 사회적 고립감·소외감에 대한 심리 문항 등이 실제 공개된 설문 문항에 포함되어 있기에, 이들이 대리적 역할을 하여 인종적 편향을 매개할 가능성은 여전히 남아 있다.

이때 대리적 역할을 하게 되는 변수들을 '대리변수'(proxy attributes)라고 하고, 이를 통해 차별이 결과적으로 나타날 때 명시적 차별이 아니라 '대리 차별'이라고 부른다. 이러한 가능성과 관련하여, 시카고대의 Nathan Srebro 교수는 "'예측적 동등성'은 실상 '최적 차별'에 해당한다." (Angwin, J. et.al. 2022)고 비판하였다.

COMPAS 도구를 둘러싼 양측의 기준을 동등하게 비교하여 판단하기는 쉽지 않다. 그렇다면 두 유형의 공정성 기준들을 동시에 충족하게 할 수는 없을까? 흑백 집단 간 기저 재범률(base rate)이 다른 상황에서, COMPAS가 예측적 균등성을 유지하면서 동시에 위양성률과 위음성률을 여러 집단에 유사하게 조정하는 것은 수학적으로 불가능하다. 위양성률을 균등하게 맞추면, 위음성률은 반대 방향으로 변하고 예측적 균등성도 변한다. 또, 이에 따라 정확도 또한 변동한다. 결론적으로, 하나의 도메인에 적절하게 적용할 수 있는 공정성 기준들이 몇 개 있다고 할지라도 상이한 공정성 기준들을 동시에 충족시킬 수는 없다.

인공지능 분야에서 논의되는 공정성 기준들(Verma et al. 2018)은 다음의 20가지가 기본 유형이다. 예측 기반의 집단공정성, 조건부 통계적 동등성, 예측 및 실제 결과 기반의 예측적 동등성, 위양성률 균형, 위음성률 균형, 동등 확률, 조건부 사용 정확도 동등성, 전체 정확도 동등성, 대우 동등성. 그리고 예측확률 및 실제결과 기반의 테스트공정성, 보정성, 양성집단 균형, 음성집단 균형. 유사성 기반의 인과적 차별, 무지를 통한 공정성, 인식을 통한 공정성. 인과추론 기반의 반사실적 공정성, 미해결된 차별 없음, 대리 차별 금지, 공정 추론. 이들 기준들 또한 동시에 충족시키기는 어려운 기준들이며 해당 사안과 맥락에 따라 다르게 적용되기에 그 맥락을 파악하는

'인간이 부여하는 가치'에 따라 다르게 선택될 수밖에 없다.

이러한 공정성 기준 설정의 난제뿐만 아니라, 도구 자체가 가진 근본적인 예측력과 신뢰도에 대해서도 심각한 의문이 제기되고 있다. 알고리즘의 기술적 공정성을 검토하기에 앞서, 그 도구의 정확성 검토가 우선시되어야 하기 때문이다. 재범 예측 도구 아홉 가지의 유효성을 검토한 선행 연구에 따르면, COMPAS를 포함한 8가지 도구가 정확한 예측에 실패(Geraghty & Woodhams 2015)했으며, 9가지 폭력 예측 도구에 대한 메타 분석 결과(Yang M, et al. 2010)는 이들 기법의 예측 정확도가 중간 수준에 불과하다는 점을 보여준다.

따라서 이러한 기술이 예방적 구금과 같은 형사 사법 의사결정의 일부 근거로 사용되는 것도 조심스럽게 접근할 필요성이 있다. 또한, "어떤 기준으로 알고리즘 편향 및 데이터 편향을 완화할 것인가"는 순수하게 기술적인 문제가 아니라, 어떤 유형의 오류가 더 심각한 해악인가에 대한 인간의 규범적 판단이 선행되는 문제이다. 결국 인공지능 공정성 논의는 더 완벽한 알고리즘의 개발만으로는 도달하기 어렵다.

우리나라의 인공지능기본법은 인공지능의 결정이 개인의 권리에 중대한 영향을 줄 경우 그 근거를 설명하도록 하는 '설명 의무'를 명시하고 있다. 하지만 본문에서 살펴본 COMPAS 사례는 기술적 설명만으로는 해결되지 않는 공정성의 '수학적·윤리적 모순'이 존재함을 시사한다. "어떤 유형의 오류를 우리 사회가 더 용인할 것인가"의 문제는 알고리즘이 대신 결정해 줄 수 없다. 수학적으로 양립 불가한 공정성 기준들 사이에서 하나를 선택하는 행위는 본질적으로 가치 기반의 정치적이고 윤리적인 결정으로, 가치 판단의 문제를 공론화하여 사회적 합의를 이루어 나가는 과정이 요구된다.

참고자료

1. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016) "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." *ProPublica*. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
2. W. Dieterich, C. Mendoza, T. Brennan (2016) "COMPAS risk scales: Demonstrating accuracy equity and predictive parity" (*Technical Report*, Northpointe Inc., 2016).
3. Dressel, J., & Farid, H. (2018) "The accuracy, fairness, and limits of predicting recidivism". *Science advances*, 4(1), eaao5580.
4. K. A. Geraghty, J. Woodhams (2015) "The predictive validity of risk assessment tools for female offenders: A systematic review," *Aggression and Violent Behaviour*. 21, 25.
5. Yang M. et.al. (2010) "The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools." *Psychological Bulletin*. 136, 740 - 767.
6. Verma, S., & Rubin, J. (2018) "Fairness Definitions Explained." In *2018 IEEE/acm international workshop on software fairness* (fairware) (pp. 1-7). IEEE.

통계로 바라보는 세상이야기

신동헌 | 도서출판 지일북스 대표

웹소설 산업 잠입기 :AI가 남긴 웹소설 산업의 과제

한국출판문화산업진흥원의 「2024년 웹소설 산업 현황 실태조사」에 따르면, 국내 웹소설 산업 규모는 2022년에는 약 1조 390억 원에서 2024년에는 약 1조 3,500억 원까지 성장한 것으로 추정되고 있는데요. 특히, 평소 월 1회 이상 이용하는 디지털 콘텐츠를 살펴보면, 웹소설 이용 경험은 40.6%로 10명 중 약 4명이 웹소설을 읽어본 경험이 있다고 응답했습니다. 웹소설 플랫폼을 대상으로 한 조사에 따르면, AI 서비스를 도입한 경험이 있는 플랫폼은 15.4%에 불과했습니다. 반면, 84.6%는 AI 도입 경험이 없는 것으로 나타났습니다. AI 서비스를 도입한 가장 큰 이유로는 '콘텐츠 생성 효율 증대(48.0%)'가 꼽혔고, 그다음으로는'데이터 분석과 독자 맞춤형 추천(32.0%)'이 뒤를 이었습니다.

데이터로 본 국민 운동 습관

문화체육관광부의 「2025년 국민생활체육조사」를 살펴보면, 국민의 40.5%가 참여하며 압도적 1위를 차지한 종목은 바로 '걷기(속보 포함)'였습니다. 17.5%의 보디빌딩(헬스), 17.1%의 등산이 그 뒤를 이었습니다. 특히 달리기는 2024년 종목을 구분하여 조사하기 시작한 이래로 꾸준히 상승 곡선을 그리며, 최근의 러닝 열풍을 직관적으로 보여주고 있습니다. 성별과 연령에 따라 즐겨 하는 운동은 달랐습니다. 여성의 40대 이상과 10대는 '걷기', 2030 세대는 '요가 및 필라테스'에 가장 많이 참여하는 것으로 나타난 반면 남성의 경우 10대는 '축구', 20대부터 40대까지는 '헬스', 50대는 '등산', 그 이후 연령대에서는 '걷기'에 가장 많이 참여하고 있습니다. 문화체육관광부의 「스포츠산업조사」에 따르면, 코로나 시기 이후 매년 스포츠 시설업과 스포츠 용품업의 사업체 수와 매출액이 꾸준히 증가하며 2024년 기준 역대 최대 규모를 기록했습니다.

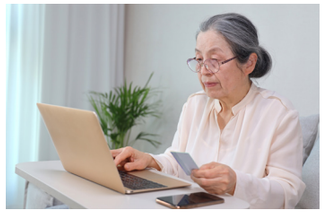
통계로 살펴보는 러닝 트렌드

러닝 인구가 눈에 띄게 늘어나면서, '1,000만 러너 시대'에 접어들었다는 말도 나오고 있습니다. 문화체육관광부의 「2025년 국민생활체육조사」에 따르면, 주로 참여하는 체육활동 중 '달리기'를 선택한 비율이 2024년 4.8%에서 2025년 7.7%로 2.9%p 상승했습니다. 걷기, 등산, 자전거 등 전통적으로 인기 있던 운동들이 주춤하는 사이 러닝은 가장 지속적인 성장세를 보이며 '대세 운동'으로 자리 잡고 있습니다. 코로나19 이후 밀폐된 공간을 피하고 혼자, 시간 제약 없이 할 수 있는 운동을 찾으면서 러닝이 빠르게 유행하기 시작했습니다. 스마트워치와 러닝 앱, 커뮤니티의 확산으로 러닝은 단순한 운동을 넘어 '자기 관리, 기록, 소통'의 방식이 되었다고 합니다. 그런데, 왜 하필 러닝일까요? 러닝은 걷기보다 빠르지만 부담은 적고, 운동화 한 켤레만 있으면 되니 자전거보다 간편합니다. 이러한 이유로 러닝은 비교적 '지속 가능하다'고 느끼는 운동입니다.

데이터로 본 디저트 소비 열풍

두바이 쫀득 쿠키, 일명 '두쫀쿠'가 열풍이던 시기가 있었습니다. 사람들은 디저트에 열광하는데, 유행은 왜 오래 지속되지 못할까요? 국가데이터처의 「가계동향조사」에 따르면 식료품, 비주류음료, 식사 등의 식품 관련 소비는 2021년 75만 원에서 2025년 89만 5천 원으로 최근 5년간 꾸준한 상승세를 보였습니다. 특히 2022년에는 전년 대비 4만 7천 원 증가하며 가장 큰 상승폭을 기록했습니다. 한국소비자원의 「2025 한국의 소비생활지표」에 따르면 소비자가 가장 중요하게 생각하는 소비생활 분야는 '식품·외식' 29.0%, '금융·보험' 10.8%, '주거·가정' 10.6% 순으로 나타났습니다. 특히, 2023년 조사에서 28.7%로 1위를 기록했던 식품·외식 분야는 2025년 조사에서도 종합 순위 1위를 유지한 것으로 나타났습니다. 이제는 의식주가 아닌 '식금주'라는 용어까지 등장하면서 식품·외식은 소비생활에서 가장 중요한 분야로 여겨지고 있습니다.

누구나 연결된 세상, 클릭과 스크롤 사이의 새로운 문해력



디지털 리터러시(디지털 문해력)는 디지털 기기를 활용해 필요한 정보를 찾고, 원하는 작업을 수행할 수 있는 지식과 능력을 의미합니다. 과학기술정보통신부의 「2024년 디지털정보격차 실태조사」에 따르면, 20대와 30대는 각각 139.8%, 136.1%로 전반적으로 높은 수준을 보인 반면, 60대는 63.4%, 70대 이상은 26.8%로 큰 차이를 보였습니다. 디지털 교육은 어떻게 이루어져야 할까요? 한국교육학술정보원의 「2024년 초·중등 디지털 전환 현황 및 인식조사」에 따르면, 응답자의 59.8%가 '디지털 기반 교과 연계 학습 활동'이 강화되어야 한다고 답했으며, '디지털 리터러시(문해력) 및 시민성 교육'(49.6%), '디지털 기반 콘텐츠 활용 및 생산 방법'(40.5%)이 뒤를 이어, 학교 현장에서의 디지털 활용 역량 강화의 필요성이 확인되었습니다. 디지털 문해력은 모든 국민이 갖춰야 할 기본 역량으로 자리 잡고 있습니다.

제로 칼로리? 제로 슈가?

건강보험심사평가원의 「당뇨병 진료현황발표」에 따르면, 2019년 대비 2023년 국내 당뇨병 환자 수와 총 진료비가 모두 증가했습니다. 환자 수는 2019년 3,229천 명에서 2023년 3,829천 명을 기록했고, 진료비도 환자수 증가에 따라 2019년 9,357억 원에서 2023년 1조 1,765억 원으로 상승했습니다. 많은 사람들이 제로(Zero)에 주목하고 있습니다. 일반적으로 제로 칼로리는 열량이 '거의' 없다는 뜻으로, 100ml당 4kcal 미만에 해당하며, 제로 슈가는 설탕(당류)을 거의 포함하지 않은 제품을 의미하며, 100ml당 당류가 0.5g 미만인 경우에 해당합니다. 제로 제품이라 하더라도 소량의 당류나 감미료가 포함될 수 있습니다. 대체당은 일반 설탕보다 칼로리가 훨씬 낮아 체중 관리에 도움이 되고 혈당 조절 및 당뇨 예방에도 효과적이지만, 장기간 다량 섭취할 경우 심장마비나 고혈압과 같은 위험을 높일 수 있다는 결과도 보고되었습니다.

늦어지는 결혼과 첫 출산

국가데이터처의 「인구동향조사」에 따르면, 1995년 이후 약 30년간 우리나라 혼인 건수는 전반적으로 감소 흐름을 보였습니다. 1995년 약 39만 9천 건이었던 혼인 건수는 이후 지속적인 하락세를 이어가며 2022년에는 약 19만 2천 건으로 최저치를 기록했습니다. 1995년 남성의 경우 25-29세 연령대의 혼인율이 가장 높았으며, 인구 천 명당 93.9명이 결혼한 것으로 나타났으나 20대 남성의 혼인율은 지속적으로 감소하고 있습니다. 반면 35-39세 연령대는 1995년 대비 21.7% 증가하여 가장 큰 상승 폭을 보이고 있으며, 2025년 기준 혼인율이 가장 높은 연령대는 30-34세로 나타나고 있습니다. 여성은 남성보다 높은 20대 혼인율을 보였으나, 이 역시 1995년에 비해 감소한 흐름을 보이고 있습니다. 1995년에는 20-24세 혼인율이 73.6%, 25-29세가 76.1%로 나타났으나, 2025년에는 각각 7.6%와 44.3%으로 크게 감소했습니다.

일과 가정, 맞벌이의 일상화

국가데이터처의 「2024년 하반기 지역별고용조사」에 따르면, 유배우 가구 중 맞벌이 가구 비중은 2021년 45.9%에서 2022년 46.1%, 2023년 48.2%로 꾸준히 증가했습니다. 2024년에는 48.0%로 하락했습니다. 최근 4년간 45% 수준을 상회하며, 유배우 가구의 절반 가까이가 맞벌이 가구로 나타났습니다. 이는 맞벌이 가구가 우리 사회에서 큰 비중을 차지하고 있음을 보여줍니다. 연령별 맞벌이 가구 비중을 살펴보면, '30~39세'가 61.5%로 맞벌이 가구 비중이 가장 높은 것으로 나타났습니다. 40~49세가 59.2%, 50~59세가 58.0%로 뒤를 이었습니다. 가정과 일을 병행하는 세대일수록 맞벌이 비중이 높은 것을 알 수 있습니다. 가구주 직업을 기준으로 맞벌이 가구 비율을 살펴보면, 농림어업 숙련 종사자가 79.8%로 가장 많은 비율을 차지했고, 그 뒤를 이어 서비스 종사자가 66.8%, 판매 종사자가 66.1% 순으로 나타났습니다.

최저임금의 변화 속 대학생들이 아르바이트하는 이유는?

최저임금위원회의 「연도별 최저임금 결정현황」에 따르면, 2015년 5,580원, 2016년 6,030원, 2017년 6,470원, 2018년 7,530원, 2019년 8,350원, 2020년 8,590원, 2021년 8,720원, 2022년 9,160원, 2023년 9,620원, 2024년 9,860원이었고, 2025년 10,030원으로 만 원대에 진입한 이래 올해는 전년 대비 2.9% 인상한 10,320원으로 시행되고 있습니다. 2024년 알바천국의 설문조사에 따르면 '스스로 돈을 벌어보고 싶어서'가 59.0%로 가장 높은 비중을 차지했습니다. 이어 '등록금이나 여행비 등 목돈을 마련하기 위해' 아르바이트를 한다는 응답도 37.7%로 나타났으며, 경제적 이유가 주요 동기임을 보여줍니다. 이외에도 '아르바이트 자체를 경험해 보고 싶어서(37.5%)' 또는 '로망을 실현하고 싶어서(7.5%)' 지원했다는 응답도 확인할 수 있었습니다.

소도시 여행 열풍과 로컬힙

최근 국내에서는 소도시 여행 열풍이 일어나고 있습니다. 이런 흐름은 유명한 관광 도시에 가기보다는 로컬에서의 개인적인 경험과 의미에 집중하는 로컬리즘과, 이를 힙(hip)하다고 느끼는 현대인의 감정이 어우러진 '로컬힙'의 결과물이기도 합니다. 강원 동해시, 경남 남해군, 강원 양양군, 전남 고흥군은 모두 인구 10만 명이 되지 않는 소도시입니다. 한국관광데이터랩의 「지역별 관광 현황」에 따르면, 네 지역 모두 2024년 대비 2025년 방문자 수가 증가했습니다. 이 가운데 경남 남해군은 2024년에 약 839만 명이던 방문자가 2025년에는 약 906만 명으로 약 8% 상승하며 가파른 증가세를 보였습니다.

특히 이런 소도시 여행의 인기 상승의 이유를 두고 한국관광데이터랩은 「2025 관광컨설팅 이슈발굴」 보고서에서 '로컬리즘'이라고 설명하고 있습니다. 로컬리즘은 로컬에 대한 경험이나 로컬 푸드, 공정관광 등으로 여행의 관심사가 이동 및 확대되는 현상을 말합니다.

우리 국민의 복지 체감은?

지난 4월 23일 국회 본회의에서 「장애인권리보장법」이 통과되며 우리 사회 장애 정책의 근본적 전환을 예고하는 새로운 법적 틀이 마련되었습니다. 민주사회를 위한 변호사모임은 24일 성명을 내고 "10년의 논의가 마침내 결실을 맺었다"며 이번 입법을 역사적 진전으로 평가했습니다. 국가데이터처의 「2025년 사회조사」에 따르면, 우리 국민의 전반적인 생활 여건 만족도는 2023년 39.1%에서 40.0%로 상승한 것으로 나타났습니다. 분야별로 살펴보면, 사회보장제도가 좋아졌다고 응답한 비율이 2023년 42.7%에서 43.7%로 상승했으며, 문화 및 여가 생활 여건이 좋아졌다고 응답한 비율 또한 2023년 대비 2025년 40.9%로 1.5%p 상승한 것으로 나타난 반면, 보건의료 서비스는 좋아졌다고 응답한 비율이 45.7%에서 42.0%로 다소 하락하였습니다. 이는 국민들이 일상에서 체감하는 여가와 문화 활동 및 사회보장제도의 질이 점차 높아졌음을 의미합니다.

슬기로운 의류 생활 방법은?

기후에너지환경부에서 발표한 「전국 폐기물발생 및 처리 현황」에 따르면, 폐기물의 발생량은 2020년도에 37만 톤을 넘었습니다. 이후 2022년도에 소폭 감소했지만, 다음 해인 2023년도에는 최고치인 39.8만 톤을 기록했습니다. 효율적으로 의류 생활을 즐기는 방법은 무엇일까요? 먼저, 비영리재단 다시입다연구소에서 제안하는 '21% 파티'입니다. 연구소의 조사에 따르면, 옷을 사놓고 입지 않는 옷의 비율이 21%인데, 21%의 안 입는 옷을 행사를 통해 다시 되살리자는 취지의 행사입니다. 두 번째 방법은 아름다운가게입니다. 아름다운가게는 물건의 재사용과 재순환을 실천하는 비영리재단으로, 기부된 물품을 재판매하여 얻은 수익금으로 어려운 이웃을 돕고 있습니다. 아름다운가게의 「2024 참여와 나눔 보고서」에 따르면, 2024년 한 해에 아름다운가게에 기부된 물품 2,698만 점 중 약 62%가 의류였다고 합니다. 끝으로 중고마켓을 이용하는 방법도 추천드립니다.

"나만의 조합으로 소비하다" 확산하는 토핑경제의 시대

토핑경제란 기본적인 제품이나 서비스에 소비자가 직접 다양한 옵션을 추가해 자신의 개성과 취향을 반영하는 소비 형태를 말합니다. 더스쿠프가 발표한 「고명처럼 얹어라! 지비츠가 보여준 토핑경제」에 따르면, 산업별 토핑경제 확산 사례로, 먼저, 패션 산업에서는 크록스의 '지비츠' 꾸미기 액세서리가 인기를 끌며 크록스 전체 매출에서 차지하는 비중이 2022년 8.0%에서 2023년 17.0%로 증가했으며, 식음료 산업에서는 요거트 아이스크림 브랜드 '요아정'이 대표적인 사례인데요, 50여 가지 토핑을 제공하며 나만의 조합을 즐길 수 있어 큰 인기를 끌었습니다. 실제로 매출 변화를 살펴보면 2023년 51억 원에서 2024년 471억 원으로 9배가량 증가했습니다. IT·디지털 산업에서도 스마트폰 케이스, 노트북 커버, 태블릿 액세서리 등 개인 맞춤형 디자인 서비스가 확대되고 있으며 일상 속에서도 개성과 취향을 표현하는 수단으로 자리 잡았습니다.

중고거래에서 시작된 놀이문화

문화체육관광부의 「2025년 국민여가활동조사」에서 우리 국민 만 15세 이상을 대상으로 조사한 결과에 따르면, 56.6%는 '혼자서' 여가활동을 보낸다고 응답했고, 29.4%는 '가족과 함께', 11.6%는 '친구나 연인'과 함께 즐긴다고 나타났습니다. 이에 비해 낯선 타인과 즉흥적으로 어울리는 경도 놀이의 확산은 기존의 여가 방식과는 매우 다른 양상으로 보입니다. 온라인·비대면 소통이 일상이 된 현대 사회를 통해 경도 놀이 유행의 배경을 살펴볼 수 있습니다. 한국언론진흥재단의 「2024 소셜미디어 이용자 조사」에 따르면, 이용자의 약 절반인 48.1%가 인스타그램을 매일 이용하는 것으로 나타났으며, 2024년 국가데이터처 「사회조사」와 「2025 국민여가활동조사」를 보면, 많은 성인들이 일상 속에서 스트레스를 느끼며 살아가고 있다는 걸 알 수 있습니다. 이런 흐름 속에서 '경도 놀이'는 부담 없이 즐길 수 있는 스트레스 해소 방법 중 하나가 되는 것입니다.

관객이 완성하는 전시의 시대

문화체육관광부의 「2024년 국민문화예술활동조사」에 따르면, 국민의 63.0%가 지난 1년간 한 번 이상 문화예술행사를 관람한 것으로 나타났습니다. 코로나19의 영향이 본격화된 2020년 이후 관람률이 감소했습니다. 특히 2021년에는 33.6%까지 떨어지며 최저치를 기록했는데, 이후 점차 회복세를 보이며 2024년에는 63.0%까지 다시 올라왔습니다. 관람 횟수를 살펴보면, 2024년 기준으로 전 국민의 1인당 평균 관람 횟수는 2.6회로 나타났습니다. 반면 문화예술행사를 한 번 이상 관람한 사람들만 보면 평균 4.1회로 더 높아졌음을 알 수 있습니다. 연도별 관람 횟수를 살펴보면, 전체 국민의 평균 관람 횟수는 2021년 1.4회로 크게 줄었다가 이후 꾸준히 증가해 2024년 2.6회까지 회복되었습니다. 문화예술행사 관람자 기준으로는 2020년 5.1회에서 2024년 4.1회로 다소 감소했지만, 여전히 문화예술을 꾸준히 즐기는 관람층이 있음을 보여줍니다.

많이 찾는 문화예술행사는?

여러분은 일상에서 어떤 문화생활을 하고 계신가요? 문화체육관광부의 「2024년 국민문화예술활동조사」에 따르면, '영화'의 관람률이 57.0%로 가장 높게 나타났으며, 이어 '대중음악/연예'가 14.6%, '뮤지컬' 6.4%, '연극' 5.9%, '미술전시회' 5.6% 순으로 조사되었습니다. 이를 통해 문화예술행사 분야 별로 관람 경험에 차이가 있으며, 일부 분야에 관람이 집중되는 경향을 보였습니다. 같은 조사에 따르면, 대중음악/연예 관람률은 매해 증가하며, '2021년' 1.3%에서 '2024년' 14.6%로 꾸준히 상승한 것으로 나타났습니다. 지난 3월 21일 저녁 8시 광화문 광장, 3년 9개월의 공백을 깨고 화려하게 복귀한 BTS의 'ARIRANG'이 전 세계 팬들을 사로잡았습니다. 전 세계적인 K-팝 아티스트들의 해외 공연과 글로벌 음원차트 성과, 국내 연예 콘텐츠의 해외 플랫폼 확산 등은 우리나라의 대중음악과 연예가 국경을 넘어 주목받고 있습니다.

생성형 AI를 마켓 리서치에 활용하기

구자룡 | 밸류바인 대표

기존의 시장 조사(Market Research)가 수개월의 시간과 막대한 비용을 들여 '사람'의 뒤를 쫓았다면, 이제는 인공지능(AI)이 그 자리를 대신하며 단 몇 분 만에 통찰을 제시하는 시대가 되었다. 최근의 변화는 단순히 분석의 속도가 빨라진 수준을 넘어, 리서치의 패러다임 자체가 바뀌고 있다. 이번 칼럼에서는 생성형 AI가 마켓 리서치의 지형을 어떻게 뒤흔들고 있는지, 그리고 일반인들도 실무에서 이를 어떻게 활용할 수 있는지 구체적으로 살펴보고자 한다.

1. 리서치도 이제 생성형 AI 시대

시장 조사는 오랫동안 기업의 중요한 의사 결정 도구였다. 신제품을 출시하기 전에 소비자가 무엇을 원하는지 확인하고, 광고 캠페인을 집행하기 전에 어떤 메시지가 더 설득력 있는지 검토하며, 고객 만족도 조사를 통해 서비스 개선 방향을 찾는 일이 모두 시장 조사의 영역이었다. 과거에는 이러한 조사를 수행하기 위해 조사 설계자, 조사 업체, 설문 시스템, 통계 분석가가 필요했다. 시간도 적지 않게 걸렸다. 설문지를 설계하고, 응답자를 모집하고, 데이터를 수집하고, 통계 분석을 거쳐 분석 보고서를 작성하는 데 몇 주에서 몇 달이 걸리는 경우도 흔했다.

그러나 생성형 AI가 등장하면서 마켓 리서치의 방식이 빠르게 바뀌고 있다. 생성형 AI는 단순히 설문 문항을 대신 써주는 도구에 머물지 않는다. 조사 목적을 정리하고, 가설을 만들고, 응답자 특성을 설계하고, 질문지를 구성하고, 인터뷰 질문을 만들고, 수집된 응답을 요약하며, 보고서 초안까지 작성한다. 더 나아가 실제 사람이 아닌 '가상 고객 또는 합성 고객(Synthetic Customer)'을 만들어 초기 제품 아이디어나 광고 메시지에 대한 반응을 시뮬레이션하는 단계까지 확장되고 있다.

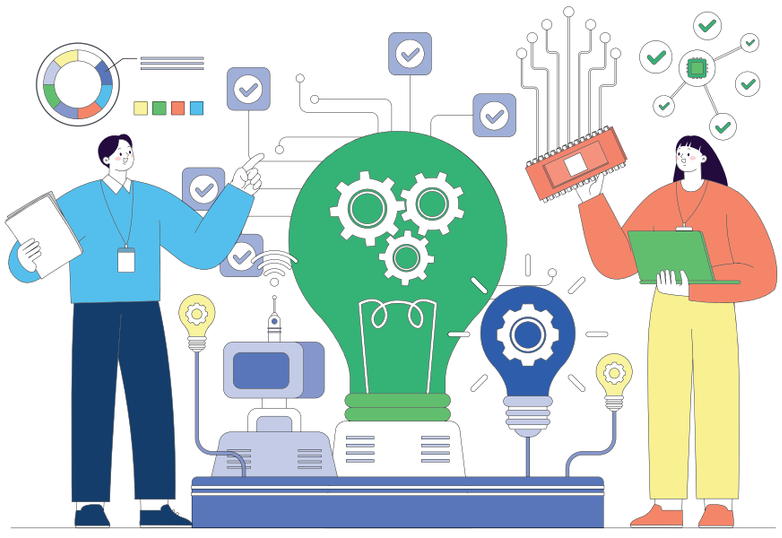
[그림1] 가상 고객 인터뷰(제미나이에서 이미지 생성)

이 변화는 기존 시장 조사의 가치를 부정하는 것이 아니다. 오히려 시장 조사의 본질을 더 분명하게 드러낸다. 빅데이터는 고객이 "무엇을 했는지"는 보여주지만, "왜 그렇게 했는지"를 설명하는 데 한계가 있다. 서베이와 인터뷰는 이 '왜'를 확인하기 위한 대표적인 조사 방법이다. 생성형 AI는 바로 이 과정, 즉 질문을 만들고, 응답을 해석하고, 의미를 정리하는 과정을 더 빠르고 넓게 확장하는 도구가 되고 있다.

하버드비즈니스리뷰(HBR)는 생성형 AI가 시장 조사에서 네 가지 기회를 제공한다고 설명한다. 첫째, 기존 조사 업무를 더 빠르고 저렴하게 수행하게 한다. 둘째, 합성 데이터와 합성 고객을 활용해 일부 조사 방식을 대체하거나 보완한다. 셋째, 기존 데이터가 부족해 직관에 의존했던 영역에 새로운 근거를 제공한다. 넷째, 디지털 트윈과 같은 새로운 형태의 고객 실험을 가능하게 한다. 또한 170명 이상의 시장 조사 실무자와 이용자를 대상으로 한 조사에서 45%가 이미 생성형 AI를 데이터와 인사이트 업무에 활용하고 있고, 또 다른 45%가 향후 활용할 계획이라고 답했다.

물론 여기서 중요한 점은 "AI가 시장 조사를 완전히 대체한다"라는 식의 단순한 결론이 아니다. 오히려 더 현실적인 결론은 "AI가 시장 조사의 초기 단계와 반복 작업을 크게 줄여주고, 사람은 더 중요한 판단에 집중하게 된다"라는 것이다. 질문지를 처음부터 끝까지 사람이 혼자 작성하던 시대에서, 이제는 AI가 초안을 만들고 사람이 검토하는 방식으로 바뀌고 있다. 수십 개의 제품 콘셉트를 한 번에 사람에게 물어보기 어려웠던 시대에서, 이제는 가상 고객에게 먼저 물어본 뒤 가능성 높은 아이디어만 실제 고객에게 검증하는 방식이 가능해지고 있다.

따라서 생성형 AI 시대의 마켓 리서치는 '조사를 안 해도 되는 시대'가 아니라, '더 자주, 더 빠르게, 더 넓게 조사할 수 있는 시대'라고 보는 것이 정확하다. 문제는 도구의 성능이 아니라 사용자의 판단력이다. AI가 제안한 결과를 그대로 믿으면 위험하지만, 올바른 조사 설계와 검증 절차 속에 넣으면 매우 강력한 리서치 보조 도구가 된다.



2. AI를 활용한 조사 설계 방법

시장 조사의 첫 단계는 조사 설계다. 조사 설계란 무엇을 밝히기 위해, 누구에게, 어떤 방식으로, 어떤 질문을 던질 것인지를 정하는 과정이다. 좋은 시장 조사는 좋은 질문에서 출발한다. 질문이 모호하면 응답도 모호해지고, 응답이 모호하면 분석 결과도 흔들린다. 따라서 AI를 활용하더라도 조사 설계의 기본 원칙은 달라지지 않는다. 다만 그 과정을 훨씬 효율적으로 수행할 수 있다.

AI를 활용한 조사 설계는 크게 다섯 단계로 정리할 수 있다. 첫째, 비즈니스 문제를 명확히 정의한다. 예를 들어 "신제품으로 건강기능식품을 출시하고 싶다"라는 표현은 아직 조사 문제가 아니다. "40~50대 직장인이 피로 회복용 건강기능식품을 선택할 때 가장 중요하게 보는 요인은 무엇인가?"처럼, 조사 가능한 질문으로 바꾸어야 한다. 둘째, 조사 목적을 설정한다. 인지도 파악인지, 구매 의향 확인인지, 가격 민감도 분석인지, 고객 불만 탐색인지에 따라 질문 방식이 달라진다. 셋째, 가설을 세운다. 예를 들어 "소비자는 기능성보다 원료의 신뢰성을 더 중요하게 생각한다"와 같은 가설을 만들 수 있다. 넷째, 필요한 정보 목록을 작성한다. 다섯째, 조사 방법과 대상자를 결정한다.

이 과정에서 생성형 AI는 매우 유용하다. 조사 목적을 입력하면 AI는 예상 가설, 핵심 질문, 응답자 조건, 문항 구조, 분석 방법을 제안할 수 있다. 예를 들어 다음과 같이 요청할 수 있다.

"건강기능식품 신제품 개발을 위해 40~50대 직장인을 대상으로 소비자 조사를 하려고 한다. 조사 목적은 피로 회복 제품에 대한 구매 동기, 기대 효능, 가격 민감도, 기존 제품 불만을 파악하는 것이다. 이 조사를 위한 조사 설계서를 작성해 줘. 조사 목적, 핵심 가설, 조사 대상, 주요 질문, 분석 방법을 포함해 줘."

이런 프롬프트를 사용하면 AI는 조사 설계서의 초안을 빠르게 작성해 준다. 하지만 여기서 끝내면 안 된다. AI가 만든 조사 설계서는 '초안'이지 '최종안'이 아니다. 특히 조사의 목적, 표본의 대표성, 응답자 조건, 실제 현업에서 필요한 의사결정 기준은 사람이 반드시 검토해야 한다. 오픈서베이의 자료에서도 조사 설계는 가설 수립, 필요한 정보 목록, 조사 방법론과 대상자 결정의 순서로 진행되며, AI가 자료 탐색과 아이데이션(Ideation)에는 도움을 줄 수 있지만 부문별 이해관계와 내부 맥락 조율은 여전히 담당자의 역할로 남는다고 설명한다.

질문지 작성에서도 AI는 강력하다. AI는 객관식, 리커트 척도, 순위형, 서술형 문항을 목적에 맞게 제안한다. 예를 들어 브랜드 만족도를 측정하려면 "전반적 만족도", "품질 만족도", "가격 만족도", "재구매 의향", "추천 의향"과 같은 문항을 구성할 수 있다. 신제품 콘셉트 조사를 하려면 "콘셉트 이해도", "새로움", "구매 의향", "가격 적절성", "우려 사항" 등을 포함할 수 있다.

생성형 AI를 조사 설계에 활용할 때 가장 좋은 방식은 "초안 생성 → 비판적 검토 → 수정 요청 → 파일럿 테스트 → 최종 확정"의 순서다. AI에게 처음부터 완벽한 질문지를 요구하기보다, 여러 차례 대화를 통해 문항을 다듬어야 한다. 특히 "중복 문항을 제거해 줘", "응답자에게 어렵게 느껴질 수 있는 표현을 쉬운 말로 바꿔 줘", "유도 질문이 있는지 검토해 줘", "각 문항이 어떤 분석 목적과 연결되는지 표로 정리해 줘"와 같은 후속 요청이 필요하다.

3. AI 가상 고객 사전 인터뷰 방법

최근 마켓 리서치에서 주목받는 변화 중 하나는 가상 고객, 즉 합성 고객의 등장이다. 합성 고객은 실제 고객 데이터를 바탕으로 특정 고객 유형을 가상으로 만든 뒤, 그 고객이 실제 사람처럼 질문에 답하도록 설계한 AI 기반 페르소나다. 단순한 상상 속 인물이 아니라 연령, 직업, 소득, 생활 방식, 구매 경험, 가치관, 브랜드 태도 등을 반영해 만든 시뮬레이션 고객이다. 합성 응답자는 객관식 응답뿐 아니라 선택 이유나 감정적 배경에 대해서도 질문할 수 있다.

가상 고객 인터뷰의 장점은 다음과 같다. 첫째, 매우 빠르다. 실제 고객 인터뷰는 응답자를 모집하고 일정을 조율하고 인터뷰를 진행해야 한다. 반면 가상 고객 인터뷰는 몇 분 안에 여러 명의 고객 반응을 얻을 수 있다. 둘째, 비용이 덜 든다. 셋째, 초기 아이디어를 걸러내는 데 유용하다. 아직 완성되지 않은 제품 콘셉트나 광고 문구를 실제 고객에게 바로 보여주기가 부담스러운 경우, 가상 고객에게 먼저 반응을 물어볼 수 있다. 넷째, 다양한 고객 유형을 동시에 비교할 수 있다. 예를 들어 20대 미혼 직장인, 30대 맞벌이 부부, 50대 건강 관심층, 60대 은퇴자를 각각 가상 고객으로 만들고 같은 질문에 대한 반응을 비교할 수 있다.

실무에서 가상 고객 사전 인터뷰는 다음과 같이 진행할 수 있다. 먼저 조사하고 싶은 제품이나 서비스의 콘셉트를 정리한다. 예를 들어 “수면의 질을 개선하는 기능성 음료”라고 하자. 다음으로 타깃 고객을 정의한다. “수면 부족을 경험하는 30대 직장인 여성”, “야근이 잦은 40대 남성 관리자”, “건강관리에 관심이 많은 50대 여성”처럼 구체적으로 설정한다. 세 번째로 각 고객의 페르소나를 만든다. 페르소나는 이름, 나이, 직업, 생활 패턴, 건강 고민, 구매 습관, 브랜드 태도, 가격 민감도 등을 포함해야 한다. 네 번째로 인터뷰 질문을 던진다. “이 제품을 처음 봤을 때 어떤 느낌이 드는가?”, “구매를 망설이게 하는 요인은 무엇인가?”, “기존 제품과 비교해 어떤 점이 차별적으로 느껴지는가?”, “어느 정도 가격이면 구매를 고려하겠는가?”와 같은 질문이다. 마지막으로 응답을 비교 분석하여 공통된 기대, 불만, 우려, 구매 장벽을 정리한다.

그리고 가상 고객과의 심층 인터뷰를 위한 스크립트도 AI 기반으로 작성할 수 있다. 예를 들어, 챗GPT에 “가상의 고객과 사전 인터뷰를 진행하고자 한다. 가상 고객인 **‘박지연**(Ji-yeon Park, 28세, UX 디자이너)’이라는 페르소나를 대상으로, 새로운 제품 개발(콘셉트: 수면의 질을 개선하는 기능성 음료)에 대한 피드백을 받는 상황을 가정하여 인터뷰 질문지를 작성해 줘”라고 요청하면 된다.

이때 중요한 점은 가상 고객을 ‘진짜 고객’으로 착각하지 않는 것이다. 가상 고객은 실제 고객을 대신하는 최종 판단자가 아니다. 초기 가설을 만들고, 질문을 더 정교하게 다듬고, 실제 조사 전에 위험 요인을 미리 찾는 도구다. HBR 관련 자료에서도 LLM 기반 합성 고객은 제품 아이디어를 빠르게 테스트하고 약한 아이디어를 조기에 걸러내는 데 유용하지만, 인간 조사를 완전히 대체해서는 안 된다고 경고한다. 특히 감정적 요인, 인구통계별 차이, 신흥 시장 요인에서는 실제 사람의 미묘한 반응을 충분히 반영하지 못할 수 있다.

Synthetic Interview 시뮬레이션

주제: 수면의 질을 개선하는 기능성 음료

인터뷰 대상 페르소나

- 이름: 박지연(Ji-yeon Park)
- 나이: 28세
- 직업: UX 디자이너
- 라이프스타일:
  - 디지털 기기 사용 시간이 길다
  - 업무 몰입도가 높아 밤늦게까지 화면을 보는 일이 많다
  - 자기관리에는 관심이 있지만 루틴이 아주 안정적이지는 않다
  - 건강기능 제품에 관심은 있으나, 과장된 효능 표현에는 냉소적이다
  - 감성적 경험과 실질적 효용을 둘 다 중요하게 생각한다

---

1. 도입: 수면 문제의 생활 맥락 파악

사회자

평소 하루 일과가 어떻게 흘러가나요? 특히 밤 시간대 루틴이 궁금합니다.

지연

출근하는 날은 보통 7시쯤 일어나고, 회사에서 일하다가 저녁 먹고 집에 와도 완전히 일이 끝난 느낌은 아니에요. UX 디자이너라 그런지 머릿속에서 계속 화면이나 흐름을 정리하게 되거든요. 집에 와서도 핸드폰 보거나, 레퍼런스 조금 더 보거나, 유튜브를 보다가 자는 경우가 많아요.
자려고 누워도 머리가 바로 꺼지는 느낌은 아니고, 몸은 피곤한데 생각은 계속 돌아가는 느낌이 있어요.

[그림2] 가상 고객 대상 인터뷰 스크립트와 답변 내용(챗GPT에서 생성)

가상 고객 인터뷰의 품질을 높이려면 세 가지가 필요하다. 첫째, 페르소나를 막연하게 만들지 않아야 한다. “30대 여성”보다 “서울 거주, 34세, 맞벌이, 초등학생 자녀 1명, 평일 저녁 시간이 부족하고, 건강기능식품은 온라인 리뷰를 보고 구매하는 직장인 여성”이 훨씬 좋은 페르소나다. 둘째, 제품 정보를 충분히 제공해야 한다. AI는 제공받은 정보 안에서 답을 만들기 때문에 제품의 기능, 가격, 사용 상황, 경쟁 제품 정보를 구체적으로 알려줘야 한다. 셋째, 반드시 반대 질문을 던져야 한다. “이 제품의 장점은 무엇인가?”만 물으면 긍정적인 답변이 많이 나온다. “왜 구매하지 않을 것 같은가?”, “가장 믿기 어려운 표현은 무엇인가?”, “광고 문구 중 과장되어 보이는 부분은 무엇인가?”처럼 부정적 반응을 의도적으로 끌어내야 한다.

결국 가상 고객 사전 인터뷰의 핵심은 답을 얻는 것이 아니라, 실제 고객에게 물어볼 더 좋은 질문을 발견하는 것이다. 가상 고객은 시장의 정답이 아니라 리서치 설계의 예행연습이다.

4. AI를 활용한 서베이 자동화

AI를 활용한 서베이 자동화는 크게 네 영역에서 이루어진다. 첫째, 설문지 자동 생성이다. 둘째, 온라인 설문 도구와의 연계다. 셋째, 응답 과정의 대화형 전환이다. 넷째, 응답 데이터의 자동 분석과 보고서 작성이다.

가장 기본적인 활용은 설문 문항 생성이다. 조사 목적, 대상자, 분석 목적을 입력하면 AI는 문항 초안을 제안한다. 예를 들어 "신제품 커피 브랜드의 구매 의향 조사를 위한 구글 설문지 문항을 만들어 줘"라고 요청하면 AI는 브랜드 인지도, 커피 구매 빈도, 선호 맛, 가격 수용도, 패키지 선호도, 구매 의향, 추천 의향 등을 포함한 설문지를 작성해 준다. 이후 이 문항을 구글 폼, 네이버 폼, 카카오 비즈니스 폼 등에 입력해 설문을 만들 수 있다.

두 번째는 설문 제작 과정의 자동화다. 실무자는 AI가 만든 문항을 그대로 복사해 온라인 설문 도구에 입력할 수 있다. 더 나아가 구글 앱스 스크립트나 API를 활용하면 AI가 생성한 설문 문항을 구글 폼 구조에 맞게 자동으로 변환할 수도 있다. 전문 개발자가 아니더라도 AI에게 "이 문항들을 구글 폼에 입력하기 쉬운 표 형태로 정리해 줘. 문항 번호, 질문, 질문 유형, 선택지, 필수 응답 여부를 포함해 줘"라고 요청하면 설문 제작 시간이 크게 줄어든다.

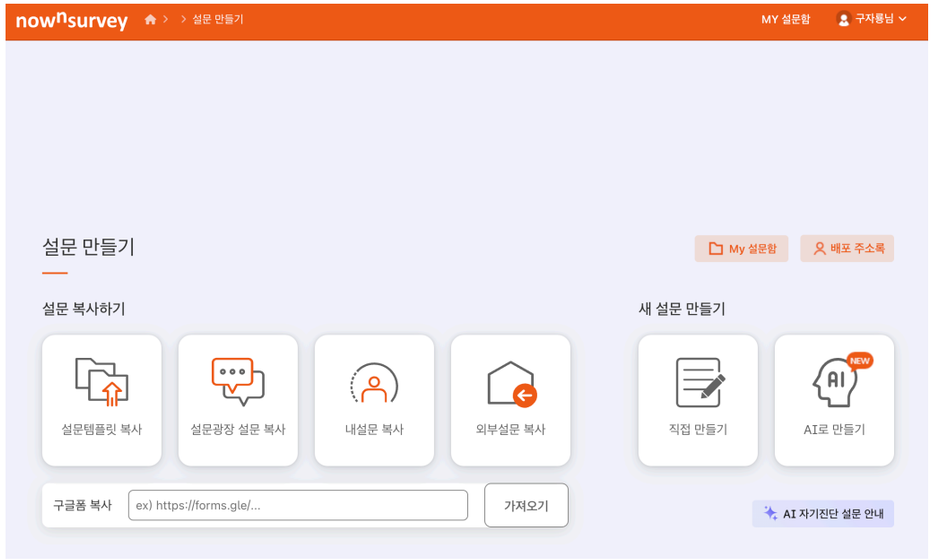
세 번째는 대화형 서베이다. 전통적인 설문조사는 응답자가 정해진 문항에 순서대로 답하는 방식이다. 하지만 대화형 서베이는 응답자의 답변에 따라 AI가 후속 질문을 던진다. 예를 들어 응답자가 "가격이 비싸다"라고 답하면 AI는 "어느 정도 가격이면 적절하다고 느끼십니까?"라고 추가로 묻는다. 응답자가 "포장이 불편하다"라고 답하면 "어떤 상황에서 가장 불편했습니까?"라고 물을 수 있다. 이는 정량 조사의 규모와 정성 조사의 깊이를 결합하는 방식이다. HBR에 의하면 아웃셋 AI(Outset.ai)와 같은 AI 모더레이션 플랫폼이 응답자의 이전 답변을 바탕으로 새로운 질문을 동적으로 제시해 자동화된 설문조사의 속도와 전통적 인터뷰의 깊이를 결합한다고 설명한다.

네 번째는 응답 데이터 분석 자동화다. 설문 응답이 수집되면 AI는 기술 통계 분석, 교차 분석, 상관 분석, 집단 간 차이 검정, 텍스트 응답 요약, 주요 키워드 추출, 감성 분석, 보고서 작성까지 지원할 수 있다. 예를 들어 "응답 데이터를 분석해서 핵심 발견 5가지를 정리하고, 마케팅 시사점과 신제품 개발 방향을 제안해 줘"라고 요청할 수 있다. 서술형 응답이 많을 때 "불만 요인을 제품, 가격, 유통, 커뮤니케이션, 사용 경험으로 분류해 줘"와 같이 요청하면 된다.

다만 서베이 자동화에는 위험도 있다. 자동화가 쉬워질수록 부실한 질문지도 쉽게 만들어진다. 조사 목적이 불명확한 상태에서 문항만 많이 만들면 데이터는 쌓이지만, 쓸모 있는 인사이트는 나오지 않는다. 또한 AI가 생성한 분석 결과를 검산하지 않으면 잘못된 해석이 보고서에 포함될 수 있다. 예를 들어 평균 차이가 작고 통계적으로 유의하지 않은데도 "여성이 남성보다 만족도가 높다"라는 식으로 과도하게 해석할 수 있다. 따라서 서베이 자동화의 목표는 '사람을 빼는 것'이 아니라 '사람이 더 중요한 판단에 집중하게 하는 것'이어야 한다.

AI 서베이 자동화의 실무 절차는 다음과 같이 정리할 수 있다. 문제 정의, 조사 목적 설정, 가설 수립, 질문지 초안 생성, 문항 검수, 파일럿 테스트, 온라인 설문 배포, 응답 데이터 수집, AI 기반 분석, 사람의 해석과 검증, 최종 보고서 작성이다. 이 절차에서 AI는 초안 작성자, 보조 분석가, 요약가, 보고서 편집자의 역할을 한다. 그러나 조사 책임자는 여전히 사람이다.

그리고 AI를 활용해 설문 기획, 생성, 분석까지 자동화하여 리서치 효율을 높인 서비스로 오픈서베이와 나우앤서베이가 있다. 오픈서베이는 고도화된 데이터 분석과 사전 예측에, 나우앤서베이는 생성형 AI 기반의 간편한 설문 제작과 실시간 분석에 강점이 있다. 이러한 서비스들은 기술적 장벽을 낮추고, 설문 조사 과정을 훨씬 빠르고 정확하게 만들어 기업들의 데이터 기반 의사 결정을 지원한다.



[그림3] 나우앤서베이의 AI로 설문 만들기

5. AI 마켓 리서치의 미래

AI 마켓 리서치의 미래는 더 빠른 조사, 더 많은 실험, 더 정교한 고객 시뮬레이션으로 나아갈 것이다. 앞으로 기업은 하나의 신제품 아이디어만 조사하지 않고, 수십 개의 콘셉트와 가격 조합, 광고 메시지, 패키지 디자인을 동시에 검토하게 될 가능성이 높다. 실제 고객 조사는 여전히 중요하지만, 그 전에 AI를 활용한 사전 검토가 일상화될 것이다. 즉, 시장 조사의 순서가 "사람에게 먼저 묻고 분석한다"에서 "AI로 먼저 걸러내고 사람에게 검증한다"로 바뀔 것이다.

HBR에 의하면, 합성 고객이 전통적 인간 연구를 완전히 대체하지는 못하지만, 제품 아이디어를 빠르게 탐색하고 수많은 콘셉트를 초기 필터로 검토하는 데 유용하다고 설명한다. 특히 기업 내부의 과거 설문 데이터와 고객 데이터를 활용해 자체 고객 시뮬레이터를 만들면 더 정교한 초기 인사이트를 얻을 수 있다. 그러나 초기 트렌드 탐지 이상의 의사결정에서는 여전히 인간 조사가 필수라고 강조한다.

미래의 시장 조사는 세 가지 방향으로 발전할 가능성이 높다. 첫째, 리서치의 민주화다. 과거에는 전문 조사기관만 수행할 수 있었던 조사 설계와 분석을 이제 일반 실무자도 AI와 온라인 도구를 활용해 수행할 수 있다. 둘째, 리서치의 상시화다. 연 1~2회 대규모 조사를 하는 방식에서 벗어나, 필요할 때마다 짧고 빠르게 고객 반응을 확인하는 방식이 늘어날 것이다. 셋째, 리서치의 개인화다. 전체 고객 평균만 보는 것이 아니라 세그먼트별, 페르소나별, 심지어 개별 고객의 디지털 트윈을 활용해 반응을 예측하는 방향으로 진화할 것이다.

하지만 이 변화에는 반드시 경계해야 할 지점이 있다. AI는 그럴듯한 답을 잘 만든다. 바로 이 점이 장점이자 위험이다. AI가 생성한 가상 고객 응답은 실제 고객의 목소리가 아니다. 실제 데이터를 바탕으로 한 시뮬레이션일 뿐이다. 따라서 AI 리서치 결과를 그대로 시장의 진실로 받아들이는 것은 위험하다. 특히 신제품 출시, 가격 결정, 브랜드 포지셔닝처럼 큰 비용이 들어가는 의사결정에서는 반드시 실제 고객 데이터와 교차 검증해야 한다.

AI 마켓 리서치의 핵심 원칙은 "대체가 아니라 결합"이다. AI는 빠르게 가설을 만들고, 초안을 구성하고, 약한 아이디어를 걸러내고, 응답을 요약하는 데 강하다. 사람은 맥락을 이해하고, 질문의 품질을 판단하고, 결과의 의미를 해석하고, 최종 의사결정의 책임을 진다. 이 둘을 분리해서 생각하면 AI는 위험한 자동화 도구가 되지만, 결합해서 활용하면 매우 강력한 시장 이해 도구가 된다.

결국 생성형 AI 시대의 마켓 리서치 역량은 도구를 얼마나 잘 다루는가에 그치지 않는다. 더 중요한 것은 무엇을 물어야 하는지 아는 능력이다. 좋은 질문이 없으면 좋은 데이터도 없다. 좋은 데이터가 없으면 좋은 분석도 없다. 좋은 분석이 없으면 좋은 전략도 없다. 생성형 AI는 질문을 대신 만들어줄 수 있지만, 어떤 질문이 중요한지는 사람이 판단해야 한다. 이선 몰릭 교수가 제안했듯이 생성형 AI 시대에 인간의 고유한 능력과 AI의 능력을 융합하여 상호 보완적인 파트너십을 구축하는 공동 지능(Co-Intelligence)을 목표로 해야 한다. 그리고 생성형 AI로 마켓 리서치를 더 쉽게 만들지만, 결코 가볍게 만들어서는 안 된다. 마켓 리서치는 여전히 고객을 이해하기 위한 신중한 과정이다. 달라진 것은 도구이고, 변하지 않는 것은 고객을 제대로 이해해야 한다는 원칙이다.

참고문헌

구자룡(2024).『AI 데이터 분석』. 커뮤니케이션북스.

구자룡(2024).『데이터 마인드 기르는 습관』. 좋은습관연구소.

구자룡(2025). "AI를 활용하여 서베이 및 고객 데이터 분석하기". 통계의창.

구자룡(2025).『챗GPT로 시작하는 데이터 리터러시』. 마들렌북.

나우앤서베이. 'AI 자기진단 설문 서비스' 소개서. https://www.nownsurvey.com/

오픈서베이(2024). "고객 조사에 활용하는 생성형 AI 기반 MI 접근법". https://blog.opensurvey.co.kr/research-tips/genai-mi-3/

이명식·구자룡·양석준(2017).『마케팅 리서치(개정판)』. 형설출판사.

이선 몰릭(2025). 듀얼 브레인. 신동숙 옮김. 상상스퀘어. Ethan Mollick(2024). Co-Intelligence: Living and Working with AI, Portfolio.

James Brand, Ayelet Israeli and Donald Ngwe(2025). "Larger, Faster, Cheaper: The Future of Market Research with AI". Harvard Business Review. https://d3.harvard.edu/larger-faster-cheaper-the-future-of-market-research-with-ai/

James Brand, Ayelet Israeli and Donald Ngwe(2025). "Using Gen AI for Early-Stage Market Research". Harvard Business Review. https://hbr.org/2025/07/using-gen-ai-for-early-stage-market-research

Jeremy Korst, Stefano Puntoni and Olivier Toubia(2025). "How Gen AI Is Transforming Market Research". Harvard Business Review. https://hbr.org/2025/05/how-gen-ai-is-transforming-market-research

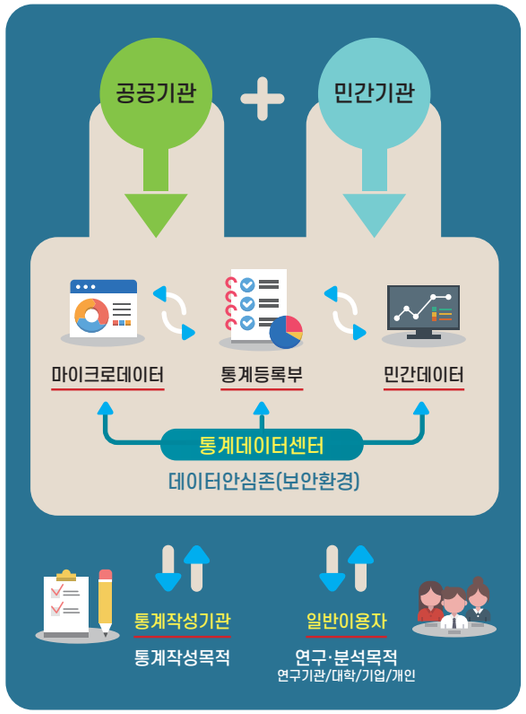
SDC
Statistics
Data Center

행정통계자료와 민간자료를 한곳에!

통계데이터센터 서비스

통계데이터센터가 새로운 서비스로 정보화 사회를 선도합니다.

국가데이터처가 제공하는 통계등록부, 마이크로데이터, 민간데이터 등 다양한 자료를
한 곳에서 연계 및 분석이 가능한 통계데이터센터(SDC)



1 분석 플랫폼 제공 서비스

- 분석시스템 · 통계패키지 제공
- 통계자료(통계등록부 · 통계기초자료) 및 민간자료, 이용자 반입자료 연계 · 분석

2 전문가 분석지원 서비스

- 분석 경험이 없는 이용자를 위한 데이터 분석 지원
- 센터 이용 상담 및 데이터 분석 자문

3 주문형 분석서비스

- 시간 및 거리상 센터 방문이 어렵거나 직접 자료분석을 하기 힘든 이용자를 위한 서비스
- 센터 이용자료를 활용하여 연계 · 분석 후 이용자가 원하는 형태로 결과를 제공

4 명부 서비스(승인통계대상)

- 분석센터로 방문하여 자료분석 및 표본설계를 통해 데이터 반출

5 이용자 교육 서비스

- 이용자 교육 홈페이지 운영
- 통계분석 프로그램 및 분석사례 교육
- 매년 통계데이터 활용대회 개최

국가데이터처, 정부부처, 지방자치단체, 연구기관 등

모든 기관의 마이크로데이터를 한 곳으로

보다 심도 있고 다양한 분석을 원한다면
지금 바로 MDIS를 클릭해 보세요.

■ 서비스 소개 (2026년 1월 기준)

가. 서비스명 : 마이크로데이터통합서비스(MDIS, mdis.mods.go.kr)
나. 제공 통계 수 : 174개 작성기관 통계 총 402종 제공(국가데이터처 50종)
다. 제공 형태 : 마이크로데이터(조사대상에 따라 인구, 사업체, 가구 기반 자료)

기준		통계명
국가데이터처 (50종)	인구·가구 (15종)	가계금융복지조사, 가계동향조사, 가구소비실태조사, 경제활동인구조사, 국내인구이동통계, 녹색생활조사, 사망원인통계, 사회조사, 생활시간조사, 이민자 체류실태 및 고용조사, 인구동향조사, 인구총조사, 주택총조사, 지역별고용조사, 초중고 사교육비조사
	사업체 (11종)	건설업조사, 경제총조사, 광업제조업조사, 기업활동조사, 도소매업조사, 서비스업조사, 서비스업총조사, 운수업조사, 전국사업체조사, 전문과학기술서비스업조사, 프랜차이즈조사
	농림어업 (12종)	농가경제조사, 농림어업조사, 농림어업총조사, 농산물생산비조사, 농어업법인조사, 농업면적조사, 농작물생산조사, 세종시특별센서스, 양곡소비량조사, 어가경제조사, 어류양식동향조사, 어업생산동향조사
	행정통계 (12종)	귀농어 · 귀촌인통계, 기업생멸행정통계, 생애단계별행정통계, 신혼부부통계, 연금통계, 영리법인기업체행정통계, 육아휴직통계, 일자리이동통계, 일자리행정통계, 임금근로일자리동향행정통계, 주택소유통계, 소득이동통계
통계작성기관 (352종)		전국다문화가족실태조사, 가족실태조사, 노인실태조사, 자동차주행거리통계, 직종별사업체노동력조사, 국민여가활동조사, 외래관광객조사, 가공식품소비자태도조사 외 344종

■ 서비스 내용

가. 구분 : 자료의 민감성 정도에 따라 공공용, 인가용으로 구분 운영

나. 서비스 방법
- 공공용
  - 다운로드 서비스 : MDIS 포털에서 직접 무료로 다운로드
  - 온라인분석 서비스 : MDIS 포털 내 분석 시스템에서 무료로 추출 · 편집 · 분석
- 인가용
  - 원격접근 서비스 : 승인 후 이용자가 집 · 사무실 등에서 국가데이터처 서버 접속 후 활용
  - 센터 서비스 : 승인 후 이용자가 분석센터에 방문하여 활용
  - 주문형 서비스 : 승인 후 이용자에게 맞춤형으로 제공

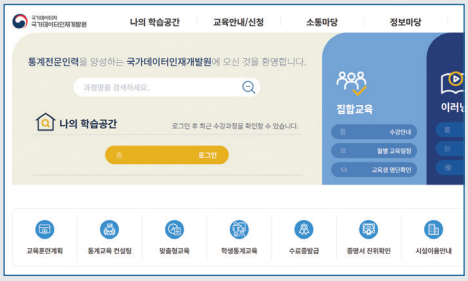
다. 수수료
- 공공용 : 무료
- 인가용 : 유료

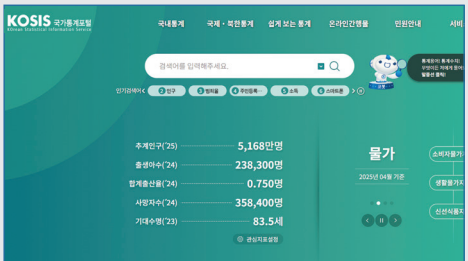
■ 문의

- 연락처 : (재) 한국통계진흥원
- 전화 : (02) 3457-9700, FAX :(02) 515-0240
- 주소 : (06097) 서울특별시 강남구 선릉로 612, 6층
- E-mail : mdis.kspi@gmail.com

국가데이터처에서 국가통계를 활용하세요!

국가데이터처는 통계개발·활용·교육에 필요한 모든 정보와 도움을 제공합니다.
다양한 국가통계정보 제공 사이트를 활용하세요.





국가데이터인재개발원

dshi.mods.go.kr

국내 유일의 국가통계교육 전문기관

통계 작성 및 활용 전문통계과정,
기관맞춤형과정, e-러닝 과정

통계데이터센터

data.mods.go.kr

행정통계자료와 민간자료를 한곳에

행정통계자료(통계등록부), 민간자료의
연계·융합이 가능한 데이터 플랫폼

MDIS

mdis.mods.go.kr

원하는 자료를 직접 분석 및 요청

다운로드 서비스/온라인분석 서비스 선택 시
공공용 마이크로데이터를 무료로 분석 활용 가능

KOSIS

kosis.kr

국가통계 쉽게 찾기

국내, 국제, 북한의 주요 통계를
한 곳에 모아 알기 쉽게 분류해 제공

SGIS

sgis.mods.go.kr

지도 위 통계정보 살펴보기

인구, 가구, 주택, 사업체 통계 등 각종 통계를
지도(GIS) 위에서 한눈에 파악

국가데이터처
국가데이터인재개발원