

인공지능(AI) 기반 조사데이터 품질 제고 방안 연구

이기호

오미애·조용찬·염아림·김은주·윤열매·최호식



사람을
생각하는
사람들



KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS



한국 보건사회연구원
KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS



■ 연구진

연구책임자	이기호	한국보건사회연구원 부연구위원
공동연구진	오미애	한국보건사회연구원 연구위원
	조용찬	한국보건사회연구원 연구원
	염아림	한국보건사회연구원 전문원
	김은주	한국보건사회연구원 선임전문원
	윤열매	한국보건사회연구원 전문원
	최호식	서울시립대학교 교수

연구보고서 2025-04

인공지능(AI) 기반 조사데이터 품질 제고 방안 연구

발행일 2025년 3월
발행인 신영석
발행처 한국보건사회연구원
주소 [30147] 세종특별자치시 시청대로 370
세종국책연구단지 사회정책동(1~5층)
전화 대표전화: 044)287-8000
홈페이지 <http://www.kihasa.re.kr>
등록 1999년 4월 27일(제2015-000007호)
인쇄처 (주)에이치에이엔컴퍼니

© 한국보건사회연구원 2025
ISBN 979-11-7252-083-0 [93510]
<https://doi.org/10.23060/kihasa.a.2025.04>

발|간|사

최근 몇 년간 인공지능(AI) 기술은 유례없는 속도로 발전하고 있으며, 다양한 산업 분야에서 급속히 채택되고 있다. 자연어 처리, 기계학습, 예측 분석, 컴퓨터 비전 등 인공지능(AI) 기술은 이제 전 세계 디지털 전환 전략의 핵심으로 자리 잡고 있으며, 정부, 기업, 연구기관 등은 인공지능(AI) 시스템을 활용해 효율성을 높이고, 의사결정을 자동화하며, 방대한 데이터를 통해 인사이트를 도출하고 있다.

인공지능(AI) 기술이 데이터 분석 및 통계 처리에 폭넓게 활용되면서, 조사데이터의 신뢰성과 정확성을 제고하기 위한 기술적 접근에 대한 관심이 커지고 있다. 대규모 패널 데이터를 다루는 경우, 입력 오류나 정보 누락, 논리적 비일관성 등의 문제가 빈번하게 발생하는데, 이러한 오류를 효과적으로 해결하지 못하면 연구 결과의 타당성이 심각하게 훼손될 수 있다. 그러나 기존의 데이터 품질관리 방식은 대부분 수작업에 의존하고 있어서 오류 검출 및 수정에 많은 시간과 인력이 소요되는 한계가 있다.

이에 본 연구에서는 조사데이터의 품질을 향상시키기 위해 인공지능(AI) 기술을 적용하는 방안을 탐색하고자 국내외 인공지능(AI)을 활용한 통계작성 사례를 분석하고, 조사데이터의 품질관리 내부 현황을 분석하여 조사데이터 품질관리 적용 방안을 도출하였다. 또한 본 연구원에서 매년 생산하고 있는 한국의료패널 데이터를 활용하여 오류 탐지 및 정정 과정을 자동화하고, 조사데이터의 품질을 실질적으로 개선하는 실증적 분석을 수행하였다. 향후 인공지능(AI) 기술이 조사 연구 및 통계 분석에서 점차 핵심적인 역할을 담당할 것으로 예상되는바, 본 연구가 향후 데이터 품질 향상과 연구 신뢰성 강화를 위한 기초 자료로 활용될 수 있기를 기대한다.

본 연구는 한국보건사회연구원 이기호 부연구위원의 책임하에 오미애 연구위원, 조용찬 연구원, 염아림 전문원, 김은주 선임전문원, 윤열매 전문원이 원내 연구진으로 참여하였으며, 외부 연구진으로는 서울시립대학교 최호식 교수가 참여하였다. 모든 연구진의 노고에 감사드린다.

보고서 작성과 관련하여 유익한 의견을 주신 외부 전문가, 그리고 익명의 검독위원들에게도 감사의 마음을 전한다. 마지막으로, 이 보고서의 내용은 본원의 공식적인 의견이 아님을 밝힌다.

2025년 3월
한국보건사회연구원 원장
신 영 석



목 차

KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS



요 약	1
제1장 서론	5
제1절 사업의 배경 및 목적	7
제2절 연구의 내용 및 추진 방법	10
제2장 이론적 배경	13
제1절 인공지능(AI)의 개념 및 최신 기술 동향	15
제2절 데이터 오류 탐지 방법론	34
제3절 통계업무프로세스모델	51
제4절 선행연구 검토	61
제5절 소결	69
제3장 인공지능(AI)을 활용한 통계작성 사례 분석	73
제1절 주요 외국의 인공지능(AI) 활용 통계작성 사례	75
제2절 우리나라의 인공지능(AI) 활용 통계작성 사례 분석	85
제3절 소결	94
제4장 조사데이터 품질관리 내부 현황 분석	97
제1절 조사데이터 품질관리	99
제2절 한국의료패널 조사데이터 품질관리 절차 및 방법	109
제3절 소결	118



제5장 조사데이터 품질관리 적용 방안 - 한국의료패널 데이터 실증 119

제1절 데이터 오류 수정 방안	121
제2절 한국의료패널 데이터 특성	129
제3절 데이터 실증	132
제4절 소결	179

제6장 결론 및 제언 181

제1절 결론	183
제2절 제언	186

참고문헌 191

부록 199

[부록 1] 검색증강기술(RAG) 분석 결과	199
--------------------------------	-----

Abstract 279

표 목차

KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS



〈표 1-1〉 한국보건사회연구원 조사데이터 생산 현황(2016~2022년)	8
〈표 2-2〉 지도 학습 방법 및 응용 사례	22
〈표 2-3〉 비지도 학습 방법 및 응용 사례	23
〈표 2-4〉 강화 학습 방법 및 응용 사례	24
〈표 2-5〉 합성곱 신경망(CNN)과 순환 신경망(RNN) 설명 및 응용 사례	26
〈표 2-6〉 예시 데이터와 대한 동시출현 빈도표	36
〈표 2-7〉 연관규칙 $X \rightarrow Y$ 에 대한 지지도, 신뢰도, 향상도	37
〈표 2-8〉 iForest의 예제	40
〈표 2-9〉 일반통계업무프로세스(GSBPM)-‘요구사항 명세’ 단계의 하위 프로세스	53
〈표 2-10〉 일반통계업무프로세스모델(GSBPM)-‘설계’ 단계의 하위 프로세스	54
〈표 2-11〉 일반통계업무프로세스모델(GSBPM)-‘구축’ 단계의 하위 프로세스	55
〈표 2-12〉 일반통계업무프로세스모델(GSBPM)-‘수집’ 단계의 하위 프로세스	56
〈표 2-13〉 일반통계업무프로세스모델(GSBPM)-‘처리’ 단계의 하위 프로세스	57
〈표 2-14〉 일반통계업무프로세스모델(GSBPM)-‘분석’ 단계의 하위 프로세스	58
〈표 2-15〉 일반통계업무프로세스모델(GSBPM)-‘배포’ 단계의 하위 프로세스	58
〈표 2-16〉 일반통계업무프로세스모델(GSBPM)-‘평가’ 단계의 하위 프로세스	59
〈표 2-17〉 1차 데이터의 머신러닝(ML) 활용 가능성(Yung 등, 2018)	62
〈표 2-18〉 국가통계의 인공지능(AI)·빅데이터 기술 도입의 두 가지 시각(김정민 등, 2020) ·	64
〈표 2-19〉 머신러닝(ML) 활용이 가능한 통계업무프로세스 및 예상 기술군(김정민, 2021) ·	65
〈표 2-20〉 통계업무 프로세스 단계별 인공지능(AI) 적용 가능 업무(이상현, 2021)	66
〈표 2-21〉 통계청의 AI 기법 적용 국가통계사례(이상현, 2021)	67
〈표 3-1〉 주요 국가 통계기관의 머신러닝 활용 실태	75
〈표 3-2〉 머신러닝(프로젝트) 통계 주제	77
〈표 3-3〉 머신러닝(프로젝트) 활용 형태	77
〈표 3-4〉 머신러닝(프로젝트) 알고리즘 방법	78
〈표 3-5〉 유엔 유럽경제위원회(UNECE)의 머신러닝(ML) 파일럿 연구(코딩 및 분류)	79
〈표 3-6〉 유엔 유럽경제위원회(UNECE)의 머신러닝(ML) 파일럿 프로젝트(편집 및 대체)	81



〈표 3-7〉 유엔 유럽경제위원회(UNECE)의 머신러닝(ML) 파일럿 연구(이미지 분석)	82
〈표 3-8〉 국가 및 국제기구의 대형 언어모델(LLM) 도입 사례(UNECE, 2023)	84
〈표 3-9〉 머신러닝 기반 산업분류 예측 성능 평가	86
〈표 3-10〉 한류 콘텐츠 AI 분류기의 장점과 단점	91
〈표 4-1〉 통계청의 통계 품질 차원	101
〈표 4-2〉 통계청의 통계 품질 진단 항목	103
〈표 4-3〉 공공데이터 품질 지표	105
〈표 4-4〉 공공데이터 품질 진단 방법	106
〈표 4-5〉 공공데이터의 품질 지표별 진단 방법	107
〈표 4-6〉 한국의료패널 조사데이터 내검 목록(이상치 탐지 목록)	112
〈표 5-1〉 데이터 오류 유형	121
〈표 5-2〉 데이터 오류를 수정하기 위한 소프트웨어	123
〈표 5-3〉 데이터 오류를 수정하기 위한 대표적 알고리즘	124
〈표 5-4〉 시각화 도구	125
〈표 5-5〉 YData Profiling report 옵션	125
〈표 5-6〉 예제 자료에 대한 YData Profiling 실행 예시	126
〈표 5-7〉 탐색적 자료 분석과 시각화를 위한 python 패키지	128
〈표 5-8〉 한국의료패널 데이터의 조사항목과 내용	129
〈표 5-9〉 데이터 오류를 수정하기 위한 알고리즘	130
〈표 5-10〉 세대 구성 변수의 코드와 설명	133
〈표 5-11〉 데이터 크기와 변수와 수준 수	133
〈표 5-12〉 ID 및 가구원 변동 사항(ID) 변수의 코드와 설명	134
〈표 5-13〉 가구원 데이터(IND) 변수의 코드와 설명	135
〈표 5-14〉 가구 데이터(HH) 변수의 코드와 설명	135
〈표 5-15〉 데이터 병합 예시	137
〈표 5-16〉 사전 검수 질의문	138
〈표 5-17〉 시각화 그래프에 활용되는 툴	146



〈표 5-18〉 변수와 파생변수 및 json 표현	152
〈표 5-19〉 가구데이터(HH)의 사망 가구원 포함 여부(DEATH_YN) 변수에 대한 ‘보기 문항 내용’과 ‘사용 주의사항’ 변수의 내용을 반영한 json 자료 구조	153
〈표 5-20〉 검수 질의문에 대한 논리적 오류 검토	154
〈표 5-21〉 분포 인지 RAG 기반 검수 질의문 추천 알고리즘	156
〈표 5-22〉 단계 1에서의 RAG 1 이상치 탐지 기준 생성 예시	158
〈표 5-23〉 단계 2에서의 RAG 2 이상치 범위 생성 예시	159
〈표 5-24〉 유형 1) 문항 번호 4	164
〈표 5-25〉 유형 2) 문항 번호 27-1	165
〈표 5-26〉 유형 2) 문항 번호 27-2	166
〈표 5-27〉 유형 2) 문항 번호 27-3	167
〈표 5-28〉 유형 2) 문항 번호 27-4	168
〈표 5-29〉 유형 3) 문항 번호 11	169
〈표 5-30〉 유형 4) 문항 번호 15	170
〈표 5-31〉 유형 5) 문항 번호 24	171
〈표 5-32〉 유형 6) 문항 번호 9	172
〈표 5-33〉 유형 7) 문항 번호 39	173
〈표 5-34〉 유형 8) 문항 번호 3	174
〈표 5-35〉 생성된 탐지 기준 적용 가능성 검토	177
〈부표 1〉 유형 1) 문항 번호 1	199
〈부표 2〉 유형 1) 문항 번호 2	200
〈부표 3〉 유형 1) 문항 번호 20	201
〈부표 4〉 유형 1) 문항 번호 29	202
〈부표 5〉 유형 1) 문항 번호 30	203
〈부표 6〉 유형 2) 문항 번호 8	204
〈부표 7〉 유형 2) 문항 번호 25-1	205
〈부표 8〉 유형 2) 문항 번호 25-2	206

〈부표 9〉 유형 2) 문항 번호 25-3	207
〈부표 10〉 유형 2) 문항 번호 25-4	208
〈부표 11〉 유형 2) 문항 번호 26-1	209
〈부표 12〉 유형 2) 문항 번호 26-2	210
〈부표 13〉 유형 2) 문항 번호 28-1	211
〈부표 14〉 유형 2) 문항 번호 28-2	212
〈부표 15〉 유형 2) 문항 번호 44-1	213
〈부표 16〉 유형 2) 문항 번호 44-2	214
〈부표 17〉 유형 2) 문항 번호 44-3	215
〈부표 18〉 유형 2) 문항 번호 44-4	216
〈부표 19〉 유형 2) 문항 번호 45-1	217
〈부표 20〉 유형 2) 문항 번호 45-2	218
〈부표 21〉 유형 2) 문항 번호 50	219
〈부표 22〉 유형 2) 문항 번호 53-1	220
〈부표 23〉 유형 2) 문항 번호 53-2	221
〈부표 24〉 유형 2) 문항 번호 53-3	222
〈부표 25〉 유형 2) 문항 번호 53-4	223
〈부표 26〉 유형 2) 문항 번호 57-1	224
〈부표 27〉 유형 2) 문항 번호 57-2	225
〈부표 28〉 유형 2) 문항 번호 57-3	226
〈부표 29〉 유형 2) 문항 번호 57-4	227
〈부표 30〉 유형 2) 문항 번호 57-5	228
〈부표 31〉 유형 2) 문항 번호 59-1	229
〈부표 32〉 유형 2) 문항 번호 59-2	230
〈부표 33〉 유형 3) 문항 번호 10	231
〈부표 34〉 유형 4) 문항 번호 3	232
〈부표 35〉 유형 4) 문항 번호 5	233



〈부표 36〉 유형 4) 문항 번호 14	234
〈부표 37〉 유형 4) 문항 번호 17	235
〈부표 38〉 유형 5) 문항 번호 16	236
〈부표 39〉 유형 5) 문항 번호 47	237
〈부표 40〉 유형 5) 문항 번호 49	238
〈부표 41〉 유형 6) 문항 번호 6	239
〈부표 42〉 유형 6) 문항 번호 7	240
〈부표 43〉 유형 6) 문항 번호 12	241
〈부표 44〉 유형 6) 문항 번호 13	242
〈부표 45〉 유형 6) 문항 번호 18	243
〈부표 46〉 유형 6) 문항 번호 19	244
〈부표 47〉 유형 6) 문항 번호 21	245
〈부표 48〉 유형 6) 문항 번호 22	246
〈부표 49〉 유형 6) 문항 번호 23	247
〈부표 50〉 유형 6) 문항 번호 31	248
〈부표 51〉 유형 6) 문항 번호 32	249
〈부표 52〉 유형 6) 문항 번호 34	250
〈부표 53〉 유형 6) 문항 번호 35	251
〈부표 54〉 유형 6) 문항 번호 36	252
〈부표 55〉 유형 6) 문항 번호 37	253
〈부표 56〉 유형 6) 문항 번호 38	254
〈부표 57〉 유형 6) 문항 번호 46	255
〈부표 58〉 유형 6) 문항 번호 48	256
〈부표 59〉 유형 6) 문항 번호 51	257
〈부표 60〉 유형 6) 문항 번호 52	258
〈부표 61〉 유형 6) 문항 번호 54	259
〈부표 62〉 유형 6) 문항 번호 55	260



〈부표 63〉 유형 6) 문항 번호 56	261
〈부표 64〉 유형 6) 문항 번호 58_1	262
〈부표 65〉 유형 6) 문항 번호 58_2	263
〈부표 66〉 유형 6) 문항 번호 61	264
〈부표 67〉 유형 6) 문항 번호 62	265
〈부표 68〉 유형 6) 문항 번호 63	266
〈부표 69〉 유형 7) 문항 번호 40~43	267

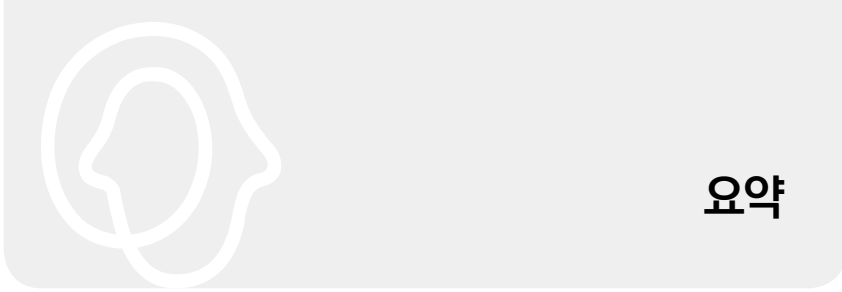
그림 목차

KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS



[그림 1-1] 연구의 내용 및 추진 방법	11
[그림 2-1] 인공지능에 대한 정의와 범주	16
[그림 2-2] 검색증강생성기술(RAG) 모델 아키텍처	45
[그림 2-3] LangChain 파이프라인 구조도	49
[그림 2-4] 일반통계업무프로세스모델(GSBPM)	52
[그림 2-5] 한국형 통계업무프로세스모델(KSBPM)	60
[그림 2-6] 머신러닝(ML) 적용 가능 분야(Beck 등, 2018)	61
[그림 2-7] 조사 연구에서 인공지능(AI) 활용 방안(이기재, 2023)	68
[그림 2-8] 인공지능을 활용한 통계 생산 방식 효율화(김정민, 2023)	68
[그림 2-9] 통계업무 프로세스 단계별 인공지능(AI) 적용 가능 업무(종합)	70
[그림 3-1] 연도별 미국 상해 및 질병 데이터 자동코딩 성공률	81
[그림 3-2] 지도 학습 기반 분류 학습을 위한 초기 기준 모델	85
[그림 3-3] 기존 시스템과 인공지능(AI) 통계분류 시스템 성능 비교(분류 예측 정확도) · 87	
[그림 3-4] 인공지능(AI) 통계분류 자동화 시스템 개념도	87
[그림 3-5] 인공지능(AI) 통계분류 모델 구축(5개 조사, 5종 분류) 및 성능평가 결과	88
[그림 3-6] 기존 시스템과 인공지능(AI) 통계분류 시스템 성능 비교(분류 예측 정확도) · 89	
[그림 3-7] 한류 콘텐츠 AI 분류기 모델 개념도	90
[그림 3-8] 한류 콘텐츠 인공지능(AI) 분류기 학습 과정	91
[그림 3-9] 규칙 기반 챗봇의 서비스 개념도	92
[그림 3-10] 대형 언어모델 기반 챗봇의 서비스 개념도	93
[그림 4-1] 한국의료패널 조사데이터 수집 과정	110
[그림 5-1] 예제 자료에 대한 YData Profiling 실행 결과	127
[그림 5-2] 유형 1과 2의 순서도	142
[그림 5-3] 유형 3과 4의 순서도	143
[그림 5-4] 유형 5와 6의 순서도	144
[그림 5-5] 유형 7과 8의 순서도	145
[그림 5-6] 샷 및 학습 전후의 이상치 탐지 예시	149

[그림 5-7] 테이블 구조 데이터와 세 변수의 3차원 텐서	150
[그림 5-8] RAG 2에서 활용되는 DB 예시	157
[그림 5-9] 생성된 규칙에 대한 2차원 결합분포와 주변분포	160
[그림 5-10] SAS 코드 생성 실험 결과	161
[그림 5-11] 이상치 범위 변경 후, SAS 코드 생성 결과	162
[그림 5-12] SAS 코드 대비 질의문의 정보 제한 예시	162
[그림 5-13] 질의문과 SAS 코드 간의 정보 불일치 예시	163
[그림 6-1] 자연어 질의문에 대한 코드 생성	188
[그림 6-2] 대형 언어모델(LLM)을 활용한 데이터 품질평가 시스템 아키텍처	189



1. 연구의 배경 및 목적

최근 인공지능(AI) 기술이 데이터 분석 및 통계 처리에서 중요한 역할을 수행하면서, 인공지능 기술을 활용한 조사데이터의 신뢰성과 정확성을 개선하는 방안에 대한 관심이 높아지고 있다. 특히, 대규모 패널 데이터를 처리할 때 발생할 수 있는 입력 오류, 정보 누락, 논리적 비일관성 등의 문제를 해결하지 않으면 연구 결과의 신뢰성이 저하될 위험이 크다. 한편 기존의 조사데이터 품질관리 방식은 수작업 중심으로 이루어지는 경우가 많아, 오류를 검출하고 수정하는 데 많은 시간과 인력이 필요하다는 한계가 있다.

본 연구에서는 한국의료패널 조사데이터 자료의 내검 업무를 중심으로 조사데이터의 품질 제고 및 업무 효율화를 위한 인공지능(AI) 기술 활용 방안을 모색하였다. 국내외 선행연구 및 사례를 분석하여 통계 생산 및 관리에 인공지능(AI) 기술을 활용할 수 있는지 그 가능성을 검토하고, 한국의료패널 데이터를 실증하여 우리 연구원의 조사데이터 품질 및 업무 효율성 제고를 위한 인공지능(AI) 기술 활용 방안을 제시하고자 한다.

2. 주요 연구 내용

본 연구의 주요 연구 결과는 다음과 같다.

첫째, 국내외 선행연구 고찰을 통해 인공지능(AI)의 개념 및 최신 기술 동향을 검토하였다. 인공지능(AI)에 관한 다양한 개념을 살펴보고, 머신러닝(ML), 딥러닝(DL), 대형 언어모델(LLM) 등 인공지능(AI) 관련 최신 기술 동향을 검토하였다. 인공지능(AI)의 개념은 진화하고 있으며, 그 기

2 인공지능(AI) 기반 조사데이터 품질 제고 방안 연구

술은 빠르게 발전하고 있는 것으로 나타났다.

둘째, 국내외 선행연구 고찰을 통해 조사데이터의 품질관리를 위한 인공지능(AI) 기반 데이터 오류 탐지 방법을 검토하였다. 구체적으로 isolation Forest 방법, AnoGan, 대형 언어모델(LLM)을 활용한 데이터 시각화, 대형 언어모델(LLM)을 활용한 데이터 오류 탐지 방법을 검토하였다.

셋째, 국내외 선행연구 고찰을 통해 인공지능(AI)을 활용한 통계 적용 가능 업무를 검토하였다. 설계 단계, 수집 단계, 처리 단계, 분석 단계, 배포 단계, 평가 단계의 다양한 통계업무에서 인공지능(AI) 적용 가능성이 검토되고 있는 것으로 나타났으며, 일부 통계업무의 경우 개념 증명 및 시험 단계를 거쳐 실무 적용 단계에 도달한 것으로 파악되었다.

넷째, 주요 외국 및 우리나라의 인공지능(AI)을 활용한 통계작성 사례를 검토하였다. 주요 외국의 경우 ‘코딩 및 분류’, ‘편집 및 대체’, ‘이미지 분석’ 업무에 머신러닝(ML) 및 딥러닝(DL) 기술의 활용 가능성을 확인할 수 있었으며, 전반적인 통계서비스 업무 개선에 대형 언어모델(LLM) 기술의 활용 가능성을 확인할 수 있었다. 한편, 우리나라의 경우 통계청뿐만 아니라 민간기관에서도 통계업무 자동화 및 효율화에 인공지능(AI) 기술을 도입한 것을 확인할 수 있었다.

다섯째, 조사데이터 품질관리 적용 방안으로 한국의료패널 데이터 실증을 통해 이상(異常)치 탐지를 위한 시각적 점검 기준 방법과 새로운 점검 기준 생성이라는 2가지 방안을 검토하였다. 우선, 이상치 탐지를 위한 시각적 점검 기준 방법은 ‘커널밀도함수(KDE), 히스토그램, Rugplot’으로 전체 데이터의 분포를 파악할 수 있고, 현재 내검 시 사용하고 있는 기준값 내에 데이터가 모두 포함되는지, 또는 현재 설정된 기준값을 조정할 필요가 있는지도 확인해볼 수 있는 그래프를 제시하였다. 다음으로, 분포 인지 검색증강생성기술(RAG) 기반 대형 언어모델(LLM) 접근법을 도입

하여 변수 간 결합분포(두 변수 간과 주변분포)를 학습시켜, 이상치의 가능성이 높은 변수 조합을 추천하는 방법론을 제안하였다. 데이터 실증 결과를 실무 담당자가 검토한 결과 시각적 점검 기준 방법의 경우 매우 유용하며, 실무에 적용 가능한 것으로 검토되었으며, 새로이 생성된 이상치 질의문 10개의 경우 5개(적용 가능 3개, 수정 후 적용 가능 2개)는 실무에 적용할 수 있는 것으로 검토되었다.

3. 결론 및 시사점

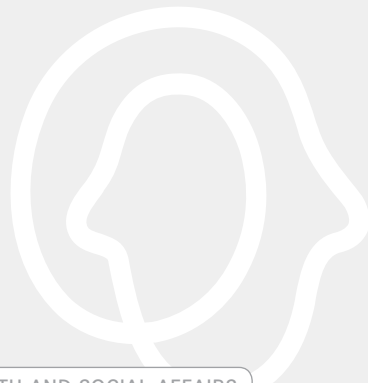
본 연구에서는 인공지능(AI) 기술이 조사데이터 품질 향상과 업무 효율성 제고에 효과적이며, 실무 적용을 위한 구체적인 방안 제시에 유의미한 도움을 줄 수 있음을 확인하였다. 또한 향후 인공지능(AI) 기술이 지속적으로 발전함에 따라 조사데이터 품질관리 방식 또한 더욱 정교해질 것으로 예상된다.

한편 본 연구에서는 향후 다양한 유형의 조사데이터에 인공지능(AI) 모델을 적용할 수 있도록 연구를 확장할 필요가 있음을 제안한다. 또한, 인공지능(AI) 모델의 성능을 지속적으로 개선하고, 보다 정밀한 데이터 정제 기법을 개발하는 것이 조사데이터의 활용도를 극대화하는 중요한 방향이 될 것이다. 인공지능(AI) 기술이 조사 연구 및 통계 분석에서 점차 핵심적인 역할을 담당하고 있는 만큼, 본 연구가 향후 데이터 품질 향상과 연구 신뢰성 강화를 위한 기초 자료로 활용될 수 있기를 기대한다.

사람을
생각하는
사람들



KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS



제 1 장

서론

제1절 사업의 배경 및 목적

제2절 연구의 내용 및 추진 방법

제 1 장 서론

제1절 사업의 배경 및 목적

전 세계적으로 2024년 가장 많이 주목받은 연구 주제 중 하나는 인공지능(AI)이다. 2024년 1월, 전 세계 각국의 정계, 관계, 재계 유력인사와 언론인, 경제학자 등이 세계 경제의 현안과 경제 문제에 대한 각종 해법 등을 함께 논의하는 세계경제포럼(World Economic Forum, WEF)의 2024년 주제는 '신뢰 재구축(Rebuilding Trust)'으로 안보, 경제성장, 인공지능(AI), 기후변화 등이 주요 이슈로 다루어졌다. 특히 인공지능(AI) 관련해서는 생성형 AI(Generative AI)의 영향과 규제 및 윤리 등에 대해 논의되었다. 또한 2024년 10월, 노벨상 발표 이후 전 세계는 다시 한 번 인공지능(AI)을 주목하게 되었다. 노벨 물리학상과 화학상을 전통적인 학자가 아닌 인공지능(AI)을 연구했거나 인공지능(AI)을 이용한 과학자가 수상하였기 때문이다.

우리나라에서도 인공지능(AI)은 단순한 기술적 차원을 넘어 산업과 사회 모든 영역에 걸친 패러다임의 변화를 촉발하고 있으며, 우리 경제의 활력 제고 및 사회문제 해결에 인공지능(AI)이 유력한 방안으로 부상하였다(관계부처 합동, 2019¹⁾). 우리나라는 2019년 12월 'IT 강국을 넘어 AI 강국으로'를 비전으로 정부 역량을 결집하여 AI 시대 미래 비전과 전략을 담은 'AI 국가전략'을 발표하였고(관계부처 합동, 2019), '대한민국 디지털 전략(2022.9)'의 후속으로 인공지능 경쟁력 강화 및 산업 육성을

1) 관계부처 합동. (2019). 인공지능 국가전략.

위해 ‘인공지능 일상화 및 산업 고도화 계획(2023.1)’, ‘대형 인공지능 경쟁력 강화 방안(2023.4)’을 연속적으로 발표·추진하고 있다. 최근에는 전 세계에서 인공지능(AI)을 가장 잘 활용하는 국가로 도약하기 위해 ‘전 국민 AI 일상화 실행계획(2023.9)’을 발표(관계부처 합동, 2023²⁾)했다.

통계 분야에서도 인공지능(AI)을 적극적으로 활용하기 위한 국내외 연구를 진행하고 있다. 2018년 독일에서는 ‘국가 및 국제기구 통계기관의 머신러닝 활용 실태조사’를 실시한 바 있으며, 유엔 유럽경제위원회(UNECE)에서는 2018년 ‘공식 통계에서 머신러닝 활용 연구’, 2019~2020년 ‘머신러닝 파일럿 연구’, 2023년 ‘대형 언어모델(LLM) 활용 연구’ 등 공식 통계 생산을 위한 인공지능(AI) 활용 가능성을 검토하고 실제 활용을 위한 파일럿 연구를 지속적으로 추진하고 있다. 우리나라에서는 주요 외국의 선행 사례 분석을 토대로 인공지능(AI) 활용 가능성을 검토하고 있다. 특히 통계청은 통계의 자료수집, 가공, 배포 단계에서 인공지능(AI)을 활용한 통계업무 자동화(Robotic Process Automation, RPA)를 추진하고 있으며, 통계행정을 지원하는 업무 자동화도 적극 추진하고 있다(이상현, 2021).

한편 우리 연구원은 국가승인통계 3종(한국의료패널조사, 한국복지패널조사, 가족과출산조사)을 비롯하여 각종 연구 수행을 위해 매년 70~110건의 조사데이터를 생산하고 있다(〈표 1-1〉 참조).

〈표 1-1〉 한국보건사회연구원 조사데이터 생산 현황(2016~2022년)

구분	2016년	2017년	2018년	2019년	2020년	2021년	2022년
조사데이터 생산(건)	79	80	76	98	116	98	74

2) 관계부처 합동. (2023). 전국민 AI 일상화 실행계획.

조사데이터는 연구의 기반이 되는 데이터로서, 데이터 오류로 인해 품질에 문제가 생기면 연구 결과의 신뢰성이 낮아지고, 이로 인해 부정확한 정책 결정이나 잘못된 연구 결과가 도출될 위험이 있다. 특히, 패널조사와 같이 대규모 조사데이터를 다루는 경우 데이터 오류가 쉽게 발생할 수 있으며, 이러한 데이터 오류가 수정되지 않으면 연구의 타당성을 저해할 수 있다. 또한 연구 종료 후 조사데이터의 외부 공개(마이크로데이터 배포) 시 이러한 데이터 오류가 발견되면 조사데이터에 대한 신뢰성 저하뿐만 아니라 연구 결과에 대한 신뢰성에 악영향을 초래할 수도 있다.

우리 연구원은 매년 한국의료패널과 한국복지패널, 2개의 패널조사를 실시하고 있으며, 조사데이터 관리부서는 조사데이터의 품질관리를 위해 조사 진행 중 및 조사 완료 후 모든 변수에 대해 입력값의 범위 점검, 변수 간 논리적 관계 점검, 동일 가구 또는 가구원의 과거 자료 비교 점검, 이상치 탐지 점검 등 자료 내검을 수행하고 있다. 특히 한국의료패널의 경우 복수 개의 변수에 대한 여러 조건들을 확인해야 하는 복잡한 점검이 수행되고, 그 기준 수가 63개에 이른다. 그런데 이러한 점검 기준을 적용한 자료 내검 방법은 데이터클리닝 담당자가 각각의 점검 기준에 대해 통계 패키지를 이용한 기술통계 분석을 수행하는 방식이어서, 데이터클리닝 담당자의 자료 내검 업무에 큰 부담이 될 뿐만 아니라 조사데이터의 품질관리에도 어려움이 있다.

이에 본 연구에서는 한국의료패널 조사데이터의 자료 내검 업무를 중심으로 조사데이터 품질 제고 및 업무 효율화를 위한 인공지능(AI) 기술 활용 방안을 모색하고자 한다. 국내외 선행연구 및 사례 분석을 통해 통계 생산 및 관리를 위한 인공지능(AI) 기술 활용 가능성을 검토하고, 한국의료패널 데이터 실증을 통해 우리 연구원 조사데이터 품질 및 업무 효율성 제고를 위한 인공지능(AI) 기술 활용 방안을 제시하고자 한다.

제2절 연구의 내용 및 추진 방법

본 사업은 앞 절에 제시한 연구 목적을 달성하기 위해 총 6개의 장으로 구성하여 전개하였다.

제1장 ‘서론’에서는 인공지능(AI) 기반 조사데이터 품질 제고 방안 연구의 배경과 본 연구의 목적을 기술하였으며, 아울러 이를 위해 본 보고서를 구성하고 있는 연구 내용과 방법을 기술하였다.

제2장 ‘이론적 배경’에서는 문헌고찰을 통해 인공지능(AI)의 개념 및 최신 기술, 데이터 오류 탐지 방법론, 통계업무프로세스모델, 선행연구 검토 결과를 기술하였다.

제3장 ‘인공지능(AI)을 활용한 통계작성 사례 분석’에서는 주요 외국과 우리나라의 통계작성에 인공지능(AI)을 활용한 사례를 조사·분석하였다.

제4장 ‘조사데이터 품질관리 내부 현황 분석’에서는 문헌고찰을 통계 정보, 조사자료, 조사데이터, 통계업무 처리 절차, 통계자료 품질관리, 공공데이터 품질관리 등 관련 개념을 정리하였고, 한국의료패널 담당자와의 인터뷰를 통해 한국의료패널 조사데이터의 자료 내검 절차 및 방법, 그리고 문제점 및 이슈 사항을 도출하였다.

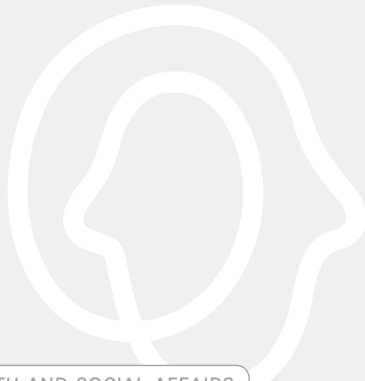
제5장 ‘조사데이터 품질관리 적용 방안’에서는 한국의료패널 데이터 실증을 통해 우리 연구원 조사데이터 품질 및 업무 효율성 제고를 위한 인공지능(AI) 활용 방안을 제시하였다.

제6장 ‘결론 및 제언’에서는 결론과 연구의 한계점, 향후 추진 과제를 제시하였다.

연구 내용 및 추진 방법을 요약하면 [그림 1-1]과 같다.

[그림 1-1] 연구의 내용 및 추진 방법

구분	연구 내용	연구 방법
제1장 서론	연구의 배경 및 목적	문헌고찰 연구진 논의
	연구의 내용 및 방법	
제2장 이론적 배경	인공지능(AI) 개념 및 최신 기술 동향	문헌연구
	데이터 오류 탐지 방법론	
	통계업무프로세스모델	
	선행연구 검토	
	소결	
제3장 인공지능(AI)을 활용한 통계작성 사례 분석	주요 외국의 인공지능(AI) 활용 통계작성 사례 분석	문헌연구 사례조사
	우리나라의 인공지능(AI) 활용 통계작성 사례 분석	
	소결	
제4장 조사데이터 품질관리 내부 현황 분석	조사데이터 품질관리	담당자 인터뷰
	조사데이터 품질관리 절차 및 방법	
	소결	
제5장 조사데이터 품질관리 적용 방안-한국의료패널 데이터 실증	한국의료패널 데이터 특성	데이터 분석
	데이터 실증	
	소결	
6장 결론 및 제언	결론	연구진 논의 전문가 자문회의
	제언	



제2장

이론적 배경

제1절 인공지능(AI)의 개념 및 최신 기술 동향

제2절 데이터 오류 탐지 방법론

제3절 일반통계업무프로세스

제4절 선행연구 검토

제5절 소결

제2장 이론적 배경

제1절 인공지능(AI)의 개념 및 최신 기술 동향

1. 인공지능(AI)의 개념

인공지능(Artificial Intelligence)의 개념은 일원화된 정의가 없으며, 연구자의 관점에 따라 다양하게 정의되었고, 관련 기술의 발전에 따라 진화하고 있다. 인공지능(AI)의 초기 개념은 앨런 튜링(Turing, 1950)³⁾이 ‘계산 기계와 지능(Computing machinery and intelligence)’에서 ‘기계는 생각할 수 있는가?(Can machines think?)’라고 한 질문과 이를 테스트하기 위해 고안한 모방게임(Imitation Game)에서 기원하였다.

한편 ‘인공지능(Artificial Intelligence, AI)’이라는 용어를 세상에 알리게 된 것은 1956년 존 매카시(John McCarthy)가 개최한 다트머스 회의(Dartmouth Dartmouth Conference)로 알려진 ‘인공지능에 관한 다트머스 여름 연구 프로젝트(Dartmouth Summer Research Project on Artificial Intelligence)’이다. 이 회의에서 인공지능(AI)이라는 용어를 처음 사용했으며, ‘기계가 언어를 사용하고, 추상화와 개념을 형성하고, 현재 인간에게만 있는 종류의 문제를 해결하며, 스스로를 개선하는 방법을 찾기 위한 시도’를 의미한다고 정의하였다(McCarthy et al., 1955).⁴⁾

3) A. M. Turing. Computing Machinery and Intelligence. Mind, 59(236), pp. 433-460.

4) J. McCarthy, M. L. Minsky, N. Rochester and C. E. Shannon. (1955). A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTEL

이후 여러 연구자들이 여러 분야에서 다양한 관점으로 인공지능(AI)에 관한 연구를 거듭하면서 인공지능(AI)에 대한 개념도 계속 진화하였다. 스튜어트 러셀 등(2009)⁵⁾은 주요 8개 인공지능(AI) 정의를, [그림 2-1]과 같이 2개 차원을 활용해 4개의 범주로 구분하여 설명하였다.

[그림 2-1] 인공지능에 대한 정의와 범주

인간처럼 생각하는(Thinking Humanly)	합리적으로 생각하는(Thinking Rationally)
- “컴퓨터가 생각하게 만들기 위한 흥미롭고 새로운 노력”... 문자 그대로 “마음을 가진 기계”(Haugeland, 1985) - “의사결정, 문제해결, 학습 등과 같은 인간의 생각이나 활동과 관련된 활동의 자동화”(Bellman, 1978)	- “계산적 모델의 이용을 통한 정식적 능력에 대한 연구”(Charniak and McDermott, 1985) - 인지하고, 추론하고, 행동하는 것이 가능하게 하는 계산들(computations)에 관한 연구(Winston, 1992)
인간처럼 행동하는(Acting Humanly)	합리적으로 행동하는(Acting Rationally)
- “인간에 의해 수행되었을 때 지능이 필요한 기능을 수행하는 기계를 만드는 기술”(Kurzweil, 1990) - “현재 인간이 더 나은 것을 어떻게 하면 컴퓨터가 할 수 있도록 만들지를 연구하는 것”(Rich and Knight, 1991)	- “컴퓨팅 지능(Computational Intelligence)은 지능적 에이전트를 설계하는 것에 대한 연구이다”(Pool et al., 1998) - “AI는... 인공물의 지능적 행동과 관련된 다.”(Nilsson, 1998)

자료: Stuart Russell and Peter Norvig. (2009). Artificial Intelligence: A Modern Approach. p.2.

첫 번째 범주인 ‘인간처럼 생각하는(Thinking Humanly)’이란 인간의 사고 과정을 모방하여 인간의 인지 과정을 이해하고 이를 시스템에 재현하려는 인지 모델링 접근 방식으로, 대표적인 연구 분야로는 문제 해결과 추론(Problem Solving and Reasoning), 기억과 학습(Memory and Learning) 등이 있다.

두 번째 범주인 ‘인간처럼 행동하는(Acting Humanly)’이란 튜링테스

LIGENCE. (http://raysolomonoff.com/dartmouth/boxa/dart564_props.pdf)
5) Stuart Russell and Peter Norvig. (2009). Artificial Intelligence: A Modern Approach. (3rd Edition).

트(turing test)와 같이 시스템이 인간처럼 행동할 수 있는지를 평가하는 접근 방식으로, 대표적인 연구 분야로는 자연어 처리(Natural Language Processing), 지식 표현(Knowledge Representation), 자동화된 추론(Automated Reasoning), 머신러닝(Machine Learning), 컴퓨터 비전(Computer Vision), 로봇공학(Robotics) 등이 있다.

세 번째 범주인 ‘합리적으로 생각하는(Thinking Rationally)’이란 인간 사고와는 무관하게 논리적 추론과 합리적 사고에 초점을 맞춰 이성적인 방식으로 문제를 해결하는 접근 방식으로, 대표적인 연구 분야로는 규칙 기반 시스템(rule-based Systems), 자동 추론(Automated Reasoning), 지식 표현 및 추론(Knowledge Representation and Reasoning), 계획 및 추론(Planning and Reasoning) 등이 있다. 다만 이 접근 방식에는 2가지 문제점이 있는데, 비형식적인 지식을 취합해 논리적 표기법에 필요한 형식으로 표현하기 어렵다는 점과 원칙적으로 문제를 해결하는 것과 실제로 문제를 해결하는 것에는 큰 차이가 있다는 점이다.

네 번째 범주인 ‘합리적으로 행동하는(Acting Rationally)’이란 최상의 결과를 달성하거나, 불확실성이 있는 경우 최상의 예상 결과를 달성하기 위해 행동하는 접근 방식으로, 대표적인 연구 분야는 에이전트 기반 시스템(Agent-Based Systems), 강화 학습(Reinforcement Learning), 계획 및 탐색(Planning and Search), 의사결정 이론(Decision Theory) 등이 있다.

앞에서 설명한 4개 범주에 대한 설명에서 알 수 있듯이 인공지능(AI)의 개념은 연구자들의 관점에 따라 매우 다양하다는 것을 알 수 있다. 초기의 연구자들은 우측 범주인 ‘합리적으로 생각하는(Thinking Rationally)’ 또는 ‘합리적으로 행동하는(Acting Rationally)’ 시스템으로 접근하는 방식이었던 반면, 신경 회로망(Neural Networks)과 딥러닝(Deep

Learning)이 등장한 이후에는 좌측 범주인 ‘인간처럼 생각하는(Thinking Humanly)’ 또는 ‘인간처럼 행동하는(Acting Humanly)’ 시스템으로 접근하는 방식으로 변화하는 추세이다(박승규, 2018).⁶⁾

한편 인공지능(AI)의 개념은 목적에 따라 분류되기도 한다. 1999년 존 설(John Searle)은 인공지능(AI)을 목적 측면에서 특정한 문제의 해결을 위해 필요한 지능을 의미하는 ‘약 인공지능(Weak AI)’과 인간의 지능과 유사하거나 더 나은 지능을 의미하는 ‘강 인공지능(Strong AI)’으로 구분하였다(김명철, 양기철, 2017).⁷⁾ ‘약 인공지능(Weak AI)’은 특정 문제를 해결하기 위한 단일 기능을 지녔으며, 인간이 직접 설계해야 하고 확장이 어렵다는 특징이 있는 반면, ‘강 인공지능(Strong AI)’은 컴퓨터에 인간 수준의 지성을 구현하는 것으로 스스로 학습하고 확장이 용이한 특징이 있다. 다만 ‘약 인공지능(Weak AI)’에서 ‘강 인공지능(Strong AI)’으로 발전하였다는 것을 의미하는 것은 아니다.

6) 박승규. (2018). 인공지능 기술동향. 주간기술동향, (1861).

7) 김명철, 양기철. (2017). 4차 산업혁명과 인공지능. 한국엔터테인먼트산업학회 학술대회 논문집, 58-63.

2. 주요 국가의 인공지능(AI) 법률에서 정의하는 인공지능(AI)

주요 국가에서는 인공지능(AI)에 대한 통합적·포괄적 법률을 제정하고 있으며, 이 법률에서 인공지능(AI)에 대한 개념을 정의하고 있다.

미국에서는 2020년 제정된 「2020 국가 인공지능 이니셔티브법(National Artificial Intelligence Initiative Act of 2020)」의 SEC. 5002. (3)에서 ‘인공지능이란 인간이 정의한 목적의 주어진 세트를 위해 예측을 하고, 추천을 하고, 실제 또는 가상 환경에 영향을 주는 결정을 할 수 있는 기계 기반형 시스템’이라고 정의하였다(세계법제정보센터, 2021).⁸⁾

유럽연합(EU)은 2024년 5월 21일 세계 최초로 인공지능을 포괄적으로 규제하는 법률인 「인공지능법(Artificial Intelligence Act)」을 제정하였다. 해당 법 제3조 제1항(Article 3.(1))에서 ‘인공지능 시스템이란 배포 이후에 적응성을 보이고 명시적 또는 묵시적 목표를 위하여 물리 환경이나 가상 환경에 영향을 미칠 수 있는 예측, 콘텐츠, 권고나 결정 등의 산출물을 생성하는 방법을 뜻하며, 입력을 통하여 추론할 수 있는 다양한 수준의 자율성을 가지고 작동하도록 설계된 기계 기반 시스템’이라고 정의하였다.⁹⁾

중국 정부는 기본법 제정을 통한 종합적·포괄적인 인공지능(AI) 규제 대신, 새로운 인공지능(AI) 현안에 즉각 대응하여 관련 규범을 제정하는 방식을 택하고 있다¹⁰⁾(법무부, 2024.10). 따라서 중국은 인공지능(AI)에

8) 세계법제정보센터. (2023). 국가 인공지능 구상법 2020(National Artificial Intelligence Initiative Act of 2020) 번역본. https://world.moleg.go.kr/web/wli/lgsInfoReadPage.do?CTS_SEQ=50782&AST_SEQ=313에서 2024.05.23. 발췌

9) 세계법제정보센터. (2024). 인공지능법(Artificial Intelligence Act) 완역본. https://world.moleg.go.kr/web/wli/lgsInfoReadPage.do?CTS_SEQ=51409&AST_SEQ=93에서 2024.05.23. 발췌

10) 법무부. (2024. 10). 최신 글로벌 인공지능(AI) 규범 동향 리포트-유럽·미국·중국·일본

대한 법률에서 정하는 일원화된 정의는 없는 것으로 나타났다.

일본은 기존 법률인 개인정보보호법 등을 활용하면서도 비즈니스 분야에서 발생할 신규 법률 분쟁의 경우 중국과 유사하게 인공지능(AI) 현안에 따라 신규 지침을 제작·활용하는 동향을 보이고 있다(법무부, 2024.10). 일본의 경우도 중국과 마찬가지로 인공지능(AI)에 대한 법률에서 정하는 일원화된 정의는 없는 것으로 나타났다.

우리나라는 21대 국회에서 인공지능 관련 산업 육성 및 인력양성 등에 초점을 맞춘 12건의 인공지능 관련 법률안이 제안되었으나 좀더 심도 깊은 논의가 필요하다는 판단에 22대 국회로 법률안 제정을 미루게 되었다. 2025년 1월 ‘인공지능의 건전한 발전을 지원하고 인공지능 사회의 신뢰 기반 조성에 필요한 기본적인 사항을 규정함으로써 국민의 권익과 존엄성을 보호하고 국민의 삶의 질 향상과 국가 경쟁력을 강화하는 데 이바지할 수 있는 대한민국 인공지능의 새로운 기준을 마련’하고자 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」을 제정하였다. 동법 제2조 제1항에 따르면, ‘인공지능이란 ‘학습, 추론, 지각, 판단, 언어의 이해 등 인간이 가진 지적 능력을 전자적 방법으로 구현한 것’으로 정의하였다.

등 주요 국가들의 제정 규범을 중심으로, 해외규제 모니터링 제5호.

3. 인공지능(AI) 관련 최신 기술 동향

인공지능(AI)의 한 분야인 머신러닝(Machine Learning, ML), 딥러닝(Deep Learning, DL)은 의료, 금융, 자율 주행, 자연어 처리를 포함한 광범위한 분야에 혁명을 일으키는 혁신적인 기술로 등장했다(Rane et al. 2024). 이에 머신러닝(ML)과 딥러닝(DL) 기술에 대해 살펴보고자 한다.

가. 머신러닝(ML)

머신러닝(ML)은 컴퓨터가 데이터를 통해 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야를 말하며, 데이터를 이용하여 모델을 훈련시키고, 훈련된 모델을 이용하여 예측을 수행하는 지도 학습(Supervised Learning), 주어진 데이터를 기반으로 분류를 수행하는 비지도 학습(Unsupervised Learning), 그리고 주어진 환경에서 에이전트가 현재의 상태를 인식하여 행동하고 이에 따른 보상으로 최적화 과정을 수행하는 강화 학습(Reinforcement Learning)의 3가지로 분류할 수 있다(조민호, 2021).¹¹⁾

지도 학습은 머신러닝의 초석으로, 각 입력이 출력 레이블과 쌍을 이루는 레이블이 지정된 데이터 세트에 대한 훈련 모델이 포함되며, 일반적인 알고리즘에는 선형 회귀(linear regression), 로지스틱 회귀(logistic regression), SVM(support vector machines) 및 신경망(neural networks)이 포함된다(Rane et al., 2024¹²⁾). 최근 지도형 머신러닝의 발

11) 조민호. (2021). 인공지능의 역사, 분류 그리고 발전 방향에 관한 연구. 한국전자통신학회 논문지, 16(2), 307-312.

12) Rane, N., Choudhary, S., & Rane, J. (2024). Machine Learning and Deep Learning: a Comprehensive Review on Methods, Techniques, Applications, Challenges, and Future Directions. Techniques, Applications, Challenges,

전은 랜덤 포레스트(Random Forest) 및 GBM(Gradient Boosting Machine) 같은 초매개변수 조정(hyperparameter tuning) 및 앙상블 방법(ensemble methods) 같은 기술을 통해 모델 정확성과 효율성을 향상시키는 데 중점을 두고 있으며, 특정 작업에 대해 사전 훈련된 모델을 미세 조정하는 전이 학습(transfer learning)도 특히 이미지 및 언어 처리 분야에서 주목받고 있다(Rane et al., 2024).

〈표 2-2〉 지도 학습 방법 및 응용 사례

방법/기술	설명	응용 사례
선형회귀 (linear regression)	- 종속변수와 하나 이상의 독립변수 간의 관계를 모델링하고 분석하는 통계 방법	- 예측 분석 - 추세 예측 - 재무 모델링
로지스틱 회귀 (logistic regression)	- 하나 이상의 기준을 바탕으로 이진 결과의 확률을 예측하는 데 사용되는 분류 기술	- 의료 진단 - 사기 탐지
의사결정나무 (Decision Trees)	- 예측변수를 만드는 데 사용되는 트리 구조 모델로 특성 값을 기준으로 데이터를 분기하여 결정을 내리고 결과를 예측	- 리스크 관리 - 분류 업무
Support Vector Machines(SVM)	- 데이터 포인트를 다양한 카테고리로 분류하기 위해 최적의 초평면을 찾는 지도 학습 알고리즘	- 이미지 인식 - 텍스트 분류
KNN(K-Nearest Neighbors)	- 새로운 데이터 포인트를 훈련 데이터 세트의 가장 가까운 포인트와 비교하여 분류 및 회귀에 사용되는 간단한 비모수적 알고리즘	- 추천 시스템 - 패턴 인식
Naive Bayes	- 베이즈 정리를 기반으로 예측 변수 간의 독립성을 가정하는 확률적 분류 방법	- 스팸 필터링 - 감성 분석
랜덤 포레스트 (Random Forest)	- 보다 정확하고 안정적인 예측을 위해 여러 의사결정나무를 구성하고 그 결과를 병합하는 앙상블 학습 방법	- 사기 탐지 - 특성 선택
GBM(Gradient Boosting Machine)	- 일반적으로 의사결정나무로 된 약한 모델을 연속적으로 결합하여 예측 정확도를 향상시키는 방법	- 웹 검색 순위 - 추천 시스템

자료: Rane et al. (2024). Machine Learning and Deep Learning. p.8.

and Future Directions (May 31, 2024).

비지도 학습은 자율학습이라고도 하며, 알고리즘은 레이블이 지정된 출력이 없는 데이터 세트의 패턴을 식별하는 것이다(Rane et al., 2024). K-평균(mean), 계층적 클러스터링(hierarchical clustering), DBSCAN 같은 클러스터링 기술은 데이터 구조를 밝히는 데 자주 사용된다. 주성분 분석(Principal Component Analysis) 및 t-SNE(t-Distributed Stochastic Neighbor Embedding)와 같은 차원 축소 방법은 고차원 데이터를 시각화하고 노이즈를 줄이는 데 도움이 되며, 최근 추세에서는 모델이 입력 데이터의 일부를 예측하여 학습하는 자기 지도 학습과 다양한 작업에 유용한 데이터 표현 학습에 초점을 맞춘 표현 학습이 포함된다(Rane et al., 2024).

〈표 2-3〉 비지도 학습 방법 및 응용 사례

방법/기술	설명	응용 사례
K-평균 클러스터링 (K-Means Clustering)	- 데이터를 k개의 클러스터로 분할하고, 각 데이터 포인트가 가장 가까운 평균을 갖는 클러스터에 속하는 클러스터링 기법	- 고객 분할 - 이미지 압축
계층적 클러스터링 (Hierarchical Clustering)	- 기존 클러스터를 병합하거나 분할하여 클러스터 계층을 구축하는 클러스터링 방법	- SNS 분석 - 계층 데이터 분석
주성분 분석 (Principal Component Analysis)	- 데이터를 주성분이라 불리는 상관관계가 없는 변수들의 집합으로 변환하고, 각 변수가 포착하는 분산의 양에 따라 정렬하는 차원 축소 기법	- 데이터 시각화 - 노이즈 감소
독립성분 분석 (Independent Component Analysis)	- 다변량 신호를 추가적인 독립 성분 요소로 분리하는 계산 방법	- 시그널 처리 - 뇌영상
오토인코더 (Autoencoders)	- 차원 축소 또는 특성 학습을 위해 레이블이 지정되지 않은 데이터의 효율적인 코딩을 학습하는 데 사용되는 신경망 유형	- 이상 탐지 - 데이터 노이즈 제거

자료: Rane et al. (2024). Machine Learning and Deep Learning. pp.8-9.

강화 학습은 원하는 행동에 보상을 주고 원하지 않는 행동에 불이익을 주어 일련의 결정을 내리도록 모델을 훈련시키는 것으로 딥러닝을 강화 학습에 통합하여 심층 강화 학습을 구현함으로써 상당한 발전이 이루어졌다. 주목할 만한 알고리즘에는 DQN(Deep Q-Networks), Policy Gradient 방법 및 Actor-Critic 모델이 포함된다. 강화 학습의 적용 범위는 알파고(AlphaGo)와 같은 게임 플레이부터 자율 주행 및 로봇 제어와 같은 실제 작업까지 매우 다양하다. 최근 연구는 샘플 효율성, 안정성 및 강화 학습 모델의 새로운 작업으로의 전환 가능성을 향상시키는 것을 목표로 한다(Rane et al., 2024).

〈표 2-4〉 강화 학습 방법 및 응용 사례

방법/기술	설명	응용 사례
Q-러닝 (Q-Learning)	- 누적 보상을 극대화하는 정책을 사용하여 특정 상태에서의 행동 가치를 학습하는 모델 없는 강화 학습 알고리즘	- 로봇 공학 - 게임 플레이
DQN(Deep Q-Networks)	- Q-러닝과 딥러닝 신경망을 결합하여 고차원 감각 입력을 처리	- 자율주행 차량 - 게이밍
Policy Gradient Methods	- 기대 보상이 증가하는 방향으로 정책을 조정하여 정책을 직접 최적화하는 강화 학습 기술	- 로봇 제어 - 전략 게임

자료: Rane et al. (2024). Machine Learning and Deep Learning. p.9.

한편 최근에 사용되는 대부분의 머신러닝(ML)은 비지도 학습에 대한 것이며, 딥러닝(DL)도 비지도 학습에 속하는 기법이다. 최근에는 강화 학습에 대한 부분도 많이 강조되고 있는데, 그 이유는 비지도 학습에서 학습을 수행하는 과정을 기계 스스로 수행하도록 하기 위한 방안으로 고려되고 있기 때문이다(조민호, 2021).

나. 딥러닝(DL)

머신러닝(ML) 기법 중 하나인 딥러닝(DL)은 머신러닝(ML)보다 전문화된 분야이며, 여러 계층의 신경망을 활용하여 복잡한 데이터 패턴을 모델링한다. 인간의 뇌 구조에서 시냅스(Synapse)의 중첩을 흉내 낸 인공신경망(Artificial Neural Network, ANN) 알고리즘에 기반한 방법론으로 딥러닝(DL)의 기본 구조들은 1980년대에 정립되었으나, 알고리즘적으로 과적합(overfitting), 기울기 소멸(vanishing gradient) 등의 문제점이 존재하였고, 성능을 높이기 위해서는 충분한 수의 은닉계층(hidden layer)이 필요하고 그만큼 학습에 필요한 데이터와 계산량이 증가한다는 문제점을 갖고 있었다(최예림, 김관호, 2016).¹³⁾ 하지만 2000년대 중반부터 드롭아웃(Dropout), ReLU(Rectifier Neural Unit), 배치 정규화(Batch Normalization) 같은 새로운 기법들의 소개와 데이터 수집 및 연산 능력의 발달로 그 성능이 확인되고 있다(최예림, 김관호, 2016).¹⁴⁾

딥러닝 구조로는 입력 계층(Input Layer)과 출력 계층(Output Layer) 사이에 복수의 은닉 계층(Hidden Layer)이 존재하는 심층 신경망(Deep Neural Network, DNN), 은닉계층 앞에 요인 추출에 필요한 필터를 두고 필터를 함께 학습하는 합성곱 신경망(Convolutional Neural Network, CNN), 각 시간의 인공신경망을 적층해 시계열 데이터 처리가 가능한 재귀 신경망(Recurrent Neural Network, RNN) 등이 있다(최예림, 김관호, 2016).

심층 신경망(DNN)은 이미지 인식, 자연어 처리, 음성 인식 등의 분야에서 탁월한 성과를 거두었다. 하지만 최근에는 합성곱 신경망(CNN), 순

13) 최예림, 김관호. (2016). 인공지능 개요 및 적용 사례. ie 매거진, 23(2), 23-29.

14) 최예림, 김관호. (2016). 인공지능 개요 및 적용 사례. ie 매거진, 23(2), 23-29.

환 신경망(RNN)이 기계의 기능을 크게 발전시켜 이전에는 인간 지능에만 국한되었던 작업을 획기적으로 개선했다.

합성곱 신경망(CNN)은 의료 영상에서 인간 전문가와 동등한 수준 또는 그 이상의 이상 식별, 금융 분야에서는 사기 탐지 시스템 개선, 자율 분야에서는 자율 주행 자동차, 드론, 로봇 공학을 위한 고급 알고리즘 개발로 자율 기술 기능 확장 등에 활용되고 있으며, 순환 신경망(RNN)은 언어 이해 및 생성을 혁신하여 실시간 번역 및 정교한 챗봇과 같은 애플리케이션에 활용되고 있다(Rane et al., 2024).

〈표 2-5〉 합성곱 신경망(CNN)과 순환 신경망(RNN) 설명 및 응용 사례

방법/기술	설명	응용 사례
합성곱 신경망(CNN)	- 격자 형태의 데이터 구조(예: 이미지)를 처리하는 데 주로 사용되는 딥러닝 모델의 한 종류로, 특징의 공간적 계층 구조를 학습	- 이미지 및 비디오 인식 - 의료 이미지 분석
순환 신경망(RNN)	- 시간적 순서를 따라 노드 간의 연결이 방향 그래프를 형성하는 시퀀스 데이터를 위해 설계된 신경망의 한 종류	- 시계열 분석 - 언어모델링 - 음성 인식

자료: Rane et al. (2024). Machine Learning and Deep Learning. p.21.

다. 머신러닝(ML)과 딥러닝(DL)의 최신 기술 동향

머신러닝(ML)과 딥러닝(DL) 분야는 빅데이터의 가용성 증가, 컴퓨팅 성능의 발전, 보다 효율적이고 지능적인 알고리즘에 대한 수요에 의해 빠르게 진화하고 있다. 최근 주목받고 있는 머신러닝(ML)과 딥러닝(DL) 기술 및 관련 분야를 살펴보면 다음과 같다.

1) 엣지 AI(Edge AI)와 연합학습(Federated Learning)

엣지 AI와 연합학습은 최근 부상하고 있는 머신러닝(ML) 기술이다.

엣지 AI는 중앙 집중화된 클라우드 서버가 아닌 장치에서 로컬로 데이터를 처리하는 것으로, 이러한 추세는 실시간 처리, 지연 시간 단축, 향상된 개인정보 보호 및 사물인터넷(Internet of Things, IoT) 장치의 확산에 대한 수요 증가에 의해 주도되고 있다(Rane et al., 2024).

연합 학습은 데이터를 로컬화하면서 여러 분산 장치에서 모델을 학습할 수 있도록 하여 이를 보완하는데, 이 접근 방식은 데이터 개인정보 보호 및 보안을 개선하고 여러 소스의 데이터를 활용하여 보다 강력한 모델을 구축할 수 있다(Rane et al., 2024).

2) 설명 가능한 AI(Explainable AI, XAI)

머신러닝(ML) 및 딥러닝(DL) 모델이 더욱 복잡해짐에 따라 투명성과 설명 가능성에 대한 필요성이 증가하고 있으며, 설명 가능한 AI(XAI)는 이러한 모델의 의사 결정 프로세스를 인간이 이해할 수 있도록 하는 것을 목표로 한다(Rane et al., 2024). 이러한 추세는 특히 의료, 금융, 자율주행 같은 중요한 애플리케이션에서 인공지능(AI) 시스템에 대한 신뢰를 얻는 데 중요한데, SHAP(SHapley Additive exPlanations) 및 LIME(Local Interpretable Model-agnostic Explanations) 같은 기술이 인기를 얻고 있으며, 이를 통해 이해 관계자가 모델 예측을 해석하고 기본 메커니즘을 이해할 수 있다(Rane et al., 2024).

3) 전이 학습(Transfer Learning) 및 사전 학습 모델(Pre-trained Models)

대규모 데이터 세트에서 사전 학습된 모델을 활용하고 특정 작업에 맞게 미세 조정하는 것이 포함되며, 이러한 접근 방식은 광범위한 계산 리소스와 학습 시간의 필요성을 크게 단축할 수 있다(Rane et al., 2024). 주목할 만한 예로는 BERT, GPT-3, CLIP 같은 모델이 있으며, 이러한 모델은 방대한 양의 데이터에서 사전 학습되었으며 자연어 처리에서 컴퓨터 비전에 이르기까지 다양한 애플리케이션에 적용할 수 있다(Rane et al., 2024).

4) 강화 학습(Reinforcement Learning, RL)

강화 학습(RL)은 게임, 로봇, 자율 시스템 등에 적용되고 있으며, 최근에는 복잡한 환경에서 인상적인 역량을 보여준 Proximal Policy Optimization(PPO) 및 Deep Q-Networks(DQN) 같은 개선된 알고리즘에 의해 촉진되었고 강화 학습(RL)을 비지도 학습 및 모방 학습 같은 다른 학습 패러다임과 통합함으로써 잠재적인 적용 분야가 확장되고 있다(Rane et al., 2024).

5) 윤리적 AI(Ethical AI) 및 편견 완화(Bias Mitigation)

인공지능(AI) 머신러닝(ML)의 윤리적 의미에 대한 관심이 높아지면서, 편견, 공정성, 책임성에 대한 우려가 편견 탐지 및 완화 기술에 대한 연구를 주도하고 있다. 조직은 모델이 기존 편견을 영속화하거나 악화시키지

않도록 보장하기 위해 윤리적 AI 프레임워크를 점점 더 많이 채택하고 있으며, 이러한 추세는 편향된 모델이 사회적으로 상당한 영향을 미칠 수 있는 고용, 대출 및 법 집행 같은 분야에서 특히 중요하다(Rane et al., 2024).

6) 자동화된 머신러닝(Automated Machine Learning, AutoML)

자동화된 머신러닝(AutoML) 플랫폼은 머신러닝(ML)을 실제 문제에 적용하는 엔드 투 엔드(end-to-end) 프로세스를 자동화하는 것을 목표로 하고 있으며, 여기에는 데이터 전처리, 기능 엔지니어링, 모델 선택 및 하이퍼파라미터 튜닝이 포함된다. 자동화된 머신러닝(AutoML)은 이러한 작업을 자동화함으로써 비전문가가 머신러닝(ML) 모델을 효율적으로 구축할 수 있도록 하고 전문가는 모델 개발의 더 복잡한 측면에 집중할 수 있도록 한다. Google AutoML, H2O.ai, DataRobot 같은 플랫폼이 이 분야를 선도하고 있다(Rane et al., 2024).

7) 신경 구조 탐색(Neural Architecture Search, NAS)

신경망 구조 탐색(NAS)은 신경망 구조 설계를 자동화하는 데 중점을 둔 새로운 분야로, 신경망 구조 탐색(NAS) 알고리즘은 네트워크 구조를 수동으로 설계하는 대신 특정 작업에 맞는 최적의 아키텍처를 검색하는 것이다. 이 접근 방식은 수동으로 설계된 모델보다 성능이 뛰어난 새로운 아키텍처를 발견하는 데 도움이 되었다. EfficientNet 및 DARTS (Differentiable Architecture Search)와 같은 기술은 모델 성능과 효율성에서 상당한 개선을 보였습니다(Rane et al., 2024).

8) 양자 머신러닝(Quantum Machine Learning, QML)

양자 머신러닝(QML)은 양자 역학의 원리를 활용하여 고전적 알고리즘보다 특정 문제를 더 빨리 해결할 수 있는 알고리즘을 개발하는 것으로 아직 초기 단계지만 최적화, 데이터 분류 및 생성 모델에서 잠재력을 보여주었고, 양자 하드웨어가 계속 발전함에 따라 복잡한 머신러닝(ML) 문제에 접근하는 방식을 혁신할 수 있을 것으로 기대된다(Rane et al., 2024).

9) 의료용 AI(AI for Healthcare)

머신러닝(ML)과 딥러닝(DL)이 적용은 빠르게 확대되고 있는 의료 분야에서는 진단 및 개인화된 의학에서 약물 발견 및 환자 치료에 이르기까지 인공지능(AI)은 의료전달체계를 혁신하고 있다. 딥러닝(DL)은 의료 이미지를 분석하고, 환자의 결과를 예측하고, 잠재적인 치료법을 식별하는데 사용되고 있으며, 머신러닝(ML)과 웨어러블 기기를 통합하면 지속적인 건강 모니터링과 건강 문제의 조기 진단이 가능해져 환자의 결과가 개선된다(Rane et al., 2024).

10) 지속 가능한 AI

인공지능(AI) 모델, 특히 딥러닝(DL)이 환경에 미치는 영향이 분명해짐에 따라 지속 가능한 AI에 대한 관심이 증가하고 있는데, 이러한 추세에서는 인공지능(AI) 애플리케이션의 탄소 발자국(carbon footprint)을 줄이는 에너지 효율적인 알고리즘과 아키텍처를 개발하는 것이 필요하다. 모델 가지치기(model pruning), 양자화(quantization), 효율적인

신경망(efficient neural network) 같은 기술이 인공지능(AI)을 더욱 지속 가능하게 만들기 위해 탐구되고 있으며, 대규모 모델을 훈련하기 위해 재생 에너지를 사용하는 방향으로 나아가고 있다.

11) 멀티모달 러닝(Multimodal Learning)

여러 모달리티(텍스트, 이미지, 오디오 등)의 정보를 통합하는 멀티모달 러닝이 인기를 얻고 있으며, 이러한 접근 방식을 통해 모델은 보다 포괄적이고 섬세한 데이터 표현을 학습할 수 있다. 예를 들어 OpenAI의 CLIP 같은 모델은 시각과 언어를 결합하여 이미지 캡션 및 시각적 질의 응답과 같은 작업을 수행하는데, 멀티모달 러닝은 콘텐츠 생성에서 인간-컴퓨터의 상호 작용에 이르기까지 다양한 응용 분야에서 인공지능(AI) 시스템의 역량을 향상시키고 있다.

12) 금융 분야의 인공지능(AI)

금융 부문에서 머신러닝(ML)과 딥러닝(DL)은 사기 탐지, 알고리즘 거래, 위험 관리, 고객 서비스를 포함한 광범위한 애플리케이션에 활용되고 있다. 방대한 양의 재무 데이터를 분석하고 패턴을 감지하는 능력은 금융 기관이 더 나은 정보에 입각한 결정을 내리고 개인화된 서비스를 제공하는 데 도움이 되고 있고, 규제 준수(RegTech)에 인공지능(AI)을 사용하면 회사가 복잡한 규제 환경을 보다 효율적으로 탐색하는 데 큰 도움이 된다.

13) 자연어 처리(Natural Language Processin, NLP)

자연어 처리(NLP)는 트랜스포머 아키텍처에 기반한 모델로 OpenAI의 GPT-4 및 입출력을 모두 텍스트로 취하는 T5(Text-to-Text Transfer Transformer)와 같은 모델은 기계 번역, 감정 분석 및 대화형 AI와 같은 작업에 적용되고 있다. 자연어 처리(NLP)를 다른 인공지능(AI) 기술과 통합하면 음성 활성화 어시스턴트 및 자동화된 콘텐츠 생성과 같은 보다 정교한 애플리케이션도 가능해진다.

14) 사이버 보안(Cybersecurity)을 위한 인공지능(AI)

사이버 공격의 위협이 증가함에 따라 인공지능(AI)과 머신러닝(ML)은 사이버 보안을 강화하는 데 중요한 역할을 하고 있으며, 머신러닝(ML) 알고리즘은 이상을 탐지하고 잠재적 위협을 예측하며 실시간으로 보안 사고에 대응하는 데 사용되고 있다. 특히 딥러닝(DL) 모델은 사이버 위협을 나타내는 복잡한 패턴을 식별하는 데 효과적인 것으로 입증되었으며, 인공지능(AI) 기반 사이버 보안 솔루션은 민감한 데이터를 보호하고 디지털 시스템의 무결성을 보장하는 데 필수적이 되고 있다.

15) 개인화 및 추천 시스템

머신러닝(ML)과 딥러닝(DL)으로 구동되는 개인화 및 추천 시스템은 전자상거래부터 스트리밍 서비스에 이르기까지 온라인 플랫폼에서 널리 사용되고 있으며, 이러한 시스템은 사용자 행동과 선호도를 분석하여 개인화된 콘텐츠와 추천을 제공한다. 특히 협업 필터링 및 시퀀스 모델링

분야에서 심층 학습의 발전은 추천의 정확성과 관련성을 향상시키고 있으며, 개인화된 경험은 경쟁 시장에서 기업을 위한 핵심 차별화 요소가 되고 있다.

16) 대형 언어모델(Large Language Model, LLM)

대형 언어모델(LLM)은 딥러닝의 한 분야로, 방대한 양의 텍스트 데이터를 학습하여 인간과 유사한 수준의 자연어 처리(NLP) 기능을 수행하도록 설계된 모델을 의미한다(Bhunia & Jamieson, 2025). 이러한 모델은 다층 신경망을 기반으로 하며, 문맥을 이해하고 자연스러운 텍스트를 생성할 수 있는 것이 주요 특징이다.

대형 언어모델은 자기회귀 기법(Self-attention mechanism)을 활용한 Transformer 모델(Vaswani et al., 2017)을 기반으로 한다. Transformer는 문장 내 단어 간 관계를 효율적으로 분석할 수 있도록 설계되었으며, 기존 순환신경망(Recurrent Neural Network, RNN) 모델들이 겪었던 기울기 소멸(Vanishing gradient) 문제를 해결한 것이 핵심적인 차별점이다.

이러한 모델들은 단순한 텍스트 예측을 넘어 다양한 자연어 처리 작업에 적용된다. 대표적인 예로는 사전에 학습된 모델이 추가적인 훈련 없이도 새로운 작업을 수행하는 제로샷(Zero-shot) 방식과, 소량의 예제를 제공한 후 학습하여 문제 해결 능력을 향상시키는 퓨샷(Few-shot) 방식이 있다(Cyrt, P., 2024, Brown et al., 2020). 이러한 접근법을 통해, 대형 언어모델은 별도의 파인튜닝 없이도 높은 성능을 발휘할 수 있다.

제2절 데이터 오류 탐지 방법론

1. 규칙 기반 오류 확인 방법

하나의 변수가 연속형과 범주형의 유형일 때, 데이터의 오류를 검증하는 방법은 재확인, 논리적 오류 등의 유형이 있다. 두 개 이상의 변수가 조합되어 복합적으로 구성될 경우, 연속형-연속형 변수 간의 결합분포에 대한 경험누적분포(empirical cumulative distribution)인 분위수(quantile)와 연속형-범주형 변수의 경우에는 범주형 변수의 수준(level)별로 연속형 변수의 분위수를 조사한다. 문자열변수의 경우는 정규 표현식 등 입력 규칙 준수 여부를 확인할 수 있다.

한국의료패널 데이터에서 몇 가지 유형의 예를 살펴보면, [연속형-단독] 유형은 연속형 변수에 관한 방법으로, 히스토그램과 커널 확률밀도함수(Kernel Density Function, KDE)와 러그 플롯(Rug plot)으로 분포를 확인할 수 있다.

2. 연관규칙 활용 방안

연관규칙 분석(association rule analysis)은 자료에 존재하는 항목(item)들 간의 if-then 형식의 연관규칙을 찾는 방법으로서, 타겟(target) 정보가 없거나 지정되지 않은 경우에 흔히 적용할 수 있는 비지도 학습(unsupervised learning) 방법의 일종이다. 연관규칙 분석은 데이터 집합 내에서 항목 간의 연관성을 찾아내기 위한 데이터 마이닝 기법의 하나로, 장바구니 분석(market basket analysis)을 위하여 다양한 상품거래 데이터베이스 간의 관계를 찾아 숨겨진 구매 패턴의 일반규칙을

찾는 것을 목표로 고안되었다(Zhan et al., 2019). 예를 들어, 어느 슈퍼마켓에서 고객이 구입한 물건들이 담긴 장바구니의 정보가 주어졌다고 가정한다. 연관규칙 분석은 특정한 상품을 구입한 고객이 어떤 부류에 속하는지, 그들이 왜 그런 구매를 했는지 알아보기 위해서 고객들이 구매한 상품에 대한 자료를 분석하는 것이다. 이런 것들을 분석하면 효율적인 매장 진열, 패키지 상품의 개발, 교차판매 전략의 수립, 기획 상품의 결정 등에 응용할 수 있다. 구체적인 연관규칙 분석을 적용한 예로는 인터넷 쇼핑몰에서 상품의 추천, 텍스트마이닝에서 웹페이지 간의 링크에 대한 분석 등 여러 가지 문제에 적용된다(박창이 등, 2013).

데이터의 오류를 탐색하기 위한 방안으로 연관규칙을 활용하는 방안을 고려해본다. 예를 들어 변수들을 A, B, C라고 놓고, 각 변수가 가지는 값을 수준 (X1, X2), (Y1, Y2), (Z1, Z2, Z3)를 가진다고 하자. 그러면 X1, Y2, Z3에서 변수 A는 X1, B는 Y2를 가지고, C가 Z3값을 가질 때, 이 문장에 논리적 오류 여부를 포함한 판별을 수행하는 방안으로 연관규칙 분석을 적용할 수 있다. 연관규칙 분석은 지지도, 신뢰도, 향상도 등의 지표를 사용하여 평가되며, 데이터로부터 중요한 정보를 추출하는 데 활용할 수 있다.

〈표 2-6〉과 같은 5개의 데이터들의 변수들이 가지는 범주형 수준을 통해 연관규칙의 지표를 간략히 이해해본다. 이 예제의 경우 X1, Y1, Z1 등은 각 변수가 가지는 수준의 역할을 한다. 뒤의 숫자는 각 변수가 가지는 수준의 수를 나타낸다. 〈표 2-6〉의 3개의 변수별 수준 자료를 바탕으로 동시출현 빈도표를 작성할 수 있다. 모든 변수들을 범주화로 변환하는 방법을 고려할 필요가 있다. 61개 변수들을 평균 10구간으로 범주화한다. 동시출현 빈도표로부터 “변수 A가 수준 X1이면, 변수 B는 수준 Y1을 가진다”와 같은 간단한 규칙을 만들 수 있다. “If X1, then Y1”와 같은

형식으로 표현되는 규칙을 연관규칙이라 하며 모든 규칙이 유용하지는 않다.

〈표 2-6〉 예시 데이터와 대한 동시출현 빈도표

	3개의 변수(A, B, C)별 수준		X1	X2	Y1	Y2	Z1	Z2	Z3
		X1	3	0	1	2	1	1	1
		X2		3	0	1	1	0	1
1	X1, Y2, Z3	Y1			1	1	0	0	1
2	X2, Y1, Z3	Y2				2	0	0	1
3	X1, Y1, Z2	Z1					2	0	0
4	X1, Y2, Z1	Z2						1	0
5	X2, Y2, Z1	Z3							1

연관규칙이 유용한 규칙일 필요조건에 대한 측도인 지지도, 신뢰도, 향상도에 대하여 학습한다. 먼저 연관규칙 $X \rightarrow Y$ (즉 “if X, then Y”)의 지지도와 신뢰도에 대하여 예제를 통해서 알아보자. 〈표 2-6〉의 예제에서 연관규칙 “변수 X가 수준 A1이면, 변수 Y는 수준 B1을 가진다”의 지지도와 신뢰도는 동시출현 빈도표로부터 2/5와 2/4임을 알 수 있다. 지지도와 신뢰도 외에 향상도를 고려하는데, 향상도가 클수록 수준 X의 포함 여부가 수준 Y의 포함 여부에 큰 영향을 미친다. 〈표 2-7〉은 지지도, 신뢰도, 향상도에 대한 정의로 $\hat{P}$ 는 집합에 대한 확률 P 의 추정확률을 나타내며, $\hat{P}(X|Y)$ 는 집합 Y 가 주어졌을 때, 교집합 $X \cap Y$ 의 상대 비율로 추정 조건부 확률(conditional probability)을 나타낸다.

데이터 오류 탐지와 관련하여, 연관규칙 분석을 활용한 다양한 연구들이 수행되었다. Liu et al.(2011)은 연관규칙 마이닝에서 발생할 수 있는 실제로는 의미가 없는 규칙이지만 “유의미하다”고 결론 내릴 수 있는 거

짓 양성(false positives)을 제어하는 방법을 연구하였다. Thakar et al.(2021)은 범주형 데이터에 대해 Louvain 군집 분석과 연관규칙 분석을 결합한 형태로 구성된 파이프라인을 제안하고 이상치를 탐지하는 데 활용하였다. Jeong et al.(2015)은 결제 로그 분석에서 연관규칙 분석을 적용하여 이상 거래를 탐지하였다.

〈표 2-7〉 연관규칙 $X \rightarrow Y$ 에 대한 지지도, 신뢰도, 향상도

평가 지표	세부 내용
지지도 (support)	전체 데이터 중 수준 X와 수준 Y를 동시에 포함하는 데이터의 비율로 정의된다. $\hat{P}(X \cap Y) = \frac{\text{수준 } X \text{와 수준 } Y \text{를 가지는 데이터 수}}{\text{전체 데이터 수}}$
신뢰도 (confidence)	수준 X를 포함하는 데이터들 중 수준 Y를 포함하는 데이터의 비율로 다음과 같이 정의되는데 신뢰도는 조건부확률 $\hat{P}(Y X)$ 를 의미한다. $\hat{P}(Y X) = \frac{\hat{P}(X \cap Y)}{\hat{P}(X)}$ $= \frac{\text{수준 } X \text{와 수준 } Y \text{를 동시에 가지는 데이터 수}}{\text{수준 } X \text{를 가지는 데이터 수}}$
향상도 (lift)	수준 Y의 확률 대비 수준 X가 주어졌을 때의 수준 Y의 확률의 증가 비율이다. 이 값이 클수록 수준 X의 포함 여부가 항목 Y의 포함 여부에 큰 영향을 미친다. 향상도가 1보다 크면 수준 집합 간의 양의 상관관계가 있음을 나타낸다. $\frac{\hat{P}(X \cap Y)}{\hat{P}(X)\hat{P}(Y)} = \frac{\hat{P}(Y X)}{\hat{P}(Y)} = \frac{\text{신뢰도}}{\hat{P}(Y)}$ $= \frac{\text{수준 } X \text{와 } Y \text{를 포함하는 데이터 수}}{X \text{를 포함하는 데이터 수} \times Y \text{를 포함하는 데이터 수}}$

종합하면 연관규칙은 빈발한 항목들에 대한 규칙을 도출하는 방법으로 데이터 내에서의 오류를 검증하는 방식보다는 데이터 간의 상호관계를 통해 이상 패턴과 비정상 패턴을 탐지하는 데 더 적합하다고 볼 수 있다. 이는 한국의료패널 데이터에서도 적용이 가능한데, 예를 들어 환자들이

특정 질병에 걸렸을 때 활용되는 통상적인 치료법에 대한 규칙이 빈번하게 나타나고, 그것에 상응하는 치료비용의 범위를 벗어나는 비용은 높은 확률로 이상치로 간주될 수 있을 것이다. 특히, 연관규칙은 데이터의 정상적인 구조를 정의하고 데이터 간 관계 정의에 초점을 맞추기 때문에 연관규칙을 활용한 방식은 기존 방식에서 발견하기 어려웠던 잠재적인 이상 패턴이나 비정상적인 상관관계를 탐지하는 데 용이하게 활용될 수 있을 것으로 사료된다.

3. 머신러닝(ML), 딥러닝(DL) 방법

데이터셋의 오류를 탐지하는 다양한 방법 중 이상치를 검출할 수 있는 방법인 isolation Forest 방법과 대형 언어모델(LLM)을 활용하는 방법에 대해서 살펴본다.

가. isolation Forest

isolation Forest(iForest)는 비지도 학습 기반의 이상 탐지 알고리즘으로 데이터를 격리(isolation)시키는 과정을 통해 이상치를 탐지한다. isolation Forest는 이상치가 학습 시 정상 데이터로부터 빠르게 분리할 수 있다는 가정에 기반한다(Liu et al., 2008). 즉, 이상치 데이터는 정상 데이터보다 다른 데이터 포인트로부터 더 빨리 격리된다.

iForest는 랜덤 포레스트(random forest)와 유사하게, 여러 개의 결정 트리(decision trees)를 사용하여 작동한다. 하지만 각 트리는 데이터 포인트를 분류하는 것이 아니라, 데이터 포인트를 분리하는 것에 초점을 맞춘다. 트리는 데이터를 반복적으로 분할하고, 트리 깊이를 기반으로 이

상치를 식별한다. 정상 데이터는 복잡한 구조를 이루므로, 많은 분할을 거쳐야 트리에서 완전히 분리된다. 이상치는 다른 데이터와 거리가 멀거나 매우 드물기 때문에, 적은 분할만으로도 쉽게 격리된다. iForest는 각 데이터 포인트가 격리되는 데 필요한 분할 횟수를 기록하고, 그 값에 따라 이상치 여부를 판단한다.

iForest의 목적식의 이상치 점수는 격리 거리(isolation depth)를 기반으로 계산되며, 주요한 절차는 다음과 같다.

(1) 트리에서 데이터 분할: iForest는 랜덤하게 분할할 변수와 분할 지점을 선택하여 데이터를 반복적으로 나눔으로써 각 데이터는 트리 내에서 특정 깊이(depth)에서 격리된다.

(2) 격리 깊이(isolation depth): 데이터 x 가 트리에서 격리되는 데 필요한 깊이 $h(x)$ 는 이상치 여부를 판단하는 중요한 기준으로, 트리에서 하나의 데이터 x 를 분리하기 위해 필요한 평균 깊이 $E(h(x))$ 를 근사하는 통계량 $c(n) = 2[H(n) - \frac{n-1}{n}]$ 와 대비하여 판단의 근거로 활용한다. $c(n)$ 은 트리에서 데이터가 완전히 분리될 때까지 평균적으로 필요한 단계 수를 나타낸다고 할 수 있다. 여기서 n 을 데이터셋의 크기라 할 때, $H(n) = \ln(n) + 0.577215$ 이다.

(3) 이상치 점수(anomaly score): iForest는 각 데이터가 트리에서 격리된 깊이를 기준으로 이상치 점수를 계산하는데, 이상치 점수 $s(x, n) = 2^{-h(x)/c(n)}$ 로 정의된다. 점수가 1에 가까울수록 해당 포인트는 이상치일 가능성이 높고, 점수가 0에 가까울수록 정상 데이터일 가능성이 높다고 할 수 있다.

가상의 2차원의 데이터에 대해 iForest의 특징을 살펴보면 <표 2-8>과 같다. 대부분의 자료는 (10, 10) 근처에 몰려 있는 데 반해, 소수의 데이터는 (50, 50)이나 (-50, -50)과 같이 멀리 떨어져 있는 이상치를 포함

하고 있다. (10, 10) 근처에 정상 데이터가 몰려 있고, (50, 50), (-50, -50)과 같은 이상치를 추가하였다. 그리고 Isolation Forest 모델에서 contamination=0.1 옵션을 사용하여 이상치의 비율을 설정하고, 모델을 학습시켰다. 그러면 clf.predict(X) 함수를 사용해 각 데이터가 이상치인지 정상 데이터인지 예측한다.

〈표 2-8〉 iForest의 예제

가상데이터에 대한 iForest 프로그램

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import IsolationForest
# 2D 정상데이터 생성
np.random.seed(42)
X = 0.3 * np.random.randn(100, 2) + 10

# 이상치 추가
X = np.r_[X, np.array([[50, 50], [-50, -50], [30, -30]])]
# Isolation Forest 모델 학습
clf = IsolationForest(contamination=0.1)
clf.fit(X)

# 각 데이터에 대한 이상치 점수 예측
scores = clf.decision_function(X)
outliers = clf.predict(X)

# 시각화
plt.scatter(X[:, 0], X[:, 1], c=outliers, cmap='coolwarm', edgecolor='k', s=60)
plt.title("Isolation Forest Outlier Detection")
plt.show()
```

iForest의 장점은 비지도 학습으로 본 연구의 데이터와 같이 레이블이 없는 데이터에서도 이상치 탐지가 가능하다는 점이다. 고차원 데이터와 대용량 데이터에서도 효과적으로 적용이 가능하다고 할 수 있다. 트리 기반 모델이므로 계산 속도가 빠르며, 이상치를 감지하는 데 필요한 분할 횟수가 적어 효율성이 높다고 할 수 있다. 반면, iForest의 단점으로는 고

밀도 데이터에서는 효율성이 떨어질 수 있다. 이상치와 정상 데이터 간의 명확한 구분이 없거나 고밀도 지역에서 이상치가 발생하면, 정확도가 낮아지는 것으로 알려져 있다. 또한, 랜덤 분할을 사용하기 때문에 특정 데이터 분포에 의존하게 됨으로써 최적의 결과를 얻지 못할 수 있다.

나. AnoGAN

딥러닝 방법으로 데이터셋에서 이상 패턴이나 오류를 탐지하는 데 자주 사용되는 대표적인 방법으로는 autoencoder를 들 수 있다. 특히, auencoder는 선형모형인 경우에는 주성분 분석과 동일하다(박창이 외 4인, 2013). 이 방법은 입력 데이터를 압축하고 다시 복원하는 모델로, 복원 시 계량할 수 있는 재구성 오차(reconstruction error)를 최소화해 학습한다. 이상 탐지 문제에서는 정상 패턴은 복원이 잘 되지만 일반적으로 오류나 이상 패턴은 재구성 오차가 큰 일반적인 특징을 활용하여 이상 탐지에 활용한다.

한편, 적대적 학습으로 신경망을 효율적으로 학습할 수 있는 GAN (generative adversarial networks, Goodfellow et al., 2020)방법은 정상 데이터의 분포 학습을 통해 재현데이터를 생성하고 이들을 분류 문제로 다룸으로써, 관측한 데이터의 밀도함수를 추론할 수 있다. GAN에서는 fully connected layer를 CNN으로 확장한 DCGAN(Bian et al., 2019) 같은 이상값을 판별하는 연구도 많이 진행되었다. GAN에 근거한 대표적인 방법인 AnoGAN(Schlegl et al., 2017)을 간략히 살펴본다. 주로 이미지 데이터에 국한한 연구로, 한국의료패널 데이터와 같은 테이블 구조 데이터와 다소 다른 자료 구조이기는 하지만, 이상치의 라벨링 없이 이상을 탐지하는 점에서 향후 활용할 수 있을 것으로 판단된다.

AnoGAN은 이미지의 픽셀 수준에서 이상 탐지가 가능하며, 기존의 지도 학습 기반 모델과 비교하여 사전 라벨링에 의존적이지 않으며, 수동 주석(annotation)이나 라벨링 과정의 부담을 줄일 수 있다. GAN에 기초하기 때문에, 관측한 데이터에 대한 분포를 학습하는 과정은 동일하다. 새로운 이미지나 생성된 이미지가 정상 데이터 분포 내에 있는지를 계량화하는 스코어를 기반으로 이상 여부를 탐지한다.

다. 대형 언어모델(LLM)을 활용한 데이터 시각화

최근 ChatGPT 및 GPT-4 같은 대형 언어모델(LLM)이 자연어 이해에서 뛰어난 성능을 보여줌에 따라, 다양한 NLP 분야의 문제에서 활발히 활용되고 있다. 이러한 대형 언어모델(LLM)을 텍스트에서 시각화를 위한 코드를 생성하는 데(text-to-vis) 활용하는 방법도 개발되었는데, Luo et al.(2021)은 text-to-SQL 신경망을 개발하고 이와 함께 인코더-디코더 프레임워크와 신경망을 기반으로 Seq2Vis 모델을 제안하였다. Song et al.은 대규모 코드 데이터베이스에서 코드를 재활용하는 방식을 적용하여 RGVisNet이라는 프레임워크로 가장 관련성 있는 시각화를 위한 프로그램 코드 후보들을 추출하고, 이를 수정하여 사용자의 의도에 부합하는 시각화 코드를 생성했다.

자료 분석에 좀 더 직접적으로 관련 있는 대형 언어모델(LLM)을 활용한 연구로는, 데이터 시각화를 위한 python 코드를 생성하기 위하여 프롬프트를 사용한 ‘Chat2Vis’ 방법(Maddigan and Susnjak, 2023)을 들 수 있다. 대형 언어모델(LLM)에 질의하기 위해 질의문을 가공하는 일련의 방법을 프롬프트 엔지니어링(prompt engineering)이라고 일컫는데, 많은 방식 중 가장 널리 활용되는 문맥 내 학습(In-context learn-

ing, ICL)을 활용하여 시각화를 위한 코드 생성을 수행할 수 있다. 문맥 내 학습(ICL)은 대형 언어모델(LLM)을 별도의 추가 학습 없이 (learning-free) 활용하는 방식으로, 여러 입력과 출력 예제를 프롬프트로 주입하여 모델이 패턴을 학습하도록 한다.

Narayan et al.(2022)은 대형 언어모델(LLM)을 활용하여 표 형식 데이터의 속성-값(attribute-value) 쌍에서 오류를 감지하고 수정하는 방법을 제안하였다. 특히, ICL이 데이터 정제(data cleaning) 및 데이터 통합(data integration) 작업에서 얼마나 효과적인지를 조사하였으며, 그 결과를 실증적으로 제시하였다. 또한, 오류 탐지뿐만 아니라 엔티티 매칭(entity matching), 데이터 변환(data transformation) 같은 다양한 태스크에서도 ICL을 활용한 성능 개선 가능성을 실험적으로 입증하였다. Dong et al.(2023)은 프로그램으로 데이터베이스에서 가장 저변적으로 활용되는 SQL 프로그램을 생성하는 text-to-SQL 방법을 제안하였다. ICL에서 예제를 프롬프트로 구성하지 않는 제로샷(zero-shot) 기반으로 프롬프트를 구성한 후, 출력된 결과를 보정하는 방법에 대해 연구하였다. Li et al.(2024b)은 ICL 학습능력을 강화할 수 있는 데 필요한 예제들을 선별하는 방법과 답변 생성에서 참조할 데이터베이스의 스키마(scheme)를 필터링하는 방법(Prompt4Vis)을 제안하였다.

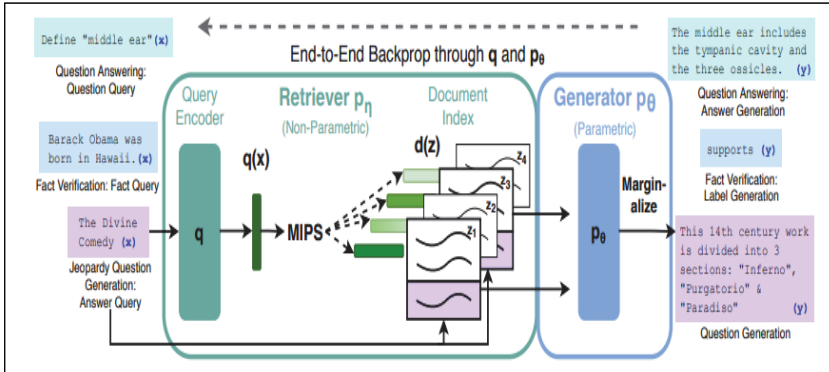
라. 대형 언어모델(LLM)을 활용한 데이터 오류 탐지

대형 언어모델(LLM)을 활용한 데이터의 오류를 탐지하는 데 필요한 기저 기술인 검색증강기술과 대형 언어모델(LLM)을 활용하여 구체적 태스크를 수행하기 위한 절차인 랑체인(langchain)에 대해 간략히 살펴본다.

1) 검색증강생성기술

검색증강생성기술(Retrieval-augmented generation, RAG)은 대형 언어모델(LLM)의 생성 능력과 정보 검색 방법을 결합하여 생성된 응답의 사실성 및 관련성을 개선하는 기술을 일컫는다. 검색증강생성기술(RAG) 아키텍처에서는, 검색 모듈이 먼저 외부 지식 기반 또는 데이터베이스에서 관련 정보를 검색하고, 이를 생성 모델에 추가적인 맥락으로 제공하여 사용한다(Lewis et al., 2020). [그림 2-2]와 같이 외부 지식을 검색하여 언어모델의 응답을 보강하는 기술이다. 기존 언어모델은 주어진 훈련 데이터에 의존하여 응답을 생성하기 때문에, 최신 정보나 특정 도메인 지식이 부족한 경우가 많았으나, 검색증강생성기술(RAG)은 검색된 문서와 언어모델을 결합하여 지식 집약적인 NLP 작업의 성능을 향상시키는 방법이다. 사전 학습된 대형 언어모델(LLM)은 새로운 정보를 업데이트하는데 많은 컴퓨팅 자원과 시간이 필요하므로 검색증강생성기술(RAG)을 통해 외부 지식을 유연하게 반영함으로써 사전 학습된 데이터에 대한 의존성과 환각 현상을 해결할 수 있다. RAG는 파라미터화된 생성 모델과 비파라미터화된 검색기(retriever)를 결합하여 외부 지식을 효과적으로 활용한다. 이를 통해 모델은 최신 정보나 특정 도메인 지식을 반영하여 보다 정확하고 풍부한 응답을 생성할 수 있다.

[그림 2-2] 검색증강생성기술(RAG) 모델 아키텍처



자료: Lewis et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, pp. 9459-9474.

검색증강생성기술(RAG)을 활용하여, 대형 언어모델(LLM)과 검색 모듈을 결합해 데이터셋 내의 잠재적 오류를 탐지하고 수정하는 접근법도 제안되고 있다. 이는 대형 언어모델(LLM)이 생성한 텍스트와 검색된 관련 정보 간의 불일치를 기반으로 오류를 탐지하는 방법이다. 지식 그래프(knowledge graph)와 대형 언어모델(LLM)을 결합하여 데이터셋의 오류를 감지하는 연구도 활발하다. 지식 그래프는 정형화된 지식 표현을 기반으로 논리적 모순이나 관계의 오류를 감지하며, 대형 언어모델(LLM)은 비정형 데이터에서의 오류를 탐지할 수 있다. 이러한 방법들은 각각 특정한 유형의 데이터 오류를 탐지하는 데 적합하며, 최신 연구에서는 이를 결합하여 종합적인 데이터셋 오류 탐지 및 수정 방법을 개발하는 추세이다.

검색증강생성기술(RAG)은 크게 두 가지 주요 요소로 구성된다. 첫 번째는 정보 검색으로 외부 지식베이스, 문서, 데이터베이스 또는 기타 저장소에서 관련 정보를 검색엔진이나 신경망 기반 임베딩 방법(DPR, FAISS)으로 검색하는 단계이다. 검색된 데이터는 대형 언어모델(LLM)의 입력으로 사용되며, 대형 언어모델(LLM)은 검색된 정보를 바탕으로 사용

자 질의에 대해 문맥에 맞는 응답을 생성한다. 이 과정에서 검색된 데이터와 사전 학습된 대형 언어모델(LLM)의 내재적 지식을 결합할 수 있다. 외부 지식베이스를 활용하여 모델이 미리 학습하지 않은 정보에 접근할 수 있으며, 대형 언어모델(LLM)의 한계(지식 갱신의 어려움)를 극복하여 확장적 사용성을 증대할 수 있다.

사전에 학습된 모델의 지식에만 의존하지 않고 최신 정보를 검색하여 신뢰성을 높일 수 있다. 검색 모듈 단계에서는 질의와 문서의 벡터를 생성하여 유사성을 비교하는 과정이 필요하며, dense passage retrieval(Karpukhin et al., 2020, DP)과 같은 신경망 기반 알고리즘과 텍스트 기반 검색을 수행할 수 있는 BM25(Amati, 2009)이나 빠르고 효율적인 임베딩 검색을 위한 오픈소스 라이브러리인 FAISS(Douze et al., 2024)을 활용할 수 있다. 생성 모듈(generator)에서는 GPT 계열, T5, BERT 등의 사전 학습된 언어모델을 사용하여 검색된 정보와 질의를 결합하여 응답을 생성한다. 마지막으로 지식베이스는 참조할 문서나 정보를 저장하는 데이터베이스로 위키피디아, 논문 데이터베이스, 기업 문서 등 다양한 외부 정보로 구성된다.

검색증강생성기술(RAG)의 목적 식은 두 가지 주요 구성 요소인 정보 검색과 응답 생성(generation)의 성능을 최적화하는 데 중점을 둔다고 할 수 있다. 검색된 정보를 응답 생성에 통합하는 과정을 반영하기 위해, 검색 단계의 확률 $P(d | q)$ 와 생성 단계의 확률 $P(a | q, d)$ 를 곱하여 최종 응답의 조건부 확률 $P(a | q)$ 를 계산한다. RAG는 검색 결과에 의존하는 생성 모델이기 때문에 검색 단계의 오류가 최소화되고, 생성 단계에서 검색된 정보를 잘 활용할수록 성능이 향상될 수 있다.

q 를 사용자 질의(query), a 를 생성된 응답(answer), d 를 검색된 문서 또는 정보(document), D 를 문서의 전체 집합이라고 놓자. $P(d | q)$ 를

검색 단계에서 문서 d 가 질의 q 와 관련될 확률, $P(a \mid q, d)$ 는 주어진 질의 q 와 문서 d 를 바탕으로 응답 a 를 생성할 확률이라고 하자. RAG 모델의 전체 목적은 다음의 조건부 확률

$$P(a \mid q) = \sum_{d \in D} P(a \mid q, d)P(d \mid q)$$

를 최대화함으로써, 질의와 가장 유사한 문서를 반환하여 정확한 정보를 검색하는 것을 목표로 한다.

다음은 정보 검색 단계로, $f(q)$, $f(d)$ 를 질의와 문서를 임베딩 공간으로 매핑하는 함수라고 놓고, $\text{sim}(f(q), f(d))$ 를 질의와 문서 간의 유사도를 측정하는 함수로 코사인 유사도와 내적이라고 할 때,

$$P(d \mid q) \propto \text{sim}(f(q), f(d))$$

를 최대화하는 d 를 추출하는 데 활용된다. T 를 응답의 토큰 총수, a_t 를 생성된 응답의 t 번째 토큰, $a_{<t}$ 를 생성된 이전 토큰 시퀀스라고 할 때, 응답 생성 단계는 확률

$$P(a \mid q, d) = \prod_{t=1}^T P(a_t \mid q, d, a_{<t})$$

를 최적화하는 과정으로, 검색된 문서와 질의를 기반으로 높은 확률의 응답 토큰 시퀀스를 생성한다. 이 태스크에서는 교차 엔트로피 손실을 사용하여 모델을 학습한다.

$$L = - \sum_{t=1}^T \log P(a_t \mid q, d, a_{<t})$$

검색 모듈은 RAG의 핵심 구성 요소로, 입력된 사용자 질의에 대해 관련성 있는 외부 문서를 검색하여 생성 모듈(generator)에 제공하는 역할을 한다. 검색 모듈은 크게 dense retrieval과 sparse retrieval로 나뉜다. 두 방식은 정보 검색의 접근 방식, 효율성, 성능 측면에서 서로 다른

특징을 가지며, 각각의 장단점에 따라 활용 가능성이 달라진다. Dense retrieval은 신경망 기반의 임베딩(embedding) 방법을 사용하여 질의와 문서를 벡터 공간에 매핑한다. 이 방식은 BERT와 RoBERTa 같은 사전 학습된 언어모델을 활용하여, 질의와 문서의 복잡한 문맥적 관계를 파악한다. Dense retrieval은 문맥적 의미를 이해하고, 복잡한 질의에서도 높은 정확도를 제공할 수 있는 강점이 있다. 하지만 고용량의 메모리 사용과 높은 연산 자원이 필요하다는 단점이 있다. 특히, 대규모 데이터베이스에서의 검색 성능을 유지하기 위해 효율적인 벡터 검색 알고리즘이 필수적으로 요구된다.

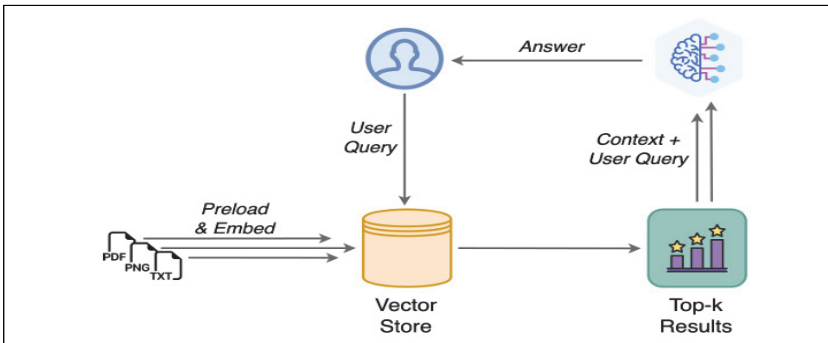
2) 랭체인

LangChain(Chase, 2023)은 언어모델을 다양한 데이터 소스와 도구와 통합하여 복잡한 작업을 수행할 수 있게 해주는 프레임워크이다. 기존의 언어모델들은 주로 특정 태스크에만 특화되었으나, 랭체인은 다양한 API와 외부 시스템과의 통합을 지원하여 더 복잡한 작업을 처리할 수 있도록 한다. 이를 통해 개발자들이 챗봇, 정보 검색, 문서 요약 등 다양한 응용 프로그램을 구축할 수 있으며, 언어모델의 활용 범위를 확장할 수 있어서 애플리케이션을 개발하기 위한 유용한 프레임워크(framework)이다. 외부 데이터와의 상호작용, 시스템 간 통합, 그리고 복잡한 워크플로우 구현을 지원함으로써 데이터 검색, 웹 API 호출, 파일 처리 등 다양한 기능을 포함한 애플리케이션을 손쉽게 구축할 수 있다. 예를 들어, 사용자는 랭체인을 활용해 데이터베이스에서 관련 정보를 검색하고, 이를 언어모델을 통해 정교하게 처리하여 사용자가 원하는 결과를 제공할 수 있다. 랭체인의 모듈화된 구조는 개발자에게 프레임워크의 특정 구성 요

소만 선택적으로 사용하거나 필요에 따라 새로운 기능을 추가할 수 있게 해줌으로써 맞춤형 애플리케이션 개발을 가능하게 한다.

랭체인(chain)의 구성 요소를 개략적으로 살펴본다. 먼저, 프롬프트 템플릿(prompt template)은 언어모델에 입력될 텍스트를 구조화하여 명확하고 일관된 응답을 생성하도록 돕는다. 이를 통해 사용자는 특정 작업에 적합한 입력 형식을 쉽게 정의할 수 있다. 예를 들어, 질문-응답 시스템에서 프롬프트는 질문을 명확히 전달하고 응답 형식을 제어할 수 있다. 다음으로, 체인(chain)은 여러 단계를 순차적으로 실행하여 복잡한 워크플로를 처리하는 구조이다. 체인을 사용하면 데이터 전처리, 언어모델 호출, 후처리 작업을 통합하여 하나의 연속적인 작업 흐름을 만들 수 있다. 예를 들어, 데이터베이스에서 관련 정보를 검색한 후, 이를 기반으로 언어모델이 답변을 생성하는 프로세스를 설계할 수 있다.

[그림 2-3] LangChain 파이프라인 구조도



자료: Mavroudis. (2024). LangChain, <hal-04817573>.

이외에, 메모리(memory)와 벡터 스토어(vector store)는 다양한 환경에서 다양한 옵션을 제공할 수 있는데, 메모리는 애플리케이션 실행 중의 상태 정보를 저장하고 관리하여 챗봇이 이전의 상호작용을 기억하고, 문맥을 유지하며, 응답을 생성할 수 있다. 예를 들어, 고객 지원 챗봇은 사용자의 이전 문의를 기억하여 더 나은 서비스를 제공할 수 있다. 벡터 스토어는 텍스트를 임베딩하여 효율적인 검색을 지원하는 저장소로, 텍스트 데이터를 벡터화하여 저장하며, 사용자의 질의에 따라 가장 관련성 높은 결과를 반환할 수 있다. 검색자는 사용자의 질의를 기반으로 벡터 스토어에서 데이터를 검색하는 역할을 하며, 임베딩된 텍스트 데이터와 질의 간의 유사도를 계산하여 가장 적합한 결과를 제공한다. [그림 2-3]과 같이, 검색증강생성기술(RAG)을 활용해 외부 지식을 결합한 응답을 생성하는 고도화된 애플리케이션을 구축할 수 있다(Mavroudis, 2024).

전술한 구성 요소들은 독립적으로 사용하거나 조합하여 활용할 수 있어, 개발자가 다양한 요구사항에 적합한 솔루션을 설계할 수 있다. 또한, 각 구성 요소는 모듈화되어 있어 필요에 따라 확장하거나 수정할 수 있다. 정리하면, 랭체인은 단순한 언어모델 활용을 넘어 복잡하고 정교한 애플리케이션 개발을 지원하는 데 최적화된 프레임워크라 할 수 있다.

제3절 통계업무프로세스모델

일반통계업무프로세스모델(Generic Statistical Business Process Model, GSBPM)은 2008년 뉴질랜드 통계청이 사용하는 업무프로세스 모델(Business Process Model)을 기반으로 UNECE/Eurostat/OECD 공동 통계메타데이터(METIS)에 의해 처음 개발되었으며, 지속적인 수정을 거쳐 2018년 일반통계업무프로세스모델(GSBPM) 5.1을 발표하였다.

일반통계업무프로세스모델(GSBPM)은 국가의 공식 통계를 생산하는데 필요한 일련의 업무 프로세스를 설명하며, 정의한 모델로 통계기관들이 통계 생산 과정을 현대화하고, 방법과 구성 요소를 공유하는 데 도움이 되는 표준 프레임워크와 용어를 제공하고 있다(UN Economic and Social Council, 2019).¹⁵⁾

일반통계업무프로세스모델(GSBPM)은 ‘요구사항 명세(Specify Needs)’, ‘설계(Design)’, ‘구축(Build)’, ‘수집(Collect)’, ‘처리(Process)’, ‘분석(Analyse)’, ‘배포(Disseminate)’, ‘평가(Evaluate)’의 총 8단계로 구성되어 있으며, 각 단계에는 [그림 2-4]와 같이 하위 프로세스가 존재한다.

15) UN Economic and Social Council. (2019). Generic Statistical Business Process Model.

[그림 2-4] 일반통계업무프로세스모델(GSBPM)

Overarching Processes							
Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1. Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs Process	7.5 Manage user support Analyse	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

자료: United Nations Economic Commission for Europe. (2019). Generic Statistical Business Process Model. p.10.

‘요구사항 명세(Specify Needs)’ 단계는 새로운 통계에 대한 필요성을 식별하고 이해관계자의 요구사항을 상세히 파악하는 단계로 ‘1.1. 요구사항 식별(Identify needs)’, ‘1.2. 요구사항 상담 및 확인(Consult and confirm needs)’, ‘1.3. 산출물 목표 설정(Establish output objectives)’, ‘1.4. 개념 식별(Identify concepts)’, ‘1.5. 데이터 가용성 검토(Check data availability)’, ‘1.6. 업무 사례 준비 및 제출(Prepare business case)’의 6개의 하위 프로세스로 구성되어 있다(〈표 2-9〉 참조).

〈표 2-9〉 일반통계업무프로세스(GSBPM)-‘요구사항 명세’ 단계의 하위 프로세스

구분	설명
1.1. 요구사항 식별	- 새로운 통계 정보에 대한 필요성 조사 - 새로운 정보요청이나 환경 변화(예: 예산 감소) - 다른 국내외 통계 기관의 유사한 데이터 생산 방식
1.2. 요구사항 상담 및 확인	- 이해관계자와 상담하여 통계에 대한 상세한 요구사항 확인
1.3. 산출물 목표 설정	- 1.2.(요구사항 상담 및 확인)에서 파악된 사용자 요구사항을 충족시키기 위한 통계 산출물의 목표 설정 - 제안된 산출물과 품질측정의 적합성에 대해 사용자와 합의 - 법적 프레임워크(예: 기밀유지 관련 등)와 가용 자원을 고려하여 산출물의 적합성과 품질 측정 기준
1.4. 개념 식별	- 사용자 관점에서 측정해야 할 필수 개념 명확화
1.5. 데이터 가용성 검토	- 현재 데이터 소스가 사용자 요구사항을 충족 - 데이터 사용 제한 사항을 포함하여 데이터 이용 가능 조건 검토
1.6. 업무 사례 준비 및 제출	- 통계 생산을 위한 승인을 얻기 위해 업무 사례 준비 및 제출

자료: United Nations Economic Commission for Europe. (2019). Generic Statistical Business Process Model. p.11-12.(요약)

‘설계(Design)’ 단계는 통계 산출물을 설계하는 단계로 ‘2.1. 산출물 설계(Design outputs)’, ‘2.2. 변수 설명 설계(Design variable descriptions)’, ‘2.3. 수집 방법 설계(Design collection)’, ‘2.4. 모집단 및 표본 설계(Design frame and sample)’, ‘2.5. 처리 및 분석 설계(Design processing and analysis)’, ‘2.6. 통계 생산과 워크플로우 설계(Design production systems and workflow)’의 6개의 하위 프로세스로 구성되어 있다(〈표 2-10〉 참조).

〈표 2-10〉 일반통계업무프로세스모델(GSBPM)-‘설계’ 단계의 하위 프로세스

구분	설명
2.1. 산출물 설계	- 사용자 요구사항을 충족시키는 통계 산출물의 상세 설계
2.2. 변수 설명 설계	- 수집할 데이터와 통계 변수에 대한 상세한 정의와 설명
2.3. 수집 방법 설계	- 수집할 데이터와 통계 변수에 대한 상세한 정의와 설명 - 기존의 통계 표준과의 일치 여부 확인
2.4. 모집단 및 표본 설계	- 표본 추출에 사용될 모집단을 설계하거나 기존 모집단 수정 - 표본 크기, 할당, 선택 방법 등 표본 설계
2.5. 처리 및 분석 설계	- 수집된 데이터의 처리, 편집, 코딩, 검증 절차 설계 - 통계적 분석 방법과 절차 정의
2.6. 통계 생산과 워크플로우 설계	- 전체 통계 생산 과정의 워크플로우를 설계 - 필요한 기술 인프라 및 도구 정의

자료: United Nations Economic Commission for Europe. (2019). Generic Statistical Business Process Model. p.13-15.(요약)

‘구축(Build)’ 단계는 통계 생산 시스템을 실행 가능한 형태로 구현하고, 모든 구성 요소가 작동하는지 확인하는 단계로 ‘3.1. 수집 도구 구축 또는 재사용(Reuse or Build Collect instrument)’, ‘3.2. 프로세스 구성 요소 구축 또는 재사용(Reuse or build processing and analysis components)’, ‘3.3. 배포 구성 요소 구축 또는 개선(Reuse or build dissemination components)’, ‘3.4. 워크플로우 구성(Configure workflows)’, ‘3.5. 통계생산시스템 테스트(Test production system)’, ‘3.6. 통계업무 프로세스 테스트(Test statistical business process)’, ‘3.7. 통계생산시스템 확정(Finalise production systems)’ 등 7개의 하위 프로세스로 구성되어 있다(표 2-11 참조).

〈표 2-11〉 일반통계업무프로세스모델(GSBPM)-‘구축’ 단계의 하위 프로세스

구분	설명
3.1. 수집 도구 재사용 또는 구축	- ‘2. 설계 단계’에서 정의된 사양에 따라 데이터 수집 도구를 개발
3.2. 프로세스 구성 요소 재사용 또는 구축	- 데이터 처리, 편집, 변환, 추정 등을 위한 서비스와 절차 개발 - 기존 도구의 재사용 가능성 검토 및 필요시 새로운 도구 개발
3.3. 배포 구성 요소 재사용 또는 구축	- 종이 출판, 웹 서비스, 오픈 데이터, 마이크로데이터 접근 등 다양한 배포 방식을 구축 - 사용자 요구사항에 부합하는 맞춤형 배포 솔루션 개발
3.4. 워크플로우 구성	- 데이터 수집부터 배포까지 전체 통계 생산 과정 워크플로우 구성 - 시스템 간 데이터 흐름과 처리 순서 정의 및 설정
3.5. 통계생산시스템 테스트	- 구축된 서비스, 워크플로우, 시스템의 기술적 테스트 - 새로운 프로그램과 기존 루틴의 적합성 검증
3.6. 통계업무 프로세스 테스트	- 전체 통계 생산 과정을 소규모로 시범 운영 - 수집 도구, 데이터 처리, 분석 과정이 예상대로 작동하는지 확인
3.7. 통계생산시스템 구축 완료	- 테스트를 거친 시스템을 실제 운영 환경 적용 - 시스템 문서화, 사용자 교육, 생산환경으로의 이전 등 수행

자료: United Nations Economic Commission for Europe. (2019). Generic Statistical Business Process Model. p.15-17.(요약)

‘수집(Collect)’ 단계는 ‘4.1. 모집단 및 표본 생성(Create frame and select sample)’, ‘4.2. 수집 준비(Set up collection)’, ‘4.3. 수집 실행(Run collection)’, ‘4.4. 수집 마무리(Finalise collection)’의 4개의 하위 프로세스로 구성되어 있다(〈표 2-12〉 참조).

〈표 2-12〉 일반통계업무프로세스모델(GSBPM)-‘수집’ 단계의 하위 프로세스

구분	설명
4.1. 모집단 및 표본 생성	- 수집에 사용할 모집단 프레임을 생성하고 유지 관리 - 설계 단계에서 결정된 방법론에 따라 표본 선정 - 필요한 경우 하위 표본 추출
4.2. 수집 준비	- 수집 직원 교육 - 머신러닝(지도 학습) 기술을 사용한 교육시스템 - 수집 리소스(노트북, 수집 앱, API 등) 사용 가능 확인 - 데이터를 요청하고 수신하기 위한 수집 시스템 구축 - 수집할 데이터 보안 보장 - 수집 도구 준비(설문조사 목적을 설명하는 편지 또는 브로셔 등) - 자료 번역(해당 국가에서 다른 언어를 사용하는 경우)
4.3. 수집 실행	- 다양한 방법(예: 개인 면접, 전화 조사, 온라인 설문 등)을 통해 데이터 수집 - 응답자의 문의에 대응하고, 필요시 추가 정보 제공 - 수집 진행 상황을 모니터링하고 관리
4.4. 수집 마무리	- 수집된 데이터의 안전성 확인 - 수집 과정에서 사용된 자료 정리 - 수집 담당자들로부터 피드백을 수집하여 향후 개선에 활용 - 수집 과정의 품질을 평가

자료: United Nations Economic Commission for Europe. (2019). Generic Statistical Business Process Model. p.17-19.(요약)

‘처리(Process)’ 단계는 수집된 원시 데이터를 분석 가능한 형태로 변환하는 과정으로 데이터의 품질, 일관성, 정확성을 보장하기 위한 다양한 기술과 방법이 적용되며, ‘5.1. 데이터 통합(Integrate data)’, ‘5.2.분류 및 코딩(Classify and code)’, ‘5.3. 데이터 검토 및 유효성 검사(Review and validate)’, ‘5.4. 편집 및 대체(Edit and impute)’, ‘5.5. 신규 변수 및 단위 도출(Derive new variables and units)’, ‘5.6. 가중치 계산 (Calculate weights)’, ‘5.7. 집계 계산(Calculate aggregates)’, ‘5.8. 최종 데이터 파일 완성(Finalize data files)’ 등 8개의 하위 프로세스로 구성되어 있다(〈표 2-13〉 참조).

〈표 2-13〉 일반통계업무프로세스모델(GSBPM)-‘처리’ 단계의 하위 프로세스

구분	설명
5.1. 데이터 통합	- 여러 출처에서 수집된 데이터를 하나의 데이터 세트로 통합 - 다양한 수집 모드로 얻은 데이터를 결합 - 필요시 데이터 연결(record linkage) 기술 적용
5.2. 분류 및 코딩	- 수집된 데이터를 적절한 분류 체계에 따라 분류하고 코딩
5.3. 데이터 검토 및 유효성 검사	- 데이터의 잠재적 문제, 오류, 불일치 확인 - 사전에 정의된 유효성 규칙에 따라 데이터 검증 - 이상치나 오류가 있는 데이터를 식별하고 처리
5.4. 편집 및 대체	- 누락되거나 신뢰할 수 없는 데이터 처리 - 통계적 방법을 사용하여 누락된 값 대체 - 데이터 편집 규칙을 적용하여 일관성 유지
5.5. 신규 변수 및 단위 도출	- 기존 변수를 기반으로 신규 변수를 계산하거나 도출 - 필요시 새로운 통계 단위 생성
5.6. 가중치 계산	- 표본 설계와 응답률을 고려하여 가중치 계산 - 필요시 보정 계수 적용
5.7. 집계 계산	- 데이터 요약 및 집계 측정값을 계산 - 필요한 통계량(예: 평균, 중앙값, 지수 등) 산출
5.8. 최종 데이터 파일 완성	- 분석 준비가 완료된 최종 데이터 파일 생성 - 데이터 파일의 일관성과 완전성 확인 - 필요시 데이터 파일 승인 절차

자료: United Nations Economic Commission for Europe. (2019). Generic Statistical Business Process Model. p.19-22.(요약)

‘분석(Analyse)’ 단계는 수집되고 처리된 데이터를 심층적으로 검토하고 해석하는 단계로, ‘6.1. 산출물 준비(Prepare draft outputs)’, ‘6.2. 산출물 검증(Validate outputs)’, ‘6.3. 산출물 해석 및 설명 (Interpret and explain outputs)’, ‘6.4. 공개 통제 적용 (Apply disclosure control)’의 4개의 하위 프로세스로 구성되어 있다(〈표 2-14〉 참조).

〈표 2-14〉 일반통계업무프로세스모델(GSBPM)-‘분석’ 단계의 하위 프로세스

구분	설명
6.1. 산출물 준비	- 수집 및 처리된 데이터를 바탕으로 초기 통계 결과물 생성 - 표, 그래프, 지표 등 다양한 형태의 산출물 준비
6.2. 산출물 검증	- 생성된 산출물의 정확성과 일관성 검증 - 이전 주기의 결과나 다른 관련 통계와 비교하여 결과의 타당성 확인
6.3. 산출물 해석 및 설명	- 통계 결과의 의미를 심층 해석 - 결과에 대한 설명을 작성하고, 주요 사항 강조
6.4. 공개 통제 적용	- 개인정보 보호 필요시 공개 통제 방법 적용 - 민감한 정보가 노출되지 않도록 데이터를 익명화

자료: United Nations Economic Commission for Europe. (2019). Generic Statistical Business Process Model. p.22-23.(요약)

‘배포(Disseminate)’ 단계는 통계 정보를 최종 사용자에게 효과적으로 전달하여 생산된 통계의 가치가 실현되고, 정책 결정, 연구, 비즈니스 등 다양한 분야에서 활용될 수 있도록 하는 단계로 ‘7.1. 산출물 시스템 업데이트(Update output systems)’, ‘7.2. 배포 결과물 생산(Produce dissemination products)’, ‘7.3. 배포 결과물 공표 관리(Manage release of dissemination products)’, ‘7.4. 배포 결과물 홍보(Promote dissemination products)’, ‘7.5. 사용자 지원 관리(Manage user support)’의 5개의 하위 프로세스로 구성되어 있다(〈표 2-15〉 참조).

〈표 2-15〉 일반통계업무프로세스모델(GSBPM)-‘배포’ 단계의 하위 프로세스

구분	설명
7.1. 산출물 시스템 업데이트	- DB와 온라인 데이터 저장소를 신규 데이터로 업데이트 - 웹사이트, 통계 포털 등 배포 플랫폼을 준비
7.2. 배포 결과물 작성	- 보고서, 보도자료, 통계표, 인포그래픽 등 다양한 형태의 배포 관련 산출물 자료 작성 - 사용자의 요구에 맞춘 맞춤형 데이터 제품 생산
7.3. 배포 결과물 공표 관리	- 배포 일정을 관리하고 자료의 품질을 확인 - 공개 전 최종 검토 및 승인 절차
7.4. 배포 결과물 홍보	- 통계 생산물의 가치와 유용성을 알리는 마케팅 활동을 수행 - 소셜 미디어, 뉴스레터 등 다양한 채널을 통해 생산물을 홍보
7.5. 사용자 지원 관리	- 사용자 문의에 대응하고 필요한 지원을 제공

자료: United Nations Economic Commission for Europe. (2019). Generic Statistical Business Process Model. p.24-25.(요약)

‘평가(Evaluate)’ 단계는 통계 생산 과정의 지속적인 개선을 위한 중요한 피드백 메커니즘을 제공하며, 이를 통해 향후 통계 생산의 품질과 효율성을 높이기 위한 단계로 ‘8.1. 평가 자료 수집(Gather evaluation inputs)’, ‘8.2. 평가 수행(Conduct evaluation)’, ‘8.3. 개선 계획 수립(Agree an action plan)’의 3개의 하위 프로세스로 구성되어 있다(〈표 2-16〉 참조).

〈표 2-16〉 일반통계업무프로세스모델(GSBPM)-‘평가’ 단계의 하위 프로세스

구분	설명
8.1. 평가 자료 수집	- 사용자의 의견, 프로세스 메타데이터, 시스템 보고서 등을 포함한 통계 생산 과정 전반에 걸친 품질 지표와 피드백 수집
8.2. 평가 수행	- 수집된 자료를 바탕으로 통계 생산 과정을 종합적 평가 - 효율성, 비용 효과성, 통계 품질 등 다양한 측면을 검토
8.3. 개선 계획 수립	- 평가 결과를 바탕으로 개선이 필요한 영역 식별 - 향후 통계 생산 과정 개선을 위한 구체적인 행동 계획을 수립

자료: United Nations Economic Commission for Europe. (2019). Generic Statistical Business Process Model. p.25-26.(요약)

한편 국제 표준인 일반통계업무프로세스모델(GSBPM)을 우리나라 환경에 맞게 재정의한 한국형 통계업무 표준 프로세스 모델인 통계업무프로세스모델(Korea Statistical Business Process Model, KSBPM)은 2010년 10월에 통계업무프로세스모델(KSBPM) 1.0, 2012년 9월에는 일반통계업무프로세스모델(GSBPM) V4.0에 기반하여 기획-설계-구축-수집-처리-분석-배포-보관-평가 9단계로 구성된 통계업무프로세스모델(KSBPM) 2.0을 도출하였다(이의규, 2018).¹⁶⁾

16) 이의규. (2018). 표준 자료처리 모델 GSDEMs의 소개 및 시사점. 2018년 상반기 연구 보고서 제1권. 통계개발원.

[그림 2-5] 한국형 통계업무프로세스모델(KSBPM)

기획	설계	구축	수집	처리	분석	배포	보관	평가
수요 파악	신용률 설계	수집 도구 구현	자료 수집대상 설정	자료통합	신용률 작성	공표자료 점검 및 직제	자료보관규칙 정의	평가 계획 수립
수요 확정	항목 설정	차리시스템 개발, 개선	자료 수집 준비	분류 및 코딩	신용률 검증	공표 신용률 작성	자료 보존 및 관리	수행 및 보고서 작성
신용률표 수립	수집 방법 설계	배포시스템 개발, 개선	자료 수집 진행	자료검토 및 보완	신용률 해석 및 설명 작성	자료 배포 관리		개선과제도출 및 실행계획수립
통계적 개념 정립	모집단 및 표본설계	업무 절차 설정	수집자료 점검 및 원료	결측치 처리	정보보존 및 공개법위 검토	배포 촉진		
자료 가용성 검토	자료처리 방법 설계	시스템 점검		신규변수(항목) 및 단위 도출	신용률 확정	이용자그룹 관리		
통계 작성 계획안 수립	통계 작성 체계 설계	작성 과정 통합 점검		가중치 계산				
		통계 작성 체계 확정		집계				
				자료 집계 및 확정				

자료: 이의규. (2018). 표준 자료처리 모델 GSDEMs의 소개 및 시사점. p.52.

한국형 통계업무프로세스모델(KSBPM)은 우리나라의 국가(승인) 통계에 최적화된 통계업무프로세스모델이다. 하지만 통계업무 프로세스 중 인공지능(AI) 적용 가능성을 검토한 해외 선행연구는 물론 국내 선행연구에서도 국제 표준인 일반통계업무프로세스모델(GSBPM)을 활용하여 설명하고 있다. 이에 본 연구에서는 일반통계업무프로세스모델(GSBPM)을 활용하여 인공지능(AI) 적용 가능성을 검토하고자 한다.

제4절 선행연구 검토

1. 국외 선행연구

Beck et al.(2018)은 독일 연방통계청(Destatis) 국가 및 국제 통계 생산기관의 머신러닝(ML) 활용 실태조사 결과를 토대로 일반통계업무프로세스(GSBPM) 단계별로 머신러닝(ML)의 적용 가능 분야를 검토한 결과, 설계 단계(2.4), 수집 단계(4.1), 처리 단계(5.1, 5.2, 5.3, 5.4, 5.5), 분석 단계(6.3), 보급 단계(7.5), 평가 단계(8.2)에서 적용 가능하다고 분류하였다([그림 2-6] 참조).

[그림 2-6] 머신러닝(ML) 적용 가능 분야(Beck 등, 2018)

Overarching Processes							
Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysts components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysts	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs Process	7.5 Manage user support Analyse	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

자료: Beck, M., Dumpert, F., & Feuerhake, J. (2018). Machine learning in official statistics. p.10.

한편, 유엔 유럽경제위원회(United Nations Economic Commission for Europe, UNECE) 머신러닝팀의 Yung 등(2018)¹⁷⁾은 국가 통계 생산에 있어서 머신러닝(ML)의 중요성을 설명하고, 1차 데이터(확률 표본)와 2차 데이터(등록 데이터, 빅데이터 및 비확률 표본 등)에서 머신러닝(ML)의 활용 가능성을 설명하였다. Yung 등(2018)은 일반통계업무프로세스모델(GSBPM)의 8개 단계와 그 하위 프로세스를 이용하여 1차 데이터에서 머신러닝(ML) 활용 가능성을 설명하였다(〈표 2-17〉 참조).

〈표 2-17〉 1차 데이터의 머신러닝(ML) 활용 가능성(Yung 등, 2018)

과업	머신러닝 기술	GSBPM 절차
자료연계(Record linkage)	클러스터링(Clustering)	2.4, 5.1
코딩(Coding)	분류(Classification)	2.4, 4.3, 5.2
이상치 탐지(Outlier detection)	클러스터링(Clustering)	2.4, 4.3, 5.1, 6.2
층화(Stratification)	분류(Classification)	4.1, 4.3, 5.4, 5.6
추정(Estimation)	회귀/분류(Regression/classification)	4.3
대체(Imputation)	회귀/분류(Regression/classification)	5.4
교정(Calibration)	회귀/분류(Regression/classification)	5.6
공개 통제(Disclosure control)	회귀/분류(Regression/classification)	6.4

자료: Yung, W. et al. (2018). The Use of Machine Learning in Official Statistics. p.8.

2019년 유엔 유럽경제위원회(UNECE) 공식 통계 현대화를 위한 고위급 그룹(High-Level Group for the Modernisation of Official Statistics, HLG-MOS)은 머신러닝(ML)의 공식 통계 활용을 촉진하고, 통계 기관의 경험을 공유하기 위해 2019년에 머신러닝 프로젝트를 시작하였다(UNECE, 2021).¹⁸⁾ HLG-MOS의 머신러닝(ML) 프로젝트에서는

17) Yung, W., Karkimaa, J., Scannapieco, M., Barcaroli, G., Zardetto, D., Ruiz Sanches, J.A., Braaksma, B., Buelens, B. and Burger, J. (2018) The Use of Machine Learning in Official Statistics. UNECE.

18) United Nations Economic Commission for Europe. (2021). Machine Learning for Official Statistics. United Nations.

그동안의 연구 경험과 잠재적인 부가가치가 높을 것으로 여겨지는 데이터 분류 및 코딩, 편집 및 대체, 이미지 분석의 3개 분야 25개의 파일럿 연구를 진행하였다. 일부 보고서를 제출하지 않아 결과를 확인할 수 없는 연구도 있었으나 전반적으로 3개 분야의 머신러닝(ML) 연구는 매우 성공적인 것으로 평가된다.

한편, 2022년 말 ChatGPT-3.5의 등장으로 촉발된 대형 언어모델(LLM)의 혁신적인 잠재력은 통계 분야에도 영향을 미쳤다. 이에 2023년 유엔 유럽경제위원회(UNECE) 공식 통계 현대화를 위한 고위급 그룹(HLG-MOS)은 대형 언어모델(LLM)에 대한 백서 초안을 작성하였다(UNECE, 2023).¹⁹⁾ 공식 통계 분야에서 대형 언어모델(LLM)의 도입은 아직 초창기이기는 하지만 이 백서에서는 5개의 도입 사례를 통해 통계 기관의 다양한 업무 프로세스에서 효율성을 높이고, 사용자의 편의성을 향상시키는데 기여할 수 있다는 것을 설명하였다(UNECE, 2023). 또한 이 백서에서는 대형 언어모델(LLM)을 구현할 때 발생하는 위험과 완화 조치, 통계 기관이 대형 언어모델(LLM)을 추진함에 있어 고려해야 할 사항 등을 제시하였다(UNECE, 2023).

2. 국내 선행연구

김정민 등(2020)²⁰⁾은 국가통계 부문의 인공지능(AI)·빅데이터 기술 도입 목적을 <표 2-18>과 같이 신규 통계의 생산, 통계 생산 프로세스의 현대화 등 두 가지 시각으로 구분하고 세부 적용 분야를 제안하였다.

19) United Nations Economic Commission for Europe. (2023). Large Language Models for Official Statistics. United Nations.

20) 김정민, 전이슬, 황유림. (2020). AI기술의 국가통계 활용 사례 및 국내 도입 촉진 방안. 이슈리포트 IS-098. 소프트웨어정책연구소.

〈표 2-18〉 국가통계의 인공지능(AI)·빅데이터 기술 도입의 두 가지 시각(김정민 등, 2020)

구분	세부 적용 분야
신규 통계 생산	- 대체(Replace): 기존 통계를 완전 또는 부분 대체하는 수단 - 부가(Addition): 조사 통계 방식으로 도출이 어려운 부문에 대해 데이터 분석에 기반을 둔 우회적 해결 - 인사이트(Insight): 완전히 새로운 분야의 통계 데이터 제시
통계 생산 프로세스의 현대화	- 정확성(Accuracy): 통계의 정확도 향상 - 생산성(Productivity): 업무 효율화를 위한 자동화 기법 - 연결성(Connectivity) : 연결 가능한 데이터들의 상호 보완적 활용 - 비용(Price): 통계 생산의 비용 절감을 위해 활용

자료: 김정민, 전이슬, 황유림. (2020). AI기술의 국가통계 활용 사례 및 국내 도입 촉진 방안. p.23.

또한 김정민 등(2020)은 해외 문헌 고찰을 통해 머신러닝(ML) 활용이 가능하다고 판단되는 통계업무프로세스와 과업을 〈표 2-19〉와 같이 제안하였다.

〈표 2-19〉 머신러닝(ML) 활용이 가능한 통계업무프로세스 및 예상 기술군(김정민, 2021)

코드 (1레벨)	명칭	코드 (2레벨)	명칭	과업 정의	도입 가능 기술군
2	설계	2.4	모집단 및 표본 설계	- 모집단을 식별 및 지정 - 적합한 샘플링 기준 및 방법론 결정	- 분류 - 군집화
4	수집	4.1	모집단 생성 및 샘플링	- 2.4에서 지정된 모집단을 생성 - 2.4에서 지정된 샘플링 수	- 분류
		4.3	수집 진행	- 실사 진행 및 후속 조치 점검 - 수기 데이터 입력 및 현장 작업 관리 등	- 분류 - 군집화 - 회귀
5	처리	5.1	자료 통합	- 데이터의 출처가 2개 이상일 시 통합 - 검증된 비통계 데이터로 설문 결과 일부를 대체하는 과정도 이에 해당	- 군집화
		5.2	분류 및 코딩	- 설문 결과 데이터의 분류 및 코딩	- 분류
		5.3	검토 및 검증	- 데이터를 검사하여 이상치, 항목 무응답, 잘못된 코딩 등 잠재적 문제, 오류, 불일치 사항을 식별	- 분류 - 군집화
		5.4	편집 및 대체	- 부정확 데이터 및 누락, 신뢰할 수 없는 데이터를 새로운 값으로 대체 및 제거	- 군집화 - 회귀 - 분포 추정
		5.6	가중치 계산	- 설계 단계(2)에서 정의한 가중치를 조사 결과에 적용함으로써 모집단에 대한 통계 결과를 도출 - 설문 결과의 무응답 조정 및 변수의 정규화 등	- 분류 - 회귀
6	분석	6.2	산출물 검증	- 품질 측정 프레임워크에 따라 통계 품질을 검증	- 군집화
		6.3	산출물 해석 및 설명 작성	- 결과가 초기 기대치를 얼마나 잘 반영하는지 평가 - 다양한 관점에서 통계 결과를 해석 - 심층 통계 분석 등	- (미정)주)
		6.4	정보보호 및 정보공개 검토	- 데이터 배포 과정의 기밀성 훼손 방지 - 필요에 따라 데이터 비식별화 처리 수행	- 분류 - 회귀
7	배포	7.5	이용자 지원 관리	- 통계 서비스에 대한 사용자 질의 및 요청에 대해 합의된 마감일 내에 응답을 제공	- (미정)
8	평가	8.2	평가 실시	- 8.1 단계에서 확보한 평가 피드백과 목표한 벤치마킹 대상(만약 존재한다면)과 결과를 비교한 후 평가보고서 작성	- (미정)

주(미정): 2차 분석의 영역으로서 예상 기술군을 특정하지 않음.

자료: 김정민, 전이슬, 황유림. (2020). AI기술의 국가통계 활용 사례 및 국내 도입 촉진 방안. pp.28-30.

한편 김정민(2021)²¹⁾은 2020년도의 연구 결과(김정민 등, 2020)와 함께 국가통계 AI 도입 활성화 방안으로 인공지능(AI) 국가통계 생산 절차 수립, 인공지능(AI) 기술 표준화, 인공지능(AI) 평가 기준 마련, 국가 통계의 인공지능(AI) 기술 활용 지침 개발 등을 담당하는 인공지능(AI) 표준 수립기관(안)을 제안하였다. 또한 인공지능(AI)·빅데이터 기술을 활용하여 소프트웨어(SW) 분야의 새로운 국가통계를 개발하기 위한 실현 가능한 아이디어로 인공지능(AI)·빅데이터를 활용한 ‘디지털산업 분류 방안’과 ‘전자공시 데이터 기반 디지털전환 지수 산출 방안’을 제안하였다(김정민, 2021)

이상현(2021)²²⁾은 전통적인 통계작성 과정에 인공지능(AI)을 이용한 통계편제 방법을 도입하거나 사무자동화 등에 참고할 수 있도록 통계 생산 표준 프로세스 단계별 인공지능(AI) 적용 가능 업무를 제안하였다.

〈표 2-20〉 통계업무 프로세스 단계별 인공지능(AI) 적용 가능 업무(이상현, 2021)

단계	하위 부문	인공지능(AI) 적용 가능 업무
2. 설계	2.4 모집단 및 표본 설계	- 샘플링 구조 식별작업 * 샘플링 구조 설계 정보의 코딩 및 품질검증
	4.1 자료수집 대상 설정	- 샘플링 효율화를 위한 모집단 층화 작업
4. 수집	4.3 자료수집 진행	- 자료수집 활동의 관리업무 효율화 - 이상치 파악, 오류 수정 등
5. 가공	5.1 자료 통합	- 자료 소스나 수집 방법이 상이한 자료들의 매칭과 통합
	5.2 분류 및 코딩	- 입력자료의 분류 및 코드화(기호, 숫자)
	5.4 결측치 처리	- 누락자료 등의 대체 방법 선택, 자료 대체
	5.6 가중치 계산	- 무응답 조정을 위한 가중치 사용
6. 분석	6.2 산출물 검정	- 최종 산출물의 이상치 검색 등 통계 품질 검증
	6.4 정보공개 범위 설정	- 자료배포 과정의 기밀성 훼손 방지 * 필요시 데이터 비식별화
7. 배포	7.5 이용자 그룹 관리	- 통계서비스 이용자 질의에 대한 응답
8. 평가	8.2 평가 실시	- 피드백 자료와 벤치마킹 대상 비교 평가

자료: 이상현. (2021). AI를 활용한 국내의 통계편제 현황. p.46.

21) 김정민. (2021). 국가통계 AI도입 활성화 방안 연구.

22) 이상현. (2021). AI를 활용한 국내의 통계편제 현황. 국민계정리뷰, 2021(2). 한국은행.

또한 이상현(2021)은 우리나라 통계청에서 추진하고 있는 인공지능(AI) 적용 국가통계 사례를 <표 2-21>과 같이 정리하여 제시하였다. 통계청의 추진 사례는 앞에서 설명한 인공지능(AI) 적용 가능 업무와는 다소 차이가 있었다. 다만 2020년 9월까지의 통계청의 추진 사례를 정리하였음에도 불구하고 인공지능(AI) 적용 가능 업무로 제안한 업무의 상당 부분이 시험편제 중이거나 완성 단계에 도달되었음을 알 수 있다.

<표 2-21> 통계청의 AI 기법 적용 국가통계사례(이상현, 2021)

단계	하위 부문	인공지능(AI) 적용 검토 업무	비고
4. 수집	4.3 자료수집 진행	- 다출처 데이터 수집 (AI 활용한 온라인 가격자료 수집)	
	4.4 수집자료 점검	- 데이터 처리 및 내검 개선 (머신러닝 기반 통계적 내검 및 자료처리)	
5. 가공	5.2 분류 및 코딩	- 산업 및 직업분류 코드 (설문상 응답 분석 및 분류 코드 자동 매칭)	완성
	5.4 결측치 처리	- 데이터 정제 (무응답 대체 시 머신러닝 또는 딥러닝 적용)	
6. 분석	6.3 산출물 해석 및 설명자료 작성	- 통계 방법의 효율적 대체 (머신러닝 예측모형을 시산 결과에 적용)	
7. 배포	7.4 배포 촉진	- 통계서비스 개선 (AI 기반 챗봇시스템 구축 및 결과 분석)	완성

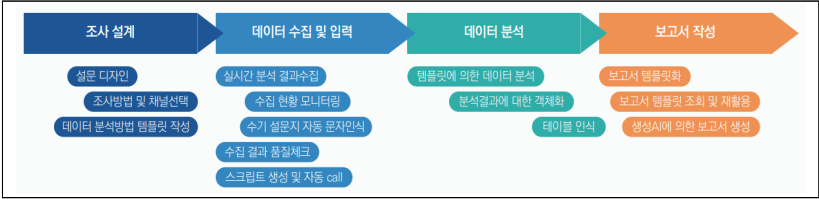
자료: 이상현. (2021). AI를 활용한 국내의 통계편제 현황. p.47.

이기재(2023)²³⁾는 인공지능(AI)과 Chat-GPT 기법을 사용하면 조사 연구의 효율성, 정확성 및 효과가 개선되고 궁극적으로 더 나은 통찰력과 정보에 입각한 의사결정이 가능할 것이라고 전망하였으며, [그림 2-7]과 같이 조사 연구에서 인공지능(AI) 활용 방안을 제안하였다.

23) 이기재. (2023). 조사연구에서 인공지능(AI) 활용방안. 미래정책포커스, 2023(여름호), pp.52-53.

68 인공지능(AI) 기반 조사데이터 품질 제고 방안 연구

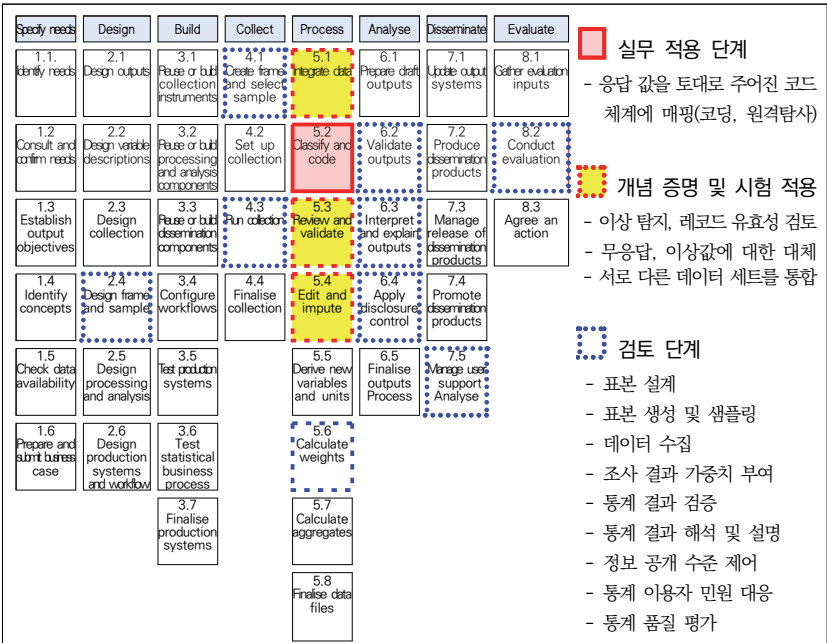
[그림 2-7] 조사 연구에서 인공지능(AI) 활용 방안(이기재, 2023)



자료: 이기재, (2023). 조사연구에서 인공지능(AI) 활용방안. 미래정책포커스, 2023(여름호), p.53.

김정민(2024)은 통계 생산 방식 효율화 측면에서 우리나라 국가통계 분야의 인공지능(AI) 활용을 ‘실무 적용 단계’, ‘개념 증명 및 시험 적용 단계’, ‘검토 단계’로 정리하여 발표하였다.

[그림 2-8] 인공지능을 활용한 통계 생산 방식 효율화(김정민, 2023)



자료: 김정민, (2024). 통계분야의 AI 도입-국내외 사례 및 향후 과제. 제13회 국가통계발전포럼 자료집, p.155. (재작성)

제5절 소결

제2장에서는 인공지능의 개념 및 최신 기술 동향, 데이터 오류 탐지 방법론, 통계업무프로세스모델, 인공지능(AI)을 활용한 선행연구를 검토하였다.

인공지능(AI)의 개념은 연구자의 관점에 따라 또는 국가별 법률에 따라 다소 차이는 있으나 인간처럼 이해(학습과 지각, 언어의 이해 등)하고, 실제 환경이나 가상 환경에 영향을 미칠 수 있도록 인간처럼 행동(예측, 9추론, 판단 등) 하는 기계 기반 시스템으로 정의할 수 있다.

인공지능(AI) 기술 중 머신러닝(ML)과 딥러닝(DL)에 대해 살펴보았다. 머신러닝(ML)은 지도 학습, 비지도 학습, 강화 학습의 개념과 기술과 응용 사례를 설명하였다. 머신러닝(ML) 기법 중 하나인 딥러닝(DL)은 합성곱 신경망(CNN)과 순환 신경망(RNN)의 개념과 기술과 응용 사례를 설명하였다. 다음으로 최근 주목받고 있는 머신러닝(ML)과 딥러닝(DL) 기술 및 관련 분야로 옛지 AI, 연합학습, 설명 가능한 AI(XAI), 전이 학습 및 사전 학습 모델, 강화 학습(RL), 윤리적 AI 및 편견 완화, 자동화된 머신러닝(AutoML), 신경 구조 탐색(NAS), 양자 머신러닝(QML), 의료용 AI, 지속 가능한 AI, 멀티모달 러닝, 금융 분야의 인공지능(AI), 자연어 처리(NLP), 사이버 보안을 위한 인공지능(AI), 개인화 및 추천 시스템, 대형 언어모델(LLM)에 대해 간략히 살펴보았다.

한국의료패널 조사데이터 같은 테이블 구조 데이터에 대한 오류를 탐지하기 위해 규칙 기반 오류 확인 방법, 연관규칙 활용 방안, 머신러닝(ML) 및 딥러닝(DL) 방법, 대형 언어모델(LLM) 등 다양한 방법론에 대한 활용 방안도 검토하였다.

인공지능(AI) 적용이 가능한 통계업무 프로세스를 살펴보기에 앞서, 국

가 통계를 생산하는 데 필요한 일련의 업무 프로세스를 정의할 필요가 있다. 이에 국제 표준인 일반통계업무프로세스모델(GSBPM)과 한국형 통계업무프로세스모델(KSBPM)을 살펴보았다. 한국형 통계업무프로세스 모델(KSBPM)이 우리나라 국가통계 생산 업무를 설명하는 데 최적화된 통계업무프로세스모델이지만 인공지능(AI)을 활용한 통계업무프로세스 가능성을 검토한 국내외 선행연구 및 적용 사례에서는 일반통계업무프로세스모델(GSBPM)을 활용하였기에, 본 고에서는 일반통계업무프로세스 모델(GSBPM)을 사용하였다.

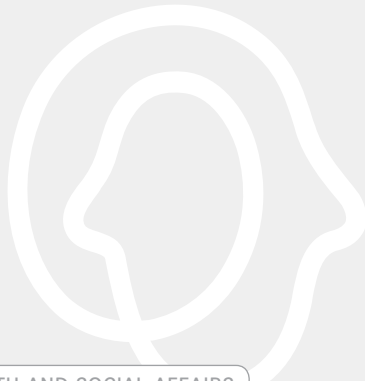
마지막으로 통계업무 프로세스 중 인공지능(AI)을 적용할 수 있는 업무를 살펴보기 위해 국내외 선행연구를 검토하였다. 연구자에 따라 다소 차이는 있으나 종합하면 [그림 2-9]와 같다.

[그림 2-9] 통계업무 프로세스 단계별 인공지능(AI) 적용 가능 업무(종합)

Overarching Processes							
Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design production and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs Process	7.5 Manage support Analyse	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

자료: 저자 작성

한편, 우리나라의 경우 초기부터 독일이나 유엔 유럽경제이사회(UNECE)의 사례를 지속적으로 검토하고 있다. 현재 국가통계 분야에서 인공지능(AI) 활용 수준이 일부 업무에서는 실무 적용 단계(5.2) 또는 개념 증명 및 시험 단계(5.1, 5.3, 5.4)에 도달한 것으로 파악되었다.



제3장

인공지능(AI)을 활용한 통계작성 사례 분석

제1절 주요 외국의 인공지능(AI) 활용 통계작성 사례
제2절 우리나라의 인공지능(AI) 활용 통계작성 사례 분석
제3절 소결

제3장 인공지능(AI)을 활용한 통계작성 사례 분석

제1절 주요 외국의 인공지능(AI) 활용 통계작성 사례

1. 국가 및 국제기구 통계기관의 머신러닝 활용 실태

2018년 독일 연방통계청(Destatis)은 국가 통계기관에서 머신러닝(M)을 활용하는 실태를 조사하고자 EU 회원국 27개 국가, 유럽자유무역연합(EFTA) 4개 국가, 일부 비유럽 국가 6개(호주, 캐나다, 이스라엘, 일본, 뉴질랜드, 미국) 총 37개 국가 및 2개 국제기구(OECD, Eurostat)를 대상으로 설문조사를 실시하여, 25개 국가 2개 국제기구에서 총 136개의 머신러닝(ML) 프로젝트를 진행하고 있다고 응답 받았다(〈표 3-1〉 참조).

국가별로는 캐나다가 36개로 가장 많았고, 미국 20개, 네덜란드가 16개 순으로 나타났다.

〈표 3-1〉 주요 국가 통계기관의 머신러닝 활용 실태

기관	프로젝트 수
캐나다(Statistics Canada)	36
네덜란드(Statistics Netherlands)	16
미국(U.S. Bureau of Labor Statistics)	11
뉴질랜드 통계청(Stats NZ)	9
미국(U.S. Department of Agriculture NASS)	7
호주 통계청(Australian Bureau of Statistics)	6
스위스(Federal Statistical Office of Switzerland)	6
프랑스(INSEE)	5
루마니아(National Institute of Statistics Romania)	4
오스트리아(Statistics Austria)	4

기관	프로젝트 수
포르투갈(Statistics Portugal)	4
스페인(Statistics Spain (INE))	3
스웨덴(Statistics Sweden)	3
EU(Eurostat)	2
룩셈부르크(STATEC)	2
핀란드(Statistics Finland)	2
아일랜드(Statistics Iceland)	2
폴란드(Statistics Poland)	2
미국(Bureau of Economic Analysis(USA))	1
라트비아(Central Statistical Bureau of Latvia)	1
아일랜드(Central Statistics Office of Ireland)	1
헝가리(Hungarian Central Statistical Office)	1
이탈리아(Italian National Institute of Statistics)	1
일본(National Statistics Center, Japan)	1
OECD	1
영국(ONS)	1
벨기에(Statistics Belgium)	1
덴마크(Statistics Denmark)	1
노르웨이(Statistics Norway)	1
미국(U.S. Census Bureau)	1
합계	136

자료: Beck, M., Dumpert, F., & Feuerhake, J. (2018). Machine learning in official statistics. p.8.

136개 머신러닝(ML) 프로젝트의 ‘통계 주제’는 교차통계가 26개, 노동시장이 15개, 비즈니스통계 15개, 가구통계 14개, 농업통계 10개, 인구조사 8개, 분기분류 7개, 가격통계 5개, 교통통계 4개, 기타 32개로 나타났다(〈표 3-2〉 참조).

〈표 3-2〉 머신러닝(프로젝트) 통계 주제

구분	프로젝트 수
교차통계(Cross-statistical)	26
노동시장(Labour market)	15
비즈니스통계(Business statistics)	15
가구통계(Household statistics)	14
농업통계(Agricultural statistics)	10
인구조사(Census)	8
분기분류(Branch classification)	7
가격통계(Price statistics)	5
교통통계(Traffic statistics)	4
기타(Other)	32
합계	136

자료: Beck, M., Dumpert, F., & Feuerhake, J. (2018). Machine learning in official statistics. p.7.

‘머신러닝(ML) 활용 형태’는 다중응답 방식으로 조사하여 총 161개의 응답을 받았으며, 이 중 분류(Classification)가 78개로 가장 많았고, 대체(Imputation) 22개, 마이크로데이터 연결(Microdata linking) 15개, 클러스터링(Clustering) 9개, 텍스트 분석(Text analysis) 8개, 회귀분석(Regression) 6개, 신분 확인(Identification) 4개, 차원 축소(Dimension) 2개, 기타 17개로 나타났다(〈표 3-3〉 참조).

〈표 3-3〉 머신러닝(프로젝트) 활용 형태

구분	프로젝트 수
분류(Classification)	78
대체(Imputation)	22
마이크로데이터 연결(Microdata linking)	15
클러스터링(Clustering)	9
텍스트 분석(Text analysis)	8
회귀분석(Regression)	6
신분 확인(Identification)	4
차원 축소(Dimension)	2
기타(Other)	17
합계	161

자료: Beck, M., Dumpert, F., & Feuerhake, J. (2018). Machine learning in official statistics. p.7.

다중응답으로 조사한 결과 ‘사용된 머신러닝(ML) 알고리즘’은 총 171개라는 응답을 받았으며, 이 중 랜덤 포레스트(Random forest)가 37개로 가장 많았고, 신경망(Neural networks) 22개, 서포트 벡터 머신(Support vector machine) 22개, 의사결정나무(Decision tree methods) 20개, 최근접 이웃 접근법 (Nearest-neighbour approaches) 12개, 베이저안 접근법(Bayesian approaches) 6개, 자연어 처리(Natural language processing) 5개, 클러스터 방식(Cluster method) 2개, 기타 45개로 나타났다(〈표 3-4〉 참조).

〈표 3-4〉 머신러닝(프로젝트) 알고리즘 방법

구분	프로젝트 수
랜덤 포레스트(Random forest)	37
신경망(Neural networks)	22
서포트 벡터 머신(SVM)	22
의사결정나무(Decision tree methods)	20
최근접 이웃 접근법(Nearest-neighbour approaches)	12
베이저안 접근법(Bayesian approaches)	6
자연어 처리(Natural language processing)	5
클러스터 방식(Cluster method)	2
기타(Other)	45
합계	171

자료: Beck, M., Dumpert, F., & Feuerhake, J. (2018). Machine learning in official statistics. p.7.

독일 연방통계청(Destatis)의 국가 및 국제 통계 생산기관의 머신러닝(ML) 활용 실태조사는 주요 국가 및 국제기구의 통계 생산기관에서 인공지능(AI)을 활용한 통계작성 실태를 파악하였다는 데 그 의미가 있지만 다음과 같은 한계점도 갖고 있다. 첫째, 2018년에 조사가 실시된 이래로 상당 기간이 경과되어 기존 프로젝트의 지속 여부 및 최신 동향을 파악하기 어렵다. 둘째, 2018년 이후 추진된 신규 프로젝트 동향을 파악하기 어렵다. 셋째, 2018년 이후 주목받고 있는 최신 인공지능(AI) 활용 동향을 파악하기 어렵다.

2. 유엔 유럽경제위원회(UNECE) 머신러닝 파일럿 프로젝트

유엔 유럽경제위원회(UNECE)의 ‘공식 통계 현대화를 위한 고위급 그룹 머신러닝 프로젝트(HLG-MOS Machine Learning Project)’에서는 머신러닝(ML)을 활용한 ‘코딩 및 분류’, ‘편집 및 대체’, ‘이미지 분석’의 3가지 주제로 25개 파일럿 연구를 수행하였다.

가. 코딩 및 분류²⁴⁾

유엔 유럽경제위원회(UNECE)는 코딩 및 분류를 주제로 9개의 파일럿 프로젝트를 수행했다(〈표 3-5〉 참조).

〈표 3-5〉 유엔 유럽경제위원회(UNECE)의 머신러닝(ML) 파일럿 연구(코딩 및 분류)

구분	프로젝트
멕시코 국립통계 및 지리연구소	자연어 처리를 사용한 직업 및 경제 활동 코딩
캐나다 통계청	산업 및 직업 코딩
벨기에 프란더스 통계청	트위터 데이터의 감성 분석
세르비아 통계청	노동력 조사에서 수집된 경제 활동에 대한 텍스트로 기술된 데이터 코딩
미국 노동통계국	업무상 상해 및 질병 코딩
폴란드 통계청	목적에 따른 개인 소비의 유럽 분류(ECOICOP)에 대한 생산 설명
국제통화기금(IMF)	IMF 시계열 카탈로그를 사용한 자동 코딩
아이슬란드	사회 통계 조사에서 직업 및 산업의 자동 코딩
노르웨이	머신러닝을 이용한 표준 산업 코드 분류

자료: 저자 작성

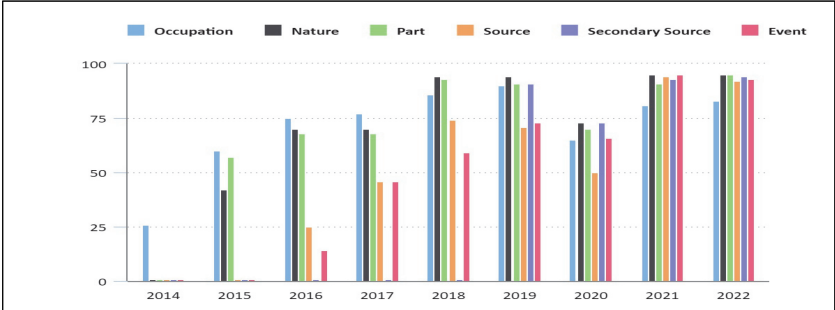
24) Claus Sthamer. (2020). UNECE-HLG-MOS Machine Learning Project-Classification and Coding Theme Report. <https://statswiki.unece.org/spaces/ML/pages/285216478/WP1+-+Theme+1+Coding+and+Classification+Report>

전통적이고 수동적인 코딩 및 분류 프로세스는 많은 자원을 소모하고 오류를 발생시킬 수 있다. 코딩 및 분류 작업에 머신러닝(ML) 도입 시, 데이터 세트의 크기에 관계없이 빠른 분석 속도와 우수한 예측 결과를 얻을 수 있는 것으로 나타났으며, 코딩 및 분류는 머신러닝(ML)을 도입하는 데 최적화된 활용처로 추천된다.

한편, 머신러닝(ML)을 도입한 결과 그 정확도가 활용 가능한 수준까지 고도화되기 위해서는 지속적으로 학습 데이터를 구축 및 축적해야 한다는 점을 강조했으며, 코드 분류를 인간이 직접 수행했을 때와의 비용 편익에서는 국가별 통계조직의 기계학습 기술 역량, 시간 경과에 따른 소요 비용 변화 등의 변수가 존재하므로, 단기적인 효율보다는 장기적인 관점에서 비용 효율이 발생할 것이라 예상하였다.

머신러닝(ML)을 활용한 코딩 및 분류의 대표적인 사례는 미국 노동부 노동통계국(Bureau of Labor Statistics, BLS)의 업무상 상해 및 질병 코딩이다. 미국 노동부 노동통계국(BLS)은 매년 업무상 상해 및 질병 설문조사(Survey of Occupational Injuries and Illnesses, SOII)를 통해 약 300,000건의 데이터를 수집하고 있으나, 설문조사 중 서술형 응답을 표준 직업분류와 직업성 질환 및 상해 분류로 수동 코딩하는 데 연간 약 25,000시간이 소요되었다. 머신러닝을 활용한 방식은 서술형 응답을 분석해 각 응답에 가장 적합한 코드 6개(occupation, nature, part, source, secondary source, event)를 추천해줌으로써 자동 분류하는 방식이다. 이를 위해 로지스틱 회귀(2018년 이전), 심층 신경망(2019년 이후)을 사용하고 있다. 업무상 상해 및 질병 설문조사(SOII)의 근로자 상해 응답에 대한 자동 코드 분류는 매년 지속적으로 향상되었으며, 2021~2022년의 경우 성공률이 92% 수준으로 크게 향상된 것으로 나타났다([그림 3-1] 참조).

[그림 3-1] 연도별 미국 상해 및 질병 데이터 자동코딩 성공률



자료: 미국 노동통계국(BLS) 홈페이지. <https://www.bls.gov/iif/automated-coding.htm>에서 2024.08.24. 추출

나. 편집 및 대체²⁵⁾

유엔 유럽경제위원회(UNECE)는 편집 및 대체를 주제로 8개의 머신러닝(ML) 파일럿 연구를 수행했다(〈표 3-6〉 참조).

〈표 3-6〉 유엔 유럽경제위원회(UNECE)의 머신러닝(ML) 파일럿 프로젝트(편집 및 대체)

구분	파일럿 연구
이탈리아	개인 기본 등록부의 변수 "교육 수준"의 대체
폴란드	폴란드 거주자의 여행 참여에 대한 표본 조사에서의 대체
독일	대체를 위한 머신러닝
벨기에	머신러닝을 이용한 에너지 밸런스 통계의 초기 추정
캐나다	에너지 밸런스의 초기 추정을 위한 시계열 모델
영국	생활비 및 식품 조사 소득 데이터 편집
이탈리아	이탈리아 공공행정 등록부 편집
이탈리아	NSI에서 데이터 편집 클리닝을 위한 머신러닝

자료: 저자 작성

25) Florian Dumpert. (2020). UNECE-HLG-MOS Machine Learning Project-Edit and Imputation Theme Report. <https://statswiki.unece.org/spaces/ML/pages/290358735/WP1+-+Theme+2+Edit+and+Imputation+Report>

해당 과업은 조사 과정에서 빈번히 발생하는 잘못된 데이터 또는 결측값(missing value)을 보완함으로써 통계 수치에 직접적인 영향을 끼칠 수 있다. 이 점을 고려해서인지 유엔 유럽경제위원회(UNECE)는 머신러닝(ML)에 기반한 편집 및 표본 대체의 완전 자동화를 추천하지 않는 것으로 나타났으며, 머신러닝(ML) 기술을 특정 의심 데이터에 관한 의사결정을 지원하는 보조 수단으로 사용하는 것을 권장한다. 예컨대 기계학습은 오기가 의심되는 데이터를 추천하는 용도로 활용할 수 있을지언정, 편집 및 표본 대체의 대상이 되는 수치의 결정은 반드시 관련 전문가의 노하우가 요구된다고 강조하고 있다.

다. 이미지 분석²⁶⁾

유엔 유럽경제위원회(UNECE)는 이미지 분석을 주제로 5개의 머신러닝(ML) 파일럿 연구를 수행했다(〈표 3-7〉 참조).

〈표 3-7〉 유엔 유럽경제위원회(UNECE)의 머신러닝(ML) 파일럿 연구(이미지 분석)

구분	파일럿 연구
호주	- 주소 등록 자동 이미지 인식(AIR) 모델
네덜란드	- 이미지에서 통계 정보 학습: 개념 증명
스위스	- Arealstatistik Deep Learning (ADELE)
멕시코	- 비인구 조사 연도의 도시 지역 매핑을 위한 Landsat 위성 데이터 활용
UNECE	- 위성 데이터와 머신러닝을 활용한 공식 통계 생산을 위한 일반 파이프라인

자료: 저자 작성

26) Abel Coronado and Jimena Juárez. (2020). UNECE-HLG-MOS Machine Learning Project-Imagery Theme Report. <https://statswiki.unece.org/spaces/ML/pages/290358675/WP1+-+Theme+3+Imagery+Analysis+Report>

각국의 이미지 분석 실험은 딥러닝(DL) 기반의 위성사진 분석에 집중되는 경향을 보였다. 주로 자국의 영토에 대한 지적조사 및 주소 등록을 자동화하는 데 목적을 두었다. 유엔 유럽경제위원회(UNECE)는 보고서를 통해 해당 프로젝트에 참여한 국가별 제안 모델과 각각의 분석 정확도를 공개하였다. 실험 결과 호주의 경우 합성곱 신경망(Convolution Neural network, CNN)을 분석 데이터에 맞게 커스터마이징한 모델을 통해 96.9%의 위성 이미지 분석 정확도를 달성하였으며, 그 외 네덜란드, 스위스의 경우 서로 다른 합성곱 신경망(CNN) 모델을 활용하였으나 각각 74.0%, 90% 수준의 정확도를 달성하여 보완을 계획하고 있는 것으로 나타났다. 멕시코의 경우 딥러닝(DL)이 아닌 엑스트라 트리 모델(Extra-Tree model)을 사용하였는데, 93.87%라는 높은 정확도를 달성한 점이 특징적이었다.

한편, 유엔 유럽경제위원회(UNECE)는 딥러닝 기반의 학습이 특수한 하드웨어(ex: 고성능 컴퓨터)의 사용을 강제하는 측면이 있고, 대량의 기밀 데이터를 학습시키는 과정에서 보안 문제가 발생할 수 있다고 지적하였다. 이를 해결하기 위해서는 고성능 하드웨어를 내부 데이터센터에 통합하는 방안을 강구해야 한다고 지적했다.

호주²⁷⁾를 제외한 나머지 국가들은 아직 개념 증명 단계에 있으며, 호주만은 이미 프로젝트를 통계 생산 단계로 옮겨 현재 해당 기관에서 운영 중이다.

27) 호주 통계청. ABS Address Register, Users' Guide. <https://www.abs.gov.au/statistics/research/abs-address-register-users-guide> 보완 예정

3. 공식 통계를 위한 대형 언어모델(LLM)

유엔 유럽경제위원회(UNECE)의 공식 통계 현대화를 위한 고위급 그룹(HLG-MOS)은 2023년 공식 통계 생산을 대형 언어모델(LLM) 활용 방안에 대한 아이디어를 얻고자 국가 및 국제기구의 도입 사례를 조사하였다(〈표 3-8〉 참조).

〈표 3-8〉 국가 및 국제기구의 대형 언어모델(LLM) 도입 사례(UNECE, 2023)

사례	설명
호주 통계청 (ABS)	- 통계분류 정의 업데이트 시 대형 언어모델(LLM)을 활용하여 직업 별 수행 작업 목록 초안을 생성 → 분석가의 업무 효율성을 향상
아일랜드 통계청 (CSO Ireland)	- SAS 코드를 R 코드로 변환하는 도구 개발 → 코드 변환 작업 간 소화 및 자동화
국제통화기금 (IMF)	- StatGPT를 개발 → 자연어 기반 인터페이스를 통해 사용자가 IMF 통계 데이터에 쉽게 접근하고 활용할 수 있도록 지원
캐나다 통계청 (Statistics Canada)	- 대형 언어모델(LLM)을 사용하여 통계 데이터 테이블에 대한 텍스트 설명 초안을 자동으로 생성 → 보고서 작성 프로세스 효율화
국제결제은행 (BIS)	- 대형 언어모델(LLM)을 활용하여 메타데이터 편집 작업을 자동화 하고 품질을 향상시킴

자료: UNECE. (2023). Large Language Models for Official Statistics. p.22-41. (요약)

유엔 유럽경제위원회(UNECE)의 대형 언어모델(LLM) 보고서는 아직 프리프린트(pre-print) 수준이지만, 공식 통계를 위한 대형 언어모델(LLM)의 활용은 통계기관의 운영에 중요한 역할을 할 것이라는 데는 의심의 여지가 없다고 보고하고 있다. 통계기관의 이메일 작성이나 회의록 작성 같은 일상적인 업무에서부터 통계업무 프로세스 같은 전문적인 업무에 이르기까지 업무 효율성을 향상하는 데 대형 언어모델(LLM)이 활용될 수 있을 것으로 예상된다.

국가 및 국제기구의 대형 언어모델(LLM) 사례에서 보았듯이 SAS에서 R로의 코드 변환(아일랜드 통계청), 통계분류시스템 업데이트(호주통계

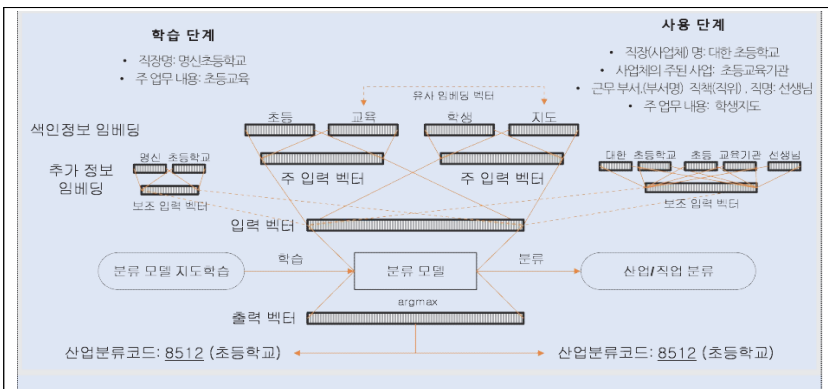
청), 보고서 작성(캐나다 통계청), 자연어 기반의 데이터 검색 및 메타데이터 편집(국제통화기금, 국제결제은행) 등 다양한 활용 사례가 조사되고 있으며, 향후 더 많은 통계기관에서 더 다양한 업무에 대형 언어모델(LLM)을 활용하려는 시도가 있을 것으로 예상된다.

제2절 우리나라의 인공지능(AI) 활용 통계작성 사례 분석

1. 인공지능(AI) 통계분류 자동화

통계개발원은 2020년 ‘인공지능(AI) 통계분류 가능성 검토’를 위해 통계청의 지역별 고용조사 2회차 약 520,000건의 산업분류 문항 텍스트 자료에 대해서 인공지능(AI) 통계분류 예측 정확도에 대한 시험분석을 진행하였다(임경민, 2024).

[그림 3-2] 지도 학습 기반 분류 학습을 위한 초기 기준 모델



자료: 임경민. (2024). AI 통계분류 결과분석 및 실무활용성 제고방안 연구. p.5.

지도 학습 기반 머신러닝 방법론을 적용하여 시산한 결과, 대분류 수준에서는 95.9%, 중분류 수준에서는 93.8%, 소분류 수준에서는 91.4%, 세분류 수준에서는 88.9%의 분류 예측 정확도 결과가 도출되었다. 이는 제한된 일부 자료에 대한 분석이지만 비교적 높은 수준의 분류 정확도와 빠른 수행 속도를 보여, 향후 인공지능(AI) 통계분류 자동화의 가능성을 확인할 수 있었다(임경민, 2023).

〈표 3-9〉 머신러닝 기반 산업분류 예측 성능 평가

모델	분류	대분류 (21)	중분류 (77)	소분류 (232)	세분류 (495)
기본 모델 (FastTest + FCNN, epoch=200,lr=0.05)	정확도(F1)	0.959	0.938	0.914	0.889
	학습시간(s)	67.0	143.0	271.2	393.0
	분류수행속도(iter/s)	79691.9	81395.6	39805.3	21773.8

자료: 임경민. (2024). AI 통계분류 결과분석 및 실무활용성 제고방안 연구. p.6.

2020년에 기초연구를 통해 인공지능(AI) 통계분류 가능성을 확인한 후 2021년에는 기존 통계분류시스템(규칙 베이스 기반 분류모델)과 인공지능(AI) 통계분류를 비교하는 연구를 수행한 결과, 인공지능(AI) 분류 예측이 기존 규칙 기반 사례 사전 자동코딩에 비해 정확도 및 커버리지에서 더 우월할 결과를 보임을 확인할 수 있었다(임경민, 2024).

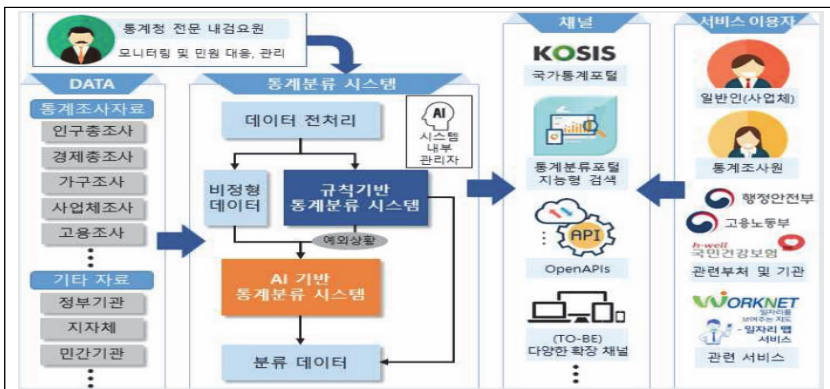
[그림 3-3] 기존 시스템과 인공지능(AI) 통계분류 시스템 성능 비교(분류 예측 정확도)

<기존 규칙 기반 자동코딩시스템 성능>						<AI 기반 통계분류 예측결과와 비교>						
구분	총계	미적용 데이터	자동코딩 활용	공표 정보 일치	공표 정보 불일치	평가 데이터	분류	정밀도 (Precision)	재현율 (Recall)	F-1 Score (Accuracy)	정확도 (Accuracy)	비고
공중 분류	1,187,855 (100%)	314,751	873,104	710,955	162,149	전체 데이터	공중	0.88	0.88	0.88	0.88	전체적인 정확도 상승
		26.50% 총 데이터 대비	73.50% 총 데이터 대비	81.43% 총 데이터 대비	18.57% 총 데이터 대비		발주자	0.88	0.88	0.88	0.88	
발주자 분류		505,989	681,866	595,844	86,022	현 시스템 분류 결과 = 최종결과인 경우	공중	0.96	0.95	0.96	0.95	자동코딩 시스템과 유사성
		42.60% 총 데이터 대비	57.40% 총 데이터 대비	87.38% 총 데이터 대비	12.62% 총 데이터 대비		발주자	0.95	0.95	0.95	0.95	
종류	분류개수(종)	분류코드 구성		사례사진/색인DB 구성		현 시스템 분류 결과 없는 경우	공중	0.81	0.82	0.81	0.82	사전 작업 감소 정도
공중 분류	64	* 만족되지 않은 3자리 코드		* 규칙(조건: 128,336개 (48,969개 고유색인어))			발주자	0.83	0.82	0.82	0.82	
발주자 분류	96	* 중앙정부기관(25) 5자리 * 지방자치단체(24) 4자리 * 기타(4) 5으로 시작하는 3자리 * 민간부문(43) 8로 시작하는 3자리 * 제조업(25), 비제조업(18)		* 규칙(조건: 40,652개 (82,365개 고유색인어))		현 시스템 분류 결과 ≠ 최종결과인 경우	공중	0.67	0.62	0.63	0.62	오답률 개선 정도
							발주자	0.77	0.74	0.73	0.74	

자료: 임경민. (2024). 통계 분야의 AI 도입 사례-AI기반 통계분류 자동화. p.165.

2020~2021년의 기초연구를 통해 AI 기반 통계분류 자동화의 가능성을 확인한 후, 통계청은 추가 후속 연구 및 실무 활용을 위해 2022년에 AI 통계분류 자동화 시스템 개발 연구를 추진하였다(임경민, 2024).

[그림 3-4] 인공지능(AI) 통계분류 자동화 시스템 개념도



자료: 임경민. (2024). AI 통계분류 결과분석 및 실무활용성 제고방안 연구. p.5.

2022년의 인공지능(AI) 통계분류 자동화 시스템은 대규모 데이터를 처리하는 5개 조사(인구주택총조사, 전국사업체조사, 지역별고용조사,

건설업조사, 온라인쇼핑조사)의 5종 분류(산업·직업 분류, 공중·발주자 분류, 상품군별 분류)를 대상으로 개발되었다.

5개 조사의 5종 분류 관련 총 1억 6,000~7,000 만 건의 텍스트를 중복 제거 등의 전처리 정제작업을 거쳐 학습 및 평가 데이터로 사용하였으며, AI 분류 예측 정확도 평가 결과는 [그림 3-5]과 같다.

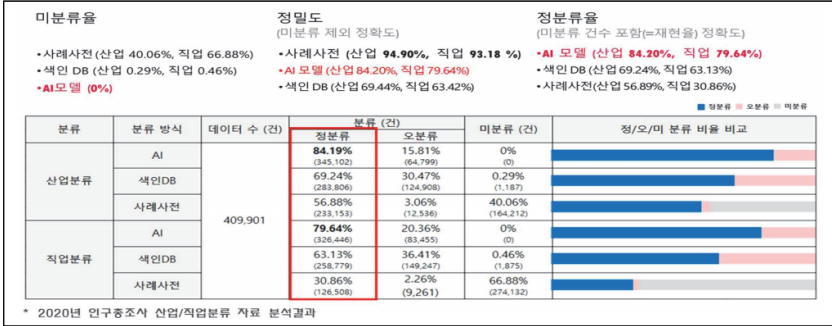
[그림 3-6]은 기존 시스템과 인공지능(AI) 통계분류 시스템의 분류 예측 정확도를 비교한 결과이다. 미분류 건수를 포함하는 정분류율의 경우 인공지능(AI) 통계분류 시스템이 기존 시스템에 비해 월등히 뛰어난 것으로 나타났다.

[그림 3-5] 인공지능(AI) 통계분류 모델 구축(5개 조사, 5종 분류) 및 성능평가 결과

분류 구분	관련 부서	적용조사 (주기)	데이터양 (주/예측)	기초 데이터 이수 건수	데이터 전처리	데이터 정제	데이터 분리	학습/평가 품질 (Accuracy / F-1)	구축횟수 (최종학습일)
한국표준 산업 (10차)	고용 통계과	지역별 고용조사 (연 2회)	약 43만 (6개월 간격)	총 347만 (44~51회 중 8회차 19~22년 자료)	결측: 109만 중복: 32만	122만	학습: 110만 검증: 12만 평가: 43만	검증: 92.74% / 0.9271 평가: 94.17% / 0.9414	4회 (12/14)
	경제 총조사과	전국사업체 조사 (연 1회)	약 800만 (1년 간격)	총 2,594만 (18~21년)	결측: 484만 중복: 553만	834만	학습: 751만 검증: 83만 평가: 723만	검증: 85.04% / 0.8508 평가: 86.48% / 0.8623	9회 (01/11)
	인구 총조사과	인구총조사 (5년)	약 1,000만 (5년 간격 전기구원 통합)	총 981만 (2020 인구총조사 기초데이터)	처리: 571만 중복: 40만 결측: 65개	410만	학습: 369만 검증: 41만	검증: 86.01% / 0.8610 평가: 85.60% / 0.8551	6회 (11/06)
한국표준 직업 (7차)	고용 통계과	지역별 고용조사 (연 2회)	약 43만 (6개월 간격)	총 347만 (44~51회 중 8회차 19~22년 4개년 자료)	결측: 109만 중복: 32만	122만	학습: 110만 검증: 12만 평가: 43만	검증: 91.44% / 0.9141 평가: 92.67% / 0.9259	5회 (12/14)
	인구 총조사과	인구총조사 (5년)	약 1,000만 (5년 간격 전기구원 통합)	총 981만 (2020 인구총조사 기초데이터)	처리: 571만 중복: 40만 결측: 65	410만	학습: 369만 검증: 41만	검증: 79.52% / 0.7915 평가: 79.25% / 0.7896	6회 (11/06)
건설업 공중	산업 통계과	건설업 조사 (연 1회)	평균 180만 (1년 간격)	총 883만 (17~21년 전수)	중복: 250만	633만	학습: 192.1만 검증: 21.3만 평가: 92.1만 (21년, 내림 대상)	검증: 84.22% / 0.8411 평가: 85.03% / 0.8482	8회 (01/04)
건설업 발주자	산업 통계과	건설업 조사 (연 1회)	평균 180만 (1년 간격)	총 883만 (17~21년 전수)	중복: 294만 결측: 5만	589만	학습: 192.6만 검증: 21.4만 평가: 92.2만 (21년, 내림 대상 Y)	검증: 91.81% / 0.9177 평가: 89.32% / 0.8937	6회 (01/04)
물품통관 분류	서비스업 통항과	온라인소 핑 통항 (월 1회)	평균 230만 (1월 간격)	총 9,730만 (21 ~ 22년 2년 자료)	중복: 8,315만 (21' 01.~22' 08.)	상품군: 93만 (22년 04월) HS코드: 1,182만 (21' 01.~22' 08.)	학습: 26만 검증: 2.9만 평가: 9.3만 (상품군, 22년 04 월)	검증: 96.54% / 0.9653 평가: 96.86% / 0.9685	4회 (12/02)

자료: 임경민. (2024). 통계 분야의 AI 도입 사례-AI기반 통계분류 자동화. p.169.

[그림 3-6] 기존 시스템과 인공지능(AI) 통계분류 시스템 성능 비교(분류 예측 정확도)



자료: 임경민. (2024). 통계 분야의 AI 도입 사례-AI기반 통계분류 자동화. p.170.

통계청은 인공지능(AI) 통계분류 예측 결과를 활용하여 내검 효율성을 제고하고, 학습데이터 갱신 및 재학습을 통해 인공지능 통계분류 예측 정확도를 향상시킬 계획이다.

2. 해외 한류 콘텐츠 침해 실태조사

한류 콘텐츠의 인기가 전 세계로 확산되어 온라인을 통해 해외 불법유통 사이트에 한류 콘텐츠 불법복제물이 지속적으로 확산됨에 따라 해외 한류 콘텐츠 침해 현황을 계량적으로 파악하기 위해 2022년부터 매년 ‘해외 한류 콘텐츠 침해 실태조사’를 실시하고 있다(한국저작권보호원, 2023).²⁸⁾

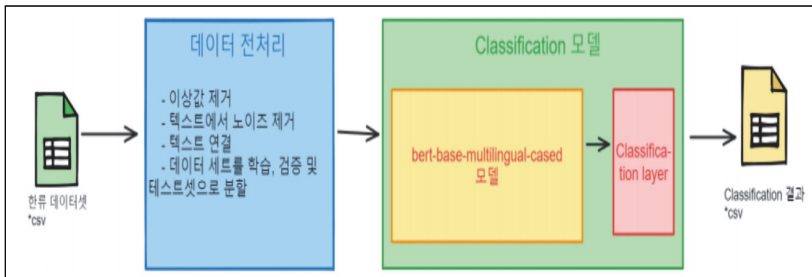
해외에서 유통되는 한류 콘텐츠의 불법복제물 규모를 파악하기 위해 한국저작권보호원에서 보유하고 있는 불법유통 사이트 목록 19,775개 (2023년 기준)를 장르별, 제공 언어별, 트래픽 순위별 등으로 특성을 파악하고, 실태조사 표본 설계를 위해 전체 모집단 사이트 생멸 여부 확인 및 각 사이트 정보를 업데이트하였으며, 최종 정제가 완료된 모집단(리스

28) 한국저작권보호원. (2023). 2023 해외 한류콘텐츠 침해 실태조사-영화·방송·웹툰 분야 불법유통 사이트 중심.

트)을 기준으로 통계전문가를 통한 표본 설계를 진행하였다(한국저작권보호원, 2023).

데이터 수집은 텍스트를 기반으로 하여 게시물 정보를 수집하는 크롤링²⁹⁾ 방식을 사용하였고, 게시물 데이터의 정제는 한류 콘텐츠 분류의 정확도와 속도를 제고하기 위해 조사 수행사인 (주)한국리서치에서 개발한 한류 콘텐츠 AI 분류기 모델로 수행하였다(한국저작권보호원, 2023).

[그림 3-7] 한류 콘텐츠 AI 분류기 모델 개념도

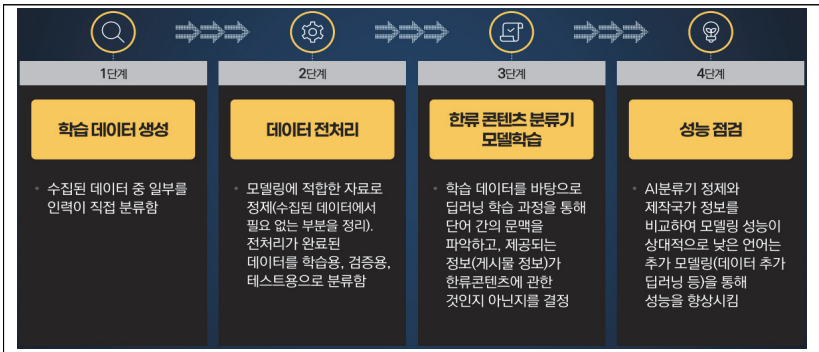


자료: 한국리서치. (2024). 한국리서치 창립 46주년 기념 고객 세미나 발표자료집. p.175.

AI 분류기 모델링 딥러닝을 위한 데이터셋을 구축하기 위해 올해 수집된 웹툰 및 영상 데이터 중 일부를 인력이 정제하여 각 장르별로 한류 콘텐츠와 비한류 콘텐츠 각 2,000개 이상의 데이터셋을 구축하였고, 한류 콘텐츠 AI 분류기는 사전에 분류된 한류 콘텐츠와 비한류 콘텐츠를 바탕으로 딥러닝 학습 과정을 통해 단어 간의 문맥을 파악하고, 제공되는 정보(게시물 정보)가 한류 콘텐츠에 관한 것인지 아닌지를 파악하여 최종적으로 한류 콘텐츠를 분류하였다.

29) 크롤링: 웹 페이지를 그대로 가져와서 데이터를 추출해내는 방식으로 사람이 일일이 해당 사이트의 정보를 검색하는 것이 아니라 컴퓨터 프로그램에 미리 입력된 방식에 따라 정보를 찾아 추출하는 작업을 반복 수행하여 방대한 자료를 단시간에 수집이 가능함(한국저작권보호원, 2023).

[그림 3-8] 한류 콘텐츠 인공지능(AI) 분류기 학습 과정



자료: 한국리서치. (2024). 한국리서치 창립 46주년 기념 고객 세미나 발표자료집. p.177.

한류 콘텐츠 AI 분류기의 장점과 단점을 정리하면, <표 3-10>과 같다.

<표 3-10> 한류 콘텐츠 AI 분류기의 장점과 단점

장점	단점
- (일관성) 인력이 분류할 경우 사람마다 한류를 분류하는 기준이 다를 수 있으며 분석 데이터가 많아질수록 피로도에 따른 오류 발생 가능한 부분을 일관된 기준으로 정제가 가능함 - (다양성) 80여 개 이상의 언어로 된 광범위한 콘텐츠를 분류하기 위해 다국어에 적합한 사전 학습된 머신러닝 모델 BERT multilingual³⁰⁾을 과업에 맞게 파인튜닝(fine-tuning)하였기 때문에 단일 모델로 여러 언어를 동시에 처리할 수 있음 - (고성능) 대규모 데이터셋에서 사전에 충분히 훈련(BERT multilingual)된 모델을 바탕으로 하기 때문에 성능이 보장됨	- (소스 소모) 큰 모델 사이즈와 복잡한 구조로 인해 많은 계산 리소스가 소요됨 - (데이터 불균형) 특정 언어에 대한 데이터가 부족할 경우 성능 저하 발생함 * 다만, 제공 언어별 모델링 성능이 떨어진 경우, 제공 언어별 추가 모델링 업데이트를 통해 성능향상이 가능함

자료: 한국저작권보호원. (2023). 2023 해외 한류콘텐츠 침해 실태조사-영화·방송·웹툰 분야 불법 유통 사이트 중심. p.8. (요약)

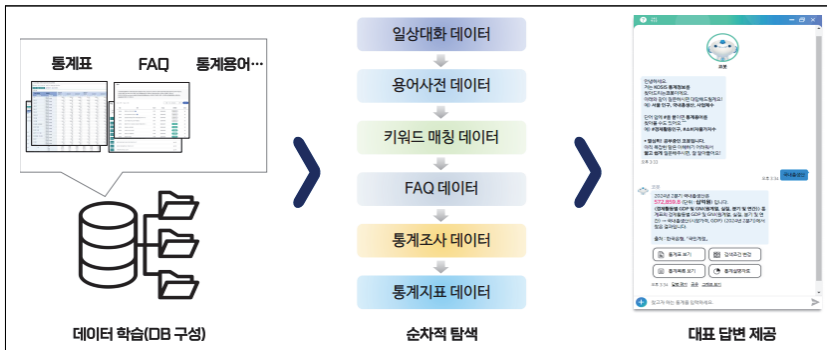
30) BERT(Bidirectional Encoder Representations from Transformers) Multilingual: 여러 언어를 동시에 처리할 수 있는 사전 훈련된 모델임. 이 모델은 100개 이상의 언어에 대한 훈련 데이터를 사용하여 만들어졌으며, 단일 모델로서 다양한 언어에 대한 텍스트 분석 작업을 수행할 수 있음. 다양한 자연어 처리 작업에 대해 fine-tuning이 가능하며, 이를 통해 특정 작업에서 좋은 성능을 얻을 수 있음. 특히, 문맥을 양방향으로 고려하는 덕분에 문맥에 따라 단어의 의미를 더 정확하게 파악할 수 있음. (한국저작권위원회, 2023. p.8에서 발췌)

한류 콘텐츠 인공지능(AI) 분류는 통계청이 아닌 민간기관에서 인공지능(AI)을 활용해 통계업무 프로세스 실무에 성공적으로 적용하였다는 점에서 시사하는 바가 크다.

3. 일상언어로 묻고 답하는 초거대 AI 통계 챗봇

통계청은 2021년 4월 국민이 원하는 통계 정보에 대해 바로 답해주는 인공지능(AI) 플랫폼 서비스인 ‘코봇’ 서비스를 개시하였다. ‘코봇’ 서비스는 방대한 통계 데이터로의 접근 편의성을 개선하기 위해 도입한 규칙 기반 챗봇으로, 통계표, FAQ, 통계용어 등의 학습 데이터로부터 이용자의 질의에 따라 순차적으로 정보에 접근하여 대표 답변을 제시하였다(그림 3-9) 참조).

[그림 3-9] 규칙 기반 챗봇의 서비스 개념도



자료: 신우람. (2024). 일상언어로 묻고 답하는 초거대 AI 통계챗봇. p.195.

하지만 규칙 기반 챗봇인 ‘코봇’은 정확한 통계용어를 사용하는 경우에는 이용자가 원하는 답변을 제공하지만, 자연어로 질의하는 경우에는 질문의 의도를 파악하는 데 실패하여 이용자가 원하는 통계를 제공하지 못

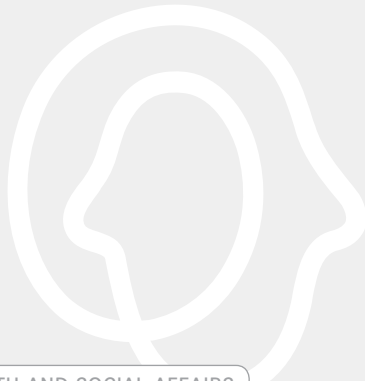
제3절 소결

제3장에서는 주요 외국 및 우리나라의 인공지능(AI)을 활용한 통계작성 사례를 검토하였다.

우선, 주요 외국의 인공지능(AI)을 활용한 통계작성 사례를 살펴보면 다음과 같다. 2018년 독일 연방통계청(Destatis)이 실시한 국가 통계기관 머신러닝(M) 활용 실태조사에서는 많은 국가들이 통계작성을 위해 인공지능(AI)을 활용하려는 시도를 확인할 수 있었고, 통계작성 주제와 활용형태, 인공지능(AI) 방법 등을 확인할 수 있었다. 유엔 유럽경제위원회(UNECE)의 머신러닝(ML) 파일럿 프로젝트에서는 인공지능(AI) 활용에 적합한 ‘코딩 및 분류’, ‘편집 및 대체’, ‘이미지 분석’ 업무에 인공지능(AI)을 활용한 프로젝트 사례를 검토하였다. 또한 유엔 유럽경제위원회(UNECE)의 공식 통계 생산을 위한 대형 언어모델(LLM) 활용 방안에서는 최근 가장 주목받고 있는 대형 언어모델(LLM)을 활용한 국가 및 국제기구의 통계작성 시도를 검토하였다. 주요 외국의 인공지능(AI)을 활용한 통계작성 사례에서 검토한 유엔 유럽경제위원회(UNECE) 보고서 및 프로젝트는 그 사례가 유럽 국가에만 제한된 것이 아니라 미국, 캐나다, 호주, 뉴질랜드, 멕시코 등의 통계작성 사례를 포함하고 있어 주요 국가의 통계작성 사례를 검토할 수 있었다.

우리나라의 인공지능(AI)을 활용한 통계작성 사례를 살펴보면, 가장 체계적으로 연구되고 실무 적용 단계에 이른 분야는 통계분류 자동화이다. 일찍이 통계청을 중심으로 인공지능(AI)을 활용한 통계분류 가능성을 검토하고 정확도를 높이기 위한 지속적인 노력을 통해 5개 조사의 5종 분류가 실무 적용 단계에 이르렀으며, 앞으로는 더 많은 조사에 더 많은 분류에 활용되고, 분류 정확도가 보다 더 정교화될 것으로 기대된다. 또한 해

외의 한류 콘텐츠 침해 실태조사 사례에서는 민간기관의 통계작성을 위한 자료수집 및 통계분류를 위한 인공지능(AI) 활용 시도를 검토하였다. 향후 민간기관에서도 이러한 시도가 더 확대될 것으로 여겨진다. 한편, 통계청에서 개발 중인 초거대 AI 통계 챗봇 사례는 대형 언어모델(LLM)을 활용하여 통계서비스 업무 개선 가능성을 높이는 좋은 사례이며, 앞으로는 우리나라도 더 많은 통계업무에 대형 언어모델(LLM)이 활용될 것으로 여겨진다.



제4장

조사데이터 품질관리 내부 현황 분석

제1절 조사데이터 품질관리

제2절 한국의료패널 조사데이터 품질관리 절차 및 방법

제3절 소결

제4장 조사데이터 품질관리 내부 현황 분석

제1절 조사데이터 품질관리

1. 통계자료, 조사자료, 조사데이터 개념

조사데이터는 학술적·법률적으로 일원화된 정의는 없으나 사용자에게 따라 협의적으로는 통계조사 원자료(raw data)를 의미하는 용어로 사용하기도 하고, 광의적으로는 조사의 전 과정에서 수집된 모든 데이터를 의미하는 용어로 사용하기도 한다. 또한 조사데이터는 그 의미가 명확하게 정의되지 않았기 때문에 사용자에게 따라서는 통계자료, 조사자료 등과 같은 용어와 혼용되어 사용하기도 한다. 이에 통계자료, 조사자료, 조사데이터의 개념을 우선 살펴보고자 한다.

통계자료는 「통계법」 제3조 제2호에서 ‘통계작성기관이 통계의 작성을 위하여 수집·취득 또는 사용한 자료(데이터베이스 등 전산자료를 포함한다)로서 통계등록부를 제외한 것을 말한다’고 정의하고 있다.³¹⁾ 또한 통계청의 「통계승인업무 처리지침」 제2조 각호에서는 통계승인업무에 필요한 용어를 정의하고 있는데, 동 지침에 따른 용어 정의에 따르면, 통계자료는 조사통계, 보고통계, 가공통계 작성을 위해 조사·수집된 자료를 의미한다(통계청, 2023).³²⁾

조사자료도 조사데이터와 마찬가지로 학술적·법률적으로 일원된 정의

31) 통계법, [시행 2024. 3. 26.] [법률 제20408호, 2024. 3. 26., 일부개정]

32) 통계승인업무 처리지침 [시행 2023. 12. 05.] [통계청예규 제265호, 2023.12.05., 일부개정]

는 없다. 다만 한국보건사회연구원에서는 조사자료의 체계적인 관리를 위해 마련된 「조사자료 관리지침」³³⁾ 제2조 제1호에서 ‘조사자료란 연구원의 연구·사업 수행을 목적으로 연구원 내·외부의 면접조사, 전화조사, 우편조사 및 웹조사 등을 통하여 생산된 기계 판독이 가능한 모든 형태의 자료’로 용어를 정의하고 있다. 즉, 동 지침에서는 연구를 위한 각종 조사 과정에서 생성된 조사표, 부호화 지침서, 마이크로데이터, 녹취록 및 녹음 파일 등 모든 형태의 자료를 조사자료로 정의하고 있다(박성준 외, 2023).

앞에서 살펴본 통계자료, 조사자료, 조사데이터의 정의를 토대로 조사데이터는 통계자료의 일부분이며, 조사자료와는 매우 유사한 의미를 갖고 있다.

2. 통계 품질관리

앞에서 정의한 바와 같이 조사데이터는 통계 정보의 일부분이다. 따라서 통계 품질관리 기준과 진단 항목을 살펴보고 조사데이터의 품질관리에 대해 논의하고자 한다.

통계 품질관리란 통계 이용자가 통계를 사용하는 데 적합하도록 생산하는 방법뿐만 아니라 이용자에게 만족을 주면서 가장 경제적인 방법으로 통계를 작성·보급·관리하기 위한 모든 수단을 통합한 체계를 말한다(통계청, 2024).

통계청에서는 통계자료 품질관리를 위해 <표 4-1>과 같이 관련성, 정확성, 시의성/정시성, 비교성/일관성, 접근성/명확성의 5가지로 통계 품질의 차원을 정의하고, 통계 품질 수준이 어느 정도인지를 측정하여 각

33) 조사자료 관리지침. [시행 2024.12.31.]

차원의 품질 수준을 높이기 위해 통계를 어떻게 개선해야 하는지 그 방향을 제시하고 있다(통계청, 2024, pp.10-12).

〈표 4-1〉 통계청의 통계 품질 차원

구분	정의	개선 방향
관련성	- 통계자료가 포괄 범위와 개념, 내용 등에 있어서 이용자의 요구사항을 충족하는 정도를 의미하며, 통계 이용자에게 얼마나 의미 있고 유용한 통계를 작성하여 제공하고 있는가와 관련된 개념임.	- 통계를 작성하는 과정에서 통계의 목적을 명확히 설정하고 이를 달성하기 위하여 이용자 파악, 전문가 자문회의, 이용자 만족도 조사 등 이용자의 요구를 파악하고 반영함.
정확성	- 측정하고자 하는 모집단의 특성이나 크기를 얼마나 근사하게 측정했는가를 의미하며, 대부분의 통계는 알 수 없는 참값을 추정하게 되는데, 정확성은 미지의 참값과 추정된 값과의 근접성에 관한 개념임. * (조사통계) 포괄 범위, 표본추출, 응답 및 무응답, 작성 과정 등에 의해서 오차가 발생할 수 있음. * (가공통계) 투입자료인 다른 표본 조사나 총조사 자료의 오류나 포괄 범위, 조사시기, 평가 방법 등의 불일치 등에 의해 오차가 발생할 수 있음.	- 표본오차의 크기, 비표본오차의 정도 등을 검토함으로써 알 수 있음.
시의성/정시성	- (시의성) 작성 기준 시점과 결과 공표 시점 간의 차이를 나타내는 통계의 현실 반영도와 관련된 개념임 - (정시성) 예고된 공표 시기를 정확히 준수하고 있는가에 대한 개념임.	- (시의성) 작성 기준 시점과 결과 발표 시점이 근접할수록 시의성이 높은 통계임. - (정시성) 통계 이용자들이 통계의 공표 일정을 사전에 알 수 있도록 사전 예고제를 실시하는데 이러한 사전공표 일정을 정확히 준수할수록 정시성이 높은 통계임.
비교성/일관성	- (비교성) 시간 또는 공간이 달라도 통계 자료가 동일한 개념, 분류, 측정도구, 측정 과정 및 기초 자료 등을 기준으로 집계되어 서로 비교가 가능한지를 나타냄. - (일관성) 동일한 경제·사회 현상에 대해 서로 다른 기초 자료나 작성 방법, 작성 주기 (공표 주기)에 의해 작성된 통계자료들이 얼마나	- (비교성) 특정 통계에 대하여 다른 나라, 다른 도시 또는 다른 연도의 자료와 비교가 가능한지를 보는 것으로, 국제 비교성을 높이기 위해서는 국제적인 기준 및 분류, 평가 방법 등의 적용이 필요함. 또한 작성 주기가 부정기 또는 장기인 경우는 담당자 변동, 환경 변화 등으로 과거 조사와 개념, 조사항목, 조사 방법 등이 달라져 시간적 비

구분	정의	개선 방향
	유사한가를 의미함. 예를 들어, 잠정자료와 확정자료, 연간자료와 분기(월) 자료, 각 통계조사와 국민계정은 서로 다른 자료원과 작성 방법에 의해 작성될 수 있으나 서로 유사한 결과를 보인다면 일관성이 높다고 할 수 있음.	교성이 낮아질 수 있으므로 특히 유의하여야 함. - (일관성) 예를 들어 잠정자료와 확정자료, 연간 자료와 분기(월) 자료, 각 통계조사와 국민계정은 서로 다른 자료원과 작성 방법에 의해 작성될 수 있으나 서로 유사한 결과를 보인다면 일관성이 높다고 할 수 있음.
접근성/ 명확성	- (접근성) 이용자가 통계자료에 얼마나 쉽게 접근할 수 있는지에 대한 정도를 의미함. - (명확성) 통계가 어떻게 만들어졌는지에 대한 정보제공 수준을 의미함.	- (접근성) 통계자료의 DB화, 간행물 및 보도자료 홈페이지 게시, 모바일 메시지를 통한 속보 전송 등 다양한 방법으로 통계자료를 제공하고, 자료를 쉽게 찾을 수 있도록 검색 기능을 제공함. - (명확성) 다양한 매체를 통해 제공되는 통계에 대한 이해를 돕기 위해 통계를 작성하는 과정, 자료 이용 방법, 마이크로 데이터 이용 방법, 통계 이용상의 유의사항 등에 대한 충분한 설명자료(메타데이터)를 제공함.

자료: 통계청. (2024). 2024년 정기통계 품질진단 매뉴얼. p.11-12.(요약)

통계 품질 진단은 통계자료가 정확한지, 시의성은 있는지, 유용한 통계인지, 이용자들이 쉽게 접근할 수 있는지, 이용자들이 쉽게 분석하고 활용할 수 있는지를 가늠해 보기 위해 통계작성 전 과정을 진단하는 업무이다(통계청, 2024).

통계청 조사통계의 통계 품질 진단 항목을 통계 품질과 작성 절차(통계 작성 기획, 통계설계, 자료수집, 통계처리 및 분석, 통계공표, 관리 및 이용자 서비스, 통계 기반 및 개선)에 따라 분류하면 <표 4-2>와 같다.

〈표 4-2〉 통계청의 통계 품질 진단 항목

구분	통계 진단 항목	관련성	정확성	시의성/ 정시성	비교성/ 일관성	접근성/ 명확성
통계 작성 기획	1. 조사방법	●				
	2. 조사 및 공표 주기	●				
	3. 조사 일정 및 일정별 수행업무 제시	●				
	4. 통계작성 문서화	●				
	5. 통계 연혁	●				
	6. 통계의 작성 목적	●				
	7. 유형별 자료 이용자 관리	●				
	8. 이용자 의견 수렴	●				
통계 설계	1. 조사항목			●		
	2. 적용 분류체계			●		
	3. 조사표 구성		●			
	4. 조사표 설계 및 변경 절차나 방법의 적절성	●				
	5. 조사표 변경 이력	●				
	6. 목표 모집단과 조사 모집단		●			
	7. 조사 모집단(전수조사) 또는 표본추출틀(표본조사)		●			
	8. 표본 설계 방법 및 결과		●			
	9. 표본관리		●			
자료 수집	1. 조사방법		●			
	2. 조사원 채용 및 처우		●			
	3. 조사원 교육훈련		●			
	4. 조사원 업무량		●			
	5. 조사업무 흐름도		●			
	6. 조사준비 및 준비조사		●			
	7. 조사항목별 조사 방법		●			
	8. 조사 관리		●			
	9. 조사 질의응답 체계		●			
	10. 조사(또는 응답) 대상		●			
	11. 무응답 대처		●			
	12. 표본 대체		●			
	13. 사후조사		●			
	14. 행정자료 활용 목적 및 내용	●				
	15. 활용 행정자료의 특성 및 입수체계		●			
통계 처리 및 분석	1. 자료코딩		●			
	2. 자료입력		●			
	3. 자료 내검		●			
	4. 주요 항목무응답 실태		●			
	5. 항목무응답 대체		●			
	6. 단위무응답 실태		●			
	7. 가중치 조정		●			
	8. 통계추정 산출식 및 내용		●			
	9. 표본오차 추정 방법 및 결과(표본조사)		●			

구분	통계 진단 항목	관련성	정확성	시의성/ 정시성	비교성/ 일관성	접근성/ 명확성
	10. 지수 유형 및 산출식		●			
	11. 지수 가중치 및 갱신		●			
	12. 지수 개편		●			
	13. 디스플레이터		●			
	14. 계절 조정	●				
	15. 행정자료의 매칭 방법		●			
통계공 표, 관리 및 이용자 서비스	1. 공표통계 해석 방법	●				
	2. 공표통계 일치성		●			
	3. 조사 대상 기간/조사 기준시점과 공표시기			●		
	4. 공표일정			●		
	5. 통계 작성 방법의 비교성				●	
	6. 시계열 비교성				●	
	7. 국제 간 비교성				●	
	8. 동일 영역 통계와 일관성				●	
	9. 잠정치와 확정치의 일관성				●	
	10. 통계의 이용자 서비스					●
	11. 통계 설명자료 제공					●
	12. 마이크로데이터 생성·관리		●			●
	13. 마이크로데이터 서비스					
	14. 마이크로데이터 일치율		●			
	15. 자료 수집, 처리 및 보관 과정의 비밀보호	●				
	16. 공표자료의 비밀보호	●				
	17. 자료 보안 및 접근제한	●				
통계 기반 및 개선	1. 기획 및 분석 인력, 사업예산		●			
	2. 통계위탁 조사		●			
	3. 통계 품질관리 및 개선		●			

자료: 통계청. (2024). 2024년 정기통계품질진단 매뉴얼. p.53-54.(정리)

한국보건사회연구원의 조사데이터는 공공데이터이다. ‘공공데이터 품질관리 매뉴얼’에서는 <표 4-3>과 같이 공공기관의 데이터 품질을 측정하는 기준으로 준비성, 완전성, 일관성, 정확성, 보안성, 적시성, 유용성의 7개 지표와 지표별 세부 특성을 반영한 24개 세부 지표를 제시하고 있다(한국지능정보사회진흥원, 2018).³⁴⁾

34) 한국지능정보사회진흥원. (2018). 공공데이터 품질관리 매뉴얼.

〈표 4-3〉 공공데이터 품질 지표

구분	지표 정의	세부 지표
준비성	- 공공데이터의 품질관리를 위해 기본적으로 관리해야 하는 정책, 규정, 조직, 절차 등을 마련하고, 최신의 내용으로 충실하게 관리되는지를 측정하는 지표	- 관리지표 - 내용 충실
완전성	- 공공데이터의 저장소인 데이터베이스를 구축함에 있어 논리적인 설계와 물리적인 구조를 갖추고, 업무요건에 맞게 데이터가 저장되는지를 측정하는 지표	- 논리모델 - 식별자 - 물리구조 - 속성 의미
일관성	- 같은 의미를 갖는 데이터는 논리적 속성 단위, 물리적 컬럼 단위에서 일관된 이름과 형식을 갖도록 표준을 준수하고 있는지, 공공 데이터의 공동 활용을 위해 공유·연계하는 데이터는 누락이 없이 상호 간의 일관성을 유지하는지를 측정하는 지표	- 속성 - 표준 - 중복값 - 연계값
정확성	- 정확한 데이터 제공을 위해 데이터의 입력 단계부터 오류가 입력되지 않도록 하고, 저장된 데이터가 정의된 기준에 맞게 유효한 값의 범위와 형식으로 되어 있는지, 저장된 데이터가 현실에 가장 가까운 최신 값을 반영하고 있는지를 측정하는 지표	- 입력값 - 업무규칙 - 범위·형식 - 참조 관계 - 계산식
보안성	- 지속적인 품질 확보를 위해 운영되는 데이터의 관리 주체가 관리하고 있는지, 권한에 따른 데이터 접근이 적절히 통제되고 개인정보 등 중요 데이터에 대해서는 암호화 등 보안 조치가 이루어져 있는지를 측정하는 지표	- 오픈쉽 - 접근 제한 - DB 보호
적시성	- 사용자가 만족하는 수준의 응답시간이 확보되고 있는지, 사용자의 데이터 요구에 따른 수집·처리·제공까지의 절차가 체계적으로 관리되고 있는지를 측정하는 지표	- 응답시간 - 데이터 제공 - 최신값
유용성	- 사용자가 만족하는 수준의 충분한 정보가 제공되고 있는지, 정보 접근 시 사용자의 편의성이 확보되고 있는지, 사용자의 정보 이용에 따른 만족 수준을 높이도록 노력하고 있는지를 측정하는 지표	- 충분 - 접근 - 활용

자료: 한국지능정보사회진흥원. (2018). 공공데이터 품질관리 매뉴얼. p.78.

공공데이터의 품질 진단 방법은 품질 지표들을 어떤 방법으로 측정할 것인가에 대한 기준이 되는 진단 방법으로 〈표 4-4〉와 같이 프로파일링, 체크리스트, 업무규칙 진단, 비정형 실측의 4가지 방법을 제시하고 있다.

〈표 4-4〉 공공데이터 품질 진단 방법

품질 진단 방법		방법 설명
프로파 일링	값 진단	- 데이터 값의 유효성, 정확성 등 데이터 값 자체의 오류를 분석하는 방법 * 컬럼 분석, 날짜 분석, 패턴 분석, 코드 분석 등을 통해 데이터 값의 정확성을 중심으로 진단
	구조 진단	- 논리적 데이터 구조의 오류로 인한 일관성, 정합성 등을 확보하지 못하는 결함을 분석하고 진단하는 방법 * 표준화 수준(코드, 도메인 등), 테이블 구조, 정규화 수준, 컬럼 및 관계 정의 등 데이터의 구조적 결함 측정
체크리스트 (인터뷰·설문)		- 전반적인 데이터 품질관리 수준과 지표별 데이터 품질 수준을 체크리스트(설문 또는 인터뷰)를 통해 진단하는 방법
업무규칙 진단		- 법, 규정에 정의된 업무 기준(산출식)에 근거하여 데이터가 관리되고 있는지를 진단하는 방법 * 업무규칙(Business Rule, BR)을 준수하고 있는지에 관한 측정 스크립트(SQL 등)를 실행하여 오류 값을 추출
비정형 실측		- 문서, 이미지, 동영상 등 정형화되어 있지 않는 정보를 사람이 직접 확인(실측)을 통하여 오류 여부를 진단하는 방법 * 별도 도구 없이 직접 정보를 조회하거나 해당 문서를 수기로 확인 등

자료: 한국지능정보사회진흥원. (2018). 공공데이터 품질관리 매뉴얼. p.79.

공공데이터의 품질 지표별 진단 방법은 <표 3-13>과 같다.

<표 4-5> 공공데이터의 품질 지표별 진단 방법

지표		품질 진단 방법		
		프로파일링 (값·구조 진단)	업무규칙 진단	체크리스트 (인터뷰·설문)
준비성	관리요소			●
	내용 충실			●
완전성	논리모델			●
	식별자	●	●	●
	물리구조	●	●	●
	속성 의미	●		●
	속성	●	●	●
일관성	표준	●	●	●
	중복값		●	●
	연계값		●	●
	입력값	●	●	●
정확성	업무규칙		●	●
	범위·형식	●	●	●
	참조 관계	●	●	●
	계산식		●	●
	오너쉽			●
보안성	접근 제한			●
	DB 보호			●
적시성	응답시간			●
	데이터 제공			●
	최신값			●
유용성	충분			●
	접근			●
	활용			●

자료: 한국지능정보사회진흥원. (2018). 공공데이터 품질관리 매뉴얼. p.80.

4. 조사데이터의 품질관리

조사데이터는 통계자료이며, 공공데이터이다. 광의적 의미의 조사데이터 품질관리는 조사데이터의 수집, 처리 및 분석, 저장 및 관리, 공개 등 데이터의 생성부터 보관 및 공개에 이르는 전 과정에서 데이터의 품질 수준을 높이기 위한 업무이다. 다만 본 보고서에서 다루는 조사데이터 품질관리의 범위는 자료 수집 및 처리 과정에서 발생할 수 있는 데이터 오류를 탐지하고 정확성을 개선하는 자료 내검 업무로 정의하고자 하며, 그 이유는 다음과 같다.

첫째, 조사데이터는 연구의 기반이 되는 데이터로서, 데이터의 오류 및 이상치를 제거하는 자료의 내검 과정이 조사데이터의 품질관리에 가장 많은 영향을 미치기 때문이다. 조사데이터의 오류로 인해 문제가 생기면 연구 결과의 신뢰성이 낮아지고, 이로 인해 부정확한 정책 결정이나 잘못된 연구 결과가 도출될 위험이 있다. 특히, 대규모의 조사데이터를 다루는 경우 조사데이터에서 오류가 쉽게 발생할 수 있으며, 이러한 오류가 수정되지 않으면 연구의 타당성을 저해할 수 있다. 또한 조사데이터 배포(외부 공개) 시 조사데이터의 오류는 연구 결과의 신뢰성에 악영향을 미칠 수 있다.

둘째, 인공지능(AI)을 활용한 연구에서 자료의 내검은 조사데이터의 품질관리에 가장 적합한 업무이기 때문이다. 선행연구 및 사례조사를 통해 인공지능(AI) 활용에 가장 적합한 영역은 ‘코딩 및 분류’, ‘편집 및 대체’ 영역으로 파악되었으며, 한국의료패널 조사데이터 품질관리 중 반복적으로 발생하는 가장 큰 문제점이 자료의 내검 과정에서 발생하였기 때문이다.

제2절 한국의료패널 조사데이터 품질관리 절차 및 방법

1. 한국의료패널의 개요

한국의료패널은 「통계법」 제18조(통계작성의 승인)에 따라 작성된 국가승인통계(제920012호)이며, 국가 보건의료체계의 대응성·접근성 향상과 효율화를 위한 정책 수행의 기초 정보로서 질환, 의료이용, 의약품 복용, 의료비 지출 및 지출원, 건강 관련 인식 및 행태 등에 관한 개인 및 가구 단위의 통계를 생산하고 관련 요인을 규명하고자 구축되었다(이수형 외, 2024).

한국의료패널조사는 2007년에 ‘한국의료패널 구축을 위한 기초 연구’를 시작으로, 2008년에는 한국의료패널 예비조사 및 가구 유치조사를 수행하였다. 2009년에는 의료이용 행태와 의료비 지출 규모에 관한 정보뿐만 아니라 의료이용 및 의료비 지출에 영향을 미치는 요인들을 포괄하여 심층적으로 분석할 수 있는 패널 데이터를 구축하는 데 주요 목적을 두고 제1기 한국의료패널 조사를 시작하였다(한국보건사회연구원, 2023).³⁵⁾ 한편, 제1기 한국의료패널은 12년간의 조사로 조사 가구의 피로도가 누적되고, 패널 마모가 발생함에 따라 2014년부터 개편 방향을 논의하기 시작하였다. 2016년, 2018년에 제2기 한국의료패널 구축 방안 연구를 수행하였고, 2018년에 대표성과 경시성의 동시 확보를 위해 (기간 확정) 고정 패널 방식으로 제2기 한국의료패널을 개편하였다(한국보건사회연구원, 2023).

한국의료패널의 조사 내용은 가구 조사와 가구원 조사로 구분할 수 있다. 가구 조사는 가구 및 가구원 일반사항, 경제활동 상태, 소득(가구원,

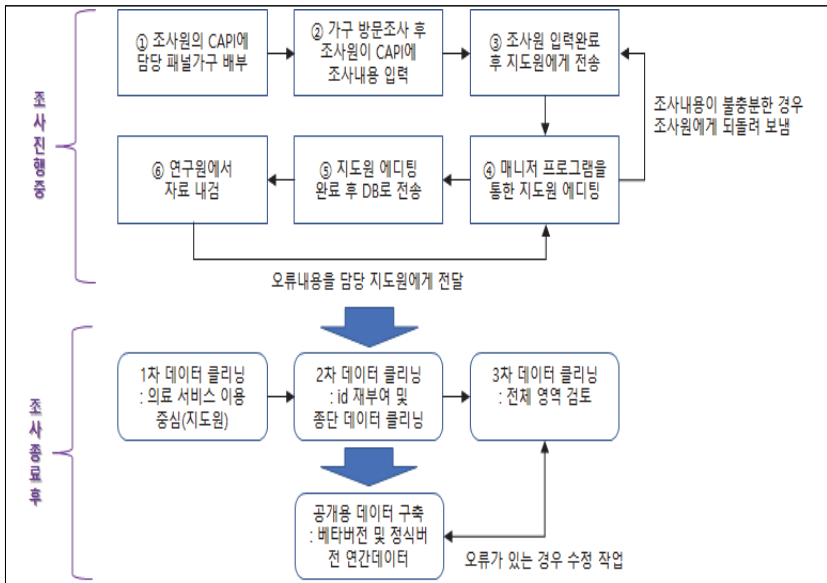
35) 한국보건사회연구원. (2023). 한국의료패널 통계정보보고서.

가구), 부채 및 자산, 지출, 만성질환, 일반의약품 및 의료 관련 지출, 의료서비스 이용(응급/입원/외래/건강검진), 장기요양 및 기타 돌봄, 민간 의료보험 영역의 조사항목으로 구성되어 있으며, 가구원 조사는 건강 수준, 건강생활 습관, 의료 접근성, 상용치료원, 민간의료보험 영역의 조사항목으로 구성되어 있다(한국보건사회연구원, 2023).

2. 한국의료패널 조사데이터 품질관리 절차 및 방법

한국의료패널의 조사데이터를 수집하는 과정에서 수행되는 조사데이터의 품질관리는 [그림 4-1]과 같이 조사 진행 중과 조사 종료 후로 구분할 수 있다.

[그림 4-1] 한국의료패널 조사데이터 수집 과정



자료: 내부 자료

‘조사 진행 중’의 과정을 살펴보면 다음과 같다. ① 조사원의 CAPI 시스템으로 담당 패널가구 기초데이터가 배부된다. ② 가구를 방문조사한 후에 조사원이 CAPI 시스템에 조사 내용을 입력하는데, 이때 CAPI 시스템에 입력 범위 및 설문 로직 등을 설정해 놓은 조사항목은 조사 내용이 사전에 설정된 입력 값의 범위를 벗어났거나 설문 로직에 위배되는 경우에 오류가 발생하여 경고 메시지를 띄우고 입력이 불가하게 된다. 다만, 입력 범위 및 설문 로직은 조사의 특성을 고려하여 모든 조사항목에 설정되는 것은 아니다. ③ 조사원이 입력 완료 후에 조사데이터를 지도원에게 전송한다. ④ 조사팀별로 데이터 내검을 전문으로 하는 숙련된 지도원은 조사원이 입력한 CAPI 내용을 열람하여 조사 내용이 불충분하거나 오류가 발견된 경우, 조사원에게 수정 지시 혹은 직접 응답자에게 보충 조사를 실시한다. ⑤ 지도원이 데이터 내검 완료 후에 DB로 전송하면, ⑥ 연구원에서는 데이터클리닝 담당자가 조사지도원이 데이터 내검을 완료한 자료를 데이터로 추출, 전산 내검을 실시하고, 조사팀의 지도원에게 오류 사항을 전달하여 조사팀의 보충조사 등을 유도하는 식으로 응답 내용의 정합성을 확보하고 있다. 조사 영역별로 데이터클리닝 담당자가 데이터상의 모든 변수를 전산 내검하게 되며, 조사 진행 중에 통상 5회 이상의 전산 내검을 실시하게 된다.

‘조사 종료 후’의 과정을 살펴보면, 1차 데이터클리닝은 지도원이 의뢰 서비스 이용을 중심으로 수행하며, 2차 데이터클리닝은 ID 재부여 및 종단 데이터클리닝을 중심으로, 3차 데이터클리닝은 전체 영역 검토로 연구원 데이터클리닝 담당자가 수행한다. 조사 진행 중과 마찬가지로 조사팀에게 오류 사항을 전달하지만 자료의 오류는 데이터상에서 수정을 하고, 종단 데이터의 특성상 시간 불변성 변수의 경우 해당 패널의 지난 차수 데이터를 참고하기도 한다.

앞에서 설명한 바와 같이 한국의료패널의 조사데이터 품질관리를 위한 데이터 내검은 CAPI 시스템, 조사원, 지도원, 연구원의 데이터클리닝 담당자에 의해 수행된다. 특히 데이터 내검 과정 중 가장 중요한 역할을 수행하는 것은 한국의료패널 조사데이터를 최종적으로 내검하는 연구원의 데이터클리닝 담당자이다. 연구원의 데이터클리닝 담당자는 기본적인 범위의 내검이나 설문 로직 내검뿐 아니라 CAPI 시스템에 구현하기 복잡하고 조사원 또는 지도원이 발견하기 어려운 연속형 변수의 이상치 탐색이나 다년도 검토가 필요한 데이터의 내검을 수행해야 하며, 최종적으로는 연구를 위한 연구용 데이터셋과 외부 공개를 위한 배포용 데이터셋을 생산하여 제공해야 하기 때문이다.

한국의료패널 조사데이터의 경우, 연구원의 데이터클리닝 담당자가 수행하는 자료 내검은 연속형 변수 60개와 다년도 검토가 필요한 변수 3개, 총 63개가 있다(〈표 4-6〉 참조).

〈표 4-6〉 한국의료패널 조사데이터 내검 목록(이상치 탐지 목록)

구분	설명	비고
소득1	- [근로소득] 이상치 재확인 * 고소득 근로소득자의 경우 “경제활동 상태”의 [주된 경제활동 상태]와 [직업분류] 재확인 * 육아휴직인 경우 근로소득 0일 수 있음	- 이상치 탐지
	- “가구원 일반사항 2”의 [의료보장 형태]가 ‘⑤ 의료급여(1, 2종) 세대주’ 또는 ‘⑥ 의료급여 세대원’일 때 [근로소득] 재확인 연 근로소득이 2,000만 원을 초과하거나 50만 원 미만인 경우	- 이상치 탐지
	- [사업소득] 이상치 재확인 * 상위 1% 등 고소득 사업소득자의 경우 “경제활동 상태”의 [주된 경제활동 상태]와 [직업분류]를 재확인 * [사업소득]은 마이너스(-)가 있을 수 있으므로 모름(-9)과 잘 구분하여야 함	- 이상치 탐지
	- “가구원 일반사항 2”의 [의료보장 형태]가 ‘⑤ 의료급여(1, 2종) 세대주’ 또는 ‘⑥ 의료급여 세대원’일 때 [사업소득] 재확인 * 연 사업소득이 1,000만 원 이상인 경우	- 이상치 탐지
소득2	- 작년(이전) 소득과 비교 재확인 * 별도로 예러 체크는 하지 않음	- 다년도 검토
	- [재산소득_금융소득] 이상치 재확인 * 연 금융소득이 0원 초과 10만 원 미만인 경우 또는 1억 이상인 경우	- 이상치 탐지

구분	설명	비고
	- [재산소득_부동산/동산 소득] 이상치 재확인 * 연 부동산/동산 소득이 0원 초과 10만 원 미만인 경우 또는 2억 5천만 원 초과인 경우	- 이상치 탐지
	- [재산소득_기타소득] 이상치 재확인 * 연 기타소득이 0원 초과 10만 원 미만인 경우, 또는 1000만 원 초과인 경우	- 이상치 탐지
	- [공적 이전 소득_정부 및 지방자치단체] 이상치 재확인	- 이상치 탐지
	- [공적연금]이 '① 유'인 경우, 가구원 구성원의 만 나이 재확인 * ex) 공적연금을 받는 데 가구 구성원의 나이가 모두 25세 미만이면 오류 * 공적연금(국민연금, 공무원·군인연금, 사학연금, 보훈연금, 별정직우체국연금 등)	- 이상치 탐지
	- [기초연금]이 '① 유'인 경우, 만 나이가 65세 이상인 가구원이 없으면 오류 * 기준연도, 가구 내 수급자 수에 따라 금액이 상이함 * 2023년 기준 최대금액: 노인 1인 388만 원, 2인 776만 원, 부부노인 621만 원, 3인 1163만 원	- 이상치 탐지
	- [건강보험] 소득 이상치 재확인	- 이상치 탐지
	- [사적이전] 소득 이상치 재확인	- 이상치 탐지
부채 및 자산	- [부채] 이상치 재확인	- 이상치 탐지
	- [자산] 이상치 재확인	- 이상치 탐지
지출	- [지출_소비지출과 지출_비소비지출의 합이 월평균 소득보다 큰 경우 재확인 * 분가하여 신규로 조사된 가구는 소득합계가 적어서 비교값이 차이날 수 있음	- 이상치 탐지
일반 의약품 및 의료 관련 지출	- [일반의약품 및 의약품] 이상치 재확인	- 이상치 탐지
	- [한약 및 첨약] 이상치 재확인	- 이상치 탐지
	- [건강보조식품_복용] 이상치 재확인	- 이상치 탐지
	- [건강보조식품_선물] 이상치 재확인	- 이상치 탐지
	- [시력 보조용품] 이상치 재확인	- 이상치 탐지
	- [청력 보조용품] 이상치 재확인	- 이상치 탐지
	- [의료기기 및 의료용품] 이상치 재확인	- 이상치 탐지
	- [일반의약품 및 의료 관련 지출]의 총합이 [소비지출]보다 큰 경우 오류	- 이상치 탐지
만성질환 이환	- [전체 질환별 만성질환 유무] 개인의 만성질환 이환 변화 비교 재확인	- 다년도 검토
의료 서비스 이용(공통)	- 교통수단이 '③ 대중교통'인데 [교통비]가 5,000원 초과, 250원 미만이면 재확인 * 7세 이하, 65세 이상, 장애가 있는 경우 제외	- 이상치 탐지
	- 교통수단이 '④ 택시'인데 [교통비]가 10만 원 초과, 1800원 미만이면 재확인	- 이상치

114 인공지능(AI) 기반 조사데이터 품질 제고 방안 연구

구분	설명	비고
	* 장애가 있는 경우 제외	탐지
	- 교통수단이 '⑤ 기차/비행기/고속(시외)버스'인데 [교통비]가 20만 원 초과, 1,800원 미만이면 재확인	- 이상치 탐지
	* 7세 이하 제외	
	- 교통수단이 '⑦ 병원구급차' 또는 '⑧ 그 외 구급차'인데 [교통비]가 20만 원 초과, 3만 원 미만이면 재확인	- 이상치 탐지
	- [의료비_수납 금액] 이상치 재확인(ex. 응급, 입원 5천 원 미만)	- 이상치 탐지
	- [의과 입원 치료 내역]이 '① 수술 및 시술'인데 [의료비_수납 금액]이 10만 원 미만인 경우 재확인	- 이상치 탐지
	- [의료비_법정본인부담금] 1,000만 원 초과, 100원 미만이면 재확인 의료급여 환자 제외	- 이상치 탐지
	- [의료비_비급여 금액] 1,000만 원 초과, 100원 미만이면 재확인	- 이상치 탐지
	- [의료비_합인, 절삭 등 금액]이 마이너스인 경우 영수증 재확인(ex. 1천 원 이상)	- 이상치 탐지
	- [의료비_합인, 절삭 등 금액]이 큰 경우 영수증 재확인(ex. 15만 원 이상) * 의료급여 환자, 자동차보험 환자, 산재보험 환자 제외	- 이상치 탐지
의료 서비스 이용(응급)	- [의료비_보험자부담금] 8,000만 원 초과, 1천 원 미만이면 재확인	- 이상치 탐지
	* 의료급여 환자 제외	
의료 서비스 이용(입원)	- [약제비] 500원 미만이면 재확인	- 이상치 탐지
	* 의료급여 환자, 65세 이상 제외	
의료 서비스 이용(응급)	- [소요시간] 이상치 재확인(8시간 초과, 10분 미만)	- 이상치 탐지
	* 본인이 해당 병원에 근무하는 경우가 있을 수 있음	
의료 서비스 이용(외래)	- [유급간병 비용]이 30만 원 이상 만 원 미만(95 제외) 재확인	- 이상치 탐지
의료 서비스 이용(외래)	- [외래 관련 주질환]과 [연령(해당 연도-생년)] 관계 재확인 * 10세 미만의 협심증, 심근경색증	- 이상치 탐지
	- [외래 관련 주질환]과 [연령(해당 연도-생년)] 관계 재확인 * 15세 미만의 자궁경부암, 임신 및 출산 관련	- 이상치 탐지
	- [외래 관련 주질환]과 [연령(해당 연도-생년)] 관계 재확인 * 20세 미만의 알코올성 간질환, 통풍, 등통증, 유방암(등통증은 나오기도 함)	- 이상치 탐지
	- [외래 관련 주질환]과 [연령(해당 연도-생년)] 관계 재확인 * 30세 미만의 치매, 백내장 재확인	- 이상치 탐지
	- [외래 관련 주질환]과 [연령(해당 연도-생년)] 관계 재확인 * 40세 미만의 무릎골관절염, 무릎 외 골관절염	- 이상치 탐지
장기요양 및 기타 돌봄 서비스 이용	- [(방문/보호)재가서비스, 유급 간병인, 가사도우미, 기타 돌봄(비용)] 이상치 재확인(300만 원 이상)	- 이상치 탐지
	- [장기요양시설, 양로원 등 공동주거시설, 기타 시설(비용)] 이상치 재확인(200만 원 이상)	- 이상치 탐지
	- [산후조리원 비용] 이상치 재확인(10만 원 미만, 500만 원 이상)	- 이상치 탐지
	- 일 평균 [산후조리원 비용](비용(만 원)/이용기간(일)) 이상치 재확인(10만 미만, 50만 원 이상)	- 이상치 탐지
	- [산후도우미 비용] 이상치 재확인(10만 원 미만, 500만 원 이상)	- 이상치 탐지
	- 일 평균 [산후도우미 비용](비용(만 원)/이용기간(일)) 이상치 재확인(10	- 이상치

구분	설명	비고
	만 원 미만, 500만 원 이상)	탐지
민간의료 보험	- [보험료] 이상치 재확인(모르는 경우 -9) * [보험료]가 '0원' 인 경우 예러 → [보험료 납부 여부]가 관계없이 계약 상 월 보험료를 작성 * 보험료가 '1~10' 값인 경우 예러 → '원' 단위인데 '만 원' 단위로 단위 착각	- 이상치 탐지
	- [보장 유형]이 '⑦ 종신/연금/변액/저축보험'인 경우 이상치 확인 * 지침: "⑦ 종신/연금/변액/저축보험"이 포함되는 경우, 주계약(종신/연 금/변액/저축보험)을 제외한 의료특약 보험료만을 기입함	- 이상치 탐지
	- [보험금 수령액] 재확인 * 보험금 수령액이 '0~100' 값인 경우 예러 → '원' 단위인데 '만 원' 단 위로 단위 착각	- 이상치 탐지
건강생활 습관	- [신체활동별 비용] 이상치 재확인	- 이상치 탐지
	- [걷기-최근 일주일 하루 걸은 시간] 이상치 재확인 * 응답값이 명확하게 입력된 값인 경우, 오타인지 재확인(보통 응답할 때 5단위 혹은 10단위로 응답하므로, 11분, 17분, 23분.... 등의 경우 재확인함)	- 이상치 탐지
	- [(필련형) 전자담배 사용 비용] 이상치 재확인	- 이상치 탐지
	- [(니코틴 포함 액상) 전자담배 사용 비용] 이상치 재확인	- 이상치 탐지
	- [금연 방법별 비용] 이상치 재확인(100만 원 초과)	- 이상치 탐지
	- [금주 시작 시기(년/월)] 이상치 재확인	- 이상치 탐지
	- [신장], [몸무게] 이상치 재확인(연령과 연결해서 확인)	- 이상치 탐지
건강 수준	- 지난 차수와 비교하여 [신장], [몸무게] 불일치 재확인 * 키 5cm 이상, 몸무게 15kg 이상 차이날 때 재확인 요청	- 다년도 검토
	- [와병 여부]가 '① 예'일 때 [와병일수] 이상치 재확인 * [와병일수]가 365일 이상인 경우 확인 * [와병일수] 전체 빈도 확인 * [와병일수] 큰 값 확인(300일 이상)	- 이상치 탐지
	- [결근결석 여부]가 '① 예'일 때 [결근결석일수] 이상치 재확인 * [결근결석일수]가 365일 이상인 경우 확인 * [결근결석일수] 전체 빈도 확인 * [결근결석일수] 큰 값 확인(ex. 200일 이상: 2021년 법정공휴일 64 일, 총휴일 113일)	- 이상치 탐지
	- [프리젠티즘 여부]가 '① 예'일 때 [프리젠티즘 일수] 이상치 재확인 * [결근결석일수]가 365일 이상인 경우 확인 * [결근결석일수] 전체 빈도 확인 * [결근결석일수] 큰 값 확인(ex. 200일 이상: 2021년 법정공휴일 64 일, 총휴일 113일)	- 이상치 탐지

자료: 저자 작성

3. 한국의료패널 조사데이터의 품질관리 이슈

데이터클리닝 담당자 인터뷰를 통해 현행 한국의료패널 조사데이터의 품질관리 문제점 및 이슈를 파악하였다.

첫째, 변수 간 논리적 관계 오류의 문제점이다. 유관 변수 간 논리적 불일치가 발생할 수 있으며, 이를 최소화할 수 있는 로직 개발이 필요하다. 또한 이러한 로직을 한국의료패널의 CAPI 시스템에 구현한다면 조사 진행 중에 발생할 수 있는 입력 오류를 최소화하고 조사원이나 지도원에 의한 자료 내검을 개선할 수 있을 것이다. 다만, 유관 변수 간 논리적 불일치가 발생하는 오류를 개선하기 위한 로직 개발과 이를 CAPI 시스템에 구현하는 작업은 이미 추진하고 있으며 상당 부분 반영된 것으로 파악되었다.

둘째, 연속형 변수의 이상치 탐지 문제점이다. 데이터 입력 시 항목을 선택하는 명목형 변수의 경우 데이터 이상치가 발생할 가능성이 없으나 연속형 변수의 데이터 입력 시 데이터 이상치가 빈번하게 발생할 수 있어, 연속형 변수의 이상치 탐지를 위한 개선이 필요하다. 특히 복수 개의 변수에 대한 여러 조건들을 확인해야 하는 연속형 변수의 이상치 탐지는 데이터클리닝 담당자의 자료 내검 업무에 큰 부담이 되고 있다. 또한 데이터클리닝 담당자는 연속형 변수의 이상치 탐지에 있어 현재는 통계 패키지(SAS)를 활용하는 기술통계 분석 방법을 사용하고 있으나 보다 직관적으로 이상치를 판단할 수 있는 시각화된 분석 방법이 필요하다.

셋째, 의료서비스 이용 데이터(이용금액) 입력 오류의 문제점이다. 영수증 등 의료서비스 이용 기록을 토대로 데이터를 입력할 때 데이터(이용금액) 입력 오류가 가장 많이 발생하고 있다. 만약 영수증 등에 표기된 의료서비스 이용 데이터를 자동으로 입력할 수 있다면 수기 입력에 따른 입

력 오류를 크게 개선할 수 있을 것이다. 따라서 영수증 등 의료서비스 이용 이미지 분석을 통해 의료서비스 이용 데이터를 자동으로 입력할 수 있는 방안이 필요하다.

넷째, 질병코드 분류의 문제점이다. 질병코드 분류는 의무기록사 등 전문인력이 텍스트로 작성한 의료서비스 이용 내용을 분석하여 질병코드를 분류하는 것으로, 많은 인력과 시간이 소모된다. 따라서 텍스트로 작성한 의료서비스 이용 내용을 분석하여 질병코드를 자동 분류할 수 있는 개선 방안이 마련되어야 할 것이다. 다만 질병코드 분류는 한국의료패널 1기 이후 분류하고 있지 않은 비공개 변수이다.

다섯째, 이상치 탐지를 위한 새로운 점검 기준 개발의 문제점이다. 현재 한국의료패널은 63개의 점검 기준을 마련해 이상치 탐지를 수행하고 있다. 데이터클리닝 담당자는 한국의료패널 조사데이터 품질관리를 위해 신규 변수뿐만 아니라 기존 변수의 이상치 탐지를 위한 새로운 점검 기준을 개발하고자 노력하고 있다. 하지만 새로운 점검 기준 개발은 오랫동안 한국의료패널 조사데이터를 다루었던 데이터클리닝 담당자에게도 매우 어려운 작업이며, 도전이다. 따라서 한국의료패널 조사데이터의 새로운 점검 기준 개발을 위한 보다 개선된 방안이 필요하다.

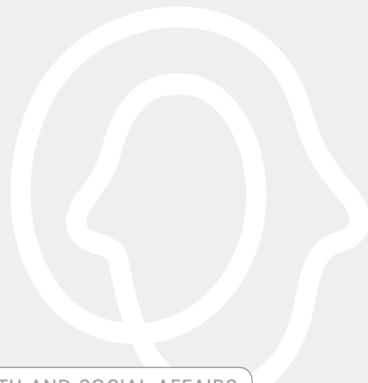
상기 5개 문제점 중 변수 간 논리적 관계 오류의 문제점은 현재 개선 작업을 진행하고 있으며, 의료서비스 이용 데이터(이용금액) 입력 오류 문제점과 질병코드 분류의 문제점은 한국의료패널 조사데이터의 품질에 영향을 주는 문제점이지만 향후 연구가 필요한 부분이다. 따라서 본 보고서에서는 연속형 변수의 이상치 탐지를 위한 시각화된 분석 방법과 새로운 점검 기준의 개발 및 검토에 초점을 맞춰 데이터 검증을 실시하고자 한다.

제3절 소결

제4장에서는 조사데이터 품질관리, 조사데이터 품질관리 절차 및 방법, 문제점 및 이슈 사항을 검토하였다.

조사데이터 품질관리에서는 통계 정보, 조사자료, 조사데이터 등에 대한 개념과 통계청의 통계 품질관리 및 행정안전부의 공공데이터 품질관리에 대해 살펴보았으며, 본 보고서에서는 조사데이터의 품질관리 범위를 자료 수집 및 처리 과정에서 발생할 수 있는 데이터 오류를 탐지하고 정확성을 개선하는 자료 내검 업무로 그 범위를 협의적으로 정의하였다.

한편, 한국의료패널 담당자 인터뷰를 통해 한국의료패널 조사데이터의 품질관리와 관련하여 변수 간 논리적 관계 오류의 문제점, 연속형 변수의 이상치 탐지의 문제점, 영수증 등 의료서비스 이용 입력 오류의 문제점, 텍스트로 작성된 의료서비스 이용 내용을 토대로 한 질병코드 분류의 문제점, 새로운 점검 기준 개발이라는 5개 문제점과 개선 방안을 검토하였다. 이 중에서 한국의료패널 조사데이터 품질관리에 있어서 핵심적인 문제점으로 연속형 변수의 이상치 탐지의 문제점과 새로운 점검 기준 개발 2개를 선정하여, 제5장의 ‘데이터 실증’에서 데이터 실증을 통해 조사데이터 품질관리 적용 방안을 논의하고자 한다.



제5장

조사데이터 품질관리 적용 방안 - 한국의료패널 데이터 실증

제1절 데이터 오류 수정 방안

제2절 한국의료패널 데이터 특성

제3절 데이터 실증

제4절 소결

제5장 조사데이터 품질관리 적용 방안 - 한국의료패널 데이터 실증

데이터의 오류 유형과 해당 오류에 대한 수정 알고리즘, 방법론 등을 파악한다. 이후, 한국의료패널 데이터에 대해서 개괄적인 특징을 확인하고, 변수별로 분포를 살펴보는 방법에 대해서 살펴보고, 다양한 방법을 적용하여 결과를 고찰하여 보기로 한다.

제1절 데이터 오류 수정 방안

1. 데이터 오류 유형

연구에서 분석 대상인 한국의료패널 데이터 같은 설문조사 데이터에서 변수 간의 논리적 오류를 확인하는 방법은 데이터의 일관성, 상호연관성, 그리고 논리적 규칙에 맞는지를 검사하는 과정이 필요하다. 사전에 정의된 변수 간의 논리적 관계에 기반한 규칙을 기준으로 오류를 찾아낼 수 있다. <표 5-1>은 몇 가지 오류 유형을 나타낸다.

<표 5-1> 데이터 오류 유형

유형	내용
논리적 오류	나이(age)가 18세 미만인 사람이 운전 면허증 소지 여부 변수에서 "소지함"으로 응답하는 것은 논리적으로 오류일 수 있다.
	"예/아니오"로 답할 수 없는 질문에 대해 이중 응답 검증으로 모순된 답변을 찾아낼 수 있다. 예를 들어, "흡연 여부"에서 "흡연함"으로 답한 사람이 "흡연 시도 경험"에서 "없음"이라고 답하면 논리적 오류가 있을 수 있다.

유형	내용
범주형 변수 간의 상관성 확인	교차 분석(contingency table)을 통해 범주형 변수 간의 관계를 확인할 수 있다. 예를 들어, 결혼 상태(marital status)와 자녀 수(children)가 서로 일치하는지 검토. 결혼 상태가 “미혼”인데 자녀 수가 2명인 경우 논리적 오류일 수 있다.
정상 범위 내에서의 값 확인	각 변수의 값이 현실적 범위 내에 있는지 확인할 필요성이 있다. 예를 들어, 나이(age)가 200세로 기록된 경우, 이는 오류일 가능성이 크다고 할 수 있다. 히스토그램, 박스플롯 등의 시각화를 사용해 변수들의 분포를 확인하고, 극단적인 값이나 불가능한 값을 식별할 수 있으며, 경험누적분포 추정을 통해 분위수에 기반한 이상치 검출을 수행할 수 있다.
수치 변수 간의 상관관계 분석	상관 분석을 통해 수치형 변수들 간의 관계를 평가할 수 있다. 예를 들어, 총소득과 소비 지출 간의 상관성을 확인하여 소비 지출이 총소득보다 큰 경우를 논리적 오류로 볼 수 있다.
시간 관련 데이터 확인	날짜 또는 시간이 포함된 데이터인 시계열 데이터 속성을 가지는 경우, 순서를 확인할 필요성이 있다. 예를 들어, 설문조사 응답자의 졸업 연도가 입학 연도보다 이전이라면 오류라 할 수 있겠다.
결측치와 비정상 값 처리	결측치(NA)와 숫자 아닌 값(NaN) 또는 정의역에서 벗어난 값(예: 성별이 “남성”과 “여성” 외의 다른 값으로 기록된 경우)을 찾아 처리할 수 있다. 또한, 응답자들이 논리적으로 응답하지 못한 항목으로, 예를 들어, 현재 시점을 2024년이라고 할 때, 생년월일이 2025년으로 기록된 경우를 들 수 있다.

자료: 저자 작성

2. 오류 수정 알고리즘

테이블 구조의 데이터에서 변수 간 오류를 탐지하고 품질을 평가하는 방법에는 통계적 방법, 기계학습 기반 이상 탐지, 규칙 기반 데이터 정리, AI 및 대형 언어모델(LLM) 활용, 시각적 분석 도구(〈표 5-2〉 참조) 등 다양한 방법이 연구되고 있다. 이 절에서는 데이터 오류를 수정하기 위한 알고리즘 연구의 대표적인 연구 및 기법을 살펴본다.

〈표 5-2〉 데이터 오류를 수정하기 위한 소프트웨어

소프트웨어	내용
OpenRefine	대규모의 데이터를 정리하고 변환하는 데 유용한 도구로, 오타, 형식 불일치, 중복 항목 등의 문제를 자동으로 감지하고 수정할 수 있는 기능을 제공함(https://openrefine.org/)
HoloClean	HoloClean은 PyTorch와 PostgreSQL을 기반으로 구축된 데이터를 정리 및 보강하는 통계적 추론을 수행하는 머신러닝 시스템으로 사용 가능한 품질 규칙, 가치 상관 관계, 참조 데이터 및 기타 여러 신호를 활용하여 데이터 생성 프로세스를 정확하게 포착하는 확률적 모델을 구축하여 오류를 수정할 수 있음(http://www.holoclean.io/)

자료: 저자 작성

가. 데이터 이상 탐지 기법

데이터의 품질을 평가하기 위해 이상 탐지(anomaly detection) 알고리즘을 적용할 수 있다. 통계 기반 이상 탐지 방법은 데이터의 분포를 분석하여 정상 범위를 설정하고, 이 범위를 벗어나는 값을 이상치로 식별하는 방식이다. 예를 들어, 평균과 표준편차를 이용해 데이터의 범위를 설정한 후, 이를 초과하거나 미달하는 값을 이상치로 판단할 수 있다. 더 나아가, 일반화된 통계적 기계 학습 방법들은 데이터의 분포 추정을 중요한 과정으로 다루며, 보다 정교한 이상 탐지 모델을 구축하는 데 활용될 수 있다

분포 추정을 위한 훈련 데이터에는 다른 데이터와 거리가 먼 관찰치로 정의된 이상치가 포함되어 있다고 가정한다. 이상치는 비정상적이거나 이례적인 관찰치라 할 수 있으므로, 비지도 학습 방법에 근거한 이상치 감지 방법은 훈련 데이터에서 소수개의 이상치 데이터가 저밀도 영역에 있다고 가정한다. 대표적인 방법인 밀도 기반 이상 탐지(DBSCAN, Ester et al., 1996) 방법은 이상치는 밀집 군집을 형성하기 어려운 이상치의 특징을 활용하여 비정상적인 데이터 패턴을 감지한다.

한국의료패널 데이터 같은 테이블 구조 데이터(tabular data)의 품질 평가를 위한 머신러닝 방법은, 데이터에 내재된 관계나 규칙을 학습하고 이를 기반으로 이상치나 오류를 탐지하는 알고리즘이다. <표 5-3>은 데이터 오류를 수정하기 위한 규칙 기반, 머신러닝 기반, AI 기반의 알고리즘의 특징을 나타낸다.

<표 5-3> 데이터 오류를 수정하기 위한 대표적 알고리즘

알고리즘	내용
규칙 기반 기법	사전에 정의된 비즈니스 규칙이나 도메인 지식을 활용하여 데이터의 일관성 및 정확성을 평가하는 방법으로, 예를 들어, 고객 데이터에서 우편번호와 도시명이 일치하는지, 나이와 생년월일 간의 일관성이 유지되는지를 평가할 수 있다.
Isolation Forest	비지도 학습 기반의 이상 탐지 알고리즘으로, 데이터 포인트를 반복적으로 분할하여 예외적으로 드문 패턴을 가진 데이터를 감지한다.
AI를 이용한 자동 오류 탐지	LLM을 활용해 테이블 데이터 내에서 테이블 구조에서 발생하는 비논리적 패턴을 탐지하는 데 사용할 수 있다. Tyen et al.(2024)에서는 LLM이 논리적 오류를 탐지하지 못하지만, 오류 위치가 주어지면 오류를 효과적으로 수정할 수 있음을 확인하였다. 이는 데이터 테이블에서도 오류 위치 정보를 제공하거나 별도의 작은 오류 탐지 모델을 도입하여 LLM을 활용한 자동 수정이 가능하다는 것을 시사한다.

자료: 저자 작성

나. 시각적 오류 탐지도구

데이터를 실시간으로 확인하고 분석하여 그 데이터를 기반으로 빠르게 의사결정을 내릴 수 있다. 변수 유형(예: 연속형, 범주형)에 따라 데이터를 시각적으로 표현하여 결측치나 이상값을 발견할 수 있다. 시각적 표현은 데이터를 쉽게 이해하고 분석할 수 있도록 하여, 데이터 분석 경험은 사용자도 유용하게 활용할 수 있다. 다양한 데이터 소스를 통합적으로 관리하며, 분석 속도를 높이는 데 기여할 수 있다. 이 절에서는 탐색적 자료 분석 및 데이터 시각화를 효과적으로 제공해주는 툴인 YData Profiling에 대해 살펴본다.

〈표 5-4〉 시각화 도구

소프트웨어	내용
Tableau	시각화 도구를 사용하여 시각화하고 패턴과 이상값을 탐지할 수 있음
Pandas Profiling	Python 라이브러리인 Pandas Profiling은 분석 및 시각적으로 요약 해주는 도구로, 결측값, 분포, 상관관계 등을 한눈에 파악할 수 있어 데이터 품질을 쉽게 평가할 수 있음 https://github.com/HoloClean/HoloClean

자료: 저자 작성

YData Profiling은 데이터셋을 간단하고 빠르게 분석하고 이를 시각적으로 요약하여 보여주는 python 라이브러리로, 최소값, 최대값, 평균과 같은 컬럼에 대한 기본적인 요약 통계와 더불어 상관관계, 결측치, 이상치 탐지의 기초 자료를 제공한다.

〈표 5-5〉 YData Profiling report 옵션

항목	내용
개요	데이터셋의 행(row)과 열(column) 수, 결측값 수, 데이터 타입 등 데이터의 전반적인 요약
변수	각 변수별 분포, 평균, 최솟값, 최댓값, 표준편차 같은 요약 통계량 및 데이터 분포 그래프 제공
상관관계	변수별 상관계수행렬 및 열그림(heatmap) 제공
결측값 분석	Pearson, Spearman 등의 상관계수를 이용해 변수 간의 상관관계 분석 및 시각화 자료 제공
샘플 데이터	실행 데이터셋의 일부 샘플 row 출력

자료: 저자 작성

또한, 히스토그램 같은 시각적 자료를 통해 데이터의 분포를 쉽게 확인할 수 있으며, 프로파일링 결과를 HTML 파일로 저장할 수 있다. 주요 기능(〈표 5-6〉 참조) 중, 변수별 요약은 각 변수에 대해 다음과 같은 세부 정보가 제공된다. 수치형 변수는 분포, 평균, 최솟값, 최댓값, 백분위수, 결측값 비율, 이상값 등을 제공한다. 범주형 변수는 각 범주의 빈도수, 결측값, 고유값 등을 제공한다. 데이터의 결측값을 사전에 규약된 규칙에 따라 자동으로 탐지하여 결측값이 많은 열에 대해 경고를 표시하며, 결측

값 비율을 시각화한다. 이상치를 탐지하여 리포트하고, 데이터 정제 작업을 할 수 있다. 상관분석을 통해 변수 간의 관계를 시각적으로 나타낼 수 있다. Pearson, Spearman, Kendall 상관계수를 제공하며, 변수 간의 상관관계가 높은 경우 해당 사항을 리포트한다.

예제를 통해 사용법을 살펴보기로 한다. pip 명령어를 사용하여 설치하고, 'example.csv' 데이터를 불러온다. 로드한 example 데이터에 대해 자동화된 Profiling report를 생성하고 이를 Jupyter Notebook 내에서 HTML 형식으로 저장한다.

〈표 5-6〉 예제 자료에 대한 YData Profiling 실행 예시

```
#YData Profiling 설치
pip install ydata-profiling
#또는 conda install -c conda-forge ydata-profiling
#패키지 호출
from ydata_profiling import ProfileReport
import pandas as pd
# 데이터 불러오기
df = pd.read_csv('example.csv')
# 프로파일링 생성
profile = ProfileReport(df, title="Data Profiling Report")
# Jupyter Notebook 내에서 보고서 확인
profile.to_notebook_iframe()
# HTML 보고서 저장
profile.to_file("output_report.html")
```

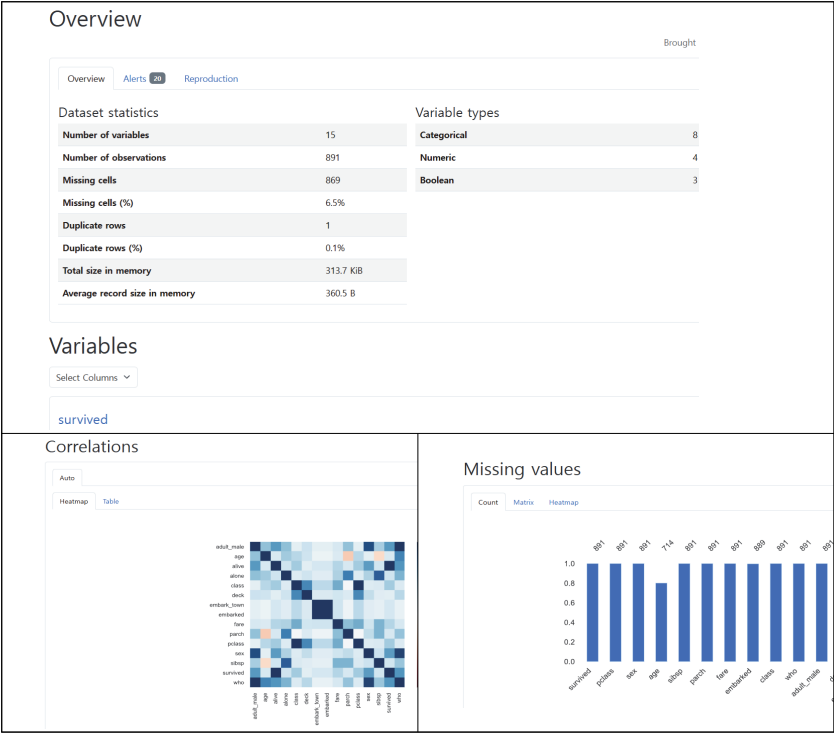
```
# 상관관계 분석을 활성화한 프로파일링 예시
import seaborn as sns
import matplotlib.pyplot as plt
df = sns.load_dataset('titanic')

df = sns.load_dataset('titanic')
profile = ProfileReport(df, title="Titanic Dataset Profiling Report", explorative=True,
                        correlations={
                            "pearson": {"calculate": True}, # Pearson 상관계수
                            "spearman": {"calculate": True}, # Spearman 상관계수
                            "kendall": {"calculate": False},
                            "phi_k": {"calculate": False}, # phi_k 상관계수
                        })
profile = ProfileReport(df, title="Titanic Dataset", explorative=True)
profile.to_file("titanic_report.html")
```

자료: 저자 작성

[그림 5-1]은 출력된 "output_report.html"에 대한 결과의 일부를 나타낸다.

[그림 5-1] 예제 자료에 대한 YData Profiling 실행 결과



자료: 저자 작성

또한, 분석에 불필요한 변수를 제외하고 보고서를 생성할 수 있다. 예를 들어, ID와 같은 유일한 식별자는 분석에서 제외할 수 있다.

```
# 특정 열 제외
profile = ProfileReport(df, title="Titanic Dataset Profiling Report",
explorative=True, vars={"num": {"reject": ["passenger_id"]}})
```

대규모 데이터셋을 처리할 때는 일부 데이터만 샘플링하여 프로파일링 보고서를 생성하는 방법이 유용할 수 있다.

```
# 데이터의 10%만 샘플링하여 프로파일링
profile = ProfileReport(df.sample(frac=0.1), title="Sampled Titanic Dataset
Report", explorative=True)
```

Pandas Profiling과 유사한 데이터 탐색 및 보고서 생성 도구 역할을 하는 패키지의 리스트는 <표 5-7>과 같다.

<표 5-7> 탐색적 자료 분석과 시각화를 위한 python 패키지

패키지	주요 기능	특징
Sweetviz	데이터의 요약 통계 및 시각적 보고서 생성	두 데이터프레임 간의 비교 기능 제공
dtale	대화형 웹 UI를 통해 데이터 분석 및 시각화	데이터 필터링, 통계 정보, 시각화
Autoviz	데이터의 중요한 정보를 자동 시각화	대규모 데이터셋에서도 빠르게 시각화 생성
Lux	데이터 탐색 시 다양한 시각화 옵션을 자동 제안	Pandas와 결합하여 사용 가능

자료: 저자 작성

제2절 한국의료패널 데이터 특성

한국의료패널은 개인 및 가구 단위로 매년 반복 조사되며, 이를 통해 의료 서비스 이용의 변화를 추적할 수 있다. 한국의료패널 데이터는 의료 서비스 이용과 건강 상태에 대한 중요한 데이터를 제공하는 조사로, 이를 통해 한국의 보건의료 정책과 연구에 큰 기여를 하고 있다고 할 수 있다. 이 데이터는 패널 방식으로 수집되기 때문에 시간에 따른 변화 추적이 가능하며, 연구자와 정책 입안자에게 매우 중요한 자료이다. 따라서 데이터의 품질관리 및 효율적인 공개를 위해 지속적인 연구와 관리 체계 개선이 필요할 수 있다.

한국의료패널 데이터에 대한 조사항목과 특징에 대해 살펴본다. 한국의료패널 데이터는 매년 반복 조사(panel survey) 형태로 진행되는데, 패널 조사 방식으로, 동일한 가구와 개인을 대상으로 매년 조사를 반복함으로써 시간에 따른 변화를 추적할 수 있는 장점을 가진다. 가구, 개인, 의료 이용 등의 다양한 수준에서 세부적인 정보를 제공한다(〈표 5-8 참조〉).

〈표 5-8〉 한국의료패널 데이터의 조사항목과 내용

조사항목	내용
가구	가구 구성, 경제적 상태, 주거 형태 등 가구 단위의 정보 및 가구의 의료 비용 부담, 의료 서비스 접근성 등의 정보
개인	개인의 기본 정보: 나이, 성별, 교육 수준, 직업, 소득 수준, 건강 상태 등 건강 행태: 흡연, 음주, 운동 습관 등의 건강 관련 행동 만성질환 여부: 고혈압, 당뇨병 등 주요 만성질환의 유무와 관리 여부 건강보험: 개인의 건강보험 및 의료 보장성 관련 정보
의료 이용	외래 진료 및 입원: 진료받은 날짜, 진료받은 의료기관의 유형, 진료과, 진료 횟수, 입원 기간 등 의료 비용: 진료비, 본인 부담 비용, 건강보험 적용 비용 등 의약품 사용: 처방받은 의약품의 종류, 사용 기간, 비용
의료비 지출	개인 및 가구가 지출한 의료비와 이를 충당하기 위한 자금원(보험, 보조금 등)을 조사하며, 미용 목적의 의료 서비스나 대체 치료 등

자료: 저자 작성

한국의료패널 조사는 매년 동일한 항목을 조사하는 고정 항목과 당해 연도의 정책적 필요나 연구 주제에 따라 추가되는 변동 항목으로 구성된다.

〈표 5-9〉 데이터 오류를 수정하기 위한 알고리즘

항목	내용
고정	인구통계학적 특성(나이, 성별, 교육 수준 등) 의료 이용 현황(외래 방문, 입원, 응급실 방문 등) 가구의 소득 및 지출 구조 건강보험 및 의료보장 관련 정보 만성질환 여부 및 관리 상태
변동	특정 연도에 중점적으로 다루는 주제나 연구에 필요한 항목이 포함되며, 예를 들어, 감염병 유행 시기에는 해당 내용이 추가될 수 있다.

자료: 저자 작성

한국의료패널 데이터는 연구 목적으로 사용될 수 있으며, 연구자들에게는 마이크로데이터 형태로 제공된다. 데이터를 사용하려면 한국의료패널 데이터베이스에 신청해야 하며, 연구자들은 승인받은 후 데이터를 이용할 수 있다. 한국의료패널 데이터는 다양한 분야에서 연구 목적으로 활용되며, 특히 의료정책, 보건의료 서비스 연구, 의료비용 분석 등에 자주 사용된다. 예를 들어, 시간이 지남에 따라 의료서비스 이용 패턴이 어떻게 변화하는지와 같은 의료서비스 이용 추세 분석을 수행할 수 있다. 또한, 건강보험 제도 연구를 통한 국민건강보험의 적용 범위와 보장성을 분석하여 정책 개선에 기여할 수 있다. 한편, 의료비 지출이 가구 경제에 미치는 영향을 평가하고, 이를 바탕으로 의료비 부담 연구와 관련된 정책을 제안하는 데 활용되기도 한다.

한국의료패널 데이터는 방대한 양의 정보를 제공하지만, 다음과 같은 한계점이 존재한다. 의료 이용이나 건강 상태 등의 정보는 주로 응답자가 직접 제공하는 자기 보고(self-reported)의 특징을 가지는 자료로, 주관적인 평가가 포함될 수 있다. 또한, 일부 취약 계층이나 의료 접근성이 낮

은 인구의 데이터는 특정 계층의 정보 부족으로 충분히 반영되지 못할 수 있다. 이를 개선하기 위해 데이터 관리 부서의 품질관리 및 데이터 가공 절차가 더욱 강화될 필요가 있으며, 자동화된 데이터 분석 도구와 시각화를 통해 연구자들의 접근성을 높여야 한다.

제3절 데이터 실증

본 절에서는 한국의료패널 데이터에서 설문조사의 질문과 변수별 분포 및 변수 간 관계를 분석하여 내포된 데이터 오류를 탐지하는 방안을 서술한다. 먼저, 각 변수의 오류 탐지에 활용되는 규칙을 유형별로 분류하고, 유형별 규칙이 적용된 결과와 관련된 데이터 분포를 입체적으로 분석하는 방법을 살펴본다. 다음으로, 대형 언어모델(LLM)을 적용하여 자연어 질의로부터 구조적 질의 언어(SQL) 또는 파이썬(Python) 코드를 생성하는 방안을 모색하고, 이를 통해 순서도를 따라 기존 규칙을 자동으로 처리할 수 있는 방법을 탐색한다. 마지막으로, 대형 언어모델(LLM)을 활용하여 기존 규칙을 보완하거나 새로운 규칙을 탐색하는 연구를 수행한다.

1. 메타정보와 검수 질의문

메타정보는 세대 구성 외 5개의 그룹으로 구성되어 있는데, ID 및 가구원 변동 사항(ID), 가구 데이터(HH), 가구원 데이터(IND), 부가조사_건강정보 이해능력(C_IND), 의료서비스 이용(MS), 민간의료보험(PHI) ID 및 가구원 변동 사항(ID)으로 분류되어 있다. <표 5-10>은 세대 구성 변수군에 관한 그룹코드를 나타낸다.

〈표 5-10〉 세대 구성 변수의 코드와 설명

구분	코드	설명
1세대	11	1인 가구
	12	부부(응답자+배우자)
	13	응답자+형제자매
	14	기타 1세대 가구
2세대	21	부부+자녀
	22	편부모+자녀
	23	부부+양친
	24	부부+편부모
	25	기타 2세대 가구
3세대	31	부부+자녀+양친
	32	부부+자녀+편부모
	33	편조부모+편부모+자녀
	34	기타 3세대 가구
4세대	41	4세대 가구

자료: 저자 작성

나머지 5개의 정보는 가구주와의 관계(20종), 산업분류(21종), 직업분류(52종), 만성질환 코드(2019~2020년, 30종, 2021년, 31종, ‘어깨관절질환’ 추가), 의과 입원/외래 관련 주질환 코드(2019~2010년, 50종, 2021년 51종, ‘알코올성 간질환’ 추가)로 구성되어 있다.

〈표 5-11〉 데이터 크기와 변수와 수준 수

구분	2020 조사 데이터 크기	2021 조사 데이터 크기	2022 조사 데이터 크기	변수 수	수준 수
ID 및 가구원 변동 사항(ID)	14741	13530	12874	22	54
가구 데이터(HH)	6748	6217	5907	76	204
가구원 데이터(IND)	16587	14844	13799	335	1317
부가조사_건강정보 이해능력(C_IND)				22	136
의료서비스 이용(MS)	260989	244370	270795	195	845
민간의료보험(PHI)	21538	21345	21392	25	71

자료: 저자 작성

〈표 5-12〉는 가구원 변동 사항(ID), 가구 데이터(HH), 가구원 데이터(IND), 부가조사_건강정보 이해능력(C_IND), 의료서비스 이용(MS), 민간의료보험(PHI) ID 및 가구원 변동 사항(ID)에 대한 총 변수 수와 총 수준 수를 나타낸다.

〈표 5-12〉 ID 및 가구원 변동 사항(ID) 변수의 코드와 설명

변수명	한글 변수명	보기 문항 내용	사용 주의사항	2019		2020		2021	
				빈도	(%)	빈도	(%)	빈도	(%)
가구 및 가구원 ID									
HID	원가구 ID	000000		14,741	100	13,530	100	12,874	100
M1	가구생성 차수	(01) 2020년 1차 조사		14,741	100	13,530	100	12,874	100
M2	가구분리 번호	0		14,741	100	13,530	100	12,874	100
HHID	가구식별 번호	000000000	HID(6자리)+ M1(2자리)+ M2(1자리)	14,741	100	13,530	100	12,874	100
PID	가구원 번호	00		14,741	100	13,530	100	12,874	100
N1	가구원 진입차수	(01) 2020년 1차 조사		14,741	100	13,530	100	12,874	100
HPID	가구원 식별번호	00000000000	HHID(9자리)+ PID(2자리)	14,741	100	13,530	100	12,874	100
PIDWON	가구원 고유번호	00000000	HID(6자리)+ PID(2자리)	14,741	100	13,530	100	12,874	100
가구원 변동 사항									
CH1	가구원 변동 유무	(0) 변동 사항 없음	부가 가구로 조사된 원가구원 포함			13,290	98.23	12,626	98.07
		(11) 신규 가구원				114	0.84	109	0.85
		(12) 신규+사망 가구원				0	0.00	1	0.01
		(13) 신규+부가 가구원 (중략)				0	0.00	1	0.01
		(53) 합가+부가 가구원				0	0.00	0	0.00
...		...							

자료: 저자 작성

〈표 5-13〉 가구원 데이터(IND) 변수의 코드와 설명

변수명	한글 변수명	보기 문항 내용	사용 주의사항	2019		2020		2021	
				빈도	(%)	빈도	(%)	빈도	(%)
HHID	가구식별 번호	000000000	HID(6자리)+ M1(2자리)+M2 (1자리)	16,587	100	14,844	100	13,799	100
PIDWON	가구원 고유번호	000000000	HID(6자리)+ PID(2자리)	16,587	100	14,844	100	13,799	100
I_WGC	가구원 가중치	유효	가중치만 존재하고 응답값은 나타나지 않음	14,741	88.87	13,530	(91.15)	12,874	(93.30)
	가중치 횡단 (모집단)	가중치만 존재하고 응답값은 나타나지 않음		1,846	11.13	1,314	(8.85)	925	(6.70)
I_WGL	가구원 가중치	유효	가중치만 존재하고 응답값은 나타나지 않음			13,530	(91.15)	12,874	(93.30)
	가중치 종단 (모집단)	가중치만 존재하고 응답값은 나타나지 않음				1,314	(8.85)	925	(6.70)
I_WSC	가구원 가중치	유효	가중치만 존재하고 응답값은 나타나지 않음	14,741	88.87	13,530	(91.15)	12,874	(93.30)
	가중치 횡단 (표본)	가중치만 존재하고 응답값은 나타나지 않음		1,846	11.13	1,314	(8.85)	925	(6.70)
I_WSL	가구원 가중치	유효	가중치만 존재하고 응답값은 나타나지 않음			13,530	(91.15)	12,874	(93.30)
	가중치 종단 (표본)	가중치만 존재하고 응답값은 나타나지 않음				1,314	(8.85)	925	(6.70)

자료: 저자 작성

〈표 5-14〉 가구 데이터(HH) 변수의 코드와 설명

변수명	한글 변수명	보기 문항 내용	사용 주의사항	2019		2020		2021	
				빈도	(%)	빈도	(%)	빈도	(%)
HHID	가구식별 번호	000000000	HID(6자리)+M1 (2자리)+M2(1자리)	6,748	100	6,217	100	5,907	100
가중치									
H_WGC	가구 가중치	유효		6,689	(99.13)	6,190	(99.57)	5,878	(99.51)
	가중치만 횡단 (모집단)	존재하고 응답값은 나타나지 않음		59	(0.87)	27	(0.43)	29	(0.49)
H_WSC	가구 가중치	유효		6,689	(99.13)	6,190	(99.57)	5,878	(99.51)
	가중치만 횡단 (표본)	존재하고 응답값은 나타나지 않음		59	(0.87)	27	(0.43)	29	(0.49)
가구 일반 사항									
DATE_Y	조사일자 (년)	()년		6,689	100	6,190	(100.00)	6,190	(100.00)
DATE_M	조사일자 (월)	()월		6,689	100	6,190	(100.00)	6,190	(100.00)
DATE_D	조사일자 (일)	()일		6,689	100	6,190	(100.00)	6,190	(100.00)

자료: 저자 작성

전체 63개 질의문에 대한 비율(anomaly_ratio)은 <표 5-16>과 같다. 확률의 분포로부터 최소, 최대값은 [0, 19.943]이다(단위: %). 표에서 유형은 후술할 질의문들의 특징에 따른 순서도의 종류를 나타낸다.

오류 탐지를 위한 훈련 데이터로 활용할 때 가능한 모든 정보를 활용하기 위해서는 데이터 구조를 확인할 필요가 있다. 6개의 파일을 병합하여 정보를 통합적으로 활용하기 위해, 각 파일의 행을 고유하게 식별할 수 있는 기본 키(primary key)와 파일 간 매핑 가능한 가구별 및 개인별 키를 식별한다. ID 파일은 가구 및 가구원의 고유 식별 정보를 포함하며, 사망이나 가구원 구성 변경 등 변동 사항을 관리한다. 가구를 식별하기 위한 키인 HHID와 가구원을 식별하기 위한 키인 PIDWON이 포함되어 있으며, PIDWON은 ID 파일 내에서 중복되지 않는다.

HH 파일은 가구 단위 정보를 담고 있으며, 주요 항목으로는 가중치, 가구 일반 사항, 소득, 부채 및 자산, 지출 등이 포함된다. 가구 구분 키인 HHID가 포함되어 있으며, HHID는 HH 파일 내에서 중복되지 않는다. IND 파일은 가구원 단위의 정보를 포함하며, 주요 항목으로는 가중치, 가구원 일반 사항, 경제활동 상태, 소득 등이 있다. 가구 구분 키 HHID와 가구원 구분 키 PIDWON을 포함하며, PIDWON은 IND 파일 내에서 중복되지 않는다.

MS 파일은 의료 이용 건별 정보를 포함하며, 개인별 의료 이용 공통 사항, 응급 서비스 이용, 입원 서비스 이용, 외래 서비스 이용 등으로 구성된다. 가구 구분 키 HHID와 가구원 구분 키 PIDWON이 포함되며, 개인별 및 가구별 의료 이용 내역이 복수로 존재할 수 있어 HHID와 PIDWON이 중복될 수 있다. PHI 파일은 가구별 민간의료보험 가입 정보를 포함하며, 가구 구분 키 HHID, 가구별 보험 일련번호 PHI_N, 피보험자 모두의 PIDWON을 포함하고 있다. 가구별로 복수의 민간의료보험

에 가입되어 있을 수 있어 HHID가 중복으로 등장할 수 있다.

데이터 테이블 간 병합 과정에서는 left join을 사용하여 기준 파일의 모든 데이터를 유지하면서 필요한 정보를 추가하거나, right join을 통해 부차 파일의 데이터를 최대한 포함할 수 있도록 한다. 또한, 두 파일 간 공통 데이터를 중심으로 병합하는 경우 inner join을 사용할 수 있으며, 모든 데이터를 포함해야 하는 경우 full outer join을 활용하여 데이터 손실을 방지한다. 이러한 다양한 병합 방법을 적절히 적용함으로써 가구 별 및 개인별 정보를 효과적으로 검색하고 활용할 수 있다. 또한, 병합 시 결합되는 유형에 따라 결합되지 않는 정보의 손실이 발생되지 않게 주의할 필요가 있음을 시사한다.

〈표 5-15〉 데이터 병합 예시

병합된 파일이름	기준 파일	결합 대상 파일	병합 기준 key	병합 방법	설명
병합 1	ID	HH	HHID	left	ID 파일의 가구 데이터를 기준으로 HH 파일의 가구 단위 정보를 추가
병합 2	HH	IND	HHID	left	HH 파일의 가구 데이터를 기준으로 IND 파일의 가구원 정보를 연결
병합 3	IND	MS	PIDWON	right	MS 파일의 의료 이용 기록을 기준으로 IND 파일의 개인정보를 연결하며, 의료 데이터 손실을 방지
병합 4	병합 3	PHI	HHID	right	HH 파일의 가구 데이터를 기준으로 PHI 파일의 민간의료보험 정보를 연결

자료: 저자 작성

〈표 5-16〉 사전 검수 질의문

idx	category	검수 질의문	유형	데이터 크기	anomaly	anomaly_ratio (%)
1	소득 1	[근로소득] 이상치 재확인 *고소득 근로소득자의 경우 “경제활동 상태”의 [주된 경제활동 상태]와 [직업분류] 재확인 *육아휴직인 경우 근로소득 0일 수 있음	1	16587	23	0.139
2		“가구원 일반사항 2”의 [의료보장 형태]가 ‘⑤ 의료급여(1, 2종) 세대주’ 또는 ‘⑥ 의료급여 세대원’일 때 [근로소득] 재확인 *연근로소득이 2,000만 원을 초과하거나 50만 원 미만인 경우	1	16587	0	0.000
3		[사업소득] 이상치 재확인 *상위 1% 등 고소득 사업소득자의 경우 “경제활동 상태”의 [주된 경제활동 상태]와 [직업분류]를 재확인 *[사업소득]은 마이너스(-)가 있을 수 있으므로 모름(-9)과 잘 구분하여야 함	8	16587	13	0.078
4		“가구원 일반사항 2”의 [의료보장형태]가 ‘⑤ 의료급여(1, 2종) 세대주’ 또는 ‘⑥ 의료급여세대원’일 때 [사업소득] 재확인 * 연사업소득이 1000만 원 이상인 경우	1	16587	3	0.018
5	소득 2	[재산소득_금융소득] 이상치 재확인 *연금융소득이 0원 초과 10만 원 미만인 경우 또는 1억 이상인 경우	4	6748	101	1.497
6		[재산소득_부동산/동산 소득] 이상치 재확인 *연부동산/동산소득이 0원 초과 10만 원 미만인 경우 또는 2억 5천만 원 초과인 경우	6	6748	2	0.030
7		[재산소득_기타소득] 이상치 재확인 *연기타소득이 0원 초과 10만 원 미만인 경우, 또는 1,000만 원 초과인 경우	6	6748	3	0.045
8		[재산소득_금융소득] & [재산소득_부동산/동산 소득]이 0원이 아닌데 [자산]이 0원인 경우, 또는 [자산] 이상치 재확인 *부동산/동산 소득 중 보증금 여부를 재확인할 수 없으므로 조사원 교육 시 강조 필요	2	6748	5	0.074
9		[공적 이전 소득_정부 및 지방자치단체] 이상치 재확인	6	6748	119	1.764
10		[공적연금]이 ‘① 유’인 경우, 가구원 구성원의 만 나이 재확인 *ex) 공적연금을 받는데 가구 구성원의 나이가 모두 25세 미만이면 오류 *공적연금(국민연금, 공무원·군인연금, 사학연금, 보훈연금, 별정직 우체국연금 등)	3	16587	0	0.000
11		[기초연금]이 ‘① 유’인 경우, 만 나이가 65세 이상인 가구원이 없으면 오류 *기준연도, 가구 내 수급자 수에 따라 금액 이상이하 *2023년 기준 최대금액: 노인 1인 388만 원, 2인 776만 원, 부부노인 621만 원, 3인 1,163만 원	3	16587	30	0.181

idx	category	검수 질의문	유형	데이터 크기	anomaly	anomaly_ratio (%)
12	부채 및 자산	[건강보험] 소득 이상치 재확인	6	6748	10	0.148
13		[사적이전] 소득 이상치 재확인	6	6748	3	0.045
14		[부채] 이상치 재확인	4	6748	9	0.133
15		[자산] 이상치 재확인	4	6748	5	0.074
16	지출	[지출_소비지출]과 [지출_비소비지출]의 합이 월평균 소득보다 큰 경우 재확인 *분가하여 신규로 조사된 가구는 소득 합계가 적어서 비교값이 차이 날 수 있음	5	6748	3	0.045
17	일반의약품 및 의료 관련 지출	[일반의약품 및 의약품] 이상치 재확인	4	6748	22	0.326
18		[한약 및 첨약] 이상치 재확인	6	6748	25	0.371
19		[건강보조식품_복용] 이상치 재확인	6	6748	81	1.200
20		[건강보조식품_선물] 이상치 재확인	1	6748	25	0.371
21		[시력 보조용품] 이상치 재확인	6	6748	0	0.000
22		[청력 보조용품] 이상치 재확인	6	6748	0	0.000
23		[의료기기 및 의료용품] 이상치 재확인	6	6748	0	0.000
24		[일반의약품 및 의료 관련 지출]의 총합이 [소비지출]보다 큰 경우 오류	5	6748	0	0.000
25		교통수단이 '③ 대중교통'인데 [교통비]가 5,000원 초과, 250원 미만이면 재확인 *7세 이하, 65세 이상, 장애가 있는 경우 제외	2	260989	37	0.014
26		교통수단이 '④ 택시'인데 [교통비]가 10만 원 초과, 1,800원 미만이면 재확인 *장애가 있는 경우 제외	2	260989	322	0.123
27	의료 서비스 이용 (공통)	교통수단이 '⑤ 기차/비행기/고속(시외)버스'인데 [교통비]가 20만 원 초과, 1,800원 미만이면 재확인 *7세 이하 제외	2	260989	2059	0.789
28		교통수단이 '⑦ 병원구급차' 또는 '⑧ 그 외 구급차'인데 [교통비]가 20만 원 초과, 3만 원 미만이면 재확인	2	260989	15	0.006
29		[의료비_수납 금액] 이상치 재확인(ex. 응급, 입원 5천 원 미만)	1	260989	40	0.015
30		[의과 입원 치료 내용]이 '① 수술 및 시술'인데 [의료비_수납 금액]이 10만 원 미만인 경우 재확인	1	260989	31	0.012
31		[의료비_법정본인부담금] 1,000만 원 초과, 100원 미만이면 재확인 *의료급여 환자 제외	6	260989	19	0.007
32		[의료비_비급여 금액] 1,000만 원 초과, 100원 미만이면 재확인	6	260989	70	0.027
33		[의료비_할인, 절삭 등 금액]이 마이너스인 경우 영수증 재확인(ex. 1천 원 이상)	6	260989	169	0.065
34		[의료비_할인, 절삭 등 금액]이 큰 경우 영수증 재확인(ex. 15만 원 이상) *의료급여 환자, 자동차보험 환자, 산재보험 환자 제외	6			
35		[의료비_보험자부담금] 8,000만 원 초과, 1천 원 미만이면 재확인 *의료급여 환자 제외	6	260989	67	0.026

140 인공지능(AI) 기반 조사데이터 품질 제고 방안 연구

idx	category	검수 질의문	유형	데이터 크기	anomaly	anomaly_ratio (%)
36		[약제비] 500원 미만이면 재확인 *의료급여 환자, 65세 이상 제외	6	260989	17	0.007
37	의료서비스 이용(응급)	[소요 시간] 이상치 재확인(8시간 초과, 10분 미만) *본인이 해당 병원에 근무하는 경우가 있을 수 있음	6	260989	34	0.013
38	의료서비스 이용(입원)	[유급 간병 비용]이 30만 원 이상 만 원 미만(95 제외) 재확인	6	260989	3	0.001
39	의료서비스 이용(외래)	[외래 관련 주질환]과 [연령(해당 연도-생년)] 관계 재확인 *10세 미만의 협심증, 심근경색증	7	260989	0	0.000
40		[외래 관련 주질환]과 [연령(해당 연도-생년)] 관계 재확인 *15세 미만의 자궁경부암, 임신 및 출산 관련	7	260989	0	0.000
41		[외래 관련 주질환]과 [연령(해당 연도-생년)] 관계 재확인 *20세 미만의 알코올성 간질환, 통풍, 등통증, 유방암(등통증은 나오기도 함)	7	260989	0	0.000
42		[외래 관련 주질환]과 [연령(해당 연도-생년)] 관계 재확인 *30세 미만의 치매, 백내장	7	260989	0	0.000
43		[외래 관련 주질환]과 [연령(해당 연도-생년)] 관계 재확인 *40세 미만의 무릎 골관절염, 무릎 외 골관절염	7	260989	0	0.000
44	장기요양 및 기타 돌봄 서비스	[(방문/보호)재가서비스, 유급 간병인, 가사도우미, 기타 돌봄(비용)] 이상치 재확인(300만 원 이상)	2	16587	0	0.000
45		[장기요양시설, 양로원 등 공동주거시설, 기타 시설(비용)] 이상치 재확인(200만 원 이상)	2	16587	0	0.000
46		[산후조리원 비용] 이상치 재확인(10만 원 미만, 500만 원 이상)	6	16587	1	0.006
47		일 평균 [산후조리원 비용](비용(만 원)/이용 기간(일)) 이상치 재확인(10만 미만, 50만 원 이상)	5	16587	4	0.024
48		[산후도우미 비용] 이상치 재확인(10만 원 미만, 500만 원 이상)	6	16587	2	0.012
49		일 평균 [산후도우미 비용](비용(만 원)/이용 기간(일)) 이상치 재확인(10만 원 미만, 500만 원 이상)	5	16587	26	0.157
50	민간의료 보험 1	[보험료] 이상치 재확인(모르는 경우 -9) *[보험료]가 '0원'인 경우 에러→ [보험료 납부 여부]에 관계없이 계약상 월 보험료를 작성 *보험료가 '1~10' 값인 경우 에러→'원' 단위인데 '만 원' 단위로 단위 착각	2	21538	22	0.102
51	민간의료 보험 2	[보장 유형]이 '⑦ 종신/연금/변액/저축보험'인 경우 이상치 확인 *지침: "⑦ 종신/연금/변액/저축보험"이 포함되는 경우, 주계약(종신/연금/변액/저축보험)을 제외한 의료특약 보험료만을 기입함	6			
52	민간의료 보험 3	[보험금 수령액] 재확인 *보험금 수령액이 '0~100' 값인 경우 에러→'원' 단위인데 '만 원' 단위로 단위 착각	6	21538	0	0.000
53	건강생활 습관	[신체활동별 비용] 이상치 재확인	2	16587	0	0.000

idx	category	검수 질문문	유형	데이터 크기	anomaly	anomaly_ratio (%)
54	건강 수준	[걷기-최근 일주일 하루 걸은 시간] 이상치 재확인 *응답값이 명확하게 입력된 값인 경우, 오타인지 재확인(보통 응답할 때 5단위 혹은 10단위로 응답하므로 예: 11분, 17분, 23분.... 등)	6	16587	3308	19.943
55		[(결련형) 전자담배 사용 비용] 이상치 재확인	6	16587	0	0.000
56		(니코틴 포함 액상) 전자담배 사용 비용 이상치 재확인	6	16587	0	0.000
57		[금연 방법별 비용] 이상치 재확인(100만 원 초과)	2	16587	0	0.000
58		[금주 시작 시기(년/월)] 이상치 재확인	6	16587	0	0.000
59		[신장], [몸무게] 이상치 재확인(연령과 연결해서 확인)	2	16587	0	0.000
60		지난 차수와 비교하여 [신장], [몸무게] 불일치 재확인 *키 5cm 이상, 몸무게 15kg 이상 차이 날 때 재확인 요청	unk now n			
61		키 5cm 이상, 몸무게 15kg 이상 차이 날 때 재확인 요청[와병 여부]가 '① 예'일 때 [와병일수] 이상치 재확인 *[와병일수]가 365일 이상인 경우 확인 *[와병일수] 전체 빈도 확인 *[와병일수] 큰 값 확인(300일 이상)	6	16587	0	0.000
62		[결근결석 여부]가 '① 예'일 때 [결근결석일수] 이상치 재확인 *[결근결석일수]가 365일 이상인 경우 확인 *[결근결석일수] 전체 빈도 확인 *[결근결석일수] 큰 값 확인(ex.200일 이상: 2021년 법정 공휴일 64일, 총휴일 113일)	6	16587	1	0.006
63		[프리젠티즘 여부]가 '① 예'일 때 [프리젠티즘 일수] 이상치 재확인 *[결근결석일수]가 365일 이상인 경우 확인 *[결근결석일수] 전체 빈도 확인 *[결근결석일수] 큰 값 확인(ex. 200일 이상: 2021년 법정 공휴일 64일, 총휴일 113일)	6	16587	3	0.018

자료: 저자 작성

2. 검수 질문의 유형

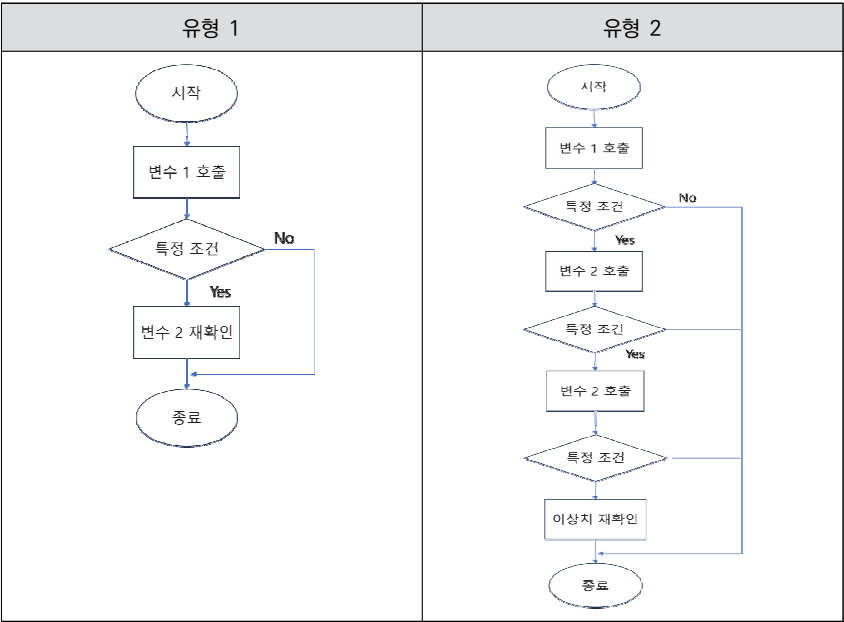
전체 63개의 검수 질문문의 특징을 살펴보고자 순서도를 활용하여 유형별로 구분하여, 8개의 유형별로 구분된 질문문의 순서도를 나타낸다.

[그림 5-2]는 유형 1과 2를 나타낸다. 유형 1은 변수 1을 조건을 통해 1차 필터링한 후, 변수 2를 재확인하는 구조로 이루어져 있는데 1~5번, 10번, 20번이 이에 해당된다. 특정 조건은 '해당 변수가 특정 범주에 해

당하는지?’ 또는 ‘해당 변수가 특정 범위에 해당하는지?’를 1차적으로 필터링하게 되고, 해당하는 행들에 대하여 2차 변수를 통해 최종 이상치를 탐지하게 된다.

유형 2는 복수 개의 변수(2~3개)에 대한 여러 조건들을 확인한 후 부합하면 이상치를 재확인한다. 질의문 8번, 25~29번, 44~45번, 53번, 57번, 59번이 이에 해당된다. 특정 조건은 ‘복수 개의 변수 값에 대하여 동시에 살펴봄으로써 이상치를 확인하는 방법’과 ‘특정 변수가 해당되는 특정 범주의 타 변수 값의 이상치를 살펴보는 방법’, ‘동일한 경계 값을 기준으로 복수의 변수의 이상치를 탐지하는 방법’, 마지막으로 각기 다른 기준을 복수 개의 변수에 적용함으로써 이상치를 확인하는 방법들이 있다.

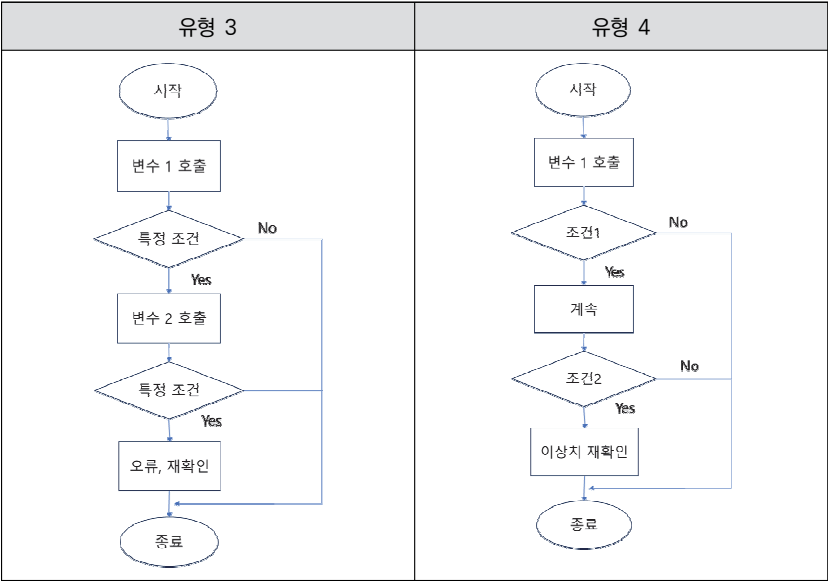
[그림 5-2] 유형 1과 2의 순서도



자료: 저자 작성

유형 3은 여러 변수에 대해 각각의 조건이 부합하면 오류로 판정한다. 특정 조건은 ‘연금 수령 나이나 수술 및 시술에 대한 의료비 수납 금액’ 같은 통상적으로 어느 정도 기준치가 정해져 있는 조건들이 해당한다. 유형 4는 같은 변수에 복수의 조건을 통해 부합 시 이상치를 재확인한다. 질의문 14, 15, 17, 46, 48, 50번에 해당된다. 본 순서도 유형에서는 ‘부채와 자산’ 같은 수치형 기준에 ‘OR 논리형 연산자’가 복합적으로 작용하여 이상치를 탐지하게 된다.

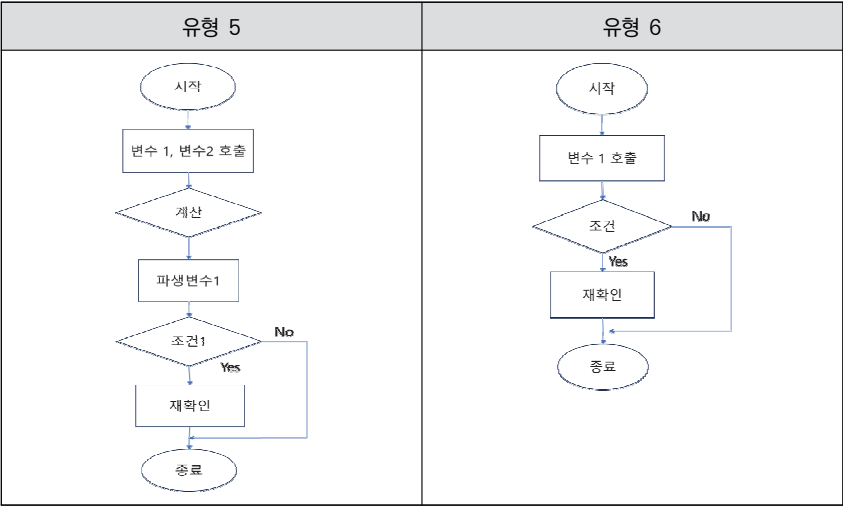
[그림 5-3] 유형 3과 4의 순서도



자료: 저자 작성

[그림 5-4]는 유형 5와 6을 나타낸다. 유형 5는 파생변수를 통해 조건 적용 후, 부합 시 재확인한다. 질의문 16, 24, 47, 49, 60번이 이에 해당된다. 본 순서도 유형에서는 복수 개의 변수값들을 활용하여 파생변수를 생성하고, 이렇게 생성된 파생변수에 수치형 비교나 특정 변수와의 일대일 대소 비교를 통해 에러를 탐지하게 된다. 유형 6은 하나의 변수에 조건이 부합하면 변수를 재확인한다. 질의문 6~9, 12, 13, 18, 19, 21~23, 31~38, 51, 52, 54~56, 58, 61~63번이 이에 해당된다. 본 순서도 유형에서는 주로 단일 변수에 대하여 수치형 비교를 통해 이상치를 탐지하게 된다.

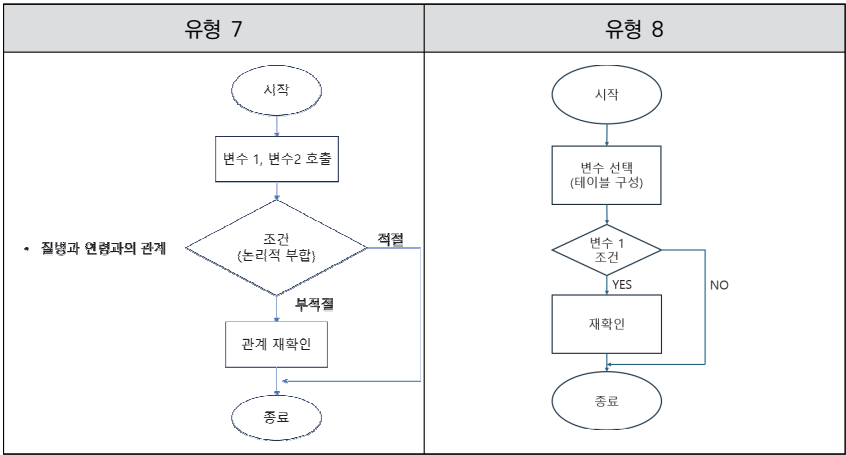
[그림 5-4] 유형 5와 6의 순서도



자료: 저자 작성

[그림 5-5]는 유형 7과 8을 나타낸다. 유형 7은 복수의 변수에 논리적 조건을 적용하여 관계가 부적절할 경우에 양 변수의 관계를 재확인한다. 질의문 39~43번이 해당된다. 본 순서도 유형에서는 복수 개의 변수에 대하여 ‘주 질병에 해당할 수 있는 나이’ 같은 부분들을 확인함으로써 이상치를 탐지하게 된다. 유형 8은 단일 변수 조건에 부합하여 이상치로 판단되는 경우에 다른 변수도 재확인한다. 질의문 3, 21, 23번이 이에 해당된다. 본 순서도 유형에서는 특정 기준치 값이나 변수의 분위수를 통해 이상치로 판단될 경우 타 변수를 재확인하도록 한다.

[그림 5-5] 유형 7과 8의 순서도



자료: 저자 작성

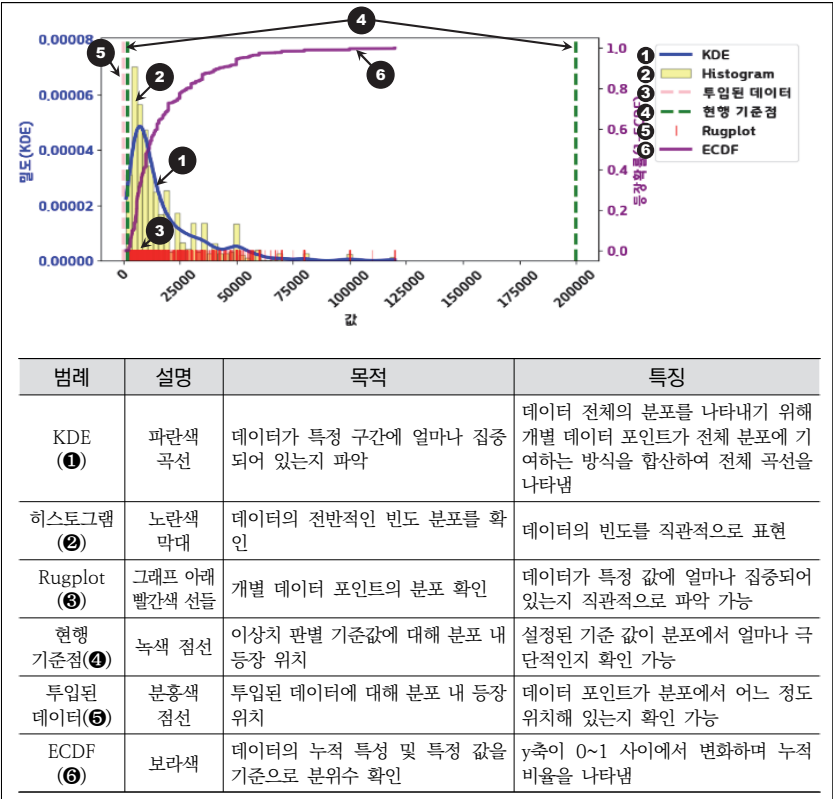
3. 유형별 적용 결과

기존 규칙에 따른 분포와 검수를 위한 8가지 유형의 질의문별 결과를 제시한다. 전체 데이터에서 이상치를 파악하기 위해 시각화 기법을 활용하고, 이상치 범위에 포함된 데이터 테이블을 함께 제공한다. 시각화 그

래프는 커널확률밀도함수(kernel density estimation, KDE), 히스토그램(histogram), ECDF(empirical cumulative distribution function), 그리고 rugplot을 사용하여, 이상치가 존재하더라도 데이터의 전체적인 분포를 직관적으로 파악할 수 있도록 하였다.

또한, <표 5-17>에서 범례에 포함된 상세한 내용을 확인할 수 있으며, 현행 기준점과 새로 유입된 데이터의 등장 확률(1-ECDF)을 계산하여 해당 데이터가 이상치인지 정상값인지 판단하는 데 도움을 주도록 하였다.

<표 5-17> 시각화 그래프에 활용되는 툴



자료: 저자 작성

총 8종의 각 유형별 2020년 조사 데이터에 적용한 결과를 살펴본다. <표 5-17>은 유형 1에 속하는 문항 번호 4번 변수에 대한 실행 결과 예시이다. [의료보장형태(HEALTH_INS)]에 해당하는 첫 번째 변수에 대해 ‘[의료보장형태]가 5 또는 6인 경우’로 필터링을 진행한 후, 두 번째 변수인 [사업소득(I_INC2)]의 수치적 분포와 시각적 분포를 확인한다.

4. 대형 언어모델(LLM)을 활용한 검수 질의문 생성

가. 관련 연구 검토

tabular data에 관련된 연구 배경과 언어모델을 활용하여 테이블 구조 데이터(tabular data)에 대해 질의응답(QA), 추론(reasoning)을 수행한 연구들을 간략히 소개하고, 한국의료패널 데이터의 태스크에 대해 적절한 방법을 모색해본다.

대형 언어모델(LLM)에 대한 제로샷 및 퓨샷 학습이 발전함에 따라, 이상 탐지 및 분포 외(out of distribution, OOD) 탐지 방법으로서 사전 학습된 대형 언어모델(LLM)을 방대한 훈련 없이 활용하는 연구가 진행되고 있다. 이러한 변화는 기존의 이상 탐지 및 OOD 탐지의 정의에서 활용되는 전통적인 학습과 평가 방법에 관한 절차가 항상 적용되지 않을 수도 있음을 시사한다. 대형 언어모델(LLM)을 활용한 이상 감지 연구에서의 정의를 간략히 살펴본다(Xu and Ding, 2024).

확률분포 P^{out} 에 대해 covariate-shift를 가지는 확률분포 P^{in} 를 고려하고, 각 샘플 x_i 는 확률분포 P^{in} 또는 P^{out} 를 따른다고 가정하자. 주어진 평가데이터 $D_{test} = \{x_1, \dots, x_n\}$ 에 대한 대형 언어모델(LLM)에 기반한 이상 탐지 방법은 사전 학습된 대형 언어모델(LLM)으로 탐지 모델 $f_{LLM}(\cdot)$ 을 구축하여 각 샘플 $x' \in D_{test}$ 가 P^{out} 에 속하는지 예측하는 것에 있다. 입력

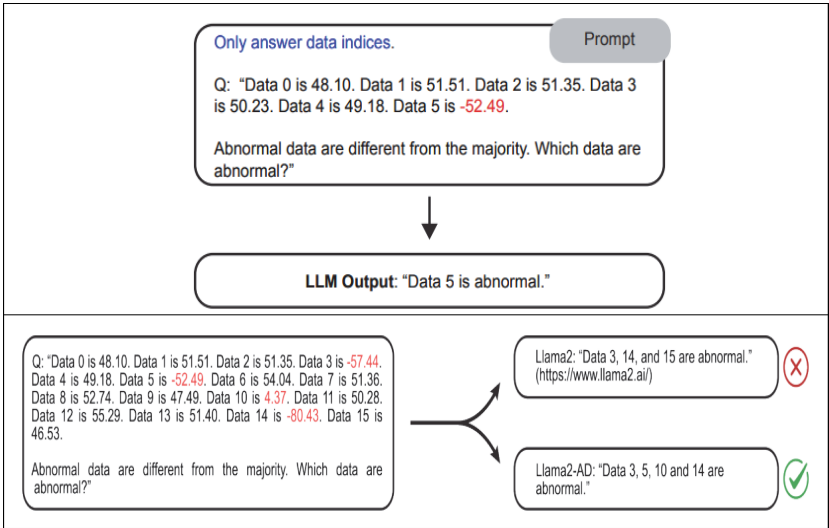
X 와 프롬프트 P 로부터 구조화된 프롬프트를 $\tilde{X} = \text{Template}(X, P)$ 로 구성하고, $Y = \text{LLM}(\tilde{X})$ 를 얻는다. 반면, OOD 탐지는 객체 수준에서 알려진 ID 클래스 중 어느 것에도 속하지 않는 사례를 식별하는 데 초점을 맞춘다고 할 수 있다.

본 고에서는 대형 언어모델(LLM)을 튜닝하기보다는 다양한 프롬프트 기법을 활용하여 탐지를 수행한다. 이상 탐지를 위한 적절한 프롬프트를 설계하기 위해 역할 기반 프롬프트, ICL, 연쇄 추론(CoT) 등의 다양한 프롬프트 기법을 조합하여 효과적인 프롬프트 템플릿을 만들고 있다. SIGLLM(Alnegheimish et al., 2024)의 연구는 대형 언어모델(LLM)을 활용한 시계열 이상 탐지를 위해 특정 역할 기반 지시를 포함한 프롬프트를 사용하여 주어진 데이터에서 이상 요소를 직접 식별하였다.

튜닝을 수행하는 대형 언어모델(LLM) 방법은 대형 언어모델(LLM)을 추가 학습시키거나 파인튜닝하여 다운스트림 탐지 작업에 적응시키는 방법이다. 그러나 전체 대형 언어모델(LLM)을 파인튜닝하는 것은 연산 비용이 많이 든다. 보다 효율적인 방법으로 PEFT(parameter-efficient fine-tuning)를 들 수 있다. 예를 들어, Li et al.(2024a)은 대형 언어모델(LLM)을 활용하여 이상 탐지 방법을 제안하였다. 사전 훈련된 대형 언어모델(LLM)이 추가적인 모델 조정 없이도 배치 수준에서 이상치를 탐지할 수 있음을 보여준다([그림 5-6] 참고). 실제 자료와 생성한 자료를 증강하여 만든 자료를 활용하여 이상치 탐지 성능을 개선하였다. 이상값이 포함된 가상의 합성데이터(synthetic data)를 만들어 대형 언어모델(LLM)의 학습 데이터로 활용하고, Llama2와 Mistral을 LoRA(Low-rank adaptation, Hu et al., 2022) 방식으로 파인튜닝하였다. 테이블 데이터를 대형 언어모델(LLM)이 이해할 수 있는 텍스트 형식(json, markdown 등)으로 변환하여 데이터 직렬화(serialization)를 하고, “이상값을 감지해줘” 같은 명확한 질문으로 프롬프트를 구성하였다. [그림 5-6]은 한 번

수의 5개의 데이터에 대한 데이터를 텍스트화(데이터 직렬화)하고 “Abnormal data are different from the majority. Which data are abnormal?”의 질의문에 대한 abnormality 유무에 대한 판별을 대형 언어모델(LLM)로 출력한 것을 나타낸다.

[그림 5-6] 샷 및 학습 전후의 이상치 탐지 예시

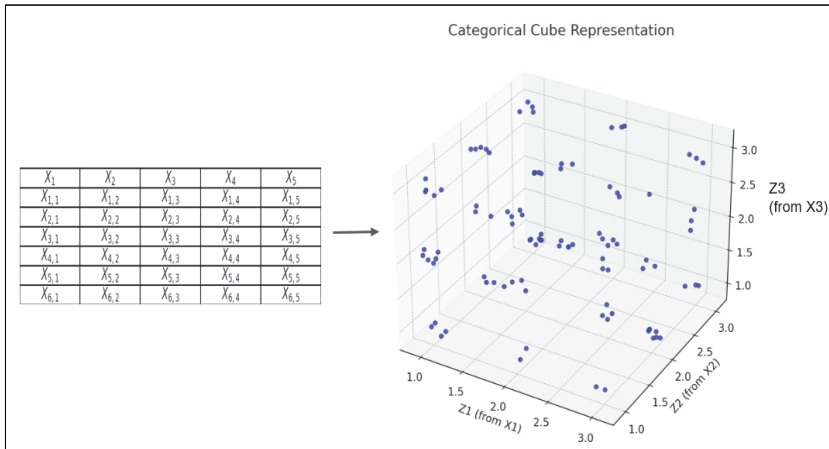


자료: Li et al. (2024a). Anomaly detection of tabular data using llms. arXiv preprint arXiv:2406.16308

앞에서 소개한 대부분의 연구들은 데이터 구조 데이터(tabular data)로부터 표(table)를 생성하는 태스크라기보다 요약된 통계표(table)로부터 표의 여러 셀(cell)의 내용들이 포함된 질의에 대한 추론을 수행하는 태스크(reasoning task)를 다룬다고 할 수 있다. 반면, 본 고에서의 분석 대상이 되는 한국의료패널 데이터는 기본적으로 요약된 통계표라기보다는 원시자료에 가까운 2차원 텐서(tensor)인 행렬자료 구조를 띠고 있다고 할 수 있으므로, 새로운 방안이 필요하다고 할 수 있다.

이러한 목적에 부합하는 태스크를 수행할 수 있는 방법을 고안하여 적용해보기로 한다. 우선, 데이터에서 검수되는 샘플들을 선별하는 과정을 정식화해보기로 하자. 예를 들어, x_1 을 연속형인 원변수에서 l 개의 구간으로 구성된 수준(level) $a_1, a_2, \dots, a_l$ 을 가진 순서범주형으로 변환된 경우라고 놓고, x_2 를 m 개의 수준, $b_1, b_2, \dots, b_m$ 인 범주형 변수라 놓자. n 을 샘플 수, p 를 변수 수라고 할 때, $n \times p$ 크기를 가진 테이블 구조 데이터로부터 임의의 $l \times m$ 크기의 표를 생성할 때, 이 중 빈도수(혹은 확률)가 0 이라고 예상되었던 특정 셀이 일정의 확률을 가질 때, 이 셀의 결합 조건 ($x_1 = a_l, x_2 = b_m$)에 속하는 샘플들이 이상치 확인 대상 혹은 검수 대상이 될 수 있음을 의미한다고 할 수 있다.

[그림 5-7] 테이블 구조 데이터와 세 변수의 3차원 텐서



자료: 저자 작성

3개 이상의 변수로 확장할 경우를 고려해보자. 이는 2차원 이상의 텐서(tensor) 구조를 가진다고 볼 수 있는데, 셀은 여러 변수들이 결합이 되는 텐서라 할 수 있다. 예를 들어, 세 개 변수인 경우 큐브 형태의 텐서가 된다. <그림 5-7>은 $n \times p$ 크기의 테이블 구조 데이터로부터 $l_1 \times l_2 \times l_3$ 크기의 텐서를 추출하는 과정을 나타낸다. 결국, 연속형 데이터가 포함된 경우, 이를 구간을 가지는 범주형 변수로 변환해야 하며, 이 과정에서 구간 경계를 설정하기 위한 밀도 추정 등 데이터 분포에 대한 고려가 필요하다. 이는 대형 언어모델(LLM)이 문맥을 통해 이러한 분포를 이해하도록 하는 최적화 연산이 요구됨을 시사한다.

나. 자료 구조

여러 데이터가 산재되어 있어, 고유 ID를 기준으로 데이터를 취합하는 것이 어렵다. 또한, ID를 기준으로 데이터를 병합할 경우, 변수의 정보가 없는 데이터에 대해서는 대량의 NA가 기계적으로 저장됨을 의미한다. <표 5-18>은 변수와 파생변수 자료와 json 표현을 나타내는데, 대형 언어모델(LLM)의 입력은 텍스트 구조 형태를 다루는 대형 언어모델(LLM)의 입력으로, 테이블 구조 데이터는 json 표현과 같이 데이터 직렬화(serialization)가 필요하다고 할 수 있다.

〈표 5-18〉 변수와 파생변수 및 json 표현

tabular 구조				
ID	v1	v2	v11	v12
101	1	5.3	0.1	0.9
102	0	3.2	NA	NA
103	1	4.8	0.5	1.2
104	0	2.7	NA	NA

<pre>import pandas as pd import json # tabular 데이터 data = { "id": [101, 102, 103, 104], "v1": [1, 0, 1, 0], "v2": [5.3, 3.2, 4.8, 2.7], "v11": [0.1, None, 0.5, None], "v12": [0.9, None, 1.2, None] } df = pd.DataFrame(data) # v1 값에 따라 Json 구성 def row_to_json(row): result={"id":row["id"], "v1":row["v1"], "v2":row["v2"]} if row["v1"]==1: #v1=1인 경우에만 v11, v12 추가 result["v11"] = row["v11"] result["v12"] = row["v12"] return result json_data=[row_to_json(row) for _, row in df.iterrows()] # Json 파일로 저장 with open("output.json", "w") as json_file: json.dump(json_data, json_file, indent=4)</pre>	<pre>[{ "id": 101, "v1": 1, "v2": 5.3, "v11": 0.1, "v12": 0.9}, { "id": 102, "v1": 0, "v2": 3.2}, { "id": 103, "v1": 1, "v2": 4.8, "v11": 0.5, "v12": 1.2}, { "id": 104, "v1": 0, "v2": 2.7}]</pre>
--	---

v1 값에 따라 조건적 필드인 v11과 v12가 json 객체에 포함되므로, v11과 v12가 없는 경우 해당 필드는 제외하여 간결하게 표현됨을 확인할

수 있다. 또한, 변수의 메타정보를 모형에 반영하고자 할 경우 json 형식은 텍스트 형식으로 표현 시 비교우위점을 가질 수 있는데, 예를 들어 ‘v1’을 ‘보기 문항 내용’이라고 할 경우, 메타정보는 <표 5-19>와 같다.

<표 5-19> 가구데이터(HH)의 사망 가구원 포함 여부(DEATH_YN) 변수에 대한
‘보기 문항 내용’과 ‘사용 주의사항’ 변수의 내용을 반영한 json 자료 구조

(0) 사망 가구원이 포함되지 않은 가구 (1) 사망 가구원이 포함된 가구	{“사망 가구원이 포함된 가구”로 응답 시, 사망 가구원의 경우 설문 영역별로 대표 응답자가 대리 응답함,
--	---

자료: 저자 작성

다. 검수 질의문의 확률분포

평가를 위해 수행한 60개의 검수 질의문에 대한 확률분포는 다음과 같다. 범위는 0과 19.94%이며, 최대값 19.94%인 검토 질의문은 57번으로 다음과 같으며, 재확인 용도의 질문으로 파악되었다.

[건강-최근 일주일 하루 걸은 시간] 이상치 재확인 * 응답값이 명확하게 입력된 값인 경우, 오타인지 재확인(보통 응답할 때 5단위 혹은 10단위로 응답하므로, 11분, 17분, 23분, 등의 경우 재확인함.)
--

Q1은 0%, Q2는 0.0136%, Q3는 0.09026%로 나타났다. 기존 검수 질의문의 누적 75%가 0.09026%에 해당하며, 이는 대부분의 검수 질의문이 약 0.001 이하의 희박한 검수 빈도를 가진다는 것을 의미한다.

라. 분포 인지 RAG 기반 알고리즘

본 연구에서는 테이블 구조 데이터에서 여러 변수 간의 교차빈도표를 분석하여, 특정 셀의 확률이 0으로 예상되었음에도 불구하고 0이 아닌 경우로 의심되는 셀을 탐색하고자 한다. 앞에서 언급한 바와 같이, 이러한 문제를 ‘희귀한 테이블 구조 내의 텐서(rare tabular tensor)’ 탐지 문제로 정의할 수 있다.

먼저, 언어모델이 검수 질의문의 논리적 오류를 탐지할 수 있는지를 확인하기 위해 ChatGPT에 간략히 질의하고 답변을 받아보았다. <표 5-20>의 결과에 따르면, 논리적 오류 검토 시 일부 예시 질의문에서 오류를 탐지할 수 있는 것으로 나타났다. 이에 따라, 이러한 오류 탐지 문제를 해결하기 위한 방법으로 검수 질의문과 유사한 질의문을 생성하는 방식을 적용해보기로 한다

<표 5-20> 검수 질의문에 대한 논리적 오류 검토

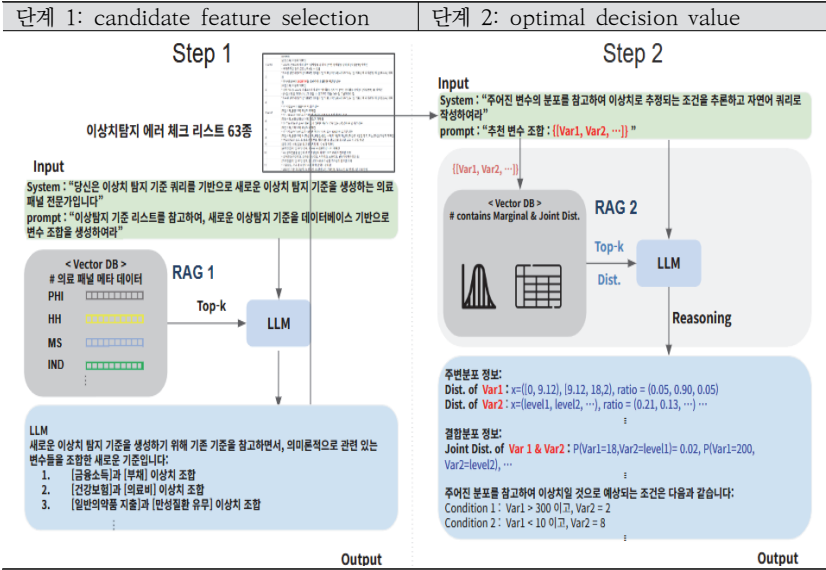
Q	[재산소득_금융소득]은 0원이 아니고 [재산소득_부동산/동산 소득]도 0이 아닌데 [자산]이 0원인 경우는 논리적 오류가 있나요?	[지출_소비지출]과 [지출_비소비지출]의 합이 월평균 소득보다 클 수 있는가요?
A	네, 논리적 오류가 있습니다. 자산은 사람이 소유하고 있는 재산의 총합을 의미합니다. 이에는 부동산, 동산, 금융 자산 등이 포함됩니다. 질문에 따르면 "[재산소득_금융소득]"과 "[재산소득_부동산/동산 소득]"이 0이 아닌 상황에서 자산이 0원이라는 것은 논리적으로 맞지 않습니다. 소득이 발생한다는 것은 그 소득의 근거가 되는 자산이 존재한다는 것을 의미합니다. 예를 들어, 금융소득이 발생했다면 그 소득은 금융 자산(예: 예금, 주식 등)에서 나온 것이며, 부동산/동산 소득이 발생했다면 이는 해당 부동산이나 동산이 존재해야 가능한 일입니다. 따라서 금융 소득과 부동산/동산 소득이 0이 아닌데 자산이 0인 것은 잘못된 데이터 입력이거나 계산상의 오류일 가능성이 큼니다.	네, [지출_소비지출]과 [지출_비소비지출]의 합이 월평균 소득보다 클 수 있습니다. 이는 다음과 같은 이유로 발생할 수 있습니다. 가계가 저축해둔 돈이나 기존 자산(예: 예금, 주식, 부동산 처분)을 지출에 활용할 수 있습니다. 이런 경우, 당장의 소득보다 지출이 더 많아질 수 있습니다. 대출 및 신용카드 사용으로 지출을 늘릴 수 있습니다. 이는 소득보다 더 많은 지출을 가능하게 만듭니다. 일시적인 지출 증가로, 가구가 일시적으로 큰 지출(예: 병원비, 결혼식 비용 등)을 감당해야 하는 경우, 소득보다 지출이 더 커질 수 있습니다. 또는 외부에서 지원받은 돈이나 가족/친구로부터의 금전적 도움으로 인해 소득 이상으로 지출할 수 있습니다. 따라서 소득보다 지출이 더 클 수 있으며, 이는 비정상적이거나 오류가 있는 것이 아니라 다양한 재정적 수단을 활용하는 상황에서 발생할 수 있는 일입니다.

자료: 저자 작성

논리적 오류 검토 시 예시 질의문에 대한 논리적 오류를 일부 탐지할 수 있는 것으로 조사되었다. 이러한 오류를 탐지하는 문제에 대한 예시로 검수 질의문과 유사한 질의문을 생성하는 방법을 적용해보기로 한다. 이를 위하여, 첫 번째로, 한국의료패널 데이터의 조사 목적 등 관련된 언어적 특징을 문맥으로 반영하기 위하여 자기주도학습(self-supervised learning)을 수행하였다. 자기주도학습을 위한 학습 데이터셋은 입력을 변수설명으로, 출력은 변수명(한글명)으로 설정하였다. <표 5-21>은 2단계로 이루어진 본 연구에 적용할 알고리즘을 나타낸다. 단계 1에서는 유사한 질의문들과 유사한 질의들을 추출하기 위해 ChatGPT를 활용하여 가구원데이터 관련 변수 334종과 관련된 메타정보를 반영하기 위해, 변수들의 주변분포(marginal distribution)를 RAG 기법에서 활용되는 벡터 데이터베이스(벡터 DB)로 구성하였다.

다음으로, 언어모델을 적용하기 위한 랭체인을 <표 5-21>과 같이 구성하였으며, 두 단계에 대한 프롬프트 예시와 랭체인을 나타낸다. 먼저, ChatGPT에 적절한 프롬프트를 구성하여 유사한 질의문을 생성하기 위하여, 검수 질의문 60여 개를 instruction 예제로 제시하고, 유사한 질의문을 생성하는 태스크를 수행하는 ICL 방법을 적용하였다. 단계 2에서는 유사한 질의문에서 출현한 feature들에 대한 분위수 값을 생성하기 위한 값들을 학습하기 위해 해당 변수들의 주변분포와 결합분포(joint distribution)을 벡터 DB로 구축하였다. 단계 1의 주변분포 파악용 벡터 DB와 단계 2의 결합분포 인지용 벡터 DB([그림 5-8] 참조)를 각각 RAG_marginal, RAG_joint로 명명하기로 한다. 생성된 유사 질문리스트에 대해 전체 데이터셋에서의 확률을 산출하였다.

〈표 5-21〉 분포 인지 RAG 기반 검수 질의문 추천 알고리즘



〈표 5-21〉은 RAG를 활용한 새로운 이상치 질의문 추천의 과정을 나타내며, 두 개의 단계로 구성되어 있다. 단계 1에서는 이상치 탐지 기준 리스트를 기준으로 이상치가 존재할 수 있는 새로운 변수 조합을 추천하며, 단계 2에서는 단계 1에서 추천된 변수 조합을 기반으로 이상치로 예상되는 변수값의 범위(또는 조건)를 생성한다.

단계 1에서

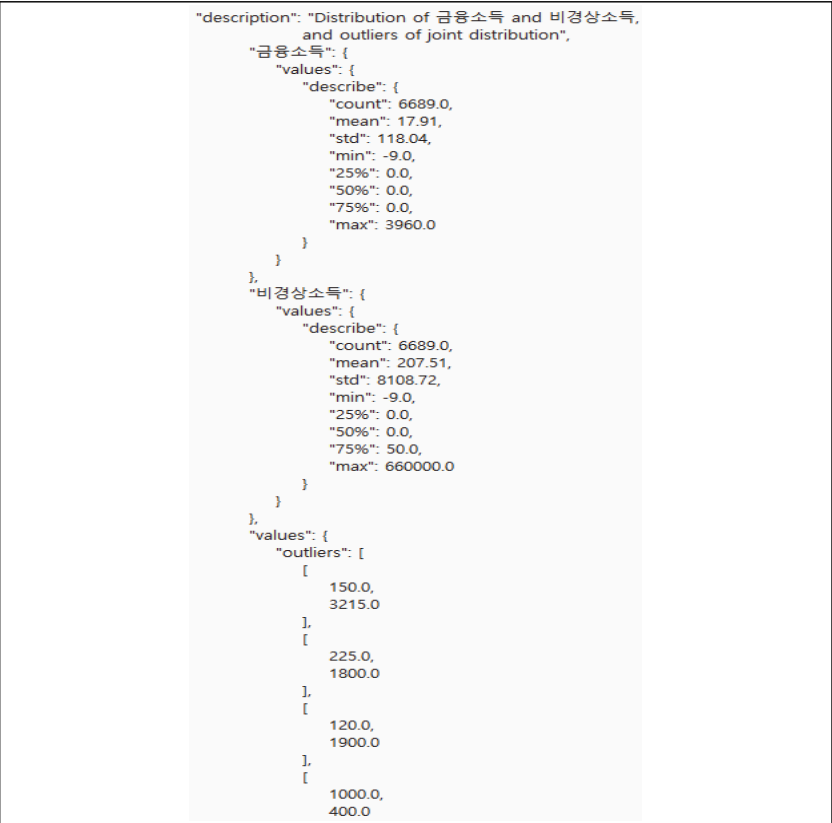
주어진 이상치 탐지 기준 리스트를 참고하여, 새로운 이상치 기준을 데이터베이스 기반으로 변수 조합을 생성하여라.

와 같은 프롬프트가 주어지면, 기존에 정의된 63종의 이상 탐지 기준을 참고하여 RAG 1이 벡터 DB에서 가장 관련성이 높은 이상치 탐지 기준을 검색한다. 대형 언어모델(LLM)은 의미적으로 적절한 새로운 조합을 도출하며, 이 과정에서 데이터의 의미적 특성을 반영한 맞춤형 이상치 탐

지 변수 조합을 자동 생성하는 데 활용되었다.

단계 2는 앞선 단계에서 추천된 변수 조합으로부터 RAG_marginal, RAG_joint 내에서 유사성이 높은 documents를 찾고(RAG 2, 그림 5-8 참조), 유사한 documents에서 서술된 내용을 바탕으로 분포 인지 기반의 cut-off-value를 생성한다.

[그림 5-8] RAG 2에서 활용되는 DB 예시



자료: 저자 작성

〈표 5-22〉 단계 1에서의 RAG 1 이상치 탐지 기준 생성 예시

In put	위 이상치 탐지 기준 리스트 60개를 참고하되, 리스트에는 없는 새로운 이상치 탐지 기준을 벡터 데이터베이스를 기반으로 한글 변수명을 통해 2개의 의미론적인 변수 조합을 생성하여라.
Out put	다음은 제공된 리스트를 기반으로 새롭게 생성한 이상치 탐지 기준입니다. 각 기준은 서로 다른 변수를 의미 있게 조합하여 설정하였습니다: 1. [근로소득] & [사업소득] 조합: [근로소득]과 [사업소득]이 모두 0으로 보고되었으나, 개인이 고용 상태 또는 자영업자로 등록되어 있는 경우 오류 가능성이 있습니다.
	2. [재산소득_금융소득] & [재산소득_부동산/동산 소득] 조합: 두 소득 유형이 모두 매우 높게 보고되었으나(예: 금융소득 > 1억, 부동산/동산 소득 > 2억), [자산]이 낮거나 0으로 보고된 경우 불일치가 있을 수 있습니다.
	3. [공적연금] & [기초연금] 조합: 개인이 [공적연금]과 [기초연금]을 모두 수령하지만, 나이가 일반적인 수급 연령(예: 65세 미만)에 맞지 않는 경우 오류로 간주될 수 있습니다.
	4. [의료비_수납 금액] & [의료비_비급여 금액] 조합: [의료비_수납 금액]이 매우 낮고 [의료비_비급여 금액]이 매우 높은 경우, 카테고리의 오해 또는 보고 오류일 수 있습니다.
	5. [교통비] & [의료비_수납 금액] 조합: 의료 방문을 위한 [교통비]가 [의료비_수납 금액]에 비해 매우 높은 경우, 교통비 보고에 오류가 있을 수 있습니다.
	6. [보험료] & [보험금 수령액] 조합: [보험료]가 지속적으로 높게 보고되지만 [보험금 수령액]이 0이거나 매우 낮은 경우, 보험 청구 또는 보고에 문제가 있을 수 있습니다.
	7. [신체활동별 비용] & [건강보조식품_복용] 조합: [신체활동별 비용]이 높지만 [건강보조식품_복용]이 0인 경우, 건강 관련 지출 습관에 불일치가 있을 수 있습니다.
	8. [와병일수] & [결근결석일수] 조합: [와병일수]가 높지만 [결근결석일수]가 낮은 경우, 질병 보고와 실제 결근 사이에 불일치가 있을 수 있습니다.
	9. [소득_소비지출] & [소득_비소비지출] 조합: [소득_소비지출]과 [소득_비소비지출]의 합이 보고된 소득을 크게 초과하는 경우, 소득 또는 지출 보고에 오류가 있을 수 있습니다.
	10. [장기요양 및 기타 돌봄 서비스 비용] & [산후조리원 비용] 조합: 두 비용이 동시에 높게 보고된 경우, 돌봄 관련 비용의 중복 또는 분류 오류가 있을 수 있습니다.
	이 기준들은 서로 다른 변수 간의 논리적 관계를 통해 잠재적인 불일치나 오류를 식별하는 데 목적이 있습니다.

자료: 저자 작성

〈표 5-23〉 단계 2에서의 RAG 2 이상치 범위 생성 예시

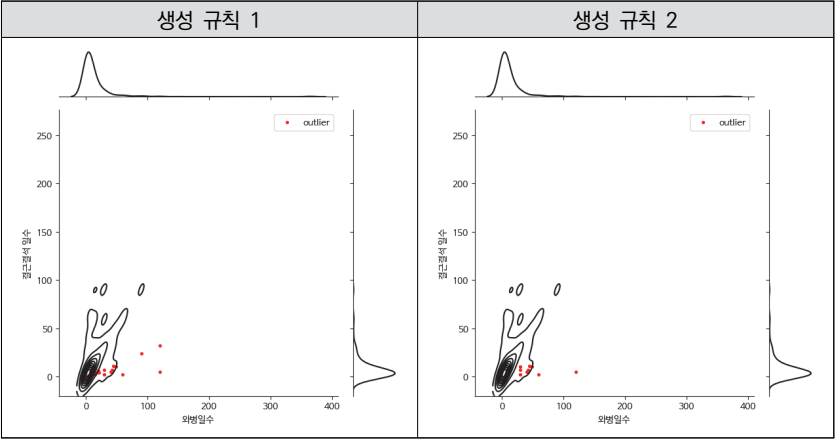
Input	8번 변수 조합 기반으로 분포를 확인하고, 확인한 정보를 바탕으로 이상치 조건(cut-off value)을 새로 추론하여라.
Output	주변분포 정보: Dist. of [와병일수]: $x = ([0, 10.34), [10.34, 26.28), [26.28, \infty))$, ratio = (0.68, 0.25, 0.07)[1] Dist. of [결근결석일수]: $x = ([0, 11.59), [11.59, 32.93), [32.93, \infty))$, ratio = (0.70, 0.25, 0.05)[2] 결합분포 정보: Joint Dist. of [와병일수] & [결근결석일수]: $P(\text{와병일수}=120, \text{결근결석일수}=5)=0.01$, $P(\text{와병일수}=60, \text{결근결석일수}=90)=0.01$, ...[3] 주어진 분포를 참고하여 이상치일 것으로 예상되는 조건은 다음과 같습니다: Condition 1: [와병일수] > 3 * [결근결석일수] Condition 2: [와병일수] > 26.28 (상위 7%) 이고, [결근결석일수] < 11.59 (하위 70%) 이러한 조건은 [와병일수]가 상대적으로 높지만 [결근결석일수]가 낮은 경우를 이상치로 간주하여, 질병 보고와 실제 결근 사이의 불일치를 탐지하는 데 사용됩니다.

자료: 저자 작성

이러한 절차가 가지는 장점은 데이터 자체의 확률분포뿐만 아니라, 외부의 자료, 기관의 통계자료 등으로 참고하는 document를 확장 및 수정할 수 있다는 점이다. 추후, 정보가 확장되고, 최신화됨에 따라 강건한 이상치 탐지 질의문을 생성할 수 있을 것으로 기대한다.

[그림 5-9]는 생성된 질의문이 한국의료패널 데이터에서 보이는 확률을 나타낸다. 생성 규칙 1은 와병일수(질병이나 상해로 인해 일상적인 활동을 거의 하루 종일 또는 만나질 이상 못하는 일수)가 결근결석일수의 3배 이상인 조건으로 31건(0.1869%)이며, 생성 규칙 2는 와병일수가 26.28일 초과이면서 결근결석일수가 11.59일 미만인 분포 기반의 조건으로 11건(0.0663%)이다. 실제로 희소하게 일어난 조건들이었으며, 이상치일 가능성이 클 것으로 예측된다.

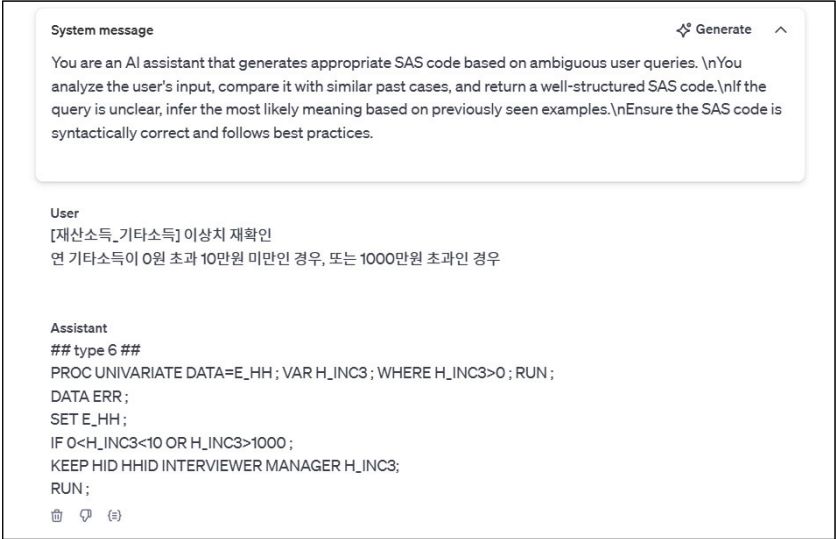
[그림 5-9] 생성된 규칙에 대한 2차원 결합분포와 주변분포



자료: 저자 작성

최종적으로, 60개의 질의문과 해당 SAS 스크립트 명령문을 각각 입력과 출력으로 설정한 훈련 데이터를 활용하여 대형 언어모델(LLM)을 파인 튜닝하였으며, 이를 통해 질의에 대한 코드 생성 태스크인 query-to-SAS_code를 수행할 수 있도록 하였다. [그림 5-10]은 질의문을 입력으로 받아 SAS 프로그램을 생성한 결과를 나타낸다. 대형 언어모델(LLM)은 자연어로 된 질의문을 SAS 코드 스크립트로 변환할 수 있으며, 본 절에서는 예러 체크리스트에 포함된 질의문과 SAS 코드를 기반으로 학습된 언어모델의 수행 결과를 [그림 5-10]에 제시하였다.

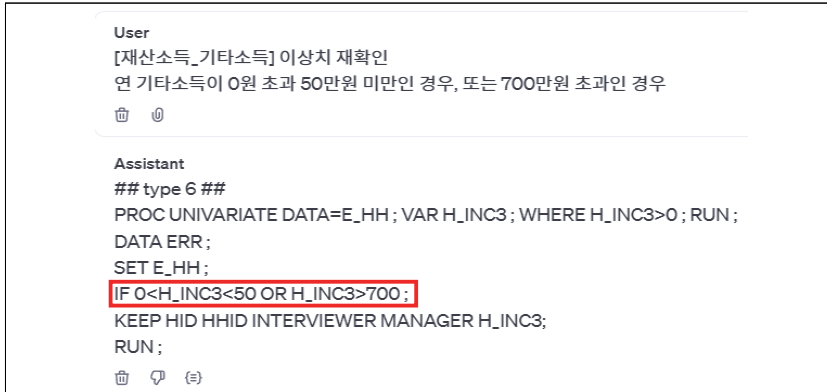
[그림 5-10] SAS 코드 생성 실험 결과



자료: 저자 작성

학습한 언어모델은 한국의료패널 데이터의 영문 변수명과 SAS 코드 형식을 학습하고, 이를 활용하여 SAS 코드를 작성할 수 있었다. 특히, [그림 5-11]에서 ‘연 기타소득’의 이상치 범위가 바뀐 경우에서도 강건한 코드 생성 능력을 확인하였다.

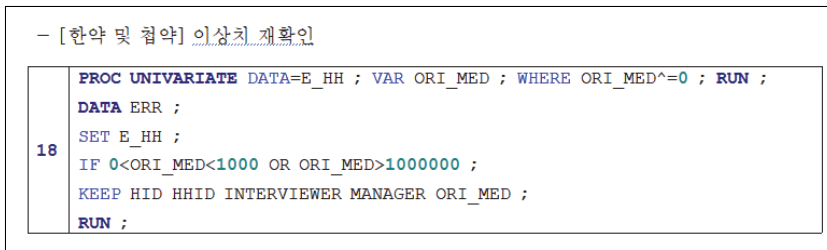
[그림 5-11] 이상치 범위 변경 후, SAS 코드 생성 결과



자료: 저자 작성

그러나 사전 검수 질의문인 에러 체크리스트에 수록된 63종 이외의 이상치 검토 질의문에 대해서는 실행되지 않는 코드를 실행하는 모습을 보였으며, 이는 지도 학습에서 학습데이터의 부족이 원인으로 판단된다. 또한, 에러 체크리스트에 수록된 질의문과 SAS 코드 사이의 정보가 서로 일관되지 않는다는 문제가 있었다.

[그림 5-12] SAS 코드 대비 질의문의 정보 제한 예시



자료: 저자 작성

특히, 예를 들어 [그림 5-12]에서와 같이 여러 체크리스트에 명시된 질의문만으로는 SAS 코드 생성에 필요한 충분한 정보가 제공되지 않거나, [그림 5-13]과 같이 제공된 정보가 정확하지 않은 경우가 있어, 코드 생성 과정에서 어려움이 발생했다. 이러한 문제는 추후 연구를 통해 개선할 필요가 있을 것이다.

[그림 5-13] 질의문과 SAS 코드 간의 정보 불일치 예시

-	“가구원 일반사항2”의 [의료보장 형태]가 ‘⑤ 의료급여(1,2종)세대주’ 또는 ‘⑥ 의료급여 세대원’일 때 [근로소득] 재확인
*	연 근로소득이 2000만원을 초과하거나 50만원 미만인 경우
2	<pre>PROC UNIVARIATE DATA=E_IND ; VAR I_INC2 ; WHERE HEALTH_INS IN (5,6) ; RUN ; DATA ERR ; SET E_IND ; IF I_INC2>=1000 AND HEALTH_INS IN (5,6) ; KEEP HID HHID PID PIDWON INTERVIEWER MANAGER AGE I_INC2 HEALTH_INS; RUN ;</pre>

자료: 저자 작성

〈표 5-24〉 유형 1) 문항 번호 4

요점	“가구원 일반사항 2”의 [의료보장 형태]가 ‘⑤ 의료급여(1, 2종) 세대주’ 또는 ‘⑥ 의료급여 세대원’일 때 [사업소득] 재확인 <pre># [파일이름, 조건, 타겟변수, 타겟변수의 현행 기준, 새로 확인할 데이터] input_list_4 = ['E_IND', 'HEALTH_INS in [5,6]', 'I_INC2', [1000], 800]</pre>
세상	실행결과 <pre>input_list = input_list_4 type1_2_main(input_list=input_list, var_num=4)</pre> 데이터 통계 정보 조건 : HEALTH_INS in [5,6] I_INC2데이터 분포: Count: 14 Mean: 595.71 Std: 613.80 Min: 100.00 25%: 192.50 50%: 300.00 75%: 675.00 Max: 2000.00 Percentile 위치: 1000 : 하위 85.71%에 해당 Percentile 위치: 800 : 하위 78.57%에 해당 문항번호 4 kde이상치 탐지 (등장확률(1-ECDF) = 0.2143)

자료: 저자 작성

〈표 5-25〉 유형 2) 문항 번호 27-1

입력1 (방문)	교통수단이 '⑤ 기차/비행기/고속(시외)버스'인데 [교통비]가 20만 원 초과, 1,800원 미만이면 재확인(* 7세 이하 제외) <pre>case1(병원 방문 교통수단, 방문 교통비) • M_TRANS1==5 # [파일이름, 조건1, 조건2, 타겟변수, 타겟변수 형형 기준, 새로 확인 할 데이터] input_list_27_case1 = ["E_MS", "AGE>7", "M_TRANS1==5", ["M_TRANS1_EXP1", "M_TRANS1_EXP2"], [[1800, 200000], [1800, 200000]], 300]</pre>
출력 1-1 (본인 방문)	실행결과 <pre>input_list = input_list_27_case1 type2_3_main(input_list=input_list, var_num=27) ✓ 0.9%</pre> 데이터 통계 정보 조건 : AGE>7 and M_TRANS1==5 M_TRANS1_EXP1데이터 분포: Count: 2425 Mean: 18109.91 Std: 17762.08 Min: 1300.00 25%: 6100.00 50%: 11300.00 75%: 24400.00 Max: 120000.00 Percentile 위치: 1800 : 하위 0.04%에 해당 200000 : 하위 100.00%에 해당 Percentile 위치: 300 : 하위 0.00%에 해당 <Figure size 300x500 with 0 Axes> 문항번호 27 kde이상치 탐지 (등장확률(1-ECDF) = 1.0000)

자료: 저자 작성

〈표 5-26〉 유형 2) 문항 번호 27-2

입력1 (방문)	교통수단이 '⑤ 기차/비행기/고속(시외)버스' 인데 [교통비]가 20만 원 초과, 1,800원 미만이면 재확인(* 7세 이하 제외) case1(병원 방문 교통수단, 방문 교통비) • M_TRANS1==5 # [파일이름, 조건1, 조건2, 타겟변수, 타겟변수 형형 기준, 새로 확인 할 데이터] input_list_27_case1 = ['E_MS', 'AGE>7', 'M_TRANS1==5', ['M_TRANS1_EXP1', 'M_TRANS1_EXP2'], [[1800, 200000], [1800, 200000]], 300]
출력 1-2 (동반자 방문)	데이터 통계 정보 조건 : AGE>7 and M_TRANS1==5 M_TRANS1_EXP2데이터 분포: Count: 2425 Mean: 4714.23 Std: 14031.29 Min: 0.00 25%: 0.00 50%: 0.00 75%: 0.00 Max: 120000.00 Percentile 위치: 1800 : 하위 82.76%에 해당 200000 : 하위 100.00%에 해당 Percentile 위치: 300 : 하위 82.76%에 해당 (figure size 300x500 with 0 Axes> 문항번호 27 kde이상치 탐지 (등강확률(1-ECDF) = 0.1724)

자료: 저자 작성

〈표 5-27〉 유형 2) 문항 번호 27-3

입력2 (귀가)	교통수단이 '⑤ 기차/비행기/고속(시외)버스'인데 [교통비]가 20만 원 초과, 1,800원 미만이면 재확인(* 7세 이하 제외) case2(병원 귀가 교통수단, 귀가 교통비) • M_TRANS2==5 <pre># [파일이름, 조건1, 조건2, 타겟변수, 타겟변수 형형 기문, 새로 확인 할 데이터] input_list_27_case2 = ['E_MS', 'AGE>7', 'M_TRANS2==5', ['M_TRANS2_EXP1', 'M_TRANS2_EXP2'], [[1800, 200000], [1800, 200000]], 300]</pre>
출력 2-1 (본인 귀가)	<pre>input_list = input_list_27_case2 type2_3_main(input_list=input_list, var_num=27)</pre> ✓ 0.9% 데이터 통계 정보 조건 : AGE>7 and M_TRANS2==5 M_TRANS2_EXP1데이터 분포: Count: 2369 Mean: 18027.06 Std: 17836.67 Min: 1300.00 25%: 6050.00 50%: 11000.00 75%: 24800.00 Max: 120000.00 Percentile 위치: 1800 : 하위 0.04%에 해당 200000 : 하위 100.00%에 해당 Percentile 위치: 300 : 하위 0.00%에 해당 Figure size 300x500 with 0 Axes 문항번호 27 kde이상치 탐지 (동강확률(1-ECDF) = 1.0000)

자료: 저자 작성

유형 2에 속하는 문항 번호 27번 문항에 대한 실행 결과로, 방문일 때
와 귀가일 때로 케이스를 나누어 판단해야 한다. 방문의 경우 [나이
(AGE)]>7과 [방문교통수단(M_TRANS1)]=5에 해당하는 조건으로 필터

링을 한 뒤, [본인 방문 교통비(M_TRANS1_EXP1)]와 [동반자 방문 교통비(M_TRANS1_EXP2)] 각각의 분포를 확인한다. 귀가 케이스도 같은 방법을 적용하여 이상치를 확인한다.

〈표 5-28〉 유형 2) 문항 번호 27-4

입력2 (귀가)	교통수단이 '⑤ 기차/비행기/고속(시외)버스'인데 [교통비]가 20만 원 초과, 1,800원 미만이면 재확인(* 7세 이하 제외) <pre>case2(병원 귀가 교통수단, 귀가 교통비) • M_TRANS2==5</pre> # [파일 이름, 조건1, 조건2, 타겟 변수, 타겟 변수 유형, 기존, 새로 확인 할 데이터] input_list_27_case2 = [['E_MS', 'AGE>7', 'M_TRANS2==5', ['M_TRANS2_EXP1', 'M_TRANS2_EXP2'], [[1800, 200000], [1800, 200000]], 300]
출력 2-2 (동반자 귀가)	데이터 통계 정보 조건 : AGE>7 and M_TRANS2==5 M_TRANS2_EXP2 데이터 분포: Count: 2369 Mean: 4543.12 Std: 13953.27 Min: 0.00 25%: 0.00 50%: 0.00 75%: 0.00 Max: 120000.00 Percentile 위치: 1800 : 하위 83.58%에 해당 200000 : 하위 100.00%에 해당 Percentile 위치: 300 : 하위 83.58%에 해당 Figure size 300x500 with 0 Axes 문항번호 27 kde이상치 탐지 (등장확률(1-ECDF) = 0.1642)

자료: 저자 작성

유형 3에 속하는 문항 번호 11번 변수에 대한 실행 결과이다. 이 문항의 경우, 확실한 오류를 확인할 수 있는 '연금 수령 나이', '수급 금액'이 정해져 있어 해당 오류 조건에 해당하는 데이터를 확인할 수 있도록 오류 데이터 테이블을 반환한다.

〈표 5-29〉 유형 3) 문항 번호 11

[기초연금]이 '㉠'인 경우, 만 나이가 65세 이상인 가구원이 없으면 오류

입력

```
# [과일이름, 조건1, 조건2, ..., 조건n]
input_list_11 = ["E_IND",
                 "(CNT.isna() and H_INC4_1_YN > 0)",
                 "(CNT == 1 and H_INC4 > 389)",
                 "(CNT == 2 and H_INC4 > 622)",
                 "(CNT == 3 and H_INC4 > 1164)"]
]
```

문항번호 11

오류확인-----

필터링 조건 : (CNT.isna() and H_INC4_1_YN > 0)

Empty DataFrame

Columns: [HHID, H_WGC, H_WSC, DATE_Y, DATE_M, DATE_D, REGION1, REGION2, FAM_N, IND_N, DEATH_YN, ALL_CON, Index: []

[0 rows x 77 columns]

출력

필터링 조건 : (CNT == 1 and H_INC4 > 389)

	HHID	H_WGC	H_WSC	DATE_Y	DATE_M	DATE_D	REGION1	REGION2	FAM_N	IND_N	...	HEXP2_1	HEXP2_2
115	135235011	1384	0	2020	7	17	30	1	3	2	...	NaN	NaN
138	143487011	2169	1	2020	6	18	41	1	2	1	...	NaN	NaN
153	145020011	627	0	2020	7	24	43	1	2	1	...	1	1
160	145254011	1800	1	2020	6	25	42	2	2	1	...	NaN	NaN
178	153412011	785	0	2020	7	15	45	1	2	1	...	NaN	NaN
251	173024011	1199	0	2020	7	22	47	1	2	1	...	1	1
305	185024011	1195	0	2020	7	15	27	1	2	1	...	NaN	NaN
337	194398011	5857	2	2020	7	27	41	2	3	1	...	NaN	NaN
450	224463011	605	0	2020	8	3	46	2	2	1	...	1	1
457	225071011	4268	1	2020	8	7	41	1	2	1	...	NaN	NaN
464	225673011	2323	1	2020	7	14	46	2	3	2	...	NaN	NaN
477	233118011	7106	2	2020	7	2	11	1	4	1	...	1	1
487	233484011	1515	1	2020	8	7	48	1	2	1	...	1	1
491	234215011	1450	0	2020	6	14	44	1	3	2	...	NaN	NaN

[14 rows x 77 columns]

필터링 조건 : (CNT == 2 and H_INC4 > 622)

	HHID	H_WGC	H_WSC	DATE_Y	DATE_M	DATE_D	REGION1	REGION2	FAM_N	IND_N	...	HEXP2_1	HEXP2_2
111	135091011	945	0	2020	9	23	30	1	3	3	...	NaN	NaN
257	173504011	603	0	2020	6	23	31	1	3	2	...	NaN	NaN

[2 rows x 77 columns]

필터링 조건 : (CNT == 3 and H_INC4 > 1164)

Empty DataFrame

Columns: [HHID, H_WGC, H_WSC, DATE_Y, DATE_M, DATE_D, REGION1, REGION2, FAM_N, IND_N, DEATH_YN, ALL_CON, Index: []

[0 rows x 77 columns]

자료: 저자 작성

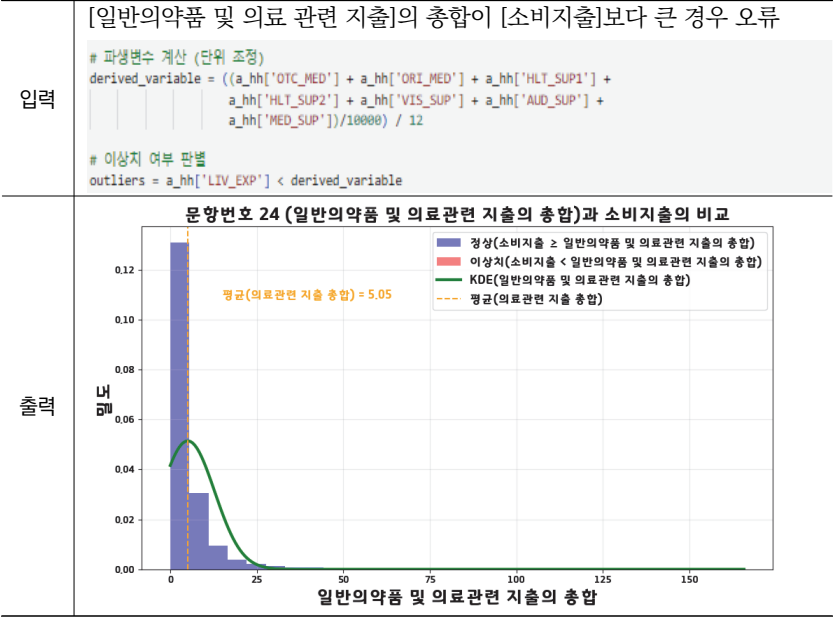
유형 4에 속하는 문항 번호 15번에 대한 실행 결과로, 유형 4는 단일 변수에 대하여 or 논리연산자로 구분되는 복수의 기준치를 다룬다. [자산 (PROP)] 변수에 대해서 1000000인 경우에 대해 1차로 이상치를 재확인 하고 분포 중 하위 5번째에 해당하는 530000인 경우에도 이상치를 확인 하기 때문에 PROP의 분포와 이상치 기준의 수치적 분포, 시각적 분포를 확인할 수 있다.

〈표 5-30〉 유형 4) 문항 번호 15

입력	[자산] 이상치 재확인 + Code + Notebook # [파일이름, 타겟변수, 타겟변수의 현행 기준점, 새로 확인할 데이터] input_list_15 = ['E_HI', 'PROP', [1000000, 530000], 100]
출력	실행 결과 input_list = input_list_15 type4_1_main(input_list=input_list, var_num=15) 데이터 통계 정보 조건 : 조건 x PROP데이터 분포: Count: 6689 Mean: 29273.92 Std: 44418.72 Min: -9.00 25%: 6000.00 50%: 18000.00 75%: 35300.00 Max: 777050.00 Percentile 위치: 1000000 : 하위 100.00%에 해당 530000 : 하위 99.93%에 해당 Percentile 위치: 100 : 하위 3.05%에 해당 문항번호 15 kde이상치 탐지 (등장확률(1-ECDF) = 0.9670)

자료: 저자 작성

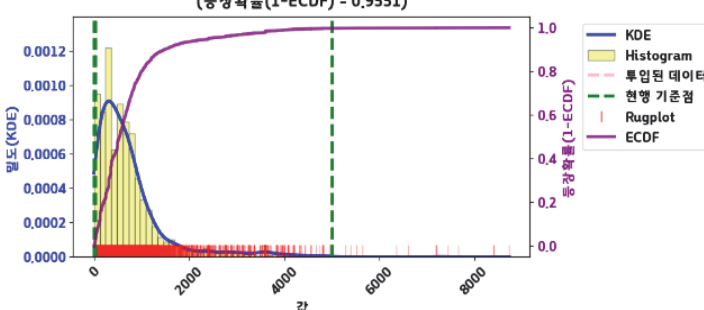
〈표 5-31〉 유형 5) 문항 번호 24



자료: 저자 작성

유형 6에 속하는 문항 번호 9번 변수에 대한 실행 결과 예시이다. 유형 6은 단일 변수에 대해서 조건 범위에 포함되면 이상치로 판단한다. 9번 문항에서는 [공적 이전 정부 소득_정부 및 지방자치단체(H_INC4)] 단일 변수에 대해 조건 범위에 포함되는 데이터를 재확인한다.

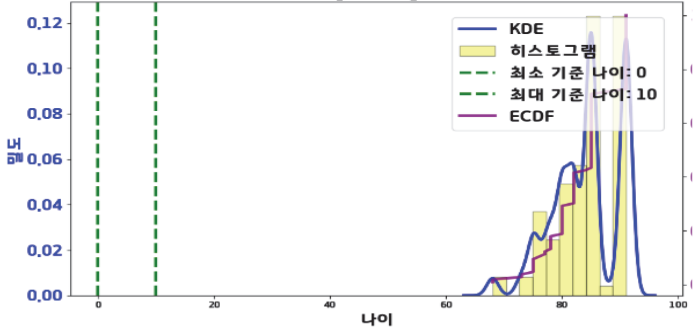
〈표 5-32〉 유형 6) 문항 번호 9

입력	[공적 이전 소득_정부 및 지방자치단체] 이상치 재확인 <pre>data2 = E_HH['H_INC4'].dropna() bound2 = [0, 20, 5000] input_data = data2 # 전체 데이터 data_to_confirm = 30 # 이상치인지 아닌지 판단할 데이터 conditions=bound2 # 현재 고려하는 이상치 조건에 해당하는 값</pre>
출력	데이터 통계 정보 조건 : 조건 x H_INC4데이터 분포: Count: 5074 Mean: 734.24 Std: 822.80 Min: -9.00 25%: 261.25 50%: 528.00 75%: 881.00 Max: 8736.00 Percentile 위치: 0 : 하위 0.18%에 해당 20 : 하위 2.64%에 해당 5000 : 하위 99.72%에 해당 Percentile 위치: 30 : 하위 3.94%에 해당 문항번호 9 kde이상치 탐지 (통강확률(1-ECDF) = 0.9551) 

자료: 저자 작성

유형 7에 속하는 문항 번호 39번 변수에 대한 실행 결과 예시이다. [외래 관련 주질환(OUMED_DZ1, OUMED_DZ2, OUMED_DZ3)]과 [나이(AGE)] 변수의 관계를 파악하여 이상치를 확인한다.

〈표 5-33〉 유형 7) 문항 번호 39

입력	[외래 관련 주질환]과 [연령(해당년도-생년)] 관계 재확인 <pre># 질환 및 기준 설정 diseases = { "협심증": 5, "심근경색증": 6, "자궁경부암": 31, "임신 및 출산 관련": 34, "알코올성 간질환": 11, "통풍": 21, "당뇨병": 24, "유방암": 30, "치매": 36, "백내장": 45, "무릎관절염": 18, "무릎외관절염": 19 } age_criteria = { 5: (0, 10), 6: (0, 10), 31: (0, 15), 34: (0, 15), 11: (0, 20), 21: (0, 20), 24: (0, 20), 30: (0, 20), 36: (0, 30), 45: (0, 30), 18: (0, 40), 19: (0, 40) } # 질병 코드 필터링 filtered_data = data[(data['OUMED_DZ1'] == disease_code) (data['OUMED_DZ2'] == disease_code) (data['OUMED_DZ3'] == disease_code)]</pre>
출력 1-1 (협심증)	협심증 데이터 통계 정보 데이터 분포: Count: 106, Mean: 84.04, Std: 5.67 Min: 68.00, 25%: 80.00, 50%: 85.00 75%: 91.00, Max: 91.00 Percentile 위치: 85.0: 하위 57.08%에 해당 협심증 - 나이 분포 [코드: 5] 

자료: 저자 작성

〈표 5-34〉는 유형 8에 속하는 문항 번호 3번 변수에 대한 실행 결과이다. [사업소득(I_INC2)] 변수를 호출하여 분포를 확인한 후, 이상치 조건에 해당하는 경우 [경제활동 참여 상태(ECO1)], [직업분류(ECO9)]를 포함한 타 변수도 고려하여 재확인할 수 있도록 한다.

〈표 5-34〉 유형 8) 문항 번호 3

입력	[사업소득] 이상치 재확인 (* 상위 1% 등 고소득 사업소득자의 경우 “경제활동 상태”의 [주된 경제활동 상태]와 [직업분류] 를 재확인, * [사업소득]은 마이너스(-)가 있을 수 있으므로 모름(-9)과 잘 구분하여야 함) <pre># [파일이름, 타겟변수, 타겟변수의 현형 기준점, 새로 확인할 데이터] input_list_3 = ['E_IND', 'I_INC2', [0, 50, 20000], 800]</pre>
출력	실행결과 <pre>input_list = input_list_3 typed_1_main(input_list=input_list, var_num=3) ✓ 0.4s</pre> 데이터 통계 정보 조건 : 조건 x I_INC2데이터 분포: Count: 2042 Mean: 2379.91 Std: 2249.51 Min: -500.00 25%: 757.50 50%: 1800.00 75%: 3200.00 Max: 24000.00 Percentile 위치: 0 : 하위 0.56%에 해당 50 : 하위 1.35%에 해당 20000 : 하위 99.95%에 해당 Percentile 위치: 800 : 하위 26.00%에 해당 문항번호 3 kde이상치 탐지 (동장확률(1-ECDF) = 0.7316) Empty DataFrame Columns: [I_ID, P_IDOWN, I_INC2_YN, I_INC2, ECO9, ECO1] Index: []

자료: 저자 작성

5. 실무 적용 가능성 검토 결과

대형 언어모델(LLM)을 활용하여 한국의료패널 데이터를 기반으로 이상치 탐지 및 검수 질의문 생성을 수행한 실증 연구 결과를 바탕으로, 실제 실무 환경에서 해당 방법론이 적용될 수 있는 가능성을 평가하고, 실무 적용 시 고려해야 할 사항을 정리하였다.

가. 시각적 기법 활용

한국의료패널 조사데이터를 내검할 때, 범주형 변수에 대해서는 이상치나 예외 사항에 대해 비교적 파악하기 쉽지만, 연속형 변수에 대해서는 데이터 수치만으로 이상치를 파악하기 쉽지 않다. 그러므로 데이터의 전체적인 분포를 직관적으로 파악할 수 있는 시각화 기법을 활용한다면 이상치를 탐지하고 판단 기준값을 설정하는 데 도움이 될 것이다.

데이터 실증에서 시각적 기법 출력 결과로 제시한 그래프를 보면 ‘커널 밀도함수(KDE), 히스토그램, Rugplot’으로 전체 데이터의 분포를 파악할 수 있고, 현재 내검 시 사용하고 있는 기준값(녹색 점선) 내에 데이터가 모두 포함되는지, 또는 현재 설정된 기준값(녹색 점선)을 조정할 필요가 있는지도 확인할 수 있다. 추가적으로 ‘새로운 기준값’이 그래프에 함께 표시된다면 유용한 정보가 될 것이다.

연속형 변수의 이상치를 판단하는 기준값은 매년 데이터 담당자가 데이터의 분포 등을 고려하여 조정하므로 시각적 기법을 활용한다면 매우 유용할 것으로 기대된다. 다만 기준값을 설정할 때에는 분포만으로 판단하지 않고, 변수와 관련된 제도나 예외 사항 등 변수의 특성을 고려하여 조정하여야 한다.

나. 생성된 탐지 기준 활용

내검 목록은 조사데이터의 품질관리에 가장 중요한 요소 중 하나로서, 2기 한국의료패널 조사데이터 내검 목록을 2020년 1차 조사부터 2024년 5차 조사가 진행된 이후에도 매년 업데이트 중이다. 조사 진행 중에 확인할 수 있는 내검 목록이 정교할수록 데이터 품질을 확보할 수 있으므로 내검 목록을 지속적으로 관리하여야 한다.

데이터 실증에서 RAG를 활용한 새로운 이상치 질의문 생성 예시를 살펴본 결과, 10개 중 절반은 바로 실무에 적용할 수 있을 것으로 보였으며, 추후 더 자세한 정보로 학습시킨다면 실무에 적용 가능한 많은 질의문을 확보하여 내검 목록을 고도화할 수 있을 것으로 기대된다.

다. SAS 코드 생성 결과 활용

한국의료패널 조사데이터 내검 시 SAS 통계프로그램을 활용하고 있는데, 생성된 탐지 기준은 대형 언어모델(LLM)을 통해 SAS 코드 스크립트로 변환되므로 바로 실무에서 활용할 수 있는 장점이 있다.

그리고 기존 내검 목록에 있는 질의문 또한 대형 언어모델(LLM)을 통해 SAS 코드 스크립트로 변환하는 작업을 실시한다면, 기존 SAS 코드와 비교하여 질의문의 정보가 부족한 경우에 내용을 추가하는 등 내검 목록을 업데이트하는 데 도움이 될 수 있을 것이다.

〈표 5-35〉 생성된 탐지 기준 적용 가능성 검토

구분	생성된 탐지 기준	활용 ^{주)}
1	[근로소득]과 [사업소득]이 모두 0으로 보고되었으나, 개인이 고용 상태 또는 자영업자로 등록되어 있는 경우 오류 가능성이 있습니다.	△
2	[재산소득_금융소득], [재산소득_부동산/동산 소득] 두 소득 유형이 모두 매우 높게 보고되었으나(예: 금융소득 > 1억, 부동산/동산 소득 > 2억), [자산]이 낮거나 0으로 보고된 경우 불일치가 있을 수 있습니다.	○
3	개인이 [공적연금]과 [기초연금]을 모두 수령하지만, 나이가 일반적인 수급 연령(예: 65세 미만)에 맞지 않는 경우 오류로 간주될 수 있습니다.	○
4	[의료비_수납 금액]이 매우 낮고 [의료비_비급여 금액]이 매우 높은 경우, 카테고리의 오해 또는 보고 오류일 수 있습니다.	×
5	의료 방문을 위한 [교통비]가 [의료비_수납 금액]에 비해 매우 높은 경우, 교통비 보고에 오류가 있을 수 있습니다.	×
6	[보험료]가 지속적으로 높게 보고되지만 [보험금 수령액]이 0이거나 매우 낮은 경우, 보험 청구 또는 보고에 문제가 있을 수 있습니다.	×
7	[신체활동별 비용]이 높지만 [건강보조식품_복용]이 0인 경우, 건강 관련 지출 습관에 불일치가 있을 수 있습니다.	×
8	[와병일수]가 높지만 [결근결석일수]가 낮은 경우, 질병 보고와 실제 결근 사이에 불일치가 있을 수 있습니다.	△
9	[소득_소비지출]과 [소득_비소비지출]의 합이 보고된 소득을 크게 초과하는 경우, 소득 또는 지출 보고에 오류가 있을 수 있습니다.	○
10	[장기요양 및 기타 돌봄 서비스 비용], [산후조리원 비용] 두 비용이 동시에 높게 보고된 경우, 돌봄 관련 비용의 중복 또는 분류 오류가 있을 수 있습니다.	×

주: ○(적용 가능), △(수정하여 적용 가능), ×(적용 불가)

자료: 저자 작성

실무 적용 가능성에 대해 살펴본 결과, 기존 수작업 검수 프로세스를 자동화하여 시각적으로 직관적인 이상치 탐지가 가능할 것으로 기대되며, 질의문 추천 및 SAS 코드 생성으로 검수 속도를 향상할 수 있을 것이다. 특히 반복적인 검수 작업 자동화에 효과적일 것으로 기대되며, 내검의 전반적인 프로세스를 고도화할 수 있을 것이다. 이러한 수행 과정의 완성도를 통해 조사데이터 품질을 제고할 수 있을 것으로 기대된다.

다만 초기에는 더 많은 비용과 인력이 필요할 수 있으며, 성능을 최적화하기까지 시간이 걸릴 수 있다. 또한 한국의료패널 조사데이터는 의료 관련 정보로서 개인정보 및 민감정보 보호를 위한 조치 등의 보안 문제를 고려하여 접근하여야 한다. 이를 위하여 실무에 일부 적용하여 실제 내검 작업에서 효율성이 있는지 검증하는 등 지속적인 연구 및 개선이 필요할 것이다.

제4절 소결

제5장에서는 조사데이터 품질관리 적용 방안으로 한국의료패널 데이터의 실증을 통해 데이터 오류 수정 방안, 한국의료패널 데이터의 특성을 검토하고 데이터 실증을 수행하였다.

데이터 오류 수정 방안에서는 한국의료패널 데이터의 오류 유형을 정리하고, 데이터 이상 탐지 기법, 시각적 오류 탐지 도구 등 오류 수정 알고리즘을 검토하였다.

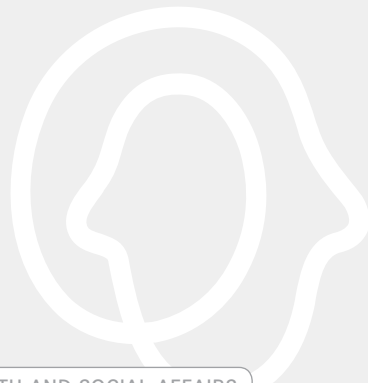
데이터 실증은 이상치 탐지를 위한 시각적 점검 기준 방법과 새로운 점검 기준 생성 2가지를 검토하였다. 첫째, 이상치 탐지를 위한 시각적 점검 기준 출력 결과는 ‘커널밀도함수(KDE), 히스토그램, Rugplot’으로 전체 데이터의 분포를 파악할 수 있고, 현재 내검 시 사용하고 있는 기준값(녹색 점선) 내에 데이터가 모두 포함되는지, 또는 현재 설정된 기준값(녹색 점선)을 조정할 필요가 있는지도 확인해볼 수 있는 그래프를 제공하였다. 실무 담당자 검토 결과, 기존 방식과 비교해 직관적으로 이상치 탐지를 확인할 수 있다는 점에 매우 유용했다. 다만 추가적으로 ‘새로운 (추천) 기준값’이 그래프에 함께 표시된다면 보다 유용한 정보가 될 것으로 여겨진다.

둘째, 분포 인지 RAG 기반 알고리즘을 활용해 생성된 새로운 이상치 질의문 10개를 제시하였다. 실무 담당자 검토 결과, 이 중 5개(적용 가능 3개, 수정 후 적용 가능 2개)는 실무에 적용할 수 있는 것으로 검토되었으며, 추후 더 자세한 정보로 학습시킨다면 실무에 적용 가능한 많은 질의문을 확보하여 내검 목록을 고도화할 수 있을 것으로 기대된다.

사람을
생각하는
사람들



KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS



제6장

결론 및 제언

제1절 결론

제2절 제언

제6장 결론 및 제언

제1절 결론

본 연구에서는 인공지능을 활용하여 조사데이터의 품질을 향상시키는 방안에 대해 모색하였다. 국내외 인공지능(AI) 적용 사례를 검토하고, 데이터 품질관리의 효율성을 높이기 위해 딥러닝을 위시한 인공지능(AI) 기술을 활용할 수 있는 방안을 제시하였다. 더 나아가 한국의료패널 데이터를 이용해 실증 연구를 수행하였고, 이를 통해 조사데이터 품질을 개선하고, 통계 생산의 자동화 및 신뢰성을 강화하는 인공지능(AI) 기반 접근법의 실효성을 평가하였다.

본 연구의 결과를 요약하면 다음과 같다.

첫째, 인공지능(AI)의 개념 및 인공지능 관련 최신 기술 동향을 검토하였다. 인공지능(AI)에 대한 다양한 개념을 살펴보았으며, 국내외 법률에서 규정하고 있는 인공지능의 정의를 검토하였다. 또한 인공지능(AI)의 최신 기술 동향에서는 머신러닝(ML)과 딥러닝(DL)에 대한 최신 기술 동향을 검토하였다.

둘째, 조사데이터의 품질관리를 위한 인공지능(AI) 기반 데이터 오류 탐지 방법론에 대해 검토하였다. 데이터의 오류를 탐지하는 다양한 방법 중 이상치를 검출할 수 있는 방법인 isolation Forest 방법, AnoGan, 대형 언어모델(LLM)을 활용한 데이터 시각화, 대형 언어모델(LLM)을 활용한 데이터 오류 탐지 방법 등에 대해서 검토하였다.

셋째, 통계업무 프로세스 중 인공지능(AI) 적용 가능성을 검토하기 위해 국제 표준 및 한국형 통계업무프로세스모델을 검토하였다. 한국형 통

제업무프로세스모델(KSBPM)은 우리나라의 국가(승인) 통계에 최적화된 통계업무프로세스모델이지만 인공지능(AI) 적용 가능성을 검토한 국내외 선행연구에서 국제 표준인 일반통계업무프로세스모델(GSBPM)을 활용하여 설명하고 있어서, 일반통계업무프로세스모델(GSBPM)을 활용하여 인공지능(AI) 적용 가능성을 검토하였다.

넷째, 통계작성에 있어서 인공지능(AI) 적용 가능성을 검토한 국내외 선행연구를 검토하였다. 설계 단계(2.4), 수집 단계(4.1, 4.3), 처리 단계(5.1, 5.2, 5.3, 5.4, 5.5, 5.6), 분석 단계(6.2, 6.3, 6.4), 배포 단계(7.5), 평가 단계(8.2)의 다양한 업무에서 인공지능(AI) 적용 가능성이 검토되고 있는 것으로 나타났다.

다섯째, 국내외 인공지능(AI)을 활용한 통계작성 사례를 조사하였다. 주요 해외 국가(미국, EU 등) 및 한국에서 AI가 통계 생산과 데이터 품질 개선에 어떻게 적용되고 있는지를 살펴보고, 외국에서는 머신러닝(ML)과 대형 언어모델(LLM)을 활용한 자동화된 데이터 분류, 오류 탐지, 통계 작성 사례가 증가하고 있으며, 한국에서도 통계청을 중심으로 AI 기반 데이터 정제 및 통계 자동화(RPA) 도입이 진행 중임을 조사하였다.

여섯째, 조사데이터의 품질을 유지하고 향상시키기 위한 기존 절차와 문제점에 대해 검토하였다. 한국의료패널 데이터를 중심으로 품질관리 프로세스를 정리하고, 데이터의 오류 탐지 및 수정 과정에서 발생하는 한계점을 검토하였다. 이를 통해 기존의 조사데이터 품질관리 방식이 수작업 중심으로 비효율적일 수 있으며, 인공지능(AI) 기술을 활용한 개선 방안의 필요성을 검토하였다.

일곱째, 데이터 실증에서는 이상치 탐지를 위한 시각적 점검 기준 방법인 ‘커널밀도함수(KDE), 히스토그램, Rugplot’으로 전체 데이터의 분포를 파악할 수 있고, 현재 내검 시 사용하고 있는 기준값 내에 데이터가 모

두 포함되는지, 또는 현재 설정된 기준값을 조정할 필요가 있는지도 확인해볼 수 있는 그래프를 제시하였다. 또한, 분포 인지 검색증강생성기술(RAG) 기반 대형 언어모델(LLM) 접근법을 도입하여 변수 간 결합분포(두 변수 간과 주변분포)를 학습시켜서, 이상치 가능성이 높은 변수 조합을 추천하는 방법론을 제안하였다. 이를 통해 새로운 이상치 질의문 10개를 생성해 제시하였다. 데이터 실증 결과에 대한 실무 담당자 검토 결과, 시각적 점검 기준 방법의 경우 매우 유용하며 실무에 적용 가능한 것으로 검토되었으며, 새로이 생성된 이상치 질의문 10개의 경우 5개(적용 가능 3개, 수정 후 적용 가능 2개)는 실무에 적용할 수 있는 것으로 검토되었다. 이를 통해 인공지능(AI) 기술이 조사데이터의 품질 향상과 업무 효율성 제고에 효과적이며, 실무 적용을 위한 구체적인 방안 제시에도 유의미한 도움을 줄 수 있음을 보여주었다.

제2절 제언

본 연구는 한국의료패널 데이터에서 수집하는 데이터를 대상으로 대상자의 응답에 대해 근거 중심의 탐색적 데이터 분석 및 데이터 오류 탐지에 적용할 수 있는 머신러닝(ML)과 딥러닝(DL) 방법을 살펴보고, 데이터 내의 논리적 오류를 파악할 수 있는 인공지능 기술의 활용 가능성을 파악하는 것을 목표로 하였다.

대형 언어모델(LLM)이 데이터 변수 간의 논리적 오류를 탐지하기 위해서는 전체 변수들 간의 결합분포를 학습할 필요성이 있는데, 이번 연구에서는 주변분포와 두 변수들 간의 결합분포를 문맥 정보로 이해할 수 있는 분포인지 검색증강생성기술(RAG) 기반 대형 언어모델(LLM) 방법을 고안하여 적용하였다. 실험 결과, 대형 언어모델(LLM)을 활용하여 기존의 이상치 탐지 질의어를 기반으로 이상치 가능성이 존재하는 새로운 변수 조합을 추천하였으며, 추천된 변수 조합의 확률분포를 바탕으로 이상치 범위(또는 조건)를 추론하였다.

한편, 기존에 운영 중인 여러 데이터에 대한 오류 규칙 갱신, 최신화가 지속적으로 발생할 수 있기 때문에, 대형 언어모델(LLM)을 활용하여 데이터 오류관리체계의 도입이 필요할 것으로 생각되며, 적합한 RPA 체계와 알고리즘 개발이 필요할 수 있다.

본 연구의 한계점은 다음과 같다. 연구에서는 세 개 이상의 변수 간 결합 분포를 반영하지 못하였으며, 이를 고려할 경우 보다 의미론적으로 우수한 결과를 얻을 수 있을 것으로 기대된다. 이를 해결하기 위한 방안으로, BERT 등의 모델에서 활용된 MLM(masked language model) 학습 방식을 적용할 수 있다. 즉, 임의의 masked token을 생성하고 이를 추론하는 방식을 활용하면, 주어진 조건에 대한 조건부 결합분포를 학습할

수 있다. 이러한 기법을 적용하면, 데이터 내 변수 간의 구조적 상관성을 보다 정밀하게 이해할 수 있으며, 더욱 정교한 추론 방법을 구축할 수 있을 것으로 기대된다.

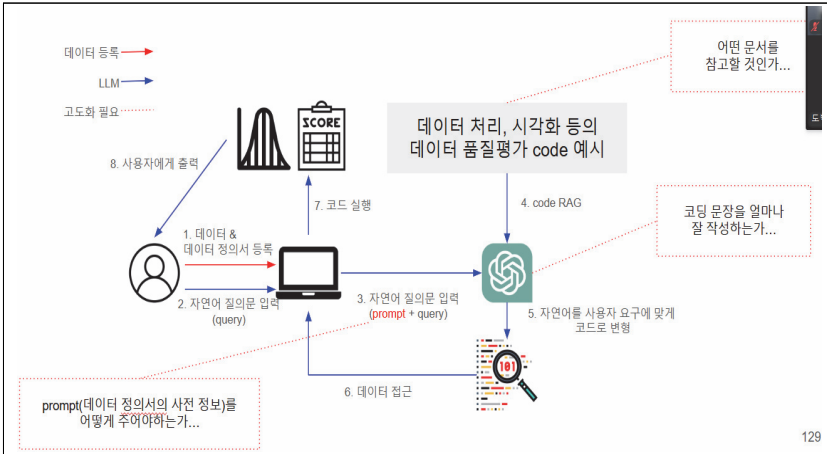
대형 언어모델(LLM)은 형식 불일치나 누락된 값과 같은 일반적인 오류 패턴을 자동으로 식별할 수 있다. 이러한 능력은 단순히 데이터를 처리하는 것을 넘어, 문맥적 이해를 통해 데이터의 품질까지 평가할 수 있음을 시사한다. 이를 바탕으로, 데이터 품질 분석 및 오류 감지 방법론이 더욱 확장될 수 있을 것으로 기대된다.

검수 질의문 생성 방안을 개선하고, 피드백(feedback) 과정을 최적화하는 방법은 다음과 같다.

대형 언어모델(LLM)을 사용하여 데이터 변수 간의 논리적 오류 또는 입력 오류를 감지하는 것은 데이터 품질평가에 대한 효과적인 접근 방식이 될 수 있으며, 특히 데이터 구조를 파악할 수 있는 추론 기술을 동시에 활용할 때 효과가 증대될 수 있을 것으로 기대한다. 한국의료패널 데이터는 실제 자료의 분포를 대형 언어모델(LLM)에 투입하여 오류가 있는 자료를 식별할 수 있을 것으로 기대하나, 모든 자료를 자연어 형태의 문장으로 구성하여 주입할 경우, 대형 언어모델(LLM)이 처리할 수 있는 토큰의 수는 엄청나게 증가하게 되므로 사전에 이를 고려할 필요성이 있다.

품질평가 규칙 기반에 대한 전문가의 의견을 반영하여 의심되는 변수 조건들의 다양한 조합에 대한 피드백을 수집하고, 이를 사전화하는 방안을 고려할 수 있다. 또한, Active learning을 적용하여 초기에는 소수의 데이터 라벨링을 수행한 후, 점진적으로 대상 범위를 확대하는 학습 방식인 incremental learning 기법을 활용하면, 데이터 관계 오류 탐지에 효과적으로 적용할 수 있다. 이를 통해, 새로운 규칙 집합을 지속적으로 확장할 수 있을 것으로 기대된다.

[그림 6-1] 자연어 질의문에 대한 코드 생성



자료: 저자 작성

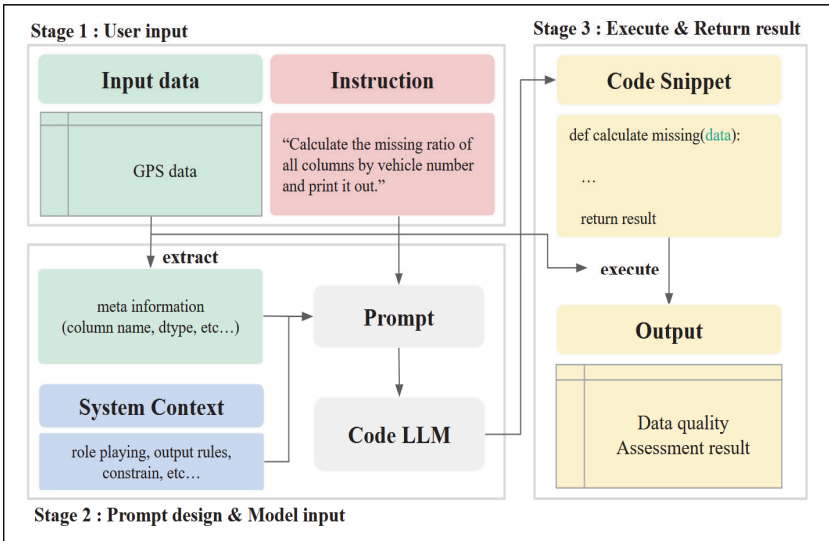
대형 언어모델(LLM)을 활용한 데이터 품질평가 시스템을 설계해보고 이를 간략히 검토해본다. [그림 6-2]는 개략적인 자연어 질의문에 대한 코드 생성의 랭체인을 나타낸다.

일반화된 데이터 품질평가 시스템은 총 3단계로 구성되어 있으며, 각 단계에서 이루어지는 과정은 다음과 같다. 단계 1에서는 품질평가를 받을 데이터와 품질평가를 수행할 지시문을 사용자로부터 입력받으며, 단계 2는 사용자로부터 입력받은 정보를 기반으로 프롬프트가 구성된 후 대형 언어모델(LLM)에 전달되며, instruction과 메타정보(LLM에 데이터원본의 유출 방지하기 위해 데이터의 메타정보만 추출하여 활용), system context(LLM에 의해 출력된 문장이 원활하게 실행되기 위한 출력 제약과 역할 부여 등)로 구성된다. 단계 3은 Code LLM이 출력한 문장을 실행하고 결과를 출력하며, 이 과정에서 사전에 입력받은 데이터를 활용한다.

설계한 아키텍처는 사용자가 입력한 데이터의 메타정보만 대형 언어모

텔(LLM)의 입력값에 활용되므로 원본 데이터의 유출을 방지할 수 있으며, 사용자의 요구사항을 자동화된 프로세스를 통해 충족시킬 수 있을 것으로 기대한다.

[그림 6-2] 대형 언어모델(LLM)을 활용한 데이터 품질평가 시스템 아키텍처



자료: 저자 작성

한편 본 연구에서는 데이터클리닝 담당자의 전통적인 자료 내검 방법과 인공지능(AI)을 활용한 자료 내검 방법의 업무 효율성 등 성능 측정을 비교 분석하지 않았다. 이는 데이터클리닝 담당자가 전통적인 자료 내검 방법을 통해 완료된 조사데이터를 다시 반복해서 업무를 수행하는 데 따른 부담이 있었기 때문이다. 따라서 향후 연구에서는 기존 방식과 개선 방식의 업무 효율성 및 데이터 정확성을 구체적으로 측정하여 비교하는 연구가 필요할 것으로 여겨진다.



- 관계부처 합동. (2019). 인공지능 국가전략.
- 관계부처 합동. (2023). 전국민 AI 일상화 실행계획.
- 김명철, 양기철. (2017). 4차 산업혁명과 인공지능. 한국엔터테인먼트산업학회 학술대회 논문집, 58-63.
- 김정민, 전이슬, 황유림. (2020). AI 기술의 국가통계 활용 사례 및 국내 도입 촉진 방안. 이슈리포트 IS-098. 소프트웨어정책연구소.
- 김정민. (2021). 국가통계 AI도입 활성화 방안 연구.
- 박성준, 신정우, 조용찬. (2023). 보건복지 조사데이터의 체계적 관리 및 활용 방안 연구 - 한국보건사회연구원 조사자료 관리 사례를 중심으로. 한국보건사회연구원.
- 박승규. (2018). 인공지능 기술동향. 주간기술동향, (1861).
- 박창이, 김용대, 김진석, 송종우, 최호식. (2013). R을 이용한 데이터마이닝, 교우사.
- 법무부. (2024). 최신 글로벌 인공지능(AI) 규범 동향 리포트-유럽·미국·중국·일본 등 주요 국가들의 제정 규범을 중심으로. 해외규제 모니터링 제5호.
- 세계법제정보센터. (2023). 국가 인공지능 구상법 2020(National Artificial Intelligence Initiative Act of 2020) 번역본. https://world.moleg.go.kr/web/wli/lgsInfoReadPage.do?CTS_SEQ=50782&AST_SEQ=313 (2024.05.23. 발췌).
- 세계법제정보센터. (2024). 인공지능법(Artificial Intelligence Act) 완역본. https://world.moleg.go.kr/web/wli/lgsInfoReadPage.do?CTS_SEQ=51409&AST_SEQ=93 (2024.05.23. 발췌).
- 신우람. (2024). 일상언어로 묻고 답하는 초거대 AI 통계챗봇 [발표자료]. 제13회 국가통계발전포럼.

- 이기재. (2023). 조사연구에서 인공지능(AI) 활용방안. 미래정책포커스, 2023 (여름호), 52-53.
- 이상현. (2021). AI를 활용한 국내외 통계편제 현황. 국민계정리뷰, 2021(2). 한국은행.
- 이수형, 박은자, 전진아, 박나영, 이해정, 강희정, 오수진, 문석준, 김은주, 염아림, 윤열매, 현승재. (2024). 2022년 한국의료패널 기초분석보고서(II). 한국보건사회연구원.
- 이의규. (2018). 표준 자료처리 모델 GSDEMs의 소개 및 시사점. 2018년 상반기 연구보고서 제1권. 통계개발원.
- 임경민. (2024). AI 통계분류 결과분석 및 실무활용성 제고방안 연구. 통계개발원.
- 조민호. (2021). 인공지능의 역사, 분류 그리고 발전 방향에 관한 연구. 한국전자통신학회 논문지, 16(2), 307-312.
- 조사자료 관리지침. (2024.12.31.). [시행 2024. 12. 31.].
- 최예림, 김관호. (2016). 인공지능 개요 및 적용 사례. IE 매거진, 23(2), 23-29.
- 통계법. (2024.03.26.). [시행 2024. 3. 26.], [법률 제20408호, 2024. 3. 26., 일부개정].
- 통계승인업무 처리지침. (2023.12.05.). [시행 2023. 12. 05.], [통계청예규 제265호, 2023. 12. 05., 일부개정].
- 한국리서치. (2024). 한국리서치 창립 46주년 기념 고객 세미나 발표자료집. 한국리서치.
- 한국보건사회연구원. (2023). 한국의료패널 통계정보보고서. 한국보건사회연구원.
- 한국저작권보호원. (2023). 2023 해외 한류콘텐츠 침해 실태조사-영화·방송·웹툰 분야 불법유통 사이트 중심.
- 한국지능정보사회진흥원. (2018). 공공데이터 품질관리 매뉴얼.
- 호주 통계청. (n.d.). ABS Address Register, Users' Guide. <https://www.abs.gov.au/statistics/research/abs-address-register-users-guide>
- A. M. Turing. (1950). Computing machinery and intelligence. Mind, 59(236), 433-460.

- Abel Coronado & Jimena Juárez. (2020). UNECE-HLG-MOS Machine Learning Project - Imagery Theme Report. <https://statswiki.unece.org/spaces/ML/pages/290358675/WP1+++Theme+3+Imagery+Analysis+Report>
- Alnegheimish, S., Nguyen, L., Berti-Equille, L., & Veeramachaneni, K. (2024). Large language models can be zero-shot anomaly detectors for time series? arXiv preprint arXiv:2405.14755.
- Amati, G. (2009). BM25. In *Encyclopedia of Database Systems* (pp. 257-260). Springer US, Boston, MA.
- Beck, M., Dumpert, F., & Feuerhake, J. (2018). Machine learning in official statistics. arXiv preprint arXiv:1812.10422. <https://arxiv.org/abs/1812.10422>
- Bian, Y., Wang, J., Jun, J. J., & Xie, X.-Q. (2019). Deep convolutional generative adversarial network (DCGAN) models for screening and design of small molecules targeting cannabinoid receptors. *Molecular Pharmaceutics*, 16(11), 4451-4460.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)* (Article No. 159, pp. 1-25).
- Chase, H. (2023). LangChain LLM app development framework.
- Claus Sthamer. (2020). UNECE-HLG-MOS Machine Learning Project - Classification and Coding Theme Report. <https://statswiki.unece.org/spaces/ML/pages/285216478/WP1+++Theme+1+Coding+and+Classification+Report>

- Cyrta, P. (2024). Zero-shot approach using Bielik LLM for emotion recognition in Polish. In Proceedings of the PolEval 2024 Workshop. Retrieved from <http://poleval.pl/files/poleval2024.pdf#page=85>
- Dong, X., Zhang, C., Ge, Y., Mao, Y., Gao, Y., Chen, L., Lin, J., & Lou, D. (2023). C3: Zero-shot Text-to-SQL with ChatGPT. CoRR abs/2307.07306.
- Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., & Jégou, H. (2024). The Faiss library. arXiv preprint arXiv:2401.08281.
- Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the second ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 226-231).
- Florian Dumpert. (2020). UNECE-HLG-MOS Machine Learning Project - Edit and Imputation Theme Report. <https://statswiki.unece.org/spaces/ML/pages/290358735/WP1---+Theme+2+Edit+and+Imputation+Report>
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In Proceedings of the International Conference on Learning Representations (ICLR 2022).
- J. McCarthy, M. L. Minsky, N. Rochester, & C. E. Shannon. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. <http://raysolomonoff.com/dartmouth/boxa/dart>

564_props.pdf

- Jeong, S., Kim, H., Shin, Y., Lee, T., & Kim, H. (2015). A survey of fraud detection research based on transaction analysis and data mining technique. *Journal of the Korea Institute of Information Security & Cryptology*, 25(6), 1525-1540.
- Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769-6781).
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* 33 (pp. 9459-9474).
- Li, A., Zhao, Y., Qiu, C., Kloft, M., Smyth, P., Rudolph, M., & Mandt, S. (2024a). Anomaly detection of tabular data using LLMs. *arXiv preprint arXiv:2406.16308*.
- Li, S., Chen, X., Song, Y., Song, Y., & Zhang, C. (2024b). Prompt4Vis: Prompting large language models with example mining and schema filtering for tabular data visualization. *arXiv preprint arXiv:2402.07909*.
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining* (pp. 413-422). IEEE.
- Liu, G., Zhang, H., & Wong, L. (2011). Controlling false positives in association rule mining. *Proceedings of the VLDB Endowment*, 5(2), 145-156.

- Luo, Y., Tang, N., Li, G., Chai, C., Li, W., & Qin, X. (2021). Synthesizing natural language to visualization (NL2VIS) benchmarks from NL2SQL benchmarks. In Proceedings of the SIGMOD '21: International Conference on Management of Data (pp. 1235-1247). ACM.
- Maddigan, P., & Susnjak, T. (2023). Chat2VIS: Generating data visualizations via natural language using ChatGPT, Codex, and GPT-3 large language models. IEEE Access, 11, 45181-45193.
- Mavroudis, V. (2024). LangChain. < hal-04817573 > .
- Narayan, A., Arora, S., Yang, B., Eyuboglu, S., & Ré, C. (2022).Language models enable simple systems for generating structured views of heterogeneous data lakes. arXiv preprint arXiv:2304.09433. <https://arxiv.org/abs/2304.09433>
- Rane, N., Choudhary, S., & Rane, J. (2024). Machine learning and deep learning: A comprehensive review on methods, techniques, applications, challenges, and future directions. Techniques, Applications, Challenges, and Future Directions.
- Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. Information Processing in Medical Imaging, 10265, 146-157. <https://arxiv.org/abs/1703.05921>
- Stuart Russell & Peter Norvig. (2009). Artificial intelligence: A modern approach. (3rd ed.).
- Thakar, S. G., Martinez, D., Pogoriler, J., Varadarajan, N., et al. (2024). Interleukin-15-armoured GPC3 CAR T cells for patients with solid cancers. Nature, 626, 509-517. <https://doi.org/10.1038/s41586-024-08261-8>

- UN Economic and Social Council. (2019). Generic statistical business process model.
- United Nations Economic Commission for Europe. (2019). Generic Statistical Business Process Model. United Nations.
- United Nations Economic Commission for Europe. (2021). Machine learning for official statistics. United Nations.
- United Nations Economic Commission for Europe. (2023). Large language models for official statistics. United Nations.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* 30.
- Xu, R., & Ding, K. (2024). A survey of large language models for anomaly and out-of-distribution detection. *arXiv preprint arXiv:2409.01980*.
- Yung, W., Karkimaa, J., Scannapieco, M., Barcaroli, G., Zardetto, D., Ruiz Sanches, J. A., Braaksma, B., Buelens, B., & Burger, J. (2018). The use of machine learning in official statistics. UNECE.
- Zhan, F., Zhu, X., Zhang, L., & Wang, X. (2019). Summary of association rules. In *IOP Conference Series: Earth and Environmental Science*, 252(3), 032219. <https://doi.org/10.1088/1755-1315/252/3/032219>



[부록 1] 검색증강기술(RAG) 분석 결과

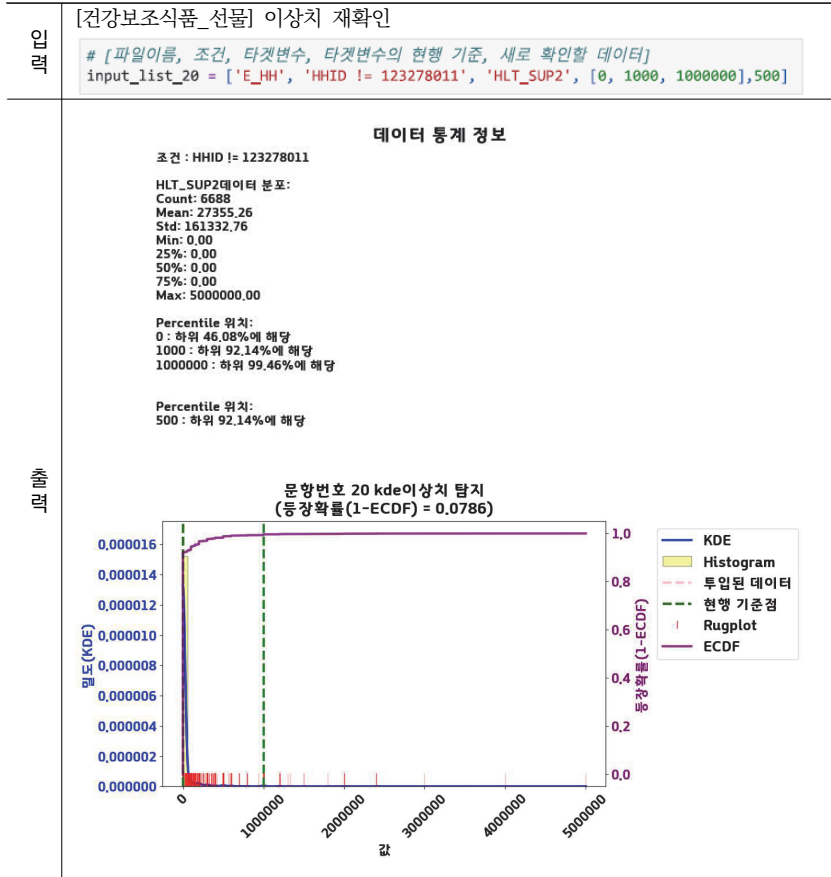
〈부표 1〉 유형 1) 문항 번호 1

요약	[근로소득] 이상치 재확인 <pre># [파일이름, 조건, 타겟변수, 타겟변수의 현행 기준, 새로 확인할 데이터] input_list_1= ['E_IND', [None, 'I_INC1_M>10'], ['I_INC1', 'I_INC1'], [[30000], [100]]]</pre>
세상	<pre>input_list = input_list_1 type1_1_main(input_list=input_list, var_num=1)</pre> 데이터 통계 정보 조건 : 조건 x I_INC1데이터 분포: Count: 5316 Mean: 2735.92 Std: 2421.43 Min: 0.00 25%: 1080.00 50%: 2400.00 75%: 3600.00 Max: 50000.00 Percentile 위치: 30000 : 하위 99.96%에 해당 Percentile 위치: 800 : 하위 19.62%에 해당 문항번호 1 kde이상치 탐지 (동장확률(1-ECDF) = 0.7998)

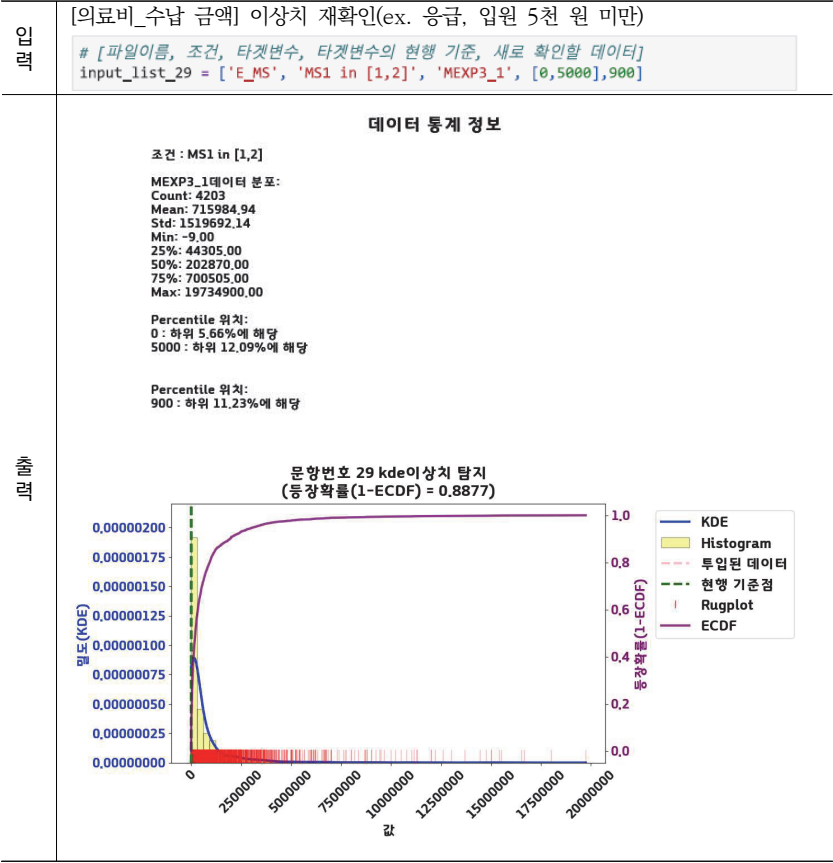
〈부표 2〉 유형 1) 문항 번호 2

요 요 요	요 요 요	“가구원 일반사항 2”의 [의료보장 형태]가 ‘⑤ 의료급여(1, 2종) 세대주’ 또는 ‘⑥ 의료급여 세대원’일 때 [근로소득] 재확인 # [파일 이름, 조건, 타겟변수, 타겟변수의 현행 기준, 새로 확인할 데이터] input_list_2 = ['E_IND', 'HEALTH_INS in [5,6]', 'I_INC2', [1000], 800]
	요 요 요	데이터 통계 정보 조건 : HEALTH_INS in [5,6] I_INC2데이터 분포: Count: 14 Mean: 585.71 Std: 613.80 Min: 100.00 25%: 192.50 50%: 300.00 75%: 675.00 Max: 2000.00 Percentile 위치: 1000 : 하위 85.71%에 해당 Percentile 위치: 800 : 하위 78.57%에 해당 문항번호 2 kde이상치 탐지 (동장확률(1-ECDF) = 0.2143)

〈부표 3〉 유형 1) 문항 번호 20



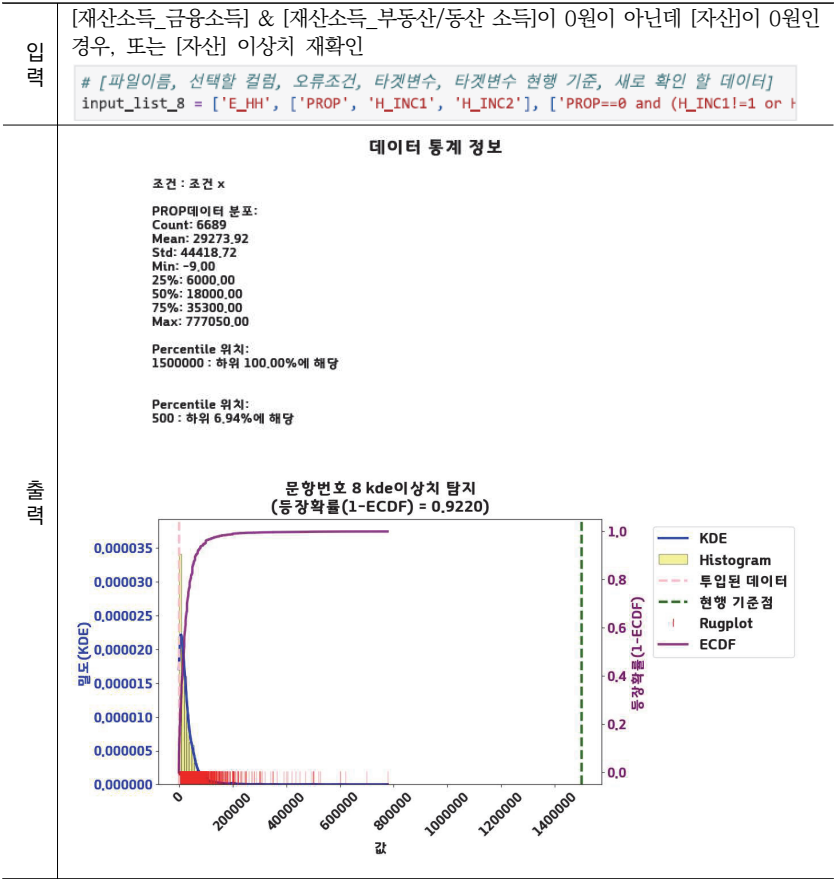
〈부표 4〉 유형 1) 문항 번호 29



〈부표 5〉 유형 1) 문항 번호 30

입 력	[의과 입원 치료 내용]이 '① 수술 및 시술'인데 [의료비_수납 금액]이 10만 원 미만 인 경우 재확인 # [파일을, 조건, 타겟변수, 타겟변수의 현행 기준, 새로 확인할 데이터] input_list_30 = ['E_MS', 'INMED_TRT==1', 'MEXP3_1', [0,100000],10000]
	데이터 통계 정보 조건 : INMED_TRT==1 MEXP3_1데이터 분포: Count: 999 Mean: 1702108.24 Std: 2209890.68 Min: -9.00 25%: 330620.00 50%: 912910.00 75%: 2272216.00 Max: 16662170.00 Percentile 위치: 0 : 하위 1.05%에 해당 100000 : 하위 5.01%에 해당 Percentile 위치: 10000 : 하위 2.90%에 해당 문항번호 30 kde이상치 탐지 (등강확률(1-ECDF) = 0.9710)

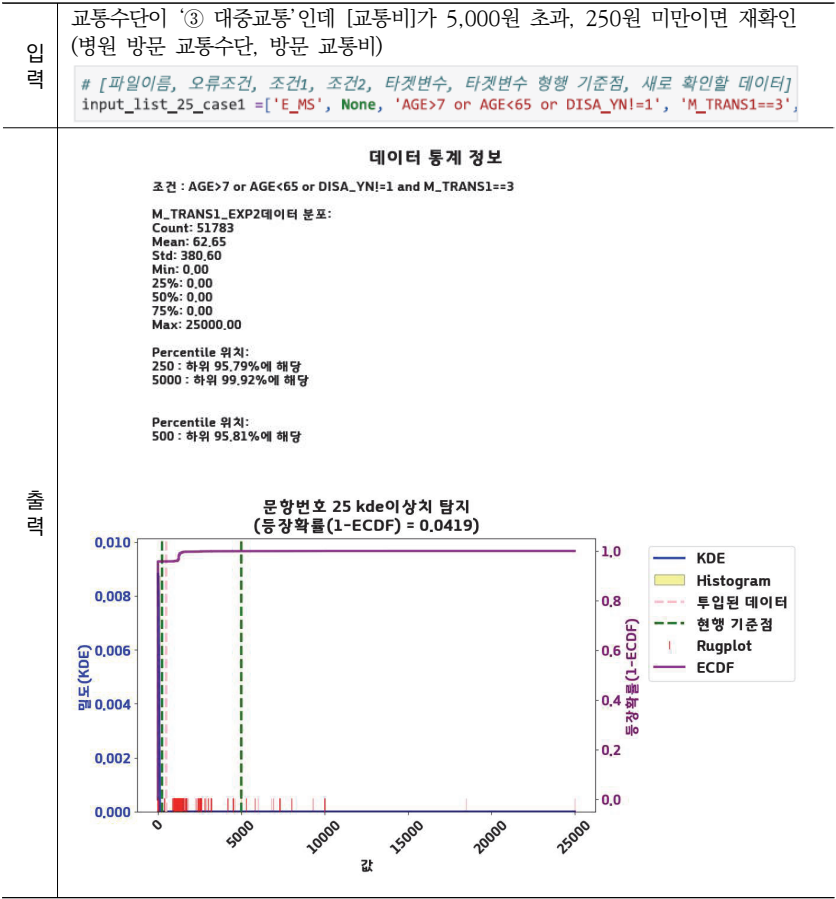
〈부표 6〉 유형 2) 문항 번호 8



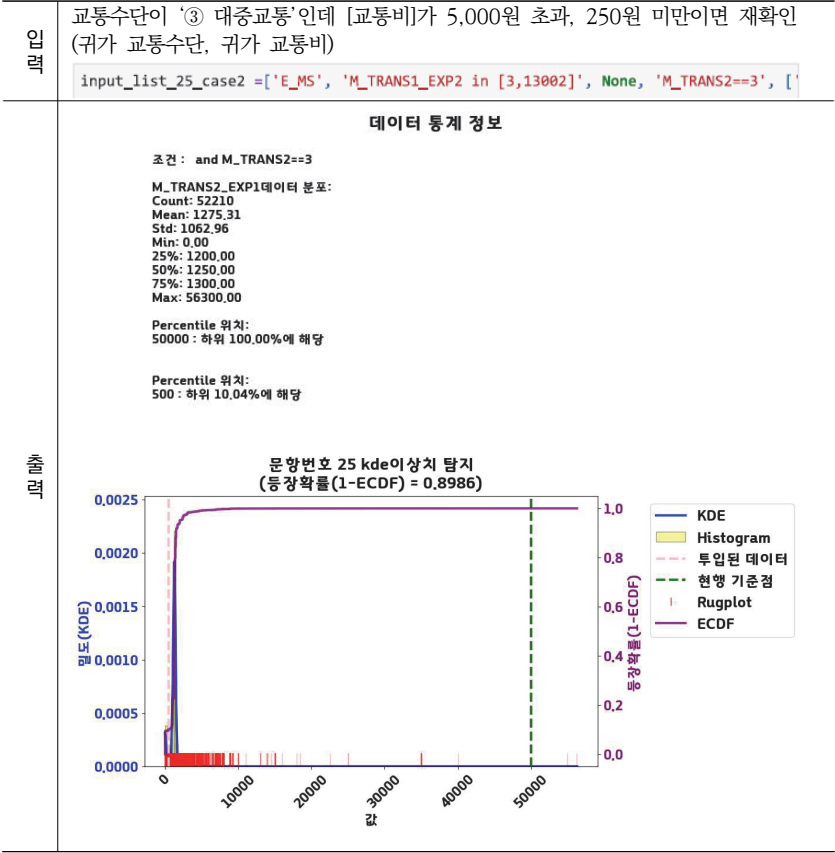
〈부표 7〉 유형 2) 문항 번호 25-1

인원	교통수단이 '③ 대중교통'인데 [교통비]가 5,000원 초과, 250원 미만이면 재확인 (병원 방문 교통수단, 방문 교통비) # [파일이름, 오류조건, 조건1, 조건2, 타겟변수, 타겟변수 형행 기준점, 새로 확인할 데이터] input_list_25_case1=['E_MS', None, 'AGE>7 or AGE<65 or DISA_YN!=1', 'M_TRANS1==3',
세정	데이터 통계 정보 조건 : AGE>7 or AGE<65 or DISA_YN!=1 and M_TRANS1==3 M_TRANS1_EXP1데이터 분포: Count: 51783 Mean: 1272.56 Std: 999.16 Min: 0.00 25%: 1200.00 50%: 1250.00 75%: 1300.00 Max: 39000.00 Percentile 위치: 250 : 하위 9.57%에 해당 5000 : 하위 99.11%에 해당 Percentile 위치: 500 : 하위 10.11%에 해당 문항번호 25 kde이상치 탐지 (등강확률(1-ECDF) = 0.8979)

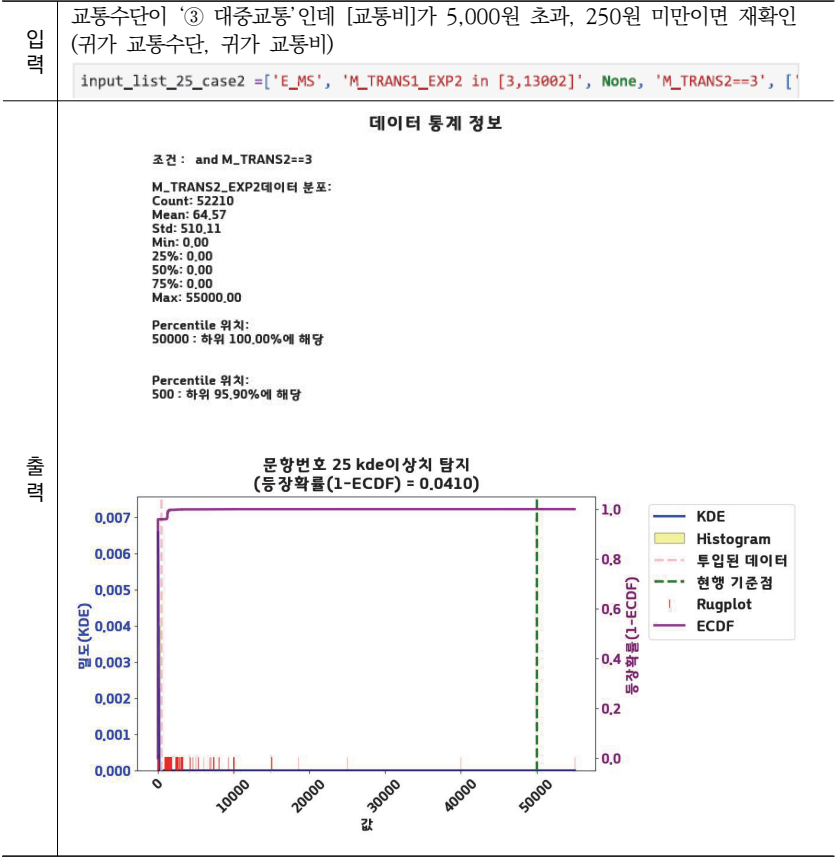
〈부표 8〉 유형 2) 문항 번호 25-2



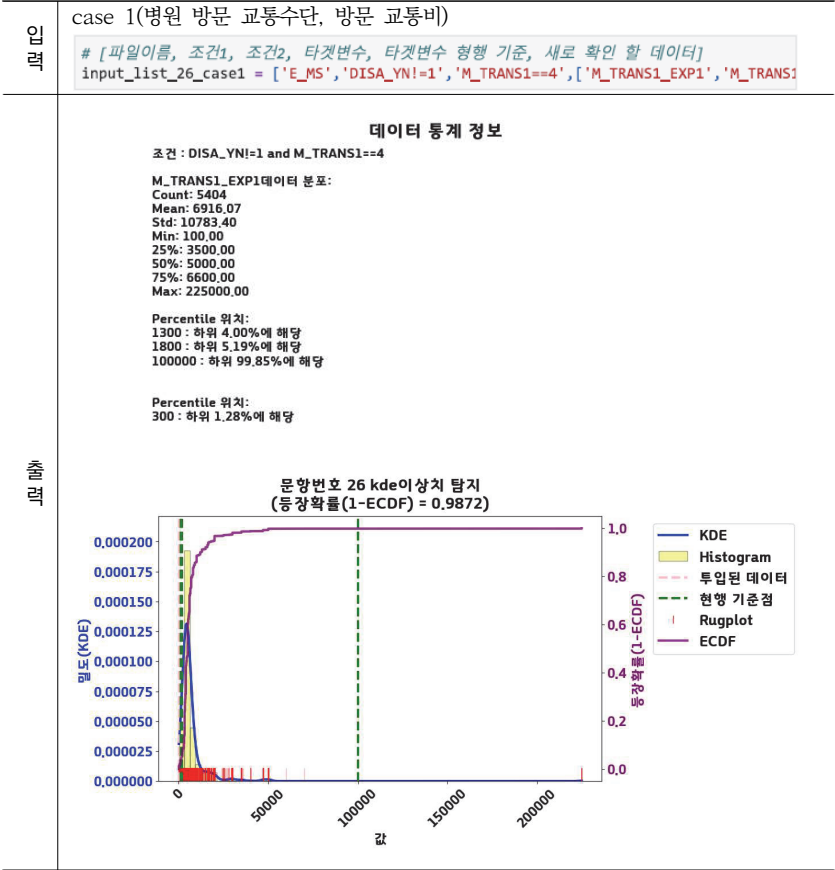
〈부표 9〉 유형 2) 문항 번호 25-3



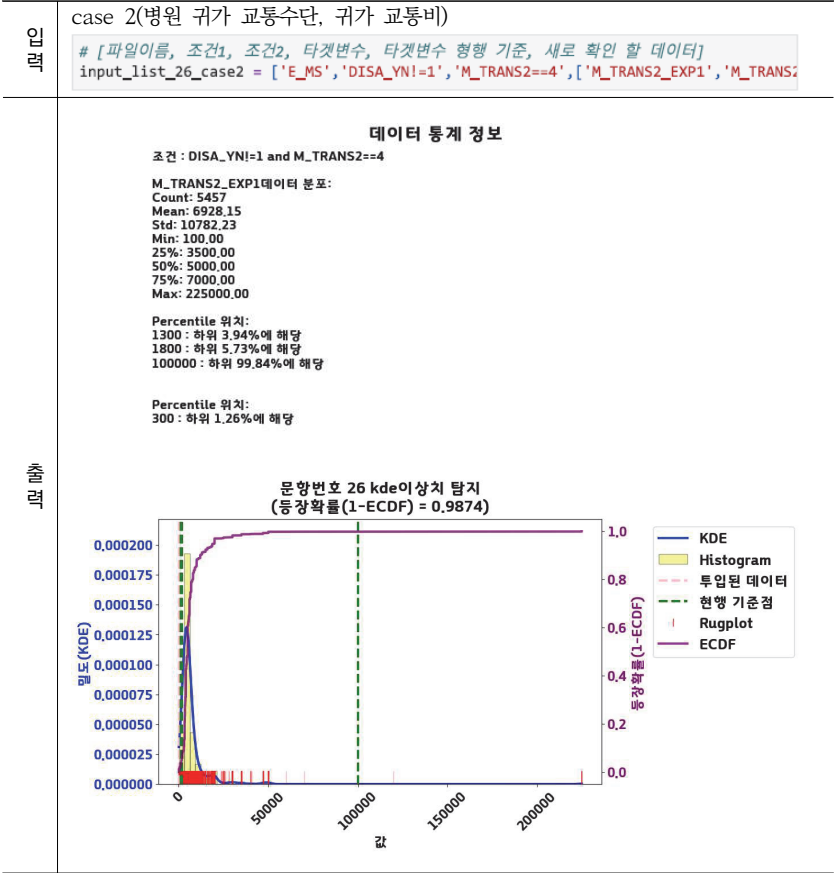
〈부표 10〉 유형 2) 문항 번호 25-4



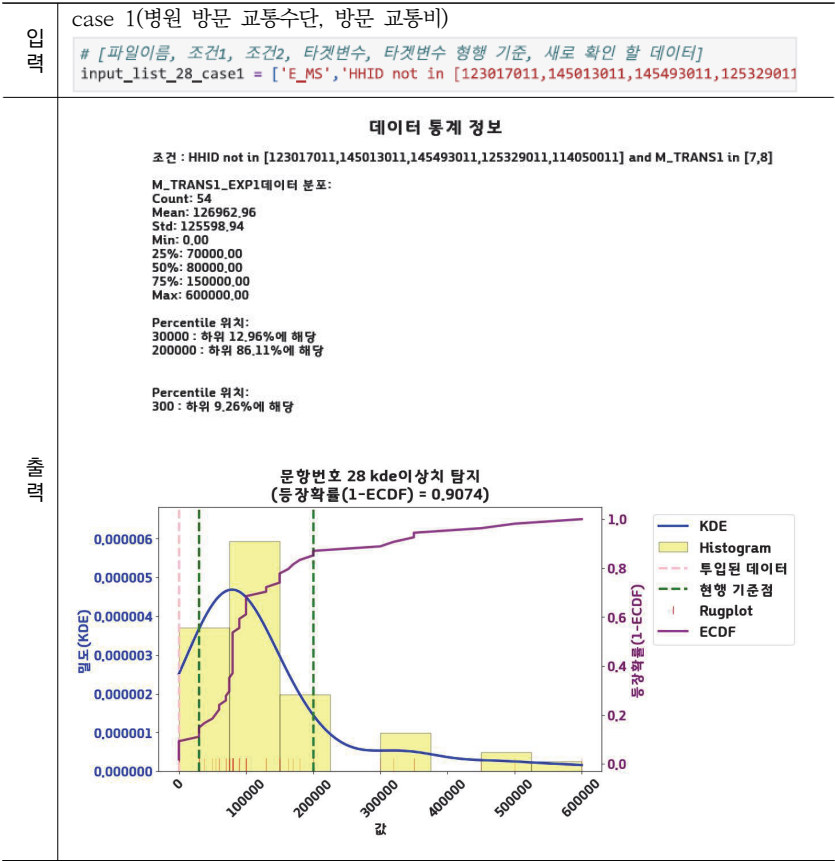
〈부표 11〉 유형 2) 문항 번호 26-1



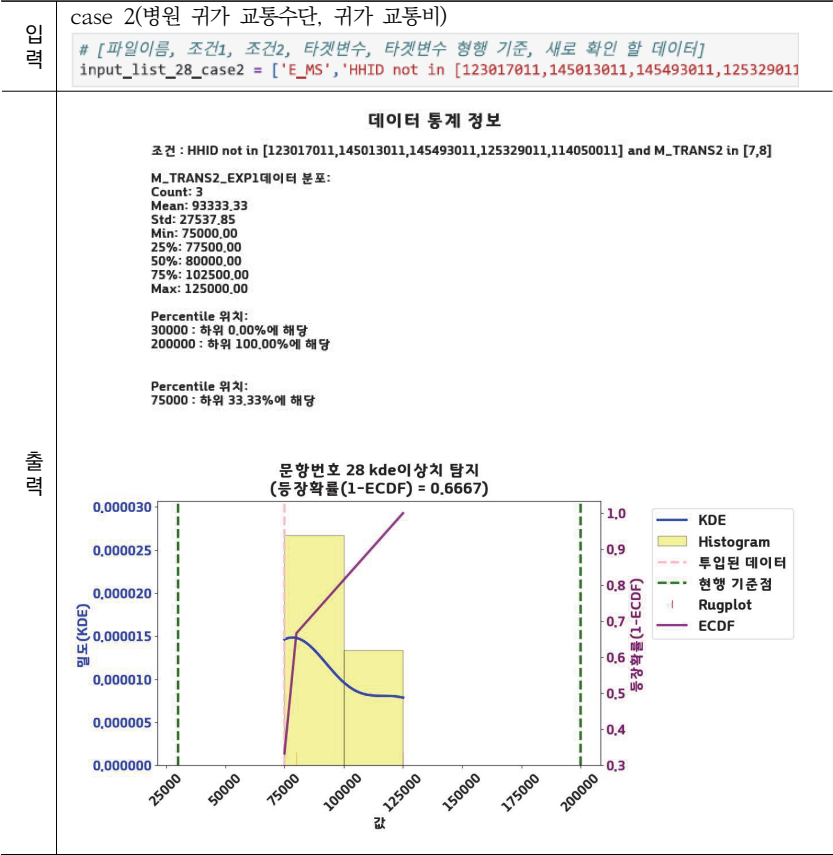
〈부표 12〉 유형 2) 문항 번호 26-2



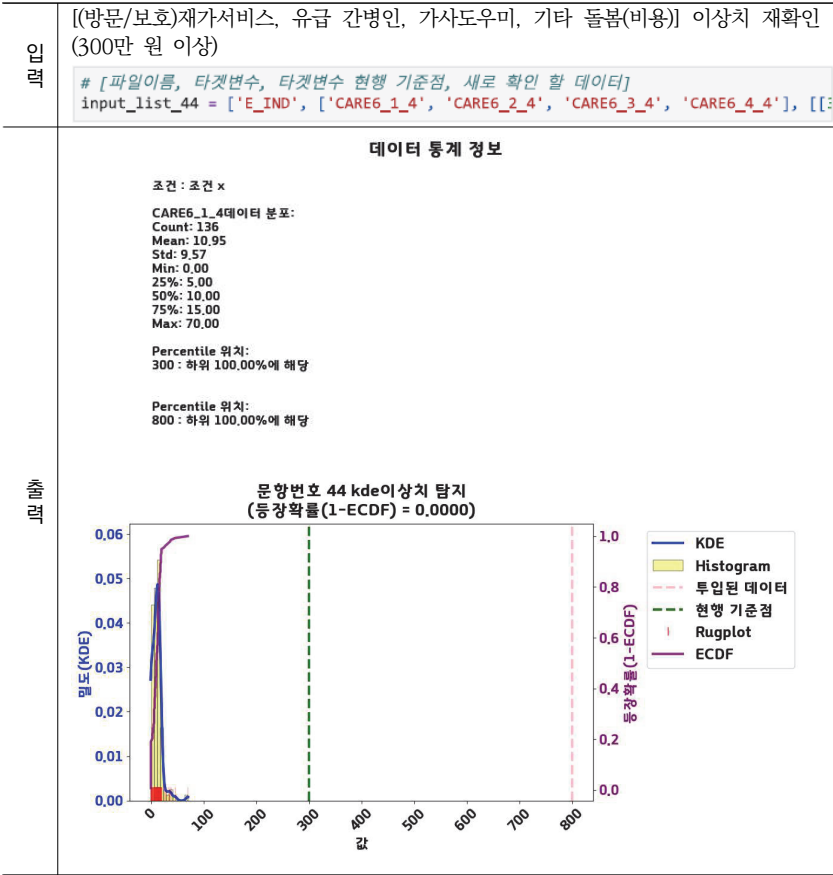
〈부표 13〉 유형 2) 문항 번호 28-1



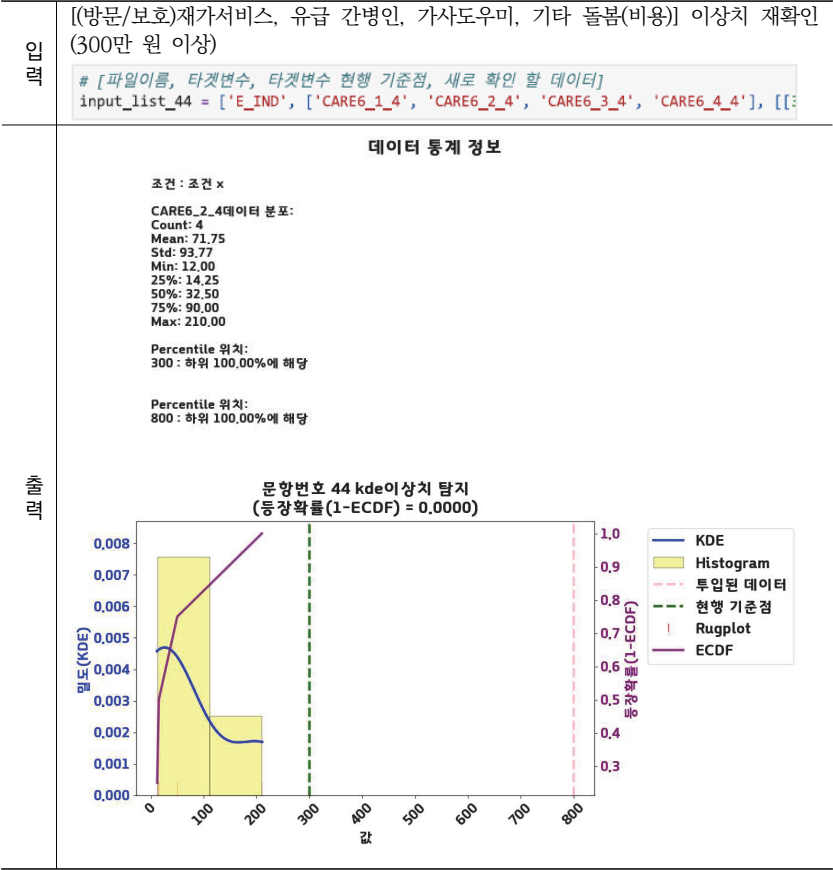
〈부표 14〉 유형 2) 문항 번호 28-2



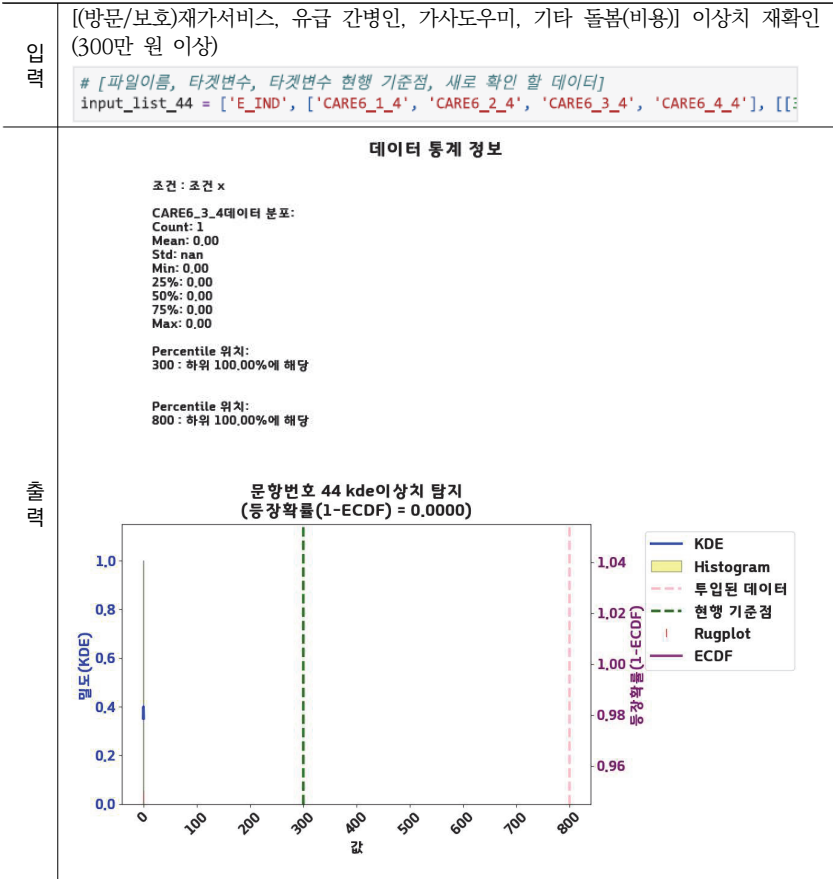
〈부표 15〉 유형 2) 문항 번호 44-1



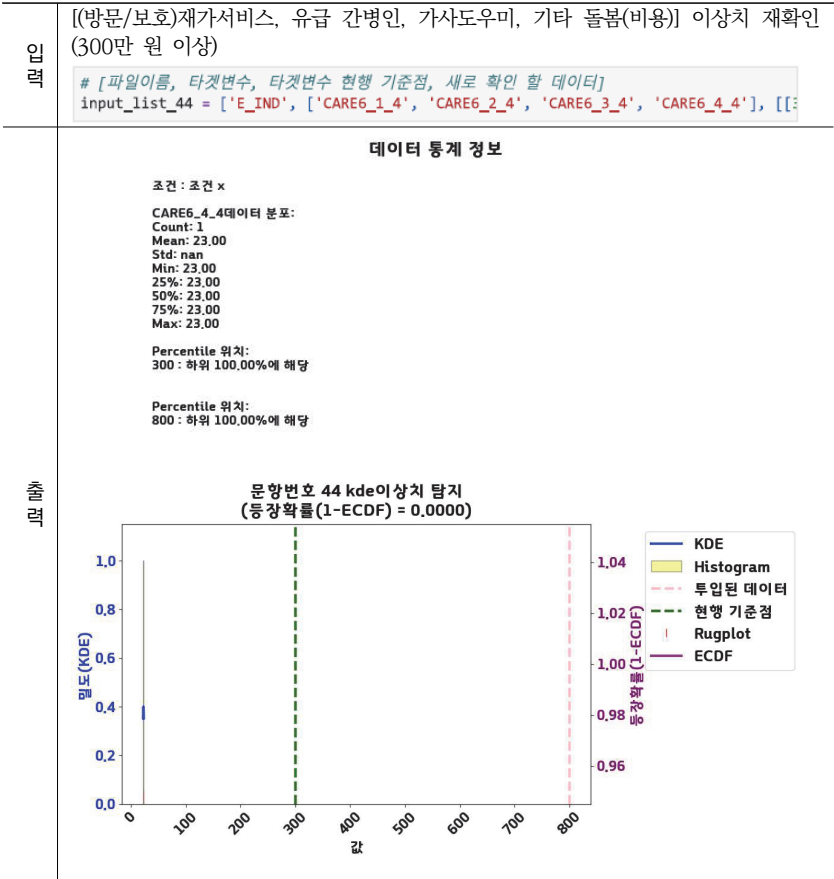
〈부표 16〉 유형 2) 문항 번호 44-2



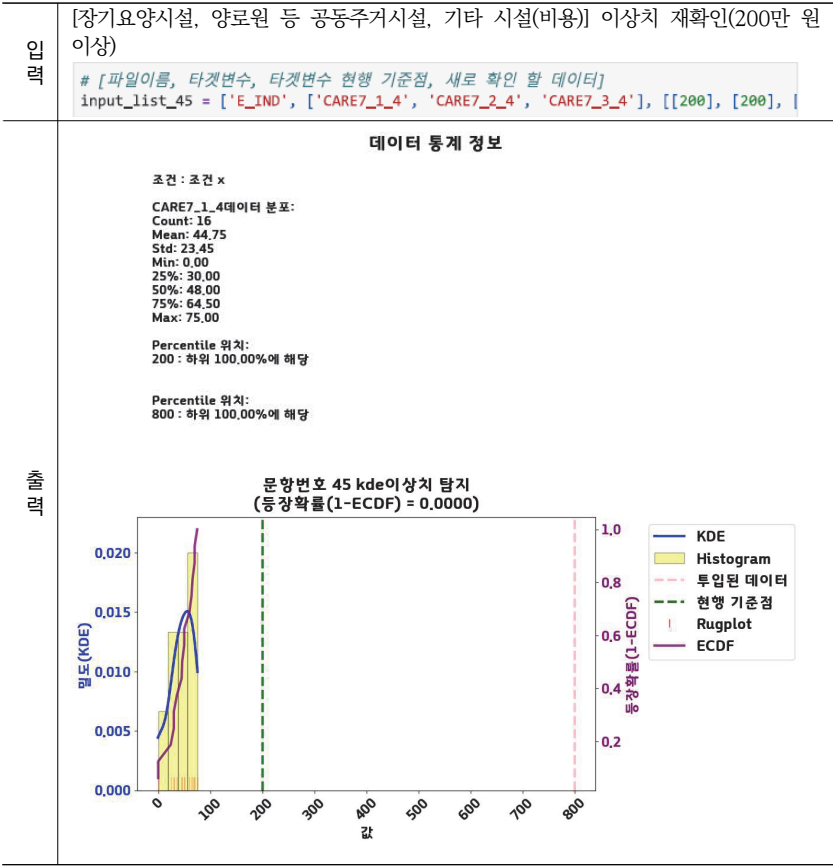
〈부표 17〉 유형 2) 문항 번호 44-3



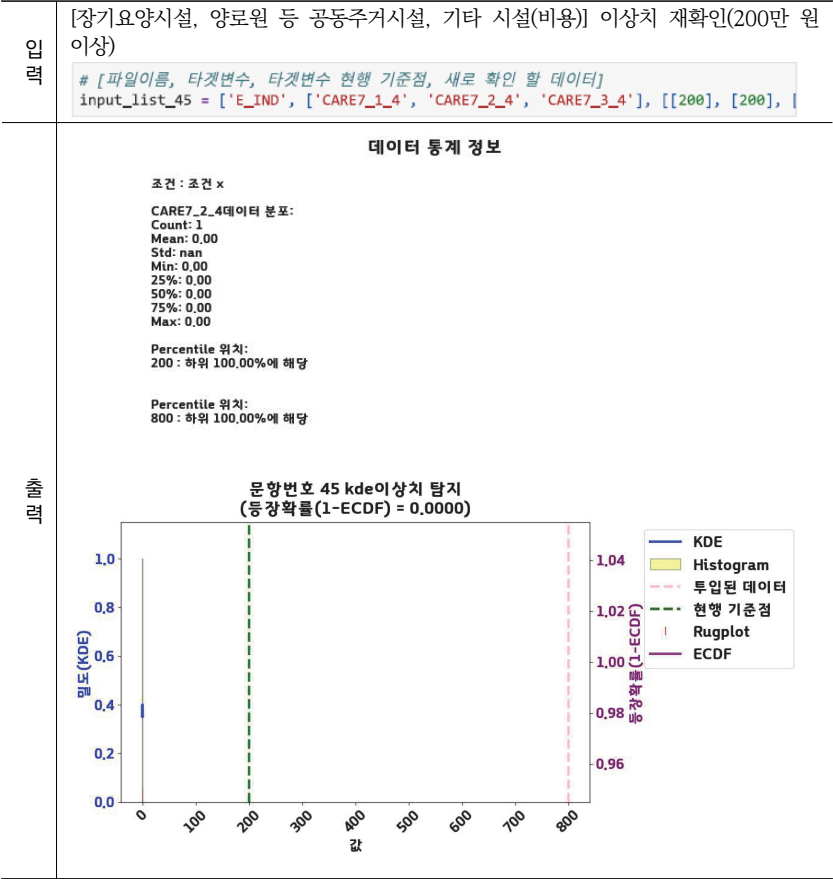
〈부표 18〉 유형 2) 문항 번호 44-4



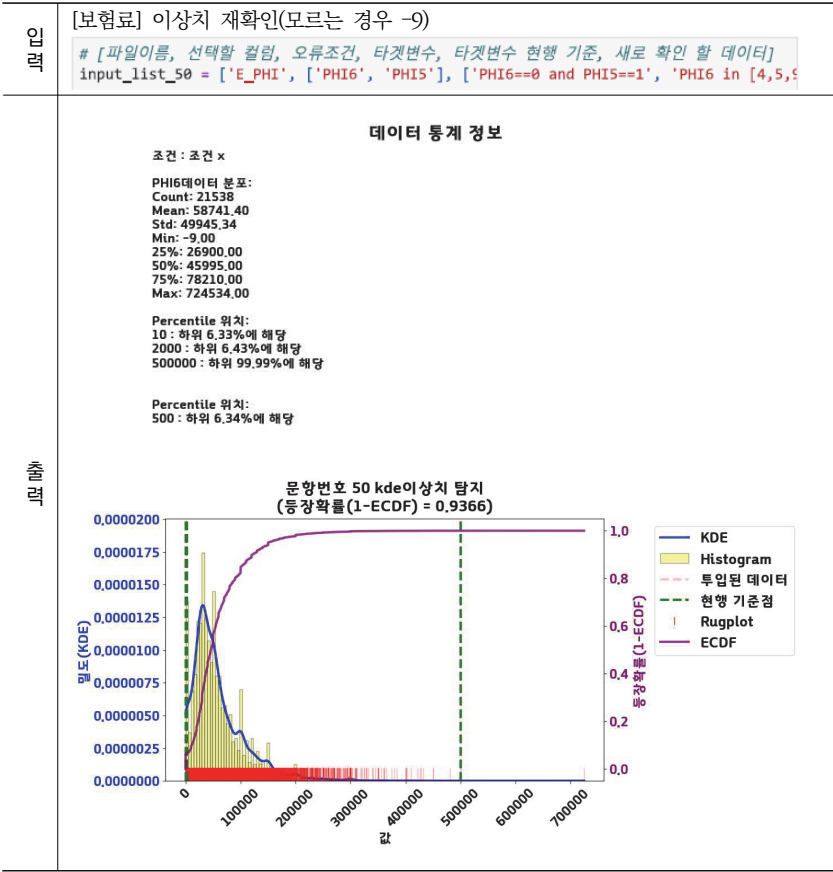
〈부표 19〉 유형 2) 문항 번호 45-1



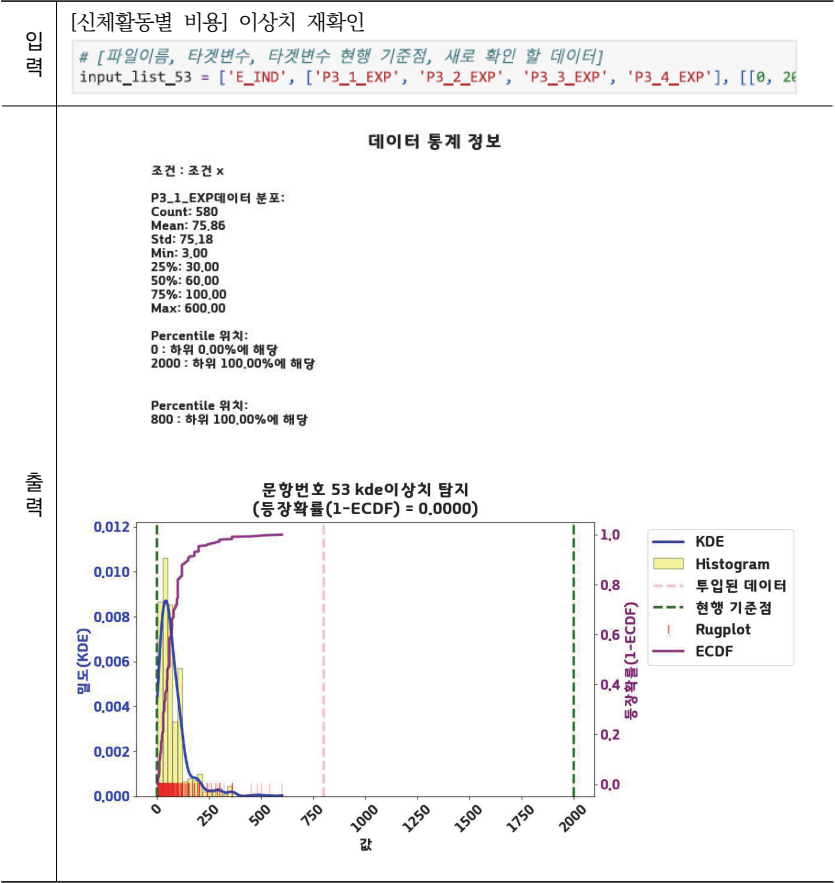
〈부표 20〉 유형 2) 문항 번호 45-2



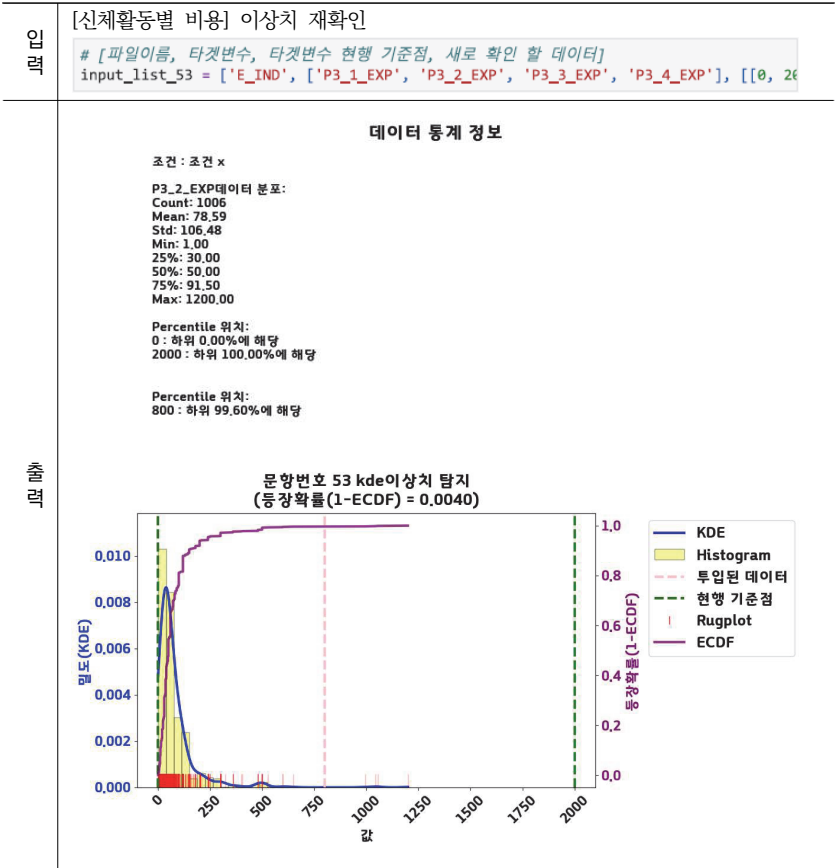
〈부표 21〉 유형 2) 문항 번호 50



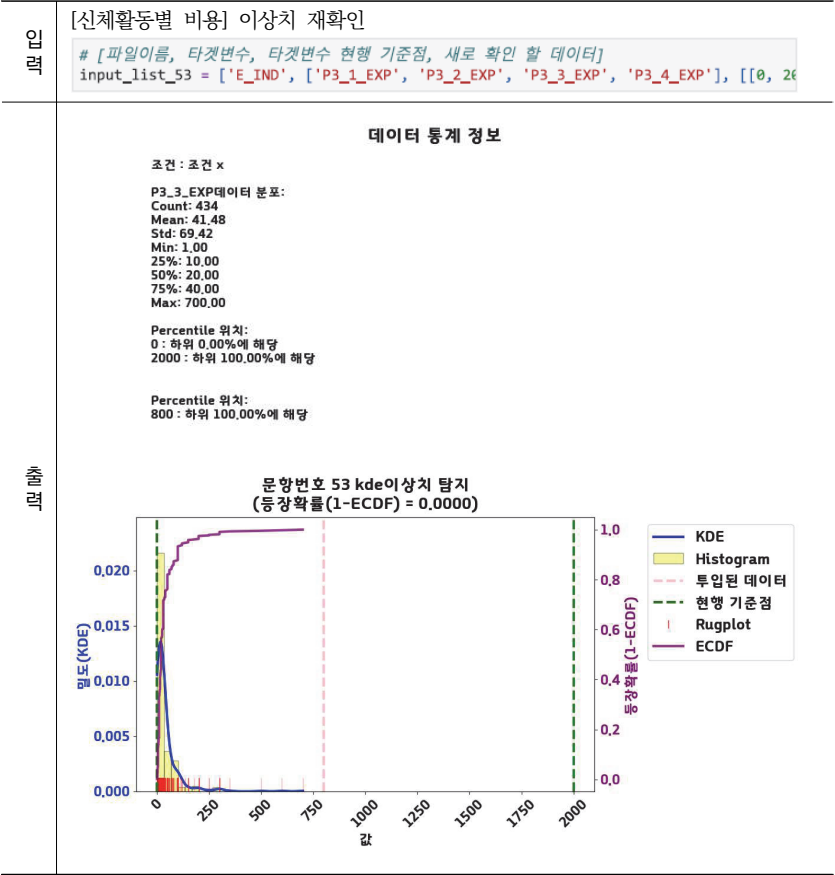
〈부표 22〉 유형 2) 문항 번호 53-1



〈부표 23〉 유형 2) 문항 번호 53-2



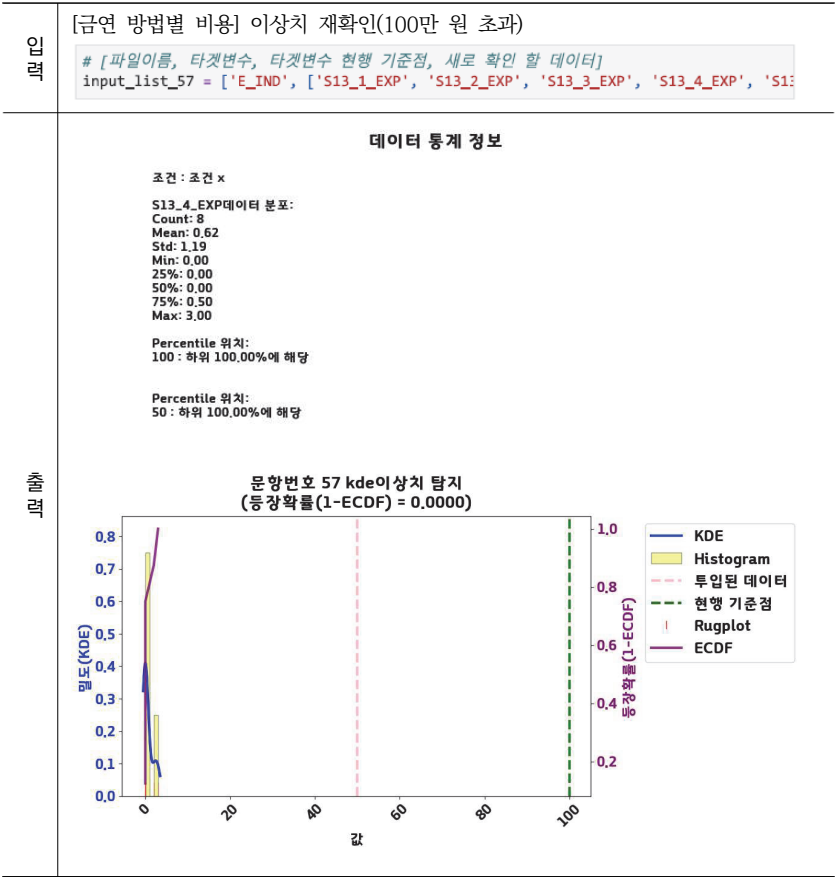
〈부표 24〉 유형 2) 문항 번호 53-3



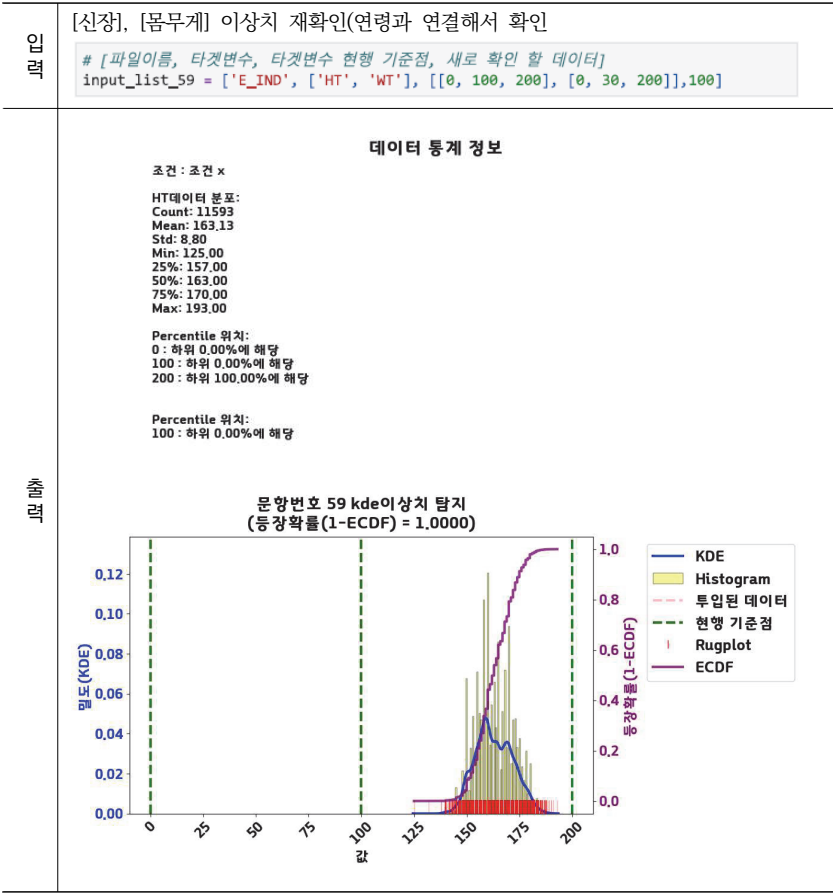
〈부표 25〉 유형 2) 문항 번호 53-4

요점	[신체활동별 비용] 이상치 재확인 # [파일이름, 타겟변수, 타겟변수 현행 기준점, 새로 확인 할 데이터] input_list_53 = ['E_IND', ['P3_1_EXP', 'P3_2_EXP', 'P3_3_EXP', 'P3_4_EXP'], [[0, 2000], [0, 2000]]]
----	--

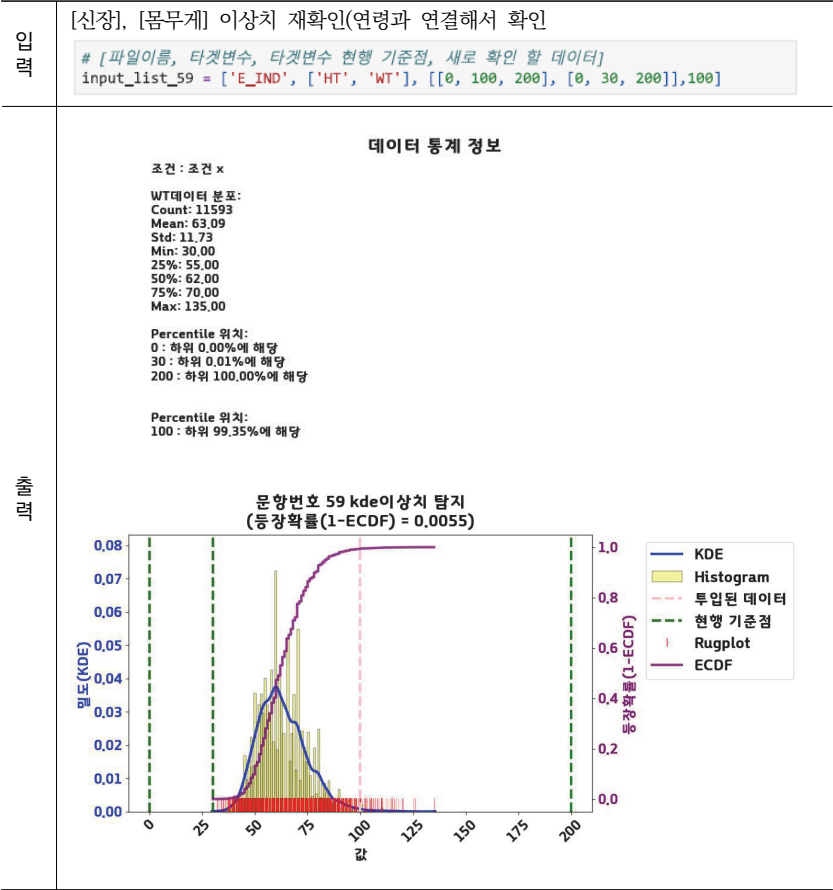
〈부표 29〉 유형 2) 문항 번호 57-4



〈부표 31〉 유형 2) 문항 번호 59-1



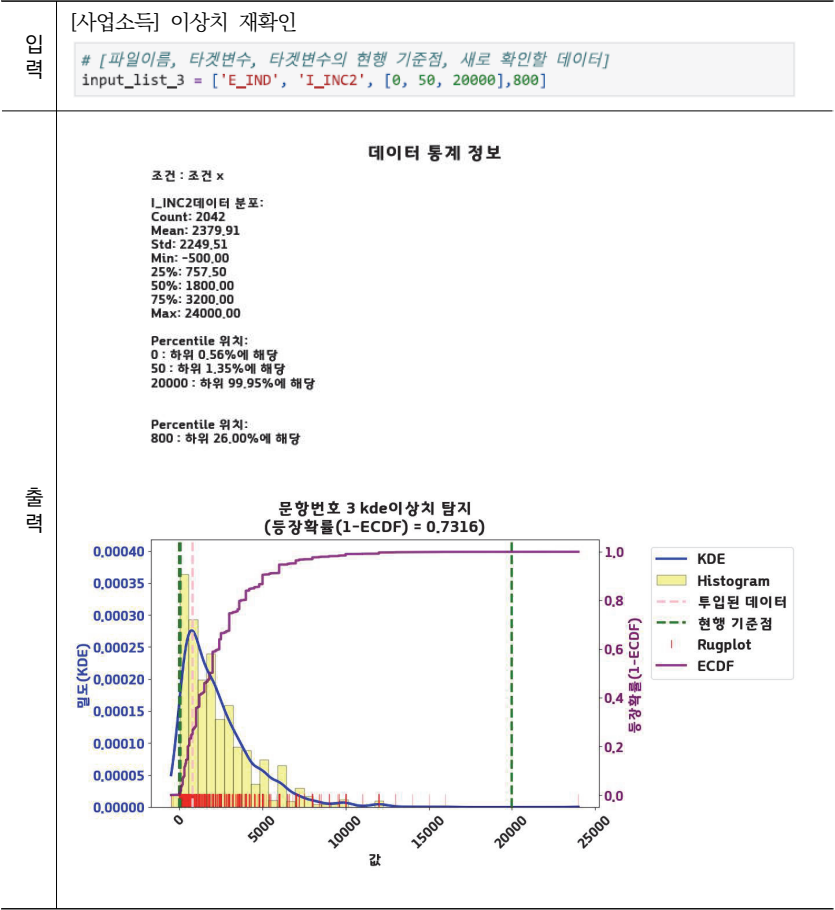
〈부표 32〉 유형 2) 문항 번호 59-2



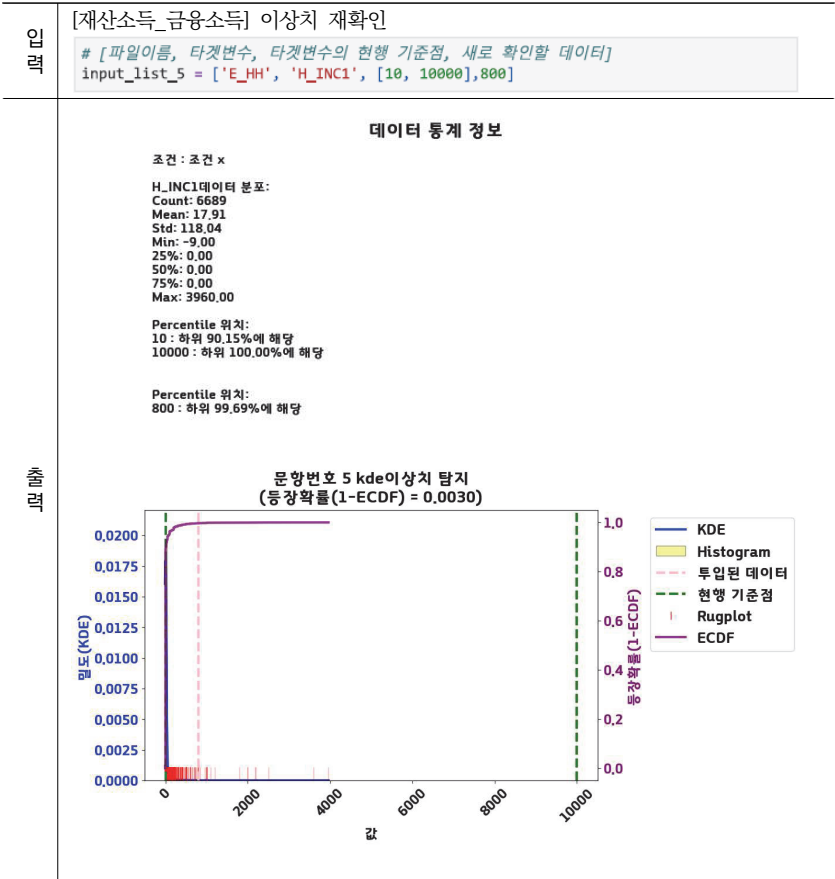
〈부표 33〉 유형 3) 문항 번호 10

입력	[공적연금]이 '① 유'인 경우, 가구원 구성원의 만 나이 재확인 # [파일이름, 조건] input_list_10 = ['E_IND', '(H_INC4_1_YN==1) and (AGE25 in [None,0])']
출력	type3_1_main(input_list=input_list_10,var_num=10) 문항번호 10 오류확인----- 필터링 조건 : (H_INC4_1_YN==1) and (AGE25 in [None,0]) Empty DataFrame Columns: [HHID, H INC4 1 YN, AGE25] Index: [] end-----

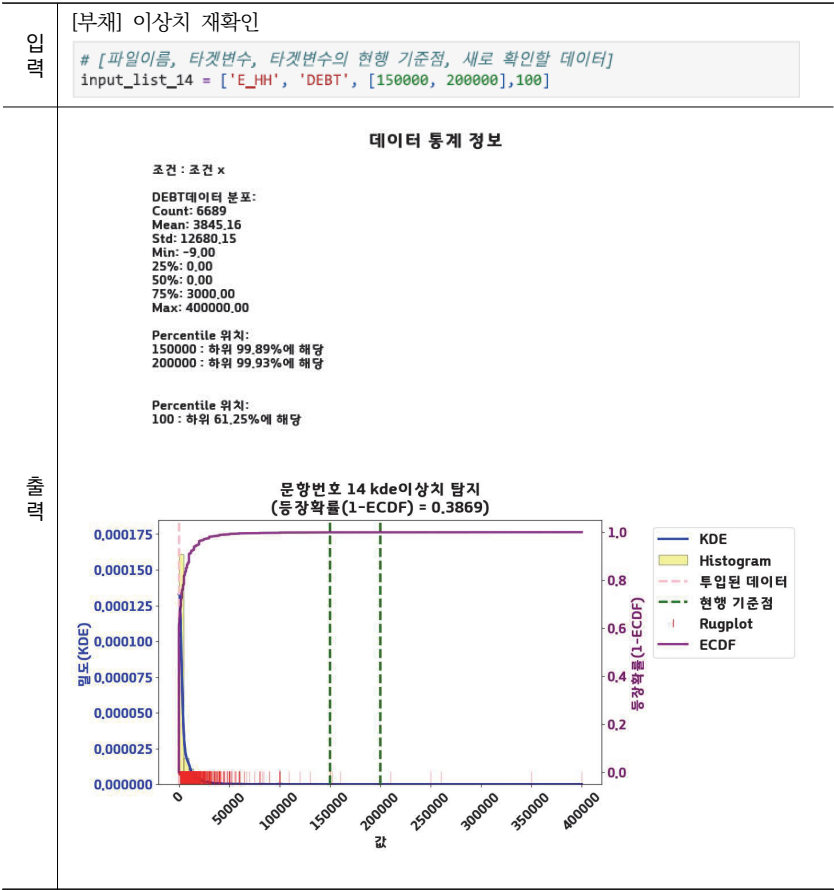
〈부표 34〉 유형 4) 문항 번호 3



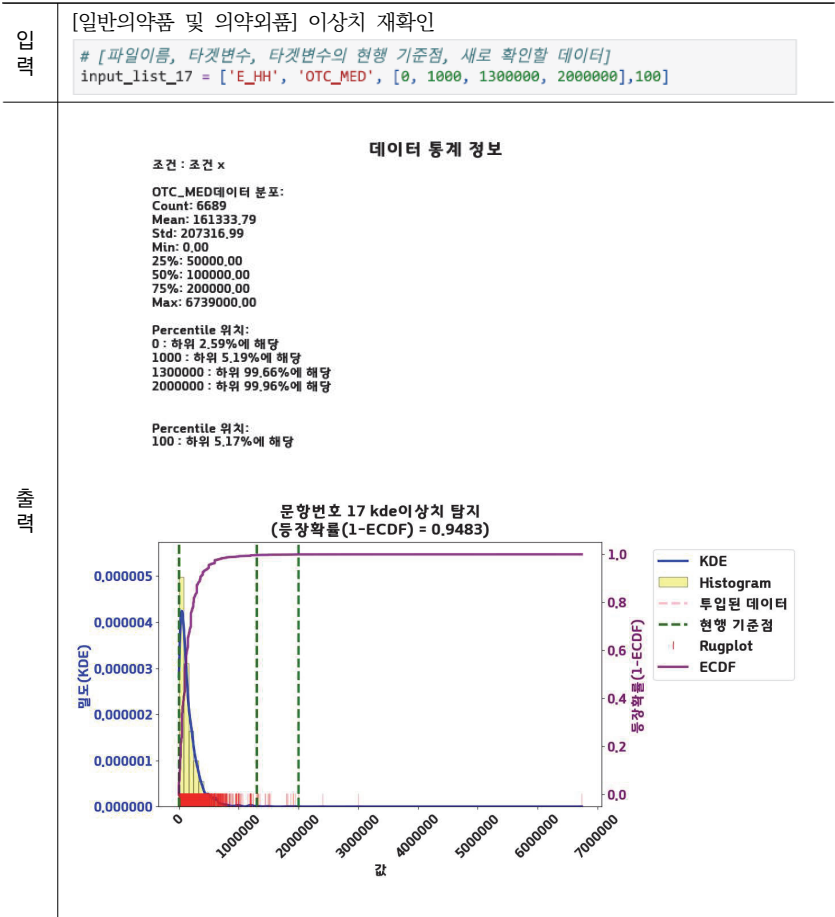
〈부표 35〉 유형 4) 문항 번호 5



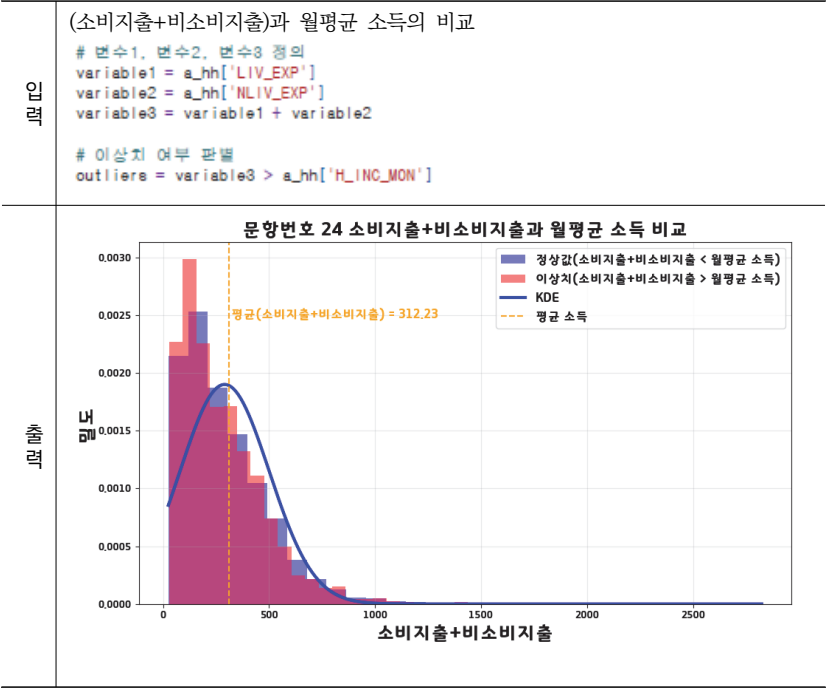
〈부표 36〉 유형 4) 문항 번호 14



〈부표 37〉 유형 4) 문항 번호 17



〈부표 38〉 유형 5) 문항 번호 16



〈부표 39〉 유형 5) 문항 번호 47

인간 자원

일 평균 [산후조리원 비용] 이상치 확인

```

# fit한 데이터에 대해 밀도 추정 계산
x_min, x_max = data.min() - 0.5, data.max() + 0.5
x_val = np.linspace(x_min, x_max, 1000).reshape(-1, 1)
log_density = kde.score_samples(x_val)
density = np.exp(log_density)

# bins 계산
if sqrt_bin and bin==None:
    bins = sqrt_bins(data)
elif sqrt_bin==False and bin==None:
    print('bin을 입력하지 않아 30개로 진행함')
    bins = 30

# ECDF 계산
sorted_data = np.sort(data.flatten())
eodf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data)

# 특정 값(data_to_confirm)의 ECDF 값 계산
eodf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data)
print(f'{data_to_confirm} 등장 확률(1-ECDF): {1-eodf_value:.3f}')

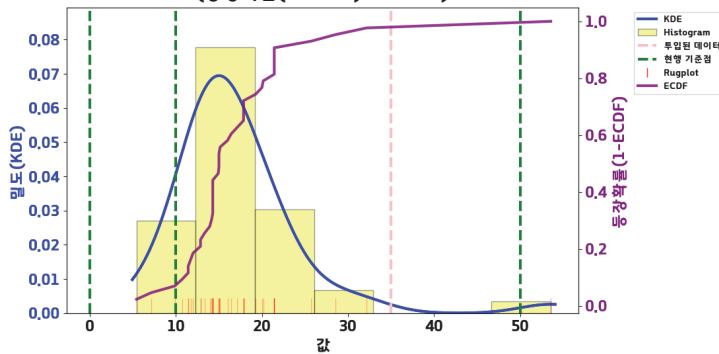
title_font_size = 20
font_size = 17
width = 3

fig, ax1 = plt.subplots(figsize=(10, 6))

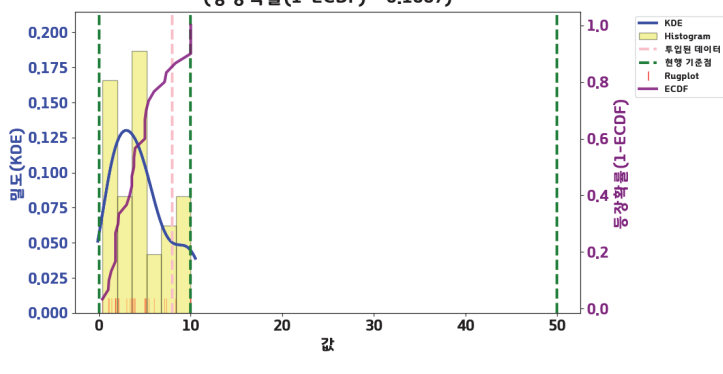
```

데이터

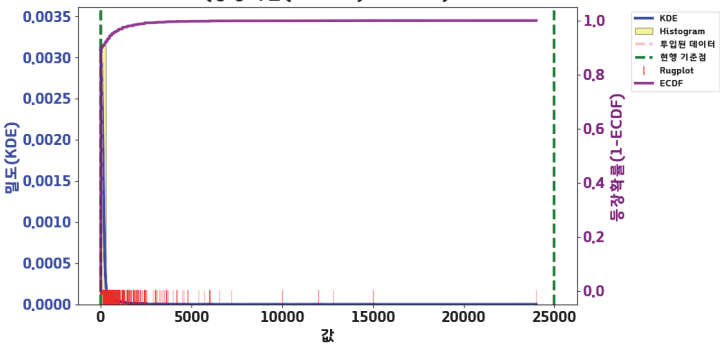
문항번호 47 kde이상치 탐지
(등장확률(1-ECDF) = 0.0233)



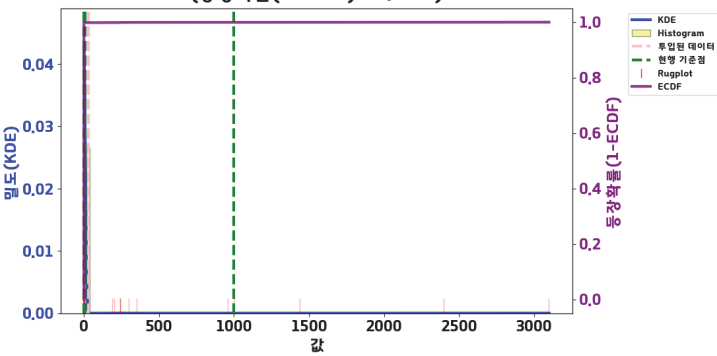
〈부표 40〉 유형 5) 문항 번호 49

인원	일 평균 [산후도우미 비용] 이상치 확인 <pre># fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vals = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vals) density = np.exp(log_density) # bins 계산 if sqrt_bin and bin==None: bins = sqrt_bins(data) elif sqrt_bin==False and bin==None: print('bin을 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_oconfirm)의 ECDF 값 계산 ecdf_value = (sorted_data <= data_to_oconfirm).sum() / len(sorted_data) print(f'{data_to_oconfirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6))</pre>
세점	문항번호 49 kde이상치 탐지 (등장확률(1-ECDF) = 0.1667) 

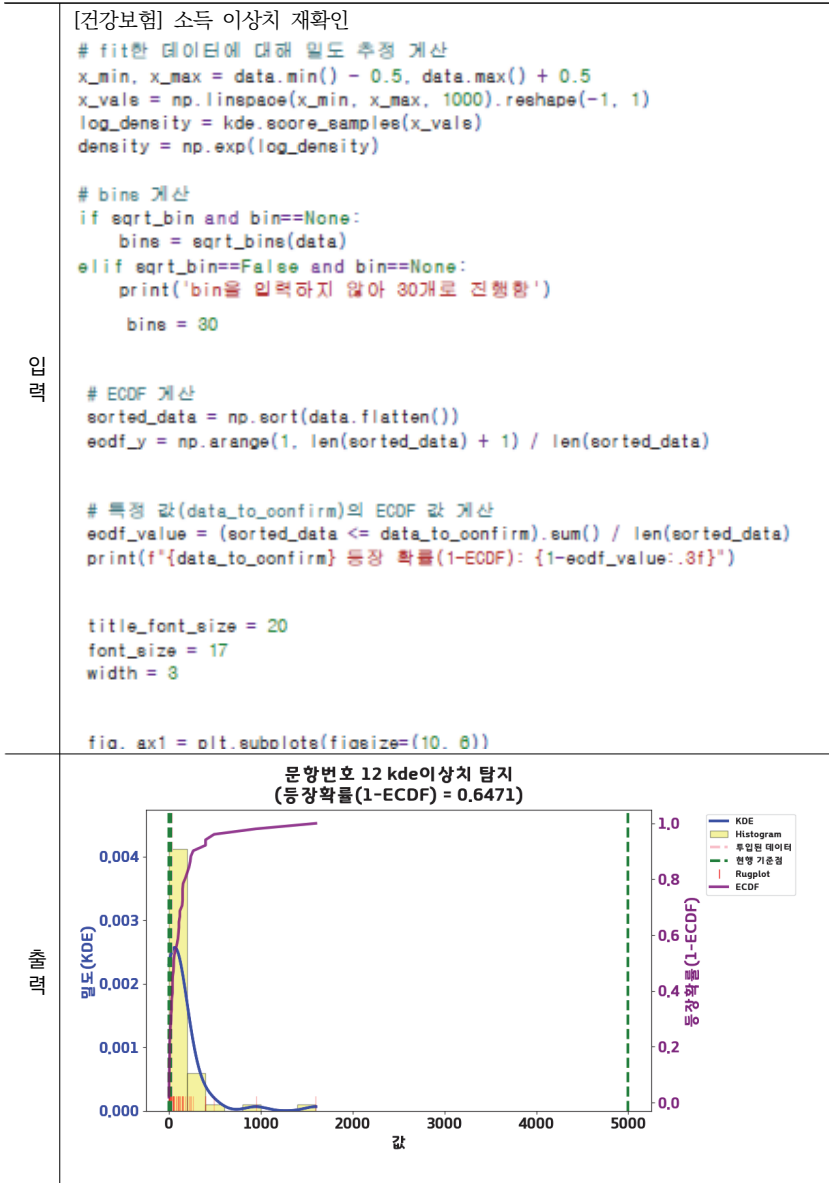
〈부표 41〉 유형 6) 문항 번호 6

코드	<pre> [재산소득_부동산/동산 소득] 이상치 재확인 # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vale = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vale) density = np.exp(log_density) # bins 계산 if eqrt_bin and bin==None: bins = eqrt_bins(data) elif eqrt_bin==False and bin==None: print('bin을 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) eodf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_oconfirm)의 ECDF 값 계산 eodf_value = (sorted_data <= data_to_oconfirm).sum() / len(sorted_data) print(f'{data_to_oconfirm} 등장 확률(1-ECDF): {1-eodf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6)) </pre>
결과	문항번호 6 kde이상치 탐지 (등장확률(1-ECDF) = 0.1032) 

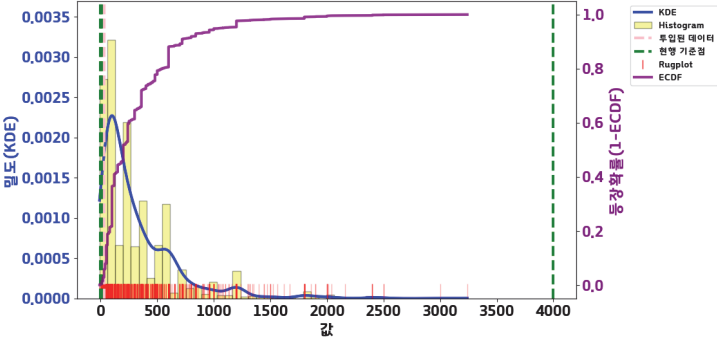
〈부표 42〉 유형 6) 문항 번호 7

인 적	<pre>[재산소득_기타소득] 이상치 재확인 # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vals = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vals) density = np.exp(log_density) # bins 계산 if sqrt_bin and bin==None: bin = sqrt_bins(data) elif sqrt_bin==False and bin==None: print('bin을 입력하지 않아 30개로 진행함') bin = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) eodf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_oconfirm)의 ECDF 값 계산 eodf_value = (sorted_data <= data_to_oconfirm).sum() / len(sorted_data) print(f'{data_to_oconfirm} 등장 확률(1-ECDF): {1-eodf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6))</pre>
세 계	문항번호 7 kde이상치 탐지 (등장확률(1-ECDF) = 0.0018) 

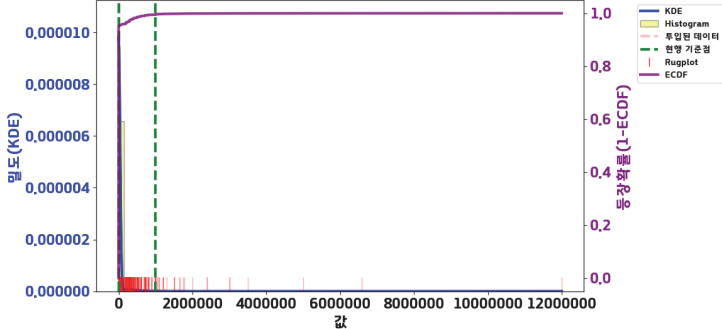
〈부표 43〉 유형 6) 문항 번호 12



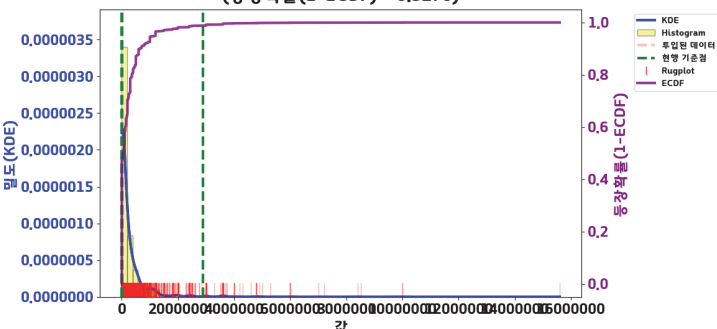
〈부표 44〉 유형 6) 문항 번호 13

코딩	[사적이전] 소득 이상치 재확인 <pre># fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vals = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vals) density = np.exp(log_density) # bins 계산 if eqrt_bin and bin==None: bins = eqrt_bins(data) elif eqrt_bin==False and bin==None: print('bins를 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) eodf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_confirm)의 ECDF 값 계산 eodf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data) print(f'{data_to_confirm} 등장 확률(1-ECDF): {1-eodf_value:.3f}')</pre> title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6))
해설	문항번호 13 kde이상치 탐지 (등장확률(1-ECDF) = 0.9425)  Legend: - KDE - Histogram - 투입된 데이터 - 한양 기준점 - Rugplot - ECDF

〈부표 45〉 유형 6) 문항 번호 18

인기	[한약 및 침약] 이상치 재확인 <pre> # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_val = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_val) density = np.exp(log_density) # bins 계산 if sqrt_bin and bin==None: bin = sqrt_bins(data) elif sqrt_bin==False and bin==None: print('bin을 입력하지 않아 30개로 진행함') bin = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_confirm)의 ECDF 값 계산 ecdf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data) print(f'{data_to_confirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6)) </pre>
세화	문항번호 18 kde이상치 탐지 (등장확률(1-ECDF) = 0.0507) 

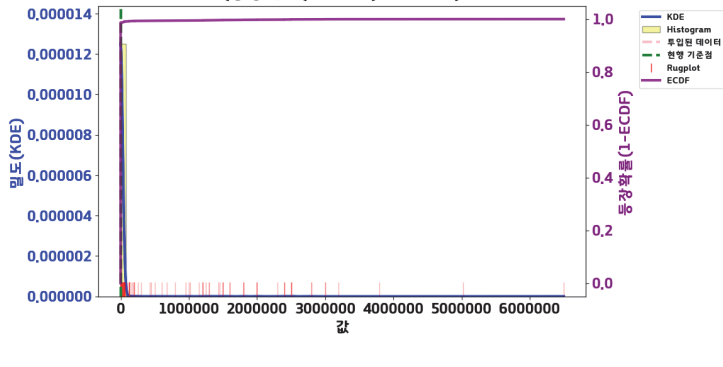
〈부표 46〉 유형 6) 문항 번호 19

인원	<pre>[건강보조식품_복용] 이상치 재확인 # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vale = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vale) density = np.exp(log_density) # bins 계산 if eqrt_bin and bin==None: bins = eqrt_bins(data) elif eqrt_bin==False and bin==None: print('bins를 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) eodf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_oconfirm)의 ECDF 값 계산 eodf_value = (sorted_data <= data_to_oconfirm).sum() / len(sorted_data) print(f'{data_to_oconfirm} 등장 확률(1-ECDF): {1-eodf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6))</pre>
내용	문항번호 19 kde이상치 탐지 (등장확률(1-ECDF) = 0.5270) 

〈부표 47〉 유형 6) 문항 번호 21

코드	<pre> [건강보조식품_복용] 이상치 재확인 # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vale = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vale) density = np.exp(log_density) # bins 계산 if eqrt_bin and bin==None: bine = eqrt_bine(data) elif eqrt_bin==False and bin==None: print('bin을 입력하지 않아 30개로 진행함') bine = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) eodf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_oconfirm)의 ECDF 값 계산 eodf_value = (sorted_data <= data_to_oconfirm).sum() / len(sorted_data) print(f'{data_to_oconfirm} 등장 확률(1-ECDF): {1-eodf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6)) </pre>
결과	문항번호 21 kde이상치 탐지 (등장확률(1-ECDF) = 0.3145)

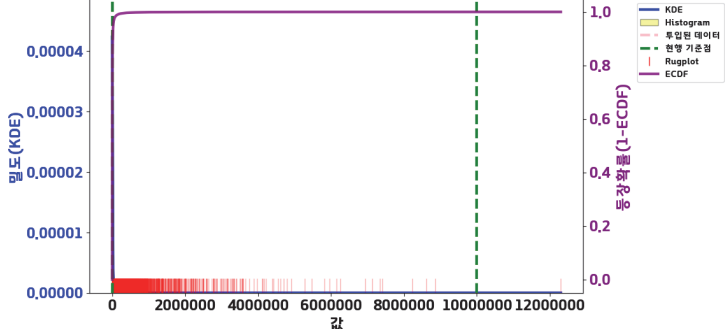
〈부표 48〉 유형 6) 문항 번호 22

근거	[청력 보조용품] 이상치 재확인 <pre># fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vals = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vals) density = np.exp(log_density) # bins 계산 if sqrt_bin and bin==None: bins = sqrt_bins(data) elif sqrt_bin==False and bin==None: print('bins을 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_confirm)의 ECDF 값 계산 ecdf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data) print(f"{data_to_confirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}") title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6))</pre>
해설	문항번호 22 kde이상치 탐지 (등장확률(1-ECDF) = 0.0133) 

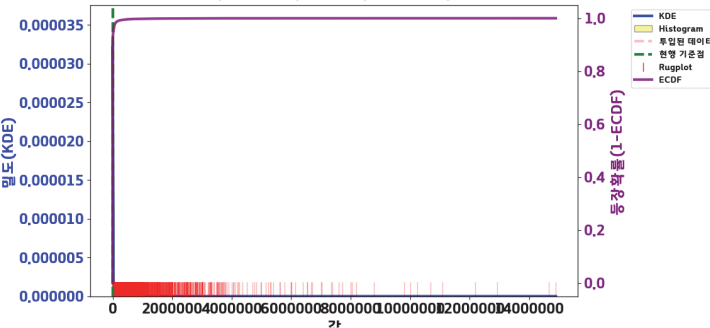
〈부표 49〉 유형 6) 문항 번호 23

코드	<pre> [의료기기 및 의료용품] 이상치 재확인 # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_val = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_val) density = np.exp(log_density) # bins 계산 if eqrt_bin and bin==None: bins = eqrt_bins(data) elif eqrt_bin==False and bin==None: print('bins를 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_confirm)의 ECDF 값 계산 ecdf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data) print(f'{data_to_confirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6)) </pre>
예시	문항번호 23 kde이상치 탐지 (등장확률(1-ECDF) = 0.1102)

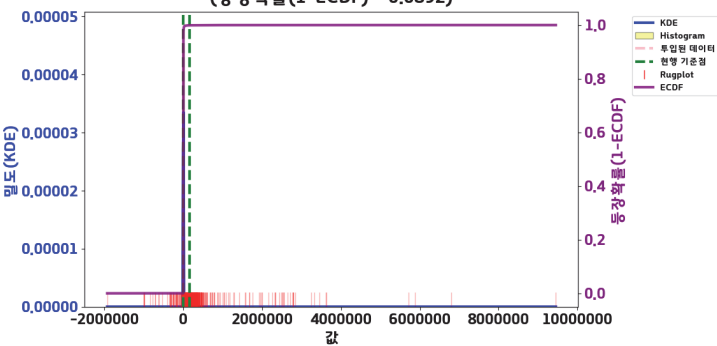
〈부표 50〉 유형 6) 문항 번호 31

인 간 전	[의료기기 및 의약품] 이상치 재확인 <pre data-bbox="265 405 992 1081"># fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vals = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vals) density = np.exp(log_density) # bins 계산 if sqrt_bin and bin==None: bins = sqrt_binst(data) elif sqrt_bin==False and bin==None: print('bin을 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_confirm)의 ECDF 값 계산 ecdf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data) print(f'{data_to_confirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6))</pre>
세 계 전	문항번호 31 kde이상치 탐지 (등장확률(1-ECDF) = 0.5524) 

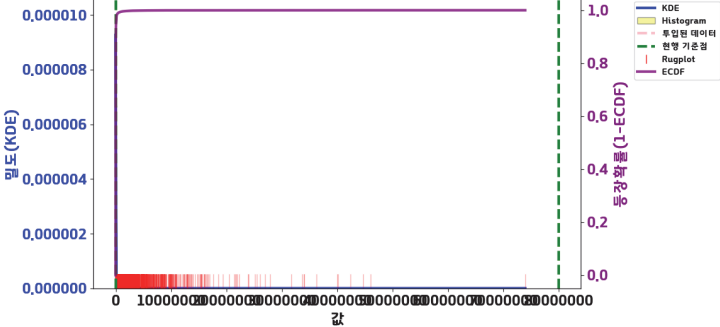
〈부표 51〉 유형 6) 문항 번호 32

인 전	[의료비 비급여 금액] 재확인 <pre> # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vale = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vale) density = np.exp(log_density) # bins 계산 if sqrt_bin and bin==None: bins = sqrt_bins(data) elif sqrt_bin==False and bin==None: print('bins를 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_confirm)의 ECDF 값 계산 ecdf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data) print(f'{data_to_confirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6)) </pre>
세 정	문항번호 32 kde이상치 탐지 (등장확률(1-ECDF) = 0.1041) 

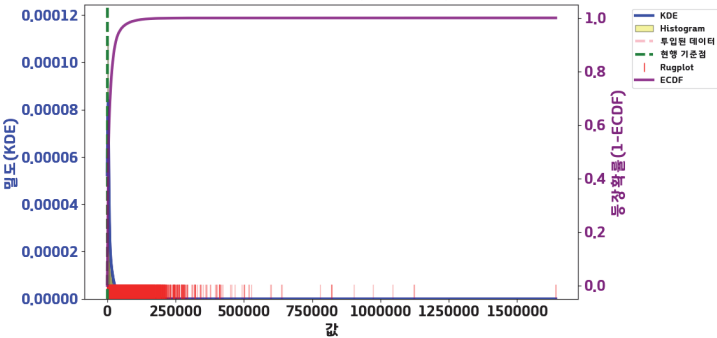
〈부표 52〉 유형 6) 문항 번호 34

인원	<pre>[의료비 할인, 절삭 등 금액]이 큰 경우 영수증 재확인 # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vals = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vals) density = np.exp(log_density) # bins 계산 if eqrt_bin and bin==None: bins = eqrt_bins(data) elif eqrt_bin==False and bin==None: print('bins을 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_confirm)의 ECDF 값 계산 ecdf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data) print(f'{data_to_confirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6))</pre>
세제	문항번호 34 kde이상치 탐지 (등장확률(1-ECDF) = 0.0892) 

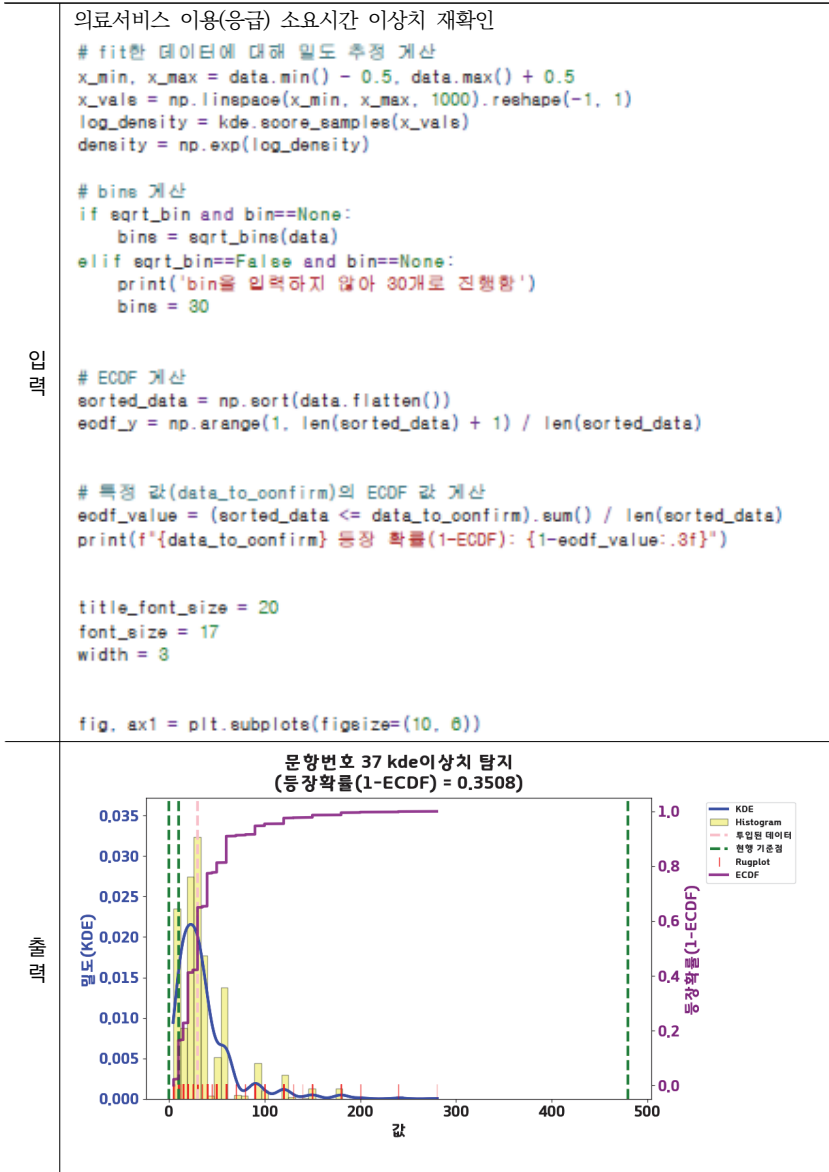
〈부표 53〉 유형 6) 문항 번호 35

기 본 판	[의료비_보험자부담금] <pre> # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vale = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vale) density = np.exp(log_density) # bins 계산 if sqrt_bin and bin==None: bins = sqrt_bins(data) elif sqrt_bin==False and bin==None: print('bins를 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) eodf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_oconfirm)의 ECDF 값 계산 eodf_value = (sorted_data <= data_to_oconfirm).sum() / len(sorted_data) print(f'{data_to_oconfirm} 등장 확률(1-ECDF): {1-eodf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6)) </pre>
세 상 판	문항번호 35 kde이상치 탐지 (등장확률(1-ECDF) = 0.5883) 

〈부표 54〉 유형 6) 문항 번호 36

연구절차	약제비 재확인 <pre># fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vals = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vals) density = np.exp(log_density) # bins 계산 if eqrt_bin and bin==None: bins = eqrt_bins(data) elif eqrt_bin==False and bin==None: print('bin을 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_confirm)의 ECDF 값 계산 ecdf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data) print(f"{data_to_confirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}") title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6))</pre>
결과	문항번호 36 kde이상치 탐지 (등장확률(1-ECDF) = 0.9571) 

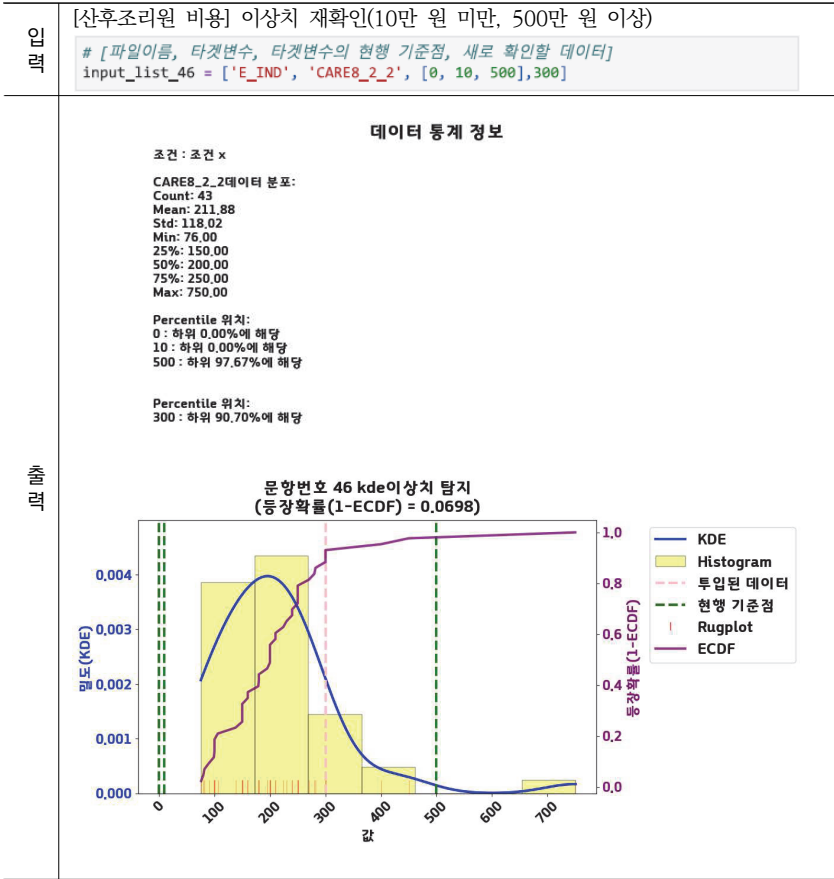
〈부표 55〉 유형 6) 문항 번호 37



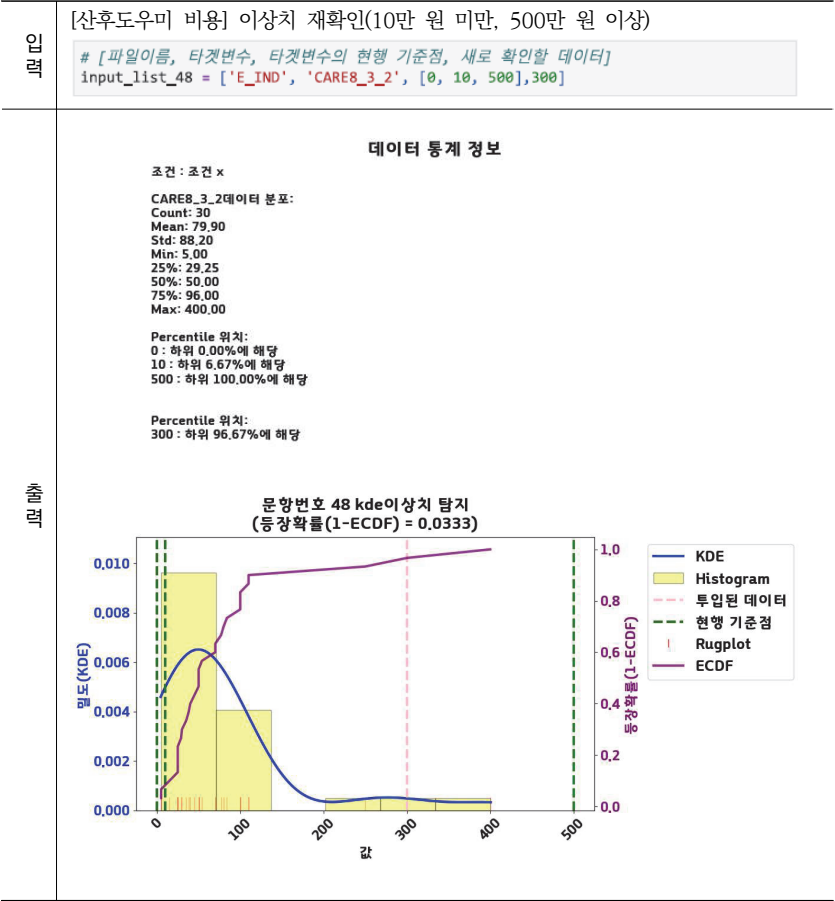
〈부표 56〉 유형 6) 문항 번호 38

코드	<pre>의료서비스 이용(입원) [유급 간병 비용] 재확인 # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vals = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vals) density = np.exp(log_density) # bins 계산 if eqrt_bin and bin==None: binc = eqrt_binc(data) elif eqrt_bin==False and bin==None: print('bin을 입력하지 않아 30개로 진행함') binc = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_confirm)의 ECDF 값 계산 ecdf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data) print(f'{data_to_confirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6))</pre>
예시	문항번호 38 kde이상치 탐지 (등장확률(1-ECDF) = 0.9919)

〈부표 57〉 유형 6) 문항 번호 46



〈부표 58〉 유형 6) 문항 번호 48



〈부표 59〉 유형 6) 문항 번호 51

```

[보장 유형] == 7번(중신/연금/변액/저축보험)

# fit한 데이터에 대해 밀도 추정 계산
x_min, x_max = data.min() - 0.5, data.max() + 0.5
x_vals = np.linspace(x_min, x_max, 1000).reshape(-1, 1)
log_density = kde.score_samples(x_vals)
density = np.exp(log_density)

# bins 계산
if eqrt_bin and bin==None:
    bins = eqrt_bins(data)
elif eqrt_bin==False and bin==None:
    print('bin을 입력하지 않아 30개로 진행함')
    bins = 30

# ECDF 계산
sorted_data = np.sort(data.flatten())
ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data)

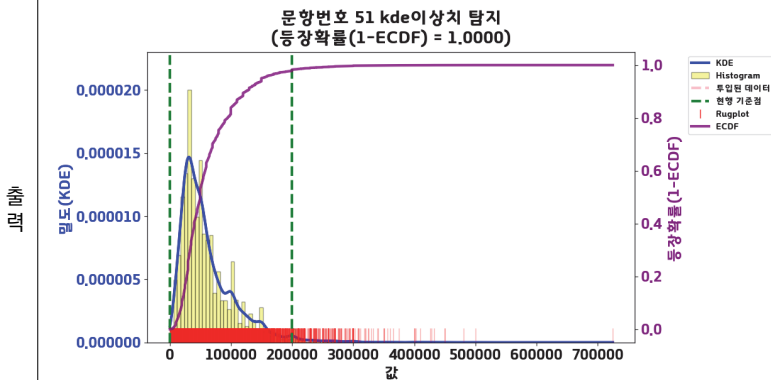
# 특정 값(data_to_confirm)의 ECDF 값 계산
ecdf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data)

print(f'{data_to_confirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}')

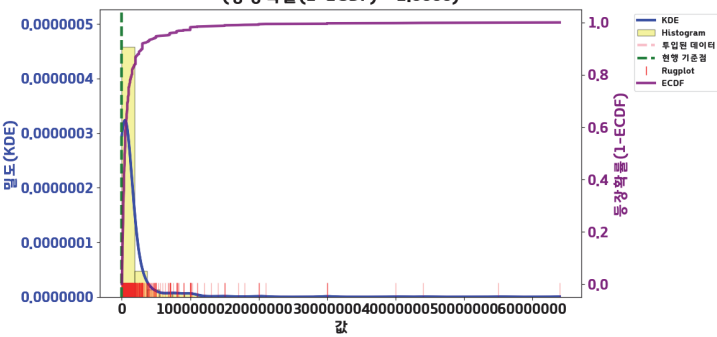
title_font_size = 20
font_size = 17
width = 3

fig, ax1 = plt.subplots(figsize=(10, 6))

```



〈부표 60〉 유형 6) 문항 번호 52

인적	<pre>[보험금 수령액] 재확인 # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_val = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_val) density = np.exp(log_density) # bins 계산 if sqrt_bin and bin==None: bins = sqrt_bins(data) elif sqrt_bin==False and bin==None: print('bin을 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_confirm)의 ECDF 값 계산 ecdf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data) print(f'{data_to_confirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}')</pre> <pre>title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 8))</pre>
세부	문항번호 52 kde이상치 탐지 (등장확률(1-ECDF) = 1.0000) 

〈부표 61〉 유형 6) 문항 번호 54

[견기-최근 일주일 하루 걸은 시간] 이상치 재확인

```
# fit한 데이터에 대해 밀도 추정 계산
x_min, x_max = data.min() - 0.5, data.max() + 0.5
x_vals = np.linspace(x_min, x_max, 1000).reshape(-1, 1)
log_density = kde.score_samples(x_vals)
density = np.exp(log_density)

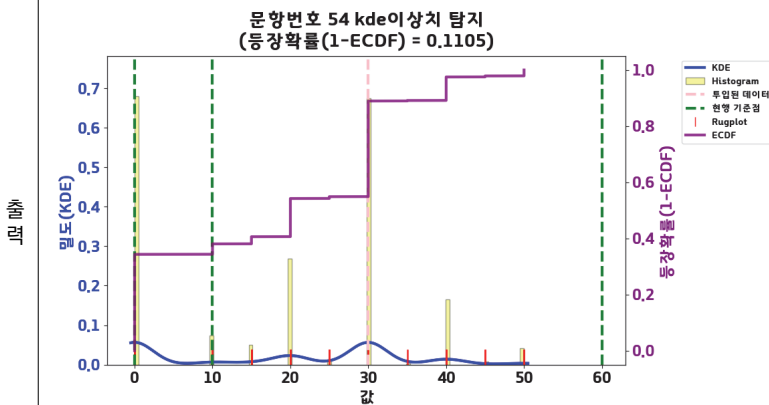
# bins 계산
if sqrt_bin and bin==None:
    bins = sqrt_bins(data)
elif sqrt_bin==False and bin==None:
    print('bins를 입력하지 않아 30개로 진행함')
    bins = 30

# ECDF 계산
sorted_data = np.sort(data.flatten())
ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data)

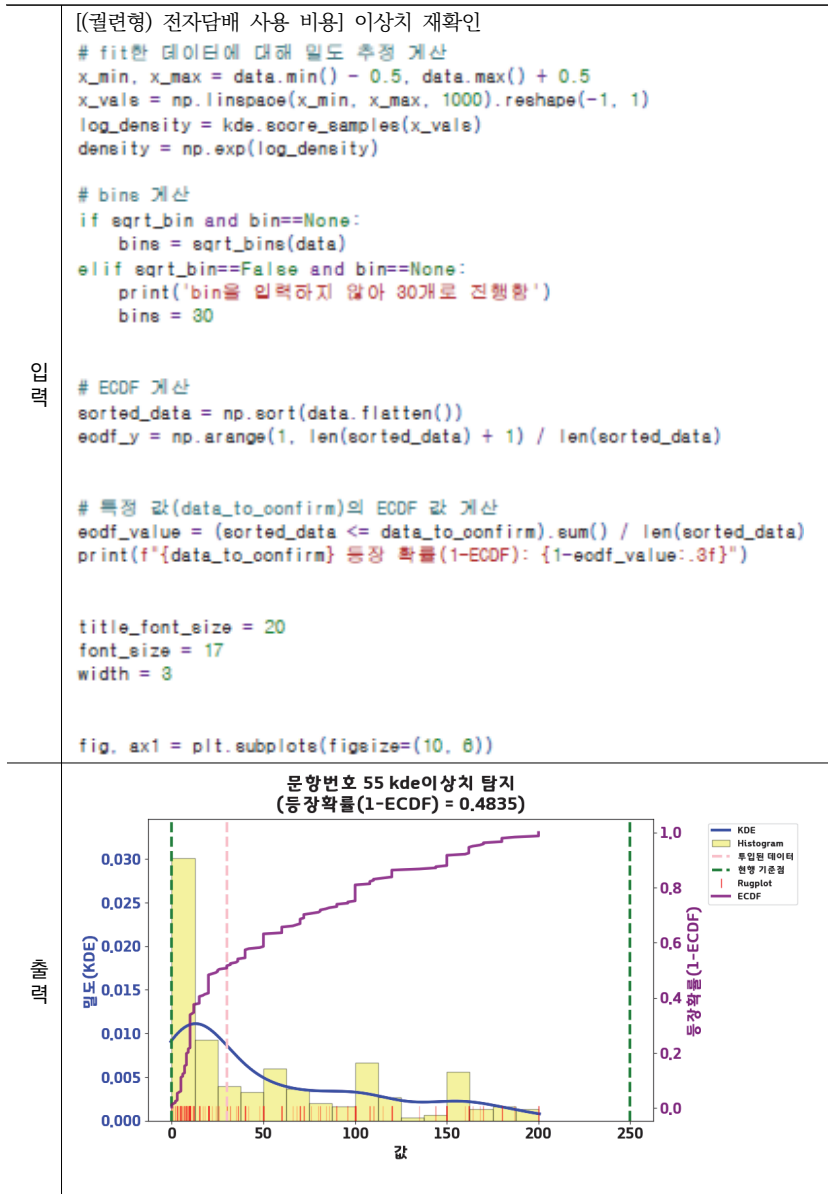
# 특정 값(data_to_confirm)의 ECDF 값 계산
ecdf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data)
print(f'{data_to_confirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}')

title_font_size = 20
font_size = 17
width = 3

fig, ax1 = plt.subplots(figsize=(10, 6))
```



〈부표 62〉 유형 6) 문항 번호 55



〈부표 63〉 유형 6) 문항 번호 56

```

[[니코틴 포함 액상] 전자담배 사용 비용] 이상치 재확인
# fit한 데이터에 대해 밀도 추정 계산
x_min, x_max = data.min() - 0.5, data.max() + 0.5
x_vale = np.linspace(x_min, x_max, 1000).reshape(-1, 1)
log_density = kde.score_samples(x_vale)
density = np.exp(log_density)

# bins 계산
if eqrt_bin and bin==None:
    bins = eqrt_bins(data)
elif eqrt_bin==False and bin==None:
    print('bin을 입력하지 않아 30개로 진행함')
    bins = 30

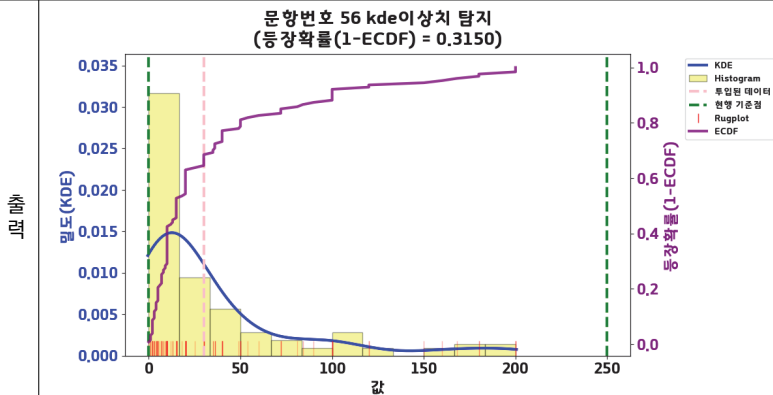
# ECDF 계산
sorted_data = np.sort(data.flatten())
eodf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data)

# 특정 값(data_to_oconfirm)의 ECDF 값 계산
eodf_value = (sorted_data <= data_to_oconfirm).sum() / len(sorted_data)
print(f'{data_to_oconfirm} 등장 확률(1-ECDF): {1-eodf_value:.3f}')

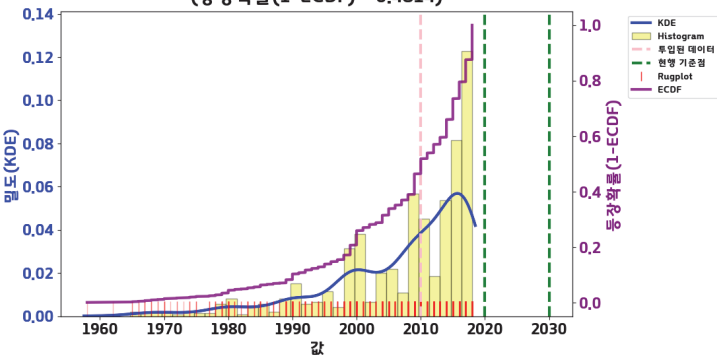
title_font_size = 20
font_size = 17
width = 3

fig, ax1 = plt.subplots(figsize=(10, 6))

```



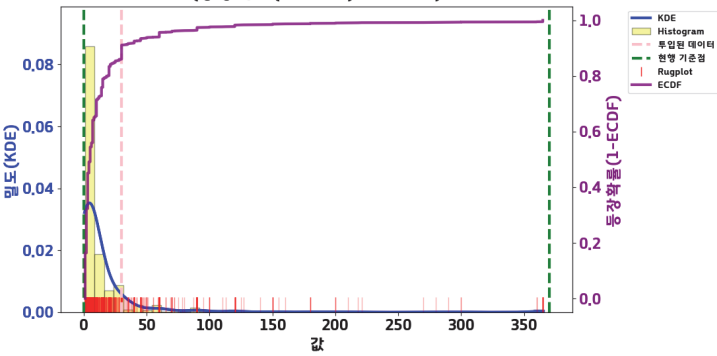
〈부표 64〉 유형 6) 문항 번호 58_1

인원	<pre>금주 시작 시기(년) 이상치 재확인 # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vals = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vals) density = np.exp(log_density) # bins 계산 if sqrt_bin and bin==None: bins = sqrt_bins(data) elif sqrt_bin==False and bin==None: print('bin을 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) eodf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_confirm)의 ECDF 값 계산 eodf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data) print(f'{data_to_confirm} 등장 확률(1-ECDF): {1-eodf_value:.3f}") title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6))</pre>
세례	문항번호 58_1 kde이상치 탐지 (등장확률(1-ECDF) = 0.4814) 

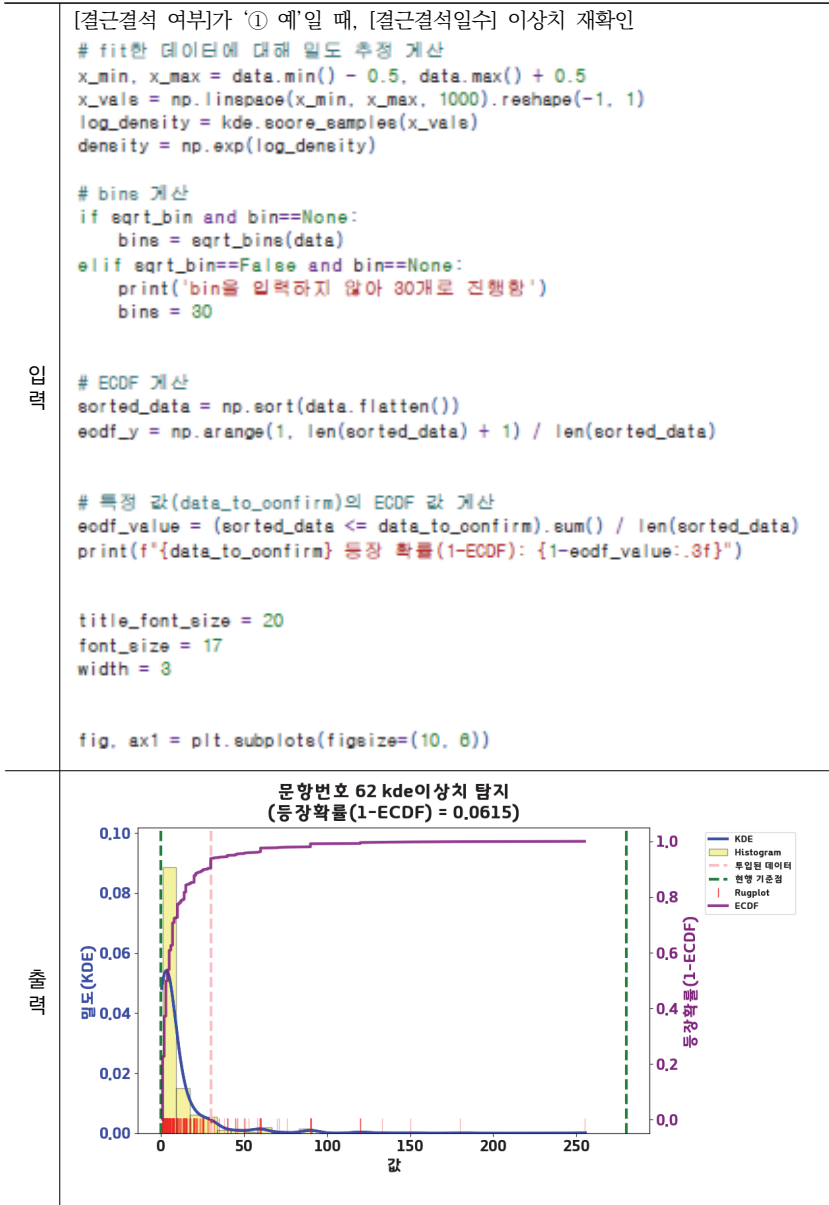
〈부표 65〉 유형 6) 문항 번호 58_2

인제	<pre> 금주 시작 시기(월) 이상치 재확인 # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vals = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vals) density = np.exp(log_density) # bins 계산 if eqrt_bin and bin==None: bins = eqrt_bins(data) elif eqrt_bin==False and bin==None: print('bins를 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_confirm)의 ECDF 값 계산 ecdf_value = (sorted_data <= data_to_confirm).sum() / len(sorted_data) print(f'{data_to_confirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6)) </pre>
제출	문항번호 58_2 kde이상치 탐지 (등장확률(1-ECDF) = 0.1766) The figure displays a KDE plot with a histogram and rugplot. The x-axis represents the data values (값) from 0 to 12. The left y-axis shows the density (밀도(KDE)) from 0.0 to 1.4. The right y-axis shows the occurrence probability (등장확률(1-ECDF)) from 0.0 to 1.0. The plot includes a blue KDE curve, a yellow histogram, a red rugplot, and a purple ECDF step function. A vertical dashed red line is at x=6, and two vertical dashed green lines are at x=0 and x=12.

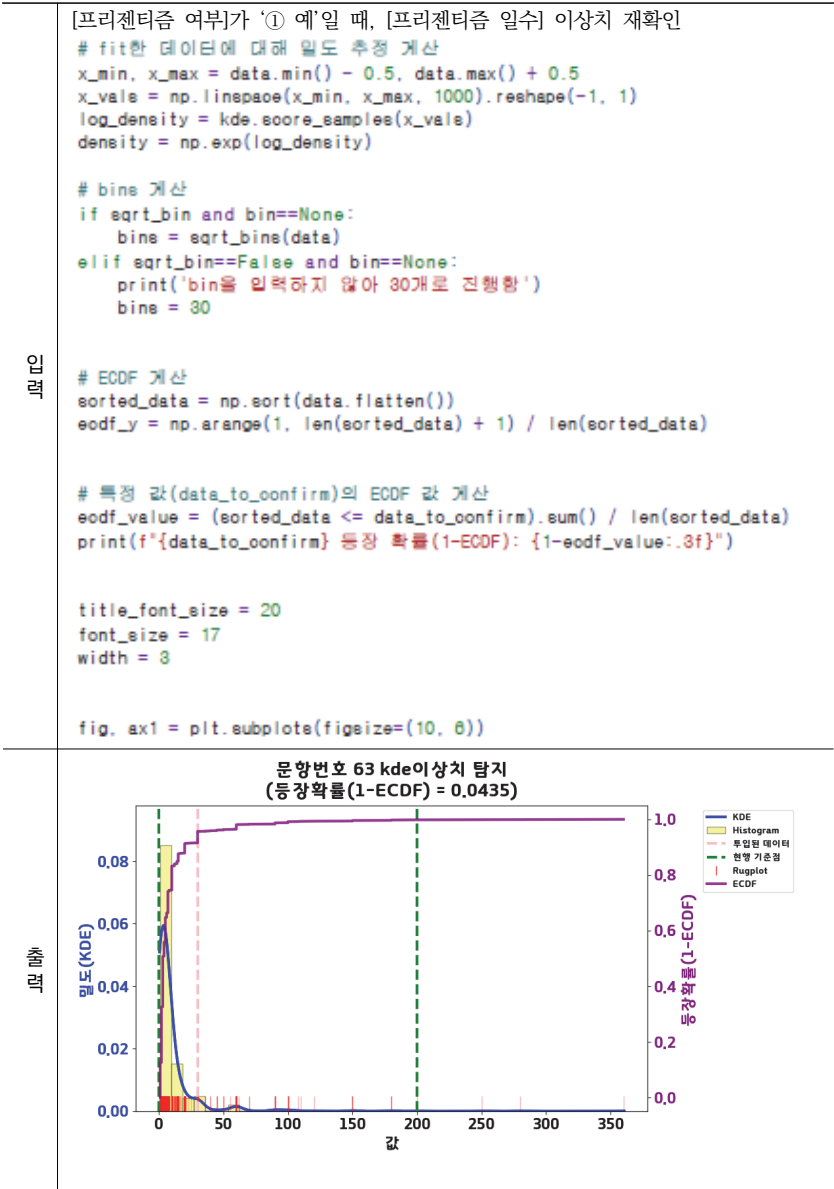
〈부표 66〉 유형 6) 문항 번호 61

코드	<pre>[와병 여부] == 1(예) 일 경우, [와병일수] 이상치 재확인 # fit한 데이터에 대해 밀도 추정 계산 x_min, x_max = data.min() - 0.5, data.max() + 0.5 x_vals = np.linspace(x_min, x_max, 1000).reshape(-1, 1) log_density = kde.score_samples(x_vals) density = np.exp(log_density) # bins 계산 if sqrt_bin and bin==None: bins = sqrt_bins(data) elif sqrt_bin==False and bin==None: print('bin을 입력하지 않아 30개로 진행함') bins = 30 # ECDF 계산 sorted_data = np.sort(data.flatten()) ecdf_y = np.arange(1, len(sorted_data) + 1) / len(sorted_data) # 특정 값(data_to_oconfirm)의 ECDF 값 계산 ecdf_value = (sorted_data <= data_to_oconfirm).sum() / len(sorted_data) print(f'{data_to_oconfirm} 등장 확률(1-ECDF): {1-ecdf_value:.3f}') title_font_size = 20 font_size = 17 width = 3 fig, ax1 = plt.subplots(figsize=(10, 6))</pre>
예시	문항번호 61 kde이상치 탐지 (등장확률(1-ECDF) = 0.0891)  Legend: - KDE - Histogram - 투입된 데이터 - 임계 기준점 - Rugplot - ECDF

〈부표 67〉 유형 6) 문항 번호 62



〈부표 68〉 유형 6) 문항 번호 63



〈부표 69〉 유형 7) 문항 번호 40~43

```

[프리젠테이션 여부가 '① 예'일 때, [프리젠테이션 일수] 이상치 재확인

# 질환 및 기준 설정
diseases = {
    "협심증": 5, "심근경색증": 6, "자궁경부암": 31, "임신 및 출산 관련": 34,
    "알코올성 간질환": 11, "통풍": 21, "등통증": 24, "유방암": 30,
    "치매": 38, "백내장": 45, "무릎관절염": 18, "무릎외관절염": 19
}

age_criteria = {
    5: (0, 10), 6: (0, 10), 31: (0, 15), 34: (0, 15),
    11: (0, 20), 21: (0, 20), 24: (0, 20), 30: (0, 20),
    38: (0, 30), 45: (0, 30), 18: (0, 40), 19: (0, 40)
}

def calculate_quantile(data):
    describe = data.describe()
    return describe

def calculate_percentile(data, conditions):
    percentile_results = []
    for value in conditions:
        percentile = percentileofscore(data, value, kind='rank')
        percentile_results.append((value, percentile))
    return percentile_results

def visualize_disease_with_info_separated(data, disease_name, disease_code, age_range):
    # 질병 코드 필터링
    filtered_data = data[
        (data['OUMED_D21'] == disease_code) |
        (data['OUMED_D22'] == disease_code) |
        (data['OUMED_D28'] == disease_code)
    ]

    # 나이 데이터 추출 및 결측값 제거
    ages = filtered_data['AGE'].dropna()

    # 빈 데이터 처리
    if ages.empty:
        print(f"{disease_name} (코드: {disease_code})에 대한 데이터가 없습니다.")
        return

    # KDE 계산
    kde = KernelDensity(kernel='gaussian', bandwidth=1.0).fit(ages.values.reshape(-1, 1))
    x_val = np.linspace(ages.min() - 5, ages.max() + 5, 1000).reshape(-1, 1)
    log_density = kde.score_samples(x_val)
    density = np.exp(log_density)

    # ECDF 계산
    sorted_ages = np.sort(ages)
    ecdf_y = np.arange(1, len(sorted_ages) + 1) / len(sorted_ages)

    # 데이터 통계와 Percentile 결과 계산
    describe = calculate_quantile(ages)
    percentiles = calculate_percentile(ages, [np.median(ages)])

    state_text = (
        f"데이터 분포:\n"
        f"Count: {describe['count']:.0f}, Mean: {describe['mean']:.2f}, Std: {describe['std']:.2f}\n"
        f"Min: {describe['min']:.2f}, 25%: {describe['25%']:.2f}, 50%: {describe['50%']:.2f}\n"
        f"75%: {describe['75%']:.2f}, Max: {describe['max']:.2f}\n"
        f"Percentile 위치:\n"
    )

    for value, percentile in percentiles:
        state_text += f"{value}: 하위 {percentile:.2f}%에 해당\n"

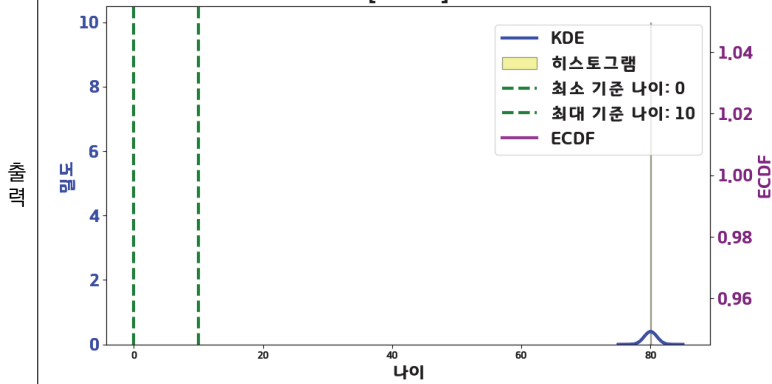
```

심근경색증 데이터 통계 정보

데이터 분포:
 Count: 1, Mean: 80.00, Std: nan
 Min: 80.00, 25%: 80.00, 50%: 80.00
 75%: 80.00, Max: 80.00

Percentile 위치:
80.0: 하위 100.00%에 해당

심근경색증 - 나이 분포
[코드: 6]

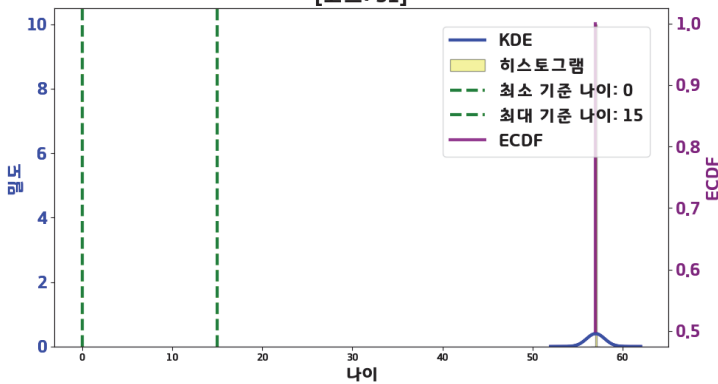


자궁경부암 데이터 통계 정보

데이터 분포:
Count: 2, Mean: 57.00, Std: 0.00
Min: 57.00, 25%: 57.00, 50%: 57.00
75%: 57.00, Max: 57.00

Percentile 위치:
57.0: 하위 75.00%에 해당

자궁경부암 - 나이 분포
[코드: 31]

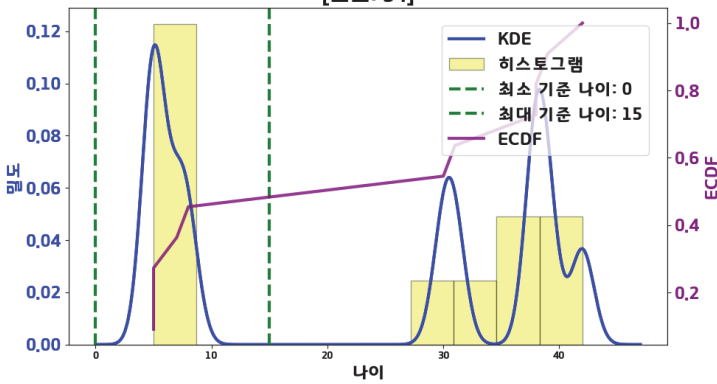


임신 및 출산 관련 데이터 통계 정보

데이터 분포:
Count: 11, Mean: 22.55, Std: 16.22
Min: 5.00, 25%: 6.00, 50%: 30.00
75%: 38.00, Max: 42.00

Percentile 위치:
30.0: 하위 54.55%에 해당

임신 및 출산 관련 - 나이 분포
[코드: 34]

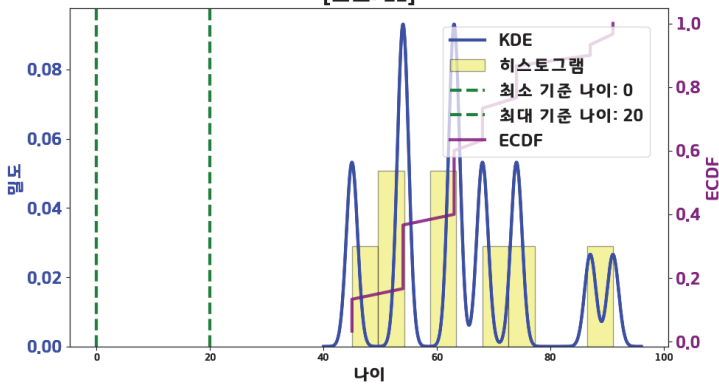


통풍 데이터 통계 정보

데이터 분포:
Count: 30, Mean: 64.10, Std: 13.30
Min: 45.00, 25%: 54.00, 50%: 63.00
75%: 72.50, Max: 91.00

Percentile 위치:
63.0: 하위 50.00%에 해당

통풍 - 나이 분포
[코드: 21]

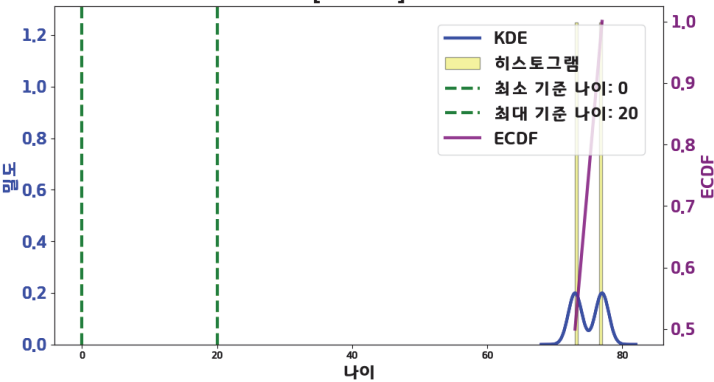


등통증 데이터 통계 정보

데이터 분포:
Count: 2, Mean: 75.00, Std: 2.83
Min: 73.00, 25%: 74.00, 50%: 75.00
75%: 76.00, Max: 77.00

Percentile 위치:
75.0: 하위 50.00%에 해당

등통증 - 나이 분포
[코드: 24]



유방암 데이터 통계 정보

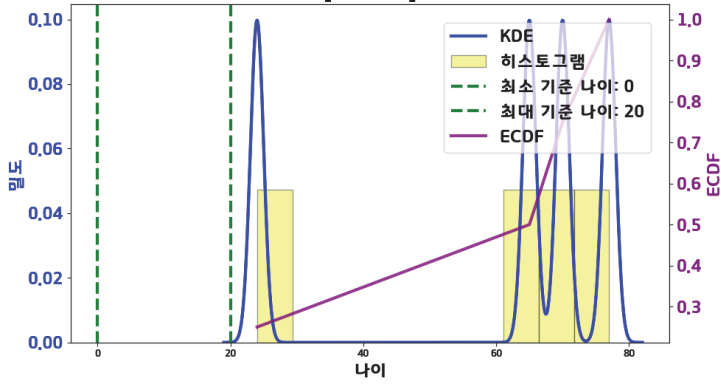
데이터 분포:

Count: 4, Mean: 59.00, Std: 23.85
 Min: 24.00, 25%: 54.75, 50%: 67.50
 75%: 71.75, Max: 77.00

Percentile 위치:

67.5: 하위 50.00%에 해당

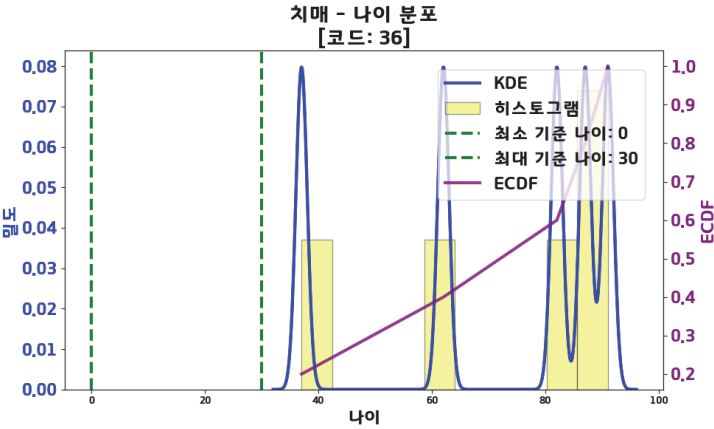
유방암 - 나이 분포
 [코드: 30]



치매 데이터 통계 정보

데이터 분포:
Count: 5, Mean: 71.80, Std: 22.42
Min: 37.00, 25%: 62.00, 50%: 82.00
75%: 87.00, Max: 91.00

Percentile 위치:
82.0: 하위 60.00%에 해당



백내장 데이터 통계 정보

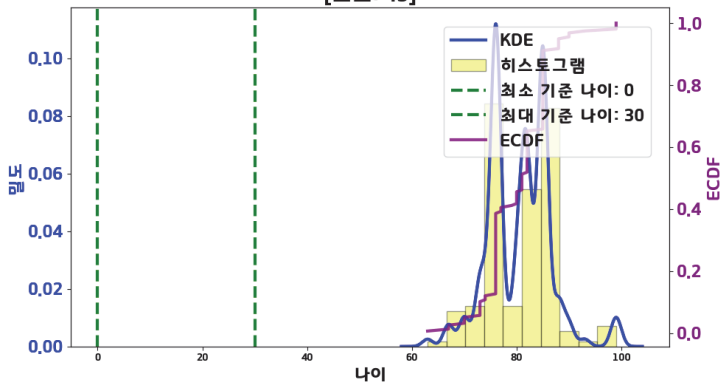
데이터 분포:

Count: 158, Mean: 80.49, Std: 6.07
 Min: 63.00, 25%: 76.00, 50%: 81.00
 75%: 85.00, Max: 99.00

Percentile 위치:

81.0: 하위 48.73%에 해당

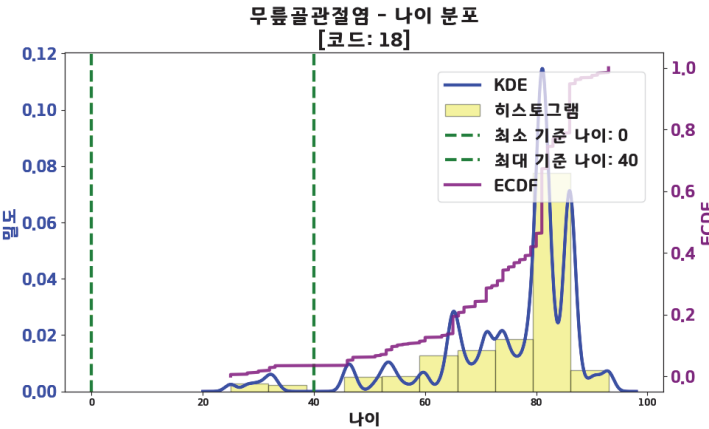
백내장 - 나이 분포
 [코드: 45]



무릎관절염 데이터 통계 정보

데이터 분포:
Count: 807, Mean: 75.16, Std: 13.31
Min: 25.00, 25%: 71.00, 50%: 81.00
75%: 83.00, Max: 93.00

Percentile 위치:
81.0: 하위 56.88%에 해당



무릎외골관절염 데이터 통계 정보

데이터 분포:

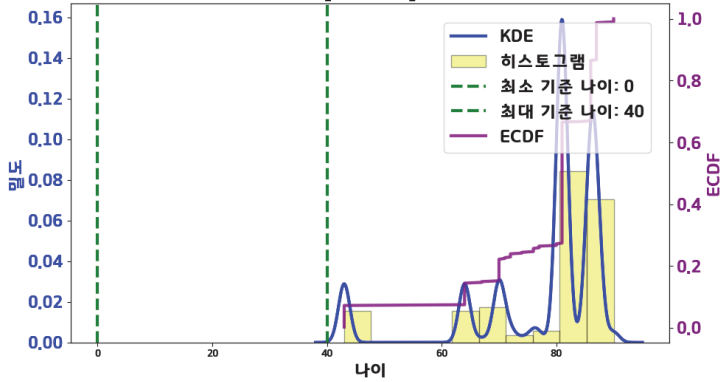
Count: 401, Mean: 77.70, Std: 11.71

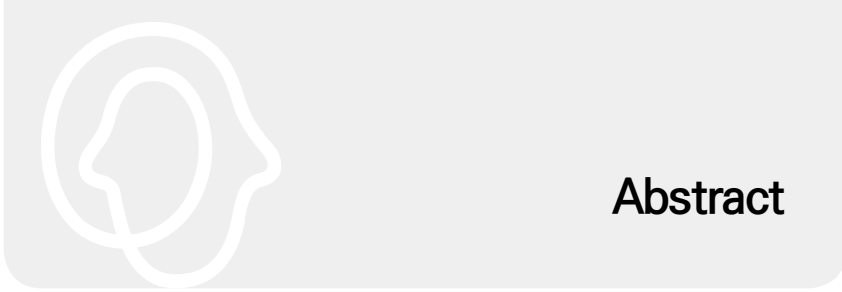
Min: 43.00, 25%: 76.00, 50%: 81.00

75%: 86.00, Max: 90.00

Percentile 위치:

81.0: 하위 47.01%에 해당

무릎외골관절염 - 나이 분포
[코드: 19]



Abstract

Strategies for Enhancing the Quality of Survey Data Using Artificial Intelligence(AI)

Project Head: Lee, Sangjung

This study aims to explore the application of artificial intelligence (AI) technology to improve the quality of survey data. As AI continues to play an increasingly vital role in data analysis and statistical processing, there is a growing need to enhance the accuracy and reliability of survey data. In particular, large-scale panel datasets often contain input errors, missing values, and logical inconsistencies, which, if left unaddressed, can undermine the credibility of research findings. Traditional methods of survey data quality management heavily rely on manual processes, making error detection and correction labor-intensive and time-consuming. To address these challenges, this study applies AI-based techniques to automate the identification and rectification of data errors, thereby improving overall data quality through empirical analysis.

AI technology has been widely utilized across various industries and research fields for data preprocessing, anomaly detection, and pattern recognition, offering greater precision and efficiency compared to conventional manual methods.

Co-Researchers: Oh, Miae · Cho, Yongchan · Yeom, Ah-rim · Kim, Eunjoo · Yoon, Yealmae · Choi, Hosik

Specifically, machine learning and natural language processing (NLP) techniques enable automatic detection and correction of errors that may go unnoticed by researchers, ultimately enhancing the reliability of analytical outcomes. This study examines how AI can be leveraged to improve survey data quality and assesses the effectiveness of AI-based data quality management models through an empirical study using data from the Korea Health Panel Survey.

In this study, AI-driven error detection algorithms were employed to automatically classify different types of errors in survey data, while machine learning techniques were utilized to identify anomalies. Furthermore, the effectiveness of AI-recommended data correction methods was compared with traditional manual data-cleaning approaches to evaluate the impact of AI on data quality enhancement. The findings indicate that AI-based quality management significantly improves the accuracy of error detection while reducing both the time and cost associated with the data-cleaning process.

Ensuring the reliability of survey data is not only critical for researchers but also for policymakers. High-quality data provides a solid foundation for sound policy decisions and facilitates more precise analyses. In particular, health and welfare panel data are essential for assessing long-term social changes and evaluating policy effectiveness, making data quality management a crucial task. This study focuses on applying AI tech-

nology to enhance survey data reliability and presents practical strategies for real-world implementation.

As AI technology continues to advance, the approaches used for survey data quality management are expected to become increasingly sophisticated. This study empirically demonstrates the impact of AI-driven automation on survey data quality and highlights the need for further research to expand AI applications across diverse survey data types. Additionally, continuous improvements in AI model performance and the development of more refined data-cleaning techniques will be essential for maximizing the utility of survey data. Given the growing role of AI in statistical analysis and survey research, this study serves as a foundational resource for future advancements in data quality enhancement and research credibility.

Key words: artificial intelligence, survey data, quality control