

2026년 연구보고서

기업통계등록부 정비사업 AI 산업 자동분류 도입방안 연구

2026. 9.



국가데이터처
국가데이터연구원

연구보고서 2026-03

기업통계등록부 정비사업 AI 산업 자동분류 도입방안 연구

최하은 · 임경민 · 이규호



Ministry of Data and Statistics
Data and Statistics Research
Institute

발간사

국가데이터·통계는 사회와 경제의 변화를 정확히 이해하고 미래를 준비하기 위한 핵심 자산입니다. 최근 인공지능(AI)을 비롯한 디지털 기술의 발전과 데이터 활용 환경의 변화는 국가데이터·통계의 생산과 관리, 활용 방식에도 새로운 혁신을 요구하고 있습니다.

국가데이터연구원은 이러한 변화에 능동적으로 대응하여 국가데이터·통계의 정확성과 활용성을 높이고, 데이터에 기반한 정책 수립을 지원하기 위한 연구를 지속적으로 수행해 왔습니다.

2026년 연구보고서에는 국가데이터·통계의 체계적 관리와 인공지능(AI) 기반 통계 생산 혁신, 그리고 우리 사회와 경제의 주요 변화를 심층적으로 분석한 연구 성과를 담았습니다.

첫째, 국가데이터·통계를 체계적으로 관리하고 통계생산의 효율성을 높이기 위한 연구를 강화하였습니다. 국가데이터·통계에서 사용되는 개념과 정보를 일관성 있게 관리하기 위해 국제표준인 GSIM을 활용할 수 있는 방안을 연구하였으며, 또한, 사망 원인통계의 원사인 분류와 기업통계등록부 정비과정의 산업분류에 인공지능(AI) 기술 적용 가능성을 연구하여, 전문적인 판단과 많은 시간이 필요한 통계분류 업무의 정확성과 효율성을 높이기 위한 방안을 제시하였습니다.

둘째, 다양한 데이터를 연계·활용하여 청년층의 경제적 이동과 삶의 변화를 입체적으로 파악하고자 하였습니다. 통계등록부를 활용하여 청년층의 경제적 자립과 지역 이동의 특성을 이해할 수 있는 소득 변화와 이동성을 분석하였습니다. 이러한 연구는 청년정책과 지역정책을 보다 정교하게 설계하기 위한 기초자료로 활용될 수 있을 것입니다.

셋째, 국민의 삶과 지속가능한 발전을 종합적으로 측정하기 위한 연구를 수행하였습니다. 주요국의 삶의 질 지표와 지속가능발전목표(SDG) 지표체계를 비교·분석하여 각국의 지표 구성과 활용방식을 살펴보고, 우리나라 지표체계의 발전 방향과 정책적 활용도를 높이기 위한 시사점을 도출하였습니다. 이는 사회 발전을 경제적 성과뿐만 아니라 삶의 질과 지속가능성의 관점에서 폭넓게 진단하는 데 기여할 것으로 기대합니다.

이번 연구들은 국가데이터·통계의 생산과 관리에 새로운 기술과 국제표준을 적극적으로 접목하는 한편, 다양한 데이터를 활용하여 사회·경제 현상을 심층적으로 이해하려는 노력의 결과입니다.

국가데이터연구원은 앞으로도 최신 기술과 다양한 데이터를 적극 활용하여 국가데이터·통계의 전문성과 신뢰성을 높이고, 시대적 변화에 선제적으로 대응할 수 있는 연구를 이어가겠습니다.

본 연구보고서가 통계 생산자와 이용자 모두에게 실질적인 도움이 되고, 데이터에 기반한 정책 수립과 의사결정에 유용한 자료로 활용되기를 기대합니다.

많은 관심과 성원을 부탁드립니다.

2026년 9월

국가데이터연구원 연구기획실장



목 차

제1장 서론	1
제2장 기업통계등록부 자료처리 현황	2
제1절 기업통계등록부 개요	2
제2절 기업통계등록부 정비사업	4
제3절 기업통계등록부 정비사업의 산업분류	8
제3장 AI통계분류시스템 활용 현황	13
제1절 AI통계분류시스템 개요	13
제2절 AI통계분류시스템 방법론	15
제3절 AI통계분류시스템 운영 모델	20
제4장 AI통계분류시스템 적용 분석	22
제1절 기업통계등록부 정비사업 자료 검토	22
제2절 AI 산업 자동분류 정확도 분석	25
제3절 AI통계분류시스템 실무 활용 방안	31
제5장 결론 및 시사점	39
제1절 연구요약	39
제2절 향후 과제	40
참고문헌	42
Abstract	43

요 약

본 연구는 국가데이터처 행정자료관리과가 매년 수행하는 기업통계등록부 정비사업의 산업분류 코드 부여 업무에 AI 산업 자동분류를 효과적으로 도입하기 위한 방안을 제시하는 것을 목적으로 한다.

기업통계등록부 정비사업은 연간 약 130만 개에 달하는 신규등록 사업체의 사업자등록자료를 대상으로, 매년 새로 채용되는 약 130명의 정비담당자가 전화조사 등 수작업으로 사업체 정보를 검토·보완하는 사업이다. 이 중 산업분류가 가장 빈번하게 수정되는 항목으로, 담당 인력의 경험과 역량에 따라 산업분류 부여에 소요되는 시간과 정확도에 편차가 발생한다는 문제가 있었다. 이에 행정자료관리과는 이러한 편차를 줄이고 정비 업무의 효율성을 높이기 위해 AI통계분류시스템의 산업 자동분류 기능 도입을 검토하였다.

본 연구에서는 2025년 정비사업 자료를 활용해 AI 산업분류의 추론 성능을 분석하였다. 분석 결과, 원자료의 ‘업종·업태’를 활용한 AI 추론 1순위 산업분류와 등록분류 간 일치율은 71.5%였으나, 정비 과정에서 작성된 ‘사업내용’을 활용한 경우 93.1%로 나타났다. 또한 AI 모델의 학습데이터 작성 방식으로 ‘사업내용’을 보다 구체적으로 작성한 결과, 기존 방식의 90.3%에 비해 변경 방식에서는 96.8%로 일치율이 상승하였다. 이를 통해 AI 산업분류의 성능은 입력 정보의 내용과 작성 방식에 따라 달라질 수 있으며, 구체적이고 명확하게 작성한 ‘사업내용’이 AI 추론 성능 향상에 유용한 정보로 활용될 수 있음을 확인하였다.

이러한 분석 결과는 기업통계등록부 정비사업의 산업분류 부여 업무에 AI통계분류시스템을 활용할 경우, 업무 효율성과 산업분류의 일관성을 높일 가능성이 있음을 보여준다. 이에 본 연구에서는 AI 산업 자동분류의 실무 활용을 위해 내검 시스템 효율화, AI통계분류시스템 홈페이지(A code) 활용, 그리고 기업통계등록부 정비 시스템에 실시간 API 연계를 통한 실무 적용 방안을 제시하였다.

주요 용어: AI통계분류시스템, 기업통계등록부 정비사업, 산업분류

제 1 장

서 론

최근 인공지능(AI) 기술은 문서 이해, 자연어 처리, 대규모 언어모델(LLM) 등을 중심으로 빠르게 발전하며 산업 전반의 업무 방식을 근본적으로 변화시키고 있다. 정부도 이러한 흐름에 대응해 2026년 ‘인공지능 기본계획(2026~2028)’을 수립하고 AI 3대 강국(G3) 도약을 국가 전략 목표로 제시하고, 국가 AI컴퓨팅 센터 구축 등 인프라 확충에 대규모 예산을 투입하며 공공부문 전반의 AI 전환(AI Transformation)을 추진하고 있다. 각 부처와 공공기관 역시 반복적이고 규칙 기반적인 수작업 업무를 AI로 대체, 보조하는 방향으로 업무 프로세스를 재설계하는 추세이며, 이는 대량의 텍스트 자료를 다루는 행정 및 통계 분야에서 특히 두드러지게 나타나고 있다.

국가통계 분야도 예외는 아니다. 유엔 유럽경제위원회(UNECE)는 공식통계 현대화를 위한 고위급그룹(HLG-MOS)을 중심으로 머신러닝 활용 시범사업을 추진해 왔으며, 최근에는 대형 언어모델(LLM)을 공식통계 생산에 접목하려는 논의도 확대되고 있다. 여러 국가의 통계기관 또한 조사자료의 통계분류 코딩 과정에 AI 기술을 도입하여 처리 속도와 정확도를 높이려는 시도를 이어가고 있으며, 2022년에 구축하여 꾸준히 운영되어 온 국가데이터처의 AI통계분류시스템은 이러한 국제적 흐름 속에서 해외 통계기관과의 협력 회의에서 우수 사례로 소개되는 등 대외적으로도 주목받고 있다.

이러한 대내외적 AI 활용 확산 흐름과 다년간 축적된 시스템 운영 경험을 바탕으로, 국가데이터연구원과 행정자료관리과는 기업통계등록부 정비사업에도 AI통계분류시스템을 도입할 필요성에 공감하였다. 이에 본 연구는 기업통계등록부 정비사업의 산업분류 부여 업무 현황과 AI통계분류시스템의 활용 현황을 살펴보고, 실제 정비사업 자료를 활용한 AI 추론 정확도 분석을 통해 기업통계등록부 정비사업에 AI 산업 자동분류의 실무 적용 가능성을 검증하는 것을 목적으로 한다.

이를 위해 기업통계등록부 정비사업의 자료처리 체계와 산업분류 부여 방식(제2장), AI통계분류시스템의 방법론 및 운영 모델(제3장)을 검토하고, 정비사업 자료를 이용한 AI 추론 정확도로 정비사업의 효과 및 ‘사업내용’ 작성 방식 개선 효과를 실증적으로 분석하였다(제4장). 이러한 분석 결과를 토대로 AI 산업 자동분류 도입의 시사점과 향후 실무 활용 과제를 제시하고자 한다(제5장).

제 2 장

기업통계등록부 자료처리 현황

제1절 기업통계등록부 개요¹⁾

기업통계등록부(SBR, Statistical Business Registers)는 국가데이터처가 각종 경제통계 작성을 지원하기 위해 다양한 기관의 행정자료와 조사자료를 연계하여 구축한 기업 및 사업체 단위의 모집단 자료이다. 기업통계등록부는 전(全)산업에 걸쳐 공공 및 민간, 영리 및 비영리 기관과 관계없이 행정자료와 통계조사를 통해 수집된 모든 기업 즉, 물리적 장소가 있는 등록·무등록 사업체, 물리적 장소가 없는 등록사업체, 비활동 사업체를 포괄한다. 따라서 산업 및 지역별 표본조사 추출틀로 활용할 수 있으며, 다른 데이터와의 추가 연계를 통해 다양한 분석 목적으로도 이용이 가능하다.

유럽 등 많은 선진국에서 ‘통계등록부’의 중요성을 인식하고 활용을 확대하고 있는 추세²⁾로, 이제 행정자료는 통계생산에 있어 가장 중요한 원천자료가 되고 있다. 이에, 국가데이터처는 2017년부터 연간 기업통계등록부 서비스를 시작하여 매월 기업의 개·폐업 정보를 활용하여 명부를 작성하고, 종사자, 매출액 등 관련 내용을 수록하고 있다.

또한 기업의 모집단 변화를 더욱 시의성 있게 파악할 수 있도록 2019년부터 분기, 2023년부터 월간 기업통계등록부 서비스를 시작하였고, 2026년 12월부터는 연간 기업통계등록부(잠정)도 제공 예정으로, 계속해서 자료의 활용도를 높이기 위해 노력하고 있다.

<표 1> 연간, 분기, 월간 기업통계등록부 비교

구분	월간	분기	연간(잠정)	연간(확정)
기준시점	월말	분기말	기준연도 12월	기준연도 12월
제공시기	익월 말	익분기 익익월	기준연도 익년 12월	기준연도 익익년 4월
명부	행정자료	행정자료	행정+조사자료	행정+조사자료
종사자	기업 단위	기업 단위	기업, 사업체 단위	기업, 사업체 단위
매출액	기업 단위	기업 단위	기업, 사업체 단위	기업, 사업체 단위

1) 국가데이터처 「2026년 기업통계등록부 이용자 안내서」 내용을 토대로 정리함

2) (EU) 스웨덴, 핀란드, 네덜란드 등의 국가에서 경제통계 생산을 위해 기업등록부 구축

(미국) 국세청 등 행정자료를 기반으로 구축하고, 센서스 조사 자료를 통해 보완

(호주) 국세청의 사업자신고자료를 기반으로 구축, 경제통계 작성에 활용

(영국) 국세청, 기업등록청 등의 행정자료 및 민간자료를 이용하여 구축

<표 2> 기업통계등록부 구축 시 활용자료

기관	행정자료명	구분		구축항목
		연간	분기월간	
국세청	사업자등록자료	○	○	사업체, 기업 기본사항
	법인세 자료	○	○	
	부가가치세 자료	○	○	
	사업소득신고(복식기장, 간편장부)	○	○	기업 매출액, 재무정보
	추계소득금액계산서	○	○	
	온라인 판매대행 및 결제대행 자료	○	○	
	근로소득지급명세서	○		
	사업소득지급명세서	○		
	일용근로소득지급명세서	○	○	기업 종사자 수
4대보험	일용근로내역확인서	○	○	외국인 종사자 수
	건강보험자료_직장 가입자	○	○	건강보험 보수 총금액
	국민연금자료_사업장 가입자	○	○	
	고용·산재보험자료_피보험자 및 상실자	○	○	
	산재보험자료_근로자 고용	○	○	
국가데이터처	전국사업체조사(또는 경제총조사)	○		조사-행정 연계
	기업활동조사	○		매출액 및 종사자 배분
	광업제조업조사	○		경제통계조사 응답현황표
	서비스업조사	○		
	건설업조사	○		
	운수업조사	○		
	프랜차이즈조사	○		경제통계조사 응답현황표
	광공업동향조사	○		
	서비스업동향조사	○		
	온라인쇼핑동향조사	○		
	기업 및 사업체의 공간정보	○		좌표, 격자정보
	기업통계등록부 정비 결과	○	○	산업분류, 주소 등 보완
공정거래위원회	공시대상 기업집단 리스트	○		기업규모
산업통상부	중견기업 리스트	○		
	소재·부품·장비 전문기업	○		소·부장 전문기업 여부 및 소·부장 부문
한국에너지공단	에너지사용 및 온실가스배출 실태조사	○		에너지 분야 항목
	신재생에너지산업 실태조사	○		특수산업분류, 종사자 수, 매출액, 투자액 등
중소벤처기업부	중소기업 밀집지역		○	밀집지역 유형 및 명칭
	중소기업지원사업정보	○		중앙부처·지방자치단체의 부문별 지원금액
	뿌리기술 전문기업	○		뿌리기술 전문기업 여부, 공정기술, 업종, 핵심기술명
소상공인시장 진흥공단	전통시장 및 상권 정보	○		전통시장·상권 명칭, 면적
국토부	산업단지 공간정보	○		산업단지명, 면적, 구분코드
지식재산처	지식재산권 정보	○		특허, 실용신안, 디자인, 상표 보유 수
관세청	기업별 수출입 연간 총 교역액	○		수입액 및 수출액 구간
한국전력공사	전력정보	○		전기사용량, 전기사용요금, 전기사용개월수

제2절 기업통계등록부 정비사업³⁾

1. 추진 배경

연간 기업통계등록부는 행정자료에 전국사업체조사 등 조사 내용을 반영하여 명부가 포괄적이거나, 기준년도와 서비스 시점 간 15개월의 시차(기준년도 익익년 3월에 구축)가 발생하여 활용에 시의성이 떨어진다는 한계가 있었다. 이러한 기업통계등록부의 정확성 및 시의성 제고 요구에 따라 2019년부터 행정자료로만 구축하는 분기 기업통계등록부를 서비스하기로 하였고, 이를 위해서는 연간 발생하는 약 130만 개의 신규 등록 사업체의 사업자등록자료를 정확하고 시의성 있게 검증·보완하는 작업이 필요했다. 따라서 매월 발생하는 신생 및 변경 사업체의 기본정보를 점검하기 위해 2019년부터 국가데이터처 행정자료관리과에서 매년 ‘기업통계등록부 정비사업’을 실시하게 되었다.

2. 사업 내용

1) 정비대상

기업통계등록부 정비사업은 기준년도 전년 4분기부터 금년 3분기까지 국세청에 신규 등록한 사업체를 대상으로 실시한다. 그중 정비시점에 폐업, 정비 실효성이 낮은 산업(개인 부동산임대업, 운송업, 인적용역 등), 미운영 사업체 등은 제외하며, 기존 사업체 중 산업분류 등의 재확인이 필요한 사업체를 일부 포함하기도 한다.

<표 3> 신규진입 및 폐업 등을 반영한 모집단의 변화

1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월	구분	
												A. 신규 진입 사업체 (약 130여만개) 기업통계등록부 정비대상	중점점검 대상
												B. 존속 사업체 (약 1,000여만개)	일부 점검
												C. 폐업 사업체 (약 90여만개)	점검 제외

3) 국가데이터처 행정자료관리과 「2026년 기업통계등록부 정비 세부추진계획」 내용을 토대로 정리함

2) 정비방법

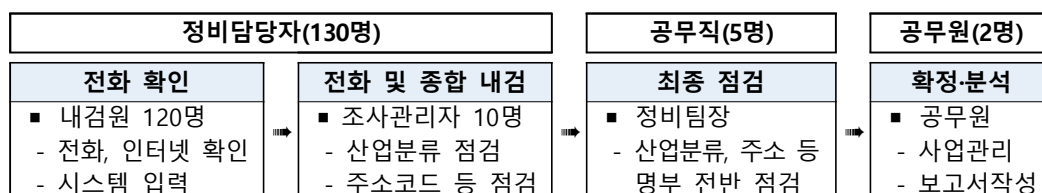
정비 방법은 크게 두 가지로, 대표자 등 사업체 운영 전반에 대해 응답이 가능한 자를 통해 전화 면접 가이드 등 절차에 따라 전화조사를 우선적으로 진행한다. 전화 확인이 안 되는 경우(전화번호가 없거나 오류), 사업체 정보 확인이 가능한 참고 사이트(홈페이지, 로드뷰, 배달앱, 블로그 등)를 활용하여 정비하거나, 전국사업체조사 자료가 있는 사업체인 경우, 잦은 조사로 인한 민원 방지를 위해 조사결과를 활용한다.

3) 추진체계

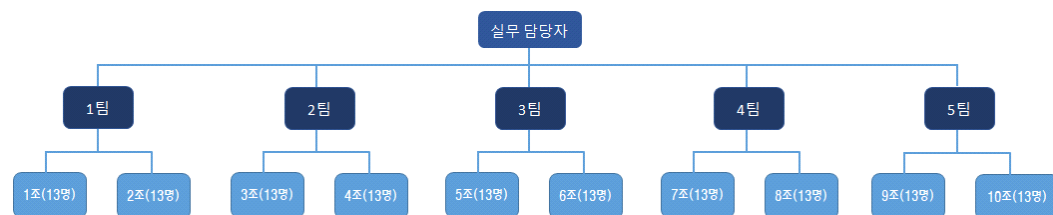
행정자료관리과는 기업통계등록부 정비사업을 위해 매년 약 9개월간 정비담당자 130명을 채용한다. 공무원 근로자 5명이 임차사무실에 상주하며 정비담당자 인력 및 사업 관리를 수행하며, 정비담당자 130명이 전화점검을 시행하는 방식으로 진행된다.

정비담당자 130명 중 내검원 120명이 매월 신규등록 사업체의 사업장 기본정보 및 산업분류 등 7개 항목을 확인(1차 내검4)하고 조사관리자 10명(조장)이 조사대상 중 일부를 뽑아5) 조사 내용을 재검토(2차 내검)하면, 공무원 근로자(팀장)가 최종적으로 정비자료를 점검한다.

<표 4> 기업통계등록부 정비사업 추진체계



<그림 1> 행정자료관리실 업무 체계도



- 실무담당자: 공무원 2명
- 팀장(1팀~5팀): 공무직 5명 - 팀당 1명
- 조장(1조~10조): 정비담당자 중 조사관리자 10명 - 조당 1명
- 조원(1조~10조): 정비담당자 중 내검원 120명 - 조당 12명

- 4) 사업장 기본정보 등 기존 자료에 수정·보완할 부분이 있는지 내용을 검토하는 작업을 말함
- 5) 작업완료, 일부확인, 폐업 사업체 중 내검규칙에 해당하는 사업체와 그 외 랜덤으로 뽑은 사업체

4) 정비 항목 및 내용

정비사업을 통해 사업체명, 전화번호, 주소(소재지), 사업체 운영 장소(종사자 유무, 간판유무), 사업내용(산업분류), 개업일자, 대표자(성별, 연령대) 등 7개 항목을 점검한다. 이 중 필수항목은 주소, 사업체 운영 장소, 산업분류이며, 필수항목이 모두 확인되어야 '작업완료', 필수항목 중 한 가지 항목이라도 확인이 되어야 '일부확인' 상태가 된다.

<표 5> 기업통계등록부 정비 항목 및 내용

번호	항목	내용
1	사업체명	○ 사업체명 확인
2	전화번호	○ 대표전화인 경우 해당 사업장 전화번호 조사 ○ 해당 사업체 현황에 대한 응답이 가능한 전화번호 확인 ○ 전화번호가 오류인 경우 인터넷으로 검색하여 확인
3	주소	○ 도로명주소 및 행정구역분류코드를 조사 ※ 도로명주소를 모르는 경우 지번주소로 조사
4	사업체 운영장소	① 무인매장 - 고객이 종사자의 도움없이 또는 최소한의 도움으로 상품 또는 서비스를 선택하고, 대금은 무인결제 시스템(온라인 송금, 동전 투입 등도 포함)으로 지불하여 상품 구매 또는 서비스를 받을 수 있는 물리적 공간 (예) 무인카페, 코인노래방, 셀프 사진관·빨래방·세차장, 무인 독서실, 무인 소매점(아이스크림, 문방구, 반려동물 용품 등) ② 상점·상가, 공장, 사무실 등 → 4-1. 종사자 유·무 항목으로 이동 - 소비자가 실제 사업장을 방문하여 상품, 서비스를 구매할 수 있는 사업체 또는 제품 제조 사업체(개인 운수업 포함) ③ 그 외 → 4-2. 간판 유·무 항목으로 이동 - 가정집(전자 상거래 등 가구 내 사업체가 존재) - 노점, 이동차량, 기타
4-1	종사자 유·무	○ 종사자 유·무 - 상근 근로자 유·무만 확인(종사자수는 조사 제외)
4-2	간판 유·무	○ 간판 유·무 - 육안으로 확인 가능한 배너, 입간판, 현판 등의 간판 유·무 확인 - 간판명은 조사 제외
5	산업분류	○ 주로 영위하는 사업(주사업)의 산업분류를 파악
6	개업일자	○ 개업일자 확인 (단, 등록일자와 개업일자의 차이가 6개월 이상인 경우에 한해 조사)
7	대표자	※ 사업자등록증상의 대표자 정보를 조사 ○ (성명) 대표자 성명 확인 ○ (성별) 대표자 성별을 남·여로 구분 ○ (연령대) 대표자 출생연도를 기준으로 연령대 기입 (예) 20대, 30대... - 연령대가 이상치(20대 미만, 80대 이상)인 경우 재확인

3. 2025년 기업통계등록부 정비결과

1) 정비 대상 및 응답 결과

2025년 신규등록 사업체 1,240천 개 중 정비 실효성이 낮은 505천 개(40.7%) 사업체를 제외한 735천 개(59.3%) 사업체가 정비대상으로, 그중 전체 또는 일부를 확인한 사업체가 508천 개(69.2%), 확인불가·폐업 등 미확인 사업체가 226천 개(30.8%)로 나타났다.

<표 6> 정비대상 사업체 현황

구분	신규등록 (A+B)	정비대상 (A)	제외대상(B)				
			폐업 ^①	부동산임대 ^②	운송 ^③	인적용역등 ^④	기타 ^⑤
사업체수(천개)	1,239.7	734.6	41.5	94.3	54.3	84.9	229.7
구성비(%)	100.0	59.3	3.3	7.6	4.4	6.9	18.5

- ① 해당기간 중 신규등록을 했으나, 정비 명부 작성 시점에 이미 폐업상태인 사업체
 ② 개인사업체 중 부동산 임대 및 공급업(681**), 법인사업체 중 부동산 임대업(6811*)이면서 전화번호 중복
 ③ 일반화물, 개별화물, 용달화물, 택시운송 등
 ④ 인적용역(특수고용형태종사자, 기타종사자) 및 단체(수익사업이 아닌 친목성격의 단체)
 ⑤ 태양력발전, 건설장비, 통신판매(아파트주소), 전화번호 누락, 정비 효율성 제고를 위한 정비 제외 산업 분류 55개, 사업자등록 이후 6개월(180일) 초과 개업 사업체, 산업분류 정비를 위해 변동사항이 적은 개인이면서 47911, 47612를 '23년 4분기 정비에서 일시 제외 등

<표 7> 응답 현황

구분	합계 (A+B)	확인 사업체(A)			미확인 사업체(B)				
		소계	전체	일부	소계	불응 ^①	확인불가 ^②	폐업 ^③	기타 ^④
사업체수(천개)	734.6	508.4	479.4	29.0	226.3	1.8	123.3	30.4	70.8
구성비(%)	100.0	69.2	65.3	3.9	30.8	0.2	16.8	4.1	9.6

- ① 전체항목에 대해 응답 거부
 ② 전화번호 오류, 부재 등으로 전화 확인 및 인터넷 확인불가
 ③ 정비대상 명부 생성 시점에 폐업
 ④ 휴업, 개업 전, 국방시설(군부대 등)

2) 정비 현황

확인 사업체 508천 개 중 사업체 정보를 수정한 사업체는 242천 개(47.6%)이며, 세부 항목별로는 산업분류 177천 건(34.9%), 전화번호 81천 건(15.9%) 순으로 나타났다.

<표 8> 확인 사업체 수정 현황

구분	확인 사업체	수정 사업체	세부 항목별 수정 건(천건)					
			소계*	사업체명	전화번호	주소	산업분류	대표자
사업체수(천개)	508.4	242.2	285.0	1.7	80.9	23.5	177.2	1.9
구성비(%)	100.0	47.6	-	0.3	15.9	4.6	34.9	0.4

* 1개 사업체에서 다수의 세부항목 수정이 발생할 수 있음

제3절 기업통계등록부 정비사업의 산업분류

1. 한국표준산업분류⁶⁾

1) 목적

한국표준산업분류는 산업 관련 통계자료의 정확성, 비교성을 확보하기 위해 작성된 것으로 유엔의 국제표준산업분류에 기초하여 작성되었다. 또한, 산업 활동에 의한 통계자료의 수집, 제표, 분석 등을 위해서 활동 분류 및 범위를 제공하기 위한 것으로 통계법 제22조(표준분류)에 따라 모든 통계작성기관이 이를 의무적으로 사용하도록 규정하고 있다. 통계 작성 목적 이외에도 일반 행정 및 산업정책 관련 법령에서 적용 대상 산업영역을 한정하는 기준 등으로 활용되고 있다.

2) 개정 주기

이에 한국표준산업분류는 1963년 제정된 이래 국내 산업구조 및 환경 변화, 개정 수요·국제분류 기준 등을 반영하기 위해 열 차례의 개정을 통해 보완·발전되어 왔으나, 국가데이터처는 그동안 부정기적으로 개정 되어왔던 산업분류를 가장 최근에 이루어진 제11차 개정(2024. 1. 1. 고시)부터 대규모 통계조사(총조사)에 시의성 있게 활용할 수 있도록 5년마다의 정기개정으로 변경하였다.

제11차 개정에서는 미래·성장 산업을 중심으로 분류 항목을 신설 또는 세분하였으며, 상대적으로 비중이 감소한 산업은 분류를 통합하였다. 또한 대국민·관계기관 수렴 의견, 다수 민원 및 규제 개선요청 등 개정 수요와 국제표준산업분류 기준을 반영하였다.

3) 분류 기준

산업분류는 생산단위(사업체단위, 기업체단위 등)가 주로 수행하고 있는 산업 활동을 그 유사성에 따라 체계적으로 유형화한 것으로 다음과 같은 분류 기준에 의해 적용된다.

- ① 산출물(생산된 재화 또는 제공된 서비스)의 특성
 - 산출물의 물리적 구성 및 가공 단계, 산출물의 수요처, 산출물의 기능 및 용도
- ② 투입물의 특성
 - 원재료, 생산 공정, 생산기술 및 시설 등
- ③ 생산활동의 일반적인 결합형태

6) 국가데이터처 「한국표준산업분류 제11차 개정 해설서(2024)」 내용을 토대로 정리함

4) 분류 구조

산업분류는 대분류(알파벳 문자, Section), 중분류(2자리 숫자, Division), 소분류(3자리 숫자, Group), 세분류(4자리 숫자, Class), 세세분류(5자리 숫자, Sub-Class) 총 5단계로 구성되며, 부호 처리를 할 경우에는 아라비아 숫자만 사용하도록 했다. 권고된 국제분류 ISIC Rev.4를 기본체계로 하였으나, 국내 실정을 고려하여 국제분류의 각 단계 항목을 분할, 통합 또는 재그룹화하여 독자적으로 분류 항목과 분류 부호를 설정하였다.

<표 9> 한국표준산업분류 예시

대분류	중분류	소분류	세분류	세세분류
I	56	561	5611	56112
숙박 및 음식점업	음식점 및 주점업	음식점업	한식음식점업	한식면요리전문점

<표 10> 제11차 개정 한국표준산업분류 단계별 분류 항목 수

대분류(중분류)	중분류	소분류	세분류	세세분류
A 농업, 임업 및 어업(01~03)	3	8	21	33
B 광업(05~08)	4	7	10	11
C 제조업(10~34)	25	85	182	480
D 전기, 가스, 증기 및 공기조절 공급업(35)	1	3	5	10
E 수도, 하수 및 폐기물 처리, 원료 재생업(36~39)	4	6	14	19
F 건설업(41~42)	2	8	16	46
G 도매 및 소매업(45~47)	3	20	62	186
H 운수 및 창고업(49~52)	4	11	19	47
I 숙박 및 음식점업(55~56)	2	4	11	30
J 정보통신업(58~63)	6	12	26	45
K 금융 및 보험업(64~66)	3	8	14	29
L 부동산업(68)	1	2	4	12
M 전문, 과학 및 기술 서비스업(70~73)	4	14	20	50
N 사업시설 관리, 사업지원 및 임대 서비스업(74~76)	3	11	22	32
O 공공행정, 국방 및 사회보장 행정(84)	1	6	10	28
P 교육서비스(85)	1	7	17	33
Q 보건업 및 사회복지 서비스업(86~87)	2	6	9	25
R 예술, 스포츠 및 여가관련 서비스업(90~91)	2	4	17	43
S 협회 및 단체, 수리 및 기타 개인 서비스업(94~96)	3	8	18	41
T 가구 내 고용활동 및 자가 소비 생산활동(97~98)	2	3	3	3
U 국제 및 외국기관(99)	1	1	1	2
총 21개	77개	234개	501개	1,205개

2. 산업분류 정비 현황

1) 국세청의 사업자등록자료

기업통계등록부 정비사업은 국세청 사업자등록자료를 토대로 작업을 수행한다. 그중 산업분류 정비를 위해서는 <그림 2>의 2. 사업장 현황 자료를 활용하며, 각 사업체의 ‘주업태’, ‘주종목’, ‘주업종 코드’가 기업통계등록부 정비자료로 입수된다. 사업자등록 자료의 ‘주업태’, ‘주종목’ 항목은 기업통계등록부 정비자료의 ‘업태’, ‘업종’ 항목으로 바로 활용이 가능하나, ‘주업종 코드’는 별도의 변환 과정을 거치게 된다.

사업자등록자료의 ‘업종코드’는 국세청이 사업 형태(업태)와 제공 품목·서비스(종목)를 세무 관리용으로 분류한 6자리 코드로, 이는 앞에서 설명한 한국표준산업분류와 다른 분류체계이다. 그러나 2019년 국세청에서 한국표준산업분류 제10차 개정(2017.1.)에 따라 한 개의 업종코드에 대응하는 산업분류가 다수인 경우, 산업분류에 맞춰 업종을 세분화하는 등 산업분류와 대응하도록 업종코드를 개편하여 현재 연계표를 통해 한국표준산업분류 코드(세세분류, 5자리 숫자)로 자료 입수가 가능하다.

<그림 2> 사업자등록 신청서 양식

■ 부가가치세법 시행규칙 [별지 제4호서식] <개정 2026. 3. 20.> 홈택스(www.hometax.go.kr)에서도 신청할 수 있습니다.

사업자등록 신청서(개인사업자용) (법인이 아닌 단체의 고유번호 신청서)

※ 사업자등록의 신청 내용은 영구히 관리되며, 납세 성실도를 검증하는 기초자료로 활용됩니다.
아래 해당 사항을 사실대로 작성하시기 바라며, 신청서에 본인이 자필로 서명해 주시기 바랍니다.
※ []에는 해당하는 곳에 √ 표를 합니다. (알록)

접수번호	처리기간	2일(보정 기간은 불산입)
------	------	----------------

1. 인적사항

상호(단체명)	연	락	(사업장 전화번호)
성명(대표자)		처	(주소지 전화번호)
주민등록번호			(휴대전화번호)
(단체)부동산등기용등록번호			(팩스 번호)
사업장(단체) 소재지			총 호
사업장이 주소지인 경우 주소지 이전 시 사업장 소재지 자동 정정 신청 ([]여, []부)			

2. 사업장 현황

업 종	주업태	주종목	주생산 요소	주업종 코드	개업일	종업원 수
	부업태	부종목	부생산 요소	부업종 코드		

사이버물 명칭 사이버물 도메인

사업장 구분	자가	타가	사업장을 빌려준 사람			임대차 명세		
	면적	면적	성명 (법인명)	사업자 등록번호	주민(법인) 등록번호	임대차 계약기간	(전세) 보증금	월세 (차임)
	m ²	m ²				~	원	원

2) 산업분류 정비 작업

기업통계등록부 정비사업 정비담당자는 산업분류를 정비할 때 사업자등록자료의 기존 산업분류(업종, 업태, 산업분류코드)를 기반으로 질의하여 산업분류 변경 여부를 판단한다. 기존 산업분류가 공란인 경우, 검색을 통해 사전정보 확인 후 질의하며 사업체 응답 내용에 따라 산업분류를 분류한다.

<그림 3>은 실제 기업통계등록부 정비 시스템 화면이다. 기존 산업분류는 ‘김밥 및 기타 간이 음식점업’이었으나, 정비 작업으로 ‘제과점업’으로 변경된 것을 확인할 수 있다. 정비담당자는 ‘색인검색’을 통해 색인어 또는 산업분류코드를 검색하여 11차 산업분류를 입력할 수 있으며, 10차 산업분류는 11차와 1:1인 경우 자동 맵핑된다.

<그림 3> 기업통계등록부 정비 시스템(산업분류)

산업분류

기준 : 56191 김밥 및 기타 간이 음식점업 **산업분류가 변경되었습니다**

주산업 : (11차) 56150 제과점업 색인검색 색인초기화 ☐ 확인불가

주산업 : (10차) 56191 제과점업 색인검색 색인초기화 ☐ 확인불가

11차 산업분류 코드 검색

초기화 조회

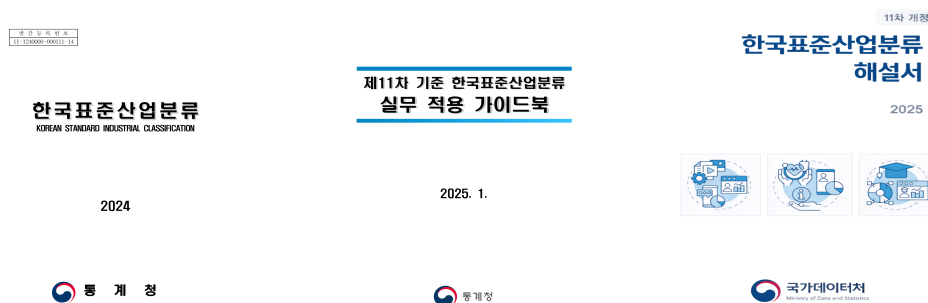
산업분류 색인 검색 제과점 운영

산업분류 목록 (1건)

번호	산업분류	산업분류명	대분류항목	소분류항목	세분류항목	세세분류항목	5자리	4자리	3자리	비고
1	56150	제과점 운영	음식점 및 ...	음식점업	제과점업	제과점업	선택	선택	선택	확인

이외에도 한국표준산업분류의 제정·개정 및 운영을 담당하는 통계기준과가 발간한 ‘한국표준산업분류 해설서(총 960페이지)’ 또는 행정자료관리과에서 자체적으로 관리하는 ‘한국표준산업분류 종합색인표’, ‘SBR 산업분류 사례집’ 등의 자료를 활용하기도 한다.

<그림 4> 한국표준산업분류 참고자료



또한 통계분류포털(kssc.mods.go.kr) 홈페이지에서 찾고자 하는 분류를 검색할 수도 있겠으나, 정비담당자는 업무망(내부망)에서 정비 작업을 수행하므로 궁금한 내용이 있을 때마다 망 전환을 통해 인터넷망(외부망)에서 검색하는 것이 불편할 수밖에 없다.

<그림 5> 통계분류포털 홈페이지



제2장 제2절 3에서 살펴봤듯이, 기업통계등록부 정비사업을 통해 점검하는 항목 중 가장 많이 수정되는 항목이 ‘산업분류(2025년 정비사업 세세분류 기준 34.9%)’로, 기업통계등록부 정비사업에서 산업분류를 정비하는 업무가 매우 중요하다고 볼 수 있다. 그러나 전문교육 실시에도 매년 새로 채용되는 정비담당자 개인의 역량에 따라 산업분류 코드 부여에 소요되는 시간과 정확도에 편차가 발생할 수밖에 없는 것이 현실이다.

<표 11> 2025년 기업통계등록부 정비사업 산업분류 수정 현황

구분	확인 사업체	산업분류 수준별 변경 사업체(천개)				
		대분류	중분류	소분류	세분류	세세분류
사업체수(천개)	508.4	45.3	76.0	114.2	134.8	177.2
구성비(%)	100.0	8.9	15.0	22.5	26.5	34.9

이에 기업통계등록부 정비사업을 담당하는 행정자료관리과에서는 정비담당자의 산업분류 부여 업무의 효율성 및 정확성 제고를 위해 기업통계등록부 정비에 AI 산업 자동분류 도입을 추진하게 되었다. 따라서 다음 장에서는 국가데이터연구원에서 현재 운영 중인 AI통계분류시스템의 구조와 기업통계등록부 정비사업에 적용할 수 있는 산업분류 모델의 특성을 검토하려고 한다.

제 3 장

AI통계분류시스템 활용 현황

제1절 AI통계분류시스템 개요

1. 추진 배경 및 필요성

모든 통계조사는 공포 집계를 위해 자유로운 텍스트 형식의 조사자료를 조사별로 사용하는 분류체계에 따라 코드로 분류하는 작업이 필요하다. 하지만 대량의 조사자료 처리를 위해 그 작업에 다수의 인력이 투입되어야 하므로 이러한 수작업 중심의 내검 작업은 업무 효율에 한계가 있다. 분류체계가 방대하고 조사마다 달라, 조사원, 내검원 등 담당자별로 경험과 지식수준에 따라 편차가 발생할 수밖에 없기 때문이다.

따라서 국가데이터처에서는 수작업 분류 업무를 효율화하기 위해 규칙·사례 기반의 자동코딩 시스템(사례사전, 색인DB 등)을 구축하여 활용하였다. ‘사례사전’은 분류를 한정할 수 있는 단어들의 조합으로 이루어진 규칙 베이스에 따라 코드를 부여하는 방식이며, ‘색인DB’는 조사자료에서 어절 단위로 추출한 색인어를 색인DB에서 검색하여 검색된 색인어의 빈도 정보로 검색 점수를 계산하여 코드를 부여하는 방식이다.

각 방식이 조금씩 다를 수는 있으나, 모두 공통적으로 등록된 사전 지식을 기반으로 분류코드를 부여하는 시스템으로, 조사 응답이 미리 등록해 둔 키워드 규칙 또는 사례 사전과 일치하게 되면 자동으로 코드를 부여하게 된다. 하지만 표현이 조금만 달라져도 코드를 부여하지 않아, 분류의 성능을 높이려면 규칙과 사례 사전을 사람이 계속해서 관리해야 한다는 번거로움이 있었다.

최신의 인공지능(AI) 기술을 도입하여 이러한 기존 시스템의 한계점을 극복하기 위해 국가데이터연구원에서는 2020~2021년 2년 간의 기초연구7)를 통해 AI 기술을 활용한 통계분류 자료처리의 가능성 및 기존 자동코딩 시스템 대비 향상된 분류 성능을 확인하였다. 이를 토대로 2022년 과학기술정보통신부의 ‘디지털 공공서비스 혁신 프로젝트’ 지원사업에 선정되어 확보한 예산(15.5억 원)으로 ‘AI통계분류시스템’을 구축하였다.

7) 임경민, 김혜란, 오교중(2020) “데이터 과학 기법의 국가통계 시험적용 분석”, 국가데이터연구원
 임경민, 김혜란(2021) “인공지능 기반 통계분류 자동코딩 시스템 구축 연구”, 국가데이터연구원
 임경민, 김혜란(2021) “인공지능 기반 산업분류 자동화 처리 연구”, 국가데이터연구원

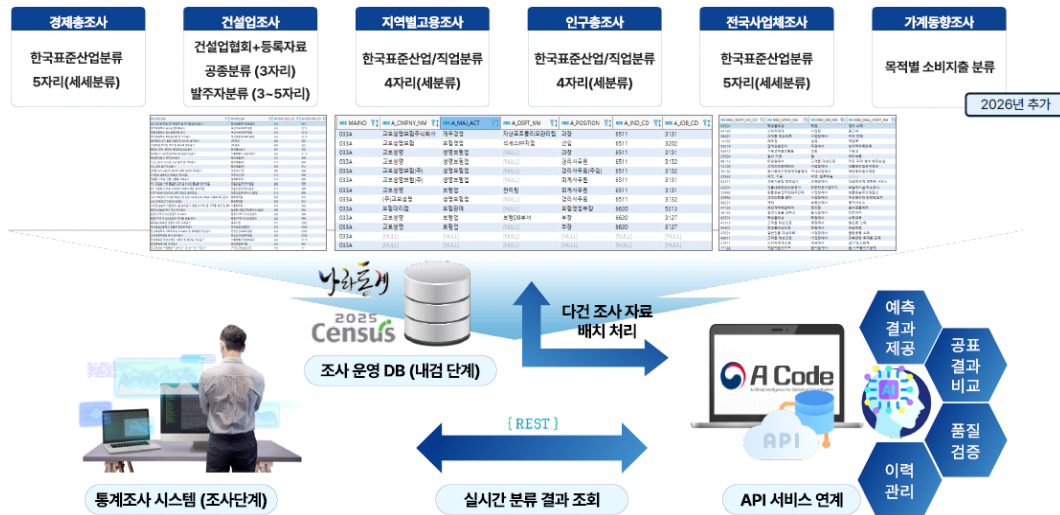
2. 시스템 현황 및 운영

AI통계분류시스템은 최초 구축 이후에도 여러 연구⁸⁾를 통해 꾸준히 발전되었으며, 그 결과 현재 국가데이터처의 여러 조사(인구주택총조사, 경제활동인구조사, 가계동향조사, 지역별고용조사, 이민자체류실태및고용조사, 경제총조사, 전국사업체조사, 건설업조사 등)의 자료처리 및 내검 작업에 직·간접적으로 활용되고 있다.

본 시스템은 국가데이터처의 통계 생산 시스템인 나라통계시스템과 연동되어 조사원, 내검원 및 조사담당자에게 실시간 API를 통해 AI 추론 결과를 제공하고 있다. 이 외 다른 자체 시스템도 API 서비스 신청을 통해 실시간 API로 바로 연동하여 AI통계분류시스템을 이용할 수 있어, 활용 범위를 계속 확장할 수 있는 시스템이다.

학습 및 서비스를 위한 하드웨어(HW)는 GPU 서버(A30 2식, A100 7식)와 국가정보자원관리원 클라우드 환경의 WEB, WAS 장비 등을 활용하였고, 시스템 운영 소프트웨어(SW)는 여러 대의 서버를 하나의 통합 시스템으로 운영하는 쿠버네티스(kubernetes)를 활용하였다. 또한 새로운 학습데이터 즉, 조사별 최신 자료가 공표될 때마다 모델을 재학습하는 과정을 거치며 시스템 성능을 높이기 위한 노력을 계속하고 있다.

<그림 6> 2026년도 AI통계분류시스템 유지관리 사업 업무 내용



- 8) 임경민(2023) “AI 통계분류 결과 분석 및 실무 활용성 제고 방안 연구”, 국가데이터연구원
 김건수, 임경민(2023) “인공지능 통계분류 자동화 확대 적용 연구”, 국가데이터연구원
 임경민(2024) “인공지능 기반 행정자료 품질관리 도입방안 연구”, 국가데이터연구원
 임경민(2024) “AI 통계분류를 위한 가계동향조사 학습데이터 개선방안 연구”, 국가데이터연구원
 임경민(2025) “가계동향조사 AI 통계분류 실무활용 연구”, 국가데이터연구원
 이규호, 임경민(2025) “AI기반 건설경기동향 자동분류 코딩 연구”, 국가데이터연구원

제2절 AI통계분류시스템 방법론

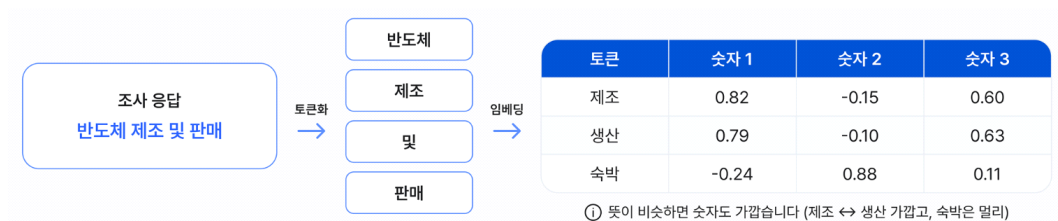
AI통계분류시스템은 국가데이터처 공표자료를 지도학습한 기계학습 모델이 분류하는 “지도학습 기반의 기계학습(Machine Learning)” 방식으로 운용되는 시스템이다.

1. 입력 처리 단계

1) 토큰화(Tokenization) 및 임베딩(Embedding)

먼저 입력 내용을 의미 단위(토큰)로 나누는 작업이 수행된다. 예를 들어, “반도체 제조 및 판매”를 입력하면 “반도체”, “제조”, “및”, “판매”로 나뉜다. 이러한 토큰을 컴퓨터가 계산할 수 있는 숫자 정보(임베딩 벡터)로 변환하여 입력 내용의 의미 정보를 파악한다. 뜻이 비슷한 표현은 비슷한 수치의 임베딩 벡터를 가지게 되는데, 입력을 이루는 토큰의 문맥(순서) 정보를 통해 입력 전체의 의미 정보를 수치로 다룰 수 있다.

<그림 7> 조사 응답이 숫자 벡터로 바뀌는 과정



2) 사전학습 언어모델(Pretrained Language Model)

입력 속 각 단어가 다른 단어에 얼마나 관련성이 있는지를 가중치로 계산하여 저장하는 작업을 ‘어텐션(Attention)’이라고 하며, 어텐션을 여러 층으로 쌓아 문장 속 모든 단어의 관계(문맥 정보)를 한 번에 파악하는 구조를 ‘트랜스포머(Transformer)’라고 한다.

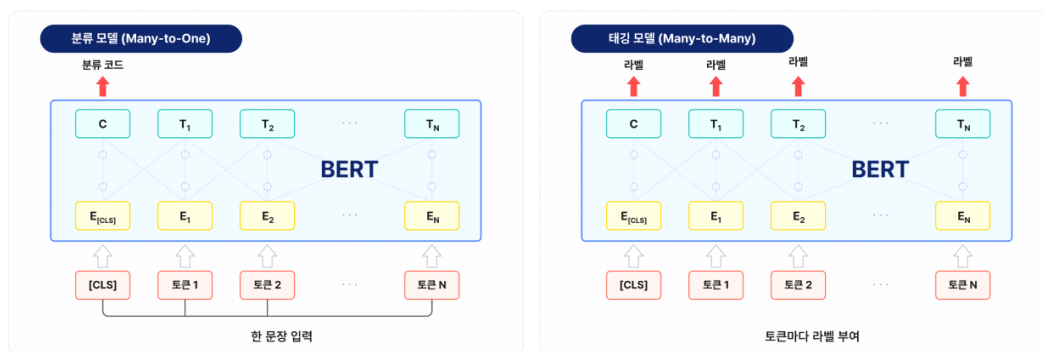
사전학습 언어모델은 트랜스포머의 인코더, 디코더 구조를 기반으로 입력 문장에 대한 의미 이해나 생성 기능에 집중하여 만들어진 딥러닝 모델이다. 다양한 분야의 말뭉치를 대량으로 학습해 기본적인 언어 이해 능력을 갖추고(사전학습), 이후 조사자료로 분류에 특화하면(미세조정) 학습 시간이 짧아지고 분류 정확도가 높아지게 된다.

AI통계분류시스템에 적용하는 언어모델 세대가 발전할 때마다 그 개선점이 분류 정확도와 처리 효율로 이어지므로, 한국어 특화, 분류 성능, 처리 효율을 기준으로 실제 공표자료로 학습, 평가한 결과를 검토하여 적합한 공개 언어모델을 선택한다.

3) 언어모델별 특징

BERT는 문장의 일부 단어를 가리고 맞추는 방식(마스킹)으로 학습하는 모델이다. 앞뒤를 모두 봐야 풀 수 있어 양쪽 방향의 맥락을 함께 이해하게 되는데, 이러한 양방향 이해는 응답 전체 맥락을 봐야 하는 분류 작업에 특히 유리하다. 따라서 BERT는 문장 분류, 태깅, 질의응답 등 여러 작업에 쓰이며, 통계분류는 이 중 ‘문장 분류, 태깅’ 방식(응답 전체 또는 토큰마다 코드 부여)을 사용한다.

<그림 8> BERT의 분류 모델 및 태깅 모델



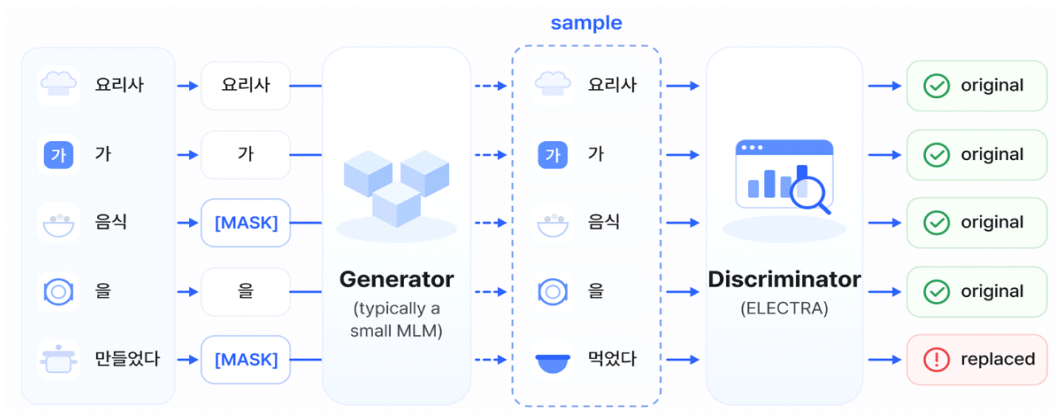
RoBERTa는 BERT와 구조는 동일하나, 학습 방법을 개선해 더 많은 데이터로 오래 학습하고 도움이 적은 학습 목표(‘다음 문장 예측’)를 제거해 이해력을 높인 모델이다. 현재 AI통계분류시스템에 적용한 KLUE-RoBERTa-large 모델은 KLUE 평가셋과 대규모 한국어를 학습해 뉴스 등 격식체에 강하다는 특성이 있다.

<그림 9> BERT 대비 RoBERTa의 학습 방법 개선

학습 요소	BERT	RoBERTa
마스킹 방식	→ 정적 (고정)	→ 동적 (매번 새로 마스킹)
다음 문장 예측 (NSP)	→ 사용	→ 제거 (불필요)
학습 데이터	→ 16GB	→ 160GB (약 10배)
학습량	→ 기본	→ 더 큰 배치 · 더 오래

ELECTRA는 생성기가 일부 단어를 그럴듯하게 바꿔치기하면, 판별기가 모든 토큰에서 원래 것인지, 바뀐 것인지 판별하는 “치환 탐지” 방식으로 학습하여 같은 계산으로도 더 빠르고 촘촘하게 학습이 가능한 모델이다. 현재는 다른 모델로 교체되었지만 기존 AI통계분류시스템에 적용했던 KcELECTRA 모델은 온라인 댓글 등 구어체도 학습해 오타자, 비표준 표현에 강하다는 특성이 있다.

<그림 10> ELECTRA 치환 탐지 학습 구조



T5는 번역, 분류, 요약 등 다양한 작업에서 고른 성능을 내기 위해 인코더와 디코더를 모두 사용하여 만들어진 다중 작업 학습(Multi-task Learning) 모델이다. 현재 AI통계분류시스템에 적용한 pko-t5 모델은 한국어 WordPiece 토크나이저의 미등록어(OOV) 문제를 해결하였으며, 한국어 자료만으로 사전학습해 한국어에 특화되었다는 특성이 있다.

<그림 11> T5 과제 통합 처리 구조



종합적으로 현재 AI통계분류시스템은 RoBERTa 계열의 KLUE-RoBERTa-large 모델과 T5 계열의 pko-t5 모델(기존 ELECTRA 계열의 KcELECTRA 모델에서 변경)을 운영하고 있으며, 통계분류 성능 개선을 위해 계속해서 최신 기술의 모델을 도입하고 있다.

<그림 12> AI통계분류시스템 한국어 특화 모델 적용 현황



2. 분류 모델 학습 단계

분류 모델은 미세조정(Fine-tuning), 시퀀스 분류(Sequence Classification), 계층형 분류 체계(Hierarchical Classification)의 학습 과정을 거치게 된다.

1) 미세조정

미세조정은 일반 언어 지식만 갖고 있는 사전학습 언어모델에 조사별 공표자료(입력 및 정답 코드)를 지도학습하여 통계분류 작업을 위한 특화 모델로 만드는 과정이다. 이러한 방법으로 기존의 자동코딩 시스템처럼 규칙을 일일이 만들지 않아도 학습데이터 관리를 통해 분류 정확도의 유지·개선이 가능하다.

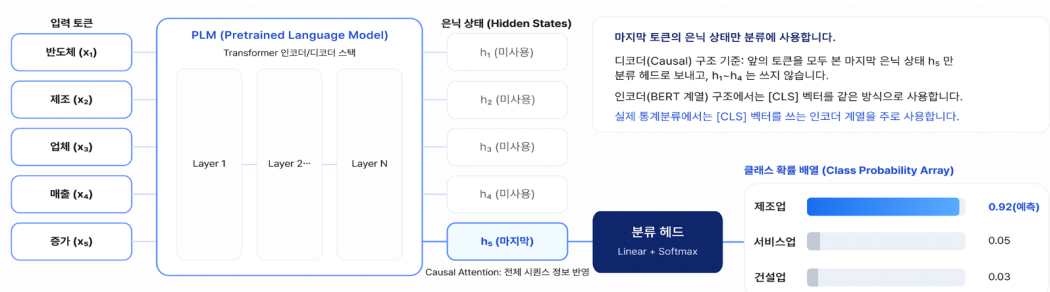
<그림 13> 미세조정으로 만드는 통계분류 특화 모델



2) 시퀀스 분류 학습

시퀀스 분류 학습은 입력 문장을 토큰 시퀀스로 인코딩한 후, 사전학습 언어모델에서 문장 전체를 대표하는 토큰의 은닉 상태를 분류 헤드에 전달하여 클래스별 확률 점수를 산출하는 방법이다. 예를 들어, “반도체 제조업체 매출 증가”라는 입력 문장은 시퀀스 분류 학습 구조를 거쳐 제조업 0.92, 서비스업 0.05, 건설업 0.03의 확률 점수를 얻게 된다. 이러한 학습 방법은 가변 길이의 입력을 고정 차원 벡터로 압축해 처리하므로, 문장 전체 맥락을 반영한 안정적인 분류가 가능하다.

<그림 14> 시퀀스 분류 학습 구조



3) 계층형 분류체계 학습

계층형 분류체계 학습은 시퀀스 분류 학습 방법에서의 분류 헤드 Linear 레이어를 고도화하여 분류체계의 단계적인 계층 구조(대분류 → 중분류 → 소분류 → 세분류 → 세세분류) 정보를 구분하여 학습에 활용하는 방법이다. 예를 들면, “반도체 제조”에 대해 제조업(C) → 전자 부품, 컴퓨터, 영상, 음향 및 통신장비 제조업(26) → 반도체 제조업(261) → 전자집적회로 제조업(2611) → 메모리용 전자집적회로 제조업(26111)의 형식으로 학습하는 것이다. 이러한 학습 방법은 한 번에 수천 개의 코드에서 고르지 않고, 단계마다 좁혀 나가기 때문에 방대한 분류체계에서도 정확하고 안정적인 학습이 가능하다.

<그림 15> 계층형 분류 흐름



3. 분류 모델 추론 단계

입력 처리 단계와 분류 모델 학습 단계를 거쳐 AI통계분류시스템은 학습에 활용한 ‘사전학습 언어모델(PLM)’과 ‘미세조정(Fine-tuning)된 분류 모델’을 이용하여 새로운 입력에 대한 분류 결과를 제공한다. 분류 결과는 코드와 함께 AI 예측 확률을 제공하고, 담당자는 이 수치를 보고 기계학습(ML) 모델의 예측을 얼마나 믿을지 판단하면 된다.

<그림 16> 조사 응답에서 분류코드까지의 처리 흐름



이러한 언어모델 기반의 분류 방식은 기업통계등록부 정비사업에서 정비담당자가 확보하는 사업체의 ‘사업내용’과 같이 비정형 텍스트로 작성된 정보를 분석하여 적절한 산업분류를 추론하는 데 활용할 수 있다. 따라서 AI통계분류시스템에서 운영하는 모델 중 기업통계등록부 정비사업에 바로 적용할 수 있는 모델이 무엇인지 살펴보고 해당 모델을 적용하여 AI 산업 자동분류 도입 가능성을 검토하려고 한다.

제3절 AI통계분류시스템 운영 모델

1. AI 분류 추론 모델

현재 AI통계분류시스템은 5개 조사에서 총 8개 모델을 운영하고 있다.

먼저 전국사업체조사 모델은 2023년 기준 조사자료(2024년 조사)를 학습하여 제11차 한국표준산업분류(KSIC)를 세세분류(5자리)까지 추론한다. 얼마 전 조사가 완료된 2025년 기준 경제총조사(2026년 조사)의 경우, 해당 모델에 최신 자료인 2024년 기준 조사자료(2025년 조사)를 학습하여 AI 추론 결과를 제공할 예정이다.

지역별고용조사 모델은 2022년 상반기부터 2025년 하반기 조사자료를 학습하여 제11차 한국표준산업분류(KSIC)와 제8차 한국표준직업분류(KSCO)를 각각 세분류(4자리)까지 추론하는 2개의 모델로 구성된다.

인구총조사 모델은 다른 모델들과 달리 연구를 통해 구축되지 않고 2025년 인구총조사의 자료처리를 위해 구축된 모델로, 재코딩한 2020년 조사자료에 2023~2024년 지역별고용조사 자료를 추가 학습하여 지역별고용조사 모델과 같이 2개의 모델로 구성된다.

가계동향조사 모델은 2024년 1분기부터 2026년 1분기 조사자료를 학습하여 한국표준목적별개별소비지출분류(COICOP-K)를 세분류(4자리)까지 추론한다.

마지막으로 건설업조사 모델은 2021년부터 2023년 기준 조사자료(2022~2024년 조사)를 학습하여 공사종류분류(KCS)는 세부공종(3자리)까지, 발주자분류(PPO)는 세부발주자(부문에 따라 3~5자리)까지 추론하는 2개의 모델로 구성된다.

<표 12> AI통계분류시스템 운영 모델

조사	분류 모델	분류 단계	학습데이터
전국사업체조사	산업분류(KSIC 11차)	세세분류(5자리)	2023년 기준 전국사업체조사
지역별고용조사	산업분류(KSIC 11차) 직업분류(KSCO 8차)	세분류(4자리)	2022년 상반기~2025년 하반기 지역별고용조사
인구총조사	산업분류(KSIC 11차) 직업분류(KSCO 8차)	세분류(4자리)	2020년 인구총조사 2023~2024년 지역별고용조사
가계동향조사	목적별소비지출분류 (COICOP-K)	세분류(4자리)	2024년 1분기~2026년 1분기 가계동향조사
건설업조사	공사종류분류(KCS) 발주자분류(PPO)	세부공종(3자리) 세부발주자(3~5자리)*	2021~2023년 기준 건설업조사

* 중앙정부기관(5자리), 지방자치단체(4자리), 기타·민간부문(3자리)에 따라 다름

2. 산업분류 추론 모델

앞서 살펴봤듯이, AI통계분류시스템에서 산업분류 추론에 활용할 수 있는 모델은 전국사업체조사 모델, 지역별고용조사 모델, 인구총조사 모델 총 3가지로 볼 수 있다. 그중에서 산업분류를 세세분류(5자리)까지 조사하는 기업통계등록부 정비사업에 활용할 수 있는 모델은 바로 ‘전국사업체조사 모델(2023년 기준 조사자료 학습)’이다.

전국사업체조사 모델은 조사표의 ‘사업체명’, ‘사업자등록번호 또는 법인등록번호’, ‘사업체 구분’, ‘사업의 종류(주사업)’, ‘산업분류부호’ 등의 항목을 학습하여 추론한다.

<그림 17> 전국사업체조사 조사표



승인번호
전 국 : 제101037호

2024년 기준

사업체조사 조사표

조사표 구분



• 이 조사는 경제 정책수립 및 관련 산업 연구를 위해 통계법 제17조, 제18조에 따라 실시하는 국가지정통계(제101037호)입니다.
 • 해당 사업체는 통계법 제25조, 제26조, 제32조에 따라 응답의무가 있으며, 이에 응하지 않을 경우 과태료가 부과될 수 있습니다.
 • 조사된 자료는 통계법 제33조에 따라 통계목적외로만 사용되고, 사업체의 비밀은 엄격히 보호되니 성실하게 응답해 주시기 바랍니다.

행정구역분류부호		조사구분호		사업체 일련번호		★ 사업체고유번호		변동부호	운영장소
				건물	사업체				
① 사업체명 (간판명) ()				② 사업장 대표자		성명		성별	연령대
								① 남 ② 여	대
③ 소재지	도	로	명	읍·면 도로명 지하 건물(본번) - 건물(부번)					
	주	소		빌딩·시장·상가 동 층 호					
	지	번		읍·면·(법정)동 리 번지 호 통 반					
	주	소		공단·단지·택지 블록(B) 로트(L)					
주 소 부정확 사유 * 도로명 주소가 부정확한 경우 해당 사유를 모두 선택 ☐ 도로명 없음 ☐ 노점 ☐ 도로명 주소 불명확 ☐ 주소와 지도 불일치									
④ 창설연월				⑤ 사업자 등록번호		* 미가입 사유			
						☐ 미등록 ☐ 응답거부			
⑥ 조직형태				* ② 회사법인과 ③ 회사이외법인인 경우만 법인등록번호와 사업체 구분을 작성					
				법인등록번호					
				사업체 구분		☐ 단독사업체 ☐ 본사, 본점 등 ☐ 공장, 지사(점), 영업소 등			
⑦ 사업의 종류	[투입] 무엇을 가지고 (영업장소, 원재료 등)		[과정] 어떤 방법으로 (생산 및 영업활동)		[산출] 생산·제공하였는가 (최종 제화 및 용역)		매출액 비중(%)		산업분류부호
	주 사업		배합사료 공장에서 각종 재료를 혼합하여		반려동물용 사료 제조		100		1 0 9 0 1
	부사업 1								
	부사업 2								
	10801 배합사료 제조업 (11차)		▶ (개념) 가금류를 포함한 가축용, 어류 양식용, 애완동물용, 실험 동물용 각종 배합사료를 생산하는 산업활동 ▶ (11차) 10901 반려동물용사료 제조업, 10902 배합사료 제조업 ▶ (예시) 반려애완동물 사료, 반려애완동물 기호식품, 실험동물용 배합사료 / 가축용 또는 어류용 배합사료, 소, 젓소, 돼지, 닭, 말, 토끼, 사슴, 개, 고양이, 물고기, 관상어, 담수어, 해산어						
* 산업분류가 105(곡물가공품, 전분·전분제품 제조업), 108(기타 식물 제조업), 11(음료 제조업)인 사업체만 응답 쌀 사용 여부 식품이나 음료를 만드는 데 원재료로 쌀을 사용하고 있습니까? ① 예 ② 아니오									

제 4 장

시통계분류시스템 적용 분석

제1절 기업통계등록부 정비사업 자료 검토

1. 정비자료

기업통계등록부 정비사업은 국세청의 사업자등록자료를 기반으로 수행한다. 따라서 정비 전 자료는 사업자등록자료 원자료이며, 정비 후 자료는 정비담당자들이 전화조사를 통해 확인한 자료라고 볼 수 있다. 그러나 실제 행정자료관리과에서 기업통계등록부 정비를 위해 관리하는 데이터(테이블)는 작업배정 테이블, 정비결과 테이블, 최종 테이블 총 3가지이다. 각 테이블 구성에 대한 이해를 돕기 위해 아래의 2025년 정비사업 정비대상 사업체 수치를 예로 들어 설명하려고 한다.

<표 13> 2025년 기업통계등록부 정비사업 정비대상 사업체

신규등록 사업체 1,240천 개			
정비대상 사업체 735천 개(59.3%)			정비제외 사업체 505천 개(40.7%)
작업배정 테이블	정비결과 테이블	최종 테이블	
735천 개 + α	735천 개 + α	735천 개	

1) 작업배정 테이블

작업배정 테이블은 국세청에서 최초 입수한 신규등록 사업체의 사업자등록자료 중 정비대상 사업체 자료로, 정비담당자에게 업무량 배정을 위해 사용하는 테이블이다. 정비대상 사업체는 신규등록 사업체 중 정비시점에 폐업, 정비의 실효성이 낮은 산업(개인 부동산임대업, 운송업, 인적용역 등), 미운영 사업체 등은 제외하고, 기존 사업체 중 산업분류 등의 재확인이 필요한 사업체를 일부 포함한다.

예를 들어, 2025년 4분기부터 2026년 3분기까지의 신규등록 사업체 대상으로 실시된 2025년 정비사업에서 신규등록 사업체 1,240천 개 중 정비대상 사업체는 735천 개이다. 따라서 작업배정 테이블에는 735천 개 사업체의 원자료가 있으나, 전화조사를 담당하는 내검원에게 배정된 사업체와 전화조사 자료의 내검을 담당하는 조사관리자에게 배정된 사업체가 중복으로 포함되어 있어, 실제 데이터 개수는 735천 개보다 많다.

2) 정비결과 테이블

정비결과 테이블은 정비작업이 완료된 정비대상 사업체 자료로, 정비담당자들의 작업 결과를 집계한 테이블이다. 따라서 해당 테이블도 작업배정 테이블과 마찬가지로 735천 개 사업체의 정비자료가 있으나, 내검원의 전화조사와 조사관리자의 내검 자료가 중복으로 포함되어 있어, 실제 데이터 개수는 735천 개보다 많다.

결과적으로 작업배정 테이블과 정비결과 테이블의 실제 데이터 개수는 735천 개보다 많고 그 수가 동일하다는 공통점이 있으나, 산업분류 정비 관점에서 가장 큰 차이점은 ‘사업내용’의 유무이다. 작업배정 테이블(정비 전)에는 원자료의 ‘업종’과 ‘업태’가 들어 있으나, 정비결과 테이블(정비 후)에는 정비담당자가 조사한 ‘사업내용’이 들어있다.

3) 최종 테이블

최종 테이블은 모든 작업을 마친 후 기업통계등록부 구축에 활용되는 최종자료로, 전화조사 자료와 내검 자료가 둘 다 있는 사업체의 경우에는 최종적으로 내검 자료를 반영한 테이블이다. 따라서 해당 테이블은 테이블 중 유일하게 사업체 수, 실제 데이터 수 모두 정비대상 사업체 수와 같은 735천 개다.

결국 정비결과 테이블에서 정비대상 사업체의 중복 데이터를 없앴 것으로, 정비결과 테이블과 레이아웃, 형식 등 내용은 동일하나, 실제 데이터 개수에서만 차이가 있다.

<표 14> 작업배정 테이블, 정비결과 테이블, 최종 테이블 비교

구분	작업배정 테이블	정비결과 테이블	최종 테이블
내용	사업자등록 원자료	전화조사 및 내검 결과	전화조사 및 내검 결과
목적	기간제 업무량 배정	정비 결과 취합	기업통계등록부 구축
데이터 수	정비대상 사업체 + α (사업체 중복 있음)	정비대상 사업체 + α (사업체 중복 있음)	정비대상 사업체 (사업체 중복 없음)
주요 항목	업종, 업태	사업내용	사업내용

2. 사업내용 작성 방법

국가데이터연구원과 행정자료관리과는 본 연구과제를 추진하기 전인 2025년부터 기업통계등록부 정비사업에 AI통계분류시스템을 도입하기 위한 실무적인 협의를 이어왔다. 그중 AI통계분류시스템을 통해 기업통계등록부의 산업분류를 효과적으로 추론하기 위한 사전작업으로 추진한 것이 바로 ‘사업내용’ 작성 방법 변경이다.

정비결과 자료(정비결과 또는 최종 테이블) 중 ‘사업내용’이 산업분류 추론을 위한 입력값이 되므로, AI통계분류시스템이 추론을 잘하기 위해서는 ‘사업내용’을 충실하게 작성하는 것이 중요하다. 또한 앞에서 살펴봤듯이, 기업통계등록부 정비사업의 산업분류 추론에 전국사업체조사 모델(산업세세분류)을 적용할 예정이므로 해당 모델이 데이터를 학습한 대로 입력값을 넣어야 추론의 정확도를 높이는 데에 효과가 있을 수밖에 없다.

따라서 행정자료관리과에서는 2025년 5월 기준자료부터 AI통계분류시스템의 전국 사업체조사 모델 학습데이터와 같이 ‘사업내용’을 작성하도록 작성 방법을 변경하였다. 즉, 정비담당자에게 ‘사업내용’에 <그림 18>의 전국사업체조사 조사표와 같이 ‘무엇을 가지고(영업장소, 원재료 등)’, ‘어떤 방법으로(생산 및 영업활동)’, ‘생산·제공하였는가(최종 재화 및 용역)’ 등에 대해 기존보다 자세하게 작성하도록 했다.

<그림 18> 전국사업체조사 조사표(사업의 종류)

	[투입] 무엇을 가지고 (영업장소, 원재료 등)	[과정] 어떤 방법으로 (생산 및 영업활동)	[산출] 생산·제공하였는가 (최종 재화 및 용역)	매출액 비중(%)	산업분류부호
주 사 업	배합사료 공장에서	각종 재료를 혼합하여	반려동물용 사료 제조	100	1 0 9 0 1

아래의 두 그림은 실제 정비담당자가 기업통계등록부 정비 시스템에 ‘사업내용’을 입력한 화면이다. <그림 19>와 같이 기존에는 ‘사업내용’을 비교적 간단하게 작성했으나, 작성 방법 변경 이후에는 <그림 20>와 같이 전국사업체조사 방식(투입, 과정, 산출)대로 더 자세하게 작성한 것을 확인 할 수 있다.

<그림 19> 기존 사업내용 작성 예시(2025년 2월 기준자료)

등록일자	20250204	개업일자	20250205 	☐ 확인불가	기준 : 20250205
업태	도매 및 소매업	업종	전기기계설비		
사업내용	11차: 금속컨베이어(롤러베드)장치 설치공사 10차: 금속컨베이어(롤러베드)장치 설치공사				

<그림 20> 변경 사업내용 작성 예시(2025년 7월 기준자료)

등록일자	20250731	개업일자	20250725 	☐ 확인불가	기준 : 20250725
업태	전문, 과학 및 기술서비스업	업종	인테리어 디자인업		
사업내용	11차: 인테리어 디자인업체가 실내 공간 구성을 기획, 설계하여 실내 인테리어 디자인 10차: 인테리어 디자인업체가 실내 공간 구성을 기획, 설계하여 실내 인테리어 디자인				

제2절 AI 산업 자동분류 정확도 분석

1. 분석 대상 및 방향

기업통계등록부 정비사업 데이터에 대한 AI 산업 자동분류 정확도를 효과적으로 분석하기 위해 다음과 같이 분석 대상 및 방향을 설정하려고 한다.

1) 2025년 기업통계등록부 정비사업 대상

기업통계등록부 정비사업은 사업년도 기준 익년 4분기부터 금년 3분기까지의 신규 등록 사업체 대상으로 수행된다. 같은 해에 동일한 정비 지침, 동일한 정비담당자를 통해 정비된 사업체를 분석하는 것이 시계열 일관성을 유지할 수 있으므로 가장 최근에 완료된 ‘2025년 정비사업(2024년 4분기 ~ 2025년 3분기)’ 사업체 대상으로 분석하려고 한다.

전체 기간 사업체에 대한 분석과 더불어, AI통계분류시스템을 통해 ‘기존 방식으로 작성한 사업내용(2024년 10월 ~ 2025년 4월 기준자료)’으로 산업분류를 추론한 결과와 ‘변경 방식으로 작성한 사업내용(2025년 5월 ~ 9월 기준자료)’으로 산업분류를 추론한 결과를 비교 분석하여 ‘사업내용’ 작성 방법 변경에 따른 효과를 살펴볼 수 있다.

<표 15> 기업통계등록부 정비사업 자료 시점

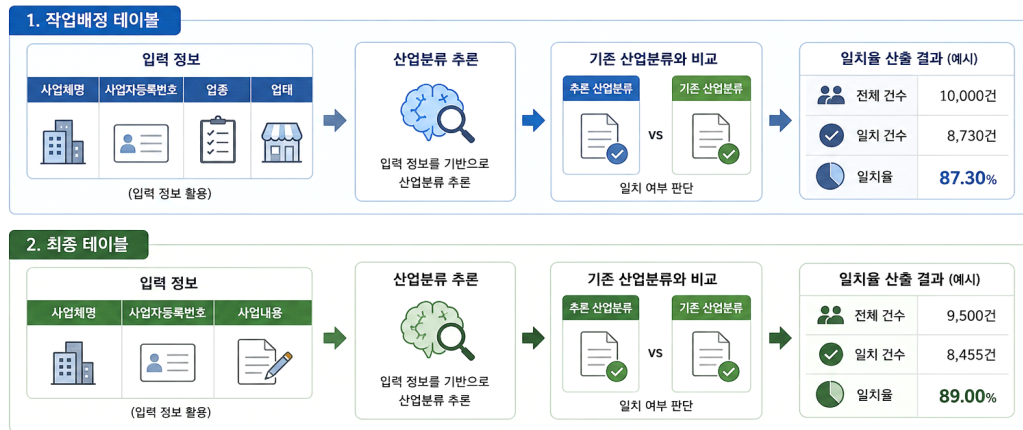
2024년						2025년											
3분기			4분기			1분기			2분기			3분기			4분기		
7월	8월	9월	10월	11월	12월	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
2024년 정비			2025년 정비사업												2026년 정비		
기존 방식으로 '사업내용' 입력									변경 방식으로 '사업내용' 입력								

2) 작업배정 테이블 및 최종 테이블

제4장 제1절 1에서 살펴봤듯이, 정비사업 관련 데이터는 작업배정 테이블, 정비결과 테이블, 최종 테이블이 있다. 이 중 정비결과 테이블은 사실상 최종 테이블과 동일한 데이터이기 때문에 분석에서 제외하고, 정비 전 원자료인 작업배정 테이블과 정비 후 결과인 최종 테이블만 분석하려고 한다.

AI통계분류시스템을 통해 작업배정 테이블의 ‘사업체명’, ‘사업자등록번호’, ‘업종’, ‘업태’ 항목으로 산업분류를 추론한 결과와 최종 테이블의 ‘사업체명’, ‘사업자등록번호’, ‘사업내용’ 항목으로 산업분류를 추론한 결과를 비교하면, 원자료를 바탕으로 사업체의 ‘사업내용’을 확인하는 작업 즉, 기업통계등록부 정비사업의 효과를 살펴볼 수 있다.

<그림 21> 작업배정 테이블 및 최종 테이블 분석 방법



3) 사업내용이 있는 사업체

하지만 최종 테이블에 있는 모든 사업체에 ‘사업내용’이 기재되어 있는 것은 아니다. 2025년 정비사업 정비대상 사업체 735천 개 중 정비담당자가 정비한 확인 사업체는 508천 개(69.2%), 미확인 사업체(불응, 폐업 등)는 226천 개(30.8%)이다. 미확인 사업체는 사업체 관련 정보가 전혀 정비되지 않았기에 당연히 ‘사업내용’이 기재되어 있지 않다. 그러나 확인 사업체 중에서도 모든 항목을 정비한 ‘전체확인’ 사업체가 아닌, ‘일부확인’ 사업체의 경우에는 ‘사업내용’이 없는 사업체가 존재한다.

2025년 정비사업 최종 테이블 기준으로 확인 사업체 508,371개 중 482,076개 사업체 (94.8%)에만 ‘사업내용’이 기재되어 있는데, 이는 ‘전체확인’ 사업체 479,420개 전체와 ‘일부확인’ 사업체 28,951개 중 일부 ‘사업내용’이 있는 2,656개 사업체를 합한 수치이다. 이와 더불어 ‘사업내용’이 있는 482,076개 사업체 중 ‘사업내용’이 기존 방식으로 작성된 사업체는 273,069개(56.6%), 변경 방식으로 작성된 사업체는 209,007개(43.4%)이다.

따라서 앞서 1)과 2)에서 언급했던 기업통계등록부 정비사업의 효과 및 ‘사업내용’ 작성 방법 변경에 따른 효과를 알아보기 위해 2025년 정비사업 대상 사업체 중 최종 테이블을 기준으로 ‘사업내용’이 있는 482,076개 사업체 대상으로만 분석하려고 한다.

<표 16> 기업통계등록부 정비사업 대상 사업체

구분	정비대상 사업체	확인 사업체	사업내용 있는 사업체
2025년 정비사업	734,641개(100.0%)	508,371개(100.0%)	482,076개(100.0%)
- 기존 방식 자료	412,975개(56.2%)	286,729개(56.4%)	273,069개(56.6%)
- 변경 방식 자료	321,666개(43.8%)	221,642개(43.6%)	209,007개(43.4%)

* ‘전체확인’ 사업체 479,420개 전체, ‘일부확인’ 사업체 28,951개 중 ‘사업내용’이 있는 2,656개

2. 분석 결과

분석 결과에서 ‘정확도’는 AI 추론 산업분류와 ‘등록분류’가 일치한 비율로, 이때의 ‘등록분류’는 사람이 부여한 산업분류이다. ‘등록분류’ 자체도 일부 오류가 존재할 수 있어 AI 추론 산업분류와 일치하지 않는다고 해서 항상 AI가 오답인 것은 아니지만, 해당 분석은 ‘등록분류’가 정답임을 가정하여 수행하였다.

1) 정비사업 효과 - 작업배정 테이블 vs 최종 테이블 비교 분석

기업통계등록부 정비사업의 효과는 원자료인 작업배정 테이블과 정비결과인 최종 테이블의 AI 추론 결과 비교 분석을 통해 확인할 수 있다. 해당 분석은 최종 테이블 기준으로 ‘사업내용’이 있는 482,076개 사업체의 사업자등록번호로 사업체를 특정하여 작업배정 테이블에서도 482,076개 사업체를 추출하였다. 이를 통해 동일한 사업체에 대해 기업통계등록부 정비사업 전, 후 분석 결과를 비교할 수 있다.

먼저 추론 정확도를 살펴보면, AI 1순위 정확도는 작업배정 테이블이 71.5%, 최종 테이블이 93.1%로, 정비 후 자료가 원자료보다 21.6%p 더 높았다. 게다가 AI 1~3순위 정확도도 작업배정 테이블이 88.8%, 최종 테이블이 98.0%로, 정비 후 자료가 원자료보다 9.2%p 더 높았다. 따라서 원자료보다 정비담당자가 정비를 마친 자료가 AI통계분류 시스템으로 산업분류를 추론하는 데 더 잘 맞는 데이터임을 알 수 있다.

또한 작업배정 테이블의 산업분류 중 세세분류 5자리가 아닌 2, 3, 4자리인 경우가 4,208개 사업체(전체의 0.9%)에서 발견되었는데, 이 또한 AI 추론 1순위와의 비교에서 불일치로 처리하면 작업배정 테이블의 AI 1순위 정확도는 71.1%이 되기도 한다. 앞서 제시한 작업배정 테이블의 AI 1순위 정확도는 위의 경우를 제외한 값으로, 추론 정확도 측면에서는 0.4%p의 미미한 차이로 비교의 의미가 없을 수 있으나, 이는 원자료의 산업분류를 그대로 쓰기가 구조적으로 완전하지 않음을 보여준다.

<표 17> 작업배정 테이블과 최종 테이블 분석 결과

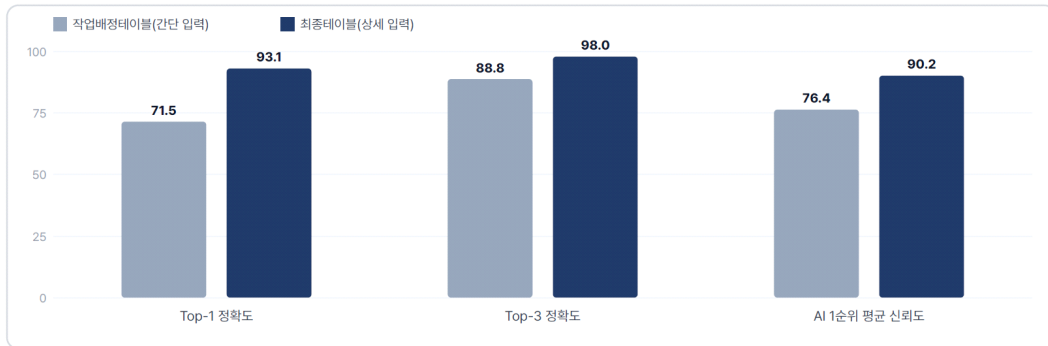
구분	사업체 수	AI 1순위 정확도*	AI 1~3순위 정확도**	AI 1순위 평균 신뢰도***	입력값 평균 글자 수
작업배정 테이블 (정비 전 원자료)	482,076개	71.5%	88.8%	76.4%	14.3자
최종 테이블 (정비 후 자료)	482,076개	93.1%	98.0%	90.2%	24.0자

* AI 추론 1순위 산업분류와 등록분류(작업배정 또는 최종 테이블의 산업분류)가 일치한 비율

** AI 추론 1순위~3순위 산업분류 안에 등록분류(작업배정 또는 최종 테이블의 산업분류)가 포함된 비율

*** AI 추론 1순위 산업분류에 부여한 AI 스스로의 확신도(0~100%) - 전체 사업체의 1순위 신뢰도 평균

<그림 22> 원자료와 정비결과 비교 분석



또한 AI 추론을 위한 입력값 기준으로 살펴보면, 작업배정 테이블의 ‘업종, 업태’ 평균 글자 수는 14.3자, 최종 테이블의 ‘사업내용’ 평균 글자 수는 24.0자로, 입력 정보가 상세해지자 AI 1순위 평균 신뢰도가 76.4%에서 90.2%로, 13.8%p 증가했다. 이를 통해 AI통계분류시스템을 통해 산업분류를 추론할 때 입력값을 상세히 적어 넣을수록 AI가 스스로에게 부여한 확신도가 상승한다는 것을 확인할 수 있다.

그렇다면 원자료의 등록분류는 AI 추론 1순위 산업분류와 일치하지 않았으나, 새로 정비 후 등록분류와 AI 추론 1순위 산업분류가 일치하게 된 124,072개 사업체에 대해 어떠한 요인으로 불일치에서 일치가 되었는지 정확도 상승 요인을 살펴보고자 한다.

놀랍게도 AI 추론 정확도 상승 요인이 정비작업을 통해 등록분류가 바뀐 정비효과(34.7%), 등록분류는 동일하나 ‘업종, 업태’보다 더 자세하게 작성된 ‘사업내용’으로 AI 추론이 달라진 입력 상세화 효과(35.7%), 둘 다(29.6%)로 거의 30%씩 균등하게 나뉜다. 즉, AI 추론 정확도는 어느 한 요인 때문이 아니라, 정비사업을 통해 정답(등록분류)과 입력(사업내용)이 함께 정리되며 맞물린 결과로 상승한 것이다.

결국, 정비담당자가 전화조사를 통해 원자료의 ‘등록분류’가 맞는지 확인하는 작업, 원자료의 ‘업종, 업태’보다 ‘사업내용’을 자세하게 작성하는 작업 모두 정비사업이므로 기업통계등록부 정비사업이 AI 산업분류 추론에 효과가 있음을 확인할 수 있다.

<표 18> AI 추론 정확도 상승 요인

구분	합계	등록분류만 바뀐* (정비효과)	AI 추론만 바뀐** (입력 상세화 효과)	둘 다 바뀐***
사업체 수	124,072개	43,022개	44,337개	36,713개
비중	100.0%	34.7%	35.7%	29.6%

* 작업배정 테이블의 등록분류에서 최종 테이블의 등록분류로 바뀜

** 작업배정 테이블의 등록분류와 최종 테이블의 등록분류가 동일하나, AI 추론 산업분류가 달라짐

*** 작업배정 테이블의 등록분류에서 최종 테이블의 등록분류로 바뀌고, AI 추론 산업분류도 달라짐

2) 사업내용 작성 방법 효과 - 기존 방식 vs 변경 방식 비교 분석

‘사업내용’ 작성 방법의 효과는 최종 테이블의 ‘사업내용’이 있는 482,076개 사업체 중 기존 방식으로 ‘사업내용’이 작성된 273,069개 사업체와 변경 방식으로 ‘사업내용’이 작성된 209,007개 사업체의 AI 추론 결과 비교 분석을 통해 확인할 수 있다.

먼저 추론 정확도를 살펴보면, AI 1순위 정확도는 기존 방식이 90.3%, 변경 방식이 96.8%로, 작성 방법 변경 후 자료가 변경 전 자료보다 6.5%p 더 높았다. 게다가 AI 1~3순위 정확도도 기존 방식이 97.1%, 변경 방식이 99.2%로, 작성 방법 변경 후 자료가 100%에 가까운 높은 정확도를 보이며 새로운 방식으로 작성된 ‘사업내용’이 AI통계분류 시스템으로 산업분류를 추론하는 데 더 잘 맞는 데이터임을 알 수 있다.

앞의 분석처럼 같은 시점의 동일한 사업체 대상으로 분석한 결과가 아니므로 정비 숙련도 변화 등 다른 요인의 효과를 완전히 배제하기는 어렵지만, ‘사업내용’ 입력이 평균 14.7자에서 36.6자로 상세해진 것이 정확도 상승의 유력한 요인으로 볼 수 있다.

<표 19> 기존 방식과 변경 방식 분석 결과

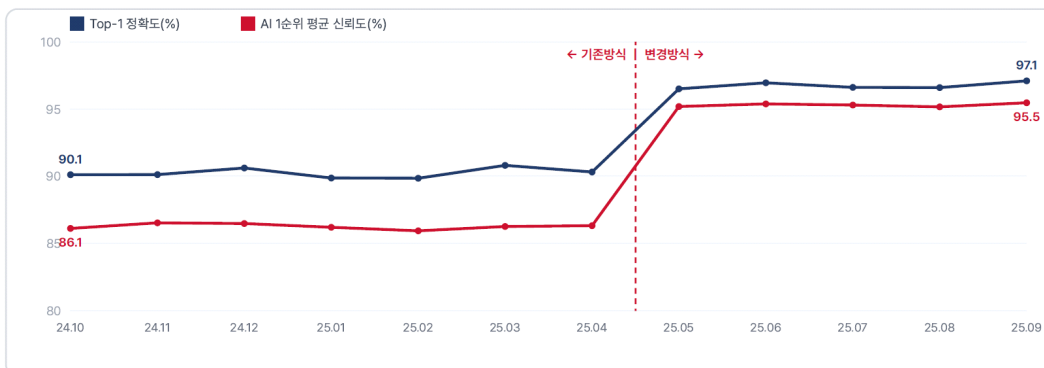
구분	사업체 수 (비중)	AI 1순위 정확도*	AI 1~3순위 정확도**	AI 1순위 평균 신뢰도***	입력값 평균 글자 수
최종 테이블 (2024.10.~2025.9.)	482,076개 (100.0%)	93.1%	98.0%	90.2%	24.0자
기존 방식 (2024.10.~2025.4.)	273,069개 (56.4%)	90.3%	97.1%	86.3%	14.7자
변경 방식 (2025.5.~2025.9.)	209,007개 (43.4%)	96.8%	99.2%	95.3%	36.6자

* AI 추론 1순위 산업분류와 등록분류(최종 테이블의 산업분류)가 일치하는 비율

** AI 추론 1순위~3순위 산업분류 안에 등록분류(최종 테이블의 산업분류)가 포함된 비율

*** AI 추론 1순위 산업분류에 부여한 AI 스스로의 확신도(0~100%) - 전체 사업체의 1순위 신뢰도 평균

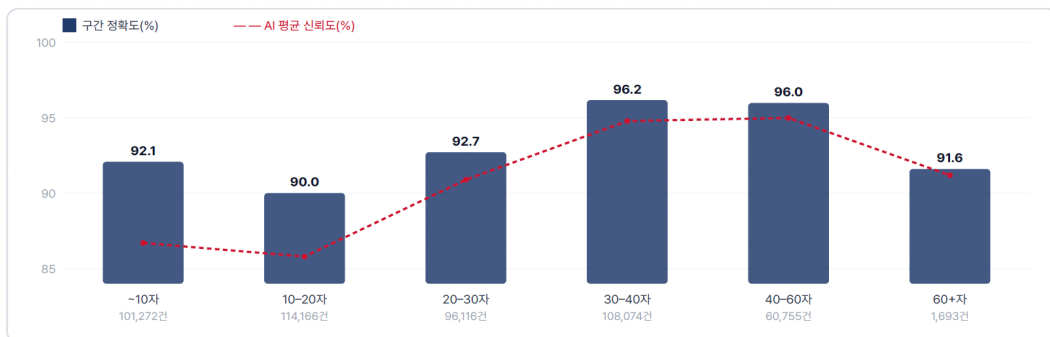
<그림 23> 기존 방식과 변경 방식 비교 분석



그렇다면 정말로 ‘사업내용’을 상세하게 적는 것이 효과가 있는지 확인하기 위해 최종 테이블 전체를 ‘사업내용’ 글자 수 구간으로 나누어 살펴보았다. <그림 24>를 보면, ‘사업내용’이 적당히 길수록(약 30~60자) AI 1순위 정확도와 평균 신뢰도가 함께 높은 것을 확인할 수 있다. 다만, 가장 짧은 구간(~10자)이 10~20자 구간보다 높고, 가장 긴 구간(60+자)에서는 AI 1순위 정확도와 평균 신뢰도가 다시 내려가는 모습을 보인다.

이는 단순히 글자 수 자체가 정확도와 신뢰도를 보장하는 것이 아니라, ‘사업내용’에 무엇을 적었는지 그 내용이 중요하다는 것을 의미한다. 따라서 ‘사업내용’의 글자 수만 늘리는 것이 아닌, 변경 방식처럼 더 많은 정보를 담고 있는 ‘문장형 입력’으로의 개선 방향성은 AI 산업분류 추론에 효과가 있음을 확인할 수 있다.

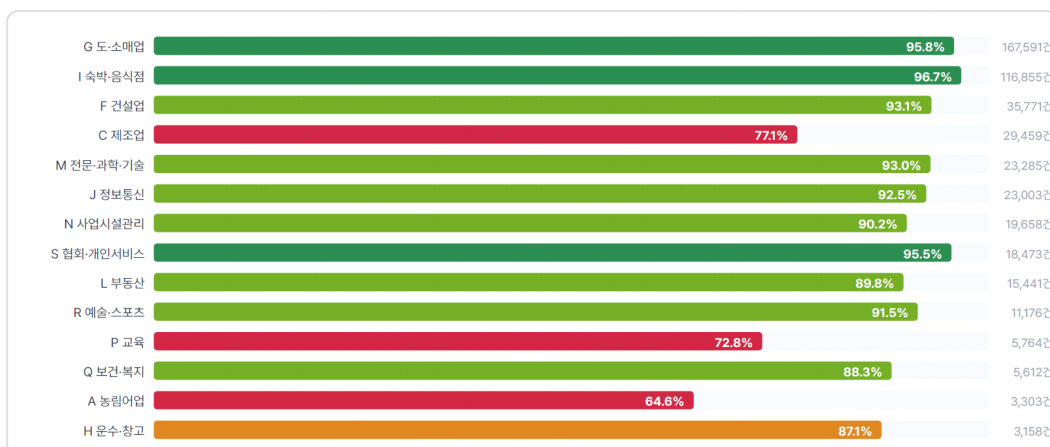
<그림 24> 글자 수 구간별 정확도 및 신뢰도



3) 산업분류별 정확도

최종 테이블 ‘사업내용’이 있는 사업체의 산업대분류별 정확도를 살펴보면, 숙박·음식점업, 도·소매업 등에서 높고, 농림어업, 교육서비스업 등에서 낮은 것으로 나타났다.

<그림 25> 산업분류별 정확도(사업체 많은 분류 상위)



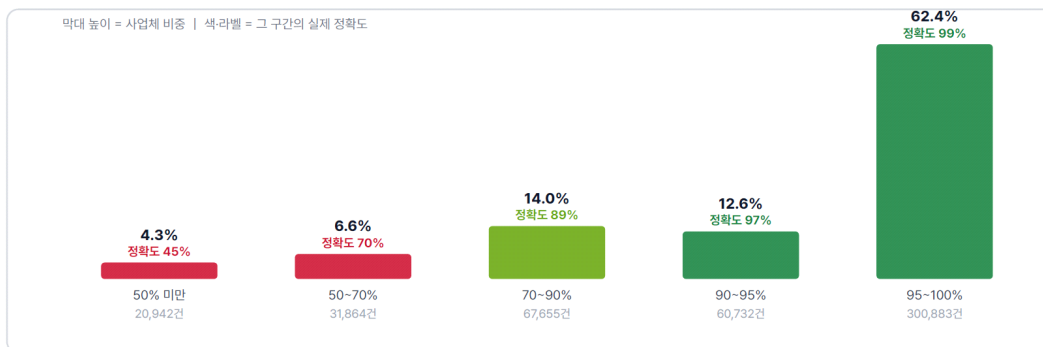
제3절 AI통계분류시스템 실무 활용 방안

1. AI 산업분류 추론 결과 활용

제2절에서는 등록분류와 AI 추론 산업분류의 정확도 위주로 분석했지만, 이번에는 신뢰도 관점에서 살펴보려고 한다. ‘신뢰도’는 AI가 산업분류를 추론할 때 그 결과를 얼마나 확신하는지 스스로 0~100%의 범위로 확신 정도를 매긴 값이다. 따라서 신뢰도가 AI 추론 산업분류가 정답일 확률 그 자체는 아니지만, 신뢰도가 높을수록 등록분류와 AI 추론 산업분류가 일치할 가능성이 높아지게 된다.

2025년 정비사업 최종 테이블의 ‘사업내용’이 있는 482,076개 사업체 중 AI 1순위 신뢰도가 95~100%인 사업체는 300,883개(62.4%)로, 이들의 AI 1순위 정확도는 무려 99.2%에 달한다. 이러한 신뢰도의 특성을 바탕으로, 적절한 신뢰도 기준을 설정하여 산업분류 정비 후 내검 과정에서 신뢰도가 높은 경우는 AI 추론 산업분류로 자동 확정 하고, 신뢰도가 낮은 경우에만 산업분류가 제대로 부여되었는지 검토하는 방식으로 정비 사업의 산업분류 내검 시스템을 효율화하여 운영이 가능하다.

<그림 26> AI 1순위 신뢰도 구간별 AI 1순위 정확도



<표 20> 자동확정 신뢰도 기준별 정확도

자동 확정 기준	자동 확정 비율	자동 확정분 정확도	인력 검토 비율
신뢰도 95% 이상	62.4%	99.2%	37.6%
신뢰도 90% 이상	75.0%	98.8%	25.0%
신뢰도 80% 이상	84.3%	98.0%	15.7%
신뢰도 70% 이상	89.0%	97.2%	11.0%

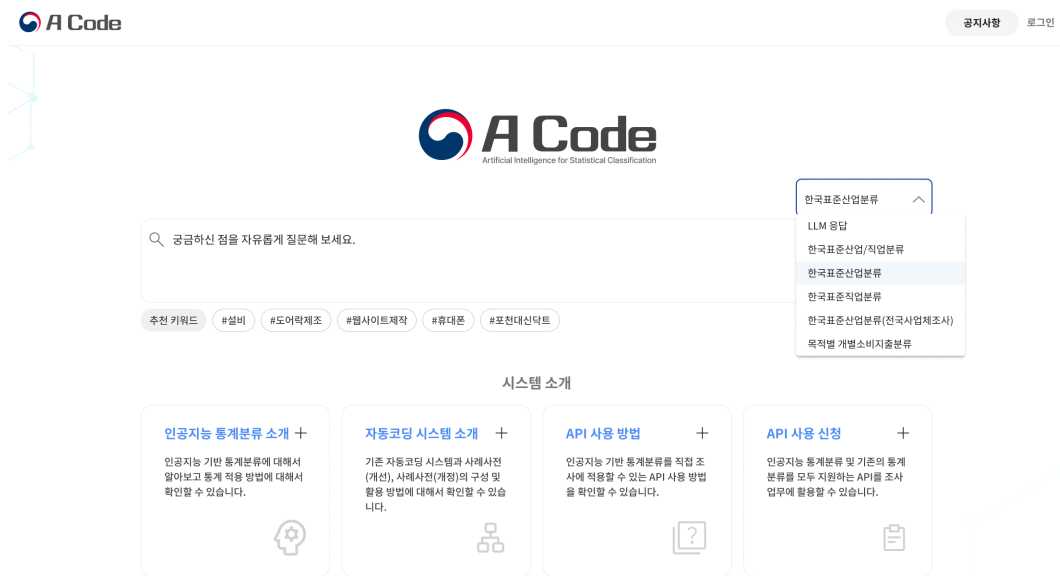
※ 예) 신뢰도 90% 이상을 자동 확정하면, 전체의 75%를 정확도 98.8%로 자동 처리, 나머지 25%만 검토

2. AI통계분류시스템 홈페이지(A Code) 활용

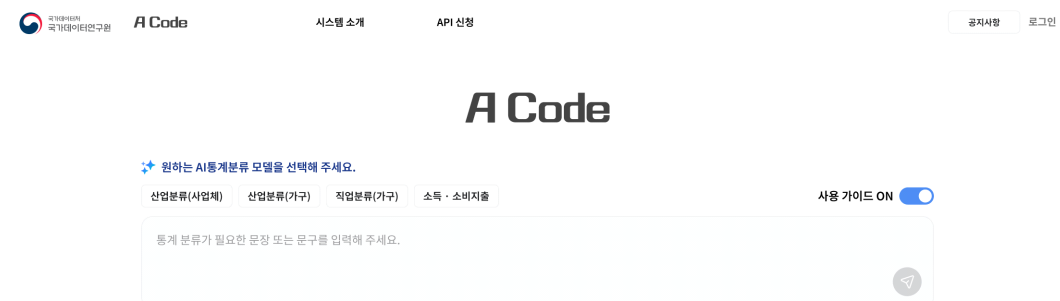
정비담당자가 산업분류를 검토할 때 모호한 사례가 있을 때마다 내부망에서 접속할 수 있는 AI통계분류시스템 홈페이지를 활용하는 방법도 있다. 본 연구의 일환으로 기업통계등록부 정비사업의 정비담당자가 편리하게 활용할 수 있도록 기존 2022년에 구축한 AI통계분류시스템 홈페이지(이하 A Code)를 전면적으로 개편하였다.

기존 홈페이지에서도 원하는 분류의 AI 모델을 선택할 수는 있으나, 각 모델에 대한 설명이 따로 기재되어 있지 않아 A Code에 대한 아무 배경지식이 없는 이용자가 사용하기는 어려웠다. 따라서 A Code를 처음 활용하는 초보자도 어려움이 없도록 불필요한 항목은 제거하고 직관적인 디자인으로 변경하였으며, AI 모델별 설명 및 입력 예시를 제시하는 등 사용자 친화적으로 구성하여 시스템 활용 편의성을 개선하였다.

<그림 27> 기존 AI통계분류시스템 홈페이지 초기화면



<그림 28> 개선 AI통계분류시스템 홈페이지 초기화면



A Code에서 검색을 지원하는 AI통계분류 모델은 산업분류(사업체), 산업분류(가구), 직업분류(가구), 소득·소비지출로 총 4가지인데, 이중 기업통계등록부 정비사업에서는 ‘산업분류(사업체)’ 모델을 활용하면 된다. 이는 전국사업체조사 자료를 학습한 모델로, 제11차 산업분류를 세세분류(5자리)로 제공하므로 정비사업에 활용하기에 적합하다.

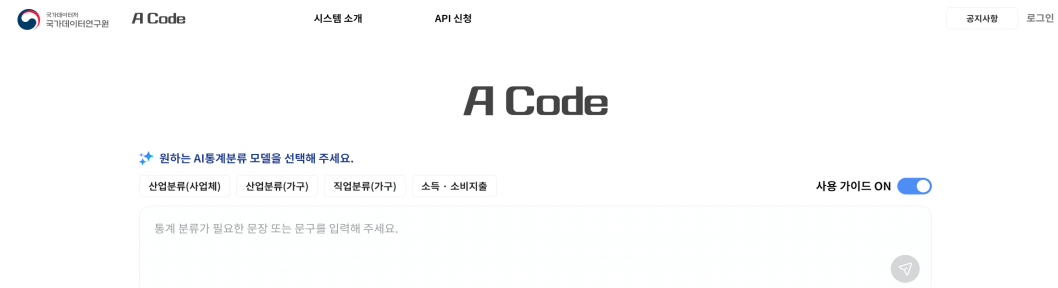
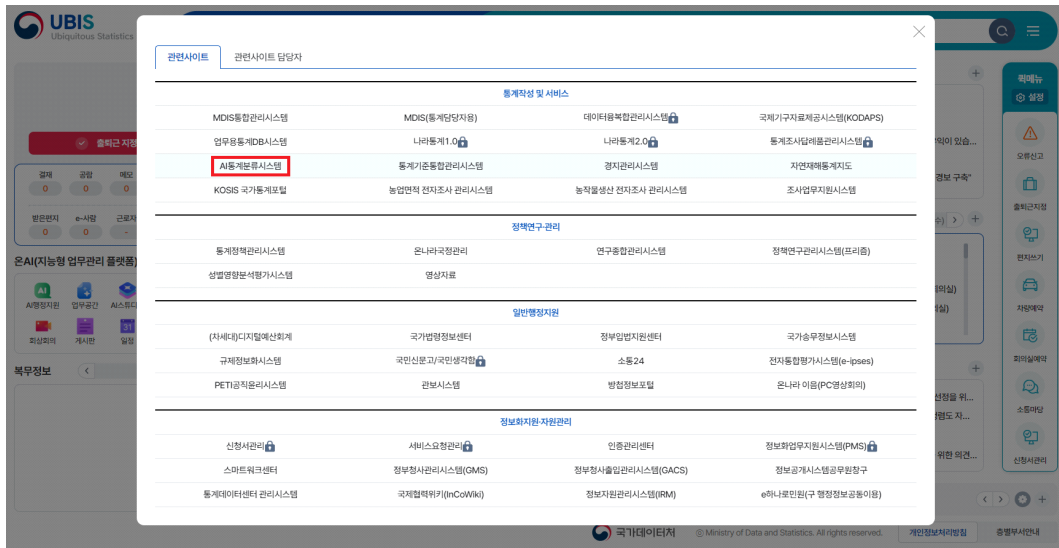
그 외 산업분류(가구)와 직업분류(가구) 모델은 지역별고용조사 자료를 학습한 모델로, 제11차 산업분류와 제8차 직업분류를 세분류(4자리)로 제공하며, 소득·소비지출 모델은 가계동향조사 자료를 학습한 모델로, 목적별 소비지출분류를 세분류(4자리)로 제공한다.

기업통계등록부 정비사업 정비담당자가 A Code를 활용하는 방법은 이리하다.

1) AI통계분류시스템 접속

내부망 UBIS 상단의 관련사이트 목록에서 AI통계분류시스템을 찾아 누르면, 손쉽게 사이트에 접속할 수 있다. 또한, 정비담당자가 주로 사용하게 될 ‘통계분류 검색 기능’은 A Code 홈페이지 별도의 로그인 없이 이용이 가능하다.

<그림 29> UBIS 관련사이트 목록 및 A Code 초기화면



2) 산업분류(사업체) 모델 선택 후 입력

AI통계분류 모델 중 산업분류(사업체) 모델을 선택하면, 모델 선택버튼 하단에 해당 모델의 사용 가이드가 표시된다. 이 모델은 전국사업체조사 항목 중 ‘사업체명’, ‘무엇을 가지고’, ‘어떤 방법으로’, ‘생산·제공하였는가’의 내용을 학습하여, 제11차 산업분류를 세세분류(5자리)까지 제공한다. 따라서 이용자가 학습데이터와 비슷한 형식으로 검색하여 보다 정확한 추론 결과를 얻을 수 있도록 안내 문구와 입력 예시를 기재하였다.

또한 모델 선택 후에만 검색이 가능하도록 모델을 선택하지 않으면 검색이 되지 않게 입력창을 닫아두었으며, A code 이용자의 숙련도 및 사용 편의 등 필요에 따라 우측 상단의 버튼을 통해 사용 가이드를 활성화(ON) 또는 비활성화(OFF)할 수 있다.

<그림 30> 산업분류(사업체) 모델 사용 가이드 - 활성화



<그림 31> 산업분류(사업체) 모델 사용 가이드 - 비활성화



3) 분류 결과 확인

이용자가 입력한 내용은 입력창 하단에 위치하게 되며 AI통계분류시스템은 AI 추천 1순위부터 3순위까지의 결과를 제공한다. AI 추천 1순위~3순위 각 카드 우측 상단에는 해당 분류의 ‘신뢰도’ 즉, AI가 해당 분류를 추론할 때 그 결과를 얼마나 확신하는지 AI 스스로 0~100%까지의 범위로 확신 정도를 매긴 값을 제시한다.

또한 각 카드를 누르면 해당 세세분류(5자리)에 대한 설명자료가 나오는데, 분류 결과 하단의 버튼 선택을 통해 대분류부터 세분류까지의 설명 및 예시도 확인할 수 있다. 따라서 검색 후 분류 결과의 내용을 읽어보고 어떤 분류가 적합한지 판단할 수 있으며, 우측 맨 하단의 ‘검색 입력창으로 돌아가기 버튼’을 통해 입력창으로 돌아갈 수 있다.

<그림 32> AI 추천 1순위 ~ 3순위 분류 결과

The screenshot displays the 'A Code' AI classification system interface. At the top, there's a navigation bar with 'A Code', '시스템 소개', and 'API 신청' links. Below this, a search bar contains the text '한식 음식점'. The main area shows three AI-recommended categories with their respective confidence scores: '한식 일반 음식점업' (56111) at 94.2%, '한식 요리 전문점' (56113) at 3.3%, and '기타 주점업' (56219) at 0.9%. Each category has a button to view the detailed explanation. The interface also includes a 'LLM으로 분류 결과 검증' button and a '검색 입력창으로 돌아가기' button.

4) LLM 분류 검증 결과 확인

추가로 AI 추천 3순위 카드 하단의 ‘LLM으로 분류 결과 검증’ 버튼을 누르면 생성형 LLM 모델로 추론한 결과도 확인 할 수 있다. 이는 2026년 11월 완료 예정인 다른 연구 과제⁹⁾에서 추진 중인 내용으로, 기본적인 틀은 생성형 LLM 모델의 종합 판단을 먼저 제시하고 각 후보 분류에 대한 적합도, 근거, 대안 등 검토 내용을 나열한다.

9) 통계분야에 특화된 생성형 LLM 구축방안 연구(임경민, 최하은, 이규호, 2026)

유의할 점은 LLM 분류 결과가 기존 분류 결과와 항상 같은 것은 아니라는 것이다. 기존 분류 결과는 머신러닝(ML) 모델로 추론한 것이기 때문에 LLM 모델 추론 결과와 다르게 나타날 수 있다. 또한 LLM 모델은 추론할 때마다 결과가 달라므로 분류 결과 확인 후 다른 답변을 원하면 ‘LLM으로 다시 검증’ 버튼을 통해 재검증할 수 있다.

<그림 33> LLM 분류 검증 결과

A Code

시스템 소개
API 신청

공시사항
로그인

A Code

산업분류(사업제)
산업분류(가구)
작업분류(가구)
소독·소비지출

사용 가이드 ON ON

전국사업체조사 자료를 학습한 모델로, 제11차 산업분류를 세세분류(5자리)로 제공합니다.
"사업체명 + 무엇을 가지고 + 어떤 방법으로 + 생산·제공하였는가"를 입력하면 보다 정확한 결과를 제공합니다.

예시 1. "주식회사 OO 경영 컨설팅업체가 경영 문제에 관하여 경영자문 및 자문"
예시 2. "OO 건축공사현장에서 전기배선 공사를 전문적으로 수행 전기공사업"

한식 음식점

↗

☞ 한식 음식점

★ 아래 카드를 누르면 각 분류에 대한 설명을 확인하실 수 있습니다.

AI 추천 1순위

신뢰도 94.2%

(56111) 한식 일반 음식점업

대분류 중분류 소분류 세분류 세세분류

AI 추천 2순위

신뢰도 3.3%

(56113) 한식 육류 요리 전문점

대분류 중분류 소분류 세분류 세세분류

AI 추천 3순위

신뢰도 0.9%

(56219) 기타 주점업

대분류 중분류 소분류 세분류 세세분류

LLM으로 다시 검증

분류 결과
LLM 분류 검증 결과

종합 판단

- 검토 필요: 시스템이 제안한 후보 중 '56111 한식 일반 음식점업'은 적합하지만, 나머지 후보들은 입력된 '한식 음식점'이라는 모호한 범위를 지나치게 세부화하거나 성격이 다른 업종을 포함하고 있어 재검토가 필요합니다. 특히 '56219 기타 주점업'은 음식점업이 아닌 주점업이므로 입력 문의의 의도와 다를 가능성이 높습니다.

분류 검토

한국표준산업분류 (KSIC)

- 후보 1: 56111 한식 일반 음식점업
 - 적합도: 적함
 - 근거: 입력된 '한식 음식점'의 가장 포괄적이고 전형적인 정의에 부합함. 백반, 찌개, 탕 등 일반적인 한식 서비스를 제공하는 산업활동을 포함함.
 - 대안: 없음
- 후보 2: 56113 한식 육류 요리 전문점
 - 적합도: 부분 적함
 - 근거: 한식 음식점의 하위 유형 중 하나이나, 입력문이 특정 메뉴(육류 구이 등)를 명시하지 않았으므로 이를 상위 개념인 '한식 음식점'의 대표 분류로 바로 지정하기에는 제한적임.
 - 대안: 56111 한식 일반 음식점업 (입력문의 포괄성을 고려할 때 더 적절)
- 후보 3: 56219 기타 주점업
 - 적합도: 부적함
 - 근거: 해당 코드는 술을 주된 목적으로 하는 주점업으로, '음식점'을 검색한 입력 의도와는 성격이 다름. 한식 음식점 내에서 주류를 판매할 수는 있으나, 산업 분류상 주점업은 별도로 구분됨.
 - 대안: 없음

참고 안내

- 한식 음식점은 제공하는 요리의 종류(면류, 해산물 요리 등)와 운영 방식에 따라 세부 코드가 달라집니다.
- 만약 입력된 '한식 음식점'이 냉면, 칼국수 등 면 요리를 전문으로 한다면 56112(한식 면 요리 전문점)로, 한국식 찻집이나 생선구이집 형태라면 56114(한식 해산물 요리 전문점)로 분류될 수 있습니다.
- 또한, 집계 시점 없이 이동식으로 운영하는 경우(포장마차 등)는 56142(이동 음식점업)로 분류되므로, 보다 정확한 분류를 위해서는 주요 메뉴(면, 해산물, 육류 등)나 운영 형태(고정식, 이동식)에 대한 추가 정보가 필요합니다.

※ 생성형 AI 검증 결과는 참고용입니다. 최종 판단은 분류해설서·공식 자료로 확인해 주세요.

3. 기업통계등록부 정비 시스템에 실시간 API 연계

기업통계등록부 정비사업에 AI 산업 자동분류를 도입하는 방안으로 앞에서 서술한 내검 시스템 효율화, AI통계분류시스템 홈페이지 활용 등도 있으나, 결국에는 정비 담당자가 사용하는 기업통계등록부 정비 시스템에 AI 산업분류 추론 결과를 실시간 API로 받아 직접적으로 활용하는 것이 최종 실무 활용 단계라고 볼 수 있다.

현재 나라통계시스템을 통해 실시간 API로 AI통계분류시스템 추론 결과를 활용하는 조사는 대표적으로 경제활동인구조사, 가계동향조사 중 2가지이다. 경제활동인구조사는 직업분류 검색에는 기존 자동코딩 시스템(색인DB)을 유지하고 있으나, 산업분류 검색에 대해서는 AI 추론 결과를 활용하고 있다.

가계동향조사는 가계부 항목 분류 검색에 AI 추론 결과와 기존 자동코딩 시스템(사례사전)을 병행하여 활용하고 있으며 본청 내검 단계에서 AI 추론 결과를 참고하는 정도였으나, 2026년 7월부터는 시스템에서도 항목 분류 일치 내검 자료로 활용하고 있다.

<그림 34> 경제활동인구조사 산업분류 및 직업분류 검색 화면

코드	분류항목명	확률
1 8610	병원	97.94%
2 8620	의원	1.955%
3 7310	수의업	0.01%
4 8690	기타 보건업	0.008%
5 8630	공중 보건 의료업	0.007%

코드	분류항목명	빈도수	방식
1 2440	간호사	100.0%	색인DB
2 4331	수의사 보조원	42.58%	색인DB
3 2495	간호조무사	11.52%	색인DB
4 2464	치과위생사	11.52%	색인DB

<그림 35> 가계동향조사 가계부 항목 분류 검색 화면

AI예측결과

No.	구분	AI입력코드	AI항목명	예측점수(100점 기준)
1	지출	1592	한과 및 기타과자	99.946
2	지출	7515	외식비	0.009
3	지출	1565	초콜릿	0.008

사례사전 검색결과

※확장검색시, ▾버튼을 눌러 검색항목(키워드) 확인바랍니다.

No.	구분	입력코드	항목명	내용	검색항목
1	지출	1592	한과 및 기타과자	◆ 한국의 전통과자 및 위의 항목에 분류되지 않은 당류 및 과자류 ◆ 비스킷, 스낵과자에 초콜...	[과자]

<그림 36> 가계동향조사 항목분류 일치내검(AI) 화면

항목분류 일치내검(AI) 홈 > 내검 > 항목분류 일치내검(AI)

지정항목코드: 사무소코드: 확인결과: 전체

조사년월: 2026 05 구분: 전체 조사구번호: 담당자명:

총 31건

No.	☐	조사년월	사무소명	사무소코드	조사구	구역번호	가계번호	가구번호	담당자명	가계부내	내용	금액	입력코드	항목명	시코드 1순위	시항목명 1순위	시코드 2순위	시항목명 2순위	확인결과
1	☐	202605								20260521	원부	15,000	1184	기타연금수...	1195	기타수선등...	1184	기타연금수...	1: 시코드1순위
2	☐	202605								20260522	가구주 국만...	250,000	0411	가구주공작...	9114	직장가입 국...	9112	지역가입 국...	0: 기존코드
3	☐	202605								20260503	행(러일지즈...	25,960	1092	식행 및 기타...	1085	케이...	1092	식행 및 기타...	0: 기존코드
4	☐	202605								20260503	윌타미버드...	14,500	1690	죽석·동결식품	1060	국수류	1690	죽석·동결식품	0: 기존코드
5	☐	202605								20260504	차량용기...	6,000	3610	식기류	3690	기타주방용품	3610	식기류	0: 기존코드
6	☐	202605								20260522	의료대리프...	22,020	4060	보건의료 소...	4090	기타보건약...	4060	보건의료 소...	0: 기존코드
7	☐	202605								20260523	의료대리프...	22,020	4060	보건의료 소...	4090	기타보건약...	4060	보건의료 소...	0: 기존코드
8	☐	202605								20260505	요양이 카페...	20,000	7560	주점·카페...	5290	반려동물 관...	7560	주점·카페...	0: 기존코드
9	☐	202605								20260508	피어노악원...	7,200	5407	참고서 및 책...	5437	잡지 및 기타...	5425	기타종이서적	0: 기존코드
10	☐	202605								20260515	자석배위...	4,800	3790	기타소형공...	3690	기타주방용품	2725	수선재료 및 ...	2: 시코드2순위
11	☐	202605								20260522	동배 고유가...	200,000	0641	가구주기타...	0621	가구주기초...	0622	2번가구원가...	0: 기존코드
12	☐	202605								20260522	첫째 고유가...	200,000	0641	가구주기타...	0622	2번가구원가...	0621	가구주기초...	0: 기존코드
13	☐	202605								20260523	연문	8,340	1340	함의	5018	시창기 콘넨...	1340	함의	0: 기존코드
14	☐	202605								20260523	여의당	13,100	1690	죽석·동결식품	7515	외식비	1186	여의	2: 시코드2순위
15	☐	202605								20260524	연문	7,750	1340	함의	5018	시창기 콘넨...	1340	함의	0: 기존코드
16	☐	202605								20260507	원배용기...	1,000	3610	식기류	3690	기타주방용품	3610	식기류	0: 기존코드
17	☐	202605								20260518	군고구미...	5,000	7515	외식비	1495	채소가공품	7515	외식비	1: 시코드1순위
18	☐	202605								20260522	엘리트...	6,400	1690	죽석·동결식품	7515	외식비	1492	기타채소	0: 기존코드
19	☐	202605								20260523	변전통	2,000	3610	식기류	3690	기타주방용품	3610	식기류	0: 기존코드

추가로 기존에 자동코딩 시스템(사례사전, 색인DB)을 활용하고 AI 추론 결과는 본청 내검 단계에서 참고 자료로만 활용했던 지역별고용조사도 AI통계분류시스템의 단계적 도입을 추진 중이다. 올해 하반기부터 나라통계시스템에서 산업 및 직업분류 검색 시 제시하는 색인DB 1~3순위를 AI 추론 1~3순위로 대체할 예정(시계열 유지 등을 위해 사례사전은 기존과 동일하게 유지)이며, 내년 하반기에는 자동코딩 시스템에서 AI통계분류시스템으로 전면 전환 예정이다.

비록 기업통계등록부 정비사업은 나라통계시스템이 아닌 데이터융복합관리시스템을 사용하지만, 이와 동일한 방식으로 AI통계분류시스템 추론 결과를 실시간 API를 통해 제공받아 기업통계등록부 정비 시스템에서 직접적으로 활용할 수 있다.

본 연구 결과를 토대로 AI통계분류시스템 도입을 면밀하게 검토하여 <그림 37>처럼 현재 정비담당자가 정비 시스템에서 색인검색을 하는 방식과 같이 산업분류 입력에서 검색을 통해 AI 추론 결과를 바로 확인할 수 있도록 기업통계등록부 정비 시스템에 AI통계분류시스템을 정식 탑재하는 방안을 고려해 볼 수 있을 것이다.

<그림 37> 현재 기업통계등록부 정비 시스템 산업분류 검색 화면

산업분류

기존 : 56191 김밥 및 기타 간이 음식점업 산업분류가 변경되었습니다

주산업 : (11차) ☐ 확인불가

주산업 : (10차) ☐ 확인불가

11차 산업분류 코드 검색

· 산업분류 색인 검색

● 산업분류 목록 [1건]

번호	산업분류	산업분류명	대분류항목	소분류항목	세분류항목	세세분류항목	5자리	4자리	3자리	비고
1	56150	제과점 운영	음식점 및 ...	음식점업	제과점업	제과점업	선택	선택	선택	확인

제 5 장

결론 및 시사점

제1절 연구요약

본 연구는 국가데이터처 행정자료관리과에서 수행하는 기업통계등록부 정비사업의 산업분류 부여 업무에 AI통계분류시스템을 도입하기 위한 기초자료를 마련하고, 실제 정비사업 자료를 활용하여 AI 산업 자동분류의 성능과 실무 적용 가능성을 분석하였다.

기업통계등록부는 다양한 경제통계 작성과 표본조사의 모집단 자료로 활용되는 국가 통계의 중요한 기반자료로서, 수록 정보의 정확성과 최신성을 지속적으로 관리하는 것이 중요하다. 특히, 산업분류는 사업체의 주요 경제활동을 파악하고 산업별 통계 생산의 기준이 되는 핵심 정보로, 기업통계등록부 정비사업에서 중요한 정비 항목에 해당한다. 이에 본 연구에서는 기업통계등록부 정비 과정에서 확보되는 사업체 정보를 AI통계분류 시스템에 적용하여 AI 산업분류 성능을 분석하고, 입력자료의 특성과 AI 추론 결과의 정확도 및 신뢰도를 중심으로 실무 활용 가능성을 검토하였다.

첫째, 기업통계등록부 정비 과정에서 확보한 ‘사업내용’은 AI 산업분류의 추론 성능 향상에 중요한 역할을 하는 것으로 나타났다. 동일한 사업체를 대상으로 정비 전 작업 배정 테이블과 정비 후 최종 테이블을 비교한 결과, AI 추론 1순위 산업분류와 등록분류 간 일치율은 작업배정 테이블에서 71.5%, 최종 테이블에서 93.1%로 나타났다. 이는 원 자료의 업종·업태 정보에 비해 정비 과정에서 확인한 ‘사업내용’이 AI가 사업체의 실제 경제활동을 추론하는 데 보다 유용한 정보로 활용될 수 있음을 보여준다.

둘째, ‘사업내용’의 작성 방식에 따라 AI 산업분류 추론 성능에 차이가 나타났다. 기존 방식으로 작성된 ‘사업내용’을 활용한 경우, AI 추론 1순위 산업분류와 등록분류 간 일치율은 90.3%였으나, 전국사업체조사의 작성 방식에 맞추어 ‘사업내용’을 보다 구체적으로 작성한 변경 방식에서는 96.8%로 나타났다. 또한 ‘사업내용’의 평균 글자 수도 기존 방식보다 증가하였으며, 일정 수준 이상 구체적으로 작성된 ‘사업내용’에서 상대적으로 높은 분류 성능이 나타났다. 이는 AI 산업분류를 위한 입력자료를 구성할 때 단순히 입력 텍스트의 길이를 늘리는 것보다 사업체의 실제 경제활동을 구체적이고 명확하게 설명하는 것이 중요함을 보여준다.

셋째, AI가 제시하는 신뢰도와 산업분류 결과의 일치율 간에 관련성이 있음을 확인하였다. 최종 테이블의 ‘사업내용’이 있는 사업체를 대상으로 분석한 결과, AI 추론 1순위 신뢰도가 높을수록 등록분류와의 일치율도 높아지는 경향이 나타났다. 특히, 신뢰도 90% 이상인 사업체는 전체의 75.0%를 차지하였으며, 해당 집단의 AI 1순위 산업분류와 등록분류 간 일치율은 98.8%로 나타났다. 이는 AI의 신뢰도 정보가 산업분류 결과를 검토하고 활용하는 과정에서 유의미한 판단자료가 될 수 있음을 보여준다.

종합하면, 본 연구는 실제 기업통계등록부 정비사업 자료를 활용하여 AI 산업 자동분류의 성능과 활용 가능성을 실증적으로 검토하였다는 데 의의가 있다. 분석 결과, 정비 과정에서 확보되는 ‘사업내용’의 구체성과 충실도는 AI 산업분류 성능에 영향을 미치는 주요 요인으로 나타났으며, AI가 제시하는 산업분류 후보와 신뢰도 역시 산업분류 부여 업무를 지원하는 정보로 활용될 가능성을 확인하였다. 이러한 결과는 향후 AI통계분류시스템을 기업통계등록부 정비사업에 활용하기 위한 기초자료로 활용될 수 있을 것으로 판단된다.

제2절 향후 과제

본 연구 결과는 AI통계분류시스템의 활용을 단순히 산업분류 결과를 자동으로 제시하는 기능에 한정하기보다, 입력자료의 품질관리, 담당자의 의사결정 지원 및 통계자료의 품질관리와 연계하여 종합적으로 검토할 필요가 있음을 시사한다. 이에 향후 AI통계분류시스템의 활용성과 안정성을 높이기 위해 다음과 같은 과제를 검토할 필요가 있다.

첫째, AI 산업분류에 활용되는 입력자료의 품질을 지속적으로 관리할 필요가 있다. AI 산업분류의 성능은 모델 자체의 기술적 성능뿐만 아니라, 입력되는 정보의 구체성과 충실성에도 영향받을 수 있다. 따라서 향후 AI 산업분류의 활용을 확대하기 위해서는 사업체의 실제 경제활동을 충실히 반영할 수 있도록 ‘사업내용’ 작성 기준과 자료 입력 방식을 지속적으로 관리할 필요가 있다. 특히, 정비 담당자나 자료 수집 방식에 따라 정보의 작성 수준에 차이가 발생하지 않도록 작성 기준을 구체화하고, 산업분류 추론에 필요한 정보가 일관되게 축적될 수 있는 자료 관리 체계를 마련할 필요가 있다.

둘째, AI를 담당자의 의사결정을 지원하는 도구로 활용하는 방안을 지속적으로 발전시킬 필요가 있다. AI가 제시하는 산업분류 후보와 신뢰도는 담당자가 최종적으로 산업분류를 결정하는 과정에서 참고할 수 있는 추가적인 정보가 될 수 있다. 따라서 AI가 사람의 판단을 일률적으로 대체하기보다는, 담당자가 보다 효율적이고 일관된 판단을 할 수 있도록 지원하는 방향으로 활용될 필요가 있다. 특히, 통계분류와 같이 사례별 특성과 입력 내용을 종합적으로 판단해야 하는 업무에서는 사람의 전문적 판단과 AI의 분석 결과를 상호 보완적으로 활용하는 방식이 중요하다.

셋째, AI 추론 결과를 기업통계등록부 정비자료의 품질관리에도 활용할 필요가 있다. AI 추론 결과와 등록분류가 일치하지 않는 사례에는 AI의 오류뿐만 아니라 기존 등록분류에 대한 추가적인 확인이 필요한 경우도 포함될 수 있다. 따라서 불일치 사례를 단순히 AI 오분류 사례로 처리하기보다, 기존 등록분류의 오류 가능성을 점검하고 추가적인 확인이 필요한 사례를 식별하는 정보로 활용할 필요가 있다. 향후 불일치 사례의 유형과 원인을 체계적으로 분석한다면 AI통계분류시스템을 산업분류 부여 업무를 지원하는 도구를 넘어, 기존 통계자료의 오류 가능성을 탐지하고 품질점검 대상을 선정하는 지원 도구로 발전시킬 수 있을 것으로 기대된다.

종합하면, 향후 AI통계분류시스템의 활용은 단순히 AI의 분류 정확도를 높이는 데 그치지 않고, 양질의 입력자료를 지속적으로 확보하고, 담당자의 판단을 효과적으로 지원하며, AI와 기존 정보 간의 차이를 통계자료의 품질관리로 연결하는 방향으로 발전시킬 필요가 있다. 향후 제4장에서 제시한 구체적인 실무 활용 방안을 실제 업무에 적용하고, 운영 결과를 지속적으로 검토, 보완한다면 AI통계분류시스템은 기업통계등록부 정비사업의 업무 효율성을 높이는 동시에 산업분류 부여 업무의 품질과 일관성을 향상하는 데에 기여할 수 있을 것으로 기대한다.

참고문헌

- 국가데이터처(2026) 「2026년 기업통계등록부 이용자 안내서」
국가데이터처 행정자료관리과(2026) 「2026년 기업통계등록부 정비 세부추진계획」
국가데이터처(2024) 「한국표준산업분류 제11차 개정 해설서」
국가데이터처(2026) “AI통계분류시스템 유지관리 사업 착수보고회” 발표자료
임경민, 김혜란, 오교중(2020) “데이터 과학 기법의 국가통계 시험적용 분석”, 국가데이터연구원
임경민, 김혜란(2021) “인공지능 기반 통계분류 자동코딩 시스템 구축 연구”, 국가데이터연구원
임경민, 김혜란(2021) “인공지능 기반 산업분류 자동화 처리 연구”, 국가데이터연구원
임경민(2023) “AI 통계분류 결과 분석 및 실무 활용성 제고 방안 연구”, 국가데이터연구원
김건수, 임경민(2023) “인공지능 통계분류 자동화 확대 적용 연구”, 국가데이터연구원
임경민(2024) “인공지능 기반 행정자료 품질관리 도입방안 연구”, 국가데이터연구원
임경민(2024) “AI 통계분류를 위한 가계동향조사 학습데이터 개선방안 연구”, 국가데이터연구원
임경민(2025) “가계동향조사 AI 통계분류 실무활용 연구”, 국가데이터연구원
이규호, 임경민(2025) “AI기반 건설경기동향 자동분류 코딩 연구”, 국가데이터연구원

Abstract**Introduction of AI-based Automatic Industry Classification for the Statistical Business Register Maintenance Project****Ha-eun Choi, Kyung-min Lim, Kyu-ho Lee**

This study aims to propose a plan for effectively introduce AI-based automatic industry classification into the industry classification code assignment task of the statistical business register maintenance project, which the Administrative Data Management Division of the Ministry of Data and Statistics annually carries out.

The maintenance project involves reviewing and updating business registration information for approximately 1.3 million newly registered establishments each year, through manual work including data collection via telephone, by about 130 newly hired maintenance staff members. Industry classification, which usually leads to the most frequent modification process, requires different workload and shows different accuracy levels depending on the staff member's experience and competence. To reduce this gap and enhance business efficiency, the Administrative Data Management Division reviewed the introduction of AI-assisted automatic industry classification system.

Using data from the 2025 maintenance project, this research analyzed the inference performance of the AI-based industry classification system. The analysis results showed that the match rate between the first assigned industry classification code of the AI-assisted inference task using 'business type' in raw data and the registered classification code stood at 71.5%. On the other hand, it rose to 93.1% when using 'business information' found during the maintenance process. In addition, when reflecting detailed 'business information' found from AI model's learning data, the match rate rose from 90.3% when using the existing method to 96.8%. These findings indicate that AI-driven industry classification performance depends on the content and format of the input information, and that specific and clear 'business information' can serve as useful data to improve AI-based inference performance.

These results imply that the application of AI-based industry classification system to the industry classification assignment task in the statistical business register maintenance project can improve business efficiency and consistency in industry classification. Therefore, to facilitate the practical application of the AI-based automatic industry classification, this research suggested practical implementation strategies such as streamlining the editing system, utilizing the AI-driven statistical classification system website (A code), and applying real-time API integration with the statistical business register maintenance system.

Key words: AI-based industry classification system, statistical business register maintenance project, industry classification

연구진

- 최하은(국가데이터처 국가데이터연구원 데이터과학연구팀 주무관)
- 임경민(국가데이터처 국가데이터연구원 데이터과학연구팀 사무관)
- 이규호(국가데이터처 국가데이터연구원 데이터과학연구팀 주무관)

* 연구진의 소속 및 직급은 연구과제 완료 시 기준임을 알려드립니다.

연구보고서 2026-03

기업통계등록부 정비사업 시 산업 자동분류 도입방안 연구

인 쇄	2026년 9월
발 행	2026년 9월
발 행 인	백지선
발 행 처	국가데이터처 국가데이터연구원 35220 대전광역시 서구 한밭대로 713 Tel.(042)366-7100 Fax.(042)366-7123
홈페이지	https://mods.go.kr/dsri/
ISSN(Online)	2733-4120



국가데이터처
국가데이터연구원