컴퓨터교육학회 논문지 2026년 제29권 제3호
https://doi.org/10.32431/kace.2026.29.3.005

다변량 민감 속성 기반 반사실적 공정성 평가와 XAI 기법을 활용한 대학생 학업성취도 예측의 편향 분석

Counterfactual Fairness and XAI-based Bias Attribution in Academic Performance Prediction Using Multivariate Sensitive Features

문동수†
Dongsoo Moon†

요약

본 연구는 기존 선행연구와 차별화하여, 반사실적 조합(counterfactual instances)을 기반으로 예측 모델의 공정성을 정량적으로 평가하고, 그 원인을 민감 속성 중심의 설명 가능한기여도 분석을 통해 식별하고자 하였다. 특히, 단순히 예측 결과의 불평등을 측정하는 데 그치지 않고, 모델이 특정 민감 속성 변화에 따라 예측을 반전(Flip)시키는지를 분석하고, 그 경계 근처에서 민감 속성이 어떤 기여를 하는지를 Counterfactual SHAP 및 iBreakDown 기반 설명 가능한 인공지능 기법으로 시각화 및 설명하였다. 이는 고등교육 정책 수립 및 조기 이탈 대응에 있어 데이터 기반 알고리즘의 책임성과 신뢰성을 높이는 정량적 근거로 활용될 수 있다.

실험에서는 대학생의 학업 성취 수준(졸업, 중도탈락, 재학)을 예측하는 AutoML 기반 머신러닝 모델 중 Stacked Ensemble이 가장 높은 예측력을 보여 기준 모델로 활용되었으며, 민감 속성으로 부모의 학력 및 직업을 설정하였다. 상위 조합(top-n=3)과 20개 샘플을 대상으로 한 반사실적 데이터 실험에서 Flip 발생률은 16.9%로 측정되었으며, 특히 '아버지의 학력'이 예측 반전에 큰 영향을 주는 것으로 나타났다. 더불어 SHAP 및 iBreakDown 기법을 활용한 분석 결과, 해당 속성의 상대적 기여도가 타 속성보다 약 1.5배 이상 높은 영향력을 보이며 예측 반전에 결정적 역할을 하는 것으로 확인되었다. 이러한 결과는 학습 알고리즘의 편향 원인을 추론하고 공정성을 개선하기 위한 실질적 분석 도구로 활용 가능함을 시사한다.

주제어 반사실적 공정성, 설명가능 인공지능, 학업성취 예측, 민감 속성 기여도 분석, AutoML 기반 앙상블 학습

ABSTRACT

This study aims to quantitatively evaluate the fairness of predictive models using counterfactual instances and to identify potential sources of bias through interpretable attribution analysis focused on sensitive attributes. Unlike prior studies that merely assess disparities in prediction outcomes, this research analyzes whether a model reverses its prediction (i.e., Flip) in response to changes in sensitive attributes and visualizes their contributions near decision boundaries using explainable AI techniques, such as Counterfactual SHAP and iBreakDown.

These analyses offer empirical grounds to enhance the accountability and trustworthiness of data-driven algorithms, particularly in formulating higher education policies and preventing early student dropouts. Among several AutoML-based machine learning models tested to predict university students' academic outcomes (Graduated, Dropout, Enrolled), the Stacked Ensemble model demonstrated the highest predictive performance and was adopted as the reference model. The sensitive attributes were defined as parental education and occupation.

In counterfactual experiments involving top-n (n=3) combinations across 20 samples, the Flip rate was measured at 16.9%, with 'Father's Qualification' emerging as a key variable contributing to prediction reversals. Furthermore, SHAP and iBreakDown analyses revealed that this attribute exerted approximately 1.5 times greater impact compared to other features, indicating its decisive role in altering model outputs. These findings suggest the effectiveness of counterfactual analysis as a diagnostic tool for detecting algorithmic bias and enhancing fairness in predictive modeling.

Keywords Counterfactual Fairness, Explainable Artificial Intelligence (XAI), Academic Achievement Prediction, Sensitive Attribute Contribution Analysis, AutoML-based Ensemble Learning

†정회원 성균관대학교 일반대학원 교과교육학과 컴퓨터교육 전공 박사

논문투고 2025년 07월 30일
심사완료 2025년 11월 03일
게재확정 2025년 12월 23일
발행일자 2026년 03월 22일

1. 서론

최근 교육 데이터 분석(Educational Data Mining, EDM)과 학습 분석(Learning Analytics, LA)은 인공지능 기반 예측 모델을 통해 학습자의 학업 성취도를 조기에 파악하고, 맞춤형 개입 전략을 수립하는 방향으로 발전하고 있다. 특히 학업 중단 위험이 있는 학생을 조기에 식별하고, 성과 저하 원인을 진단하여 교육 현장에 실질적 도움을 주는 분석 기술에 대한 수요가 증가하고 있다.

기계학습 기반의 성취 예측 모델은 높은 예측 정확도를 제공하지만, 여전히 두 가지 중요한 한계를 안고 있다. 첫째는 예측 결과의 설명가능성(explainability) 부족이며, 둘째는 특정 집단에 불리한 결과를 초래할 수 있는 예측의 편향(bias) 가능성이다. 블랙박스 형태의 고성능 모델은 학습자가 왜 '성공' 혹은 '실패'로 분류되었는지에 대한 정당한 설명을 제공하지 못하며, 이는 예측 결과의 수용성을 저해할 수 있다. 동시에 성별, 출신 지역, 부모의 직업 및 학력과 같은 민감 속성(sensitive attributes)이 예측에 과도한 영향을 미칠 경우, 모델은 공정하지 않은 판단을 지속적으로 재생산할 우려가 있다.

이러한 문제를 해결하기 위한 방안으로 설명 가능한 인공지능(eXplainable AI, XAI) 기법과 공정성 기반 모델 평가(fairness auditing)가 동시에 주목받고 있다[1, 2]. 특히 최근의 정책 및 기술 프레임에서는 반사실적 공정성(Counterfactual Fairness) 개념이 중요한 평가 기준으로 부각된다. 반사실적 공정성이란, 어떤 개인이 민감 속성을 달리 가졌더라도 예측 결과가 달라지지 않아야 한다는 원칙에 기반하며, 예측 과정에서의 구조적 편향을 정량화할 수 있는 강력한 도구로 작용한다[3-13].

이와 관련하여 Cornacchia 외(2023)은 반사실적 추론(counterfactual reasoning)을 기반으로, 예측값의 변화만이 아니라 예측 경로상의 영향 요인들까지 분석할 것을 제안하며, 예측모형의 구조적 불공정성을 해석 가능한 방식으로 평가할 수 있도록 하였다[4]. 또한, Ferrara(2023)는 AI 모델 개발의 전 생애주기(AI-SDLC)에 공정성 점검 체계를 통합하는 방법론을 제시하며[2], 민감 속성을 통한 결과 차이뿐 아니라 모델 학습 이전 단계에서의 구조적 편향을 사전에 감지·완화하는 절차를 강조하였다[5]. 더불어, NIST SP 1270에서는 공정성은 통계적 정의 이상의 것이며, 사회적·응용적 맥락에서 역동적으로 정의되어야 한다"고 명시하며, 특히 데이터 적합성, 이해관계자의 관점, 사용 목적 등을 반영한 평가 체계의 중요성을 강조하고 있다[14].

본 연구는 김용우와 우상미(2024)의 학업 성취도 예측 모델 및 설명 가능한 인공지능 기반 해석 연구를 기반으로 설명가능성과 공정성을 동시에 고려한 분석들을 확장·적용하고자 한다[15]. 기존 연구에서는 H2O AutoML을 활용하여 다양한 예측 모델의 성능을 비교하고 SHAP, ALE, IBreakDown 등의 기법을 통해 주요 영향 요인을 설명하였다.

H2O AutoML은 다양한 머신러닝 알고리즘에 대해 자동으로 학습, 검증, 튜닝을 수행하고 성능이 우수한 모델을 선별하는 자동화된 머신러닝 프레임워크이다.

본 연구는 이를 기반으로 다변량 민감 속성에 따른 반사실적 공정성 분석을 추가하고, 예측 시 편향 발생에 기여한 속성을 설명 가능한 인공지능 관점에서 정량적으로 평가함으로써 예측 모델의 편향 문제를 보다 구조적으로 다음과 같이 접근하고자 한다.

첫째, 포르투갈 대학생 데이터를 활용하여 H2O AutoML 기반의 Random Forest, XGBoost, Stacked Learner 모델을 구성하고 동일 하이퍼파라미터 조건 하에 성능을 비교한다.

둘째, "부모별 직업과 학력" 4개의 다변량 민감 속성을 기준으로 반사실적(counterfactual) 데이터를 생성하여, 예측값이 바뀌는 사례(Flip)를 추출한다.

셋째, 예측 Flip이 발생한 사례에 대해 설명 가능한 인공지능 기법(SHAP, iBreakdown)을 적용함으로써, 각 민감 속성이 예측 결과 변화에 어느 정도 기여했는지 정량적으로 해석하고, 공정성 위반 가능성을 분석한다.

이를 통해 본 연구는 기존 학업성취 예측 연구들이 간과했던 구조적 편향과 민감 속성의 영향력을 설명 가능한방식으로 분석하고자 하며, 다음과 같은 연구 질문(Research Questions)을 중심으로 진행된다.

·RQ1: 다변량 민감 속성을 기반으로 생성된 반사실적 샘플에서 예측 Flip이 발생한 경우, 어떤 민감 속성이 예측에 가장 크게 기여하는가?

·RQ2: 설명 가능한 인공지능 기법을 활용하여 예측 결과 변화(Flip)에 기여한 속성을 정량적으로 설명할 수 있는가?

·RQ3: 동일한 하이퍼파라미터 조건 하에 훈련된 Random Forest, XGBoost, Stacked Learner 모델 중 어떤 모델이 공정성과 설명가능성 측면에서 가장 우수한가?

·RQ4: 민감 속성 기반 편향은 실제 교육개입 전략 수립에 어떤 통찰을 제공할 수 있는가?

본 연구는 단지 예측 성능을 높이는 데 그치지 않고, 교육 맥락에서 예측 결과의 정당성, 신뢰성, 공정성을 확보하는 방향으로 인공지능 기술의 적용을 확장하며, AI의 윤리적 활용을 위한 실천적 연구 사례로서 의의가 있다.

2. 연구 배경

2.1 학습 분석 및 학업성취 예측 모델

학습 분석과 교육 데이터 분석은 학습자의 행동, 맥락, 성과 데이터를 분석하여 학습을 최적화하는 데 목적을 둔 대표적인 학제 간 연구 분야이다 [16]. 학습 분석은 학습자의 성과 예측, 중도포기 탐지, 행동 패턴 모델링, 인지 상태 분석 등 다양한 예측 과제를 포함한다[17].

국내외 연구에서는 입학 성적, 수강 이력, 학사경고 여부, 부모의 학력 및 직업 등 다양한 특성을 활용하여 학업 성취

를 예측하는 모델들이 개발되었다[18, 19]. 특히, 김용우와 우상미(2024)는 포르투갈 대학생 데이터를 기반으로 GBM, RF, FNN 등 다양한 분류 모델을 비교하고, AutoML 기반의 최적 성능 모델을 도출하였다[15]. 본 연구는 성취도 예측에서 고성능 모델의 가능성을 제시했으나, 예측 결과에 대한 구조적 편향 분석은 포함되지 않았다.

2.2 설명 가능한 인공지능의 교육 분야 적용 사례

설명 가능한 인공지능은 모델의 결정 과정을 인간이 이해할 수 있도록 설명하는 기술로, 교육 분야에서 학습자의 예측 결과를 해석하고, 교사나 관리자에게 인사이트를 제공하는 수단으로 활용되고 있다[20, 21].

설명 가능한 인공지능 기법 중 SHAP, ALE, Permutation Importance는 다양한 예측 모델에 적용 가능하며, 특히 개별 관측치 기반의 해석(iBreakdown, SHAP force plot)은 학습자 개개인에 대한 설명에 유용하다. 김용우 & 우상미(2024)는 교육 현장의 다양한 요구를 반영하여 SHAP, iBreakdown, ALE, Permutation Importance 등 다수의 기법을 적용함으로써, 설명 가능한 인공지능을 단일 설명에 국한하지 않고 복수의 관점에서 통합적으로 활용할 수 있음을 제안하였다[15].

그러나 기존 연구들은 대부분 설명 가능성에만 초점을 두었으며, 설명된 요인이 공정한가에 대한 분석은 상대적으로 부족하다. 본 연구는 이러한 설명 기법을 공정성 분석과 결합함으로써, 설명 가능한 공정성(eXplainable Fairness)을 구현하고자 한다.

2.3 반사실적 공정성 및 편향 측정 기법

반사실적 공정성은 “어떤 개인이 민감 속성만 달랐더라도 동일한 예측을 받았을 것”이라는 전제하에 공정성을 평가하는 이론적 틀이다[3]. 이는 단순한 통계적 평등 지표(statistical parity)보다 인과적이며 개별 예측 단위 수준에서의 편향 여부를 측정할 수 있다는 장점이 있다.

Wachter et al.(2017)은 GDPR에 부합하는 모델 설명을 위해 Counterfactual Explanation의 법적·실천적 정당성을 제시하였으며[22], Cornacchia et al.(2021)은 예측 결과뿐 아니라 모델 내 의사결정 경로 상 민감 속성의 개입 경로까지 추적해야 한다는 점을 강조하였다[4]. 이는 단순히 Flip 발생 여부를 넘어, 어떤 속성이 Flip에 가장 크게 기여했는지를 설명할 수 있어야 진정한 반사실적 공정성 달성이 가능함을 시사한다.

한편, Ferrara(2024)는 AI-SDLC 기반의 공정성 점검 프레임을 제시하며, 모델 학습 단계뿐 아니라 설계 및 배포 단계에서도 민감 속성에 기반한 편향이 구조적으로 발생할 수 있음을 경고하였다[2]. NIST SP 1270에서는 공정성을 맥락 특이적(context-specific) 개념으로 정의하고, 데이터와 응용 목적에 맞는 공정성 정의가 필요하다고 보았다[14].

2.4 민감 속성과 예측 편향의 정의 및 분석 전략

AI 모델에서의 예측 편향은 학습 데이터 내 특정 그룹의 과소 대표성, 구조적 불균형, 혹은 민감 속성의 과도한 개입에 의해 발생한다[23]. 민감 속성이란 예측 대상과 직접 관련되지 않아야 하며, 성별, 연령, 인종, 부모의 직업·학력 등이 포함될 수 있다.

김용우와 우상미(2024)는 부모의 학력, 직업과 같은 사회경제적 배경 변수가 예측 정확도에 영향을 미친다는 점을 XAI 분석을 통해 시사하였다[15]. 하지만 해당 연구는 민감 속성의 영향을 공정성 관점에서 다루지 않았기 때문에, 본 연구에서는 해당 속성들이 예측값 변화(Flip)에 어떠한 기여를 하였는지 반사실적 기반 샘플을 활용하여 분석하고자 한다.

이러한 분석을 위해 본 연구는 민감 속성을 조작한 반사실적 샘플을 생성하고, 예측값이 변경되는 경우를 추출한 뒤, SHAP 기반의 Counterfactual SHAP 분석과 iBreakDown 설명 가능한 인공지능 기법을 통해 예측 변화에 기여한 속성의 영향력 차이를 정량화함으로써, 다변량 민감 속성 기반 공정성 해석을 시도한다.

3. 연구 방법론 (Methodology)

3.1 데이터세트 및 변수 구성

본 연구에 사용된 데이터 세트는 김용우와 우상미(2024)가 활용한 포르투갈 고등교육기관 대학생 성적 데이터 세트를 기반으로 했으며[15], 포르투갈의 대학에 등록한 학생들에 관한 데이터 세트로 Realinho, Machado, Baptista, and Martins(2022)에서 사용된 데이터이다[24]. 해당 데이터는 총 4,424명의 고등교육 성적을 포함하고 있으며, 대상 데이터 세트는 전처리(결측값 제거 등) 이후 총 4,424건으로 구성되었으며, 각 클래스별 분포는 Graduate(2,209건, 49.9%), Dropout(1,421건, 32.1%), Enrolled(794건, 18.0%)로 확인되었다. 종속 변수(target variable)는 학업결과(Target)로 이진 분류 문제로 정의했다.

특성 변수(feature variables)는 성적, 출석률, 등록정보, 재정 상황, 부모의 직업 및 학력 등 총 34개 변수로 구성되어 있고, 이 중 부모의 학력, 부모의 직업 4개 변수는 “민감 속성”으로 지정했다. 이들은 개인의 특성을 나타내며 예측값에 비정보적으로 가지게 되는 집단성과 포함성을 가지고 있어, 모델의 결정에 결정적 사용이 될 가능성이 있다.

3.2 반사실적 데이터 생성

공정성 범위 방식의 테스트를 위해, 문화적 민감성(부모의 학력과 직업)을 변경한 다변량 반사실적 샘플을 생성했다. 이때 기호 비민감 속성은 원래 값을 유지하며, 각 관측치에 대해 가능한 민감 속성 조합으로 대체된 새로운 샘플을 생성했다. 예측 결과가 변경된 경우(“prediction(x) ≠

prediction(x_cf)")를 Flip 사례로 정의하고, 이를 공정성 위반 가능성이 있는 사례로 간주했다.

민감 속성 조합의 수는 각 속성에서 상위 N개의 고유값만 사용하는 방식으로 제한하였으며, 이를 통해 계산 효율성과 현실성을 동시에 확보하였다. 실험 결과는 Table 1과 같다.

Table 1. Flip incidence rate by top-n combinations

Top-n	Flip 발생률(%)	Top-n	Flip 발생률(%)
3	2.60	4	3.80
5	5.07	6	5.21

본 연구에서 Top-n=5 기준을 채택한 이유는 다음과 같다. 민감 속성 4개(부모의 학력과 직업)에 대해 고빈도값 기준의 다변량 조합을 생성하는 과정에서, Top-n 값이 6 이상으로 증가할 경우 조합 수가 기하급수적으로 증가하게 되어, 실험의 계산 복잡도가 급격히 상승하고 연산 안정성이 저하되는 문제가 발생하였다.

따라서 실험이 안정적으로 수행 가능한 범위 내에서 분석적 타당성과 계산 효율성을 모두 확보할 수 있는 수준으로 Top-n을 5로 설정하였다. 이 기준은 예측 민감도 및 Flip 패턴의 경향을 파악하는 데에도 충분한 수준으로 판단되며, 향후 자원 확장이 가능한 환경에서는 Top-n=10, Top-n=50, 또는 Top-n=100 이상의 조합 확장도 고려할 수 있다.

본 연구는 각 민감 속성에 대해 고유값 분포 상위 5개(top-n=5)를 사용하여 반사실적 데이터를 구성하였다. 이는 전체 민감 속성 공간을 포괄하되, 데이터의 대표성을 유지하면서 실현 가능한 예측 조합 수 내에서 Flip 발생률을 분석하기 위함이다. 이 방식은 민감 속성 공간을 전반적으로 포괄하면서도, 실제 교육현장 적용 가능성과 연산 효율성을 동시에 확보하고자 하는 전략이다. Flip 발생률은 5.07%로 측정되었으며, 이는 공정성 분석의 타당성과 현장 적용성을 모두 고려한 적절한 기준으로 판단한다.

Table 2. Academic outcome class distribution

Class Label	Class Meaning	Sample Count	Proportion (%)
0	Dropout	1,137	32.1
1	Enrolled	635	17.9
2	Graduate	1,767	50.0

본 데이터 세트는 3개의 학업 결과 클래스(Dropout, Enrolled, Graduate)로 구성되어 있으며, 전체 분포는 Table 2와 같다. Graduate클래스가 전체의 약 50%를 차지하는 반면, Enrolled 클래스는 18% 수준으로 상대적으로 적게 분포되어 있다. 이러한 불균형은 학습 모델의 성능 및 공정성에 영향을 줄 수 있기 때문에, 본 연구에서는 mean per class error, MSE, RMSE 등의 지표를 사용하여 평가하고, 반사실적 공정성 기반의 분석을 통해 구조적 편향 가능성을 검토하였다.

각 테스트 샘플에 대한 훈련 데이터(X_train)에서 해당 민감 속성의 고유값 조합을 기반으로 가능한 모든 대체 샘플을 생성하고, 이를 AutoML을 통해 도출된 최고 성능 모델에 입력하여 예측 결과의 변화 여부를 평가하였다.

3.3 예측 목표 모델 구성 및 학습

목표 모델은 H2O의 AutoML 기능을 활용하여 동일한 조건 하에 3가지의 모델(Random Forest (RF), XGBoost, Stacked Ensemble (base: RF + XGB))을 학습시키고, 상위 성능을 보인 모델 1개를 Table 3과 같이 선정했다. AutoML 과정에서는 다음과 같은 조건을 설정했다.

Table 3. Performance evaluation metrics

Model ID	Mean per class error	LogLoss	RMSE	MSE
StackedEnsemble_BestOfFamily_2_AutoML_1	0.293	0.548	0.423	0.179
StackedEnsemble_AllModels_1_AutoML_1	0.293	0.543	0.421	0.177
StackedEnsemble_AllModels_3_AutoML_1	0.293	0.545	0.421	0.177
StackedEnsemble_BestOfFamily_4_AutoML_1	0.294	0.546	0.422	0.178
StackedEnsemble_BestOfFamily_3_AutoML_1	0.297	0.544	0.422	0.178
StackedEnsemble_BestOfFamily_1_AutoML_1	0.299	0.546	0.423	0.179
StackedEnsemble_AllModels_2_AutoML_1	0.299	0.543	0.421	0.177
XGBoost_grid_1_AutoML_1_model_2	0.301	0.568	0.424	0.180
XGBoost_grid_1_AutoML_1_model_3	0.301	0.567	0.423	0.179
GBM_5_AutoML_1	0.301	0.554	0.424	0.180

이 연구에서는 다중 클래스 분류 문제의 특성과 불균형한 클래스 분포를 고려하여 성능 평가 지표를 구성하였다.

· 평가 지표 : Mean_mean_per_class_error, logloss, RMSE, MSE

· 자체 검사 : 5-fold cross-validation

Mean per class error는 각 클래스의 분류 오류율을 평균한 지표로, 클래스 불균형 상황에서도 모델의 전반적인 성능을 균형 있게 평가할 수 있다. 본 연구에서는 AUCPR 대신 해당 지표를 AutoML 정렬 기준으로 사용하였다. LogLoss는 모델이 예측한 클래스 확률 분포의 신뢰도를 평가하는 지표로, 예측이 실제 클래스에 가까울수록 손실값이 낮아진다. RMSE는 예측값과 실제값 사이의 평균적인 오차

크기를 직관적으로 보여주는 지표이며, 모델의 전반적인 예측 안정성을 확인할 수 있다. MSE는 전체 오차의 제곱 평균으로, 예측에서 발생하는 오류의 제곱합을 통해 예측 안정성을 정량화한다. 선정된 모델에는 AutoML이 동일한 하이퍼파라미터 조건을 적용하여, 동일 조건 하에서 성능 비교가 가능하도록 했다.

4. 실험 및 분석 결과 (Experiments and Results)

4.1 모델별 성능 평가

선정된 최고 성능 모(StackedEnsemble_BestOfFamily_2_AutoML_1)의 테스트 데이터 기준 78%의 정확도를 기록하였으며, Fig. 1의 혼동행렬 결과와 Table 4의 정량 지표를 통해 그 성능을 확인할 수 있다.

클래스별 F1-score는 Dropout 0.78, Enrolled 0.49, Graduate 0.87로, Graduate 클래스에서 가장 높은 예측 성능을 보였다. 반면 Enrolled 클래스는 상대적으로 낮은 F1-score(0.49)를 기록했으며, 이는 샘플 수의 불균형(159건) 및 다른 클래스와의 경계 모호성에 기인한 것으로 해석된다.

평균 관점에서 Macro 평균 F1-score는 0.71로 모든 클래스를 균등하게 고려한 모델의 전반적 성능을 나타내며, Weighted 평균 F1-score는 0.77로, 특히 샘플 수가 많은 Graduate 클래스(442건)의 영향이 반영된 결과로 볼 수 있다. 이러한 클래스 간 성능 차이는 향후 공정성 분석 및 반사실적 기반 평가 시 고려되어야 할 주요 요소로 작용한다.

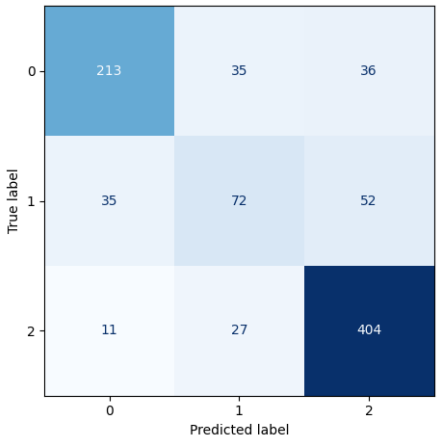

Figure 1. Confusion matrix of the best AutoML mode

Table 4는 Stacked Ensemble 기반의 최종 선정 모델이 테스트 데이터에 대해 나타낸 분류 성능 지표를 요약한 것이다. 전체 정확도는 0.78로, 3개의 클래스(Class 0: Dropout, Class 1: Enrolled, Class 2: Graduate)에 대해 균형 잡힌 예측 성능을 보인다.

Table 4. Confusion matrix of the best-performing model

Class	Precision	Recall	F1-score	Support
Class 0 (Dropout)	0.82	0.75	0.78	284
Class 1 (Enrolled)	0.54	0.45	0.49	159
Class 2 (Graduate)	0.82	0.91	0.87	442
정확도 (Accuracy)			0.78	885
Macro Avg	0.73	0.71	0.71	885
Weighted Avg	0.77	0.78	0.77	885

최종 선정된 스태킹 앙상블 모델(StackedEnsemble_BestOf Family_2_AutoML_1)은 다중 분류 문제에서 높은 정확도(78.1%)와 낮은 평균 클래스 오류율(29.7%)을 보였다. 특히, Dropout 및 Graduate 클래스의 예측은 높은 일관성을 보였으나, Enrolled 클래스는 상대적으로 높은 오류율을 Table 5와 같이 확인할 수 있다.

Table 5. Summary of performance metrics (cross-validation)

Metric	Value
LogLoss	0.547
RMSE	0.422
MSE	0.178
mean_per_class_error	0.296
mean_per_class_accuracy	0.704
top-1 Accuracy	0.781
Top-3 Hit Ratio	1.000

4.2 민감 속성 변경에 따른 예측 Flip 사례 통계

본 연구에서는 AutoML에서 생성된 후보 모델 중 테스트 세트 기준 상대적으로 가장 우수한 성능을 보인 모델(StackedEnsemble_BestOfFamily)을 기반으로 민감 속성 변화에 따른 반사실적 데이터를 구성하였다. 각 민감 속성별 고유값 중 상위 5개(top-n=5)를 선택하여 다변량 조합을 생성하고, 이를 활용한 반사실적 샘플을 통해 모델 예측의 민감성과 공정성을 분석하고자 하였다.

다변량 조합 방식의 경우 속성 수와 고유값 수에 따라 조합 수가 기하급수적으로 증가하게 되며, 이는 분석 시간과 계산 복잡도를 비선형적으로 증가시킨다. 따라서 본 연구에서는 실험의 초기 단계에서 다양한 민감 속성 조합에 따른 예측 민감성을 정성적으로 확인하고자, 분석 가능성과 계산 안정성이 확보된 범위 내에서 대표성 있는 3개의 테스트 샘플을 선정하였다. 이들 샘플은 학습 데이터의 주요 클래스 분포를 반영할 수 있도록 계층적 추출(stratified sampling) 방식으로 선정되었으며, 각 샘플에 대해 동일한 방식의 반사실적 조합이 적용되었다.

본 연구에서의 Top-3 Hit Ratio는 단순 정확도 지표가 아닌, 민감 속성 조합 변경에 따라 예측 클래스가 변화하는지(Flip 여부)를 확인하고, 변화가 전체 클래스 공간(3개 클래스) 내에서 안정적으로 유지되는지를 평가하기 위한 보조

지표로 사용되었다. 이는 모델의 예측 민감도와 편향된 클래스 이동 경향을 간접적으로 점검하는 데 목적이 있다.

이후 실험에서는 배치 예측 및 top-n 조절 전략을 병행하여 반사실적 공간을 점진적으로 확장하였으며, 예측 결과 및 설명 가능성 분석의 경향성이 유지되는지를 검증함으로써 소규모 샘플 기반 분석이 실제 민감도 평가에 있어 유의미한 패턴을 도출할 수 있음을 확인하였다. 이는 제한된 자원 하에서도 반사실적 기반의 공정성 평가가 타당한 초기 진단 수단이 될 수 있음을 시사한다.

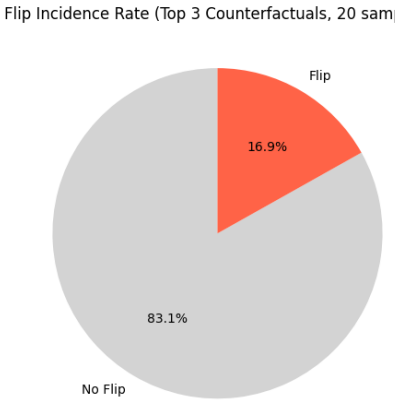

Figure 2. Flip incidence rate (pie chart visualization)

그 결과, Fig. 2와 같이 선택된 3개 샘플에 대해 총 N개의 반사실적 조합을 생성하고 예측을 비교한 결과, 약 16.9%의 Flip 발생률을 확인하였다. 이는 민감 속성의 변화가 모델 예측에 유의미한 영향을 미침을 시사하며, 공정성 분석의 실효성을 갖춘 실험 설정으로 판단된다.

특히 본 실험에서는 Top-n=5 기준으로 생성된 다변량 반사실적 조합을 기반으로 Flip 발생률을 측정하였다. 해당 기준은 민감 속성 공간의 대표성을 유지하면서도, 조합 수 증가에 따른 계산 복잡도를 제어하여 연산 안정성과 실험의 반복 가능성을 확보할 수 있는 적정 수준으로 설정되었다.

이 범위는 예측 모델의 민감도 및 예측 경계 인근에서의 Flip 발생 경향을 파악하기에 충분한 해상도를 제공하며, 향후 자원이 확장된 환경에서는 Top-n=10, Top-n=50 등의 조합을 통한 정밀도 향상도 가능할 것으로 기대된다.

4.3 민감 속성별 기여도 분석

본 연구에서는 학업성취도 예측모델의 공정성 검증을 위하여 반사실적 기반 설명 가능한 인공지능 분석을 수행하였다. 특히, 예측 결과가 변화한 사례(Flip Sample)를 중심으로, 개별 민감 속성이 예측 변경에 어느 정도 기여하였는지를 정량적으로 해석하고자 하였다. 본 절의 연구에서는 각 민감 속성에 대해 고유값 분포 상위 3개(topn=3)를 사용하여 반사실적 데이터를 구성 및 플립율을 계산하고 이를 분석하기 위해 SHAP 기반의 Counterfactual SHAP 분석과 iBreakDown기법을 적용하여, 각 민감 속성의 SHAP 값 차이(원본 - 반사실적)의 누적 기여도를 도출하였다.

분석 결과, 예측이 변화한 샘플에서 가장 큰 영향을 미친 요인은 '아버지의 학력'으로 확인되었다. 해당 속성은 전반적으로 예측 변경을 유발하는 방향(음의 기여도)으로 작용하였으며, 다른 속성과 비교했을 때 상대적으로 가장 큰 SHAP 변화폭을 보였다. 반면, '어머니의 직업'과 '어머니의 학력'은 예측을 유지하거나 변화에 대한 기여가 제한적인 수준에 머물렀다. 일부 사례에서 '아버지의 직업' 역시 미미한 영향을 미쳤으나, 전체적인 기여도는 낮은 편이었다.

이러한 결과는 모델의 예측이 특정 민감 속성, 특히 아버지의 학력 수준에 과도하게 의존하고 있을 가능성을 시사한다. 민감 속성이 실제로 예측값에 큰 영향을 미친다는 것은 공정성 관점에서 차별적 요소가 모델에 내재되어 있을 가능성을 의미하며, 후속적으로 이러한 편향을 완화하기 위한 알고리즘 개선이나 속성 비중 조정이 필요하다.

본 연구에서는 이를 정량화하기 위하여 각 속성의 상대적 기여도를 정규화하였으며, SHAP 기반의 Counterfactual SHAP 분석과 iBreakDown 분석 결과에 따르면 '아버지의 학력'은 모든 민감 속성 중 예측 변화에 대한 가장 큰 정량적 약 1.4배 수준의 기여도를 나타냈으며, 이러한 정성적·정량적 분석은 모델의 불공정성 탐지 및 설명 가능한인공지능 체계 설계에 유의미한 시사점을 제공한다.

4.4 Flip 발생과 민감 속성 기여도의 관계

반사실적 샘플에서 예측이 변화한 사례(Flip Sample)를 중심으로, 민감 속성 변화가 예측 결과에 어떤 영향을 미쳤는지를 해석 가능하게 분석하기 위해 SHAP 기법을 적용하였다. 이를 통해 예측 변경에 기여한 주요 속성을 도출하고, 모델의 잠재적 편향 구조를 시각적으로 검토하고자 하였다.

특히, Flip이 발생한 반사실적 샘플과 원본 샘플 간의 SHAP 값 차이를 누적적으로 계산하여 시각화한 결과, 'Father's qualification' 속성이 다른 속성에 비해 가장 큰 음의 SHAP 차이(기여도 감소)를 나타냈다. 이는 해당 속성이 모델의 결정 경계(Decision Boundary)에 민감하게 작용하고 있으며, 예측 결과의 변화에 직접적인 영향을 주고 있음을 시사한다.

이러한 결과는 다음과 같은 해석을 가능하게 한다.

· 예측이 변경된 샘플들(Flip Sample)은 민감 속성 중 일부가 결정 경계 근처에서 작용하고 있다는 증거이며, 이는 해당 속성이 모델의 편향(Bias)에 기여하고 있을 수 있음을 시사한다.

· 특히, 아버지의 학력 수준은 모델의 예측에서 비선형적으로 작용하며, 학력 수준이 높은 경우에는 긍정적인 예측(예: 졸업)의 확률이 증가하고, 낮은 경우에는 부정적인 예측(예: 중도탈락)으로 이동하는 경향을 보였다.

· 반면, '어머니의 직업'이나 '아버지의 직업'은 상대적으로 Flip 발생과의 연관성이 낮은 기여도를 보였으며, 이는 해당 속성이 모델의 예측에 미치는 영향이 제한적임을 나타낸다.

이러한 분석은 단순히 민감 속성이 예측에 미치는 절대적인 영향력을 넘어, 민감 속성의 변화가 예측 결과를 어떻게, 얼마나 바꾸는지를 정량적으로 이해하는 데 도움을 준다. 더불어 Flip이 발생한 경우에만 국한하여 기여도를 분석함으로써, 모델의 잠재적 불공정성과 차별 가능성을 보다 정밀하게 진단할 수 있는 방법론적 기초를 제공한다.

4.5 대표 사례에 대한 XAI 시각화 및 해석

본 절에서는 예측 결과가 반사실적 조합에 따라 변화(Flip)한 사례를 대상으로, 모델의 판단 근거를 설명 가능한 인공지능 기법(XAI)을 통해 분석하였다. 특히, Counterfactual SHAP 기반의 기여도 분석과 iBreakDown 시각화를 함께 활용함으로써, 민감 속성이 예측 결과에 어떠한 영향을 미치는지를 정량적·직관적으로 해석하고자 하였다.

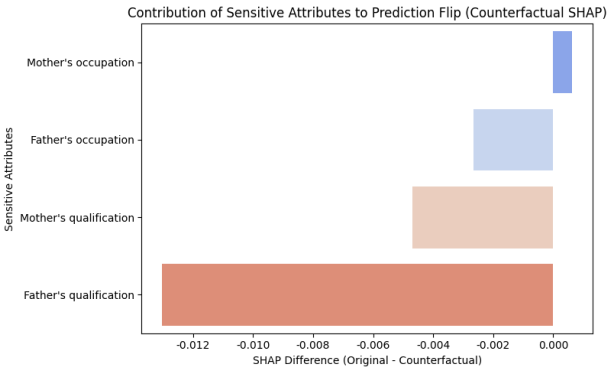

Figure 3. Counterfactual SHAP–based contribution analysis

우선, Fig. 3은 Flip이 발생한 대표 샘플에 대해 Counterfactual SHAP 값을 시각화하여 민감 속성의 상대적 기여도를 확인하였다. 분석 결과, '아버지의 학력'과 '어머니의 학력' 속성의 SHAP 값 차이가 예측 변화에 결정적인 영향을 미쳤으며, 반사실적 조합에 따라 해당 속성의 기여도가 크게 감소하거나 반전되는 양상을 보였다. 예를 들어, 실제 예측값이 'Graduate'였던 한 샘플은 '아버지의 학력'을 낮은 수준으로 변경한 반사실적 조합에서 'Dropout'으로 예측이 변경되었으며, 해당 속성의 SHAP 기여도 역시 0.15에서 −0.05로 감소하였다.

반면, '어머니의 직업'이나 '아버지의 직업'은 일부 샘플에서 SHAP 기여도가 미미하거나 변화폭이 작아 예측에 큰 영향을 주지 않는 것으로 나타났다. 이는 모델이 특정 민감 속성에 대해 상대적으로 높은 민감도를 보이며, 예측 경계 근처에 위치한 경우 해당 속성이 예측 결과를 변화시키는 데 주요한 역할을 할 수 있음을 시사한다.

또한, Flip 사례에 대한 추가 분석을 위해 iBreakDown 시각화 기법을 적용하였다. 이 방식은 예측값이 기본 예측점에서 각 특성의 기여도 변화에 따라 어떻게 누적적으로 이동하는지를 시각적으로 표현해준다. iBreakDown은 특정 인스턴스(단일 관측치) 예측에 대한 개별 설명변수의 기여도를 평가하는 방법이며, 계산에 사용되는 각 설명변수의 채택 순서에 따라 기여도에 차이가 발생한다[15, 25]. 상호작용에 취약한 특성 때문에 SHAP와는 상이한 결괏값을 Fig. 4와 같이 발생한다.

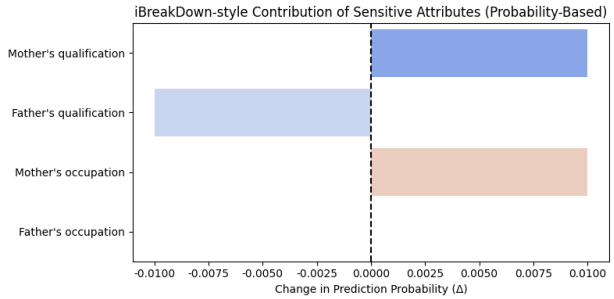

Figure 4. iBreakDown plot showing sensitivity contribution

그럼에도 불구하고 분석 결과, 예측 변화가 발생한 사례에서도 Counterfactual SHAP와 같이 '아버지의 학력'의 변화에 의해 전체 예측값이 급격히 하락하거나 상승하는 구간이 존재함을 확인할 수 있었다. 이는 해당 속성이 예측 반전을 유발하는 핵심 인자로 작용했음을 보다 직관적으로 보여준다.

예를 들어, 한 Flip 사례에서 원래 예측된 클래스는 'Enrolled'였으나, 아버지의 학력이 낮은 수준으로 설정된 반사실적 조합에서는 'Dropout'으로 변경되었으며, 이때 해당 속성의 iBreakDown 기여도 변화가 전체 누적 변화의 46.9% 이상을 차지하였다.

이상의 분석을 통해 다음과 같은 시사점을 도출할 수 있다.

· 모델은 민감 속성 중에서도 일부 속성(특히 부모의 학력)에 대해 높은 예측 민감도를 보이며, 이는 예측의 반사실적 공정성 관점에서 잠재적인 편향 요소로 작용할 수 있다.

· SHAP 기반의 기여도 시각화는 단일 예측 결과에 대한 설명뿐만 아니라, 반사실적 변화에 따른 예측 경로의 이해에도 유용한 도구로 활용될 수 있다. 예를 들어, Graduate로 예측되었던 원본 샘플이 '아버지의 학력'을 하향 조정한 반사실적 조건에서 Dropout으로 변경된 사례를 분석한 결과, 해당 속성의 SHAP 값은 +0.153에서 −0.235로 변화하여 총 0.388만큼 음의 방향으로 이동하였다. 이는 예측값에 대한 긍정적 기여가 부정적 기여로 전환되었음을 의미하며, 해당 민감 속성의 변화가 예측 반전에 결정적으로 작용했음을 수치적으로 보여준다. Fig. 3에 제시된 누적 SHAP 차이 시각화는 이러한 값을 시각적으로 표현한 것으로, 속성별 기여도 변화의 방향성과 크기를 직관적으로 파악할 수 있게 한다.

· iBreakDown 시각화는 예측에 영향을 준 주요 속성들의 기여도를 누적적으로 구간별 분해하여, 예측 결과 형성 과정을 단계적으로 설명할 수 있도록 한다. Flip 사례에 적용된 iBreakDown plot에서는 '아버지의 학력'의 변

화가 전체 예측 점수 변화에서 가장 큰 비중을 차지하며, 약 28.5%p에 해당하는 부정적 기여를 유도한 것으로 나타났다. 반면, 다른 민감 속성들은 예측 변화에 미친 영향이 상대적으로 작았으며, 예측 경로의 결정에 있어 주요한 변수로 작용하지 않았다. 이러한 분석은 단일 예측값의 형성과 Flip 발생 간의 인과 경로를 정량적으로 파악할 수 있는 기반을 제공한다.

이러한 시각화 기반의 설명 방식은 단순한 예측 성능 지표를 넘어, 민감 속성에 대한 모델의 반응과 그 기여도를 종합적으로 이해할 수 있는 중요한 근거를 제공한다. 특히, 반사실적 공정성 확보를 위한 후속 조치(예: 민감 속성 기반 규칙 제어, Re-Rraining 기준 수립 등) 수립 시, 해당 분석 결과는 실증적 판단 근거로 활용 가능하다.

5. 결론 (Conclusion)

본 연구는 대학생의 학업 성취 예측 문제를 다루는 머신러닝 모델을 대상으로, 민감 속성 기반의 반사실적 공정성 평가와 설명 가능한 인공지능 기법을 적용하여 예측 결과의 공정성과 해석 가능성을 종합적으로 분석하였다. 특히 부모의 학력과 직업을 민감 속성으로 설정하고, 해당 속성의 조합에 따른 예측 결과 변화 및 설명력을 정량적·시각적으로 검토하였다.

우선, 다양한 분류 모델을 자동 탐색하는 H2O AutoML을 활용하여 성능 비교를 수행한 결과, Stacked Ensemble 모델이 가장 우수한 예측력을 보이는 것으로 확인되었다. 해당 모델은 LogLoss 0.547, RMSE 0.422, mean_per_class_accuracy 70.4%, top-1 정확도 78.1%, Top-3 Hit Ratio 1.0으로 학습 및 교차검증 데이터 모두에서 일관되고 안정적인 성능을 나타냈으며, 이후의 반사실적 공정성 분석 및 설명 가능한 인공지능 해석의 기준 모델로 활용되었다.

반사실적 데이터는 민감 속성들의 상위 고유값 조합(TOP-n=5)을 기반으로 생성되었으며, 실험의 계산 복잡도와 해석의 실효성을 고려하여 일부 샘플에 대해 제한된 조합(예: TOP-n=3, sample size=20)을 설정하였다. 분석 결과, 민감 속성 값이 변화했을 때 예측 클래스가 달라지는 Flip 발생률은 약 16.9%로 측정되었으며, 이는 모델이 일부 조합에 대해 예측 결과를 변경할 가능성이 있음을 의미한다. 특히, 부모의 학력 속성('아버지의 학력', '어머니의 학력')이 예측 반전에 상대적으로 큰 영향을 미치는 것으로 나타났다.

이러한 예측 변화의 원인을 보다 정밀하게 파악하기 위해 SHAP 기반의 Counterfactual Explanation 분석을 수행하였다. 원본 샘플과 반사실적 샘플 간의 SHAP 값 차이를 비교한 결과, Flip이 발생한 경우 특정 민감 속성의 기여도가 크게 변동함을 확인할 수 있었다. 또한, iBreakDown 기법을 통해 개별 예측 사례의 기여도를 시각화함으로써, 예측 결과에 대한 인사이트를 직관적으로 제공함으로써 민감 속성이 예측에 미치는 평균적 영향력도 평가할 수 있었다.

또한 본 연구는 단순히 민감 속성 변화에 따른 예측 Flip 빈도를 정량적으로 측정하는 데 그치지 않고, 특정 속성이 예측 변화에 미치는 구조적 원인에 대한 해석적 논의까지 확대하였다.

특히, 분석 결과 부모의 학력이 예측 결과에 높은 영향을 미친 주요 요인으로 나타났는데, 이는 해당 속성이 학생의 교육 성취와 밀접하게 연결된 사회경제적 배경을 반영하기 때문이다. 예를 들어, 부모 학력이 높을수록 학생에게 제공될 수 있는 학습 자원, 진학 지도, 정서적 지지 수준이 상향될 가능성이 존재하며, 이는 실제 학업 이탈 위험도와도 상관관계를 가질 수 있다.

반면, 직업 속성은 고정된 직종보다는 소득 수준이나 근무 안정성 등과의 매개관계가 존재할 수 있어 단독 영향력보다는 조합 효과(multivariate interaction)를 통해 반사실적 공정성에 영향을 미친 것으로 해석된다.

이처럼 단일 속성의 독립적 영향뿐만 아니라, 다변량 민감 속성 간의 결합 구조를 반영한 공정성 분석은 실제 교육 현장의 복잡한 사회적 맥락을 보다 정밀하게 설명할 수 있으며, 향후 정책적 개입이나 개인화된 교육 지원 방안을 설계하는 데 실질적 시사점을 제공할 수 있다.

본 연구의 주요 결론은 다음과 같다. 첫째, 높은 예측 정확도를 보이는 모델이라 하더라도, 민감 속성의 변화에 따라 예측이 쉽게 달라질 수 있으며, 이는 모델의 공정성에 대한 추가적인 검토가 필요함을 시사한다. 둘째, SHAP, iBreakDown를 활용한 설명 가능한 인공지능 기법은 모델 예측의 내부 메커니즘을 해석하고, 민감 속성이 예측에 미치는 영향을 정량적으로 설명하는 데 효과적인 도구임을 확인하였다. 셋째, 반사실적 데이터를 기반으로 한 공정성 분석은 AI 모델의 책임성(responsibility)과 투명성(transparency) 확보에 실질적으로 기여할 수 있으며, 특히 교육, 고용, 복지 등 사회적 영향이 큰 분야에서의 사전 검증 도구로 활용 가능성이 높다.

향후 연구에서는 단순한 사후 해석에 그치지 않고, 학습 과정 자체에 공정성을 내재화하는 사전 처리(Pre-processing) 또는 학습 중 처리(In-processing) 기법을 통합하는 전략이 필요하다. 예를 들어, 편향 제거를 위한 reweighing(가중치 재할당), adversarial debiasing(공정성 학습방식) 등의 알고리즘을 반사실적 평가와 연계함으로써 공정성을 강화할 수 있다. 또한, 반사실적 데이터 생성 시 단순한 조합 기반 접근을 넘어서, 변수 간의 구조적 인과관계를 반영하는 구조적 인과 모델(Structural Causal Models, SCM) 기반의 의미 기반 반사실적 생성 알고리즘 도입도 중요한 연구 방향이 될 수 있다. 더불어, 본 연구는 Flip이 발생한 사례에 한정하여 민감 속성의 영향력을 분석하였으나, 향후 연구에서는 Flip이 발생하지 않은 샘플(Non-Flip group)에도 동일한 해석 기법을 적용하여 양 군 간의 영향도 차이를 비교함으로써, 예측의 안정성과 공정성 수준을 더욱 정밀하게 진단할 수 있을 것이다.

끝으로, 반사실적 평가 결과와 설명 가능한 인공지능의 해석 결과를 사용자 친화적 인터페이스를 통해 시각적으로 제공하고, 학습자, 교사, 정책 입안자 등이 이를 직관적으로 이해할 수 있도록 하는 공정성 대시보드 시스템의 개발도 실용적인 후속 연구로 고려할 수 있을 것이다. 이러한 연구는 교육뿐 아니라 공공 인공지능 시스템 전반의 윤리성과 신뢰성 제고에 기여할 수 있을 것으로 기대된다.

참고문헌

[1] Khosravi, H. et al. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3, 100074. https://doi.org/10.1016/j.caeai.2022.100074

[2] Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci, 6*(1), 3. https://doi.org/10.3390/sci6010003

[3] Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, 30.

[4] Cornacchia, A., et al. (2023). Auditing fairness under unawareness through counterfactual reasoning. *Inf. Process. Manag., 60*(2), 103224. https://doi.org/10.1016/j.ipm.2022.103224

[5] Moon, D., & Ahn, S. (2025). Metrics and Algorithms for Identifying and Mitigating Bias in AI Design: A Counterfactual Fairness Approach. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3556082

[6] Ma, J., Guo, R., Wan, M., Yang, L., Zhang, A., & Li, J. (2022). Learning fair node representations with graph counterfactual fairness. *Proceedings of the fifteenth ACM international conference on web search and data mining*, 695-703. https://doi.org/10.1145/3488560.3498391

[7] Ma, J., Guo, R., Zhang, A., & Li, J. (2023). Learning for counterfactual fairness from observational data. Proceedings of the 29th *ACM SIGKDD conference on knowledge discovery and data mining*, 1620-1630. https://doi.org/10.1145/3580305.3599408

[8] Shao, P., Wu, L., Zhang, K., Lian, D., Hong, R., Li, Y., & Wang, M. (2024). Average user-side counterfactual fairness for collaborative filtering. *ACM Transactions on Information Systems*, 42(5), 1-26. https://doi.org/10.1145/3656639

[9] Anthis, J., & Veitch, V. (2023). Causal context connects counterfactual fairness to robust prediction and group fairness. *Advances in neural information processing systems*, 36, 34122-34138.

[10] Wang, Z., Chu, Z., Blanco, R., Chen, Z., Chen, S. C., & Zhang, W. (2024). Advancing graph counterfactual fairness through fair representation learning. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 40-58. Cham: Springer Nature Switzerland.

[11] Cheong, J., Kalkan, S., & Gunes, H. (2022). Counterfactual fairness for facial expression recognition. *European Conference on Computer Vision*, 245-261. Cham: Springer Nature Switzerland.

[12] Yin, Z., Wang, Z., & Zhang, W. (2024). Improving fairness in machine learning software via counterfactual fairness thinking. *Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings*, 420-421. https://doi.org/10.1145/3639478.3643531

[13] Zhou, Z., Liu, T., Bai, R., Gao, J., Kocaoglu, M., & Inouye, D. I. (2024). Counterfactual fairness by combining factual and counterfactual predictions. *Advances in Neural Information Processing Systems*, 37, 47876-47907.

[14] Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. *Special Publication (NIST SP)*, National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1270

[15] Kim, Y., & Woo, S. (2024). On Use of eXplainable Artificial Intelligence to Explain Student Dropout. *Journal of The Korean Association of Artificial Intelligence Education*, 5(2), 12-25. https://doi.org/10.52618/aied.2024.5.2.2

[16] Romero, C., & Sebastian, V. (2020). Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 10(3), e1355. https://doi.org/10.1002/widm.1355

[17] Du, X. et al. (2021). A systematic meta-review and analysis of learning analytics research. *Behaviour & Information Technology*, 40(1), 49-62. https://doi.org/10.1080/0144929X.2019.1669712

[18] Moon, K., Kim, J., & Lee, J. (2021). Early Prediction Model of Student Performance Based on Deep Neural Network Using Massive LMS Log Data. *Journal of The Korea Contents Association*, 21(10), 1-10. https://doi.org/10.5392/JKCA.2021.21.10.001

[19] Kim, R. (2023). Exploring predictors of academic underachievement of college students based on learning analytics: Focusing on cumulative academic variables before the start of the semester. *The Journal of Learner-Centered Curriculum and Instruction*, 23(7), 121-136. https://doi.org/10.22251/jlcci.2023.23.7.121

[20] Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

[21] Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44-65. https://doi.org/10.1080/10618600.2014.907095

[22] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL &*

Tech., 31(2), 841.
[23] Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1-81.
[24] Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). *Predicting student dropout and academic success. Data*, 7(11), 146. https://doi.org/10.3390/data7110146
[25] Staniak, M., & Biecek, P. (2018). Explanations of model predictions with live and breakDown packages. *The R Journal*, 10(2), 395-409. https://doi.org/10.32614/RJ-2018-072

문동수

· 2002년 서울과학기술대학교 컴퓨터공학과(공학학사)
· 2006년 한국외국대학교 전자계산교육전공 (교육학석사)
· 2025년 성균관대학교 교과교육학과 컴퓨터교육전공 (교육학박사)

관심분야 : 인공지능 윤리, AI(머신러닝, 딥러닝), 컴퓨터 비전
m1d2s3@gmail.com