연구보고서 2025-49

보건복지분야 데이터 드리프트 (Data Drift) 사례 및 관리방안 연구

오미애
이정란·안수인

사람을 생각하는 사람들

KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

한국보건사회연구원
KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

▌연구진

연구책임자	**오미애**	한국보건사회연구원 선임연구위원
공동연구진	**이정란**	노키아벨랩 박사
	안수인	한국보건사회연구원 전문연구원

연구보고서 2025-49

보건복지분야 데이터 드리프트(Data Drift) 사례 및 관리방안 연구

발 행 일	2025년 12월
발 행 인	신 영 석
발 행 처	한국보건사회연구원
주 소	[30147]세종특별자치시 시청대로 370 세종국책연구단지 사회정책동(1~5층)
전 화	대표전화: 044)287-8000
홈페이지	http://www.kihasa.re.kr
등 록	1999년 4월 27일(제2015-000007호)
인 쇄 처	(사)아름다운사람들

ISBN 979-11-7252-124-0 [93510]
https://doi.org/10.23060/kihasa.a.2025.49

발|간|사

2025년 7월 대통령 직속 국정기획위원회의 AI 태스크포스(TF) 발족과 함께 보건복지분야에서도 AI와 통계 분석 기반 의사결정 시스템의 도입이 빠르게 확산되고 있다. 이에 따라 시스템의 안정성과 장기적 운영 가능성을 보장하기 위해 데이터 드리프트(Data Drift) 현상에 대한 체계적인 연구와 대응이 필요한 시점이다. 데이터 드리프트는 기계학습 모델에 사용된 데이터의 통계적 특성이 시간 경과에 따라 변화하는 것을 의미하며, 보건복지분야에서는 인구구조 변화, 복지서비스 대상 확대 등으로 인해 두드러지게 나타나고 있다.

데이터 드리프트의 탐지와 관리·모니터링 방안에 주목해야 하는 이유는 빠르게 변하고 있는 사회 현상과 폭발적으로 축적되는 데이터의 양, 분석 기술의 발전이 복합적으로 작용하면서 데이터 기반 의사결정의 불확실성과 리스크가 증대되고 있기 때문이다. 본 연구는 데이터 드리프트의 유형별 특성을 검토하고 탐지 방법들의 장단점을 체계적으로 정리하며, 공공 데이터 기반 시뮬레이션을 수행하여 데이터 기반 보건복지 행정의 신뢰성과 지속 가능성을 강화할 수 있는 실질적 방안을 제안하였다.

보건복지분야의 데이터 드리프트 관리 방안 마련으로 보다 나은 보건복지서비스 제공과 나아가 정책 효과성 제고로 이어지길 기대한다.

2025년 12월

한국보건사회연구원 원장

신 영 석

목 차

표 목차

그림 목차

요약

1. 연구의 배경 및 목적

2025년 7월 대통령 직속 국정기획위원회의 AI 태스크포스(TF) 발족과 함께 보건복지분야에서도 AI와 통계 분석 기반 의사결정 시스템의 도입이 빠르게 확산되고 있다. 이에 따라 시스템의 안정성과 장기적 운영 가능성을 보장하기 위해 데이터 드리프트(Data Drift) 현상에 대한 체계적인 연구와 대응이 필요한 시점이다. 데이터 드리프트는 기계학습 모델에 사용된 데이터의 통계적 특성이 시간 경과에 따라 변화하는 것을 의미하며, 보건복지분야에서는 인구구조 변화, 복지서비스 대상 확대 등으로 인해 두드러지게 나타나고 있다. 국제적으로 코로나19 팬데믹 이후 환자 집단 특성과 데이터 분포 급변으로 예측 모델 성능이 크게 하락한 사례가 보고되었다. 데이터 드리프트의 탐지와 관리·모니터링 방안에 주목해야 이유는 빠르게 변하고 있는 사회 현상과 폭발적으로 축적되는 데이터의 양, 분석 기술의 발전이 복합적으로 작용하면서 데이터 기반 의사결정의 불확실성과 리스크가 증대되고 있기 때문이다.

데이터 드리프트 관련 국내 연구는 기술적 방법론에 국한되고 이공계 분야 중심으로 이루어지고 있으며, 실제 공공 마이크로데이터를 활용한 심층 적용 사례는 드물다. 본 연구는 데이터 드리프트의 유형별 특성을 검토하고 탐지 방법들의 장단점을 체계적으로 정리하며, 공공 데이터 기반 시뮬레이션을 수행하여 데이터 기반 보건복지 행정의 신뢰성과 지속 가능성을 강화할 수 있는 실질적 방안을 제안하고자 한다.

2. 주요 연구 내용

제2장 데이터 드리프트 대응 기술 및 최신 방법론 고찰에서는 데이터 드리프트 유형별 특성과 탐지 방법을 제시하였다. 특성 드리프트 및 공변량 드리프트는 입력 데이터의 통계적 특성이 변하는 것으로, 복지수급자의 연령별 분포 변화, 가구구성 변화 등이 이에 해당된다. 개념 드리프트는 입력 변수와 목표 변수 간 관계가 변화하는 것으로, 코로나19로 인한 의료서비스 수요-공급 관계 변화가 대표적 사례이다. 사전확률 드리프트는 목표 변수의 분포가 변하나 입력-타깃 관계는 유지되는 현상이며, 라벨 드리프트는 분류 기준 자체가 변화하는 것으로 고혈압 진단 기준 변경 등이 이에 해당한다. 기타에는 파이프라인 드리프트, 표본선택 드리프트, 도메인 드리프트, 시간적 드리프트, 모델 드리프트 등을 살펴보았다.

탐지 방법으로는 통계 기반 방법(PSI, KS 검정, KL/JS 발산, Chi-square/t/Mann-Whitney 검정, IV), 시퀀셜 분석, 시계열 기반 방법(Page-Hinkley 테스트, EWMA 차트, DDM, EDDM, HDDM, ADWIN, FHDDM, ECDD), 머신러닝 및 딥러닝 기반 방법(CIDD-ADODNN, UDD, Meta-ADD/LSTMDD/WSCDD/CD-BTM S, Autoencoder 기반, DriftLens) 등을 기술하였다. 각 방법의 공식, 장단점, 활용 분야를 정리하여 드리프트 유형에 따라 적절한 탐지 기법을 선택할 수 있도록 하였다.

제3장 보건복지분야 데이터 드리프트 시뮬레이션에서는 실제 공공 데이터를 활용하여 세 가지 드리프트 유형별 분석을 진행하였다.

가구구성 변화(특성 드리프트)에서는 통계청 자료로 1980년부터 2024년까지 가구원수별 가구구성 변화를 분석하였다. 평균 가구원수를 비교해보면 1980년에는 4.5명에서 2024년에 2.2명으로 감소하였다. 기준연도는 1990, 2000, 2010, 2015년으로 선정하고 데이터 드리프트 탐

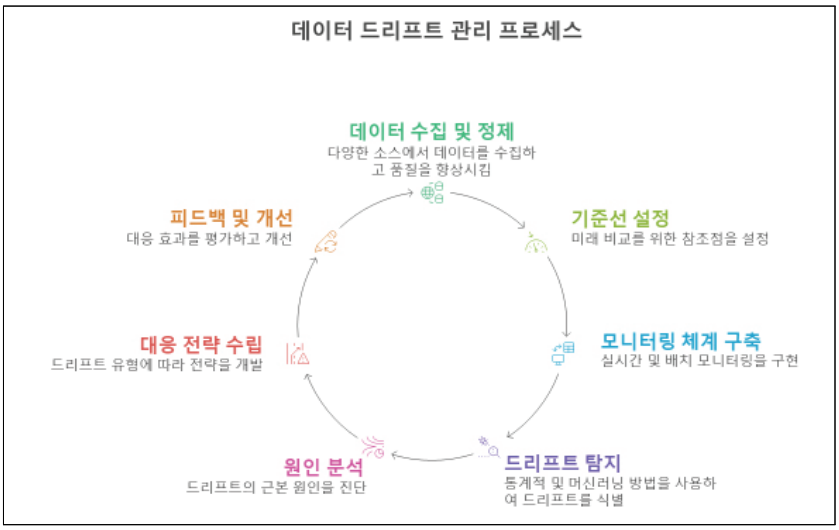
지 방법은 PSI, IV, KS를 사용하여 측도값을 제시하였다. 기준연도를 1990년으로 설정한 경우 2010년부터 PSI, IV, KS 모두 매우 높게 나타났으며, 이는 1990년대 4인 가구 중심에서 2010년 2인 가구로, 이후 1인 가구 증가로의 전환을 반영한다. 기준연도를 현재와 가까운 시점으로 할 경우 데이터 드리프트 단기적 변화에 대한 탐지력이 높아짐을 확인하였다.

고혈압 기준 변화 분석(라벨 드리프트)에서는 고혈압 진단 기준이 140/90mmHg에서 130/80mmHg로 변경될 경우의 영향을 분석하였다. 동일한 혈압 수치도 기준값 변화에 따라 다른 진단 결과를 초래하기 때문에 분류체계 자체의 변화가 모델 성능에 미치는 영향을 확인하였다.

복지사각지대 데이터 분석(모델 드리프트)에서는 가상데이터를 생성하여 입력 변수 분포 변화와 예측 확률 분포 변화를 PSI와 JSD로 분석하였다. 위기정보별 대상자 비율 변화, 드리프트 발생 변수 식별, 예측확률 분포의 변화를 통해 모델 드리프트 현상을 보여주었다.

제4장 데이터 드리프트 관리 방안에서는 데이터 드리프트 관리를 위한 7단계 프로세스를 ① 데이터 수집 및 전처리(데이터 확보, 품질 점검, 스키마 검증), ② 기준선 설정(기준 데이터셋 정의 및 임계값 결정), ③ 모니터링(데이터 분포와 모델 성능 실시간 추적 및 알림), ④ 드리프트 탐지(다양한 지표와 알고리즘을 활용한 이상 징후 포착), ⑤ 원인 분석(내·외부 요인별 영향도 평가), ⑥ 대응 전략 수립(데이터 보정, 모델 재학습 및 파라미터 튜닝), ⑦ 피드백 및 개선(성과 평가 결과를 반영한 자동화된 순환 체계 구축)으로 구성하였다. 그리고 각 단계별 구체적인 점검 항목과 실무 체크리스트를 통해 활용성을 높이고자 하였다. 제3장에서 분석한 세 가지 드리프트 유형(특성 드리프트, 라벨 드리프트, 모델 드리프트)에 대해 7단계 프로세스를 구체적으로 적용함으로써 유형별 모니터링 방안을 제시하였다.

〔요약그림 1〕 데이터 드리프트 관리 프로세스 도식화



자료: 저자가 작성한 내용을 Napkin.ai를 활용하여 시각화 하였음

3. 결론 및 시사점

제5장 결론 및 시사점에서는 이 연구의 내용을 요약하고 시사점을 제시하였다. 시사점은 다층적 접근의 필요성, 보건복지분야 특수성을 고려한 맞춤형 전략, 인간 중심 접근과 기술적 솔루션의 조화, 정책적 차원의 체계 구축으로 구분하였다. 데이터 드리프트의 효과적인 탐지와 관리를 위해서는 공변량 드리프트, 개념 드리프트 등 다양한 유형에 대한 적합한 탐지 방법을 여러 측도로 적용하고, AI 기반의 검증된 최신 기술을 활용한 모니터링 체계가 구축되어야 한다. 보건복지분야는 데이터 프라이버시와 보안이 매우 중요한 만큼, 보건복지분야의 특수성을 고려한 맞춤형 전략도 필요하다. 데이터 드리프트 탐지를 위해 지속적인 학습과 모니터링으로 시스템을 개선하되, 전문가의 판단과 개입이 함께 이루어져야 할 것이다. 데이터 드리프트 대응을 위한 데이터 품질관리 시스템 구축, 정

기적 모델 성능평가, 적응형 프로그램 설계를 통해 변화하는 환경에 효과적으로 대응해야 하며, 정책 입안자, 연구자, 실무자 등 다양한 이해관계자와의 협력을 통해 드리프트의 근본 원인을 해결하는 노력이 필요하다.

본 연구는 실제 공공 데이터를 활용한 데이터 드리프트 구체적 사례 분석과 실무 적용 가능한 7단계 관리 프로세스를 제시함으로써, 보건복지 분야 특성을 반영한 실효성 있는 관리방안을 마련하였다는 점에서 의의가 있다.

주요 용어: 보건복지, 데이터 드리프트, 기술, 모니터링 방안

제1장

서론

제 1 장 서론

제1절 연구의 배경 및 목적

2025년 7월, 대통령 직속 국정기획위원회는 인공지능(AI) 태스크포스(TF)를 공식 발족하였다. 이 TF는 '모두의 AI'를 핵심 비전으로 설정하고, 국가 차원의 AI 혁신 로드맵을 설계하며 다양한 세부 사업을 단계적으로 추진하고 있다. 이러한 정책 환경 변화와 맞물려 보건복지분야에서도 AI와 통계 분석을 기반으로 한 의사결정 시스템의 도입이 빠르게 확산되고 있다. 이에 따라 시스템의 안정성과 장기적인 운영 가능성을 보장하기 위해 데이터의 품질 관리, 특히 데이터 드리프트(Data Drift) 현상에 대한 체계적인 연구와 대응이 시급한 과제로 대두되고 있다.

"Drift"는 기계학습 모델을 학습하는 데 사용된 데이터의 통계적 특성이 시간이 지남에 따라 변경되는 것을 의미하며, 데이터 드리프트(Data Drift)는 모델이 적용되는 데이터의 변화하는 분포를 나타낸다. 특히 사회정책 분야에서는 인구구조 변화, 기후변화 그리고 복지서비스 대상의 확대 등으로 인해 데이터 드리프트 현상이 두드러지게 나타나고 있다. 이러한 변화는 기존의 데이터 분석 모델의 정확성과 신뢰성에 큰 영향을 미치고 있다. 예를 들어, 공공 건강 분석 분야를 살펴보면, 최근의 팬데믹으로 인한 인구 이동 패턴의 변화가 질병 발생 예측 모델의 정확도에 상당한 영향을 미치고 있다. 사회 복지 프로그램에서도 급격한 경제적 조건 변화로 인해 복지 급여 수급 자격의 평가 시스템에 대한 편향성 문제가 대두될 수 있다. 이러한 데이터 드리프트는 새로운 건강 정책이나 사회 복지 규정의 도입, 대중의 행동 변화 그리고 팬데믹이나 경제 침체와 같

은 외부 요인들로 인해 발생한다.

특히 주목할 만한 점은 데이터 드리프트가 모델 성능 저하, 자원 할당의 비효율성 그리고 윤리적 형평성 문제를 야기할 수 있다는 것이다. 예를 들어, 복지 프로그램을 이용하는 인구의 인구통계학적 변화나 공공 건강 트렌드의 변화는 기존 모델의 예측 정확도를 크게 떨어뜨릴 수 있다. 또한 코로나19 팬데믹의 충격으로 생활고를 겪는 근로자 및 자영업자 급증으로 긴급복지지원제도의 사회 복지 신청 건수 증가와 같은 상황은 기존 모델의 예측 능력을 무력화시킬 수 있다.

국제적으로도 데이터 드리프트의 중요성은 이미 입증되었는데, 코로나19 팬데믹 이후 환자 집단 특성과 데이터 분포가 급변하면서, 여러 의료기관에서 기존 예측 모델의 성능이 크게 하락한 사례가 보고되었다(Kore et al., 2024). 예를 들어, 영국 NHS는 2020년 4월 응급실 이용 건수가 전년 동월 대비 약 57% 감소하는 등 환자 흐름과 질병 패턴이 급격히 변했고, 이로 인해 팬데믹 기간 중 입원 예측 모델의 성능이 이전보다 현저히 떨어졌다(Duckworth et al., 2021). 데이터 드리프트를 적시에 탐지·관리하지 못하면 정책 결정의 신뢰도 저하, 예측 오류 확대, 서비스 품질 저하, 대상자 선정의 불공정성 등 심각한 문제가 발생할 수 있다.

국내에서도 데이터 드리프트 탐지와 대응을 주제로 한 연구가 증가하고 있으나, 상당수가 기술적 방법론에 국한되고 이공계 분야 중심으로 이루어지고 있다. 또한 활용되는 데이터는 실제 공공 마이크로데이터보다 비공개 행정자료(유호범, 2022)나 제한된 범위의 데이터셋에 편중되는 경향이 있으며(Kwon & Baek, 2020; 이정욱, 2021; 이예은·이태진, 2023; 최옥주·김유경, 2024; Park et al., 2024; 강현우·남덕윤, 2025; 나경민·김도형·이영호, 2025), 실제 공공 데이터를 활용한 심층 적용 사례는 드물다.

반면 EU, 미국, 영국, 호주 등 주요국은 데이터 드리프트 조기 탐지와 실시간 모니터링 체계 도입을 적극 추진하고 있으며, WHO(2021)의 「글로벌 디지털 헬스 전략(2020-2025)」에서도 데이터 품질 관리, 개인정보 보호, 데이터 무결성 및 신뢰성 확보를 핵심 가치로 제시하고 있다.

현재 공공 데이터를 활용한 데이터 드리프트 연구는 여전히 제한적이고, 그마저도 대부분이 단순 분포 변화 관찰이나 공변량·개념 드리프트 탐지 수준에 머무르고 있다(이상연 외, 2023; Kang, Lee, & Kang, 2024; Kim et al., 2024). 따라서 보건복지분야 정책결정의 신뢰성을 높이기 위해서는 실제 행정 데이터와 마이크로데이터를 활용한 체계적 분석과 실효성 있는 관리방안 마련이 필요하다.

이를 위해 지속적인 모니터링 시스템 구축, 정기적인 모델 재학습, 데이터 품질 관리 강화가 필수적이다. 또한 설명 가능한 모델을 사용하고 공정성을 고려한 알고리즘을 도입하여 의도하지 않은 편향을 최소화해야 한다. 더불어 정책 입안자와 의료 제공자 등 다양한 이해관계자들과의 협력을 통해 드리프트의 근본적인 원인을 해결하는 노력도 필요하다.

결론적으로 데이터 드리프트 방법론에 대한 체계적인 검토와 대응은 현대 사회의 복잡한 변화에 효과적으로 대처하기 위한 필수적인 과제가 되었다. 이는 단순히 기술적인 문제를 넘어서, 사회정책의 효과성과 형평성을 보장하기 위한 중요한 연구 분야로 자리잡고 있다. 본 연구에서는 데이터 드리프트의 유형별 특성을 검토하고, 이를 탐지하는 방법들과 그 방법들별 장단점을 체계적으로 정리하고자 하였다. 그리고 공공 데이터를 기반으로 시뮬레이션을 수행하였다. 또한 관리 전략과 정책적 시사점을 도출함으로써, 데이터 기반 보건복지 행정의 신뢰성과 지속 가능성을 강화할 수 있는 실질적 방안을 제안하고자 한다.

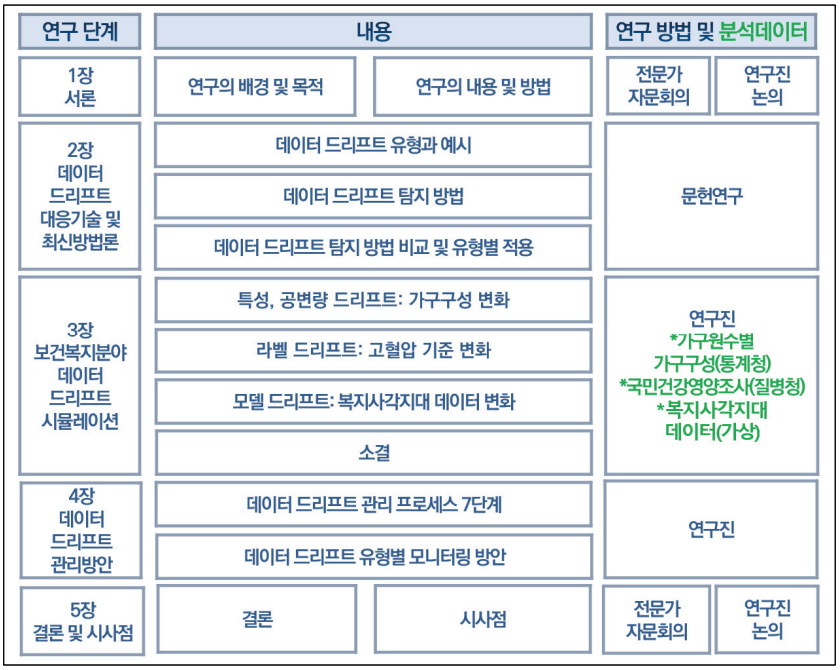
제2절 연구의 내용 및 방법

보고서 「보건복지분야 데이터 드리프트(Data Drift) 사례 및 관리방안 연구」의 주요 연구 내용은 다음과 같다. 제2장에서는 데이터 드리프트 유형과 예시, 데이터 드리프트 탐지 방법, 데이터 드리프트 탐지 방법 비교 및 유형별 적용사례를 검토하였다. 제3장에서는 보건복지분야 데이터 드리프트 분석 결과를 데이터 드리프트 유형별로 제시하였다. 매크로데이터를 활용한 가구구성 변화에 따른 특성, 공변량 드리프트 분석, 마이크로데이터를 활용한 고혈압 기준 변화에 따른 라벨 드리프트 분석, 가상의 복지사각지대 데이터를 활용한 모델 드리프트 분석으로 구체적인 드리프트 사례를 제시하고자 하였다. 제4장에서는 데이터 드리프트 관리 프로세스를 7단계로 제시하였고, 각 단계별 검토사항도 기술하였다. 그리고 3장의 데이터 드리프트 유형별로 모니터링 방안을 제시하였다. 마지막으로 제5장에서는 결론과 시사점을 제시하였다.

이 보고서 작성을 위해 국내외 문헌 연구, 전문가 자문회의, 데이터 분석(R/Python) 등 다양한 방법을 활용하였다.

연구 수행 과정 개요는 다음과 같다.

〔그림 1-1〕 연구 수행 과정 개요



자료: 저자 작성.

제2장

데이터 드리프트(Data Drift) 대응 기술 및 최신 방법론 고찰

제 2 장 데이터 드리프트(Data Drift) 대응 기술 및 최신 방법론 고찰

제1절 데이터 드리프트 유형과 예시

앞서 언급한 것처럼, “Drift”는 기계학습 모델을 학습하는데 사용된 데이터의 통계적 특성이 시간이 지남에 따라 변경되는 것을 의미하며, 데이터 드리프트(Data Drift)의 정의는 학습 모델이 적용되는 데이터가 변화할 때 그 분포를 의미한다. 이제 사회 공공복지 및 보건정책에서의 데이터 드리프트 유형과 특징을 살펴보자. 본 절에서는 보건복지 데이터 분석 과정에서 나타날 수 있는 주요 드리프트 유형을 체계적으로 분류하고, 이해를 돕기 위한 실제 정책·행정 사례를 제시한다.

1. 특성 드리프트(Feature Drift) 혹은 공변량 드리프트(Covariate Drift)

특성 혹은 공변량 드리프트는 시간 경과에 따른 입력 데이터의 통계적 특성이 변하는 것을 의미한다(Webb et al., 2016). 입력 데이터의 특성이 변하는 동안 입력과 반응 변수 사이의 관계는 변하지 않기 때문에 ‘입력 드리프트’(Input Drift)라고도 한다. 다시 말해, 모델을 처음 구축했을 때와 비교하여 모델 학습에 사용된 독립변수(특성)들의 분포가 변화했음을 의미한다. 중요한 점은 입력 특성과 목표 변수(모델이 예측하려는 대상) 간의 관계는 동일하게 유지된다는 것인데, 모델은 여전히 동일한 결과를 예측하려 하지만, 입력 데이터가 달라진 경우이다. 변수 간의 근본적인 관계는 변하지 않았지만, 공변량 드리프트는 모델이 학습하지 않은

데이터에 노출되기 때문에 모델의 성능에 부정적인 영향을 미칠 수 있다. 일반적으로 특성 드리프트는 사용자 행동 변화, 시장 상황 또는 새로운 제품이나 기능의 도입 등 다양한 요인으로 인해 발생할 수 있다. 예를 들어, 사회 정책 수립과 관련된 주제들에서는 다음과 같은 입력 데이터의 특성 변화가 발생할 수 있다.

- □ 복지수급자의 연령별, 소득별 분포 변화: 과거 vs. 현재의 연령분포 혹은 소득 분포 변화 비교

- □ 의료서비스 이용 패턴의 변화: 과거 vs. 현재의 외래진료, 입원진료, 원격진료 비율 분포 혹은 진료과목(내과, 정형외과, 기타 등등)의 분포 변화 비교

- □ 주거복지 수요 특성의 변화: 과거 vs. 현재의 주택유형(아파트, 다세대주택, 단독주택) 혹은 가구구성(4인 가구, 2~3인 가구, 1인 가구 등)의 분포 변화 비교

구체적으로 실제 자료를 통해 나타내자면, 과거에 대비해 최근 가구의 연령별 소득 등의 분포 변화를 예로 들 수 있다. 〈표 2-1〉은 2015년 분포이고, 〈표 2-2〉는 2023년 분포이다. 이 중 가구 비중과 시장소득을 연도에 따라 나누어 비교해 보면 각각 [그림 2-1], [그림 2-2]와 같다.

〈표 2-1〉 2015년 가구주 연령대별 평균 소득 및 자산보유액

(단위: 만 원/연, 만 원, %)

구분(2015년)	30대 이하	40대	50대	60~64세	65세 이상
가구 비중	19.0	25.8	25.5	8.7	21.0
시장소득	2,853	3,036	3,409	2,699	1,309
경상소득	2,902	3,076	3,473	3,039	1,759
가처분소득	2,390	2,493	2,834	2,553	1,521
총자산	13,614	18,077	24,442	30,358	24,547
총부채	3,096	3,906	4,625	4,928	2,949
순자산	10,519	14,171	19,816	25,430	21,598
금융자산	5,158	5,475	6,448	6,698	3,841
전월세보증금	2,360	1,626	1,101	797	588
실물자산	8,456	12,602	17,994	23,660	20,706
거주주택	5,347	6,972	8,620	11,436	10,849

자료: 강신욱 외, 2016, 저소득층의 소득-자산분포를 통해 본 사회보장제도 재산기준의 개선 방향.

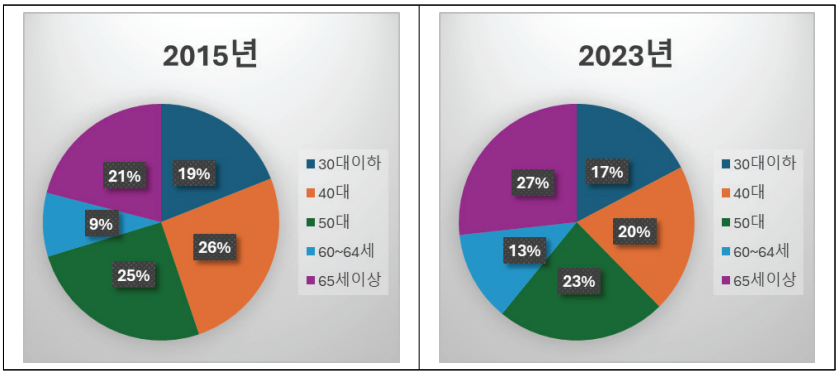
〈표 2-2〉 2023년 가구주 연령대별 평균 소득 및 자산보유액

(단위: 만 원/연, 만 원, %)

구분(2023년)	30대 이하	40대	50대	60~64세	65세 이상
가구 비중	17.3	20.3	23.3	39.1	26.7
시장소득	6,412	8,842	8,596	4,361	3,121
경상소득	6,664	9,083	8,891	5,512	4,375
가처분소득	5,455	7,182	7,125	4,680	3,782
총자산	33,615	56,122	60,452	54,836	50,714
총부채	9,937	12,531	10,715	6,206	5,174
순자산	23,678	43,590	49,737	48,630	45,540
금융자산	13,347	14,746	14,713	9,862	8,080
전월세보증금	7,275	5,225	3,116	1,791	1,514
실물자산	20,267	41,376	45,739	44,974	42,635
거주주택	12,518	25,930	26,026	24,163	22,966

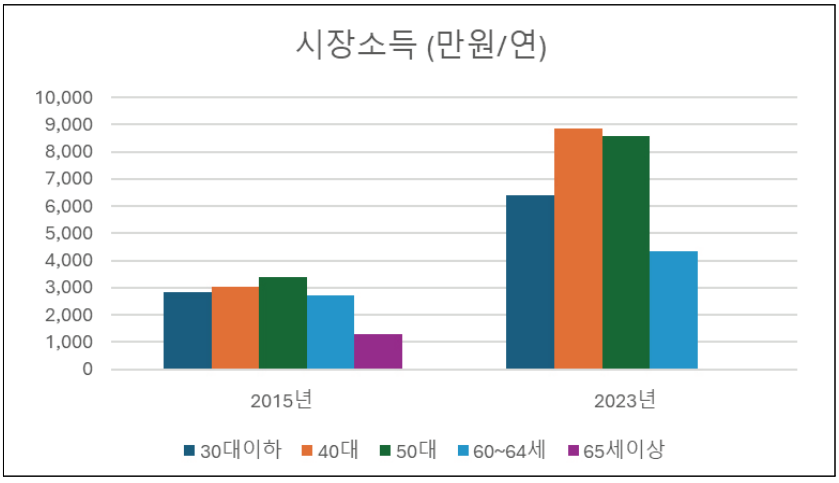
자료: 통계청, 2024, 「가계금융복지조사결과」.

〔그림 2-1〕 가구주 연령대 비율의 시간적 분포 변화



자료: 통계청, 「가계금융복지조사 결과」(2015, 2024년) 통계표를 바탕으로 저자 작성.

〔그림 2-2〕 가구주 연령대별 시장소득의 시간적 분포 변화(2015 vs. 2023)



주: 2023년은 60세 이상 전체 평균
자료: 통계청, 「가계금융복지조사 결과」(2015, 2024년) 통계표를 바탕으로 저자 작성.

2. 개념 드리프트(Concept Drift)

개념 드리프트(Concept Drift)는 입력 변수와 목표 변수 간의 관계성이 시간이 지남에 따라 변하는 상황을 의미한다(Gama et al., 2014). 이로 인해 과거 데이터로 학습된 모델이 미래 결과를 예측하는 정확도가 낮아지게 된다. 즉, 모델이 세상의 작동 방식에 대해 가정했던 내용이 더 이상 유효하지 않게 되어 성능 저하가 발생하는 것이다. 앞서 설명한 특성 혹은 공변량 드리프트를 개념 드리프트와 구별하는 것이 중요하다. 개념 드리프트에서는 입력 특성과 목표 변수 간의 관계가 변화한다. 즉, 개념 드리프트에서는 문제 자체가 변화하지만, 특성 혹은 공변량 드리프트에서는 문제는 동일하게 유지되고 입력 데이터의 분포만 변화한다. 예를 들어, 보건사회 정책 수립과 관련해서는 다음과 같은 경우들을 개념 드리프트로 생각할 수 있다.

□ 코로나19로 인한 의료서비스 수요-공급 관계 변화:

○ 변화 전: 환자 수(입력) → 필요 병상 수(출력) 관계가 선형적이고 예측 가능

○ 변화 후: 갑작스런 감염병 확산으로 동일 환자 수에도 필요 병상 수가 급증

○ 기존 모델 실패 원인: 사회적 거리두기 요구, 격리병상 필요성 등 새로운 변수 발생

□ 건강결정요인의 영향력 변화:

○ 변화 전: 소득/교육수준(입력) → 건강수준(출력)이 강한 양의 상관관계

○ 변화 후: 디지털 정보격차, 원격의료 접근성 등 새로운 요인들의 영향력 증가

○ 기존 모델 실패 원인: 전통적 사회경제적 지표만으로는 건강 불평등 설명 불가능

개념 드리프트는 일반적으로 데이터 분포가 시간에 따라 변하는 동적 환경에서 발생하여 온라인 학습 알고리즘에 어려움을 준다. 모델 성능에 미치는 영향은, 입력과 출력 간의 관계가 변함에 따라 모델의 예측 정확도가 떨어진다는 점이다. 일반적인 예로는, 의류 판매를 예측하는 모델은 여름에는 성과가 좋지만 고객 선호도의 계절적 변화로 인해 겨울에는 성과가 좋지 않을 수 있다. 또는 결혼이나 이사와 같은 일상적 이벤트 이후 사용자의 쇼핑 행동이 크게 변할 수 있다.

개념 드리프트에는 다음과 같은 유형들이 있다.

□ 갑작스러운 드리프트(Sudden(abrupt) drift): 입력과 출력 간의 관계가 빠르고 즉각적으로 변화

□ 점진적 표류(Gradual(incremental) drift): 시간이 지남에 따라 관계에서 나타나는 느리고 점진적인 변화

□ 반복적 표류(Recurring drift): 계절적 패턴처럼 초기 관찰 이후 반복적으로 나타나는 변화

3. 사전확률 드리프트(Prior Probability Drift)

사전확률 드리프트(Prior Probability Drift)는 머신 러닝 모델에서 목표(타겟) 변수의 분포가 시간이 지남에 따라 변하는 반면, 입력 특성과 타겟 변수 간의 관계는 동일하게 유지되는 현상을 말한다(Žliobaitė et al., 2016). 간단히 말해, 해당 클래스를 예측하는 특성은 변하지 않더라도, 타겟 변수에서 각 클래스의 전체 비율이 변하는 현상으로 이해할 수 있다. 보건복지분야에서 구체적으로 다음과 같은 예들을 사전확률 드리프트 현상으로 볼 수가 있다.

□ 기초생활보장 수급자 비율 변화:

○ 입력-타겟 관계 유지: 소득수준, 재산상태, 부양의무자 유무 등이 수급자격 판정에 미치는 영향은 동일

○ 사전확률 이동(가상 예시): 전체 인구 중 수급자 비율이 5%에서 8%로 증가, 2015년 소득 150만 원/재산 5천만 원 → 수급자 판정 2023년에도 동일 기준 적용되나, 전체 수급자 비율만 증가

□ 노인복지서비스 수요 증가:

○ 입력-타겟 관계 유지: 연령, 건강상태, 소득수준이 서비스 필요도 판정에 미치는 영향은 동일

○ 사전확률 이동(가상 예시): 고령화로 인해 서비스 수요자 비율이 15%에서 25%로 증가, 2015년 75세/거동불편/저소득
→ 재가서비스 대상 2023년에도 동일 기준이나, 전체 수요자 비율만 증가

□ 장애인 등록 기준 변화에 따른 분포 변화:

○ 입력-타겟 관계 유지: 장애 유형별 판정기준이 장애등급 판정에 미치는 영향은 동일

○ 사전확률 이동(가상 예시): 전체 등록 장애인 비율이 3%에서 5%로 증가 → 1급 판정 2023년에도 동일 기준이나, 전체 등록 비율만 증가

사전확률 드리프트는 모델이 학습된 분포와 다른 분포에서 작동하기 때문에 모델 정확도가 저하될 수 있다. 만약 고객 이탈을 예측하는 모델이라면, 전반적인 고객 행동이 변하여 고객 이탈 비율이 달라지면 사전확률 드리프트를 경험할 수 있다. 같은 예로, 신용 평가에서 대출을 신청하는 사람의 비율이 변하면 모델의 성능에 영향을 미칠 수 있다. 본질적으로 사전확률 드리프트는 데이터 내의 근본적인 관계가 일관되게 유지되더라도 모델링되는 문제의 전반적인 환경 변화를 강조한다. 이는 모델의 예측이 대상 변수의 새로운 분포와 일치하지 않을 수 있으므로 모델 성능이 저하되는 결과를 초래할 수 있다.

4. 라벨 드리프트(Label Drift)

라벨 이동(Label Shift) 또는 타겟 이동(Target Shift)은 라벨(분류 기준)의 정의나 해석이 시간에 따라 변화하는 것을 말한다. 이는 목표 변수 정의나 분류체계의 변화 등과 관련이 있다(Kull & Flach, 2014). 앞서 소개한 사전확률 드리프트는 목표 변수 선정 기준은 그대로지만 대상 집단의 비율이 변화하는 경우인 것에 비해, 라벨 드리프트는 분류/진단 기준 자체가 변화한 것을 말한다. 라벨 드리프트와 앞선 사전확률 드리프트

는 모두 시간 경과에 따른 머신 러닝 모델의 목표 변수 변화를 설명하지만, 그 초점은 서로 다르다. 라벨 드리프트는 모델의 예측 라벨의 output, 즉 결과적 변화를 나타내는 반면, 사전확률 드리프트는 구체적으로 대상 변수 분포의 ground truth(실제값) 변화를 나타낸다. 간단히 말해서, 라벨 드리프트는 모델이 예측하는 내용(가령, 목표 소비자에 대한 정의의 변화)에 대한 것이고, 사전확률 드리프트는 실제 현상이 하는 일(가령, 목표 소비자 행동 변화로 분포의 변화를 초래)에 대한 것이다. 라벨 드리프트는 모델의 예측이 더 이상 실제 라벨을 정확하게 반영하지 않으므로 모델 성능이 저하될 수 있는 결과를 초래할 수 있다. 라벨 드리프트의 예를 들자면, 보건복지분야에서는 질병분류체계 개정, 빈곤선 정의 변경, 장애등급제 폐지와 같은 제도 변화를 생각할 수 있다. 이해를 돕기 위해 가상의 예를 들면 다음과 같은 경우로 생각해 볼 수 있다.

□ 질병 진단 기준 변경

○ 과거: 당뇨병 진단 기준 공복혈당 140mg/dL 이상

○ 현재: 진단 기준 강화로 126mg/dL 이상으로 변경 → 같은 수치도 시기에 따라 다른 진단

□ 복지 수급 자격 기준 변경

○ 과거: 중위소득 40% 이하를 취약계층으로 분류

○ 현재: 기준 완화로 중위소득 50% 이하로 확대 → 동일한 소득 수준이어도 시기별로 다른 분류

5. 시스템 전반에 관한 기타 드리프트 유형들

가. 데이터 파이프라인 드리프트(Data Pipeline Drift)

데이터 파이프라인 드리프트는 데이터 수집·처리 방식의 변화로 파이프라인을 통과하는 데이터의 특성이 시간이 지남에 따라 변화하여 데이터 품질, 모델 성능 또는 다운스트림 프로세스에 문제를 일으킬 수 있는 현상을 말한다(Rabanser et al., 2019). 이러한 변화는 데이터의 구조, 형식 또는 통계적 속성에 영향을 미칠 수 있다. 데이터 엔지니어링 및 머신러닝 분야에서 흔히 발생하는 문제로, 데이터 무결성과 시스템 안정성을 유지하기 위해 지속적인 모니터링과 조정이 필요한 경우이다. 간단한 예를 들면 다음과 같은 경우들이 해당된다.

□ 전자의무기록 시스템 변경

○ 새로운 EMR 시스템 도입은 데이터의 입력 구조(자유로운 텍스트 형식에서 표준화된 코드 분류), 필드 구성, 결측치 처리 기준 변화를 가져오며, 통계적 특성 및 분포가 이전과 달라짐

□ 데이터 수집 프로토콜 개정

○ 새로운 프로토콜에 따라 데이터 수집 주기가 6개월에서 2개월로 단축되거나, 측정 방식이 1일 1회에서 5분 간격의 연속기록으로 달라짐

나. 표본선택 드리프트(Sample Selection Drift)

데이터 수집 대상이나 방식의 변화로 인한 드리프트를 말한다(Moreno-Torres et al., 2012). 간단한 예시로는 다음과 같은 경우를 생각할 수 있다.

□ 복지사각지대 발굴 방식 변화

○ 초기 복지사각지대 발굴시스템에서는 1종의 예측모형이 활용되었으나, 현재(2025년)에는 24종의 예측모형이 운영되고 있음

□ 국가건강검진 대상자 확대

○ 기존에는 만 54세·66세 여성이 골다공증 검진을 2회 받았으나, 2025년부터 만 60세 여성이 추가되어 총 3회(54·60·66세)에 골다공증 검진을 받을 수 있음

□ 취약계층 발굴방법의 다양화

○ 취약계층이 직접 지자체에 방문하여 복지급여를 신청하였으나, 2015년부터 운영되고 있는 복지사각지대 발굴 시스템으로 단전, 단수, 체납 등의 위기정보를 연계하여 데이터로 취약계층을 발굴하는 체계도 만들어짐. 이는 과거의 '신청주의' 한계를 극복하기 위해 보편적 발굴·지원으로 전환되었다는 데 의미가 있음

다. 도메인 드리프트(Domain Drift)

전반적 정책환경이나 사회적 맥락의 변화로 인한 드리프트로 소개되었다(Quionero-Candela et al., 2009). 간단한 예시로는 다음과 같은 경우들이 있다.

□ 복지정책 패러다임 변화

○ 과거에는 신청주의에 저소득층 지원이 중심이었다고 하면, 현재는 선제적 발굴, 데이터 기반 맞춤형 서비스로 전환되고 있고 행정 시스템의 활용도도 다양해지고 있음

□ 보건의료체계 개편

○ 과거에는 공급자 중심의 치료가 중요했다면, 이제는 소비자 중심의 의사결정 구조가 중요해지며 디지털 헬스케어, 원격진료, AI의 활용 등으로 변화하고 있음

□ 사회경제적 환경 변화

○ 과거의 기후 정책은 주로 배출 감축과 규제에 초점이 맞추어져 있었는데, 최근에는 에너지 신산업 육성, RE100(Renewable Electricity 100%) 등에 대한 관심도가 높아짐

라. 시간적 드리프트(Temporal Drift)

시간 흐름에 따른 점진적 변화로 Zliobaite(2010)에 의해 소개되었다. 입력과 목표 변수 등과 이에 대한 영향들을 명확히 구분하거나 측정하기

어려운 경우로 간단히 다음과 같은 예를 들 수 있다.

□ 계절적 복지수요 변동

○ 여름에는 냉방비 지원, 폭염 피해 등의 복지수요가 높아지는 반면에, 겨울에는 난방비 지원 등의 수요가 증가함

□ 인구고령화에 따른 변화

○ 인구 고령화에 따른 변화는 장기 요양 및 재활, 치매 등 노인성 질환 관련 서비스 이용 증가로 의료·복지 서비스 이용 패턴의 변화를 가져옴

마. 모델 드리프트(Model Drift)

학습된 예측 모형이 시간이 지남에 따라 실제 환경과 불일치하게 되면서 예측력이 저하되는 현상으로, 입력 변수 분포의 변화나 입력과 목표변수 간 관계의 변동에 의해 발생한다. 이는 특성 드리프트, 공변량 드리프트, 라벨 드리프트 등의 데이터 드리프트의 결과로 나타나는 모델 차원의 성능 저하이며, 관리하지 않으면 의사결정의 정확성에 심각한 영향을 줄 수 있으며, 다음과 같은 예를 들 수 있다.

□ 데이터 입력 특성의 변화

○ 가구구성, 소득 분포, 연령대 분포 등 학습 당시와 달라진 사회 인구학적 구조가 입력되면서 기존 모델이 새로운 패턴을 반영하지 못해 예측 정확도가 하락함

□ 외부 환경 및 제도 변화

○ 고혈압 기준 변경, 복지정책 자격 요건 변화 등 제도적 환경이 변하면 모델의 목표 변수 정의 자체가 달라지며, 이에 따라 기존 모델의 회귀계수·예측력 등이 더 이상 유효하지 않게 됨

6. 데이터 드리프트의 유형 분류 요약

본 절에서 설명한 유형들을 통계적 확률 분포 정의를 이용하여 요약 정리하면 다음과 같다.

□ 특성, 입력, 공변량 드리프트: 입력 변수 $P(X)$의 분포가 바뀌지만 조건부 분포 $P(Y|X)$는 유지됨

□ 개념 드리프트: 조건부 분포 $P(Y|X)$가 바뀌는 경우로, 모델의 의사결정 기준이 변화

□ 사전확률 드리프트: 정답 라벨 $P(Y)$의 분포 자체가 바뀌는 현상

□ 파이프라인 드리프트: 데이터 처리, 전처리, API/DB 연결 등 외부 요인으로 발생하는 드리프트

제2절 데이터 드리프트 탐지 방법

데이터 드리프트는 앞서 살펴본 유형들처럼 다양한 형태로 발생한다. 머신러닝 모델은 일반적으로 학습 데이터와 추론 데이터가 동일한 분포를 따른다고 가정한다. 하지만 실제 환경에서는 시간이 지남에 따라 데이터 분포가 변하는 현상, 즉 데이터 드리프트가 필연적으로 발생한다. 현실 세계의 AI/ML 시스템에서는 데이터 드리프트가 모델의 정확도, 공정성, 견고성 등 모델 성능 저하의 주요 원인으로 작용하기 때문에 이를 조기에 탐지하고 대응하는 것이 신뢰 가능한 AI 구축의 핵심이다. 본 절에서는 데이터 드리프트의 탐지 방법들 중 통계, 머신러닝, 딥러닝, 자동화 기반 탐지 기법 등을 소개하고, 각 유형별 적합한 탐지 방법을 검토하고자 한다.

1. 통계 기반 방법

이들은 두 시점(예: 학습 vs. 실전)의 분포 차이를 정량적으로 측정하는 방식이다.

가. Population Stability Index(PSI)

모집단 안정성 지수(Population Stability Index, PSI)는 시간 경과에 따른 또는 서로 다른 두 표본 간의 모집단 안정성을 평가하는 지표이다. 이는 변수 분포의 중요한 변화가 있었는지를 파악하는 데 도움을 주며, 머신러닝 모델 모니터링 및 기타 응용 프로그램에서 매우 중요하다. 쉽게 설명하면 PSI는 그룹의 '구성'이 얼마나 변했는지를 측정한다. 예를 들어,

2020년 데이터로 신용평가 모델을 구축했다면, PSI는 2024년의 모집단이 얼마나 다른지, 그리고 모델이 여전히 유효한지를 알려줄 수 있다. PSI는 두 데이터셋 간의 변수(소득, 나이 또는 모델의 예측값 등) 분포를 비교하는데, 다음과 같은 방식으로 작동한다.

□ 버킷팅: 변수값을 기준으로 데이터를 그룹 또는 '버킷'으로 나눔(예: 히스토그램 bin 단위)

□ 백분율 계산: 두 데이터셋에서 각 버킷 내 데이터 포인트의 백분율을 결정

□ 비교: 두 데이터셋의 각 버킷 간 백분율 차이를 계산

□ 합산: 차이를 합산하여 전체 PSI 점수를 얻음

□ 해석

PSI < 0.1: 변화가 작거나 없음, 좋은 안정성을 나타냄
0.1 < PSI < 0.25: 중간 정도의 변화, 추가 조사 필요
PSI > 0.25: 상당한 변화, 모델 재학습 또는 조정이 필요할 수 있음
즉, 값이 0.25 이상이면 큰 드리프트로 간주(Gulati, 2025)

□ 공식:

$$PSI = \Sigma\ (\text{Expected\%} - \text{Actual\%}) \times \ln(\text{Expected\%} / \text{Actual\%})$$

□ 수치형 변수의 분포 차이 측정에 효과적. Covariate Drift 감지에 적합.

나. 콜모고로프-스미르노프(Kolmogorov-Smirnov, KS) 검정

콜모고로프-스미르노프(Kolmogorov-Smirnov, KS) 검정은 두 확률 분포를 비교하는 데 사용되는 비모수 통계 검정 방법이다(Massey, 1951). 표본이 특정 분포에서 왔는지 또는 두 표본이 동일한 분포에서 왔는지 평가하는 데 사용될 수 있다. 이 검정은 비교되는 분포들의 누적분포함수(CDFs) 사이의 최대 수직 거리를 정량화한다.

- 목적: 두 데이터셋이 동일한 기저 분포에서 왔는지 또는 단일 데이터셋이 알려진 분포에서 크게 벗어나는지를 판단하는 데 도움을 줌
- 비모수적 특성: 데이터에 대해 특정 분포를 가정하지 않아, 다양한 데이터셋에 활용할 수 있음
- 작동 방식: 표본의 경험적 분포 함수(EDF)를 기준 분포의 CDF(또는 다른 표본의 EDF)와 비교함(EDF는 주어진 값보다 작거나 같은 데이터 포인트의 비율을 나타내는 계단 함수)
- KS 통계량: 검정 통계량은 두 CDF 사이의 최대 수직 거리(절대 차이)
- 해석: 더 큰 KS 통계량은 분포 간의 더 큰 차이를 나타내며, 귀무가설(분포가 동일하다는 가설)의 기각으로 이어질 수 있음
- 단일표본 vs. 이표본: 단일표본 KS 검정은 표본을 알려진 분포와 비교하고, 이표본 KS 검정은 두 개의 독립적인 표본을 비교

□ 장점: 분포에 구애받지 않고, 이해하기 상대적으로 쉬우며, 데이터 분포에 대한 가정이 필요하지 않음

□ 단점: 특히 분포의 꼬리(극단적 값들) 부분에서 차이를 감지하는 데 있어 다른 검정들보다 검정력이 낮을 수 있음

□ p-value 기준으로 분포 변화 여부를 판단

다. Kullback-Leibler(KL) Divergence(발산) 및 Jensen-Shannon(JS) 발산

KL(Kullback & Leibler, 1951), JS 발산(Lindgren, 1991)은 두 확률 분포 간의 차이를 측정하는 방법이다. 주요 차이점은 KL 발산은 비대칭적이고 무한대 값이 나올 수 있는 반면, JS 발산은 대칭적이며 0과 1 사이의 값으로 제한된다는 점이다.

□ KL: $D(P \parallel Q)$로 표기되며, 분포 P와 Q 사이의 차이를 측정, $D(P \parallel Q) \neq D(Q \parallel P)$로 방향에 따라 다른 값을 가짐(비대칭성). Q의 확률이 0이고 P의 확률이 0이 아닌 경우 무한대 값 발생(실제 응용에서 문제가 될 수 있음). P 분포의 샘플을 Q에 최적화된 코드로 인코딩할 때 필요한 추가 정보량을 측정(즉, 정보 이득을 측정함으로써 정보 이론에서 중요한 역할을 함)

□ JS: $JSD(P, Q)$로 표기. 두 분포의 혼합에 대한 KL 발산의 평균으로 계산. $JSD(P, Q) = JSD(Q, P)$로 대칭적. 0에서 1 사이의 값만 가진다는 범위 제한이 있어(밑이 2인 로그 사용 시) 해석과 비교가 용

이함. 확률이 0인 경우도 안정적으로 처리 가능하고(영 확률 처리), 실제 응용에서 더 실용적. JSD의 제곱근은 거리 측정 단위로 사용 가능

□ 실제 응용 분야: 모델 비교(서로 다른 머신러닝 모델의 예측 결과 비교), 모델 성능 평가에 활용, 데이터 드리프트 감지(시간에 따른 데이터 분포 변화 감지), JS 발산이 특히 유용함

□ 특성 선택: 가장 정보가 풍부하고 구별력 있는 특성을 식별하는데 유용한 방법으로 차원 축소에 활용

□ 공식: KL 발산: $D(P \parallel Q) = \sum P(X) \log\left(\frac{P(X)}{Q(X)}\right)$,

JS 발산: $JSD(P \parallel Q) = \frac{1}{2} D(P \parallel M) + \frac{1}{2} D(Q \parallel M)$,

여기서 $M = \frac{P+Q}{2}$

□ 활용 영역: 자연어 처리(문서 유사도 측정), 이미지 처리(GAN 학습에서의 손실 함수), 추천 시스템(사용자 선호도 변화 감지)

라. Chi square 검정/t 검정/Mann Whitney 검정

카이제곱 등의 검정 방법들은 범주형/연속형 변수 간 평균 또는 빈도 차이를 가설 검정으로 판별하는 방식들이다(Neely et al., 2003). 이들을 비교 요약하면 〈표 2-3〉과 같다.

□ 카이제곱 검정(Chi-square Test) 공식: $\chi^2 = \sum \frac{(O-E)^2}{E}$,

여기서 O는 관찰빈도, E는 기대빈도를 나타낸다. 범주형 데이터 분석의 비모수적 방법에 해당하며, 독립성 가정과 각 셀의 기대빈도 5 이상이 필요하다.

□ 활용 예시: 연령대별 만성질환 유병률 변화 감지, 소득계층별 의료서비스 이용 패턴 변화 분석, 지역별 복지서비스 수요 변화 탐지 등

□ t-검정(T-test) 공식: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, 여기서 x̄는 표본평균, s는 표준편차, n은 표본크기를 나타낸다. 연속형 데이터 분석에 유용하고, 정규성 가정과 등분산성 가정이 필요하다.

□ 보건복지 데이터 드리프트 활용 예시: 건강검진 수치의 시간적 변화 분석, 의료비 지출액의 연도별 차이 검정, 복지프로그램 효과성 평가 등

□ Mann-Whitney U 검정 공식: $U = n_1 n_2 + [\frac{n_1(n_1+1)}{2}] - R_1$,

여기서 n은 각 그룹의 표본크기, R은 순위합을 나타낸다. 비모수적 방법으로 정규성 가정이 불필요하고, 순서형/연속형 데이터에 적용 가능하다.

□ 보건복지 데이터 드리프트 활용 예시: 환자 만족도 점수 변화 감지, 의료서비스 대기시간 변화 분석, 복지서비스 접근성 점수 변화 탐지

〈표 2-3〉 범주형/연속형 변수 간 평균 또는 빈도 차이 가설 검정 방식

특성	카이제곱 검정	t-검정	Mann-Whitney U 검정
데이터 유형	범주형	연속형(정규분포)	순서형/연속형
기본가정	독립성	정규성, 등분산성	독립성
활용상황	범주 간 관계분석	평균 비교(정규분포)	중위수 비교(비정규분포)
장점	범주형 데이터 분석에 적합	정규분포 데이터에서 강력함	분포 가정이 불필요
단점	연속형 데이터 분석 불가	정규성 가정 필요	t-검정보다 검정력이 약함
표본크기	>5(각 셀)	>30(일반적)	제한 없음
귀무가설	변수 간 독립	평균이 동일	분포가 동일
결과해석	연관성 존재 여부	평균 차이	중위수/순위 차이

자료: 저자 작성.

마. Information Value(IV)

정보가치(Information Value, IV)는 독립변수가 목표 변수(주로 이진 종속변수, 예: 0 = 정상, 1 = 부도)를 설명하거나 구분하는 능력을 평가하는 지표이다. 신용평가, 리스크 모델링 등에서 변수 선택 단계에 많이 활용되며, 변수의 예측력을 정량적으로 보여준다. 즉, IV는 특정 변수가 좋은 집단과 나쁜 집단을 얼마나 잘 구분하는가를 수치화한 값이다.

- □ 그룹화(Binning): 연속형 변수는 여러 개의 구간(bin)으로 나누고, 범주형 변수는 각 범주별로 분리

- □ 비율 계산: 각 구간에서 “좋은 집단(Good)”과 “나쁜 집단(Bad)”의 비율을 계산

- □ WOE 변환: 각 구간의 비율 차이를 로그비(log odds) 형태의

WOE(Weight of Evidence)로 변환

□ 합산: 모든 구간의 $WOE \times (Good\% - Bad\%)$값을 합산하여 IV 산출

□ 해석:

- IV < 0.02: 거의 설명력이 없음(Not Predictive)
- 0.02 ≤ IV < 0.1: 약한 설명력(Weak Predictive Power)
- 0.1 ≤ IV < 0.3: 중간 정도 설명력(Medium Predictive Power)
- 0.3 ≤ IV < 0.5: 강한 설명력(Strong Predictive Power)
- IV ≥ 0.5: 과적합(overfitting) 가능성, 모델에서 제외 권장

□ 공식: $IV = \sum (Good\% - Bad\%) \times ln(\frac{Good\%}{Bad\%})$

□ 장점: 변수별 예측력을 개별적으로 평가 가능, 변수 선택 단계에서 유용. 라벨과 입력 간 관계 변화(Concept Drift 감지)에 적합

2. 시퀀셜 분석(Online Change Detection), 시계열 기반 방법

시간에 따라 데이터의 특성이 변화하는 것을 다루는 것은 데이터 마이닝과 기계학습의 핵심 문제 중 하나이다. 이러한 데이터를 마이닝하거나 학습하기 위해서는 최소한 다음 세 가지 작업에 대한 전략이 필요하다:

1. 변화가 발생하는 시점 감지
2. 어떤 예제를 유지하고 어떤 것을 잊어버릴지 결정(또는 더 일반적으로, 충분한 통계를 업데이트하며 유지)

3. 중요한 변화가 감지되었을 때 현재 모델을 수정

여기에서는 시간 스트림에서 실시간으로 드리프트 탐지하는 방식들을 소개하고자 한다.

가. Page-Hinkley(CUSUM) 테스트

Page-Hinkley 테스트(Page, 1954)는 CUSUM 테스트(Hinkley, 1971)의 변형으로, 데이터 스트림의 평균 변화를 감지하는 방법이다(Basseville & Nikiforov, 1993). 관측값과 이동 평균 간의 차이의 누적합을 모니터링하여 작동하고, 이 누적합이 미리 정의된 임계값을 초과할 때 변화가 감지된다(Mouss et al., 2004).

누적합(cumulative sum) 계산 공식

$S_t = S_{t-1} + (X_t - \mu_t - \delta)$, 여기서 S_t는 t시점의 누적합, X_t는 t시점의 관측값, μ_t는 t시점까지의 데이터 스트림의 이동 평균, δ는 실제 변화가 없을 때 변화를 방지하는 드리프트 항, 즉, 탐지하고자 하는 최소 변화량을 나타낸다.

□ 최소값(running minimum) 갱신: $mT = \min(S_i),\ i = 1, 2, \ldots, T$

PH 통계량 계산: $PHT = S_T - mT$

누적합 S_T가 미리 정의된 임계값 λ를 초과할 때 변화가 감지된다. 임계값 λ는 허용 가능한 오경보율에 따라 결정한다(즉, $PHT > \lambda$ 이면 변화 감지).

□ 공식

상향 CUSUM: $C^{+} = \max[0, X_i - (\mu_0 + K) + C_{i-1}^{+}]$

하향 CUSUM: $C^{-} = \max[0, (\mu_0 - K) - X_i + C_{i-1}^{-}]$

여기서, K는 참조값(일반적으로 $\frac{\delta}{20}$), μ_0는 목표 평균,

h는 결정 간격($C^{+} > h$ 또는 $C^{-} > h$이면 변화 감지)

이동 평균 μ_t는 점진적으로 업데이트되며, 최근 데이터 포인트에 더 많은 가중치를 주기 위해 망각 인자 α(1에 가까움)를 일반적으로 사용한다.

$\mu_t = \alpha\mu_{t-1} + (1-\alpha)X_t$

특성

테스트의 효과는 임계값 λ, 드리프트 항 δ, 망각 인자 α와 같은 매개변수의 적절한 값 설정에 크게 의존한다. 이러한 값들은 데이터의 특성과 원하는 민감도에 따라 조정되어야 한다. 높은 임계값은 오경보 가능성을 줄이지만 실제 변화 감지가 지연될 수 있다. 델타 매개변수는 작은 변화에 대한 민감도를 제어한다. Page-Hinkley 테스트는 누적합의 최소값과 최대값을 고려하여 평균의 증가와 감소 모두를 감지하도록 구성할 수 있기 때문에 양방향 테스트이기도 하다.

요약

Page-Hinkley 테스트는 관측값과 이동 평균 간의 누적 차이를 지속적으로 모니터링하여 데이터 스트림의 평균 변화를 감지하는 유용한 도구이다. 핵심 공식은 관측값, 이동 평균, 드리프트 항을 기반으로 누적합을 업데이트하는 것을 포함하며, 감지 규칙은 미리 정의된 임계값을 초과하는 것을 기반으로 한다.

활용분야

누적 오차 합이 임계값을 넘으면 경고 → 적응형 드리프트 탐지에 유용하기 때문에, 품질 관리, 시계열 분석, 데이터 스트림 모니터링, 이상 탐지, 공정 제어 등에 사용된다.

나. EWMA 차트(Exponential Weighted Moving Average)

지수 가중 이동 평균(Roberts, 1959)을 이용한 차트(Lucas & Saccucci, 1990) 방법은 시계열 데이터에서 점진적 변화를 감지하는 통계적 관리 도구로, 최근 데이터에 더 큰 가중치를 부여하고 과거 데이터는 지수적으로 감소하는 가중치를 적용하는 방법이다.

기본 통계량

$Z_t = \lambda X_t + (1-\lambda) Z_{t-1}$, 여기서 Z_t는 t시점의 EWMA 값, X_t는 t시점의 관측값, λ는 가중치 상수($0 < \lambda \leq 1$), Z_0는 초기값(보통 목표값 μ_0 사용)을 나타낸다.

관리 한계선:

$$UCL = \mu_0 + L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2t}]},$$

$$LCL = \mu_0 - L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2t}]}$$

여기서 UCL은 상한 관리선, LCL은 하한 관리선, μ_0는 공정 목표값, σ는 표준편차, L은 관리한계 승수(보통 3 사용), t는 시점을 나타낸다.

드리프트 감지 방법

기본 규칙은 Z_t가 UCL이나 LCL을 벗어나면 드리프트가 발생하는 것

으로 볼 수 있고, 보조 규칙으로는 연속된 점들의 패턴을 관찰해야 한다. 이때 매개변수로 가중치 λ를 선택하는데, 큰 값(0.2~0.4)은 급격한 변화 감지, 작은 값(0.05~0.2)은 점진적 변화를 감지하는데 사용될 수 있다. 관리한계 승수 L은 보통 L=3일 때 99.73% 신뢰구간, L=2일 때 95.45% 신뢰구간에 해당한다(Montgomery, 2009; Hunter, 1986).

[그림 2-3] EWMA를 이용한 드리프트 감지 Python 코드 예시

```python
def ewma_monitoring(data, lambda_=0.1, L=3):
    z_t = data[0]  # 초기값
    mu_0 = np.mean(data)  # 목표값
    sigma = np.std(data)  # 표준편차

    for t, x_t in enumerate(data[1:], 1):
        # EWMA 계산
        z_t = lambda_ * x_t + (1-lambda_) * z_t

        # 관리한계선 계산
        ucl = mu_0 + L * sigma * np.sqrt((lambda_/(2-lambda_)) * (1-(1-lambda_)**(2*t)))
        lcl = mu_0 - L * sigma * np.sqrt((lambda_/(2-lambda_)) * (1-(1-lambda_)**(2*t)))

        # 드리프트 감지
        if z_t > ucl or z_t < lcl:
            print(f"드리프트 감지: 시점 {t}")
```

자료: 저자 작성.

장점은 점진적 변화 감지에 효과적이고, 구현이 간단하며, 실시간 모니터링이 가능하다는 점이다. 단점은 초기값 설정에 민감하고, 급격한 변화 감지에는 상대적으로 덜 효과적이고, λ값 선정이 중요하다는 점이다(Crowder, S. V., 1987).

다. DDM(Drift Detection Method)

오류율과 표준편차를 모니터링하여 드리프트를 감지하는 방법이다(Gama et al., 2004, 2014). 구체적으로 드리프트 감지 방법(DDM)은 관찰된 오류율과 임계값의 예상 오류율의 차이를 기반으로 드리프트가

발생했는지 여부를 판별하기 위해 통계적 테스트를 한다. DDM 알고리즘은 데이터가 정상 프로세스에 의해 생성되고 시간 경과에 따른 오류율을 모니터링하여 데이터 분포의 변화를 감지할 수 있다고 가정한다. 이는 분류기의 평균 오류율과 분산을 계산하고 새 데이터 포인트가 도착할 때마다 이를 업데이트하는 것이다. 오류율이 Hoeffding 경계를 사용하여 계산된 특정 임계값을 초과하면 알고리즘은 드리프트 신호를 보낸다. 이 알고리즘은 적응형이므로 더 많은 데이터 포인트가 도착함에 따라 임계값을 조정하여 오경보를 최소화하면서 감지 정확도를 향상시킬 수 있다. DDM은 계산적으로 효율적이며 특히 갑작스러운 드리프트를 감지하는 데 실제로 좋은 성능을 보이는 것으로 나타났다(Sakurai et al., 2023).

□ 오류율(p_i) 계산: $p_i = \frac{\text{오류 개수}}{\text{전체 인스턴스 수}}$(여기서 i는 현재 시점)

□ 표준편차(s_i) 계산: $s_i = \sqrt{\frac{p_i(1-p_i)}{i}}$

□ 경계값: $\text{경고수준} = p_{\min} + 2s_{\min}$드리프트수준 $= p_{\min} + 3\, s_{\min}$

여기서, p_i는 i번째 시점의 오류율, s_i는 i번째 시점의 표준편차, $p_{\min}$는 지금까지 관찰된 최소 오류율, $s_{\min}$는 최소 오류율 시점의 표준편차를 나타낸다.

DDM는 구현이 단순하고, 계산 효율성이 높으며, 급격한 변화 감지에 효과적이라는 장점이 있고, 단점으로는 점진적 변화 감지에 둔감하고, 초기 학습 단계에서 오경보 가능성이 있다는 점이다.

라. EDDM(Early DDM)

조기 드리프트 검출법(EDDM)은 DDM 방법을 개선한 수정된 기법이다(Baena-García et al., 2006; Bifet & Gavalda, 2009). 이 기법은 데이터 분포에 상당한 변화가 발생하기 전에도 가능한 한 빨리 드리프트를 검출하도록 설계된다. EDDM은 급격한 드리프트와 점진적 드리프트를 모두 검출하는 데 효과적인 것으로 나타났다. 이 알고리즘은 각 윈도우에서 기준 분포와 관측 분포 사이의 최소 거리를 계산하고 새로운 데이터 포인트가 도착할 때마다 이를 업데이트한다. 최소 거리가 특정 임계값을 초과하면 알고리즘은 드리프트 신호를 보낸다. 임계값은 데이터 분포의 작은 변화에 민감한 스튜던트 t-검정 및 카이제곱 검정과 같은 통계적 검정을 사용하여 계산된다. EDDM은 DDM과 다른 통계적 검정을 사용하며 급격한 드리프트와 점진적 드리프트를 모두 검출하는 데 효과적인 것으로 나타났다(Sakurai et al., 2023).

□ 거리($p_i^{'}$) 계산: $p_i^{'}$=현재 오류와 이전 오류 사이의 예제수의 평균

□ 표준편차($s_i^{'}$) 계산: $s_i^{'}$=거리의 표준편차

□ 비율 계산:

$$\text{비율} = \frac{(p_i^{'} + 2s_i^{'})}{(p_{\max}^{'} + 2s_{\max}^{'})}$$

여기서, $p_i^{'}$는 현재까지의 거리 평균, $s_i^{'}$는 현재까지의 거리 표준편차, $p_{\max}^{'}$는 관찰된 최대 거리 평균, $s_{\max}^{'}$는 최대 거리 평균 시점의 표준편차를 나타낸다.

EDDM은 점진적 변화 감지에 효과적이고, 더 빠른 드리프트 감지가 가능하다는 장점이 있지만, 계산 복잡도가 더 높고, 노이즈에 더 민감하다는 단점이 있다.

마. HDDM_A(Adaptive)/HDDM_W(Window-based) (Hoeffding Drift Detection Method)

Hoeffding의 불평등을 활용하여 데이터 스트림에서 컨셉 드리프트를 감지하는 두 가지 방법이다. Hoeffding Drift Detection Method (HDDM)에 따라 생성된 드리프트 탐지 알고리즘(Frías-Blanco et al., 2015)은 Hoeffding 부등식(Hoeffding, 1963)을 사용하여 원래 DDM을 수정한다. HDDMA는 이러한 알고리즘 중 하나로, 관측된 오류율의 이동 평균을 사용하여 분류기의 실제 오류율을 추정한다. 이 알고리즘은 고정된 크기의 슬라이딩 윈도우에 대한 평균 오류율을 계산하고 이를 Hoeffding의 경계와 비교한다. 관측된 오류율이 경계를 초과하면 알고리즘은 드리프트 신호를 보낸다. HDDMW 알고리즘은 HDDMA 알고리즘과 유사하지만 관측된 오류율의 이동 가중 평균을 사용한다. 이 알고리즘은 각 데이터 포인트의 최신성을 기준으로 가중치를 할당하며, 최근 데이터 포인트일수록 가중치가 더 높다. 가중 평균은 분류기의 실제 오류율을 추정하는 데 사용되며, 이를 Hoeffding의 경계와 비교한다. HDDMW는 급격한 드리프트와 점진적인 드리프트를 모두 감지하는 데 효과적이며, 일부 시나리오에서는 DDM 및 HDDMA보다 성능이 뛰어나다.

□ Hoeffding Inequality(불평등):

$$P(|\hat{\mu}_t - \mu| \geq \epsilon) \leq 2 \cdot exp(\frac{-2n\epsilon^2}{(b-a)^2})$$

○ $\hat{\mu}_t$: 관측된 평균값, μ: 실제 평균값, n: 표본 수,
[a, b]: 관측값의 범위, ϵ: 허용 가능한 오차

□ HDDM_A(급격한 변화 감지용):

Hoeffding 경계: $\varepsilon = \sqrt{\frac{\ln(\frac{1}{\delta})}{2n}}$,

신뢰구간: $[\mu - \varepsilon, \mu + \varepsilon]$, 이동평균: $\mu w = \frac{1}{w}\sum x_i,\ i = t - w + 1$

○ 드리프트 조건: $\hat{\mu}_t - \mu_{\min} > \epsilon$

○ 허용오차 ϵ: $\epsilon = \sqrt{\frac{1}{2n} \cdot \ln(\frac{1}{\delta})}$

- $\hat{\mu}_t$: 현재까지의 평균 오류율, $\mu_{\min}$: 관측된 최소 오류율,
n: 샘플 수, δ: 신뢰 수준(일반적으로 0.001~0.01)

□ HDDM_W(점진적 변화 감지용):

가중이동평균: $\mu w = \frac{\sum w_i x_i}{\sum w_i}$,

가중치 계산: $w_i = decay(t - i)$, 신뢰구간: $[\mu w - \varepsilon, \mu w + \varepsilon]$

○ 드리프트 조건: $\hat{\mu}_1 - \hat{\mu}_2 > \epsilon$

○ 허용오차 ϵ: $\epsilon = \sqrt{\frac{1}{2} \cdot \ln(\frac{1}{\delta}) \cdot (\frac{1}{n_1} + \frac{1}{n_2})}$

- $\hat{\mu}_1$, $\hat{\mu}_2$: 각 윈도우 W_1, W_2의 평균,
n_1, n_2: 각 윈도우의 크기,

δ: 허용 오류 확률(예: 0.01)

주요 매개변수: ε는 Hoeffding 경계, δ는 신뢰수준 매개변수, n는 샘플 크기, μ는 전체 평균, μw는 윈도우 평균, w는 윈도우 크기, decay는 감쇠 계수(0 < decay < 1)를 나타낸다.

□ 드리프트 감지 규칙: 경고 수준 $|\mu w - \mu| > \beta\varepsilon$,
드리프트 수준 $|\mu w - \mu| > \alpha\varepsilon$(여기서 $\alpha > \beta$는 신뢰수준 매개변수)

HDDM_A는 급격한 변화 감지에 효과적이고, 계산 효율성이 높으며, 메모리 사용량이 적다는 장점이 있고, 단점으로는 점진적 변화 감지에 덜 효과적이고, 파라미터 설정에 민감하다는 점이 있다.

HDDM_W는 점진적 변화 감지에 효과적이고, 노이즈에 강건하며, 최근 데이터에 더 민감하게 반응한다. 단점으로는 계산 복잡도가 더 높고, 메모리 사용량이 많으며, 가중치 설정이 중요하다는 점이다.

□ 활용 분야: 온라인 학습 시스템, 데이터 스트림 모니터링, 이상 탐지, 적응형 학습 알고리즘, 실시간 의사결정 시스템

〔그림 2-4〕 HDDM를 이용한 드리프트 감지 Python 코드 예시

```python
def hddm_detection(data, delta=0.001, alpha=3, beta=2):
    n = len(data)
    epsilon = np.sqrt(np.log(1/delta)/(2*n))

    mu_total = np.mean(data)
    mu_window = np.mean(data[-window_size:])

    if abs(mu_window - mu_total) > alpha * epsilon:
        return "drift"
    elif abs(mu_window - mu_total) > beta * epsilon:
        return "warning"
    return "stable"
```

자료: 저자 작성.

바. ADWIN(Adaptive Windowing)

가변 크기의 슬라이딩 윈도우는 시간에 따라 변할 수 있는 데이터 시퀀스를 학습할 때 분포 변화와 컨셉 드리프트를 다루는 접근 방식이다. 이는 미리 고정된 크기 대신, 윈도우 자체의 데이터에서 관찰된 변화율에 따라 온라인으로 재계산되는 크기를 가진 슬라이딩 윈도우를 사용한다. 이는 사용자나 프로그래머가 변화의 시간 척도를 추측해야 하는 부담을 덜어준다. 거짓 양성과 거짓 음성 비율의 경계로서 엄격한 성능 보장을 제공한다(Bifet & Gavalda, 2007).

□ 윈도우 평균 분할기준: $|\mu W_0 - \mu W_1| > \varepsilon cut$, 새 데이터가 들어올 때마다, W를 두 개의 부분 W_0, W_1로 나누고, 이 두 부분의 평균이 통계적으로 유의미하게 다른지를 검정

○ 스트림 윈도우를 두 구간으로 나눴을 때:

- $\hat{\mu}_0$: 앞쪽 서브윈도우의 평균

- $\hat{\mu}_1$: 뒤쪽 서브윈도우의 평균

○ 드리프트 발생 조건: $\hat{\mu}_0 - \hat{\mu}_1 > \epsilon$

□ εcut 계산: $\varepsilon cut = \sqrt{(\frac{1}{m})(\frac{1}{2})\ln(\frac{4}{\delta})}$

□ 신뢰도 경계: $m = \frac{1}{(\frac{1}{n_0} + \frac{1}{n_1})}$

□ 여기서, μW_0, μW_1: 두 하위 윈도우의 평균,

m: 조화 평균, n_0, n_1: 각 하위 윈도우의 크기,

δ: 신뢰수준 매개변수, 허용 가능한 ε:

$\epsilon = \sqrt{\frac{1}{2} \cdot \ln(\frac{4 \cdot ln(n)}{\delta}) \cdot (\frac{1}{n_0} + \frac{1}{n_1})}$

- n_0, n_1: 각 서브윈도우의 샘플 수
- n=n_0+n_1: 전체 윈도우 크기
- δ: 신뢰 수준(예: 0.01)

□ 드리프트가 감지되면 윈도우 W에서 앞쪽 서브윈도우 W_0를 제거하고 새로운 데이터에 더 민감한 상태로 재설정

□ 장점: 자동 윈도우 크기 조정, 이론적 보장 제공, 파라미터 설정 최소화, 메모리 효율적 관리

□ 단점: 계산 복잡도가 높음, 초기 설정에 민감, 급격한 변화 감지에 지연 가능

〔그림 2-5〕 ADWIN을 이용한 드리프트 감지 Python 코드 예시

```python
class ADWIN:
    def __init__(self, delta=0.002):
        self.delta = delta
        self.window = []

    def update(self, value):
        self.window.append(value)
        self.check_drift()

    def check_drift(self):
        for i in range(len(self.window)):
            w0 = self.window[:i]
            w1 = self.window[i:]

            if self.detect_change(w0, w1):
                self.window = w1
                return True
        return False

    def detect_change(self, w0, w1):
        n0, n1 = len(w0), len(w1)
        if n0 == 0 or n1 == 0:
            return False

        m = 1.0/(1.0/n0 + 1.0/n1)
        epsilon = sqrt((1.0/m) * 0.5 * log(4.0/self.delta))

        return abs(np.mean(w0) - np.mean(w1)) >= epsilon
```

자료: 저자 작성.

주요 매개변수로는, δ(신뢰수준)로 작은 값은 더 엄격한 드리프트를 감지, 큰 값은 더 유연한 드리프트를 감지하고, 최소 윈도우 크기로 노이즈 필터링과 계산 효율성 조절을 할 수 있다.

□ 활용 분야: 온라인 학습, 데이터 스트림 분석, 시계열 예측, 이상 탐지, 적응형 알고리즘

□ 최적화 전략: 이진 트리 구조 사용, 요약 통계량 유지, 버킷 기반 근사, 병렬 처리 적용

□ 실제 적용 시 고려사항: 메모리 제약, 실시간 처리 요구사항, 노이즈 수준, 드리프트 유형, 계산 리소스

사. FHDDM(Fast Hoeffding Drift Detection Method) 시리즈 (FHDDMS: Stacked 버전, FHDDMS_add: 추가 개선 버전)

슬라이딩 윈도우 기반 Hoeffding 방법을 사용하여 데이터 스트림에서 컨셉트 드리프트를 고속으로 탐지하는 방법이다(Bayram, 2022).

FHDDM 기본 모형

ε: 오차 허용 범위, n: 샘플 크기, μ: 실제 평균, $\bar{X}$: 표본 평균일 때, Hoeffding 방법 $P(|\bar{X} - \mu| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$ 에서 드리프트 탐지 오차범위는 다음과 같다.

○ $|\hat{\mu}_W - \mu_{\min}| > \epsilon_H \qquad \epsilon_H = \sqrt{\frac{1}{2n} \ln\left(\frac{1}{\delta}\right)}$

□ FHDDMS(Stack 버전): 여러 개의 슬라이딩 윈도우를 스택 형태로 사용, 각기 다른 크기의 윈도우로 다양한 시간 스케일에서 드리프트 탐지

□ 드리프트 탐지: μ_w는 경고 윈도우의 평균이고, μ_r는 참조 윈도우의 평균일 때 $\text{drift} = \begin{cases} true & \text{if } \mu_w - \mu_r > \varepsilon \\ false & \text{otherwise} \end{cases}$

□ FHDDMS_add(추가 개선 버전): 적응형 임계값을 사용하여 성능 개선, 드리프트 탐지 민감도 조정 가능

□ 수정된 임계값은 δ: 신뢰도 레벨, n: 윈도우 크기일 때, $\text{threshold} = \sqrt{\frac{\ln(1/\delta)}{2n}}$

□ 특징: 계산 효율성(시간 복잡도로 빠른 처리 가능), 메모리 효율성(고정된 윈도우 크기로 일정한 메모리 사용), 실시간 처리(스트리밍 데이터에 적합), 적응성(다양한 유형의 드리프트 탐지 가능)

□ 장점: 빠른 드리프트 탐지, 낮은 지연 시간, 높은 정확도, 적은 메모리 사용량

아. ECDD(EWMA for Concept Drift Detection)

EWMA 차트로 성능(오류율 등) 모니터링 후 드리프트 발생 시 플래그 설정하는 방법이다.

□ 공식: z_t: 현재 시점의 EWMA 값, x_t: 현재 관측값, λ: 가중치 매개변수 ($0 < \lambda \leq 1$), z_{t-1}: 이전 시점의 EWMA 값일 때, EWMA

$$z_t = \lambda e_t + (1 - \lambda) z_{t-1}$$

□ 드리프트 기준: μ_0: 목표 평균, L: 관리 한계 너비 계수,
δ_z: EWMA의 표준편차일 때,

□ $|z_t - \mu_0| > L \cdot \sigma_0$

□ 특징: 점진적 변화 감지에 효과적(작은 변화도 누적되어 탐지 가능), 이동 평균의 특성(과거 데이터의 영향이 지수적으로 감소, 최근 데이터에 더 큰 가중치 부여), 파라미터 설정의 중요성(λ: 민감도 조절 - 작을수록 과거 데이터 영향 증가, L: 오탐지율과 탐지 지연시간 사이의 트레이드 오프)

□ 적용 분야: 실시간 성능 모니터링, 품질 관리, 이상치 탐지

□ 장점: 구현이 간단, 계산 효율성이 높음, 메모리 사용량이 적음

□ 한계점: 초기 파라미터 설정에 따른 성능 변동, 급격한 변화 탐지에는 상대적으로 취약

3. ML 기반(머신러닝 지도/준지도 학습 Supervised/Semi-Supervised Learning)

실제 라벨이 있을 때 사용하는 드리프트 탐지 방법이다(Hu et al., 2025).

가. CIDD-ADODNN(class imbalance with concept drift detection using Adadelta optimizer-based deep neural networks)

이 방법은 딥러닝을 ADWIN과 결합해 드리프트를 감지하는 방법이다. Adadelta로 최적화된 심층 신경망을 사용하여 불균형 데이터 스트림을 분류하고 ADWIN 알고리즘으로 드리프트를 감지한다(Priya & Uthra, 2023),

문제 정의는 다음과 같다. n개의 소스에서 수집된 입력 데이터 스트림 Soi는 Sol, So2, So3, ..., Son으로 표시된다. 소스 i는 k개의 스트림 Soik를 생성한다(즉, Soi1, Soi2, ..., Soik). 이러한 소스들의 샘플들은 완전한 스트리밍 데이터 USoi = So를 구성한다. 데이터 전처리 방법의 핵심 전제는 n개 소스로부터의 스트림 데이터 So에 대한 저장소 SR을 선언하는 것이다. 완전한 스트림 데이터에 대한 통계적 저장소 크기를 검사하는 데 두 가지 요소가 중요하다. 스트림 데이터의 불일치 정도는 각 소스에 대해 분산된 샘플 수의 차이를 보여준다. 최대 불일치 정도는 정확한 값을 추정하는 데 가능한 최소 신뢰 구간으로 이어진다. SR은 전체 샘플 크기, N은 전체 모집단, e는 신뢰 구간일 때:

저장소 크기: $SR = \frac{N}{1 + Ne^2}$,

분류기 출력: $\Theta(x_t) \in \{1, -1\}$, 라벨 할당: $y_t \in \{1, -1\}$

낮은 신뢰 구간의 경우, 데이터 샘플링 방법은 최대 인스턴스 수를 결정한다. 그렇지 않으면, 완전한 스트림 데이터를 보여주기 위해 최소한의 샘플 수가 필요한데, 샘플링 프로세스가 적용되면, 스트림 데이터 분류에서

두 클래스 문제가 유지된다. t 시간에 새로운 인스턴스 x_t를 받는 온라인 앙상블 분류기 Θ를 가정하고, 감지된 클래스 라벨은 y_t이다. 예측이 계산되면, 분류기는 x_t의 원하는 라벨을 y_t로 받는다. 따라서 예측된 라벨과 원하는 라벨은 {1, -1}을 할당한다. 앙상블 분류기 Θ의 결과는 네 개 TP, TN, FP, FN(True/False/Positive/Negative)의 클래스로 나눈다.

클래스 불균형 처리방법: Adaptive synthetic(ADASYN) 모델은 전처리된 데이터를 입력으로 받아 클래스 불균형을 처리하기 위해 실행된다. 이 모델은 학습 난이도 수준에 기반하여 서로 다른 소수 클래스 인스턴스에 대해 가중치 분포를 활용한다. 이 분포를 기반으로 소수 클래스에 대한 고유한 합성 인스턴스를 생성한다. 이는 합성 소수 클래스 오버샘플링 기법(Synthetic Minority Oversampling Technique, SMOTE), SMOTEBoost, DataBoostIM과 같은 합성 모델들의 인기로 인해 도입되었다. 이 모델은 불균형 데이터 세트로부터 학습을 수행한다.

□ 주요 특징

○ 클래스 불균형 처리를 위한 적응형 합성 데이터 생성, 학습 난이도에 기반한 가중치 분포 사용, 소수 클래스에 대한 지능적인 오버샘플링, 기존 SMOTE 계열 방법들의 장점을 활용하면서 더욱 적응적인 접근 제공

○ ADASYN 절차: 클래스 불균형 정도 계산(d = ms/ml(소수 클래스 샘플 수/다수 클래스 샘플 수)) 후 d가 임계값보다 작을 경우 다음 단계 수행:

- a) 생성할 합성 데이터 샘플 수 계산
- b) 소수 클래스의 각 샘플(xi)에 대해 K-최근접 이웃(kNN) 찾기, kNN 중 다수 클래스에 속하는 샘플 수 비율 ri 계산

- c) ri 정규화: 밀도 분포화
- d) 각 소수 클래스 샘플별 생성할 합성 데이터 수 계산
- e) 각 소수 클래스 샘플 xi에 대해 gi개의 합성 데이터 생성
- 1부터 gi까지 반복: kNN에서 무작위로 소수 클래스 샘플 선택 후 합성 데이터 인스턴스 생성

균형화된 클래스 자료에 ADWIN 적용(시퀀셜 분석(Online Change Detection), 시계열 기반 방법 내용 참고): ADWIN이 컨셉트 드리프트를 감지하면 모델이 업데이트되고 분류 작업 실행에서 분류 결과의 상당한 개선이 가능하다.

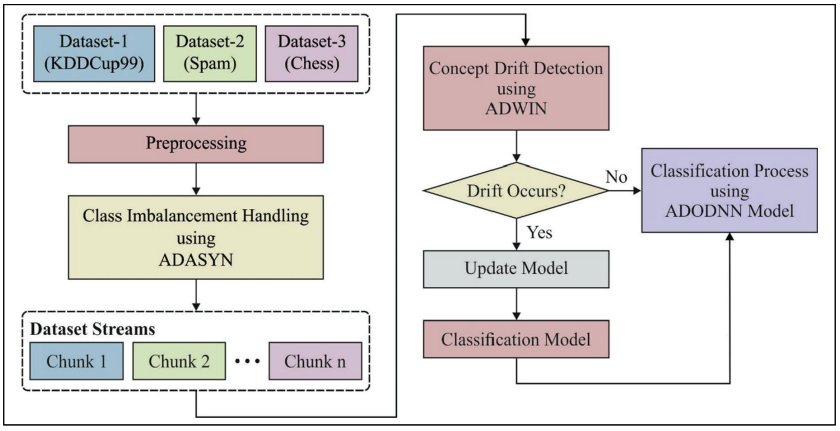
컨셉트 드리프트가 없는 경우, 모델 업데이트 없이 ADODNN(Adadelta optimizer-based DNN)으로 직접 분류를 수행한다. ADODNN은 실제 클래스 라벨을 결정하고 ADO(Adadelta optimizer) 적용으로(Zeiler, 2012) 분류 성능을 향상시킬수 있다.

DNN 기반 모델 구조: 스택형 오토인코더(stacked autoencoders, SAE)를 사용하여 컨셉트 드리프트를 분류하는 구조로 SAE와 소프트맥스 층을 통한 DNN 분류기이다. 주요 구성은 입력층(파라미터 입력), 두 개의 SAE 층, 두 개의 히든 층과 뉴런, 최종 히든 층에 소프트맥스 층 연결, 출력층(적용된 레코드의 클래스 라벨 확률 제공)이다.

ADO 적용: 딥러닝(DL) 기반 옵티마이저들은 기본적으로 사전 정의된 학습률을 가지고 있다. 하지만 실제 사례에서 DL 모델들은 비볼록 문제다. DNN 모델의 효과적인 학습률을 결정하기 위해 최대 분류 성능을 달성할 수 있는 방식으로 학습률을 계산하는 ADO가 적용된다. Adadelta는 Zeiler(2012)에 의해 개발되었고, 이 모델의 주요 목적은 분모에서 이전 제곱 그래디언트들의 누적으로 인한 학습률의 급격한 감소라는

Adagrad의 취약점을 회피하는 것이다. Adadelta는 제한된 시간 내에 처리된 현재 그래디언트를 사용하여 학습률을 측정한다. 또한 Adadelta는 이전 업데이트들을 고려하여 가속기를 적용하며, Adadelta 업데이트 규칙을 따른다(주요 특징: 사전 정의된 학습률 대신 적응적 학습률 사용, 비볼록 최적화 문제 해결, 그래디언트 이력 기반의 학습률 조정, 제한된 시간 윈도우 내의 그래디언트 고려, 이전 업데이트를 활용한 가속화 메커니즘).

[그림 2-6] Adadelta 최적화 딥러닝을 이용한 클래스 불균형 드리프트 탐색



자료: Priya & Uthra, 2023.

나. Uncertainty Drift Detection(UDD) 불확실성 드리프트 감지 (Baier et al., 2021)

딥러닝 모델의 예측 불확실도(MC Dropout, 몬테카를로 드롭아웃)(Gal & Ghahramani, 2016)를 ADWIN과 결합하여 라벨 없이도 감지하는 방법이다. 성공적인 드리프트 감지를 위해서는 실제 라벨이 전제 조건으로 필요하다. 많은 실제 응용 시나리오에서는 실제 라벨이 부족하고 이

를 획득하는 데 비용이 많이 든다. 따라서 UDD는 실제 라벨에 접근하지 않고도 드리프트를 감지할 수 있다. 이 접근 방식은 몬테카를로 드롭아웃과 결합된 심층 신경망이 제공하는 불확실성 추정치를 기반으로 한다. 시간에 따른 구조적 변화는 불확실성 추정치에 ADWIN 기법을 적용하여 감지되며, 감지된 드리프트는 예측 모델의 재학습을 트리거한다. 특히, 입력 데이터 기반 드리프트 감지와는 달리, 입력 데이터의 변화만을 감지하는 것이 아니라(이는 불필요한 재학습으로 이어질 수 있음) 현재 입력 데이터가 예측 모델의 특성에 미치는 영향을 고려한다. 회귀 및 분류 작업 모두에서 우수한 성능을 보여준다.

불확실성은 다음과 같다. ML에서 모델의 예측 확실성이 중요한데 종종 클래스 확률(예: 소프트맥스 층의 출력)이 모델의 신뢰도로 잘못 해석된다. 실제로 모델은 특정 클래스에 대한 높은 소프트맥스 출력에도 불구하고 예측에 불확실할 수 있다. 신경망은 일반적으로 보지 못한 데이터에 대한 외삽(extrapolating)에 취약하다. 따라서 모델에 특이한 데이터가 도입되면 소프트맥스 층의 출력이 오해의 소지가 있을 수 있으며, 이는 컨셉트 드리프트의 경우에 자주 발생한다. 불확실성은 크게 두 가지 유형으로 구분되는데, 우연적(aleatory) 불확실성(데이터 생성 과정의 무작위성으로 설명되는 데이터 불확실성)과 인식론적(epistemic) 불확실성(불충분한 학습 데이터로 인한 모델 불확실성)으로 나눌 수 있다. 분류 작업의 경우 불확실성은 엔트로피, 변동 비율 또는 상호 정보를 통해 정량화될 수 있다.

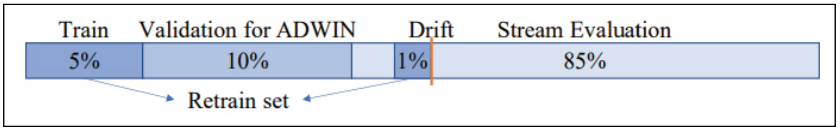
□ 몬테카를로 드롭아웃(MCD): 추론 시에 드롭아웃을 사용하여 예측에 무작위성을 도입, 여러 번의 순방향 패스를 통해 불확실성 추정치를 도출

□ 딥 앙상블(Deep Ensembles): 신경망의 최종 층을 개선하여 단일 예측이 아닌 분포적 예측 집합을 출력, 더 포괄적인 불확실성 추정을 가능하게 함

방법론을 요약하면 다음과 같다. 불확실성 드리프트 감지를 위해 신경망의 예측 불확실성을 측정하고 변화 감지를 위해 ADWIN 알고리즘을 사용한다. 회귀와 분류 작업 모두에서 개념 드리프트를 감지하고, 예측 모델과 관련된 변화만 감지한다. 불확실성 측정을 위해 몬테카를로 드롭아웃(MCD)을 사용하고, 회귀와 분류에 대한 서로 다른 방법, 즉, 분류에는 섀넌 엔트로피를, 회귀에는 경험적 분포의 분산을 사용한다. ADWIN 민감도 매개변수로 다양한 데이터 스트림에 적응한다.

□ 장점: 입력 데이터 기반 감지보다 더 효율적이며, 불필요한 재학습 감소, 회귀와 분류 작업 모두에 적용 가능, 스트림 환경에서 계산 효율성이 높음

〔그림 2-7〕 UDD 데이터 스트림 분할 방법



자료: Baier et al., 2021.

다. Meta-ADD/LSTMDD/WSCDD/CD-BTMSE 등

최신의 기타 방법들을 Hu et al.(2025)에 소개된 내용을 인용하여 아래에 간단히 설명하고자 한다.

1) Meta-ADD(Yu et al., 2022)

Active Drift Detection based on Meta learning(메다학습기반 능동 데이터 드리프트 탐지) 방법은 오프라인 사전 학습에서 드리프트 데이터를 활용하여 오류율과 관련된 메타 특성을 추출하고, 프로토타입 네트워크를 사용하여 다양한 컨셉트 드리프트 카테고리를 식별하는 방법이다. 사전 지정된 가설 검정 없이도 드리프트 유형을 감지할 수 있으나 사전 학습이 부족하고 콜드 스타트 문제를 겪을 수 있다. 갑작스러운, 점진적, 그리고 증분적 드리프트를 감지하는데 유용한 방법이다.

2) LSTMDD(Mehmood et al., 2024)

Long short-term memory data drift(장단기 메모리 데이터 드리프트)는 클라우드 컴퓨팅에서의 개념 드리프트 문제를 조사하고, 최적의 자원 활용을 가능하게 하여 드리프트의 효과적이고 조기 감지의 중요성을 강조한다. 클라우드 도메인에 맞춤화된 적절한 드리프트 감지기의 필요성을 합성 및 실제 클라우드 데이터셋을 이용하여 강조하였다. LSTM 드리프트 감지기(LSTMDD)는 수정된 버전의 장단기 메모리(LSTM)이며, 예측 오류를 주요 평가 지표로 사용한다. LSTMDD는 비가우시안 분포 클라우드 환경에서 이상 감지 성능을 향상시키도록 최적화된다. LSTMDD는 클라우드 도메인에서 점진적 및 갑작스러운 드리프트에 대해 다른 방법들보다 더 우수한 성능을 보여준다. 이는 LSTMDD와 같은 기계 학습 기술이 클라우드 컴퓨팅에서의 개념 드리프트 문제를 해결하는 유망한 접근 방식이 될 수 있으며, 이는 더 효율적인 자원 할당과 향상된 성능으로 이어질 수 있음을 시사한다.

3) WSCDD(Ma et al., 2024)

Weakly Supervised Conceptual Drift Detection method based on the Online Deep Neural Network(온라인 심층 신경망 기반의 약지도 개념 드리프트 감지 방법)은 Hedge 역전파(backpropagation) 알고리즘을 사용하여 신경망 모델을 학습시키며, 동시에 몬테카를로 방법을 사용하여 예측 불확실성을 계산한다. 이 불확실성은 개념 드리프트를 감지하기 위한 ADWIN 알고리즘의 입력으로 사용된다. 약지도 환경에서 라벨이 있는 데이터를 얻기 어렵다는 점을 고려할 때, WSCDD는 라벨이 없는 데이터에서도 개념 드리프트의 발생을 감지할 수 있다.

4) CD-BTMSE(Cai et al., 2024)

Concept drift detection model based on Bidirectional Temporal Convolutional Networks and Multi-Stacking Ensemble Learning(양방향 시간 합성곱 네트워크와 다중 스태킹 앙상블 학습을 기반으로 한 개념 드리프트 탐지 모델)은 양방향 시간 합성곱 네트워크(Bidirectional Temporal Convolutional Network, BiTCN)와 다중 스태킹 앙상블 학습 모델을 활용하여 드리프트 감지의 일반화 능력과 정확도를 향상시키면서 동시에 클래스 불균형 문제를 해결한다. BiTCN 모델을 통합함으로써, 전통적인 모델들이 일반적으로 성능이 떨어지는 시계열 데이터에서의 모델 성능을 향상시켜 기존 앙상블 방법들의 한계를 극복하는 것이 이 방법의 특징이다.

〈표 2-4〉 추가 지도학습 기반 방법론 분류

연번	드리프트 탐지 기법	유형	참고문헌
1	STAGGER	Statistical	(Schlimmer & Granger, 1986)
2	FLORA	Statistical	(Widmer, 1996)
3	CVFDT(Concept Adapting Very Fast Decision Trees)	Window-based	(Domingos & Hulten, 2000)
4	SEA(Streaming Ensemble Algorithm)	Ensemble-based	(Nick Street & Kim, 2001)
5	AWE(Accuracy Weighted Ensembles)	Ensemble-based	(Haixun Wang et al., 2003b)
6	DDM(Drift Detection Method)	Statistical	(Gama et al., 2004)
7	ACE(Adaptive Classifiers Ensemble) System	Statistical	(Nishida et al., 2005)
8	EDDM(Early Drift Detection Method)	Statistical	(Baena-Garcia et al., 2006)
9	STEPD(Statistical Test of Equal Proportions)	Statistical	(Nishida & Yamauchi, 2007)
10	ADWIN(Adaptive Windowing)	Window-based	(Bifet & Gavalda, 2007)
11	AUC(Accuracy Updated Ensembles)	Ensemble-based	(Brzezinski & Stefanowski, 2011)
12	DDM-OCI(DDM for Online Class Imbalance Learning)	Statistical	(S. Wang et al., 2013)
13	E-CVFDT(Efficient CVFDT)	Window-based	(G. Liu et al., 2013)
14	LFR(Linear Four Rates)	Statistical	(Heng Wang & Abraham, 2015)
15	FHDDM(Fast Hoeffding's Drift Detection Method)	Window-based	(Pesaranghader & Viktor, 2016)
16	SAND(semi-supervised Adaptive Novel Class Detection)	Ensemble-based	(Haque, Khan & Baron, 2016)
17	ECHO(Efficient Handling of Concept Drift and Evolution)	Ensemble-based	(Haque, Khan, Baron, et al., 2016)

연번	드리프트 탐지 기법	유형	참고문헌
18	RDDM(Reactive Drift Detection Method)	Statistical	(Barros et al., 2017)
19	HLFR(Hierarchical Linear Four Rates)	Statistical	(Yu & Abraham, 2017)
20	FPDD(Fisher Proportions Drift Detector)	Statistical	(Cabral & Barros, 2018)
21	FSDD(Fisher-based Statistical Drift Detector)	Statistical	(Cabral & Barros, 2018)
22	FTDD(Fisher Test Drift Detector)	Statistical	(Cabral & Barros, 2018)
23	MDDM(McDiarmid Drift Detection Method)	Statistical	(Pesaranghader et al., 2018)
24	DWM(Dynamic Weighted Majority)	Ensemble-based	(Jeremy Z. Kolter & Maloof, 2003)
25	Learn++	Ensemble-based	(Polikar et al., 2001)
26	Learn++. MT	Ensemble-based	(M. Muhlbaier et al., 2004)
27	Learn++. NC(New Class)	Ensemble-based	(M. D. Muhlbaier et al., 2009)
28	Learn++. NSE (Non-Stationary Environment)	Ensemble-based	(M. D. Muhlbaier & Polikar, 2007)
29	Learn++. NIE(Non-Stationary and Imbalance Environment)	Ensemble-based	(Ditzler & Polikar, 2010)
30	Learn++. CDS (Concept Drift with SMOTE)	Ensemble-based	(Ditzler & Polikar, 2013)

자료: Ali, U. & Mahmood, T., 2024.

4. 딥러닝/비지도 기반(Deep/Unsupervised Learning)

라벨이 없어도 자동으로 드리프트를 감지하는 방법들을 소개한다. 이는 라벨 없이 혹은 적은 라벨 데이터로도 데이터 분포 변화 감지가 가능하여 다양한 도메인에 적용 가능하고, 실시간 스트리밍 환경에 적합한 기법이다. 비지도 학습 기반 방법은 〈표 2-5〉에 간단히 참조하여 나열하였고, 딥러닝 기반 방법은 아래에 설명하고자 한다.

가. Autoencoder 기반

오토인코더는 병목 계층에서 인코딩된 표현을 학습하고 출력 계층에서 입력을 재생성하는 능력을 가진 딥러닝 모델로 입력 데이터를 압축(인코딩)했다가 다시 복원(디코딩)하는 신경망 구조로, 데이터의 중요 특징을 학습하여 드리프트를 감지하는 데 활용된다. 이는 간단히 재구성 오차(재구성 실패)가 일정 수준을 넘어가면 분포 변화로 간주한다. 아래는 Ali, U. & Mahmood, T.(2024)를 참조한 내용이다.

□ Yong et al.(2020a)

○ 산업 환경의 센서 데이터에서 드리프트를 감지하기 위해 베이지안 오토인코더를 사용

○ 재구성 손실, 알레토릭 및 에피스테믹 불확실성이라는 세 가지 다른 측정 방법이 드리프트 감지에 사용

□ Jaworski et al.(2018)

○ 제한된 볼츠만 머신(Restricted Boltzmann Machine, RBM)의

도움으로 생성된 합성 이진 데이터셋에 RBM을 적용하여 갑작스러운 드리프트와 점진적인 드리프트를 감지

○ 재구성 손실과 자유 에너지라는 두 가지 지표가 데이터의 드리프트를 감지하는 데 사용

○ 드리프트가 발생하는 경우 두 측정값 모두 정상 데이터와 상당한 차이를 나타냄.

□ Jaworski, Rutkowski, Angelov et al.(2020)

○ 동일한 데이터셋에 오토인코더를 적용하고 오토인코더를 사용한 재구성 오류와 교차 엔트로피를 사용하여 갑작스러운 드리프트와 점진적인 드리프트가 감지될 수 있음을 입증

○ 그러나 두 연구 모두 분류기 경계에 대한 데이터 분포 변화의 영향을 고려하지 않는 비지도 방식으로 드리프트 감지를 다룸

□ AEDDM(Autoencoder-based Drift Detection Method)(Ali, U. & Mahmood, T., 2024)

○ 배치 기반 비지도 드리프트 감지 메커니즘을 따르며 아키텍처 수준에서 세 가지 구성 요소가 있음:

- 오프라인 구성 요소: 각 클래스 데이터에 대해 오토인코더를 학습하고 임계값을 계산, 앙상블 구성 요소로 오토인코더의 순서를 정의하고(어떤 오토인코더를 레이어 1과 2에 배치할지), 온라인 구성 요소로 데이터가 배치로 도착하고 전체 배치 데이터 스트림에 대해 드리프트 감지 수행
- 오프라인/학습 단계와 오프라인 또는 학습 단계: 오토인코더

학습(라벨이 있는 데이터셋을 양성과 음성 클래스 데이터로 분리), 각 클래스별로 학습용과 검증용 데이터셋으로 분할(각 클래스에 대해 별도의 오토인코더 학습), 정상(비드리프트) 데이터의 재구성 손실 계산(각 클래스의 검증 데이터를 32개 크기의 배치로 나눔), 각 배치의 인스턴스별 재구성 오차와 배치 평균 재구성 오차 계산, 임계값 계산

- 앙상블 구성 요소: 더 낮은 배치 임계값을 가진 오토인코더를 레이어 1에 배치, 더 높은 배치 임계값을 가진 오토인코더를 레이어 2에 배치, 드리프트가 감지될 때마다 재학습 및 재구성

□ strAEm++DD(Autoencoder-based Incremental Learning Method)(Li, J., et al., 2023)

○ 증분 학습과 드리프트 감지를 결합한 방법, 라벨이 없는 스트리밍 데이터에서 이상치 탐지, 심각한 클래스 불균형 상황에서도 효과적임

나. DriftLens

DriftLens(Greco et al., 2024) 프레임워크는 오프라인 단계와 온라인 단계로 구성된다. 오프라인 단계에서는 과거(학습) 데이터로부터 기준 분포와 거리 임계값을 추정한다. 분포는 다변량 정규분포로 모델링되며 다음과 같이 계산된다. (ⅰ) 전체 배치에 대해(배치별) (ⅱ) 예측된 라벨을 조건으로(라벨별) 온라인 단계에서는 고정된 윈도우에서 데이터 스트림을 분석하여 새로운 분포와 기준 분포를 비교하고, 임계값을 사용하여 드리프트를 식별한다.

□ 데이터 모델링은 딥러닝 모델과 데이터 세트를 입력으로 받아 다변량 정규 임베딩 분포를 추정하는 과정임

○ 모델에서 임베딩을 추출하고 라벨로 보강, 임베딩의 차원을 축소, 평균 벡터 μ와 공분산 행렬 Σ를 계산하여 배치별 분포를 추정

□ 라벨별 특정 분포 추정: 임베딩을 라벨별로 그룹화, 차원 축소 수행, 각 라벨 l에 대해 라벨별 평균 μl과 공분산 Σl을 계산
(여기서 $l \in L$)

〈표 2-5〉 비지도 기반 방법들 분류

연번	드리프트 탐지 방법	유형	세부 내용	참고문헌
1	MD3(Drift Detection using Margin Density)	Partial Batch	SVM Margin Density	(Sethi & Kantardzic, 2015)
2	MD3-RS(MD3 using Random Subspace Model)	Partial Batch	Blindspot Density	(Sethi & Kantardzic, 2017)
3	Predict-Detect(Handling Adversarial Concept Drift)	Partial Batch	Disagreement Density	(Sethi & Kantardzic, 2018)
4	DDMAL (Drift Detection Method Based on Active Learning)	Partial Batch	Density of most significant Instances	(Costa et al., 2018)
5	UDetect(Unsupervised Change Detection for Activity Recognition)	Whole Batch	Classifier's Predicted Class Density	(Bashir et al., 2017)
6	NN-DVI(Nearest Neighbour based Density Variation Identification)	Whole Batch	KNN based Regional Densities	(A. Liu et al., 2018)
7	FAAD(Fast and Accurate Anomaly Detection)	Whole Batch	Anomaly Density	(Li et al., 2019)
8	SQSI-IS (Stream Quantification by Score Inspection - Instance Selection)	Whole Batch	Class Densities	(Andre G. Maletzke et al., 2019)

연번	드리프트 탐지 방법	유형	세부 내용	참고문헌
9	IKS-bdd(Incremental KS based drift detection)	Online -FRW	Feature Densities	(Dos Reis et al., 2016)
10	CD-TDS(Change Detection in Transactional Data Streams)	Online -FRW	Sample means/Edit Distance	(Koh, 2016)
11	OMV-PHT(Online Modified Version Page Hinkley Test)	Online -SRW	Change in Classifier's Confidence	(Lughofer et al., 2016)
12	NM-DDM(Nonparametric Multidimensional Drift Detection Method)	Online -SRW	Log Likelihood Ratio	(Mustafa et al., 2017)
13	Plover(On Learning Guarantees to Unsupervised Concept Drift Detection on Data Streams)	Online -SRW	Divergence	(de Mello et al., 2019)
14	DdDDA(Distribution based Drift Detection Approach)	Online FRW/ SRW	Classifiers Posterior Estimates/ Confidence	(Kim & Park, 2017)
15	An Information-Theoretic Approach to Detecting Changes in Multi-Dimensional Data Streams	Whole Batch	KL Divergence	(Dasu et al., 2006)
16	Concept Drift Detection via Competence Models	Whole Batch	Competence Distance	(N. Lu et al., 2014)
17	Detecting Change in Data Streams	Online -FRW	Relativized Discrepancy	(Kifer et al., 2004)
18	Statistical Change Detection for Multi-Dimensional Data	Whole Batch	log likelihood	(Song et al., 2007)
19	Concept Drift Detection Based on Equal Density Estimation	Online FRW /SRW	Density Estimation	(Gu et al., 2016)
20	Sync Stream (Prototype-based Learning on Concept-drifting Data Streams)	Online SRW	PCA/ Prototype -Tree	(Shao et al., 2014)

연번	드리프트 탐지 방법	유형	세부 내용	참고문헌
21	A PCA-Based Change Detection Framework for Multidimensional Data Streams	Whole Batch	PCA based Density Estimation	(Qahtan et al., 2015)
22	A pdf-Free Change Detection Test Based on Density Difference Estimation	Online -FRW	Least Square Density Difference	(Bu et al., 2018)
23	An incremental change detection test based on density difference estimation	Online -SRW	Least Square Density Difference	(Bu et al., 2017)
24	Regional Concept Drift Detection and Density Synchronized Drift Adaptation	Whole Batch	Local Drift Degree	(A. Liu et al., 2017)
25	A Concept Drift-Tolerant Case-Base Editing Technique	Whole Batch	Competence Distance	(N. Lu et al., 2016)

자료: Ali, U. & Mahmood, T., 2024.

제3절 데이터 드리프트 탐지 방법 비교 및 유형별 적용

앞선 방법들을 간단히 요약 비교하면 다음 〈표 2-6〉과 같다. 이들을 드리프트 유형별로 적합한 탐지 방법을 매칭해보면 〈표 2-7〉과 같다.

먼저 인구 구조나 사회경제적 특성 변화로 인한 Covariate Drift의 경우 PSI(Population Stability Index)와 KS 검정을 주로 활용하며, 최근에는 Autoencoder 기반의 고도화된 탐지 방법도 도입되고 있다. 이는 복지 수요 예측이나 의료 서비스 수요 변화를 파악하는 데 특히 유용하다. Prior Probability Drift는 복지 서비스 수혜 대상의 분포 변화를 설명할 수 있으며, 이는 라벨 분포 모니터링과 Chi-Square 검정을 통해 탐지할 수 있다. Y의 EWMA(Exponentially Weighted Moving Average) 차트를 활용하면 시계열적 변화도 효과적으로 포착할 수 있다. 가장 중요한 Concept Drift의 경우 DDM(Drift Detection Method)과 ADWIN(Adaptive Windowing) 같은 전통적인 방법과 함께, 최신 LSTM 기반 탐지 기법을 활용하여 복지 정책의 효과성 변화나 의료 서비스 품질 변화를 모니터링할 수 있다. 마지막으로 Pipeline Drift는 데이터 처리 과정에서 발생하는 변화를 의미하며, 결측치 체크와 스키마 검증 도구를 통해 관리된다. 이는 사회보장정보시스템의 안정적 운영과 데이터 품질 유지에 핵심적이다.

데이터 드리프트 탐지는 AI 시스템의 안정성과 성능 유지를 위한 필수 요소이다. 드리프트의 유형과 데이터 특성, 라벨 유무에 따라 적절한 탐지 기법을 선택해야 한다. 통계 기반 방식은 가볍고 간단하지만 단일 특성 위주이며, 머신러닝·딥러닝 기반 방식은 더 복잡한 패턴에 대응할 수 있다. 자동화 도구와 하이브리드 접근 방식은 대규모 생산 환경에서 매우 실용적인 솔루션을 제공한다.

〈표 2-6〉 데이터 드리프트 탐지 방법 장단점 비교

분야	대표 방법	장점	단점
통계	PSI, KS, KL, JS, Chi-square	간단·라벨 필요 없음	고차원 데이터에 취약
Seq	DDM, EDDM, ADWIN, PH, EWMA	실시간 탐지·민감도 조절 가능	파라미터 영향 큼
ML	UDD, CIDD-ADODNN, LSTMDD	라벨 있거나 예측 기반 반응 가능	라벨 필요 또는 모델 복잡
DL	Autoencoder, DriftLens	대규모 복합 데이터 대응 가능	계산 자원 많이 필요

자료: 저자 작성.

〈표 2-7〉 드리프트 유형별 탐지 기법 매핑

드리프트 유형	적합한 탐지 방법
Covariate Drift	PSI, KS 검정, Autoencoder, KL Divergence
Prior Probability Drift	라벨 분포 모니터링, Chi-Square, Y의 EWMA 차트
Concept Drift	DDM, ADWIN, 정확도/F1 기반 모니터링, UDD, LSTM 기반 탐지
Pipeline Drift	결측치 체크, 정규표현식, 체크섬, 스키마 검증 도구

자료: 저자 작성.

제3장

보건복지분야 데이터 드리프트 시뮬레이션

제3장 보건복지분야 데이터 드리프트 시뮬레이션

제1절 가구구성 변화에 따른 데이터 드리프트 분석

가구구성에서 가구원수 분포의 변화는 인구 구조, 주거 형태, 고령화 속도, 혼인 및 출산에 영향을 주는 사회 구조 전반의 변화라고 할 수 있다. 1980년대부터 2000년대까지는 가구원수별 가구 구성에서 4인 가구가 가장 높은 비중을 차지했음에 반해, 2010년에는 2인 가구가 가장 높은 비중을 차지했고 2015년 이후에는 1인 가구가 가장 높은 비중을 차지하고 있다. 평균 가구원수도 1980년에는 4.5명에서 2010년에는 2.7명, 2024년에는 2.2명으로 감소하였다. 2025년 기준 중위소득 및 급여별 선정기준 변경 발표에서도 공식 발표 기준은 4인 가구 기준이지만 가구원수별로도 제시하고 있다.

가구구성에서 가구원수 감소는 복지 수급 대상 선정 기준에서도 1인 가구 증가의 개인 특성 반영에 대한 필요성이 커지고 있고 복지 사업의 특성에도 영향을 준다. 이 절에서는 특성 드리프트, 공변량 드리프트에 속하는 가구구성 변화를 데이터 드리프트 측도로 분석하였다. 기준연도는 1990, 2000, 2010, 2015년으로 선정하여 데이터 드리프트 탐지 방법은 PSI, IV, KS를 사용하여 측도값을 제시하였다.

〈표 3-1〉 분석 활용 지표

지표	약어	해석 기준	의미
Population Stability Index	PSI	- <0.1: 안정적 - 0.1~0.25: 약한 드리프트 - >0.25: 강한 드리프트	분포 간 변화의 크기 측정
Information Value	IV	- <0.02: 거의 없음 - 0.02~0.1: 약함 - 0.1~0.3: 중간 - >0.3: 강함	분포 차이를 정보량 기준으로 측정
Kolmogorov-Smirnov	KS	- 0.1~0.2: 약한 차이 - 0.2~0.3: 중간 차이 - >0.3: 강한 차이	누적분포 간 최대 거리 측정

자료: 저자 작성.

〈표 3-2〉 가구원수별 가구구성과 평균 가구원수(1980~2017년)

(단위: 천 가구, %, 명)

연도		1980	1990	2000	2010	2015	2016	2017
가구수 (천 가구)		7,969	11,355	14,312	17,339	19,111	19,368	19,674
가구원수별 가구구성 (%)	1인	4.8	9.0	15.5	23.9	27.2	27.9	28.6
	2인	10.5	13.8	19.1	24.3	26.1	26.2	26.7
	3인	14.5	19.1	20.9	21.3	21.5	21.4	21.2
	4인	20.3	29.5	31.1	22.5	18.8	18.3	17.7
	5인	20.0	18.8	10.1	6.2	4.9	4.8	4.5
	6인 이상	29.8	9.8	3.3	1.8	1.5	1.4	1.3
평균 가구원수 (명)		4.5	3.7	3.1	2.7	2.5	2.5	2.5

자료: 통계청, 각 연도, 「인구총조사」

〈표 3-3〉 가구원수별 가구구성과 평균 가구원수(2018~2024년)

(단위: 천 가구, %, 명)

연도		2018	2019	2020	2021	2022	2023	2024
가구수 (천 가구)		19,979	20,343	20,927	21,448	21,774	22,073	22,294
가구원수별 가구구성 (%)	1인	29.3	30.2	31.7	33.4	34.5	35.5	36.1
	2인	27.3	27.8	28.0	28.3	28.8	28.8	29.0
	3인	21.0	20.7	20.1	19.4	19.2	19.0	18.8
	4인	17.0	16.2	15.6	14.7	13.8	13.3	12.7
	5인	4.3	3.9	3.6	3.3	3.1	2.9	2.7
	6인 이상	1.2	1.0	0.9	0.8	0.7	0.6	0.6
평균 가구원수 (명)		2.4	2.4	2.3	2.3	2.2	2.2	2.2

자료: 통계청, 각 연도, 「인구총조사」

기준연도를 1990년도로 선정할 경우, 2010년부터 PSI, IV, KS 모두 아주 높게 나타났다. 1990년대에는 4인 가구가 중심이었던데 반해, 2010년으로 오면서 2인 가구가 비중이 가장 많아졌기 때문이다.

〈표 3-4〉 가구구성 데이터 드리프트 분석: 기준연도 1990년

연도	PSI	IV	KS
1980	0.011	8.248	0.030
2000	0.032	5.503	0.080
2010	0.212	22.629	0.205
2015	0.328	33.072	0.250
2016	0.349	34.942	0.257
2017	0.375	37.422	0.268
2018	0.406	40.241	0.279
2019	0.444	43.571	0.292
2020	0.487	47.535	0.306

연도	PSI	IV	KS
2021	0.545	52.724	0.325
2022	0.595	57.229	0.338
2023	0.628	60.308	0.346
2024	0.662	63.095	0.355

자료: 저자 작성.

기준연도를 2000년으로 선정하였을 경우, 여전히 PSI, IV 모두 상당히 높은 편으로 나타났는데 특히 2010년 이후부터 지표가 더 높게 나타났다. 이러한 결과는 1인 가구와 핵가족이 급속히 증가한 반면 4인 이상 가구 비중은 지속적으로 하락한 가구구성의 변화를 보여준다고 볼 수 있다.

〈표 3-5〉 가구구성 데이터 드리프트 분석: 기준연도 2000년

연도	PSI	IV	KS
1980	0.067	24.635	0.094
1990	0.032	5.503	0.080
2010	0.082	7.681	0.124
2015	0.160	14.974	0.170
2016	0.175	16.332	0.177
2017	0.194	18.128	0.188
2018	0.216	20.233	0.199
2019	0.243	22.790	0.212
2020	0.274	25.720	0.226
2021	0.317	29.761	0.245
2022	0.355	33.387	0.258
2023	0.380	35.859	0.266
2024	0.407	38.252	0.274

자료: 저자 작성.

기준연도를 2010년으로 선정하였을 때는 이전 시점을 기준연도로 선정하였을 때에 비해 드리프트 수준이 낮아졌지만 여전히 PSI와 KS가 점

진적으로 증가하고 있다. 이는 1990~2000년대보다 변화의 폭이 다소 둔화되었지만 가구구성의 변화가 지속적으로 되고 있음을 의미한다.

〈표 3-6〉 가구구성 데이터 드리프트 분석: 기준연도 2010년

연도	PSI	IV	KS
1980	0.270	45.081	0.219
1990	0.212	22.629	0.205
2000	0.082	7.681	0.124
2015	0.013	1.222	0.046
2016	0.017	1.630	0.053
2017	0.023	2.222	0.063
2018	0.031	2.995	0.074
2019	0.042	4.032	0.087
2020	0.056	5.324	0.102
2021	0.076	7.287	0.120
2022	0.095	9.127	0.133
2023	0.109	10.454	0.142
2024	0.123	11.779	0.150

자료: 저자 작성.

기준연도를 2015년도로 선정하였을 경우 2020~2024년과 비교해보면 PSI와 KS가 소폭 증가하였는데, 이는 1인 가구의 비중의 지속적인 증가했기 때문이다.

〈표 3-7〉 가구구성 데이터 드리프트 분석: 기준연도 2015년

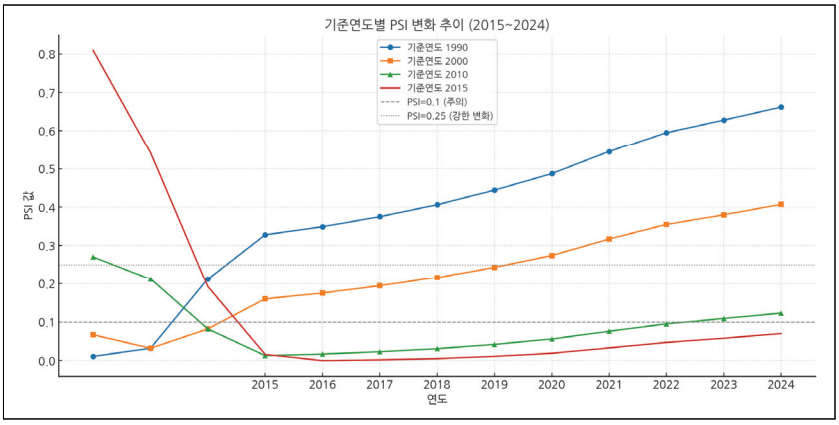
연도	PSI	IV	KS
1980	0.809	0.809	0.334
1990	0.542	0.542	0.295
2000	0.192	0.192	0.186
2010	0.016	0.016	0.052
2016	0.000	0.000	0.008

연도	PSI	IV	KS
2017	0.002	0.002	0.019
2018	0.005	0.005	0.031
2019	0.011	0.011	0.046
2020	0.019	0.019	0.062
2021	0.033	0.033	0.082
2022	0.047	0.047	0.096
2023	0.058	0.058	0.105
2024	0.070	0.070	0.115

자료: 저자 작성.

이를 종합적으로 살펴보면, 장기적 변화를 볼 수 있는 1990년의 기준은 구조적이고 급진적으로 시대 전환을 보여주고 있고, 2000~2010년 기준은 핵가족 중심에서 1~2인 가구로의 전환기, 2015년 기준은 이미 전환된 상태에서의 지속적 세분화·고령화로 볼 수 있다. 즉, 각 기준연도마다 정책 수립 시 기준 가구 단위의 재설계를 검토할 필요성이 있다.

[그림 3-1] 기준연도별 PSI 변화 추이(2015~2024년)



자료: 저자 작성.

〈표 3-8〉 기준연도별 PSI 주요 특징

기준연도	변화 추이	민감도	주요 특징
1990년	지속적이고 가파른 상승	매우 높음	2015년 이후 PSI>0.25로 급격한 변화 감지. 장기적 구조 변화에 민감
2000년	점진적 상승	중간 이상	꾸준히 상승하며 2020년 이후 변화 심화. 여전히 유의미한 구조 차이 감지
2010년	비교적 완만한 상승	낮음	2015년까지 변화 거의 없음, 이후 서서히 증가. 최근 구조와의 차이 적음
2015년	거의 변화 없음	매우 낮음	분석 구간 전체에서 가장 안정적. 가장 최근 현실 반영기준점

자료: 저자 작성.

기준연도를 현재에서 멀리 잡을수록(1990년) 변화를 더 민감하게 반영(최근 연도로 갈수록 드리프트가 누적되어 커짐)하기 때문에 장기적 추세를 감지하는 데 적합하고, 최근 기준연도(2015)의 경우에는 현시점과 유사한 상태를 반영하기 때문에 단기 예측이나 현재 진단에 적합한 측면이 있다. 따라서 복지정책, 기준중위소득, 주거지표 기준 등에서 "기준이 언제 설정되었는가"에 따라 분석의 해석이 달라질 수 있다.

데이터 모델링 관점에서는 기준 데이터의 시점이 결과 해석에 미치는 영향이 크다는 것을 보여준다.

제2절 고혈압 기준 변화에 따른 데이터 드리프트 분석

미국의 심장학회와 심장협회는 ≥140/90mmHg의 고혈압 기준을 ≥130/80mmHg으로 2017년도에 변경하였다. 진단 기준의 정의 자체가 변한 것인데, 이는 고혈압의 기준 변화는 단순한 수치 변경으로 볼 게 아니라 라벨 드리프트로 볼 수 있다. 고혈압 기준 임계값이 낮아지면 고혈압 환자수가 증가하고, 예측모형 및 연구 결과의 해석들이 달라지게 된다. 과거 기준으로 학습된 모델은 새로운 임계값을 기준으로 다시 학습해야 하며, 오즈비, 회귀계수 등 변수별 예측력이 달라지고 궁극적으로 정책 설계에서의 우선순위에도 영향을 준다.

〈표 3-9〉 고혈압 진단 기준 완화 내용(2022년 개정안)

2018년 진료 지침	수축기(mmHg)	이완기(mmHg)
정상혈압	<120	<80
주의혈압	120~129	<80
고혈압전단계	130~139	80~89
고혈압 1기	140~159	90~99
고혈압 2기	≥160	≥100
수축기단독고혈압	≥140	<90
2022년 개정안	**수축기(mmHg)**	**이완기(mmHg)**
중저위험도 고혈압(합병증 없음)	<140	<90
노인 고혈압(합병증 없음)	<140	<90
고위험도 고혈압(합병증 없음)	<130	<80
고위험도 당뇨병(합병증 없음)	<130	<80
심혈관질환(합병증 동반)	<130	<80
만성콩팥병, 알부민뇨 없음(합병증 동반)	<140	<90
알부민뇨 동반(합병증 동반)	<130	<80

자료: 최선, 2022.5.12., 한국도 고혈압 기준 강화 동참…변경 이유는?, 메디컬타임즈, https://www.medicaltimes.com/Main/News/NewsView.html?ID=1147283(검색일: 2025.9.9.).

대한고혈압학회는 2022년 춘계 학술대회에서 일부 고위험군에 한해 130/80mmHg로 상향된 고혈압 기준을 제시하였지만, 일반 기준은 과거 지침을 준용한다는 내용을 공개하였다.

이 절에서는 라벨 드리프트 현상을 살펴보기 위해 국민건강영양조사 9기 데이터를 활용하여 고혈압 임계값 기준을 ≥140/90mmHg에서 ≥130/80mmHg으로 변경하였을 때의 고혈압 유무에 미치는 요인값의 변화를 로지스틱 회귀모형으로 분석해 보고자 한다.

1. 국민건강영양조사 분석 개요

국민건강영양조사는 1995년 제정된「국민건강증진법」제16조에 근거하여 시행하는 전국 규모의 건강 및 영양조사이다. 1998년부터 2005년까지 3년 주기로 시행하였으며, 국가통계의 시의성 향상을 위해 2007년부터 매년 시행하고 있다. 본 조사의 목적은 국민의 건강수준, 건강행태, 식품 및 영양섭취 실태에 대한 국가 단위의 대표성과 신뢰성을 갖춘 통계를 산출하고, 이를 통해 국민건강증진종합계획의 목표 설정 및 평가, 건강증진 프로그램 개발 등 보건정책의 기초자료로 활용하는 것이다.[1)]

국민건강영양조사는 가구원확인조사, 건강설문조사, 검진조사, 영양조사로 이루어져있는데, 검진조사 항목의 수축기혈압, 이완기혈압으로 고혈압 유무를 확인할 수 있다. 이 절에서는 라벨 드리프트의 현상을 확인하고자 수축기혈압, 이완기혈압 정보로 고혈압 여부 종속변수를 생성하고, 성별, 나이, 소득 5분위수(가구), 교육수준, 주관적 건강인지, 월간 음주율, 현재 흡연율, 스트레스 인지율 요인을 독립변수로 선정하였다.

---

1) 질병관리청, 2024, 국민건강영양조사 제9기 1·2차년도(2022-2023) 원시자료, 이용지침서, p. 3. 조사목적 내용 발췌.

고혈압 유무 변수는 기존(≥140/90mmHg)의 임계값을 기준으로 생성한 변수는 Y1, 새로운(≥130/80mmHg) 임계값을 기준으로 생성한 변수는 Y2로 정의하였다.

〈표 3-10〉 국민건강영양조사 9기 분석 활용 변수

구분	변수명	변수 설명	
독립 변수	sex	성별	1: 남자 2: 여자
	age	나이	1~79: 1~79세 80: 80세 이상
	ho_incm5	소득 5분위수(가구)	소득 5분위수 (1: 하, 5: 상)
	edu	교육수준 재분류 코드	교육수준 재분류 (1: 초졸 이하, 4: 대졸 이상)
	D_1_1	주관적 건강인지	주관적 건강인지 (1: 매우 좋음, 5: 매우 나쁨) 무응답은 분석에서 제외
	dr_month	월간 음주율	0: 평생 비음주, 최근 1년간 월 1잔 미만 음주 1: 최근 1년간 월 1잔 이상 음주
	sm_presnt	현재 흡연율	0: 과거 흡연, 비흡연 1: 현재 흡연
	mh_stress	스트레스 인지율	0: 스트레스 적게 느낌 1: 스트레스 많이 느낌
종속 변수	Y1 (HE_sbp, HE_dbp)	고혈압 유무 (기존 기준)	≥140/90mmHg
	Y2 (HE_sbp, HE_dbp)	고혈압 유무 (신규 기준)	≥130/80mmHg

자료: 저자 작성.

국민건강영양조사 9기의 응답자는 6,929명이며 분석에 활용하는 변수에서의 무응답, 결측치를 제외하였을 때 분석 대상자수는 5,663명이었다. Y1 값이 1인 대상자 수는 674명, Y2 값이 1인 대상자 수는 1,927명

으로 고혈압 임계값 기준이 완화되었을 때, 고혈압의 대상자 수가 크게 증가함을 알 수 있다.

2. 라벨 드리프트 분석

모형 1은 고혈압 기존(≥140/90mmHg)의 임계값을 기준으로 생성한 변수 Y1을 종속변수로, 모형 2는 새로운(≥130/80mmHg) 임계값을 기준으로 생성한 변수 Y2를 종속변수로 로지스틱 회귀분석한 결과이다.

〈표 3-11〉 회귀계수 비교

변수명	모형 1 회귀계수(유의확률)	모형 2 회귀계수(유의확률)
절편	-4.60 (<2e-16) ***	-2.14 (3.30e-13) ***
성별	-0.23 (0.023) *	-0.45 (2.09e-11) ***
나이	0.049 (<2e-16) ***	0.035 (<2e-16) ***
소득 5분위수 (가구)	-0.017 (0.648)	-0.005 (0.853)
교육수준 재분류 코드	-0.049 (0.3326)	-0.011 (0.760)
주관적 건강인지	0.040 (0.436)	0.043 (0.230)
월간 음주율	0.204 (0.036) *	0.212 (0.001) **
현재 흡연율	0.139 (0.291)	0.105 (0.241)
스트레스 인지율	0.224 (0.039) *	0.083 (0.266)

주: $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
자료: 저자 작성.

두 모형의 회귀계수를 비교해보면 성별과 연령, 월간 음주율은 모두 유의하였고, 모형 2에서의 성별, 월간 음주율 영향이 더 유의미하였다. 스트레스 인지율 요인은 모형 1에서 유의미한 결과가 도출되었다. 모형 1의 절편이 더 낮게 나타났는데 이는 모형 1의 종속변수가 상대적으로 고혈압 발생 가능성이 더 낮은 집단이라는 의미이다. 이는 Y1에 비해 Y2가

완화된 진단 기준으로, 진단 혈압 기준이 낮아져서 더 많은 환자가 고혈압으로 분류됨을 뜻한다.

두 모형의 오즈비 결과에서도 회귀계수 해석 결과와 동일하게 스트레스 요인은 모형 1에서 유의미하였다.

〈표 3-12〉 오즈비 비교

변수명	모형 1(Y1) 오즈비(95% 신뢰구간)	모형 2(Y2) 오즈비(95% 신뢰구간)
(절편)	0.010 (0.004, 0.03)	0.118 (0.07, 0.21)
성별	0.80 (0.65, 0.97) *	0.64 (0.58, 0.73) ***
나이	1.05 (1.04, 1.06) ***	1.04 (1.03, 1.04) ***
소득 5분위수 (가구)	0.98 (0.91, 1.06)	1.00 (0.95, 1.05)
교육수준 재분류 코드	0.95 (0.86, 1.05)	0.99 (0.92, 1.06)
주관적 건강인지	1.04 (0.94, 1.15)	1.04 (0.97, 1.12)
월간 음주율	1.23 (1.01, 1.49) *	1.24 (1.09, 1.41)**
현재 흡연율	1.15 (0.88, 1.48)	1.11 (0.93, 1.32)
스트레스 인지율	1.25 (1.01, 1.54) *	1.09 (0.94, 1.26)

주: $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
자료: 저자 작성.

이번 로지스틱 회귀 결과는 고혈압 진단 여부(Y)라는 진단 기준이 변한, 라벨 드리프트 현상을 반영하고 있음을 보여준다. 진단 기준이 완화되면 더 많은 대상자들이 고혈압으로 분류되어 모형의 절편과 위험요인별 상대적 효과가 변한다. 라벨 드리프트 현상은 독립변수와 종속변수와의 연관성이 달라질 수 있기에, 기존 모형의 예측 성능이 저하될 수 있다.

제3절 복지사각지대 데이터 드리프트 분석

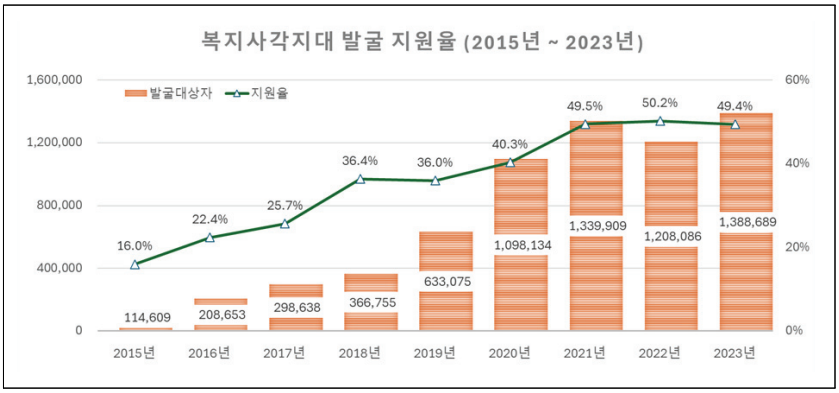
2015년 7월에 시행된 「사회보장급여의 이용·제공 및 수급권자 발굴에 관한 법률」은 정보시스템을 통해 복지사각지대를 발굴하고 지원 체계를 구축하는 근거가 되었다(김은하, 2022). 이에 따라 보건복지부는 2015년 12월 복지사각지대 소외계층을 선제적으로 발굴하고 지원하는 복지사각지대 발굴관리 시스템을 구축하였다. 해당 시스템은 복지사각지대에 놓인 취약계층의 발굴 및 지원을 위해 단전, 단수, 단가스 등의 위기 정보를 수집하고, 발굴모형을 통해 개인별 복지서비스 수혜 확률을 산출함으로써 우선순위의 발굴 대상자를 선별한다(한국사회보장정보원, 2025).

2024년 기준, 위기정보는 19개 기관으로부터 45종이 입수되고 있으며, 각 기관으로부터 입수된 데이터는 검증과정을 거친 후 분석대상자 데이터셋으로 구성된다. 분석대상자는 발굴모형을 통해 위기도(복지서비스 수혜 확률)가 산출되고, 위기 가능성이 큰 약 15만 명을 발굴대상자로 선별하여 지자체에 통보한다. 발굴 차수마다 분석대상자(2024년 기준 약 900만 명)와 발굴 대상자의 규모는 다를 수 있으며, 지자체에서는 자체적으로 발굴 대상자를 추가하여 차수별로 약 20만 명의 규모를 조사하고 있다.

복지사각지대 발굴모형 운영을 시작한 이후 발굴 대상자는 2015년 11.5만 명에서 2023년 138.8만 명으로 약 12배 이상 증가하였다. 발굴 대상자에 대한 지원율은 2015년 16.0%에서 2023년 49.4%로 3배 증가하였다. 2021년부터 발굴 대상자 절반은 복지서비스 지원이 이루어진 것으로 나타났다. 2015년 발굴시스템 구축 이후 2023년까지 665.6만 명을 조사하여 290.2만 명(43.6%)에게 공공·민간 복지서비스 지원이 이루어졌다(보건복지부, 2024).

〔그림 3-2〕 복지사각지대 발굴 지원율(2015~2023년)

(단위: 명, %)



자료: 보건복지부, 2024, 45종 위기정보 활용해 2024년 2차 복지사각지대 발굴 시행, 보도자료(3.25).

발굴 지원율의 지속적 상승은 발굴모형의 효과성을 시사하며, 이는 데이터 기반 복지정책의 성공적 사례로 평가된다. 그러나 이러한 성과의 지속가능성은 모형의 예측 정확도에 상당 부분 의존하고 있으며, 모형의 정확도는 입력 데이터의 품질과 일관성에 직결된다. 따라서 데이터 드리프트에 대한 체계적이고 지속적인 관리는 선택이 아닌 필수적인 요구사항이며, 복지 사각지대 발굴 업무의 지속적인 성과 창출을 위한 핵심 과제라 할 수 있다.

1. 복지사각지대 데이터의 특징

복지사각지대 발굴시스템은 다양한 외부기관과의 정보 연계를 통해 운영되고 있다. 2024년 기준 총 19개 협력기관으로부터 45종의 위기 관련 데이터를 수집하여 활용하고 있다. 이러한 정보 연계 범위는 지속적으로 확장되어 왔는데, 2016년 23종에서 시작하여, 2022년 34종, 2023년

39종, 2024년 45종으로 확대되었다. 수집되는 위기정보의 범위도 매우 포괄적이다. 단전, 단수, 단가스 등 기본적인 생활 인프라 관련 위기상황부터, 의료 위기, 주거 위기, 고용 위기, 경제 위기, 정신적 건강 문제, 재해재난 등 개인과 가구가 직면할 수 있는 다층적 위기상황을 종합적으로 포괄하고 있다. 이를 통해 복합적 위기에 노출된 취약계층을 보다 정확하게 식별할 수 있는 기반을 마련하고 있다.

〈표 3-13〉 복지사각지대 발굴시스템 연계정보(45종)

<table>
<tr><th colspan="2">근거: 법률(제12조 제1항 각호 및 제2항)</th><th colspan="2">근거: 시행령(제8조 제2항 별표 2 각호)</th></tr>
<tr><th>정보</th><th>보유기관</th><th>정보</th><th>보유기관</th></tr>
<tr><td>단전</td><td>한국전력공사</td><td>국민연금보험료 체납</td><td rowspan="2">건강보험공단</td></tr>
<tr><td>단수</td><td>상수도사업본부</td><td>의료 위기[1)]</td></tr>
<tr><td>단가스</td><td>도시가스사</td><td>범죄 피해</td><td>경찰청</td></tr>
<tr><td rowspan="2">초중고 교육비 지원 중
학교장 추천</td><td rowspan="2">교육부</td><td>화재 피해</td><td>소방청</td></tr>
<tr><td>재난 피해</td><td>행정안전부</td></tr>
<tr><td>건보료 체납</td><td rowspan="2">건강보험공단</td><td rowspan="2">주거 위기[2)]</td><td rowspan="2">국토교통부
한국토지주택공사
각 지방개발공사
아파트 관리사무소</td></tr>
<tr><td>건보료 부과내역</td></tr>
<tr><td>기초수급 탈락·중지</td><td rowspan="2">보건복지부</td><td rowspan="2">고용 위기[3)]</td><td rowspan="2">고용노동부
근로복지공단</td></tr>
<tr><td>복지시설 퇴소</td></tr>
<tr><td>금융연체</td><td rowspan="2">신용정보원</td><td rowspan="3">방문건강사업 대상</td><td rowspan="7">보건복지부</td></tr>
<tr><td>채무조정 중지(실효)자</td></tr>
<tr><td>통신비 체납</td><td>한국정보통신
진흥협회</td></tr>
<tr><td>노후긴급자금 대부(실버론)</td><td>국민연금공단</td><td>기저귀 분유 지원</td></tr>
<tr><td colspan="2" rowspan="6">1) ① 의료비 부담 과다
② 장기 요양
③ 중증질환 산정특례
④ 요양급여 장기 미청구
⑤ 장기요양 등급
⑥ 재난적 의료비 지원 대상</td><td>신생아 난청 지원</td></tr>
<tr><td>영양플러스 미지원</td></tr>
<tr><td>맞춤형 급여 신청</td></tr>
<tr><td>전기료 체납</td><td>한국전력공사</td></tr>
<tr><td>수도요금 체납</td><td>상수도사업본부</td></tr>
</table>

근거: 법률(제12조 제1항 각호 및 제2항)		근거: 시행령(제8조 제2항 별표 2 각호)	
정보	보유기관	정보	보유기관
2) ① 전세 기준금액 이하 ② 월세 기준금액 이하 ③ 공공임대주택 임대료 체납자 ④ 공동주택 관리비 체납자 3) ① 개별연장급여 대상자 ② 실업급여 수급자(임금체불, 폐업) ③ 비자발적 사유로 고용보험 상실 후 재취득이 없는 자 중 실업급여 미수급자 ④ 일용근로자 중 실업급여 미수급자 ⑤ 산재요양 종결 후 근로 단절 ⑥ 고용위기(최근 1년 이내 고용보험 가입 이력이 없는 대상자)		가스요금 체납	도시가스사
		자살고위험군	자살예방센터
		내원사유 자해·자살	응급의료센터
		휴·폐업자	국세청
		세대주가 사망한 가구	행정안전부
		주민등록세대원	

자료: 보건복지부, 2024, 45종 위기정보 활용해 2024년 2차 복지사각지대 발굴 시행, 보도자료(3.25).

입수 데이터는 매월 또는 격월로 입수되고 있으며, 상세 위기정보와 입수주기는 〈표 3-14〉와 같다. 연계시스템을 통해 입수된 후 통합된 위기정보들은 입수 대상자 확인 과정을 통해 개인과 가구 확인이 이루어진 후, 최종 분석대상자 데이터셋으로 구성된다.

위기 정보는 주요 정보가 45개이지만, 각 주요 정보가 발생할 경우 관련된 속성정보가 추가로 입수된다. 예를 들어, 주요 정보가 단전 여부일 때는 속성정보로 체납시작년월, 체납종료년월, 체납개월수, 중지개월수, 단전시작일자, 단전종료일자, 체납전체금액, 월평균 사용량, 1~12개월 이전 사용량 등 20개의 속성정보가 추가로 수집된다. 주요 정보와 속성정보는 약 300개의 변수로 이루어진다.

위기 정보 외에도 사회보장정보시스템에서 수집된 사회보장과 관련된 정보가 추가된다. 해당 정보는 개인영역과 가구영역, 고용 관련 정보에 해당한다. 개인영역에는 희귀/만성 질환정보, 장애정보, 수감정보, 공적연금수급정보, 소득인정액정보가 해당된다. 가구영역은 가구유형, 장애인가구유형, 가구원 희귀/만성 질환정보, 가구원 수감정보, 주거유형 및

가구유형 등이며, 고용 관련 정보는 일용직 종사, 자활근로, 노인일자리 참여 및 무직 상태 등이다(이우식 외, 2021).

최종적으로 분석대상자 데이터셋은 개인의 연령이나 성별, 거주지, 가구구성원 정보와 함께 위기 정보 및 사회보장 정보 등을 포함하여 총 350여 개의 변수로 구성된다. 이렇게 구성된 분석대상자 데이터셋은 발굴 대상자 선별을 위해 예측모형에 투입된다.

〈표 3-14〉 복지사각지대 발굴 시스템 위기정보 및 입수 주기

연번	정보명	입수 주기	보유기관
1	단전 여부	매월	한국전력공사
2	단수도 여부	격월	상수도사업본부
3	단가스 여부	격월	도시가스사
4	전기료체납 여부	매월	한국전력공사
5	국민연금체납 여부	격월	건강보험공단
6	건강보험료체납 여부	격월	건강보험공단
7	화재피해 여부	격월	소방청
8	본인부담경감대상자 여부	매년	건강보험공단
9	피부양의무자장기요양 여부	격월	건강보험공단
10	전세금액기준 이하 가구 여부	매월	국토교통부 한국토지주택공사
11	월세금액기준 이하 가구 여부	매월	국토교통부 한국토지주택공사
12	고용보험 개별연장 급여대상 여부	매월	고용노동부 근로복지공단
13	고용보험 실직사유 대상 여부	매월	고용노동부 근로복지공단
14	고용보험 비대상 여부	매월	고용노동부 근로복지공단
15	방문건강 집중관리군 여부	매월	보건복지부 한국사회보장정보원
16	기저귀 조제분유 지원대상자 여부	매월	보건복지부 한국사회보장정보원
17	신생아 난청확진자 여부	격월	보건복지부 한국사회보장정보원

연번	정보명	입수 주기	보유기관
18	자살예방관리 대상자 여부	격월	자살예방센터
19	자살시도 대상자 여부	매월	응급의료센터
20	위기학생 여부	매월	교육부
21	범죄피해 여부	격월	경찰청
22	시설 입퇴소 여부	매월	보건복지부 한국사회보장정보원
23	기초생활 긴급지원 수급탈락 여부	매월	보건복지부 한국사회보장정보원
24	공공임대 주택체납자 여부	매월	국토교통부 한국토지주택공사
25	산재요양 종결 후 근로단절자 여부	매월	고용노동부 근로복지공단
26	재난피해자 여부	매월	행정안전부
27	금융연체 대상자 여부	격월	한국신용정보원
28	의료비용 과다지출 여부	격월	건강보험공단
29	일용근로대상자여부	매월	고용노동부 한국고용정보원
30	영양플러스 미지원가구 여부	매월	보건복지부 한국사회보장정보원
31	심뇌혈관질환 대상자 여부		건강보험공단
32	휴폐업 가구 여부	격월	국세청
33	공동주택관리비 체납대상자 여부	매월	국토교통부 한국토지주택공사
34	세대주 사망 세대원 여부	격월	행정안전부
35	건강보험료 납부정보 여부	격월	건강보험공단
36	통신비 체납대상자 여부	격월	한국정보통신진흥협회

자료: 이우식 외, 2021, 복지사각지대 발굴체계 재정립 연구; 보건복지부, 2024, 45종 위기정보 활용해 2024년 2차 복지사각지대 발굴 시행, 보도자료(3.25).

2. 복지사각지대 데이터 드리프트 분석

데이터 드리프트는 시간의 경과에 따라 입력 데이터의 분포가 변화하는 현상으로, 머신러닝 모델의 성능 저하를 야기하는 주요 원인 중 하나이다. 특히 복지사각지대 발굴모형과 같이 사회경제적 환경 변화에 민감

한 모델의 경우, 외부 환경 변화, 정책 개편, 데이터 수집 과정의 변화 등으로 인해 데이터 드리프트가 빈번하게 발생할 수 있다. 이러한 드리프트는 모형의 예측 정확도를 저하시키고, 결과적으로 복지 대상자 선정의 공정성과 효율성에 부정적 영향을 미칠 수 있다.

가. 가상 데이터 생성

복지사각지대 데이터 드리프트 분석을 수행하기 위해서는 복지사각지대 발굴시스템에 입수되는 실제 데이터를 이용해야 하지만, 외부에는 데이터가 제공되지 않고 있다.이에 복지사각지대 발굴시스템 연구를 참고하여 5개년의 데이터를 유사하게 생성하여 활용하였다.

이우식 외(2020)에서는 복지사각지대 발굴모형 개발 과정에서 사용된 학습데이터 분석결과를 통해, 2016년 12월부터 2019년 3월까지 발굴한 대상자 661,566명에 대한 변수별 위험요인 보유자수를 제시하고 있다. 해당 자료가 과거의 자료이긴 하나, 이를 기반으로 새로운 데이터셋을 생성하고 데이터 드리프트를 탐지하는 데 이해를 돕기 위해 활용하는 것은 적절해 보인다. 데이터셋 생성을 위해 참고한 위기정보별 유효 대상자 수는 다음 〈표 3-15〉와 같다. 대상자 수는 661,566명이며, 해당 대상자 수가 가장 높은 변수는 '월세금액기준이하가구여부' 289,610명(43.8%), '피부양의무자장기요양여부' 162,899명(24.6%)의 순으로 나타났다. 상대적으로 해당 대상자 수가 적은 위기정보도 일부 포함되어 있다.

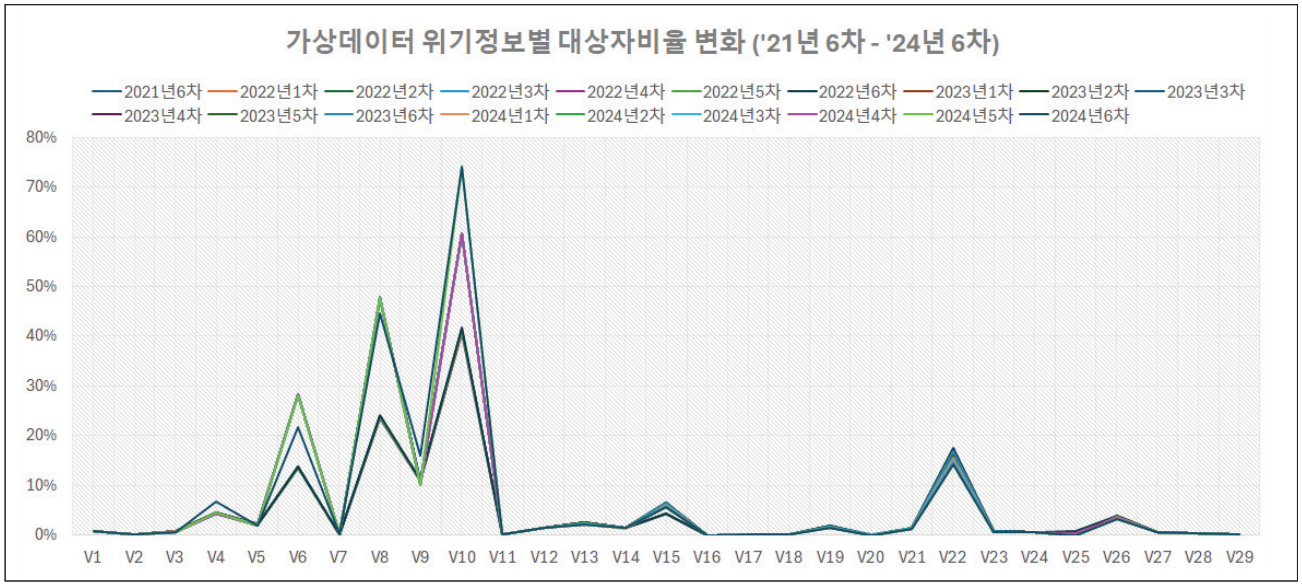
〈표 3-15〉 복지사각지대 발굴 분석대상자의 위기정보 보유자 수 현황

구분	위기정보	해당 대상자수(명)	해당 비율
전체	대상자수	661,566	100%
V1	단전여부	5,238	0.8%
V2	단수도여부	639	0.1%
V3	단가스여부	4,292	0.6%
V4	전기료체납여부	28,701	4.3%
V5	국민연금체납여부	14,445	2.2%
V6	건강보험료체납여부	92,832	14.0%
V7	화재피해여부	413	0.1%
V8	피부양의무자장기요양여부	162,899	24.6%
V9	전세금액기준이하가구여부	70,910	10.7%
V10	월세금액기준이하가구여부	289,610	43.8%
V11	고용보험개별연장급여대상여부	184	0.0%
V12	고용보험실직사유대상여부	9,233	1.4%
V13	고용보험비대상여부	16,430	2.5%
V14	방문건강집중관리군여부	9,710	1.5%
V15	기저귀조제분유지원대상자여부	28,574	4.3%
V16	신생아난청확진자여부	1	0.0%
V17	자살예방관리대상자여부	662	0.1%
V18	자살시도대상자여부	444	0.1%
V19	위기학생여부	11,907	1.8%
V20	범죄피해여부	28	0.0%
V21	시설입퇴소여부	9,326	1.4%
V22	기초생활긴급지원수급탈락여부	110,986	16.8%
V23	공공임대주택체납자여부	4,754	0.7%
V24	산재요양종결후근로단절자여부	4,604	0.7%
V25	재난피해자여부	1,563	0.2%
V26	금융연체대상자여부	25,334	3.8%
V27	의료비용과다지출가구여부	4,081	0.6%
V28	일용근로대상자여부	2,408	0.4%
V29	영양플러스미지원가구여부	462	0.1%

자료: 이우식 외, 2020, 복지사각지대 발굴관리시스템 예측모형 개선 방안 연구.

가상의 데이터셋 생성은 점진적 비율 기반 시뮬레이션을 통해 5년간(2020~2024년) 격월 30개 시점의 데이터 드리프트 시뮬레이션 데이터셋을 생성하였다. 전체 기간에서 2023년을 전후로 안정기와 드리프트 발생기로 구분하여 서로 다른 변화 패턴을 적용하였다. 시뮬레이션에서는 기본적으로 모든 변수에 대해서는 기본변동계수 ± 5%를 적용하였고, 일부 변수에 대해서는 드리프트 증가계수를 별도로 설정하여 적용하였다. 다음 그림은 생성된 가상데이터 2021년 6차부터 2024년 6차까지의 위기정보를 보유한 대상자의 비율 변화를 보여준다. 가상 데이터 생성을 위한 시뮬레이션에서 V4, V6, V8, V10, V15, V22 변수에는 드리프트 증가계수가 적용되어 다른 위기정보들에 비해 변화가 뚜렷하게 나타난다. 이는 2023년 입수 위기정보를 39종에서 44종으로 확대함과 동시에 기존 금융 연체금액 범위를 상향(보건복지부, 2023)하는 등 데이터 드리프트가 발생할 수 있는 요인이 실제 발생하였기 때문에 이를 드리프트 증가계수로 반영하고자 하였다. 전체 데이터는 29개의 변수로 구성되어 있다.

[그림 3-3] 가상데이터 위기정보별 대상자 비율 변화



자료: 저자 작성.

나. 데이터 드리프트 분석

데이터 드리프트는 모델 학습에 사용된 데이터와 운영 단계에서 입수되는 데이터의 통계적 속성이나 분포의 차이로 인해 발생한다. 이는 시간 경과에 따라 모델이 학습데이터와 다른 분포를 가진 데이터에 대해 추론을 수행하게 되면서 예측 정확도의 저하를 초래한다. 복지사각지대 발굴을 위해 활용하는 위기정보는 다양한 요인들에 의해 변화가 발생할 수 있으며, 이는 예측 가능한 요인과 불가능한 요인들이 섞여 있다.

데이터 드리프트를 산출하기 위해서는 기준 시점과 드리프트에 대한 모니터링이 수행되는 비교(현재) 시점에 대한 정의가 필요하다. 이를 위해 가상 데이터셋의 2021년 6차는 기준 시점으로, 2022년 1차부터 2024년 6차까지는 분포의 차이를 확인하기 위한 비교 시점으로 정의하였다.

PSI 산출은 29개 변수에 대해 각각 산출되며, 각 변수는 0과 1의 범주를 가지므로 다음과 같이 표현할 수 있다.

$$PSI = \sum_{i=0}^{1} (f_{a,i} - f_{e,i}) \cdot ln(\frac{f_{a,i}}{f_{e,i}})$$

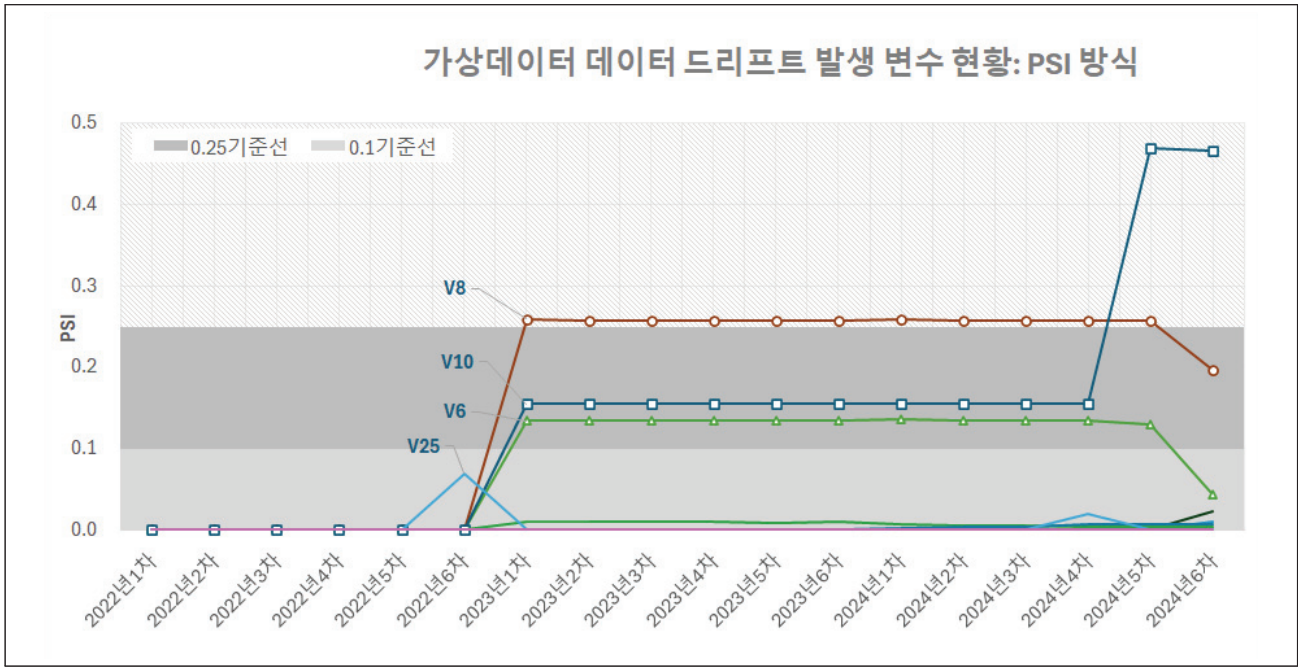
여기서 $f_{a,i}$는 비교 시점(Actual)의 i번째 구간에 속하는 데이터의 비율(%Actual)을 의미하고, $f_{e,i}$는 기준 시점(Expected)의 i번째 구간에 속하는 데이터의 비율(%Expected)을 의미한다. 분석에 사용된 변수들은 0과 1일 때의 값을 더하여 변수별 PSI를 산출하게 된다.

다음 그림은 데이터 드리프트 분석을 수행하고, 시점별 데이터 드리프트가 발생한 변수의 현황을 보여준다. 2023년 이전에는 모든 변수들의

PSI가 0.1 미만의 낮은 수준을 보여, 드리프트 발생 위험수준을 보이는 변수가 거의 없는 드리프트 안정기를 보이고 있다. 이 시기는 복지사각지대 발굴모형의 입력 데이터가 기준 시점(2021년 6차)과 유사한 분포를 유지하며 정상적으로 운영되었던 기간으로 해석된다. 2023년부터는 V6('건강보험료체납여부'), V10('월세금액기준이하가구여부'), V8('피부양의무자장기요양여부') 등 일부 변수군에서 PSI가 정상 범위를 벗어나 급격하게 증가하기 시작하였다. 특히 V8 변수는 다른 변수들보다 더 높은 상승세를 보이며 PSI가 0.3 수준까지 증가하여, 해당 변수가 외부 환경 변화에 더 민감하게 반응함을 시사한다. 이는 특정 변화가 일부 변수에 먼저 영향을 미치기 시작했음을 나타낸다. 2024년 하반기에는 V10 변수의 PSI가 0.5 근처까지 급증하여 최고점을 기록하였다. V10은 전체 분석 기간 중 가장 극적인 변화를 보인 변수로 확인되었다. 일반적으로 PSI 0.25 이상은 심각한 변화로 간주되는데, 이는 모델의 안정성에 치명적인 영향을 미칠 수 있는 수준이다. 2024년 6차에서는 다소 감소하였으나 여전히 위험 수준을 유지하고 있다.

이러한 패턴의 급격한 데이터 드리프트는 복지사각지대 발굴모형의 예측 성능에 직접적이고 심각한 영향을 미칠 수 있다. 특히 V10 변수의 PSI 0.5 수준은 해당 변수군의 분포가 기준 시점과 완전히 달라졌음을 의미하며, 이는 모형의 신뢰성을 근본적으로 훼손할 수 있는 수준이다. 따라서 V10 변수에 대해서는 즉시 긴급 대응이 필요하며, 해당 변수의 구성 요소들에 대한 전면적인 재검토가 요구된다. 또한 정상 범위를 상회하는 모든 변수들에 대해서도 단계적 대응책 마련과 지속적인 모니터링이 수행되어야 한다.

[그림 3-4] 가상데이터 데이터 드리프트 발생 변수 현황: PSI 적용



자료: 저자 작성.

PSI는 데이터 드리프트 탐지를 위한 기본적인 지표로 널리 활용되지만, 비대칭적 특성으로 인해 분포 변화의 전체적인 양상을 파악하는 데 한계가 있다. 이를 보완하기 위해 Jensen-Shannon Divergence(JSD)를 추가적으로 활용하였다. JSD는 기준 시점과 현재 시점 간의 분포 유사성을 대칭적으로 측정하는 지표로서, 두 분포와 그들의 평균 분포 간의 쿨백-라이블러 발산(Kullback-Leibler Divergence, KLD)을 기반으로 계산된다(Salem, Buschjaeger & Morik, 2012). JSD는 확률분포 간의 거리를 정보이론적 관점에서 측정하여 더욱 정교한 드리프트 분석이 가능하며, 0과 1 사이의 값을 가져 해석이 쉬운 장점이 있다.

JSD를 이용한 데이터 드리프트 분석(Dhinakaran, 2023)은 다음과 같은 과정을 따른다. 가상데이터의 V4 변수에 대해 2021년 6차(기준 시점)와 2022년 1차(현재 시점)를 비교한 결과, JSD 값이 0.000014로 나타났다. 이는 0에 매우 가까운 값으로, 기준 시점과 현재 시점 간 두 확률분포는 거의 동일하다는 것을 의미한다. 즉, V4 변수에 유의미한 데이터 드리프트가 발생하지 않았음을 나타낸다.

[그림 3-5]는 가상데이터 29개 변수(V1~V29)에 대한 시간별 JSD 변화를 보여주는 데이터 드리프트 분석 결과이다. 드리프트가 발생한 시점은 PSI 산출결과와 동일하게 나타났다. 가장 높은 JSD는 V8('피부양의무자장기요양여부'), 다음으로 V6('건강보험료체납여부')와 V10('월세금액기준이하가구여부')으로 나타났다. V10은 드리프트 발생기에는 ~0.019 수준으로 나타났으나, 급격한 변화기(2024년 5차 이후)에는 0.057로 급상승하였다. 다른 변수들 대비 약 3배 높은 수준으로 심각한 분포 변화를 보였다.

* JSD 산출

(1단계) 평균 분포(M) 계산
두 확률 분포 P와 Q의 산술 평균을 구해 새로운 평균 분포 M을 산출
- P = [0.95580, 0.04420]
- Q = [0.95558, 0.04442]
- M = (P+Q)/2 = [0.95569, 0.04431]
  · M(해당하지 않음) = (0.95580+0.95558)/2 = 0.95569
  · M(해당함) = (0.04420+0.04442)/2 = 0.04431

(2단계) KLD 계산
P와 M 사이의 KLD, Q와 M 사이의 KLD를 각각 계산
- $DKL(P \parallel M)$: P분포와 평균 분포 M 사이의 KLD

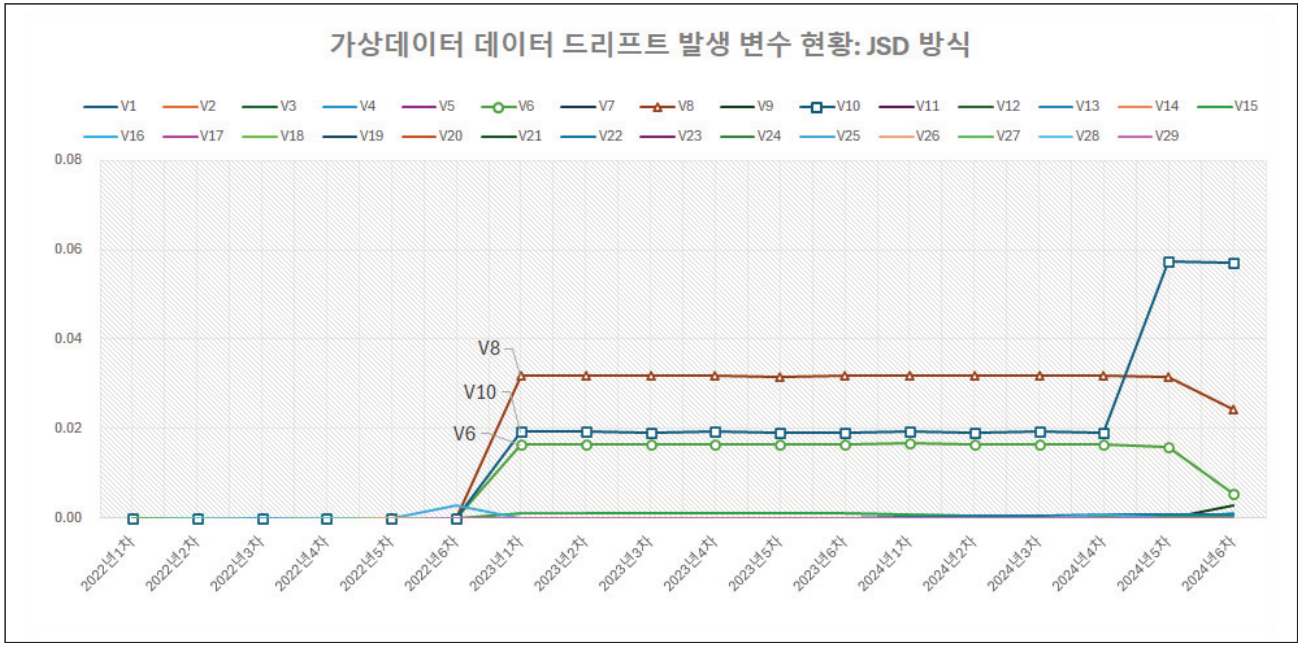
$$0.95580 \times \ln(0.95569 / 0.95580) + 0.04420 \times \ln(0.04431 / 0.04420) \approx 0.00000014$$

- $DKL(Q \parallel M)$: Q분포와 평균 분포 M 사이의 KLD

$$0.95558 \times \ln(0.95569 / 0.95558) + 0.04442 \times \ln(0.04431 / 0.04442) \approx 0.00000014$$

(3단계) 최종 JSD 계산
두 KLD 값의 평균을 계산하여 최종 JSD 값을 획득
- $JSD(P \parallel Q) = (DKL(P \parallel M) + DKL(Q \parallel M))/2$
  · JSD = (0.00000014+0.00000014)/2 = 0.00000014

[그림 3-5] 가상데이터 데이터 드리프트 발생 변수 현황: JSD 적용



자료: 저자 작성.

3. 복지사각지대 발굴모형의 모델 드리프트 평가

모델(이하 '예측') 드리프트는 머신러닝 모델의 예측 결과 분포가 시간에 따라 변화하는 현상으로, 모델의 출력값 분포 자체가 달라지는 상황을 의미한다. 복지사각지대 발굴모형은 발굴 대상자를 선정하는 과정에서 대상자별 위기도를 확률값으로 산출하여 우선순위를 결정하고, 한정된 복지자원을 가장 필요한 대상에게 효율적으로 배분할 수 있도록 지원하는 데이터 기반 의사결정 도구이다.

모형의 출력값은 0부터 1 사이의 위기도 확률값으로 표현되며, 값이 클수록 복지서비스가 시급히 필요한 대상자임을 나타낸다. 따라서 예측 분포의 변화는 복지 대상자 선정 과정의 일관성과 공정성에 직접적인 영향을 미치게 된다.

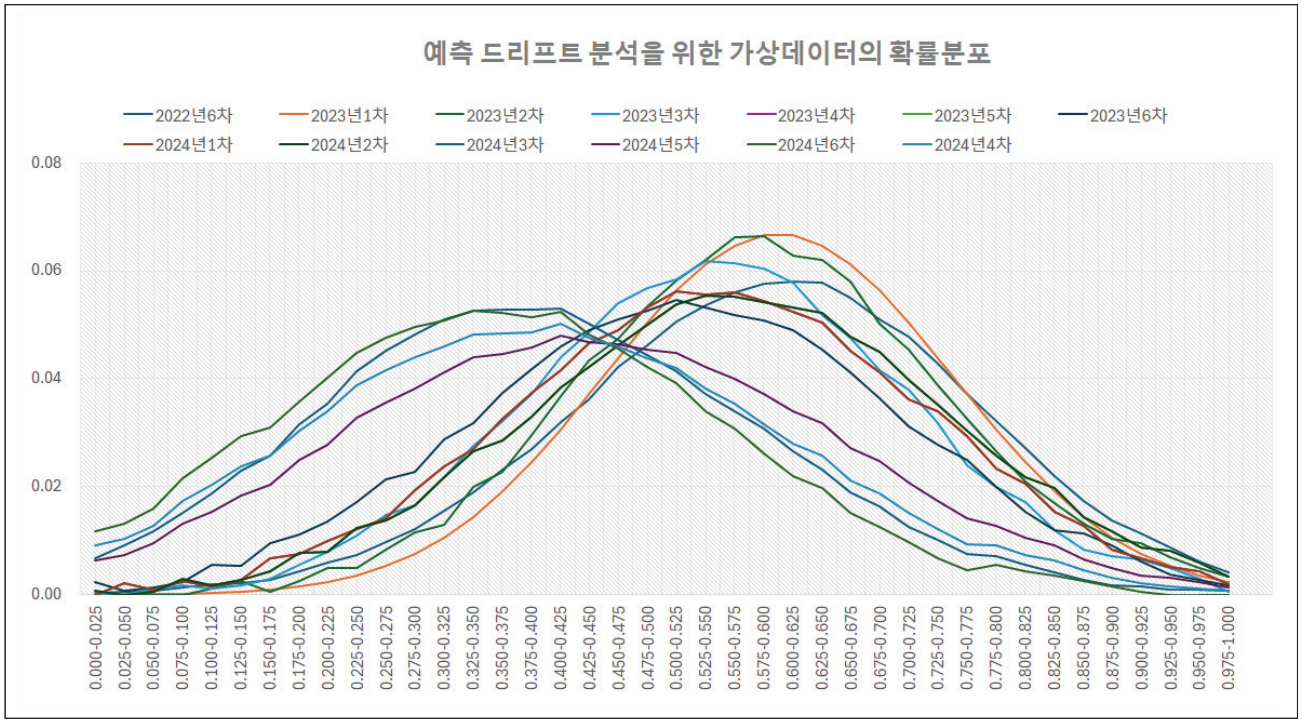
예측 모델의 PSI 산출을 위해 예측확률 0부터 1까지를 0.025 단위로 균등하게 나누어 총 40개 구간으로 설정하였다. 각 구간별 해당 대상자의 비율을 산출하여 기준 시점과 비교 시점 간의 분포 변화를 측정하였으며, PSI는 다음과 같이 계산된다.

$$PSI = \sum_{k=0}^{40} (f_{a,k} - f_{e,k}) \cdot ln(\frac{f_{a,k}}{f_{e,k}})$$

여기서 $f_{e,k}$와 $f_{a,k}$는 각각 확률구간 k에 대한 기준 시점(Expected)과 비교 시점(Actual)의 해당 대상자 비율을 의미하며, k는 40개 구간을 나타낸다. PSI 값이 0.1 미만이면 안정적, 0.1 이상 0.25 미만이면 약간의 변화, 0.25 이상이면 심각한 변화로 해석된다. 다음 [그림 3-6]은 예측모델의 시점별 확률 분포 변화를 분석하기 위한 가상 데이터를 생성한 결과

이다. 그림에서 확인할 수 있듯이, 각 시점별로 위기도 예측 확률의 분포가 서로 다른 패턴을 보이고 있다. 대부분의 분포는 0.5~0.7 구간을 중심으로 한 종 모양을 나타내지만, 시점에 따라 분포의 중심위치나 폭이 상이하게 나타난다. 특히 일부 시점에서는 분포가 더 뾰족한 형태를 보이지만, 다른 시점에서는 상대적으로 평평하고 넓은 분포를 보여주고 있다. 이러한 분포 변화는 동일한 대상자 집단에 대해서도 시점별로 모델의 예측결과가 달라질 수 있음을 시사한다. 이는 복지사각지대 발굴 과정에서 우선순위 선정의 일관성에 영향을 미칠 수 있는 요인이다.

[그림 3-6] 예측 드리프트 분석을 위한 가상데이터의 확률 분포



자료: 저자 작성.

2022년 6차를 기준 시점으로 설정하여 총 12개 시점에 대한 PSI 분석을 수행한 결과는 〈표 3-16〉과 같다. 2023년 1차(PSI=0.045), 2차(PSI=0.071), 4차(PSI=0.074), 5차(PSI=0.041)는 모두 안정적 구간에 해당하여 예측 분포가 기준 시점과 유사한 수준을 유지하였다. 반면 2023년 3차(PSI=0.103)와 6차(PSI=0.166)는 약간의 변화 구간에 해당하여 변화 시작점으로서 모니터링이 필요한 시점으로 나타났다. 2024년 1차(PSI=0.074)와 2차(PSI=0.041)에는 일시적으로 안정 구간으로 회복되었으나, 2024년 3차부터 급격한 변화가 시작되었다. 2024년 3차(PSI=1.169)에서 급격한 분포 변화가 시작되었으며, 4차(PSI=1.032)에서는 심각한 드리프트가 지속되고, 5차(PSI=0.673)에서도 여전히 높은 변화율을 보였다. 특히 2024년 6차(PSI=1.807)는 전체 분석 기간 중 가장 높은 수치를 기록하여 최대 변화점으로서 즉시 조치가 필요한 상황임을 보여준다. 이는 2024년 하반기에 예측 모델의 분포가 기준 시점 대비 극심하게 변화하였음을 의미하며, 복지사각지대 발굴 과정에서 대상자 선정 기준의 재검토가 필요함을 시사한다.

PSI 값이 0.25 이상인 심각한 드리프트 상황에서는 모델의 예측 신뢰도가 현저히 저하될 가능성이 크므로, 새로운 모델이 배포되기 전까지 기존 모델의 운영을 일시 중단하거나 예측 결과에 대한 추가적인 검증절차를 도입하는 것을 고려해야 한다.

예측 드리프트에 대한 PSI 분석 결과를 토대로 복지사각지대 발굴모형의 예측 안정성을 확보하기 위해서는 먼저, 모델 재훈련 및 성능 재평가가 즉시 수행되어야 한다. 2024년 3차부터 6차까지의 PSI 값이 모두 심각한 변화 구간에 해당하므로, 해당 기간의 데이터를 활용한 모델 재학습을 통해 현재의 분포 패턴을 반영한 새로운 예측모형을 구축할 필요가 있다. 특히 PSI 값이 1.807로 최고치를 보인 2024년 6차 데이터는 현재 운

영 환경을 가장 잘 반영하는 것으로 판단되어 모델 업데이트의 핵심 기준으로 활용되어야 한다.

〈표 3-16〉 예측 드리프트 상세 PSI 분석 결과

시점	PSI	기준 비교	해석
2023년 1차	0.045	안정적	-
2023년 2차	0.071	안정적	-
2023년 3차	0.103	주의	변화 시작점-모니터링 필요
2023년 4차	0.074	안정적	-
2023년 5차	0.041	안정적	-
2023년 6차	0.166	주의	변화 시작점-모니터링 필요
2024년 1차	0.074	안정적	-
2024년 2차	0.041	안정적	-
2024년 3차	1.169	위험	급격한 분포 변화 시작
2024년 4차	1.032	위험	심각한 드리프트 지속
2024년 5차	0.673	위험	여전히 높은 변화율
2024년 6차	1.807	위험	최대 변화점-즉시 조치 필요

자료: 저자 작성.

다음으로, 현재 운영 모델의 지속 사용 여부에 대한 검토가 필요하다. 운영 모델의 지속 사용 여부는 여러 분석 결과 및 정무적인 판단을 종합하여 이루어져야 한다.

제4절 소결

본 장에서는 특성 드리프트, 공변량 드리프트로 가구구성 변화에 따른 데이터 드리프트 분석, 라벨 드리프트로 고혈압 기준 변화에 따른 데이터 드리프트 분석, 복지사각지대 데이터 드리프트 분석을 실시하였다. 가구구성 변화에 따른 데이터 드리프트 분석은 매크로 데이터, 고혈압 기준 변화에 따른 데이터 드리프트 분석은 마이크로데이터인 국민건강영양조사로, 복지사각지대 데이터 드리프트 분석은 가상데이터를 통해 복지사각지대 발굴시스템의 데이터 드리프트 현상을 체계적으로 분석하였다.

가구구성 변화에 따른 공변량 드리프트는 기준연도 설정을 다르게 하여 PSI, IV, KS 측도로 드리프트 민감도를 계산하였고, 기준연도를 현재와 가까운 시점으로 할 경우 단기적 변화에 대한 탐지력이 높아짐을 확인하였다. 이는 모델의 성능과 직결되는 부분으로, 학습데이터의 기준 시점은 모델의 성능과 모델의 안전성, 실제 환경과의 유사성을 고려하여야 한다. 보건복지정책에서는 기준중위소득 등의 정책 기준이 변경된 시점에 따라 현재 상황에 대한 분석이 달라질 수 있다. 따라서 데이터 드리프트 연구 시 보건복지 정책 환경 변화 및 정책의 시간적 맥락, 모델의 활용 목적에 따른 접근 방식이 필요하다.

국민건강영양조사 9기 데이터를 활용한 로지스틱 회귀 분석 결과는 고혈압 진단 기준 변화에 따른 라벨 드리프트 현상이 예측 모델의 구조적 특성에 미치는 영향을 보여준다. 라벨 드리프트는 동일한 개체에서 다른 분류 결과를 얻게 되면서 모델이 학습해야 할 타겟 변수의 분포와 의미를 변화시킨다. 로지스틱 분석 결과에서도 학습된 모델의 오즈비와 변수별 예측력이 다르게 나타난 것은 라벨 드리프트가 독립변수와 종속변수 간의 연관성 자체를 변화시키기 때문이다. 라벨 드리프트는 입력 변수와 목

표 변수 간의 관계성이 시간이 지남에 따라 변하는 개념 드리프트의 한 형태로도 볼 수 있다. 고혈압 진단 기준 변화에 따른 라벨 드리프트 분석은 보건의료 분야에서 진단 기준의 변화가 단순한 임계값 조정이 아니라 모델의 예측 및 정책 의사결정에 큰 영향을 미치며 모델의 지속적인 모니터링과 업데이트의 중요성을 보여준다.

복지사각지대 데이터 드리프트 분석은 가상데이터를 활용한 시뮬레이션 분석을 통해 2023년을 기점으로 드리프트가 어떻게 발생할 수 있는지를 확인하였다. 또한 2024년 하반기 시나리오에서는 예측확률의 분포에서 극심한 변화가 나타남도 함께 시연하였다. 2015년 구축 이후 지속적으로 성과를 거두어온 복지사각지대 발굴모형이지만, 시간이 지나면서 데이터 드리프트나 예측 드리프트로 인한 성능 저하를 피할 수는 없다.

데이터 드리프트가 발생한 변수들은 중요도와 영향도에 따라 구분하여 관리해야 한다. 변수별 기여도 분석을 통해 모델 성능에 미치는 영향을 정량적으로 평가하고, 이를 바탕으로 대응 우선순위를 선정한다. 우선순위가 높은 핵심 변수에서 드리프트가 발생하였을 때 즉시 대응이 필요하며, 상대적으로 우선순위가 낮은 변수의 경우에는 모니터링 강화 수준에서 관리할 수 있다. 특히 드리프트가 발생한 변수가 전체 변수의 일정 비율을 초과하는 경우에는 모델 안정성을 검토하고, 재평가가 필요할 수 있다.

데이터 드리프트 발생 분석은 개별 변수를 대상으로 수행하지만, 복지사각지대 발굴에 활용되는 변수들은 변수 간 상관도가 높은 변수들로 구성되어 있어 변수 간 관계들도 함께 고려되어야 한다. 또한 시계열성을 가진 위기 변수들의 경우 계절성, 추세, 주기성 등을 고려한 별도의 분석이 필요하다. 일시적 변화와 구조적 변화를 구분하여 대응 방안을 달리 적용해야 하며, 구조적 변화로 판단되는 경우 해당 변수의 피처 엔지니어링 과정을 재검토하거나 새로운 파생 변수 생성 등을 고려해야 한다.

특히 데이터 드리프트에서 중요한 부분은 데이터 파이프라인의 변화를 빠르게 감지하는 것이다. 복지사각지대 위기 정보들은 연계기관에서 수신받는 정보들로, 실제 데이터 수집 과정에서 발생하는 변경사항들을 즉시 감지하기 어렵고, 변경사항들에 대한 정보를 수집하기도 쉽지 않다. 따라서 데이터 제공기관과의 협의를 통해 변경 상황이 발생한 때에는 빠르게 정보를 확보하도록 공유체계를 마련해야 한다. 입수 과정에서의 데이터 검증 과정을 강화하여 데이터 신뢰도를 지속적으로 유지하는 부분도 중요하다.

이 장에서는 가상데이터를 활용한 시뮬레이션에 기반하고 있어 실제 운영환경에서의 드리프트 패턴과는 차이가 있을 수 있다는 한계가 있다. 따라서 향후 실제 복지사각지대 데이터를 활용한 검증 연구가 필요하며, 이를 통해 본 연구에서 제시한 모니터링 체계의 실효성을 확인해야 할 것이다. 그럼에도 불구하고 복지사각지대 발굴시스템의 지속가능한 성과 창출을 위해 데이터 드리프트 관리가 필수적이라는 점을 강조하고, 체계적인 관리방안을 제시했다는 점에서 의의가 있다. 앞으로 실제 운영환경에 이러한 모니터링 체계를 적용하고 지속적으로 개선해 나간다면, 복지사각지대 해소라는 정책 목표 달성에 기여할 수 있을 것으로 기대된다.

데이터 드리프트 관리방안은 4장에서 구체적으로 데이터 드리프트 관리 프로세스에 따라 제시하고자 한다.

제4장

데이터 드리프트 관리 방안

제 4 장 데이터 드리프트 관리 방안

제1절 데이터 드리프트 관리 프로세스

최근 보건복지분야에서 인공지능과 빅데이터의 활용이 증가하면서, 데이터 드리프트 현상에 대한 체계적인 관리의 중요성이 부각되고 있다. 데이터 드리프트 관리 방안에서 중요한 것은 “왜 모니터링하는가”와 “누구를 위한 것인가”에 대한 명확한 정의가 필요하다. 모니터링의 궁극적 목적은 모형이 적용되는 업무 목표와 연계되어야 하며, 드리프트 탐지 기준과 대응 전략은 이러한 업무 목표에 기반하여 설정되어야 한다. 특히 사회보장정보시스템과 보건의료 빅데이터 플랫폼에서 수집되는 방대한 데이터의 품질 관리를 위해서는 체계적인 프로세스가 설계되어야 한다.

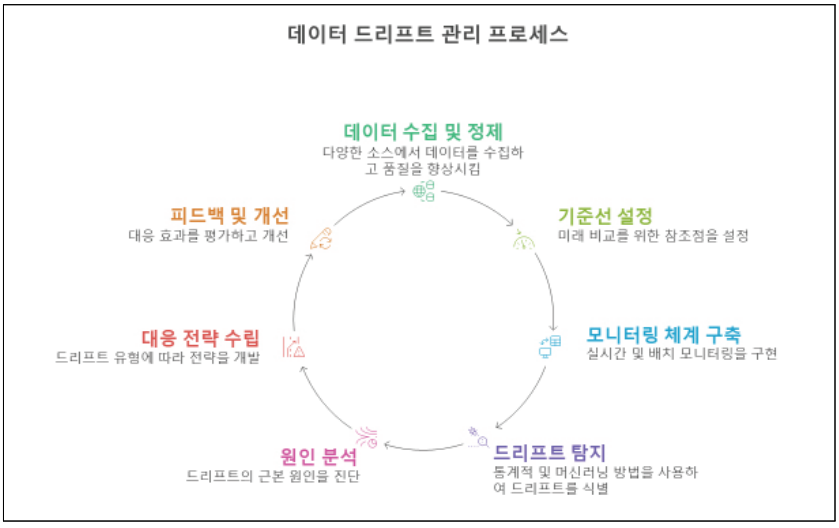
데이터 드리프트 관리방안의 전반적인 프로세스는 다음과 같은 요소로 구성할 수 있다.

첫 번째, 데이터 수집 및 전처리 단계이다. 정형데이터, 센서 데이터, 이미지 데이터 등의 다양한 형태 자료를 수집하고 통합할 수 있어야 하며, 수집 경로와 시점 등의 정보를 메타데이터로 기록한다. 서베이 조사 활용 시에는 표본설계 메타데이터(표본틀 버전, 가중치 변수 등)도 필수 관리 항목으로 지정한다. 데이터에 대한 품질관리도 함께 이루어지는데, 결측치 처리, 이상치 탐지, 중복 제거를 검토하고 데이터 구조에 대한 일관성을 검증한다.

두 번째, 기준선(Baseline) 설정이다. 데이터 드리프트 탐지를 위한 절대적 기준 마련을 위해 기준선 설정은 필요하며, PSI, KS, Chi-Square 임계값을 미리 정의한다. 기준 시점 데이터는 활용 목적에 따라 고려해야

하며 정책 시행 직전 또는 최근 1년 자료 등이 될 수 있다. 데이터 변수별 분포(평균, 표준편차, 4분위수, 범주 비율), 결측률, 성능 지표(AUC, F1-score, RMSE 등)를 기록할 수 있는 체계를 마련한다. 기준선은 단일 기준선일 수도 있지만 보건복지 데이터는 계절성·정책 변경·행정주기 영향이 큼에 따라 최근 12개월 Rolling Baseline 등의 갱신주기를 고려하여 운영하는 것이 필요할 수 있다.

[그림 4-1] 데이터 드리프트 관리 프로세스 도식화



자료: 저자가 작성한 내용을 Napkin.ai를 활용하여 시각화 하였음.

세 번째는 모니터링 단계이다. 입력 데이터(covariates), 출력 데이터(label)의 분포, 모델 성능 지표, 파이프라인 및 시스템 로그를 기록하며 대시보드 및 알림 기능으로 시각화한다. 알림 기능의 경우, 단순 알림 기능뿐만 아니라 데이터 드리프트 발생 확률을 추정하며 그 추이를 보여주는 예측 알림 방식도 설계한다. 실시간 스트리밍 모니터링과 일/주 단위 보고서를 자동 생성해주는 배치 모니터링도 고려할 수 있다.

네 번째는 드리프트 탐지 단계이며 2단계에서 설정한 기준값과 현재값을 비교한다. 공변량 드리프트는 PSI의 경우 0.25 이상, KS의 경우 0.2 이상 되면 알림 기능으로 현상을 보여주도록 한다. 모형에 대한 성능으로는 AUC를 지표로 사용할 수 있고, 기존의 타입 및 범위를 벗어났을 경우도 탐지할 수 있도록 한다. 사전확률 드리프트는 Chi-Square, EWMA 차트로 라벨 분포 변화를 보여준다. 개념 드리프트는 DDM, ADWIN, Page-Hinkley Test, LSTM 기반 탐지를 고려하며, 파이프라인 드리프트는 데이터 타입 불일치, 스키마 변경, 결측률 급증으로 변화를 탐지한다. 데이터 드리프트 탐지에서는 단일 탐지 지표보다는 다중 지표 조합을 통해 탐지의 정확성을 높일 수 있도록 하며 가짜 드리프트 탐지(False Positive)를 최소화하도록 한다.

다섯 번째 단계는 원인 분석 단계로 어떠한 이유로 데이터의 분포가 변하고 예측 성능이 하락하였는지에 대한 원인을 분석하고 기록한다. 데이터 자체로는 전처리 오류 또는 수집 파이프라인, 센서, 배치 주기가 변경되었을 수 있고, 외부 요인으로는 정책 변화(예: 복지 대상자 선정 기준 변경), 계절성, 사회적 사건(팬데믹, 경기침체 등)의 구조적 환경변화가 있을 수 있다. 개념 드리프트로 라벨 생성 규칙 변화, 목표 함수 자체의 변화도 확인해볼 필요가 있다. 특정지역, 시간대·집단에 한정된 국소적 드리프트는 분리해서 분석한다.

여섯 번째 단계는 대응 전략 수립이다. 이 단계는 데이터 드리프트를 탐지하고, 원인을 파악한 뒤, 드리프트 유형에 따라 전략을 개발하며 품질을 높이고 안정화 하기 위한 방안을 모색하는 단계이다. 데이터 측면에서는 불균형 자료(imbalanced data)에 대한 보정, 데이터 증강, 라벨 정제, 새로운 기준선에 맞춘 학습 데이터 업데이트 등이 있고, 모델 측면에서는 재학습, feature set 재구성 등이 있을 수 있다. 데이터 드리프트가

누적되면 기존 학습 모델은 현실 데이터를 제대로 반영하지 못하고 성능 저하를 일으킬 수 있다. 정기적인 재학습 주기(예: 분기별, 반기별, 혹은 성능 저하 탐지 시점)를 사전에 설정하고 최신 데이터를 반영하여 모델을 재학습하도록 한다. 단기적 충격(예: 팬데믹, 경기 불황)뿐만 아니라 장기적 추세 변화(예: 인구 구조 변화, 소득 수준 변화)도 동시에 반영할 수 있도록 재학습 시점에 데이터 커버리지를 충분히 고려해야 할 것이다. 데이터가 빠른 속도로 유입되거나 환경 변화가 빈번할 경우에는 일괄 재학습은 비용과 시간이 많이 소요된다. 대안으로 실시간 학습(Online Learning) 또는 점진적 학습(Incremental Learning) 기법을 도입하여, 변화된 데이터 패턴을 신속하게 반영할 수 있도록 한다. 운영 전략은 오프라인(offline)의 성능 검증과 온라인(online)의 테스트를 병행한다. 모델 업데이트 전에 오프라인 환경에서 기존의 기준선 데이터 및 최신 데이터를 활용하여 성능을 검증한다. 그리고 실제 운영 환경에서 발생하는 다양한 변수를 반영하기 위해 온라인 환경에서 기존 모델과 새로운 모델을 동일한 조건에서 비교 검증한다. 이러한 작업은 모델 교체로 인한 서비스 성능 저하나 예기치 못한 부작용을 최소화할 수 있다.

마지막 일곱 번째 단계인 피드백 및 개선 부분에서는 규칙 재설계 및 자동 업데이트의 기능을 담당한다. 룰 재설계는 탐지 임계값 조정(PSI=0.25 → 0.2)이나 계절성을 반영한 동적 임계값 적용 등이 검토된다. 그리고 자동화 수준을 향상시키기 위해 모니터링 → 탐지 → 원인 분석 → 안정화 → 재학습 → 배포까지 End-to-End 파이프라인을 구축하여 순환적 프로세스를 통해 점진적으로 관리 체계를 고도화한다.

다음 〈표 4-1〉에는 데이터 드리프트 관리 프로세스 각 단계별 주요 점검 항목과 체크리스트 질문을 정리하였다.

〈표 4-1〉 데이터 드리프트 관리 프로세스 단계별 주요 점검 항목

단계	주요 점검 항목	체크리스트 질문
① 데이터 수집 및 전처리	데이터 확보, 품질 관리, 스키마 검증	- 데이터 출처와 수집 주기가 명확히 정의되어 있는가? - 결측치·이상치·중복 데이터 처리가 완료되었는가? - 스키마와 메타데이터가 최신 상태로 관리되고 있는가?
② 기준선 설정	기준 데이터셋 정의, 임계값 설정	- 기준 시점 데이터(Reference Distribution)가 명확히 저장되어 있는가? - PSI, KS, Chi-Square 등 탐지 기준 임계값이 정의되어 있는가? - 성능 지표(AUC, F1-score 등)가 기준선으로 기록되어 있는가?
③ 모니터링	데이터·모델 성능 추적, 알림 시스템	- 입력 데이터 분포, 출력 분포, 성능 지표가 실시간/주기적으로 수집되는가? - 대시보드와 알림 시스템이 구축되어 있는가? - 로그가 장기 보관·분석 가능하게 관리되는가?
④ 드리프트 탐지	지표 기반 탐지, 다중 방법 적용	- Covariate Drift 탐지(PSI, KS, Autoencoder)가 수행되는가? - Prior Probability Drift 탐지(Chi-Square, EWMA)가 적용되는가? - Concept Drift 탐지(DDM, ADWIN, LSTM 등)가 적용되는가? - Pipeline Drift 탐지(스키마 변경, 결측률 급증)가 수행되는가?
⑤ 원인 분석	내부/외부 요인 파악, 모델 영향 분석	- 데이터 파이프라인 변경, 오류 여부가 점검되었는가? - 정책, 계절성, 사회적 사건 등 외부 요인이 고려되었는가? - 라벨 생성 규칙 및 목표 함수 변화 여부가 검토되었는가? - 특정 지역·시간대에 국소적 드리프트가 발생했는가?
⑥ 대응 전략 수립	데이터/모델 보정, 재학습	- 샘플 불균형 보정, 데이터 증강, 라벨 정제가 이루어졌는가? - 모델 재학습 또는 Feature set 재구성이 수행되었는가? - Offline 재검증 및 Online A/B Test가 병행되었는가?

단계	주요 점검 항목	체크리스트 질문
⑦ 피드백 및 개선	규칙 재설계, 자동화, 순환 개선	- 탐지 임계값 및 기준 데이터가 조정·갱신되었는가? - 모니터링-탐지-재학습-배포 파이프라인이 자동화되었는가? - 개선된 규칙과 모델이 다시 모니터링에 반영되었는가? - 정책적 피드백(복지 대상 기준, 서비스 공급 방식 등)에 활용되었는가?

자료: 저자가 작성한 내용을 open AI로 보완하고 저자가 연구목적에 맞게 재구성하여 작성.

이 7단계는 ① 수집 → ② 기준선 → ③ 모니터링 → ④ 탐지 → ⑤ 원인 분석 → ⑥ 대응 전략 수립 → ⑦ 피드백 순환 구조로, 단순히 “탐지”에서 끝나는 것이 아니라 원인 규명-성능 보정-지속적 개선-정책 반영까지 확장된 관리 체계라고 볼 수 있다.

이러한 드리프트 탐지 체계를 효과적으로 운영하기 위해서는 정기적인 모델 성능평가와 함께 탄력적인 정책 조정 메커니즘이 필요하다. 특히 보건복지분야의 특수성을 고려할 때, 실시간 모니터링과 신속한 대응이 가능한 적응형 프로그램 설계가 중요하다.

향후 부처 간 데이터 통합 관리체계가 고도화되면서, 이러한 드리프트 탐지 시스템은 더욱 중요해질 것으로 예상된다. 이는 단순한 기술적 과제를 넘어, 보다 효과적이고 공정한 보건복지 서비스 제공을 위한 핵심 인프라로 자리잡을 것이다.

제2절 데이터 드리프트 유형별 모니터링 방안

앞서 3장에서 특성 드리프트, 공변량 드리프트로 가구구성 변화에 따른 데이터 드리프트 분석, 라벨 드리프트로 고혈압 기준 변화에 따른 데이터 드리프트 분석, 모델 드리프트로 복지사각지대 데이터 드리프트 분석을 실시하였다. 이 절에서는 3장에서 다룬 데이터 드리프트 유형별 구체적 사례로 데이터 드리프트 관리 방안 프로세스 7단계를 제시하고자 한다.

1. 특성 드리프트 모니터링 방안

가구구성 변화는 보건복지 정책 수립 및 사회지표 산출 과정에서 직접적으로 연관되는 중요한 요인이며, 그 변화 정도를 정량적으로 평가하기 위해 PSI, KS 지표 등이 활용된다. 지표 결과값은 기준연도를 어느 시점으로 설정하느냐에 따라 변화 정도가 달라진다. 3장 1절에서 분석한 특성 드리프트를 7단계 데이터 드리프트 관리방안 프로세스로 제시하면 다음과 같다.

① 데이터 수집 및 전처리

인구총조사 등에서 확보한 데이터를 기반으로, 결측치·이상치 제거와 스키마 검증을 수행한다. 특성 드리프트는 변수 정의 변경(예: 가구원수 범주의 재구성)을 통해, 공변량 드리프트는 변수 간 상관구조 변화를 통해 탐지할 수 있다.

② 기준선 설정

각 기준연도를 Reference Distribution으로 설정한다. 특성 드리프트는 PSI, KS 지표로 단일 변수 기준 분포를 정의(예 PSI ≥ 0.25, KS ≥ 0.2)하고, 공변량 드리프트는 다변량 상관관계를 기준으로 설정하여 변화 여부를 평가한다.

③ 모니터링

중앙 대시보드에 기준연도별 PSI·KS 변화 추이를 시계열로 시각화한다. 특성 드리프트는 가구원수·세대 구성 분포 변화를 추적하고, 공변량 드리프트는 가구구성 변화가 소득·연령·주거지표 등에 미치는 관계 변화를 실시간 점검한다.

④ 드리프트 탐지

PSI ≥ 0.25, KS ≥ 0.2를 기준으로 알림을 발생시킨다. 가구구성 변화는 단일 변수로 분석하였지만, 변수가 많을 경우 특성 드리프트는 분포가 급격히 변한 변수 식별에 초점을 맞추고, 공변량 드리프트는 변수 간 상관구조 붕괴나 새로운 관계의 출현을 탐지한다.

⑤ 원인 분석

변화 원인을 내부 요인(데이터 수집 방식 변경)과 외부 요인(고령화, 핵가족화 등 사회구조적 변화)으로 구분한다. 특성 드리프트는 변수 자체의 정의·범위 변경 여부를 검증하고, 공변량 드리프트는 외부환경 요인에 따른 다변량 관계 변화를 분석한다.

⑥ 대응 전략 수립

드리프트가 발생하면 모델 재학습, Feature Engineering 갱신, 데이터 증강 등을 수행한다. 특성 드리프트는 기준 데이터셋을 최신 연도로 보정하는 방식으로 대응하고, 공변량 드리프트는 변수 간 관계 구조를 반영하는 새로운 모델링 기법을 적용한다.

⑦ 피드백 및 개선

드리프트 분석 결과를 복지정책 설계에 반영한다. 특성 드리프트는 정책 단위 기준(예: 4인 가구 기준 → 3인 가구 기준) 변경의 근거를 제공하고, 공변량 드리프트는 장기적 사회구조 변화에 따라 복지정책·주거정책·기준중위소득 산정 방식을 조정하는 데 활용된다.

〈표 4-2〉 특성 및 공변량 드리프트 관리 7단계 프로세스

단계	주요 점검 항목	특성 드리프트 대응	공변량 드리프트 대응
① 데이터 수집 및 전처리	데이터 확보, 품질 관리, 스키마 검증	가구원수, 세대 구성 변수 정의의 일관성 점검	변수 간 관계성(가구규모-연령, 가구규모-소득) 변화 감지
② 기준선 설정	기준 데이터셋 정의, 임계값 설정	PSI, KS 기준으로 변수별 기준선 구축	다변량 상관구조를 기준으로 설정
③ 모니터링	대시보드 시각화, 시계열 추적	가구원수·세대 구성의 분포 추세 모니터링	소득·연령·주거지표와의 관계 변화 모니터링
④ 드리프트 탐지	PSI, KS, Chi-Square 등 적용	기준선 대비 급격한 분포 변화 탐지	상관구조 붕괴, 새로운 패턴 등장 탐지
⑤ 원인 분석	내부/외부 요인 검토	변수 정의 변경, 조사방식 영향	사회구조적 변화(고령화, 핵가족화 등) 영향
⑥ 대응 전략 수립	데이터/모델 보정, 재학습	기준연도 갱신, Feature 보정	다변량 구조 반영, 관계 기반 모델링 개선
⑦ 피드백 및 개선	규칙 재설계, 정책 반영	가구단위 기준 재설계 근거 제공	복지·주거정책, 기준중위소득 산정 방식 개선

자료: 저자 작성.

2. 라벨 드리프트 모니터링 방안

라벨 드리프트는 예측 모델링과 정책 설계 전반에 영향을 준다. 미국의 2017년 고혈압 진단 기준 변경을 국민건강영양조사 9기 데이터로 분석한 결과 환자군의 규모가 크게 달라졌음을 알 수 있다. 로지스틱 모형 적용 결과, 라벨 정의 변경에 따라 독립변수와 종속변수 간의 관계가 달라졌음을 확인하였다. 3장 2절에서 분석한 라벨 드리프트를 7단계 데이터 드리프트 관리방안 프로세스로 제시하면 다음과 같다.

① 데이터 수집 및 전처리

국민건강영양조사 등에서 확보한 데이터를 기반으로, 결측치·이상치 제거와 스키마 검증을 수행한다. 라벨 생성 규칙이 바뀌면 Y값의 정의가 달라지기 때문에 데이터의 사전/코드북에 규칙 변경 이력을 로그로 남기며 버전 관리를 하도록 한다.

② 기준선 설정

기준 데이터셋과 임계값을 정의한다. 성능 기준이 되는 측도를 설정, Y1 기준선(기존), Y2 기준선(변경)으로 두 개의 기준선에 대한 관리를 한다.

③ 모니터링

라벨 기준 변경 전후의 유병률을 시계열적으로 모니터링하면 실제 질병 발생 증가와 단순 분류 기준 변경을 구분할 수 있다. 대시보드에서는 Y1과 Y2를 함께 제시하여 기준값 변경이 환자 수, 인구집단별 유병률에 미치는 영향을 실시간으로 확인하도록 한다.

④ 드리프트 탐지

라벨 정의 변경에 따른 양성률 급증과 계수/오즈비 변동을 드리프트 신호로 사용할 수 있다. 동일 데이터셋에서 로지스틱 회귀 계수 비교(모형1 vs. 모형2)하고, 가능도비 검정 결과로 드리프트를 탐지한다.

⑤ 원인 분석

변화 원인을 내부 요인(데이터 수집 방식 변경)과 외부 요인(가이드라인 및 정책 변화)으로 구분한다. 여기에서는 기준 변경의 배경과 의학적 정당성을 함께 검토해야 하며, 라벨 규칙 변경(진단 임계값 하향)이 독립변수–종속변수 관계를 어떻게 재구성하였는지에 대해 분석한다.

⑥ 대응 전략 수립

라벨 드리프트 발생 시, Y2 기준으로 재학습한 모형 2를 운영에 반영하도록 한다. 변수 가중치를 재산정하고, 필요시 클래스 가중 조정 작업도 검토한다. 이전 모형 1은 레퍼런스로 유지하고 점진적 학습으로 전환 비용을 최소화할 수도 있다.

⑦ 피드백 및 개선

기준 변경 시 자동 경보 시스템을 통해 연구자와 정책결정자에게 영향을 알리고, 복지·보건 지표 설계 시 기준 변화의 효과를 명확히 구분하여 설명하는 체계가 필요하다. 예를 들어 고혈압 환자 증가가 실제 질환 악화 때문인지, 진단 기준 완화 때문인지를 명확히 구분하여 피드백 및 개선 방안이 도출되어야 한다. 기준 변경 효과를 하나의 주기로 간주하여 재학습 주기/알림 임계값을 사후 조정하도록 하며 규칙을 재설계한다.

〈표 4-3〉 라벨 드리프트 관리 7단계 프로세스

단계	주요 점검 항목	특성 드리프트 대응
① 데이터 수집 및 전처리	데이터 확보, 품질 관리, 스키마 검증	라벨 생성 규칙(≥140/90, ≥130/80)을 구분하여 명확히 버전 관리. Y1, Y2 파생 변수 생성 및 로그 기록
② 기준선 설정	기준 데이터셋 정의, 임계값 설정	Y1과 Y2 기준선을 이중으로 관리. 양성률 및 성능 지표를 각각 저장
③ 모니터링	대시보드 시각화, 시계열 추적	Y1과 Y2의 분류율, 절편, 변수별 유의성 변화를 대시보드로 시각화
④ 드리프트 탐지	다중 지표 자동 탐지	모형 1과 모형 2의 회귀계수, 오즈비 차이 검정. 라벨 양성률 급변을 탐지 기준으로 활용
⑤ 원인 분석	내부/외부 요인 검토	진단 기준 변경(의학 가이드라인)이 외부 요인임을 규정. 변수군별 영향 변화 분석
⑥ 대응 전략 수립	데이터/모델 보정, 재학습	Y2 기준으로 재학습하여 변수 가중치 재산정. 점진적 학습 적용
⑦ 피드백 및 개선	규칙 재설계, 정책 반영	현상 변화에서 기준 변경 효과를 명확히 구분. 환자 수 변화 및 정책 우선순위 변화 보고

자료: 저자 작성.

3. 모델 드리프트 모니터링 방안

복지사각지대 발굴 시스템은 다양한 데이터의 수집과 2개월 주기의 운영, 24종의 모형이 활용되고 있는 복잡한 체계로 구성되어 있다. 복지사각지대 발굴모형이 안정적으로 운영되기 위해서는 데이터 드리프트와 예측 드리프트를 함께 관리할 수 있는 통합 모니터링 체계가 필수적이다. 3장 3절에서 분석한 모델 드리프트를 7단계 데이터 드리프트 관리방안 프로세스로 제시하면 다음과 같다.

① 데이터 수집 및 전처리

복지사각지대 발굴모형은 다양한 외부기관으로부터 데이터를 주기적으로 수집하기 때문에, 이 과정에서 스키마 일관성 검증(변수명·타입·단위)과 데이터 품질 관리(결측·이상치 처리)가 반드시 수행되어야 한다. 특히

변수별 분포가 기관별·시점별로 다르게 나타날 수 있으므로, 데이터 입수 즉시 데이터 드리프트 탐지 절차가 자동 실행되도록 설계해야 한다. 이를 통해 입력단계에서 발생할 수 있는 구조적 오류와 초기 드리프트를 사전에 차단한다.

② 기준선 설정

복지사각지대 발굴모형에서는 (1) 데이터 입력단계의 기준선(변수 분포)과 (2) 출력단계의 기준선(위기도 확률 분포)을 각각 설정해야 한다. PSI, KS, Chi-Square 등 지표별 임계값을 명확히 정의하고, 발굴 차수별로 관리하여, 차수 간 비교가 가능하게 해야 한다. 이를 통해 새로운 데이터가 들어왔을 때 기준선 대비 변화 정도를 객관적으로 측정할 수 있도록 한다.

③ 모니터링

중앙집중식 대시보드를 통해 데이터와 예측 결과의 상태를 상시 모니터링한다. 대시보드에는 전체 모델 상태 요약, 단계별 드리프트 현황 및 시계열 추이, 각 단계 간 상관관계 분석 결과가 포함되어야 한다. 사용자의 역할에 따라 요약 현황을 제공하거나, 필요시 상세 분석으로 다운이 가능하도록 설계한다. 특히 위기도 확률값은 핵심 지표이므로, 가능한 한 실시간에 가까운 수준으로 변화 감지가 필요하다.

④ 드리프트 탐지

드리프트 탐지는 입력단계와 출력단계로 구분해 구축한다. 입력단계는 외부기관 데이터의 변수별 분포 변화를 탐지하여 데이터 드리프트를 식별한다. 출력단계는 발굴모형이 산출하는 위기도 예측 결과의 분포 변화

를 모니터링하여 예측 드리프트를 탐지한다.
각 탐지 결과는 상호 연계 분석을 수행해 드리프트 발생 경로와 영향 범위를 추적할 수 있어야 한다. PSI 기준으로 0.1 미만은 정상, 0.1~0.25 미만은 주의, 0.25 이상은 위험으로 구분하여 단계별 알림을 제공한다.

⑤ 원인 분석

드리프트가 발생했을 때는 내부 요인(데이터 파이프라인 오류, 변수 정의 변경)과 외부 요인(사회환경 변화, 정책 변경)을 구분하여 원인을 분석한다. 예를 들어 국가재난상황, 경기침체, 팬데믹, 복지정책 변경과 같은 외부 사건이 특정 시점에 집중적으로 드리프트를 유발할 수 있다. 또한 데이터 입수 주기 차이(월별, 분기별, 연 단위)가 드리프트에 영향을 줄 수 있으므로, 각 기관별 주기 특성을 고려한 원인 분석이 필요하다.

⑥ 대응 전략 수립

드리프트 감지 심각도별 차등 알림 체계를 구축한다. PSI 기준 0.1 미만은 정상 상태로 별도 알림 없이 발생 로그만 기록하고, 0.1 이상 0.25 미만은 주의 단계로 담당자에게 알림을 발송하며, 0.25 이상은 위험 단계로 즉시 SMS 및 전화 알림을 통해 긴급 대응을 요청해야 한다. 알림 발생 시에는 드리프트 정도, 영향변수, 예상 원인, 권장 대응 방안 등의 정보를 포괄적으로 제공하고, 신속한 의사결정을 지원하도록 한다.

또한 머신러닝 기반의 이상 탐지 모델을 활용하여 드리프트 발생을 사전에 예측하는 조기 경보 시스템을 구축한다. 과거 드리프트 발생 패턴과 외부 환경 변화 간의 상관관계를 학습하여 드리프트 발생 가능성이 높은 시점을 미리 식별하고, 예방적 조치를 취할 수 있도록 한다.

⑦ 피드백 및 개선

발굴모형은 차수별로 운영이 되므로, 차수별 발굴모형의 드리프트 분석 보고서를 자동 생성하고, 트렌드 분석과 예측, 개선 권고사항 등을 포함하여 지속적인 모델 관리가 가능하도록 한다. 이 단계에서는 드리프트 관리 체계를 모델 생명주기 관리와 연계한다. 드리프트 탐지 기록을 축적하여 주기적인 패턴을 분석하고, 이를 바탕으로 재훈련 주기를 최적화한다. 또한 드리프트에 강건한 모델 아키텍처(예: 앙상블, 적응형 학습모형)로 개선 방향을 도출하도록 한다.

〈표 4-4〉 모델 드리프트 관리 7단계 프로세스

단계	주요 점검 항목	모델 드리프트 대응
① 데이터 수집 및 전처리	데이터 확보, 품질 관리, 스키마 검증	입력데이터 분포 변화 자동 탐지, 초기 드리프트 차단
② 기준선 설정	기준 데이터셋 정의, 임계값 설정	입력/출력 단계별 기준선 설정 및 임계값 관리
③ 모니터링	대시보드 시각화, 시계열 추적	모델 상태 요약, 시계열 추이, 상관관계 분석 제공
④ 드리프트 탐지	다중 지표 자동 탐지	입력단계(데이터)와 출력단계(예측) 분포 변화를 구분 탐지
⑤ 원인 분석	내부/외부 요인 검토	사회환경 변화·정책 변경과 데이터 파이프라인 오류 구분
⑥ 대응 전략 수립	데이터/모델 보정, 재학습	재학습, 점진적 학습 적용
⑦ 피드백 및 개선	규칙 재설계, 정책 반영	조기 경보 시스템 구축, 재훈련 주기 최적화, 강건한 모델 아키텍처 개선

자료: 저자 작성.

제5장

결론 및 시사점

제 5 장 결론 및 시사점

보건복지분야의 데이터 드리프트 현상은 코로나19로 인한 의료서비스 수요-공급 관계 변화, 장애인 등록 기준 변화에 따른 분포 변화, 질병 진단 기준 변경, 복지 수급 자격 기준 변경 등 다양하게 나타나고 있다. 전에도 이러한 변화는 지속적으로 발생해왔지만, 현재 데이터 드리프트의 탐지와 관리·모니터링 방안에 특히 주목해야 하는 이유는 빠르게 변하고 있는 사회 현상과 폭발적으로 축적되는 데이터의 양, 분석 기술의 발전이라고 생각한다. 보건복지분야에서 데이터 분포 변화가 예측 성능 및 업무 효율에 미치는 영향은 막대하며, 이를 탐지하고 탐지 측도로 정도를 파악하여 향후 의사결정의 기초자료로 활용하는 것은 정책의 신뢰성을 확보하는 주요 수단이 될 수 있다. 즉, 보건복지분야에서 데이터 드리프트 현상은 정책 수립과 서비스 제공에 있어 핵심적인 도전 과제로 대두되고 있다. 따라서 본 과제를 통해 데이터 드리프트의 대응 기술 및 방법론 정리, 데이터 드리프트 유형별 데이터 분석, 데이터 드리프트 관리 프로세스를 정립하고자 하였다.

이 연구에서는 데이터 드리프트의 주요 유형을 특성 드리프트, 개념 드리프트, 사전확률 드리프트, 라벨 드리프트, 기타(파이프 라인, 샘플링, 시간적, 모델 드리프트)로 정리하였다. 그리고 각 유형별로 적용 가능한 통계적 지표(PSI, KS, KL, JS, IV 등), 시계열 기반(Page-Hinkley, EWMA, DDM, EDDM 등) 및 머신러닝/딥러닝 기반 알고리즘(LSTMDD, ADWIN, Page-Hinkley 등)의 특징과 원리, 파이썬(Python)/R 도구 활용 방법을 구

체적으로 제시하였다. 각 데이터 드리프트 탐지 방법들은 민감도와 적합성이 다르기 때문에, 여러 지표를 함께 사용해야 신뢰성을 확보할 수 있다.

보건복지분야 데이터 드리프트 시뮬레이션에서는 드리프트 유형별로 데이터 분석 결과와 이슈를 짚어보았다. 가구의 가구원수 변화는 특성 드리프트에 속하며, 기준연도를 변화시켜가며 PSI·IV·KS 분석으로 시계열 변동을 살펴보았다. 2017년 미국의 고혈압 기준 변화(140/90mmHg → 130/80mmHg)는 우리나라에도 영향을 미치는 바, 이는 라벨 드리프트 사례에 속한다. 국민건강영양조사 데이터로 고협압 기준을 변화시켜 데이터 분석을 실시하였고 라벨이 변경되었을 때 주요 요인도 바뀜을 확인하였다. 그리고 복지사각지대 데이터는 가상의 데이터를 생성하여 PSI·JSD를 적용하였고, 위기 정보별 대상자 비율 변화, 데이터 드리프트 발생 변수를 식별하고, 예측 성능 저하 가능성을 확인하는 등 모델 드리프트의 사례를 제시하였다.

데이터 드리프트 관리를 위한 방안은 7단계로 다음과 같이 제시하였다. ① 데이터 수집 및 전처리(데이터 확보, 품질 관리, 스키마 검증), ② 기준선 설정(기준 데이터셋 정의, 임계값 설정), ③ 모니터링(데이터·모델 성능 추적, 알림 시스템), ④ 드리프트 탐지(지표 기반 탐지, 다중 방법 적용), ⑤ 원인 분석(내부/외부 요인 파악, 모델 영향 분석), ⑥ 대응 전략 수립(데이터/모델 보정, 재학습), ⑦ 피드백 및 개선(규칙 재설계, 자동화, 순환 개선)으로 관리 방안 단계를 세분화 하였다. 각 단계별로 주요 점검 항목과 체크리스트 질문을 제시하여 실무에 적용할 수 있도록 하였다. 또한 3장에서 분석한 특성 드리프트, 라벨 드리프트, 모델 드리프트 유형에 7단계 프로세스를 적용하여 유형별 모니터링 방안을 제시하였다. 이러한 정교화된 절차는 실제 운영 환경에서도 안정적이고 일관된 모델 관리를 가능하게 한다.

본 연구는 기존 연구의 한계였던 기술적 방법론 중심 접근과 차별화하여, 데이터 드리프트 대응 기술과 최신 방법론을 정리하고, 실제 공공 데이터를 활용한 사례분석과 실무 적용 가능한 7단계 관리 프로세스를 제시하였다는 점에서 의의가 있다.

본 연구를 통해 도출된 주요 시사점과 향후 발전 방향을 다음과 같이 제시하고자 한다.

첫째, 데이터 드리프트의 효과적인 탐지와 관리를 위해서는 다층적 접근이 필요하다. 특성 드리프트, 사전확률 드리프트, 개념 드리프트, 파이프라인 드리프트 등 다양한 유형의 드리프트에 대해 각각 적합한 탐지 방법을 적용해야 한다. 특히 머신러닝, AI 기반 등 검증된 최신 기술을 활용한 모니터링 체계의 구축이 시급하다.

둘째, 보건복지분야의 특수성을 고려한 맞춤형 드리프트 관리 전략이 필요하다. 보건복지분야에서는 정책 개편, 자격 기준 변경, 신규 제도 도입 등이 빈번하게 발생하며, 이러한 제도적 변화는 데이터 분포에 직접적인 영향을 미친다. 예를 들어, 특정 복지급여의 수급 기준이 완화되면 해당 변수의 유효비율이 증가하고, 이는 드리프트로 탐지될 수 있다. 드리프트 탐지 결과를 해석할 경우 제도적 변화에 기인한 것인지, 데이터 품질 이슈에 기인한 것인지를 구분하여 접근할 필요가 있으며, 기술 지표(PSI, KS) 외에 수급자 수 증감률, 예산 소요 변화율, 정책 타깃 집단 변동률과 같은 정책 영향 지표를 드리프트 관리 프로세스에 통합할 필요가 있다. 데이터 프라이버시와 보안 문제에 대한 철저한 대비도 필요하다. 특히 의료정보와 같은 민감한 개인정보를 다루는 만큼, 데이터 보호와 활용의 균형을 맞추는 것이 중요하다. 이를 위해 한국 보건의료 환경에 적합한 데이터 드리프트 관리 방안도 필요하다. 데이터 드리프트 관리 방안은 분야의 특수성과 드리프트 유형을 고려하여 차별화되어야 한다. 추가적

으로 시나리오 기반 실증적 대응 방안을 미리 계획하여 실제 현장에서 발생 가능한 구체적인 상황에 대한 예측을 고려할 수 있도록 할 수 있다. 코로나19와 같은 이벤트의 경우 여러 종류의 드리프트가 동시다발적으로 발생하기 때문에, 통계적 탐지와 더불어 모니터링 주기 단축, 룰 베이스 일시 전환 등의 계획이 제시될 수 있다.

셋째, 인간 중심의 접근(Human-in-the-loop)과 기술적 솔루션의 조화가 중요하다. 지속적 학습과 모니터링을 통해 시스템을 개선하되, 전문가의 판단과 개입이 적절히 이루어져야 한다. 지표마다 민감도와 스케일이 다르기 때문에 지표 간 결과가 상충되었을 때 실무자가 어떻게 판단해야 하는지에 대한 기준도 필요하다. 실무자가 명확한 판단을 할 수 있는 구체적인 지표 결합 및 의사결정 로직도 전문가의 판단으로 이루어져야 할 것이다.

마지막으로, 정책적 차원에서 견고한 프레임워크 구축과 탄력적인 서비스 전달 모델의 개발이 요구된다. 데이터 품질관리 시스템 구축, 정기적 모델 성능평가, 그리고 적응형 프로그램 설계를 통해 변화하는 환경에 효과적으로 대응해야 한다.

이러한 종합적인 접근을 통해 보건복지분야의 데이터 드리프트 관리는 더욱 고도화될 수 있으며, 이는 궁극적으로 더 나은 보건복지서비스 제공과 정책 효과성 제고로 이어질 것으로 기대된다.

참고문헌

강신욱, 노대명, 이소정, 양난주, 김근혜. (2016). 저소득층의 소득-자산분포를 통해 본 사회보장제도 재산기준의 개선 방향. 한국보건사회연구원.

강현우, 남덕윤. (2025). 모델 잡음 강건성 향상을 위한 Spiking Neuron 적용 가능성 고찰. 정보과학회 컴퓨팅의 실제 논문지, 31(5), 234-240.

김은하. (2022). 빅데이터 정보시스템 활용 현황과 과제: 복지 사각지대 발굴 시스템을 중심으로. 보건복지포럼, (313), 24-34. https://doi.org/10.23062/2022.11.3.

나경민, 김도형, 이영호. (2025). VROFed: 부분 분산 감소를 이용한 최적화된 공정한 연합학습. 한국정보통신학회논문지, 29(1), 34-44.

보건복지부. (2023). 위기정보 44종으로 확대, 복지사각지대 발굴 촘촘해진다. 보도자료(4.24).

보건복지부. (2024). 45종 위기정보 활용해 2024년 2차 복지사각지대 발굴 시행. 보도자료(3.25).

유호범. (2022). AI 기반의 공공재정 FDS 신속대응·탐지체계 구축 사례. 한국행정학회 동계학술발표논문집, 2022, 1217-1228.

이상연, 조은성, 전성현, 홍석철. (2023). 국민건강보험 빅데이터를 활용한 주요 만성질환 발생률 예측모형의 개발과 활용. 보험학회지, 133, 23-48.

이예은, 이태진. (2023). MLOps를 위한 효율적인 AI 모델 드리프트 탐지방안 연구. Journal of Internet Computing & Services, 24(5).

이우식, 김선월, 최솔지, 이인수, 조아라, 강동욱. (2021). 복지사각지대 발굴체계 재정립 연구 -운영 프로세스 중심으로. 한국사회보장정보원.

이우식, 김인수, 최솔지, 박규범, 황진섭. (2020). 복지사각지대 발굴관리시스템 예측모형 개선 방안 연구. 한국사회보장정보원.

이정욱. (2021). 딥 러닝을 이용한 단일 카메라 SLAM에서 스케일 드리프트 감소. 제어로봇시스템학회 논문지, 27(8), 518-527.

질병관리청. (2024). 국민건강영양조사 제9기 1·2차년도(2022-2023). 원시자료 이용지침서. 오송: 질병관리청. https://knhanes.kdca.go.kr/.

최선. (2022.5.12.). 한국도 고혈압 기준 강화 동참…변경 이유는?. 메디컬타임즈. https://www.medicaltimes.com/Main/News/NewsView.html?ID=1147283(검색일: 2025.9.9.).

최옥주, 김유경. (2024). 머신러닝 드리프트에 대한 2-단계 데이터 품질 평가 방법론. 한국소프트웨어감정평가학회논문지, 20(1), 75-85.

통계청, (2015).「가계금융복지조사결과」2015년 보도자료 부록 통계표. 대전: 통계청(현 국가데이터처).

통계청. (2024).「가계금융복지조사결과」2024년 보도자료 부록 통계표. 대전: 통계청(현 국가데이터처).

통계청. (1970-2017).「인구총조사」. 대전: 통계청(현 국가데이터처).

한국사회보장정보원. (2025). 복지사각지대발굴 소개. http://ssis.or.kr(검색일: 2025.8.15).

Ali, U. & Mahmood, T. (2024). A novel framework for concept drift detection using autoencoders for classification problems in data streams. https://www.qeios.com/read/ZU17S4/pdf.

Baena-García, M. et al. (2006). Early drift detection method. 4th ECML PKDD International Workshop.

Baier, L., Schlör, T., Schöffer, J. & Kühl, N. (2021). Detecting concept drift with neural network model uncertainty. arXiv preprint arXiv:2107.01873.

Basseville, M. & Nikiforov, I. V. (1993). Detection of abrupt changes: theory and application. Prentice Hall.

Bayram, F., Ahmed, B.S. & Kassler, A. (2022). From concept drift to model degradation: An overview on performance-aware drift detectors. Knowledge-Based Systems, 245, p.108632.

Bifet, A. & Gavalda, R. (2007). Learning from Time-Changing Data with Adaptive Windowing. SDM, 7, 443-448.

Bifet, A. & Gavalda, R. (2009). Adaptive learning from evolving data streams. IDA 2009.

Cai, S., Zhao, Y., Hu, Y., Wu, J., Wu, J., Zhang, G., & Sosu, R. N. A. (2024). CD-BTMSE: A concept drift detection model based on bidirectional temporal convolutional network and multi-stacking ensemble learning. Knowledge-Based Systems, 294, 111681.

Candela, J. Q., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). Dataset shift in machine learning. The MIT Press, 1, 5.

Crowder, S. V. (1987). A simple method for studying run-length distributions of exponentially weighted moving average charts. Technometrics, 29(4), 401-407.

Dhinakaran, A. (2023). Step-by-Step Tutorial on Jensen-Shannon Divergence and Applications in Machine Learning(검색일: 2025.8.15).

Duckworth, C., Chmiel, F. P., Burns, D. K., Zlatev, Z. D., White, N. M., Daniels, T. W., Michael Kiuber, & Boniface, M. J. (2021). Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. Scientific reports, 11(1), 23017.

Frías-Blanco, I., et al. (2015). Online and non-parametric drift detection methods based on Hoeffding's bounds. IEEE Transactions on Knowledge and Data Engineering, 27(3), 810-823.

Gal, Y. & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. ICML.

Gama, J., Medas, P., Castillo, G., & Rodrigues, P. (2004). Learning with drift detection. SBIA 2004.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM computing surveys (CSUR), 46(4), 1-37.

Garcia Moreno-Torres, J. (2013). Dataset shift in classification: terminology, benchmarks and methods. Universidad de Granada.

Greco, S., Vacchetti, B., Apiletti, D. & Cerquitelli, T. (2024). Unsupervised Concept Drift Detection from Deep Learning Representations in Real-time. arXiv preprint arXiv:2406.17813.

Hinkley, D. V. (1971). Inference about the change-point from cumulative sum tests. Biometrika, 58(3), 509-523.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association, 58(301), 13-30.

Hu, L., Lu, Y. & Feng, Y. (2025). Concept Drift Detection Based on Deep Neural Networks and Autoencoders. Applied Sciences, 15(6), 3056.

Hunter, J. S. (1986). The exponentially weighted moving average. Journal of Quality Technology, 18(4), 203-210.

Jayita Gulati. (April 17, 2025). in Practical Machine Learning, Detecting & Handling Data Drift in Production. https://machinelearning mastery.com/detecting-handling-data-drift-in-production/.

Kang, T., Lee, Y., & Kang, M. (2024). Impact of COVID-19 on healthcare utilization among chronic disease patients in South Korea. Preventive Medicine Reports, 41, 102680.

Kim, S., Sung, H. K., Lee, J., Ko, E., & Kim, S. J. (2024). Trends in emergency department visits for emergency care-sensitive conditions before and during the COVID-19 pandemic: a nationwide study in Korea, 2019–2021. Clinical and Experimental

Emergency Medicine, 11(1), 88.

Kore, A., Abbasi Bavil, E., Subasri, V., Abdalla, M., Fine, B., Dolatabadi, E., & Abdalla, M. (2024). Empirical data drift detection experiments on real-world medical imaging data. Nature communications, 15(1), 1887.

Kull, M. & Flach, P. (2014, September). Patterns of dataset shift. In First international workshop on learning over multiple contexts (LMCE) at ECML-PKDD (Vol. 5).

Kullback, S. & Leibler, R. A. (1951). On Information and Sufficiency. The Annals of Mathematical Statistics, 22(1), 79-86.

Kwon, Y. S. & Baek, M. S. (2020). Development and validation of a quick sepsis-related organ failure assessment-based machine-learning model for mortality prediction in patients with suspected infection in the emergency department. Journal of clinical medicine, 9(3), 875.

Li, J., et al. (2023). Autoencoder-based Anomaly Detection in Streaming Data with Incremental Learning and Concept Drift Adaptation.

Lindgren, B. E. S. (1991). Some Properties of Jensen-Shannon Divergence and Mutual Information.

Lucas, J. M. & Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: properties and enhancements. Technometrics, 32(1), 1-12.

Hu, L., Lu, Y. & Feng, Y. (2025). Concept Drift Detection Based on Deep Neural Networks and Autoencoders. Applied Sciences, 15(6), 3056.

Massey Jr, F.J. (1951). The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association, 46(253), 68-78.

Mehmood, T., Latif, S., Jamail, N. S. M., Malik, A., & Latif, R. (2024). LSTMDD: an optimized LSTM-based drift detector for concept drift in dynamic cloud computing. PeerJ Computer Science, 10, e1827.

Montgomery, D. C. (2009). Introduction to Statistical Quality Control, 6th Edition. Wiley.

Mouss, H., Mouss, D., Mouss, N., & Sefouhi, L. (2004). Test of Page-Hinkley, an approach for fault detection in an agro-alimentary production system. IEEE Conference on Control Applications.

Neely, J.G., Hartman, J.M., Forsen Jr, J.W. & Wallace, M.S. (2003). Tutorials in clinical research: Ⅷ. Understanding comparative statistics (contrast)—part B: Application of T-test, Mann-Whitney U, and Chi-Square. The Laryngoscope, 113(10), 1719-1725.

Page, E. S. (1954). Continuous Inspection Schemes. Biometrika, 41(1/2), 100-115.

Park, S. W., Yeo, N. Y., Kang, S., Ha, T., Kim, T. H., Lee, D., ... & Heo, J. (2024). Early prediction of mortality for septic patients visiting emergency room based on explainable machine learning: a real-world multicenter study. Journal of Korean Medical Science, 39(5).

Priya, S. & Uthra, R.A. (2023). Deep learning framework for handling concept drift and class imbalanced complex decision-making on streaming data. Complex & Intelligent Systems, 9(4), 3499-3515.

Rabanser, S., Günnemann, S. & Lipton, Z. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. Advances in Neural Information Processing Systems, 32.

Roberts, S. W. (1959). Control Chart Tests Based on Geometric Moving Averages. Technometrics, 1(3), 239-250.

Sakurai, G.Y., Lopes, J.F., Zarpelão, B.B. & Barbon Junior, S. (2023). Benchmarking change detector algorithms from different concept drift perspectives. Future Internet, 15(5), p.169.

Salem, M., Buschjaeger, S., & Morik, K. (2012). Anomaly detection in network traffic using Jensen-Shannon divergence. Computer Communications, 50, 33-44.

Webb, G.I., Hyde, R., Cao, H., Nguyen, H.L. & Petitjean, F. (2016). Characterizing concept drift. Data Mining and Knowledge Discovery, 30(4), 964-994.

WHO. (2021). Global strategy on digital health 2020-2025.

Yu, H., Zhang, Q., Liu, T., Lu, J., Wen, Y., & Zhang, G. (2022). Meta-ADD: A meta-learning based pre-trained model for concept drift active detection. Information Sciences, 608, 996-1009.

Zeiler, M.D. (2012). Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.

Žliobaitė, I. (2010). Learning under concept drift: an overview. arXiv preprint arXiv:1010.4784.

Abstract

Study on Cases and Management Strategies of Data Drift in the Health and Welfare Sector

Project Head: Oh, Miae

Data drift refers to the phenomenon where the statistical properties of the data used to train machine learning models shift over time. In the health and welfare sector, this is becoming increasingly prominent, driven by rapid demographic transitions and the continuous expansion of welfare service eligibility. The necessity for robust detection, management, and monitoring of data drift arises from the growing uncertainty and risks in data-driven decision-making, which are exacerbated by the complex interplay of evolving social trends, the exponential accumulation of data, and advancements in analytical technologies.

This study examines the unique characteristics of various types of data drift and provides a systematic evaluation of the strengths and weaknesses of different detection methodologies. Furthermore, through simulations utilizing public administrative datasets, the research proposes practical measures to enhance the reliability and sustainability of data-driven governance within the health and welfare domain.

Key words: Health and Welfare, Data Drift, Technology, Monitoring Strategies

Co-Researchers: Lee, Jeongran·An, Suin