

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

PHẠM THỊ TỔNG

MỘT SỐ PHƯƠNG PHÁP ĐẢM BẢO TÍNH CÔNG BẰNG
CHO CÁC HỆ THỐNG HỌC MÁY
TRONG LĨNH VỰC GIÁO DỤC

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Hà Nội - 2026

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

PHẠM THỊ TỔ NGA

MỘT SỐ PHƯƠNG PHÁP ĐẢM BẢO TÍNH CÔNG BẰNG
CHO CÁC HỆ THỐNG HỌC MÁY
TRONG LĨNH VỰC GIÁO DỤC

Chuyên ngành: Khoa học máy tính

Mã số: 9480101.01

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

NGHIÊN CỨU SINH

TẬP THỂ CÁN BỘ HƯỚNG DẪN



PHẠM THỊ TỔ NGA GS.TS. NGUYỄN ĐỨC ANH PGS.TS. PHẠM NGỌC HÙNG

XÁC NHẬN CỦA ĐƠN VỊ ĐÀO TẠO

Hà Nội - 2026

Mục lục

0.1. Đặt vấn đề	1
0.1.1. Bối cảnh nghiên cứu về công bằng cho hệ thống học máy trong lĩnh vực giáo dục	1
0.1.2. Các thách thức và khoảng trống nghiên cứu	4
0.2. Mục tiêu, đối tượng, phương pháp, và phạm vi nghiên cứu	6
0.2.1. Mục tiêu nghiên cứu	6
0.2.2. Đối tượng và phương pháp nghiên cứu	7
0.2.3. Phạm vi nghiên cứu	7
0.3. Các đóng góp chính của luận án	10
0.4. Bố cục của luận án	13

Chương 1. TỔNG QUAN VỀ VIỆC ĐẢM BẢO TÍNH CÔNG BẰNG CHO CÁC HỆ THỐNG HỌC MÁY TRONG LĨNH VỰC GIÁO DỤC 14

1.1. Trí tuệ nhân tạo và học máy trong giáo dục	14
1.1.1. Vai trò của trí tuệ nhân tạo và học máy trong giáo dục	15
1.1.2. Đánh giá hiệu suất của các mô hình học máy	16
1.2. Công bằng trong các hệ thống học máy	18
1.2.1. Khái niệm về công bằng trong các hệ thống học máy	18
1.2.2. Khái niệm thiên vị trong học máy	20
1.2.3. Thuộc tính nhạy cảm trong nghiên cứu về công bằng	22
1.2.4. Độ đo công bằng trong các hệ thống học máy	24
1.2.5. Các hướng tiếp cận đảm bảo tính công bằng cho các hệ thống học máy	27

1.2.6. Những thách thức trong việc đảm bảo tính công bằng trong các hệ thống học máy.....	28
1.3. Sinh dữ liệu tổng hợp trong đảm bảo tính công bằng cho các hệ thống học máy.....	30
1.3.1. Các khái niệm.....	30
1.3.2. Kỹ thuật sinh dữ liệu tổng hợp.....	31
1.3.3. Thách thức và cơ hội của dữ liệu tổng hợp trong việc đảm bảo tính công bằng.....	33
1.4. Tổng quan nghiên cứu về công bằng cho các hệ thống học máy ứng dụng trong lĩnh vực giáo dục.....	34
1.5. Mối quan hệ giữa công bằng và hiệu suất của các hệ thống học máy	36
1.5.1. Sự đánh đổi giữa công bằng và hiệu suất.....	36
1.5.2. Các hướng tiếp cận để xử lý đánh đổi.....	37
1.6. Câu hỏi nghiên cứu.....	38
1.7. Tổng kết chương.....	39
Chương 2. PHƯƠNG PHÁP ĐẢM BẢO TÍNH CÔNG BẰNG NHỜ LOẠI BỎ SỰ PHỤ THUỘC VÀO CÁC THUỘC TÍNH NHẠY CẢM	41
2.1. Giới thiệu.....	41
2.2. Phương pháp Fairedu.....	43
2.2.1. Nguyên lý hoạt động.....	43
2.2.2. Kiến trúc tổng thể.....	44
2.2.3. Thuật toán.....	46
2.2.4. Ví dụ minh họa.....	49
2.3. Thực nghiệm.....	51
2.3.1. Dữ liệu.....	51

2.3.2. Lựa chọn mô hình học máy	54
2.3.3. Thiết lập thực nghiệm	55
2.3.4. Chỉ số đánh giá.....	55
2.4. Kết quả thực nghiệm	56
2.4.1. Thiên vị hệ thống trong dữ liệu giáo dục.....	57
2.4.2. Ảnh hưởng của mô hình học máy đến mức độ công bằng.....	58
2.4.3. Khả năng xử lý đồng thời nhiều thuộc tính nhạy cảm của FairEdu..	60
2.4.4. Mối quan hệ giữa công bằng và hiệu suất dự đoán	64
2.5. Thảo luận	65
2.5.1. Thảo luận các phát hiện chính từ kết quả thực nghiệm.....	65
2.5.2. Hạn chế của nghiên cứu.....	67
2.5.3. Ý nghĩa và ứng dụng thực tiễn.....	70
2.6. Tổng kết chương.....	71
Chương 3. PHƯƠNG PHÁP ĐẢM BẢO TÍNH CÔNG BẰNG NHỜ CÂN BẰNG DỮ LIỆU DỰA TRÊN KỸ THUẬT SINH DỮ LIỆU TỔNG HỢP	73
3.1. Giới thiệu	74
3.2. Phương pháp DPF	75
3.2.1. Nguyên lý hoạt động.....	76
3.2.2. Kiến trúc tổng thể.....	77
3.2.3. Thuật toán	82
3.2.4. Ví dụ minh họa.....	83
3.3. Thực nghiệm	85
3.3.1. Dữ liệu	85

3.3.2. Lựa chọn mô hình học máy và kỹ thuật sinh dữ liệu tổng hợp..	86
3.3.3. Thiết lập thực nghiệm	89
3.3.4. Chỉ số đánh giá.....	90
3.4. Kết quả thực nghiệm	91
3.4.1. Đặc điểm của dữ liệu tổng hợp trong lĩnh vực giáo dục.....	91
3.4.2. Hiệu quả của phương pháp DPF trong cải thiện công bằng giao thoa	94
3.4.3. Tác động của DPF đến hiệu suất dự đoán của mô hình	101
3.5. Thảo luận	102
3.5.1. Phân tích và tổng hợp các phát hiện chính.....	102
3.5.2. Thảo luận về dữ liệu tổng hợp và khả năng áp dụng	104
3.5.3. Hạn chế của nghiên cứu	105
3.5.4. Ý nghĩa và ứng dụng thực tiễn.....	106
3.6. Tóm tắt chương	106
Chương 4. PHƯƠNG PHÁP ĐẢM BẢO TÍNH CÔNG BẰNG HAI CHIỀU CHO DỮ LIỆU DẠNG BẢNG – CHỈ SỐ ĐÁNH ĐỐI GIỮA CÔNG BẰNG VÀ HIỆU SUẤT	108
4.1. Giới thiệu	109
4.2. Phương pháp FaireduPlus	110
4.2.1. Ý tưởng chính.....	110
4.2.2. Kiến trúc tổng thể.....	111
4.2.3. Thuật toán FaireduPlus	114
4.3. Thực nghiệm	115
4.3.1. Dữ liệu, mô hình học máy, kỹ thuật sinh dữ liệu tổng hợp, và độ đo đánh giá.....	115
4.3.2. Thiết lập thực nghiệm	117

4.3.3. Chỉ số đánh giá.....	117
4.4. Kết quả.....	122
4.4.1. Mức độ cải thiện tính công bằng đồng thời của FairEduPlus ..	122
4.4.2. Tác động của FairEduPlus đến hiệu suất dự đoán.....	124
4.4.3. Đánh giá mối quan hệ đánh đổi giữa công bằng và hiệu suất ..	128
4.5. Thảo luận.....	131
4.5.1. Phân tích và tổng hợp các phát hiện chính.....	131
4.5.2. Hạn chế của nghiên cứu.....	132
4.5.3. Hàm ý thực tiễn.....	134
4.6. Tóm tắt chương.....	135
PHỤ LỤC.....	164
Chương A. Các Bảng tổng hợp về tổng quan về nghiên cứu về đảm bảo tính công bằng trong học máy.....	165
Chương B. Các bảng bổ sung cho phân tích so sánh Fairedu và LTDD trong Chương 3.....	178
Chương C. Các bảng chi tiết về dữ liệu, kết quả công bằng, và hiệu suất theo từng chỉ số khi đánh giá cấu hình DPF với các cấu hình tham chiếu Chương 4.....	183

Danh sách hình vẽ

1	Cây nghiên cứu về tính công bằng cho các hệ thống học máy trong giáo dục liên quan đến luận án.	10
2	Mối quan hệ giữa các chương đề xuất phương pháp trong luận án.	11
1.1	Minh họa về các giai đoạn can thiệp công bằng trong các hệ thống học máy [1].	27
1.2	Các phương pháp sinh dữ liệu tổng hợp	31
1.3	Quy trình sinh dữ liệu tổng hợp [2]	33
2.1	Kiến trúc tổng thể của phương pháp Fairedu.	46
2.2	So sánh chỉ số $ 1 - DI $ giữa các thuộc tính nhạy cảm trên các tập dữ liệu	58
3.1	Kiến trúc tổng thể của phương pháp DPF	77
3.2	Tổng quan ba cấu hình thực nghiệm đánh giá hiệu quả DPF	89
4.1	Kiến trúc tổng thể của phương pháp FaireduPlus.	113
4.2	Tổng quan các cấu hình thực nghiệm đánh giá hiệu quả FaireduPlus	118

Danh sách bảng

1.1	Các định nghĩa về công bằng trong hệ thống học máy	21
1.2	Phân loại các thuộc tính nhạy cảm thường dùng trong giáo dục	23
2.1	Bảng mã hóa các thuộc tính nhạy cảm	52
2.2	So sánh các chỉ số công bằng đối với các thuộc tính nhạy cảm trong các thuật toán học máy khác nhau	59
2.3	Bảng so sánh giữa Fairedu và các phương pháp hiện có	61
2.4	Bảng so sánh chỉ số công bằng giữa Fairedu với các phương pháp LTDD và Origin với mô hình <i>Hồi quy logistic</i>	63
2.5	Bảng so sánh chỉ số công bằng giữa Fairedu với các phương pháp LTDD và Origin với mô hình <i>Rừng ngẫu nhiên</i>	63
2.6	Bảng so sánh chỉ số công bằng giữa Fairedu với các phương pháp LTDD và Origin với mô hình <i>Cây quyết định</i>	64
3.1	Bảng xác định nhóm con của DNU	83
3.2	Bảng xác định nhóm con sau khi sinh dữ liệu của DNU	84
3.3	Tổng quan chi tiết về các bộ dữ liệu và phân phối nhóm giao nhau	87
3.4	Độ biến thiên của chỉ số DI qua các lần chạy lặp lại (Logistic Regression)	92
3.5	Minh họa độ biến thiên của các chỉ số công bằng và hiệu suất qua các lần chạy lặp lại sử dụng kỹ thuật CTGAN	92
3.6	Minh họa tổng quan về các tập dữ liệu tổng hợp sử dụng kỹ thuật phân nhóm	93
3.7	Tổng hợp số lượt thắng về chỉ số công bằng theo từng cấu hình	95
3.8	Tổng hợp số lần thắng về kết quả công bằng theo từng mô hình	96
3.9	Biểu đồ heatmap so sánh các chỉ số công bằng của phương pháp DPF trên các bộ dữ liệu thực nghiệm	100
3.10	Số lượt thắng theo từng cấu hình trên hai tiêu chí ACC và Recall	101

4.1	Tóm tắt các bộ dữ liệu sử dụng trong nghiên cứu	116
4.2	So sánh chỉ số công bằng của FaireduPlus với các cấu hình đối sánh khi sử dụng CTGAN	124
4.3	So sánh chỉ số công bằng của FaireduPlus với các cấu hình đối sánh khi sử dụng LLM	125
4.4	So sánh các chỉ số hiệu suất (Acc và Recall) của FaireduPlus với các cấu hình thực nghiệm khác nhau	127
4.5	Comparison of fairness-performance trade-off indices ($\Delta_{trade_off}^{ACC_ 1-DI }$ and $\Delta_{trade_off}^{Recall_ 1-DI }$) across four experimental configurations using CTGAN	129
4.6	So sánh các chỉ số đánh đổi giữa công bằng và hiệu suất ($\Delta_{trade_off}^{ACC_SPD}$ và $\Delta_{trade_off}^{Recall_SPD}$) trên bốn cấu hình thực nghiệm sử dụng phương pháp sinh dữ liệu tổng hợp dựa trên LLM.	130
A.1	Danh sách các thuật toán học máy sử dụng trong các nghiên cứu chính	166
A.2	Các nghiên cứu chính theo từng loại công bằng được khảo sát	169
A.3	Các loại thiên vị khác nhau trong hệ thống học máy	170
A.4	Đặc điểm của các bộ dữ liệu được sử dụng trong nghiên cứu về đảm bảo tính công bằng giáo dục	172
A.5	Tổng hợp các phương pháp đảm bảo tính công bằng được sử dụng trong các nghiên cứu chính	174
A.6	Các kỹ thuật đánh giá công bằng và hiệu suất của học máy	176
B.1	So sánh hiệu suất (Acc và Recall) của mô hình áp dụng Fairedu với mô hình gốc và mô hình áp dụng LTDD	179
B.2	So sánh hiệu suất (Điểm số F1 và độ chính xác) của mô hình áp dụng Fairedu với mô hình gốc và mô hình áp dụng LTDD	181
C.1	Tổng quan về dữ liệu tổng hợp được sinh theo phương pháp DPF dựa trên kỹ thuật CTGAN và LLM	184
C.2	Chi tiết các chỉ số công bằng đối sánh với DPF dựa trên kỹ thuật CTGAN	187
C.3	So sánh hiệu suất của ba cấu hình thực nghiệm. Các ô xám biểu thị trường hợp thắng (W)	190

Thuật ngữ và từ viết tắt

Từ viết tắt	Từ tiếng Anh	Ý nghĩa/Tạm dịch
ABROCA	Area Between ROC Curves for different subgroups	Diện tích giữa các đường cong ROC của các nhóm khác nhau
ACC	Accuracy	Độ chuẩn xác
ADSGAN	Adversarially Debaised Synthetic GAN	Mạng GAN sinh dữ liệu tổng hợp với cơ chế giảm thiên lệch đối kháng
Adult	Adult Dataset	Bộ dữ liệu người lớn
AI	Artificial Intelligence	Trí tuệ nhân tạo
AI/ML	Artificial Intelligence/Machine Learning	Trí tuệ nhân tạo/Học máy
AOD	Average Odds Difference	Chênh lệch trung bình xác suất
CLAN	Communities with Lowly-connected Attributed Nodes	Phát hiện cộng đồng với các nút thuộc tính thuộc nhóm yếu thế
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions	Bộ dữ liệu hồ sơ quản lý người phạm tội của Hoa Kỳ
CTGAN	Conditional Tabular GAN	Mạng GAN có điều kiện cho dữ liệu bảng
Default	Default of Credit Card Clients Dataset	Bộ dữ liệu vỡ nợ của khách hàng thẻ tín dụng
DI	Disparate Impact	Tác động khác biệt
DIR	Disparate Impact Remover	Kỹ thuật loại bỏ tác động bất lợi để cân bằng phân phối thuộc tính nhạy cảm
DNN	Deep Neural Network	Mô hình học sâu
DNU	Dai Nam University Dataset	Bộ dữ liệu từ Trường Đại học Đại Nam
DPF	Data Partitioning Fairness	Phương pháp phân chia dữ liệu công bằng
DT	Decision Tree	Cây quyết định
EOD	Equal Opportunity Difference	Chênh lệch cơ hội công bằng

Từ viết tắt	Từ tiếng Anh	Ý nghĩa/Tạm dịch
F1	F1-score	Điểm số F1
FairSmote	Fair Synthetic Minority Oversampling Technique	Kỹ thuật tạo mẫu tổng hợp công bằng cho lớp thiểu số
Fairway	Fairway	Phương pháp tiền xử lý dữ liệu nhằm tăng cường công bằng trong học máy
GAN	Generative Adversarial Network	Mạng đối nghịch sinh
GB	Gradient Boosting	Tăng cường dốc
GDPR	General Data Protection Regulation	Quy định chung về bảo vệ dữ liệu
LR	Logistic Regression	Hồi quy logistic
LSTM	Long Short-Term Memory	Bộ nhớ ngắn-dài hạn
LTDD	Linear Regression based Training Data Debugging	Gỡ lỗi dữ liệu đào tạo bằng hồi quy tuyến tính
MCCM	Multiple Cooperative Classifier Model	Mô hình phân loại hợp tác nhiều thành phần
ML	Machine Learning	Học máy
NN	Neural Network	Mạng nơ ron thần kinh
OULAD	Open University Learning Analytics Dataset	Bộ dữ liệu thông tin khóa học của Đại học Mở Vương Quốc Anh
PATE-GAN	Private Aggregation of Teacher Ensembles GAN	Mạng GAN bảo toàn quyền riêng tư bằng cách giới hạn ảnh hưởng của từng mẫu dữ liệu, đảm bảo chặt chẽ tính riêng tư vi phân và nâng cao chất lượng dữ liệu tổng hợp
Precision	Precision	Độ chính xác
RBA	Reduction-based Approach	Cách tiếp cận dựa trên giảm thiểu
Recall	Recall	Độ hồi tưởng
Reweighting	Reweighting	Phương pháp gán lại trọng số để giảm thiên lệch trong dữ liệu
RF	Random Forest	Rừng ngẫu nhiên
RQ	Research Question	Câu hỏi nghiên cứu

Từ viết tắt	Từ tiếng Anh	Ý nghĩa/Tạm dịch
SDG	Synthetic Data Generation	Sinh dữ liệu tổng hợp
SPD	Statistical Parity Difference	Hiệu số chênh lệch thống kê
Student Performance	Student Performance Dataset	Bộ dữ liệu dự đoán hiệu suất học tập của sinh viên
Student Predict	Student Dropout and Academic Success Dataset	Bộ dữ liệu dự đoán bỏ học và thành công học tập
TabFairGAN	Tabular Fairness GAN	Mạng GAN sinh dữ liệu bảng tổng hợp với ràng buộc công bằng, giúp giảm thiên lệch từ các thuộc tính nhạy cảm
XAI	Explainable Artificial Intelligence	AI có khả năng giải thích

Giải thích kí hiệu

Kí hiệu	Mô tả
Δ_{trade_off}	Chỉ số đánh đổi công bằng và hiệu suất
\mathcal{X}	Tập huấn luyện
\mathbf{x}	Đầu vào
y	Đầu ra thực tế $\in \{0, 1\}$
\hat{y}	Đầu ra dự đoán $\in \{0, 1\}$
$P(\hat{Y} = 1 A = 0)$	Xác suất dự đoán thuộc lớp dương đối với nhóm ưu thế.
$P(\hat{Y} = 1 A = 1)$	Xác suất dự đoán thuộc lớp dương đối với nhóm yếu thế.
A	Thuộc tính nhạy cảm
$A = 0$	Nhóm ưu thế
$A = 1$	Nhóm yếu thế
TP	Giá trị dương tính thực (True Positive)
TN	Giá trị âm tính thực (True Negative)
FP	Giá trị dương tính giả (False Positive)
FN	Giá trị âm tính giả (False Negative)
TPR	Tỷ lệ dương tính thực (True Positive Rate, hay Recall)
FPR	Tỷ lệ dương tính giả (False Positive Rate)
$TPR_{A=0}$	Tỷ lệ dương tính thực của nhóm không đặc quyền
$TPR_{A=1}$	Tỷ lệ dương tính thực của nhóm đặc quyền
$FPR_{A=0}$	Tỷ lệ dương tính giả của nhóm không đặc quyền
$FPR_{A=1}$	Tỷ lệ dương tính giả của nhóm đặc quyền
p -value	Giá trị xác suất dùng để kiểm định giả thuyết thống kê (Wald test)
d	Tổng số thuộc tính của dữ liệu
k	Số lượng thuộc tính nhạy cảm
D_{tr}	Tập dữ liệu huấn luyện gồm các mẫu $\langle x_j, y_j \rangle$
$\langle x_j, y_j \rangle$	Mẫu dữ liệu thứ j gồm véc-tơ đặc trưng x_j và nhãn y_j
$x_j = [x_1^j, \dots, x_d^j]$	Véc-tơ đặc trưng của mẫu j với tổng cộng d thuộc tính
x_1^j, \dots, x_k^j	Các thuộc tính nhạy cảm của mẫu j
x_{k+1}^j, \dots, x_d^j	Các thuộc tính không nhạy cảm của mẫu j
x_i^{new}	véc-tơ đặc trưng mới của x_i (sau khi loại bỏ ảnh hưởng từ các thuộc tính nhạy cảm)

Kí hiệu	Mô tả
x^{te} = $[x_1^{te}, \dots, x_d^{te}]$	Véc-tơ đặc trưng của mẫu kiểm tra
\mathbb{R}^d	Không gian đặc trưng d -chiều
$\{0, 1\}$	Tập nhãn cho bài toán phân loại nhị phân (0: âm, 1: dương)
S_{ML}	Mô hình học máy được huấn luyện từ dữ liệu đã điều chỉnh
$S_{ML}(x^{te})$	Nhãn dự đoán của mô hình S_{ML} cho mẫu kiểm tra
$E_a[k+1:d]$	Mảng hệ số chệch (intercept) dùng để điều chỉnh các thuộc tính không nhạy cảm
E_{b^1}, \dots, E_{b^k}	Các mảng hệ số hồi quy tương ứng với từng thuộc tính nhạy cảm
V_1, \dots, V_k	Các véc-tơ cột ứng với thuộc tính nhạy cảm trong tập huấn luyện
V_i	Véc-tơ cột của thuộc tính không nhạy cảm thứ i
a_i	Hệ số chệch của mô hình hồi quy cho thuộc tính không nhạy cảm V_i
b_i^1, \dots, b_i^k	Hệ số hồi quy của các thuộc tính nhạy cảm khi dự đoán V_i
μ	Thành phần nhiễu trong mô hình hồi quy
$\hat{a}_i, \hat{b}_i^1, \dots, \hat{b}_i^k$	Các hệ số hồi quy ước lượng được từ dữ liệu huấn luyện sau kiểm định
D_{i1}	Nhóm con thứ i có nhãn mục tiêu $y = 1$
D_{i0}	Nhóm con thứ i có nhãn mục tiêu $y = 0$
D_{ij}	Nhóm con thứ i ứng với nhãn j , tổng cộng có 2^{k+1} nhóm
i_s, j_s	Cặp chỉ số xác định nhóm con gốc có số bản ghi nhỏ nhất
$D_{i_s j_s}$	Nhóm con gốc, được chọn bằng D_{\min}
$D'_{i_s j_s}$	Nhóm con tổng hợp tương ứng với nhóm gốc $D_{i_s j_s}$
n_0	Kích thước mục tiêu (số bản ghi mong muốn) cho nhóm con tổng hợp gốc, $n_0 > \ D_{i_s j_s}\ $
R_{mean}	Tỷ lệ trung bình giữa số bản ghi nhãn dương và nhãn âm trong tất cả các nhóm con
$\ D'_{i1}\ , \ D'_{i0}\ $	Số lượng bản ghi dự kiến của các nhóm con tổng hợp ứng với nhãn dương (1) và nhãn âm (0)
D'	Tập dữ liệu đã được cân bằng, thu được bằng cách hợp nhất tất cả các nhóm con tổng hợp D'_{ij}
D'_{ij}	Nhóm con tổng hợp ứng với nhóm D_{ij} sau khi cân bằng
O_GT	Thuộc tính Giới tính trong bộ dữ liệu Oulad
O_K.tật	Thuộc tính Khuyết tật trong bộ dữ liệu Oulad

Kí hiệu	Mô tả
SP_GT	Thuộc tính Giới tính trong bộ dữ liệu Student Performance
SP_SK	Thuộc tính Sức khỏe trong bộ dữ liệu Student Performance
SD_Nợ	Thuộc tính Tình trạng nợ trong bộ dữ liệu Student Droptout
SD_GT	Thuộc tính Giới tính trong bộ dữ liệu Student Droptout
DNU_GT	Thuộc tính Giới tính trong bộ dữ liệu DNU
DNU_Tuổi	Thuộc tính Tuổi trong bộ dữ liệu DNU
DNU_KV	Thuộc tính Khu vực trong bộ dữ liệu DNU

Lời cam đoan

Tôi xin cam đoan đây là công trình nghiên cứu do tôi thực hiện dưới sự hướng dẫn của GS. TS Nguyễn Đức Anh tại Khoa Công nghệ thông tin, trường Đại học Đông Nam Na Uy và PGS. TS. Phạm Ngọc Hùng tại Khoa Công nghệ thông tin, trường Đại học Công nghệ, Đại học Quốc gia Hà Nội. Các số liệu và kết quả trình bày trong luận án là trung thực, chưa được công bố bởi bất kỳ tác giả nào hay ở bất kỳ công trình nào khác.

Tác giả

Phạm Thị Tố Nga

Lời cảm ơn

Trước tiên tôi xin gửi lời cảm ơn chân thành và sâu sắc đến thầy giáo, GS.TS Nguyễn Đức Anh và PGS. TS. Phạm Ngọc Hùng - những người đã hướng dẫn, khuyến khích, truyền cảm hứng, chỉ bảo và tạo cho tôi những điều kiện tốt nhất từ khi bắt đầu làm nghiên cứu sinh đến khi hoàn thành luận án này.

Tôi xin chân thành cảm ơn các thầy cô giáo khoa Công nghệ thông tin, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, đặc biệt là các Thầy Cô trong Bộ môn Khoa học máy tính và Bộ môn Công nghệ Phần mềm đã tận tình đào tạo, cung cấp cho tôi những kiến thức vô cùng quý giá, đã tạo điều kiện tốt nhất cho tôi về môi trường làm việc trong suốt quá trình học tập và nghiên cứu.

Tôi xin trân trọng cảm ơn Phòng Đào tạo và Ban Giám hiệu Trường Đại học Công nghệ đã tạo điều kiện thuận lợi cho tôi trong suốt quá trình học tập và thực hiện luận án.

Tôi xin bày tỏ lòng biết ơn đến Hội đồng trường, Ban Giám hiệu và Khoa Công nghệ thông tin Trường Đại học Đại Nam đã tạo điều kiện làm việc thuận lợi để tôi tập trung học tập và thực hiện luận án.

Tôi cũng xin được gửi lời cảm ơn đến tất cả đến các thành viên trong nhóm nghiên cứu tại Phòng thí nghiệm đảm bảo chất lượng phần mềm, đặc biệt là em Đỗ Minh Khá đã đồng hành cùng tôi trên chặng đường này.

Cuối cùng, tôi xin bày tỏ lòng biết ơn vô hạn đối với cha, mẹ, chồng, các con, cùng người thân trong gia đình và bạn bè, đồng nghiệp đã luôn ủng hộ và yêu thương tôi một cách vô điều kiện. Nếu không có sự ủng hộ của của tất cả mọi người tôi không thể hoàn thành được luận án này.

Tóm tắt

Trong bối cảnh trí tuệ nhân tạo và học máy ngày càng được tích hợp sâu rộng vào các hệ thống hỗ trợ ra quyết định, vấn đề đảm bảo công bằng cho các mô hình học máy đã trở thành một yêu cầu cấp thiết. Trong lĩnh vực giáo dục, công bằng có ý nghĩa đặc biệt quan trọng vì các quyết định tự động của hệ thống có thể ảnh hưởng trực tiếp đến cơ hội học tập, đánh giá kết quả và quá trình phát triển của người học. Tuy nhiên, khi các mô hình học máy được huấn luyện trên dữ liệu mất cân bằng hoặc chứa các thuộc tính nhạy cảm như *giới tính*, *chủng tộc*, *sức khỏe*, *khu vực*, v.v., chúng dễ tạo ra kết quả không công bằng giữa các nhóm người học khác nhau.

Luận án này tập trung nghiên cứu và đề xuất các phương pháp kỹ thuật nhằm nâng cao tính công bằng của mô hình học máy ứng dụng cho dữ liệu giáo dục, đặc biệt trong trường hợp cần xem xét đồng thời nhiều thuộc tính nhạy cảm và mối quan hệ giao thoa giữa chúng. Trên cơ sở phân tích khoảng trống nghiên cứu hiện tại, luận án đề xuất ba phương pháp chính:

- *Phương pháp Fairedu* – một kỹ thuật tiền xử lý dựa trên hồi quy đa biến, có khả năng loại bỏ sự phụ thuộc giữa các biến đầu vào và nhiều thuộc tính nhạy cảm đồng thời. Phương pháp này giúp giảm rò rỉ thông tin nhạy cảm trong quá trình huấn luyện, qua đó nâng cao công bằng cho mô hình.
- *Phương pháp DPF* – một kỹ thuật cân bằng dữ liệu giữa các nhóm giao thoa bằng cách sử dụng các mô hình sinh dữ liệu tổng hợp như CTGAN và LLM. Cách tiếp cận này đảm bảo các nhóm thiểu số được đại diện công bằng hơn trong dữ liệu huấn luyện.
- *Phương pháp FaireduPlus* – một khung tiếp cận tích hợp Fairedu và DPF, xử lý công bằng hai chiều: “chiều ngang” (cân bằng dữ liệu) và “chiều dọc” (loại bỏ phụ thuộc).
- Ngoài ra, luận án còn đề xuất chỉ số đánh giá mới để định lượng mức độ đánh đổi giữa công bằng và hiệu suất của mô hình.

Các thí nghiệm được triển khai trên bốn bộ dữ liệu giáo dục (gồm ba bộ

công khai và một bộ dữ liệu thực tế từ Trường Đại học Đại Nam), với năm mô hình học máy phổ biến gồm *Hồi quy logistic*, *Cây quyết định*, *Rừng ngẫu nhiên*, *Tăng cường gradient* và *Mạng nơ ron thần kinh*. Kết quả thực nghiệm cho thấy các phương pháp đề xuất đạt hiệu quả vượt trội, cải thiện các chỉ số công bằng như “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch cơ hội công bằng*” và “*chênh lệch trung bình xác suất*” đồng thời không làm suy giảm hiệu suất dự đoán của mô hình.

Về mặt lý thuyết, luận án đóng góp một khung phương pháp toàn diện cho bài toán công bằng với đồng thời nhiều thuộc tính nhạy cảm cho các mô hình học máy trong lĩnh vực giáo dục, đồng thời giới thiệu chỉ số đánh đổi công bằng – hiệu suất mới, phản ánh rõ mối quan hệ giữa hai mục tiêu này. Về mặt thực tiễn, các phương pháp được đề xuất có thể áp dụng hiệu quả trong các hệ thống giáo dục thông minh, hỗ trợ quá trình dự đoán, đánh giá và ra quyết định một cách công bằng hơn, giúp đảm bảo cơ hội học tập công bằng cho mọi sinh viên. Kết quả của luận án không chỉ góp phần hoàn thiện cơ sở lý thuyết về công bằng cho các hệ thống học máy mà còn mở ra hướng ứng dụng thực tế cho các hệ thống trí tuệ nhân tạo đáng tin cậy, minh bạch và có trách nhiệm trong giáo dục cũng như các lĩnh vực khác có tác động xã hội sâu rộng.

GIỚI THIỆU

0.1. Đặt vấn đề

0.1.1. Bối cảnh nghiên cứu về công bằng cho hệ thống học máy trong lĩnh vực giáo dục

Các ứng dụng của Học máy (ML) và Trí tuệ nhân tạo (AI) đang ngày càng được triển khai rộng rãi trong giáo dục nhằm hỗ trợ ra quyết định, cá nhân hóa học tập, dự đoán kết quả học tập và phát hiện nguy cơ bỏ học [3, 4, 5]. Những hệ thống này góp phần nâng cao hiệu quả quản lý, hỗ trợ giảng viên trong đánh giá người học và mang lại trải nghiệm học tập cá nhân hóa cho sinh viên [6, 7, 8]. Đồng thời, các mô hình dự đoán cũng giúp nhà quản lý và giảng viên đưa ra các can thiệp kịp thời nhằm hỗ trợ người học có nguy cơ gặp khó khăn [9, 10].

Tuy nhiên, cùng với những lợi ích đó, vấn đề công bằng trong các hệ thống học máy giáo dục ngày càng được quan tâm. Công bằng gắn liền với nguyên tắc bình đẳng về cơ hội học tập và giảm khoảng cách thành tích giữa các nhóm người học [11, 12]. Nhiều nghiên cứu cho thấy các hệ thống dự đoán hoặc đánh giá học tập có thể tạo ra hoặc khuếch đại thiên lệch đối với các nhóm yếu thế, chẳng hạn theo giới tính, chủng tộc, điều kiện kinh tế – xã hội hoặc khu vực sinh sống [13, 14, 15]. Trong giáo dục, những sai lệch này không chỉ gây ra vấn đề kỹ thuật mà còn có thể ảnh hưởng trực tiếp đến động lực học tập, cơ hội nghề nghiệp và sự phát triển lâu dài của người học.

Nhiều nghiên cứu trong kỹ thuật phần mềm đã xem công bằng như một thuộc tính chất lượng quan trọng của hệ thống AI/ML [1, 16]. Một hệ thống học máy chỉ thực sự đáng tin cậy khi tính công bằng của nó được đánh giá và kiểm chứng rõ ràng [17]. Vì vậy, công bằng cần được tích hợp xuyên suốt trong toàn bộ vòng đời phát triển hệ thống, từ phân tích yêu cầu, thiết kế đến triển khai và đánh giá [18].

Trong những năm gần đây, việc khám phá, đo lường và đảm bảo tính công bằng trong các hệ thống học máy đã trở thành một hướng nghiên cứu quan trọng, thu hút sự quan tâm rộng rãi từ cộng đồng học thuật [15, 19]. Nhiều công trình tập trung phát triển các thuật toán học máy công bằng nhằm giảm thiểu sự phân biệt đối xử đối với các nhóm thiểu số [20]. Bên cạnh đó, lĩnh vực AI có khả năng giải thích (Explainable Artificial Intelligence – XAI) cũng được phát triển nhằm nâng cao tính minh bạch và khả năng diễn giải của các hệ thống AI [21, 22]. Năm 2018, bộ công cụ mã nguồn mở AI Fairness 360 được giới thiệu, cung cấp nhiều độ đo và thuật toán giảm thiên vị cho dữ liệu và mô hình [23]. Đồng thời, nhiều nghiên cứu tổng quan đã hệ thống hóa tri thức về công bằng trong AI/ML nói chung [1, 24, 25, 26, 27] cũng như các ứng dụng AI/ML trong giáo dục [28, 29].

Nhiều phương pháp kỹ thuật đã được đề xuất nhằm cải thiện công bằng trong các hệ thống học máy. Một hướng tiếp cận phổ biến là các kỹ thuật tiền xử lý nhằm giảm thiên vị trong dữ liệu huấn luyện [20]. Tiêu biểu là phương pháp Reweighting trong bộ công cụ AI Fairness 360, trong đó trọng số của các mẫu huấn luyện được điều chỉnh để cân bằng phân phối nhân giữa các nhóm nhạy cảm, bao gồm cả các nhóm giao thoa [23, 30]. Tuy nhiên, do không thay đổi phân phối dữ liệu gốc, phương pháp này khó khắc phục tình trạng thiếu đại diện của các nhóm thiểu số khi dữ liệu bị mất cân bằng nghiêm trọng. Một hướng tiếp cận khác là loại bỏ sự phụ thuộc giữa các thuộc tính không nhạy cảm và thuộc tính nhạy cảm. Chẳng hạn, Li và cộng sự đề xuất phương pháp *phương pháp gỡ lỗi dữ liệu đào tạo bằng hồi quy tuyến tính*, sử dụng hồi quy tuyến tính để tách ảnh hưởng của thuộc tính nhạy cảm khỏi các thuộc tính không nhạy cảm thông qua phần dư của mô hình hồi quy [31]. Tuy nhiên, phương pháp này chủ yếu xử lý từng cặp thuộc tính riêng lẻ và dựa trên giả định quan hệ tuyến tính.

Ngoài ra, các phương pháp sinh dữ liệu tổng hợp cũng được nghiên cứu nhằm cải thiện công bằng bằng cách tái cân bằng dữ liệu huấn luyện. Một ví dụ tiêu biểu là TabFairGAN, được Rajabi và Garibay đề xuất nhằm sinh dữ liệu bằng công bằng dựa trên kiến trúc GAN có điều kiện [32]. Phương pháp này học phân phối dữ liệu theo từng nhóm nhạy cảm và bổ sung thành phần mất mát công bằng trong quá trình huấn luyện. Tuy nhiên, TabFairGAN chủ yếu được thiết

kế cho trường hợp một thuộc tính nhạy cảm; khi số lượng thuộc tính nhạy cảm tăng, số nhóm giao thoa cũng tăng nhanh, gây khó khăn trong việc kiểm soát phân phối dữ liệu. Bên cạnh đó, Chen và cộng sự thực hiện một nghiên cứu thực nghiệm quy mô nhằm đánh giá hiệu quả của nhiều phương pháp cải thiện công bằng trong bối cảnh nhiều thuộc tính nhạy cảm và các nhóm giao thoa [33]. Kết quả cho thấy nhiều phương pháp hiện có vẫn gặp khó khăn khi xử lý công bằng giao thoa cũng như trong việc cân bằng giữa công bằng và hiệu suất dự đoán.

Một vấn đề quan trọng khác trong nghiên cứu đảm bảo tính công bằng cho các hệ thống học máy là vấn đề liên quan đến dữ liệu. Trong bối cảnh giáo dục, nơi quyền riêng tư của người học được coi trọng, việc sinh dữ liệu tổng hợp (synthetic data generation – SDG) nổi lên như một hướng tiếp cận quan trọng, vừa giải quyết thách thức về khan hiếm dữ liệu, vừa hỗ trợ bảo mật thông tin. Bằng cách tạo ra dữ liệu nhân tạo mô phỏng cấu trúc và đặc tính thống kê của dữ liệu thực, các nhà nghiên cứu có thể kiểm chứng giả thuyết, huấn luyện mô hình và đánh giá công bằng mà không phụ thuộc hoàn toàn vào dữ liệu có chứa thông tin nhạy cảm. Nhiều kỹ thuật sinh dữ liệu tổng hợp đã được triển khai trong giáo dục cũng như trong các nghiên cứu về công bằng. Trước hết, các phương pháp Bayes tận dụng tri thức miền để xây dựng mạng Bayes và bảng xác suất có điều kiện, từ đó tạo ra dữ liệu có độ tương đồng cao với dữ liệu thực và thuận lợi cho việc kiểm chứng mô hình [34]. Tiếp đến, một số nghiên cứu so sánh chỉ ra rằng các mô hình xác suất thường mang lại hiệu quả cao hơn so với các phương pháp học sâu như mạng đối nghịch sinh (Generative Adversarial Network – GAN) trong bối cảnh dữ liệu giáo dục. Điển hình là nghiên cứu của Combrink và cộng sự, trong đó mô hình xác suất đạt độ chính xác 75% so với 38% của mô hình học sâu [35]. Ngoài ra, các kỹ thuật bảo vệ quyền riêng tư cũng đóng vai trò quan trọng, vừa duy trì tính hữu ích của dữ liệu tổng hợp cho nghiên cứu, vừa giảm thiểu rủi ro rò rỉ thông tin cá nhân [36]. Bên cạnh đó, SDG vẫn đối mặt với thách thức về khả năng phản ánh chính xác kịch bản thực tế cũng như nguy cơ bị lạm dụng. Do đó, cần có sự đánh giá liên tục và cải tiến kỹ thuật để tối đa hóa giá trị ứng dụng, không chỉ phục vụ nghiên cứu giáo dục mà còn góp phần thúc đẩy công bằng trong các hệ thống học máy.

Giáo dục được lựa chọn làm miền nghiên cứu trọng tâm trong luận án vì đây là lĩnh vực có mức độ nhạy cảm cao về công bằng, nơi các mô hình học máy

tham gia trực tiếp vào những quyết định ảnh hưởng lâu dài đến cơ hội học tập và phát triển của người học. Đồng thời, đặc trưng dữ liệu giáo dục với nhiều thuộc tính nhạy cảm đồng thời và các nhóm giao thoa có quy mô nhỏ tạo ra một bối cảnh điển hình để nghiên cứu công bằng giao thoa cũng như mối quan hệ đánh đổi giữa công bằng và hiệu suất.

0.1.2. Các thách thức và khoảng trống nghiên cứu

Nghiên cứu về đảm bảo tính công bằng cho các hệ thống học máy trong giáo dục đang đối mặt với nhiều thách thức đặc thù do tính phức tạp cả về kỹ thuật lẫn bối cảnh ứng dụng. Các thách thức chính có thể được khái quát như sau.

Thứ nhất: chưa thống nhất định nghĩa và thước đo công bằng. Các khái niệm và độ đo công bằng hiện nay vừa mang tính bổ sung vừa có thể mâu thuẫn lẫn nhau, dẫn đến khó khăn trong việc lựa chọn tiêu chí đánh giá phù hợp. Một mô hình có thể thỏa mãn một tiêu chí công bằng nhưng lại vi phạm tiêu chí khác, trong khi chưa tồn tại một khung đánh giá tổng hợp có khả năng dung hòa các quan điểm này.

Thứ hai: hạn chế trong việc xử lý đồng thời nhiều thuộc tính nhạy cảm. Phần lớn các nghiên cứu hiện tại chỉ tập trung vào một thuộc tính nhạy cảm đơn lẻ, trong khi dữ liệu giáo dục thường mang tính đa chiều với nhiều yếu tố như *giới tính, chủng tộc, tình trạng khuyết tật và khu vực*. Việc bỏ qua các tương tác giữa các thuộc tính này có thể dẫn đến hiện tượng cải thiện được tính công bằng cho thuộc tính này nhưng lại vô tình làm trầm trọng hơn cho thuộc tính kia khi đánh giá công bằng.

Thứ ba: sự đánh đổi giữa công bằng và hiệu suất mô hình. Việc cải thiện công bằng thường đi kèm với sự suy giảm về độ chính xác hoặc khả năng dự đoán. Trong bối cảnh giáo dục, nơi các quyết định như dự đoán nguy cơ bỏ học hoặc đánh giá kết quả học tập có ảnh hưởng trực tiếp đến người học, việc cân bằng hai mục tiêu này trở thành một bài toán quan trọng nhưng chưa có lời giải tối ưu.

Thứ tư: hạn chế về dữ liệu giáo dục. Các bộ dữ liệu giáo dục thường có quy mô nhỏ, mất cân bằng giữa các nhóm và thiếu đại diện cho các nhóm thiểu số.

Đồng thời, các rào cản về quyền riêng tư và bảo mật thông tin cá nhân làm hạn chế khả năng chia sẻ và tái sử dụng dữ liệu. Những yếu tố này gây khó khăn cho việc xây dựng và kiểm chứng các mô hình học máy đảm bảo cả hiệu suất và công bằng.

Thứ năm: ảnh hưởng của yếu tố con người. Công bằng trong hệ thống học máy không chỉ phụ thuộc vào thuật toán mà còn chịu tác động từ các quyết định trong quá trình thu thập dữ liệu, lựa chọn thuộc tính và thiết kế mô hình. Ngoài ra, sự khác biệt trong cách hiểu về công bằng giữa các bên liên quan có thể dẫn đến các tiêu chí đánh giá không nhất quán, làm gia tăng độ phức tạp trong triển khai thực tế.

Từ các thách thức đã nêu, có thể xác định một số khoảng trống nghiên cứu chính như sau. (1) Xuất phát từ sự thiếu thống nhất về định nghĩa công bằng, hiện chưa có khung đánh giá tích hợp cho phép kết hợp và so sánh các thước đo công bằng một cách nhất quán. (2) Từ hạn chế trong xử lý nhiều thuộc tính nhạy cảm, các phương pháp hiện nay vẫn thiếu khả năng đảm bảo công bằng đa thuộc tính. (3) Liên quan đến bài toán đánh đổi, còn thiếu các cơ chế định lượng và tối ưu sự đánh đổi giữa công bằng và hiệu suất. (4) Do hạn chế dữ liệu giáo dục, việc khai thác dữ liệu tổng hợp để hỗ trợ tái cân bằng và cải thiện công bằng vẫn chưa được nghiên cứu đầy đủ. (5) Từ yếu tố con người, còn thiếu các cách tiếp cận toàn diện kết hợp giữa kỹ thuật và bối cảnh ứng dụng trong đánh giá công bằng.

Những khoảng trống này cho thấy nhu cầu phát triển các phương pháp tích hợp nhằm đồng thời xử lý đa thuộc tính nhạy cảm, tận dụng dữ liệu hiệu quả và kiểm soát tốt sự đánh đổi giữa công bằng và hiệu suất.

0.2. Mục tiêu, đối tượng, phương pháp, và phạm vi nghiên cứu

0.2.1. Mục tiêu nghiên cứu

Xuất phát từ các khoảng trống nghiên cứu đã nêu ở Mục 0.1.2, luận án hướng tới mục tiêu tổng quát là tìm ra các phương pháp đảm bảo tính công bằng cho các hệ thống học máy trong giáo dục, trong bối cảnh dữ liệu dạng bảng có nhiều thuộc tính nhạy cảm và mất cân bằng giữa các nhóm người học. Nghiên cứu tiếp cận công bằng như một yêu cầu cốt lõi, đồng thời phân tích các nguồn gốc gây thiên lệch trong dữ liệu, đặc biệt là sự phụ thuộc giữa thuộc tính không nhạy cảm và thuộc tính nhạy cảm cũng như sự thiếu đại diện của các nhóm yếu thế. Trên cơ sở đó, luận án đề xuất các giải pháp can thiệp nhằm cải thiện công bằng ngay ở khâu tiền xử lý, đồng thời duy trì hiệu suất dự báo ở mức chấp nhận được, và làm rõ mối quan hệ đánh đổi giữa công bằng và hiệu suất.

Mục tiêu cụ thể của luận án bao gồm:

- Nghiên cứu các cơ chế tiền xử lý nhằm giảm sự phụ thuộc giữa các thuộc tính không nhạy cảm vào các thuộc tính nhạy cảm, góp phần nâng cao tính công bằng của mô hình.
- Khai thác dữ liệu tổng hợp để cải thiện tính đại diện và cân bằng của các nhóm nhạy cảm trong dữ liệu huấn luyện.
- Xây dựng cách tiếp cận kết hợp các cơ chế can thiệp nhằm nâng cao công bằng mà không làm suy giảm đáng kể hiệu suất mô hình.
- Đề xuất tiêu chí đánh giá nhằm phân tích và định lượng sự đánh đổi giữa công bằng và hiệu suất.

Thông qua các mục tiêu trên, luận án làm rõ cách đảm bảo công bằng cho các hệ thống học máy trong lĩnh vực giáo dục bằng phân tích lý thuyết và thực nghiệm. Các mục tiêu này tạo nền tảng cho việc nghiên cứu, đánh giá và triển khai các giải pháp cải thiện công bằng một cách có hệ thống và đảm bảo tính khả thi.

0.2.2. Đối tượng và phương pháp nghiên cứu

Đối tượng nghiên cứu của luận án là các hệ thống học máy ứng dụng trong lĩnh vực giáo dục, đặc biệt là những mô hình khai thác dữ liệu dạng bảng để dự đoán, đánh giá và hỗ trợ ra quyết định. Các mô hình được xem xét bao gồm nhóm thuật toán học máy truyền thống như *Hồi quy logistic*, *Cây quyết định*, *Rừng ngẫu nhiên*, các mô hình tăng cường như *Tăng cường gradient* và *Mạng nơ ron thần kinh* ở mức cơ bản.

Phương pháp nghiên cứu kết hợp tiếp cận định tính và định lượng. Về định tính, luận án tiến hành phân tích, tổng hợp các công trình nghiên cứu liên quan nhằm xác định thách thức, khoảng trống và xu hướng phát triển trong việc bảo đảm tính công bằng cho hệ thống học máy trong lĩnh vực giáo dục. Về định lượng, nghiên cứu triển khai thực nghiệm trên nhiều bộ dữ liệu giáo dục phổ biến (*Student Performance*, *Student Predict Dropout*, *Oulad*, *DNU Data*) chứa từ hai thuộc tính nhạy cảm trở lên bao gồm các thuộc tính như: *giới tính*, *tuổi*, *sức khỏe*, *tình trạng nợ*, *khu vực*, v.v.. Phương pháp can thiệp nhằm đảm bảo tính công bằng được sử dụng là phương pháp tiền xử lý và so sánh thông qua các thước đo công bằng như “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*”, và “*chênh lệch cơ hội công bằng*” và các thước đo hiệu suất bao gồm “*độ chuẩn xác*”, “*độ hồi tưởng*”, “*độ chính xác*”, và “*điểm số F1*”.

0.2.3. Phạm vi nghiên cứu

Luận án tập trung nghiên cứu các hệ thống học máy sử dụng dữ liệu dạng bảng ứng dụng trong lĩnh vực giáo dục. Các mô hình được xem xét bao gồm những thuật toán học máy có giám sát truyền thống như *Hồi quy logistic*, *Cây quyết định*, *Rừng ngẫu nhiên*, cùng với các mô hình tăng cường *Tăng cường gradient* và *Mạng nơ ron thần kinh* ở mức đơn giản.

Về dữ liệu, luận án sử dụng ba bộ dữ liệu mở phổ biến trong nghiên cứu AI/ML cho lĩnh vực giáo dục, bao gồm: *Student Performance*, *Student Predict Dropout*, *Oulad*, và một bộ dữ liệu riêng được lấy từ khoa Công nghệ thông tin,

trường Đại học Đại Nam là *DNU Data*. Các bộ dữ liệu này được lựa chọn vì phản ánh đa dạng bối cảnh giáo dục và chứa nhiều thuộc tính nhạy cảm cần được xem xét trong đánh giá công bằng.

Về thuộc tính nhạy cảm, phạm vi tập trung vào các thuộc tính nhạy cảm thường gặp trong các nghiên cứu về lĩnh vực giáo dục như: *giới tính, tình trạng khuyết tật, tuổi, sức khỏe, tình trạng nợ, khu vực*.

Phạm vi nghiên cứu của luận án không chỉ được xác định theo các yếu tố kỹ thuật như dữ liệu, mô hình học máy và thuộc tính nhạy cảm, mà còn được định hướng bởi các vấn đề cốt lõi về đảm bảo tính công bằng trong các hệ thống học máy ứng dụng cho giáo dục. Trên cơ sở đó, luận án tiếp cận công bằng như một yêu cầu mang tính hệ thống, gắn liền với đặc thù dữ liệu giáo dục và mục tiêu ứng dụng thực tiễn, thay vì xem công bằng như một ràng buộc đơn lẻ trong quá trình tối ưu mô hình.

Nhằm hệ thống hóa các vấn đề nghiên cứu, các hướng tiếp cận và làm rõ mối quan hệ giữa các đóng góp của luận án, Hình 1 trình bày cây nghiên cứu tổng quát về đảm bảo công bằng cho các hệ thống AI/ML trong lĩnh vực giáo dục. Cây nghiên cứu được tổ chức theo bốn nhánh chính, tương ứng với bốn trục nghiên cứu xuyên suốt của luận án: (i) công bằng đa thuộc tính nhạy cảm, (ii) mất cân bằng dữ liệu và công bằng nhóm con, (iii) can thiệp công bằng trên dữ liệu dạng bảng, và (iv) vấn đề đánh đổi giữa công bằng và hiệu suất của mô hình. Các nhánh này đồng thời cũng xác định phạm vi và trọng tâm mà luận án tập trung khai thác.

Nhánh 1 – Công bằng đa thuộc tính nhạy cảm: Nhánh này tập trung vào vấn đề đảm bảo công bằng đồng thời cho nhiều thuộc tính nhạy cảm trong dữ liệu giáo dục, trong bối cảnh tồn tại sự phụ thuộc giữa các thuộc tính không nhạy cảm và các thuộc tính nhạy cảm. Thực tế cho thấy phần lớn các nghiên cứu trước đây chủ yếu xử lý công bằng đối với một thuộc tính nhạy cảm đơn lẻ, dẫn đến hạn chế khi áp dụng cho dữ liệu giáo dục có tính đa chiều. Trong phạm vi này, luận án tiếp cận công bằng theo hướng tiên xử lý, nhằm giảm thiểu và loại bỏ sự phụ thuộc của dữ liệu huấn luyện vào các thuộc tính nhạy cảm, qua đó cải thiện công bằng ở mức nhóm cho nhiều thuộc tính đồng thời.

Nhánh 2 – Mất cân bằng dữ liệu và công bằng nhóm con: Nhánh này tập trung vào hiện tượng mất cân bằng dữ liệu, đặc biệt là sự mất cân bằng giữa các nhóm con được hình thành từ các tổ hợp thuộc tính nhạy cảm. Đây là vấn đề phổ biến trong dữ liệu giáo dục, đặc biệt khi đồng thời xem xét nhiều thuộc tính nhạy cảm. Việc phân chia dữ liệu theo các nhóm giao thoa có thể làm xuất hiện các nhóm con có rất ít hoặc không có mẫu, gây khó khăn cho việc huấn luyện mô hình và đảm bảo công bằng. Phạm vi nghiên cứu ở nhánh này hướng tới các cơ chế can thiệp vào phân bố dữ liệu thông qua phân hoạch và sinh dữ liệu tổng hợp, nhằm cải thiện tính đại diện và cân bằng giữa các nhóm con trong dữ liệu huấn luyện.

Nhánh 3 – Can thiệp công bằng trên dữ liệu dạng bảng: Nhánh này mở rộng phạm vi nghiên cứu sang việc xử lý công bằng trong bối cảnh dữ liệu dạng bảng, nơi đồng thời tồn tại cả sự phụ thuộc vào thuộc tính nhạy cảm và hiện tượng mất cân bằng dữ liệu. Thay vì các cách tiếp cận can thiệp đơn cơ chế, luận án tập trung vào việc kết hợp nhiều cơ chế can thiệp nhằm xử lý công bằng theo cả hai chiều của dữ liệu. Cụ thể, phạm vi nghiên cứu bao gồm các can thiệp theo chiều dọc (liên quan đến mối quan hệ giữa thuộc tính không nhạy cảm và thuộc tính nhạy cảm) và can thiệp theo chiều ngang (liên quan đến phân bố dữ liệu), từ đó hình thành cách tiếp cận can thiệp hai chiều trên dữ liệu dạng bảng.

Nhánh 4 – Đánh đổi giữa công bằng và hiệu suất mô hình: Nhánh này tập trung vào vấn đề đánh giá và so sánh các phương pháp đảm bảo công bằng trong bối cảnh tồn tại sự đánh đổi giữa công bằng và hiệu suất dự báo. Thực tế cho thấy các chỉ số công bằng và hiệu suất thường được đánh giá rời rạc, gây khó khăn cho việc lựa chọn mô hình trong ứng dụng thực tiễn. Do đó, phạm vi nghiên cứu của luận án bao gồm việc xây dựng và sử dụng một chỉ số đánh giá tổng hợp, dựa trên quan điểm cải thiện tương đối và tiếp cận Pareto, nhằm hỗ trợ phân tích và so sánh các cấu hình mô hình một cách có hệ thống.

Như vậy, phạm vi nghiên cứu của luận án không chỉ giới hạn ở dữ liệu và mô hình học máy, mà còn mở rộng sang các khía cạnh liên quan đến định nghĩa, đo lường và cơ chế can thiệp nhằm đảm bảo tính công bằng. Trên cơ sở đó, luận án hướng tới việc hình thành một khung nghiên cứu có tính hệ thống cho bài toán công bằng trong các hệ thống học máy ứng dụng cho giáo dục.



Hình 1: Cây nghiên cứu về tính công bằng cho các hệ thống học máy trong giáo dục liên quan đến luận án.

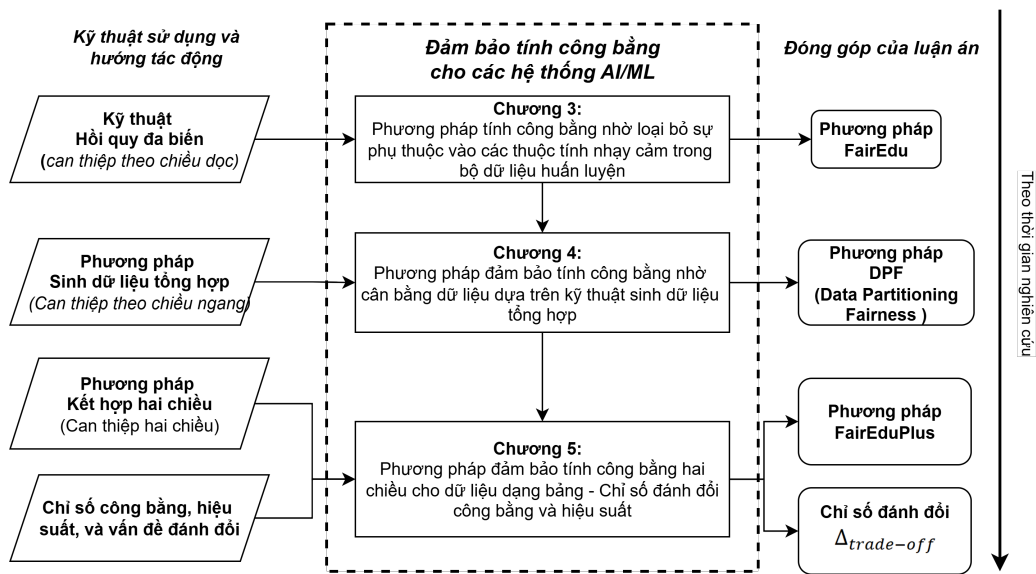
0.3. Các đóng góp chính của luận án

Dựa vào bốn mục tiêu cụ thể nêu ra trong Mục 0.2.1, luận án có bốn đóng góp chính, cụ thể như sau. *Một là, phát triển phương pháp Fairedu* nhằm nâng cao tính công bằng cho đồng thời nhiều thuộc tính nhạy cảm thông qua việc loại bỏ sự phụ thuộc của các thuộc tính đầu vào trong bộ dữ liệu huấn luyện vào các thuộc tính nhạy cảm. *Hai là, xây dựng phương pháp DPF* dựa trên kỹ thuật sinh dữ liệu tổng hợp, nhằm cân bằng phân bố giữa các tổ hợp thuộc tính nhạy cảm, qua đó đảm bảo tính công bằng trong dữ liệu huấn luyện. *Ba là, phát triển phương pháp FaireduPlus*, đảm bảo tính công bằng hai chiều cho bộ dữ liệu dạng bảng với đồng thời nhiều thuộc tính nhạy cảm. *Bốn là, đề xuất “chỉ số đánh đổi”*, nhằm định lượng mối quan hệ giữa công bằng và hiệu suất mô hình từ đó hỗ trợ so sánh và lựa chọn mô hình một cách hệ thống.

Các kết quả này đóng góp các giải pháp quan trọng để đảm bảo công bằng đồng thời với nhiều thuộc tính nhạy cảm cho các hệ thống học máy trong lĩnh vực giáo dục. Các đóng góp này không chỉ có ý nghĩa đối với lĩnh vực Khoa học

máy tính, mà còn mang giá trị đặc biệt đối với các ngành hẹp. Trong Kỹ nghệ phần mềm, công bằng được nhìn nhận như một thuộc tính chất lượng phi chức năng quan trọng cần được kiểm chứng và đảm bảo. Trong Khai phá dữ liệu giáo dục, việc đảm bảo công bằng có ý nghĩa thiết thực vì các quyết định tự động của hệ thống AI/ML có thể tác động trực tiếp đến cơ hội học tập, kết quả học tập và sự phát triển lâu dài của người học.

Điểm nổi bật là các đóng góp này không tồn tại độc lập mà có mối quan hệ kế thừa và bổ trợ lẫn nhau. Mối quan hệ này được thể hiện trong một lộ trình nghiên cứu gồm ba giai đoạn, từ nền tảng lý thuyết đến giải pháp thực nghiệm, nhằm giải quyết toàn diện bài toán công bằng cho đồng thời nhiều thuộc tính nhạy cảm cho các hệ thống học máy trong lĩnh vực giáo dục, được minh họa trong Hình 2.



Hình 2: Mối quan hệ giữa các chương đề xuất phương pháp trong luận án.

Ba giai đoạn cụ thể như sau.

Giai đoạn thứ nhất, luận án tập trung nghiên cứu các khái niệm và phương pháp liên quan đến tính công bằng cho các hệ thống học máy ứng dụng trong lĩnh vực giáo dục, với trọng tâm là các kỹ thuật tiền xử lý nhằm loại bỏ sự phụ thuộc giữa các thuộc tính đầu vào và đồng thời nhiều thuộc tính nhạy cảm có trong dữ liệu huấn luyện. Nghiên cứu đề xuất phương pháp FairEdu, một kỹ thuật tiền xử lý dựa trên hồi quy tuyến tính đa biến, can thiệp theo chiều dọc

của dữ liệu bằng cách hiệu chỉnh các thuộc tính đầu vào nhằm giảm thiểu ảnh hưởng vào các thuộc tính nhạy cảm như *giới tính, chủng tộc, tuổi, sức khỏe, tình trạng nợ, khu vực*. Kết quả thực nghiệm cho thấy Fairedu cải thiện đáng kể các chỉ số công bằng như “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*”, “*chênh lệch cơ hội công bằng*” đối với đồng thời nhiều thuộc tính nhạy cảm mà vẫn duy trì hiệu suất dự đoán ổn định (xem chi tiết Chương 2).

Giai đoạn thứ hai, luận án mở rộng nghiên cứu sang việc sử dụng kỹ thuật sinh dữ liệu tổng hợp để tạo ra các tập dữ liệu huấn luyện cân bằng hơn cho các tổ hợp của các thuộc tính nhạy cảm (phương pháp DPF) (được trình bày chi tiết trong Chương 3). Phương pháp này áp dụng các kỹ thuật sinh dữ liệu như CTGAN và LLM để cân bằng phân bố dữ liệu theo tổ hợp các thuộc tính nhạy cảm, từ đó giảm thiểu thiên lệch tiềm ẩn trong dữ liệu huấn luyện. Kết quả thực nghiệm chứng minh rằng DPF có thể cải thiện các chỉ số công bằng một cách đáng kể cho đồng thời nhiều thuộc tính nhạy cảm có trong bộ dữ liệu, mặc dù vẫn tồn tại một mức đánh đổi nhất định về hiệu suất.

Giai đoạn thứ ba, luận án đề xuất phương pháp FaireduPlus, phương pháp này là sự mở rộng của Fairedu kết hợp với DPF nhằm đảm bảo tính công bằng toàn diện cho các hệ thống học máy sử dụng dữ liệu dạng bảng, theo cả chiều dọc (loại bỏ phụ thuộc) và chiều ngang (cân bằng dữ liệu bằng SDG). FaireduPlus được triển khai với hai kỹ thuật sinh dữ liệu tiêu biểu là CTGAN và LLM, qua đó tăng cường dữ liệu cho các tổ hợp thuộc tính nhạy cảm có số lượng hạn chế (chi tiết xem trong Chương 4). Ngoài ra, trong giai đoạn này, luận án còn đề xuất một “*chỉ số đánh đổi*” nhằm hỗ trợ cân bằng giữa công bằng và hiệu suất của mô hình khi đánh giá trên nhiều độ đo công bằng và hiệu suất khác nhau. Kết quả thực nghiệm cho thấy FaireduPlus cải thiện rõ rệt các chỉ số công bằng trên nhiều mô hình học máy, đặc biệt trong các kịch bản dữ liệu chứa đồng thời nhiều thuộc tính nhạy cảm. Tuy vậy, phương pháp này vẫn đối diện một số hạn chế, chẳng hạn như sự đánh đổi với hiệu suất và hiệu quả chưa cao đối với dữ liệu có sự mất cân bằng nghiêm trọng hoặc tồn tại các nhóm giao thoa không có đủ quan sát.

0.4. Bố cục của luận án

Luận án gồm hai phần mở đầu và kết luận cùng bốn chương nội dung chính. Phần Giới thiệu trình bày bối cảnh, vấn đề nghiên cứu, mục tiêu, phạm vi và các đóng góp của luận án. Phần Kết luận tổng hợp các kết quả đạt được, phân tích những hạn chế của nghiên cứu và đề xuất các hướng phát triển trong tương lai. Bốn chương chính được sắp xếp theo trình tự từ cơ sở lý thuyết đến các đóng góp phương pháp và thực nghiệm đánh giá.

Chương 1 trình bày tổng quan về công bằng trong các hệ thống học máy trong lĩnh vực giáo dục, bao gồm các khái niệm và định nghĩa về công bằng, các phương pháp đảm bảo công bằng, kỹ thuật sinh dữ liệu tổng hợp, các tiêu chí đánh giá công bằng và hiệu suất, cũng như vấn đề đánh đổi giữa hai mục tiêu này. Trên cơ sở tổng quan, chương xác định những khoảng trống nghiên cứu và các thách thức cần giải quyết, làm nền tảng cho các đề xuất của luận án.

Ba chương tiếp theo trình bày ba phương pháp do luận án đề xuất, tương ứng với ba giai đoạn nghiên cứu được giới thiệu trong Mục 0.3.

Chương 2 đề xuất phương pháp Fairedu, một kỹ thuật tiền xử lý nhằm loại bỏ sự phụ thuộc của các thuộc tính đầu vào vào đồng thời nhiều thuộc tính nhạy cảm, qua đó cải thiện công bằng mà vẫn duy trì hiệu suất dự đoán của mô hình.

Chương 3 trình bày phương pháp DPF, sử dụng kỹ thuật sinh dữ liệu tổng hợp để cân bằng phân bố dữ liệu theo các nhóm giao thoa của thuộc tính nhạy cảm, góp phần khắc phục tình trạng mất cân bằng dữ liệu và thiếu đại diện của các nhóm yếu thế.

Chương 4 trình bày phương pháp FaireduPlus, kết hợp Fairedu và DPF nhằm cải thiện công bằng theo cả hai hướng: giảm ảnh hưởng của thuộc tính nhạy cảm và cân bằng phân bố dữ liệu. Chương cũng đánh giá toàn diện hiệu quả của phương pháp trên nhiều bộ dữ liệu và mô hình học máy, đồng thời đề xuất chỉ số đánh đổi nhằm hỗ trợ định lượng và phân tích mối quan hệ giữa công bằng và hiệu suất của mô hình.

Chương 1

TỔNG QUAN VỀ VIỆC ĐẢM BẢO TÍNH CÔNG BẰNG CHO CÁC HỆ THỐNG HỌC MÁY TRONG LĨNH VỰC GIÁO DỤC

Chương này trình bày tổng quan về vấn đề đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục. Cụ thể, chương sẽ trình bày chi tiết về vai trò của trí tuệ nhân tạo và học máy trong giáo dục, các vấn đề nổi bật trong nghiên cứu về công bằng, các phương pháp đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục cũng được đề cập. Ngoài ra, các tiêu chí đánh giá công bằng, hiệu suất của mô hình, các vấn đề đánh đổi giữa công bằng và hiệu suất cũng được trình bày chi tiết. Để làm rõ các vấn đề trên, đầu tiên, các khái niệm trí tuệ nhân tạo và học máy, cũng như việc phân biệt các thuật toán học máy được đề cập. Tiếp theo, chương trình bày tổng quan về công bằng trong hệ thống học máy, bao gồm khái niệm về công bằng và các khái niệm liên quan, đồng thời các định nghĩa phổ biến về công bằng cũng được đề cập. Sau đó, chương giới thiệu các chỉ số đo lường công bằng thường được sử dụng trong thực nghiệm, bao gồm “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*”, và “*chênh lệch cơ hội công bằng*”. Chương cũng đề cập đến các chỉ số hiệu suất mô hình phổ biến như “*độ chuẩn xác*”, “*độ chính xác*”, “*độ hồi tưởng*”, và “*điểm số F1*”. Việc phân tích mối quan hệ đánh đổi giữa hiệu suất và công bằng được đề cập trong phần cuối của chương.

1.1. Trí tuệ nhân tạo và học máy trong giáo dục

Trí tuệ nhân tạo và học máy không chỉ còn là công cụ hỗ trợ, mà đang dần trở thành nền tảng công nghệ cốt lõi trong quá trình thiết kế, triển khai và đánh giá hệ thống trong kỷ nguyên số [37]. Những công nghệ này đang làm thay đổi

cách thức tổ chức và vận hành các ứng dụng trong giáo dục [3, 4, 5]. Chúng mở ra cơ hội đổi mới trong việc cá nhân hóa quá trình học tập, dự đoán sớm nguy cơ bỏ học, đánh giá năng lực người học và hỗ trợ ra quyết định cho nhà quản lý [6, 7, 8, 38]. Để làm rõ những nội dung trên, mục này sẽ trình bày các khái niệm nền tảng bao gồm trí tuệ nhân tạo, học máy, và hệ thống học máy. Đồng thời, phần này cũng phân biệt giữa các hình thức học phổ biến trong học máy, giúp hiểu rõ cơ chế hoạt động của các mô hình học máy. Bên cạnh đó, vai trò và ứng dụng cụ thể của AI/ML trong giáo dục sẽ được phân tích thông qua một số ví dụ minh họa. Cuối cùng, các tiêu chí đánh giá hiệu suất mô hình sẽ được đề cập nhằm đảm bảo tính hiệu quả và độ tin cậy của các hệ thống trí tuệ nhân tạo/học máy trong lĩnh vực giáo dục.

1.1.1. Vai trò của trí tuệ nhân tạo và học máy trong giáo dục

Vai trò của trí tuệ nhân tạo và học máy trong giáo dục có thể khái quát ở bốn khía cạnh chính. *Thứ nhất, AI/ML là động lực của quá trình chuyển đổi số trong giáo dục.* Các công nghệ này tái định hình mô hình giáo dục truyền thống, xây dựng môi trường học tập thông minh, linh hoạt và cá nhân hóa. Nhờ khả năng xử lý dữ liệu học tập của từng cá nhân, AI/ML cho phép triển khai các phương pháp học tập thích ứng, trong đó nội dung và tiến độ được điều chỉnh theo nhu cầu và năng lực của người học [29]. *Thứ hai, AI/ML là công cụ phân tích và dự đoán mạnh mẽ.* Với năng lực xử lý dữ liệu quy mô lớn, các hệ thống ML có thể phát hiện xu hướng, dự đoán kết quả học tập và cảnh báo sớm nguy cơ bỏ học [10, 39]. Những thông tin này là cơ sở quan trọng để giảng viên và nhà quản lý thực hiện các can thiệp kịp thời và hiệu quả. *Thứ ba, AI/ML hoạt động như một người đồng hành thông minh trong dạy và học.* Đối với giảng viên, AI/ML cung cấp các công cụ chấm điểm và đánh giá tự động dựa trên xử lý ngôn ngữ tự nhiên [40, 41, 42], giảm tải khối lượng công việc hành chính và tăng thời gian cho hoạt động giảng dạy. Đối với sinh viên, các hệ thống học tập thông minh phân tích hành vi học tập để nhận diện phong cách cá nhân [43], đồng thời gợi ý môn học, nội dung hoặc lộ trình học tập phù hợp với nhu cầu riêng [44]. *Thứ tư, AI/ML thúc đẩy đổi mới phương pháp đánh giá và ra quyết định.* Các công cụ phân tích học tập cho phép đánh giá năng lực đa chiều thay vì chỉ dựa vào điểm số. Các mô hình dự đoán và hệ thống gợi ý còn hỗ trợ nhà

quản lý thiết kế chương trình đào tạo, phân bổ nguồn lực và hoạch định chính sách một cách chính xác hơn [9, 45]. Như vậy, AI/ML không chỉ góp phần đổi mới phương pháp dạy và học mà còn nâng cao năng lực quản lý giáo dục, khẳng định vai trò ngày càng quan trọng trong giáo dục hiện đại.

1.1.2. Đánh giá hiệu suất của các mô hình học máy

Việc đánh giá hiệu suất của các mô hình học máy là một bước thiết yếu trong toàn bộ quy trình xây dựng hệ thống học máy. Mục tiêu của bước này là xác định các chỉ số hiệu suất và khả năng tổng quát hóa của mô hình đối với dữ liệu mới. Trong bối cảnh giáo dục, hiệu suất của mô hình không chỉ ảnh hưởng đến tính đúng đắn của dự đoán mà còn tác động lớn đến người học và cơ sở giáo dục thông qua những quyết định quan trọng như đánh giá năng lực, xác định nguy cơ bỏ học hoặc phân nhóm người học, v.v.. [46].

Hiệu suất của mô hình học máy thường được đánh giá dựa trên mục tiêu cụ thể của bài toán, chẳng hạn như phân loại hoặc hồi quy. Do đó, các chỉ số hiệu suất có thể khác nhau tùy vào loại bài toán [47]. Trong lĩnh vực giáo dục, nhiều nghiên cứu tập trung vào các bài toán phân loại nhị phân (ví dụ: sinh viên tốt nghiệp hay không, đạt chuẩn đầu ra hay không), với các chỉ số phổ biến được sử dụng để đánh giá hiệu suất mô hình bao gồm: “*độ chuẩn xác*”, “*độ chính xác*”, “*độ hồi tưởng*”, và “*điểm số F1*” [48]. Các chỉ số này sẽ được trình bày chi tiết trong phần dưới đây.

Độ chuẩn xác (Accuracy): Là tỷ lệ tổng số dự đoán đúng (cả dương và âm) trên toàn bộ dữ liệu, dùng để đo lường mức độ tổng quát của mô hình. Giá trị độ chuẩn xác được xác định bằng Công thức 1.1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1.1)$$

Trong đó:

- *TP* (True Positive): Số trường hợp thực sự dương và được mô hình dự đoán đúng là dương,
- *TN* (True Negative): Số trường hợp thực sự âm và được mô hình dự đoán đúng là âm,

- *FP* (False Positive): Số trường hợp thực sự âm nhưng được mô hình dự đoán sai là dương, và
- *FN* (False Negative): Số trường hợp thực sự dương nhưng được mô hình dự đoán sai là âm.

Độ chính xác (*Precision*): Là tỷ lệ các trường hợp dương được mô hình dự đoán đúng trên tổng số trường hợp được dự đoán là dương. Độ chính xác phản ánh mức độ tin cậy của các dự đoán dương. Giá trị độ chính xác được xác định bằng Công thức 1.2.

$$Precision = \frac{TP}{TP + FP} \quad (1.2)$$

Độ hồi tưởng (*Recall*): Là tỷ lệ các trường hợp dương được mô hình dự đoán đúng trên tổng số trường hợp thực sự dương. Độ hồi tưởng phản ánh khả năng phát hiện đầy đủ các trường hợp dương. Giá trị độ hồi tưởng được xác định bằng Công thức 1.3.

$$Recall = \frac{TP}{TP + FN} \quad (1.3)$$

Điểm số *F1* (*F1-score*): Là trung bình điều hòa giữa độ chính xác và độ hồi tưởng. Điểm số *F1* giúp cân bằng giữa độ chính xác và độ hồi tưởng, đặc biệt hữu ích trong trường hợp dữ liệu mất cân bằng. Giá trị điểm số *F1* được xác định bằng Công thức 1.4.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (1.4)$$

Việc sử dụng kết hợp các chỉ số trên giúp đánh giá hiệu suất của mô hình một cách toàn diện, tránh sự sai lệch khi chỉ dựa vào một chỉ số đơn lẻ. Điều này đặc biệt quan trọng trong các tình huống dữ liệu mất cân bằng, vốn rất phổ biến trong các hệ thống giáo dục thực tế [46, 49].

1.2. Công bằng trong các hệ thống học máy

Các hệ thống học máy ngày càng đóng vai trò then chốt trong nhiều lĩnh vực, trong đó có lĩnh vực giáo dục [9, 37, 46]. Khi các hệ thống này tham gia trực tiếp vào các quyết định có ảnh hưởng lâu dài đến người học, chẳng hạn như đánh giá năng lực, dự đoán nguy cơ bỏ học hay hỗ trợ định hướng học tập, yêu cầu đảm bảo tính công bằng trong quá trình thiết kế và vận hành hệ thống trở nên đặc biệt quan trọng [15, 19].

Trong những năm gần đây, nhiều nghiên cứu đã tập trung hệ thống hóa các khái niệm liên quan đến công bằng trong các hệ thống học máy, làm rõ các định nghĩa nền tảng, phân loại các hướng tiếp cận đảm bảo công bằng, cũng như phân tích những thách thức và xu hướng nghiên cứu nổi bật trong lĩnh vực này. Các tổng hợp này cung cấp một khung tham chiếu quan trọng cho việc nhận diện khoảng trống nghiên cứu và định hướng phát triển các phương pháp đảm bảo công bằng cho các hệ thống học máy [14, 50].

Trên cơ sở các kết quả tổng hợp đó, trong phần tiếp theo luận án sẽ trình bày chi tiết các khái niệm, hướng tiếp cận và vấn đề cốt lõi liên quan đến công bằng trong các hệ thống học máy, làm nền tảng cho các phương pháp được đề xuất trong luận án.

1.2.1. Khái niệm về công bằng trong các hệ thống học máy

Công bằng là một khái niệm mang tính xã hội và đạo đức, hướng tới việc đảm bảo các cá nhân hoặc nhóm được đối xử bình đẳng, không bị phân biệt đối xử dựa trên các đặc điểm nhạy cảm như giới tính, chủng tộc, dân tộc, tuổi tác hoặc tình trạng khuyết tật. Trong nhiều lĩnh vực như tuyển dụng, giáo dục và tài chính, công bằng không chỉ được xem là một yêu cầu về đạo đức mà còn được quy định trong các chính sách và văn bản pháp luật nhằm bảo vệ các nhóm yếu thế [51].

Một trong những quy định có ảnh hưởng lớn đến nghiên cứu về công bằng là *quy định 80%* (80% Rule hay Four-Fifths Rule) do Ủy ban Cơ hội Việc làm

Bình đẳng Hoa Kỳ (U.S. Equal Employment Opportunity Commission - EEOC) ban hành [52]. Quy định này cho rằng một quy trình tuyển chọn có thể bị coi là có tác động bất lợi nếu tỷ lệ thuộc lớp dương của một nhóm có tỷ lệ thấp nhất (nhóm yếu thế) nhỏ hơn 80% tỷ lệ thuộc lớp dương của nhóm có tỷ lệ cao nhất (nhóm ưu thế). Quy định này đã trở thành nền tảng cho nhiều nghiên cứu về đánh giá sự công bằng được sử dụng rộng rãi trong học máy. Cần lưu ý rằng, quy định này không yêu cầu thuộc tính nhạy cảm phải là biến nhị phân. Đối với một thuộc tính nhạy cảm có nhiều hơn hai giá trị (ví dụ: chủng tộc, dân tộc hoặc khu vực), việc đánh giá công bằng vẫn được thực hiện thông qua so sánh giữa nhóm có tỷ lệ được lựa chọn cao nhất (nhóm ưu thế) và nhóm có tỷ lệ được lựa chọn thấp nhất (nhóm yếu thế). Vì vậy, trong luận án này, các thuộc tính nhạy cảm được nhị phân hóa nhằm xác định rõ hai nhóm cần so sánh, phù hợp với các thước đo công bằng được sử dụng và không làm hạn chế khả năng mở rộng của phương pháp đối với các thuộc tính nhạy cảm đa giá trị.

Khái niệm công bằng trong các hệ thống học máy đã được quan tâm trong nhiều thập kỷ qua. Các định nghĩa công bằng đã xuất hiện từ lâu trong nghiên cứu về giáo dục và tuyển dụng, một số khái niệm trước đây tương đồng với các định nghĩa hiện đang được sử dụng trong học máy [50].

Trong bối cảnh học máy, công bằng thường được hiểu là việc đảm bảo các quyết định không bị thiên vị đối với các thuộc tính nhạy cảm như *tuổi*, *chủng tộc* hoặc *giới tính* [1, 15, 50]. Hai quan điểm phổ biến là *công bằng nhóm*, tập trung so sánh kết quả giữa các nhóm xã hội khác nhau, và *công bằng cá nhân*, nhấn mạnh rằng các cá nhân tương đồng cần nhận được kết quả tương tự [17]. Tuy nhiên, nhiều nghiên cứu chỉ ra rằng hai khái niệm này thường khó đạt được đồng thời trong thực tế [53, 54].

Các khái niệm công bằng có thể được hình thức hóa thông qua nhiều định nghĩa toán học khác nhau [1, 17]. Tuy nhiên, vẫn chưa tồn tại sự đồng thuận về một định nghĩa duy nhất cho công bằng trong các hệ thống học máy [50]. Những định nghĩa phổ biến như *công bằng nhân khẩu học*, *cơ hội công bằng*, *công bằng theo xác suất*, *kiểm tra công bằng*, *công bằng qua nhận thức*, *công bằng qua không nhận thức*, *đối xử công bằng*, *công bằng phản sự kiện* và *công bằng trong miền quan hệ* phản ánh các khía cạnh công bằng khác nhau. Dựa trên mục tiêu đánh giá, các định nghĩa này thường được phân loại thành ba nhóm: *công bằng nhóm*,

công bằng cá nhân, và các khái niệm công bằng khác [1, 17]. Mặc dù nhiều định nghĩa công bằng được đề xuất, việc lựa chọn định nghĩa phù hợp cần phụ thuộc vào bối cảnh ứng dụng, đặc điểm dữ liệu và các chuẩn mực đạo đức-xã hội của từng lĩnh vực [50]

Chú ý: Trong Bảng 1.1 Y là nhãn thực tế, \hat{Y} là nhãn dự đoán của mô hình, A là thuộc tính nhạy cảm ($A = 0$ và $A = 1$ lần lượt biểu diễn hai nhóm ưu thế và yếu thế được so sánh), L là tập các thuộc tính hợp lệ và S là điểm dự đoán của mô hình. Các ký hiệu TPR , FPR , FN và FP lần lượt biểu thị tỷ lệ dự đoán đúng dương, tỷ lệ dự đoán dương giả, số lượng âm tính giả và số lượng dương tính giả.

1.2.2. Khái niệm thiên vị trong học máy

Trong nghiên cứu công bằng trong học máy, khái niệm thiên vị thường đi kèm với công bằng và có thể được xem như trạng thái đối lập của công bằng, tức là công bằng có thể được hiểu là không thiên vị [6]. Do đó, việc phát hiện và nhận diện thiên vị đóng vai trò quan trọng trong việc đánh giá mức độ công bằng của một hệ thống. Việc điều chỉnh thiên vị cụ thể sẽ giúp tiến gần hơn tới việc đạt được tính công bằng trong các hệ thống học máy [6].

Khái niệm thiên vị ngày càng được quan tâm trong các nghiên cứu về công bằng trong học máy [6, 15]. Tuy nhiên, tương tự như công bằng, hiện vẫn chưa tồn tại một định nghĩa thống nhất về thiên vị do sự đa dạng về góc nhìn lý thuyết và bối cảnh ứng dụng. Nhiều nghiên cứu đã phân loại các dạng thiên vị xuất hiện trong các giai đoạn khác nhau của hệ thống học máy như thu thập dữ liệu, huấn luyện mô hình và triển khai hệ thống. Bảng A.3 (Phụ lục A) tổng hợp các loại thiên vị phổ biến và đóng vai trò như tài liệu tham chiếu cho việc lựa chọn các độ đo và phương pháp đánh giá công bằng trong các phần tiếp theo của luận án.

Mặc dù thiên vị và công bằng thường được sử dụng song song để phân tích mức độ bất công trong các hệ thống học máy [6, 15, 20], việc lựa chọn cách tiếp cận “loại bỏ thiên vị” hay “đảm bảo công bằng” phụ thuộc vào bối cảnh ứng dụng cụ thể [20]. Trong một số lĩnh vực như giáo dục hoặc y tế, việc ưu tiên *công*

Bảng 1.1: Các định nghĩa về công bằng trong hệ thống học máy

Loại công bằng	Định nghĩa	Giải thích	Tham khảo
Công bằng theo nhóm			
Cân bằng nhân khẩu học	$P(\hat{Y} A = 0) = P(\hat{Y} A = 1)$	Xác suất dự đoán thuộc lớp dương phải bằng nhau giữa các nhóm bất kể nhân thực tế	[54, 55, 56, 57]
Cơ hội công bằng	$P(\hat{Y} = 1 A = 0, Y = 1) = P(\hat{Y} = 1 A = 1, Y = 1)$	Đảm bảo tỷ lệ đúng dương (TPR) bằng nhau giữa các nhóm	[54, 55, 58]
Xác suất công bằng	$P(\hat{Y} = 1 A = 0, Y = y) = P(\hat{Y} = 1 A = 1, Y = y)$ với $y \in \{0, 1\}$	Đảm bảo TPR và FPR tương đương giữa nhóm ưu thế và nhóm yếu thế. (ví dụ: nam và nữ)	[55, 58]
Cân bằng thống kê có điều kiện	$P(\hat{Y} L = 1, A = 0) = P(\hat{Y} L = 1, A = 1)$	Với điều kiện như nhau (ví dụ: điểm đầu vào), xác suất dự đoán phải như nhau	[54, 55, 59]
Đối xử công bằng	FN/FP bằng nhau giữa các nhóm	Tỷ lệ dự đoán sai cần được cân bằng giữa các nhóm	[55, 60]
Kiểm tra công bằng	$P(Y = 1 S = s, A = 0) = P(Y = 1 S = s, A = 1), \forall s$	Với cùng điểm dự đoán, xác suất kết quả thực tế phải như nhau giữa các nhóm	[54, 55, 61]
Công bằng cá nhân			
Công bằng qua nhận thức	Dự đoán tương tự cho cá nhân tương đồng theo một hàm đo khoảng cách phù hợp	Sinh viên có hồ sơ học tương tự nên được dự đoán như nhau, bất kể <i>giới tính</i> /sắc tộc	[54, 55, 56]
Công bằng phản sự kiện	Dự đoán không thay đổi nếu thay đổi giá trị thuộc tính nhạy cảm trong thể giới phản sự kiện	Ví dụ: kết quả dự đoán GPA không thay đổi khi đổi <i>giới tính</i> từ nữ sang nam	[55, 56]
Các khái niệm công bằng khác			
Công bằng qua không nhận thức	Không sử dụng trực tiếp thuộc tính nhạy cảm trong mô hình	Ví dụ: không sử dụng <i>giới tính</i> trong mô hình dự đoán	[54, 55, 57]
Công bằng trong miền quan hệ	Xét đến cả thuộc tính cá nhân và quan hệ xã hội giữa các cá nhân	Ví dụ: ảnh hưởng từ nhóm học tập, câu lạc bộ trong trường	[55, 62]

bằng nhóm có thể cần thiết để giảm bất bình đẳng mang tính lịch sử [63, 64], trong khi ở các lĩnh vực nhấn mạnh tính cá nhân hóa như tuyển dụng, việc giảm thiên vị ở cấp độ cá nhân trở nên quan trọng hơn [14, 18, 65]. Vì vậy, thiên vị và công bằng nên được xem là các khái niệm bổ trợ, giúp xây dựng các hệ thống học máy đáng tin cậy và có trách nhiệm.

1.2.3. Thuộc tính nhạy cảm trong nghiên cứu về công bằng

Thuộc tính nhạy cảm là các thuộc tính phản ánh đặc điểm nhân khẩu học hoặc xã hội mà theo luật pháp và chuẩn mực đạo đức cần được bảo vệ khỏi sự phân biệt đối xử [15, 66]. Trong nghiên cứu về công bằng của các hệ thống học máy, biến nhạy cảm là những thuộc tính của dữ liệu có thể dẫn đến sự phân biệt trong quá trình ra quyết định của mô hình học máy. Do đó, chúng đóng vai trò quan trọng trong việc phát hiện, đo lường và điều chỉnh thiên vị. Việc xác định thuộc tính nhạy cảm thường phụ thuộc vào bối cảnh ứng dụng và quy định pháp lý của từng quốc gia.

Các thuộc tính nhạy cảm thường được phân loại dựa trên ba tiêu chí: *định danh*, *ngữ cảnh xã hội*, và *quy định pháp lý* [52, 58, 67]. Nhóm *định danh* bao gồm các thuộc tính như *giới tính*, *chủng tộc*, quốc tịch, tôn giáo hoặc dân tộc [67]. Nhóm *ngữ cảnh xã hội* đề cập đến các thuộc tính có thể trở nên nhạy cảm trong các hệ thống ra quyết định quan trọng, ví dụ như trình độ học vấn hoặc tình trạng hôn nhân [58]. Trong khi đó, theo các quy định pháp lý như Đạo luật dân quyền Hoa Kỳ hoặc GDPR của Liên minh Châu Âu, các thuộc tính như *tuổi*, *giới tính*, *tình trạng khuyết tật* và *chủng tộc* được xem là các thông tin cá nhân cần được bảo vệ [52, 67]. Trong lĩnh vực giáo dục, các thuộc tính nhạy cảm thường được xem xét theo ba nhóm chính: *nhân khẩu học*, *địa lý – kinh tế*, và *xã hội – giáo dục*. Các nhóm này bao gồm các yếu tố như *giới tính*, *chủng tộc*, *tuổi*, *khu vực*, *mức thu nhập*, *trình độ học vấn của cha mẹ*, *sức khỏe*, hoặc những yếu tố có thể ảnh hưởng đến cơ hội tiếp cận và kết quả học tập [68]. Dựa trên tổng hợp từ các nghiên cứu liên quan, Bảng 1.2 trình bày các thuộc tính nhạy cảm thường được sử dụng trong nghiên cứu công bằng học máy trong giáo dục.

Bảng 1.2: Phân loại các thuộc tính nhạy cảm thường dùng trong giáo dục

Nhóm 1: thuộc tính nhân khẩu học	
Giới tính (Gender)	Nam, nữ hoặc phi nhị nguyên giới
Chủng tộc (Race)	Màu da, nét mặt, chất tóc dùng để phân nhóm (ví dụ: da trắng, da đen, gốc Á, v.v.)
Tình trạng khuyết tật (Disability)	Bản sắc dân tộc hoặc tình trạng khuyết tật về thể chất/tâm thần
Tuổi (Age)	Độ tuổi theo năm dương lịch
Ngôn ngữ (Language)	Ngôn ngữ mẹ đẻ hoặc ngôn ngữ sử dụng thành thạo nhất
Tôn giáo (Religion)	Niềm tin tôn giáo có thể ảnh hưởng đến cách tiếp cận giáo dục
Nhóm 2: thuộc tính địa lý và kinh tế	
Quốc gia / Xuất xứ (Country / Origin)	Quốc gia sinh sống hoặc nơi sinh ra, nguồn gốc tổ tiên
Vùng miền / Địa phương (Region, Zone) / Locality	Địa điểm cư trú (nông thôn, thành thị, vùng sâu vùng xa, v.v.)
Thu nhập (Income)	Mức thu nhập của cá nhân hoặc gia đình
Nhóm 3: thuộc tính xã hội và giáo dục	
Năm nhất (First-gen)	Sinh viên năm nhất, lần đầu vào đại học
Trình độ học vấn của cha mẹ	Mức học vấn cao nhất của cha mẹ hoặc người giám hộ
Môi trường đọc tại nhà	Sự sẵn có của sách, tài liệu đọc và hoạt động đọc tại nhà
Tình trạng sức khỏe (Health)	Tình trạng thể chất hoặc tinh thần ảnh hưởng đến học tập

Việc xác định các thuộc tính nhạy cảm là nền tảng để bảo đảm công bằng trong các hệ thống học máy ứng dụng cho giáo dục. Các bộ dữ liệu giáo dục thường bao gồm nhiều thuộc tính nhạy cảm tiềm ẩn. Việc chỉ xem xét từng yếu tố riêng lẻ sẽ không nhận diện hết các dạng bất bình đẳng trong mô hình [33]. Từ đó, nhu cầu phát triển các phương pháp bảo đảm tính công bằng cho đồng thời nhiều thuộc tính nhạy cảm trở nên đặc biệt quan trọng và là định hướng trọng tâm cho các nghiên cứu hiện nay. Hướng tiếp cận này cho phép phát hiện và xử lý các dạng thiên vị giao thoa mà các hệ thống học máy có thể tạo [69].

1.2.4. Độ đo công bằng trong các hệ thống học máy

Độ đo công bằng là các chỉ số định lượng được sử dụng để đánh giá mức độ công bằng hoặc thiên vị của một hệ thống học máy đối với các nhóm khác nhau. Các độ đo này giúp xác định liệu mô hình có tạo ra sự khác biệt về kết quả dự đoán giữa các nhóm nhạy cảm hay không, từ đó hỗ trợ việc giám sát và điều chỉnh mô hình nhằm đảm bảo các nguyên tắc công bằng được duy trì, đặc biệt trong các ứng dụng giáo dục nơi quyết định của mô hình có thể ảnh hưởng lâu dài đến người học. Nhiều độ đo công bằng đã được đề xuất, các độ đo phổ biến bao gồm “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*” và “*chênh lệch cơ hội công bằng*” [15, 66].

Trong phạm vi luận án, bài toán được phát biểu đối với phân lớp nhị phân do các thước đo công bằng được sử dụng (DI, SPD, EOD và AOD) đều được định nghĩa chủ yếu cho bài toán này. Đồng thời, các bộ dữ liệu thực nghiệm đều được chuyển đổi về hai lớp (lớp dương và lớp âm) để phù hợp với mục tiêu đánh giá và cải thiện tính công bằng. Việc mở rộng phương pháp cho bài toán phân lớp đa lớp sẽ được xem xét trong các nghiên cứu tiếp theo.

Gọi S_{ML} là một mô hình phân loại nhị phân ánh xạ một vectơ đặc trưng $\mathbf{x} = [x_1, x_2, \dots, x_d] \in \mathbb{R}^d$ thành đầu ra dự đoán nhị phân $\hat{y} \in \{0, 1\}$. Khi đó, mô hình có thể được biểu diễn theo Công thức 1.5

$$S_{ML} : \mathbb{R}^d \rightarrow \{0, 1\}. \quad (1.5)$$

trong đó $y \in \{0, 1\}$ là nhãn thực tế và $A \in \{0, 1\}$ là thuộc tính nhạy cảm dùng để phân chia đối tượng thành các nhóm (ví dụ *giới tính*) [70].

Trong đánh giá công bằng, A được sử dụng để phân chia dữ liệu thành hai nhóm: *nhóm ưu thế* và *nhóm yếu thế*. Nhóm ưu thế là nhóm có xu hướng nhận được kết quả dự đoán thuận lợi hơn, trong khi nhóm yếu thế có nguy cơ chịu bất lợi. Việc xác định hai nhóm này phụ thuộc vào ngữ cảnh cụ thể của bài toán và là cơ sở để tính toán các độ đo công bằng. Chi tiết các độ đo công bằng được trình bày dưới đây. ***Tác động khác biệt (Disparate Impact–DI)***: Là tỷ lệ giữa xác suất dự đoán kết quả thuộc lớp dương cho nhóm ưu thế và nhóm yếu

thể. Phản ánh mức độ bình đẳng về cơ hội giữa các nhóm [15]. Giá trị này được xác định bằng Công thức 1.6.

$$DI = \frac{P(\hat{Y} = 1 | A = 0)}{P(\hat{Y} = 1 | A = 1)} \quad (1.6)$$

Trong đó $P(\hat{Y} = 1 | A = 0)$ và $P(\hat{Y} = 1 | A = 1)$ lần lượt là xác suất dự đoán kết quả thuộc lớp dương cho nhóm ưu thế và xác suất dự đoán kết quả thuộc lớp dương cho nhóm yếu thế. Giá trị DI nằm trong khoảng $[0.8, 1.25]$ được xem là công bằng theo quy tắc 80% [52].

Hiệu số chênh lệch thống kê (Statistical Parity Difference–SPD):

Là giá trị chênh lệch giữa xác suất được dự đoán là thuộc lớp dương giữa hai nhóm nhạy cảm [15]. Giá trị này được xác định bằng Công thức 1.7.

$$SPD = P(\hat{Y} = 1 | A = 0) - P(\hat{Y} = 1 | A = 1) \quad (1.7)$$

Tương tự như trong Công thức 1.6, $P(\hat{Y} = 1 | A = 0)$ và $P(\hat{Y} = 1 | A = 1)$ lần lượt là xác suất dự đoán kết quả thuộc lớp dương cho nhóm ưu thế và xác suất dự đoán kết quả tích cực cho nhóm yếu thế. Giá trị SPD càng nhỏ (càng gần 0) thể hiện sự công bằng theo định nghĩa *công bằng nhân khẩu học* (xem Mục 1.2.1).

Chênh lệch trung bình xác suất (Average Odds Difference–AOD):

Là giá trị trung bình của sự khác biệt giữa tỷ lệ dương tính thực (TPR) và tỷ lệ dương tính giả (FPR) giữa hai nhóm, [15]. Giá trị này được xác định bằng Công thức 1.8.

$$AOD = \frac{1}{2} [(TPR_{A=0} - TPR_{A=1}) + (FPR_{A=0} - FPR_{A=1})] \quad (1.8)$$

Trong đó, TPR và FPR lần lượt được xác định bởi các Công thức 1.9 và 1.10.

$$TPR_A = \frac{TP_A}{TP_A + FN_A} \quad (1.9)$$

$$FPR_A = \frac{FP_A}{FP_A + TN_A} \quad (1.10)$$

Các giá trị TP , TN , FP , và FN lần lượt là các giá trị *dương tính thực*, *âm tính thực*, *dương tính giả*, và *âm tính giả* (xem Mục 1.1.2). Giá trị AOD càng gần 0 cho thấy mô hình càng công bằng theo định nghĩa *công bằng theo xác suất*[66].

Chênh lệch cơ hội công bằng (Equal Opportunity Difference–EOD): Là giá trị chênh lệch về *tỷ lệ dương tính thực* giữa hai nhóm nhạy cảm, chỉ tính trên các trường hợp có nhãn thực là dương tính [15]. Giá trị này được xác định bằng Công thức 1.11. EOD càng gần 0 cho thấy mô hình càng công bằng theo tiêu chí *cơ hội công bằng*[66].

$$EOD = TPR_{A=0} - TPR_{A=1} \quad (1.11)$$

ABROCA (Area Between ROC Curves): là một độ đo công bằng dùng để đánh giá sự khác biệt về hiệu suất phân loại giữa hai nhóm nhạy cảm thông qua so sánh các đường cong ROC. Cụ thể, ABROCA đo diện tích tuyệt đối giữa hai đường ROC ứng với hai nhóm khác nhau trong không gian FPR–TPR. Giá trị ABROCA càng nhỏ cho thấy mô hình có hiệu suất gần như nhau trên cả hai nhóm [39, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80]. Giá trị này được xác định bằng Công thức 1.12:

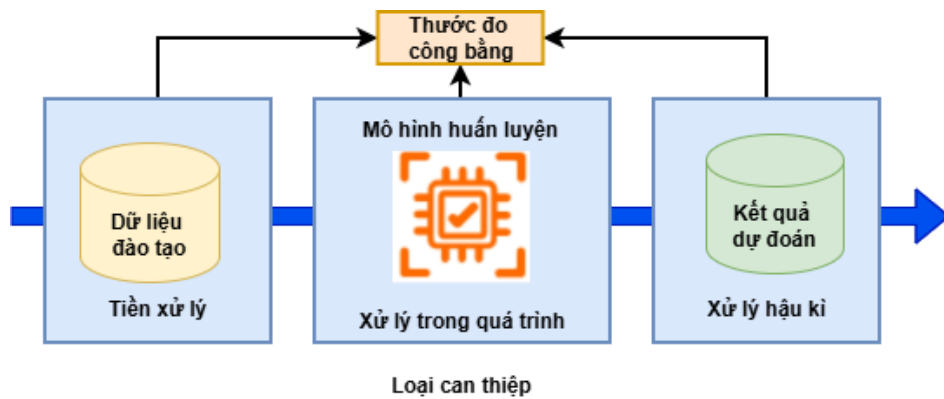
$$ABROCA = \int_0^1 |TPR_{A=0}(x) - TPR_{A=1}(x)| dx \quad (1.12)$$

Trong đó:

- $x \in [0, 1]$: là giá trị trục hoành tương ứng với False Positive Rate (FPR) và
- $TPR_{A=0}(x), TPR_{A=1}(x)$: là các giá trị True Positive Rate tương ứng tại mỗi FPR của hai nhóm nhạy cảm.

1.2.5. Các hướng tiếp cận đảm bảo tính công bằng cho các hệ thống học máy

Khung tham chiếu nhằm phân loại các phương pháp can thiệp để đảm bảo tính công bằng trong các hệ thống học máy được Caton và cộng sự đã đề xuất năm 2020 bao gồm ba giai đoạn được minh họa trong Hình 1.1 [1]. Ba giai đoạn này tương ứng với các hình thức can thiệp vào quá trình học máy, bao gồm: *trước khi huấn luyện mô hình (tiền xử lý)*, *trong quá trình huấn luyện mô hình (xử lý trong quá trình)*, và *sau khi mô hình được huấn luyện (hậu xử lý)* [81]. Mặc dù không phải mọi phương pháp đều hoàn toàn phù hợp nhưng nó cung cấp một cấu trúc nhất quán giúp hệ thống hóa các cách tiếp cận nhằm tăng cường tính công bằng trong học máy.



Hình 1.1: Minh họa về các giai đoạn can thiệp công bằng trong các hệ thống học máy [1].

Giai đoạn Tiền xử lý: Ở giai đoạn này, thiên vị thường xuất hiện trong dữ liệu và phân phối của các thuộc tính nhạy cảm [66]. Vì vậy, các phương pháp tiền xử lý tập trung điều chỉnh phân phối mẫu hoặc biến đổi dữ liệu nhằm loại bỏ sự phân biệt đối xử trong dữ liệu huấn luyện [1, 82]. Mô hình sau đó được huấn luyện trên tập dữ liệu đã được điều chỉnh. Một số hướng tiếp cận phổ biến gồm *giảm thiểu thiên vị*, *phát hiện cộng đồng công bằng*, *suy luận nhân quả công bằng*, *học biểu diễn công bằng*, và *xử lý ngôn ngữ tự nhiên công bằng* [15]. Các kỹ thuật này nhằm giảm ảnh hưởng của thiên vị trong dữ liệu [1, 24, 83, 84, 85].

Giai đoạn Xử lý trong quá trình: Trong giai đoạn này, các phương pháp can thiệp trực tiếp vào quá trình huấn luyện mô hình nhằm kiểm soát ảnh hưởng

của các thuộc tính nhạy cảm. Một cách tiếp cận phổ biến là tích hợp các độ đo công bằng vào hàm mất mát hoặc bổ sung các ràng buộc công bằng trong quá trình tối ưu [66]. Các kỹ thuật điển hình gồm *phân loại công bằng* và *hồi quy công bằng* [15]. Những phương pháp này hướng tới việc tối ưu đồng thời hiệu suất dự đoán và công bằng, đồng thời kiểm soát mức đánh đổi giữa hai yếu tố này.

Giai đoạn Hậu xử lý: Ở giai đoạn hậu xử lý, các phương pháp điều chỉnh trực tiếp đầu ra của mô hình để giảm sự khác biệt giữa các nhóm nhạy cảm mà không cần thay đổi dữ liệu huấn luyện hoặc cấu trúc mô hình [66]. Cách tiếp cận này đặc biệt hữu ích với các mô hình hộp đen. Một số kỹ thuật tiêu biểu bao gồm các phương pháp *dự đoán có cấu trúc*, chẳng hạn thuật toán RBA nhằm giảm khuếch đại thiên vị trong dự đoán [86], và các phương pháp áp dụng cho *mô hình ngôn ngữ* để phát hiện và đánh giá thiên vị trong văn bản sinh ra bởi mô hình [15, 87].

Bên cạnh ba nhóm kỹ thuật chính trên, nhiều hướng tiếp cận khác cũng được phát triển nhằm tăng cường tính công bằng trong các hệ thống học máy. Một số nghiên cứu đề xuất xây dựng các tiêu chí công bằng dưới dạng hàm lỗi để đơn giản hóa và ổn định quá trình tối ưu hóa [16]. Ngoài ra, các kỹ thuật biến đổi dữ liệu như tái cân bằng dữ liệu [88] hoặc học biểu diễn trung tính [89] cũng được sử dụng nhằm giảm ảnh hưởng của các thuộc tính nhạy cảm. Những hướng tiếp cận này mở rộng khung nghiên cứu về công bằng nhưng cũng cho thấy những thách thức trong việc lựa chọn và cân bằng các mục tiêu công bằng khác nhau.

1.2.6. Những thách thức trong việc đảm bảo tính công bằng trong các hệ thống học máy

Các nghiên cứu tổng quan gần đây cho thấy những thách thức trong đảm bảo công bằng cho hệ thống học máy có thể được phân thành bốn nhóm chính: (i) *xây dựng và hình thức hóa khái niệm công bằng*, (ii) *đánh giá công bằng theo các thuộc tính nhạy cảm*, (iii) *mối quan hệ đánh đổi giữa công bằng và hiệu suất*, và (iv) *các yếu tố liên quan đến con người và bối cảnh sử dụng hệ thống* [15, 50]. Chi tiết của từng nhóm thách thức này được trình bày và phân tích cụ thể dưới đây.

Nhóm 1. Xây dựng và hình thức hóa khái niệm công bằng: Nhóm thách thức này liên quan đến việc định nghĩa và lượng hóa khái niệm công bằng trong các hệ thống học máy. Mặc dù nhiều nghiên cứu đã đề xuất các định nghĩa khác nhau, hiện vẫn chưa có sự đồng thuận về một định nghĩa chung. Thậm chí, một số định nghĩa công bằng có thể mâu thuẫn với nhau, chẳng hạn như giữa công bằng cá nhân và công bằng nhóm. [1, 15, 50, 66, 90]. Bên cạnh đó, việc lượng hóa các khái niệm công bằng cũng trở nên phức tạp do chúng chịu ảnh hưởng bởi các yếu tố xã hội, chính trị và bối cảnh ứng dụng cụ thể.

Nhóm 2. Đánh giá công bằng theo các thuộc tính nhạy cảm: Nhóm thách thức này liên quan đến việc đánh giá công bằng dựa trên các thuộc tính nhạy cảm trong bộ dữ liệu. Phần lớn các nghiên cứu hiện nay thường xem xét công bằng đối với từng thuộc tính nhạy cảm riêng lẻ. Tuy nhiên, cách tiếp cận này có thể dẫn đến việc cải thiện công bằng cho thuộc tính này nhưng lại vô tình bất công với thuộc tính khác. Điều này gây khó khăn cho việc đánh giá và đảm bảo tính công bằng một cách toàn diện trong các mô hình AI, đặc biệt với các ứng dụng trong lĩnh vực giáo dục [37].

Nhóm 3. Mối quan hệ giữa công bằng và hiệu suất: Nhóm này tập trung những thách thức về đánh đổi giữa công bằng và hiệu suất của mô hình. Một số nghiên cứu đã chỉ ra rằng các phương pháp đảm bảo tính công bằng hiện tại có xu hướng làm giảm độ chính xác của mô hình, từ đó đặt ra vấn đề đánh đổi giữa hiệu suất và tính công bằng trong quá trình thiết kế hệ thống [17].

Nhóm 4. Yếu tố con người: Nhóm này tập trung vào những thách thức liên quan đến tác động của con người đối với sự thành công trong nghiên cứu công bằng cho các hệ thống học máy. Những tiến bộ nhanh chóng trong công nghệ đã tạo ra khoảng cách về nhận thức giữa các bên liên quan, dẫn đến yêu cầu ngày càng cao trong việc nâng cao nhận thức của con người trong thiết kế và triển khai các hệ thống học máy [1].

Từ bốn nhóm thách thức trên có thể thấy rằng, đảm bảo tính công bằng trong các hệ thống học máy không chỉ là vấn đề kỹ thuật mà còn liên quan đến các yếu tố xã hội, đạo đức và tổ chức. Việc nhận diện và giải quyết những thách thức này là nền tảng để phát triển các hệ thống học máy minh bạch, đáng tin cậy và hướng tới lợi ích con người.

1.3. Sinh dữ liệu tổng hợp trong đảm bảo tính công bằng cho các hệ thống học máy

1.3.1. Các khái niệm

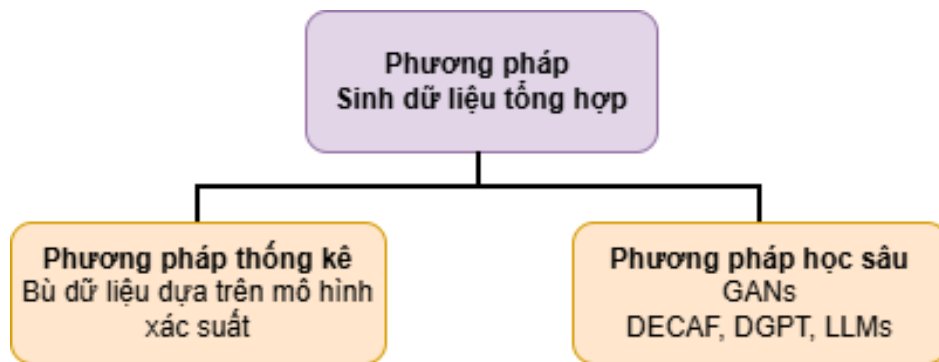
Dữ liệu tổng hợp là loại dữ liệu được tạo ra một cách nhân tạo thông qua các mô hình toán học hoặc thuật toán nhằm phục vụ cho các tác vụ cụ thể trong khoa học dữ liệu [91]. Dữ liệu tổng hợp có thể mô phỏng cấu trúc và các đặc tính thống kê của dữ liệu thực, đồng thời cung cấp tính linh hoạt cao và khả năng bảo vệ quyền riêng tư tốt hơn. Tương tự như dữ liệu thực, dữ liệu tổng hợp có thể có dạng có cấu trúc, bán cấu trúc hoặc phi cấu trúc [91]. Trong phạm vi này, nghiên cứu chỉ tập trung vào *dữ liệu bảng tổng hợp*, được tổ chức dưới dạng hàng và cột [92]. Thuật ngữ “dữ liệu tổng hợp” trong toàn bộ nghiên cứu sẽ được hiểu là dữ liệu bảng tổng hợp.

Sinh dữ liệu tổng hợp là quá trình tạo ra dữ liệu tổng hợp dựa trên phân bố và đặc tính thống kê của dữ liệu thực. Quá trình này thường sử dụng các mô hình xác suất, học sâu hoặc mạng đối kháng sinh dữ liệu như GAN nhằm đảm bảo dữ liệu sinh ra có đặc điểm tương tự dữ liệu gốc nhưng không làm lộ thông tin nhạy cảm. Trong lĩnh vực giáo dục, sinh dữ liệu tổng hợp ngày càng được ứng dụng để tăng cường dữ liệu đầu vào cho các hệ thống phân tích học tập, huấn luyện mô hình dự đoán, hoặc chia sẻ dữ liệu một cách bảo mật [36, 93].

Một nhánh quan trọng trong kỹ thuật này là *sinh dữ liệu có điều kiện*, trong đó quá trình sinh dữ liệu được kiểm soát bởi một hoặc nhiều thuộc tính cụ thể, gọi là điều kiện (ví dụ: giới tính, nhóm nguy cơ bỏ học, khu vực địa lý, v.v.). Phương pháp này cho phép tạo ra dữ liệu tổng hợp tương ứng với các nhóm mục tiêu, từ đó giúp tăng cường đại diện cho các nhóm thiểu số, khắc phục mất cân bằng dữ liệu, và đặc biệt hữu ích trong đánh giá tính công bằng của mô hình học máy [94]. Nghiên cứu sẽ đặc biệt chú trọng đến *phương pháp sinh dữ liệu tổng hợp có điều kiện*, sử dụng các mô hình như CTGAN và LLM.

1.3.2. Kỹ thuật sinh dữ liệu tổng hợp

Trong những năm gần đây, sinh dữ liệu tổng hợp đã trở thành một hướng tiếp cận phổ biến nhằm giải quyết các thách thức về quyền riêng tư, tính đại diện và công bằng trong học máy [95, 96]. Hình 4.1 minh họa sự phân loại các phương pháp sinh dữ liệu hiện nay thành hai nhóm chính, bao gồm các phương pháp thống kê và các phương pháp học sâu [96, 97, 98]. Sự phân chia này phản ánh sự khác biệt về nguyên lý kỹ thuật và phạm vi ứng dụng, đồng thời là cơ sở để lựa chọn phương pháp phù hợp với từng bài toán cụ thể.



Hình 1.2: Các phương pháp sinh dữ liệu tổng hợp

Phương pháp thống kê: Đây là các kỹ thuật sử dụng mô hình xác suất hoặc các phương pháp bù đắp dữ liệu thiếu để tạo ra dữ liệu tổng hợp có đặc điểm tương tự dữ liệu gốc. Các phương pháp này cố gắng duy trì phân phối giá trị của từng thuộc tính và mối quan hệ giữa các biến, nhờ đó dữ liệu sinh ra có thể phản ánh hợp lý cấu trúc thống kê của dữ liệu thật. Ưu điểm của các phương pháp thống kê là đơn giản, hiệu quả về mặt tính toán và phù hợp với các bộ dữ liệu nhỏ hoặc vừa [99]. Tuy nhiên, chúng thường gặp khó khăn khi mô phỏng chính xác các mối quan hệ phi tuyến hoặc dữ liệu có nhiều chiều [100].

Phương pháp học sâu: Nhóm phương pháp này sử dụng mạng nơ-ron nhân tạo để học các đặc điểm phức tạp của dữ liệu và sinh ra các bản ghi dữ liệu tổng hợp có tính chân thực cao. Một trong những hướng tiếp cận phổ biến nhất là các mô hình dựa trên Mạng đối kháng sinh (GAN) như CTGAN, ADSSGAN và PATEGAN [101]. Các mô hình này cho thấy khả năng sinh dữ liệu vừa chính xác vừa hữu ích cho các nhiệm vụ học máy [96, 100].

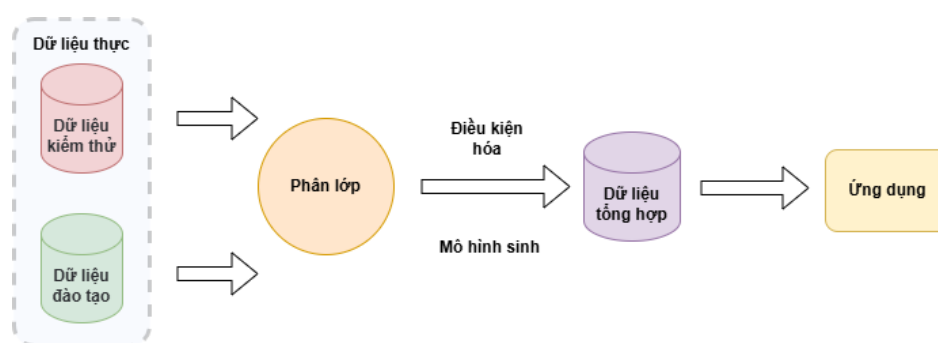
Bên cạnh GAN, các mô hình tự mã hóa biến phân (Variational Autoencoder – VAE) cũng được sử dụng rộng rãi trong sinh dữ liệu bảng. Đại diện tiêu biểu là TVAE, có khả năng học biểu diễn tiềm ẩn của dữ liệu và sinh dữ liệu ổn định hơn so với GAN. Tuy nhiên, dữ liệu sinh từ VAE đôi khi chưa phản ánh đầy đủ các trường hợp hiếm hoặc các mối quan hệ phức tạp giữa các thuộc tính [98].

Gần đây, các mô hình khuếch tán (Diffusion Models) được xem là một hướng tiếp cận đầy tiềm năng trong sinh dữ liệu tổng hợp. Các mô hình như TabDDPM cho thấy khả năng mô hình hóa phân phối dữ liệu tốt và tạo dữ liệu có chất lượng cao hơn trong một số trường hợp. Tuy nhiên, các mô hình này thường yêu cầu chi phí tính toán và thời gian huấn luyện lớn hơn so với GAN, do đó việc triển khai trong thực tế còn gặp nhiều hạn chế [98]. Bên cạnh đó, các mô hình ngôn ngữ lớn (LLMs) cũng được áp dụng để sinh dữ liệu bảng tổng hợp; các mô hình như DGPT có thể tạo dữ liệu có cấu trúc gần với dữ liệu thực và linh hoạt hơn khi mở rộng trên các tập dữ liệu lớn và đa dạng [92, 94, 102, 103].

Mỗi nhóm phương pháp sinh dữ liệu đều có những ưu điểm và hạn chế riêng. Các phương pháp thống kê có ưu điểm là đơn giản, dễ triển khai và hiệu quả trên các bộ dữ liệu nhỏ nhưng gặp khó khăn khi mô hình hóa các quan hệ phi tuyến. Các mô hình VAE và GAN có khả năng học các phân bố dữ liệu phức tạp hơn, trong đó GAN thường tạo ra dữ liệu có chất lượng cao nhưng quá trình huấn luyện dễ mất ổn định. Các mô hình khuếch tán cho chất lượng dữ liệu tốt nhưng yêu cầu chi phí tính toán lớn, trong khi LLM có ưu thế về khả năng suy luận ngữ nghĩa và tính linh hoạt trong sinh dữ liệu, song chất lượng dữ liệu phụ thuộc đáng kể vào thiết kế prompt và tính ngẫu nhiên của mô hình. Nhìn chung, các phương pháp học sâu ngày càng được ưa chuộng nhờ khả năng mô hình hóa dữ liệu phức tạp và đa chiều, mặc dù việc cân bằng giữa chất lượng dữ liệu sinh và chi phí tính toán vẫn là một thách thức [104].

Trên cơ sở phân tích các nhóm phương pháp sinh dữ liệu tổng hợp, luận án lựa chọn CTGAN và LLM làm hai kỹ thuật sinh dữ liệu trong nghiên cứu. CTGAN đại diện cho nhóm mô hình học phân phối dữ liệu bảng, phù hợp với dữ liệu hỗn hợp và dữ liệu mất cân bằng, trong khi LLM đại diện cho hướng tiếp cận mới dựa trên tri thức và thiết kế prompt, cho phép khai thác khả năng suy luận của các mô hình ngôn ngữ lớn trong sinh dữ liệu. Việc lựa chọn đồng thời hai phương pháp này cho phép so sánh và đánh giá ảnh hưởng của hai cơ

chế sinh dữ liệu khác nhau đối với hiệu quả cải thiện công bằng của mô hình học máy. Trong nghiên cứu này, CTGAN và LLM được áp dụng để sinh dữ liệu bảng tổng hợp trong bối cảnh giáo dục nhằm tăng cường tính đại diện và cân bằng dữ liệu huấn luyện cho các nhóm người học khác nhau, từ đó hỗ trợ cải thiện công bằng ngay từ giai đoạn tiền xử lý. Quy trình sinh dữ liệu tổng hợp điển hình được minh họa trong Hình 1.3, trong đó mô hình được huấn luyện trên dữ liệu thực và sau đó sinh ra các bản ghi tổng hợp có đặc tính thống kê tương tự dữ liệu gốc, phục vụ các mục tiêu như cân bằng dữ liệu, huấn luyện mô hình và bảo vệ quyền riêng tư [2].



Hình 1.3: Quy trình sinh dữ liệu tổng hợp [2]

1.3.3. Thách thức và cơ hội của dữ liệu tổng hợp trong việc đảm bảo tính công bằng

Việc sử dụng dữ liệu tổng hợp mang lại cả cơ hội và thách thức trong việc tăng cường tính công bằng cho các hệ thống học máy. Một mặt, dữ liệu tổng hợp có thể giúp giảm thiên vị và cải thiện sự đa dạng của dữ liệu huấn luyện [95]. Mặt khác, quá trình tạo dữ liệu nhân tạo cũng có thể sao chép hoặc khuếch đại các thành kiến vốn có trong dữ liệu gốc, làm suy giảm khả năng đại diện của các nhóm thiểu số [105].

Những thách thức chính khi sử dụng dữ liệu tổng hợp bao gồm: (i) sao chép thiên vị từ dữ liệu gốc, (ii) thiếu đa dạng dữ liệu, (iii) thiếu cơ chế kiểm soát công bằng, và (iv) đánh giá công bằng chưa toàn diện. Dữ liệu tổng hợp có thể kế thừa các sai lệch từ dữ liệu huấn luyện ban đầu [106]. Ngoài ra, một số mô hình sinh dữ liệu tổng hợp như GAN có xu hướng sinh ra các mẫu tương tự

nhau, dẫn đến hạn chế về tính đa dạng và khả năng đại diện cho các nhóm thiểu số [107]. Bên cạnh đó, bản thân quá trình sinh dữ liệu không tự động đảm bảo công bằng nếu không tích hợp các ràng buộc phù hợp [106]. Hơn nữa, nhiều nghiên cứu hiện nay chỉ đánh giá công bằng dựa trên một số chỉ số riêng lẻ mà chưa xem xét đầy đủ các mối quan hệ giao thoa giữa các thuộc tính nhạy cảm [108, 109]. Tuy nhiên, một số nghiên cứu cho rằng dữ liệu tổng hợp có thể được thiết kế có chủ đích để giảm thiên vị và cải thiện công bằng nếu được phát triển với định hướng phù hợp [110, 111].

1.4. Tổng quan nghiên cứu về công bằng cho các hệ thống học máy ứng dụng trong lĩnh vực giáo dục

Trong lĩnh vực giáo dục, các hệ thống học máy ngày càng được sử dụng rộng rãi để hỗ trợ hoặc thay thế con người trong các quyết định quan trọng như chấm bài luận tự động, dự đoán khả năng tốt nghiệp, tuyển sinh hoặc học tập cá nhân hóa [6]. Tuy nhiên, nhiều nghiên cứu cho thấy các hệ thống này có thể tiềm ẩn nguy cơ thiên vị, chẳng hạn quá trình chấm bài có thể bị ảnh hưởng bởi ngôn ngữ mẹ đẻ của sinh viên [60], hoặc các mô hình dự đoán kết quả học tập và tuyển sinh có thể chịu tác động của các thuộc tính nhạy cảm như *giới tính* hoặc *chủng tộc* [112, 113, 114]. Điều này cho thấy việc xem xét yếu tố công bằng là cần thiết khi áp dụng học máy trong giáo dục, bởi các quyết định này ảnh hưởng trực tiếp đến cơ hội học tập và phát triển của người học [7]. Do đó, đảm bảo công bằng trở thành điều kiện quan trọng để xây dựng niềm tin khi triển khai các hệ thống ra quyết định tự động [83, 115].

Công bằng trong học máy giáo dục đã trở thành một chủ đề nghiên cứu nổi bật trong những năm gần đây [6, 7]. Nhiều nghiên cứu tổng quan đã phân tích nguồn gốc của thiên vị, các thuộc tính nhạy cảm, thước đo công bằng và các chiến lược giảm thiểu thiên vị trong các hệ thống giáo dục [6, 7, 64]. Trên cơ sở đó, một nghiên cứu tổng quan có hệ thống được thực hiện vào năm 2023. Nghiên cứu đã khảo sát 63 công trình liên quan đến công bằng trong học máy cho giáo dục giai đoạn 2002–2023, tập trung vào các thuật toán học máy, định nghĩa công bằng, đặc điểm dữ liệu giáo dục, các phương pháp can thiệp và các

thước đo đánh giá công bằng. Kết quả khảo sát đã cung cấp cái nhìn toàn diện về sự phát triển của lĩnh vực này, đồng thời chỉ ra các xu hướng nghiên cứu, hạn chế và khoảng trống còn tồn tại. Những kết quả này đóng vai trò là cơ sở lý luận quan trọng cho việc đề xuất các phương pháp đảm bảo công bằng trong các chương tiếp theo của luận án.

Kết quả khảo sát cho thấy các nghiên cứu chủ yếu tập trung vào các bài toán dự đoán kết quả học tập, dự đoán nguy cơ bỏ học, chấm điểm tự động và hỗ trợ ra quyết định. Các mô hình được sử dụng chủ yếu là các thuật toán học máy truyền thống như *Hồi quy logistic*, *Cây quyết định* và *Rừng ngẫu nhiên* nhờ khả năng diễn giải tốt và phù hợp với đặc điểm của dữ liệu giáo dục [6, 7, 116, 117]. Điều này cho thấy khả năng diễn giải vẫn là một yêu cầu quan trọng đối với các hệ thống học máy trong giáo dục.

Kết quả khảo sát cũng cho thấy giới tính, chủng tộc, dân tộc, tuổi và quốc tịch là các thuộc tính nhạy cảm được nghiên cứu nhiều nhất [7, 50, 90]. Về đánh giá công bằng, các nghiên cứu chủ yếu áp dụng các khái niệm công bằng theo nhóm cùng các thước đo như “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch cơ hội công bằng*” và “*chênh lệch trung bình xác suất*”, kết hợp với các phương pháp tiền xử lý, trong xử lý và hậu xử lý nhằm giảm thiểu thiên vị [1, 14, 15]. Tuy nhiên, việc lựa chọn định nghĩa công bằng, thước đo và phương pháp can thiệp vẫn phụ thuộc vào từng bài toán và chưa có một giải pháp được chấp nhận rộng rãi.

Bên cạnh những kết quả đạt được, nghiên cứu tổng quan cũng chỉ ra một số hạn chế của các nghiên cứu hiện nay. Phần lớn các công trình mới chỉ xem xét từng thuộc tính nhạy cảm riêng lẻ, trong khi các nhóm giao thoa và tình trạng mất cân bằng dữ liệu chưa được nghiên cứu đầy đủ. Đồng thời, việc đánh giá chủ yếu tập trung vào công bằng hoặc hiệu suất mô hình một cách riêng rẽ, còn thiếu các phương pháp đánh giá toàn diện mối quan hệ đánh đổi giữa hai mục tiêu này [17, 37]. Những kết quả này cũng thống nhất với các thách thức chung của lĩnh vực công bằng trong học máy đã được trình bày tại Mục 0.1.2, đồng thời cho thấy trong bối cảnh giáo dục vẫn còn tồn tại các khoảng trống liên quan đến xử lý nhiều thuộc tính nhạy cảm, khắc phục mất cân bằng dữ liệu và đánh giá đồng thời công bằng với hiệu suất. Đây cũng chính là cơ sở để NCS đề xuất các phương pháp trong các chương tiếp theo của luận án.

1.5. Mối quan hệ giữa công bằng và hiệu suất của các hệ thống học máy

Mối quan hệ giữa công bằng và hiệu suất là một trong những chủ đề nghiên cứu quan trọng trong hệ thống học máy, đặc biệt khi học máy ngày càng được ứng dụng trong các lĩnh vực như giáo dục, y tế và tài chính. Quan điểm truyền thống cho rằng việc cải thiện công bằng thường làm giảm hiệu suất, nhất là độ chính xác dự đoán. Tuy nhiên, nhiều nghiên cứu gần đây cho thấy sự đánh đổi này không phải lúc nào cũng tồn tại và có thể được giảm thiểu thông qua các phương pháp kỹ thuật phù hợp. Mục này trình bày mối quan hệ đánh đổi giữa công bằng và hiệu suất, các nguyên nhân dẫn đến sự đánh đổi, cùng những hướng tiếp cận nhằm cân bằng hai mục tiêu này.

1.5.1. Sự đánh đổi giữa công bằng và hiệu suất

Có nhiều nghiên cứu đã ghi nhận sự đánh đổi rõ rệt giữa công bằng và hiệu suất của mô hình, trong đó việc tăng cường công bằng có thể dẫn đến suy giảm độ chính xác của mô hình. Trong một nghiên cứu vào năm 2024 của Nathim và cộng sự chỉ ra rằng khi áp dụng các biện pháp can thiệp nhằm cải thiện công bằng, sai lệch trong mô hình giảm trung bình 23%, trong khi độ chính xác tổng thể giảm khoảng 9% [118]. Điều này làm dấy lên câu hỏi về mức độ sẵn sàng đánh đổi hiệu suất để đạt được mức công bằng cao hơn, đặc biệt trong các lĩnh vực yêu cầu độ chính xác nghiêm ngặt như dự đoán năng lực học tập, sàng lọc người học có nguy cơ, hoặc xếp lớp. Ngược lại với quan điểm đánh đổi truyền thống, một số nghiên cứu đã chỉ ra rằng công bằng có thể được cải thiện mà không làm tổn hại, thậm chí có thể nâng cao hiệu suất mô hình. Nghiên cứu của Islam và cộng sự năm 2021 cho thấy việc điều chỉnh siêu tham số có thể đồng thời cải thiện cả hiệu suất và công bằng [119]. Các mô hình học máy khi được huấn luyện trên dữ liệu đã được xử lý để loại bỏ thiên vị thường có khả năng tổng quát hóa tốt hơn và cho kết quả ổn định hơn trên dữ liệu kiểm thử không thiên vị [120].

Đứng ở góc độ nhân quả, một số nghiên cứu gần đây đã chỉ ra rằng nguồn gốc của sự đánh đổi nằm ở các sai lệch cấu trúc trong dữ liệu huấn luyện [120, 121]. Khi các sai lệch này được khắc phục thông qua các kỹ thuật tiền xử lý, chẳng hạn như loại bỏ thành kiến thông qua hồi quy hoặc cân bằng phân phối theo nhóm, mô hình học máy không chỉ trở nên công bằng hơn mà còn có thể đạt hiệu suất cao hơn trên dữ liệu không thiên vị [120, 121]. Leininger và cộng sự thậm chí còn chỉ ra rằng một số phương pháp tiền xử lý có thể mô phỏng gần đúng một thế giới “công bằng”, nơi các thuộc tính nhạy cảm không còn ảnh hưởng đến đầu ra mô hình. Khi đó, hai mục tiêu tưởng như mâu thuẫn là công bằng và hiệu suất có thể được tối ưu song song trong thực tiễn [120].

1.5.2. Các hướng tiếp cận để xử lý đánh đổi

Mặc dù vẫn tồn tại những tình huống có sự đánh đổi giữa công bằng và hiệu suất, nhưng ngày càng có nhiều bằng chứng cho thấy hai mục tiêu này không nhất thiết loại trừ nhau. Với các phương pháp tiếp cận phù hợp như tiền xử lý dữ liệu, điều chỉnh siêu tham số, hoặc sử dụng khuôn khổ nhân quả có thể xây dựng các mô hình học máy vừa công bằng, vừa hiệu quả. Điều này đặc biệt quan trọng trong lĩnh vực giáo dục, nơi tính công bằng không chỉ là yêu cầu đạo đức, mà còn liên quan trực tiếp đến quyền tiếp cận cơ hội học tập và phát triển của người học. Với mục tiêu đạt được sự cân bằng tối ưu giữa việc giảm thiên vị và duy trì độ chính xác của mô hình. Một số giải pháp chính được ghi nhận trong các nghiên cứu gần đây có thể kể đến như: *“giảm thiểu thiên vị trong quá trình huấn luyện”*, *“tiền xử lý dựa trên nhân quả”*, và *“tối ưu hóa đa mục tiêu”*. *Kỹ thuật giảm thiểu thiên vị trong quá trình huấn luyện*: Một số phương pháp như học tập chuyển đổi với tối ưu hóa công bằng được đề xuất bởi Wang và cộng sự năm 2024 đã đề xuất việc tối ưu hóa hiệu suất mô hình trước, sau đó điều chỉnh để cải thiện công bằng. Cách tiếp cận này nhằm giảm thiểu sự đánh đổi tiêu cực giữa hai mục tiêu, bằng cách bảo toàn hiệu suất trong khi nâng cao tính công bằng [122]. *Tiền xử lý dựa trên nhân quả*: Phương pháp này tìm cách tiệm cận với công bằng bằng cách loại bỏ hoặc trung hòa ảnh hưởng nhân quả của các thuộc tính nhạy cảm. Điều này đảm bảo mô hình có thể vừa đảm bảo được công bằng, vừa đạt được hiệu suất dự đoán cao [31, 120]. *Tối ưu hóa đa mục tiêu*: Một hướng tiếp cận khác là coi công bằng và hiệu suất như hai mục

tiêu cần tối ưu đồng thời trong một bài toán đa mục tiêu. Các phương pháp tối ưu hóa ngẫu nhiên cho phép xác định biên Pareto thể hiện sự đánh đổi giữa độ chính xác và công bằng, từ đó cung cấp góc nhìn sâu sắc và nhiều sắc thái hơn về mối quan hệ giữa hai khía cạnh này [123].

Mặc dù các kỹ thuật trên cho thấy tiềm năng trong việc cân bằng giữa hiệu suất và công bằng, nhưng bản chất phức tạp của các thành kiến xã hội, cùng với yêu cầu thích ứng mô hình theo từng miền ứng dụng cụ thể, vẫn đặt ra những thách thức đáng kể. Do đó, việc thiết kế các hệ thống học máy công bằng cần được tiếp cận một cách linh hoạt, đồng thời có nhận thức rõ về bối cảnh dữ liệu và người dùng mục tiêu.

1.6. Câu hỏi nghiên cứu

Thông qua việc tổng quan và phân tích có hệ thống các công trình nghiên cứu về công bằng trong các hệ thống học máy ứng dụng cho lĩnh vực giáo dục, Chương 1 đã chỉ ra những hướng tiếp cận chủ đạo, các kết quả đạt được, cũng như những hạn chế và khoảng trống nghiên cứu còn tồn tại. Đặc biệt, các vấn đề liên quan đến dữ liệu giáo dục dạng bảng, sự tồn tại đồng thời của nhiều thuộc tính nhạy cảm, tình trạng mất cân bằng dữ liệu giữa các nhóm con, và việc đánh giá mối quan hệ đánh đổi giữa công bằng và hiệu suất của mô hình vẫn chưa được giải quyết một cách toàn diện trong các nghiên cứu hiện có.

Trên cơ sở các khoảng trống nghiên cứu được xác định, luận án tập trung làm rõ các câu hỏi nghiên cứu sau đây, đóng vai trò định hướng cho các chương nghiên cứu tiếp theo:

- *RQ1*: Trong bối cảnh dữ liệu giáo dục dạng bảng có chứa nhiều thuộc tính nhạy cảm, làm thế nào để đảm bảo tính công bằng cho đồng thời tất cả các thuộc tính mà không làm ảnh hưởng đến hiệu suất của mô hình?
- *RQ2*: Việc mất cân bằng dữ liệu cũng như thiếu dữ liệu đào tạo có ảnh hưởng như thế nào đến tính công bằng của các mô hình học máy trong giáo dục, và liệu việc sinh thêm dữ liệu tự động cùng các cơ chế can thiệp dựa trên phân bố dữ liệu có thể cải thiện công bằng cho các mô hình học máy trong lĩnh vực giáo dục?

- *RQ3*: việc kết hợp nhiều cơ chế can thiệp có giúp cải thiện công bằng cho các mô hình học máy trong lĩnh vực giáo dục một cách hiệu quả và bền vững hơn so với các cách tiếp cận đơn lẻ hay không?
- *RQ4*: Mối quan hệ đánh đổi giữa công bằng và hiệu suất của các hệ thống học máy trong lĩnh vực giáo dục được thể hiện như thế nào, và liệu có thể xây dựng một thước đo tổng hợp nhằm hỗ trợ việc đánh giá, so sánh và lựa chọn mô hình một cách có hệ thống hay không?

Các câu hỏi nghiên cứu trên được xây dựng dựa trên việc nghiên cứu có tính hệ thống đã được trình bày trong phần tổng quan, đồng thời định hướng cho toàn bộ nội dung nghiên cứu của luận án. Việc trả lời các câu hỏi này được triển khai lần lượt trong các chương tiếp theo, tương ứng với các hướng nghiên cứu và đóng góp chính của luận án về đảm bảo công bằng trong các hệ thống học máy ứng dụng cho giáo dục.

1.7. Tổng kết chương

Chương 1 đã xây dựng nền tảng lý thuyết và bức tranh tổng quan có hệ thống về vấn đề công bằng trong các hệ thống học máy, với trọng tâm là bối cảnh ứng dụng trong giáo dục. Nội dung chương không chỉ làm rõ các khái niệm cốt lõi mà còn định hình rõ các hướng tiếp cận và khoảng trống nghiên cứu, tạo cơ sở khoa học cho các phương pháp được đề xuất trong các chương tiếp theo của luận án.

Trước hết, chương đã khái quát các khái niệm nền tảng liên quan đến hệ thống học máy, bao gồm các mô hình học có giám sát và không giám sát, vai trò của trí tuệ nhân tạo và học máy trong giáo dục, cũng như các chỉ số đánh giá hiệu suất phổ biến như “*độ chuẩn xác*”, “*độ chính xác*”, “*độ hồi tưởng*” và “*điểm số F1*”. Phần này cung cấp bối cảnh kỹ thuật cần thiết để hiểu rõ mối quan hệ giữa hiệu suất dự đoán và các yêu cầu chất lượng khác của hệ thống học máy.

Tiếp theo, chương phân tích sâu các khái niệm về công bằng và thiên vị trong học máy, làm rõ sự khác biệt giữa các cách tiếp cận công bằng như công bằng theo nhóm, công bằng cá nhân và các định nghĩa công bằng dựa trên phân phối,

điều kiện hoặc phản thực. Trên cơ sở đó, chương đã hệ thống hóa các độ đo công bằng thường được sử dụng trong nghiên cứu và thực tiễn, đồng thời chỉ ra những hạn chế khi áp dụng riêng lẻ từng độ đo trong các bài toán có nhiều thuộc tính nhạy cảm.

Bên cạnh đó, chương đã phân loại và tổng hợp các nhóm phương pháp đảm bảo tính công bằng trong học máy, bao gồm các phương pháp tiền xử lý, xử lý trong quá trình huấn luyện và hậu xử lý. Phân tích này cho thấy mỗi nhóm phương pháp đều có ưu điểm và giới hạn riêng, đặc biệt trong bối cảnh dữ liệu giáo dục thường mất cân bằng, khan hiếm ở các nhóm giao thoa và chịu ảnh hưởng của nhiều yếu tố xã hội nhạy cảm. Chương cũng nhấn mạnh rằng các phương pháp đơn lẻ khó có thể xử lý triệt để các dạng thiên lệch phức tạp phát sinh khi nhiều thuộc tính nhạy cảm cùng tồn tại.

Cuối cùng, chương đã thảo luận mối quan hệ đánh đổi giữa công bằng và hiệu suất trong các hệ thống học máy. Mặc dù nhiều nghiên cứu chỉ ra rằng sự đánh đổi này là khó tránh khỏi, tổng quan cho thấy vẫn tồn tại những phương pháp có thể đạt được sự cân bằng hợp lý thông qua các chiến lược như tiền xử lý dữ liệu, loại bỏ phụ thuộc giữa các thuộc tính không nhạy cảm và thuộc tính nhạy cảm, cân bằng phân phối dữ liệu và tối ưu hóa đa mục tiêu. Những phân tích này đóng vai trò quan trọng trong việc định hướng luận án lựa chọn cách tiếp cận tập trung vào can thiệp ở mức dữ liệu và thuộc tính, thay vì chỉ điều chỉnh mô hình hoặc đầu ra.

Chương 1 đã xác lập rõ bối cảnh khoa học, các thách thức then chốt và khoảng trống nghiên cứu trong lĩnh vực công bằng cho học máy ứng dụng trong giáo dục. Các kết quả tổng hợp trong chương là nền tảng trực tiếp cho việc đề xuất và phát triển các phương pháp Fairedu, DPF và FaireduPlus trong các chương tiếp theo, cũng như cho việc xây dựng chỉ số đánh đổi giữa công bằng và hiệu suất.

Các nội dung tổng quan trong chương này đã được công bố trong hai công trình khoa học do tác giả đóng vai trò tác giả chính, bao gồm: (i) một bài viết đăng trong tuyển tập *Lecture Notes in Information Systems and Organisation* (Springer, 2023); và (ii) một bài báo công bố trên tạp chí *Journal of Systems and Software* (Q1, Elsevier, 2024).

Chương 2

PHƯƠNG PHÁP ĐẢM BẢO TÍNH CÔNG BẰNG NHỜ LOẠI BỎ SỰ PHỤ THUỘC VÀO CÁC THUỘC TÍNH NHẠY CẢM

Chương này giải quyết câu hỏi nghiên cứu *RQ1* được đặt ra trong Chương 1, tập trung vào bài toán đảm bảo tính công bằng cho các hệ thống học máy khi dữ liệu huấn luyện chứa đồng thời nhiều thuộc tính nhạy cảm và tồn tại sự phụ thuộc giữa các thuộc tính không nhạy cảm với các thuộc tính này. Để giải quyết bài toán, luận án đề xuất phương pháp tiền xử lý *FairEdu*, sử dụng hồi quy đa biến để phát hiện và loại bỏ đồng thời các mối phụ thuộc này trước khi huấn luyện mô hình học máy.

Chương gồm bốn phần chính: (i) trình bày nguyên lý hoạt động và kiến trúc tổng thể của *FairEdu*; (ii) mô tả thuật toán và khả năng xử lý đồng thời nhiều thuộc tính nhạy cảm; (iii) thiết lập thực nghiệm và đánh giá *FairEdu* trên các bộ dữ liệu *Student Performance*, *Student Predict Dropout*, *Oulad* và *DNU Data* với các mô hình *Hồi quy logistic*, *Cây quyết định* và *Rừng ngẫu nhiên*; và (iv) thảo luận kết quả, hạn chế và ý nghĩa của phương pháp.

2.1. Giới thiệu

Trong lĩnh vực giáo dục, tính công bằng là một yêu cầu quan trọng đối với các hệ thống học máy do các quyết định của mô hình có thể ảnh hưởng trực tiếp đến cơ hội học tập và sự phát triển của người học. Tuy nhiên, khi dữ liệu huấn luyện chứa thiên vị, các mô hình có thể duy trì hoặc khuếch đại sự bất bình đẳng. Như đã phân tích trong Chương 1, việc đồng thời đảm bảo tính công bằng và hiệu suất dự đoán vẫn là một thách thức lớn [59, 124, 125].

Trong số các phương pháp tiền xử lý, *phương pháp gỡ lỗi dữ liệu đào tạo bằng hồi quy tuyến tính* (LTDD) do Li và cộng sự đề xuất năm 2022 là một kỹ thuật tiêu biểu nhằm loại bỏ sự phụ thuộc giữa các thuộc tính đầu vào và thuộc tính nhạy cảm trong dữ liệu huấn luyện [31]. Phương pháp này dựa trên hồi quy tuyến tính đơn, trong đó mỗi thuộc tính không nhạy cảm X được biểu diễn như một hàm của thuộc tính nhạy cảm A , phần dư $\hat{\epsilon} = X - \hat{X}$ được sử dụng thay thế cho X trong quá trình huấn luyện nhằm loại bỏ sự phụ thuộc vào các thuộc tính nhạy cảm. Cách tiếp cận này cho phép tách biệt phần thông tin của dữ liệu không bị ảnh hưởng bởi thuộc tính nhạy cảm, từ đó giảm khả năng mô hình học được các mối quan hệ thiên lệch. Nhờ tính đơn giản, hiệu quả và khả năng áp dụng độc lập với mô hình học máy, LTDD được xem là một giải pháp tiền xử lý thực tiễn. Tuy nhiên, phương pháp này chỉ xử lý cho từng thuộc tính nhạy cảm riêng lẻ, khi tồn tại nhiều thuộc tính nhạy cảm, việc áp dụng tuần tự có thể dẫn đến xung đột trong điều chỉnh và không đảm bảo loại bỏ phụ thuộc một cách nhất quán.

Để giải quyết vấn đề công bằng cho đồng thời nhiều thuộc tính nhạy cảm, năm 2024, Chen và cộng sự đã đưa ra khái niệm công bằng giao thoa [33]. Thay vì xem xét từng thuộc tính riêng lẻ, các thuộc tính nhạy cảm được kết hợp để tạo thành các nhóm con giao thoa như “nữ – da màu – trẻ” hoặc “nam – da trắng – lớn tuổi”, và công bằng được đánh giá dựa trên độ chênh lệch lớn nhất giữa các nhóm này thông qua các chỉ số như SPD, AOD và EOD. Cách tiếp cận này cho phép phát hiện các dạng thiên lệch tiềm ẩn ở các nhóm thiểu số giao thoa, vốn thường bị bỏ sót trong các phương pháp truyền thống, đồng thời cung cấp cái nhìn chi tiết hơn về mức độ công bằng giữa các nhóm trong quần thể. Tuy nhiên, phương pháp này cũng gặp phải những thách thức đáng kể. Việc kết hợp nhiều thuộc tính nhạy cảm dẫn đến số lượng nhóm con tăng theo cấp số nhân, gây ra hiện tượng bùng nổ tổ hợp và làm gia tăng độ phức tạp tính toán. Đồng thời, nhiều nhóm con có kích thước rất nhỏ, dẫn đến tình trạng dữ liệu thưa và mất cân bằng nghiêm trọng, ảnh hưởng đến độ ổn định và độ tin cậy của mô hình.

Những hạn chế trên cho thấy các phương pháp hiện tại chưa cung cấp được một giải pháp hiệu quả để xử lý đồng thời nhiều thuộc tính nhạy cảm trong dữ liệu giáo dục dạng bảng. Để khắc phục, luận án đề xuất khái niệm “*công bằng*

đồng thời” và phương pháp tiên xử lý *Fairedu*, nhằm loại bỏ sự phụ thuộc của dữ liệu vào nhiều thuộc tính nhạy cảm cùng lúc.

Fairedu được phát triển từ *phương pháp gỡ lỗi dữ liệu đào tạo bằng hồi quy tuyến tính* nhưng mở rộng từ hồi quy tuyến tính đơn sang hồi quy đa biến, cho phép điều chỉnh đồng thời nhiều thuộc tính nhạy cảm. Phương pháp gồm ba bước chính: phát hiện phụ thuộc, ước lượng bằng hồi quy đa biến, và điều chỉnh dữ liệu để loại bỏ ảnh hưởng của thuộc tính nhạy cảm.

Kết quả thực nghiệm trên nhiều bộ dữ liệu và mô hình học máy cho thấy *Fairedu* cải thiện đáng kể các chỉ số công bằng (như DI, SPD, AOD, EOD) trong khi chỉ làm suy giảm rất nhỏ hiệu suất dự đoán. Điều này khẳng định *Fairedu* là một phương pháp đơn giản, hiệu quả và khả thi cho bài toán đảm bảo công bằng trong học máy giáo dục.

2.2. Phương pháp *Fairedu*

2.2.1. Nguyên lý hoạt động

Phương pháp *Fairedu* được đề xuất như một mở rộng tổng quát của LTDD, vừa kế thừa tính đơn giản vừa khắc phục hạn chế quan trọng của phương pháp này trong việc chỉ xử lý từng thuộc tính nhạy cảm riêng lẻ. Cụ thể, trong khi LTDD sử dụng hồi quy tuyến tính đơn biến để loại bỏ sự phụ thuộc giữa từng cặp thuộc tính đầu vào và thuộc tính nhạy cảm, thì *Fairedu* áp dụng hồi quy đa biến, cho phép xử lý đồng thời nhiều thuộc tính nhạy cảm. Nhờ đó, *Fairedu* phù hợp hơn với các bài toán công bằng giao thoa phức tạp trong lĩnh vực giáo dục.

Giả sử một mô hình học máy S_{ML} ánh xạ véc tơ đặc trưng $\mathbf{x} = [x_1, x_2, \dots, x_d] \in \mathbb{R}^d$ tới nhãn lớp $y \in \{0, 1\}$, được biểu diễn như trong Công thức 2.1.

$$S_{ML} : \mathbb{R}^d \rightarrow \{0, 1\} \quad (2.1)$$

Fairedu xử lý để loại bỏ mối quan hệ của mỗi thuộc tính không nhạy cảm x_i ($k + 1 \leq i \leq d$) với k thuộc tính nhạy cảm x_1, \dots, x_k bằng kỹ thuật hồi quy đa biến. Qua mô hình hồi quy tuyến tính đa biến, mỗi giá trị x_i tương ứng với giá trị \hat{x}_i được xác định bằng Công thức 2.2.

$$\hat{x}_i = \beta_{i0} + \beta_{i1}x_1 + \beta_{i2}x_2 + \dots + \beta_{ik}x_k, i \in \{k + 1, \dots, d\} \quad (2.2)$$

Khi đó, mỗi giá trị $x_i, i \in \{k + 1, \dots, d\}$ được biểu diễn tuyến tính theo \hat{x}_i bằng Công thức 2.3.

$$x_i = \hat{x}_i + \epsilon, i \in \{k + 1, \dots, d\} \quad (2.3)$$

Sau đó, các thuộc tính mới (đã loại bỏ thành phần phụ thuộc vào các thuộc tính nhạy cảm) được tính bằng Công thức 2.4.

$$x_i^{\text{new}} = x_i - \hat{x}_i \quad (2.4)$$

Việc loại bỏ này đồng thời được áp dụng trên cả tập huấn luyện và tập kiểm tra để đảm bảo tính công bằng trong dự đoán.

2.2.2. Kiến trúc tổng thể

Kiến trúc tổng thể của phương pháp Fairedu được minh họa trong Hình 2.1. Phương pháp được tổ chức thành hai pha chính là *pha huấn luyện* và *pha kiểm thử*. Hai pha này liên kết thông qua bộ hệ số hồi quy được ước lượng từ tập huấn luyện, nhờ đó bảo đảm quá trình loại bỏ sự phụ thuộc của các thuộc tính không nhạy cảm vào các thuộc tính nhạy cảm được thực hiện nhất quán trên cả tập huấn luyện và tập kiểm thử. Thiết kế này giúp Fairedu hoạt động như một bước tiền xử lý độc lập, có thể tích hợp với nhiều mô hình học máy khác nhau mà không cần thay đổi kiến trúc hoặc thuật toán huấn luyện của mô hình.

Trong *pha huấn luyện*, tập dữ liệu ban đầu được chia thành tập huấn luyện (85%) và tập kiểm thử (15%). Trên tập huấn luyện, Fairedu xây dựng mô hình hồi quy đa biến giữa các thuộc tính không nhạy cảm và các thuộc tính nhạy

cảm để xác định mức độ phụ thuộc giữa chúng. Đối với những thuộc tính có mối quan hệ tuyến tính có ý nghĩa thống kê (kiểm định Wald với $p\text{-value} < 0,05$), thành phần phụ thuộc vào các thuộc tính nhạy cảm sẽ được loại bỏ theo Công thức 2.4. Kết quả của bước này là tập dữ liệu huấn luyện đã được điều chỉnh cùng với bộ hệ số hồi quy $\beta_{i0}, \beta_{i1}, \dots, \beta_{ik}$, được lưu lại để sử dụng trong pha kiểm thử. Mô hình học máy sau đó được huấn luyện trên tập dữ liệu đã được xử lý.

Trong *pha kiểm thử*, bộ hệ số hồi quy thu được từ pha huấn luyện được sử dụng để loại bỏ sự phụ thuộc của các thuộc tính không nhạy cảm trên tập kiểm thử theo cùng một phép biến đổi đã áp dụng cho tập huấn luyện. Dữ liệu kiểm thử sau khi điều chỉnh được đưa vào mô hình đã huấn luyện để thực hiện dự đoán và đánh giá.

Kết quả cuối cùng được đánh giá đồng thời trên hai khía cạnh: hiệu suất dự đoán của mô hình và mức độ công bằng thông qua các thước đo công bằng tương ứng. Việc tách biệt rõ ràng giữa pha huấn luyện và pha kiểm thử, đồng thời chỉ sử dụng các hệ số hồi quy được học từ tập huấn luyện trong quá trình xử lý tập kiểm thử, giúp đảm bảo tính khách quan của quá trình đánh giá và phản ánh đúng khả năng tổng quát hóa của phương pháp Fairedu.

Tương ứng với hai pha xử lý trên, quy trình của Fairedu bao gồm năm bước: chuẩn bị dữ liệu, loại bỏ sự phụ thuộc trên tập huấn luyện, huấn luyện mô hình, loại bỏ sự phụ thuộc trên tập kiểm thử và đánh giá mô hình. Các bước được thực hiện tuần tự và độc lập với mô hình học máy, giúp Fairedu có thể tích hợp như một kỹ thuật tiền xử lý cho nhiều mô hình khác nhau. Chi tiết từng bước được trình bày dưới đây.

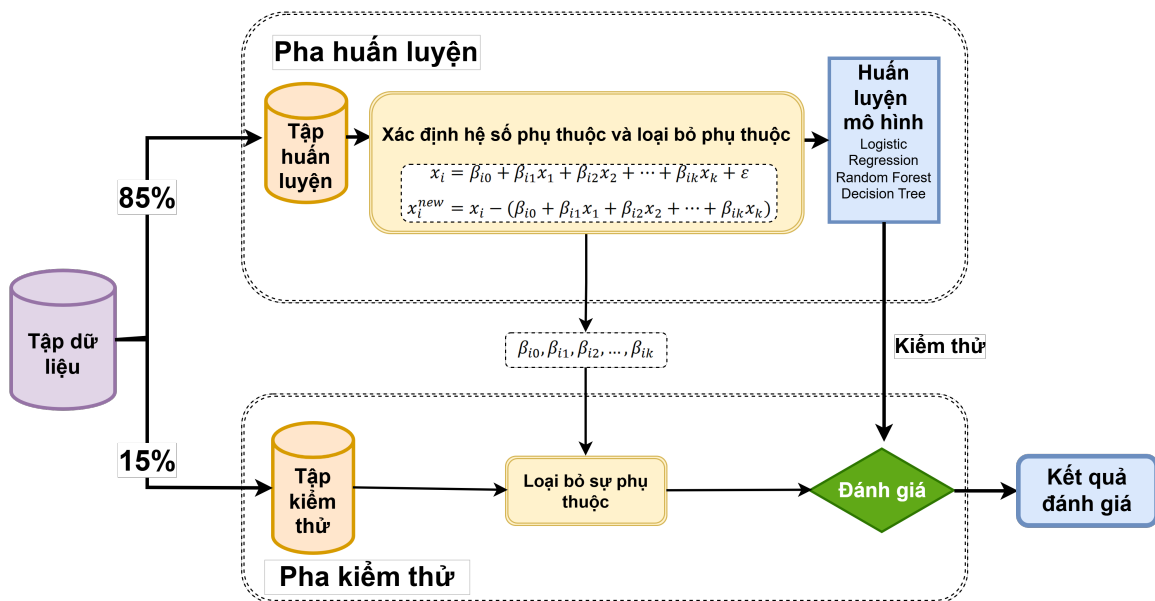
Bước 1. Chuẩn bị dữ liệu: Tập dữ liệu đầu vào bao gồm cả thuộc tính nhạy cảm và không nhạy cảm. Dữ liệu được chia thành hai phần: 85% dành cho huấn luyện và 15% dành cho kiểm tra. Nhằm đảm bảo tính ổn định và giảm phương sai mô hình, kỹ thuật xác thực chéo 10 lần được áp dụng trên tập huấn luyện.

Bước 2. Loại bỏ phụ thuộc trên tập huấn luyện: Với mỗi thuộc tính không nhạy cảm $x_i, i \in \{k+1, \dots, d\}$, tiến hành hồi quy đa biến theo các thuộc tính nhạy cảm x_1, \dots, x_k làm biến đầu vào. Nếu kết quả kiểm định Wald cho thấy $p\text{-value} < 0.05$, tức tồn tại mối liên hệ tuyến tính có ý nghĩa thống kê, thì thuộc tính x_i sẽ được điều chỉnh theo Công thức 2.4 nhằm loại bỏ ảnh hưởng từ các thuộc tính nhạy cảm.

Bước 3. Huấn luyện mô hình: Sau khi hoàn thành quá trình điều chỉnh, các thuộc tính nhạy cảm được loại bỏ khỏi tập dữ liệu. Mô hình học máy được huấn luyện trên tập dữ liệu đã điều chỉnh chỉ bao gồm các thuộc tính không nhạy cảm, nhằm tránh việc mô hình khai thác trực tiếp thông tin từ các thuộc tính nhạy cảm trong quá trình học.

Bước 4. Loại bỏ phụ thuộc trên tập kiểm tra và đánh giá mô hình: Sử dụng các hệ số hồi quy thu được từ Bước 2 để điều chỉnh các thuộc tính tương ứng trong tập kiểm tra. Điều này đảm bảo tính nhất quán giữa tập huấn luyện và kiểm tra trong việc loại bỏ thiên vị.

Bước 5. Đánh giá mô hình: Đánh giá về cả hiệu suất và tính công bằng của mô hình trên tập kiểm tra đã được xử lý trong Bước 4.



Hình 2.1: Kiến trúc tổng thể của phương pháp Fairedu.

2.2.3. Thuật toán

Thuật toán Fairedu sử dụng kỹ thuật hồi quy tuyến tính đa biến để loại bỏ sự phụ thuộc của các thuộc tính đầu vào khỏi các thuộc tính nhạy cảm trong dữ liệu huấn luyện và kiểm tra. Mục tiêu của thuật toán là điều chỉnh dữ liệu đầu vào nhằm giảm thiểu thiên vị do các thuộc tính nhạy cảm gây ra, mà không

cần thay đổi kiến trúc hoặc thuật toán của mô hình học máy. Chi tiết quy trình thực hiện thuật toán Fairedu được trình bày trong Thuật toán 2.1

Thuật toán nhận đầu vào là tập huấn luyện gồm các mẫu dữ liệu và nhãn tương ứng, trong đó mỗi mẫu bao gồm d thuộc tính, với k thuộc tính đầu tiên là các thuộc tính nhạy cảm (ví dụ: *giới tính*, *chủng tộc*) và $d - k$ thuộc tính còn lại là các thuộc tính không nhạy cảm. Thuật toán được thực hiện qua năm bước chính. *Thứ nhất*, khởi tạo các mảng lưu hệ số chệch và các hệ số hồi quy. *Thứ hai*, đối với từng thuộc tính không nhạy cảm, xây dựng mô hình hồi quy tuyến tính đa biến với các thuộc tính nhạy cảm là biến độc lập; nếu kiểm định Wald cho thấy mối quan hệ có ý nghĩa thống kê (p -value < 0.05), các hệ số hồi quy được lưu lại. *Thứ ba*, sử dụng các hệ số này để điều chỉnh tập huấn luyện bằng cách loại bỏ thành phần phụ thuộc của các thuộc tính không nhạy cảm vào các thuộc tính nhạy cảm. *Thứ tư*, huấn luyện mô hình học máy trên tập dữ liệu đã điều chỉnh. *Thứ năm*, tập kiểm thử được xử lý bằng chính các hệ số hồi quy đã ước lượng trước khi đưa vào mô hình để dự đoán kết quả.

Có thể thấy, Fairedu chỉ thực hiện hai thao tác chính là (i) ước lượng và loại bỏ sự phụ thuộc của các thuộc tính không nhạy cảm vào các thuộc tính nhạy cảm và (ii) huấn luyện mô hình trên dữ liệu đã được điều chỉnh. Do không thay đổi thuật toán học máy mà chỉ tác động lên dữ liệu đầu vào, Fairedu có thể được xem như một phương pháp tiền xử lý độc lập và dễ dàng tích hợp với nhiều mô hình học máy khác nhau. Phương pháp Fairedu có thể được áp dụng cho cả các thuộc tính định tính và định lượng, với điều kiện các thuộc tính phân loại được mã hóa thành giá trị số trước khi thực hiện hồi quy. Do đó, thuật toán có khả năng mở rộng tốt trong các hệ thống học máy sử dụng dữ liệu dạng bảng, đặc biệt trong lĩnh vực giáo dục. Điều này đảm bảo khả năng áp dụng Fairedu trên nhiều bộ dữ liệu trong giáo dục với tính chất đa dạng.

Độ phức tạp của thuật toán Fairedu. Giả sử tập dữ liệu có n mẫu, d thuộc tính và k thuộc tính nhạy cảm. Độ phức tạp của phương pháp Fairedu chủ yếu đến từ hai bước: hồi quy đa biến và điều chỉnh dữ liệu. Trong bước hồi quy, với mỗi thuộc tính không nhạy cảm (tổng cộng $d - k$ thuộc tính), thuật toán thực hiện hồi quy tuyến tính đa biến với k biến độc lập, có độ phức tạp xấp xỉ $O(n \cdot k^2)$. Do đó, tổng chi phí của bước này là $O((d - k) \cdot n \cdot k^2)$. Trong bước điều chỉnh dữ liệu, với mỗi mẫu và mỗi thuộc tính không nhạy cảm, thuật toán thực

Thuật toán 2.1 Thuật toán Fairedu dựa trên hồi quy đa biến

- 1: **Input:** Tập huấn luyện $D_{tr} = \{\langle x_1, y_1 \rangle, \dots, \langle x_n, y_n \rangle\}$ với mỗi véc tơ $x_j = [x_1^j, \dots, x_d^j]$, trong đó x_1^j, \dots, x_k^j là các thuộc tính nhạy cảm. Mẫu kiểm tra: $x^{te} = [x_1^{te}, \dots, x_d^{te}]$
- 2: **Output:** Mô hình học máy S_{ML} và nhãn dự đoán $S_{ML}(x^{te})$
- 3: Khởi tạo mảng hệ số chệch $E_a[k+1 : d] = 0$
- 4: Khởi tạo k mảng hệ số E_{b^1}, \dots, E_{b^k} có kích thước $d - k$ và gán giá trị ban đầu bằng 0
- 5: Trích xuất các véc tơ cột V_1, \dots, V_k của các thuộc tính nhạy cảm từ tập huấn luyện
- 6: **for** $i = k + 1$ **đến** d **do**
- 7: Tạo véc tơ V_i tương ứng với thuộc tính không nhạy cảm x_i
- 8: Áp dụng mô hình hồi quy:

$$V_i = a_i + b_i^1 \cdot V_1 + \dots + b_i^k \cdot V_k + \epsilon$$

- 9: Thực hiện kiểm định Wald để kiểm tra ý nghĩa thống kê của mối liên hệ
- 10: **if** p-value < 0.05 **then**
- 11: Ước lượng các hệ số $\hat{a}_i, \hat{b}_i^1, \dots, \hat{b}_i^k$
- 12: Gán vào: $E_a[i] = \hat{a}_i, E_{b^1}[i] = \hat{b}_i^1, \dots, E_{b^k}[i] = \hat{b}_i^k$
- 13: **end if**
- 14: **end for**
- 15: **for** mỗi mẫu $\langle x_j, y_j \rangle \in D_{tr}$ **do**
- 16: Loại bỏ các thuộc tính nhạy cảm khỏi x_j
- 17: **for** $i = k + 1$ **đến** d **do**
- 18: Cập nhật thuộc tính:

$$x_i^j = x_i^j - (E_a[i] + E_{b^1}[i] \cdot x_1^j + \dots + E_{b^k}[i] \cdot x_k^j)$$

- 19: **end for**
- 20: **end for**
- 21: Huấn luyện mô hình S_{ML} với tập dữ liệu huấn luyện đã điều chỉnh
- 22: Loại bỏ thuộc tính nhạy cảm khỏi mẫu kiểm tra x^{te}
- 23: **for** $i = k + 1$ **đến** d **do**
- 24: Cập nhật thuộc tính:

$$x_i^{te} = x_i^{te} - (E_a[i] + E_{b^1}[i] \cdot x_1^{te} + \dots + E_{b^k}[i] \cdot x_k^{te})$$

- 25: **end for**
 - 26: **return** S_{ML} và $S_{ML}(x^{te}) = 0$
-

hiện phép cập nhật tuyến tính với k biến, dẫn đến độ phức tạp $O(n \cdot (d - k) \cdot k)$.

Tổng thể, độ phức tạp của thuật toán Fairedu là:

$$O((d - k) \cdot n \cdot k^2 + n \cdot (d - k) \cdot k),$$

trong đó thành phần chi phối là bước hồi quy đa biến. Trong thực tế, do số lượng thuộc tính nhạy cảm k thường nhỏ, phương pháp vẫn đảm bảo tính khả thi đối với các bộ dữ liệu giáo dục dạng bảng.

So với *phương pháp gỡ lỗi dữ liệu đào tạo bằng hồi quy tuyến tính*, vốn chỉ sử dụng hồi quy tuyến tính đơn với độ phức tạp xấp xỉ $O(n)$ cho mỗi thuộc tính. Tuy nhiên, nếu áp dụng *phương pháp gỡ lỗi dữ liệu đào tạo bằng hồi quy tuyến tính* một cách tuần tự cho nhiều thuộc tính nhạy cảm (k thuộc tính), chi phí tính toán sẽ tăng lên đáng kể. Cụ thể, với mỗi thuộc tính nhạy cảm, LTDD cần thực hiện hồi quy tuyến tính đơn cho từng thuộc tính không nhạy cảm, với độ phức tạp xấp xỉ $O(n)$ cho mỗi lần hồi quy. Do có tổng cộng $d - k$ thuộc tính không nhạy cảm, chi phí cho một thuộc tính nhạy cảm là $O(n \cdot (d - k))$. Khi áp dụng lần lượt cho k thuộc tính nhạy cảm, tổng độ phức tạp sẽ là: $O(k \cdot n \cdot (d - k))$. Ngoài ra, do mỗi lần áp dụng LTDD đều điều chỉnh lại dữ liệu đầu vào, các bước xử lý tiếp theo phải thực hiện trên dữ liệu đã biến đổi, dẫn đến khả năng tích lũy sai lệch và ảnh hưởng đến tính ổn định của dữ liệu. Vì vậy, mặc dù mỗi lần áp dụng LTDD có chi phí thấp, việc lặp lại cho nhiều thuộc tính nhạy cảm không chỉ làm gia tăng độ phức tạp tính toán mà còn tiềm ẩn rủi ro về độ tin cậy của kết quả. Ngược lại, các kết quả thực nghiệm cho thấy Fairedu đạt hiệu quả vượt trội trong việc cải thiện công bằng đồng thời vẫn duy trì hiệu suất dự đoán. Điều này khẳng định Fairedu là một giải pháp hiệu quả và phù hợp cho bài toán đảm bảo công bằng với nhiều thuộc tính nhạy cảm.

2.2.4. Ví dụ minh họa

Áp dụng phương pháp Fairedu cho hệ thống dự đoán điểm tốt nghiệp dựa trên tập dữ liệu điểm của Khoa Công nghệ thông tin, Trường đại học Đại Nam bao gồm thông tin về kết quả học tập năm nhất của sinh viên, cùng với ba thuộc tính nhạy cảm là *giới tính* (nam/nữ) (x_1), *khu vực* (thành thị/nông thôn) (x_2) và *tuổi* (đúng tuổi/quá tuổi) (x_3). Mục tiêu là xây dựng một hệ thống dự đoán điểm trung bình tích lũy (GPA) khi tốt nghiệp cho sinh viên ngành CNTT, đồng thời đồng thời giảm thiểu và kiểm soát ảnh hưởng mang tính thiên lệch của các thuộc tính nhạy cảm đến kết đến dự đoán. Các bước áp dụng phương pháp Fairedu cụ thể như sau:

Bước 1. Chuẩn bị dữ liệu: Tập dữ liệu đầu vào bao gồm cả thuộc tính nhạy cảm và không nhạy cảm. Dữ liệu được chia thành hai phần: 85% dành cho huấn luyện và 15% dành cho kiểm tra.

Bước 2. *Loại bỏ phụ thuộc trên tập huấn luyện:* Trong bước này tiến hành hai bước nhỏ:

Bước 2.1. Xác định hệ số hồi quy cho các thuộc tính không nhạy cảm: Giả sử điểm Toán năm nhất (*Math*) là một thuộc tính không nhạy cảm. Ta thực hiện hồi quy tuyến tính đa biến để dự đoán *Math* dựa trên ba thuộc tính nhạy cảm theo Công thức 2.5:

$$Math = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + \epsilon \quad (2.5)$$

Trong đó:

- a : hệ số chặn,
- b_1, b_2, b_3 : các hệ số hồi quy ứng với các thuộc tính nhạy cảm x_1, x_2, x_3 , và
- ϵ : sai số

Bước 2.2. Loại bỏ ảnh hưởng từ các thuộc tính nhạy cảm: Loại bỏ sự ảnh hưởng vào các thuộc tính nhạy cảm trong biến điểm toán bằng cách thay thế điểm toán mới bằng Công thức 2.6. Khi đó $Math_{new}$ là điểm Toán đã được điều chỉnh, không còn chịu ảnh hưởng từ *giới tính, khu vực hoặc tuổi*.

$$Math_{new} = Math - (a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3) \quad (2.6)$$

Thực hiện điều chỉnh tương tự cho tất cả các thuộc tính không nhạy cảm khác trước khi đưa vào mô hình học máy.

Bước 3. *Huấn luyện mô hình trên dữ liệu đã điều chỉnh:* Sau khi loại bỏ ảnh hưởng của các thuộc tính nhạy cảm, sử dụng dữ liệu đã điều chỉnh để huấn luyện mô hình dự đoán (S_{ML}) cho GPA khi tốt nghiệp, đảm bảo đầu ra không bị thiên vị.

Bước 4. *Áp dụng điều chỉnh tương tự cho dữ liệu kiểm tra:* Đối với các mẫu trong tập kiểm tra, cũng thực hiện điều chỉnh tương tự để loại bỏ ảnh hưởng từ các thuộc tính nhạy cảm, trước khi đưa vào mô hình đã huấn luyện (S_{ML}).

Bước 5. *Đánh giá mô hình:* Đánh giá về cả hiệu suất và tính công bằng của mô hình trên tập kiểm tra đã được xử lý trong Bước 4.

Bằng cách áp dụng phương pháp Fairedu, kết quả đạt được là mô hình dự đoán điểm tốt nghiệp tại khoa CNTT trường đại học Đại Nam sẽ giảm được thiên vị với các thuộc tính nhạy cảm *giới tính*, *khu vực*, và *tuổi*. Fairedu đã góp phần thúc đẩy một phương pháp tiếp cận công bằng hơn trong bối cảnh giáo dục.

2.3. Thực nghiệm

Phần này trình bày chi tiết quá trình chuẩn bị dữ liệu và thiết lập thực nghiệm nhằm kiểm chứng hiệu quả của phương pháp FairEdu trong việc tăng cường tính công bằng cho các hệ thống học máy ứng dụng trong lĩnh vực giáo dục.

Các nội dung này lần lượt được trình bày chi tiết trong các mục tiếp theo, bao gồm: mô tả và phân tích các tập dữ liệu sử dụng (Mục 2.3.1), lựa chọn và cấu hình mô hình học máy (Mục 2.3.2), quy trình chuẩn bị dữ liệu và thiết lập thực nghiệm (Mục 2.3.3), và các chỉ số được sử dụng để đánh giá mô hình theo cả khía cạnh công bằng và độ chính xác (Mục 2.3.4).

2.3.1. Dữ liệu

Dữ liệu sử dụng cho thực nghiệm này bao gồm ba bộ dữ liệu phổ biến từ Kaggle¹ và một bộ dữ liệu được thu thập từ khoa Công nghệ Thông tin, Trường Đại học Đại Nam (DNU), Hà Nội, Việt Nam². Để đảm bảo quá trình huấn luyện ổn định trên tất cả các bộ dữ liệu, các thuộc tính dạng số được chuẩn hóa bằng phương pháp chuẩn hoá min-max. Các thuộc tính nhạy cảm có trong bộ dữ liệu như *giới tính*, *tuổi*, *tình trạng nợ*, *tình trạng khuyết tật*, *sức khỏe* và *khu vực* được mã hóa nhị phân, chi tiết có trong Bảng 2.1. Đặc điểm cụ thể và các bước tiền xử lý của từng bộ dữ liệu được trình bày dưới đây:

1. *Bộ dữ liệu dự đoán bỏ học và thành công học tập (Student Predict Dropout – SD)*. Gồm 4.425 mẫu và 34 thuộc tính, phản ánh thông tin về nhân khẩu

¹Kaggle.com

²<https://dainam.edu.vn/en>

Bảng 2.1: Bảng mã hóa các thuộc tính nhạy cảm

STT	Thuộc tính nhạy cảm	Giá trị	Mã hóa	Bộ dữ liệu
1	<i>Giới tính</i>	Nam	1	<i>SD; SP; Olad; DNU</i>
		Nữ	0	
2	<i>Sức khỏe</i>	Tốt	1	<i>SP</i>
		Khác	0	
3	<i>Tình trạng nợ</i>	Không nợ	1	<i>SD</i>
		Có nợ	0	
4	<i>Tình trạng khuyết tật</i>	Không	1	<i>Oulad</i>
		Có	0	
5	<i>Tuổi</i>	Đúng tuổi	1	<i>DNU</i>
		Quá tuổi	0	
6	<i>Khu vực</i>	Thành thị	1	<i>DNU</i>
		Nông thôn	0	

học, hoàn cảnh xã hội – kinh tế, và kết quả học tập của sinh viên đại học. Hai thuộc tính nhạy cảm được xét đến trong bộ dữ liệu gồm *giới tính* và *tình trạng nợ*. Biến mục tiêu là khả năng tốt nghiệp được nhị phân hóa, cụ thể 1 nếu sinh viên tốt nghiệp, 0 nếu sinh viên không tốt nghiệp. Bộ dữ liệu Student Dropout được giới thiệu bởi Valentim Realinho và cộng sự [126]. Trong nghiên cứu này, dữ liệu được tải từ Kaggle [127].

2. *Bộ dữ liệu dự đoán hiệu suất học tập của sinh viên (Student Performance – SP)*. Gồm 395 mẫu và 33 thuộc tính, thu thập tại hai trường trung học ở Bồ Đào Nha. Bộ dữ liệu bao gồm điểm số, đặc điểm xã hội – nhân khẩu học và thông tin hỗ trợ học tập. Hai thuộc tính nhạy cảm là *giới tính* và *sức khỏe*. Nhân kết quả được nhị phân hóa: nếu điểm số lớn hơn hoặc bằng điểm trung bình thì gán là 1, ngược lại là 0. Bộ dữ liệu Student Performance được giới thiệu bởi Paulo Cortez và cộng sự [128], phiên bản bộ dữ liệu được sử dụng được truy cập thông qua nền tảng Kaggle [129]
3. *Bộ dữ liệu thông tin khóa học của đại học Mở Vương Quốc Anh (Oulad)*. Từ bộ dữ liệu OULAD, nghiên cứu sử dụng bảng studentInfo, gồm 32.593 mẫu và 12 thuộc tính, chứa thông tin về sinh viên, khóa học và kết quả học tập. Hai thuộc tính nhạy cảm là *giới tính* và *tình trạng khuyết tật*. Biến mục tiêu *kết quả học tập* được mã hóa: “Fail” hoặc “Withdrawn” là 0, các trường hợp còn lại là 1. Bộ dữ liệu Oulad được giới thiệu bởi Kuzilek và cộng sự [130], phiên bản bộ dữ liệu được sử dụng được truy cập thông qua

nền tảng Kaggle [131]

4. *Bộ dữ liệu DNU*. Bộ dữ liệu được xây dựng trong khuôn khổ đề tài nghiên cứu khoa học cấp Trường tại trường Đại học Đại Nam năm 2022 với tiêu đề “Dự đoán kết quả học tập dựa trên kỹ thuật học máy”, trong đó tác giả luận án là một thành viên tham gia thực hiện đề tài. Đề tài đã được nghiệm thu theo quy định của Trường Đại học Đại Nam và bộ dữ liệu được phép sử dụng cho mục đích nghiên cứu. Dữ liệu được thu thập từ 13 khóa sinh viên đã ra trường, với ba chương trình đào tạo khác nhau thuộc Khoa CNTT, Trường đại học Đại Nam. Dữ liệu được trích xuất từ hệ thống quản lý đào tạo của nhà trường và đã được ẩn danh trước khi sử dụng cho nghiên cứu nhằm bảo đảm quyền riêng tư của người học. Việc thu thập dữ liệu được thực hiện trên nhiều khóa đào tạo và nhiều chương trình đào tạo khác nhau, góp phần bảo đảm tính đa dạng của các nhóm người học trong bộ dữ liệu. Dữ liệu bao các thuộc tính về thông tin cá nhân và điểm học tập của từng sinh viên trong toàn bộ chương trình đào tạo. Dữ liệu bao gồm 426 mẫu với 59 thuộc tính ban đầu. Vì dữ liệu được lấy từ 13 khóa với ba chương trình đào tạo khác nhau nên các học phần có sự khác nhau về số lượng và nội dung. Để thống nhất, các học phần đã được đánh giá là tương đương nếu có sự trùng lặp trên 70% nội dung. Chính vì vậy, sau khi chuẩn hóa, 59 thuộc tính ban đầu rút gọn còn lại 42 thuộc tính, bao gồm: sáu thuộc tính định danh sinh viên, 33 thuộc tính điểm các học phần, và ba thuộc tính tổng hợp (điểm trung bình, xếp hạng, nhãn dự đoán rủi ro). Các thuộc tính điểm số đều được quy về thang điểm 10. Ba thuộc tính nhạy cảm là *giới tính*, *khu vực* và *tuổi*. Bộ dữ liệu này được sử dụng để dự đoán kết quả học tập dựa vào điểm học tập của hai năm đầu. Ngoài ra, dữ liệu được chuẩn hóa nhằm bảo vệ quyền riêng tư; tất cả các cột nhạy cảm đều được mã hóa thành giá trị số. Các thuộc tính *giới tính*, *khu vực* và *tuổi* đều được nhị phân hóa. Các thuộc tính nhạy cảm được giữ lại không nhằm phục vụ quá trình dự đoán mà để đánh giá và phân tích mức độ công bằng của các mô hình học máy cũng như hiệu quả của các phương pháp cải thiện công bằng được đề xuất trong luận án. Biến mục tiêu là *Khả năng rủi ro* được mã hóa nhị phân là 1 nếu nhận giá trị *Không* và 0 nếu nhận giá trị *Có*.

2.3.2. Lựa chọn mô hình học máy

Việc lựa chọn các mô hình học máy trong thực nghiệm được dựa trên kết quả tổng quan nghiên cứu trình bày tại Mục ???. Cụ thể, ba mô hình phổ biến trong các ứng dụng giáo dục được sử dụng gồm: *Hồi quy logistic*, *Cây quyết định* và *Rừng ngẫu nhiên* [132, 133, 134]. Sự lựa chọn này nhằm đánh giá và so sánh mức độ công bằng cũng như hiệu quả dự đoán giữa các mô hình có độ phức tạp khác nhau trong cùng một bối cảnh.

Hồi quy logistic (LR): một mô hình thống kê thường được sử dụng cho các bài toán phân loại nhị phân, trong đó mục tiêu là dự đoán một trong hai khả năng xảy ra. Đây là một dạng phân tích hồi quy trong đó biến phụ thuộc là biến phân loại [21]. Trong nghiên cứu này, mô hình sử dụng kỹ thuật chuẩn hoá L2 với bộ giải *liblinear* – một thuật toán tối ưu hiệu quả dành cho các mô hình tuyến tính. Việc huấn luyện mô hình được giới hạn ở tối đa 100 vòng lặp để đảm bảo tính ổn định và hiệu suất tính toán.

Cây quyết định (DT): một mô hình học máy phổ biến, được sử dụng trong cả bài toán phân loại và hồi quy. Mô hình hoạt động bằng cách chia nhỏ tập dữ liệu thành các tập con dựa trên các quy tắc ra quyết định, và xây dựng một cây quyết định. Mỗi nút trong cây đại diện cho một thuộc tính, mỗi nhánh tương ứng với giá trị của thuộc tính đó, và các lá là nhãn hoặc giá trị dự đoán [135]. Cấu hình cụ thể trong nghiên cứu này là cây có độ sâu tối đa là 3 và sử dụng chỉ số độ tinh khiết Gini.

Rừng ngẫu nhiên (RF): một mô hình học máy có cấu trúc dựa trên khái niệm *Cây quyết định*, nhưng thay vì chỉ sử dụng một cây, mô hình sử dụng một tập hợp (rừng) gồm nhiều *Cây quyết định*. Mỗi cây trong rừng được xây dựng từ một tập con ngẫu nhiên của dữ liệu huấn luyện, và các thuộc tính được chọn ngẫu nhiên tại mỗi nút trong quá trình xây dựng [136]. Trong nghiên cứu này, mô hình sử dụng 100 cây (estimator), mỗi cây có độ sâu tối đa là 3, và cũng sử dụng chỉ số Gini để đánh giá độ tinh khiết.

2.3.3. Thiết lập thực nghiệm

Tập dữ liệu sau khi tiền xử lý được chia thành hai phần: 85% dùng để huấn luyện và 15% để kiểm tra. Để đánh giá hiệu quả của Fairedu, các thực nghiệm được thực hiện trên ba cấu hình gồm: (i) dữ liệu gốc không áp dụng phương pháp đảm bảo công bằng (Origin), (ii) phương pháp *phương pháp gỡ lỗi dữ liệu đào tạo bằng hồi quy tuyến tính*, và (iii) phương pháp đề xuất Fairedu. Ngoài ra, để so sánh với các phương pháp đảm bảo công bằng phổ biến khác, luận án sử dụng các kết quả đã được công bố trên cùng bộ dữ liệu đối với Reweighting và FairSMOTE. Việc sử dụng cùng bộ dữ liệu, cùng mô hình học máy và cùng các chỉ số đánh giá giúp bảo đảm tính khách quan và khả năng so sánh trực tiếp giữa các phương pháp.

Với Fairedu, quy trình thực nghiệm được tiến hành lần lượt theo các bước sau: (1) Trên tập huấn luyện, áp dụng phương pháp Fairedu để loại bỏ mối phụ thuộc giữa các thuộc tính không nhạy cảm và các thuộc tính nhạy cảm, đảm bảo dữ liệu đầu vào cho mô hình không mang thiên vị tiềm ẩn. (2) Mô hình học máy được huấn luyện trên dữ liệu đã được điều chỉnh bởi Fairedu. (3) Áp dụng cùng phép điều chỉnh (sử dụng các hệ số hồi quy thu được từ tập huấn luyện) lên tập kiểm tra để đảm bảo tính nhất quán trước khi đánh giá hiệu suất và công bằng của mô hình. Tất cả các thí nghiệm được lặp lại 100 lần và các kết quả được báo cáo dưới dạng giá trị trung bình nhằm giảm ảnh hưởng của tính ngẫu nhiên và tăng độ tin cậy của kết quả thực nghiệm. Việc đánh giá hiệu quả của các phương pháp được thực hiện dựa trên xu hướng tổng thể thông qua số lượt thắng/hòa/thua trên nhiều bộ dữ liệu, mô hình học máy, thuộc tính nhạy cảm và chỉ số công bằng.

2.3.4. Chỉ số đánh giá

Như đã trình bày trong Mục 1.2.4, đánh giá công bằng trong các mô hình học máy là một nhiệm vụ phức tạp và đa chiều, bởi thiên vị có thể phát sinh dưới nhiều hình thức và ở nhiều giai đoạn khác nhau trong quá trình huấn luyện và dự đoán. Do đó, không tồn tại một chỉ số duy nhất nào có thể phản ánh đầy đủ toàn bộ khía cạnh của công bằng. Để đảm bảo đánh giá toàn diện hơn, nghiên

cứu này sử dụng kết hợp bốn chỉ số công bằng được công nhận rộng rãi, bao gồm “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*”, và “*chênh lệch cơ hội công bằng*”. Việc kết hợp các chỉ số này không chỉ giúp đánh giá công bằng ở cấp độ nhóm mà còn góp phần phát hiện các tình huống mô hình có thể đạt được công bằng theo một chỉ số nhưng lại vi phạm nghiêm trọng theo chỉ số khác. Đây là cơ sở để đưa ra đánh giá toàn diện về hiệu quả của phương pháp FairEdu trong bối cảnh dữ liệu giáo dục có tính đa dạng và chồng chéo giữa các thuộc tính nhạy cảm.

Bên cạnh việc đánh giá công bằng, để kiểm tra mối quan hệ đánh đổi giữa công bằng và hiệu suất dự đoán, nghiên cứu cũng sử dụng bốn chỉ số phổ biến như đã trình bày trong Mục 1.1.2 để đánh giá hiệu suất của mô hình, bao gồm “*độ chuẩn xác*” và “*độ hồi tưởng*”. Sử dụng kết hợp nhiều chỉ số cho phép phản ánh hiệu suất mô hình một cách toàn diện hơn, đặc biệt trong các tình huống mất cân bằng lớp. Việc đánh giá đồng thời cả công bằng và hiệu suất cho phép xác định mức độ phù hợp của phương pháp FairEdu trong việc đạt được sự cân bằng giữa hai mục tiêu thường mâu thuẫn trong thiết kế mô hình học máy, đặc biệt là trong các ứng dụng giáo dục nhạy cảm với bất bình đẳng xã hội.

2.4. Kết quả thực nghiệm

Phần này trình bày và phân tích chi tiết các kết quả thực nghiệm nhằm đánh giá hiệu quả của phương pháp FairEdu trong việc cải thiện tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục. Các kết quả được tổ chức theo từng khía cạnh phân tích cụ thể, phản ánh các vấn đề cốt lõi đã được xác định trong quá trình thiết kế thực nghiệm.

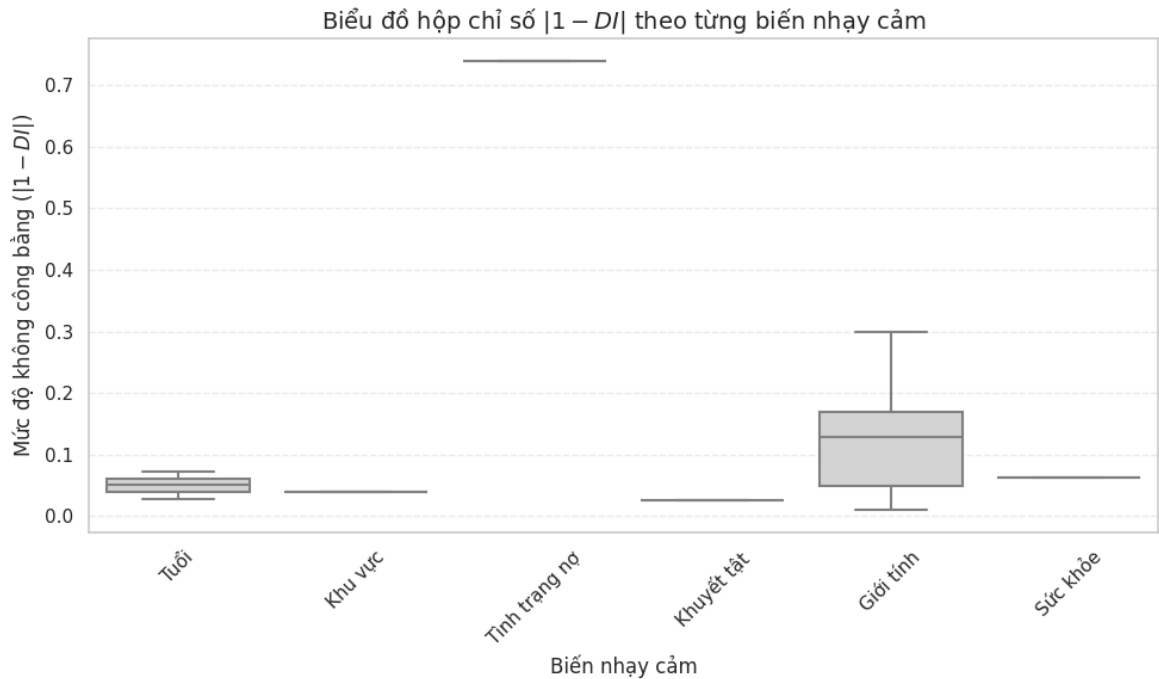
Cụ thể, Mục 2.4.1 tập trung phân tích sự tồn tại và mức độ của thiên vị hệ thống liên quan đến các thuộc tính nhạy cảm trong các tập dữ liệu giáo dục, qua đó làm rõ bối cảnh và động cơ của bài toán đảm bảo công bằng. Mục 2.4.2 đánh giá sự khác biệt về mức độ công bằng khi áp dụng các mô hình học máy khác nhau, nhằm phân tích ảnh hưởng của thuật toán học máy đến kết quả công bằng. Tiếp theo, Mục 2.4.3 phân tích khả năng xử lý đồng thời nhiều thuộc tính nhạy cảm của phương pháp FairEdu và so sánh với các phương pháp hiện có, từ

đó làm rõ hiệu quả của cách tiếp cận tiên xử lý dựa trên loại bỏ sự phụ thuộc. Cuối cùng, Mục 2.4.4 xem xét mối quan hệ giữa cải thiện công bằng và hiệu suất dự đoán của mô hình khi áp dụng FairEdu, nhằm đánh giá tính khả thi của phương pháp trong bối cảnh ứng dụng thực tiễn.

2.4.1. Thiên vị hệ thống trong dữ liệu giáo dục

Để xác định sự tồn tại của thiên vị hệ thống, nghiên cứu đo lường mức độ công bằng theo cả bốn chỉ số $|1 - DI|$, *SPD*, *AOD*, và *EOD* đối với sáu thuộc tính nhạy cảm có trên bốn bộ dữ liệu, bao gồm: *giới tính*, *tuổi*, *tình trạng khuyết tật*, *sức khỏe*, *tình trạng nợ*, và *khu vực*. Kết quả theo chỉ số $|1 - DI|$ được tổng hợp trong Hình 2.2. Trong biểu đồ này, mức độ không công bằng tương ứng với từng thuộc tính nhạy cảm (đo bằng chỉ số $|1 - DI|$) được minh họa chi tiết, trong đó giá trị càng gần 0 cho thấy mô hình càng công bằng.

Biểu đồ hộp chỉ số $|1 - DI|$ theo từng thuộc tính nhạy cảm cho thấy sự khác biệt rõ rệt về mức độ không công bằng giữa các thuộc tính. Biến *giới tính* là thuộc tính có số lượng quan sát nhiều nhất và thể hiện mức độ dao động lớn nhất, với giá trị $|1 - DI|$ dao động từ rất nhỏ (0.0097) đến tương đối cao (0.2981). Điều này cho thấy mức độ ảnh hưởng của *giới tính* đến kết quả mô hình là không đồng đều giữa các trường hợp, phản ánh sự khó kiểm soát thiên vị liên quan đến giới tính và nhấn mạnh sự cần thiết của việc xử lý công bằng cho thuộc tính này. Biến *tình trạng nợ* có giá trị $|1 - DI|$ lên tới 0.7404 — cao nhất trong toàn bộ bảng. Mặc dù chỉ xuất hiện một lần, kết quả này cho thấy đây có thể là một thuộc tính đặc biệt nhạy cảm và cần được chú ý trong các phân tích công bằng. Biến *tuổi* thể hiện mức độ không công bằng thấp hơn, với các giá trị lần lượt 0.0279 và 0.0722. Mức biến thiên nhỏ cho thấy mô hình có xu hướng ổn định hơn trong việc xử lý các biến này. Các thuộc tính khác như *tình trạng khuyết tật*, *khu vực* và *sức khỏe* đều có giá trị $|1 - DI|$ tương đối thấp (dưới 0.07), cho thấy mức độ công bằng cao hơn hoặc ít thiên vị hơn trong các tình huống được quan sát.



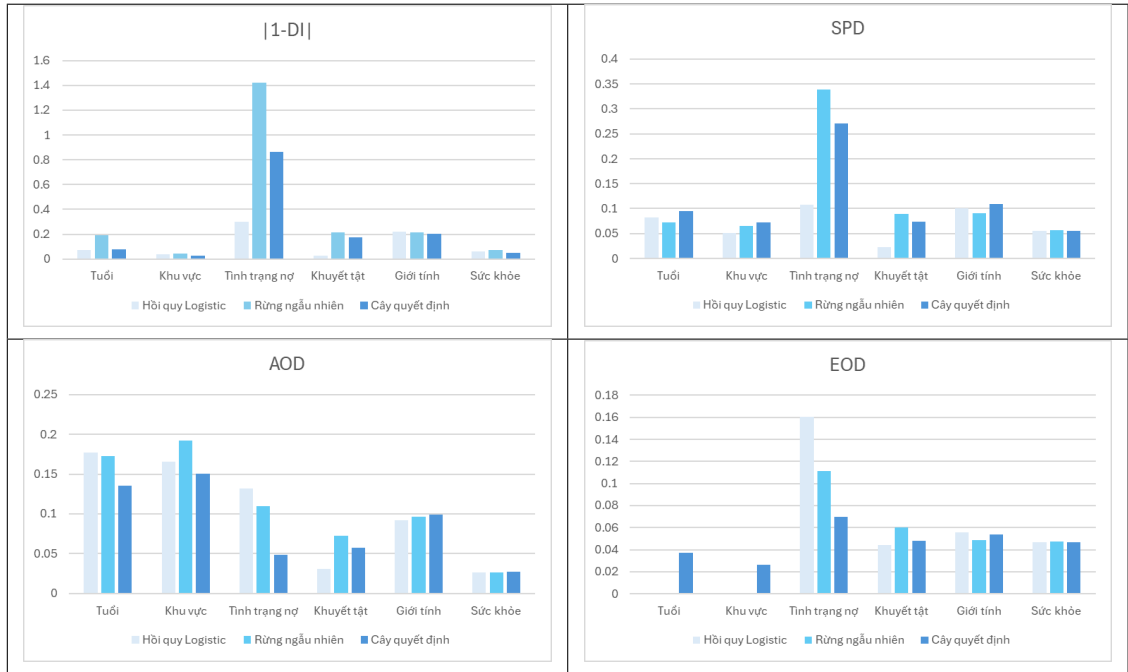
Hình 2.2: So sánh chỉ số $|1 - DI|$ giữa các thuộc tính nhạy cảm trên các tập dữ liệu

Kết quả phân tích cho thấy không thuộc tính nhạy cảm nào thể hiện mức độ thiên vị một cách nhất quán và nổi trội hơn so với các thuộc tính khác. Quan sát tương tự cũng được ghi nhận khi sử dụng các chỉ số công bằng khác, cho thấy xu hướng này không phụ thuộc vào cách đo lường. Do đó, việc xem xét toàn diện tất cả các thuộc tính nhạy cảm là cần thiết để đảm bảo đánh giá công bằng được thực hiện một cách đầy đủ và khách quan.

2.4.2. Ảnh hưởng của mô hình học máy đến mức độ công bằng

Để đánh giá mức độ ảnh hưởng của thuật toán học máy đến công bằng, thực nghiệm tiến hành so sánh bốn chỉ số công bằng phổ biến $|1 - DI|$, SPD , AOD , và EOD , tương ứng với sáu thuộc tính nhạy cảm, trên ba mô hình truyền thống gồm *Hồi quy logistic*, *Rừng ngẫu nhiên*, và *Cây quyết định*. Kết quả tổng hợp cho thấy sự khác biệt rõ rệt về mức độ công bằng giữa các thuật toán, ngay cả khi được áp dụng trên cùng một bộ dữ liệu và cùng một thuộc tính nhạy cảm. Kết quả được tổng hợp trong Bảng 2.2 với bốn biểu đồ trực quan tương ứng từng chỉ số.

Bảng 2.2: So sánh các chỉ số công bằng đối với các thuộc tính nhạy cảm trong các thuật toán học máy khác nhau



Trước hết, chỉ số $|1 - DI|$ phản ánh rõ sự khác biệt về mức độ không công bằng giữa các mô hình học máy. Trên thuộc tính *tình trạng nơ*, mô hình *Rừng ngẫu nhiên* ghi nhận giá trị cao nhất (1.4223), tiếp theo là *Cây quyết định* (0.8634) và *Hồi quy logistic* (0.2291), cho thấy sự thiên lệch rất lớn trong phân phối đầu ra đối với biến này, đặc biệt khi sử dụng các mô hình phi tuyến. Tương tự, với thuộc tính *giới tính*, cả ba mô hình đều có mức độ thiên lệch đáng kể, trong đó *Hồi quy logistic* (0.2225) và *Rừng ngẫu nhiên* (0.2172) cao hơn so với *Cây quyết định* (0.2018). Đối với thuộc tính *tuổi*, *Rừng ngẫu nhiên* thể hiện mức độ không công bằng cao nhất (0.1917), trong khi *Hồi quy logistic* (0.0722) và *Cây quyết định* (0.0788) có mức thấp hơn. Trên biến *tình trạng khuyết tật*, sự khác biệt giữa các mô hình trở nên rõ rệt hơn khi *Rừng ngẫu nhiên* (0.2173) và *Cây quyết định* (0.1738) có giá trị cao hơn đáng kể so với *Hồi quy logistic* (0.0251), cho thấy mô hình tuyến tính trong trường hợp này ít khuếch đại thiên lệch hơn. Ngược lại, với các thuộc tính như *nơi sinh* và *sức khỏe*, tất cả các mô hình đều có giá trị $|1 - DI|$ tương đối thấp, trong đó *Cây quyết định* thường đạt mức thấp nhất (lần lượt là 0.0247 và 0.0507), cho thấy khả năng duy trì công bằng tốt hơn trong các trường hợp này.

Nhìn chung, kết quả cho thấy mức độ công bằng phụ thuộc đáng kể vào cả mô hình học máy và từng thuộc tính nhạy cảm cụ thể. Đặc biệt, các mô hình phi tuyến như *Rừng ngẫu nhiên* và *Cây quyết định* có xu hướng tạo ra mức độ thiên lệch cao hơn trong một số trường hợp (như *tình trạng nơ*), trong khi *Hồi quy logistic* lại thể hiện tốt hơn ở một số thuộc tính khác (như *tình trạng khuyết tật*). Điều này cho thấy không tồn tại một mô hình đơn lẻ nào luôn đảm bảo công bằng tối ưu, nhấn mạnh sự cần thiết của các phương pháp can thiệp như FairEdu để kiểm soát thiên lệch một cách hệ thống.

2.4.3. Khả năng xử lý đồng thời nhiều thuộc tính nhạy cảm của FairEdu

Để đánh giá khả năng cải thiện công bằng của phương pháp FairEdu, nghiên cứu đã tiến hành so sánh với một số phương pháp hiện đại khác đã được công bố và sử dụng phổ biến trong nhiều nghiên cứu gần đây, bao gồm: *Cân lại dữ liệu (Reweighing)* [30], *FairSmote* [20], và *phương pháp gỡ lỗi dữ liệu đào tạo bằng hồi quy tuyến tính (LTDD)* [31]. Việc so sánh được thực hiện trên nhiều mô hình học máy tiêu biểu như *Hồi quy logistic*, *Rừng ngẫu nhiên* và *Cây quyết định*, nhằm đảm bảo tính khái quát và độ tin cậy của kết quả. Các thí nghiệm được tiến hành trên ba bộ dữ liệu thường được sử dụng trong các nghiên cứu về đảm bảo tính công bằng trong lĩnh vực giáo dục: *Oulad*, *Student Performance*, *Student Predict Dropout*, và một bộ dữ liệu riêng *DNU Data*. Các bộ dữ liệu này đều chứa ít nhất một thuộc tính nhạy cảm như *giới tính* hoặc *tình trạng khuyết tật*, là điều kiện cần thiết để đánh giá hiệu quả cải thiện công bằng.

Kết quả trong Bảng 2.3 cũng cho thấy mức độ cải thiện của FairEdu không hoàn toàn đồng đều trên tất cả các bộ dữ liệu và thuộc tính nhạy cảm. Đối với chỉ số $|1 - DI|$, FairEdu thắng hoàn toàn so với Origin (7/7), Reweighing (6/7) và FairSMOTE (5/5), nhưng chỉ đạt 4/7 trường hợp khi so sánh với LTDD. Tương tự, đối với chỉ số SPD, FairEdu tiếp tục vượt trội so với Origin (7/7), trong khi số lượt thắng so với LTDD, Reweighing và FairSMOTE lần lượt là 5/7, 4/7 và 3/5. Kết quả này cho thấy mặc dù FairEdu nhìn chung cải thiện công bằng hiệu quả hơn các phương pháp đối sánh, mức độ cải thiện vẫn phụ thuộc vào đặc điểm của từng bộ dữ liệu và từng thuộc tính nhạy cảm.

Bảng 2.3: Bảng so sánh giữa Fairedu và các phương pháp hiện có

Kỹ thuật	O_ GT	O_ K.tật	SP_ GT	SP_ SK	SD_ Nợ	DNU GT	DNU Tuổi	Thắng/ hòa/thua
	1-DI							
Origin	0.3065	0.7264	0.1461	0.0666	2.4814	0.0705	0.1242	7/0/0
LTDD	0.0390	0.0146	0.1291	0.0606	0.2612	0.0705	0.1242	4/0/3
Reweighing	0.2531	0.7061	0.1422	0.0983	1.9805	0.0727	0.0233	6/0/1
FairSmote	0.0273	0.6781	0.1422	0.0983	0.9843	-	-	5/0/0
Fairedu	0.0097	0.0251	0.1283	0.0616	0.2981	0.0136	0.0778	
Kỹ thuật	SPD							
Origin	0.1036	0.3031	0.0792	0.0564	0.4106	0.0632	0.1194	7/0/0
LTDD	0.0178	0.0217	0.0763	0.0551	0.1076	0.0632	0.1194	5/1/1
Reweighing	0.0833	0.2793	0.0682	0.0486	0.4570	0.0678	0.0221	4/0/3
FairSmote	0.0135	0.3651	0.0682	0.0486	0.3180	-	-	3/0/2
Fairedu	0.0125	0.0224	0.0756	0.0551	0.1074	0.0115	0.0688	

Ghi chú: Trong bảng này, các ô được tô màu xám biểu thị những trường hợp cho kết quả kém hơn so với FairEdu. Cột cuối cùng tổng hợp số lượt thắng/hòa/thua của từng phương pháp khi so sánh trực tiếp với FairEdu.

Nguyên nhân của sự khác biệt này là do Fairedu loại bỏ đồng thời ảnh hưởng của nhiều thuộc tính nhạy cảm thông qua mô hình hồi quy đa biến, nhờ đó xử lý được các trường hợp đồng thời tồn tại sự phụ thuộc giữa các thuộc tính nhạy cảm mà các phương pháp như LTDD chưa giải quyết được. Tuy nhiên, ở những bộ dữ liệu mà mức độ phụ thuộc giữa các thuộc tính nhạy cảm và các thuộc tính đầu vào không lớn, lợi thế của Fairedu so với các phương pháp đối sánh giảm xuống. Điều này giải thích vì sao vẫn tồn tại một số cấu hình mà Fairedu chưa đạt kết quả tốt nhất.

Nhìn chung, các kết quả chi tiết về hai chỉ số công bằng $|1 - DI|$ và SPD của các phương pháp được trình bày trong Bảng 2.3 cho thấy Fairedu không chỉ cải thiện đáng kể các chỉ số công bằng trên phần lớn các bộ dữ liệu mà còn duy trì được tính ổn định trên nhiều thuộc tính nhạy cảm khác nhau. Điều này chứng minh việc loại bỏ đồng thời sự phụ thuộc vào nhiều thuộc tính nhạy cảm là một hướng tiếp cận hiệu quả hơn so với các phương pháp chỉ xử lý từng thuộc tính riêng lẻ.

Kết quả trong Bảng 2.4 cho thấy, đối với cả hai chỉ số $|1 - DI|$ và SPD, FairEdu đạt kết quả tốt hơn đáng kể so với Origin, với tỷ lệ Thắng/Hòa/Thua

lần lượt là 8/0/1 và 9/0/0. Khi so sánh với LTDD, FairEdu cũng đạt nhiều lượt thắng hơn, với tỷ lệ Thắng/Hòa/Thua lần lượt là 5/0/4 và 6/2/1 trên tổng số chín phép so sánh. Mặc dù vẫn còn một số trường hợp LTDD đạt kết quả công bằng tốt hơn FairEdu đối với từng bộ dữ liệu hoặc từng thuộc tính nhạy cảm, kết quả tổng thể cho thấy FairEdu chiếm ưu thế trong phần lớn các cấu hình thực nghiệm.

Tương tự, đối với các mô hình Rừng ngẫu nhiên và Cây quyết định, FairEdu tiếp tục đạt nhiều lượt thắng hơn so với các phương pháp đối sánh. Kết quả chi tiết được trình bày trong Bảng 2.5 và Bảng 2.6 cho thấy FairEdu đạt từ 5 đến 7 lượt thắng trên tổng số 8 phép so sánh với LTDD và Origin. Tuy nhiên, vẫn tồn tại một số cấu hình mà LTDD cho kết quả công bằng tốt hơn, cho thấy hiệu quả của từng phương pháp còn phụ thuộc vào đặc điểm của bộ dữ liệu, mô hình học máy và thuộc tính nhạy cảm được xem xét.

Tổng hợp các kết quả trên cho thấy, xét trên toàn bộ các bộ dữ liệu, mô hình học máy, thuộc tính nhạy cảm và các chỉ số công bằng, FairEdu đạt số lượt thắng cao nhất trong các phương pháp được so sánh. Điều này khẳng định FairEdu là một phương pháp hiệu quả trong việc cải thiện công bằng cho các hệ thống học máy trong lĩnh vực giáo dục, đồng thời duy trì tính ổn định trên nhiều bộ dữ liệu và mô hình học máy khác nhau. Mặc dù vẫn còn một số cấu hình mà LTDD đạt kết quả tốt hơn, các trường hợp này không làm thay đổi xu hướng chung của toàn bộ thực nghiệm. Vì vậy, hiệu quả của FairEdu cần được đánh giá dựa trên kết quả tổng thể của toàn bộ các cấu hình thực nghiệm thay vì một số trường hợp riêng lẻ.

Tổng hợp các kết quả trên cho thấy *FairEdu* duy trì ưu thế rõ rệt so với các phương pháp so sánh trên nhiều mô hình, bộ dữ liệu và thuộc tính nhạy cảm, với mức cải thiện công bằng ổn định và có ý nghĩa thống kê. Điều này khẳng định tính hiệu quả và khả năng tổng quát của FairEdu trong việc nâng cao công bằng cho các hệ thống học máy trong lĩnh vực giáo dục.

Bảng 2.4: Bảng so sánh chỉ số công bằng giữa Fairedu với các phương pháp LTDD và Origin với mô hình *Hồi quy logistic*

Chỉ số công bằng	1-DI			SPD		
	Origin	LTDD	FairEdu	Origin	LTDD	FairEdu
<i>O_GT</i>	0.306	0.039	0.010	0.104	0.018	0.012
<i>O_K.tật</i>	0.726	0.015	0.025	0.303	0.022	0.022
<i>SP_GT</i>	0.146	0.129	0.128	0.079	0.076	0.076
<i>SP_SK</i>	0.067	0.061	0.062	0.056	0.055	0.055
<i>SD_Nợ</i>	2.481	0.261	0.298	0.411	0.108	0.107
<i>SD_GT</i>	0.816	0.141	0.740	0.287	0.070	0.240
<i>DNU_GT</i>	0.071	0.071	0.014	0.063	0.063	0.012
<i>DNU_Tuổi</i>	0.124	0.124	0.078	0.119	0.119	0.069
<i>DNU_KV</i>	0.049	0.040	0.015	0.053	0.071	0.051
Thắng/Hòa/Thua	8/0/1	5/0/4		9/0/0	6/2/1	

Ghi chú: Trong bảng này, các ô được tô màu xám và màu xanh lam lần lượt biểu thị những trường hợp cho kết quả Thua (kém hơn) và Hòa (tương đương) so với FairEdu. Dòng dưới cùng tổng hợp số lượt thắng/hòa/thua của FairEdu so với từng phương pháp tương ứng trong cột.

Bảng 2.5: Bảng so sánh chỉ số công bằng giữa Fairedu với các phương pháp LTDD và Origin với mô hình *Rừng ngẫu nhiên*

Chỉ số công bằng	1-DI			SPD		
	Origin	LTDD	FairEdu	Origin	LTDD	FairEdu
<i>O_GT</i>	0.004	0.103	0.086	0.015	0.038	0.034
<i>O_K.tật</i>	0.598	0.337	0.217	0.257	0.142	0.089
<i>SP_GT</i>	0.147	0.130	0.132	0.078	0.075	0.076
<i>SP_SK</i>	0.081	0.067	0.069	0.056	0.055	0.056
<i>SD_Nợ</i>	1.418	2.146	1.422	0.340	0.404	0.339
<i>SD_GT</i>	0.633	0.396	0.631	0.246	0.179	0.245
<i>DNU_GT</i>	0.044	0.057	0.019	0.065	0.075	0.006
<i>DNU_Tuổi</i>	0.075	0.056	0.192	0.083	0.087	0.072
<i>DNU_KV</i>	0.050	0.048	0.043	0.058	0.076	0.066
Thắng/Hòa/Thua	6/0/3	5/0/4		5/1/3	6/0/3	

Ghi chú: Trong bảng này, các ô được tô màu xám và màu xanh lam lần lượt biểu thị những trường hợp cho kết quả Thua (kém hơn) và Hòa (tương đương) so với FairEdu. Dòng dưới cùng tổng hợp số lượt thắng/hòa/thua của FairEdu so với từng phương pháp tương ứng trong cột.

Bảng 2.6: Bảng so sánh chỉ số công bằng giữa Fairedu với các phương pháp LTDD và Origin với mô hình *Cây quyết định*

Chỉ số công bằng	1-DI			SPD		
	Origin	LTDD	FairEdu	Origin	LTDD	FairEdu
<i>O_GT</i>	0.030556	0.112145	0.065006	0.0175	0.0453	0.0344
<i>O_K.tật</i>	0.226473	0.226417	0.173827	0.1015	0.1015	0.0743
<i>SP_GT</i>	0.132103	0.138954	0.124851	0.0765	0.0778	0.0767
<i>SP_SK</i>	0.052825	0.059959	0.050685	0.0564	0.0560	0.0554
<i>SD_Nợ</i>	1.360778	0.925377	0.863366	0.331947	0.279899	0.270915
<i>SD_GT</i>	0.61295	0.305683	0.581779	0.238742	0.148491	0.237721
<i>DNU_GT</i>	0.0361	0.0448	0.0354	0.0884	0.0771	0.0857
<i>DNU_Tuổi</i>	0.0621	0.0464	0.0788	0.0702	0.0854	0.0953
<i>DNU_KV</i>	0.0485	0.015	0.0247	0.0831	0.0705	0.072
Thắng/Hòa/Thua	7/0/2	6/0/3		6/0/3	5/0/4	

Ghi chú: Trong bảng này, các ô được tô màu xám và màu xanh lam lần lượt biểu thị những trường hợp cho kết quả Thua (kém hơn) và Hòa (tương đương) so với FairEdu. Dòng dưới cùng tổng hợp số lượt thắng/hòa/thua của FairEdu so với từng phương pháp tương ứng trong cột.

2.4.4. Môi quan hệ giữa công bằng và hiệu suất dự đoán

Để đánh giá tác động của phương pháp Fairedu đến hiệu suất mô hình, nghiên cứu tiến hành so sánh các chỉ số “độ chuẩn xác”, “độ chính xác”, “độ hồi tưởng” và “điểm số F1” trên ba mô hình học máy *Hồi quy logistic*, *Rừng ngẫu nhiên* và *Cây quyết định*, với bốn bộ dữ liệu và sáu thuộc tính nhạy cảm. Ba cấu hình được xem xét bao gồm: mô hình gốc (Origin), phương pháp LTDD và phương pháp Fairedu. Kết quả chi tiết được trình bày trong Phụ lục B, tại Bảng B.1 (ACC–Recall) và Bảng B.2 (F1-Score và Precision).

Kết quả thực nghiệm cho thấy Fairedu có khả năng duy trì hiệu suất dự đoán ổn định trong khi vẫn cải thiện đáng kể tính công bằng. Cụ thể, đối với “độ chuẩn xác”, Fairedu đạt kết quả cao hơn so với Origin và LTDD trong 11/27 trường hợp, với mức suy giảm tối đa chỉ 5.71%, cho thấy ảnh hưởng tiêu cực là không đáng kể. Với “độ hồi tưởng”, Fairedu vượt Origin trong 6 trường hợp và vượt LTDD trong 12 trường hợp, trong khi mức giảm lớn nhất chỉ khoảng 8%. Đối với “điểm số F1”, Fairedu đạt kết quả tốt hơn Origin trong 6 trường hợp và LTDD trong 9 trường hợp. Đáng chú ý, ở chỉ số “độ chính xác”, Fairedu thể hiện ưu thế rõ rệt khi vượt Origin trong 17 trường hợp và LTDD trong 19

trường hợp; tiêu biểu là trên mô hình Random Forest với bộ dữ liệu DNU-BP, precision tăng từ 0.929 lên 0.962 (tăng 3.5%).

Nhìn chung, các kết quả cho thấy việc áp dụng FairEdu có thể dẫn đến một mức đánh đổi hiệu suất nhất định, nhưng mức độ này là nhỏ và có thể kiểm soát. Trong bối cảnh các bài toán giáo dục và các lĩnh vực nhạy cảm, nơi yêu cầu về công bằng và đạo đức được đặt lên hàng đầu, FairEdu là một giải pháp hiệu quả giúp nâng cao công bằng mà vẫn đảm bảo hiệu suất mô hình ở mức chấp nhận được.

2.5. Thảo luận

Phần thảo luận này nhằm tổng hợp, diễn giải và đặt các kết quả thực nghiệm của Chương 2.4 trong bối cảnh nghiên cứu tổng thể của luận án. Thông qua việc phân tích các kết quả đạt được trên nhiều bộ dữ liệu giáo dục và nhiều mô hình học máy khác nhau, phần này làm rõ các vấn đề cốt lõi liên quan đến thiên vị dữ liệu, ảnh hưởng của thuật toán học máy đến công bằng, hiệu quả của phương pháp FairEdu trong xử lý đồng thời nhiều thuộc tính nhạy cảm, cũng như mối quan hệ đánh đổi giữa công bằng và hiệu suất dự đoán.

2.5.1. Thảo luận các phát hiện chính từ kết quả thực nghiệm

Mục này tập trung phân tích và tổng hợp các phát hiện chính rút ra từ kết quả thực nghiệm, qua đó làm rõ cách các vấn đề về thiên vị dữ liệu, lựa chọn mô hình học máy và phương pháp tiền xử lý ảnh hưởng đến mức độ công bằng và hiệu suất của hệ thống. Các phân tích được trình bày theo từng khía cạnh cụ thể, tương ứng với các mục tiêu nghiên cứu đã được xác định trong luận án.

Về thiên vị hệ thống trong dữ liệu giáo dục, kết quả phân tích các chỉ số công bằng trên sáu thuộc tính nhạy cảm trong các tập dữ liệu giáo dục cho thấy không có thuộc tính nào thể hiện mức độ thiên vị một cách nổi bật và nhất quán hơn các thuộc tính còn lại. Một số thuộc tính như *giới tính* có mức dao động lớn giữa các tập, phản ánh ảnh hưởng không ổn định đến kết quả mô hình và

cho thấy khó khăn trong việc kiểm soát thiên lệch liên quan đến thuộc tính này. Ngược lại, thuộc tính như *tuổi* có mức biến thiên nhỏ hơn, cho thấy mô hình có xu hướng ổn định hơn trong việc xử lý các thuộc tính này. Dù chỉ xuất hiện trong một số bộ dữ liệu, các thuộc tính như *tình trạng nợ*, *sức khỏe*, và *khuvực* vẫn ghi nhận những giá trị không công bằng khá cao trong một số trường hợp, cho thấy tiềm năng ảnh hưởng đáng kể đến công bằng tổng thể. Những quan sát trên, có thể rút ra rằng việc đánh giá công bằng không nên tập trung vào một vài thuộc tính riêng lẻ mà cần được thực hiện đồng thời trên toàn bộ các thuộc tính nhạy cảm để nhận diện đầy đủ các nguy cơ thiên vị. Điều này cho thấy, việc tìm ra một giải pháp có thể xử lý đồng thời nhiều thuộc tính nhạy cảm trong một bộ dữ liệu là rất cần thiết. Cách tiếp cận này đặc biệt cần thiết trong các bài toán học máy có yếu tố xã hội, nơi các thuộc tính nhạy cảm có thể ảnh hưởng khác nhau tùy theo ngữ cảnh và dữ liệu như trong lĩnh vực giáo dục.

Xét ảnh hưởng của mô hình học máy, kết quả thực nghiệm khẳng định rằng mức độ công bằng thay đổi đáng kể theo từng thuật toán học máy, ngay cả khi cùng áp dụng trên một bộ dữ liệu và một thuộc tính nhạy cảm. Các chỉ số công bằng như $|1 - DI|$, *SPD*, *AOD* và *EOD* đều cho thấy sự khác biệt rõ rệt giữa các mô hình. Trong nhiều trường hợp, *Hồi quy logistic* thể hiện mức công bằng cao hơn so với *Rừng ngẫu nhiên* và *Cây quyết định*. Đặc biệt, mô hình *Rừng ngẫu nhiên* thường cho mức thiên vị cao nhất trên nhiều chỉ số và thuộc tính, mặc dù đôi khi vẫn có ngoại lệ. Những kết quả này cho thấy việc lựa chọn mô hình không chỉ ảnh hưởng đến hiệu suất mà còn quyết định mức độ công bằng đầu ra, do đó cần được cân nhắc kỹ lưỡng trong các bài toán học máy có liên quan đến các thuộc tính nhạy cảm.

Đối với khả năng đảm bảo tính công bằng đồng thời nhiều thuộc tính nhạy cảm, Fairedu đã chứng minh được khả năng giảm thiểu sự phụ thuộc giữa thuộc tính đầu vào và các thuộc tính nhạy cảm, từ đó nâng cao mức độ công bằng tổng thể. Phương pháp này đặc biệt hiệu quả trong các tình huống tồn tại nhiều thuộc tính nhạy cảm nhờ khả năng điều chỉnh đồng thời nhiều quan hệ giao thoa. Kỹ thuật hồi quy tuyến tính đa biến mà Fairedu sử dụng cho phép mô hình hóa ảnh hưởng của các thuộc tính nhạy cảm đến từng thuộc tính không nhạy cảm, sau đó loại bỏ ảnh hưởng đó khỏi dữ liệu. Ví dụ, nếu điểm Toán chịu

ảnh hưởng của *giới tính* và *khu vực*, Fairedu xây dựng một mô hình hồi quy để tách phần ảnh hưởng này ra khỏi dữ liệu đầu vào, tạo ra phiên bản đã điều chỉnh. Cách tiếp cận này cũng đặc biệt hiệu quả trong việc xử lý công bằng giao thoa, tức là những tình huống mà một cá nhân nằm tại giao điểm của nhiều thuộc tính nhạy cảm (ví dụ: nữ giới da đen), từ đó cung cấp một nền tảng công bằng hơn trong toàn bộ quá trình học máy.

Xét mối quan hệ giữa công bằng và hiệu suất, kết quả thực nghiệm cho thấy Fairedu duy trì được hiệu suất mô hình ở mức ổn định, với mức suy giảm trong “*độ chuẩn xác*” và “*độ hồi tưởng*” là rất nhỏ. Cụ thể, “*độ chuẩn xác*” tăng trong 10 trường hợp, tương đương trong 15 trường hợp, và giảm trong 20 trường hợp, với mức giảm tối đa chỉ 5.71%. Với chỉ số “*độ hồi tưởng*”, phương pháp cho kết quả cao hơn trong bốn trường hợp, tương đương trong 16 trường hợp, và giảm trong 25 trường hợp, nhưng độ lệch lớn nhất vẫn dưới 10%. Mặc dù “*điểm số F1*” có sự suy giảm nhiều hơn (chỉ bảy trường hợp tăng trên tổng số 43 trường hợp), “*độ chính xác*” lại được cải thiện rõ rệt trong 20 trường hợp. Những kết quả này cho thấy Fairedu chỉ gây ra đánh đổi hiệu suất nhẹ và vẫn đảm bảo chất lượng dự đoán tốt.

Tổng thể, các kết quả thực nghiệm cho thấy FairEdu là một phương pháp tiên xử lý hiệu quả và linh hoạt trong việc cải thiện công bằng cho các hệ thống học máy ứng dụng trong giáo dục. Phương pháp không chỉ xử lý đồng thời nhiều thuộc tính nhạy cảm mà còn thích ứng tốt với các bộ dữ liệu đa chiều và các mô hình học máy khác nhau. Tuy nhiên, để đánh giá đầy đủ tính tổng quát của phương pháp, cần tiếp tục mở rộng nghiên cứu trên các tập dữ liệu đa dạng hơn và các kiến trúc mô hình học máy phức tạp hơn trong tương lai.

2.5.2. Hạn chế của nghiên cứu

Mặc dù Fairedu thể hiện tiềm năng rõ rệt trong việc cải thiện tính công bằng trên nhiều thuộc tính nhạy cảm trong các bộ dữ liệu giáo dục, vẫn tồn tại một số hạn chế liên quan đến *độ tin cậy nội tại*, *độ tin cậy bên ngoài*, *giá trị cấu trúc* và *giá trị kết luận* cần được thừa nhận [137, 138, 139]. Để đảm bảo tính hợp lệ của nghiên cứu, nghiên cứu đã tuân thủ các nguyên tắc đánh giá độ tin cậy theo Runeson [138].

Độ tin cậy nội tại: Fairedu dựa trên hồi quy tuyến tính đa biến để phát hiện và loại bỏ sự phụ thuộc giữa các thuộc tính và các thuộc tính nhạy cảm. Tuy nhiên, giả định tính tuyến tính có thể hạn chế khả năng của phương pháp trong việc nắm bắt các mối quan hệ phi tuyến vốn có trong một số bộ dữ liệu, khiến một phần thiên vị có thể chưa được loại bỏ hoàn toàn. Trong một nghiên cứu liên quan của Li và cộng sự [31], tác giả đã so sánh hồi quy tuyến tính và hồi quy đa thức, cho thấy hồi quy tuyến tính có hiệu quả vượt trội hơn đáng kể. Ngoài ra, quá trình đánh giá trong nghiên cứu này tập trung vào một số chỉ số công bằng cụ thể. Mặc dù đây là các chỉ số phổ biến và được công nhận rộng rãi, chúng có thể chưa bao quát được toàn bộ các khía cạnh công bằng trong mọi bối cảnh giáo dục. Việc lựa chọn chỉ số có thể ảnh hưởng đến kết luận nghiên cứu, và có khả năng bỏ sót các khía cạnh công bằng quan trọng khác.

Độ tin cậy bên ngoài: Việc đánh giá phương pháp Fairedu được thực hiện trên các bộ dữ liệu đặc thù của lĩnh vực giáo dục. Dù các bộ dữ liệu này được chọn để đại diện cho nhiều bối cảnh giáo dục khác nhau, chúng vẫn có thể chưa phản ánh đầy đủ sự đa dạng của các môi trường giáo dục trong thực tiễn, khiến cho khả năng áp dụng rộng rãi của Fairedu trong những bối cảnh khác vẫn còn là một dấu hỏi. Bên cạnh đó, nghiên cứu này chỉ triển khai trên các thuật toán học máy truyền thống (như *Hồi quy logistic*, *Rừng ngẫu nhiên* và *Cây quyết định*), do đó tính khái quát của phương pháp đối với các kỹ thuật hiện đại hơn như mạng nơ-ron nhân tạo hay học sâu còn hạn chế. Thêm vào đó, mặc dù Fairedu cho thấy hiệu quả trong việc giải quyết các vấn đề về công bằng trong dữ liệu giáo dục, việc mở rộng phương pháp này đến các tình huống dữ liệu lớn hoặc dữ liệu có nhiều biến phi số cũng đặt ra một số thách thức, được liệt kê như sau: *Thứ nhất, xử lý các thuộc tính phi số:* Các mô hình hồi quy tuyến tính, vốn là nền tảng của Fairedu, được thiết kế cho dữ liệu dạng số. Khi các thuộc tính nhạy cảm hoặc các biến đầu vào là dạng phân loại hoặc thứ tự, cần thực hiện các bước tiền xử lý như mã hóa one-hot hoặc mã hóa thứ bậc. Tuy nhiên, các phương pháp này có thể không phản ánh đầy đủ cấu trúc hoặc mối quan hệ tiềm ẩn trong dữ liệu. *Thứ hai, khó khăn khi mở rộng với bộ dữ liệu lớn:* Độ phức tạp tính toán của hồi quy đa biến tăng nhanh theo kích thước dữ liệu, đặc biệt khi số lượng thuộc tính hoặc số lượng bản ghi lớn. Phương pháp Fairedu yêu cầu tính toán mức độ phụ thuộc của từng biến đầu vào với các thuộc tính nhạy cảm trước khi huấn luyện, làm cho khối lượng tính toán tăng mạnh theo

quy mô dữ liệu. *Thứ ba, nhạy cảm với hiện tượng đa cộng tuyến*: Trong các bộ dữ liệu lớn, khả năng tồn tại các mối tương quan tuyến tính giữa các thuộc tính là rất cao, điều này có thể ảnh hưởng đến độ ổn định và độ chính xác của mô hình hồi quy. *Thứ tư, giới hạn trong mô hình hóa quan hệ phi tuyến*: Vấn đề này càng trở nên rõ rệt trong các bộ dữ liệu lớn, nơi các mẫu phức tạp và phi tuyến có xu hướng xuất hiện phổ biến hơn. Mặc dù Fairedu đảm bảo tính công bằng trong phạm vi mô hình tuyến tính, việc mở rộng để xử lý phi tuyến sẽ đòi hỏi tích hợp các phương pháp mô hình hóa thay thế như phương pháp kernel hoặc các kỹ thuật học sâu.

Giá trị cấu trúc: Dù Fairedu có thể xử lý đồng thời nhiều thuộc tính nhạy cảm, các tương tác phức tạp giữa các bản sắc giao thoa có thể vượt quá khả năng mô hình hóa hiện tại. Về lý thuyết, Fairedu có thể áp dụng cho cả biến rời rạc và biến liên tục. Tuy nhiên, trong lĩnh vực giáo dục, các thuộc tính nhạy cảm thường là biến rời rạc, còn thuộc tính đầu ra thường ở dạng nhị phân. Điều này có thể làm hạn chế khả năng áp dụng của mô hình trong các trường hợp phức tạp hơn về cấu trúc dữ liệu và yêu cầu công bằng.

Giá trị kết luận: Mặc dù Fairedu hướng tới mục tiêu cân bằng giữa công bằng và hiệu suất dự đoán, vẫn tồn tại khả năng phải đánh đổi trong các trường hợp đặc biệt là khi các thuộc tính nhạy cảm có sự mất cân đối cao hoặc xung đột lẫn nhau. Việc nâng cao tính công bằng cho một nhóm có thể vô tình làm ảnh hưởng đến hiệu suất hoặc công bằng của nhóm khác, ngay cả khi Fairedu đã được thiết kế để xử lý đồng thời nhiều thuộc tính. Thêm vào đó, việc điều chỉnh dữ liệu nhằm loại bỏ sự phụ thuộc có thể làm thay đổi các mối quan hệ quan trọng khác trong dữ liệu, ảnh hưởng đến khả năng diễn giải và tính hữu ích của mô hình học máy sau cùng. Trong nghiên cứu này, phần lớn các thí nghiệm được thực hiện trên các bộ dữ liệu chỉ chứa hai hoặc ba thuộc tính nhạy cảm. Trong tương lai, chúng tôi dự định mở rộng nghiên cứu này sang các bộ dữ liệu phong phú hơn với nhiều thuộc tính nhạy cảm hơn để đánh giá toàn diện hơn về độ chính xác và độ bền vững của phương pháp Fairedu.

2.5.3. Ý nghĩa và ứng dụng thực tiễn

Phương pháp Fairedu không chỉ đóng vai trò như một giải pháp kỹ thuật nhằm cải thiện tính công bằng trong các hệ thống học máy, mà còn mang lại những giá trị ứng dụng thực tiễn quan trọng, đặc biệt trong lĩnh vực giáo dục, nơi các quyết định dựa trên dữ liệu ngày càng có ảnh hưởng lớn đến quyền lợi của người học. Việc áp dụng Fairedu vào các bộ dữ liệu tuyển sinh có thể góp phần hiệu chỉnh các thiên vị tiềm ẩn liên quan đến các yếu tố như *giới tính*, *chủng tộc*, điều kiện kinh tế – xã hội, v.v. trước khi các yếu tố này tác động đến kết quả dự đoán của mô hình. Điều này giúp đảm bảo quá trình đánh giá và tuyển chọn sinh viên diễn ra một cách công bằng hơn, giảm thiểu nguy cơ mô hình học máy vô tình tái tạo hoặc khuếch đại các thiên vị xã hội có sẵn. Đây là một hướng ứng dụng có ý nghĩa đặc biệt trong bối cảnh bộ dữ liệu DNU, dữ liệu thực tế từ Trường Đại học Đại Nam đang được nghiên cứu triển khai mô hình.

Bên cạnh đó, phương pháp Fairedu có thể được áp dụng vào các hệ thống đánh giá tự động như chấm điểm bài tập, phân tích kết quả học tập hoặc dự đoán nguy cơ bỏ học. Trong các trường hợp này, Fairedu giúp giảm thiểu khả năng mô hình vô tình đưa ra kết quả bất lợi cho những nhóm sinh viên có hoàn cảnh đặc thù, chẳng hạn như người học đến từ vùng sâu vùng xa, có rào cản ngôn ngữ hoặc hạn chế trong việc tiếp cận tài nguyên học tập. Việc thực hiện hiệu chỉnh dữ liệu trước bước huấn luyện mô hình không chỉ giúp cải thiện công bằng đầu ra, mà còn góp phần nâng cao độ tin cậy và khả năng áp dụng thực tế của hệ thống.

Từ góc độ quản lý giáo dục, việc tích hợp Fairedu vào quy trình xây dựng và triển khai các mô hình học máy sẽ hỗ trợ các trường đại học hiện thực hóa các cam kết về bình đẳng, đa dạng và hòa nhập. Đây là những giá trị ngày càng được đề cao trong bối cảnh giáo dục đại học hướng tới phát triển toàn diện con người và không bỏ lại ai phía sau trong chuyển đổi số. Tuy nhiên, cần lưu ý rằng trong thực tiễn, luôn tồn tại sự đánh đổi giữa công bằng và độ chính xác. Ví dụ, trong bài toán tuyển sinh, nếu ưu tiên công bằng bằng cách điều chỉnh mô hình để đảm bảo tỷ lệ trúng tuyển tương đồng giữa các nhóm nhân khẩu học, có thể dẫn đến mức giảm nhẹ trong độ chính xác tổng thể. Ngược lại, nếu tối ưu hóa

hoàn toàn cho độ chính xác mà không kiểm soát công bằng, mô hình có thể vô tình củng cố các bất công sẵn có trong dữ liệu đầu vào. Do đó, các hệ thống học máy trong giáo dục cần được thiết kế để đồng thời theo dõi và báo cáo cả các chỉ số công bằng bên cạnh các chỉ số hiệu suất truyền thống. Trong vai trò đó, Fairerdu có thể được sử dụng như một bước tiền xử lý nhằm hiệu chỉnh dữ liệu trước khi huấn luyện mô hình, hoặc như một công cụ đánh giá độc lập nhằm đo lường mức độ công bằng của các mô hình hiện có. Việc tích hợp các phương pháp như Fairerdu vào quy trình vận hành học máy trong giáo dục không chỉ nâng cao chất lượng và tính minh bạch của hệ thống, mà còn đóng góp vào việc xây dựng một môi trường học tập công bằng và bền vững cho tất cả người học.

2.6. Tổng kết chương

Chương này đã trình bày phương pháp Fairerdu – một kỹ thuật tiền xử lý dựa trên hồi quy tuyến tính đa biến, được đề xuất nhằm nâng cao tính công bằng trong các mô hình học máy, đặc biệt trong lĩnh vực giáo dục. Phương pháp hướng tới giải quyết ba thách thức trọng tâm trong nghiên cứu về đảm bảo tính công bằng, bao gồm: (i) khả năng xử lý đồng thời nhiều thuộc tính nhạy cảm theo cách có thể mở rộng; (ii) cải thiện mức độ công bằng so với các phương pháp tiền xử lý phổ biến như Reweighting, Fair-SMOTE và LTDD; và (iii) hạn chế mức đánh đổi bất lợi giữa công bằng và hiệu suất dự đoán sau can thiệp.

Thông qua thực nghiệm trên bốn bộ dữ liệu giáo dục và ba mô hình học máy phổ biến (LR, RF, DT), Fairerdu cho thấy hiệu quả rõ rệt trong việc giảm thiểu thiên vị liên quan đến các thuộc tính nhạy cảm. Cụ thể, phương pháp đạt mức giảm đáng kể đối với các chỉ số công bằng, chẳng hạn như $|1 - DI|$ theo *giới tính* trên tập *Oulad*, *SPD* trên tập *Adult*, *AOD* theo *chủng tộc* trên tập *Compas*, và *EOD* theo *giới tính* trên tập *Oulad*. Những kết quả này cho thấy Fairerdu có khả năng làm suy giảm đáng kể mức độ phụ thuộc giữa đầu ra dự đoán và các thuộc tính nhạy cảm trong dữ liệu.

Đáng chú ý, mặc dù thực hiện điều chỉnh dữ liệu đầu vào nhằm tăng cường công bằng, Fairerdu vẫn duy trì hiệu suất mô hình ở mức ổn định, với mức thay đổi nhỏ và nằm trong ngưỡng chấp nhận được. Trong một số trường hợp, độ

chính xác thậm chí còn được cải thiện. Điều này cho thấy Fairedu không chỉ mang lại lợi ích về mặt công bằng mà còn đáp ứng tốt yêu cầu về hiệu quả kỹ thuật, qua đó thể hiện tính khả thi khi áp dụng trong các hệ thống học máy thực tế.

Bên cạnh giá trị định lượng, chương này cũng làm rõ tiềm năng ứng dụng thực tiễn của Fairedu trong các bài toán như tuyển sinh, đánh giá kết quả học tập tự động, hoặc phân tích nguy cơ bỏ học. Nhờ khả năng xử lý thiên vị giao thoa giữa nhiều thuộc tính nhạy cảm, Fairedu đặc biệt phù hợp với các hệ thống học máy áp dụng trên dữ liệu phức tạp và đa chiều, nơi yêu cầu về công bằng và minh bạch ngày càng trở nên cấp thiết. Với tính linh hoạt cao, phương pháp cũng có thể được mở rộng sang các lĩnh vực khác như tài chính, y tế hoặc thị trường lao động, nơi các quyết định do học máy hỗ trợ có tác động xã hội đáng kể.

Các kết quả và phân tích trong chương này tạo nền tảng trực tiếp cho các nghiên cứu tiếp theo của luận án, đặc biệt là các phương pháp can thiệp ở mức dữ liệu và kết hợp đa chiến lược được trình bày trong các chương sau.

Các kết quả nghiên cứu chính của chương đã được công bố trong công trình khoa học do tác giả là tác giả chính trên tạp chí *Expert Systems with Applications* (Q1, vol. 269, 2025).

Chương 3

PHƯƠNG PHÁP ĐẢM BẢO TÍNH CÔNG BẰNG NHỜ CÂN BẰNG DỮ LIỆU DỰA TRÊN KỸ THUẬT SINH DỮ LIỆU TỔNG HỢP

Như đã phân tích trong Chương 1, một trong những thách thức trung tâm của công bằng trong học máy, đặc biệt trong lĩnh vực giáo dục, xuất phát từ sự mất cân bằng và thiếu tính đại diện của dữ liệu huấn luyện. Vấn đề này trở nên nghiêm trọng hơn khi dữ liệu đồng thời chứa nhiều thuộc tính nhạy cảm, dẫn đến sự thiếu hụt nghiêm trọng các nhóm giao thoa và làm suy giảm độ tin cậy của cả quá trình huấn luyện lẫn đánh giá công bằng. Chương này tập trung trả lời RQ2 thông qua việc nghiên cứu và đề xuất một hướng tiếp cận đảm bảo công bằng dựa trên cân bằng dữ liệu huấn luyện bằng kỹ thuật sinh dữ liệu tổng hợp. Trọng tâm của chương là khảo sát và khai thác các phương pháp sinh dữ liệu hiện đại nhằm tăng cường tính đại diện cho các nhóm giao thoa có kích thước nhỏ hoặc bị thiệt thòi trong dữ liệu gốc. Bằng cách tái cấu trúc phân phối dữ liệu đầu vào, các kỹ thuật sinh dữ liệu tổng hợp cho phép giảm thiểu thiên lệch phát sinh từ dữ liệu không cân bằng và tạo điều kiện thuận lợi hơn cho việc đánh giá và cải thiện công bằng ở mức nhóm và nhóm giao thoa.

Trên cơ sở đó, chương đề xuất phương pháp *Data Partitioning Fairness (DPF)*, một khuôn khổ can thiệp theo chiều ngang nhằm chia và cân bằng lại tập dữ liệu huấn luyện theo các tổ hợp thuộc tính nhạy cảm và nhân mục tiêu. Phương pháp này tận dụng các mô hình sinh dữ liệu như GAN có điều kiện và các mô hình ngôn ngữ lớn để bổ sung dữ liệu cho các nhóm con hiếm, đồng thời cho phép kiểm soát linh hoạt phân phối dữ liệu trong bối cảnh nhiều thuộc tính nhạy cảm cùng tồn tại.

Nội dung của chương bao gồm việc trình bày khái niệm và vai trò của dữ liệu tổng hợp trong đảm bảo công bằng, tổng quan các kỹ thuật sinh dữ liệu tổng hợp hiện có, mô tả chi tiết phương pháp DPF và kiến trúc thực nghiệm,

cũng như phân tích kết quả đạt được và các thảo luận liên quan. Thông qua đó, chương nhằm xây dựng một nền tảng kỹ thuật rõ ràng và khả thi để cải thiện tính công bằng cho các mô hình học máy trên dữ liệu dạng bảng, đặc biệt trong bối cảnh giáo dục, nơi dữ liệu thường chứa nhiều yếu tố nhạy cảm, có tính riêng tư cao và tồn tại sự mất cân bằng đáng kể giữa các nhóm người học.

3.1. Giới thiệu

Trong lĩnh vực giáo dục, công bằng được hiểu là đảm bảo cơ hội học tập và đánh giá bình đẳng cho mọi người học, không bị ảnh hưởng bởi các đặc điểm cá nhân như giới tính, chủng tộc hay điều kiện kinh tế [68]. Sự phát triển của học máy trong giáo dục mang lại nhiều cơ hội đổi mới nhưng đồng thời cũng đặt ra yêu cầu cao hơn về tính đạo đức và công bằng bên cạnh các tiêu chí truyền thống như độ chính xác và hiệu suất [140, 141, 142]. Tuy nhiên, các mô hình học máy có thể tạo ra hoặc khuếch đại thiên lệch do dữ liệu huấn luyện mất cân bằng hoặc thiếu tính đại diện, khiến các nhóm ít được đại diện dễ bị đánh giá sai lệch [15]. Vấn đề này trở nên nghiêm trọng hơn khi dữ liệu chứa nhiều thuộc tính nhạy cảm, dẫn đến thiên lệch giao thoa và bất lợi đa chiều cho các nhóm thiểu số [69, 143, 144]. Mặc dù đã có nhiều nghiên cứu về công bằng đơn biến, công bằng giao thoa trong các bài toán giáo dục vẫn chưa được xem xét đầy đủ [15].

Một trong những hướng tiếp cận hiệu quả để cải thiện công bằng là điều chỉnh dữ liệu đầu vào thông qua các phương pháp tiền xử lý. Hai kỹ thuật tiêu biểu bao gồm: (i) điều chỉnh trọng số dữ liệu nhằm cân bằng phân phối giữa các nhóm giao nhau, tiêu biểu là Reweighting [30], và (ii) sinh dữ liệu tổng hợp có điều kiện để tăng cường đại diện cho các nhóm thiểu số, tiêu biểu là TabFairGAN [32].

Reweighting tính toán trọng số cho từng nhóm giao nhau giữa thuộc tính nhạy cảm và nhãn theo công thức:

$$w_{ij} = \frac{\Pr(S_i) \Pr(Y_j)}{\Pr(S_i \cap Y_j)} \quad (3.1)$$

giúp giảm thiểu ảnh hưởng của mất cân bằng dữ liệu trong quá trình huấn luyện.

Trong khi đó, TabFairGAN sử dụng kiến trúc GAN có điều kiện để sinh dữ liệu bảng, với hàm mất mát kết hợp giữa chất lượng dữ liệu và công bằng:

$$\mathcal{L}_{\text{TabFairGAN}} = \mathcal{L}_{\text{WGAN}} + \lambda \cdot \mathcal{L}_{\text{fair}}, \quad (3.2)$$

trong đó thành phần công bằng được xác định bởi:

$$\mathcal{L}_{\text{fair}} = |\mathbb{E}[\hat{y} | s = 0] - \mathbb{E}[\hat{y} | s = 1]|. \quad (3.3)$$

Ngoài ra, sinh dữ liệu tổng hợp nói chung được xem là một giải pháp hiệu quả nhằm tái cân bằng dữ liệu, kiểm soát thiên lệch và bảo vệ quyền riêng tư [95, 96, 100, 110, 111, 145].

Mặc dù các phương pháp hiện tại đã đạt được những kết quả nhất định, vẫn tồn tại các hạn chế quan trọng. Các phương pháp dựa trên trọng số như Reweighting có thể xử lý rõ ràng các nhóm giao nhau nhưng không tạo thêm dữ liệu cho các nhóm thiểu số, do đó không khắc phục được vấn đề thiếu đại diện. Ngược lại, các phương pháp sinh dữ liệu như TabFairGAN có khả năng tạo dữ liệu chất lượng cao nhưng thiếu cơ chế kiểm soát trực tiếp phân phối giữa các nhóm giao thoa, đặc biệt khi số lượng thuộc tính nhạy cảm tăng lên, dẫn đến khó đảm bảo cân bằng và ổn định. Nhìn chung, các phương pháp hiện nay chủ yếu tập trung vào công bằng đơn biến và còn hạn chế trong xử lý công bằng giao thoa một cách có kiểm soát [145].

Khoảng trống này đặt ra nhu cầu về một phương pháp có khả năng vừa xác định rõ các nhóm giao thoa, vừa chủ động cân bằng dữ liệu giữa các nhóm này. Để giải quyết vấn đề đó, nghiên cứu đề xuất phương pháp *Data Partitioning Fairness (DPF)*, kết hợp phân chia dữ liệu theo nhóm giao thoa và sinh dữ liệu tổng hợp (CTGAN, LLM) nhằm tái cân bằng dữ liệu một cách minh bạch, ổn định và có khả năng mở rộng.

3.2. Phương pháp DPF

Phương pháp đảm bảo tính công bằng nhờ cân bằng dữ liệu (*Data Partitioning Fairness* – gọi tắt là DPF) được đề xuất trong nghiên cứu này với mục tiêu cân

bằng lại phân phối dữ liệu huấn luyện giữa các nhóm giao nhau của các thuộc tính nhạy cảm và nhân mục tiêu bằng cách sử dụng các kỹ thuật sinh dữ liệu tổng hợp. DPF cho phép kiểm soát linh hoạt số lượng bản ghi sinh ra cho từng nhóm, từ đó đảm bảo sự phân bố dữ liệu công bằng hơn và có thể mở rộng áp dụng cho nhiều thuộc tính nhạy cảm cùng lúc.

3.2.1. Nguyên lý hoạt động

Giả sử tập dữ liệu đầu vào gồm k thuộc tính nhạy cảm nhị phân (x_1, x_2, \dots, x_k) và một nhân mục tiêu nhị phân $y \in \{0, 1\}$. Tất cả các bản ghi dữ liệu được chia thành $n = 2^k$ nhóm theo tổ hợp các giá trị của các thuộc tính nhạy cảm được đánh số từ 1 đến 2^k . Mỗi nhóm thứ i ($i = \overline{1, 2^k}$) sau đó lại được chia thành hai phân lớp con theo giá trị nhân mục tiêu là $y = 1$ và $y = 0$, lần lượt được ký hiệu là D_{i1} và D_{i0} . Như vậy, kết quả sau khi phân chia, tập dữ liệu sẽ có tổng cộng 2^{k+1} nhóm con D_{ij} , ($i = \overline{1, 2^k}, j \in \{0, 1\}$).

Quy trình sinh dữ liệu trong phương pháp DPF được thực hiện độc lập trên từng nhóm con D_{ij} . Đối với mỗi nhóm con, số lượng bản ghi hiện có được so sánh với một ngưỡng mục tiêu đã xác định trước. Trong trường hợp số lượng bản ghi của nhóm con nhỏ hơn ngưỡng này, các kỹ thuật sinh dữ liệu tổng hợp được áp dụng để tạo thêm các bản ghi mới nhằm đạt được tỷ lệ phân bố mong muốn. Cách tiếp cận này cho phép kiểm soát linh hoạt mức độ cân bằng dữ liệu giữa các nhóm con, đồng thời hạn chế việc sinh dữ liệu dư thừa không cần thiết.

Để đảm bảo tính ổn định và độ tin cậy của quá trình sinh dữ liệu, luận án chỉ xem xét áp dụng DPF đối với các nhóm con D_{ij} có số lượng bản ghi ban đầu từ hai mẫu trở lên. Các nhóm con có số lượng bản ghi quá ít (ví dụ chỉ có một hoặc không có bản ghi) không được đưa vào quá trình sinh dữ liệu, do không cung cấp đủ thông tin thống kê cần thiết để mô hình sinh học được phân bố dữ liệu hợp lý. Giả định này giúp tránh hiện tượng sinh dữ liệu suy diễn quá mức, đồng thời đảm bảo rằng các mẫu tổng hợp phản ánh được đặc trưng phân bố ban đầu của từng nhóm con.

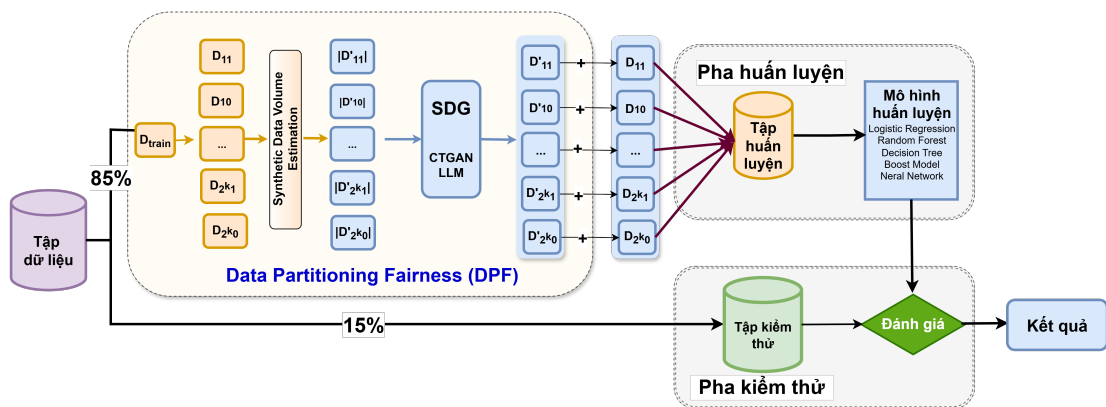
Mục tiêu của phương pháp là đảm bảo rằng tỷ lệ giữa số lượng bản ghi của mỗi cặp nhóm con D_{i1} và D_{i0} sau khi sinh dữ liệu tổng hợp phải đạt cùng một

tỷ lệ đã xác định trước, từ đó làm giảm chênh lệch phân phối nhân giữa các nhóm nhạy cảm, đồng thời giúp mô hình học máy không bị thiên lệch theo bất kỳ tổ hợp nào của thuộc tính nhạy cảm.

3.2.2. Kiến trúc tổng thể

Kiến trúc tổng thể của phương pháp DPF được minh họa trong Hình 3.1. Phương pháp gồm hai pha chính: pha huấn luyện và pha kiểm thử. Trong pha huấn luyện, tập dữ liệu được chia thành tập huấn luyện (85%) và tập kiểm thử (15%). Tập huấn luyện được đưa vào quy trình DPF, bao gồm các bước phân chia dữ liệu theo các nhóm giao thoa, xác định kích thước mục tiêu, sinh dữ liệu tổng hợp bằng CTGAN hoặc LLM, sau đó hợp nhất các nhóm để tạo tập huấn luyện cân bằng phục vụ huấn luyện mô hình học máy. Trong pha kiểm thử, tập kiểm thử được giữ nguyên và sử dụng để đánh giá mô hình theo các chỉ số hiệu suất và công bằng.

Thiết kế này giúp DPF chỉ can thiệp vào dữ liệu huấn luyện mà không thay đổi thuật toán hay kiến trúc của mô hình học máy, nhờ đó có thể tích hợp linh hoạt với nhiều mô hình khác nhau. Toàn bộ quy trình thực hiện DPF được tổ chức thành bốn bước chính: chuẩn bị dữ liệu, cân bằng dữ liệu, huấn luyện mô hình và kiểm tra, đánh giá.



Hình 3.1: Kiến trúc tổng thể của phương pháp DPF

Bước 1: Chuẩn bị và phân tách dữ liệu. Quy trình bắt đầu từ tập dữ liệu gốc, bao gồm cả các thuộc tính nhạy cảm và không nhạy cảm. Tập dữ liệu được chia thành hai phần, trong đó 85% dữ liệu được sử dụng làm tập huấn luyện D_{tr} và 15% còn lại làm tập kiểm thử D_{test} . Giả sử mỗi bản ghi trong D_{tr} được mô tả bởi k thuộc tính nhạy cảm nhị phân (x_1, \dots, x_k) và một nhãn đầu ra nhị phân Y , tập huấn luyện được phân tách thành 2^{k+1} nhóm con D_{ij} theo tất cả các tổ hợp giá trị của các thuộc tính nhạy cảm và nhãn đầu ra, với $i = \overline{1, 2^k}$ và $j \in \{0, 1\}$. Mỗi cặp (D_{i0}, D_{i1}) tương ứng với cùng một tổ hợp thuộc tính nhạy cảm nhưng khác nhau ở giá trị nhãn Y . Cách phân tách này cho phép biểu diễn tường minh các nhóm giao thoa (intersectional groups) trong dữ liệu giáo dục, vốn là nguồn gốc chính gây ra hiện tượng mất cân bằng và thiên lệch.

Sau khi phân tách, các thuộc tính nhạy cảm được loại bỏ khỏi từng tập con D_{ij} . Bước tiền xử lý này nhằm đảm bảo rằng quá trình sinh dữ liệu tổng hợp và huấn luyện mô hình ở các bước tiếp theo không trực tiếp sử dụng thông tin nhạy cảm, từ đó hạn chế nguy cơ mô hình học theo các khuôn mẫu thiên lệch.

Bước 2: Xác định nhóm con gốc và mục tiêu cân bằng. Do việc sinh dữ liệu tổng hợp đòi hỏi một số lượng mẫu tối thiểu để ước lượng phân bố dữ liệu một cách ổn định, DPF chỉ xem xét các nhóm con có số lượng bản ghi ban đầu từ hai mẫu trở lên. Các nhóm con có số lượng bản ghi quá nhỏ (ít hơn hai) được giữ nguyên và không tham gia vào quá trình sinh dữ liệu, đồng thời cặp nhóm cùng tổ hợp thuộc tính nhạy cảm tương ứng cũng được giữ nguyên để tránh suy diễn phân bố từ dữ liệu không đủ thông tin.

Trong số các nhóm con thỏa điều kiện này, thuật toán xác định nhóm con có số lượng bản ghi nhỏ nhất bằng cách tìm giá trị nhỏ nhất của số bản ghi trong các nhóm con (D_{min} theo Công thức 3.4

$$D_{min} = \min_{i=1, \dots, 2^k; j \in \{0, 1\}} \|D_{ij}\|. \quad (3.4)$$

Nhóm đạt giá trị này được chọn làm nhóm con gốc và được ký hiệu là $D_{i_s j_s}$. Tiếp theo, một kích thước mục tiêu n_0 được lựa chọn sao cho $n_0 > \|D_{i_s j_s}\|$, và đặt $\|D'_{i_s j_s}\| = n_0$. Nhóm con gốc đóng vai trò làm mốc tham chiếu để xác định kích thước mục tiêu cho các nhóm con còn lại.

Bước 3: Suy ra kích thước mục tiêu cho các nhóm giao thoa. Để duy trì sự nhất quán về phân bố nhân giữa các nhóm giao thoa, DPF tính tỷ lệ trung bình giữa số lượng mẫu nhân dương và nhân âm (R_{mean}) trong các nhóm có cùng tổ hợp thuộc tính nhạy cảm bằng Công thức 3.5.

$$R_{mean} = \frac{1}{2^k} \sum_{i=1}^{2^k} \frac{\|D_{i1}\|}{\|D_{i0}\|}. \quad (3.5)$$

Dựa trên vị trí của nhóm con gốc, kích thước mục tiêu $\|D'_{ij}\|$ của các nhóm con còn lại được suy ra theo hai trường hợp. Nếu nhóm con gốc thuộc nhân dương ($j_s = 1$), kích thước mục tiêu được xác định theo Công thức 3.6.

$$\begin{cases} \|D'_{i1}\| = \frac{\|D_{i1}\|}{\|D_{i_s j_s}\|} \cdot n_0, \\ \|D'_{i0}\| = \frac{\|D'_{i1}\|}{R_{mean}}. \end{cases} \quad (3.6)$$

Ngược lại, nếu nhóm con gốc thuộc nhân âm ($j_s = 0$), kích thước mục tiêu được xác định theo Công thức 3.7

$$\begin{cases} \|D'_{i0}\| = \frac{\|D_{i0}\|}{\|D_{i_s j_s}\|} \cdot n_0, \\ \|D'_{i1}\| = R_{mean} \cdot \|D'_{i0}\|. \end{cases} \quad (3.7)$$

Cách suy ra này cho phép mở rộng dữ liệu một cách có kiểm soát, vừa đảm bảo cân bằng giữa các nhóm giao thoa, vừa hạn chế làm sai lệch mối quan hệ giữa nhân và đặc trưng trong dữ liệu ban đầu.

Chú ý: Trong các công thức ở trên, các kí hiệu $\|D_{ij}\|$ là dùng để chỉ số lượng các bản ghi có trong các tập con D_{ij} tương ứng.

Bước 4: Sinh dữ liệu, huấn luyện và đánh giá. Sau khi xác định kích thước mục tiêu cho từng nhóm con, phương pháp DPF tiến hành sinh thêm $|D'_{ij}| - |D_{ij}|$ bản ghi mới cho mỗi tập con bằng các kỹ thuật sinh dữ liệu tổng hợp, bao gồm CTGAN và LLM.

Đối với phương pháp CTGAN, dữ liệu được phân tách theo từng nhóm con dựa trên các tổ hợp của thuộc tính nhạy cảm và nhân đầu ra. Với mỗi nhóm

con, một mô hình CTGAN riêng biệt được huấn luyện trên các thuộc tính không nhạy cảm nhằm học phân bố dữ liệu tương ứng. Trong nghiên cứu này, CTGAN được triển khai thông qua thư viện SDV với các tham số gồm 100 vòng huấn luyện (epochs), batch size bằng 500, learning rate bằng 2×10^{-4} và embedding dimension bằng 128. Sau khi huấn luyện, mô hình được sử dụng để sinh thêm các bản ghi mới cho từng nhóm con, đảm bảo đạt được kích thước mục tiêu đã xác định trong bước cân bằng dữ liệu. Quá trình này giúp duy trì cấu trúc phân bố dữ liệu trong từng nhóm, đồng thời cải thiện sự cân bằng giữa các nhóm giao thoa.

Đối với phương pháp *LLM*, dữ liệu tổng hợp được sinh thông qua cơ chế *prompt-based*. Nghiên cứu sử dụng ChatGPT-3.5 với temperature bằng 0,1, số lượng token tối đa là 4000 và kích thước mỗi lần sinh là 30 bản ghi. Quá trình sinh được thực hiện lặp nhiều lần cho đến khi đạt số lượng mẫu mục tiêu của từng nhóm con. Cụ thể, các prompt được thiết kế dựa trên mô tả thống kê của từng nhóm con, bao gồm thông tin về các thuộc tính không nhạy cảm, các bản ghi đại diện, danh sách thuộc tính cần sinh và số lượng bản ghi yêu cầu, từ đó hướng dẫn mô hình ngôn ngữ sinh ra các bản ghi dữ liệu mới phù hợp với ngữ cảnh. Sau khi sinh dữ liệu, các bước hậu xử lý được thực hiện nhằm chuẩn hóa định dạng, loại bỏ các bản ghi không hợp lệ và đảm bảo tính nhất quán của dữ liệu.

Trong nghiên cứu này, chất lượng dữ liệu tổng hợp được đánh giá gián tiếp thông qua hiệu quả của các mô hình học máy sau khi huấn luyện trên tập dữ liệu đã được bổ sung, dựa trên các chỉ số hiệu suất và công bằng. Để bảo đảm khả năng tái lập, toàn bộ cấu hình tham số, quy trình sinh dữ liệu và số lượng mẫu sinh cho từng nhóm giao thoa được giữ cố định trong các lần thực nghiệm.

Để minh họa quá trình sinh dữ liệu bằng LLM, Hộp màu xám dưới đây trình bày một ví dụ về prompt được sử dụng trong nghiên cứu. Cụ thể, các prompt được thiết kế dựa trên các bản ghi đại diện của từng nhóm con, kết hợp với danh sách các thuộc tính, các ràng buộc về định dạng và số lượng bản ghi cần sinh, từ đó hướng dẫn mô hình ngôn ngữ tạo ra các bản ghi dữ liệu mới phù hợp với đặc trưng của từng nhóm giao thoa.

Suppose you are the best data generating model in this world, we have some data samples with the following information:

sample 0: column_1 is value_1, column_2 is value_2, ...,
column_d is value_d

sample 1: column_1 is value_1, column_2 is value_2, ...,
column_d is value_d

...

sample n: column_1 is value_1, column_2 is value_2, ...,
column_d is value_d

Please note that the generated table has total [number of columns] columns of the generated data, the column names are [column list], every column should not be missed when generating the data.

Generate synthetic data samples based on the above information and your knowledge, each sample should be output on one line (do not output in multiple lines), the output format of the sample is the same as the example in this message, such as “column_name_1 is value_1”, the count of the generated data samples is [batch size].

Sau khi quá trình sinh hoàn tất, các thuộc tính nhạy cảm tương ứng được chèn lại cho từng bản ghi nhằm đảm bảo tính nhất quán theo nhóm giao thoa. Tất cả các tập con D'_{ij} được hợp nhất để tạo thành tập huấn luyện cân bằng D' . Trên tập dữ liệu này, các mô hình học máy được huấn luyện sau khi loại bỏ hoàn toàn các thuộc tính nhạy cảm. Chất lượng của dữ liệu tổng hợp được đánh giá gián tiếp thông qua hiệu quả của các mô hình học máy sau khi huấn luyện, dựa trên các chỉ số hiệu suất và các thước đo công bằng. Để bảo đảm khả năng tái lập, toàn bộ cấu hình tham số của CTGAN và LLM cũng như quy trình sinh dữ liệu được giữ cố định trong các lần thực nghiệm. Bên cạnh đó, dữ liệu được sinh trên các bộ dữ liệu giáo dục đã được ẩn danh và được thực hiện độc lập cho từng nhóm giao thoa, qua đó giảm thiểu nguy cơ rò rỉ thông tin nhạy cảm. Cuối cùng, mô hình được đánh giá trên tập kiểm thử thông qua các chỉ số hiệu suất và các thước đo công bằng nhằm đánh giá mức độ cải thiện công bằng mà DPF mang lại.

3.2.3. Thuật toán

Quy trình hoạt động của phương pháp DPF đã được trình bày chi tiết trong Mục 3.1 và được mô tả hình thức dưới dạng giả mã trong Thuật toán 3.1.

Thuật toán 3.1 DPF – Công bằng nhờ cân bằng dữ liệu

- 1: **Đầu vào:** Tập dữ liệu huấn luyện $D_{tr} = \{\langle x_1, y_1 \rangle, \dots, \langle x_n, y_n \rangle\}$, trong đó $x_j = [x_1^j, \dots, x_d^j]$ gồm k thuộc tính nhạy cảm và $d - k$ thuộc tính không nhạy cảm; $y_j \in \{0, 1\}$. Mẫu kiểm tra x^{te} .
 - 2: **Đầu ra:** D' , mô hình S_{ML} và $S_{ML}(x^{te})$.
 - 3: Gọi $n = 2^k$, chia D_{tr} thành $2n$ tập con D_{ij} theo thuộc tính nhạy cảm và nhãn.
 - 4: Với $i = 1$ đến n : xác định D_{i1} và D_{i0} .
 - 5: $D_{\min} = \min_{i,j} \|D_{ij}\|$.
 - 6: Xác định (i_s, j_s) sao cho $\|D_{i_s j_s}\| = D_{\min}$.
 - 7: Nhập $n_0 > \|D_{i_s j_s}\|$, gán $\|D'_{i_s j_s}\| = n_0$.
 - 8: $R_{\text{mean}} = \frac{1}{n} \sum_{i=1}^n \frac{\|D_{i1}\|}{\|D_{i0}\|}$.
 - 9: **if** $j_s = 1$ **then**
 - 10: **for** $i = 1$ đến n **do**
 - 11: $\|D'_{i1}\| = \frac{\|D_{i1}\|}{\|D_{i_s j_s}\|} \cdot n_0$, $\|D'_{i0}\| = \frac{\|D'_{i1}\|}{R_{\text{mean}}}$
 - 12: **end for**
 - 13: **else**
 - 14: **for** $i = 1$ đến n **do**
 - 15: $\|D'_{i0}\| = \frac{\|D_{i0}\|}{\|D_{i_s j_s}\|} \cdot n_0$, $\|D'_{i1}\| = R_{\text{mean}} \cdot \|D'_{i0}\|$
 - 16: **end for**
 - 17: **end if**
 - 18: Loại bỏ thuộc tính nhạy cảm khỏi D_{ij} .
 - 19: Khởi tạo CTGAN.
 - 20: **for** $i = 1$ đến n , $j = 0$ đến 1 **do**
 - 21: Huấn luyện CTGAN trên D_{ij} .
 - 22: Sinh dữ liệu: D'_{ij} .
 - 23: Lưu và chèn lại thuộc tính nhạy cảm.
 - 24: **end for**
 - 25: Gộp dữ liệu: $\text{combined_df} = \text{concat}(\dots)$.
 - 26: Huấn luyện S_{ML} (không dùng thuộc tính nhạy cảm).
 - 27: Tiền xử lý x^{te} và dự đoán.
 - 28: **return** D' , S_{ML} , $S_{ML}(x^{te})$. =0
-

Đánh giá độ phức tạp của thuật toán DPF. Giả sử tập huấn luyện có N bản ghi và k thuộc tính nhạy cảm. Bước phân tách dữ liệu có độ phức tạp $O(Nk)$, trong khi các bước xác định nhóm con và suy ra kích thước mục tiêu có độ phức tạp $O(2^k)$. Bước sinh dữ liệu và huấn luyện mô hình chiếm chi phí lớn nhất, với độ phức tạp xấp xỉ $O(N_{gen} \cdot C_{gen} + N' \cdot C_{model})$. Trong đó, N_{gen} là số mẫu sinh thêm, N' là kích thước dữ liệu sau cân bằng, C_{gen} là chi phí tính toán để sinh một mẫu dữ liệu tổng hợp, phụ thuộc vào mô hình sinh dữ liệu, C_{model} là chi phí tính toán để huấn luyện mô hình học máy trên một mẫu dữ liệu, phụ thuộc vào loại mô hình được sử dụng.

Tổng thể, độ phức tạp của DPF là $O(Nk + 2^k + N_{gen} \cdot C_{gen} + N' \cdot C_{model})$. Trong thực tế, với k nhỏ, chi phí chủ yếu đến từ quá trình sinh dữ liệu và huấn luyện, cho thấy DPF có khả năng mở rộng tốt trong các bài toán giáo dục.

3.2.4. Ví dụ minh họa

Ví dụ 3.1. Để minh họa hoạt động của phương pháp DPF, luận án sử dụng bộ dữ liệu DNU gồm tổng cộng 426 bản ghi với hai thuộc tính nhạy cảm là giới tính và khu vực, nhãn mục tiêu nhị phân là khả năng tốt nghiệp.

Quy trình áp dụng phương pháp DPF bao gồm các bước như sau:

Bước 1: Chuẩn bị dữ liệu. Bộ dữ liệu với hai thuộc tính nhạy cảm cùng một thuộc tính mục tiêu sẽ trên tạo ra $2^3 = 8$ tổ hợp tương ứng với tám nhóm con D_{ij} . Số lượng bản ghi trong mỗi nhóm được cho trong Bảng 3.1.

Bảng 3.1: Bảng xác định nhóm con của DNU

Nhóm	Ký hiệu	Số lượng
Nam – Thành phố – 0	D_{10}	14
Nam – Thành phố – 1	D_{11}	120
Nam – Khác – 0	D_{20}	41
Nam – Khác – 1	D_{21}	202
Nữ – Thành phố – 0	D_{30}	1
Nữ – Thành phố – 1	D_{31}	16
Nữ – Khác – 0	D_{40}	1
Nữ – Khác – 1	D_{41}	31

Đầu tiên ta rà soát các nhóm đặc biệt (có số lượng nhỏ hơn 2 bản ghi), ta thấy nhóm D_{30} và D_{40} đều nhận giá trị nhỏ nhất bằng 1 (< 2 nên sẽ được giữ nguyên. Đồng thời với đó, các nhóm ghép cặp tương ứng với chúng cũng sẽ đc giữ nguyên, bao gồm D_{31} và D_{41} .

Tiếp theo, trong số các nhóm con còn lại, nhóm D_{10} có số lượng nhỏ nhất và lớn hơn 2 là 14 nên sẽ chọn nhóm này là nhóm gốc, sau đó chọn $D'_{10} = 14 \times 20 = 280$.

Bước 2: Cân bằng dữ liệu nhờ sinh dữ liệu tổng hợp

Tính tỉ lệ trung bình R_{mean} giữa số lượng mẫu có nhãn $i1$ và nhãn $i0$ (khác 0) với $i = \overline{1,4}$:

$$R_{mean} = \frac{1}{4} \left(\frac{120}{14} + \frac{202}{41} + \frac{16}{1} + \frac{31}{1} \right) \approx 15$$

Từ nhóm chuẩn D_{10} với $j_s = 0$, tính các D_{ij} còn lại theo công thức:

$$\begin{cases} \|D'_{i0}\| = \frac{\|D_{i0}\|}{\|D_{i_s j_s}\|} \cdot n_0 \\ \|D'_{i1}\| = R_{mean} \cdot \|D'_{i0}\| \end{cases}$$

Kết quả các tập D'_{ij} được cho trong Bảng 3.2.

Bảng 3.2: Bảng xác định nhóm con sau khi sinh dữ liệu của DNU

Nhóm	Ký hiệu	SL ban đầu	SL sau khi sinh
Nam – Thành phố – 0	D_{10}	14	280
Nam – Thành phố – 1	D_{11}	120	4200
Nam – Khác – 0	D_{20}	41	820
Nam – Khác – 1	D_{21}	202	12300
Nữ – Thành phố – 0	D_{30}	1	1
Nữ – Thành phố – 1	D_{31}	16	16
Nữ – Khác – 0	D_{40}	1	1
Nữ – Khác – 1	D_{41}	31	31

Bước 3: Huân luyện và kiểm tra

Sau khi sinh dữ liệu tổng hợp cho mỗi D_{ij} để đạt kích thước mục tiêu $\|D'_{ij}\|$, các nhóm được hợp nhất thành tập huấn luyện D' . Huân luyện qua các mô hình học máy Hồi quy logistic, Rừng ngẫu nhiên, v.v.

Bước 4: Đánh giá mô hình. Đánh giá mô hình theo các chỉ số công bằng “tác động khác biệt”, “hiệu số chênh lệch thống kê”, “chênh lệch trung bình xác suất”, “chênh lệch cơ hội công bằng” và các chỉ số hiệu suất “độ chuẩn xác”, “độ hồi tưởng”.

Kết luận: Qua ví dụ này, có thể thấy dữ liệu ban đầu từ bộ DNU bị mất cân bằng nghiêm trọng giữa các nhóm giao thoa các thuộc tính nhạy cảm, đặc biệt

là các nhóm nữ và vùng nông thôn. Phương pháp DPF giúp tái cân bằng các nhóm này một cách có kiểm soát, qua đó tăng cường tính công bằng cho mô hình học máy ngay từ khâu trước quá trình mà không làm giảm hiệu suất dự đoán.

3.3. Thực nghiệm

Phần này trình bày quá trình thực nghiệm nhằm làm rõ hiệu quả của phương pháp DPF, đã được giới thiệu ở Mục 3.2. Các thí nghiệm được thiết kế để so sánh DPF với các phương pháp sinh dữ liệu tổng hợp trong bối cảnh dữ liệu giáo dục có chứa nhiều thuộc tính nhạy cảm. Nội dung thực nghiệm bao gồm mô tả bộ dữ liệu, lựa chọn mô hình, cấu hình triển khai và các chỉ số đánh giá.

3.3.1. Dữ liệu

Để triển khai thực nghiệm nhằm đánh giá hiệu quả của phương pháp DPF trong việc tăng cường tính công bằng của các mô hình học máy trong lĩnh vực giáo dục, nghiên cứu này sử dụng bốn bộ dữ liệu thực tế đại diện cho các hệ thống giáo dục và bối cảnh xã hội khác nhau. Việc lựa chọn các bộ dữ liệu được thực hiện dựa trên ba tiêu chí chính: (i) có liên quan trực tiếp đến các bài toán dự đoán kết quả học tập, (ii) bao gồm từ hai thuộc tính nhạy cảm trở lên – chẳng hạn như *giới tính*, *chủng tộc*, *tình trạng khuyết tật*, v.v., và (iii) đã được sử dụng trong các nghiên cứu học máy có xét đến yếu tố công bằng. Việc đáp ứng đồng thời ba tiêu chí này bảo đảm tính đa dạng và tính đại diện cao cho tập dữ liệu thử nghiệm, từ đó hỗ trợ phân tích toàn diện các khía cạnh liên quan đến công bằng trong mô hình hóa.

Bốn bộ dữ liệu sử dụng trong nghiên cứu này đã được trình bày chi tiết trong Mục 2.3.1, bao gồm: *Student Performance*, *Student Predict Dropout*, *Oulad*, và *DNU Data*.

Để phục vụ cho việc phân tích công bằng ở cấp độ giao nhau giữa các thuộc tính nhạy cảm, mỗi bộ dữ liệu được chia thành các nhóm con dựa trên các tổ

hợp giữa thuộc tính nhạy cảm và biến mục tiêu. Cụ thể, ba bộ dữ liệu *Student Performance*, *Student Predict Dropout*, và *Oulad*, mỗi bộ được phân thành 8 nhóm con, trong khi bộ *DNU Data* được chia thành 16 nhóm do có ba thuộc tính nhạy cảm. Thông tin chi tiết về thành phần và quy mô của các nhóm này được trình bày trong Bảng 3.3.

3.3.2. Lựa chọn mô hình học máy và kỹ thuật sinh dữ liệu tổng hợp

Trong nghiên cứu này, các thí nghiệm được tiến hành với năm mô hình học máy phổ biến thường được sử dụng trong các bài toán dự đoán trong giáo dục, bao gồm: *Hồi quy logistic*, *Rừng ngẫu nhiên*, *Cây quyết định*, *Tăng cường gradient*, và *Mạng nơ ron thần kinh* [132, 133, 134]. Các mô hình này được lựa chọn nhằm đảm bảo sự đa dạng về kiến trúc, mức độ phức tạp và khả năng học biểu diễn, từ đó giúp đánh giá toàn diện hiệu quả của các chiến lược can thiệp công bằng.

- *Hồi quy logistic*: Đây là một mô hình thống kê thường được áp dụng cho các bài toán phân loại nhị phân, trong đó biến đầu ra chỉ nhận một trong hai giá trị. Trong thực nghiệm, mô hình được huấn luyện với chuẩn hoá L2 để giảm thiểu hiện tượng quá khớp. Quá trình tối ưu sử dụng bộ giải *liblinear* – một thuật toán chuyên biệt cho các mô hình tuyến tính, có ưu điểm nhanh, tiết kiệm bộ nhớ và phù hợp với dữ liệu có số chiều lớn hoặc thưa. Số vòng lặp huấn luyện được giới hạn tối đa là 100 [21].
- *Cây quyết định*: Là mô hình đơn giản và dễ giải thích, được ứng dụng rộng rãi cho cả bài toán phân loại và hồi quy. Mô hình xây dựng một cấu trúc cây, trong đó mỗi nút trong biểu diễn một điều kiện phân tách theo thuộc tính, các nhánh biểu diễn kết quả của điều kiện, và lá cây biểu diễn nhãn dự đoán [135]. Trong cấu hình thí nghiệm, mô hình được giới hạn độ sâu tối đa là 3 và sử dụng chỉ số Gini để đo độ tinh khiết khi phân tách.
- *Rừng ngẫu nhiên*: Đây là một phương pháp học tổ hợp được xây dựng bằng cách kết hợp nhiều cây quyết định, trong đó mỗi cây được huấn luyện trên một tập con ngẫu nhiên của dữ liệu và tập thuộc tính. Cách tiếp cận này

Bảng 3.3: Tổng quan chi tiết về các bộ dữ liệu và phân phối nhóm giao nhau

STT	Bộ dữ liệu	Số TT	K.Thước	TT nhạy cảm		Kết quả	S.lượng	
				<i>G.tính</i>	<i>Sức khỏe</i>	<i>Kết quả</i>		
1	Student Performance	33	1.042	Nam	Rất tốt	Đạt	130	
				Nam	Rất tốt	Ko đạt	146	
				Nam	Khác	Đạt	79	
				Nam	Khác	Ko đạt	98	
				Nữ	Rất tốt	Đạt	139	
				Nữ	Rất tốt	Ko đạt	154	
				Nữ	Khác	Đạt	162	
				Nữ	Khác	Ko đạt	134	
2	Student Dropout Prediction	35	4.424	<i>G.tính</i>	<i>TT Nợ</i>	<i>Tốt nghiệp</i>	<i>S.lượng</i>	
				Nam	Không	Tốt nghiệp	526	
				Nam	Không	Bỏ học	817	
				Nam	Nợ	Tốt nghiệp	22	
				Nam	Nợ	Bỏ học	191	
				Nữ	Không	Tốt nghiệp	1582	
				Nữ	Không	Bỏ học	996	
				Nữ	Nợ	Tốt nghiệp	79	
Nữ	Nợ	Bỏ học	211					
3	OULAD	12	31.482	<i>G.tính</i>	<i>Khuyết tật</i>	<i>Kết quả</i>	<i>S.lượng</i>	
				Nam	Có	Đạt	522	
				Nam	Có	Ko đạt	995	
				Nam	Không	Đạt	7205	
				Nam	Không	Ko đạt	8345	
				Nữ	Có	Đạt	665	
				Nữ	Có	Ko đạt	949	
				Nữ	Không	Đạt	6263	
Nữ	Không	Ko đạt	6538					
4	DNU	11	426	<i>G.tính</i>	<i>Tuổi</i>	<i>N.Sinh</i>	<i>Kết quả</i>	<i>S.lượng</i>
				Nam	Dưới	T.thị	Đạt	87
				Nam	Dưới	T.thị	Ko đạt	11
				Nam	Dưới	Khác	Đạt	144
				Nam	Dưới	Khác	Ko đạt	17
				Nam	Trên	T.thị	Đạt	33
				Nam	Trên	T.thị	Ko đạt	3
				Nam	Trên	Khác	Đạt	58
				Nam	Trên	Khác	Ko đạt	24
				Nữ	Dưới	T.thị	Đạt	14
				Nữ	Dưới	T.thị	Ko đạt	1
				Nữ	Dưới	Khác	Đạt	21
				Nữ	Dưới	Khác	Ko đạt	1
				Nữ	Trên	T.thị	Đạt	2
Nữ	Trên	T.thị	Ko đạt	0				
Nữ	Trên	Khác	Đạt	10				
Nữ	Trên	Khác	Ko đạt	0				

giúp giảm hiện tượng quá khớp và tăng khả năng tổng quát hóa của mô hình [136]. Trong nghiên cứu này, rừng ngẫu nhiên gồm 100 cây, mỗi cây có độ sâu tối đa là 3, và sử dụng chỉ số Gini để đo độ tinh khiết trong quá trình phân tách.

- *Tăng cường gradient*: Đây là một họ các phương pháp học tổ hợp, trong đó mô hình mạnh được xây dựng bằng cách kết hợp tuần tự nhiều mô hình yếu – thường là cây quyết định. Mỗi mô hình kế tiếp tập trung vào việc khắc phục sai số do mô hình trước tạo ra [146]. Trong nghiên cứu này, cấu hình sử dụng là tăng cường theo Gradient với 100 bộ ước lượng, tốc độ học bằng 1 và độ sâu tối đa của cây là 3.
- *Mạng nơ ron thân kinh*: Đây là mô hình có khả năng học và biểu diễn các mối quan hệ phức tạp, phi tuyến tính trong dữ liệu, nhờ đó thường cho hiệu quả cao đối với các tập dữ liệu có không gian đặc trưng nhiều chiều, chẳng hạn dữ liệu giáo dục [147]. Trong nghiên cứu này, chúng tôi sử dụng mạng nơ-ron truyền thẳng gồm một lớp ẩn, với số lượng nút trong lớp ẩn bằng một nửa số thuộc tính đầu vào. Hàm kích hoạt được sử dụng là ReLU, giúp mô hình học hiệu quả hơn trên dữ liệu có tính phi tuyến.

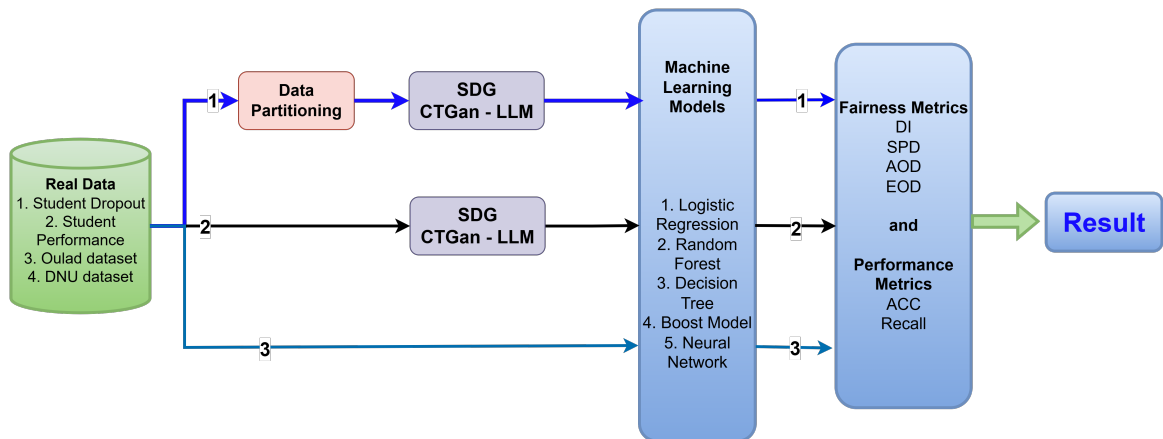
Ngoài các mô hình học máy, nghiên cứu này còn sử dụng hai kỹ thuật sinh dữ liệu tổng hợp nhằm cân bằng dữ liệu theo nhóm giao thoa của các thuộc tính nhạy cảm, bao gồm: *CTGAN* và *LLM*.

- *CTGAN*: là một biến thể của mạng sinh đối kháng có điều kiện, được thiết kế chuyên biệt cho dữ liệu dạng bảng có chứa cả biến rời rạc và biến liên tục [148]. Phương pháp này áp dụng cơ chế chuẩn hoá theo từng loại dữ liệu và lấy mẫu có điều kiện, nhờ đó mô hình hoá tốt hơn phân phối của dữ liệu, đặc biệt trong trường hợp mất cân bằng giữa các nhóm. Trong thí nghiệm, CTGAN được triển khai với các tham số: 10 vòng huấn luyện, kích thước lô dữ liệu là 500, tốc độ học 2×10^{-4} cho cả bộ sinh và bộ phân biệt, cùng với kích thước véc tơ nhúng là 128.
- *LLM*: là kỹ thuật sử dụng mô hình ngôn ngữ dung lượng lớn (trong nghiên cứu này là ChatGPT 3.5) để sinh dữ liệu mới dựa trên lược đồ và các đặc trưng thống kê của dữ liệu gốc [103]. Quá trình này được điều khiển bằng

bản hướng dẫn sinh (prompt) được thiết kế riêng cho từng tập dữ liệu, giúp phản ánh chính xác cấu trúc bảng và các ràng buộc giá trị. Trong thí nghiệm, mô hình được cấu hình với hệ số ngẫu nhiên 0.7, tham số chọn lọc hạt nhân 0.9 và độ dài đầu ra tối đa 512 từ. Sau khi sinh, dữ liệu được xử lý hậu kỳ để bảo đảm định dạng thống nhất và loại bỏ những bản ghi không hợp lệ.

3.3.3. Thiết lập thực nghiệm

Để đánh giá hiệu quả của phương pháp đề xuất trong việc tăng cường công bằng cho các mô hình học máy, nghiên cứu này thiết lập ba cấu hình thực nghiệm chính, được minh họa trong Hình 3.2. Mỗi cấu hình phản ánh một chiến lược xử lý dữ liệu khác nhau, từ không can thiệp cho đến can thiệp kép bằng phân vùng dữ liệu và sinh dữ liệu tổng hợp theo nhóm con.



Hình 3.2: Tổng quan ba cấu hình thực nghiệm đánh giá hiệu quả DPFA

– *Cấu hình 1: DPFA (Mô hình đề xuất)*

Đây là cấu hình đại diện cho phương pháp đề xuất. Dữ liệu thực được phân chia thành các nhóm con dựa trên tổ hợp của các thuộc tính nhạy cảm. Sau đó, dữ liệu tổng hợp được tạo riêng cho từng nhóm bằng kỹ thuật SDG (CTGAN hoặc LLM), đảm bảo sự đa dạng và cân bằng trong mỗi nhóm nhỏ. Dữ liệu tổng hợp toàn cục sau đó được sử dụng để huấn luyện các mô hình học máy và đánh giá theo cả chỉ số công bằng và hiệu suất.

– *Cấu hình 2: SDG (Mô hình sinh dữ liệu không phân nhóm)*

Dữ liệu thực được sử dụng để sinh dữ liệu tổng hợp bằng CTGAN hoặc LLM mà không thực hiện phân nhóm trước đó. Dữ liệu tổng hợp được huấn luyện trực tiếp trên các mô hình học máy. Cấu hình này nhằm đánh giá tác động của SDG mà không áp dụng kỹ thuật phân chia nhóm con theo thuộc tính nhạy cảm.

– *Cấu hình 3: Origin (Mô hình gốc, không can thiệp)*

Đây là cấu hình cơ sở, trong đó dữ liệu thực được sử dụng nguyên trạng mà không áp dụng bất kỳ kỹ thuật can thiệp nào (không sinh dữ liệu tổng hợp, không phân nhóm, không điều chỉnh công bằng). Cấu hình này giúp làm nền để so sánh mức độ cải thiện khi áp dụng các kỹ thuật công bằng.

Trong cả ba cấu hình thực nghiệm, nghiên cứu sử dụng bốn bộ dữ liệu giáo dục gồm *Student Performance*, *Student Predict Dropout*, *Oulad* và *DNU Data*. Trên mỗi bộ dữ liệu, năm mô hình học máy được triển khai bao gồm *Hồi quy logistic*, *Rừng ngẫu nhiên*, *Cây quyết định*, *Tăng cường gradient* và *Mạng nơ ron thần kinh*. Kết quả của các mô hình được đánh giá đồng thời theo hai nhóm tiêu chí. Nhóm chỉ số công bằng bao gồm “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*” và “*chênh lệch cơ hội công bằng*”. Nhóm chỉ số hiệu suất bao gồm “*độ chuẩn xác*” và “*độ hồi tưởng*”.

3.3.4. Chỉ số đánh giá

Các chỉ số công bằng và hiệu suất sử dụng trong chương này được kế thừa từ Mục 2.3.4 trong Chương 4, nhằm đảm bảo tính nhất quán trong toàn bộ nghiên cứu. Cụ thể, bốn chỉ số công bằng gồm “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*”, và “*chênh lệch cơ hội công bằng*”, cùng hai chỉ số hiệu suất gồm “*độ chuẩn xác*” và “*độ hồi tưởng*”.

3.4. Kết quả thực nghiệm

Phần này trình bày và phân tích các kết quả thực nghiệm nhằm đánh giá hiệu quả của phương pháp DPF trong việc cải thiện tính công bằng cho các mô hình học máy sử dụng dữ liệu giáo dục dạng bảng có chứa nhiều thuộc tính nhạy cảm. Các kết quả được tổ chức theo ba khía cạnh phân tích chính, phản ánh các mục tiêu thực nghiệm đã được xác định trong chương.

Trước khi phân tích các kết quả công bằng trong Mục 3.4, nhằm đánh giá độ ổn định của phương pháp, toàn bộ quy trình thực nghiệm bao gồm cả hai kỹ thuật sinh dữ liệu tổng hợp (SDG) là CTGAN và LLM được lặp lại ba lần một cách ngẫu nhiên và độc lập. Bảng 3.4 trình bày chi tiết kết quả đối với mô hình Hồi quy Logistic theo chỉ số DI sử dụng cả hai kỹ thuật CTGAN và LLM. Toàn bộ các kết quả sử dụng kỹ thuật CTGAN được trực quan hóa bằng biểu đồ đường trong Bảng 3.5, thể hiện mức độ biến thiên theo từng bộ dữ liệu, mô hình và thước đo.

Kết quả cho thấy sự khác biệt giữa các lần chạy là rất nhỏ trên hầu hết các cấu hình, phản ánh tính ổn định cao của phương pháp. Đặc biệt, với cách tiếp cận dựa trên LLM, nhiều trường hợp đạt phương sai bằng không ($\sigma = 0.000$), cho thấy khả năng tái lập hoàn toàn. Trong khi đó, CTGAN có mức biến thiên lớn hơn nhưng vẫn duy trì ở mức thấp, đảm bảo tính nhất quán qua các lần chạy. Những kết quả này cung cấp bằng chứng thực nghiệm quan trọng, khẳng định độ tin cậy và tính ổn định của các cải thiện công bằng đạt được bởi phương pháp DPF.

3.4.1. Đặc điểm của dữ liệu tổng hợp trong lĩnh vực giáo dục

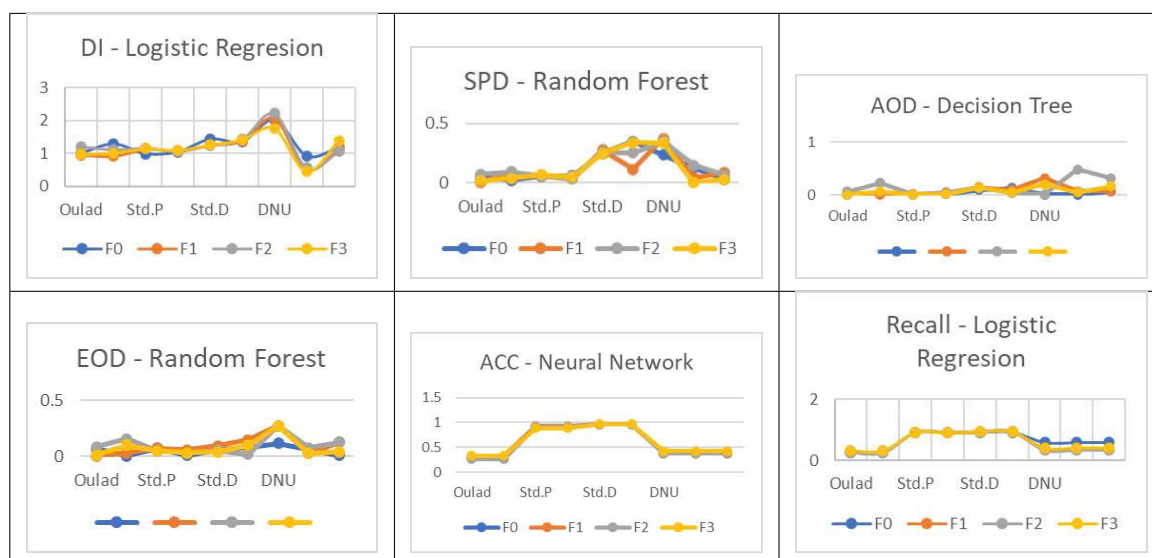
Mục này khảo sát đặc điểm của dữ liệu tổng hợp được sinh ra bởi CTGAN và LLM trong các cấu hình thí nghiệm khác nhau. Bảng 3.3 cho thấy trước khi thực hiện kỹ thuật sinh dữ liệu, hai bộ dữ liệu *Student Performance* và *Oulad* có phân bố tương đối cân bằng giữa các nhóm nhạy cảm, trong khi *Student Predict Dropout* và *DNU Data* thể hiện sự mất cân bằng rõ rệt.

Bảng 3.4: Độ biến thiên của chỉ số DI qua các lần chạy lặp lại (Logistic Regression)

TT nhảy cảm	LLM					CTGAN				
	Lần 1	Lần 2	Lần 3	T.bình	σ	Lần 1	Lần 2	Lần 3	T.bình	σ
O.GT	1.017	1.017	1.017	1.017	0.000	1.012	0.963	1.215	1.063	0.134
O.K.tật	0.956	0.956	0.956	0.956	0.000	1.292	0.929	1.120	1.114	0.182
SP.GT	1.233	1.176	1.185	1.198	0.031	0.989	1.142	1.167	1.099	0.096
SP.SK	1.088	1.069	1.065	1.074	0.012	1.045	1.098	1.073	1.072	0.027
SD.GT	1.569	1.569	1.569	1.569	0.000	1.443	1.259	1.254	1.319	0.108
SD.Nợ	1.442	1.442	1.442	1.442	0.000	1.362	1.390	1.439	1.397	0.039
DNU.GT	0.983	1.735	1.124	1.281	0.400	2.034	2.082	2.213	2.110	0.093
DNU.Tuổi	1.588	0.900	1.313	1.267	0.347	0.923	0.529	0.563	0.672	0.218
DNU.KV	2.051	1.243	2.320	1.871	0.560	1.172	1.140	1.058	1.123	0.059

Ghi chú: Kí hiệu σ là độ lệch chuẩn giữa các lần đo

Bảng 3.5: Minh họa độ biến thiên của các chỉ số công bằng và hiệu suất qua các lần chạy lặp lại sử dụng kỹ thuật CTGAN



Ghi chú: Trong bảng này, lần thực nghiệm được hiển thị với màu sắc khác nhau: F0 ứng với kết quả trung bình các lần chạy; F1, F2, F3 lần lượt hiển thị kết quả của lần chạy thứ nhất, thứ hai và thứ ba theo các chỉ số tương ứng.

Khi phân tách, mỗi bộ dữ liệu với hai thuộc tính nhảy cảm (*Student Performance*, *Oulad*, và *Student Predict Dropout*) được chia thành 8 nhóm con, trong khi bộ dữ liệu *DNU Data* với ba thuộc tính nhảy cảm được chia thành 16 nhóm con. Thuật toán DPF dựa trên tỷ lệ đầu ra thuận lợi và bất lợi trong từng nhóm giao thoa để xác định số lượng mẫu tổng hợp cần sinh cho mỗi nhóm con nhằm đạt được sự cân bằng tương đối giữa các nhóm con nhảy cảm.

Kết quả sau khi triển khai sinh dữ liệu theo chiến lược phân tách DPF cho thấy sự cải thiện rõ rệt về mức độ cân bằng tỷ lệ giữa các nhóm con. Bảng 3.6

Bảng 3.6: Minh họa tổng quan về các tập dữ liệu tổng hợp sử dụng kỹ thuật phân nhóm



minh họa trực quan kết quả này, trong đó các cột màu xanh biểu thị số lượng dữ liệu gốc của từng nhóm con, các cột màu hồng biểu thị số lượng dữ liệu được sinh bổ sung, và tổng chiều cao của mỗi cột phản ánh quy mô dữ liệu cuối cùng của nhóm sau khi áp dụng sinh dữ liệu tổng hợp. Có thể quan sát rằng, sau khi sinh dữ liệu, tỷ lệ quy mô giữa các nhóm con trong cùng một cặp trở nên tương đồng hơn, qua đó giúp giảm đáng kể sự chênh lệch về phân bố dữ liệu giữa các nhóm nhạy cảm. Chi tiết số lượng mẫu được sinh cho từng nhóm con theo từng bộ dữ liệu và từng phương pháp sinh dữ liệu được trình bày đầy đủ trong Bảng C.1 tại Phụ lục C.

Tuy nhiên, với các bộ dữ liệu thưa như *DNU Data*, một số nhóm con không chứa bất kỳ mẫu nào trong dữ liệu gốc, hoặc chỉ chứa một bản ghi duy nhất, khi đó dữ liệu không có hoặc quá ít thông tin để sinh thêm dẫn đến không sinh được hoặc sinh các bản ghi trùng nhau. Nhóm này sẽ được loại bỏ khỏi quá trình cân bằng, chỉ các nhóm con có từ hai bản ghi trở lên mới được áp dụng sinh dữ liệu tổng hợp. Ngoài ra, do chi phí sinh dữ liệu khác nhau giữa các phương pháp như CTGAN là công cụ mã nguồn mở, trong khi LLM tính phí theo token nên số lượng dữ liệu tổng hợp được điều chỉnh một cách linh hoạt: CTGAN được sử dụng để sinh đủ số lượng mẫu theo mục tiêu cân bằng, trong khi LLM chỉ được áp dụng có chọn lọc nhằm đảm bảo tính khả thi về chi phí.

Kết quả thực nghiệm cho thấy phương pháp sinh dữ liệu tổng hợp dựa trên phân tách nhóm giúp điều chỉnh đáng kể phân bố dữ liệu giữa các nhóm giao thoa, đặc biệt trên hai bộ dữ liệu mất cân bằng là *Student Predict Dropout* và *DNU Data*. Ví dụ, trong bộ *Student Predict Dropout*, nhóm *Nữ-Không nợ-Chưa Tốt nghiệp* tăng từ 996 lên 10.972 mẫu khi sử dụng CTGAN và 10.126 mẫu khi sử dụng LLM, trong khi nhóm *Nam-Nợ-Tốt nghiệp* giảm mạnh từ 191 xuống 13 (CTGAN) và 0 (LLM). Trong bộ *DNU Data*, nhóm *Nam-Quá tuổi-Khác-Tốt nghiệp* tăng từ 58 lên 525 mẫu, trong khi nhóm *Nam-Đúng tuổi-Thành phố-Chưa tốt nghiệp* gần như không thay đổi do số lượng ban đầu quá nhỏ. Những kết quả này cho thấy FaireduPlus có khả năng cải thiện đáng kể tính đại diện giữa các nhóm giao thoa, dù vẫn còn tồn tại hạn chế trong các trường hợp dữ liệu quá thưa.

Với các cấu hình không áp dụng chiến lược phân tách (SDG), dữ liệu tổng hợp được sinh trực tiếp từ toàn bộ tập dữ liệu gốc mà không chia thành các nhóm con giao thoa. Trong các thiết lập này, CTGAN và LLM được sử dụng nhằm mở rộng quy mô tập dữ liệu tổng thể, mà không nhắm trực tiếp đến việc cân bằng phân phối ở mức nhóm. Tuy nhiên, để đảm bảo tính công bằng trong so sánh thực nghiệm, tổng số lượng mẫu dữ liệu tổng hợp được sinh ra vẫn được kiểm soát và giữ tương đương với các cấu hình áp dụng kỹ thuật phân tách.

Sự khác biệt giữa cách tiếp cận sinh dữ liệu dựa trên phân tách nhóm và sinh dữ liệu trực tiếp từ tập dữ liệu tổng thể dẫn đến những thay đổi đáng kể trong phân bố dữ liệu tổng hợp và tác động khác nhau đến các chỉ số công bằng. Những ảnh hưởng này sẽ được phân tích chi tiết hơn trong các mục tiếp theo.

3.4.2. Hiệu quả của phương pháp DPF trong cải thiện công bằng giao thoa

Để đánh giá hiệu quả của phương pháp DPF so với các cấu hình đối sánh, luận án tiến hành so sánh các chỉ số công bằng (“*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*”, và “*chênh lệch cơ hội công bằng*”) của các cấu hình. Các thí nghiệm được tiến hành theo ba cấu hình DPF, SDG và Origin đã trình bày trong Mục 3.3.3.

Tổng cộng, luận án đã thực hiện thí nghiệm trên năm mô hình học máy, bốn tập dữ liệu cho ba cấu hình với hai kỹ thuật sinh dữ liệu tổng hợp (CTGAN và LLM). Các chỉ số thu được từ các thí nghiệm này là cơ sở cho các phân tích chi tiết nhằm khẳng định tính hiệu quả của DPF trong việc cải thiện tính công bằng.

3.4.2.1. Đánh giá thông qua các chỉ số công bằng

Để đánh giá tính hiệu quả của DPF qua chỉ số công bằng, các chỉ số công bằng của các cấu hình đề xuất được so sánh theo từng thuộc tính nhạy cảm tương ứng với từng bộ dữ liệu. Các chỉ số $|1 - DI|$, SPD , AOD , EOD có giá trị càng nhỏ thì càng tốt. Cấu hình nào đạt chỉ số tương ứng nhỏ nhất được xem là tốt nhất về mặt công bằng theo chỉ số đó, đồng thời cấu hình đó được ghi nhận là “thắng” các cấu hình còn lại. Chi tiết kết quả các chỉ số công bằng được cho trong Bảng C.2 tại Phụ lục C. Tổng số lần “thắng” trên toàn bộ thí nghiệm theo từng chỉ số được tổng hợp trong Bảng 3.7.

Bảng 3.7: Tổng hợp số lượt thắng về chỉ số công bằng theo từng cấu hình

Cấuhình	CTGAN				LLM			
	1-DI	SPD	AOD	EOD	1-DI	SPD	AOD	EOD
DPF	11	14	22	17	19	19	14	16
SDG	23	21	14	15	10	10	2	5
Origin	11	10	9	13	16	16	14	9

Kết quả cho thấy sự khác biệt đáng kể giữa các cấu hình và ảnh hưởng rõ rệt của kỹ thuật sinh dữ liệu tổng hợp đến hiệu quả công bằng. Với CTGAN, cấu hình SDG đạt số lượt thắng cao nhất đối với hai chỉ số phân phối $|1 - DI|$ và SPD , lần lượt là 23 và 21 lần, trong khi DPF chỉ đạt 11 và 14 lần. Ngược lại, DPF lại chiếm ưu thế ở các chỉ số sai số với 22 lần thắng đối với AOD và 17 lần đối với EOD , cao hơn SDG (14 và 15 lần). Điều này cho thấy SDG chủ yếu cải thiện sự cân bằng về phân phối giữa các nhóm, trong khi DPF hiệu quả hơn trong việc giảm chênh lệch sai số dự đoán giữa các nhóm nhạy cảm.

Đối với cấu hình sử dụng LLM, DPF đạt số lượt thắng cao nhất ở cả bốn chỉ số công bằng, gồm 19 lần đối với $|1 - DI|$, 19 lần đối với SPD , 14 lần đối với AOD và 16 lần đối với EOD . Trong khi đó, SDG chỉ đạt lần lượt 10, 10, 2 và 5

lần thắng, còn cấu hình gốc (Origin) đạt 16, 16, 14 và 9 lần. Kết quả này cho thấy việc kết hợp DPF với LLM giúp cải thiện đồng thời các chỉ số công bằng về phân phối và sai số, đồng thời mang lại hiệu quả ổn định hơn so với các cấu hình còn lại.

Tuy nhiên, DPF vẫn chưa đạt kết quả tốt nhất trong mọi trường hợp. Hiệu quả của phương pháp còn phụ thuộc vào đặc điểm của bộ dữ liệu, mức độ mất cân bằng giữa các nhóm giao thoa và chất lượng dữ liệu tổng hợp được sinh ra. Đặc biệt, khi số lượng mẫu của một số nhóm giao thoa quá nhỏ, chất lượng dữ liệu sinh có thể bị ảnh hưởng, từ đó làm giảm mức cải thiện công bằng của phương pháp.

3.4.2.2. Ảnh hưởng của mô hình học máy

Để phân tích sự thay đổi về mức độ công bằng giữa các mô hình học máy, tương tự theo cách thống kê tương tự như mục 3.4.2.1, luận án thống kê số lần mỗi mô hình đạt kết quả công bằng tốt nhất (số lượt thắng) đối với bốn chỉ số “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*” và “*chênh lệch cơ hội công bằng*” khi áp dụng hai phương pháp sinh dữ liệu tổng hợp (CTGAN và LLM). Bảng 3.8 tổng hợp số lần mỗi mô hình học máy đạt kết quả công bằng tốt nhất theo từng chỉ số và phương pháp sinh dữ liệu tổng hợp.

Bảng 3.8: Tổng hợp số lần thắng về kết quả công bằng theo từng mô hình

Chỉ số		1-DI	SPD	AOD	EOD
Mô hình	Phương pháp				
LR	CTGAN	3	5	5	2
	LLM	4	4	4	5
RF	CTGAN	1	2	5	3
	LLM	3	2	3	4
DT	CTGAN	3	3	4	6
	LLM	5	6	3	2
BM	CTGAN	3	2	4	4
	LLM	3	3	1	1
NN	CTGAN	2	2	3	3
	LLM	4	4	3	4

Đối với *Hồi quy logistic*, hai phương pháp sinh dữ liệu đều mang lại kết quả khả quan nhưng có xu hướng khác nhau. Với CTGAN, mô hình đạt số lần thắng cao ở “*hiệu số chênh lệch thống kê*” và “*chênh lệch trung bình xác suất*” (đều 5 lần), trong khi LLM cho kết quả đồng đều hơn trên cả bốn chỉ số công bằng (“*tác động khác biệt*”=4, “*hiệu số chênh lệch thống kê*”=4, “*chênh lệch trung bình xác suất*”=4 và “*chênh lệch cơ hội công bằng*”=5). Điều này cho thấy LLM giúp Logistic Regression duy trì sự cân bằng tốt hơn giữa các khía cạnh công bằng.

Đối với *Rừng ngẫu nhiên* và *Cây quyết định*, ảnh hưởng của phương pháp sinh dữ liệu thể hiện rõ hơn. Với RF, CTGAN đạt nhiều lần thắng nhất ở “*chênh lệch trung bình xác suất*” (5 lần), trong khi LLM cải thiện các chỉ số “*tác động khác biệt*” và “*chênh lệch cơ hội công bằng*”. Đối với DT, CTGAN nổi bật ở “*chênh lệch cơ hội công bằng*” (6 lần), còn LLM lại đạt số lần thắng cao hơn ở “*tác động khác biệt*” (5 lần) và “*hiệu số chênh lệch thống kê*” (6 lần). Kết quả này cho thấy cùng một mô hình nhưng các phương pháp sinh dữ liệu khác nhau có thể tối ưu các khía cạnh công bằng khác nhau.

Đối với *Tăng cường gradient* và *Mạng nơ ron thần kinh*, CTGAN mang lại kết quả tương đối ổn định, trong khi LLM cải thiện rõ hơn ở các chỉ số “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*” và “*chênh lệch cơ hội công bằng*”. Đặc biệt, Neural Network tăng số lần thắng từ 2 lên 4 đối với cả “*tác động khác biệt*” và “*hiệu số chênh lệch thống kê*”, đồng thời tăng từ 3 lên 4 đối với “*chênh lệch cơ hội công bằng*”, cho thấy LLM hỗ trợ tốt hơn đối với các mô hình có khả năng học biểu diễn phi tuyến.

Tổng hợp các kết quả cho thấy hiệu quả cải thiện công bằng không chỉ phụ thuộc vào phương pháp sinh dữ liệu mà còn phụ thuộc vào đặc điểm của từng mô hình học máy. LLM có xu hướng mang lại kết quả ổn định và đồng đều hơn trên nhiều chỉ số công bằng, đặc biệt đối với Logistic Regression và Neural Network. Trong khi đó, CTGAN vẫn phát huy ưu thế ở một số mô hình như Decision Tree và Gradient Boosting, nhất là đối với các chỉ số dựa trên sai số (AOD và EOD). Điều này cho thấy không tồn tại một kỹ thuật sinh dữ liệu tối ưu cho mọi mô hình, mà việc lựa chọn CTGAN hay LLM cần được cân nhắc theo đặc điểm của mô hình học máy và mục tiêu công bằng cần ưu tiên.

3.4.2.3. Ảnh hưởng của đặc trưng bộ dữ liệu

Bốn biểu đồ heatmap trong Bảng 3.9 minh họa giá trị của bốn chỉ số công bằng (“*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*”, và “*chênh lệch cơ hội công bằng*”) cho từng mô hình và thuộc tính nhạy cảm trên bốn bộ dữ liệu *Oulad*, *Student Performance*, *Student Predict Dropout*, và *DNU Data* sau khi áp dụng phương pháp DPF. Màu sắc càng sáng thể hiện giá trị chỉ số càng cao, tức là mức độ công bằng càng thấp. Màu tối hơn biểu thị giá trị nhỏ, nghĩa là mô hình công bằng hơn theo chỉ số đó. Các ô màu trắng phản ánh giá trị không có sẵn (NaN), thường xảy ra khi chỉ số không thể tính toán do dữ liệu của nhóm quá ít hoặc phân phối không đủ điều kiện để ước lượng đáng tin cậy.

Bộ dữ liệu *Oulad*: Hầu hết các mô hình duy trì được mức công bằng tốt với các giá trị chỉ số thấp. Đáng chú ý, *Hồi quy logistic* với thuộc tính nhạy cảm *giới tính* đạt giá trị rất nhỏ ở cả bốn chỉ số. Tuy nhiên, *Mạng nơ ron thần kinh* với thuộc tính nhạy cảm *tình trạng khuyết tật* có giá trị $|1 - DI|$ cao nhất (≈ 1.295) và EOD lớn, cho thấy phương pháp DPF chưa khắc phục hoàn toàn mất cân bằng ở trường hợp này.

Bộ dữ liệu *Student Performance*: Các giá trị chỉ số nhìn chung thấp, phản ánh hiệu quả của DPF khi dữ liệu ban đầu tương đối cân bằng. Tuy nhiên, *Tăng cường gradient* với *giới tính* có $|1 - DI|$ và SPD cao hơn rõ rệt so với các mô hình khác, cho thấy hiệu quả của DPF có thể phụ thuộc vào loại mô hình.

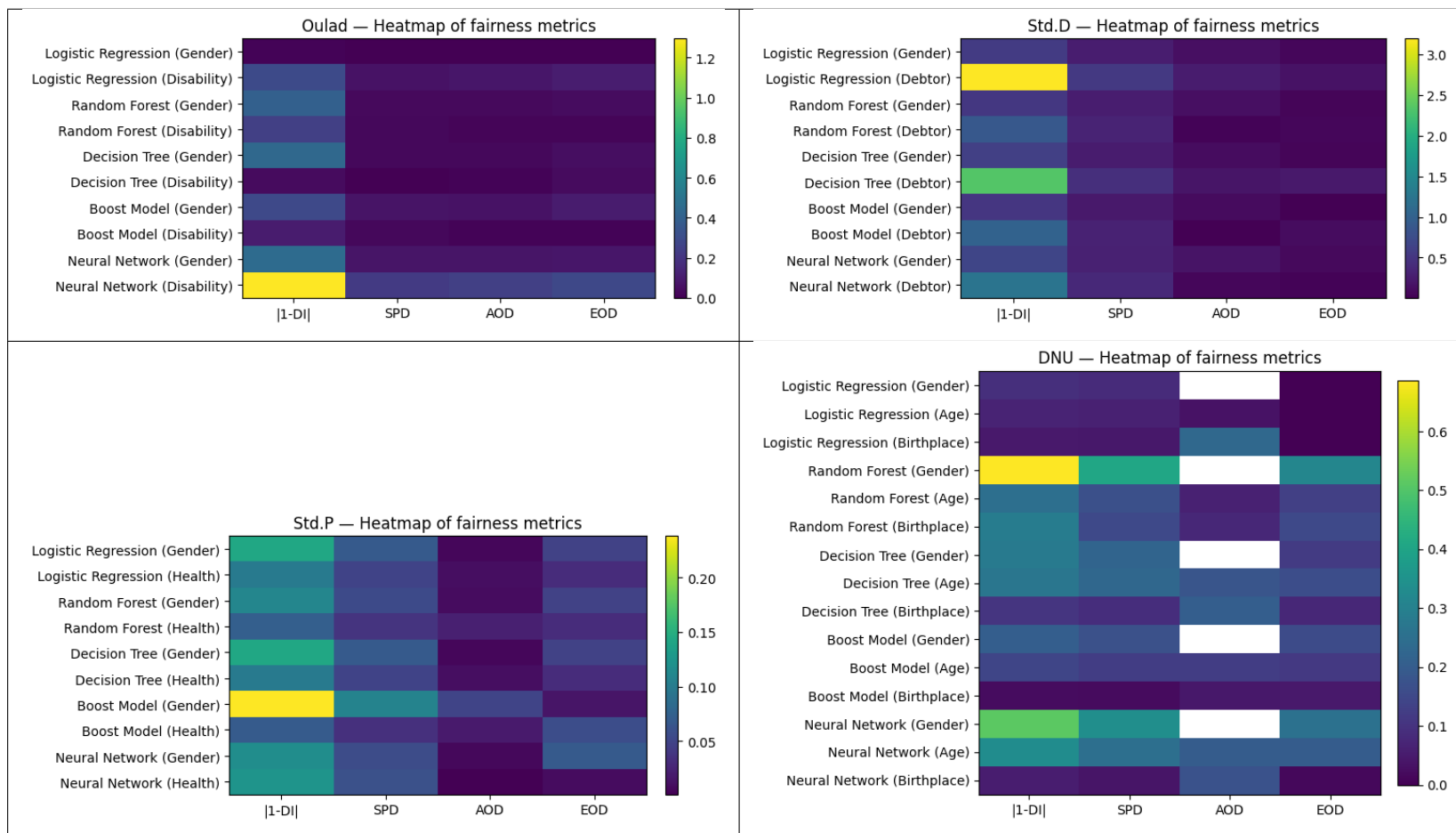
Bộ dữ liệu *Student Predict Dropout*: Mức chênh lệch giữa các nhóm nhạy cảm tăng đáng kể, đặc biệt với thuộc tính *tình trạng nơ*. *Hồi quy logistic* và *Cây quyết định* lần lượt có $|1 - DI|$ rất cao (≈ 3.193 và 2.337), vượt xa các giá trị ở các bộ dữ liệu khác. Điều này cho thấy DPF gặp khó khăn với các bộ dữ liệu có mức mất cân bằng nghiêm trọng và thuộc tính nhạy cảm có ảnh hưởng mạnh đến nhãn.

Bộ dữ liệu *DNU Data*: Các chỉ số công bằng ở mức trung bình, nhưng vẫn có sự khác biệt lớn giữa các mô hình và thuộc tính. *Random Forest* với *giới tính* có $|1 - DI|$ cao nhất (≈ 0.686), kèm SPD và EOD lớn, cho thấy mất công bằng đáng kể. Trong khi đó, *Tăng cường gradient* với *khuvực* có giá trị thấp nhất ở

hầu hết các chỉ số, phản ánh khả năng duy trì công bằng tốt hơn. Một số ô màu trắng (nhất là ở cột AOD) xuất hiện do không đủ dữ liệu hoặc phân phối xác suất đầu ra không phù hợp để tính chỉ số này cho nhóm đó.

Kết quả cho thấy DPF đạt hiệu quả cao trên các bộ dữ liệu có phân phối nhóm tương đối cân bằng (*Oulad* và *Student Performance*), nhưng hiệu quả giảm đáng kể ở các bộ dữ liệu có mất cân bằng lớn (*Student Predict Dropout* và *DNU Data*). Ngoài ra, mức cải thiện công bằng phụ thuộc không chỉ vào đặc tính của bộ dữ liệu mà còn vào loại mô hình học máy và thuộc tính nhạy cảm đang xét.

Bảng 3.9: Biểu đồ heatmap so sánh các chỉ số công bằng của phương pháp DPF trên các bộ dữ liệu thực nghiệm



3.4.3. Tác động của DPF đến hiệu suất dự đoán của mô hình

Để đánh giá tác động của phương pháp DPF lên hiệu suất dự đoán của mô hình, luận án tiến hành so sánh kết quả của ba cấu hình thực nghiệm: (i) Origin – mô hình gốc, huấn luyện trên dữ liệu ban đầu; (ii) DPF – mô hình áp dụng phương pháp đề xuất; và (iii) SDG – mô hình huấn luyện với dữ liệu tổng hợp từ phương pháp SDG. Chi tiết giá trị “độ chuẩn xác” và “độ hồi tưởng” của từng cấu hình đối với các mô hình học máy, bộ dữ liệu và thuộc tính nhạy cảm khác nhau được trình bày trong Bảng C.3 tại Phụ lục C. Các ô được tô xám thể hiện cấu hình có giá trị cao nhất (tức là chiến thắng) trong từng trường hợp so sánh.

Từ dữ liệu ở Bảng C.3, luận án thống kê số lượt thắng của từng cấu hình cho mỗi chỉ số và trình bày trong Bảng 3.10. Kết quả cho thấy cấu hình Origin đạt số lượt thắng cao nhất đối với “độ chuẩn xác” với 26 lần, tiếp theo là DPF với 13 lần và SDG với 6 lần. Ngược lại, đối với “độ hồi tưởng”, SDG chiếm ưu thế với 22 lượt thắng, trong khi Origin và DPF đạt lần lượt 12 và 11 lượt thắng.

Bảng 3.10: Số lượt thắng theo từng cấu hình trên hai tiêu chí ACC và Recall

#	Cấu hình	Số lượt thắng	
		ACC	Recall
1	DPF	13	11
2	SDG	6	22
3	Origin	26	12

Kết quả này cho thấy, về tổng thể, mô hình gốc (Origin) vẫn duy trì được ưu thế về độ chuẩn xác, trong khi phương pháp SDG có xu hướng cải thiện độ hồi tưởng tốt hơn, phản ánh khả năng phát hiện các trường hợp dương tính cao hơn nhưng đồng thời có thể làm giảm “độ chuẩn xác”. Phương pháp DPF thể hiện vai trò cân bằng khi duy trì được số lượt thắng tương đối đồng đều ở cả hai tiêu chí, với 13 lần dẫn đầu về “độ chuẩn xác” và 11 lần về “độ hồi tưởng”. Mặc dù không đạt giá trị cao nhất tuyệt đối ở hầu hết các trường hợp, DPF đạt sự cân bằng hợp lý giữa cải thiện công bằng và suy giảm hiệu suất dự đoán trong phần lớn các cấu hình thực nghiệm. Điều này cho thấy DPF là một lựa chọn khả thi trong các ứng dụng yêu cầu đồng thời duy trì hiệu suất dự đoán và giảm thiểu thiên lệch.

3.5. Thảo luận

Phần này tổng hợp và diễn giải các kết quả thực nghiệm đã trình bày trong Mục 3.4, nhằm làm rõ tác động của phương pháp DPF và các kỹ thuật sinh dữ liệu tổng hợp đối với tính công bằng và hiệu suất của các mô hình học máy trong lĩnh vực giáo dục. Thông qua việc phân tích có hệ thống trên nhiều bộ dữ liệu, nhiều mô hình học máy và các cấu hình sinh dữ liệu khác nhau, phần thảo luận cung cấp cái nhìn toàn diện về hiệu quả, giới hạn và điều kiện áp dụng của phương pháp DPF trong thực tiễn.

Các kết quả cho thấy công bằng trong học máy là một khái niệm đa chiều, chịu ảnh hưởng đồng thời bởi phân bố dữ liệu ban đầu, chiến lược can thiệp dữ liệu và đặc điểm của mô hình học máy. Việc sử dụng dữ liệu tổng hợp không chỉ tác động đến sự đại diện của các nhóm giao thoa trong dữ liệu huấn luyện, mà còn tương tác chặt chẽ với thuật toán sinh dữ liệu và kiến trúc mô hình trong việc định hình kết quả công bằng và hiệu suất. Các phân tích dưới đây làm rõ cách các yếu tố này tác động qua lại, đồng thời làm sáng tỏ những ưu điểm và hạn chế của phương pháp DPF so với các cấu hình sinh dữ liệu và mô hình học máy khác.

3.5.1. Phân tích và tổng hợp các phát hiện chính

Trước hết, kết quả phân tích đặc điểm của dữ liệu tổng hợp cho thấy chiến lược phân chia dữ liệu theo nhóm giao thoa của phương pháp DPF giúp điều chỉnh đáng kể sự phân bố giữa các nhóm nhạy cảm và nhãn đầu ra. Đối với các bộ dữ liệu có mức mất cân bằng cao như *Student Predict Dropout* và *DNU Data*, DPF cải thiện rõ rệt mức độ đại diện của các nhóm yếu thế. Ngược lại, các cách tiếp cận sinh dữ liệu trực tiếp (SDG) không hướng đến cân bằng nhóm giao thoa nên tác động đến phân phối dữ liệu thường hạn chế hơn. Kết quả này cho thấy dữ liệu tổng hợp được sinh theo chiến lược DPF có tiềm năng lớn trong việc giảm thiểu thiên lệch xuất phát từ phân bố dữ liệu gốc.

Tiếp theo, các kết quả thực nghiệm làm rõ hiệu quả của DPF trong việc cải thiện công bằng giao thoa khi so sánh với các cấu hình Origin và SDG. Khi đánh

giá thông qua các chỉ số công bằng như “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*” và “*chênh lệch cơ hội công bằng*”, kết quả cho thấy hiệu quả can thiệp phụ thuộc đáng kể vào thuật toán sinh dữ liệu được sử dụng. Với CTGAN, cấu hình SDG có xu hướng đạt kết quả tốt hơn trên các chỉ số phản ánh phân phối (“*tác động khác biệt*” và “*hiệu số chênh lệch thống kê*”), trong khi DPF thể hiện ưu thế rõ rệt trên các chỉ số sai số (“*chênh lệch trung bình xác suất*” và “*chênh lệch cơ hội công bằng*”). Đối với LLM, DPF cho thấy hiệu quả ổn định và toàn diện hơn khi dẫn đầu trên cả bốn chỉ số, phản ánh khả năng duy trì cân bằng giữa phân phối và sai số tốt hơn so với các cấu hình còn lại. Trong hầu hết các trường hợp, cấu hình Origin cho kết quả công bằng thấp nhất, qua đó khẳng định vai trò quan trọng của can thiệp dữ liệu trong cải thiện công bằng.

Bên cạnh đó, mức độ công bằng đạt được còn phụ thuộc mạnh vào sự kết hợp giữa phương pháp sinh dữ liệu và mô hình học máy. Với dữ liệu tổng hợp sinh bởi LLM, các mô hình như Logistic Regression và Neural Network thường cho mức cải thiện công bằng đồng đều hơn. Ngược lại, với CTGAN, các mô hình như Gradient Boosting hoặc Decision Tree có xu hướng duy trì kết quả ổn định hơn trong một số chỉ số. Đặc biệt, Decision Tree cho thấy sự hoán đổi ưu thế giữa nhóm chỉ số phân phối và nhóm chỉ số sai số khi thay đổi phương pháp sinh dữ liệu, nhấn mạnh tầm quan trọng của việc lựa chọn cấu hình phù hợp với mục tiêu công bằng cụ thể.

Phân tích chi tiết trên từng bộ dữ liệu và mô hình cho thấy phương pháp DPF mang lại cải thiện công bằng trong phần lớn các cấu hình, đặc biệt đối với các chỉ số “*chênh lệch trung bình xác suất*” và “*chênh lệch cơ hội công bằng*”. Tuy nhiên, vẫn tồn tại một số trường hợp suy giảm công bằng, chủ yếu xuất hiện khi dữ liệu ban đầu đã tương đối cân bằng hoặc khi số lượng nhóm con quá ít để kỹ thuật sinh dữ liệu tổng hợp phát huy hiệu quả. Điều này cho thấy DPF không phải là giải pháp tối ưu trong mọi tình huống, mà cần được áp dụng có cân nhắc dựa trên đặc điểm dữ liệu.

Cuối cùng, xét trên khía cạnh hiệu suất dự đoán, kết quả cho thấy DPF duy trì được mức hiệu suất chấp nhận được trong khi vẫn cải thiện đáng kể công bằng. Mặc dù cấu hình Origin thường đạt độ chính xác cao nhất, DPF thể hiện sự cân bằng tốt hơn giữa độ chính xác và độ hồi tưởng. Trong khi đó, SDG có

xu hướng cải thiện Recall mạnh hơn nhưng đánh đổi bằng suy giảm độ chính xác. Những quan sát này cho thấy DPF là một giải pháp trung hòa, cho phép cải thiện công bằng mà không gây suy giảm đáng kể hiệu suất, qua đó nâng cao tính khả thi của phương pháp trong các ứng dụng giáo dục thực tiễn.

3.5.2. Thảo luận về dữ liệu tổng hợp và khả năng áp dụng

Bên cạnh hiệu quả cải thiện công bằng, kết quả nghiên cứu cũng cho thấy chất lượng của dữ liệu tổng hợp đóng vai trò quan trọng đối với hiệu quả của phương pháp DPF. Trong nghiên cứu này, chất lượng dữ liệu sinh được đánh giá gián tiếp thông qua kết quả huấn luyện mô hình sau khi bổ sung dữ liệu tổng hợp. Các kết quả thực nghiệm cho thấy dữ liệu sinh từ CTGAN và LLM đều góp phần cải thiện các chỉ số công bằng trong khi vẫn duy trì hiệu suất dự đoán ở mức cạnh tranh.

Để bảo đảm khả năng tái lập, toàn bộ cấu hình tham số, quy trình sinh dữ liệu và thiết kế prompt được giữ cố định trong toàn bộ các thực nghiệm. Tuy nhiên, đối với các mô hình ngôn ngữ lớn, quá trình sinh dữ liệu vẫn mang tính xác suất, do đó kết quả giữa các lần thực hiện có thể xuất hiện những khác biệt nhỏ.

Ngoài ra, nghiên cứu sử dụng các bộ dữ liệu giáo dục đã được ẩn danh, do đó nguy cơ rò rỉ thông tin cá nhân trong quá trình sinh dữ liệu được giảm thiểu. Tuy nhiên, đối với các bộ dữ liệu có chứa thông tin nhận dạng hoặc dữ liệu nhạy cảm, cần áp dụng thêm các biện pháp bảo vệ dữ liệu trước khi sử dụng các mô hình sinh dữ liệu, đặc biệt là các mô hình ngôn ngữ lớn.

Mặc dù nghiên cứu được thực hiện trên các bộ dữ liệu giáo dục, quy trình DPF không phụ thuộc vào đặc thù của lĩnh vực này mà chủ yếu dựa trên cấu trúc dữ liệu bảng và các thuộc tính nhạy cảm. Vì vậy, phương pháp có tiềm năng được mở rộng sang các lĩnh vực khác như tài chính, y tế hoặc tuyển dụng. Tuy nhiên, khi áp dụng trên các miền dữ liệu khác nhau cần xem xét lại việc lựa chọn thuộc tính nhạy cảm, các tiêu chí công bằng và kỹ thuật sinh dữ liệu phù hợp với đặc điểm của từng bài toán.

3.5.3. Hạn chế của nghiên cứu

Mặc dù DPF cho thấy tiềm năng cải thiện công bằng giao thoa đáng kể, nghiên cứu vẫn tồn tại một số hạn chế liên quan đến độ tin cậy nội tại, độ tin cậy bên ngoài, giá trị cấu trúc và giá trị kết luận [137, 138, 139].

Độ tin cậy nội tại: Phương pháp DPF dựa trên kỹ thuật tái phân phối dữ liệu thông qua các bước sinh dữ liệu tổng hợp và hiệu chỉnh dựa trên mô hình, do đó chịu ảnh hưởng đáng kể từ chất lượng và tính đại diện của dữ liệu tổng hợp. Trong các trường hợp dữ liệu mất cân bằng nghiêm trọng hoặc tồn tại các nhóm giao thoa có rất ít mẫu, chất lượng dữ liệu sinh có thể ảnh hưởng trực tiếp đến hiệu quả của phương pháp. Cụ thể, đối với các nhóm giao thoa chỉ có 0 hoặc 1 mẫu, nghiên cứu không thực hiện sinh dữ liệu tổng hợp do không đủ thông tin để học phân bố dữ liệu một cách đáng tin cậy. Mặc dù lựa chọn này giúp hạn chế việc tạo ra dữ liệu tổng hợp thiếu chất lượng, đây cũng là một hạn chế của phương pháp vì các nhóm thiểu số rất nhỏ thường là những nhóm cần được quan tâm nhất trong các bài toán đảm bảo công bằng. Bên cạnh đó, hiệu quả của DPF cũng phụ thuộc vào kỹ thuật sinh dữ liệu được sử dụng. Mặc dù cả CTGAN và LLM đều giúp cải thiện tính công bằng trong nhiều trường hợp, kết quả thực nghiệm cho thấy hiệu quả và độ ổn định của DPF có thể khác nhau giữa các kỹ thuật sinh dữ liệu. Các chỉ số công bằng được lựa chọn trong nghiên cứu tuy phổ biến nhưng chưa bao quát toàn bộ các khía cạnh của công bằng, do đó vẫn có thể bỏ sót một số dạng thiên vị khác.

Độ tin cậy bên ngoài: Thí nghiệm được thực hiện trên các bộ dữ liệu giáo dục đặc thù, có thể chưa phản ánh đầy đủ bối cảnh thực tiễn đa dạng hơn. Khả năng tổng quát hóa sang các lĩnh vực khác hoặc các tình huống dữ liệu lớn, dữ liệu phi cấu trúc vẫn cần được kiểm chứng thêm.

Giá trị cấu trúc: DPF xử lý tốt các trường hợp có nhiều thuộc tính nhạy cảm, nhưng trong các tình huống thuộc tính nhạy cảm phức tạp hoặc dạng biến đặc biệt (phi số, phân loại đa cấp), hiệu quả của phương pháp có thể bị ảnh hưởng nếu không áp dụng các bước tiền xử lý phù hợp.

Giá trị kết luận: Mặc dù DPF đạt được sự cân bằng tốt giữa công bằng và hiệu suất trong phần lớn các cấu hình thực nghiệm, vẫn tồn tại một số trường

hợp phương pháp chưa vượt trội so với các phương pháp đối sánh. Điều này chủ yếu xuất phát từ mức độ mất cân bằng của dữ liệu, đặc điểm của các nhóm giao thoa và chất lượng của dữ liệu tổng hợp được sinh ra. Vì vậy, hiệu quả của DPF cần được tiếp tục đánh giá trên nhiều bộ dữ liệu và kỹ thuật sinh dữ liệu khác nhau nhằm nâng cao tính ổn định và khả năng tổng quát của phương pháp.

3.5.4. Ý nghĩa và ứng dụng thực tiễn

Kết quả nghiên cứu cho thấy DPF có thể trở thành một công cụ quan trọng trong việc đảm bảo tính công bằng trong các hệ thống học máy giáo dục. Việc áp dụng DPF vào các bài toán như dự đoán bỏ học, đánh giá kết quả học tập hoặc tuyển sinh có thể giúp giảm thiểu các rủi ro thiên lệch đối với những nhóm yếu thế. Không chỉ cải thiện công bằng đầu ra, phương pháp này còn đạt được sự cân bằng hợp lý giữa công bằng và hiệu suất. Điều này góp phần nâng cao khả năng ứng dụng trong môi trường thực tế. Ngoài giáo dục, DPF cũng có thể mở rộng sang các lĩnh vực như y tế, tài chính, hoặc tuyển dụng – nơi công bằng là yếu tố then chốt trong ra quyết định.

3.6. Tóm tắt chương

Chương này đã trình bày phương pháp DPF — một kỹ thuật can thiệp công bằng dựa trên chiến lược phân tách dữ liệu và sinh dữ liệu tổng hợp, nhằm giảm thiểu thiên lệch giữa các nhóm giao thoa của nhiều thuộc tính nhạy cảm trong các mô hình học máy. Phương pháp hướng tới giải quyết ba mục tiêu trọng tâm trong nghiên cứu về đảm bảo tính công bằng, bao gồm: (1) cải thiện công bằng đồng thời theo nhiều độ đo khác nhau (“*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*”, “*chênh lệch cơ hội công bằng*”); (2) duy trì hiệu suất dự đoán của mô hình (“*độ chuẩn xác*” và “*độ hồi tưởng*”) sau can thiệp; và (3) hạn chế mức đánh đổi bất lợi giữa công bằng và hiệu suất.

Các thí nghiệm được tiến hành trên bốn bộ dữ liệu giáo dục với hai kỹ thuật sinh dữ liệu tổng hợp (CTGAN và LLM) và năm mô hình học máy (*Hồi quy logistic*, *Rừng ngẫu nhiên*, *Cây quyết định*, *Tăng cường gradient*, và *Mạng nơ*

ron thần kinh). Kết quả cho thấy DPF mang lại cải thiện công bằng rõ rệt so với cấu hình dữ liệu gốc (Origin) và cấu hình chỉ áp dụng sinh dữ liệu tổng hợp trực tiếp (SDG). Trong nhiều kịch bản, DPF đạt kết quả tốt trên đồng thời nhiều độ đo công bằng và duy trì hiệu suất dự đoán ở mức ổn định. Một số trường hợp ghi nhận mức cải thiện đáng kể về công bằng, trong khi “độ chuẩn xác” và “độ hồi tưởng” chỉ thay đổi ở mức nhỏ.

Phân tích chi tiết theo mô hình và bộ dữ liệu cho thấy DPF có khả năng thích ứng tốt với các thuật toán học máy khác nhau và cho kết quả tương đối ổn định giữa các kỹ thuật sinh dữ liệu tổng hợp. Việc cân bằng dữ liệu ở mức nhóm giao thoa giúp giảm thiểu rủi ro thiên lệch do phân bố dữ liệu không đồng đều, đồng thời hạn chế hiện tượng đánh đổi quá mức giữa công bằng và hiệu suất như quan sát ở một số cấu hình đối sánh.

Bên cạnh giá trị phương pháp, chương này cũng làm rõ ý nghĩa thực tiễn của DPF trong các ứng dụng giáo dục như dự đoán nguy cơ bỏ học, hỗ trợ tuyển sinh hoặc đánh giá học tập, nơi dữ liệu thường mất cân bằng và chứa nhiều thuộc tính nhạy cảm. Ngoài ra, khả năng mở rộng của DPF sang các lĩnh vực khác như y tế, tài chính hoặc tuyển dụng cho thấy tiềm năng ứng dụng rộng rãi của phương pháp trong các hệ thống ra quyết định có yêu cầu cao về trách nhiệm xã hội.

Các kết quả và phân tích trong chương này tạo tiền đề trực tiếp cho Chương 4, nơi phương pháp DPF được kết hợp với kỹ thuật hiệu chỉnh đặc trưng nhằm xây dựng một khuôn khổ đảm bảo công bằng hai chiều và kiểm soát mối quan hệ đánh đổi giữa công bằng và hiệu suất.

Các kết quả nghiên cứu chính của chương này đã được tổng hợp trong một công trình khoa học do tác giả là tác giả chính, hiện đang trong quá trình phản biện tại tạp chí *Information and Software Technology*.

Chương 4

PHƯƠNG PHÁP ĐẢM BẢO TÍNH CÔNG BẰNG HAI CHIỀU CHO DỮ LIỆU DẠNG BẢNG – CHỈ SỐ ĐÁNH ĐỔI GIỮA CÔNG BẰNG VÀ HIỆU SUẤT

Như đã phân tích trong Chương 1, các nghiên cứu về công bằng trong học máy cho giáo dục hiện nay còn tồn tại hai khoảng trống quan trọng: (i) thiếu các phương pháp có khả năng xử lý đồng thời nhiều nguồn thiên lệch phát sinh từ cả phân phối dữ liệu và sự phụ thuộc trong không gian đặc trưng, và (ii) thiếu các thước đo định lượng rõ ràng để đánh giá mối quan hệ đánh đổi giữa cải thiện công bằng và suy giảm hiệu suất mô hình. Hai vấn đề này tương ứng với các câu hỏi nghiên cứu RQ3 và RQ4 đã được đặt ra trong phần tổng quan.

Ở các chương trước, luận án đã tiếp cận từng khía cạnh của hai vấn đề trên theo các hướng bổ sung. Chương 2 đề xuất phương pháp FairEdu, tập trung loại bỏ sự phụ thuộc giữa các thuộc tính không nhạy cảm và các thuộc tính nhạy cảm thông qua hồi quy đa biến (can thiệp theo chiều dọc). Chương 3 giới thiệu phương pháp DPF, nhằm cân bằng phân phối dữ liệu giữa các nhóm giao thoa bằng kỹ thuật sinh dữ liệu tổng hợp (can thiệp theo chiều ngang). Mặc dù mỗi phương pháp cho thấy hiệu quả nhất định, các kết quả thực nghiệm cũng cho thấy rằng việc cải thiện công bằng theo từng chiều riêng lẻ chưa đủ để giải quyết triệt để các dạng thiên lệch phức tạp trong dữ liệu giáo dục.

Trên cơ sở đó, chương này đóng vai trò trực tiếp trong việc trả lời RQ3 của luận án thông qua việc đề xuất phương pháp *FairEduPlus* – một khuôn khổ đảm bảo tính công bằng hai chiều cho dữ liệu dạng bảng. FairEduPlus kết hợp đồng thời cơ chế loại bỏ thông tin thiên lệch trong không gian đặc trưng (dựa trên FairEdu) và cơ chế cân bằng phân phối dữ liệu giữa các nhóm giao thoa (dựa trên DPF), từ đó hướng tới việc cải thiện công bằng một cách toàn diện và ổn định hơn.

Bên cạnh đó, câu hỏi RQ4 cũng được trả lời tại chương này qua việc giới thiệu một chỉ số đánh đổi giữa công bằng và hiệu suất nhằm định lượng và phân tích mối quan hệ giữa hai mục tiêu vốn thường có xu hướng đối nghịch trong học máy. Thông qua các phân tích lý thuyết và thực nghiệm, chương làm rõ mức độ cải thiện công bằng mà FairEduPlus đạt được, cũng như mức suy giảm hiệu suất tương ứng, qua đó đánh giá tính khả thi của phương pháp trong các bối cảnh giáo dục thực tiễn.

4.1. Giới thiệu

Trong các hệ thống phần mềm hiện đại, trí tuệ nhân tạo (AI) ngày càng được tích hợp sâu để hỗ trợ dự đoán, sinh nội dung và ra quyết định. Bên cạnh các đặc trưng chất lượng truyền thống như hiệu suất và độ tin cậy [140, 141, 142, 149], tính công bằng đã được xác định là một yêu cầu chất lượng cốt lõi của các hệ thống AI, do các quyết định thiên lệch có thể dẫn đến hệ quả đạo đức và phân biệt đối xử [118, 150].

Như phần tổng quan Chương 1 đã đề cập đến, trong lĩnh vực giáo dục, yêu cầu công bằng đặc biệt quan trọng vì các hệ thống AI thường tham gia trực tiếp vào các quyết định có tác động lâu dài đến người học, như dự đoán nguy cơ bỏ học, đánh giá năng lực hoặc hỗ trợ phân bổ nguồn lực. Một trong những thách thức trọng tâm đã được chỉ ra là sự tồn tại đồng thời của nhiều thuộc tính nhạy cảm trong dữ liệu giáo dục, dẫn đến các dạng thiên lệch giao thoa phức tạp và khó kiểm soát. Việc đảm bảo công bằng trong bối cảnh này không chỉ đòi hỏi các phương pháp can thiệp ở mức mô hình, mà còn yêu cầu dữ liệu huấn luyện phải có tính đại diện đầy đủ cho các nhóm giao thoa. Tuy nhiên, trong thực tế, nhiều nhóm con hiếm khi xuất hiện trong dữ liệu gốc, làm hạn chế khả năng đánh giá và cải thiện công bằng một cách đáng tin cậy [151, 152]. Dữ liệu tổng hợp vì vậy được xem là một hướng tiếp cận tiềm năng, vừa giúp bổ sung dữ liệu cho các nhóm yếu thế, vừa giảm thiểu rủi ro về quyền riêng tư [95, 96, 100, 110, 111, 145].

Trong các chương trước, luận án đã tiếp cận vấn đề công bằng từ hai hướng bổ sung. Chương 2 đề xuất phương pháp FairEdu, một kỹ thuật tiền xử lý dựa trên hồi quy đa biến nhằm loại bỏ sự phụ thuộc giữa các thuộc tính không nhạy

cảm và các thuộc tính nhạy cảm trong dữ liệu huấn luyện, tương ứng với cơ chế can thiệp theo chiều dọc. Phương pháp này cho thấy hiệu quả trong việc giảm thiên lệch do phụ thuộc ẩn, nhưng còn hạn chế trong các trường hợp dữ liệu mất cân bằng nghiêm trọng giữa các nhóm giao thoa. Trong khi đó, Chương 3 tập trung vào can thiệp theo chiều ngang thông qua phương pháp DPF, nhằm cân bằng phân phối dữ liệu giữa các nhóm giao thoa bằng kỹ thuật sinh dữ liệu tổng hợp.

Trên cơ sở những phân tích và kết quả đó, chương này đề xuất phương pháp *FairEduPlus*, một khuôn khổ can thiệp hai chiều nhằm giải quyết đồng thời hai vấn đề cốt lõi đã được xác định trong tổng quan: (i) làm thế nào để cải thiện công bằng trong bối cảnh dữ liệu có nhiều thuộc tính nhạy cảm và phân bố không cân bằng, và (ii) làm thế nào để kiểm soát mối quan hệ đánh đổi giữa cải thiện công bằng và suy giảm hiệu suất của mô hình học máy. *FairEduPlus* tích hợp việc cân bằng phân phối dữ liệu giữa các nhóm giao thoa thông qua DPF với việc loại bỏ thông tin thiên lệch tiềm ẩn trong không gian đặc trưng thông qua *FairEdu*, từ đó hướng tới một giải pháp công bằng toàn diện hơn.

Nội dung của chương tập trung phân tích ba khía cạnh chính: mức độ cải thiện công bằng khi áp dụng *FairEduPlus* trên các kịch bản dữ liệu giáo dục khác nhau; tác động của phương pháp đến hiệu suất dự đoán của mô hình học máy; và khả năng định lượng mối quan hệ đánh đổi giữa công bằng và hiệu suất thông qua chỉ số đánh đổi được đề xuất. Thông qua các phân tích này, chương đóng vai trò trực tiếp trong việc làm sáng tỏ các vấn đề đã được đặt ra trong phần tổng quan, đồng thời đánh giá tính hiệu quả và tính khả thi của *FairEduPlus* trong các bối cảnh giáo dục thực tiễn.

4.2. Phương pháp *FaireduPlus*

4.2.1. Ý tưởng chính

Nghiên cứu này hướng tới việc phát hiện và giảm thiểu thiên lệch của các mô hình học máy ngay từ giai đoạn tiền xử lý, đặc biệt với dữ liệu dạng bảng. Điểm mới của phương pháp đề xuất là can thiệp đồng thời theo hai chiều dữ liệu:

Chiều ngang – điều chỉnh phân phối dữ liệu: sử dụng kỹ thuật DPF 3 nhằm cân bằng số lượng mẫu giữa các nhóm giao thoa được tạo bởi tổ hợp các thuộc tính nhạy cảm và nhãn đầu ra. Việc cân bằng này giúp phân phối các kết quả thuận lợi và bất lợi đồng đều hơn giữa các nhóm.

Chiều dọc – điều chỉnh đặc trưng: áp dụng kỹ thuật Fairedu để loại bỏ sự phụ thuộc của các thuộc tính không nhạy cảm vào các thuộc tính nhạy cảm, từ đó giảm ảnh hưởng của các yếu tố nhạy cảm đến dự đoán của mô hình.

Phương pháp FaireduPlus kết hợp sức mạnh của hai hướng tiếp cận này: (i) DPF để tái cân bằng dữ liệu ở mức phân phối, và (ii) Fairedu để khử bỏ thông tin thiên lệch ở mức đặc trưng. Quá trình này vừa đảm bảo tính công bằng đồng thời, vừa duy trì hiệu suất dự đoán của mô hình. Đồng thời, kỹ thuật này can thiệp ở giai đoạn tiền xử lý nên độc lập với mô hình, do đó có thể áp dụng cho nhiều thuật toán học máy khác nhau mà không cần chỉnh sửa cấu trúc hay quy trình huấn luyện.

Với giả định dữ liệu có k thuộc tính nhạy cảm (x_1, \dots, x_k) và nhãn nhị phân $y \in \{0, 1\}$, tập dữ liệu được chia thành 2^{k+1} nhóm con theo tổ hợp các giá trị của thuộc tính nhạy cảm và nhãn. Ở bước cân bằng, phương pháp DPF sinh bổ sung dữ liệu cho các nhóm con, nhằm đảm bảo tỷ lệ giữa nhãn dương và nhãn âm được duy trì đồng đều trong từng tổ hợp thuộc tính nhạy cảm, sau đó hợp nhất lại thành một tập dữ liệu thống nhất. Tiếp theo, bước điều chỉnh áp dụng phương pháp Fairedu để loại bỏ sự phụ thuộc của các thuộc tính không nhạy cảm vào các thuộc tính nhạy cảm, từ đó hoàn thiện tập dữ liệu công bằng trước khi huấn luyện mô hình.

4.2.2. Kiến trúc tổng thể

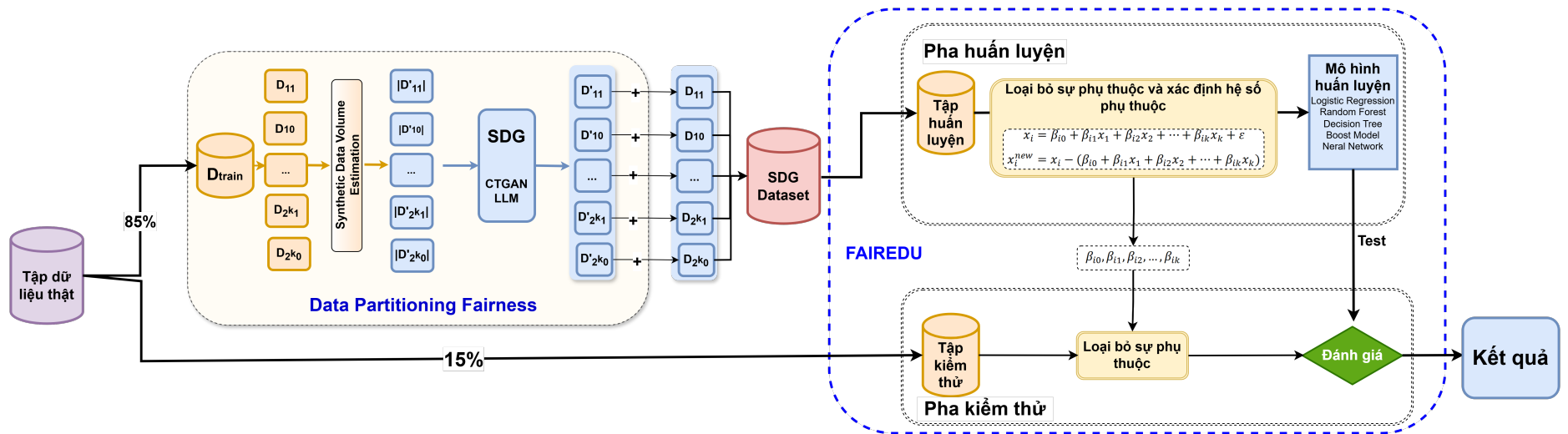
Kiến trúc tổng thể của phương pháp FaireduPlus được minh họa trong Hình 4.1. Quy trình bắt đầu bằng việc chia tập dữ liệu thành tập huấn luyện (85%) và tập kiểm thử (15%). Trên tập huấn luyện, phương pháp DPF phân chia dữ liệu thành các nhóm giao thoa theo tổ hợp các thuộc tính nhạy cảm và nhãn đầu ra, sau đó sinh dữ liệu tổng hợp bằng CTGAN hoặc LLM nhằm cân bằng số lượng mẫu giữa các nhóm và tạo thành tập dữ liệu huấn luyện mới

(SDG Dataset). Bộ dữ liệu này tiếp tục được đưa vào phương pháp Fairedu để xác định và loại bỏ sự phụ thuộc giữa các thuộc tính không nhạy cảm và các thuộc tính nhạy cảm trước khi huấn luyện mô hình học máy. Trong pha kiểm thử, tập kiểm thử được điều chỉnh bằng các hệ số hồi quy thu được trong pha huấn luyện nhằm đảm bảo tính nhất quán giữa hai pha, sau đó được sử dụng để đánh giá mô hình theo các chỉ số hiệu suất và công bằng.

Kiến trúc này kết hợp hai cơ chế cải thiện công bằng bổ trợ cho nhau. DPF đóng vai trò cân bằng dữ liệu theo các nhóm giao thoa thông qua kỹ thuật sinh dữ liệu tổng hợp, trong khi Fairedu tiếp tục giảm ảnh hưởng của các thuộc tính nhạy cảm trên bộ dữ liệu đã được tăng cường. Nhờ đó, FaireduPlus vừa cải thiện tính đại diện của dữ liệu huấn luyện, vừa giảm sự phụ thuộc của mô hình vào các thuộc tính nhạy cảm, góp phần nâng cao tính công bằng mà vẫn duy trì hiệu suất dự đoán.

Trên cơ sở kiến trúc tổng thể trên, quy trình của FaireduPlus gồm ba bước chính: (i) phân chia và cân bằng dữ liệu bằng DPF, (ii) điều chỉnh công bằng bằng Fairedu và (iii) kiểm thử, đánh giá mô hình. Chi tiết từng bước được trình bày như sau:

- **Bước 1 – Phân chia và cân bằng dữ liệu (DPF):** Tập dữ liệu được chia thành 85% dữ liệu huấn luyện và 15% dữ liệu kiểm thử. Tập huấn luyện được phân tách thành 2^{k+1} nhóm giao thoa theo tổ hợp các thuộc tính nhạy cảm và nhãn đầu ra. Sau đó, DPF sử dụng CTGAN hoặc LLM để sinh dữ liệu tổng hợp, cân bằng số lượng mẫu giữa các nhóm và tạo thành tập dữ liệu huấn luyện mới.
- **Bước 2 – Điều chỉnh công bằng (Fairedu):** Trên tập dữ liệu đã được cân bằng, Fairedu xác định và loại bỏ sự phụ thuộc giữa các thuộc tính không nhạy cảm và các thuộc tính nhạy cảm bằng hồi quy tuyến tính đa biến. Bộ dữ liệu sau khi điều chỉnh được sử dụng để huấn luyện các mô hình học máy.
- **Bước 3 – Kiểm thử và đánh giá:** Tập kiểm thử được điều chỉnh bằng các hệ số hồi quy thu được trong pha huấn luyện nhằm đảm bảo tính nhất quán giữa hai pha. Cuối cùng, mô hình được đánh giá trên tập kiểm thử thông qua các chỉ số hiệu suất và công bằng.



Hình 4.1: Kiến trúc tổng thể của phương pháp FaireduPlus.

4.2.3. Thuật toán FaireduPlus

Phương pháp FaireduPlus được thiết kế nhằm nâng cao tính công bằng của các mô hình học máy bằng cách kết hợp hai hướng can thiệp: (i) cân bằng phân phối giữa các nhóm giao thoa của thuộc tính nhạy cảm và nhân (theo chiều ngang), và (ii) loại bỏ sự phụ thuộc của các thuộc tính không nhạy cảm vào các thuộc tính nhạy cảm (theo chiều dọc). Chi tiết quy trình thực hiện thuật toán FaireduPlus được mô tả trong Thuật toán 4.1.

Thuật toán 4.1 FaireduPlus

- 1: **Input:** Tập huấn luyện $D_{tr} = \{\langle x_1, y_1 \rangle, \dots, \langle x_n, y_n \rangle\}$ với $x_j = [x_1^j, \dots, x_d^j]$; x_1^j, \dots, x_k^j là các thuộc tính nhạy cảm, x_{k+1}^j, \dots, x_d^j là các thuộc tính không nhạy cảm, $y_j \in \{0, 1\}$. Tập kiểm thử D_{test} ;
 - 2: **Output:** Tập dữ liệu công bằng D' , mô hình ML S_{ML} , và nhân dự đoán $S_{ML}(x^{te})$;
 - 3: Xác định $n = 2^k$, phân D_{tr} thành $2n$ nhóm $D_{10}, D_{11}, \dots, D_{n0}, D_{n1}$ theo mọi tổ hợp nhị phân của k thuộc tính nhạy cảm và nhân Y ;
 - 4: Tìm nhóm nhỏ nhất: $D_{\min} = \min_{i=1, \dots, n; j=0, 1} \|D_{ij}\|$, lưu chỉ số (i_s, j_s) ;
 - 5: Chọn $n_0 > \|D_{i_s j_s}\|$ và đặt $\|D'_{i_s j_s}\| = n_0$;
 - 6: Tính $R_{\text{mean}} = \frac{1}{n} \sum_{i=1}^n \frac{\|D_{i1}\|}{\|D_{i0}\|}$ với $\|D_{i0}\| \neq 0$;
 - 7: **if** $j_s = 1$ **then**
 - 8: **for** $i = 1$ **to** n **do**
 - 9: $\|D'_{i1}\| = \frac{\|D_{i1}\|}{\|D_{i_s j_s}\|} n_0$;
 - 10: $\|D'_{i0}\| = \frac{\|D'_{i1}\|}{R_{\text{mean}}}$;
 - 11: **end for**
 - 12: **else**
 - 13: **for** $i = 1$ **to** n **do**
 - 14: $\|D'_{i0}\| = \frac{\|D_{i0}\|}{\|D_{i_s j_s}\|} n_0$;
 - 15: $\|D'_{i1}\| = R_{\text{mean}} \cdot \|D'_{i0}\|$;
 - 16: **end for**
 - 17: **end if**
 - 18: Loại bỏ thuộc tính nhạy cảm khỏi từng D_{ij} ;
 - 19: **Sinh dữ liệu tổng hợp:**
 - 20: **for** $i = 1$ **to** n **do**
 - 21: **for** $j = 0$ **to** 1 **do**
 - 22: Áp dụng CTGAN hoặc LLM để sinh thêm $\|D'_{ij}\| - \|D_{ij}\|$ mẫu cho D_{ij} (sau khi loại thuộc tính nhạy cảm), rồi gắn lại các thuộc tính nhạy cảm tương ứng;
 - 23: **end for**
 - 24: **end for**
 - 25: Hợp nhất tất cả D'_{ij} thành tập huấn luyện mới D'_{tr} ;
 - 26: Áp dụng Fairedu trên D'_{tr} để điều chỉnh đặc trưng (loại bỏ phụ thuộc vào thuộc tính nhạy cảm);
 - 27: Huấn luyện mô hình S_{ML} trên dữ liệu đã điều chỉnh;
 - 28: Trên tập kiểm thử D_{test} , áp dụng cùng quy trình loại bỏ phụ thuộc như với tập huấn luyện;
 - 29: Đánh giá S_{ML} theo các chỉ số công bằng ($|1 - DI|$, SPD, AOD, EOD) và hiệu suất (Accuracy, Recall);
 - 30: **return** $D'_{tr}, S_{ML}, S_{ML}(x^{te})$; =0
-

Độ phức tạp của thuật toán FaireduPlus. Giả sử tập dữ liệu có n mẫu, d thuộc tính và k thuộc tính nhạy cảm. Thuật toán FaireduPlus kết hợp hai thành phần chính: (i) cân bằng dữ liệu dựa trên DPF và (ii) điều chỉnh phụ thuộc đặc trưng theo phương pháp Fairedu.

Phần DPF có độ phức tạp $O(nk + 2^k + n_{gen} \cdot C_{gen})$, trong đó n_{gen} là số mẫu sinh thêm và C_{gen} là chi phí sinh một mẫu dữ liệu tổng hợp.

Phần Fairedu có độ phức tạp:

$$O((d - k) \cdot n' \cdot k^2 + n' \cdot (d - k) \cdot k),$$

trong đó $n' = n + n_{gen}$ là kích thước dữ liệu sau khi cân bằng.

Do đó, tổng thể độ phức tạp của FaireduPlus có thể được biểu diễn như sau:

$$O(nk + 2^k + n_{gen} \cdot C_{gen} + (d - k) \cdot n' \cdot k^2 + n' \cdot (d - k) \cdot k).$$

Trong thực tế, với số lượng thuộc tính nhạy cảm k nhỏ, chi phí tính toán chủ yếu đến từ quá trình sinh dữ liệu và bước hồi quy đa biến của Fairedu. Điều này cho thấy FaireduPlus vẫn đảm bảo khả năng mở rộng và tính khả thi trong các bài toán học máy giáo dục.

4.3. Thực nghiệm

4.3.1. Dữ liệu, mô hình học máy, kỹ thuật sinh dữ liệu tổng hợp, và độ đo đánh giá

Nghiên cứu sử dụng bốn bộ dữ liệu giáo dục thực tế đã được mô tả chi tiết trong Mục 2.3.1 của Chương 2, bao gồm: *Student Performance*, *Student Predict Dropout*, *Oulad* và *DNU Data*, với các thuộc tính nhạy cảm *giới tính*, *sức khỏe*, *tuổi*, *tình trạng nợ*, *tình trạng khuyết tật* và *khu vực*. Quy trình tiền xử lý, phương pháp phân chia dữ liệu dựa trên tổ hợp giao thoa giữa thuộc tính nhạy cảm và nhãn mục tiêu, cũng như các thuộc tính đầu vào, được áp dụng thống nhất như trong Chương 3.

Các thí nghiệm được thực hiện trên năm mô hình học máy gồm *Hồi quy logistic*, *Cây quyết định*, *Rừng ngẫu nhiên*, *Tăng cường gradient* và *Mạng nơ ron thần kinh* với cấu hình huấn luyện, tham số điều chỉnh và tiêu chí dừng được giữ nguyên như trong Mục 3.3.2 của Chương 3. Bên cạnh đó, hai kỹ thuật sinh dữ liệu tổng hợp là *CTGAN* và *LLM* cũng được áp dụng theo các thiết lập đã mô tả trong Mục 3.3.2. Việc duy trì đồng nhất toàn bộ quy trình, dữ liệu, cấu hình mô hình và thông số sinh dữ liệu nhằm loại bỏ ảnh hưởng của các yếu tố ngoài phương pháp đề xuất, từ đó đảm bảo rằng mọi khác biệt quan sát được trong kết quả đều phản ánh trực tiếp hiệu quả của FaireduPlus đối với cả tính công bằng và hiệu suất dự đoán.

Về độ đo đánh giá, chương này sử dụng các độ đo công bằng và hiệu suất phổ biến đã trình bày chi tiết ở Chương 1, trong Mục 1.1.2 và 1.2.4. Cụ thể, về tính công bằng, bốn độ đo được sử dụng gồm “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*”, “*chênh lệch cơ hội công bằng*” phản ánh cả sự khác biệt ở mức nhóm và công bằng có điều kiện theo dự đoán. Về hiệu suất dự đoán, hai độ đo chuẩn được sử dụng gồm “*độ chuẩn xác*”, “*độ hồi tưởng*”, cho phép phân tích rõ ràng sự đánh đổi giữa công bằng và hiệu suất.

Bảng 4.1: Tóm tắt các bộ dữ liệu sử dụng trong nghiên cứu

STT	Bộ dữ liệu	Số TT	Số mẫu	Thuộc tính nhạy cảm	Nhãn mục tiêu	Đặc điểm phân phối
1	<i>Student Performance</i> (SPdt)	33	1.042	<i>giới tính, sức khỏe</i>	Đạt / Không đạt	Mất cân đối vừa phải giữa các nhóm <i>giới tính-sức khỏe</i>
2	<i>Student Predict Dropout</i> (SPredt)	35	4.424	<i>giới tính, tình trạng nợ</i>	Tốt nghiệp / Bỏ học	Chênh lệch lớn, một số nhóm rất nhỏ
3	<i>Oulad</i> (OLdt)	12	31.482	<i>giới tính,</i>	Đạt / Không đạt	Nhóm chiếm tỷ lệ nhỏ
4	<i>DNU Data</i>	11	426	<i>giới tính, khu vực</i>	Đạt / Không đạt	Nhiều nhóm giao thoa kích thước rất nhỏ hoặc rỗng

4.3.2. Thiết lập thực nghiệm

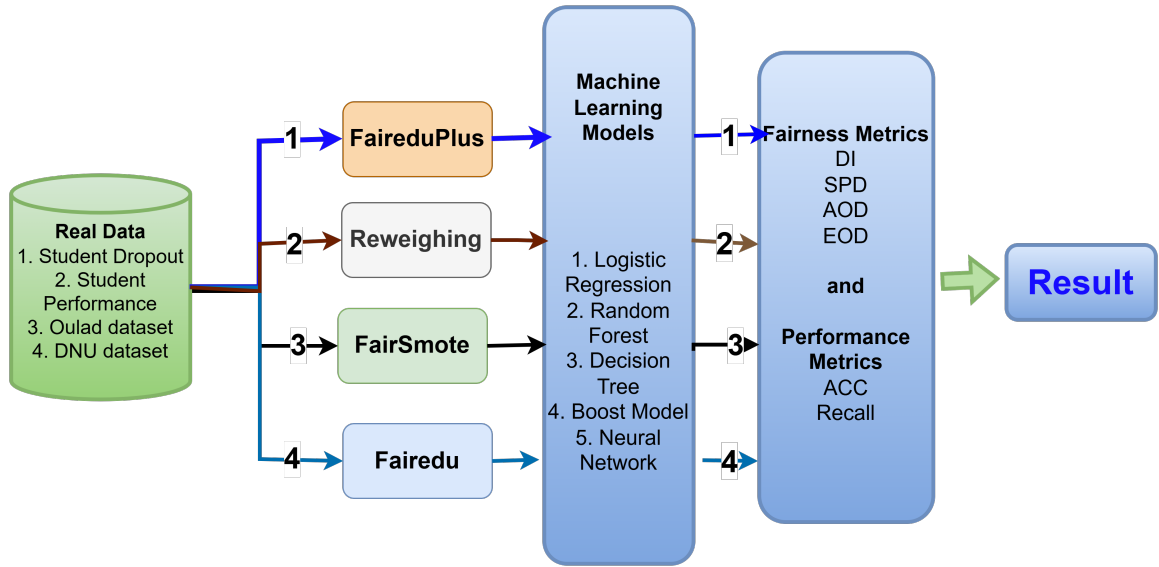
Nhằm tiến hành đánh giá một cách hệ thống hiệu quả của các can thiệp công bằng trên dữ liệu dạng bảng, thực nghiệm được tiến hành trên bốn cấu hình được minh họa như trong Hình 4.2. Mỗi cấu hình được xây dựng có chủ đích để khảo sát tác động của các cách kết hợp khác nhau giữa dữ liệu thực và dữ liệu tổng hợp trong việc giảm thiểu thiên lệch. Bốn cấu hình thực nghiệm bao gồm: FaireduPlus (1), Reweighting (2), FairSmote (3), và Fairedu (4).

Để sinh dữ liệu tổng hợp, chúng tôi sử dụng hai kỹ thuật khác nhau: CTGAN và mô hình ngôn ngữ lớn (LLM).

- Đối với CTGAN, chúng tôi sử dụng mô hình CTGAN của thư viện SDV (phiên bản v0.18.0) với 10 epoch, kích thước batch bằng 500, tốc độ học $2e-4$ cho cả hai mạng, và vector embedding 128 chiều. Cấu hình này cho phép chuẩn hóa theo từng chế độ và lấy mẫu có điều kiện đối với các thuộc tính rời rạc và mất cân bằng. Một mô hình CTGAN riêng biệt được huấn luyện cho từng nhóm con, với số lượng mẫu sinh ra được điều chỉnh phù hợp với nhu cầu của từng nhóm.
- Đối với sinh dữ liệu dựa trên LLM, chúng tôi sử dụng ChatGPT 3.5, với cấu trúc prompt được thiết kế phù hợp với lược đồ và các đặc trưng thống kê của dữ liệu gốc. Quá trình sinh dữ liệu dựa trên prompt được thực hiện với các tham số: temperature = 0.7, độ dài tối đa 512 token và top- p = 0.9. Sau khi sinh, dữ liệu được hậu xử lý nhằm đảm bảo tính nhất quán về định dạng và loại bỏ các bản ghi không hợp lệ về mặt cấu trúc.

4.3.3. Chỉ số đánh giá

Các chỉ số công bằng và hiệu suất sử dụng trong chương này được kế thừa từ Mục 2.3.4, nhằm đảm bảo tính nhất quán trong toàn bộ nghiên cứu. Cụ thể, bốn chỉ số công bằng gồm “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*”, và “*chênh lệch cơ hội công bằng*”, cùng hai chỉ số hiệu suất gồm “*độ chuẩn xác*” và “*độ hồi tưởng*”.



Hình 4.2: Tổng quan các cấu hình thực nghiệm đánh giá hiệu quả FaireduPlus

Ngoài ra, để đánh giá sự đánh đổi giữa công bằng và hiệu suất, luận án đã so sánh mức độ cải thiện công bằng với mức suy giảm hiệu suất tương ứng, dựa trên quan điểm đa mục tiêu lấy cảm hứng từ Pareto [153]. Trên cơ sở xác định và lượng hóa hai đại lượng cơ bản: *chỉ số cải thiện công bằng* và *chỉ số suy giảm hiệu suất*, từ đó xác định một chỉ số mới gọi tên là *chỉ số đánh đổi tổng hợp*. Chỉ số này cho phép so sánh trực tiếp mức độ cải thiện công bằng đạt được với chi phí phải trả về mặt hiệu suất một cách nhất quán giữa các cấu hình can thiệp khác nhau, từ đó hỗ trợ việc đánh giá và lựa chọn các cấu hình mô hình một cách có cơ sở.

Ghi chú: Trong các định nghĩa dưới đây, *NewTechnique* biểu thị phương pháp đề xuất và *Baseline* tương ứng với cấu hình huấn luyện ban đầu không áp dụng bất kỳ kỹ thuật đảm bảo công bằng nào.

Định nghĩa 1. [*Chỉ số cải thiện công bằng*] Chỉ số cải thiện công bằng $\Delta_{fairness}$ phản ánh mức độ thay đổi về công bằng khi áp dụng một kỹ thuật mới so với kỹ thuật cơ sở *Fairedu*. Chỉ số này được xác định bởi Công thức 4.1 như sau:

$$\Delta_{fairness} = \frac{FairnessMetric_{New_Technique} - FairnessMetric_{Baseline}}{FairnessMetric_{Baseline}} \quad (4.1)$$

trong đó $FairnessMetric_{New_Technique}$ là giá trị của một độ đo công bằng (ví dụ:

|1-DI|, SPD, AOD, EOD) khi sử dụng kỹ thuật mới, còn $FairnessMetric_{Baseline}$ là giá trị tương ứng khi áp dụng kỹ thuật cơ sở.

Theo định nghĩa này, giá trị của $\Delta_{fairness}$ càng nhỏ (càng âm) thì mức cải thiện công bằng càng lớn. Cụ thể, $\Delta_{fairness}$ được diễn giải như sau:

- $\Delta_{fairness} < 0$: kỹ thuật mới cải thiện tính công bằng so với cấu hình cơ sở;
- $\Delta_{fairness} > 0$: tính công bằng bị suy giảm khi áp dụng kỹ thuật mới;
- $\Delta_{fairness} = 0$: mức độ công bằng không thay đổi so với cấu hình cơ sở.

Định nghĩa 2. [*Chỉ số suy giảm hiệu suất*] chỉ số suy giảm hiệu suất $\Delta_{performance}$ phản ánh mức độ thay đổi về hiệu suất dự đoán của mô hình khi áp dụng một kỹ thuật mới so với kỹ thuật cơ sở. Chỉ số này được xác định bởi Công thức 4.2 như sau:

$$\Delta_{performance} = \frac{Performance_{New_Technique} - Performance_{Baseline}}{Performance_{Baseline}} \quad (4.2)$$

trong đó $Performance_{New_Technique}$ là giá trị của một độ đo hiệu suất (ví dụ: Accuracy hoặc Recall) khi sử dụng kỹ thuật mới, còn $Performance_{Baseline}$ là giá trị tương ứng của kỹ thuật cơ sở.

Theo định nghĩa này, giá trị của $\Delta_{performance}$ càng lớn (càng dương) thì mức cải thiện hiệu suất càng cao. Cụ thể, $\Delta_{performance}$ được diễn giải như sau:

- $\Delta_{performance} > 0$: kỹ thuật mới cải thiện hiệu suất so với cấu hình cơ sở;
- $\Delta_{performance} < 0$: hiệu suất bị suy giảm khi áp dụng kỹ thuật mới;
- $\Delta_{performance} = 0$: hiệu suất không thay đổi so với cấu hình cơ sở.

Định nghĩa 3. [*Chỉ số đánh đổi tổng hợp*] Cho $\Delta_{fairness}$ là mức cải thiện công bằng và $\Delta_{performance}$ là mức thay đổi hiệu suất của mô hình so với cấu hình cơ sở, chỉ số đánh đổi tổng hợp được xác định theo Công thức 4.3.

$$\Delta_{trade.off} = \Delta_{performance} - \Delta_{fairness} \quad (4.3)$$

Diễn giải:

- $\Delta_{trade_off} > 0$: mức cải thiện công bằng lớn hơn mức suy giảm hiệu suất, cho thấy phương pháp tiến theo hướng đánh đổi thuận lợi;
- $\Delta_{trade_off} = 0$: mức cải thiện công bằng và suy giảm hiệu suất đạt trạng thái cân bằng;
- $\Delta_{trade_off} < 0$: mức suy giảm hiệu suất lớn hơn mức cải thiện công bằng, cho thấy sự đánh đổi không thuận lợi.

Chỉ số đánh đổi được xây dựng trên cơ sở hai chỉ số thay đổi tương đối của công bằng và hiệu suất, do đó có ưu điểm là không phụ thuộc vào đơn vị đo của từng chỉ số thành phần và cho phép so sánh giữa các cấu hình thực nghiệm khác nhau. Tuy nhiên, do $\Delta_{fairness}$ và $\Delta_{performance}$ đều được xác định theo tỷ lệ thay đổi so với cấu hình cơ sở, trong các trường hợp giá trị của cấu hình cơ sở nhỏ, các biến động nhỏ của công bằng hoặc hiệu suất có thể dẫn đến sự thay đổi tương đối lớn của chỉ số. Thực nghiệm cũng ghi nhận một số cấu hình có giá trị chỉ số đánh đổi biến động đáng kể do sự thay đổi lớn của một trong hai thành phần. Vì vậy, chỉ số này cần được diễn giải kết hợp với các chỉ số công bằng và hiệu suất gốc nhằm đảm bảo đánh giá đầy đủ hiệu quả của phương pháp.

Trên cơ sở định nghĩa trên, nghiên cứu tiến hành tính toán chỉ số đánh đổi tổng hợp Δ_{trade_off} cho từng cặp chỉ số hiệu suất và công bằng, nhằm đánh giá một cách định lượng mức độ hài hòa giữa hai mục tiêu này trong các cấu hình can thiệp khác nhau.

Cụ thể, khi sử dụng “độ chuẩn xác” làm chỉ số hiệu suất, chỉ số đánh đổi được xác định tương ứng với từng độ đo công bằng như sau:

- Với “độ chuẩn xác” và $|1 - DI|$:

$$\Delta_{trade_off}^{ACC_|1-DI|} = \Delta_{performance}^{ACC} - \Delta_{fairness}^{|1-DI|} \quad (4.4)$$

- Với “độ chuẩn xác” và “hiệu số chênh lệch thống kê”:

$$\Delta_{trade_off}^{ACC_SPD} = \Delta_{performance}^{ACC} - \Delta_{fairness}^{SPD} \quad (4.5)$$

- Với “độ chuẩn xác” và “chênh lệch trung bình xác suất”:

$$\Delta_{trade_off}^{ACC_AOD} = \Delta_{performance}^{ACC} - \Delta_{fairness}^{AOD} \quad (4.6)$$

- Với “độ chuẩn xác” và “chênh lệch cơ hội công bằng”:

$$\Delta_{trade_off}^{ACC_EOD} = \Delta_{performance}^{ACC} - \Delta_{fairness}^{EOD} \quad (4.7)$$

Tương tự, nghiên cứu cũng áp dụng cùng cách tiếp cận để tính toán chỉ số đánh đổi tổng hợp khi sử dụng “độ hồi tưởng” làm chỉ số hiệu suất, tương ứng với bốn độ đo công bằng:

$$\Delta_{trade_off}^{Recall_|1-DI|}, \quad \Delta_{trade_off}^{Recall_SPD}, \quad \Delta_{trade_off}^{Recall_AOD}, \quad \Delta_{trade_off}^{Recall_EOD},$$

Để làm rõ ý nghĩa của chỉ số đánh đổi tổng hợp Δ_{trade_off} , xét một ví dụ cụ thể với chỉ số hiệu suất “độ chuẩn xác” và độ đo công bằng “hiệu số chênh lệch thống kê”.

Ví dụ 4.1. Giả sử cấu hình cơ sở (FairEdu) đạt:

$$SPD_{Baseline} = 0.20, \quad ACC_{Baseline} = 0.80.$$

Xét một cấu hình mới (ví dụ FairEduPlus) cho kết quả:

$$SPD_{New} = 0.10, \quad ACC_{New} = 0.78.$$

Khi đó, mức thay đổi công bằng và hiệu suất được tính như sau:

$$\Delta_{fairness}^{SPD} = \frac{0.10 - 0.20}{0.20} = -0.50, \quad \Delta_{performance}^{ACC} = \frac{0.78 - 0.80}{0.80} = -0.025.$$

Chỉ số đánh đổi tổng hợp tương ứng là:

$$\Delta_{trade_off}^{ACC_SPD} = \Delta_{performance}^{ACC} - \Delta_{fairness}^{SPD} = -0.025 - (-0.50) = 0.475 > 0.$$

Kết quả này cho thấy, mặc dù hiệu suất dự đoán giảm nhẹ (2.5%), mức cải thiện công bằng đạt được là đáng kể (50%), dẫn đến một giá trị Δ_{trade_off} dương. Điều này phản ánh một sự đánh đổi thuận lợi, trong đó lợi ích về công bằng vượt trội so với chi phí phải trả về hiệu suất.

Ngược lại, nếu một cấu hình mới chỉ cải thiện công bằng ở mức rất nhỏ trong khi gây suy giảm hiệu suất lớn, chỉ số Δ_{trade_off} sẽ nhận giá trị âm, cho thấy sự đánh đổi không hợp lý và kém hấp dẫn trong thực tiễn.

4.4. Kết quả

Phần này trình bày và phân tích các kết quả thực nghiệm nhằm đánh giá hiệu quả của phương pháp FairEduPlus trong việc cải thiện công bằng và duy trì hiệu suất dự đoán của các mô hình học máy trong lĩnh vực giáo dục. Các kết quả được tổng hợp từ bốn cấu hình theo thiết lập thực nghiệm đã mô tả tại Mục 4.3. Luận án tập trung phân tích cả hai khía cạnh: (i) mức độ cải thiện công bằng, (ii) tác động của FairEduPlus đến hiệu suất của mô hình, và (iii) sự đánh đổi giữa công bằng và hiệu suất dự đoán của phương pháp FairEduPlus.

Ghi chú: Trong các phần trình bày dưới đây, các Bảng 4.2, 4.3, 4.4, 4.5 và 4.6 được định dạng thống nhất như sau: các ô màu xám biểu thị các cấu hình có kết quả kém hơn so với FairEduPlus, trong khi các ô màu xanh nhạt thể hiện các cấu hình có hiệu suất tương đương với FairEduPlus. Hàng cuối của mỗi bảng tổng hợp số lượng trường hợp Thắng/Hòa/Thua (W/T/L) của FairEduPlus so với các phương pháp đối sánh tương ứng trên từng chỉ số công bằng.

4.4.1. Mức độ cải thiện tính công bằng đồng thời của FairEduPlus

Để đánh giá mức độ cải thiện công bằng của FairEduPlus so với các phương pháp hiện có, nghiên cứu tiến hành so sánh bốn chỉ số công bằng gồm DI, SPD, AOD và EOD trên bốn cấu hình thực nghiệm, bao gồm FairEduPlus, Reweighting, FairSMOTE và FairEdu, như được mô tả trong Mục 4.3.2. Kết quả so sánh chi

tiết của FaireduPlus khi kết hợp với các phương pháp sinh dữ liệu tổng hợp dựa trên CTGAN và LLM lần lượt được trình bày trong Bảng 4.2 và 4.3.

Với cấu hình sử dụng CTGAN, các thống kê W/T/L trong Bảng 4.2 cho thấy FaireduPlus khi kết hợp với CTGAN vượt trội một cách nhất quán so với các phương pháp đối sánh trên hầu hết các chỉ số công bằng khi xét theo tỷ lệ phần trăm. Cụ thể, đối với chỉ số $|1 - DI|$, FaireduPlus đạt tỷ lệ thắng 75.6% so với Reweighting (34/45), 75.6% so với FairSMOTE (34/45) và 60.0% so với Fairedu (27/45). Các kết quả này cho thấy FaireduPlus hoạt động vượt trội đáng kể so với các cấu hình cơ sở. Đối với chỉ số SPD, FaireduPlus đạt tỷ lệ thắng dao động từ 62.2% (28/45) so với Fairedu đến 84.4% (38/45) so với FairSMOTE. Những kết quả này cung cấp bằng chứng thực nghiệm rõ ràng về khả năng cải thiện công bằng vượt trội của FaireduPlus khi sử dụng phương pháp sinh dữ liệu tổng hợp dựa trên CTGAN.

Tương tự, với cấu hình sử dụng LLM, các thống kê W/T/L trong Bảng 4.3 cho thấy FaireduPlus khi kết hợp với phương pháp sinh dữ liệu tổng hợp dựa trên LLM vẫn duy trì lợi thế tổng thể về công bằng, mặc dù mức độ vượt trội không rõ rệt như trong cấu hình CTGAN. Cụ thể, đối với chỉ số $|1 - DI|$, FaireduPlus đạt tỷ lệ thắng 66.7% (30/45) so với Reweighting, 53.3% (24/45) so với FairSMOTE và 53.3% (24/45) so với Fairedu. Xu hướng tương tự cũng được quan sát đối với chỉ số SPD, trong đó FaireduPlus đạt tỷ lệ thắng 55.6% (25/45) so với Reweighting, 62.2% (28/45) so với FairSMOTE và 48.9% (22/45) so với Fairedu.

Các kết quả trên cho thấy FaireduPlus có khả năng cải thiện công bằng một cách hiệu quả và nhất quán trên nhiều chỉ số khác nhau (DI, SPD, AOD và EOD), trong đó mức cải thiện đặc biệt rõ rệt đối với các chỉ số $|1 - DI|$ và SPD. Bên cạnh đó, FaireduPlus đạt hiệu quả cao nhất khi được kết hợp với CTGAN, đồng thời vẫn duy trì lợi thế tổng thể khi sử dụng phương pháp sinh dữ liệu dựa trên LLM, mặc dù mức độ cải thiện trong trường hợp này kém hơn so với CTGAN.

Bảng 4.2: So sánh chỉ số công bằng của FaireduPlus với các cấu hình đối sánh khi sử dụng CTGAN

Chỉ số công bằng		[I-DI]				SPD				AOD				EOD					
Mô hình	TT nhạy cảm	F.Plus	R.W	F.S	F.edu	F.Plus	R.W	F.S	F.edu	F.Plus	R.W	F.S	F.edu	F.Plus	R.W	F.S	F.edu		
LR	Oulad	GT	0.063	0.253	0.027	0.017	0.003	0.083	0.014	0.006	0.017	0.081	0.010	0.003	0.002	0.116	0.038	0.018	
		K.tật	0.114	0.706	0.678	0.044	0.018	0.279	0.365	0.016	0.005	0.267	0.347	0.008	0.040	0.294	0.349	0.034	
	Std.P	GT	0.099	0.142	0.142	0.142	0.043	0.068	0.068	0.068	0.005	0.005	0.005	0.005	0.048	0.048	0.048	0.048	
		SK	0.072	0.098	0.098	0.098	0.023	0.049	0.049	0.049	0.010	0.010	0.010	0.010	0.031	0.031	0.031	0.031	
	Std.D	GT	0.319	0.401	0.607	0.215	0.164	0.209	0.278	0.122	0.084	0.080	0.144	0.002	0.026	0.047	0.076	0.016	
		Nợ	0.397	1.52	0.887	0.305	0.172	0.421	0.314	0.155	0.053	0.092	0.015	0.079	0.043	0.033	0.045	0.069	
	DNU	GT	1.110	0.073	0.341	0.475	0.312	0.068	0.254	0.258	0.092	0.250	0.089	0.086	0.267		0.196	0.173	
		Tuổi	0.328	0.087	0.231	0.000	0.042	0.083	0.188	0.000	0.071	0.133	0.117	0.032	0.034		0.168	0.064	
		KV	0.123	0.023	0.159	0.023	0.017	0.022	0.136	0.013	0.004	0.133	0.283	0.018	0.009		0.099	0.035	
	RF	Oulad	GT	0.010	0.383	0.049	0.183	0.003	0.088	0.023	0.067	0.001	0.086	0.019	0.063	0.051	0.111	0.034	0.082
			K.tật	0.093	0.613	0.331	0.174	0.019	0.177	0.162	0.070	0.005	0.164	0.141	0.045	0.001	0.169	0.132	0.024
		Std.P	GT	0.110	0.142	0.117	0.117	0.054	0.068	0.056	0.056	0.006	0.005	0.006	0.006	0.048	0.048	0.068	0.068
SK			0.069	0.098	0.069	0.069	0.034	0.049	0.034	0.034	0.001	0.010	0.022	0.022	0.009	0.031	0.055	0.055	
Std.D		GT	0.549	0.411	0.505	0.386	0.250	0.219	0.257	0.190	0.113	0.101	0.136	0.057	0.055	0.050	0.069	0.029	
		Nợ	0.248	0.84	0.692	1.659	0.114	0.323	0.287	0.408	0.020	0.041	0.017	0.107	0.019	0.042	0.039	0.034	
DNU		GT	0.311	0.513	0.967	0.553	0.237	0.339	0.492	0.356	0.071	0.065	0.206	0.075	0.118	0.255	0.412	0.275	
		Tuổi	0.062	0.229	0.379	0.206	0.042	0.167	0.229	0.146	0.013	0.149	0.106	0.138	0.025	0.098	0.213	0.075	
		KV	0.038	0.145	0.064	0.012	0.029	0.111	0.037	0.008	0.003	0.215	0.031	0.146	0.006	0.096	0.061	0.041	
DT		Oulad	GT	0.007	0.007	0.058	0.169	0.003	0.003	0.027	0.061	0.001	0.001	0.023	0.058	0.008	0.007	0.037	0.075
			K.tật	0.067	0.204	0.2	0.134	0.026	0.088	0.097	0.053	0.003	0.068	0.075	0.033	0.018	0.065	0.060	0.026
		Std.P	GT	0.135	0.159	0.159	0.142	0.066	0.078	0.078	0.068	0.005	0.015	0.015	0.005	0.007	0.028	0.028	0.048
	SK		0.031	0.077	0.077	0.098	0.017	0.04	0.04	0.049	0.010	0.022	0.022	0.010	0.029	0.007	0.022	0.031	
	Std.D	GT	0.547	0.638	0.505	0.285	0.239	0.265	0.248	0.148	0.097	0.127	0.119	0.019	0.037	0.047	0.050	0.006	
		Nợ	0.349	1.237	0.699	1.010	0.161	0.342	0.279	0.326	0.039	0.133	0.000	0.097	0.008	0.219	0.042	0.131	
	DNU	GT	0.372	0.21	1.269	0.180	0.271	0.139	0.559	0.122	0.105	0.035	0.245	0.008	0.157	0.055	0.490	0.016	
		Tuổi	0.108	0.031	0.423	0.125	0.083	0.021	0.229	0.083	0.007	0.037	0.110	0.089	0.014	0.125	0.220	0.179	
		KV	0.109	0.127	0.112	0.465	0.089	0.079	0.059	0.244	0.064	0.105	0.042	0.129	0.129	0.123	0.085	0.257	
	BM	Oulad	GT	0.115	0.276	0.01	0.360	0.030	0.096	0.005	0.124	0.029	0.094	0.001	0.121	0.029	0.113	0.014	0.136
			K.tật	0.168	0.522	0.615	0.410	0.046	0.217	0.331	0.172	0.031	0.201	0.309	0.159	0.027	0.203	0.298	0.176
		Std.P	GT	0.098	0.201	0.151	0.176	0.051	0.094	0.076	0.082	0.005	0.036	0.023	0.026	0.048	0.041	0.014	0.062
SK			0.011	0.045	0.016	0.017	0.006	0.023	0.009	0.009	0.010	0.029	0.060	0.047	0.031	0.080	0.057	0.057	
Std.D		GT	0.258	0.32	0.14	0.424	0.156	0.154	0.198	0.170	0.061	0.025	0.069	0.043	0.021	0.019	0.036	0.048	
		Nợ	2.067	1.516	1.539	1.730	0.401	0.373	0.355	0.346	0.157	0.114	0.129	0.066	0.164	0.089	0.164	0.025	
DNU		GT	0.195	1.269	0.388	0.037	0.146	0.559	0.112	0.031	0.110	0.245	0.033	0.238	0.165	0.490	0.067	0.102	
		Tuổi	0.029	0.125	0.2	0.077	0.021	0.063	0.063	0.063	0.028	0.010	0.021	0.036	0.016	0.020	0.041	0.140	
		KV	0.042	0.026	0.163	0.034	0.030	0.012	0.054	0.028	0.013	0.001	0.037	0.036	0.041	0.003	0.073	0.006	
NN		Oulad	GT	0.019	0.487	0.039	0.047	0.005	0.13	0.02	0.017	0.008	0.128	0.017	0.014	0.009	0.164	0.044	0.035
			K.tật	0.152	0.567	0.63	0.010	0.032	0.196	0.346	0.004	0.055	0.178	0.326	0.020	0.088	0.173	0.323	0.044
		Std.P	GT	0.086	0.142	0.135	0.110	0.043	0.068	0.066	0.054	0.013	0.005	0.001	0.009	0.001	0.048	0.028	0.048
	SK		0.077	0.098	0.102	0.073	0.04	0.049	0.051	0.037	0.001	0.010	0.010	0.022	0.007	0.031	0.007	0.031	
	Std.D	GT	0.423	0.505	0.578	0.451	0.226	0.24	0.247	0.220	0.099	0.105	0.109	0.085	0.054	0.057	0.074	0.061	
		Nợ	0.552	1.279	0.815	1.279	0.255	0.374	0.276	0.374	0.005	0.131	0.032	0.090	0.036	0.158	0.081	0.058	
	DNU	GT	0.77	0.439	0.513	0.186	0.261	0.305	0.339	0.092	0.083	0.007	0.065	0.084	0.227	0.235	0.255	0.169	
		Tuổi	0.167	0.167	0.333	0.043	0.063	0.125	0.25	0.021	0.094	0.006	0.199	0.040	0.011	0.122	0.199	0.081	
		KV	0.221	0.009	0.056	0.404	0.063	0.007	0.04	0.154	0.031	0.061	0.174	0.080	0.061	0.012	0.015	0.161	
	Win/Tie			34/1	34/1	27/0		35/1	38/1	27/0		30/3	30/3	29/0		28/3	33/4	32/0	
	Lose			10	10	18		9	6	18		12	12	16		14	8	13	

4.4.2. Tác động của FairEduPlus đến hiệu suất dự đoán

Để đánh giá tác động của FaireduPlus đến hiệu suất của mô hình, luận án đã tiến hành so sánh các chỉ số hiệu suất, bao gồm “độ chuẩn xác” (ACC) và “độ hồi tưởng” (Recall) của FaireduPlus với các phương pháp đối sánh, bao gồm: Reweighting, FairSMOTE và Fairedu. Kết quả chi tiết được trình bày trong Bảng 4.4.

Bảng 4.3: So sánh chỉ số công bằng của FaireduPlus với các cấu hình đối sánh khi sử dụng LLM

Chỉ số công bằng			1-DI				SPD				AOD				EOD			
Model	Dataset	S.var	F.Plus	R.W	F.S	F.edu	F.Plus	R.W	F.S	F.edu	F.Plus	R.W	F.S	F.edu	F.Plus	R.W	F.S	F.edu
LR	Oulad	GT	0.017	0.253	0.027	0.017	0.033	0.083	0.014	0.006	0.003	0.081	0.010	0.003	0.096	0.116	0.038	0.018
		K.tật	0.044	0.706	0.678	0.044	0.293	0.279	0.365	0.016	0.008	0.267	0.347	0.008	0.397	0.294	0.349	0.034
	Std.P	GT	0.198	0.142	0.142	0.142	0.021	0.068	0.068	0.068	0.018	0.005	0.005	0.005	0.030	0.048	0.048	0.048
		SK	0.074	0.098	0.098	0.098	0.017	0.049	0.049	0.049	0.022	0.010	0.010	0.010	0.045	0.031	0.031	0.031
	Std.D	GT	0.569	0.401	0.607	0.215	0.268	0.209	0.278	0.122	0.119	0.080	0.144	0.002	0.077	0.047	0.076	0.016
		Ng	0.442	1.52	0.887	0.305	0.461	0.421	0.314	0.155	0.078	0.092	0.015	0.079	0.059	0.033	0.045	0.069
	DNU	GT	0.281	0.073	0.341	0.475	0.051	0.068	0.254	0.258	0.167	0.250	0.089	0.086	0.013		0.196	0.173
		Tuổi	0.267	0.087	0.231	0.000	0.063	0.083	0.188	0.000	0.003	0.133	0.117	0.032	0.015		0.168	0.064
		KV	0.871	0.023	0.159	0.023	0.025	0.022	0.136	0.013	0.063	0.133	0.283	0.018	0.037		0.099	0.035
	RF	Oulad	GT	0.141	0.383	0.049	0.183	0.031	0.088	0.023	0.067	0.050	0.086	0.019	0.063	0.059	0.111	0.034
K.tật			0.12	0.613	0.331	0.174	0.206	0.177	0.162	0.070	0.027	0.164	0.141	0.045	0.290	0.169	0.132	0.024
Std.P		GT	0.117	0.142	0.117	0.117	0.029	0.068	0.056	0.056	0.005	0.005	0.006	0.006	0.021	0.048	0.068	0.068
		SK	0.069	0.098	0.069	0.069	0.011	0.049	0.034	0.034	0.010	0.010	0.022	0.022	0.059	0.031	0.055	0.055
Std.D		GT	0.528	0.411	0.505	0.386	0	0.219	0.257	0.190	0.098	0.101	0.136	0.057	0.052	0.050	0.069	0.029
		Ng	1.735	0.84	0.692	1.659	0	0.323	0.287	0.408	0.107	0.041	0.017	0.107	0.052	0.042	0.039	0.034
DNU		GT	0.264	0.513	0.967	0.553	0.068	0.339	0.492	0.356	0.065	0.065	0.206	0.075	0.026	0.255	0.412	0.275
		Tuổi	0.05	0.229	0.379	0.206	0.229	0.167	0.229	0.146	0.037	0.149	0.106	0.138	0.069	0.098	0.213	0.075
		KV	2.174	0.145	0.064	0.012	0.093	0.111	0.037	0.008	0.273	0.215	0.031	0.146	0.035	0.096	0.061	0.041
DT		Oulad	GT	0.169	0.007	0.058	0.169	0.003	0.003	0.027	0.061	0.058	0.001	0.023	0.058	0.010	0.007	0.037
	K.tật		0.134	0.204	0.2	0.134	0.088	0.088	0.097	0.053	0.033	0.068	0.075	0.033	0.214	0.065	0.060	0.026
	Std.P	GT	0.121	0.159	0.159	0.142	0.009	0.078	0.078	0.068	0.005	0.015	0.015	0.005	0.025	0.028	0.028	0.048
		SK	0.017	0.077	0.077	0.098	0.028	0.04	0.04	0.049	0.010	0.022	0.022	0.010	0.039	0.007	0.022	0.031
	Std.D	GT	0.616	0.638	0.505	0.285	0.265	0.265	0.248	0.148	0.115	0.127	0.119	0.019	0.047	0.047	0.050	0.006
		Ng	1.503	1.237	0.699	1.010	0.342	0.342	0.279	0.326	0.106	0.133	0.000	0.097	0.219	0.219	0.042	0.131
	DNU	GT	0.157	0.21	1.269	0.180	0.102	0.139	0.559	0.122	0.088	0.035	0.245	0.008	0.054	0.055	0.490	0.016
		Tuổi	0	0.031	0.423	0.125	0.104	0.021	0.229	0.083	0.126	0.037	0.110	0.089	0.144	0.125	0.220	0.179
		KV	0.012	0.127	0.112	0.465	0.14	0.079	0.059	0.244	0.285	0.105	0.042	0.129	0.190	0.123	0.085	0.257
	BM	Oulad	GT	0.36	0.276	0.01	0.360	0.089	0.096	0.005	0.124	0.121	0.094	0.001	0.121	0.109	0.113	0.014
K.tật			0.41	0.522	0.615	0.410	0.124	0.217	0.331	0.172	0.159	0.201	0.309	0.159	0.222	0.203	0.298	0.176
Std.P		GT	0.029	0.201	0.151	0.176	0.061	0.094	0.076	0.082	0.053	0.036	0.023	0.026	0.041	0.041	0.014	0.062
		SK	0.04	0.045	0.016	0.017	0.06	0.023	0.009	0.009	0.073	0.029	0.060	0.047	0.027	0.080	0.057	0.057
Std.D		GT	0.467	0.32	0.14	0.424	0.231	0.154	0.198	0.170	0.050	0.025	0.069	0.043	0.019	0.019	0.036	0.048
		Ng	1.344	1.516	1.539	1.730	0.411	0.373	0.355	0.346	0.041	0.114	0.129	0.066	0.198	0.089	0.164	0.025
DNU		GT	0.186	1.269	0.388	0.037	0.342	0.559	0.112	0.031	0.096	0.245	0.033	0.238	0.226	0.490	0.067	0.102
		Tuổi	0.2	0.125	0.2	0.077	0.063	0.063	0.063	0.063	0.094	0.010	0.021	0.036	0.059	0.020	0.041	0.140
		KV	0.411	0.026	0.163	0.034	0.009	0.012	0.054	0.028	0.092	0.001	0.037	0.036	0.009	0.003	0.073	0.006
NN		Oulad	GT	0.111	0.487	0.039	0.047	0.047	0.13	0.02	0.017	0.038	0.128	0.017	0.014	0.128	0.164	0.044
	K.tật		0.016	0.567	0.63	0.010	0.25	0.196	0.346	0.004	0.017	0.178	0.326	0.020	0.342	0.173	0.323	0.044
	Std.P	GT	0.034	0.142	0.135	0.110	0.003	0.068	0.066	0.054	0.027	0.005	0.001	0.009	0.032	0.048	0.028	0.048
		SK	0.022	0.098	0.102	0.073	0.014	0.049	0.051	0.037	0.042	0.010	0.010	0.022	0.083	0.031	0.007	0.031
	Std.D	GT	0.497	0.505	0.578	0.451	0.24	0.24	0.247	0.220	0.107	0.105	0.109	0.085	0.056	0.057	0.074	0.061
		Ng	0.963	1.279	0.815	1.279	0.374	0.374	0.276	0.374	0.020	0.131	0.032	0.090	0.157	0.158	0.081	0.058
	DNU	GT	0.18	0.439	0.513	0.186	0.136	0.305	0.339	0.092	0.014	0.007	0.065	0.084	0.046	0.235	0.255	0.169
		Tuổi	0.091	0.167	0.333	0.043	0	0.125	0.25	0.021	0.056	0.006	0.199	0.040	0.026	0.122	0.199	0.081
		KV	0.953	0.009	0.056	0.404	0.02	0.007	0.04	0.154	0.154	0.061	0.174	0.080	0.009	0.012	0.015	0.161
	Win/Tie				30/0	25/3	24/0		25/7	28/2	22/0		23/0	25/0	17/0		19/4	22/0
Lose				15	17	21		13	15	23		22	20	28		19	23	22

Với cấu hình sử dụng CTGAN, mặc dù FaireduPlus đạt được kết quả vượt trội về công bằng, phương pháp này vẫn duy trì hiệu suất dự đoán ở mức cạnh tranh. Đối với chỉ số ACC, FaireduPlus đạt tỷ lệ thắng 40.0% (8/20) so với Reweighing, 50.0% (10/20) so với FairSMOTE và 45.0% (9/20) so với Fairedu. Đối với Recall, lợi thế trở nên rõ rệt hơn. FaireduPlus ghi nhận tỷ lệ thắng 60.0% (12/20) so với Reweighing, 45.0% (9/20) so với FairSMOTE và 55.0% (11/20) so với Fairedu.

Với cấu hình sử dụng LLM, FaireduPlus thể hiện hiệu suất vượt trội so với các cấu hình đối sánh. Cụ thể, đối với ACC, phương pháp này đạt tỷ lệ thắng 65.0% (13/20) so với Reweighing, 70.0% (14/20) so với FairSMOTE và 65.0%

(13/20) so với Fairedu. Lợi thế này càng trở nên rõ rệt hơn đối với Recall, khi FaireduPlus đạt tỷ lệ thắng 75.0% (15/20) so với Reweighing, 80.0% (16/20) so với FairSMOTE và 85.0% (17/20) so với Fairedu.

Hơn nữa, tỷ lệ thắng cao hơn đối với Recall cho thấy FaireduPlus đặc biệt phù hợp với các mô hình học máy trong bối cảnh giáo dục, nơi việc xác định chính xác các học sinh có nguy cơ thường quan trọng hơn so với những cải thiện nhỏ về độ chính xác tổng thể.

Bảng 4.4: So sánh các chỉ số hiệu suất (Acc và Recall) của FaireduPlus với các cấu hình thực nghiệm khác nhau

KT sinh DL		CTGAN								LLM							
Hiệu suất		ACC				Recall				ACC				Recall			
Mô hình	Dữ liệu	F.Plus	R.W	F.S	F.edu	F.Plus	R.W	F.S	F.edu	F.Plus	R.W	F.S	F.edu	F.Plus	R.W	F.S	F.edu
LR	Oulad	0.567	0.588	0.585	0.577	0.442	0.453	0.591	0.442	0.572	0.588	0.585	0.577	0.602	0.453	0.591	0.442
	Std.P	0.936	0.943	0.943	0.943	0.919	0.919	0.919	0.919	0.805	0.943	0.943	0.943	0.977	0.919	0.919	0.919
	Std.D	0.736	0.825	0.833	0.798	0.942	0.968	0.956	0.933	0.839	0.825	0.833	0.798	0.959	0.968	0.956	0.933
	DNU	0.495	0.938	0.797	0.688	0.464	1.000	0.821	0.643	0.891	0.938	0.797	0.688	0.982	1.000	0.821	0.643
RF	Oulad	0.574	0.575	0.578	0.584	0.487	0.338	0.554	0.478	0.569	0.575	0.578	0.584	0.595	0.338	0.554	0.478
	Std.P	0.936	0.943	0.936	0.936	0.919	0.919	0.907	0.907	0.853	0.943	0.936	0.936	0.977	0.919	0.907	0.907
	Std.D	0.840	0.797	0.801	0.831	0.921	0.959	0.962	0.933	0.813	0.797	0.801	0.831	0.950	0.959	0.962	0.933
	DNU	0.906	0.781	0.672	0.766	0.750	0.768	0.625	0.750	0.901	0.781	0.672	0.766	0.982	0.768	0.625	0.750
DT	Oulad	0.580	0.582	0.576	0.580	0.468	0.504	0.553	0.468	0.571	0.582	0.576	0.580	0.617	0.504	0.553	0.468
	Std.P	0.943	0.936	0.936	0.943	0.942	0.930	0.930	0.919	0.777	0.936	0.936	0.943	0.953	0.930	0.930	0.919
	Std.D	0.849	0.839	0.821	0.822	0.918	0.912	0.959	0.927	0.839	0.839	0.821	0.822	0.912	0.912	0.959	0.927
	DNU	0.875	0.766	0.609	0.813	0.911	0.750	0.554	0.786	0.687	0.766	0.609	0.813	0.964	0.750	0.554	0.786
BM	Oulad	0.571	0.576	0.582	0.578	0.479	0.468	0.591	0.479	0.580	0.576	0.582	0.578	0.484	0.468	0.591	0.479
	Std.P	0.943	0.917	0.873	0.911	0.930	0.895	0.884	0.884	0.847	0.917	0.873	0.911	0.977	0.895	0.884	0.884
	Std.D	0.824	0.815	0.819	0.822	0.912	0.886	0.860	0.825	0.848	0.815	0.819	0.822	0.892	0.886	0.860	0.825
	DNU	0.703	0.554	0.422	0.859	0.554	0.609	0.339	0.893	0.599	0.554	0.422	0.859	0.643	0.609	0.339	0.893
NN	Oulad	0.567	0.583	0.583	0.580	0.469	0.405	0.603	0.451	0.578	0.583	0.583	0.580	0.582	0.405	0.603	0.451
	Std.P	0.936	0.943	0.943	0.936	0.930	0.919	0.930	0.919	0.811	0.943	0.943	0.936	0.965	0.919	0.930	0.919
	Std.D	0.822	0.837	0.846	0.840	0.965	0.953	0.921	0.956	0.841	0.837	0.846	0.840	0.953	0.953	0.921	0.956
	DNU	0.453	0.781	0.781	0.257	0.393	0.786	0.768	0.554	0.870	0.781	0.781	0.257	0.946	0.786	0.768	0.554
Thắng/Hòa			8/0	10/0	9/2		12/0	9/0	11/4		13/0	14/0	13/0		15/0	16/0	17/0
Thua			12	10	9		8	11	5		7	6	7		5	4	3

4.4.3. Đánh giá mối quan hệ đánh đổi giữa công bằng và hiệu suất

Để trả đánh giá mối quan hệ đánh đổi giữa công bằng và hiệu suất của FaireduPlus, luận án tiến hành tính toán *chỉ số cải thiện công bằng*, *chỉ số suy giảm hiệu suất* và *chỉ số đánh đổi tổng hợp*, như đã được định nghĩa trong Mục 4.3.3, nhằm đánh giá mối quan hệ đánh đổi giữa công bằng và hiệu suất trên các cấu hình can thiệp khác nhau.

Các Bảng 4.5 và 4.6 trình bày chỉ số đánh đổi tổng hợp (Δ_{trade_off}) cho bốn cấu hình trong hai thiết lập CTGAN và LLM. Giá trị càng cao thể hiện mức đánh đổi giữa công bằng và hiệu suất càng tốt.

Kết quả trong Bảng 4.5 cho thấy, chỉ số $\Delta^{ACC_{1-DI}|trade_off}$ của FaireduPlus đạt tỷ lệ thắng dao động từ 60.0% đến 75.6% (27–34 trên tổng số 45 phép so sánh). Tương tự, chỉ số $\Delta^{Recall_{1-DI}|trade_off}$ ghi nhận tỷ lệ thắng từ 60.0% đến 73.3% (27–33 trên tổng số 45 phép so sánh). Những kết quả này cho thấy FaireduPlus đạt được sự cân bằng tốt hơn giữa công bằng và hiệu suất so với các cấu hình đối sánh tương ứng.

Tương tự, kết quả trong Bảng 4.6 cho thấy, chỉ số $\Delta^{ACC_{sPD}|trade_off}$ của FaireduPlus đạt tỷ lệ thắng 66.7% (30/45) so với Reweighing, 66.7% (30/45) so với FairSMOTE và 53.3% (24/45) so với Fairedu. Tương tự, đối với chỉ số $\Delta^{Recall_{sPD}|trade_off}$, FaireduPlus đạt tỷ lệ thắng 68.9% (31/45) so với Reweighing, 66.7% (30/45) so với FairSMOTE và 55.6% (25/45) so với Fairedu.

Bên cạnh xu hướng chung này, một số cấu hình thực nghiệm ghi nhận giá trị của chỉ số đánh đổi biến động tương đối lớn do sự thay đổi đáng kể của một trong hai thành phần $\Delta_{fairness}$ hoặc $\Delta_{performance}$. Vì vậy, trong nghiên cứu này, chỉ số đánh đổi được sử dụng như một thước đo tổng hợp hỗ trợ và luôn được diễn giải kết hợp với các chỉ số công bằng và hiệu suất gốc nhằm đảm bảo đánh giá đầy đủ hiệu quả của phương pháp.

Nhìn chung, kết quả thực nghiệm cho thấy FaireduPlus đạt được sự đánh đổi thuận lợi giữa công bằng và hiệu suất dự đoán trong phần lớn các cấu hình đánh giá.

Bảng 4.5: Comparison of fairness-performance trade-off indices ($\Delta_{trade_off}^{ACC_|1-DI|}$ and $\Delta_{trade_off}^{Recall_|1-DI|}$) across four experimental configurations using CTGAN

Chỉ số công bằng			$ 1 - DI - ACC$				$ 1 - DI - Recall$			
Model	Dataset		F.Plus	R.W	F.S	F.edu	F.Plus	R.W	F.S	F.edu
LR	Oulad	Gen	75.8%	17.4%	90.6%	92.6%	72.7%	13.1%	116.2%	87.8%
		Dis	80.8%	2.7%	6.1%	92.0%	77.7%	-1.5%	31.6%	87.3%
	Std.P	Gen	76.3%	66.7%	66.7%	66.7%	76.8%	66.5%	66.5%	66.5%
		Heal	-7.9%	-46.3%	-46.3%	-46.8%	-7.5%	-46.5%	-46.5%	-47.0%
	Std.D	Gen	48.3%	48.8%	24.4%	68.3%	67.3%	60.2%	33.6%	79.0%
		Deb	38.7%	-88.4%	-9.9%	57.3%	57.7%	-77.0%	-0.7%	67.9%
	DNU	Gen	-1518.8%	1.1%	-394.8%	-597.0%	-1527.6%	-3.5%	-401.5%	-609.5%
		Tuổi	-197.7%	37.6%	-89.0%	76.7%	-206.5%	33.0%	-95.8%	64.3%
		BP	-44.1%	86.1%	-39.1%	58.0%	-52.9%	81.5%	-45.9%	45.6%
	RF	Oulad	Gen	-157.4%	-9721.4%	-1156.7%	-4601.8%	-155.3%	-9750.4%	-1141.4%
Dis			83.5%	-3.4%	44.3%	71.6%	85.5%	-32.5%	59.6%	70.1%
Std.P		Gen	25.7%	4.6%	20.9%	20.7%	25.5%	3.7%	19.5%	19.3%
		Heal	15.0%	-20.3%	15.0%	14.5%	14.9%	-21.1%	13.6%	13.0%
Std.D		Gen	14.9%	31.4%	17.1%	39.5%	17.9%	44.0%	29.5%	44.9%
		Deb	84.2%	37.1%	48.1%	-16.4%	87.2%	49.8%	60.5%	-11.0%
DNU		Gen	-344.5%	-644.3%	-1300.0%	-702.2%	-366.1%	-650.9%	-1309.1%	-708.9%
		Tuổi	48.9%	-93.1%	-220.3%	-76.9%	27.2%	-99.6%	-229.5%	-83.6%
		BP	66.1%	-33.4%	20.1%	72.3%	44.4%	-40.0%	11.0%	65.6%
DT		Oulad	Gen	77.4%	77.9%	-90.0%	-453.6%	67.6%	74.5%	-82.8%
	Dis		70.7%	10.7%	11.3%	41.3%	60.9%	7.3%	18.4%	31.4%
	Std.P	Gen	-1.0%	-19.9%	-19.9%	-6.5%	1.5%	-17.9%	-17.9%	-6.5%
		Heal	42.5%	-45.3%	-45.3%	-85.1%	45.0%	-43.4%	-43.4%	-85.1%
	Std.D	Gen	14.4%	-1.7%	17.8%	53.9%	16.4%	0.9%	28.0%	60.3%
		Deb	78.0%	11.5%	48.8%	26.1%	80.0%	14.1%	59.0%	32.5%
	DNU	Gen	-662.0%	-344.0%	-2517.9%	-277.8%	-665.2%	-351.5%	-2528.3%	-286.8%
		Tuổi	-54.3%	40.7%	-520.9%	-84.8%	-57.5%	33.1%	-531.3%	-93.8%
		BP	-198.1%	-258.6%	-235.3%	-1164.8%	-201.3%	-266.2%	-245.7%	-1173.9%
	BM	Oulad	Gen	59.4%	5.9%	96.9%	-22.3%	60.6%	4.0%	119.5%
Dis			41.5%	-77.4%	-107.9%	-39.2%	42.7%	-79.3%	-85.3%	-39.3%
Std.P		Gen	16.9%	-77.1%	-37.7%	-55.5%	16.5%	-78.6%	-35.6%	-57.6%
		Heal	88.4%	38.0%	73.7%	76.6%	87.9%	36.4%	75.7%	74.5%
Std.D		Gen	61.7%	51.6%	78.3%	37.4%	69.8%	57.7%	80.8%	35.4%
		Deb	-8.8%	19.0%	18.3%	8.7%	-1.0%	24.8%	20.5%	6.3%
DNU		Gen	-270.2%	-2181.1%	-640.3%	25.5%	-284.5%	-2173.7%	-648.1%	31.3%
		Tuổi	50.9%	-39.0%	-110.8%	30.1%	37.1%	-31.1%	-118.2%	36.6%
		BP	-43.2%	-14.4%	-408.4%	-5.3%	-57.0%	-6.6%	-415.9%	1.2%
NN		Oulad	Gen	93.3%	0.1%	92.0%	89.8%	100.2%	-10.1%	125.6%
	Dis		70.4%	-0.1%	-11.2%	97.7%	77.2%	-10.3%	22.5%	98.2%
	Std.P	Gen	38.8%	0.1%	5.0%	21.6%	40.7%	0.1%	6.3%	22.3%
		Heal	21.0%	0.3%	-3.7%	24.7%	22.9%	0.3%	-2.5%	25.3%
	Std.D	Gen	14.5%	0.0%	-13.4%	11.0%	18.1%	0.7%	-17.2%	11.6%
		Deb	55.0%	0.0%	37.3%	0.4%	58.7%	0.6%	33.5%	1.0%
	DNU	Gen	-117.4%	0.0%	-16.9%	-9.5%	-6.1%	238.7%	214.1%	196.2%
		Tuổi	-42.2%	-0.2%	-99.8%	6.8%	69.1%	238.5%	131.2%	212.5%
		BP	-2317.8%	3.2%	-502.0%	-4310.8%	-2206.4%	241.9%	-271.0%	-4105.1%
	Win/Tie				32/0	34/0	27/0		33/0	32/0
Lose				13	11	18		12	13	18

Bảng 4.6: So sánh các chỉ số đánh đổi giữa công bằng và hiệu suất ($\Delta_{trade_off}^{ACC_SPD}$ và $\Delta_{trade_off}^{Recall_SPD}$) trên bốn cấu hình thực nghiệm sử dụng phương pháp sinh dữ liệu tổng hợp dựa trên LLM.

Chỉ số công bằng			SPD-ACC				SPD-Recall			
Model	Dataset		F.Plus	R.W	F.S	F.edu	F.Plus	R.W	F.S	F.edu
LR	Oulad	Gen	65.5%	20.1%	85.9%	92.2%	95.5%	15.9%	111.5%	87.4%
		Dis	0.5%	7.8%	-21.1%	92.7%	30.5%	3.6%	4.5%	88.0%
	Std.P	Gen	59.5%	14.8%	14.8%	14.6%	80.5%	14.6%	14.6%	14.4%
		Heal	55.7%	13.4%	13.4%	14.1%	76.7%	13.1%	13.1%	13.9%
	Std.D	Gen	34.4%	47.1%	31.2%	65.0%	43.2%	58.5%	40.4%	75.6%
		Deb	-61.0%	-48.8%	-10.6%	40.7%	-52.3%	-37.3%	-1.4%	51.4%
	DNU	Gen	-4.6%	-34.1%	-429.4%	-449.0%	-5.9%	-38.8%	-436.2%	-461.5%
		Tuổi	9.4%	-13.9%	-179.6%	76.7%	8.2%	-18.6%	-186.4%	64.3%
		BP	52.3%	63.1%	-167.7%	51.7%	51.0%	58.5%	-174.5%	39.2%
	RF	Oulad	Gen	-108.6%	-487.5%	-53.7%	-344.5%	-83.2%	-516.6%	-38.3%
Dis			17.9%	30.3%	36.6%	73.5%	43.3%	1.2%	52.0%	72.0%
Std.P		Gen	54.5%	14.1%	28.8%	28.5%	69.7%	13.3%	27.4%	27.1%
		Heal	72.1%	13.8%	39.9%	39.3%	87.2%	13.0%	38.5%	37.9%
Std.D		Gen	98.4%	32.0%	21.3%	44.7%	108.0%	44.6%	33.7%	50.1%
		Deb	98.4%	-34.9%	-19.7%	-65.4%	108.0%	-22.3%	-7.3%	-60.0%
DNU		Gen	-8.5%	-438.2%	-685.3%	-465.9%	-6.4%	-444.8%	-694.4%	-472.6%
		Tuổi	-179.8%	-117.9%	-204.2%	-94.0%	-177.7%	-124.4%	-213.4%	-100.7%
		BP	-64.2%	-108.0%	7.9%	68.3%	-62.1%	-114.6%	-1.3%	61.6%
DT		Oulad	Gen	82.2%	84.1%	-50.4%	-240.4%	102.5%	80.7%	-43.3%
	Dis		11.7%	13.6%	3.5%	48.0%	32.0%	10.2%	10.7%	38.1%
	Std.P	Gen	71.6%	-2.1%	-2.1%	11.5%	93.1%	-0.2%	-0.2%	11.5%
		Heal	33.4%	29.1%	29.1%	14.4%	54.9%	31.0%	31.0%	14.4%
	Std.D	Gen	22.6%	22.6%	25.5%	55.7%	25.2%	25.2%	35.7%	62.0%
		Deb	-40.7%	-40.7%	-16.6%	-36.2%	-38.1%	-38.1%	-6.3%	-29.9%
	DNU	Gen	-137.3%	-205.9%	-1098.0%	-165.4%	-114.6%	-213.4%	-1108.4%	-174.5%
		Tuổi	-73.4%	53.7%	-260.5%	-30.2%	-50.7%	46.1%	-270.9%	-39.3%
		BP	-189.0%	-65.3%	-44.7%	-370.8%	-166.3%	-72.9%	-55.1%	-379.9%
	BM	Oulad	Gen	15.3%	8.0%	95.5%	-18.5%	16.0%	6.1%	118.1%
Dis			-18.0%	-107.1%	-214.6%	-64.0%	-17.3%	-109.0%	-192.0%	-64.1%
Std.P		Gen	6.5%	-32.8%	-12.1%	-16.2%	21.7%	-34.3%	-10.0%	-18.3%
		Heal	5.8%	67.4%	82.9%	87.8%	21.0%	65.9%	85.0%	85.6%
Std.D		Gen	6.9%	34.6%	17.0%	28.8%	9.6%	40.7%	19.4%	26.7%
		Deb	-15.2%	-8.3%	-2.6%	0.3%	-12.7%	-2.5%	-0.5%	-2.0%
DNU		Gen	-381.6%	-668.4%	-101.2%	50.6%	-375.3%	-661.0%	-109.1%	56.5%
		Tuổi	13.8%	9.1%	-4.6%	41.4%	20.6%	17.0%	-12.1%	47.9%
		BP	50.5%	41.8%	-27.6%	52.9%	57.3%	49.7%	-35.1%	59.4%
NN		Oulad	Gen	62.4%	-1.6%	84.3%	86.0%	92.3%	-11.9%	118.0%
	Dis		-41.6%	-10.4%	-94.9%	97.4%	-11.7%	-20.6%	-61.2%	97.9%
	Std.P	Gen	81.7%	0.2%	3.2%	19.4%	100.6%	0.2%	4.4%	20.1%
		Heal	57.3%	-0.8%	-4.9%	23.0%	76.2%	-0.9%	-3.7%	23.6%
	Std.D	Gen	0.3%	-0.2%	-2.0%	8.6%	0.5%	0.5%	-5.8%	9.2%
		Deb	0.4%	-0.1%	27.2%	0.4%	0.6%	0.6%	23.4%	1.0%
	DNU	Gen	66.8%	0.0%	-11.1%	2.9%	363.2%	238.7%	219.9%	208.6%
		Tuổi	111.4%	0.0%	-100.0%	16.3%	407.8%	238.7%	131.0%	221.9%
		BP	-189.6%	-5.4%	-502.0%	-2283.7%	106.8%	233.3%	-271.0%	-2078.1%
	Win/Tie				30/0	30/0	24/0		31/0	30/0
Lose				15	15	21		14	15	20

4.5. Thảo luận

4.5.1. Phân tích và tổng hợp các phát hiện chính

Phần này tổng hợp và diễn giải các kết quả thực nghiệm đã trình bày ở Mục 4.4, nhằm đánh giá hiệu quả mà phương pháp FaireduPlus đạt được theo các mục tiêu nghiên cứu, bao gồm: (i) mức độ cải thiện công bằng, (ii) tác động của FaireduPlus đến hiệu suất của mô hình, và (iii) sự đánh đổi giữa công bằng và hiệu suất dự đoán của phương pháp FaireduPlus. Các kết quả thực nghiệm cho thấy *FaireduPlus* không chỉ đạt được hiệu quả cải thiện công bằng mà còn duy trì tính ổn định và khả năng tổng quát hóa trong nhiều bối cảnh dữ liệu khác nhau.

(i) *Mức độ cải thiện công bằng.* Kết quả thực nghiệm cho thấy FaireduPlus đạt được sự cải thiện công bằng một cách nhất quán trên nhiều chỉ số khác nhau, bao gồm DI, SPD, AOD và EOD. Trong đó, mức cải thiện đặc biệt rõ rệt đối với các chỉ số $|1 - DI|$ và SPD, cho thấy phương pháp có khả năng giảm thiểu đáng kể sự chênh lệch phân phối giữa các nhóm nhạy cảm. Ngoài ra, FaireduPlus đạt hiệu quả cao nhất khi được kết hợp với phương pháp sinh dữ liệu dựa trên CTGAN, trong khi cấu hình sử dụng LLM vẫn duy trì lợi thế tổng thể nhưng mức cải thiện có kém hơn. Điều này gợi ý rằng cơ chế tái tạo phân phối dữ liệu của CTGAN có thể hỗ trợ tốt hơn cho mục tiêu công bằng, trong khi LLM mang lại sự ổn định cao hơn trong quá trình sinh dữ liệu.

(ii) *Tác động đến hiệu suất mô hình.* Mặc dù tập trung vào cải thiện công bằng, FaireduPlus vẫn duy trì hiệu suất dự đoán ở mức cạnh tranh so với các phương pháp đối sánh trên cả hai chỉ số Acc và Recall. Đáng chú ý, lợi thế của phương pháp trở nên rõ rệt hơn đối với chỉ số Recall, đặc biệt trong cấu hình sử dụng LLM. Điều này cho thấy FaireduPlus có khả năng cải thiện việc phát hiện các trường hợp dương (ví dụ: sinh viên có nguy cơ), một yếu tố có ý nghĩa thực tiễn quan trọng trong các bài toán giáo dục, nơi việc bỏ sót các đối tượng cần hỗ trợ có thể dẫn đến hậu quả đáng kể.

(iii) *Sự đánh đổi giữa công bằng và hiệu suất.* Kết quả thực nghiệm cho thấy FaireduPlus đạt được sự cân bằng tốt giữa công bằng và hiệu suất dự đoán. So với các phương pháp đối sánh, FaireduPlus không chỉ cải thiện các chỉ số công bằng mà còn duy trì hiệu suất mô hình ở mức cạnh tranh, từ đó đạt được mức đánh đổi thuận lợi hơn trong phần lớn các cấu hình thực nghiệm. Để hỗ trợ đánh giá đồng thời hai mục tiêu này, luận án đề xuất chỉ số đánh đổi tổng hợp, phản ánh mức cải thiện công bằng so với sự thay đổi của hiệu suất mô hình. Chỉ số có ưu điểm là đơn giản, dễ tính toán và hỗ trợ so sánh giữa các phương pháp trong cùng điều kiện thực nghiệm. Tuy nhiên, chỉ số hiện mới phản ánh xu hướng đánh đổi tương đối và cần được tiếp tục kiểm chứng trên nhiều bộ dữ liệu và kịch bản thực nghiệm khác nhau nhằm đánh giá tính tổng quát và khả năng áp dụng trong thực tiễn.

Tổng hợp lại, các phát hiện cho thấy FaireduPlus là một phương pháp toàn diện, không chỉ cải thiện công bằng một cách hiệu quả mà còn duy trì hiệu suất và tính ổn định của mô hình, đặc biệt phù hợp với các bài toán học máy trong lĩnh vực giáo dục, nơi yêu cầu đồng thời cả tính công bằng và độ tin cậy của hệ thống.

4.5.2. Hạn chế của nghiên cứu

Mặc dù FaireduPlus cho thấy hiệu quả trong việc cải thiện công bằng giao thoa đồng thời duy trì hiệu suất của mô hình, nghiên cứu vẫn còn tồn tại một số hạn chế liên quan đến độ tin cậy nội tại, độ tin cậy bên ngoài, giá trị cấu trúc và giá trị kết luận [137, 138, 139].

Độ tin cậy nội tại: Các biện pháp đã được thực hiện nhằm đảm bảo độ tin cậy nội tại, bao gồm chuẩn hóa cấu hình mô hình, sử dụng nhiều bộ dữ liệu và áp dụng quy trình huấn luyện thống nhất. Tuy nhiên, quá trình sinh dữ liệu tổng hợp bằng CTGAN và LLM vẫn mang tính ngẫu nhiên, do đó có thể tạo ra sự biến động nhất định giữa các lần thực nghiệm. Mặc dù các thí nghiệm đã được lặp lại với nhiều giá trị khởi tạo khác nhau nhằm giảm ảnh hưởng của yếu tố ngẫu nhiên, hiệu quả của FaireduPlus vẫn phụ thuộc vào chất lượng dữ liệu tổng hợp được sinh ra. Trong các trường hợp dữ liệu mất cân bằng nghiêm trọng hoặc các nhóm giao thoa có rất ít mẫu, chất lượng dữ liệu sinh có thể ảnh hưởng đến mức độ cải thiện công bằng của phương pháp. Ngoài ra, các siêu tham số của

CTGAN và LLM được lựa chọn dựa trên tài liệu tham khảo và quá trình hiệu chỉnh sơ bộ, chưa được tối ưu toàn diện, do đó vẫn có thể ảnh hưởng đến kết quả thực nghiệm.

Độ tin cậy bên ngoài: Nghiên cứu được đánh giá trên bốn bộ dữ liệu giáo dục thực tế với quy mô và đặc điểm nhân khẩu học khác nhau, bao gồm cả dữ liệu mở và dữ liệu thu thập từ nhiều quốc gia. Điều này góp phần nâng cao khả năng khái quát của phương pháp trong lĩnh vực giáo dục. Tuy nhiên, hiệu quả của FaireduPlus trên các lĩnh vực khác như tài chính, y tế hoặc tư pháp vẫn cần được kiểm chứng thêm. Bên cạnh đó, các thuộc tính nhạy cảm ngoài phạm vi nghiên cứu hiện tại, chẳng hạn như tôn giáo hoặc điều kiện kinh tế - xã hội, cũng cần được xem xét trong các nghiên cứu tiếp theo.

Giá trị cấu trúc: Công bằng là một khái niệm đa chiều và chịu ảnh hưởng của bối cảnh ứng dụng. Trong nghiên cứu này, bốn chỉ số công bằng phổ biến gồm DI, SPD, AOD và EOD được sử dụng để đánh giá công bằng theo nhóm, kết hợp với hai chỉ số hiệu suất là “độ chuẩn xác” và “độ hồi tưởng”. Đồng thời, luận án đề xuất chỉ số đánh đổi nhằm hỗ trợ đánh giá đồng thời giữa công bằng và hiệu suất. Mặc dù cách tiếp cận này phù hợp với nhiều nghiên cứu trước, nó vẫn chưa phản ánh đầy đủ các khía cạnh khác của công bằng như công bằng cá nhân, công bằng nhân quả hoặc công bằng theo thời gian. Ngoài ra, chỉ số đánh đổi hiện tại chủ yếu phản ánh xu hướng đánh đổi giữa công bằng và hiệu suất, đồng thời có thể kém ổn định khi giá trị gốc của chỉ số công bằng hoặc hiệu suất tiến gần 0 do sử dụng mức cải thiện tương đối. Trong tương lai, nghiên cứu sẽ tiếp tục hoàn thiện chỉ số này thông qua các kỹ thuật chuẩn hóa, xử lý giá trị ngoại lai và các công thức đánh đổi ổn định hơn.

Giá trị kết luận: Toàn bộ thí nghiệm được thực hiện trên các thư viện và bộ dữ liệu công khai; các cấu hình mô hình, tham số và quy trình thực nghiệm đều được ghi lại nhằm hỗ trợ khả năng tái lập. Tuy nhiên, do đặc tính phi xác định của các mô hình sinh dữ liệu, đặc biệt là LLM, việc tái lập hoàn toàn kết quả có thể xuất hiện những sai khác nhỏ. Ngoài ra, mặc dù FaireduPlus cho thấy hiệu quả cải thiện công bằng trong phần lớn các cấu hình thực nghiệm, vẫn tồn tại một số trường hợp phương pháp chưa vượt trội so với các phương pháp đối sánh. Điều này cho thấy hiệu quả của FaireduPlus còn phụ thuộc vào đặc điểm của dữ liệu, mức độ mất cân bằng giữa các nhóm giao thoa và chất lượng của kỹ thuật sinh dữ liệu được sử dụng.

4.5.3. Hàm ý thực tiễn

Các kết quả thực nghiệm của chương này mang lại một số hàm ý thực tiễn quan trọng đối với việc thiết kế, triển khai và đánh giá các hệ thống học máy hướng đến công bằng trong lĩnh vực giáo dục.

Thứ nhất, FairEduPlus cho thấy tính khả thi cao như một giải pháp tiên xử lý có thể triển khai trong các hệ thống thực tế, đặc biệt trong những bối cảnh yêu cầu đồng thời cải thiện công bằng và duy trì hiệu suất dự đoán. Kết quả thực nghiệm cho thấy FairEduPlus không nhằm tối ưu đơn lẻ từng chỉ số, mà đạt được sự cân bằng ổn định giữa các độ đo công bằng giao thoa và hiệu suất (“*độ chuẩn xác*”, “*độ hồi tưởng*”) trên nhiều bộ dữ liệu, mô hình học máy và cấu hình sinh dữ liệu khác nhau. Điều này đặc biệt phù hợp với các hệ thống hỗ trợ ra quyết định trong giáo dục, nơi việc suy giảm hiệu suất quá mức có thể làm giảm độ tin cậy và khả năng chấp nhận của người dùng cuối.

Thứ hai, các kết quả về số lượt thắng và chỉ số đánh đổi cho thấy việc kết hợp sinh dữ liệu tổng hợp với chiến lược phân tách nhóm đóng vai trò then chốt trong cải thiện công bằng giao thoa. So với các cấu hình chỉ sinh dữ liệu trực tiếp, các phương pháp có cân bằng tường minh giữa các nhóm con giúp đạt được nhiều trường hợp đánh đổi thuận lợi hơn giữa công bằng và hiệu suất. Hàm ý thực tiễn ở đây là các hệ thống học máy trong giáo dục nên ưu tiên các chiến lược sinh dữ liệu có kiểm soát theo nhóm, thay vì chỉ mở rộng dữ liệu một cách tổng quát.

Thứ ba, nghiên cứu cho thấy không tồn tại một kỹ thuật sinh dữ liệu tổng hợp duy nhất phù hợp cho mọi bối cảnh. CTGAN thường mang lại kết quả ổn định hơn, trong khi các phương pháp dựa trên mô hình ngôn ngữ lớn thể hiện tính linh hoạt cao nhưng nhạy cảm hơn với đặc điểm dữ liệu và cấu hình mô hình. Do đó, trong thực tiễn triển khai, việc lựa chọn kỹ thuật sinh dữ liệu cần dựa trên đánh giá thực nghiệm cụ thể, kết hợp với các bước hiệu chỉnh đặc trưng như trong FairEduPlus để tránh làm gia tăng thiên lệch ngoài mong muốn.

Cuối cùng, việc sử dụng dữ liệu tổng hợp trong FairEduPlus mang lại lợi ích bổ sung về mặt bảo vệ quyền riêng tư và tuân thủ quy định, đặc biệt trong bối cảnh dữ liệu giáo dục thường chứa nhiều thông tin nhạy cảm và khó chia sẻ.

Bằng cách giảm sự phụ thuộc trực tiếp vào dữ liệu gốc của các nhóm yếu thế, FairEduPlus không chỉ góp phần cải thiện công bằng mà còn hỗ trợ xây dựng các hệ thống học máy có trách nhiệm, minh bạch và phù hợp hơn với các yêu cầu pháp lý và đạo đức trong giáo dục hiện nay.

4.6. Tóm tắt chương

Chương này đã trình bày phương pháp *FaireduPlus* như một khuôn khổ tiền xử lý nhằm đảm bảo công bằng hai chiều cho dữ liệu dạng bảng. Phương pháp kết hợp hai hướng can thiệp bổ sung: (i) cân bằng phân phối dữ liệu giữa các nhóm giao thoa thông qua chiến lược phân tách và sinh dữ liệu tổng hợp (DPF), và (ii) loại bỏ sự phụ thuộc tiềm ẩn giữa các thuộc tính không nhạy cảm và các thuộc tính nhạy cảm thông qua cơ chế hiệu chỉnh đặc trưng kế thừa từ phương pháp *Fairedu*. Cách tiếp cận tích hợp này cho phép xử lý đồng thời các dạng thiên lệch xuất phát từ cả phân bố dữ liệu và cấu trúc đặc trưng.

Các thí nghiệm được tiến hành trên bốn bộ dữ liệu giáo dục với nhiều mô hình học máy và kỹ thuật sinh dữ liệu khác nhau cho thấy *FaireduPlus* mang lại cải thiện công bằng giao thoa một cách ổn định trên hầu hết các độ đo, đồng thời hạn chế đáng kể mức suy giảm hiệu suất dự đoán. So với các cấu hình chỉ áp dụng sinh dữ liệu tổng hợp hoặc chỉ hiệu chỉnh đặc trưng, *FaireduPlus* thể hiện khả năng cân bằng tốt hơn giữa hai mục tiêu vốn thường xung đột là công bằng và hiệu suất, đặc biệt khi xét theo các chỉ số đánh đổi tổng hợp được đề xuất trong chương.

Kết quả phân tích cũng cho thấy hiệu quả của *FaireduPlus* phụ thuộc vào đặc điểm dữ liệu và kỹ thuật sinh dữ liệu được sử dụng. Trong nhiều kịch bản, CTGAN mang lại tính ổn định cao hơn, trong khi các phương pháp dựa trên mô hình ngôn ngữ lớn thể hiện tính linh hoạt nhưng cần được kết hợp với các bước hiệu chỉnh phù hợp để tránh làm gia tăng thiên lệch. Những quan sát này nhấn mạnh vai trò của việc lựa chọn chiến lược can thiệp dựa trên thực nghiệm và bối cảnh ứng dụng cụ thể.

Tổng thể, *FaireduPlus* được chứng minh là một phương pháp tiên xử lý có tính thực tiễn và khả năng mở rộng, phù hợp cho các bài toán học máy trong lĩnh vực giáo dục, nơi dữ liệu thường mất cân bằng, thưa thớt và chứa nhiều thuộc tính nhạy cảm. Phương pháp này góp phần làm rõ cách tiếp cận tích hợp nhằm cải thiện công bằng giao thoa trước giai đoạn huấn luyện mô hình, đồng thời cung cấp cơ sở cho các nghiên cứu tiếp theo về đánh giá và kiểm soát mối quan hệ đánh đổi giữa công bằng và hiệu suất.

Các kết quả nghiên cứu chính của chương này đã được tổng hợp trong một công trình khoa học do tác giả là tác giả chính, hiện đang trong quá trình phản biện tại tạp chí *Information and Software Technology*.

KẾT LUẬN

Luận án tập trung nghiên cứu và phát triển các phương pháp nhằm đảm bảo tính công bằng trong các mô hình học máy ứng dụng cho lĩnh vực giáo dục. Toàn bộ nội dung nghiên cứu được triển khai xuyên suốt bốn chương chính, tương ứng với bốn hướng đóng góp khoa học có mối liên hệ chặt chẽ và kế thừa lẫn nhau.

Chương 1 trình bày tổng quan toàn diện về vấn đề công bằng trong các hệ thống trí tuệ nhân tạo và học máy, với trọng tâm là bối cảnh giáo dục. Nội dung chương hệ thống hóa khung lý thuyết nền tảng, làm rõ các khái niệm cốt lõi liên quan đến công bằng và thiên lệch, phân loại các hướng tiếp cận đảm bảo công bằng, đồng thời phân tích những thách thức và khoảng trống nghiên cứu còn tồn tại. Các phân tích này đóng vai trò định hướng quan trọng cho việc xác lập các mục tiêu và phương pháp nghiên cứu của luận án.

Trên cơ sở đó, Chương 2 đề xuất phương pháp *Fairedu*, một kỹ thuật tiền xử lý dựa trên hồi quy đa biến nhằm loại bỏ sự phụ thuộc giữa các thuộc tính nhạy cảm và các thuộc tính đầu vào trong dữ liệu huấn luyện. Phương pháp này cho phép xử lý đồng thời nhiều thuộc tính nhạy cảm và đã được kiểm chứng thực nghiệm trên các bộ dữ liệu giáo dục thực tế. Kết quả cho thấy *Fairedu* giúp cải thiện đáng kể các chỉ số công bằng, đồng thời vẫn duy trì hiệu suất dự đoán của mô hình ở mức chấp nhận được.

Chương 3 tập trung khai thác các kỹ thuật sinh dữ liệu tổng hợp để xử lý vấn đề mất cân bằng phân phối trong dữ liệu, từ đó đề xuất phương pháp *DPF*. Phương pháp này hướng tới cân bằng dữ liệu giữa các nhóm giao thoa của nhiều thuộc tính nhạy cảm, qua đó giảm thiểu thiên lệch phát sinh do thiếu tính đại diện. Các kết quả thực nghiệm cho thấy *DPF* đạt được sự cải thiện công bằng rõ rệt so với các cấu hình tham chiếu, đồng thời thể hiện khả năng duy trì hiệu suất và tính mở rộng trong các kịch bản dữ liệu khác nhau.

Trên nền tảng của hai phương pháp trên, Chương 4 đề xuất *FaireduPlus* như một khuôn khổ đảm bảo công bằng hai chiều, kết hợp giữa can thiệp theo chiều dọc (loại bỏ phụ thuộc) và can thiệp theo chiều ngang (cân bằng phân phối dữ

liệu). Phương pháp này cho phép xử lý hiệu quả các tập dữ liệu chứa đồng thời nhiều thuộc tính nhạy cảm và mất cân bằng nghiêm trọng. Thực nghiệm cho thấy *FaireduPlus* không chỉ cải thiện công bằng một cách ổn định mà còn duy trì hiệu suất dự đoán ở mức hợp lý. Bên cạnh đó, chương này còn đề xuất và bước đầu kiểm chứng một *chỉ số đánh đổi tổng hợp* nhằm hỗ trợ định lượng mối quan hệ giữa công bằng và hiệu suất, qua đó hỗ trợ so sánh và lựa chọn mô hình trong các bối cảnh ra quyết định thực tiễn. Mặc dù chỉ số này còn cần tiếp tục được hoàn thiện và đánh giá trên nhiều bối cảnh khác nhau, kết quả ban đầu cho thấy tiềm năng hỗ trợ đánh giá sự đánh đổi giữa hai mục tiêu quan trọng của các hệ thống học máy.

Tổng thể, luận án đã đóng góp một chuỗi giải pháp có tính hệ thống, từ phân tích lý thuyết, đề xuất phương pháp đến đánh giá thực nghiệm, nhằm nâng cao tính công bằng cho các mô hình học máy trong lĩnh vực giáo dục. Các kết quả đạt được không chỉ có ý nghĩa về mặt học thuật mà còn mang lại giá trị ứng dụng thực tiễn trong việc phát triển các hệ thống AI đáng tin cậy, có trách nhiệm và hướng tới người học.

Luận án đã mang đến những đóng góp quan trọng cả về lý thuyết và thực tiễn. Điểm cốt lõi xuyên suốt của nghiên cứu là tập trung giải quyết bài toán đảm bảo tính công bằng cho dữ liệu với đồng thời nhiều thuộc tính nhạy cảm – một thách thức chưa được quan tâm đầy đủ trong các nghiên cứu trước đây. Các phương pháp được đề xuất đều hướng đến xử lý vấn đề này theo những cách tiếp cận khác nhau: (i) *Fairedu* điều chỉnh dữ liệu huấn luyện nhằm loại bỏ sự phụ thuộc giữa thuộc tính nhạy cảm và biến không nhạy cảm, giúp nâng cao tính công bằng ngay cả khi xét nhiều thuộc tính nhạy cảm đồng thời; (ii) *DPF* khai thác dữ liệu tổng hợp để cân bằng phân phối giữa các nhóm giao nhau, từ đó khắc phục hiện tượng mất cân bằng dữ liệu khi số lượng thuộc tính nhạy cảm tăng lên; (iii) *FaireduPlus* tích hợp sức mạnh của hai hướng tiếp cận trên, vừa cân bằng dữ liệu bằng kỹ thuật sinh dữ liệu tổng hợp, vừa điều chỉnh sự phụ thuộc, tạo nên một giải pháp toàn diện, dung hòa công bằng và hiệu suất. Bên cạnh đó, luận án còn đóng góp ở khía cạnh phương pháp luận khi đề xuất và bước đầu kiểm chứng một “chỉ số đánh đổi” nhằm hỗ trợ đánh giá sự cân bằng giữa tính công bằng và hiệu suất của mô hình. Chỉ số này là cơ sở ban đầu cho các nghiên cứu tiếp theo hướng tới xây dựng các thước đo đánh đổi ổn định và

toàn diện hơn. Các kết quả thực nghiệm đã chứng minh rằng ba phương pháp này không chỉ cải thiện rõ rệt mức độ công bằng, mà còn duy trì hiệu suất dự đoán ổn định, đồng thời khẳng định tính khả thi và khả năng mở rộng trong thực tiễn giáo dục – lĩnh vực vốn đặc trưng bởi dữ liệu khan hiếm, mất cân bằng và chứa nhiều thiên lệch.

Bên cạnh những kết quả đạt được, luận án vẫn còn tồn tại một số hạn chế nhất định. *Thứ nhất*, các phương pháp đề xuất mới chỉ được kiểm chứng trên dữ liệu dạng bảng và các mô hình học máy truyền thống, trong khi chưa được đánh giá trên dữ liệu phi cấu trúc, dữ liệu quy mô lớn cũng như các mô hình học sâu và kiến trúc Transformer, do đó chưa phản ánh đầy đủ tính tổng quát. *Thứ hai*, hiệu quả của các phương pháp DPF và FaireduPlus còn hạn chế khi xử lý các bộ dữ liệu mất cân bằng nghiêm trọng hoặc tồn tại các nhóm giao thoa có rất ít hoặc không có mẫu, do thiếu dữ liệu đại diện cho những trường hợp hiếm. Nghiên cứu chủ yếu đánh giá dữ liệu tổng hợp thông qua hiệu quả của mô hình học máy và các chỉ số công bằng, chưa đánh giá một cách hệ thống các tiêu chí như fidelity giữa dữ liệu sinh và dữ liệu gốc cũng như khả năng tái lập của các mô hình sinh dữ liệu, đặc biệt đối với các mô hình ngôn ngữ lớn. *Thứ ba*, chỉ số đánh đổi giữa công bằng và hiệu suất mới dừng ở mức đề xuất và thử nghiệm ban đầu, đồng thời có thể kém ổn định trong một số trường hợp do sử dụng mức cải thiện tương đối. Bên cạnh đó, việc đánh giá hiệu quả của các phương pháp chủ yếu dựa trên thống kê thắng/thua, chưa kết hợp đầy đủ các kiểm định thống kê và các chỉ số kích thước ảnh hưởng (effect size) để đánh giá mức độ khác biệt giữa các phương pháp. *Thứ tư*, các phương pháp hiện vẫn đòi hỏi nhiều bước can thiệp thủ công và hiệu chỉnh theo từng bộ dữ liệu cụ thể, chưa hình thành một khung giải pháp tổng quát và tự động để triển khai trên nhiều lĩnh vực ứng dụng.

Những hạn chế nêu trên không chỉ cho thấy giới hạn của phạm vi nghiên cứu hiện tại, mà còn mở ra nhiều hướng đi mới để tiếp tục phát triển và hoàn thiện các phương pháp trong tương lai. Trên cơ sở những kết quả đã đạt được trong việc đánh giá và cải thiện tính công bằng của các hệ thống học máy thông qua các phương pháp tiền xử lý như Fairedu, DPF và FaireduPlus, luận án đồng thời đề xuất một số định hướng nghiên cứu tiềm năng nhằm nâng cao hiệu quả và mở rộng khả năng ứng dụng trong thực tiễn. Các định hướng này có thể được xem xét trong những nghiên cứu tiếp theo, bao gồm:

1. *Mở rộng phạm vi dữ liệu và mô hình áp dụng*: Trong phạm vi hiện tại, các phương pháp Fairedu, DPF, và FaireduPlus chủ yếu được thử nghiệm trên các bộ dữ liệu giáo dục với các mô hình truyền thống như *Hồi quy logistic*, *Cây quyết định* và *Rừng ngẫu nhiên*. Trong tương lai, cần tiến hành đánh giá trên nhiều bộ dữ liệu trong các lĩnh vực khác nhau như y tế, tài chính, bao gồm cả dữ liệu lớn và dữ liệu phi cấu trúc. Đồng thời, việc tích hợp với các mô hình hiện đại như *Học sâu* hoặc *Transformer-based* sẽ giúp kiểm chứng tính hiệu quả trong những kiến trúc phức tạp hơn, từ đó đánh giá mức độ tổng quát và khả năng mở rộng của phương pháp. Bên cạnh đó, cần nghiên cứu ảnh hưởng của hiện tượng đa cộng tuyến giữa các thuộc tính nhạy cảm và tích hợp các kỹ thuật xử lý phù hợp nhằm nâng cao tính ổn định của quá trình ước lượng và mở rộng khả năng áp dụng của phương pháp trên các bộ dữ liệu có cấu trúc phức tạp. Bên cạnh đó, cần nghiên cứu tích hợp các kỹ thuật hiệu chỉnh đa kiểm định như Bonferroni hoặc False Discovery Rate (FDR) trong quá trình xác định các thuộc tính cần điều chỉnh nhằm giảm nguy cơ sai số loại I và nâng cao độ tin cậy của phương pháp Fairedu. Ngoài ra, cần nghiên cứu mở rộng các phương pháp đề xuất cho các bài toán phân lớp đa lớp nhằm mở rộng phạm vi ứng dụng của phương pháp.
2. *Phát triển và hoàn thiện các chỉ số công bằng và phương pháp đánh giá*: Các thước đo hiện tại như “*tác động khác biệt*”, “*hiệu số chênh lệch thống kê*”, “*chênh lệch trung bình xác suất*” và “*chênh lệch cơ hội công bằng*” tuy được sử dụng phổ biến nhưng vẫn còn hạn chế trong việc phản ánh đầy đủ các dạng thiên vị phức tạp, đặc biệt khi tồn tại đồng thời nhiều thuộc tính nhạy cảm. Một hướng phát triển quan trọng là đề xuất và kiểm thử các chỉ số công bằng mới, đồng thời hoàn thiện “*chỉ số đánh đổi*” như một công cụ hỗ trợ lựa chọn mô hình tối ưu giữa công bằng và hiệu suất. Ngoài ra, cần nghiên cứu các phương pháp đánh giá chặt chẽ hơn thông qua việc kết hợp các kiểm định thống kê, khoảng tin cậy và các chỉ số kích thước ảnh hưởng (effect size), nhằm nâng cao độ tin cậy và tính toàn diện trong việc so sánh hiệu quả giữa các phương pháp đảm bảo công bằng. Các nghiên cứu tiếp theo cũng sẽ xem xét áp dụng các kiểm định phi tham số như Wilcoxon signed-rank, các phương pháp bootstrap, đồng thời tích hợp các kỹ thuật hiệu chỉnh đa kiểm định như Bonferroni hoặc False Discovery Rate (FDR)

trong quá trình xác định các thuộc tính cần điều chỉnh của Fairedu nhằm giảm nguy cơ sai lầm và nâng cao độ tin cậy thống kê của phương pháp.

3. *Nghiên cứu giải pháp sinh dữ liệu tổng hợp nhằm cân bằng dữ liệu cho các tình huống đặc biệt:* Một hướng quan trọng là phát triển và tích hợp các kỹ thuật sinh dữ liệu tổng hợp tiên tiến nhằm bổ sung dữ liệu cho các trường hợp thiên lệch trầm trọng hoặc thiếu vắng những tình huống giao thoa quan trọng. Việc này không chỉ giúp cân bằng phân phối dữ liệu mà còn đảm bảo sự hiện diện đầy đủ hơn của các nhóm thiểu số, qua đó nâng cao hiệu quả can thiệp công bằng. Đồng thời, cần xây dựng các tiêu chí đánh giá chất lượng dữ liệu tổng hợp như fidelity, privacy leakage và khả năng tái lập nhằm bảo đảm độ tin cậy khi ứng dụng các mô hình sinh dữ liệu trong các bài toán công bằng.
4. *Xây dựng khung giải pháp công bằng tổng quát và tự động:* Hiện tại, DPF và FaireduPlus vẫn đòi hỏi sự can thiệp thủ công đáng kể trong quá trình xử lý và hiệu chỉnh. Một định hướng dài hạn là phát triển một khung làm việc công bằng tổng quát có khả năng tự động phát hiện, điều chỉnh và đánh giá công bằng trên nhiều loại dữ liệu và mô hình khác nhau. Khung giải pháp này sẽ mở rộng khả năng ứng dụng của DPF và FaireduPlus trong các hệ thống học máy quy mô lớn, đặc biệt tại những lĩnh vực có yêu cầu nghiêm ngặt về đạo đức và công bằng như giáo dục, y tế, tuyển dụng và cấp tín dụng.

Những hướng đi trên không chỉ góp phần khắc phục các hạn chế hiện tại của luận án mà còn tạo tiền đề cho các nghiên cứu chuyên sâu hơn về công bằng trong học máy và trí tuệ nhân tạo. Việc tiếp tục mở rộng, chuẩn hóa và hoàn thiện các giải pháp công bằng sẽ là nền tảng quan trọng để phát triển các hệ thống học máy minh bạch, đáng tin cậy và vì con người, đặc biệt trong những lĩnh vực có tác động xã hội sâu rộng.

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN TỚI LUẬN ÁN

1. [Nga1] **Pham, N.**, Pham-Ngoc, H., Nguyen-Duc, A. (2023). *Fairness Requirement in AI Engineering – A Review on Current Research and Future Directions*. In: Gupta, V., Rubalcaba, L., Gupta, C., Hanne, T. (eds) Sustainability in Software Engineering and Business Information Management. SSEBIM 2022. Lecture Notes in Information Systems and Organisation, vol 62. Springer, Cham. https://doi.org/10.1007/978-3-031-32436-9_1.
2. [Nga2] **N. Pham**, P. N. Hung, and A. Nguyen-Duc, *Fairness for machine learning software in education: A systematic mapping study*, J. Syst. Softw., p. 112244, Oct. 2024, doi: 10.1016/j.jss.2024.112244 (Q1)
3. [Nga3] **N. Pham**, M. K. Do, T. V. Dai, P. N. Hung, and A. Nguyen-Duc, *FAIREDU: A Multiple Regression-Based Method for Enhancing Fairness in Machine Learning Models for Educational Applications*, Expert Syst. Appl., Dec. 2024, doi.org/10.1016/j.eswa.2024.126219 (Q1)
4. [Nga4] **N. Pham**, Minh Kha Do, Quang Trung Doan, Pekka Abrahamsson, Anh Nguyen-Duc, Pham Ngoc Hung, *FaireduPlus: Enhancing Intersectional Fairness in Education Focused Machine Learning Using Synthetic Data*, submitted to IST (Q1)

Danh mục này gồm 04 công trình.

Tài liệu tham khảo

- [1] Simon Caton and Christian Haas. Fairness in Machine Learning: A Survey. *arXiv:2010.04053 [cs, stat]*, October 2020. URL <http://arxiv.org/abs/2010.04053>. arXiv: 2010.04053.
- [2] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. Machine Learning for Synthetic Data Generation: A Review, June 2024. arXiv:2302.04062 [cs].
- [3] X. Zhai and et al. A review of artificial intelligence (ai) in education from 2010 to 2020. *Complexity*, 2021:1–18, Apr 2021. doi: 10.1155/2021/8812542.
- [4] L. Chen, P. Chen, and Z. Lin. Artificial intelligence in education: A review. *IEEE Access*, 8:75264–75278, 2020. doi: 10.1109/ACCESS.2020.2988510.
- [5] Alp Dulundu. Ai in education: Benefits and concerns. *Next Generation Journal for The Young Researchers*, 8:81, 11 2024. doi: 10.62802/3fr4f412.
- [6] Ryan S. Baker and Aaron Hawn. Algorithmic bias in education. 32(4):1052–1092, 2022. ISSN 1560-4306. doi: 10.1007/s40593-021-00285-9. URL <https://doi.org/10.1007/s40593-021-00285-9>.
- [7] René F. Kizilcec and Hansol Lee. Algorithmic fairness in education, 2021. URL <https://arxiv.org/abs/2007.05443>.
- [8] L. Paquette, J. Ocumpaugh, Z. Li, A. Andres, and R. Baker. Who’s learning? using demographics in edm research. *J. Educ. Data Min.*, 12(3):1–30, Oct 2020. doi: 10.5281/zenodo.4143612.
- [9] K. Zhang and A. Aslan. Ai technologies for education: Recent research and future directions. *Comput. Educ. Artif. Intell.*, 2:100025, Jun 2021. doi: 10.1016/j.caeai.2021.100025.
- [10] R. Yu, Q. Li, C. Fischer, S. Doroudi, and D. Xu. Towards accurate and fair prediction of college success: Evaluating different sources of student data. Technical report, International Educational Data Mining Society, 2020. URL <https://eric.ed.gov/?ft=on&q=decent&id=ED608066>.
- [11] OECD. Equity and quality in education: Supporting disadvantaged students and schools, 2012. URL https://www.oecd.org/en/publications/equity-and-quality-in-education_9789264130852-en.html.

- [12] Education 2030: Incheon declaration and framework for action for the implementation of sustainable development goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all - UNESCO digital library, 2016. URL <https://unesdoc.unesco.org/ark:/48223/pf0000245656>.
- [13] Weijie Jiang and Zachary A. Pardos. Towards Equity and Algorithmic Fairness in Student Grade Prediction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 608–617, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8473-5. doi: 10.1145/3461702.3462623. URL <https://dl.acm.org/doi/10.1145/3461702.3462623>.
- [14] D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3):51:1–51:44, Feb 2022. doi: 10.1145/3494672. URL <https://doi.org/10.1145/3494672>.
- [15] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), July 2021. ISSN 0360-0300. doi: <https://dl.acm.org/doi/10.1145/3457607>.
- [16] Naman Goel, Mohammad Yaghini, and Boi Faltings. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 116, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278722. URL <https://doi.org/10.1145/3278721.3278722>.
- [17] Yuriy Brun and Alexandra Meliou. Software fairness. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2018, page 754–759, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355735. doi: 10.1145/3236024.3264838. URL <https://doi.org/10.1145/3236024.3264838>.
- [18] Lionel P. Robert, Casey Pierce, Liz Marquis, Sangmi Kim, and Rasha Alahmad and. Designing fair AI for managing employees in organizations: a review, critique, and design agenda. 35(5):545–575. doi: 10.1080/07370024.2020.1735391. URL <https://doi.org/10.1080/07370024.2020.1735391>. Publisher: Taylor & Francis_eprint: <https://doi.org/10.1080/07370024.2020.1735391>.
- [19] P. Angelov, E. Soares, R. Jiang, N. Arnold, and P. Atkinson. Explainable artificial intelligence: An analytical review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, Jul 2021. doi: 10.1002/widm.1424.

- [20] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in Machine Learning Software: Why? How? What to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 429–440, August 2021. doi: 10.1145/3468264.3468537. URL <https://doi.org/10.1145/3468264.3468537>.
- [21] F. K. Došilović, M. Brčić, and N. Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, May 2018. doi: 10.23919/MIPRO.2018.8400040.
- [22] Usman Gohar, Sumon Biswas, and Hriday Rajan. Towards Understanding Fairness and its Composition in Ensemble Machine Learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 1533–1545, May 2023. doi: 10.1109/ICSE48619.2023.00133. URL <https://ieeexplore.ieee.org/document/10172501>.
- [23] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, October 2018. URL <http://arxiv.org/abs/1810.01943>. arXiv:1810.01943 [cs].
- [24] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM J. Responsib. Comput.*, 1(2):11:1–11:52, June 2024. doi: 10.1145/3631326. URL <https://dl.acm.org/doi/10.1145/3631326>.
- [25] Zhenpeng Chen, Jie M. Zhang, Max Hort, Mark Harman, and Federica Sarro. Fairness Testing: A Comprehensive Survey and Analysis of Trends. *ACM Transactions on Software Engineering and Methodology*, 2024. ISSN 1049-331X. doi: 10.1145/3652155. URL <https://dl.acm.org/doi/10.1145/3652155>. Just Accepted.
- [26] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):35:1–35:27, 2023. ISSN 1556-4681. doi: 10.1145/3551390. URL <https://dl.acm.org/doi/10.1145/3551390>.
- [27] Zeyu Tang, Jiji Zhang, and Kun Zhang. What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective. *ACM Computing*

- Surveys*, 55(13s):299:1–299:37, 2023. ISSN 0360-0300. doi: 10.1145/3597199. URL <https://dl.acm.org/doi/10.1145/3597199>.
- [28] Sumon Biswas and Hriday Rajan. Do the Machine Learning Models on a Crowd Sourced Platform Exhibit Bias? An Empirical Study on Model Fairness. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 642–653, November 2020. doi: 10.1145/3368089.3409704. URL <https://doi.org/10.1145/3368089.3409704>. arXiv:2005.12379 [cs, stat].
- [29] Ekene Francis Okagbue, Ujunwa Perpetua Ezechikulo, Tosin Yinka Akintunde, Mustapha Bala Tsakuwa, Samuel Nchekwubemchukwu Ilokanulo, Kosiso Modest Obiasoanya, Chidiebere Emeka Ilodibe, and Cheick Amadou Tidiane Ouattara. A comprehensive overview of artificial intelligence and machine learning in education pedagogy: 21 years (2000–2021) of research indexed in the scopus database. *Social Sciences Humanities Open*, 8(1):100655, 2023. ISSN 2590-2911. doi: <https://doi.org/10.1016/j.ssaho.2023.100655>. URL <https://www.sciencedirect.com/science/article/pii/S2590291123002607>.
- [30] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, October 2012. ISSN 0219-3116. doi: 10.1007/s10115-011-0463-8. URL <https://doi.org/10.1007/s10115-011-0463-8>.
- [31] Yanhui Li, Linghan Meng, Lin Chen, Li Yu, Di Wu, Yuming Zhou, and Baowen Xu. Training Data Debugging for the Fairness of Machine Learning Software. In *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*, pages 2215–2227, May 2022. doi: 10.1145/3510003.3510091. URL <https://ieeexplore.ieee.org/document/9794106>. ISSN: 1558-1225.
- [32] Amirarsalan Rajabi and Ozlem Ozmen Garibay. Tabfairgan: Fair tabular data generation with generative adversarial networks, 2021. URL <https://arxiv.org/abs/2109.00666>.
- [33] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. Fairness improvement with multiple protected attributes: How far are we? In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, ICSE '24*, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400702174. doi: 10.1145/3597503.3639083. URL <https://doi.org/10.1145/3597503.3639083>.
- [34] Pallavi Singh, Ayisha Necholi, and Wilfrido Moreno. Synthetic data generation for engineering education: A bayesian approach. In *2023 IEEE 3rd International*

- Conference on Advanced Learning Technologies on Education Research (ICALTER)*, pages 1–4, 2023. doi: 10.1109/ICALTER61411.2023.10372935.
- [35] Aryan Pathare, Ramchandra Mangrulkar, Kartik Suvarna, Aryan Parekh, Govind Thakur, and Aruna Gawade. Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. *International Journal of Information Management Data Insights*, 3(2):100177, 2023. ISSN 2667-0968. doi: <https://doi.org/10.1016/j.ijime.2023.100177>. URL <https://www.sciencedirect.com/science/article/pii/S2667096823000241>.
- [36] Jill-Jênn Vie, Tomas Rigaux, and Sein Minn. Privacy-preserving synthetic educational data generation. In Isabel Hilliger, Pedro J. Muñoz-Merino, Tinne De Laet, Alejandro Ortega-Arranz, and Tracie Farrell, editors, *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*, pages 393–406, Cham, 2022. Springer International Publishing. ISBN 978-3-031-16290-9.
- [37] Xuesong Zhai, Xiaoyan Chu, Ching Sing Chai, Morris Siu Yung Jong, Andreja Istenic, Michael Spector, Jia-Bao Liu, Jing Yuan, and Yan Li. A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, 2021(1):8812542, 2021. ISSN 1099-0526. doi: 10.1155/2021/8812542. URL <https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/8812542>. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/8812542>.
- [38] O. N. Kulkarni, V. Patil, V. K. Singh, and P. K. Atrey. Accuracy and fairness in pupil detection algorithm. In *2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM)*, pages 17–24, Oct 2021. doi: 10.1109/BigMM52142.2021.00011.
- [39] M. Verger, S. Lallé, F. Bouchet, and V. Luengo. Is your model ‘madd’? a novel metric to evaluate algorithmic fairness for predictive student models. Jul 2023. doi: 10.5281/zenodo.8115786.
- [40] Brent Bridgeman, Catherine Trapani, and Yigal Attali. Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Appl. Meas. Educ.*, 25(1):27–40, Jan 2012. doi: 10.1080/08957347.2012.635502.
- [41] B. Bridgeman, C. Trapani, and Y. Attali. Considering fairness and validity in evaluating automated scoring. Technical report, Jan 2009.
- [42] B. Clauser, M. Kane, and D. Swanson. Validity issues for performance-based tests scored with computer-automated scoring systems. *Appl. Meas. Educ. - APPL MEAS EDUC*, 15:413–432, Oct 2002. doi: 10.1207/S15324818AME1504-05.

- [43] Yu Song. Research on learning behavior detection based on deep learning. In Fred Paas, Srikanta Patnaik, and Taosheng Wang, editors, *Recent Trends in Educational Technology and Administration*, pages 278–287, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-60777-6.
- [44] Yu Hsuan Wu and Eric Hsiaokuang Wu. Ai-based college course selection recommendation system: Performance prediction and curriculum suggestion. In *2020 International Symposium on Computer, Consumer and Control (IS3C)*, pages 79–82, 2020. doi: 10.1109/IS3C50286.2020.00028.
- [45] Zhenpeng Chen, Jie M. Zhang, Max Hort, Mark Harman, and Federica Sarro. Fairness testing: A comprehensive survey and analysis of trends. *ACM Trans. Softw. Eng. Methodol.*, 33(5), June 2024. ISSN 1049-331X. doi: 10.1145/3652155. URL <https://doi.org/10.1145/3652155>.
- [46] Zoltán Somogyi. *Performance Evaluation of Machine Learning Models*, pages 87–112. Springer International Publishing, Cham, 2021. ISBN 978-3-030-60032-7. doi: 10.1007/978-3-030-60032-7_3. URL https://doi.org/10.1007/978-3-030-60032-7_3.
- [47] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O’Reilly Media, 2nd edition, 2019. ISBN 9781492032618.
- [48] Alex J Bowers and Xuejun Zhou. Receiver operating characteristic (roc) area under the curve (auc): A diagnostic measure for evaluating the accuracy of predictors in education research. *Journal of Education for Students Placed at Risk (JESPAR)*, 24(1):20–46, 2019. doi: 10.1080/10824669.2018.1523734.
- [49] V. Jagan Raja, Dhanamalar M, Gautam Solaimalai, D. Leela Rani, P. Deepa, and R G Vidhya. Machine learning revolutionizing performance evaluation: Recent developments and breakthroughs. In *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, pages 780–785, 2024. doi: 10.1109/ICSCSS60660.2024.10625103.
- [50] Ben Hutchinson and Margaret Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* ’19*, page 49–58, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287600. URL <https://doi.org/10.1145/3287560.3287600>.
- [51] John Rawls. *A theory of justice*. Cambridge, Mass., Belknap Press of Harvard University Press. ISBN 978-0-674-88010-8 978-0-674-88014-6. URL http://archive.org/details/theoryofjustice0000rawl_s617.

- [52] U.S. Equal Employment Opportunity Commission (EEOC). Uniform guidelines on employee selection procedures. <https://www.ecfr.gov/current/title-29/subtitle-B/chapter-XIV/part-1607>, 1978. Adopted jointly by EEOC, Civil Service Commission, Department of Labor, and Department of Justice.
- [53] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. How Do Fairness Definitions Fare?: Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106, Honolulu HI USA, January 2019. ACM. ISBN 978-1-4503-6324-2. doi: 10.1145/3306618.3314248. URL <https://dl.acm.org/doi/10.1145/3306618.3314248>.
- [54] S. Verma and J. Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness, FairWare '18*, pages 1–7, New York, NY, USA, May 2018. Association for Computing Machinery. doi: 10.1145/3194770.3194776.
- [55] K. Zhang and A. Aslan. Ai technologies for education: Recent research and future directions. *Comput. Educ. Artif. Intell.*, 2:100025, Jun 2021. doi: 10.1016/j.caeai.2021.100025.
- [56] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 4069–4079, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- [57] T. Mester. Statistical bias types explained (with examples) - part1, Accessed: Nov. 13, 2023. URL <https://data36.com/statistical-bias-types-explained/>.
- [58] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16*, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.
- [59] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17*, pages 797–806, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098095. URL <https://doi.org/10.1145/3097983.3098095>.
- [60] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal

- justice risk assessments: The state of the art. *Sociol. Methods Res.*, 50(1):3–44, Feb 2021. doi: 10.1177/0049124118782533.
- [61] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 52:153–163, 2016.
- [62] G. Farnadi, B. Babaki, and L. Getoor. Fairness in relational domains. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ACM, pages 108–114, New Orleans LA USA, Dec 2018. doi: 10.1145/3278721.3278733.
- [63] Joan Casas-Roma and Jordi Conesa. A literature review on artificial intelligence and ethics in online learning. pages 111–131. January 2021. ISBN 978-0-12-823410-5. doi: 10.1016/B978-0-12-823410-5.00006-1.
- [64] Lin Li, Lele Sha, Yuheng Li, Mladen Raković, Jia Rong, Srecko Joksimovic, Neil Selwyn, Dragan Gašević, and Guanliang Chen. Moral Machines or Tyranny of the Majority? A Systematic Review on Predictive Bias in Education. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, LAK2023, pages 499–508, New York, NY, USA, March 2023. Association for Computing Machinery. ISBN 978-1-4503-9865-7. doi: 10.1145/3576050.3576119. URL <https://doi.org/10.1145/3576050.3576119>.
- [65] Centre for Data Ethics and Innovation. Review into bias in algorithmic decision-making, 2020. URL <https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making>.
- [66] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
- [67] Regulation (EU) 2016/679 of the European Parliament and of the Council. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>, 04 2016. Accessed: 2025-06-24.
- [68] OECD. *Equity in Education: Breaking Down Barriers to Social Mobility*. OECD Publishing, 2018. URL <https://doi.org/10.1787/9789264073234-en>.
- [69] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018. URL <https://proceedings.mlr.press/v81/buolamwini18a.html>.

- [70] Tom M. Mitchell. *Machine Learning*. McGraw-Hill Science/Engineering/Math, New York, NY, USA, March 1997. ISBN 0070428077. URL <https://www.cs.cmu.edu/~tom/mlbook.html>. Hardcover, Dimensions: 0.75 x 10.00 x 6.50 inches.
- [71] S. Hutt, M. Gardner, A. L. Duckworth, and S. D’Mello. Evaluating fairness and generalizability in models predicting on-time graduation from college applications. In *EDM*, 2019.
- [72] Lele Sha, Mladen Rakovic, Alexander Whitelock-Wainwright, David Carroll, Victoria M. Yew, Dragan Gasevic, and Guanliang Chen. Assessing Algorithmic Fairness in Automatic Classifiers of Educational Forum Posts. In Ido Roll, Danielle McNamara, Sergey Sosnovsky, Rose Luckin, and Vania Dimitrova, editors, *Artificial Intelligence in Education*, pages 381–394, Cham, 2021. Springer International Publishing. ISBN 978-3-030-78292-4. doi: 10.1007/978-3-030-78292-4_31.
- [73] J. Gardner, C. Brooks, and R. Baker. Evaluating the fairness of predictive student models through slicing analysis. Feb 2019. doi: 10.1145/3303772.3303791.
- [74] O. B. Deho, S. Joksimovic, L. Liu, J. Li, C. Zhan, and J. Liu. Assessing the fairness of course success prediction models in the face of (un)equal demographic group distribution. In *Proceedings of the Tenth ACM Conference on Learning @ Scale, L@S ’23*, pages 48–58. Association for Computing Machinery, Jul 2023. doi: 10.1145/3573051.3593381.
- [75] L. Sha, M. Raković, A. Das, D. Gašević, and G. Chen. Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Trans. Learn. Technol.*, 15(4):481–492, Aug 2022. doi: 10.1109/TLT.2022.3196278.
- [76] S. Riazy, K. Simbeck, and V. Schreck. Fairness in learning analytics: Student at-risk prediction in virtual learning environments. In *Proceedings of the 12th International Conference on Computer Supported Education*, pages 15–25, Prague, Czech Republic, 2020. SCITEPRESS - Science and Technology Publications. doi: 10.5220/0009324100150025.
- [77] H. Jeong, M. D. Wu, N. Dasgupta, M. Médard, and F. Calmon. Who gets the benefit of the doubt? racial bias in machine learning algorithms applied to secondary school math education. In *Neural Inf. Process. Syst. NeurIPS 2021 Workshop Math AI Educ. MATHAI4ED*.
- [78] Lele Sha, Dragan Gašević, and Guanliang Chen. Lessons from debiasing data for fair and accurate predictive modeling in education. 228:120323. ISSN

- 0957-4174. doi: <https://doi.org/10.1016/j.eswa.2023.120323>. URL <https://www.sciencedirect.com/science/article/pii/S0957417423008254>.
- [79] Nathalie Rzepka, Linda Fernsel, Hans-Georg Müller, Katharina Simbeck, and Neils Pinkwart. Unbias me! Mitigating Algorithmic Bias for Less-studied Demographic Groups in the Context of Language Learning Technology, June 2023. URL <https://osf.io/qa9vz>.
- [80] Nathalie Rzepka, Katharina Simbeck, Hans-Georg Müller, and Niels Pinkwart. Fairness of In-session Dropout Prediction:. In *Proceedings of the 14th International Conference on Computer Supported Education*, pages 316–326, Online Streaming, — Select a Country —, 2022. SCITEPRESS - Science and Technology Publications. ISBN 978-989-758-562-3. doi: 10.5220/0010962100003182. URL <https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010962100003182>.
- [81] Annette Markham and Elizabeth Buchanan. *Ethical Decision-Making and Internet Research Recommendations from the AoIR Ethics Working Committee (Version 2.0)*. 01 2012.
- [82] Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized data pre-processing for discrimination prevention, 2017. URL <https://arxiv.org/abs/1704.03354>.
- [83] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. volume 32, New York, NY, USA, May 2023. Association for Computing Machinery. doi: 10.1145/3583561. URL <https://doi.org/10.1145/3583561>.
- [84] Ninareh Mehrabi, Fred Morstatter, Nanyun Peng, and Aram Galstyan. Debiasing community detection: the importance of lowly connected nodes. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '19*, page 509–512, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368681. doi: 10.1145/3341161.3342915. URL <https://doi.org/10.1145/3341161.3342915>.
- [85] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/brunet19a.html>.

- [86] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms, 2019. URL <https://arxiv.org/abs/1905.12843>.
- [87] Shikha Bordia and Samuel R. Bowman. Identifying and reducing gender bias in word-level language models, 2019. URL <https://arxiv.org/abs/1904.03035>.
- [88] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6, 2009. doi: 10.1109/IC4.2009.4909197.
- [89] Tao Zhang, Tianqing Zhu, Jing Li, Mengde Han, Wanlei Zhou, and Philip S. Yu. Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination. *IEEE Transactions on Knowledge and Data Engineering*, 34(4): 1763–1774, 2022. doi: 10.1109/TKDE.2020.3002567.
- [90] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, pages 329–338. Association for Computing Machinery, Jan 2019. doi: 10.1145/3287560.3287589.
- [91] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic data - what, why and how? *CoRR*, abs/2205.03257, 2022. doi: 10.48550/ARXIV.2205.03257. URL <https://doi.org/10.48550/arXiv.2205.03257>.
- [92] Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. Large Language Models(LLMs) on Tabular Data: Prediction, Generation, and Understanding – A Survey, June 2024. arXiv:2402.17944 [cs].
- [93] Qinyi Liu, Ronas Shakya, Jelena Jovanovic, Mohammad Khalil, and Javier de la Hoz-Ruiz. Ensuring privacy through synthetic data generation in education. 56(3):1053–1073. ISSN 1467-8535. doi: 10.1111/bjet.13576. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13576>. _eprint: <https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13576>.
- [94] Mohammad Khalil, Farhad Vadiée, Ronas Shakya, and Qinyi Liu. Creating Artificial Students that Never Existed: Leveraging Large Language Models and CTGANs for Synthetic Data Generation, January 2025.
- [95] Mohamed Ashik Shahul Hameed, Asifa Mehmood Qureshi, and Abhishek Kaushik. Bias Mitigation via Synthetic Data Generation: A Review. *Electronics*,

- 13(19):3909, January 2024. ISSN 2079-9292. doi: 10.3390/electronics13193909. Number: 19 Publisher: Multidisciplinary Digital Publishing Institute.
- [96] Alvaro Figueira and Bruno Vaz. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics*, 10(15):2733, January 2022. ISSN 2227-7390. doi: 10.3390/math10152733. Number: 15 Publisher: Multidisciplinary Digital Publishing Institute.
- [97] Fida K. Dankar, Mahmoud K. Ibrahim, and Leila Ismail. A Multi-Dimensional Evaluation of Synthetic Data Generators. *IEEE Access*, 10:11147–11158, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3144765. Conference Name: IEEE Access.
- [98] WeiKang Liu, Yanchun Zhang, Hong Yang, and Qinxue Meng. A Survey on Differential Privacy for Medical Data Analysis. *Annals of Data Science*, 11(2): 733–747, April 2024. ISSN 2198-5812. doi: 10.1007/s40745-023-00475-3.
- [99] Christer Samuelsson and Sanja Štajner. 255statistical methods: Fundamentals. In *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 06 2022. ISBN 9780199573691. doi: 10.1093/oxfordhb/9780199573691.013.40.
- [100] Emmanouil Panagiotou, Arjun Roy, and Eirini Ntoutsi. Synthetic Tabular Data Generation for Class Imbalance and Fairness: A Comparative Study, September 2024. arXiv:2409.05215 [cs].
- [101] James Jordon, Jinsung Yoon, and M. Schaar. PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. September 2018.
- [102] Dr Gopalan and Dr Sivasubramanian. Crafting Artificial Insights: Mastering Synthetic Data for Model Training. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 08:1–6, November 2024. doi: 10.55041/IJSREM19586.
- [103] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language Models are Realistic Tabular Data Generators, April 2023. arXiv:2210.06280 [cs].
- [104] Ruxue Shi, Yili Wang, Mengnan Du, Xu Shen, Yi Chang, and Xin Wang. A comprehensive survey of synthetic tabular data generation, 2025. URL <https://arxiv.org/abs/2504.16506>.
- [105] Paul Tiwald, Alexandra Ebert, and Daniel T. Soukup. Representative fair synthetic data, 2021. URL <https://arxiv.org/abs/2104.03007>.

- [106] Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. Synthetic data in AI: challenges, applications, and ethical implications. *CoRR*, abs/2401.01629, 2024. doi: 10.48550/ARXIV.2401.01629. URL <https://doi.org/10.48550/arXiv.2401.01629>.
- [107] Kevin Fang, Vaikkunth Mugunthan, Vayd Ramkumar, and Lalana Kagal. Overcoming challenges of synthetic data generation. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 262–270, 2022. doi: 10.1109/BigData55660.2022.10020479.
- [108] Bhushan Chaudhari, Himanshu Chaudhary, Aakash Agarwal, Kamna Meena, and Tanmoy Bhowmik. Fairgen: Fair synthetic data generation, 2022. URL <https://arxiv.org/abs/2210.13023>.
- [109] Boris van Breugel and Mihaela van der Schaar. Beyond privacy: Navigating the opportunities and challenges of synthetic data, 2023. URL <https://arxiv.org/abs/2304.03722>.
- [110] David Pujol, Amir Gilad, and Ashwin Machanavajjhala. PreFair: Privately Generating Justifiably Fair Synthetic Data, March 2023. arXiv:2212.10310 [cs].
- [111] Lan Jiang, Clara Belitz, and Nigel Bosch. Synthetic Dataset Generation for Fairer Unfairness Research. In *Proceedings of the 14th Learning Analytics and Knowledge Conference, LAK '24*, pages 200–209, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400716188. doi: 10.1145/3636555.3636868.
- [112] H. Anderson, A. Boodhwani, and R. Baker. Assessing the fairness of graduation predictions. In *12th International Conference on Educational Data Mining*, 2019.
- [113] Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. Fairea: a model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2021*, pages 994–1006, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8562-6. doi: 10.1145/3468264.3468565. URL <https://doi.org/10.1145/3468264.3468565>.
- [114] Jie M. Zhang and Mark Harman. "Ignorance and Prejudice" in Software Fairness. In *Proceedings of the 43rd International Conference on Software Engineering, ICSE '21*, pages 1436–1447, Madrid, Spain, 2021. IEEE Press. ISBN 978-1-4503-9085-9. doi: 10.1109/ICSE43902.2021.00129. URL <https://doi.org/10.1109/ICSE43902.2021.00129>.

- [115] Mengdi Zhang and Jun Sun. Adaptive Fairness Improvement Based on Causality Analysis, September 2022. URL <http://arxiv.org/abs/2209.07190>. arXiv:2209.07190 [cs].
- [116] M. Du, F. Yang, N. Zou, and X. Hu. Fairness in deep learning: A computational perspective. *IEEE Intell. Syst.*, 2021. doi: 10.1109/mis.2020.3000681.
- [117] J. Vasquez Verdugo, X. Gitiaux, C. Ortega, and H. Rangwala. Faired: A systematic fairness analysis approach applied in a higher educational context. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, LAK22, pages 271–281. Association for Computing Machinery, Mar 2022. doi: 10.1145/3506860.3506902.
- [118] Khalida Walid Nathim, Nada Abdulkareem Hameed, Saja Abdulfattah Salih, Nada Adnan Taher, Hayder Mahmood Salman, and Dmytro Chornomordenko. Ethical ai with balancing bias mitigation and fairness in machine learning models. In *2024 36th Conference of Open Innovations Association (FRUCT)*, pages 797–807, 2024. doi: 10.23919/FRUCT64283.2024.10749873.
- [119] Rashidul Islam, Shimei Pan, and James R. Foulds. Can we obtain fairness for free? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 586–596, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462614. URL <https://doi.org/10.1145/3461702.3462614>.
- [120] Charlotte Leininger, Simon Rittel, and Ludwig Bothmann. Overcoming fairness trade-offs via pre-processing: A causal perspective. *CoRR*, abs/2501.14710, 2025. doi: 10.48550/ARXIV.2501.14710. URL <https://doi.org/10.48550/arXiv.2501.14710>.
- [121] Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. Does enforcing fairness mitigate biases caused by subpopulation shift?, 2021. URL <https://arxiv.org/abs/2011.03173>.
- [122] Xiaoyang Wang, Chia-Hsuan Chang, and Christopher C. Yang. Achieving equity via transfer learning with fairness optimization. *IEEE Access*, 12:195229–195241, 2024. doi: 10.1109/ACCESS.2024.3519465.
- [123] Suyun Liu and Luis Nunes Vicente. Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach. 19(3):513–537. ISSN 1619-6988. doi: 10.1007/s10287-022-00425-z. URL <https://doi.org/10.1007/s10287-022-00425-z>.
- [124] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological*

- Methods & Research*, 50(1):3–44, February 2021. ISSN 0049-1241, 1552-8294. doi: 10.1177/0049124118782533. URL <http://journals.sagepub.com/doi/10.1177/0049124118782533>.
- [125] Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.
- [126] Valentim Realinho, Jorge Machado, Luís Baptista, and Mónica V. Martins. Predicting student dropout and academic success. 7(11):146. ISSN 2306-5729. doi: 10.3390/data7110146. URL <https://www.mdpi.com/2306-5729/7/11/146>.
- [127] TheDevastator. Predict students’ dropout and academic success, 2022. URL <https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention>. Dataset.
- [128] P. Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. 2008. URL <https://api.semanticscholar.org/CorpusID:16621299>.
- [129] UCI student performance dataset: By paulo cortez. URL <https://www.kaggle.com/datasets/rxshark/uci-student-performance-dataset-by-paulo-cortez>.
- [130] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. Open University Learning Analytics dataset. *Scientific Data*, 4(1):170171, November 2017. ISSN 2052-4463. doi: 10.1038/sdata.2017.171. URL <https://www.nature.com/articles/sdata2017171>.
- [131] Open university learning analytics dataset (OULAD), 2022. URL <https://www.kaggle.com/datasets/anlgrbz/student-demographics-online-education-dataoulad>.
- [132] Latika Kharb and Prateek Singh. Role of Machine Learning in Modern Education and Teaching. In *Impact of AI Technologies on Teaching, Learning, and Research in Higher Education*, pages 99–123. IGI Global, 2021. ISBN 978-1-79984-763-2. doi: 10.4018/978-1-7998-4763-2.ch006. URL [https://www.igi-global.com/chapter/role-of-machine-learning-in-modern-education-and-tea](https://www.igi-global.com/chapter/role-of-machine-learning-in-modern-education-and-teaching/www.igi-global.com/chapter/role-of-machine-learning-in-modern-education-and-tea) 261497.
- [133] Hui Luan and Chin-Chung Tsai. A Review of Using Machine Learning Approaches for Precision Education. *Educational Technology & Society*, 24

- (1):250–266, 2021. ISSN 1176-3647. URL <https://www.jstor.org/stable/26977871>. Publisher: International Forum of Educational Technology & Society.
- [134] Semen M. Levin . Review of Machine Learning Models for Application in Adaptive Learning for Higher Education Student. *International Journal For Multidisciplinary Research*, 6(2):15481, March 2024. ISSN 2582-2160. doi: 10.36948/ijfmr.2024.v06i02.15481. URL <https://www.ijfmr.com/research-paper.php?id=15481>.
- [135] Vitor G. Costa and Carlos E. Pedreira. Recent advances in decision trees: an updated survey. *Artificial Intelligence Review*, 56:4765–4800, 2023. doi: 10.1007/s10462-022-10275-5. URL <https://doi.org/10.1007/s10462-022-10275-5>.
- [136] Simon Bernard, Laurent Heutte, and Sebastien Adam. On the selection of decision trees in random forests. In *2009 International Joint Conference on Neural Networks*, pages 302–307, 2009. doi: 10.1109/IJCNN.2009.5178693.
- [137] Trista Hollweck. Robert k. yin. (2014). case study research design and methods (5th ed.). thousand oaks, ca: Sage. 282 pages. *The Canadian Journal of Program Evaluation*, 30, 03 2016. doi: 10.3138/cjpe.30.1.108.
- [138] Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2): 131–164, 2009. doi: <https://doi.org/10.1007/s10664-008-9102-8>.
- [139] D. S. Cruzes and T. Dyba. Recommended steps for thematic synthesis in software engineering. In *2011 International Symposium on Empirical Software Engineering and Measurement*, page 275–284, Sep 2011. doi: 10.1109/ESEM.2011.36.
- [140] Mohammed Djameleddine Belgoumri, Mohamed Reda Bouadjenek, Sunil Aryal, and Hakim Hacid. Data Quality in Edge Machine Learning: A State-of-the-Art Survey. 2024. doi: 10.48550/ARXIV.2406.02600. Publisher: arXiv Version Number: 1.
- [141] Aaditya Bhatia, Dayi Lin, Gopi Krishnan Rajbahadur, Bram Adams, and Ahmed E. Hassan. Data Quality Antipatterns for Software Analytics, August 2024. arXiv:2408.12560 [cs].
- [142] Kaveen Hiniduma, Suren Byna, Jean Luca Bez, and Ravi Madduri. AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI, June 2024. arXiv:2406.19256 [cs].
- [143] A. Mahmud. Racial disparities in student outcomes in british higher education: Examining mindsets and bias. *Teaching in Higher Education*, page 254–269, Jul 2020. doi: 10.1080/13562517.2020.1796619.

- [144] Chenguang Pan and Zhou Zhang. Examining the algorithmic fairness in predicting high school dropouts. In *Educational Data Mining*, 2024. URL <https://api.semanticscholar.org/CorpusID:273692031>.
- [145] Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks, November 2021. arXiv:2110.12884 [cs].
- [146] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.
- [147] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 2nd edition, 1998. ISBN 9780132733502.
- [148] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. *Modeling tabular data using conditional GAN*. Curran Associates Inc., Red Hook, NY, USA, 2019.
- [149] Gilberto Recupito, Raimondo Rapacciuolo, Dario Di Nucci, and Fabio Palomba. Unmasking Data Secrets: An Empirical Investigation into Data Smells and Their Impact on Data Quality. In *2024 IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, pages 53–63, April 2024.
- [150] Padmaja Pulivarthy and Pawan Whig. Bias and Fairness Addressing Discrimination in AI Systems. In *Ethical Dimensions of AI Development*, pages 103–126. IGI Global Scientific Publishing, 2025. ISBN 9798369341476. doi: 10.4018/979-8-3693-4147-6.ch005.
- [151] Maassoumeh Javadi, Mandy S. M. Chung, Nushin G. Fard, and Mohammed Miskat. A Data Governance Literature Review in Education Sector. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1662–1669, October 2023. doi: 10.1109/CSCI62032.2023.00275. ISSN: 2769-5654.
- [152] Polydorou Eleni. Towards a Secure and Privacy Compliant Framework for Educational Data Mining. In Selmin Nurcan, Andreas L. Opdahl, Haralambos Mouratidis, and Aggeliki Tsohou, editors, *Research Challenges in Information Science: Information Science and the Connected World*, pages 534–541, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-33080-3. doi: 10.1007/978-3-031-33080-3_35.
- [153] Kalyanmoy Deb. *Multi-objective Optimisation Using Evolutionary Algorithms: An Introduction*, pages 3–34. Springer London, London, 2011. ISBN 978-0-

- 85729-652-8. doi: 10.1007/978-0-85729-652-8_1. URL https://doi.org/10.1007/978-0-85729-652-8_1.
- [154] Francesco Di Carlo, Nazanin Nezami, Hadis Anahideh, and Abolfazl Asudeh. FairPilot: An Explorative System for Hyperparameter Tuning through the Lens of Fairness, April 2023. URL <http://arxiv.org/abs/2304.04679>. arXiv:2304.04679 [cs].
- [155] Q. Hu and H. Rangwala. Towards fair educational data mining: A case study on detecting at-risk students. Technical report, International Educational Data Mining Society, 2020. URL <https://eric.ed.gov/?q=bias+real+life+examples&id=ED608050>.
- [156] J. Gardner, R. Yu, Q. Nguyen, C. Brooks, and R. Kizilcec. Cross-institutional transfer learning for educational models: Implications for model performance, fairness, and equity. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1664–1684. Association for Computing Machinery, Jun 2023. doi: 10.1145/3593013.3594107.
- [157] C. Kung and R. Yu. Interpretable models do not compromise accuracy or fairness in predicting college success. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*, L@S '20, pages 413–416. Association for Computing Machinery, Aug 2020. doi: 10.1145/3386527.3406755.
- [158] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio du Pin Calmon. Optimized Score Transformation for Consistent Fair Classification. October 2021. doi: 10.48550/arXiv.1906.00066. URL <http://arxiv.org/abs/1906.00066>. arXiv:1906.00066 [cs, math, stat].
- [159] C. Belitz, L. Jiang, and N. Bosch. Automating procedurally fair feature selection in machine learning. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 379–389, Jul 2021. doi: 10.1145/3461702.3462585.
- [160] Nazanin Nezami, Parian Haghghat, Denisa Gándara, and Hadis Anahideh. Assessing Disparities in Predictive Modeling Outcomes for College Student Success: The Impact of Imputation Techniques on Model Performance and Fairness. *Education Sciences*, 14(2):136, January 2024. ISSN 2227-7102. doi: 10.3390/educsci14020136. URL <https://www.mdpi.com/2227-7102/14/2/136>.
- [161] Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond Adult and COMPAS: Fair Multi-Class Prediction via Information Projection. *Advances in Neural Information Processing Systems*, 35:38747–38760, December

2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/fd5013ea0c3f96931dec77174eaf9d80-Abstract-Conference.html.
- [162] Hansol Lee and René F. Kizilcec. Evaluation of Fairness Trade-offs in Predicting Student Success. *Fairness, Accountability, and Transparency in Educational Data Cyberspace*, 2020.
- [163] D. Gándara, H. Anahideh, M. P. Ison, and A. Tayal. Inside the black box: Detecting and mitigating algorithmic bias across racialized groups in college student-success prediction. Technical report, arXiv, Jan 2023.
- [164] O. B. Deho, S. Joksimovic, J. Li, C. Zhan, Jixue Liu, and L. Liu. Should learning analytics models include sensitive attributes? explaining the why. *IEEE Trans. Learn. Technol.*, 16(4):560–572, Aug 2023. doi: 10.1109/TLT.2022.3226474.
- [165] A. Loukina, N. Madnani, and K. Zechner. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10, Aug 2019. doi: 10.18653/v1/W19-4401.
- [166] X. Li, D. Song, M. Han, Y. Zhang, and R. F. Kizilcec. On the limits of algorithmic prediction across the globe. In *ArXiv*, 2021.
- [167] V. Grari, B. Ruf, S. Lamprier, and M. Detyniecki. Achieving fairness with decision trees: An adversarial approach. *Data Sci. Eng.*, 5(2):99–110, Jun 2020. doi: 10.1007/s41019-020-00124-2.
- [168] V. Bayer, M. Hlosta, and M. Fernandez. Learning analytics and fairness: Do existing algorithms serve everyone equally? In *Lecture Notes in Computer Science*, volume 12749, pages 71–75. Springer International Publishing, 2021. doi: 10.1007/978-3-030-78270-12.
- [169] P. Manisha and S. Gujar. A neural network framework for fair classifier. Technical report, arXiv, Nov 2018.
- [170] S. Doroudi and E. Brunskill. Fairer but not fair enough on the equitability of knowledge tracing. In *Proceedings of the 9th International Conference on Learning Analytics Knowledge*, pages 335–339, Mar 2019. doi: 10.1145/3303772.3303838.
- [171] F. Marcinkowski, K. Kieslich, C. Starke, and M. Lünich. Implications of ai (un-)fairness in higher education admissions: the effects of perceived ai (un-)fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* ’20*, pages 122–130. Association for Computing Machinery, Jan 2020. doi: 10.1145/3351095.3372867.

- [172] C. Li, W. Xing, and W. Leite. Yet another predictive model? fair predictions of students' learning outcomes in an online math learning platform. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, LAK21, pages 572–578. Association for Computing Machinery, Aug 2021. doi: 10.1145/3448139.3448200.
- [173] E. Gómez, C. Shui Zhang, L. Boratto, M. Salamó, and M. Marras. The winner takes it all: Geographic imbalance and provider (un)fairness in educational recommender systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1808–1812. ACM, Jul 2021. doi: 10.1145/3404835.3463235.
- [174] Sebastian Tschitschek, Maria Knobelsdorf, and Adish Singla. Equity and fairness of bayesian knowledge tracing, 2022. URL <https://arxiv.org/abs/2205.02333>.
- [175] Noah Arthurs and AJ Alvero. *Whose Truth is the "Ground Truth"? College Admissions Essays and Bias in Word Vector Evaluation Methods*. July 2020.
- [176] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. Aequitas: A Bias and Fairness Audit Toolkit. April 2019. doi: 10.48550/arXiv.1811.05577. URL <http://arxiv.org/abs/1811.05577>. arXiv:1811.05577 [cs].
- [177] A. Mashhadi, A. Zolyomi, and J. Quedado. A case study of integrating fairness visualization tools in machine learning education. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, pages 1–7, New York, NY, USA, Aug 2022. Association for Computing Machinery. doi: 10.1145/3491101.3503568.
- [178] Y. N. Elglaly and Y. Liu. Promoting machine learning fairness education through active learning and reflective practices. *ACM SIGCSE Bull.*, 55(3):4–6, Jul 2023. doi: 10.1145/3610585.3610589.
- [179] B. N. Larson. Gender as a variable in natural-language processing: Ethical considerations. In *EthNLP@EACL*, 2017. doi: 10.18653/v1/W17-1601.
- [180] M. T. Islam, A. Fariha, and A. Meliou. Through the data management lens: Experimental analysis and evaluation of fair classification. *ArXiv*, 2021.
- [181] Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. FFB: A Fair Fairness Benchmark for In-Processing Group Fairness Methods. arXiv, June 2024. doi: 10.48550/arXiv.2306.09468. URL <http://arxiv.org/abs/2306.09468>. arXiv:2306.09468 [cs].

- [182] M. Madaio, S. L. Blodgett, E. Mayfield, and E. Dixon-Román. Beyond ‘fairness:’ structural (in)justice lenses on ai for education. Technical report, arXiv, Nov 2021.
- [183] R. Yu, H. Lee, and R. F. Kizilcec. Should college dropout prediction models include protected attributes? In *Proceedings of the Eighth ACM Conference on Learning @ Scale*, pages 91–100, Virtual Event Germany, Jun 2021. doi: 10.1145/3430895.3460139.
- [184] O. Sahlgren. The politics and reciprocal (re)configuration of accountability and fairness in data-driven education. *Learn. Media Technol.*, 48(1):95–108, Jan 2023. doi: 10.1080/17439884.2021.1986065.
- [185] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Front. Big Data*, 2, 2019. URL <https://www.frontiersin.org/articles/10.3389/fdata.2019.00013>.
- [186] A. Nematzadeh, G. L. Ciampaglia, F. Menczer, and A. Flammini. How algorithmic popularity bias hinders or promotes quality. *Sci. Rep.*, 8(1):15951, Oct 2018. doi: 10.1038/s41598-018-34203-2.
- [187] Michael Yee, Anindya Roy, Meghan Perdue, Consuelo Cuevas, Keegan Quigley, Ana Bell, Ahaan Rungta, and Shigeru Miyagawa. AI-assisted analysis of content, structure, and sentiment in MOOC discussion forums. *Frontiers in Education*, 8, September 2023. ISSN 2504-284X. doi: 10.3389/feduc.2023.1250846. URL <https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2023.1250846/full>. Publisher: Frontiers.
- [188] S. Akgun and C. Greenhow. Artificial intelligence in education: Addressing ethical challenges in k-12 settings. *AI Ethics*, 2(3):431–440, Aug 2022. doi: 10.1007/s43681-021-00096-7.
- [189] E. Okur, S. Aslan, N. Alyuz, A. Arslan Esme, and R. S. Baker. Role of socio-cultural differences in labeling students’ affective states. In *Artificial Intelligence in Education*, volume 10947 of *Lecture Notes in Computer Science*, pages 367–380. Springer International Publishing, 2018. doi: 10.1007/978-3-319-93843-1-27.
- [190] Z. Wang, K. Zechner, and Y. Sun. Monitoring the performance of human and automated scores for spoken responses. *Lang. Test.*, 35(1):101–120, Jan 2018. doi: 10.1177/0265532216679451.
- [191] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’18, page 335–340, New York, NY,

- USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278779. URL <https://doi.org/10.1145/3278721.3278779>.
- [192] G. Fenu, R. Galici, and M. Marras. Experts' view on challenges and needs for fairness in artificial intelligence for education. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, and V. Dimitrova, editors, *Artificial Intelligence in Education*, volume 243–255 of *Lecture Notes in Computer Science*. Springer International Publishing, 2022. doi: 10.1007/978-3-031-11644-5-20.
- [193] D. Sculley, J. Snoek, A. Wiltschko, and A. Rahimi. Winner's curse? on pace, progress, and empirical rigor. Jun 2018. URL <https://openreview.net/forum?id=rJWF0Fywf>.
- [194] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. Jul 2017. doi: 10.48550/arXiv.1707.00075.
- [195] P. S. Thomas, B. Castro da Silva, A. G. Barto, S. Giguere, Y. Brun, and E. Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, Nov 2019. doi: 10.1126/science.aag3311.
- [196] F. Xiang, X. Zhang, J. Cui, M. Carlin, and Y. Song. Algorithmic bias in a student success prediction models: Two case studies. In *2022 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, pages 310–315, Dec 2022. doi: 10.1109/TALE54877.2022.00058.
- [197] K. Holstein, J. W. Vaughan, H. Daumé III, M. Dudík, and H. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proc. 2019 CHI Conf. Hum. Factors Comput. Syst.*, pages 1–16, May 2019. doi: 10.1145/3290605.3300830.

Phụ lục A

Các Bảng tổng hợp về tổng quan về nghiên cứu về đảm bảo tính công bằng trong học máy

Phần phụ lục này trình bày các bảng tổng hợp chi tiết được sử dụng làm cơ sở cho phân tích trong Chương 1. Nội dung tập trung vào việc hệ thống hóa các nghiên cứu tiêu biểu, thuật toán học máy, loại thiên vị, bộ dữ liệu, và phương pháp đảm bảo công bằng đã được khảo sát trong lĩnh vực hệ thống học máy ứng dụng cho giáo dục. Các bảng này đóng vai trò như nguồn tham chiếu nền tảng, giúp làm rõ bức tranh tổng quan về cách các công trình trước đây định nghĩa, đo lường và xử lý vấn đề công bằng trong trí tuệ nhân tạo/học máy.

Cụ thể:

- **Bảng A.1:** Danh sách các thuật toán học máy được sử dụng phổ biến trong các nghiên cứu chính, bao gồm các mô hình hồi quy, cây quyết định, mạng nơ-ron, và các biến thể nâng cao.
- **Bảng A.2:** Phân loại các nghiên cứu chính theo từng loại công bằng được khảo sát (ví dụ: công bằng nhóm, công bằng cá nhân, công bằng quá trình).
- **Bảng A.3:** Tổng hợp các loại thiên vị thường gặp trong hệ thống học máy, như thiên vị dữ liệu, thiên vị mô hình và thiên vị xã hội.
- **Bảng A.4:** Trình bày đặc điểm chi tiết của các bộ dữ liệu được sử dụng trong nghiên cứu về công bằng giáo dục, bao gồm nguồn gốc, kích thước, thuộc tính nhạy cảm và biến mục tiêu.
- **Bảng A.5:** Tổng hợp các phương pháp đảm bảo công bằng được sử dụng trong các nghiên cứu chính, chia theo ba nhóm tiếp cận: tiền xử lý, trong mô hình, và hậu xử lý.
- **Bảng A.6:** Liệt kê các kỹ thuật đánh giá công bằng và hiệu suất của học máy, bao gồm các chỉ số như SPD, DI, AOD, EOD, Accuracy, Recall và F1-score.

Bảng A.1: Danh sách các thuật toán học máy sử dụng trong các nghiên cứu chính

Thuật toán / Mô hình	Định nghĩa	Tài liệu tham khảo	Số lượng nghiên cứu
Hồi quy Logistic	Mô hình thống kê phân tích kết quả nhị phân bằng cách ước lượng xác suất sử dụng hàm logistic	[10, 39, 40, 41, 42, 72, 73, 74, 80, 90, 112, 117, 154, 155, 156, 157, 158, 159, 160, 161]	21
Rừng ngẫu nhiên (Random Forest)	Thuật toán có giám sát sử dụng phương pháp học tổ hợp từ nhiều cây quyết định	[10, 71, 72, 74, 112, 117, 154, 157, 160, 161, 162, 163, 164, 165, 166]	15
Cây quyết định (Decision Tree)	Mô hình ra quyết định dựa trên cấu trúc cây phân cấp gồm các quy tắc và kết quả	[39, 73, 76, 79, 80, 90, 117, 154, 157, 158, 159, 160, 167]	13
Máy véc tơ hỗ trợ (SVM)	Mô hình học có giám sát phân loại dữ liệu bằng cách tìm siêu mặt phẳng phân tách tối ưu	[73, 74, 90, 112, 117, 154, 157, 160]	8
Naive Bayes	Bộ phân loại xác suất dựa trên định lý Bayes với giả định độc lập mạnh giữa các thuộc tính	[39, 73, 76, 90, 157, 159]	6
Học tăng cường (Gradient Boosting)	Kỹ thuật học tổ hợp tối ưu mô hình dự đoán bằng cách thêm dần các bộ học yếu để giảm lỗi	[74, 156, 161, 164, 168]	5
K láng giềng gần nhất (KNN)	Phương pháp phi tham số dùng để phân loại và hồi quy dựa trên các điểm dữ liệu gần nhất	[39, 79, 80, 117, 160]	5
Còn tiếp trang sau			

Bảng A.1 – tiếp theo từ trang trước

Thuật toán / Mô hình	Định nghĩa	Tài liệu tham khảo	Số lượng nghiên cứu
Bộ nhớ ngắn dài hạn (LSTM)	Mạng nơ-ron hồi tiếp có khả năng học phụ thuộc dài hạn nhờ các ô bộ nhớ	[13, 72, 73]	3
Mạng nơ-ron (NN)	Mô hình học máy mô phỏng hoạt động của não người	[156, 169]	2
Mô hình phân loại hợp tác (MCCM)	Tích hợp nhiều bộ phân loại hoạt động cùng nhau nhằm cải thiện độ chính xác và độ tin cậy	[7, 155]	2
Bayesian Knowledge Tracing (BKT)	Mô hình xác suất theo dõi kiến thức người học theo thời gian dựa trên suy luận Bayes	[170?]	2
Perceptron nhiều lớp (MLP)	Mạng nơ-ron nhân tạo có nhiều lớp nút để học các mối quan hệ phức tạp	[76, 79]	2
RANDOM, PERFECT, META	Mô hình RANDOM sử dụng mẫu ngẫu nhiên từ phân phối chuẩn. PERFECT chỉ có một thuộc tính. META chỉ dựa trên thông tin nhân khẩu học	[165]	1
Hồi quy OLS	Phương pháp thống kê ước lượng các tham số chưa biết bằng cách tối thiểu hóa tổng bình phương sai số	[171]	1
Thuật toán Seldonian	Khung thuật toán đảm bảo các ràng buộc công bằng và an toàn được đáp ứng trong quá trình ra quyết định	[172]	1
Còn tiếp trang sau			

Bảng A.1 – tiếp theo từ trang trước

Thuật toán / Mô hình	Định nghĩa	Tài liệu tham khảo	Số lượng nghiên cứu
Năm thuật toán lọc cộng tác (FCF)	Gồm UserKNN, ItemKNN, BPR, BiasedMF, SVD++	[173]	1
Mô hình nhân tố cộng gộp (AFM)	Mô hình thống kê dùng để hiểu mối quan hệ giữa biến và kết quả đầu ra	[170]	1
Deep Knowledge Tracing (DKT)	Kỹ thuật học máy dùng trong phân tích dữ liệu giáo dục và mô hình hóa kiến thức người học	[174]	1
Bayesian- Bayesian Knowledge Tracing (B2KT)	Phát triển mở rộng của BKT sử dụng khung Bayesian phân cấp	[174]	1
Word2vec Skip- Gram (nhúng từ)	Mô hình học máy dùng để tạo biểu diễn véc tơ cho từ trong không gian ngữ nghĩa	[175]	1
Học sâu (DL)	Mạng nơ-ron có nhiều lớp ẩn tự động học thuộc tính từ dữ liệu phức tạp bằng kích hoạt phi tuyến và lan truyền ngược	[80]	1

Bảng A.2: Các nghiên cứu chính theo từng loại công bằng được khảo sát

Loại công bằng		Tài liệu tham khảo	Số nghiên cứu
Công bằng theo nhóm	Công bằng nhóm tổng quát	[10, 38, 39, 41, 50, 72, 74, 76, 79, 80, 112, 116, 154, 157, 161, 163, 164, 173, 176, 177, 178, 179, 180]	24
	Công bằng theo xác suất	[6, 13, 73, 76, 77, 112, 117, 154, 158, 160, 161, 166, 177, 178]	15
	Công bằng nhân khẩu học	[6, 13, 72, 73, 74, 78, 116, 117, 162, 167, 168, 169, 177, 181]	14
	Cơ hội công bằng	[6, 13, 112, 117, 160, 162, 163, 181]	8
	Đối xử công bằng	[117]	1
Công bằng cá nhân	Công bằng cá nhân tổng quát	[7, 50, 71, 73, 74, 155, 176, 177, 180, 182]	11
	Công bằng qua không nhận thức	[7, 13, 71, 183, 184]	5
	Công bằng phản sự kiện	[7, 71]	2
	Công bằng quy trình	[159, 170]	2
Công bằng khác	Công bằng thống kê	[6, 117, 154, 158, 160, 165, 184]	7
	Công bằng nhóm con tổng quát	[50, 156, 168]	3
	Độ chính xác thủ tục có điều kiện	[165, 177]	2
	Công bằng xử lý	[165]	1
	Độ chính xác sử dụng có điều kiện	[165]	1
Tiếp tục trang sau			

Bảng A.2 – tiếp theo từ trang trước

Loại công bằng		Tài liệu tham khảo	Số nghiên cứu
	Độ chính xác tổng hợp	[165]	1

Bảng A.3: Các loại thiên vị khác nhau trong hệ thống học máy

Loại thiên vị	Giải thích	Tham khảo
Thiên vị lịch sử	Là những bất công và vấn đề xã hội-kỹ thuật đã tồn tại trong thế giới thực và có thể xâm nhập vào dữ liệu ngay cả khi việc lấy mẫu và chọn thuộc tính là hoàn hảo.	[59, 113, 125]
Thiên vị đại diện	Xuất hiện từ cách chúng ta định nghĩa và lấy mẫu từ quần thể.	[20, 113]
Thiên vị đo lường	Phát sinh từ cách chọn, sử dụng và đo lường các thuộc tính cụ thể.	[20, 113]
Thiên vị đánh giá	Xuất hiện trong quá trình đánh giá mô hình.	[20, 113]
Thiên vị tổng hợp	Xảy ra khi rút ra kết luận sai cho một nhóm con dựa trên quan sát từ nhóm khác, hoặc khi giả định sai về quần thể ảnh hưởng đến kết quả mô hình.	[20, 113]
Thiên vị quần thể	Phát sinh khi các đặc điểm thống kê, nhân khẩu học và hành vi người dùng trong dữ liệu khác với quần thể mục tiêu ban đầu.	[20, 185]
Nghịch lý Simpson	Xu hướng trong các nhóm con có thể bị đảo ngược khi các nhóm được gộp lại, dẫn đến thiên vị khi phân tích dữ liệu dị thể.	[20, 55]
Nguy hiểm dữ liệu dọc	Xảy ra khi dữ liệu cắt ngang được xử lý như dữ liệu dọc, dễ dẫn đến thiên vị do nghịch lý Simpson.	[20]
Thiên vị lấy mẫu	Phát sinh khi quá trình lấy mẫu các nhóm không phải ngẫu nhiên.	[20]

Bảng A.3 – tiếp theo

Loại thiên vị	Giải thích	Tham khảo
Thiên vị hành vi	Phát sinh từ sự khác biệt trong hành vi người dùng trên các nền tảng, bối cảnh hoặc tập dữ liệu khác nhau.	[20, 185]
Thiên vị trong sản xuất nội dung	Xuất hiện do sự khác biệt về cấu trúc, từ vựng, ngữ nghĩa và cú pháp trong nội dung do người dùng tạo ra.	[20, 185]
Thiên vị liên kết	Phát sinh khi các thuộc tính mạng xã hội từ kết nối hoặc tương tác người dùng không phản ánh đúng hành vi thực.	[20, 185]
Thiên vị thời gian	Do sự thay đổi trong hành vi hoặc đặc điểm quần thể theo thời gian.	[20, 185]
Thiên vị phổ biến	Các mục phổ biến dễ được hiển thị nhiều hơn, nhưng mức độ phổ biến này có thể bị thao túng bởi đánh giá giả hoặc bot.	[20, 186]
Thiên vị thuật toán	Thiên vị không đến từ dữ liệu đầu vào mà phát sinh từ chính thuật toán.	[20, 54]
Thiên vị do tương tác người dùng	Là dạng thiên vị không chỉ quan sát được trên web mà còn bị kích hoạt bởi giao diện người dùng hoặc hành vi thiên vị của chính người dùng.	[20, 54]
Thiên vị hiển thị	Phát sinh từ cách thông tin được trình bày ảnh hưởng đến cách người dùng tiếp nhận.	[20, 54]
Thiên vị xếp hạng	Người dùng có xu hướng nhấp nhiều hơn vào kết quả được xếp hạng đầu tiên vì cho rằng đó là quan trọng nhất.	[20]
Thiên vị xã hội	Hành động hoặc nội dung từ người khác có thể ảnh hưởng đến đánh giá và quyết định của người dùng.	[20, 54]
Thiên vị phát sinh	Xuất hiện sau khi hệ thống được triển khai và tương tác với người dùng thật, thường do thay đổi về quần thể, văn hóa hoặc tri thức xã hội.	[20, 56]

Bảng A.3 – tiếp theo

Loại thiên vị	Giải thích	Tham khảo
Thiên vị do tự chọn	Là dạng con của thiên vị chọn mẫu, khi đối tượng nghiên cứu tự nguyện tham gia dẫn đến sai lệch.	[20, 57]
Thiên vị do thiếu biến	Xảy ra khi một hoặc nhiều biến quan trọng bị bỏ sót khỏi mô hình.	[20, 57]
Thiên vị nhân-quả	Phát sinh từ ngộ nhận rằng tương quan ngụ ý quan hệ nhân quả.	[20, 57]
Thiên vị người quan sát	Khi người nghiên cứu vô thức áp kỳ vọng chủ quan của mình lên kết quả nghiên cứu.	[20, 57]
Thiên vị tài trợ	Phát sinh khi kết quả nghiên cứu bị báo cáo thiên vị để phục vụ lợi ích của nhà tài trợ.	[20, 57]

Bảng A.4: Đặc điểm của các bộ dữ liệu được sử dụng trong nghiên cứu về đảm bảo tính công bằng giáo dục

Tên bộ dữ liệu	Tài liệu tham khảo	Loại dữ liệu	Nguồn dữ liệu	Số lượng nghiên cứu
Điểm số sinh viên từ các học phần đại học	[10, 74, 76, 78, 155, 162, 164, 166, 183]	Đóng	Không có thông tin	9
Khóa học MOOC trong lĩnh vực STEM	[73, 78, 79, 80, 157, 173, 187]	Đóng	Không có thông tin	7
Dữ liệu học sinh trung học Mỹ (K-12)	[77, 78, 163, 188]	Mở	https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018140	4
Tiếp tục trang sau				

Bảng A.4 – tiếp theo từ trang trước

Tên bộ dữ liệu	Tài liệu tham khảo	Loại dữ liệu	Nguồn dữ liệu	Số lượng nghiên cứu
Điểm GPA của sinh viên tại Chile, Mỹ, v.v.	[13, 112, 117, 189]	Đóng	Không có thông tin	4
Điểm ngoại ngữ	[165, 190]	Đóng	Không có thông tin	2
Bộ dữ liệu khuôn mặt (CelebA)	[38, 181]	Mở	https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html	2
Dữ liệu ELS và IPEDS	[154, 160]	Mở	(https://www.icpsr.umich.edu/web/ICPSR/studies/4275); (https://surveys.nces.ed.gov/Iped/)	2
Dữ liệu từ ETS (Educational Testing Service)	[41]	Đóng	Không có thông tin	1
Bài luận tuyển sinh tại hệ thống đại học công lập Mỹ	[175]	Đóng	Không có thông tin	1
Dữ liệu từ National Student Clearinghouse (NSC)	[71]	Đóng	Không có thông tin	1
Dữ liệu PISA tại 65 quốc gia	[166]	Đóng	Không có thông tin	1
Ảnh từ Flickr	[38]	Mở	https://www.flickr.com/photos/tags/images/	1

Tiếp tục trang sau

Bảng A.4 – tiếp theo từ trang trước

Tên bộ dữ liệu	Tài liệu tham khảo	Loại dữ liệu	Nguồn dữ liệu	Số lượng nghiên cứu
Bộ dữ liệu học tập của Đại học Mở Anh (OULAD)	[39]	Mở	https://analyse.kmi.open.ac.uk/open_dataset	1
Khảo sát HSLS (High School Longitudinal Study)	[161]	Đóng	Không có thông tin	1
Bộ dữ liệu ENEM (Brazil)	[161]	Mở	https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/enem	1

Bảng A.5: Tổng hợp các phương pháp đảm bảo tính công bằng được sử dụng trong các nghiên cứu chính

Phương pháp	Mô tả	Nghiên cứu chính	Số lượng
Tính toán chỉ số chênh lệch	Phương pháp này huấn luyện một bộ phân loại không thiên vị bằng cách đảm bảo rằng kết quả dự đoán không phụ thuộc vào thuộc tính nhạy cảm. Mục tiêu kép của mô hình là: (1) tối đa hóa khả năng dự đoán chính xác nhãn đầu ra, và (2) đồng thời huấn luyện một mô hình đối kháng để làm giảm khả năng suy ra thuộc tính nhạy cảm từ các dự đoán của mô hình chính [191]	[10, 38, 39, 76, 157, 164, 165, 169, 173, 183]	10
Tiếp tục ở trang sau			

Bảng A.5 – tiếp từ trang trước

Phương pháp	Mô tả	Nghiên cứu chính	Số lượng
Kỹ thuật cân bằng lớp (CBTs)	Tăng cường các lớp có ít mẫu hoặc giảm bớt các lớp có nhiều mẫu để đạt được sự cân bằng trong dữ liệu huấn luyện	[72, 78, 156, 160, 181, 192]	6
Sử dụng công cụ phát hiện thiên vị	Sử dụng các công cụ có sẵn để phát hiện phân biệt đối xử và giảm thiểu thiên vị như AI Fairness 360, Aequitas, Google Analogy Test Set (GATS), SMOTE và các thước đo công bằng	[13, 71, 161, 166, 166, 167]	4
Phân tích lát cắt	Phân tích hiệu suất theo các chiều hoặc nhóm khác nhau trong dữ liệu [193]	[71, 172]	2
Huấn luyện mô hình học đối kháng	Kỹ thuật nhằm học các biểu diễn sâu không thiên vị từ dữ liệu có thiên vị [194]. Mục tiêu là các biểu diễn này có thể dự đoán tốt nhãn của nhiệm vụ chính nhưng lại không thể dự đoán thuộc tính nhạy cảm [116]	[13, 167]	2
Thuật toán Seldonian	Phương pháp huấn luyện mô hình học máy nhằm đảm bảo thỏa mãn các ràng buộc đạo đức như công bằng hoặc an toàn, bên cạnh việc tối ưu hiệu suất. Nó sử dụng các công cụ thống kê để kiểm soát rủi ro và hạn chế vi phạm ràng buộc do nhiễu dữ liệu [195]	[172]	1
Tiếp tục ở trang sau			

Bảng A.5 – tiếp từ trang trước

Phương pháp	Mô tả	Nghiên cứu chính	Số lượng
Thuật toán FairProjection	FairProjection là một thuật toán có thể song song hóa, dùng để điều chỉnh đầu ra của các mô hình học máy nhằm đảm bảo tính công bằng giữa các nhóm. Thuật toán sử dụng phương pháp ADMM để tối ưu hóa và cung cấp đảm bảo về độ phức tạp mẫu cũng như tốc độ hội tụ, từ đó đảm bảo kết quả cuối cùng vừa công bằng vừa hiệu quả	[161]	1

Bảng A.6: Các kỹ thuật đánh giá công bằng và hiệu suất của học máy

Loại đánh giá		Tài liệu tham khảo	Số lượng
Thiết lập thí nghiệm	Siêu tham số (Hyperparameter)	[10, 38, 39, 71, 75, 76, 77, 78, 79, 112, 117, 154, 155, 156, 160, 161, 163, 164, 165, 166, 167, 169, 170, 171, 175, 176, 183, 187, 196]	29
	Thiết lập mặc định (Default setting)	[8, 74, 80, 90, 157, 162, 174]	7
Đối sánh mô hình	So sánh thực nghiệm với mô hình gốc	[10, 13, 39, 72, 73, 74, 75, 78, 79, 80, 90, 112, 154, 155, 156, 157, 158, 160, 161, 162, 163, 164, 165, 166, 167, 169, 170, 175, 176, 181, 183, 187, 196]	33
	So sánh với dữ liệu thực (Ground truth)	[40, 41, 42, 171, 174, 175]	6
Tiếp tục trang sau			

Bảng A.6 – tiếp theo từ trang trước

Loại đánh giá		Tài liệu tham khảo	Số lượng
Phương pháp đánh giá	Xác thực chéo (Cross-validation)	[10, 39, 76, 77, 78, 79, 90, 117, 154, 156, 157, 158, 160, 161, 163, 164, 166, 167, 169, 175, 176, 183, 187]	23
	Khác	[13, 38, 74, 80, 155, 174]	6
	So sánh giữa người và máy	[40, 41, 42, 165, 171]	5
	Phân tích theo lát cắt (Slicing analysis)	[71, 73, 112]	3
Kết quả đánh giá	Cải thiện công bằng	[13, 39, 40, 41, 72, 73, 75, 76, 77, 78, 79, 157, 158, 160, 162, 165, 166, 169, 170, 171, 174, 175, 197]	23
	Cải thiện cả công bằng và hiệu suất	[8, 71, 154, 156, 161, 164, 167, 176, 181, 196]	10
	Khác	[10, 42, 74, 80, 112, 155, 163, 182, 183]	9

Phụ lục B

Các bảng bổ sung cho phân tích so sánh Fairedu và LTDD trong Chương 3

Phần phụ lục này cung cấp các bảng chi tiết bổ sung cho các phân tích đã trình bày trong Chương 3.4. Nội dung tập trung vào việc so sánh giữa ba cấu hình: mô hình gốc (Origin), mô hình áp dụng LTDD, và mô hình áp dụng Fairedu. Các bảng này minh họa rõ ràng sự khác biệt về cả công bằng và hiệu suất khi áp dụng Fairedu, với các giá trị thống kê được kiểm định ở mức ý nghĩa $p_value < 0.05$.

Cụ thể:

- **Bảng B.1** So sánh hiệu suất (Acc và Recall) giữa mô hình gốc, mô hình áp dụng LTDD và mô hình áp dụng Fairedu. Các ô xám thể hiện trường hợp Fairedu thắng (W) hoặc hòa (T).
- **Bảng B.2** So sánh hiệu suất (F1-score và Precision) giữa mô hình gốc, mô hình áp dụng LTDD và mô hình áp dụng Fairedu. Các ô xám thể hiện trường hợp Fairedu thắng (W) hoặc hòa (T).

Bảng B.1: So sánh hiệu suất (Acc và Recall) của mô hình áp dụng Fairedu với mô hình gốc và mô hình áp dụng LTDD

Mô hình	Chỉ số	ACC			Recall		
	Phương pháp	Origin	LTDD (%change)	Fairedu (%change)	Origin	LTDD (%change)	Fairedu (%change)
LR	O-GT	0.588	0.585(-0.3%)	0.582(-0.6%)	0.473	0.468(-0.5%)	0.453(-2.%)
RF	O-GT	0.58	0.581(0.1%)	0.58(0.%)	0.482	0.465(-1.7%)	0.478(-0.4%)
DT	O-GT	0.578	0.579(0.1%)	0.578(0.%)	0.518	0.479(-3.9%)	0.492(-2.6%)
LR	O-Disa	0.588	0.583(-0.5%)	0.582(-0.6%)	0.473	0.458(-1.5%)	0.453(-2.%)
RF	O-Disa	0.58	0.579(-0.1%)	0.58(0.%)	0.482	0.483(0.1%)	0.478(-0.4%)
DT	O-Disa	0.578	0.578(0.%)	0.578(0.%)	0.518	0.518(0.%)	0.492(-2.6%)
LR	SP-GT	0.935	0.935(0.%)	0.936(0.1%)	0.913	0.912(-0.1%)	0.913(0.%)
RF	SP-GT	0.93	0.938(0.8%)	0.935(0.5%)	0.914	0.912(-0.2%)	0.909(-0.5%)
DT	SP-GT	0.932	0.932(0.%)	0.928(-0.4%)	0.908	0.908(0.%)	0.91(0.2%)
LR	SP-SK	0.935	0.935(0.%)	0.936(0.1%)	0.913	0.913(0.%)	0.913(0.%)
RF	SP-SK	0.93	0.937(0.7%)	0.935(0.5%)	0.914	0.911(-0.3%)	0.909(-0.5%)
DT	SP-SK	0.932	0.931(-0.1%)	0.928(-0.4%)	0.908	0.907(-0.1%)	0.91(0.2%)
LR	SD-Nợ	0.843	0.821(-2.2%)	0.819(-2.4%)	0.885	0.859(-2.6%)	0.805(-8.%)
RF	SD-Nợ	0.827	0.827(0.%)	0.825(-0.2%)	0.88	0.885(0.5%)	0.878(-0.2%)
DT	SD-Nợ	0.819	0.818(-0.1%)	0.817(-0.2%)	0.869	0.883(1.4%)	0.886(1.7%)
LR	SD-GT	0.843	0.827(-1.6%)	0.819(-2.4%)	0.885	0.876(-0.9%)	0.805(-8.%)
RF	SD-GT	0.827	0.824(-0.3%)	0.825(-0.2%)	0.88	0.897(1.7%)	0.878(-0.2%)
							Tiếp trang sau

Bảng B.1 – tiếp theo

Mô hình	Chỉ số	ACC			Recall		
	Phương pháp	Origin	LTDD (%change)	Fairedu (%change)	Origin	LTDD (%change)	Fairedu (%change)
DT	SD-GT	0.819	0.812(-0.7%)	0.817(-0.2%)	0.869	0.903(3.4%)	0.886(1.7%)
LR	DNU_GT	0.907	0.917(1.1%)	0.934(2.98%)	1	0.996(-0.4%)	0.969(-3.1%)
RF	DNU_GT	0.941	0.93(-1.17%)	0.932(-0.96%)	1	0.999(-0.1%)	0.99(-1.%)
DT	DNU_GT	0.93	0.925(-0.54%)	0.891(-4.19%)	0.978	0.928(-5.11%)	0.928(-5.11%)
LR	DNU_Tuổi	0.91	0.925(1.65%)	0.934(2.64%)	1	0.969(-3.1%)	0.969(-3.1%)
RF	DNU_Tuổi	0.938	0.932(-0.64%)	0.932(-0.64%)	1	0.99(-1.%)	0.99(-1.%)
DT	DNU_Tuổi	0.942	0.934(-0.85%)	0.891(-5.41%)	0.977	0.973(-0.41%)	0.928(-5.02%)
LR	DNU_KV	0.908	0.912(0.44%)	0.934(2.86%)	1	1.(0.%)	0.969(-3.1%)
RF	DNU_KV	0.941	0.94(-0.11%)	0.932(-0.96%)	1	1.(0.%)	0.99(-1.%)
DT	DNU_KV	0.945	0.947(0.21%)	0.891(-5.71%)	0.981	0.982(0.1%)	0.928(-5.4%)
W		11	11		6	12	

Bảng B.2: So sánh hiệu suất (Điểm số F1 và độ chính xác) của mô hình áp dụng Fairedu với mô hình gốc và mô hình áp dụng LTDD

Mô hình	Chỉ số	F1-Score			Precision		
	Phương pháp	Original	LTDD (%change)	Fairedu (%change)	Original	LTDD (%change)	Fairedu (%change)
RF	O_GT	0.516	0.507(-1.7%)	0.513(-0.5%)	0.555	0.559(0.7%)	0.556(0.1%)
DT	O_GT	0.532	0.514(-3.3%)	0.519(-2.4%)	0.549	0.554(0.9%)	0.551(0.3%)
LR	O_K.tật	0.516	0.505(-2.1%)	0.502(-2.7%)	0.568	0.564(-0.7%)	0.563(-0.8%)
RF	O_K.tật	0.516	0.516(0%)	0.513(-0.5%)	0.555	0.555(0%)	0.556(0.1%)
DT	O_K.tật	0.532	0.532(0%)	0.519(-2.4%)	0.549	0.549(0%)	0.551(0.3%)
LR	SP_GT	0.93	0.931(0.1%)	0.931(0.1%)	0.95	0.952(0.2%)	0.952(0.2%)
RF	SP_GT	0.926	0.934(0.8%)	0.934(0.8%)	0.94	0.957(1.8%)	0.957(1.8%)
DT	SP_GT	0.927	0.925(-0.2%)	0.925(-0.2%)	0.949	0.944(-0.5%)	0.944(-0.5%)
LR	SP_SK	0.93	0.931(0.1%)	0.931(0.1%)	0.95	0.952(0.2%)	0.952(0.2%)
RF	SP_SK	0.926	0.933(0.7%)	0.934(0.8%)	0.94	0.957(1.8%)	0.957(1.8%)
DT	SP_SK	0.927	0.926(-0.1%)	0.925(-0.2%)	0.949	0.949(0%)	0.944(-0.5%)
LR	SD_Nợ	0.849	0.827(-2.5%)	0.821(-3.2%)	0.816	0.799(-2.%)	0.816(0%)
RF	SD_Nợ	0.835	0.836(0.1%)	0.833(-0.2%)	0.795	0.793(-0.2%)	0.793(-0.2%)
DT	SD_Nợ	0.827	0.829(0.2%)	0.829(0.2%)	0.79	0.782(-1%)	0.779(-1.3%)
LR	SD_GT	0.849	0.835(-1.6%)	0.825(-2.8%)	0.816	0.797(-2.3%)	0.828(1.4%)
RF	SD_GT	0.835	0.835(0%)	0.833(-0.2%)	0.795	0.782(-1.6%)	0.793(-0.2%)

Tiếp trang sau

Bảng B.2 – tiếp theo

Mô hình	Chỉ số	F1-Score			Precision		
	Phương pháp	Original	LTDD (%change)	Fairedu (%change)	Original	LTDD (%change)	Fairedu (%change)
DT	SD_GT	0.827	0.827(0%)	0.829(0.2%)	0.79	0.764(-3.2%)	0.779(-1.3%)
LR	DNU_GT	0.963	0.963(0%)	0.953(-1%)	0.929	0.929(0%)	0.944(1.6%)
RF	DNU_GT	0.963	0.963(0%)	0.962(-0.1%)	0.929	0.929(0%)	0.962(3.5%)
DT	DNU_GT	0.95	0.95(0%)	0.808(-14.9%)	0.918	0.912(-0.6%)	0.926(0.8%)
LR	DNU_Tuổi	0.963	0.963(0%)	0.953(-1%)	0.929	0.929(0%)	0.944(1.6%)
RF	DNU_Tuổi	0.963	0.963(0%)	0.962(-0.1%)	0.929	0.929(0%)	0.962(3.5%)
DT	DNU_Tuổi	0.95	0.953(0.3%)	0.808(-14.9%)	0.918	0.944(2.8%)	0.926(0.8%)
LR	DNU_BP	0.963	0.963(0%)	0.953(-1%)	0.929	0.929(0%)	0.944(1.6%)
RF	DNU_BP	0.963	0.963(0%)	0.962(-0.1%)	0.929	0.929(0%)	0.962(3.5%)
DT	DNU_BP	0.95	0.95(0%)	0.808(-14.9%)	0.918	0.919(0.1%)	0.926(0.8%)
W		6	9		17	19	

Phụ lục C

Các bảng chi tiết về dữ liệu, kết quả công bằng, và hiệu suất theo từng chỉ số khi đánh giá cấu hình DPF với các cấu hình tham chiếu Chương 4

Phụ lục này cung cấp các bảng chi tiết liên quan đến thí nghiệm đánh giá cấu hình DPF so với các cấu hình tham chiếu đã trình bày trong Chương 3.

Cụ thể, các bảng được trình bày trong phần này bao gồm:

- **Bảng C.1:** Tổng quan về dữ liệu tổng hợp được sinh ra theo phương pháp DPF, sử dụng hai kỹ thuật CTGAN và LLM. Bảng thể hiện rõ quy mô, tỷ lệ và phân phối dữ liệu giữa các nhóm giao thoa sau khi sinh.
- **Bảng C.2:** Trình bày chi tiết các chỉ số công bằng của cấu hình DPF và các cấu hình đối sánh khi sử dụng kỹ thuật sinh dữ liệu CTGAN.
- **Bảng C.3:** So sánh hiệu suất của ba cấu hình thực nghiệm trên các chỉ số *Accuracy* và *Recall*. Các ô được tô xám biểu thị trường hợp DPF đạt kết quả tốt hơn (W) so với các cấu hình tham chiếu.

Bảng C.1: Tổng quan về dữ liệu tổng hợp được sinh theo phương pháp DPF dựa trên kỹ thuật CTGAN và LLM

STT	Bộ dữ liệu	Thuộc tính nhạy cảm			Kết quả	Số lượng dữ liệu				
		<i>Giới tính</i>	<i>SK</i>			<i>Dự đoán</i>	<i>Thực tế</i>	<i>Train</i>	<i>CTGAN</i>	<i>LLM</i>
1	Student Performance	Nam	Tốt		1	130	104	104	49	
		Nam	Tốt		0	146	123	106	45	
		Nam	Khác		1	79	68	68	32	
		Nam	Khác		0	98	85	65	25	
		Nữ	Tốt		1	139	120	120	56	
		Nữ	Tốt		0	154	135	129	59	
		Nữ	Khác		1	162	132	132	62	
		Nữ	Khác		0	134	120	171	94	
		Total					1,042	887	895	422
		2	Student Dropout Predict	<i>Giới tính</i>	<i>Nợ</i>		<i>Kết quả</i>	<i>Thực tế</i>	<i>Train</i>	<i>CTGAN</i>
Nam	Không				Tốt nghiệp	526	438	2,190	2,002	
Nam	Không				Không	816	700	3,169	2,892	
Nam	Nợ				Tốt nghiệp	22	20	100	91	
Nam	Nợ				Không	191	164	13	0	
Nữ	Không				Tốt nghiệp	1,582	1,338	6,690	6,115	

(Còn tiếp trang sau...)

STT	Bộ dữ liệu	Thuộc tính nhạy cảm			Kết quả	Số lượng dữ liệu			
		Nữ	Không		Không	996	846	10,972	10,126
		Nữ	Nợ		Tốt nghiệp	79	71	355	324
		Nữ	Nợ		Không	211	183	444	399
		Total				4,424	3,760	23,933	21,949
3	Oulad	<i>Giới tính</i>	<i>K.tật</i>		<i>Kết quả</i>	<i>Thực tế</i>	<i>Train</i>	<i>CTGAN</i>	<i>LLM</i>
		Nam	Không		1	522	458	458	198
		Nam	Không		0	995	852	349	8
		Nam	Có		1	7,205	6,095	6,095	2,636
		Nam	Có		0	8,345	7,114	8,866	4,331
		Nữ	Không		1	665	558	558	241
		Nữ	Không		0	949	790	673	258
		Nữ	Có		1	6,263	5,309	5,309	2,296
		Nữ	Có		0	6,538	5,584	8,335	4,385
				Total				31,482	26,760
		<i>Giới tính</i>	<i>Tuổi</i>	<i>BP</i>	<i>Dự đoán</i>	<i>Thực tế</i>	<i>Train</i>	<i>CTGAN</i>	<i>LLM</i>
		1	1	1	1	87	72	144	144
		1	1	1	0	11	8	12	12
		1	1	0	1	144	120	290	290

(Còn tiếp trang sau...)

STT	Bộ dữ liệu	Thuộc tính nhạy cảm			Kết quả	Số lượng dữ liệu			
		1	1	0	0	17	15	23	23
		1	0	1	1	33	32	54	54
		1	0	1	0	3	3	5	5
		1	0	0	1	58	47	525	525
		1	0	0	0	24	21	32	32
		0	1	1	1	14	12	42	42
		0	1	1	0	1	1	3	3
		0	1	0	1	21	19	35	35
		0	1	0	0	1	1	3	3
		0	0	1	1	2	2	3	3
		0	0	1	0	0	0	0	0
		0	0	0	1	10	9	14	14
		0	0	0	0	0	0	0	0
		Total				426	362	1,187	1,187

Bảng C.2: Chi tiết các chỉ số công bằng đối sánh với DPF dựa trên kỹ thuật CTGAN

Chỉ số công bằng		1-DI			SPD			AOD			EOD		
Mô hình	TT nhạy cảm	Origin	DPF	SDG	Origin	DPF	SDG	Origin	DPF	SDG	Origin	DPF	SDG
LR	O_GT	0.307	0.012	0.309	0.104	0.003	0.198	0.100	0.000	0.201	0.119	0.002	0.178
	O_K.tật	0.726	0.292	0.157	0.303	0.061	0.087	0.294	0.080	0.073	0.318	0.102	0.081
	SP_GT	0.416	0.142	0.192	0.079	0.068	0.092	0.027	0.005	0.025	0.049	0.048	0.007
	SP_SK	0.067	0.098	0.156	0.056	0.049	0.077	0.028	0.010	0.014	0.048	0.031	0.018
	SD_GT	0.816	0.556	0.515	0.411	0.268	0.264	0.185	0.143	0.146	0.190	0.057	0.073
	SD_Nợ	0.816	3.193	1.355	0.287	0.540	0.417	0.139	0.252	0.105	0.129	0.158	0.036
	DNU_GT	0.071	0.093	0.017	0.049	0.085	0.017	0.206			0.000	0	0.000
	DNU_Tuổi	0.130	0.067	0.063	0.070	0.063	0.063	0.144	0.033	0.167	0.000	0	0.000
	DNU_KV	0.124	0.048	0.023	0.053	0.045	0.023	0.159	0.233	0.100	0.000	0	0.000
RF	O_GT	0.004	0.398	0.143	0.015	0.035	0.078	0.015	0.033	0.080	0.018	0.045	0.057
	O_K.tật	0.598	0.243	0.247	0.257	0.026	0.129	0.245	0.018	0.115	0.254	0.019	0.115
	SP_GT	0.147	0.110	0.086	0.078	0.054	0.043	0.029	0.009	0.020	0.048	0.048	0.068
	SP_SK	0.081	0.073	0.045	0.056	0.037	0.023	0.029	0.022	0.034	0.048	0.031	0.055
	SD_GT	0.633	0.520	0.520	0.340	0.258	0.258	0.112	0.135	0.134	0.109	0.048	0.042
	SD_Nợ	1.418	0.881	1.062	0.246	0.325	0.359	0.104	0.036	0.058	0.092	0.057	0.048
	DNU_GT	0.071	0.686	0.113	0.065	0.407	0.102	0.228			0.000	0.314	0.059
	DNU_Tuổi	0.130	0.250	0.133	0.083	0.167	0.125	0.151	0.064	0.249	0.000	0.129	0.030
	DNU_KV	0.124	0.288	0.002	0.058	0.151	0.002	0.176	0.076	0.032	0.000	0.152	0.003
	O_GT	0.031	0.436	0.002	0.018	0.029	0.003	0.020	0.029	0.002	0.020	0.050	0.015

(Còn tiếp trang sau...)

Chỉ số công bằng		1-DI			SPD			AOD			EOD		
Mô hình	TT nhạy cảm	Origin	DPF	SDG	Origin	DPF	SDG	Origin	DPF	SDG	Origin	DPF	SDG
	O_K.tật	0.227	0.040	0.252	0.101	0.003	0.088	0.085	0.015	0.098	0.071	0.038	0.111
	SP_GT	0.132	0.142	0.142	0.076	0.068	0.068	0.029	0.005	0.005	0.050	0.048	0.048
	SP_SK	0.053	0.098	0.098	0.056	0.049	0.049	0.028	0.010	0.010	0.049	0.031	0.031
	SD_GT	0.613	0.616	0.419	0.332	0.253	0.233	0.115	0.110	0.120	0.117	0.047	0.036
	SD_Nợ	1.361	2.337	0.646	0.239	0.431	0.288	0.100	0.184	0.033	0.087	0.219	0.030
	DNU_GT	0.049	0.283	0.000	0.048	0.220	0.000	0.206			0.000	0.118	0.000
	DNU_Tuổi	0.072	0.268	0.000	0.070	0.229	0.000	0.144	0.181	0	0.000	0.161	0.000
	DNU_KV	0.037	0.105	0.000	0.053	0.090	0.000	0.159	0.205	0	0.000	0.076	0.000
BM	O_GT	0.295	0.284	0.065	0.105	0.067	0.032	0.102	0.065	0.034	0.121	0.097	0.011
	O_K.tật	0.295	0.099	0.160	0.105	0.027	0.078	0.102	0.013	0.065	0.121	0.013	0.060
	SP_GT	0.113	0.238	0.183	0.070	0.107	0.090	0.030	0.050	0.042	0.048	0.014	0.089
	SP_SK	0.071	0.069	0.087	0.069	0.034	0.049	0.038	0.018	0.106	0.056	0.057	0.080
	SD_GT	0.688	0.516	0.059	0.242	0.227	0.029	0.096	0.095	0.003	0.091	0.008	0.063
	SD_Nợ	1.908	1.008	0.698	0.349	0.310	0.215	0.147	0.007	0.168	0.173	0.108	0.196
	DNU_GT	0.057	0.204	1.000	0.077	0.169	0.078	0.215			0.020	0.157	0.098
	DNU_Tuổi	0.129	0.143	0.500	0.130	0.125	0.104	0.117	0.124	0.009	0.025	0.114	0.116
	DNU_KV	0.036	0.023	0.349	0.075	0.020	0.008	0.134	0.043	0.083	0.019	0.047	0.032
	O_GT	0.487	0.454	0.283	0.128	0.073	0.178	0.128	0.071	0.181	0.164	0.078	0.155
	O_K.tật	0.567	1.295	0.096	0.178	0.222	0.052	0.178	0.244	0.061	0.173	0.282	0.029
	SP_GT	0.142	0.117	0.241	0.068	0.056	0.116	0.005	0.006	0.061	0.048	0.068	0.021
	SP_SK	0.098	0.124	0.005	0.049	0.060	0.003	0.010	0.001	0.053	0.031	0.009	0.055

NN

(Còn tiếp trang sau...)

Chỉ số công bằng		1-DI			SPD			AOD			EOD		
Mô hình	TT nhạy cảm	Origin	DPF	SDG	Origin	DPF	SDG	Origin	DPF	SDG	Origin	DPF	SDG
	SD_GT	0.505	0.668	0.665	0.240	0.303	0.284	0.105	0.172	0.166	0.057	0.084	0.125
	SD_Nợ	1.279	1.230	1.483	0.374	0.378	0.388	0.131	0.062	0.096	0.158	0.039	0.004
	DNU_GT	0.439	0.513	0.000	0.305	0.339	0.000				0.235	0.255	0
	DNU_Tuổi	0.167	0.333	0.000	0.125	0.250	0.000	0.006	0.199	0.000	0.122	0.199	0
	DNU_KV	0.009	0.056	0.000	0.007	0.040	0.000	0.061	0.174			0.015	0

Bảng C.3: So sánh hiệu suất của ba cấu hình thực nghiệm. Các ô xám biểu thị trường hợp thắng (W)

Chỉ số hiệu suất			Độ chuẩn xác			Độ hồi tưởng		
MH	D.liệu	TT	Origin	DPF	SDG	Origin	DPF	SDG
LR	Oulad	G.tính	0.641	0.566	0.567	0.568	0.268	0.624
		K.tật	0.658	0.566	0.567	0.508	0.268	0.624
	Std.P	G.tính	0.662	0.943	0.955	0.576	0.919	0.942
		S.khỏe	0.641	0.943	0.955	0.568	0.919	0.942
	Std.D	G.tính	0.658	0.810	0.795	0.508	0.953	0.965
		Nợ	0.662	0.810	0.795	0.576	0.953	0.965
	DNU	G.tính	0.81	0.953	0.891	0.228	1.000	0.875
		Tuổi	0.808	0.953	0.891	0.236	1.000	0.875
		N.sinh	0.821	0.953	0.891	0.365	1.000	0.875
RF	Oulad	G.tính	0.588	0.556	0.560	0.453	0.140	0.574
		K.tật	0.58	0.556	0.560	0.478	0.140	0.574
	Std.P	G.tính	0.578	0.936	0.930	0.492	0.919	0.907
		S.khỏe	0.935	0.936	0.930	0.913	0.919	0.907
	Std.D	G.tính	0.93	0.804	0.804	0.909	0.953	0.953
		Nợ	0.932	0.804	0.804	0.91	0.953	0.953
	DNU	G.tính	0.935	0.75	0.875	0.913	0.714	0.946
		Tuổi	0.93	0.75	0.875	0.909	0.714	0.946
		N.sinh	0.932	0.75	0.875	0.91	0.714	0.946
DT	Oulad	G.tính	0.907	0.547	0.565	0.996	0.106	0.489
		K.tật	0.941	0.547	0.565	0.999	0.106	0.489
	Std.P	G.tính	0.93	0.943	0.943	0.977	0.919	0.919
		S.khỏe	0.91	0.943	0.943	0.977	0.919	0.919
	Std.D	G.tính	0.938	0.851	0.779	0.962	0.912	0.971
		Nợ	0.942	0.851	0.779	0.973	0.912	0.971
DT	DNU	G.tính	0.908	0.891	0.875	0.982	0.893	1.000
		Tuổi	0.941	0.891	0.875	0.982	0.893	1.000
		Nơi sinh	0.945	0.891	0.875	0.982	0.893	1.000
	Oulad	G.tính	0.580	0.567	0.552	0.481	0.322	0.535
		K.tật	0.580	0.567	0.552	0.481	0.322	0.535
		G.tính	0.909	0.911	0.892	0.901	0.884	0.895

Bảng C.3 – tiếp theo

Chỉ số hiệu suất			Độ chuẩn xác			Độ hồi tưởng		
MH	D.liệu Std.P	TT	Origin	DPF	SDG	Origin	DPF	SDG
GB	Std.P	S.khỏe	0.909	0.911	0.892	0.901	0.884	0.895
		G.tính	0.830	0.819	0.592	0.85	0.895	0.591
	DNU	Nợ	0.828	0.819	0.592	0.85	0.895	0.591
		G.tính	0.952	0.781	0.172	0.929	0.857	0.089
		Tuổi	0.959	0.781	0.172	0.929	0.857	0.089
		N.sinh	0.959	0.781	0.172	0.929	0.857	0.089
NN	Oulad	G.tính	0.583	0.563	0.563	0.451	0.243	0.617
		K.tật		0.563	0.563	0.451	0.243	0.617
	Std.P	G.tính	0.943	0.936	0.904	0.919	0.907	0.907
		S.khỏe		0.936	0.904	0.919	0.907	0.907
	Std.D	G.tính	0.837	0.827	0.804	0.947	0.962	0.904
		Nợ		0.827	0.804	0.947	0.962	0.904
	DNU	G.tính	0.781	0.781	0.875	0.232	0.768	1.000
		Tuổi		0.781	0.875	0.232	0.768	1.000
		N.sinh		0.781	0.875	0.232	0.768	1.000

Số: 612 /QĐ-ĐHCN

Hà Nội, ngày 30 tháng 3 năm 2026

QUYẾT ĐỊNH
Về việc thành lập Hội đồng đánh giá luận án tiến sĩ

HIỆU TRƯỞNG
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Căn cứ Quyết định số 3568/QĐ-ĐHQGHN ngày 08 tháng 10 năm 2014 của Giám đốc Đại học Quốc gia Hà Nội (ĐHQGHN) về việc ban hành quy định về Tổ chức và hoạt động của các đơn vị thành viên và các đơn vị trực thuộc ĐHQGHN;

Căn cứ Nghị quyết số 985/NQ-HĐT ngày 08 tháng 11 năm 2022 của Hội đồng trường Trường Đại học Công nghệ (ĐHCN) về việc ban hành Quy chế tạm thời về Tổ chức và hoạt động của Trường ĐHCN;

Căn cứ Thông tư số 08/2017/TT-BGDĐT ngày 04 tháng 04 năm 2017 của Bộ Giáo dục và Đào tạo về việc ban hành quy chế tuyển sinh và đào tạo trình độ tiến sĩ;

Căn cứ Quyết định số 4555/QĐ-ĐHQGHN ngày 24 tháng 11 năm 2017 của Giám đốc ĐHQGHN về việc ban hành Quy chế đào tạo tiến sĩ tại ĐHQGHN và các Quyết định số 1787/QĐ-ĐHQGHN ngày 10 tháng 06 năm 2019 và Quyết định số 2311/QĐ-ĐHQGHN ngày 11/08/2020 của Giám đốc ĐHQGHN về việc bổ sung một số nội dung của Quy chế đào tạo tiến sĩ tại ĐHQGHN;

Căn cứ Biên bản đánh giá tổng thể luận án ở đơn vị chuyên môn và xác nhận chỉnh sửa luận án của nghiên cứu sinh (NCS) Phạm Thị Tố Nga;

Căn cứ đơn xin bảo vệ luận án trước Hội đồng đánh giá luận án tiến sĩ của NCS Phạm Thị Tố Nga và nhận xét của tập thể cán bộ hướng dẫn;

Căn cứ Công văn số 131/CNTT-SĐH ngày 29 tháng 12 năm 2025 của Khoa Công nghệ thông tin về việc đề nghị xem xét việc đánh giá luận án cho NCS Phạm Thị Tố Nga;

Theo đề nghị của Trường phòng Công tác sinh viên.

QUYẾT ĐỊNH:

Điều 1. Thành lập Hội đồng đánh giá luận án tiến sĩ của nghiên cứu sinh Phạm Thị Tố Nga, sinh ngày 27/10/1978 tại Phú Thọ, là nghiên cứu sinh khóa QH-2021 (K28), ngành đào tạo Khoa học máy tính, mã số 9480101.

Tên đề tài luận án: Một số phương pháp đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục

Cán bộ hướng dẫn:

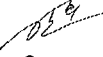
CBHD chính: GS. TS. Nguyễn Đức Anh, Trường Đại học Đông Nam Na Uy

CBHD phụ: PGS. TS. Phạm Ngọc Hùng, Trường Đại học Công nghệ, ĐHQGHN

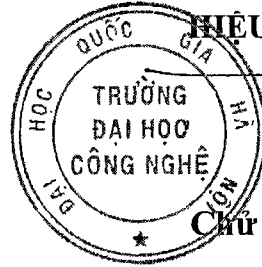
Danh sách các thành viên Hội đồng kèm theo quyết định này.

Điều 2. Hội đồng đánh giá luận án tiến sĩ của nghiên cứu sinh Phạm Thị Tố Nga thực hiện theo đúng Quy chế Đào tạo tiến sĩ ở Đại học Quốc gia Hà Nội và các quy định hiện hành. Hội đồng tự giải thể sau khi hoàn thành nhiệm vụ.

Điều 3. Trưởng phòng Hành chính Quản trị và Tổ chức Cán bộ, Đào tạo, Công tác sinh viên, Thủ trưởng các đơn vị liên quan, các thành viên trong Hội đồng và nghiên cứu sinh Phạm Thị Tố Nga chịu trách nhiệm thi hành quyết định này./.

Nơi nhận: 

- Như Điều 3;
- Lưu: VT, CTSV, NTH.12.



HIỆU TRƯỞNG

Chữ Đức Trình

DANH SÁCH HỘI ĐỒNG ĐÁNH GIÁ LUẬN ÁN TIẾN SĨ

(Kèm theo Quyết định số: 612 /QĐ-ĐHCN ngày 30 tháng 3 năm 2026 của
Hiệu trưởng Trường Đại học Công nghệ)

STT	Chức danh khoa học, Họ và tên	Chuyên ngành- Lĩnh vực chuyên môn	Cơ quan công tác	Trách nhiệm trong Hội đồng
1	PGS. TS. Võ Đình Hiếu	Công nghệ thông tin	Trường ĐH Công nghệ, ĐHQGHN	Chủ tịch
2	TS. Nguyễn Văn Sơn	Công nghệ thông tin	Trường ĐH Công nghệ, ĐHQGHN	Thư ký
3	PGS. TS. Phạm Văn Cường	Khoa học máy tính	Học viện Công nghệ Bưu chính Viễn thông	Phản biện
4	PGS. TS. Trần Đăng Hưng	Công nghệ thông tin	ĐH Công nghiệp Hà Nội	Phản biện
5	PGS. TS. Lê Hồng Phương	Khoa học máy tính	Trường ĐH Khoa học Tự nhiên, ĐHQGHN	Phản biện
6	PGS. TS. Phan Xuân Hiếu	Công nghệ thông tin	Trường ĐH Công nghệ, ĐHQGHN	Ủy viên
7	TS. Nguyễn Thị Minh Huyền	Công nghệ thông tin	Trường ĐH Khoa học Tự nhiên, ĐHQGHN	Ủy viên

Danh sách này gồm 07 người./

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ

Họ và tên người nhận xét: Võ Đình Hiếu..... Học hàm, học vị: PGS, TS
Chuyên ngành: Công nghệ công tin
Cơ quan công tác: Trường ĐH Công nghệ.....
Họ và tên nghiên cứu sinh: Phạm Thị Tố Nga
Tên đề tài luận án: Một số phương pháp đảm bảo tính công bằng cho các hệ thống học
máy trong lĩnh vực giáo dục
Ngành đào tạo: Khoa học máy tính..... Mã số: 9480101.01

Ý KIẾN NHẬN XÉT¹

1. Tính cấp thiết, ý nghĩa khoa học và thực tiễn của đề tài luận án

Đề tài mang tính cấp thiết cao do hệ thống học máy (AI/ML) đang được ứng dụng rộng rãi trong giáo dục nhưng lại tiềm ẩn nguy cơ tạo ra sự mất công bằng đối với các nhóm yếu thế, làm ảnh hưởng trực tiếp đến động lực và cơ hội học tập của người học.

Về ý nghĩa khoa học, luận án đóng góp một khung phương pháp toàn diện giúp giải quyết bài toán công bằng cho đồng thời nhiều thuộc tính nhạy cảm (công bằng giao thoa), đồng thời đề xuất chỉ số mới giúp định lượng sự đánh đổi giữa công bằng và hiệu suất mô hình.

Về ý nghĩa thực tiễn, các giải pháp đề xuất có thể tích hợp trực tiếp vào các hệ thống giáo dục thông minh (như dự đoán bỏ học, chấm điểm, tuyển sinh) để ra quyết định công bằng hơn, thúc đẩy xây dựng các hệ thống AI minh bạch, đáng tin cậy và có trách nhiệm.

2. Sự không trùng lặp của đề tài nghiên cứu so với các công trình, luận văn, luận án đã công bố trong và ngoài nước; tính trung thực, rõ ràng và đầy đủ trong trích dẫn tài liệu tham khảo

Chưa thấy sự trùng lặp trong các nội dung nghiên cứu với các công trình khác trong và ngoài nước. Tài liệu tham khảo được trích dẫn đầy đủ, rõ ràng.

3. Mức độ tổng quan các vấn đề lý luận của đề tài

Mức độ tổng quan các vấn đề lý luận liên quan đến chủ đề của luận án được thực hiện rất khá tốt và có hệ thống. NCS đã tổng hợp khá toàn diện các khái niệm cốt lõi về học máy, tính công bằng, thiên vị và các nhóm thuộc tính nhạy cảm đặc thù trong giáo dục. Dựa trên việc khảo sát bài bản 63 công trình nghiên cứu, luận án đã phân tích chi tiết các độ đo đo lường, ba giai đoạn can thiệp kỹ thuật và ứng dụng của dữ liệu tổng hợp. Đặc biệt, việc thảo luận về sự đánh đổi giữa tính công bằng và hiệu suất dự đoán đã giúp xác định cụ thể các khoảng trống nghiên cứu hiện tại.

4. Sự phù hợp giữa tên đề tài với nội dung, giữa nội dung với chuyên ngành và mã số ngành

Tên đề tài, nội dung, và chuyên ngành và mã số ngành phù hợp

5. Độ tin cậy và tính hiện đại của phương pháp đã sử dụng để nghiên cứu

Luận án có tính hiện đại thể hiện ở chỗ ứng dụng các kỹ thuật học sâu tiên tiến nhất như CTGAN và Mô hình ngôn ngữ lớn (LLM) trong sinh dữ liệu tổng hợp, nhằm giải quyết bài toán công bằng giao thoa.

Về độ tin cậy, thiết kế thực nghiệm được thực hiện chặt chẽ. Việc triển khai đánh giá trên 4 bộ dữ liệu thực tế cùng 5 mô hình học máy khác nhau đã khẳng định độ tin cậy và khả năng tổng quát hóa của phương pháp.

6. Kết quả nghiên cứu mới của tác giả; đóng góp mới cho sự phát triển khoa học chuyên ngành; đóng góp mới phục vụ sản xuất, kinh tế, xã hội, an ninh, quốc phòng và đời sống. Ý nghĩa khoa học, giá trị và độ tin cậy của những kết quả đó

Luận án có các kết quả sau:

- Phương pháp Fairedu: Kỹ thuật tiền xử lý bằng hồi quy đa biến nhằm nâng cao tính công bằng cho đồng thời nhiều thuộc tính nhạy cảm, thông qua việc loại bỏ sự phụ thuộc của các đặc trưng đầu vào vào các thuộc tính nhạy cảm trong bộ dữ liệu huấn luyện.

- Phương pháp DPF (Data Partitioning Fairness): Sử dụng các mô hình sinh dữ liệu tổng hợp (như CTGAN và LLM) để cân bằng phân bố dữ liệu huấn luyện giữa các nhóm giao thoa của thuộc tính nhạy cảm, giúp giải quyết triệt để vấn đề mất cân bằng và thiếu hụt dữ liệu ở các nhóm yếu thế.

- Phương pháp FaireduPlus: Khung tiếp cận đảm bảo tính công bằng toàn diện "hai chiều" cho dữ liệu dạng bảng, kết hợp cả chiều ngang (cân bằng dữ liệu bằng DPF) và chiều dọc (loại bỏ phụ thuộc bằng Fairedu).

- Đề xuất "Chỉ số đánh đổi": Xây dựng một thước đo tổng hợp mới giúp định lượng rõ ràng mối quan hệ đánh đổi giữa việc cải thiện tính công bằng và sự suy giảm hiệu suất dự đoán của mô hình.

- Ý nghĩa khoa học: Các kết quả này đóng góp một khung phương pháp để giải quyết bài toán công bằng giao thoa (xử lý đồng thời nhiều thuộc tính nhạy cảm) – một khoảng trống nghiên cứu trước đây.

Về mặt thực tiễn, các phương pháp đề xuất mang lại ý nghĩa cho xã hội và đời sống, đặc biệt trong lĩnh vực giáo dục:

- Đảm bảo bình đẳng giáo dục: Các giải pháp có thể tích hợp trực tiếp vào hệ thống giáo dục thông minh (như hệ thống tuyển sinh, chấm điểm tự động, dự đoán nguy cơ bỏ học) để hiệu chỉnh thiên vị tiềm ẩn do giới tính, khu vực, hay điều kiện kinh tế. Điều này hỗ trợ bảo vệ quyền lợi và cơ hội học tập công bằng cho mọi sinh viên.

- Khả năng mở rộng ứng dụng: Những phương pháp này (như DPF, FaireduPlus) không chỉ giới hạn trong giáo dục mà còn có tiềm năng mở rộng sang các hệ thống ra quyết định quan trọng trong kinh tế, xã hội như y tế, tài chính, tín dụng và tuyển dụng lao động – nơi yêu cầu rất cao về sự minh bạch và trách nhiệm xã hội.

Các kết quả có độ tin cậy cao.

7. Ưu điểm và nhược điểm về nội dung, kết cấu và hình thức của luận án

- Về kết cấu: Luận án có bố cục chặt chẽ, logic và hệ thống, bao gồm 6 chương chính cùng các phần phụ lục chi tiết. Trình tự các chương được sắp xếp liền mạch: từ việc đặt vấn đề, tổng quan cơ sở lý luận, xác định khoảng trống nghiên cứu, cho đến việc đề xuất lần lượt ba phương pháp từ đơn lẻ đến tích hợp (Fairedu, DPF, FaireduPlus), và cuối cùng là đánh giá, kết luận. Mối quan hệ giữa các chương cũng được sơ đồ hóa rõ ràng để minh họa cho tiến trình nghiên cứu.

- Về nội dung: Hàm lượng khoa học của luận án cao, giải quyết khoảng trống nghiên cứu về công bằng giao thoa (xử lý đồng thời nhiều thuộc tính nhạy cảm) trong dữ liệu giáo dục.

- Về hình thức: Hình thức trình bày phù hợp với một luận án tiến sĩ. Các khái niệm, chữ viết tắt được định nghĩa rõ ràng. Luận án sử dụng hiệu quả các bảng biểu tổng hợp, sơ đồ kiến trúc thuật toán, biểu đồ hộp và biểu đồ heatmap để trực quan hóa kết quả, giúp người đọc dễ tiếp nhận các dữ liệu kỹ thuật phức tạp.

Nhược điểm và hạn chế của luận án:

- Một số nội dung bị lặp (định nghĩa công bằng, thách thức trong nghiên cứu)

8. Nội dung luận án đã được công bố trên tạp chí, tuyển tập công trình hội nghị khoa học và giá trị khoa học của những công trình đã công bố

1. Pham, N., Pham-Ngoc, H., & Nguyen-Duc, A. (2023). Fairness requirement in AI engineering – A review on current research and future directions. In V. Gupta, L. Rubalcaba, C. Gupta, & T. Hanne (Eds.), *Sustainability in Software Engineering and Business Information Management. SSEBIM 2022 (Vol. 62)*. Springer, Cham. https://doi.org/10.1007/978-3-031-32436-9_1.

2. Pham, N., Hung, P. N., & Nguyen-Duc, A. (2024). Fairness for machine learning software in education: A systematic mapping study. *Journal of Systems and Software*, 112244. <https://doi.org/10.1016/j.jss.2024.112244>.

3. Pham, N., Do, M. K., Dai, T. V., Hung, P. N., & Nguyen-Duc, A. (2024). FAIREDU: A multiple regression-based method for enhancing fairness in machine learning models for educational applications. *Expert Systems with Applications*.
<https://doi.org/10.1016/j.eswa.2024.126219>.

4. Pham, N., Do, M. K., Doan, Q. T., Abrahamsson, P., Nguyen-Duc, A., & Hung, P. N. (n.d.). FaireduPlus: Enhancing intersectional fairness in education focused machine learning using synthetic data. Manuscript submitted for publication to *Information and Software Technology (IST)*.

Các nghiên cứu đã được xuất bản trong các tạp chí /hội nghị có uy tín. Nội dung của các công bố phù hợp với nội dung luận án và có giá trị khoa học.

9. Luận án có đáp ứng các yêu cầu đối với luận án tiến sĩ không? Luận án có thể đưa ra Hội đồng đánh giá luận án tiến sĩ được hay không

Luận án đáp ứng các yêu cầu đối với luận án tiến sĩ. Luận án có thể đưa ra Hội đồng đánh giá luận án tiến sĩ.

10. Kết luận chung cần khẳng định mức độ đáp ứng yêu cầu đối với một luận án tiến sĩ theo quy định của Đại học Quốc gia Hà Nội; bản tóm tắt luận án phản ánh trung thành nội dung cơ bản của luận án hay không

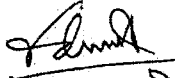
Luận án đáp ứng yêu cầu đối với một luận án tiến sĩ theo quy định của Đại học Quốc gia Hà Nội. Bản tóm tắt luận án phản ánh trung thành nội dung cơ bản của luận án.

11. Các ý kiến khác (nếu có).

Hà Nội, ngày tháng năm 20...

XÁC NHẬN CỦA CƠ QUAN CÔNG TÁC

NGƯỜI NHẬN XÉT
(Ký và ghi rõ họ tên)


Vũ Thị Thiết

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ

Họ và tên người nhận xét: Nguyễn Văn Sơn Học hàm, học vị: TS.

Chuyên ngành: Công nghệ thông tin

Cơ quan công tác: Trường Đại học Công nghệ, ĐHQGHN

Họ và tên nghiên cứu sinh: Phạm Thị Tố Nga

Tên đề tài luận án: Một số phương pháp đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục

Ngành đào tạo: Khoa học máy tính Mã số: 9480101.01

Ý KIẾN NHẬN XÉT

1. Tính cấp thiết, ý nghĩa khoa học và thực tiễn của đề tài luận án

- Đề tài có tính cấp thiết và thời sự cao trong bối cảnh các hệ thống học máy/AI ngày càng được sử dụng trong giáo dục, như hỗ trợ tuyển sinh, đánh giá kết quả học tập, dự đoán nguy cơ bỏ học và hỗ trợ ra quyết định.
- Việc nghiên cứu bảo đảm tính công bằng cho các mô hình học máy có ý nghĩa khoa học rõ ràng, gắn với các hướng nghiên cứu về AI đáng tin cậy, AI có trách nhiệm và khai phá dữ liệu giáo dục.
- Về thực tiễn, các phương pháp đề xuất có tiềm năng hỗ trợ giảm thiên lệch đối với các nhóm người học khác nhau, góp phần tăng tính minh bạch, công bằng và trách nhiệm trong các hệ thống giáo dục thông minh.

2. Sự không trùng lặp của đề tài nghiên cứu; tính trung thực, rõ ràng và đầy đủ trong trích dẫn tài liệu tham khảo

- Theo hiểu biết của người nhận xét và trên cơ sở các tài liệu được cung cấp, chưa thấy dấu hiệu trùng lặp của đề tài và các kết quả chính của luận án với các công trình, luận văn, luận án đã công bố trong và ngoài nước.
- Các tài liệu tham khảo được trích dẫn tương đối đầy đủ, rõ ràng và phù hợp với nội dung nghiên cứu. Danh mục tài liệu phong phú, bao quát các hướng liên quan đến công bằng trong học máy, AI trong giáo dục và sinh dữ liệu tổng hợp.
- NCS cần tiếp tục rà soát để bảo đảm một số tài liệu tham khảo có đủ thông tin xuất bản và thống nhất định dạng trích dẫn.

3. Mức độ tổng quan các vấn đề lý luận của đề tài

- Luận án đã tổng quan khá hệ thống các khái niệm liên quan đến học máy, tính công bằng, thuật toán nhạy cảm, các độ đo công bằng, các giai đoạn can thiệp kỹ thuật và vấn đề đánh đổi giữa công bằng và hiệu suất mô hình.

- Luận án xác định được khoảng trống nghiên cứu quan trọng: nhiều nghiên cứu trước đây tập trung vào từng thuộc tính nhạy cảm riêng lẻ, trong khi thực tế giáo dục thường tồn tại đồng thời nhiều thuộc tính nhạy cảm có tác động giao thoa.
- Phân tổng quan sẽ tốt hơn nếu giảm bớt các mô tả khái niệm quá cơ bản, hạn chế tính liệt kê và tăng thêm phân tích so sánh giữa các nhóm phương pháp liên quan.

4. Sự phù hợp giữa tên đề tài với nội dung, giữa nội dung với chuyên ngành và mã số ngành

- Tên đề tài phản ánh đúng trọng tâm nghiên cứu của luận án là đề xuất và đánh giá các phương pháp nhằm bảo đảm tính công bằng cho hệ thống học máy trong lĩnh vực giáo dục.
- Nội dung nghiên cứu, phương pháp tiếp cận, thực nghiệm và các kết quả chính phù hợp với chuyên ngành Khoa học máy tính và mã số ngành đào tạo.

5. Độ tin cậy và tính hiện đại của phương pháp đã sử dụng để nghiên cứu

- Phương pháp nghiên cứu phù hợp với một luận án theo định hướng Khoa học máy tính ứng dụng: khảo sát tổng quan, xác định khoảng trống, đề xuất phương pháp, thiết kế thực nghiệm và đối sánh kết quả.
- Luận án sử dụng nhiều bộ dữ liệu giáo dục, nhiều mô hình học máy và nhiều độ đo công bằng/hiệu suất, nhờ đó tăng độ tin cậy và khả năng kiểm chứng của kết quả.
- Các kỹ thuật như hồi quy đa biến để loại bỏ phụ thuộc, sinh dữ liệu tổng hợp bằng CTGAN/LLM và phân tích đánh đổi công bằng–hiệu suất thể hiện tính hiện đại của hướng tiếp cận.
- Tuy nhiên, luận án nên làm rõ hơn kiểm định thống kê cho các so sánh thực nghiệm, chất lượng dữ liệu sinh tổng hợp, khả năng tái lập khi sử dụng LLM và độ ổn định của chỉ số đánh đổi.

6. Kết quả nghiên cứu mới của tác giả; đóng góp mới cho sự phát triển khoa học chuyên ngành; đóng góp mới phục vụ sản xuất, kinh tế, xã hội, an ninh, quốc phòng và đời sống; ý nghĩa khoa học, giá trị và độ tin cậy của những kết quả đó

- Luận án có bốn kết quả/dóng góp chính:
 - Đề xuất phương pháp Fairedu, một kỹ thuật tiền xử lý dựa trên hồi quy đa biến nhằm giảm sự phụ thuộc giữa các đặc trưng đầu vào và đồng thời nhiều thuộc tính nhạy cảm.
 - Đề xuất phương pháp DPF nhằm xử lý mất cân bằng giữa các nhóm giao thoa của thuộc tính nhạy cảm thông qua phân nhóm dữ liệu và sinh data tổng hợp.
 - Xây dựng phương pháp FaireduPlus, kết hợp hai cơ chế xử lý theo “chiều dọc” và “chiều ngang”, hướng tới bảo đảm công bằng toàn diện hơn cho dữ liệu dạng bảng trong giáo dục.
 - Đề xuất chỉ số đánh đổi nhằm hỗ trợ định lượng quan hệ giữa cải thiện công bằng và suy giảm/duy trì hiệu suất mô hình.

- Các kết quả trên có ý nghĩa khoa học ở việc giải quyết bài toán công bằng giao thoa đa thuộc tính trong bối cảnh giáo dục; đồng thời có giá trị thực tiễn đối với các hệ thống hỗ trợ ra quyết định trong giáo dục.
- Độ tin cậy của các kết quả được củng cố thông qua thực nghiệm trên nhiều bộ dữ liệu, mô hình và độ đo. Nhưng, NCS cần thảo luận kỹ hơn các trường hợp phương pháp đề xuất chưa vượt trội hoặc phụ thuộc mạnh vào kỹ thuật sinh dữ liệu.

7. Ưu điểm và nhược điểm về nội dung, kết cấu và hình thức của luận án

- Ưu điểm:
 - Luận án có cấu trúc rõ ràng, logic; ba chương phương pháp có quan hệ kế thừa và bổ sung cho nhau, không phải các kết quả rời rạc.
 - Khối lượng thực nghiệm tương đối phong phú, có sử dụng nhiều bộ dữ liệu, nhiều mô hình học máy, nhiều độ đo công bằng và hiệu suất.
 - LA thể hiện được tư duy nghiên cứu có hệ thống: xác định khoảng trống, đề xuất phương pháp, đánh giá thực nghiệm, thảo luận hạn chế và định hướng.
 - Các công bố khoa học liên quan có chất lượng tốt, trong đó có công bố trên các tạp chí uy tín.
- Một số hạn chế/góp ý cần tiếp thu, chỉnh sửa:
 - Nên rút gọn một số tiêu đề chương/mục và các phần giải thích khái niệm quá cơ bản để luận án cô đọng hơn.
 - Phần tổng quan nên tăng tính phân tích, so sánh và tổng hợp, thay vì chủ yếu mô tả hoặc liệt kê các công trình liên quan.
 - Cần bổ sung hoặc làm rõ hơn kiểm định thống kê, khoảng tin cậy hoặc effect size khi so sánh các phương pháp, tránh chỉ dựa vào thống kê thắng/thua.
 - Cần thảo luận sâu hơn các trường hợp phương pháp đề xuất không ổn định hoặc thua phương pháp đối sánh với biên độ đáng kể; đặc biệt là sự phụ thuộc của DPF vào kỹ thuật sinh dữ liệu.
 - Đối với chỉ số đánh đổi, nên làm rõ tính chất toán học, độ ổn định số học và cách xử lý các giá trị ngoại lai hoặc trường hợp mẫu số nhỏ.
 - Cần mô tả chi tiết hơn quy trình sinh dữ liệu bằng LLM/CTGAN, thiết kế prompt, đánh giá chất lượng data sinh, vấn đề tái lập và nguy cơ rò rỉ thông tin.
 - Cần thống nhất ký hiệu trong các thuật toán, chú thích đầy đủ các thành phần trong công thức, rà soát lỗi chính tả, trình bày danh mục từ viết tắt/ký hiệu thuận tiện hơn cho tra cứu.

8. Nội dung luận án đã được công bố trên tạp chí, tuyển tập công trình hội nghị khoa học và giá trị khoa học của những công trình đã công bố

- NCS có 03 công trình đã công bố và 01 công trình đang gửi phản biện liên quan đến nội dung luận án; trong đó có các công trình trên tạp chí quốc tế uy tín.
- Các công bố phản ánh các nội dung chính của luận án, bao gồm khảo sát về tính công bằng trong học máy cho giáo dục và các phương pháp đề xuất như Fairedu/FaireduPlus.

- Các công trình đã công bố có giá trị khoa học, phù hợp với yêu cầu công bố đối với luận án tiến sĩ ngành Khoa học máy tính.

9. Luận án có đáp ứng các yêu cầu đối với luận án tiến sĩ không? Luận án có thể đưa ra Hội đồng đánh giá luận án tiến sĩ được hay không

- Luận án đáp ứng các yêu cầu cơ bản về nội dung, hình thức, khối lượng nghiên cứu, kết quả công bố và đóng góp khoa học đối với một LÁT ngành KHMT;
- Đồng ý để NCS Phạm Thị Tố Nga đưa luận án ra Hội đồng đánh giá LÁT.

10. Kết luận chung cần khẳng định mức độ đáp ứng yêu cầu đối với một luận án tiến sĩ theo quy định của Đại học Quốc gia Hà Nội; bản tóm tắt luận án phản ánh trung thành nội dung cơ bản của luận án hay không

- Luận án đáp ứng yêu cầu đối với một luận án tiến sĩ theo quy định của Đại học Quốc gia Hà Nội.
- Bản tóm tắt luận án về cơ bản phản ánh trung thành nội dung, cấu trúc, các kết quả nghiên cứu chính và đóng góp của luận án.
- NCS nên tiếp thu các góp ý của Hội đồng để hoàn thiện bản luận án cuối cùng, đặc biệt ở các phần trình bày thực nghiệm, phân tích kết quả, kiểm định thống kê, chỉ số đánh đổi và một số vấn đề hình thức.

11. Các ý kiến khác (nếu có)

- NCS có thể xem xét mở rộng nghiên cứu theo hướng tích hợp ràng buộc công bằng vào quá trình huấn luyện mô hình, thay vì chỉ tập trung vào tiền xử lý dữ liệu.
- Các hướng nghiên cứu tiếp theo nên quan tâm thêm đến mô hình học sâu, đánh giá dữ liệu sinh tổng hợp về fidelity, privacy và reproducibility, cũng như xử lý các nhóm giao thoa rất nhỏ hoặc rộng trong dữ liệu giáo dục.

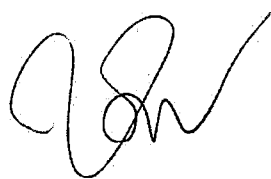
Hà Nội, ngày tháng năm 2026

**XÁC NHẬN CỦA CƠ QUAN CÔNG
TÁC**

(Ký và ghi rõ họ tên)

NGƯỜI NHẬN XÉT

(Ký và ghi rõ họ tên)



TS. Nguyễn Văn Sơn

BẢN NHẬN XÉT PHẢN BIỆN LUẬN ÁN TIẾN SĨ

Tên đề tài luận án: *Một số phương pháp đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục.*

Ngành: Khoa học máy tính

Mã số: 9480101

Nghiên cứu sinh: Phạm Thị Tố Nga

Họ và tên người nhận xét: Phạm Văn Cường; Học hàm, Học vị: PGS. TS.

Ngành: Khoa học máy tính

Cơ quan công tác: Học viện Công nghệ Bưu chính Viễn thông

Ý KIẾN NHẬN XÉT

- Về các ý kiến, nội dung tiếp thu, sửa chữa, giải trình của nghiên cứu sinh về nhận xét của phản biện độc lập;**
NCS đã tiếp thu các ý kiến, sửa chữa, và giải trình về nhận xét của phản biện độc lập.
- Về tính cấp thiết, thời sự, ý nghĩa khoa học và thực tiễn của đề tài luận án;**
Đề tài luận án có tính cấp thiết, thời sự, ý nghĩa khoa học và thực tiễn trong các lĩnh vực trí tuệ nhân tạo, đạo đức AI trong giáo dục, khai phá dữ liệu và học máy; hệ thống AI công bằng v.v..
- Kết luận về sự không trùng lặp (hay trùng lặp) của đề tài nghiên cứu so với các công trình, luận văn, luận án đã công bố ở trong và ngoài nước; tính trung thực, rõ ràng và đầy đủ trong trích dẫn tài liệu tham khảo;**
Luận án không trùng lặp so với các công trình, luận văn, luận án đã công bố ở trong và ngoài nước; tính trung thực, rõ ràng và đầy đủ trong trích dẫn tài liệu tham khảo.
- Mức độ tổng quan các vấn đề lý luận của đề tài;**
Mức độ tổng quan các vấn đề lý luận của đề tài được trình bày tương đối khái quát về phát triển các phương pháp đảm bảo tính công bằng cho các mô hình học máy trong giáo dục.
- Về sự phù hợp giữa tên đề tài với nội dung, giữa nội dung với ngành và mã số ngành;**
Luận án phù hợp giữa tên đề tài với nội dung, giữa nội dung với ngành khoa học máy tính và mã số 9.48.01.01;

6. Sự hợp lý, độ tin cậy và tính hiện đại của các phương pháp đã sử dụng để nghiên cứu:

Phương pháp nghiên cứu kết hợp tiếp cận định tính và định lượng. Luận án tiến hành phân tích, tổng hợp các công trình nghiên cứu liên quan nhằm xác định thách thức, khoảng trống và xu hướng phát triển trong việc bảo đảm tính công bằng cho hệ thống học máy trong lĩnh vực giáo dục. Về định lượng, nghiên cứu triển khai thực nghiệm trên nhiều bộ dữ liệu giáo dục phổ biến (Student Performance, Student Predict Dropout, Oulad, DNU Data) chứa từ hai thuộc tính nhạy cảm trở lên bao gồm các thuộc tính như: giới tính, tuổi, sức khỏe, tình trạng nợ, khu vực, v.v.. Phương pháp can thiệp nhằm đảm bảo tính công bằng được sử dụng là phương pháp tiền xử lý và so sánh thông qua các thước đo công bằng như "tác động khác biệt", "hiệu số chênh lệch thống kê", "chênh lệch trung bình xác suất", và "chênh lệch cơ hội công bằng" và các thước đo hiệu suất bao gồm "độ chuẩn xác", "độ hồi tưởng", "độ chính xác", và "điểm số F1". Phương pháp này tương đối phù hợp.

7. Về những kết quả nghiên cứu mới của tác giả; đóng góp mới cho sự phát triển khoa học ngành; đóng góp mới phục vụ sản xuất, kinh tế, xã hội, an ninh, quốc phòng và đời sống. Đánh giá ý nghĩa, giá trị khoa học và độ tin cậy của những kết quả đó;

Một là luận án đề xuất một kỹ thuật tiền xử lý dựa trên hồi quy bội (FAIREDU) nhằm loại bỏ sự phụ thuộc giữa thuộc tính nhạy cảm và các thuộc tính khác trong dữ liệu huấn luyện;

Hai là luận án đề xuất một phương pháp đảm bảo tính công bằng dựa trên cân bằng dữ liệu bằng việc xử lý mất cân bằng phân phối giữa các nhóm giao nhau nhờ các mô hình sinh dữ liệu (PDF);

Ba là luận án xây dựng phương pháp giải quyết sự mất cân bằng phân phối giữa các nhóm dữ liệu giao nhau và loại bỏ sự phụ thuộc giữa thuộc tính nhạy cảm và không nhạy cảm dựa trên sự kết hợp giữa FAIREDU và DPF; Đồng thời luận án cũng đề xuất một chỉ số đánh đổi nhằm định lượng sự đánh đổi và cân bằng giữa công bằng và hiệu suất mô hình nhằm hỗ trợ so sánh và lựa chọn mô hình phù hợp.

8. Nhận xét về ưu điểm và nhược điểm về nội dung, kết cấu và hình thức của luận án:

Ưu điểm:

Luận án được cấu trúc thành sáu chương, trong đó các chương được sắp xếp theo trình tự logic từ cơ sở lý thuyết đến các đóng góp phương pháp và kết quả nghiên cứu. Chương 1 giới thiệu bối cảnh nghiên cứu, các khái niệm cơ bản, những vấn đề còn tồn tại trong các phương pháp hiện có, đồng thời nêu bật mục tiêu và đóng góp chính của luận án. Chương 2 cung cấp cái nhìn tổng quan về công bằng trong các hệ thống học máy trong lĩnh vực giáo dục. Nội dung chương đề cập đến định nghĩa và các khái niệm liên quan đến công bằng, các phương pháp đảm bảo tính công bằng trong học máy, kỹ thuật sinh dữ liệu tổng hợp và ứng dụng trong nghiên cứu về đảm bảo tính công bằng, các tiêu chí đánh giá công bằng và hiệu suất, cùng với vấn đề đánh đổi giữa hai khía cạnh này. Các chương gồm Chương 3, Chương 4, và Chương 5 trình bày chi tiết ba phương pháp do luận án đề xuất. Mối quan hệ giữa ba chương này chính là mối quan

hệ của ba giai đoạn được trình bày trong Mục 1.3. Trong đó, Chương 3 giới thiệu phương pháp Fairedu, một kỹ thuật tiền xử lý nhằm loại bỏ sự phụ thuộc của các thuộc tính đầu vào vào đồng thời nhiều thuộc tính nhạy cảm, qua đó cải thiện công bằng mà vẫn duy trì hiệu suất mô hình. Chương 4 trình bày phương pháp DPF, sử dụng kỹ thuật sinh dữ liệu tổng hợp để cân bằng phân bố dữ liệu theo tổ hợp các thuộc tính nhạy cảm, khắc phục tình trạng mất cân bằng và thiếu đại diện trong các nhóm yếu thế. Chương 5 giới thiệu phương pháp FaireduPlus, một sự kết hợp giữa Fairedu và DPF nhằm đảm bảo tính công bằng toàn diện các hệ thống học máy sử dụng dữ liệu dạng bảng, với khả năng can thiệp đồng thời theo cả chiều dọc và chiều ngang của bộ dữ liệu. Bên cạnh đó, chương này cũng đề xuất một chỉ số đánh đổi nhằm định lượng sự đánh đổi và cân bằng giữa công bằng và hiệu suất mô hình, qua đó hỗ trợ so sánh và lựa chọn mô hình phù hợp. Cuối cùng, Chương 6 trình bày phần kết luận của luận án, trong đó tổng hợp các kết quả nghiên cứu chính, chỉ ra những hạn chế còn tồn tại, đồng thời đề xuất các hướng nghiên cứu tiếp theo.

Các đóng góp chính của luận án tương đối rõ ràng, lô gic và hợp lý. Các công bố của luận án có chất lượng tốt.

Những điểm cần chỉnh sửa, bổ sung:

- Nhiều đoạn trong luận án được trình bày quá dài (ví dụ trang 13).
- Mục 2.3.2 nên khảo sát các mô hình phổ biến được dùng trong sinh dữ liệu (không chỉ có GAN và LLM) và phân tích rõ hơn các ưu, nhược điểm của mỗi mô hình sinh.
- Giải thích rõ các thành phần và mối liên hệ, các luồng xử lý giữa các khối kiến trúc FAIREDU (hình 3.1 trang 53)
- Phân tích từng bước và phân tích độ phức tạp của thuật toán 3.1 trang 54
- Cần mô tả về bộ dữ liệu DNU: quá trình thu thập, ý nghĩa từng thuộc tính, tính đa dạng, tính bảo đảm công bằng dữ liệu ...
- Bổ sung các luận giải chi tiết hơn các kết quả của bảng 3.3 trang 67.
- Phân tích rõ và so sánh mức độ nhạy cảm giữa hiện tượng đa cộng tuyến vs. mô hình hóa quan hệ phi tuyến?
- Luận giải chi tiết lý do lựa chọn các mô hình CTGAN và LLM để sinh dữ liệu; mô tả quá trình huấn luyện và quá trình sinh cũng như đánh giá chất lượng từ dữ liệu sinh bởi các mô hình này.
- Bổ sung các luận giải chi tiết hơn các kết quả của bảng 4.7, 4.8 trang 99 và 100.
- Mô tả chi tiết kiến trúc tổng thể của phương pháp DPF (hình 4.1 trang 93)
- Giải thích các bước và phân tích, đánh giá thuật toán công bằng nhờ cân bằng dữ liệu DPF (thuật toán 4.1)
- Trình bày rõ hơn các kỹ thuật prompt được sử dụng để sinh dữ liệu với mô hình chatGPT 3.5.
- Giải thích các bước và phân tích, đánh giá thuật toán FAIREDUPLUS.
- Nên cập nhật một số tài liệu tham khảo đã cũ.

9. **Đánh giá về sự phù hợp của nội dung luận án đã được công bố trên tạp chí, tuyển tập công trình hội nghị khoa học và giá trị khoa học của các công trình đã công bố;**
Nội dung luận án đã được công bố trên 4 bài báo trên tạp chí, hội thảo quốc tế uy tín; các công trình đã công bố phù hợp với các nội dung trong luận án.
10. **Luận án có đáp ứng các yêu cầu đối với luận án tiến sĩ hay không? Luận án có thể đưa ra bảo vệ để nhận học vị tiến sĩ được hay không?**
Luận án đáp ứng các yêu cầu đối với luận án tiến sĩ; luận án có thể đưa ra bảo vệ để nhận học vị tiến sĩ.
11. **Kết luận chung cần khẳng định mức độ đáp ứng yêu cầu đối với đối với một luận án tiến sĩ theo quy định của Đại học Quốc gia Hà Nội; Bản tóm tắt luận án có phản ánh trung thành nội dung cơ bản của luận án hay không?**
Luận án đáp ứng được các yêu cầu đối với đối với một luận án tiến sĩ theo quy định của Đại học Quốc gia Hà Nội; Bản tóm tắt luận án có phản ánh trung thành với các nội dung cơ bản của luận án.
12. **Các ý kiến khác (nếu có) :** không.

Hà Nội, ngày tháng năm 2026

NGƯỜI NHẬN XÉT

(Ký và ghi rõ họ tên)



Phạm Văn Cường

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ

Họ và tên người nhận xét: Trần Đăng Hưng Học hàm, học vị: PGS, TS

Chuyên ngành: Khoa học máy tính

Cơ quan công tác: Trường CNTT&TT, Đại học Công nghiệp Hà Nội

Họ và tên nghiên cứu sinh: Phạm Thị Tô Nga

Tên đề tài luận án: Một số phương pháp đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục.

Ngành đào tạo: Khoa học máy tính Mã số: 9480101

NỘI DUNG NHẬN XÉT

1. Tính cần thiết, thời sự, ý nghĩa khoa học và thực tiễn của đề tài luận án

Trong những năm gần đây, sự phát triển mạnh mẽ của trí tuệ nhân tạo và học máy đã tạo ra những chuyên biến sâu sắc trong lĩnh vực giáo dục. Các mô hình học máy ngày càng được ứng dụng rộng rãi trong nhiều hoạt động như tuyển sinh, đánh giá năng lực người học, dự đoán kết quả học tập, cá nhân hóa lộ trình học tập hay hỗ trợ ra quyết định trong quản lý giáo dục. Tuy nhiên, bên cạnh những lợi ích, việc ứng dụng học máy cũng đặt ra nhiều thách thức, trong đó nổi bật là vấn đề đảm bảo tính công bằng. Các mô hình học máy thường được huấn luyện trên dữ liệu lịch sử, có thể chứa đựng những định kiến xã hội, bất bình đẳng kinh tế, xã hội, giới tính, vùng miền hoặc dân tộc. Nếu không được thiết kế và kiểm soát chặt chẽ, các mô hình này có nguy cơ ảnh hưởng đến sự công bằng giữa các đối tượng khác nhau trong giáo dục.

Luận án này nghiên cứu bài toán đảm bảo tính công bằng cho các mô hình học máy trong giáo dục mang ý nghĩa không chỉ về mặt kỹ thuật mà còn về đạo đức, xã hội và chính sách. Thứ nhất, nó góp phần nâng cao độ tin cậy và tính minh bạch của các hệ thống trí tuệ nhân tạo được sử dụng trong nhà trường. Thứ hai, việc đảm bảo công bằng giúp thúc đẩy sự bình đẳng trong tiếp cận giáo dục, tạo điều kiện cho mọi người học.

Vì vậy, nghiên cứu bài toán đảm bảo tính công bằng cho các mô hình học máy trong lĩnh vực giáo dục là có tính cấp thiết trong bối cảnh chuyển đổi số hiện nay. Đây không chỉ là một vấn đề kỹ thuật cần được giải quyết, mà còn là một nhiệm vụ mang ý nghĩa xã hội sâu sắc, góp phần bảo đảm công bằng giáo dục.

2. Sự không trùng lặp của đề tài nghiên cứu so với các công trình đã công bố ở trong và ngoài nước; tính trung thực, rõ ràng và đầy đủ trong trích dẫn tài liệu tham khảo.

Theo tìm hiểu của người đọc, nội dung của luận án chưa phát hiện trùng lặp với nội dung của các bài báo/luận án trong và ngoài nước. Luận án trích dẫn 209 tài liệu tham khảo bằng tiếng Anh. Các trích dẫn đầy đủ và trung thực.

3. Kết quả nghiên cứu mới của tác giả; đóng góp mới cho sự phát triển khoa học chuyên ngành. Ý nghĩa khoa học, giá trị và độ tin cậy của những kết quả đó.

Về cơ bản luận án có các đóng góp sau:

- Luận án đã đề xuất phương pháp tiên xử lý dữ liệu FairEdu nhằm giảm sự phụ thuộc giữa các thuộc tính nhạy cảm và các thuộc tính còn lại trong dữ liệu huấn luyện. Việc lựa chọn các thuộc tính nhạy cảm được thực hiện phù hợp với bối cảnh của từng tập dữ liệu. Kết quả thực nghiệm trên nhiều thuật toán học máy khác nhau cho thấy phương pháp có hiệu quả trong việc cải thiện tính công bằng của mô hình.
- Luận án đề xuất phương pháp tiên xử lý dữ liệu dựa trên tăng cường dữ liệu (DPF), trong đó dữ liệu được phân nhóm theo các tổ hợp giá trị của thuộc tính nhạy cảm và được tăng cường theo quy mô của từng nhóm. Cách tiếp cận này góp phần xử lý vấn đề mất cân bằng dữ liệu giữa các nhóm và đạt được sự cân bằng hợp lý giữa tính công bằng và hiệu năng mô hình trong thực nghiệm.
- Trên cơ sở hai phương pháp trên, luận án xây dựng phương pháp FairEduPlus nhằm kết hợp ưu điểm của FairEdu và DPF. Phương pháp này vừa giải quyết sự mất cân bằng phân phối giữa các nhóm giao nhau, vừa loại bỏ sự phụ thuộc giữa các thuộc tính nhạy cảm và không nhạy cảm. Kết quả cho thấy cách tiếp cận này có khả năng cải thiện tính công bằng một cách ổn định hơn.

4. Nội dung luận án đã được công bố trên tạp chí, tuyển tập công trình hội nghị khoa học và giá trị khoa học của các công trình công bố.

NCS đã công bố 03 bài báo khoa học và 01 bài đang trong quá trình phản biện, trong đó NCS đứng tên đầu cả 4 bài. Trong số 3 bài thì có 2 bài thuộc ISI Q1 và 1 bài hội nghị quốc tế. Các bài báo công bố có nội dung phù hợp với nội dung của luận án.

5. Ưu điểm và những điểm cần cải tiến

Ưu điểm:

- + Luận án nghiên cứu bài toán có ý nghĩa trong lĩnh vực khai phá dữ liệu giáo dục, trong đó vấn đề nghiên cứu đảm bảo tính công bằng cho các thuật toán học máy dựa trên việc điều chỉnh dữ liệu huấn luyện là một bài toán thú vị và có tính nhân văn.
- + Luận án được trình bày công phu, diễn đạt logic và dễ hiểu, các nội dung được viết khá trong sáng; kịch bản và kết quả thực nghiệm khá đa dạng.
- + Luận án có 02 công bố có chất lượng trên các tạp chí ISI Q1.

Một số nhận xét và các điểm cần cải tiến:

- Luận án chỉ nên để bốn chương, gồm chương tổng quan và ba chương là ba đề xuất chính của luận án; Chương 1 và Chương 6 nên để là Phần Mở đầu và Phần Kết luận.
- Chương 2 nên được viết theo kiểu luận án Khoa học máy tính, đưa ra các định nghĩa ngắn gọn và có phân tích thêm các nghiên cứu liên quan đến chủ đề “đảm bảo tính công bằng cho các mô hình học máy trong giáo dục”.
- Cần phát biểu ngắn gọn và đầy đủ hơn về các khái niệm cơ bản: “tính công bằng”, “thuộc tính nhạy cảm”,....
- Nên bỏ Mục 2.4.1 và Mục 2.4.2, các thống kê trong hai mục này không có nhiều ý nghĩa.
- Mặc dù luận án được thiết kế khá logic và nhiều thực nghiệm, tuy nhiên việc hiệu quả của các đề xuất chỉ được thể hiện khi so sánh với dữ liệu gốc; sẽ tốt hơn nếu có thể so sánh với các phương pháp/mô hình đảm bảo tính công bằng khác.
- Nên cung cấp các bảng kết quả liên quan đến độ chính xác (F1) của mô hình sau khi điều chỉnh dữ liệu huấn luyện; vì mục tiêu cuối cùng của các mô hình học máy là độ chính xác càng cao càng tốt, nên mô hình phải vừa đảm bảo tính công bằng vừa phải đảm bảo độ chính xác thì đề xuất mới có ý nghĩa.


- Cần xem lại kí hiệu của các thuộc tính nhạy cảm trong các Thuật toán 3.1, 4.1, 5.1 cho thống nhất, hiện tại đang kí hiệu không thống nhất (x_k^j hay x_k).
- Trong trường hợp số lượng thuật toán nhạy cảm nhiều, khi đó dữ liệu sau khi điều chỉnh liệu còn đảm bảo được phân bố như trong dữ liệu gốc? có ràng buộc gì đối với tham số k và d trong các thuật toán của luận án không?
- Chương 4, làm rõ hơn quá trình sinh thêm dữ liệu từ các bộ dữ liệu sau khi phân nhóm, số lượng sinh bao nhiêu và cách sinh dữ liệu như thế nào cho mỗi nhóm. Đồng thời nên bình luận thêm đối với những trường hợp mà thuộc tính nhạy cảm không có giá trị nhị phân thì như thế nào?
- Các giải pháp trích chọn đặc trưng (feature extraction) trong học máy có đảm bảo được tính công bằng trong các mô hình khai phá dữ liệu giáo dục không?

6. Kết luận

- Luận án đáp ứng đầy đủ yêu cầu về nội dung và hình thức đối với một luận án TS theo quy định của ĐHQGHN.
- Bản tóm tắt phản ánh trung thực nội dung luận án.
- Đồng ý cho NCS đưa luận án ra bảo vệ tại HĐ cấp ĐHQG để nhận học vị TS.

Hà Nội, ngày 4 tháng 5 năm 2026

Người nhận xét



Trần Đăng Hưng

Ngày ký: 06/05/2026 11:23
Người ký: Trần Đăng Hưng
Nơi ký: Đại học Công nghiệp Hà Nội

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ

Họ và tên người nhận xét: Lê Hồng Phương Học hàm, học vị: PGS. TS.
Chuyên ngành: Khoa học Máy tính
Cơ quan công tác: Trường ĐH Khoa học Tự nhiên, ĐHQGHN
Họ và tên nghiên cứu sinh: Phạm Thị Tố Nga
Tên đề tài luận án: Một số phương pháp đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục
Ngành đào tạo: Khoa học Máy tính Mã số: 9480101.01

Ý KIẾN NHẬN XÉT¹

- Tính cấp thiết, ý nghĩa khoa học và thực tiễn của đề tài luận án**
 - Đề tài là cấp thiết và có ý nghĩa khoa học và thực tiễn.
- Sự không trùng lặp của đề tài nghiên cứu so với các công trình, luận văn, luận án đã công bố trong và ngoài nước; tính trung thực, rõ ràng và đầy đủ trong trích dẫn tài liệu tham khảo**
 - Đề tài nghiên cứu không trùng lặp.
 - Các trích dẫn là rõ ràng và trung thực.
- Mức độ tổng quan các vấn đề lý luận của đề tài**
 - Luận án thiên về khía cạnh cải tiến kỹ thuật và ứng dụng. Luận án có kết cấu 6 chương theo mô hình chuẩn (giới thiệu – tổng quan – 3 chương phương pháp đề xuất – kết luận), cộng với 3 phụ lục chi tiết.
 - Đây là kết cấu hợp lý, mạch lạc; mỗi chương phương pháp (3, 4, 5) đều theo khung thống nhất giới thiệu → phương pháp → thực nghiệm → kết quả → thảo luận → hạn chế → tổng kết, giúp người đọc dễ theo dõi và so sánh giữa các chương.

Ba phương pháp đề xuất có quan hệ kế thừa rõ ràng:

- Fairedu (chương 3): tiên xử lý theo “chiều dọc” – dùng hội quy đa biến để loại bỏ phụ thuộc giữa đặc trưng và nhiều thuộc tính nhạy cảm đồng thời (mở rộng từ LTDD đơn biến).
- DPF (chương 4): can thiệp theo “chiều ngang” – cân bằng phân bố nhóm giao thoa bằng dữ liệu tổng hợp (CTGAN, LLM).
- FaireduPlus (chương 5): kết hợp cả hai chiều, cộng với đề xuất “chỉ số đánh đổi” công bằng–hiệu suất.

Đây là một thiết kế nghiên cứu logic, có tính tích lũy, không phải ba mảnh ghép rời rạc.

4. Sự phù hợp giữa tên đề tài với nội dung, giữa nội dung với chuyên ngành và mã số ngành

- Tên đề tài phù hợp với nội dung.
- Nội dung và phương pháp của luận án phù hợp với chuyên ngành Khoa học Máy tính, mã số 9480101.01.

5. Độ tin cậy và tính hiện đại của phương pháp đã sử dụng để nghiên cứu

- Phương pháp sử dụng là tin cậy và hiện đại.

6. Kết quả nghiên cứu mới của tác giả; đóng góp mới cho sự phát triển khoa học chuyên ngành; đóng góp mới phục vụ sản xuất, kinh tế, xã hội, an ninh, quốc phòng và đời sống. Ý nghĩa khoa học, giá trị và độ tin cậy của những kết quả đó

Luận án có bốn đóng góp, được công bố trong 2 bài báo Q1 (ESWA, JSS) và 1 bài đang gửi (IST) là tín hiệu tốt về việc cộng đồng đã thẩm định một phần kết quả. Điểm mới đáng ghi nhận:

- Giải quyết bài toán công bằng giao thoa đa thuộc tính trong giáo dục – một khoảng trống thực sự, vì phần lớn nghiên cứu trước (Reweighing, FairSMOTE, LTDD) chỉ xử lý từng thuộc tính nhạy cảm riêng lẻ.
- Tổng quát hóa LTDD từ hội quy đơn biến sang đa biến – đơn giản về ý tưởng nhưng giải quyết đúng hạn chế cốt lõi của phương pháp gốc.
- Tích hợp LLM như một kỹ thuật sinh dữ liệu tổng hợp song song với CTGAN – khả thi sự.
- “Chỉ số đánh đổi” công bằng–hiệu suất là ý tưởng có giá trị thực tiễn (hỗ trợ lựa chọn mô hình), tuy về mặt phương pháp luận vẫn còn ở mức sơ khởi ban đầu.

7. Ưu điểm và nhược điểm về nội dung, kết cấu và hình thức của luận án

Luận án có các ưu điểm sau:

- Thực nghiệm đa dạng: 4 bộ dữ liệu (3 mở + 1 dữ liệu thật DNU), 5 mô hình học máy, 2 kỹ thuật sinh dữ liệu, 4 độ đo công bằng, so sánh với 3-4 baseline hiện đại (Reweighing, FairSMOTE, LTDD). Nhiều bảng/biểu đồ (Bảng 3.3–3.6, 4.4–4.10, 5.2–5.6, cộng phụ lục).

- Tự phê bình nghiêm túc: mỗi chương đều có mục "Hạn chế nghiên cứu" áp dụng khung Runeson (độ tin cậy nội tại/bên ngoài, giá trị cấu trúc, giá trị kết luận) – hiếm thấy ở mức độ chi tiết này trong luận án ứng dụng, thể hiện tư duy khoa học tốt.
- Tính ứng dụng rõ: gắn với dữ liệu thật của một trường đại học, có thảo luận hàm ý thực tiễn cho tuyển sinh, chấm điểm tự động, dự đoán bỏ học.
- Kết quả được báo cáo trung thực cả khi không thuận lợi cho phương pháp đề xuất (ví dụ Bảng 4.10: Origin vẫn thắng Accuracy nhiều nhất, DPF không áp đảo tuyệt đối) – không "thối phồng" kết quả.

Những hạn chế và thiếu sót cần khắc phục: 8 điểm chính.

a) Tính nhất quán nội tại của Fairedu (Chương 3)

- Khi đối chiếu các Bảng 3.4–3.6, Fairedu không nhất quán thắng LTDD. Ví dụ ở mô hình hồi quy logistic, thuộc tính SD_GT: Fairedu = 0.740 trong khi LTDD chỉ = 0.141 (Bảng 3.4) — Fairedu thua khá xa, không chỉ "thua nhẹ". Với rừng ngẫu nhiên, SD_GT: Fairedu = 0.631 vs LTDD = 0.396; DNU_Tuổi: Fairedu = 0.192 vs LTDD = 0.056 – đều là những khoảng cách lớn.
- Tỷ lệ thắng/hòa/thua tổng hợp (ví dụ 5/0/4) che lấp việc một số trường hợp thua có biên độ rất lớn. Luận án nên báo cáo thêm độ lớn chênh lệch (effect size), không chỉ đếm lượt thắng/thua, và lý giải tại sao Fairedu lại kém ổn định hơn LTDD chính trên thuộc tính "tình trạng nợ"/"giới tính" ở các tập dữ liệu thưa – đây có lẽ là dấu hiệu hồi quy đa biến gặp đa cộng tuyến hoặc quá khớp khi kết hợp nhiều biến nhạy cảm tương quan với nhau, hoặc tập dữ liệu kích thước nhỏ.

b) Thiếu kiểm định thống kê chặt chẽ cho so sánh thắng/thua

- Toàn luận án dùng cách đếm "thắng/hòa/thua" dựa trên so sánh điểm đơn lẻ giữa các cấu hình (không phải trên nhiều lần lặp có kiểm định ý nghĩa thống kê, trừ một vài trường hợp ở chương 3 có nhắc $p < 0.05$ cho kiểm định Wald khi ước lượng hệ số hồi quy, chứ không phải cho so sánh hiệu quả giữa phương pháp).
- Việc này khiến các con số như "27/45 thắng" có thể bị ảnh hưởng bởi khác biệt rất nhỏ, không có ý nghĩa thực tế. Nên bổ sung kiểm định Wilcoxon signed-rank hoặc khoảng tin cậy bootstrap cho các so sánh tổng hợp.

c) Chỉ số đánh đổi (trade-off index) ở Chương 5

- Bảng 5.5–5.6 cho thấy chỉ số đánh đổi có những giá trị cực đoan bất thường: -9750.4%, -4601.8%, -2517.9%, -2317.8%... Những giá trị phần trăm âm hàng nghìn lần cho thấy công thức tính (rất có thể là cải thiện tương đối chia cho baseline gần 0) bị bất ổn định khi mẫu số (mức công bằng/hiệu suất gốc) tiến gần 0 – đây là lỗi kinh điển của các chỉ số "cải thiện tương đối". Hệ quả là: Các thông kê thắng/thua dựa trên chỉ số này (Bảng 5.5, 5.6) có thể bị chi phối bởi vài giá trị ngoại lai cực đoan, làm méo mó kết luận tổng thể. Luận án chưa làm rõ liệu các trường hợp giá trị "khủng" này có được loại trừ, hay dùng trung vị thay cho trung bình khi tổng hợp không.
- Việc đề xuất chỉ số đánh đổi là một trong 4 đóng góp chính, nhưng chính chương 5 cũng tự nhận "*chỉ số đánh đổi mới dừng ở mức khái niệm và thử nghiệm ban*

đầu” – nên đây là điểm cần làm rõ thêm về tính ổn định số học của công thức, không chỉ về phạm vi áp dụng.

d) Mâu thuẫn giữa kết luận tổng thể và dữ liệu chi tiết ở Chương 4 (DPF)

- Bảng 4.7 cho thấy với CTGAN, cấu hình SDG (không phân tách nhóm) thắng DPF ở chỉ số |I-DI| (23 so với 11) và SPD (21 so với 14) – DPF chỉ thắng ở AOD, EOD. Với LLM thì DPF mới áp đảo toàn diện. Điều này có nghĩa hiệu quả của DPF phụ thuộc mạnh vào kỹ thuật sinh dữ liệu được dùng kèm, chứ DPF không phải lúc nào cũng vượt trội – đây là một phát hiện thú vị nhưng phần thảo luận tổng kết chương có xu hướng diễn giải nghiêng về phía thuận lợi cho DPF hơn là nhấn rõ sự phụ thuộc/điều kiện này.
- Tương tự, Bảng 4.10 cho thấy Origin (không can thiệp) thắng Accuracy nhiều nhất (26/45), DPF chỉ 13/45 – cho thấy can thiệp công bằng vẫn phải trả giá về hiệu suất rõ ràng hơn so với cách luận án mô tả (“duy trì hiệu suất cạnh tranh”).

d) Vấn đề dữ liệu thừa/nhóm giao thoa rộng chưa được giải quyết triệt để

- Cả ba phương pháp đều thừa nhận hạn chế khi nhóm còn có 0 hoặc 1 mẫu (DNU Data) – các nhóm này bị loại khỏi quá trình cân bằng.
- Đây chính là nhóm dễ bị bỏ sót nhất trong thực tế (nhóm thiểu số cực nhỏ), nên về bản chất phương pháp “bò qua” đúng nơi cần công bằng nhất – một nghịch lý cần được thảo luận sâu hơn, ví dụ dùng kỹ thuật tăng cường dữ liệu chuyên biệt, hoặc gộp nhóm thông minh hơn là loại bỏ.

e) Thiếu kiểm soát đa kiểm định

- Thuật toán Fairedu dùng kiểm định Wald với ngưỡng $p < 0.05$ cho mỗi đặc trưng không nhạy cảm và mỗi thuộc tính nhạy cảm, lặp lại trên nhiều bộ dữ liệu/mô hình mà không có điều chỉnh Bonferroni hay FDR – có nguy cơ dương tính giả khi xác định “phụ thuộc có ý nghĩa thống kê” cần điều chỉnh.
- Gợi ý: thực hiện điều chỉnh FDR và kiểm định thống kê.

f) Phạm vi mô hình và dữ liệu còn hẹp

- Chỉ kiểm chứng trên mô hình truyền thống (logistic regression, decision tree, random forest, gradient boosting, mạng nơ-ron đơn giản), không có mô hình sâu/transformer. Với bối cảnh năm 2025-2026 thì đây là một giới hạn về tính kịp thời.

g) Tính minh bạch của dữ liệu sinh tổng hợp bởi LLMs

- Luận án không nêu rõ cách thiết kế prompt, cách kiểm định chất lượng/độ trung thực của dữ liệu sinh ra (fidelity, privacy leakage) — đây là khoảng trống về tái lập (reproducibility)
- LLM thay đổi nhanh, ảnh hưởng tính tái lập được của các thí nghiệm.

h) Một số lỗi kỹ hiệu và nhất quán

- Kí hiệu véc-tơ toán học: nên dùng ngoặc nhọn thay cho ngoặc vuông (p.51)
- Một số lỗi chính tả (p. 67)

8. Nội dung luận án đã được công bố trên tạp chí, tuyển tập công trình hội nghị khoa học và giá trị khoa học của những công trình đã công bố

- Luận án có 4 công bố khoa học.
- Các công trình đáp ứng yêu cầu chất lượng.

9. Luận án có đáp ứng các yêu cầu đối với luận án tiến sĩ không? Luận án có thể đưa ra Hội đồng đánh giá luận án tiến sĩ được hay không

- Luận án cơ bản đáp ứng yêu cầu, có thể đưa ra Hội đồng đánh giá luận án tiến sĩ.

10. Kết luận chung cần khẳng định mức độ đáp ứng yêu cầu đối với một luận án tiến sĩ theo quy định của Đại học Quốc gia Hà Nội: bản tóm tắt luận án phản ánh trung thành nội dung cơ bản của luận án hay không

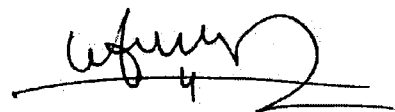
- Luận án giải quyết một vấn đề có ý nghĩa thực tiễn và học thuật rõ ràng (công bằng da thuộc tính trong học máy trong giáo dục), có khối lượng thực nghiệm đủ, kể thừa hợp lý qua 3 phương pháp, đã được kiểm định một phần qua các bài báo khoa học.
- Tuy nhiên, các bảng kết quả chi tiết cho thấy mức độ ưu việt của các phương pháp đề xuất không đồng nhất và đôi khi mâu thuẫn với phần diễn giải tổng hợp, đặc biệt là “chỉ số đánh đổi” ở chương 5 có vấn đề về tính ổn định số học cần được làm rõ trước khi công bố rộng.
- Tác giả nên bổ sung kiểm định thống kê chặt chẽ hơn cho các so sánh, xử lý vấn đề ngoại lai trong chỉ số đánh đổi, và thảo luận sâu hơn về các trường hợp phương pháp đề xuất thua kém phương pháp cơ bản với biên độ lớn, thay vì chỉ trình bày tỷ lệ thắng/thua tổng hợp.

11. Các ý kiến khác (nếu có). Không

Hà Nội, ngày 20 tháng 6 năm 2026

XÁC NHẬN CỦA CƠ QUAN CÔNG TÁC

NGƯỜI NHẬN XÉT
(Ký và ghi rõ họ tên)



Lê Hồng Phương

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ

Họ và tên người nhận xét: Phan Xuân Hiếu..... Học hàm, học vị: PGS.TS

Chuyên ngành: Khoa học thông tin và máy tính

Cơ quan công tác: Khoa CNTT, Trường ĐH Công nghệ, ĐHQGHN

Họ và tên nghiên cứu sinh: Phạm Thị Tố Nga

Tên đề tài luận án: *Một số phương pháp đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục*

Ngành đào tạo: Khoa học máy tính..... Mã số: 9480101.01

Ý KIẾN NHẬN XÉT¹

1. Tính cấp thiết, ý nghĩa khoa học và thực tiễn của đề tài luận án

Đảm bảo tính công bằng trong các mô hình học máy nói chung và cho lĩnh vực giáo dục nói riêng là vấn đề được quan tâm trong những năm qua. Khi các mô hình học máy hiện đại ngày càng cho hiệu suất cao trên các tập dữ liệu thì việc xem xét chúng ở góc độ “công bằng” trở nên quan trọng, đặc biệt là cho các lĩnh vực ứng dụng mà sự phân loại, đoán nhận của các mô hình chịu ảnh hưởng bởi các thuộc tính nhạy cảm.

Luận án này đặt trọng tâm nghiên cứu vào việc đảm bảo tính công bằng cho các mô hình học máy trong miền ứng dụng giáo dục bằng nhiều giải pháp khác nhau, ở các giai đoạn khác nhau bao gồm lại bỏ tính phụ thuộc, sinh dữ liệu tổng hợp để đảm bảo tính cân bằng dữ liệu .v.v.

Các đề xuất giải pháp dựa trên các nghiên cứu liên quan của cộng đồng và dựa trên thực nghiệm. Đề tài và các đóng góp có ý nghĩa khoa học và tiềm năng áp dụng thực tiễn.

2. Sự không trùng lặp của đề tài nghiên cứu so với các công trình, luận văn, luận án đã công bố trong và ngoài nước; tính trung thực, rõ ràng và đầy đủ trong trích dẫn tài liệu tham khảo

Không có sự trùng lặp theo hiểu biết của người đọc.

3. Mức độ tổng quan các vấn đề lý luận của đề tài

Luận án xem xét tổng thể vấn đề về đảm bảo tính công bằng trong học máy nói chung và cho miền giáo dục nói riêng. Nhìn chung, luận án đảm bảo tính tổng quan ở nhiều khía cạnh như sau:

- Phân tích các nghiên cứu liên quan và tìm khoảng trống nghiên cứu. NCS đã khảo sát 63 bài báo liên quan đến chủ đề công bằng học máy cho lĩnh vực giáo dục.
 - Khảo sát kỹ lưỡng các khái niệm “công bằng” theo nhiều góc độ khác nhau
 - Mô tả các “độ đo công bằng” dựa trên các nghiên cứu đã có và mô tả cụ thể để áp dụng cho các nội dung nghiên cứu của luận án.
 - Các lập luận và đề xuất giải pháp cho các nội dung nghiên cứu có tính khái quát.
4. Sự phù hợp giữa tên đề tài với nội dung, giữa nội dung với chuyên ngành và mã số ngành
- Tên đề tài phù hợp với nội dung luận án.
 - Nội dung luận án phù hợp với chuyên ngành và mã số ngành đào tạo.
5. Độ tin cậy và tính hiện đại của phương pháp đã sử dụng để nghiên cứu
- Phương pháp nghiên cứu phù hợp và có tính hiện đại.
6. Kết quả nghiên cứu mới của tác giả; đóng góp mới cho sự phát triển khoa học chuyên ngành; đóng góp mới phục vụ sản xuất, kinh tế, xã hội, an ninh, quốc phòng và đời sống. Ý nghĩa khoa học, giá trị và độ tin cậy của những kết quả đó

Luận án có 04 đóng góp chính như sau:

- Đề xuất phương pháp Fairedu nhằm nâng cao tính công bằng cho đồng thời nhiều thuộc tính nhạy cảm thông qua việc loại bỏ sự phụ thuộc của các thuộc tính đầu vào trong dữ liệu huấn luyện vào các thuộc tính nhạy cảm.
- Đề xuất phương pháp DPF dựa trên kỹ thuật sinh dữ liệu tổng hợp, nhằm cân bằng phân bố giữa các tổ hợp thuộc tính nhạy cảm, qua đó đảm bảo tính công bằng trong dữ liệu huấn luyện.
- Phát triển phương pháp FaireduPlus, đảm bảo tính công bằng hai chiều cho bộ dữ liệu dạng bảng với đồng thời nhiều thuộc tính nhạy cảm.
- Đề xuất “chỉ số đánh đổi” nhằm định lượng mối quan hệ giữa công bằng và hiệu suất mô hình học máy làm cơ sở so sánh và lựa chọn mô hình một cách có hệ thống.

Các đóng góp của NCS về cơ bản là hợp lý. Tuy nhiên một số điểm cần làm rõ thêm (xem các nhận xét và góp ý ở mục 11 dưới đây).

7. Ưu điểm và nhược điểm về nội dung, kết cấu và hình thức của luận án

Ưu điểm:

- Luận án trình bày tương đối rõ ràng.
- Hình 1.1 (trang 10) giúp làm sáng tỏ cấu trúc các vấn đề nghiên cứu và các giải pháp đề xuất của luận án. Tương tự cho hình 1.2 (mối quan hệ giữa các chương)
- Luận án có 2 công bố trên các tạp chí ISI uy tín

Nhược điểm:

- Xem các nhận xét và góp ý ở mục 11 dưới đây.

8. Nội dung luận án đã được công bố trên tạp chí, tuyển tập công trình hội nghị khoa học và giá trị khoa học của những công trình đã công bố

Luận án có 03 công trình đã công bố và 01 công trình đang được gửi phản biện. Tất cả các công trình NCS đều là tác giả đầu. Cụ thể các công trình như sau:

- Công trình số 1: Bài báo tại hội nghị quốc tế năm 2022 về “Fairness requirement in AI engineering” <= công trình này không liên quan nhiều đến nội dung chính của luận án, chỉ liên quan ở mức chủ đề, đó là tính công bằng (fairness)
- Công trình số 2: Bài báo ở tạp chí Journal of Systems and Software (tạp chí Q1) về “Fairness for machine learning software in education: A systematic mapping study” <= công trình này có chủ đề gần với luận án, đó là tính công bằng cho học máy trong lĩnh vực giáo dục, tuy nhiên đây là bài báo dạng khảo cứu (survey) chứ chưa phải công bố nội dung thực hiện của luận án. Tuy vậy, đây cũng là một nỗ lực tốt của nghiên cứu sinh.
- Công trình số 3: Bài báo ở tạp chí Expert Systems with Applications (tạp chí Q1) với tiêu đề “FAIREDU: A Multiple Regression-Based Method for Enhancing Fairness in Machine Learning Models for Educational Applications” <= công bố này liên quan trực tiếp đến nội dung nghiên cứu của luận án.
- Công trình số 4: Đang được submit vào tạp chí Q1 với tiêu đề “FaireduPlus: Enhancing Intersectional Fairness in Education Focused Machine Learning Using Synthetic Data” <= công trình này cũng liên quan trực tiếp đến nội dung nghiên cứu của luận án.

Các công bố của NCS đáp ứng yêu cầu của luận án tiến sĩ ngành khoa học máy tính. NCS cùng tập thể hướng dẫn đã hướng đến các công bố khoa học vào những tạp chí uy tín, chất lượng cao.

9. Luận án có đáp ứng các yêu cầu đối với luận án tiến sĩ không? Luận án có thể đưa ra Hội đồng đánh giá luận án tiến sĩ được hay không

Luận án đáp ứng được yêu cầu của luận án tiến sĩ chuyên ngành khoa học máy tính. Đồng ý đưa ra Hội đồng đánh giá luận án tiến sĩ.

10. Kết luận chung cần khẳng định mức độ đáp ứng yêu cầu đối với một luận án tiến sĩ theo quy định của Đại học Quốc gia Hà Nội; bản tóm tắt luận án phản ánh trung thành nội dung cơ bản của luận án hay không

- Luận án đáp ứng đầy đủ các quy định về công bố khoa học, cấu trúc, hình thức và nội dung của luận án theo quy định của ĐHQG Hà Nội.
- Bản tóm tắt phản ánh trung thành nội dung của luận án.

11. Các ý kiến khác (nếu có).

- **Tiêu đề các chương, mục hiện còn hơi dài dòng:**
 - o Tiêu đề Mục 1.1.2 (Các thách thức và khoảng trống nghiên cứu trong đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục) => nên để gọn lại thành “Các thách thức và khoảng trống nghiên cứu” là đủ.
 - o Tiêu đề Chương 2 có thể để ngắn gọn lại như sau: “Tổng quan về đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục” (bỏ cụm “trí tuệ nhân tạo”)
 - o Tương tự còn nhiều mục khác và tiêu đề Chương 3 cũng còn khá dài dòng, nên làm gọn bớt.
- Mục 2.1.1. (Các khái niệm) mô tả các khái niệm như “trí tuệ nhân tạo”, “học máy”, “dữ liệu”, “thuật toán học máy”, “mô hình học máy” ... không thật sự cần thiết. Đây là các khái niệm cơ sở gần như độc giả nào của luận án này cũng đã nắm bắt được. Vì thế, có thể xem xét bỏ mục này để rút ngắn độ dài của luận án.
- **Về các công thức và mô hình hoá:**
 - o Cần chú thích ý nghĩa của các thành phần trong các công thức ở Bảng 2.1 (Các định nghĩa về công bằng trong hệ thống học máy, trang 22). Ví dụ: Y là gì, Y mũ là gì, A là gì ...? Dự đoán “tích cực” có phải là dự đoán vào “lớp dương” (positive)? Nếu vậy thì không nên dịch là “tích cực” mà dịch là “lớp dương”.
 - o Tại sao chỉ phát biểu cho bài toán phân lớp nhị phân? (công thức 2.5, trang 25)
 - o Tại sao thuộc tính A lại chỉ có trường hợp nhị phân, tức A chia dữ liệu thành hai nhóm (*nhóm ưu tiên và nhóm không ưu tiên*)? Tại sao không tổng quát hoá A chia dữ liệu thành k nhóm? (trang 25)
 - o Ở độ đo về “Hiệu số chênh lệch thống kê” (SPD, công thức 2.7, trang 26): “Giá trị SPD càng nhỏ (càng gần 0) thể hiện sự công bằng theo định nghĩa công bằng nhân khẩu học” => SPD có thể có giá trị âm? Nếu vậy thì phải viết là “trị tuyệt đối của SPD càng nhỏ (càng gần 0) ...”
 - o Công thức tính “Chênh lệch trung bình xác suất” (AOD) ở trang 26 có cần lấy giá trị tuyệt đối của hiệu TPR và trị tuyệt đối của hiệu FPR của hai nhóm trước khi cộng vào nhau hay không?
 - o Hình 3.1 (“Kiến trúc tổng thể của phương pháp Fairedu”, trang 53): nên làm rõ hai ý: (1) loại bỏ các thuộc tính nhạy cảm, và (2) loại bỏ sự phụ thuộc giữa các thuộc tính không nhạy cảm đối với các thuộc tính nhạy cảm. Hiện hình 3.1 mới mô tả ý thứ (2).
- **Các phát biểu cần xem xét:**
 - o “Đây là vấn đề phổ biến trong dữ liệu giáo dục, nơi các nhóm yếu thế thường có số lượng mẫu rất nhỏ, gây khó khăn cho việc huấn luyện mô hình và đảm bảo công bằng.” (trang 9, mục 1.2.3 – Phạm vi nghiên cứu, nhánh 2) <= Tại sao “các nhóm yếu thế” lại thường “có số lượng mẫu rất nhỏ”? Ví dụ một tập dữ liệu về sinh viên thì có khi 70% sinh viên đến từ nông thôn so với 30% sinh viên đến từ thành phố. Có thể xem sinh viên đến từ nông thôn là nhóm yếu thế, nhưng số lượng mẫu của họ lại áp đảo? “Yếu thế” không hoàn toàn đồng nghĩa với “thiểu số”, thậm chí, theo nguyên lý 80-20 (Pareto) thì nhóm “yếu thế” có khi còn chiếm số đông.
- **Lỗi hành văn và chính tả:**

- “Nghiên cứu về đảm bảo tính công bằng cho các hệ thống học máy trong giáo dục đang đối mặt với nhiều thách thức đặc thù, phản ánh sự phức tạp của cả về kỹ thuật lẫn ứng dụng.” (trang 4)
- “Nghiên cứu các cơ chế tiên xử lý nhằm giảm phụ thuộc giữa đặc trưng và thuộc tính nhạy cảm, hỗ trợ công bằng đa nhóm.” (trang 6, mục tiêu cụ thể số 1) <= “đặc trưng” và “thuộc tính” khác gì nhau?
- “Đảm bảo TPR và FPR bằng nhau giữa các nhóm (ví dụ: nam và nữ)” (trang 22, bảng 2.1) => những cụm như thế này rất dễ gây hiểu nhầm.
- “Để giải quyết vấn đề này, chương trình bày phương pháp FairEdu, một cách tiếp cận tiên xử” (trang 48) <= “... chương này trình ...”

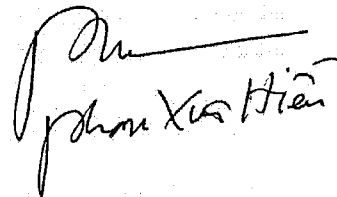
Các câu hỏi:

- Tại sao NCS không đề xuất phương pháp nào thuộc “Giai đoạn xử lý trong quá trình”, ví dụ tích hợp việc đảm bảo tính công bằng vào hàm loss, hàm tối ưu của mô hình học máy? Điều này có ưu và nhược điểm gì? Có những khó khăn gì xét trong ngữ cảnh luận án của NCS?
- Các độ đo “công bằng” ở trên chủ yếu tính toán dựa trên tập dữ liệu đã gắn nhãn. Có cơ chế nào để phát hiện ra các tập dữ liệu đang bị gắn nhãn có yếu tố thiên vị (nghĩa là ngay từ khi gắn nhãn đã không công bằng)?
- Trong phương pháp FairEdu, tại sao lại giả định rằng các thuộc tính không nhạy cảm có phụ thuộc tuyến tính (đa biến) với các thuộc tính nhạy cảm? Mối quan hệ tuyến tính này có tính phổ quát không? Ví dụ trong một tập dữ liệu nào đó, “thu nhập” sẽ có quan hệ phi tuyến với “tuổi tác” (tuổi trung niên thường có thu nhập lớn hơn so với tuổi trẻ và tuổi già).

Hà Nội, ngày tháng năm 2026

XÁC NHẬN CỦA CƠ QUAN CÔNG TÁC

NGƯỜI NHẬN XÉT
(Ký và ghi rõ họ tên)



Lưu ý

- Nhận xét của phản biện không cần nêu lại kết cấu và nội dung các chương của luận án, không trình bày theo thứ tự chương mục của luận án mà cần trình bày theo các vấn đề hay kết quả mới mà luận án đạt được.
- Những thiếu sót và tồn tại về nội dung, hình thức của luận án (nếu có) phải được nêu cụ thể, rõ ràng và ngắn gọn.

TRÍCH QUY CHẾ ĐÀO TẠO TIẾN SĨ TẠI ĐẠI HỌC QUỐC GIA HÀ NỘI

BẢN NHẬN XÉT LUẬN ÁN TIẾN SĨ

Tên đề tài luận án: “*Một số phương pháp đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục*”.

Chuyên ngành: Khoa học máy tính Mã số: 9480101.01

Nghiên cứu sinh: Phạm Thị Tố Nga

Họ và tên người nhận xét: Nguyễn Thị Minh Huyền

Học hàm, Học vị: Tiến sĩ

Chuyên ngành: Tin học

Cơ quan công tác: Trường ĐH Khoa học Tự nhiên, ĐHQG Hà Nội

Ý KIẾN NHẬN XÉT

- Về tính cấp thiết, thời sự, ý nghĩa khoa học và thực tiễn của đề tài luận án:
Đề tài luận án tập trung nghiên cứu vấn đề là đảm bảo công bằng cho các mô hình học máy ứng dụng trong lĩnh vực giáo dục với nhiều thuộc tính nhạy cảm. Đây là đề tài có tính thời sự, có ý nghĩa khoa học và thực tiễn, liên quan tới hai hướng nghiên cứu đang rất được quan tâm hiện nay: AI đáng tin cậy (*Trustworthy AI*) và ứng dụng AI trong giáo dục (*AI in Education*).
- Kết luận về sự không trùng lặp (hay trùng lặp) của kết quả nghiên cứu của luận án với kết quả của các công trình khoa học của các tác giả khác; tính trung thực, rõ ràng và đầy đủ trong trích dẫn tài liệu tham khảo:
Đề tài và kết quả của luận án không trùng lặp với các công trình nghiên cứu khác. Theo hiểu biết của người phản biện, các trích dẫn trong luận án là trung thực, tương đối rõ ràng và đầy đủ.
- Về sự phù hợp giữa tên đề tài với nội dung, giữa nội dung với chuyên ngành và mã số chuyên ngành:
Đề tài luận án phù hợp với mã số chuyên ngành.
- Sự hợp lý, độ tin cậy và tính hiện đại của các phương pháp đã sử dụng để nghiên cứu:
Luận án thiên về nghiên cứu thực nghiệm và phát triển giải pháp ứng dụng. Phương pháp nghiên cứu có hệ thống, phù hợp với mục tiêu đề ra. Các giải pháp được kiểm chứng và đối sánh bằng thực nghiệm trên nhiều bộ dữ liệu. Phương pháp nghiên cứu là hợp lý, có tính tin cậy.
- Về những kết quả nghiên cứu mới của tác giả; nêu rõ những kết quả mới; đánh giá mức độ đóng góp cho sự phát triển khoa học chuyên ngành; đóng góp mới phục vụ sản xuất, kinh

tế, xã hội, an ninh, quốc phòng và đời sống. Đánh giá ý nghĩa, giá trị khoa học và độ tin cậy của những kết quả đó:

Luận án có 4 đóng góp chính:

- 1) Đề xuất Fairedu là mở rộng LTDD bằng cách thay hồi quy đơn biến bằng hồi quy đa biến để loại bỏ đồng thời ảnh hưởng của nhiều thuộc tính nhạy cảm, qua đó cải thiện công bằng trong khi vẫn duy trì hiệu suất mô hình.
Về mặt học thuật, Fairedu là mở rộng tự nhiên từ hồi quy đơn biến sang hồi quy đa biến. Đóng góp chính nằm ở phương pháp ứng dụng, giúp giải quyết được bài toán nhiều thuộc tính nhạy cảm.
- 2) Xây dựng phương pháp hỗ trợ phân chia dữ liệu công bằng DPF (*Data Partitioning Fairness*) bằng cách chia dữ liệu theo các nhóm giao thoa và sử dụng các công cụ như CTGAN và LLM để sinh dữ liệu tổng hợp cho các nhóm thiếu dữ liệu nhằm tái cân bằng dữ liệu có kiểm soát.
Đây là đóng góp có tính thực tiễn cao, giúp xử lý được những vấn đề thiên lệch do số nhóm giao thoa có nhiều.
- 3) Đề xuất FaireduPlus nhằm đảm bảo công bằng hai chiều: chiều ngang (cân bằng dữ liệu) và chiều dọc (loại bỏ phụ thuộc) bằng cách kết hợp hai cơ chế DPF và Fairedu.
Đề xuất này giúp giải quyết đồng thời hai nguồn gốc của thiên lệch, kết hợp hai kỹ thuật đề xuất trước đó.
- 4) Đề xuất chỉ số đánh đổi để phân tích quan hệ công bằng - hiệu suất.
Chỉ số này đã được diễn giải ý nghĩa qua công thức và sử dụng trong đánh giá thực nghiệm.

6. Nhận xét về ưu điểm và nhược điểm về nội dung, kết cấu và hình thức của luận án:

Về nội dung, luận án đã xác định được khoảng trống nghiên cứu, dựa trên khảo sát cho thấy phần lớn các nghiên cứu hiện nay tập trung vào một thuộc tính nhạy cảm riêng lẻ, trong khi thực tế thường tồn tại đồng thời nhiều thuộc tính nhạy cảm có tác động giao thoa. Từ đó, luận án đề xuất một hệ thống giải pháp tương đối hoàn chỉnh, theo một chuỗi logic: Fairedu xử lý phụ thuộc giữa thuộc tính nhạy cảm và đặc trưng dữ liệu; DPF xử lý mất cân bằng giữa các nhóm giao thoa; FairEduPlus kết hợp hai hướng tiếp cận trên. Luận án đánh giá trên nhiều bộ dữ liệu, nhiều mô hình học máy và nhiều độ đo công bằng khác nhau. Điều này giúp tăng độ tin cậy của kết quả thực nghiệm và chứng minh được khả năng áp dụng của các phương pháp đề xuất.

Nội dung luận án thiên về thực nghiệm, dựa trên mở rộng và kết hợp các kỹ thuật đơn giản đã có, cùng với cải tiến quy trình xử lý dữ liệu. Tính mới về mặt lý thuyết của các phương pháp còn chưa thật nổi bật. Về chỉ số đánh đổi đề xuất trong luận án, NCS có thể làm rõ hơn về tính chất toán học cũng như ưu điểm của chỉ số này so với các chỉ số đã có.

Về kết cấu, luận án có cấu trúc rõ ràng, hợp lí, bố cục cân đối giữa các chương. Ngoài phần giới thiệu, phần tổng quan và phần kết luận, mỗi chương đều được tổ chức thành 6 mục: Giới thiệu; Phương pháp; Thực nghiệm; Kết quả; Thảo luận; Tổng kết chương. Các chương trình bày kết quả nghiên cứu được xây dựng theo trình tự logic, mỗi chương giải quyết một vấn đề cụ thể và là tiền đề cho chương tiếp theo. NCS cũng nêu rõ các hạn chế trong từng nghiên cứu.

Luận án sẽ có chất lượng tốt hơn nếu phần tổng quan được trình bày kĩ hơn, bớt mang tính liệt kê mà tập trung vào phân tích, so sánh và tổng hợp các công trình này.

Về mặt trình bày, luận án được trình bày tương đối cẩn thận. Phần tổng quan có thể rút gọn một số nội dung, ví dụ như các khái niệm cơ bản ở 2.1.1. Phần diễn giải kết quả thực nghiệm ở các chương thiên về mô tả con số mà chưa phân tích sâu nguyên nhân dẫn tới sự khác biệt giữa các phương pháp.

Phần danh sách các từ viết tắt, các kí hiệu cần trình bày theo thứ tự ABC để tiện tra cứu.

Tài liệu tham khảo phong phú, tuy nhiên một số tài liệu còn thiếu thông tin xuất bản.

- Đánh giá về sự phù hợp của nội dung luận án với nội dung của các công trình đã được công bố trên tạp chí, tuyển tập công trình hội nghị khoa học; nhận xét về giá trị khoa học của các công trình đã công bố?

Luận án có 3 công trình đã công bố [Nga1-3], trong đó có 1 báo cáo hội nghị và 2 bài báo tạp chí Q1, NCS là tác giả đầu tiên của cả 3 công trình.

- Kết luận về sự đáp ứng các yêu cầu đối với luận án tiến sĩ hay không đáp ứng? Bản tóm tắt luận án có phản ánh trung thành nội dung cơ bản của luận án hay không? Luận án có thể đưa ra bảo vệ để nhận học vị tiến sĩ được hay không?

Luận án của NCS Phạm Thị Tô Nga đáp ứng các yêu cầu đối với một luận án tiến sĩ, có thể đưa ra bảo vệ để nhận học vị tiến sĩ. Bản tóm tắt cần chỉnh sửa để có thể tham khảo như tài liệu độc lập (ví dụ tóm tắt không có câu hỏi nghiên cứu rq, tuy nhiên trong các phần sau lại dùng tới kí hiệu rq này)

- Các ý kiến khác (nếu có).

Hà Nội, ngày tháng năm 2026

NGƯỜI NHẬN XÉT

Nguyễn Thị Minh Huyền

Hà Nội, ngày 25 tháng 6 năm 2026

**QUYẾT NGHỊ
CỦA HỘI ĐỒNG ĐÁNH GIÁ LUẬN ÁN TIẾN SĨ**

Căn cứ Quyết định số 612/QĐ-ĐHCN ngày 30/3/2026 của Hiệu trưởng Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội về việc thành lập Hội đồng đánh giá luận án tiến sĩ, Hội đồng đánh giá luận án tiến sĩ của nghiên cứu sinh Phạm Thị Tố Nga, sinh ngày 27/10/1987, đã họp vào hồi 14h00 ngày 25/6/2026 tại Trường Đại học Công nghệ.

Tên đề tài luận án: Một số phương pháp đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục

Ngành đào tạo: Khoa học máy tính

Mã số: 9480101

Sau khi nghe nghiên cứu sinh (NCS) trình bày kết quả phân biện độc lập luận án, trình bày tóm tắt luận án tiến sĩ; các phân biện đọc nhận xét; Thư ký Hội đồng đọc bản tổng hợp các nhận xét khác; NCS trả lời các câu hỏi của Hội đồng và đại biểu tham dự; đại diện tập thể hướng dẫn phát biểu ý kiến, Hội đồng đã họp, trao đổi ý kiến và thống nhất kết luận như sau:

1. Về ý nghĩa khoa học của luận án, các kết quả của luận án:

Luận án nghiên cứu vấn đề bảo đảm tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục. Đây là vấn đề có tính thời sự, có ý nghĩa khoa học và thực tiễn rõ ràng trong bối cảnh các hệ thống trí tuệ nhân tạo và học máy ngày càng được ứng dụng trong tuyển sinh, đánh giá kết quả học tập, dự đoán nguy cơ bỏ học, cá nhân hóa học tập và hỗ trợ ra quyết định trong quản lý giáo dục.

Về mặt khoa học, luận án tập trung vào bài toán công bằng giao thoa khi đồng thời tồn tại nhiều thuộc tính nhạy cảm, là hướng nghiên cứu quan trọng trong AI đáng tin cậy và AI có trách nhiệm. Các kết quả của luận án góp phần mở rộng cách tiếp cận bảo đảm công bằng trong học máy từ xử lý một thuộc tính nhạy cảm sang xử lý nhiều thuộc tính nhạy cảm và các nhóm giao thoa trong dữ liệu giáo dục.

Về mặt thực tiễn, các phương pháp đề xuất có tiềm năng hỗ trợ xây dựng các hệ thống giáo dục thông minh minh bạch hơn, hạn chế thiên lệch đối với các nhóm người học khác nhau, góp phần thúc đẩy bình đẳng về cơ hội học tập và nâng cao trách nhiệm xã hội trong ứng dụng AI vào giáo dục.

2. Về tính hiện đại, hợp lý và độ tin cậy của phương pháp nghiên cứu mà NCS đã sử dụng:

Luận án sử dụng phương pháp nghiên cứu phù hợp với một luận án tiến sĩ ngành Khoa học máy tính theo định hướng thực nghiệm và phát triển phương pháp: khảo sát tổng quan, xác định khoảng trống nghiên cứu, đề xuất phương pháp, thiết kế thực nghiệm, đối sánh với các phương pháp liên quan và phân tích kết quả.

Các phương pháp được đề xuất có tính hiện đại, bao gồm kỹ thuật tiền xử lý dựa trên hồi quy đa biến, phân nhóm dữ liệu theo các tổ hợp thuộc tính nhạy cảm, sinh dữ liệu tổng hợp bằng CTGAN và mô hình ngôn ngữ lớn, cùng với phân tích quan hệ đánh đổi giữa công bằng và hiệu suất mô hình.



Độ tin cậy của kết quả được củng cố thông qua thực nghiệm trên nhiều bộ dữ liệu giáo dục, nhiều mô hình học máy và nhiều thước đo đánh giá cả về tính công bằng và hiệu suất. Tuy nhiên, Hội đồng ghi nhận NCS cần tiếp tục làm rõ thêm một số khía cạnh như kiểm định thống kê cho các so sánh thực nghiệm, chất lượng và khả năng tái lập của dữ liệu sinh tổng hợp, cũng như độ ổn định của chỉ số đánh đổi trong một số trường hợp đặc biệt.

3. Về các kết quả mới của luận án. Giá trị của các kết quả này trong lĩnh vực khoa học chuyên ngành về mật lý thuyết và ứng dụng:

- Đề xuất được phương pháp Fairedu, một kỹ thuật tiên xử lý dựa trên hồi quy đa biến nhằm giảm sự phụ thuộc giữa các thuộc tính không nhạy cảm và đồng thời nhiều thuộc tính nhạy cảm trong dữ liệu huấn luyện, qua đó góp phần cải thiện tính công bằng của mô hình học máy trong bối cảnh dữ liệu giáo dục.
- Đề xuất được phương pháp DPF (Data Partitioning Fairness), trong đó dữ liệu được phân chia theo các nhóm giao thoa của thuộc tính nhạy cảm và sử dụng kỹ thuật sinh dữ liệu tổng hợp để cân bằng dữ liệu huấn luyện, giúp xử lý vấn đề mất cân bằng và thiếu đại diện của một số nhóm người học.
- Xây dựng được phương pháp FaireduPlus, kết hợp hai cơ chế xử lý: cân bằng phân bố giữa các nhóm giao thoa theo “chiều ngang” và giảm phụ thuộc giữa đặc trưng với thuộc tính nhạy cảm theo “chiều dọc”, hướng tới bảo đảm công bằng toàn diện hơn đối với dữ liệu dạng bảng trong giáo dục.

Các kết quả trên có giá trị khoa học trong việc mở rộng và hệ thống hóa hướng nghiên cứu bảo đảm công bằng cho các mô hình học máy với nhiều thuộc tính nhạy cảm. Về ứng dụng, các kết quả có thể tham khảo và phát triển cho các hệ thống hỗ trợ ra quyết định trong giáo dục như cảnh báo sớm nguy cơ bỏ học, hỗ trợ đánh giá người học, tuyển sinh và quản lý đào tạo. Các kết quả nghiên cứu của NCS đã được công bố trong các công trình khoa học có chất lượng, phù hợp với nội dung của luận án.

4. Những thiếu sót về nội dung và hình thức của luận án:

Bên cạnh các kết quả đạt được, Hội đồng thống nhất ghi nhận một số điểm NCS cần tiếp thu, chỉnh sửa và hoàn thiện trong bản luận án cuối cùng:

- Cần phân tích sâu hơn các kết quả thực nghiệm, nhất là các trường hợp phương pháp đề xuất chưa vượt trội so với phương pháp đối sánh hoặc phụ thuộc mạnh vào kỹ thuật sinh dữ liệu được sử dụng; nên bổ sung kiểm định thống kê, khoảng tin cậy hoặc độ lớn ảnh hưởng khi so sánh các phương pháp.
- Đối với dữ liệu sinh tổng hợp bằng CTGAN và LLM, cần mô tả rõ hơn lý do lựa chọn mô hình sinh, quy trình huấn luyện/sinh dữ liệu, thiết kế prompt, cách đánh giá chất lượng dữ liệu sinh, tính tái lập, cũng như các vấn đề liên quan đến fidelity, privacy và khả năng rò rỉ thông tin.
- Đối với chỉ số đánh đổi công bằng - hiệu suất, cần làm rõ hơn tính chất toán học, độ ổn định số học và cách xử lý các giá trị ngoại lai hoặc các trường hợp mẫu số nhỏ.
- Cần rà soát và chỉnh sửa việc hình thức/trình bày của luận án theo yêu cầu của Hội đồng.
- Lược bỏ các nội dung trùng lặp, kiến thức cơ bản trong luận án để đảm bảo tính nhất quán.



5. Kết quả kiểm phiếu

- Số phiếu hợp lệ: 7 phiếu
- Số phiếu không hợp lệ: 0 phiếu
- Số phiếu đánh giá ĐẠT: 7 phiếu
- Số phiếu đánh giá KHÔNG ĐẠT: 0 phiếu

6. Kết luận của Hội đồng:

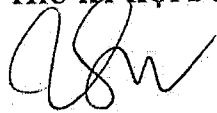
- Luận án đáp ứng yêu cầu đối với một luận án tiến sĩ theo quy định của Đại học Quốc gia Hà Nội. Bản tóm tắt luận án về cơ bản phản ánh trung thành nội dung, cấu trúc và các kết quả nghiên cứu chính của luận án.
- Đề nghị NCS nghiêm túc tiếp thu các ý kiến góp ý của Hội đồng, hoàn thiện luận án và các hồ sơ liên quan theo quy định trước khi thực hiện các thủ tục tiếp theo.
- Trên cơ sở kết quả kiểm phiếu, Hội đồng thống nhất đề nghị công nhận kết quả đánh giá luận án tiến sĩ cho nghiên cứu sinh Phạm Thị Tố Nga theo quy định hiện hành.

7. Đề nghị khen thưởng (nếu luận án đặc biệt xuất sắc)

Không đề nghị khen thưởng.

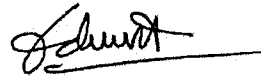
Quyết nghị này được 07/07 thành viên của Hội đồng nhất trí thông qua./.

THƯ KÝ HỘI ĐỒNG



TS. Nguyễn Văn Sơn

CHỦ TỊCH HỘI ĐỒNG



PGS. TS. Võ Đình Hiếu

**XÁC NHẬN CỦA ĐƠN VỊ ĐÀO TẠO
TL. HIỆU TRƯỞNG
KT. TRƯỞNG PHÒNG ĐÀO TẠO
PHÓ TRƯỞNG PHÒNG**



Nguyễn Huy Tiệp



Hà Nội, ngày 10 tháng 07 năm 2026

BẢN XÁC NHẬN ĐÃ SỬA CHỮA CÁC THIẾU SÓT CỦA LUẬN ÁN

Ông/bà Phạm Thị Tố Nga được công nhận là nghiên cứu sinh (NCS) theo Quyết định số 25/QĐ-CTSV ngày 14 / 01 /2022 của Hiệu trưởng Trường Đại học Công nghệ. Thực hiện Quyết định số 612/QĐ-ĐHCN ngày 30/03/2026 của Hiệu trưởng trường Đại học Công nghệ về việc thành lập Hội đồng đánh giá luận án tiến sĩ của NCS Phạm Thị Tố Nga. Hội đồng đã họp ngày 25/06 /2026, (có biên bản kèm theo). Theo Quyết nghị của Hội đồng, NCS phải bổ sung và sửa chữa các điểm sau đây trong luận án:

1. Cần phân tích sâu hơn các kết quả thực nghiệm, nhất là các trường hợp phương pháp đề xuất chưa vượt trội so với phương pháp đối sánh hoặc phụ thuộc mạnh vào kỹ thuật sinh dữ liệu được sử dụng; nên bổ sung kiểm định thống kê, khoảng tin cậy hoặc độ lớn ảnh hưởng khi so sánh các phương pháp.
2. Đối với dữ liệu sinh tổng hợp bằng CTGAN và LLM, cần mô tả rõ hơn lý do lựa chọn mô hình sinh, quy trình huấn luyện/sinh dữ liệu, thiết kế prompt, cách đánh giá chất lượng dữ liệu sinh, tính tái lập, cũng như các vấn đề liên quan đến fidelity, privacy và khả năng rò rỉ thông tin.
3. Đối với chỉ số đánh đổi công bằng - hiệu suất, cần làm rõ hơn tính chất toán học, độ ổn định số học và cách xử lý các giá trị ngoại lai hoặc các trường hợp mẫu số nhỏ.
4. Cần rà soát và chỉnh sửa việc hình thức/trình bày của luận án theo yêu cầu của Hội đồng.
5. Lược bỏ các nội dung trùng lặp, kiến thức cơ bản trong luận án để đảm bảo tính nhất quán.

Ngày 07/07/2026, NCS Phạm Thị Tố Nga đã nộp bản luận án và tóm tắt luận án có chỉnh sửa. Chúng tôi nhận thấy rằng nội dung, hình thức của luận án và tóm tắt luận án đã được sửa chữa, bổ sung theo các điểm trên của Quyết nghị (Bản giải trình sửa chữa của NCS kèm theo).

Xin trân trọng cảm ơn./.

THƯ KÝ HỘI ĐỒNG



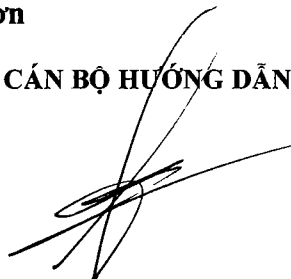
Nguyễn Văn Sơn

NGHIÊN CỨU SINH



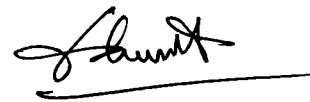
Phạm Thị Tố Nga

CÁN BỘ HƯỚNG DẪN



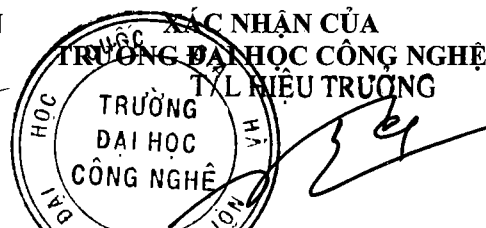
Phạm Ngọc Hùng

CHỦ TỊCH HỘI ĐỒNG



Võ Đình Hiếu

XÁC NHẬN CỦA
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ
TYL HIỆU TRƯỞNG
TRƯỜNG
ĐẠI HỌC
CÔNG NGHỆ
HÀ NỘI



KT. TRƯỞNG PHÒNG CÔNG TÁC SINH VIÊN

PHÓ TRƯỞNG PHÒNG

Nguyễn Huy Tiệp

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Hà Nội, ngày 06 tháng 07 năm 2026

**BẢN GIẢI TRÌNH
CỦA NGHIÊN CỨU SINH VỀ VIỆC SỬA CHỮA, BỔ SUNG LUẬN ÁN**

Họ và tên NCS: Phạm Thị Tố Nga

Tên đề tài luận án: Một số phương pháp đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục.

Chuyên ngành: Khoa học máy tính

Mã số: 9480101.01

Sau khi họp Hội đồng đánh giá luận án tiến sĩ, Nghiên cứu sinh đã nghiêm túc thực hiện việc nghiên cứu, tiếp thu, sửa chữa, bổ sung luận án theo Quyết nghị của Hội đồng và có các giải trình như sau:

STT	Ý kiến của Hội đồng	Kết quả chỉnh sửa	Vị trí chỉnh sửa trong luận án
Đề nghị chỉnh sửa của Hội đồng			
1	Cần phân tích sâu hơn các kết quả thực nghiệm, nhất là các trường hợp phương pháp đề xuất chưa vượt trội so với phương pháp đối sánh hoặc phụ thuộc mạnh vào kỹ thuật sinh dữ liệu được sử dụng; nên bổ sung kiểm định thống kê, khoảng tin cậy hoặc độ lớn ảnh hưởng khi so sánh các phương pháp.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung và làm rõ phần phân tích kết quả thực nghiệm theo hướng không chỉ mô tả kết quả mà còn phân tích nguyên nhân của các trường hợp phương pháp chưa vượt trội, làm rõ ảnh hưởng của đặc điểm dữ liệu, mức độ mất cân bằng, số lượng thuộc tính nhạy cảm và kỹ thuật sinh dữ liệu (CTGAN/LLM) đến hiệu quả của từng phương pháp. Đồng thời, luận án đã bổ sung phần thảo luận về các hạn chế của cách đánh giá hiện tại và đưa vào phần hướng nghiên cứu tiếp theo định hướng kết hợp các kiểm định thống kê, khoảng tin cậy và các chỉ số kích thước ảnh hưởng (effect size) nhằm nâng cao độ tin cậy của việc so sánh các phương pháp.	Các mục kết quả thực nghiệm và thảo luận trong Chương 2 (trang 56-69, Chương 3 (trang 91-104) và Chương 4 (trang 122-133). Phần Kết luận (trang 137-141)
2	Đối với dữ liệu sinh tổng hợp bằng CTGAN và LLM, cần mô tả rõ hơn lý do lựa chọn mô hình sinh, quy trình huấn luyện/sinh dữ liệu,	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung chi tiết quy trình sinh dữ liệu bằng CTGAN và LLM, bao gồm lý do lựa chọn mô hình, kiến trúc tổng thể, các tham số thực nghiệm, quy trình sinh dữ liệu và thiết kế prompt. Đồng thời, luận án bổ sung phần	Chương 3 mới, Mục Kiến trúc tổng thể (trang 77-81); Mục Hạn chế của nghiên cứu (trang 105-106);

	thiết kế prompt, cách đánh giá chất lượng dữ liệu sinh, tính tái lập, cũng như các vấn đề liên quan đến fidelity, privacy và khả năng rò rỉ thông tin.	thảo luận về chất lượng dữ liệu tổng hợp, khả năng tái lập, các vấn đề liên quan đến fidelity, privacy leakage và những hạn chế của phương pháp; đồng thời xác định đây là hướng nghiên cứu cần tiếp tục hoàn thiện trong tương lai.	Phân Kết luận và hướng nghiên cứu tiếp theo (trang 137-142)
3	Đối với chỉ số đánh giá công bằng - hiệu suất, cần làm rõ hơn tính chất toán học, độ ổn định số học và cách xử lý các giá trị ngoại lai hoặc các trường hợp mẫu số nhỏ.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần thảo luận làm rõ phạm vi áp dụng và những hạn chế của chỉ số đánh giá, trong đó nêu rõ chỉ số mới dừng ở mức đề xuất và thử nghiệm ban đầu, có thể kém ổn định trong một số trường hợp do sử dụng mức cải thiện tương đối. Đồng thời, luận án đã bổ sung định hướng nghiên cứu tiếp theo về việc chuẩn hóa chỉ số, xử lý giá trị ngoại lai và xây dựng các công thức đánh giá ổn định hơn.	Mục 4.3.3. Chỉ số đánh giá (trang 120) Mục Hạn chế của nghiên cứu (trang 132-133); Phân Kết luận và hướng nghiên cứu tiếp theo (trang 137-142)
4	Cần rà soát và chỉnh sửa việc hình thức/trình bày của luận án theo yêu cầu của Hội đồng. - Bỏ phần 2.1.1 cũ (Các khái niệm về học máy, AI). Đổi tiêu đề	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và chỉnh sửa toàn bộ luận án theo hướng cô đọng, thống nhất và đúng quy cách trình bày. Cụ thể, NCS đã lược bỏ các mục không cần thiết (Mục 2.1.1 cũ), rút gọn tiêu đề các chương và mục, bổ sung giải thích các ký hiệu trong công thức, thống nhất thuật ngữ và cách diễn đạt, sắp xếp danh mục từ viết tắt theo thứ tự ABC, chuẩn hóa và cập nhật danh mục tài liệu tham khảo, đồng thời rà soát, chỉnh sửa các lỗi hành văn và chính tả trong toàn bộ luận án.	Toàn bộ luận án
5	Lược bỏ các nội dung trùng lặp, kiến thức cơ bản trong luận án để đảm bảo tính nhất quán.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát, tái cấu trúc và rút gọn nội dung luận án theo hướng tập trung vào các đóng góp khoa học. Cụ thể, Chương Tổng quan được viết lại theo hướng cô đọng hơn, lược bỏ các kiến thức cơ sở và các nội dung trùng lặp, đồng thời tăng cường phân tích các nghiên cứu liên quan và khoảng trống nghiên cứu. Cấu trúc luận án cũng được điều chỉnh theo hướng tinh gọn, chuyển Chương 1 thành phần Giới thiệu và Chương 6 thành phần Kết luận, bảo đảm	Phân Giới thiệu (trang 1-13); Chương 1 (mới) (trang 14-40); Phân Kết luận (trang 137-142); Toàn bộ luận án.

		tính logic và nhất quán của toàn bộ luận án.	
1. PGS. TS Võ Đình Hiếu – Chủ tịch Hội đồng			
1	Một số nội dung bị lặp (định nghĩa công bằng, thách thức trong nghiên cứu)	NCS xin trân trọng cảm ơn ý kiến góp ý của Thầy. NCS đã rà soát toàn bộ luận án, lược bỏ các nội dung trình bày trùng lặp về khái niệm công bằng và các thách thức trong nghiên cứu. Các khái niệm nền tảng được trình bày đầy đủ tại Chương 2; ở các chương sau chỉ giữ lại những nội dung cần thiết và bổ sung dẫn chiếu đến các mục đã trình bày trước nhằm tránh lặp lại, đồng thời bảo đảm tính mạch lạc của luận án.	Chương 1. Mục 1.4. Tổng quan nghiên cứu về công bằng cho các hệ thống học máy ứng dụng trong lĩnh vực giáo dục (trang 34, 35)
2. TS Nguyễn Văn Sơn – Ủy viên Thư kí			
1	Nên rút gọn một số tiêu đề chương/mục và các phần giải thích khái niệm quá cơ bản để luận án cô đọng hơn.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát toàn bộ luận án, rút gọn một số tiêu đề chương, mục theo hướng ngắn gọn và phản ánh đúng nội dung trình bày. Đồng thời, NCS lược bỏ hoặc tinh giản các phần giải thích những khái niệm quá cơ bản, chỉ giữ lại các nội dung cần thiết phục vụ trực tiếp cho nghiên cứu nhằm nâng cao tính cô đọng và mạch lạc của luận án.	Đổi tên Chương 1 và Chương 6 thành Phần Giới thiệu và Kết luận: Trang 1 và 137. Đổi tên các mục trong Chương 2 cũ (Giờ là Chương 1): Trang 14-40
2	Cần thống nhất ký hiệu trong các thuật toán, chú thích đầy đủ các thành phần trong công thức, rà soát lỗi chính tả, trình bày danh mục từ viết tắt/ký hiệu thuận tiện hơn cho tra cứu.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và thống nhất hệ thống ký hiệu sử dụng trong toàn bộ các thuật toán và công thức của luận án; bổ sung chú thích đầy đủ cho các ký hiệu và tham số; đồng thời kiểm tra, chỉnh sửa các lỗi chính tả và lỗi trình bày. Danh mục từ viết tắt và ký hiệu cũng được sắp xếp, bổ sung nhằm giúp người đọc thuận tiện hơn trong quá trình tra cứu.	Mục chú ý (trang 20) Bảng 1.1: Các định nghĩa về công bằng trong hệ thống học máy, Trang 21
3	Phần tổng quan nên tăng tính phân tích, so sánh và tổng hợp, thay vì chủ yếu mô tả hoặc liệt kê các công trình liên quan.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và chỉnh sửa Chương Tổng quan theo hướng tăng cường phân tích, so sánh và tổng hợp các công trình nghiên cứu, làm rõ ưu điểm, hạn chế và khoảng trống nghiên cứu của từng nhóm phương pháp. Trên cơ sở đó, NCS nhấn mạnh hơn cơ sở khoa học và tính cần thiết của các phương pháp được đề xuất trong luận án.	Viết lại phần tổng quan trang 34, 35

4	Cần bổ sung hoặc làm rõ hơn kiểm định thống kê, khoảng tin cậy hoặc effect size khi so sánh các phương pháp, tránh chỉ dựa vào thống kê thắng/thua.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung vào phần hạn chế của luận án nội dung về việc các kết quả thực nghiệm hiện chủ yếu được đánh giá dựa trên thống kê thắng/thua, chưa kết hợp các kiểm định thống kê và các chỉ số đánh giá bổ sung để đánh giá mức độ khác biệt giữa các phương pháp. Đồng thời, trong phần hướng nghiên cứu tiếp theo, NCS đã bổ sung định hướng nghiên cứu hoàn thiện phương pháp đánh giá thông qua việc kết hợp các kiểm định thống kê, khoảng tin cậy và các chỉ số kích thước ảnh hưởng (effect size), nhằm nâng cao độ tin cậy và tính toàn diện trong việc so sánh hiệu quả của các phương pháp đảm bảo công bằng.	Phần Kết luận, Mục Hạn chế và hướng nghiên cứu trong tương lai (trang 139-141)
5	Cần thảo luận sâu hơn các trường hợp phương pháp đề xuất không ổn định hoặc thua phương pháp đối sánh với biên độ đáng kể; đặc biệt là sự phụ thuộc của DPF vào kỹ thuật sinh dữ liệu.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần thảo luận trong mục hạn chế của phương pháp DPF và FaireduPlus, làm rõ rằng hiệu quả của các phương pháp có thể bị ảnh hưởng bởi mức độ mất cân bằng của dữ liệu, số lượng mẫu trong các nhóm giao thoa và chất lượng của dữ liệu tổng hợp được sinh ra. Đồng thời, NCS cũng bổ sung nội dung về sự phụ thuộc của DPF và FaireduPlus vào kỹ thuật sinh dữ liệu (CTGAN và LLM), cũng như đề xuất hướng nghiên cứu tiếp theo nhằm phát triển các kỹ thuật sinh dữ liệu ổn định và có khả năng bảo toàn tốt hơn phân phối dữ liệu, từ đó nâng cao tính ổn định và khả năng tổng quát của các phương pháp đề xuất.	Chương 3. Phần hạn chế của DPF (trang 105); Phần hạn chế của FaireduPlus (trang 132-133); Phần hướng nghiên cứu tiếp theo (trang 133)
6	Đối với chỉ số đánh đổi, nên làm rõ tính chất toán học, độ ổn định số học và cách xử lý các giá trị ngoại lai hoặc trường hợp mẫu số nhỏ.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần giải thích về cơ sở xây dựng, ý nghĩa và cách diễn giải của chỉ số đánh đổi tổng hợp, đồng thời làm rõ đặc tính của chỉ số khi được xây dựng từ hai chỉ số thay đổi tương đối về công bằng và hiệu suất. Bên cạnh đó, trong phần đánh giá kết quả, NCS đã bổ sung thảo luận đối với các trường hợp chỉ số đánh đổi có giá trị biến động lớn và làm rõ rằng chỉ số này được sử dụng như một thước đo hỗ trợ, cần	Chương 4. Mục chỉ số đánh đổi: Trang 115; Mục 4.4.3. Đánh giá mối quan hệ đánh đổi giữa công bằng và hiệu suất: trang 128; Mục Thảo luận: (trang 131-133)

		tính logic và nhất quán của toàn bộ luận án.	
1. PGS. TS Võ Đình Hiếu – Chủ tịch Hội đồng			
1	Một số nội dung bị lặp (định nghĩa công bằng, thách thức trong nghiên cứu)	NCS xin trân trọng cảm ơn ý kiến góp ý của Thầy. NCS đã rà soát toàn bộ luận án, lược bỏ các nội dung trình bày trùng lặp về khái niệm công bằng và các thách thức trong nghiên cứu. Các khái niệm nền tảng được trình bày đầy đủ tại Chương 2; ở các chương sau chỉ giữ lại những nội dung cần thiết và bổ sung dẫn chiếu đến các mục đã trình bày trước nhằm tránh lặp lại, đồng thời bảo đảm tính mạch lạc của luận án.	Chương 1. Mục 1.4. Tổng quan nghiên cứu về công bằng cho các hệ thống học máy ứng dụng trong lĩnh vực giáo dục (trang 34, 35)
2. TS Nguyễn Văn Sơn – Ủy viên Thư kí			
1	Nên rút gọn một số tiêu đề chương/mục và các phần giải thích khái niệm quá cơ bản để luận án cô đọng hơn.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát toàn bộ luận án, rút gọn một số tiêu đề chương, mục theo hướng ngắn gọn và phản ánh đúng nội dung trình bày. Đồng thời, NCS lược bỏ hoặc tinh giản các phần giải thích những khái niệm quá cơ bản, chỉ giữ lại các nội dung cần thiết phục vụ trực tiếp cho nghiên cứu nhằm nâng cao tính cô đọng và mạch lạc của luận án.	Đổi tên Chương 1 và Chương 6 thành Phần Giới thiệu và Kết luận: Trang 1 và 137. Đổi tên các mục trong Chương 2 cũ (Giờ là Chương 1): Trang 14-40
2	Cần thống nhất ký hiệu trong các thuật toán, chú thích đầy đủ các thành phần trong công thức, rà soát lỗi chính tả, trình bày danh mục từ viết tắt/ký hiệu thuận tiện hơn cho tra cứu.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và thống nhất hệ thống ký hiệu sử dụng trong toàn bộ các thuật toán và công thức của luận án; bổ sung chú thích đầy đủ cho các ký hiệu và tham số; đồng thời kiểm tra, chỉnh sửa các lỗi chính tả và lỗi trình bày. Danh mục từ viết tắt và ký hiệu cũng được sắp xếp, bổ sung nhằm giúp người đọc thuận tiện hơn trong quá trình tra cứu.	Mục chú ý (trang 20) Bảng 1.1: Các định nghĩa về công bằng trong hệ thống học máy, Trang 21
3	Phần tổng quan nên tăng tính phân tích, so sánh và tổng hợp, thay vì chủ yếu mô tả hoặc liệt kê các công trình liên quan.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và chỉnh sửa Chương Tổng quan theo hướng tăng cường phân tích, so sánh và tổng hợp các công trình nghiên cứu, làm rõ ưu điểm, hạn chế và khoảng trống nghiên cứu của từng nhóm phương pháp. Trên cơ sở đó, NCS nhấn mạnh hơn cơ sở khoa học và tính cần thiết của các phương pháp được đề xuất trong luận án.	Viết lại phần tổng quan trang 34, 35

4	Cần bổ sung hoặc làm rõ hơn kiểm định thống kê, khoảng tin cậy hoặc effect size khi so sánh các phương pháp, tránh chỉ dựa vào thống kê thắng/thua.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung vào phần hạn chế của luận án nội dung về việc các kết quả thực nghiệm hiện chủ yếu được đánh giá dựa trên thống kê thắng/thua, chưa kết hợp các kiểm định thống kê và các chỉ số đánh giá bổ sung để đánh giá mức độ khác biệt giữa các phương pháp. Đồng thời, trong phần hướng nghiên cứu tiếp theo, NCS đã bổ sung định hướng nghiên cứu hoàn thiện phương pháp đánh giá thông qua việc kết hợp các kiểm định thống kê, khoảng tin cậy và các chỉ số kích thước ảnh hưởng (effect size), nhằm nâng cao độ tin cậy và tính toàn diện trong việc so sánh hiệu quả của các phương pháp đảm bảo công bằng.	Phân Kết luận, Mục Hạn chế và hướng nghiên cứu trong tương lai (trang 139-141)
5	Cần thảo luận sâu hơn các trường hợp phương pháp đề xuất không ổn định hoặc thua phương pháp đối sánh với biên độ đáng kể; đặc biệt là sự phụ thuộc của DPF vào kỹ thuật sinh dữ liệu.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần thảo luận trong mục hạn chế của phương pháp DPF và FaireduPlus, làm rõ rằng hiệu quả của các phương pháp có thể bị ảnh hưởng bởi mức độ mất cân bằng của dữ liệu, số lượng mẫu trong các nhóm giao thoa và chất lượng của dữ liệu tổng hợp được sinh ra. Đồng thời, NCS cũng bổ sung nội dung về sự phụ thuộc của DPF và FaireduPlus vào kỹ thuật sinh dữ liệu (CTGAN và LLM), cũng như đề xuất hướng nghiên cứu tiếp theo nhằm phát triển các kỹ thuật sinh dữ liệu ổn định và có khả năng bảo toàn tốt hơn phân phối dữ liệu, từ đó nâng cao tính ổn định và khả năng tổng quát của các phương pháp đề xuất.	Chương 3. Phần hạn chế của DPF (trang 105); Phần hạn chế của FaireduPlus (trang 132-133); Phần hướng nghiên cứu tiếp theo (trang 133)
6	Đối với chỉ số đánh đổi, nên làm rõ tính chất toán học, độ ổn định số học và cách xử lý các giá trị ngoại lai hoặc trường hợp mẫu số nhỏ.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần giải thích về cơ sở xây dựng, ý nghĩa và cách diễn giải của chỉ số đánh đổi tổng hợp, đồng thời làm rõ đặc tính của chỉ số khi được xây dựng từ hai chỉ số thay đổi tương đối về công bằng và hiệu suất. Bên cạnh đó, trong phần đánh giá kết quả, NCS đã bổ sung thảo luận đối với các trường hợp chỉ số đánh đổi có giá trị biến động lớn và làm rõ rằng chỉ số này được sử dụng như một thước đo hỗ trợ, cần	Chương 4. Mục chỉ số đánh đổi: Trang 115; Mục 4.4.3. Đánh giá mối quan hệ đánh đổi giữa công bằng và hiệu suất: trang 128; Mục Thảo luận: (trang 131-133)

		được diễn giải kết hợp với các chỉ số công bằng và hiệu suất gốc để đánh giá đầy đủ hiệu quả của phương pháp. Ngoài ra, NCS cũng bổ sung phần thảo luận về phạm vi áp dụng, những hạn chế của chỉ số và định hướng nghiên cứu tiếp theo nhằm tiếp tục hoàn thiện và kiểm chứng thước đo này trên nhiều bộ dữ liệu và kịch bản thực nghiệm khác nhau.	
7	Cần mô tả chi tiết hơn quy trình sinh dữ liệu bằng LLM/CTGAN, thiết kế prompt, đánh giá chất lượng data sinh, vấn đề tái lập và nguy cơ rò rỉ thông tin, thảo luận về việc áp dụng giải pháp sang các lĩnh vực khác ngoài giáo dục.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung mô tả chi tiết hơn quy trình sinh dữ liệu bằng CTGAN và LLM, bao gồm quy trình thực hiện, tham số thực nghiệm và thiết kế prompt. Đồng thời, NCS bổ sung phần thảo luận về chất lượng dữ liệu tổng hợp, khả năng tái lập kết quả, nguy cơ rò rỉ thông tin khi sử dụng dữ liệu tổng hợp và các lưu ý khi áp dụng các mô hình sinh dữ liệu. Ngoài ra, luận án cũng mở rộng phần thảo luận về khả năng áp dụng phương pháp đề xuất sang các lĩnh vực khác như tài chính, y tế và tuyển dụng, đồng thời chỉ ra những điều chỉnh cần thiết khi triển khai trên các miền dữ liệu khác nhau.	Chương 3. Mục 3.2.2 Kiến trúc tổng thể (trang 77-81) Bổ sung Mục 3.5.2. Thảo luận về dữ liệu tổng hợp và khả năng áp dụng: Trang 104
3. PGS.TS Phạm Văn Cường – Ủy viên Phản biện			
1	Nhiều đoạn trong luận án được trình bày quá dài (ví dụ trang 13).	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. Trên cơ sở góp ý của Hội đồng, NCS đã rà soát và chỉnh sửa toàn bộ luận án, đặc biệt tại các vị trí có đoạn văn dài, bằng cách chia tách thành các đoạn ngắn hơn, lược bỏ các nội dung diễn giải chưa cần thiết và sắp xếp lại ý theo hướng logic hơn. Nhờ đó, nội dung luận án trở nên cô đọng, dễ theo dõi và nâng cao tính mạch lạc trong trình bày.	Phần Giới thiệu. Mục 0.4. Bố cục của luận án (trang 13) và các vị trí liên quan trong toàn bộ luận án.
2	- Mục 2.3.2 nên khảo sát các mô hình phổ biến được dùng trong sinh dữ liệu (không chỉ có GAN và LLM) và phân tích rõ hơn các ưu, nhược điểm	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã mở rộng Mục 1.3.2 bằng cách bổ sung tổng quan về các nhóm phương pháp sinh dữ liệu tổng hợp phổ biến, bao gồm các phương pháp thông kê, mô hình tự mã hóa biến phân (VAE), mô hình khuếch tán (Diffusion Models), bên cạnh các mô hình GAN và LLM. Đồng thời, luận án đã bổ sung	Chương 1. Mục 1.3.2. Kỹ thuật sinh dữ liệu tổng hợp (trang 31-32)

	của mỗi mô hình sinh.	phân tích ưu điểm, hạn chế và phạm vi áp dụng của từng nhóm phương pháp. Trên cơ sở đó, NCS làm rõ hơn lý do lựa chọn CTGAN và LLM làm hai kỹ thuật sinh dữ liệu được sử dụng trong nghiên cứu.	
3	- Giải thích rõ các thành phần và mối liên hệ, các luồng xử lý giữa các khối kiến trúc FAIREDU (hình 3.1 trang 53)	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung và chỉnh sửa phần mô tả kiến trúc tổng thể của phương pháp Fairedu theo hướng làm rõ hai pha xử lý (pha huấn luyện và pha kiểm thử), chức năng của từng khối trong kiến trúc, mối liên hệ và luồng xử lý dữ liệu giữa các thành phần. Đồng thời, NCS bổ sung phần giải thích về việc sử dụng bộ hệ số hồi quy được ước lượng trên tập huấn luyện để xử lý tập kiểm thử nhằm đảm bảo tính nhất quán của phương pháp và tránh rò rỉ thông tin trong quá trình đánh giá mô hình.	Chương 2. Mục 2.2.2. Kiến trúc tổng thể (trang 44-46).
4	- Phân tích từng bước và phân tích độ phức tạp của thuật toán 3.1 trang 54	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần diễn giải Thuật toán 3.1 theo từng bước thực hiện nhằm làm rõ quy trình xử lý dữ liệu, từ giai đoạn ước lượng các hệ số hồi quy, điều chỉnh dữ liệu đến huấn luyện và dự đoán. Đồng thời, luận án đã bổ sung phân tích độ phức tạp tính toán của từng giai đoạn và độ phức tạp tổng thể của thuật toán, đồng thời so sánh với phương pháp LTDD để làm rõ hơn hiệu quả tính toán và khả năng áp dụng của Fairedu trong trường hợp có nhiều thuộc tính nhạy cảm.	Chương 2. Mục 2.2.3. Thuật toán (trang 47-50)
5	- Cần mô tả về bộ dữ liệu DNU: quá trình thu thập, ý nghĩa từng thuộc tính, tính đa dạng, tính bảo đảm công bằng dữ liệu ...	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung mô tả chi tiết hơn về bộ dữ liệu DNU, bao gồm nguồn gốc và quy trình thu thập dữ liệu từ hệ thống quản lý đào tạo, quá trình chuẩn hóa và ẩn danh dữ liệu nhằm bảo đảm quyền riêng tư của người học. Bên cạnh đó, luận án làm rõ ý nghĩa của các nhóm thuộc tính, đặc điểm đa dạng của bộ dữ liệu thông qua việc thu thập từ nhiều khóa học và chương trình đào tạo khác nhau, đồng thời giải thích vai trò của các thuộc tính nhạy cảm trong việc	Chương 2. Mục 2.3.1. Dữ liệu (trang 52-53)

		đánh giá và cải thiện tính công bằng của mô hình học máy.	
6	- Bổ sung các luận giải chi tiết hơn các kết quả của bảng 3.3 trang 67.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần phân tích và thảo luận chi tiết hơn đối với các kết quả trong Bảng 3.3 cũ (hiện nay được cập nhật là Bảng 2.3), làm rõ xu hướng cải thiện của Fairedu trên từng chỉ số công bằng và từng bộ dữ liệu, đồng thời phân tích nguyên nhân của các trường hợp phương pháp đạt hoặc chưa đạt kết quả vượt trội so với các phương pháp đối sánh. Bên cạnh đó, luận án cũng bổ sung phần luận giải về ý nghĩa của các kết quả thực nghiệm, qua đó khẳng định hiệu quả của Fairedu trong việc xử lý đồng thời nhiều thuộc tính nhạy cảm và cải thiện tính công bằng của mô hình học máy.	Chương 2. Mục 2.4.3. Khả năng xử lý đồng thời nhiều thuộc tính nhạy cảm của FairEdu (trang 60-62)
7	- Phân tích rõ và so sánh mức độ nhạy cảm giữa hiện tượng đa cộng tuyến vs. mô hình hóa quan hệ phi tuyến?	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần thảo luận về ảnh hưởng của hiện tượng đa cộng tuyến và các quan hệ phi tuyến đối với phương pháp Fairedu trong phần hạn chế của nghiên cứu, làm rõ tác động của từng yếu tố đến khả năng áp dụng của mô hình hồi quy tuyến tính đa biến. Đồng thời, luận án cũng bổ sung định hướng nghiên cứu tiếp theo theo hướng mở rộng Fairedu bằng các mô hình phi tuyến và nghiên cứu các kỹ thuật xử lý đa cộng tuyến nhằm nâng cao tính ổn định và khả năng áp dụng của phương pháp trên các bộ dữ liệu có cấu trúc phức tạp.	Chương 2. Mục 2.5.2. Hạn chế của nghiên cứu (trang 67-69) Phần Kết luận, hướng nghiên cứu trong tương lai: 1. Mở rộng phạm vi dữ liệu và mô hình áp dụng (Trang 140)
8	- Luận giải chi tiết lý do lựa chọn các mô hình CTGAN và LLM để sinh dữ liệu; mô tả quá trình huấn luyện và quá trình sinh cũng như đánh giá chất lượng từ dữ liệu sinh bởi các mô hình này.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần luận giải về cơ sở lựa chọn CTGAN và LLM làm hai kỹ thuật sinh dữ liệu tổng hợp trong nghiên cứu, làm rõ ưu điểm, hạn chế và sự phù hợp của từng mô hình đối với bài toán sinh dữ liệu bằng nhằm cải thiện công bằng. Đồng thời, luận án bổ sung mô tả chi tiết quy trình huấn luyện và sinh dữ liệu của CTGAN và LLM, bao gồm các tham số thực nghiệm, thiết kế prompt và quy trình hậu xử lý dữ liệu.	Chương 3. Mục 3.2.2 Kiến trúc tổng thể (Trang 77-81) Bổ sung Mục 3.5.2. Thảo luận về dữ liệu tổng hợp và khả năng áp dụng: Trang 101

		Bên cạnh đó, NCS bổ sung phần thảo luận về chất lượng dữ liệu tổng hợp, khả năng tái lập kết quả, nguy cơ rò rỉ thông tin và các lưu ý khi sử dụng dữ liệu tổng hợp. Ngoài ra, luận án cũng mở rộng phần thảo luận về khả năng áp dụng phương pháp đề xuất sang các lĩnh vực khác như tài chính, y tế và tuyển dụng, đồng thời chỉ ra những điều chỉnh cần thiết khi triển khai trên các miền dữ liệu khác nhau.	
9	- Bổ sung các luận giải chi tiết hơn các kết quả của bảng 4.7, 4.8 trang 99 và 100.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần phân tích và thảo luận chi tiết hơn đối với các kết quả trong Bảng 4.7 và Bảng 4.8, làm rõ xu hướng cải thiện về công bằng và hiệu suất của phương pháp đề xuất theo từng mô hình học máy và từng phương pháp sinh dữ liệu. Đồng thời, luận án bổ sung các luận giải về nguyên nhân của những khác biệt giữa CTGAN và LLM, ý nghĩa của các kết quả đạt được cũng như phân tích các trường hợp phương pháp chưa đạt kết quả vượt trội do ảnh hưởng của đặc điểm dữ liệu và chất lượng dữ liệu tổng hợp. Qua đó, làm rõ hơn ưu điểm, hạn chế và phạm vi áp dụng của phương pháp đề xuất.	Chương 3. Mục 3.4.2.1. Đánh giá thông qua các chỉ số công bằng (trang 95-96) 3.4.2.2. Ảnh hưởng của mô hình học máy (trang 96-97)
10	- Mô tả chi tiết kiến trúc tổng thể của phương pháp DPF (hình 4.1 trang 93)	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung mô tả chi tiết kiến trúc tổng thể của phương pháp DPF, làm rõ hai pha huấn luyện và kiểm thử, chức năng của từng thành phần, luồng xử lý dữ liệu từ phân chia dữ liệu, phân vùng theo nhóm giao thoa, sinh dữ liệu tổng hợp, hợp nhất dữ liệu đến huấn luyện và đánh giá mô hình. Đồng thời, NCS cũng giải thích rõ mối liên hệ giữa các thành phần trong kiến trúc nhằm giúp người đọc dễ theo dõi quy trình xử lý và vai trò của từng bước trong việc cải thiện công bằng.	Chương 3. Mục 3.2.2. Kiến trúc tổng thể (trang 77-81)
11	- Giải thích các bước và phân tích, đánh giá thuật toán công bằng nhờ cân	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần diễn giải chi tiết các bước thực hiện của Thuật toán 4.1, làm rõ mục đích và vai trò của từng bước trong quá trình phân chia dữ	Chương 3. Mục 3.2.2. Kiến trúc tổng thể (trang 77-81) và 3.2.3.

	bảng dữ liệu DPF (thuật toán 4.1)	liệu theo nhóm giao thoa, sinh dữ liệu tổng hợp và cân bằng dữ liệu huấn luyện. Đồng thời, NCS bổ sung phân phân tích độ phức tạp tính toán, đánh giá vai trò của từng giai đoạn đối với việc cải thiện công bằng, cũng như thảo luận về khả năng áp dụng và những hạn chế của thuật toán trong các trường hợp dữ liệu mất cân bằng nghiêm trọng hoặc có số lượng mẫu rất nhỏ.	Thuật toán (trang 82-84)
12	- Trình bày rõ hơn các kỹ thuật prompt được sử dụng để sinh dữ liệu với mô hình chatGPT 3.5.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung mô tả chi tiết quy trình sinh dữ liệu bằng ChatGPT 3.5, bao gồm các tham số thực nghiệm, chiến lược thiết kế prompt, cấu trúc prompt, ví dụ prompt sử dụng trong thực nghiệm và quy trình sinh dữ liệu theo từng nhóm giao thoa. Qua đó, làm rõ hơn cách thức sinh dữ liệu tổng hợp và nâng cao khả năng tái lập của nghiên cứu.	Chương 3. Mục 3.2.2. Kiến trúc tổng thể (trang 77-81)
13	- Giải thích các bước và phân tích, đánh giá thuật toán FAIREDUPLUS.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần mô tả chi tiết các bước của thuật toán FaireduPlus, làm rõ vai trò của từng giai đoạn và mối liên hệ giữa cơ chế sinh dữ liệu tổng hợp theo nhóm giao thoa với cơ chế loại bỏ sự phụ thuộc của các thuộc tính nhạy cảm. Đồng thời, luận án bổ sung phân phân tích độ phức tạp tính toán, đánh giá vai trò của từng thành phần và hiệu quả của toàn bộ phương pháp trong việc cải thiện công bằng đồng thời duy trì hiệu suất dự đoán của mô hình.	Chương 4. Mục 4.2.2. Kiến trúc tổng thể (trang 111-113)
14	- Nên cập nhật một số tài liệu tham khảo đã cũ.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát toàn bộ danh mục tài liệu tham khảo, loại bỏ hoặc thay thế một số tài liệu đã cũ hoặc ít phù hợp, đồng thời bổ sung các tài liệu mới, đặc biệt là các công trình công bố trong giai đoạn 2024-2026. Sau khi chỉnh sửa, danh mục tài liệu tham khảo được tinh gọn từ 209 xuống còn 197 tài liệu, đồng thời bảo đảm tính cập nhật và phù hợp hơn với nội dung nghiên cứu của luận án.	Mục Tài liệu tham khảo (trang 143-164)

4. PGS.TS Trần Đăng Hưng – Ủy viên Phản biện

1	<p>Luận án chỉ nên để bốn chương, gồm chương tổng quan và ba chương là ba đề xuất chính của luận án; Chương 1 và Chương 6 nên để là Phần Mở đầu và Phần Kết luận.</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã điều chỉnh lại cấu trúc luận án theo hướng cô đọng và phù hợp hơn với quy định đối với luận án tiến sĩ. Cụ thể, Chương 1 được chuyển thành phần <i>Giới thiệu</i>, Chương 6 được chuyển thành phần <i>Kết luận</i>. Nội dung chính của luận án được tổ chức thành bốn chương, gồm một chương <i>Tổng quan</i> và ba chương trình bày ba đề xuất chính của luận án. Việc điều chỉnh này giúp cấu trúc luận án mạch lạc hơn, làm nổi bật các đóng góp khoa học và bảo đảm phù hợp với đặc thù của luận án trong lĩnh vực Khoa học máy tính.</p>	<p>Đổi tên Chương 1 và Chương 6 thành Phần Giới thiệu và Kết luận: Trang 1 và 137. Đổi tên các mục trong Chương 2 cũ (Giờ là Chương 1): Trang 14-40</p>
2	<p>- Chương 2 nên được viết theo kiểu luận án Khoa học máy tính, đưa ra các định nghĩa ngắn gọn và có phân tích thêm các nghiên cứu liên quan đến chủ đề “đảm bảo tính công bằng cho các mô hình học máy trong giáo dục”.</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã chỉnh sửa Chương 2 cũ (giờ là Chương 1) theo hướng phù hợp hơn với cấu trúc của luận án thuộc lĩnh vực Khoa học máy tính, trong đó rút gọn các định nghĩa cơ bản, tập trung trình bày những nội dung trực tiếp phục vụ cho bài toán nghiên cứu. Đồng thời, NCS đã viết lại phần tổng quan theo hướng tăng cường phân tích, so sánh các nghiên cứu liên quan đến bảo đảm tính công bằng cho các mô hình học máy trong giáo dục, làm rõ khoảng trống nghiên cứu và cơ sở khoa học cho ba phương pháp được đề xuất trong luận án.</p>	<p>Chương 1 (mới) – Mục 1.4. Tổng quan nghiên cứu về công bằng cho các hệ thống học máy ứng dụng trong lĩnh vực giáo dục (trang 34–35).</p>
3	<p>- Cần phát biểu ngắn gọn và đầy đủ hơn về các khái niệm cơ bản: “tính công bằng”, “thuộc tính nhạy cảm”,....</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và chỉnh sửa các khái niệm cơ bản theo hướng ngắn gọn, rõ ràng và đầy đủ hơn về mặt khoa học. Bên cạnh việc chuẩn hóa các khái niệm như <i>tính công bằng</i> và <i>thuộc tính nhạy cảm</i>, NCS đã bổ sung phần giới thiệu về khái niệm công bằng trong bối cảnh xã hội, đạo đức và pháp lý, đồng thời trình bày quy định 80% (Four-Fifths Rule) như một nền tảng quan trọng cho các nghiên cứu đánh giá công bằng trong học máy. Các thuật ngữ cũng được thống nhất sử dụng xuyên suốt toàn bộ luận án.</p>	<p>Chương Mục 1.2.1. Khái niệm về công bằng trong các hệ thống học máy (trang 18–19).</p>

4	<p>- Nên bỏ Mục 2.4.1 và Mục 2.4.2, các thống kê trong hai mục này không có nhiều ý nghĩa.</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã tiếp thu ý kiến và chỉnh sửa lại toàn bộ mục tổng quan nghiên cứu. Cụ thể, các Mục 2.4.1 và 2.4.2 trong bản cũ đã được lược bỏ; đồng thời NCS tái cấu trúc toàn bộ Mục 2.4 theo hướng cô đọng hơn, loại bỏ các nội dung thống kê chưa thực sự cần thiết và tập trung phân tích các nghiên cứu liên quan đến bảo đảm tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục. Sau khi chỉnh sửa, nội dung được tổ chức lại thành Mục 1.4. Tổng quan nghiên cứu về công bằng cho các hệ thống học máy ứng dụng trong lĩnh vực giáo dục, qua đó làm rõ khoảng trống nghiên cứu và cơ sở khoa học cho các phương pháp đề xuất trong luận án.</p>	<p>Chương 1. Mục 1.4. Tổng quan nghiên cứu về công bằng cho các hệ thống học máy ứng dụng trong lĩnh vực giáo dục (trang 34, 35)</p>
5	<p>- Mặc dù luận án được thiết kế khá logic và nhiều thực nghiệm, tuy nhiên việc hiệu quả của các đề xuất chỉ được thể hiện khi so sánh với dữ liệu gốc; sẽ tốt hơn nếu có thể so sánh với các phương pháp/mô hình đảm bảo tính công bằng khác.</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. Thực tế, các thực nghiệm trong luận án đã được thiết kế để so sánh phương pháp đề xuất với dữ liệu gốc (Origin) cũng như các phương pháp bảo đảm công bằng tiêu biểu khác như LTDD, Reweighing, FairSMOTE và SDG. Trong bản chỉnh sửa, NCS đã làm rõ hơn nội dung này trong phần thiết lập thực nghiệm, đồng thời bổ sung các phân tích và thảo luận chi tiết hơn về kết quả so sánh nhằm làm nổi bật hiệu quả, ưu điểm và những hạn chế của phương pháp đề xuất so với các phương pháp đối sánh.</p>	<p>Chương 2. Mục 2.3.3. Thiết lập thực nghiệm (trang 55) 3.3.3. Thiết lập thực nghiệm (trang 89-90) 4.3.2. Thiết lập thực nghiệm (trang 117)</p>
6	<p>- Nên cung cấp các bảng kết quả liên quan đến độ chính xác (F1) của mô hình sau khi điều chỉnh dữ liệu huấn luyện; vì mục tiêu cuối cùng của các mô hình học máy là độ chính xác càng cao càng tốt, nên mô hình phải vừa đảm bảo tính công bằng vừa phải đảm bảo độ chính xác thì</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. Trong bản chỉnh sửa, ngoài các chỉ số Accuracy và Recall, NCS đã bổ sung kết quả đánh giá theo Precision và F1-score đối với phương pháp Fairedu nhằm cung cấp cái nhìn đầy đủ hơn về ảnh hưởng của phương pháp đề xuất đến hiệu năng dự đoán. Các kết quả này được trình bày trong Phụ lục B (Bảng B.2), so sánh giữa mô hình gốc, mô hình áp dụng LTDD và mô hình áp dụng Fairedu. Đối với các phương pháp DPF và FaireduPlus, luận án chủ yếu đánh giá hiệu năng dự đoán thông qua các chỉ số</p>	<p>Phụ lục B – Bảng B.2 (bổ sung kết quả Precision và F1-score của Fairedu). (trang 181, 182)</p>

	đề xuất mới có ý nghĩa.	Accuracy và Recall. Trong thời gian tới, NCS sẽ tiếp tục mở rộng nghiên cứu, bổ sung đánh giá bằng các chỉ số Precision và F1-score nhằm có cái nhìn toàn diện hơn về sự đánh đổi giữa công bằng và hiệu năng của các phương pháp đề xuất.	
7	- Cần xem lại kí hiệu của các thuộc tính nhạy cảm trong các Thuật toán 3.1, 4.1, 5.1 cho thống nhất, hiện tại đang kí hiệu không thống nhất (<i>xkj</i> hay <i>xk</i>).	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và thống nhất hệ thống ký hiệu sử dụng trong các Thuật toán 2.1, 3.1 và 4.1. Cụ thể, các thuộc tính nhạy cảm và không nhạy cảm đều được biểu diễn theo cùng một quy ước ký hiệu, bảo đảm tính nhất quán giữa các thuật toán và tránh gây nhầm lẫn trong quá trình theo dõi luận án.	Chương 2. Mục Thuật toán 2.1 (trang 48) Thuật toán 3.1 (trang 82) Thuật toán 4.1 (trang 114)
8	- Chương 4, làm rõ hơn quá trình sinh thêm dữ liệu từ các bộ dữ liệu sau khi phân nhóm, số lượng sinh bao nhiêu và cách sinh dữ liệu như thế nào cho mỗi nhóm. Đồng thời nên bình luận thêm đối với những trường hợp mà thuộc tính nhạy cảm không có giá trị nhị phân thì như thế nào?	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung và làm rõ quy trình sinh dữ liệu tổng hợp trong phương pháp DPF thông qua phân mô tả chi tiết kiến trúc tổng thể, trong đó trình bày cụ thể các bước phân chia dữ liệu theo nhóm giao thoa, cách xác định số lượng mẫu cần sinh cho từng nhóm và quy trình sinh dữ liệu bằng CTGAN hoặc LLM trước khi hợp nhất thành tập dữ liệu huấn luyện mới. Bên cạnh đó, NCS cũng bổ sung phần thảo luận về trường hợp thuộc tính nhạy cảm không nhận giá trị nhị phân trong mục khái niệm về công bằng, làm rõ rằng bài toán công bằng được xây dựng trên việc so sánh giữa nhóm ưu thế và nhóm yếu thế; do đó, đối với thuộc tính nhạy cảm có nhiều hơn hai giá trị, việc đánh giá công bằng vẫn được thực hiện thông qua hai nhóm có tỷ lệ được lựa chọn cao nhất và thấp nhất, không làm hạn chế khả năng áp dụng của phương pháp đề xuất.	Chương 3. Mục 3.2.2. Kiến trúc tổng thể (trang 78–81). Mục 1.2.1. Khái niệm về công bằng trong các hệ thống học máy (trang 18–19).
5. PGS.TS Lê Hồng Phương – Ủy viên Phản biện			
1	a) Tính nhất quán nội tại của Fairedu (Chương 3)		
	- Khi đối chiếu các Bảng 3.4–3.6, Fairedu không nhất quán thẳng LTDD.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và bổ sung phần phân tích kết quả thực nghiệm trong Chương 3 cũ (Chương 2 mới) theo	Chương 2. Mục 2.4.3. Khả năng xử lý đồng thời

	<p>Ví dụ ở mô hình hồi quy logistic, thuộc tính SD_GT: $Fairedu = 0.740$ trong khi LTDD chỉ $= 0.141$ (Bảng 3.4) — $Fairedu$ thua khá xa, không chỉ “thua nhẹ”. Với rừng ngẫu nhiên, SD_GT: $Fairedu = 0.631$ vs LTDD $= 0.396$; DNU_Tuổi: $Fairedu = 0.192$ vs LTDD $= 0.056$ – đều là những khoảng cách lớn.</p>	<p>hướng chi tiết và khách quan hơn. Cụ thể, luận án đã làm rõ rằng $Fairedu$ không đạt kết quả tốt nhất trong mọi cấu hình thực nghiệm; vẫn tồn tại một số trường hợp LTDD cho kết quả công bằng tốt hơn đối với từng bộ dữ liệu hoặc từng thuộc tính nhạy cảm. Tuy nhiên, khi đánh giá trên toàn bộ các bộ dữ liệu, mô hình học máy, thuộc tính nhạy cảm và các chỉ số công bằng, $Fairedu$ vẫn đạt số lượt thắng cao hơn các phương pháp đối sánh. NCS cũng đã điều chỉnh cách diễn giải kết quả theo hướng nhấn mạnh xu hướng tổng thể của toàn bộ thực nghiệm thay vì khẳng định tính vượt trội dựa trên từng trường hợp riêng lẻ, qua đó phản ánh khách quan hơn hiệu quả và khả năng tổng quát của phương pháp đề xuất.</p>	<p>nhiều thuộc tính nhạy cảm của FairEdu (trang 60-62)</p>
	<p>- Tỷ lệ thắng/hòa/thua tổng hợp (ví dụ 5/0/4) che lấp việc một số trường hợp thua có biên độ rất lớn. Luận án nên báo cáo thêm độ lớn chênh lệch (effect size), không chỉ đếm lượt thắng/thua, và lý giải tại sao $Fairedu$ lại kém ổn định hơn LTDD chính trên thuộc tính “tình trạng nợ”/”giới tính” ở các tập dữ liệu thua – đây có lẽ là dấu hiệu hồi quy đa biến gặp đa cộng tuyến hoặc quá khớp khi kết hợp nhiều biến nhạy cảm tương quan với nhau, hoặc tập dữ liệu kích thước nhỏ.</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung và chỉnh sửa phần thảo luận kết quả thực nghiệm theo hướng khách quan hơn, làm rõ rằng hiệu quả của $Fairedu$ cần được đánh giá trên xu hướng tổng thể của toàn bộ các cấu hình thực nghiệm, đồng thời phân tích thêm một số trường hợp $Fairedu$ chưa đạt kết quả tốt nhất so với LTDD. Bên cạnh đó, trong phần hạn chế của nghiên cứu, NCS đã bổ sung nội dung thảo luận về các yếu tố có thể ảnh hưởng đến tính ổn định của phương pháp, như kích thước dữ liệu nhỏ, mức độ mất cân bằng giữa các nhóm nhạy cảm và đặc điểm của dữ liệu đầu vào khi áp dụng hồi quy đa biến. Các thước đo phản ánh mức độ khác biệt giữa các phương pháp (như effect size) sẽ được NCS tiếp tục nghiên cứu và bổ sung trong các công trình tiếp theo nhằm đánh giá toàn diện hơn hiệu quả của phương pháp đề xuất.</p>	<p>Chương 2. Mục 2.4.3. Khả năng xử lý đồng thời nhiều thuộc tính nhạy cảm của FairEdu (trang 60-62) Mục 2.5.2. Hạn chế của nghiên cứu (trang 67-69).</p>
<p>2</p>	<p>b) Thiếu kiểm định thống kê chặt</p>		

	chế cho so sánh thắng/thua		
	<p>- Toàn luận án dùng cách đếm “thắng/hòa/thua” dựa trên so sánh điểm đơn lẻ giữa các cấu hình (không phải trên nhiều lần lặp có kiểm định ý nghĩa thống kê, trừ một vài trường hợp ở chương 3 có nhắc $p < 0.05$ cho kiểm định Wald khi ước lượng hệ số hồi quy, chứ không phải cho so sánh hiệu quả giữa phương pháp). Việc này khiến các con số như “27/45 thắng” có thể bị ảnh hưởng bởi khác biệt rất nhỏ, không có ý nghĩa thực tế. Nên bổ sung kiểm định Wilcoxon signed-rank hoặc khoảng tin cậy bootstrap cho các so sánh tổng hợp.</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. Trong luận án, phương pháp thống kê Thắng/Hòa/Thua được sử dụng nhằm đánh giá xu hướng tổng thể của các phương pháp trên nhiều bộ dữ liệu, mô hình học máy, thuộc tính nhạy cảm và chỉ số công bằng, thay vì chỉ dựa trên một cấu hình riêng lẻ. Để giảm ảnh hưởng của tính ngẫu nhiên, các thực nghiệm được lặp lại nhiều lần và kết quả được báo cáo dưới dạng giá trị trung bình. Trong bản chỉnh sửa, NCS đã bổ sung phần thảo luận làm rõ ý nghĩa và phạm vi áp dụng của phương pháp thống kê này. Đồng thời, NCS tiếp thu ý kiến của Hội đồng và xem việc bổ sung các kiểm định thống kê như Wilcoxon signed-rank hoặc khoảng tin cậy bootstrap là một hướng nghiên cứu tiếp theo nhằm đánh giá toàn diện hơn mức độ khác biệt giữa các phương pháp.</p>	<p>Chương 2. Mục 2.3.3. Thiết lập thực nghiệm (trang 55) Phần Kết luận, Hướng phát triển trong tương lai, ý 2. Phát triển và hoàn thiện các chỉ số công bằng và phương pháp đánh giá (trang 140-141)</p>
3	c) Chỉ số đánh đổi (trade-off index) ở Chương 5		
	<p>- Bảng 5.5–5.6 cho thấy chỉ số đánh đổi có những giá trị cực đoan bất thường: -9750.4%, -4601.8%, -2517.9%, -2317.8%... Những giá trị phân trăm âm hàng nghìn lần cho thấy công thức tính (rất có thể là cải</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và bổ sung phần thảo luận về chỉ số đánh đổi trong Chương 5. Cụ thể, luận án đã làm rõ rằng chỉ số đánh đổi được xây dựng dựa trên mức cải thiện tương đối nên có thể xuất hiện các giá trị lớn trong những trường hợp giá trị gốc của chỉ số công bằng hoặc hiệu suất rất nhỏ. NCS đã bổ sung phân nhận xét về ảnh hưởng của hiện tượng này đối với việc diễn giải kết quả và coi đây là một hạn chế của chỉ số hiện tại.</p>	<p>Chương 4. Mục 4.5.2. Hạn chế của nghiên cứu (trang 132, 133)</p>

	<p>thiện tương đối chia cho baseline gần 0) bị bất ổn định khi mẫu số (mức công bằng/hiệu suất gốc) tiến gần 0 – đây là lỗi kinh điển của các chỉ số “cải thiện tương đối”. Hệ quả là: Các thống kê thắng/thua dựa trên chỉ số này (Bảng 5.5, 5.6) có thể bị chi phối bởi vài giá trị ngoại lai cực đoan, làm méo mó kết luận tổng thể. Luận án chưa làm rõ liệu các trường hợp giá trị “khủng” này có được loại trừ, hay dùng trung vị thay cho chung bình khi tổng hợp không.</p>	<p>Đồng thời, luận án bổ sung định hướng nghiên cứu tiếp theo nhằm phát triển các công thức đánh đổi ổn định hơn, kết hợp các phương pháp tổng hợp như trung vị, xử lý giá trị ngoại lai và các kỹ thuật chuẩn hóa thích hợp.</p>	
	<p>- Việc đề xuất chỉ số đánh đổi là một trong 4 đóng góp chính, nhưng chính chương 5 cũng tự nhận “<i>chỉ số đánh đổi mới dừng ở mức khái niệm và thử nghiệm ban đầu</i>” – nên đây là điểm cần làm rõ thêm về tính ổn định số học của công thức, không chỉ về phạm vi áp dụng.</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung và làm rõ hơn phần thảo luận về chỉ số đánh đổi trong Chương 5. Cụ thể, luận án nhấn mạnh rằng đây là chỉ số được đề xuất ở mức thử nghiệm ban đầu nhằm hỗ trợ đánh giá sự đánh đổi giữa công bằng và hiệu suất của mô hình. Đồng thời, NCS đã bổ sung phần phân tích về hạn chế của chỉ số, trong đó làm rõ khả năng kém ổn định khi giá trị gốc của chỉ số công bằng hoặc hiệu suất tiến gần 0 do sử dụng mức cải thiện tương đối. Trên cơ sở đó, luận án đề xuất hướng nghiên cứu tiếp theo nhằm hoàn thiện chỉ số thông qua các kỹ thuật chuẩn hóa, xử lý giá trị ngoại lai và xây dựng các công thức đánh đổi ổn định hơn.</p>	<p>Chương 4. Mục 4.5.2. Hạn chế của nghiên cứu (trang 132, 133) Phần Kết luận (trang 137- 141)</p>
4	<p>d) Mâu thuẫn giữa kết luận tổng thể và dữ liệu chi tiết ở Chương 4 (DPF)</p>		

	<p>- Bảng 4.7 cho thấy với CTGAN, cấu hình SDG (không phân tách nhóm) thắng DPF ở chỉ số 1-DI (23 so với 11) và SPD (21 so với 14) – DPF chỉ thắng ở AOD, EOD. Với LLM thì DPF mới áp đảo toàn diện. Điều này có nghĩa hiệu quả của DPF phụ thuộc mạnh vào kỹ thuật sinh dữ liệu được dùng kèm, chứ DPF không phải lúc nào cũng vượt trội – đây là một phát hiện thú vị nhưng phần thảo luận tổng kết chương có xu hướng diễn giải nghiêng về phía thuận lợi cho DPF hơn là nhấn rõ sự phụ thuộc/điều kiện này.</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và điều chỉnh phần thảo luận kết quả thực nghiệm trong Chương 4 cũ (Chương 3 mới) theo hướng khách quan và đầy đủ hơn.</p> <p>Cụ thể, luận án đã làm rõ rằng hiệu quả của DPF phụ thuộc đáng kể vào kỹ thuật sinh dữ liệu tổng hợp được sử dụng. Kết quả thực nghiệm cho thấy với CTGAN, DPF chủ yếu cải thiện các chỉ số liên quan đến sai số dự đoán (AOD và EOD), trong khi SDG đạt kết quả tốt hơn ở các chỉ số phân phối (1-DI và SPD). Ngược lại, khi kết hợp với LLM, DPF thể hiện ưu thế trên cả bốn chỉ số công bằng.</p>	<p>Chương 3. Mục 3.4.2.1. Đánh giá thông qua các chỉ số công bằng (trang 95-96)</p>
	<p>- Tương tự, Bảng 4.10 cho thấy Origin (không can thiệp) thắng Accuracy nhiều nhất (26/45), DPF chỉ 13/45 – cho thấy can thiệp công bằng vẫn phải trả giá về hiệu suất rõ ràng hơn so với cách luận án mô tả (“duy trì hiệu suất cạnh tranh”).</p>	<p>Bên cạnh đó, NCS cũng bổ sung phần phân tích về sự đánh đổi giữa công bằng và hiệu suất, làm rõ rằng việc cải thiện công bằng có thể làm giảm Accuracy trong một số cấu hình thực nghiệm. Vì vậy, kết luận của luận án được điều chỉnh theo hướng nhấn mạnh khả năng đạt được sự cân bằng hợp lý giữa công bằng và hiệu suất thay vì khẳng định duy trì hiệu suất trong mọi trường hợp.</p>	<p>Chương 3. Mục 3.4.3. Tác động của DPF đến hiệu suất dự đoán của mô hình (trang 101)</p>
<p>5</p>	<p>đ) Vấn đề dữ liệu thừa/nhóm giao thoa rộng chưa</p>		

	được giải quyết triệt để		
	<p>- Cả ba phương pháp đều thừa nhận hạn chế khi nhóm con có 0 hoặc 1 mẫu (DNU Data) – các nhóm này bị loại khỏi quá trình cân bằng.</p> <p>- Đây chính là nhóm dễ bị bỏ sót nhất trong thực tế (nhóm thiếu số cực nhỏ), nên về bản chất phương pháp "bỏ qua" đúng nơi cần công bằng nhất – một nghịch lý cần được thảo luận sâu hơn, ví dụ dùng kỹ thuật tăng cường dữ liệu chuyên biệt, hoặc gộp nhóm thông minh hơn là loại bỏ</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần thảo luận làm rõ hạn chế của DPF và FaireduPlus đối với các nhóm giao thoa có số lượng mẫu quá ít (0 hoặc 1 mẫu). Luận án giải thích rằng các nhóm này không được đưa vào quá trình sinh dữ liệu nhằm tránh tạo ra dữ liệu tổng hợp thiếu độ tin cậy do không đủ thông tin để học phân bố dữ liệu. Đồng thời, NCS nhấn mạnh đây là một hạn chế của phương pháp hiện tại và bổ sung trong phần hướng nghiên cứu tiếp theo các định hướng phát triển kỹ thuật sinh dữ liệu chuyên biệt và các chiến lược gộp nhóm phù hợp nhằm hỗ trợ tốt hơn các nhóm thiếu số có rất ít mẫu.</p>	<p>Chương 3. Mục 3.5.3. Hạn chế của nghiên cứu (trang 105)</p> <p>Phần Kết luận, ý 3. Nghiên cứu giải pháp sinh dữ liệu tổng hợp nhằm cân bằng dữ liệu cho các tình huống đặc biệt (trang 141)</p>
6	e) Thiếu kiểm soát đa kiểm định		
	<p>Thuật toán Fairedu dùng kiểm định Wald với ngưỡng $p < 0.05$ cho mỗi đặc trưng không nhạy cảm và mỗi thuộc tính nhạy cảm, lặp lại trên nhiều bộ dữ liệu/mô hình mà không có điều chỉnh Bonferroni hay FDR – có nguy cơ dương tính giả khi xác định “phụ thuộc có ý nghĩa thống kê” cần điều chỉnh.</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần thảo luận về hạn chế của phương pháp Fairedu, trong đó làm rõ rằng việc xác định các thuộc tính cần điều chỉnh dựa trên kiểm định Wald với ngưỡng ý nghĩa thống kê $p < 0.05$ có thể chịu ảnh hưởng của hiện tượng đa kiểm định khi áp dụng trên nhiều thuộc tính và nhiều bộ dữ liệu. Đồng thời, trong phần hướng nghiên cứu tiếp theo, NCS đã bổ sung định hướng nghiên cứu kết hợp các kỹ thuật hiệu chỉnh đa kiểm định như Bonferroni hoặc False Discovery Rate (FDR), cùng với các kiểm định thống kê như Wilcoxon signed-rank, bootstrap và các chỉ số kích thước ảnh hưởng (effect size), nhằm nâng cao độ tin cậy thống kê</p>	<p>Phần Kết luận, ý 2. Phát triển và hoàn thiện các chỉ số công bằng và phương pháp đánh giá (trang 140)</p>

	Gợi ý: thực hiện điều chỉnh FDR và kiểm định thống kê	và tính toàn diện trong việc đánh giá hiệu quả của các phương pháp đề xuất.	
7	f) Phạm vi mô hình và dữ liệu còn hẹp		
	Chỉ kiểm chứng trên mô hình truyền thống (logistic regression, decision tree, random forest, gradient boosting, mạng nơ-ron đơn giản), không có mô hình sâu/transformer. Với bối cảnh năm 2025-2026 thì đây là một giới hạn về tính kịp thời.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung và làm rõ phạm vi áp dụng của các phương pháp đề xuất trong phần hạn chế của nghiên cứu. Cụ thể, luận án nhấn mạnh rằng các thực nghiệm hiện được thực hiện trên các mô hình học máy truyền thống nhằm đánh giá tính tổng quát của các phương pháp tiền xử lý. Đồng thời, trong phần hướng nghiên cứu tiếp theo, NCS đã bổ sung định hướng mở rộng thực nghiệm trên các mô hình học sâu và các kiến trúc Transformer hiện đại nhằm kiểm chứng khả năng áp dụng của các phương pháp trong bối cảnh phát triển AI hiện nay.	Phần Kết luận, ý Hạn chế (trang 141) và ý 1. Mở rộng phạm vi dữ liệu và mô hình áp dụng (trang 139)
8	g) Tính minh bạch của dữ liệu sinh tổng hợp bởi LLMs		
	Luận án không nêu rõ cách thiết kế prompt, cách kiểm định chất lượng/độ trung thực của dữ liệu sinh ra (fidelity, privacy leakage) — đây là khoảng trống về tái lập (reproducibility). LLM thay đổi nhanh, ảnh hưởng tính tái lập được của các thí nghiệm.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. Trong bản chỉnh sửa, NCS đã bổ sung chi tiết quy trình sinh dữ liệu bằng CTGAN và LLM, bao gồm thiết kế prompt, các tham số thực nghiệm và quy trình sinh dữ liệu nhằm nâng cao khả năng tái lập của nghiên cứu. Đồng thời, luận án đã bổ sung phần thảo luận về chất lượng dữ liệu tổng hợp, nguy cơ rò rỉ thông tin và các hạn chế khi sử dụng LLM, trong đó làm rõ rằng việc cập nhật liên tục của các mô hình ngôn ngữ lớn có thể ảnh hưởng đến khả năng tái lập hoàn toàn các thực nghiệm. Bên cạnh đó, NCS cũng bổ sung trong phần hạn chế và hướng nghiên cứu tiếp theo định hướng nghiên cứu các phương pháp đánh giá chuyên biệt đối với dữ liệu tổng hợp như fidelity, privacy leakage và các tiêu chí đánh giá chất lượng dữ liệu nhằm nâng cao độ tin cậy của phương pháp.	Phần Kết luận, ý Hạn chế (trang 139) và ý 3. Nghiên cứu giải pháp sinh dữ liệu tổng hợp nhằm cân bằng dữ liệu cho các tình huống đặc biệt (trang 141)

6. PGS.TS Phan Xuân Hiếu – Ủy viên Hội đồng

6. PGS.TS Phan Xuân Hiếu – Ủy viên Hội đồng			
1	Tiêu đề các chương, mục hiện còn hơi dài dòng:		
	o Tiêu đề Mục 1.1.2 (Các thách thức và khoảng trống nghiên cứu trong đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục) => nên để gọn lại thành “Các thách thức và khoảng trống nghiên cứu” là đủ.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và điều chỉnh toàn bộ hệ thống tiêu đề chương, mục theo hướng ngắn gọn, súc tích và thống nhất hơn, bao gồm việc rút gọn tên Chương 2 cũ (Chương 1 mới), Chương 3 cũ (Chương 2 mới) và các mục có tiêu đề dài theo đúng tinh thần góp ý của Hội đồng. Đồng thời, NCS đã viết lại Chương 2 (cũ), chuyển thành Chương 1 (mới) theo hướng cô đọng hơn, lược bỏ các khái niệm cơ sở không thực sự cần thiết, chỉ giữ lại những nội dung liên quan trực tiếp đến bài toán đảm bảo tính công bằng trong học máy, qua đó làm nổi bật hơn khoảng trống nghiên cứu và cơ sở khoa học cho các phương pháp đề xuất.	Phản Giới thiệu, . Tiêu đề mục 0.1.2. Các thách thức và khoảng trống nghiên cứu (trang 4) Chương 1. (trang 14 - 40)
	Tiêu đề Chương 2 có thể để ngắn gọn lại như sau: “Tổng quan về đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục” (bỏ cụm “trí tuệ nhân tạo”). Tương tự còn nhiều mục khác và tiêu đề Chương 3 cũng còn khá dài dòng, nên làm gọn bớt.		
2	- Mục 2.1.1. (Các khái niệm) mô tả các khái niệm như “trí tuệ nhân tạo”, “học máy”, “dữ liệu”, “thuật toán học máy”, “mô hình học máy” ... không thật sự cần thiết. Đây là các khái niệm cơ sở gần như độc giả nào của luận án này cũng đã nắm bắt được. Vì thế, có thể xem xét		

	bỏ mục này để rút ngắn độ dài của luận án.		
3	- Về các công thức và mô hình hoá: Cần chú thích ý nghĩa của các thành phần trong các công thức ở Bảng 2.1 (Các định nghĩa về công bằng trong hệ thống học máy, trang 22). Ví dụ: Y là gì, Y mũ là gì, A là gì ...? Dự đoán “tích cực” có phải là dự đoán vào “lớp dương” (positive)? Nếu vậy thì không nên dịch là “tích cực” mà dịch là “lớp dương”.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và bổ sung chú thích đầy đủ ý nghĩa của các ký hiệu sử dụng trong các công thức tại Bảng 1.1, như Y (nhãn thực), Y [^] (nhãn dự đoán) và A (thuộc tính nhạy cảm), nhằm giúp người đọc dễ theo dõi và tránh nhầm lẫn. Đồng thời, NCS đã thống nhất sử dụng thuật ngữ "lớp dương" (positive class) thay cho "dự đoán tích cực" trong toàn bộ luận án để bảo đảm chính xác về mặt thuật ngữ và phù hợp với cách diễn đạt phổ biến trong học máy.	Chương 1. Mục 1.2.1. Khái niệm về công bằng trong các hệ thống học máy. Bổ sung Chú ý cho bảng định nghĩa về công bằng (trang 18-19) Toàn bộ luận án
4	o Tại sao chỉ phát biểu cho bài toán phân lớp nhị phân? (công thức 2.5, trang 25)	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung làm rõ phạm vi áp dụng của mô hình hóa và các công thức trong luận án. Cụ thể, luận án tập trung vào bài toán phân lớp nhị phân do các thước đo công bằng được sử dụng (DI, SPD, EOD và AOD) chủ yếu được định nghĩa cho bài toán này, đồng thời các bộ dữ liệu thực nghiệm đều được chuyển đổi về hai lớp để thống nhất với mục tiêu nghiên cứu. NCS cũng bổ sung nội dung thảo luận rằng việc mở rộng phương pháp cho các bài toán phân lớp đa lớp là một hướng nghiên cứu tiếp theo.	Chương 1. Mục 1.2.4. Độ đo công bằng trong các hệ thống học máy (trang 24-26) Phần Kết luận, ý 1. Mở rộng phạm vi dữ liệu và mô hình áp dụng (trang 140)
5	Tại sao thuộc tính A lại chỉ có trường hợp nhị phân, tức A chia dữ liệu thành hai nhóm (nhóm ưu tiên và nhóm không ưu tiên)? Tại sao không tổng quát	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần giải thích trong Mục 1.2.1. <i>Khái niệm về công bằng trong các hệ thống học máy</i> (trang 19) nhằm làm rõ cơ sở của việc phát biểu bài toán đối với hai nhóm so sánh. Trong đó chỉ ra rằng thuộc tính nhạy cảm không nhất thiết phải là biến	Mục 1.2.1. Khái niệm về công bằng trong các hệ thống học máy (trang 18, 19)

	<p>hoá A chia dữ liệu thành k nhóm? (trang 25)</p>	<p>nhị phân. Đối với các thuộc tính nhạy cảm có nhiều hơn hai giá trị, việc đánh giá công bằng vẫn được thực hiện thông qua so sánh giữa nhóm có tỷ lệ dương tính cao nhất (nhóm ưu thế) và nhóm có tỷ lệ dương tính thấp nhất (nhóm yếu thế).</p>	
6	<p>o Ở độ đo về “Hiệu số chênh lệch thống kê” (SPD, công thức 2.7, trang 26): “Giá trị SPD càng nhỏ (càng gần 0) thể hiện sự công bằng theo định nghĩa công bằng nhân khẩu học” => SPD có thể có giá trị âm? Nếu vậy thì phải viết là “trị tuyệt đối của SPD càng nhỏ (càng gần 0) ...”</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã bổ sung phần giải thích cho các thước đo SPD và AOD nhằm làm rõ cơ sở xây dựng công thức. Cụ thể, trong luận án, các công thức được phát biểu trên cơ sở quy định 80% của EEOC, trong đó thuộc tính nhạy cảm được xác định theo hai nhóm: nhóm ưu thế (có tỷ lệ được lựa chọn cao nhất) và nhóm yếu thế (có tỷ lệ được lựa chọn thấp nhất). Do đó, các hiệu số trong công thức SPD và AOD luôn được tính theo cùng một thứ tự (nhóm ưu thế trừ nhóm yếu thế), nên không cần sử dụng giá trị tuyệt đối trong công thức. Đồng thời, NCS đã chỉnh sửa phần diễn giải của chỉ số SPD theo hướng thống nhất với công thức, làm rõ rằng trong phạm vi luận án, SPD luôn nhận giá trị không âm do được tính theo thứ tự giữa nhóm ưu thế và nhóm yếu thế, vì vậy giá trị càng gần 0 thì mức độ công bằng càng cao.</p>	<p>Chương 1. Mục 1.2.1. Khái niệm về công bằng trong các hệ thống học máy (trang 18-26)</p>
7	<p>Công thức tính “Chênh lệch trung bình xác suất” (AOD) ở trang 26 có cần lấy giá trị tuyệt đối của hiệu TPR và trị tuyệt đối của hiệu FPR của hai nhóm trước khi cộng vào nhau hay không?</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. Để làm rõ đầy đủ quy trình của phương pháp FairEdu, NCS đã bổ sung phần mô tả tại Bước 3 (<i>Huấn luyện mô hình</i>) nhằm nhấn mạnh rằng sau khi loại bỏ sự phụ thuộc giữa các thuộc tính không nhạy cảm và các thuộc tính nhạy cảm, các thuộc tính nhạy cảm cũng được loại bỏ khỏi tập dữ liệu trước khi huấn luyện mô hình. Việc bổ sung này giúp làm rõ hai bước xử lý của FairEdu được trình bày trong Hình 3.1 cũ (Hình 2.1 mới).</p>	
8	<p>o Hình 3.1 (“Kiến trúc tổng thể của phương pháp FairEdu”, trang 53): nên làm rõ hai ý: (1) loại bỏ các thuộc tính nhạy cảm, và (2) loại bỏ sự phụ thuộc giữa các thuộc tính không nhạy cảm đối với các thuộc tính nhạy cảm. Hiện</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. Để làm rõ đầy đủ quy trình của phương pháp FairEdu, NCS đã bổ sung phần mô tả tại Bước 3 (<i>Huấn luyện mô hình</i>) nhằm nhấn mạnh rằng sau khi loại bỏ sự phụ thuộc giữa các thuộc tính không nhạy cảm và các thuộc tính nhạy cảm, các thuộc tính nhạy cảm cũng được loại bỏ khỏi tập dữ liệu trước khi huấn luyện mô hình. Việc bổ sung này giúp làm rõ hai bước xử lý của FairEdu được trình bày trong Hình 3.1 cũ (Hình 2.1 mới).</p>	<p>Chương 2. Mục 2.2.2. Kiến trúc tổng thể, Bước 3 (<i>Huấn luyện mô hình</i>) (trang 46)</p>

	hình 3.1 mới mô tả ý thứ (2).		
9	- Các phát biểu cần xem xét:		
	o “Đây là vấn đề phổ biến trong dữ liệu giáo dục, nơi các nhóm yếu thế thường có số lượng mẫu rất nhỏ, gây khó khăn cho việc huấn luyện mô hình và đảm bảo công bằng.” (trang 9, mục 1.2.3 – Phạm vi nghiên cứu, nhánh 2) <= Tại sao “các nhóm yếu thế” lại thường “có số lượng mẫu rất nhỏ”? Ví dụ một tập dữ liệu	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã chỉnh sửa nội dung tại Mục 0.2.3 (<i>Phạm vi nghiên cứu</i>) nhằm diễn đạt chính xác hơn. Thay vì khẳng định các nhóm yếu thế thường có số lượng mẫu rất nhỏ, luận án đã làm rõ rằng hiện tượng này chủ yếu xảy ra khi đồng thời xem xét nhiều thuộc tính nhạy cảm, dẫn đến các nhóm giao thoa có rất ít hoặc không có mẫu. Đây cũng chính là động lực cho việc đề xuất các phương pháp DPF và FaireduPlus nhằm xử lý tình trạng mất cân bằng dữ liệu giữa các nhóm giao thoa.	Phần Giới thiệu. Mục 0.2.3 <i>Phạm vi nghiên cứu, nhánh 2 (trang 9)</i>
10	Lỗi hành văn và chính tả:		
	o “Nghiên cứu về đảm bảo tính công bằng cho các hệ thống học máy trong giáo dục đang đối mặt với nhiều thách thức đặc thù, phản ánh sự phức tạp của cả về kỹ thuật lẫn ứng dụng.” (trang 4)	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và chỉnh sửa câu văn tại trang 4 nhằm bảo đảm diễn đạt chính xác, mạch lạc và đúng ngữ pháp. Đồng thời, NCS cũng rà soát toàn bộ luận án để chỉnh sửa các lỗi hành văn, chính tả và thống nhất cách diễn đạt. “Nghiên cứu về đảm bảo tính công bằng cho các hệ thống học máy trong giáo dục đang đối mặt với nhiều thách thức đặc thù do tính phức tạp cả về kỹ thuật lẫn bối cảnh ứng dụng.”	Phần Giới thiệu. Mục 0.1.2. Các thách thức và khoảng trống nghiên cứu (trang 4-5)
	o “Nghiên cứu các cơ chế tiền xử lý nhằm giảm phụ thuộc giữa đặc trưng và thuộc tính nhạy cảm, hỗ trợ công bằng đa nhóm.” (trang 6, mục tiêu cụ thể số 1) <= “đặc trưng”	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và thống nhất cách sử dụng thuật ngữ trong toàn bộ luận án. Cụ thể, cụm từ “đặc trưng” trong mục tiêu nghiên cứu đã được thay bằng “các thuộc tính không nhạy cảm” nhằm thống nhất với thuật ngữ “thuộc tính nhạy cảm” được sử dụng xuyên suốt trong luận án.	Phần Giới thiệu. Mục 0.2.1. Mục tiêu nghiên cứu (trang 6) và toàn bộ luận án.

	và “thuộc tính” khác g. nhau?		
	o “Đảm bảo TPR và FPR bằng nhau giữa các nhóm (ví dụ: nam và nữ)” (trang 22, bảng 2.1) => những cụm như thế này rất dễ gây hiểu nhầm	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã chỉnh sửa phần diễn giải tại Bảng 1.1 theo hướng tổng quát hơn, bổ sung nội dung “Đảm bảo TPR và FPR tương đương giữa nhóm ưu thế và nhóm yếu thế”.	Chương 1. Bảng 1.1: Các định nghĩa về công bằng trong hệ thống học máy (trang 21)
	o “Đề giải quyết vấn đề này, chương trình bày phương pháp FairEdu, một cách tiếp cận tiên xử” (trang 48) <= “... chương nay tr.nh ...”	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã chỉnh sửa câu văn tại phần mở đầu Chương 2 theo hướng diễn đạt rõ ràng, mạch lạc hơn, thay cách diễn đạt “chương trình bày” bằng “trong chương này, luận án đề xuất phương pháp FairEdu”, qua đó làm nổi bật đóng góp của nghiên cứu và bảo đảm tính chuẩn mực trong diễn đạt khoa học.	Chương 2 (trang 41)
7. TS Nguyễn Thị Minh Huyền - Ủy viên Hội đồng			
1	<p>Phần tổng quan có thể rút gọn một số nội dung, ví dụ như các khái niệm cơ bản ở 2.1.1.</p> <p>Phần diễn giải kết quả thực nghiệm ở các chương thiên về mô tả con số mà chưa phân tích sâu nguyên nhân dẫn tới sự khác biệt giữa các phương pháp.</p>	<p>NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và viết lại Chương 1 theo hướng cô đọng hơn, lược bỏ các khái niệm cơ sở không thực sự cần thiết (như Mục 2.1.1 cũ), đồng thời rút gọn và tái cấu trúc phần tổng quan, tập trung vào phân tích các nghiên cứu liên quan, khoảng trống nghiên cứu và cơ sở khoa học cho các phương pháp đề xuất.</p> <p>NCS đã rà soát và bổ sung phần phân tích kết quả thực nghiệm trong các chương nghiên cứu theo hướng làm rõ nguyên nhân dẫn đến sự khác biệt giữa các phương pháp, thay vì chỉ mô tả các kết quả định lượng. Cụ thể, luận án đã bổ sung các phân tích về ảnh hưởng của đặc điểm dữ liệu, mức độ mất cân bằng giữa các nhóm, số lượng thuộc tính nhạy cảm, kỹ thuật sinh dữ liệu và cơ chế hoạt động của từng phương pháp nhằm giải thích các kết quả thực nghiệm và làm nổi bật ưu điểm cũng như hạn chế của các phương pháp đề xuất.</p>	<p>Chương 1 (mới), (trang 14-40).</p> <p>kết quả thực nghiệm trong Chương 2 (trang 56-65, Chương 3 (trang 91-101) và Chương 4 (trang 122-130).</p>

2	Phân danh sách các từ viết tắt, các kí hiệu cần trình bày theo thứ tự ABC để tiện tra cứu.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và sắp xếp lại toàn bộ danh sách các từ viết tắt và các ký hiệu theo thứ tự bảng chữ cái (ABC) nhằm tạo thuận lợi cho việc tra cứu, đồng thời thống nhất cách trình bày trong toàn bộ luận án.	Bảng thuật ngữ và từ viết tắt (trang ix-xi)
3	Tài liệu tham khảo phong phú, tuy nhiên một số tài liệu còn thiếu thông tin xuất bản.	NCS xin trân trọng cảm ơn ý kiến góp ý của Hội đồng. NCS đã rà soát và chuẩn hóa toàn bộ danh mục tài liệu tham khảo, bổ sung các thông tin xuất bản còn thiếu. Bên cạnh đó, danh mục tài liệu tham khảo cũng được tinh giản từ 209 xuống còn 197 tài liệu (lược bỏ các tài liệu đã quá cũ) nhằm nâng cao tính cập nhật của luận án.	Tài liệu tham khảo (trang 143-164)

Trên đây là giải trình chi tiết của nghiên cứu sinh về việc tiếp thu, chỉnh sửa, bổ sung luận án theo các ý kiến đóng góp của các thành viên tham gia cũng như các nhà khoa học tham dự.

Xin trân trọng cảm ơn.

CÁN BỘ HƯỚNG DẪN

PGS.TS/Phạm Ngọc Hùng

NGHIÊN CỨU SINH

Phạm Thị Tố Nga

THÔNG TIN VỀ LUẬN ÁN TIẾN SĨ

1. Họ và tên nghiên cứu sinh: Phạm Thị Tố Nga..... 2. Giới tính: Nữ
3. Ngày sinh: 27/10/1978 4. Nơi sinh: Phú Thọ
5. Quyết định công nhận nghiên cứu sinh số 25/QĐ-CTSV ngày 14 tháng 1 năm 2022 của Hiệu trưởng Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.
6. Các thay đổi trong quá trình đào tạo: Không
(ghi các hình thức thay đổi và thời gian tương ứng)
7. Tên đề tài luận án: Một số phương pháp đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục
(tên luận án chính thức đề nghị đánh giá luận án)
8. Ngành đào tạo: Khoa học máy tính 9. Mã số: 9480101
10. Cán bộ hướng dẫn khoa học: *(ghi rõ chức danh khoa học, học vị, họ và tên)*
 - Hướng dẫn chính: GS.TS Nguyễn Đức Anh - Trường Đại học Đông Nam Na Uy
 - Hướng dẫn phụ: PGS.TS. Phạm Ngọc Hùng - Trường Đại học Công nghệ - ĐHQGHN
11. Tóm tắt các **kết quả mới** của luận án:
 - Mục đích và đối tượng nghiên cứu của luận án: Luận án hướng tới mục tiêu đề xuất các phương pháp đảm bảo tính công bằng cho các hệ thống học máy (ML) trong lĩnh vực giáo dục, nơi dữ liệu thường ở dạng bảng và chứa nhiều thuộc tính nhạy cảm. Trọng tâm nghiên cứu là nâng cao công bằng cho đồng thời nhiều thuộc tính nhạy cảm mà vẫn duy trì hiệu suất mô hình, đồng thời làm rõ mối quan hệ đánh đổi giữa công bằng và hiệu suất nhằm hỗ trợ lựa chọn mô hình tối ưu. Luận án tập trung phát triển ba phương pháp chính: FairEdu (loại bỏ phụ thuộc giữa dữ liệu và thuộc tính nhạy cảm), DPF (cân bằng phân phối bằng dữ liệu tổng hợp), và FairEduPlus (tích hợp hai hướng can thiệp để đảm bảo công bằng hai chiều). Ngoài ra, chỉ số đánh đổi giữa công bằng và hiệu suất cũng được đề xuất nhằm định lượng mối quan hệ này một cách hệ thống. Đối tượng nghiên cứu là các mô hình học máy ứng dụng trong giáo dục, bao gồm Logistic Regression, Decision Tree, Random Forest, Gradient Boosting và Neural Network cơ bản, được huấn luyện trên dữ liệu chứa nhiều thuộc tính nhạy cảm như giới tính, chủng tộc, tuổi, sức khỏe, tình trạng nợ và nơi ở.

- Phương pháp nghiên cứu: Luận án kết hợp phương pháp nghiên cứu lý thuyết và triển khai thực nghiệm các kết quả thu được so với các kết quả mới nhất trên thế giới. Đây là phương pháp nghiên cứu hiện đại, có độ tin cậy cao và đang được sử dụng phổ biến.

- Các kết quả chính và kết luận: Luận án đã nghiên cứu và phát triển các phương pháp đảm bảo tính công bằng cho các mô hình học máy trong giáo dục, nơi dữ liệu thường dạng bảng, chứa nhiều thuộc tính nhạy cảm và mất cân bằng. Bốn kết quả chính đạt được gồm: (1) Phương pháp FairEdu – kỹ thuật tiên xử lý dựa trên hồi quy đa biến, loại bỏ sự phụ thuộc giữa thuộc tính nhạy cảm và dữ liệu huấn luyện, giúp cải thiện công bằng khi xét đồng thời nhiều thuộc tính. (2) Phương pháp DPF (Data Partitioning Fairness) – sử dụng kỹ thuật sinh dữ liệu tổng hợp (SDG) để cân bằng phân phối giữa các nhóm giao thoa, tăng tính đại diện của dữ liệu, từ đó nâng cao tính công bằng cho các mô hình học máy ngay từ khâu tiên xử lý. (3) Phương pháp FairEduPlus – tích hợp FairEdu và DPF thành cơ chế can thiệp hai chiều, vừa hiệu chỉnh dữ liệu, vừa cân bằng phân phối, đảm bảo công bằng toàn diện mà vẫn duy trì hiệu suất ổn định. (4) Chỉ số đánh đổi – công cụ định lượng mối quan hệ giữa công bằng và hiệu suất mô hình. Kết quả thực nghiệm trên bốn bộ dữ liệu giáo dục cho thấy các phương pháp đề xuất giảm rõ rệt chênh lệch công bằng (DI, SPD, AOD, EOD) trong khi vẫn giữ hiệu suất ở mức cao. Về ý nghĩa, luận án đóng góp một khung phương pháp luận mới cho nghiên cứu công bằng trong học máy, có khả năng mở rộng sang các lĩnh vực nhạy cảm khác như y tế hoặc tuyển dụng. Tuy nhiên, các phương pháp hiện mới được kiểm chứng trên dữ liệu dạng bảng và cần được tự động hóa để áp dụng rộng rãi hơn trong tương lai.

12. Khả năng ứng dụng trong thực tiễn: *(nếu có)*

Các phương pháp đề xuất như FairEdu, DPF, và FairEduPlus đều có thể trực tiếp áp dụng cho các mô hình học máy trong lĩnh vực giáo dục có sử dụng dữ liệu đào tạo dạng bảng nhằm kiểm tra và nâng cao tính công bằng cho hệ thống.

13. Những hướng nghiên cứu tiếp theo: *(nếu có)*

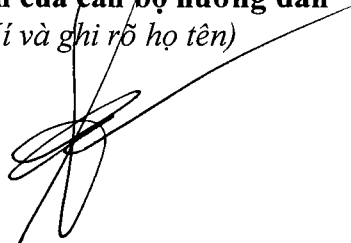
Trên cơ sở các kết quả đạt được, luận án đề xuất một số định hướng nghiên cứu tiếp theo nhằm nâng cao hiệu quả và mở rộng khả năng ứng dụng trong thực tiễn, bao gồm:

- Mở rộng phạm vi dữ liệu và mô hình, bao gồm các lĩnh vực ngoài giáo dục (y tế, tài chính) và các mô hình hiện đại như deep learning, Transformer, nhằm đánh giá khả năng tổng quát và mở rộng của phương pháp.
- Phát triển các thước đo công bằng toàn diện hơn, đặc biệt cho trường hợp đa thuộc tính nhạy cảm, đồng thời hoàn thiện chỉ số đánh đổi để hỗ trợ lựa chọn mô hình trong thực tiễn.
- Nâng cao kỹ thuật sinh dữ liệu tổng hợp cho các tình huống mất cân bằng nghiêm trọng hoặc dữ liệu thưa, nhằm cải thiện đại diện cho các nhóm giao thoa.
- Xây dựng khung công bằng tự động, có khả năng phát hiện, điều chỉnh và đánh giá công bằng trên nhiều loại dữ liệu và mô hình, giảm phụ thuộc vào can thiệp thủ công.

14. Các công trình đã công bố có liên quan đến luận án:

1. Pham, N., Pham-Ngoc, H., Nguyen-Duc, A. (2023). *Fairness Requirement in AI Engineering – A Review on Current Research and Future Directions*. In: Gupta, V., Rubalcaba, L., Gupta, C., Hanne, T. (eds) *Sustainability in Software Engineering and Business Information Management*. SSEBIM 2022. Lecture Notes in Information Systems and Organisation, vol 62. Springer, Cham. https://doi.org/10.1007/978-3-031-32436-9_1.
2. N. Pham, P. N. Hung, and A. Nguyen-Duc, *Fairness for machine learning software in education: A systematic mapping study*, *J. Syst. Softw.*, p. 112244, Oct. 2024, doi: 10.1016/j.jss.2024.112244 (Q1)
3. N. Pham, M. K. Do, T. V. Dai, P. N. Hung, and A. Nguyen-Duc, *FAIREDU: A Multiple Regression-Based Method for Enhancing Fairness in Machine Learning Models for Educational Applications*, *Expert Syst. Appl.*, Dec. 2024, doi.org/10.1016/j.eswa.2024.126219 (Q1)
4. N. Pham, Minh Kha Do, Quang Trung Doan, Pekka Abrahamsson, Anh Nguyen-Duc, Pham Ngoc Hung, *FaireduPlus: Enhancing Intersectional Fairness in Education Focused Machine Learning Using Synthetic Data*, submitted to IST (Q1)

Ngày tháng năm 2026
Xác nhận của cán bộ hướng dẫn
(Kí và ghi rõ họ tên)



Phạm Ngọc Hùng

Ngày tháng năm 2026
Nghiên cứu sinh
(Kí và ghi rõ họ tên)



Phạm Thị Tố Nga

INFORMATION ON DOCTORAL THESIS

1. Full name : Pham Thi To Nga 2. Sex: Female
3. Date of birth: 27/10/1978 4. Place of birth: Phu Tho
5. Admission decision number: 25/QĐ-CTSV. Dated 01/14/2022
6. Changes in academic process: Non
(List the forms of change and corresponding times)
7. Official thesis title: Several Methods for Ensuring Fairness in Machine Learning Systems in Education
8. Major: Computer Science..... 9. Code: 9480101
10. Supervisors:
 1. Prof. Dr. Nguyen Duc Anh – University of South-Eastern Norway
 2. Assoc. Prof. Dr. Pham Ngoc Hung – University of Engineering and Technology, VNU
11. Summary of the **new findings** of the thesis:
 - Research Objectives and Scope: The dissertation aims to propose effective methods for ensuring fairness in machine learning (ML) systems applied to educational data, which are typically tabular and contain multiple sensitive features. The focus is to enhance fairness simultaneously across several sensitive features while maintaining model performance, and to clarify the trade-off relationship between fairness and accuracy to support optimal model selection. Three main methods are developed: (1) FairEdu – a preprocessing method that removes dependencies between sensitive attributes and non-sensitive features; (2) DPF (Data Partitioning Fairness) – a data-level balancing method that uses synthetic data generation (SDG) to equalize the representation of intersectional groups; and (3) FairEduPlus – a hybrid framework integrating both approaches to achieve two-dimensional fairness (distributional and relational). In addition, (4) a trade-off index is proposed to systematically quantify the balance between fairness and predictive performance. The research objects are ML models commonly used in educational applications, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and basic Neural Networks, trained on datasets containing multiple sensitive variables such as gender, ethnicity, age, health status, debt condition, and residence.

- **Research Methods:** The dissertation combines theoretical analysis with empirical experiments, comparing the proposed methods against the most recent state-of-the-art approaches worldwide. This is a modern and reliable research methodology, widely adopted in contemporary AI and data science studies.

- **Main Results and Conclusions:** The dissertation investigates and develops fairness-enhancing methods for machine learning models applied in education, where data are tabular, imbalanced, and contain multiple sensitive features. Four major results are achieved: (1) **FairEdu Method:** A multivariate regression-based preprocessing technique that removes statistical dependencies between sensitive and non-sensitive attributes, effectively improving fairness across multiple sensitive variables. (2) **DPF (Data Partitioning Fairness):** A method leveraging synthetic data generation (SDG) to rebalance the distribution among intersectional subgroups, thereby increasing the representativeness of the training data, and consequently enhancing fairness in machine learning systems from the pre-processing stage. (3) **FairEduPlus Method:** A comprehensive, two-dimensional approach integrating FairEdu and DPF to simultaneously adjust data features and group distributions, achieving overall fairness without significant performance loss. (4) **Trade-off Index:** A new quantitative measure that evaluates the balance between model fairness and predictive accuracy. Experiments on four educational datasets demonstrate that the proposed methods substantially reduce fairness disparities (measured by DI, SPD, AOD, and EOD) while maintaining high model performance. In terms of contribution, the dissertation introduces a systematic methodological framework for fairness assessment in ML, with potential extensions to other sensitive domains such as healthcare and recruitment. However, the proposed methods have been validated primarily on tabular data and still require further automation to be applied widely in large-scale, real-world systems.

12. Practical applicability, if any:

The proposed methods, including FairEdu, DPF, and FairEduPlus, can be directly applied to machine learning models in the educational domain that use tabular training data to evaluate and enhance the fairness of the system.

13. Further research directions, if any:

Based on the achieved results, the dissertation proposes several future research directions to enhance effectiveness and expand practical applicability, including:

- Extending the scope of datasets and models, including domains beyond education (e.g., healthcare, finance) and modern architectures such as deep learning and Transformer-based models, to evaluate generalizability and scalability.
- Developing more comprehensive fairness metrics, especially for multi-sensitive attribute settings, and further refining the trade-off index to support practical model selection.

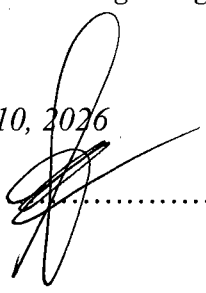
- Improving synthetic data generation techniques for highly imbalanced or sparse data scenarios, to enhance representation of intersectional subgroups.
- Designing an automated fairness framework capable of detecting, mitigating, and evaluating fairness across various data types and models, thereby reducing reliance on manual intervention.

14. Thesis-related publications:

1. Pham, N., Pham-Ngoc, H., Nguyen-Duc, A. (2023). *Fairness Requirement in AI Engineering – A Review on Current Research and Future Directions*. In: Gupta, V., Rubalcaba, L., Gupta, C., Hanne, T. (eds) *Sustainability in Software Engineering and Business Information Management*. SSEBIM 2022. Lecture Notes in Information Systems and Organisation, vol 62. Springer, Cham. https://doi.org/10.1007/978-3-031-32436-9_1.
2. N. Pham, P. N. Hung, and A. Nguyen-Duc, *Fairness for machine learning software in education: A systematic mapping study*, J. Syst. Softw., p. 112244, Oct. 2024, doi: 10.1016/j.jss.2024.112244 (Q1)
3. N. Pham, M. K. Do, T. V. Dai, P. N. Hung, and A. Nguyen-Duc, *FAIREDU: A Multiple Regression-Based Method for Enhancing Fairness in Machine Learning Models for Educational Applications*, Expert Syst. Appl., Dec. 2024, doi.org/10.1016/j.eswa.2024.126219 (Q1)
4. N. Pham, Minh Kha Do, Quang Trung Doan, Pekka Abrahamsson, Anh Nguyen-Duc, Pham Ngoc Hung, *FaireduPlus: Enhancing Intersectional Fairness in Education Focused Machine Learning Using Synthetic Data*, submitted to IST (Q1)

Date: July 10, 2026

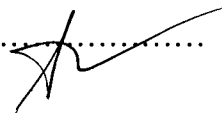
Signature:



Full name: Pham Ngoc Hung

Date: July 10, 2026

Signature:



Full name: Pham Thi To Nga

TRÍCH YẾU LUẬN ÁN

a) Tóm tắt mở đầu:

- Tên tác giả: Phạm Thị Tố Nga
- Tên luận án: Một số phương pháp đảm bảo tính công bằng cho các hệ thống học máy trong lĩnh vực giáo dục
- Ngành khoa học của luận án: Khoa học máy tính Mã số: 9480101.01
- Tên đơn vị đào tạo: Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

b) Nội dung bản trích yếu:

- Mục đích và đối tượng nghiên cứu của luận án: Luận án hướng tới mục tiêu đề xuất các phương pháp đảm bảo tính công bằng cho các hệ thống học máy (ML) trong lĩnh vực giáo dục, nơi dữ liệu thường ở dạng bảng và chứa nhiều thuộc tính nhạy cảm. Trọng tâm nghiên cứu là nâng cao công bằng cho đồng thời nhiều thuộc tính nhạy cảm mà vẫn duy trì hiệu suất mô hình, đồng thời làm rõ mối quan hệ đánh đổi giữa công bằng và hiệu suất nhằm hỗ trợ lựa chọn mô hình tối ưu. Luận án tập trung phát triển ba phương pháp chính: FairEdu (loại bỏ phụ thuộc giữa dữ liệu và thuộc tính nhạy cảm), DPF (cân bằng phân phối bằng dữ liệu tổng hợp), và FairEduPlus (tích hợp hai hướng can thiệp để đảm bảo công bằng hai chiều). Ngoài ra, chỉ số đánh đổi giữa công bằng và hiệu suất cũng được đề xuất nhằm định lượng mối quan hệ này một cách hệ thống. Đối tượng nghiên cứu là các mô hình học máy ứng dụng trong giáo dục, bao gồm Logistic Regression, Decision Tree, Random Forest, Gradient Boosting và Neural Network cơ bản, được huấn luyện trên dữ liệu chứa nhiều thuộc tính nhạy cảm như giới tính, chủng tộc, tuổi, sức khỏe, tình trạng nợ và nơi ở.

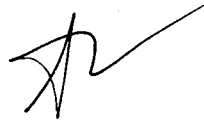
- Phương pháp nghiên cứu: Luận án kết hợp phương pháp nghiên cứu lý thuyết và triển khai thực nghiệm các kết quả thu được so với các kết quả mới nhất trên thế giới. Đây là phương pháp nghiên cứu hiện đại, có độ tin cậy cao và đang được sử dụng phổ biến.

- Các kết quả chính và kết luận: Luận án đã nghiên cứu và phát triển các phương pháp đảm bảo tính công bằng cho các mô hình học máy trong giáo dục, nơi dữ liệu thường dạng bảng, chứa nhiều thuộc tính nhạy cảm và mất cân bằng. Bốn kết quả chính đạt được gồm: (1) Phương pháp FairEdu – kỹ thuật tiền xử lý dựa trên hồi quy đa biến, loại bỏ sự phụ thuộc giữa thuộc tính nhạy cảm và dữ liệu huấn luyện, giúp cải thiện công bằng khi xét đồng thời nhiều thuộc tính. (2) Phương pháp DPF (Data Partitioning Fairness) – sử dụng kỹ thuật sinh dữ liệu tổng hợp (SDG) để cân bằng phân phối giữa các nhóm giao thoa, tăng tính đại diện của dữ liệu. (3) Phương pháp FairEduPlus – tích hợp FairEdu và DPF thành cơ chế can thiệp hai chiều, vừa hiệu chỉnh dữ liệu, vừa cân bằng phân phối, đảm bảo công bằng toàn diện mà vẫn duy trì hiệu suất ổn định. (4) Chỉ số đánh đổi – công cụ định lượng mối quan hệ giữa công bằng và hiệu suất mô hình. Kết quả thực nghiệm trên bốn bộ dữ liệu giáo dục cho thấy các phương pháp đề xuất giảm rõ rệt chênh lệch công bằng (DI, SPD, AOD, EOD) trong khi vẫn giữ hiệu suất ở mức cao. Về ý nghĩa, luận án đóng góp

một khung phương pháp luận mới cho nghiên cứu công bằng trong học máy, có khả năng mở rộng sang các lĩnh vực nhạy cảm khác như y tế hoặc tuyển dụng. Tuy nhiên, các phương pháp hiện mới được kiểm chứng trên dữ liệu dạng bảng và cần được tự động hóa để áp dụng rộng rãi hơn trong tương lai.

Hà Nội, ngày 10 tháng 7 năm 2026

NGHIÊN CỨU SINH
(Ký và ghi rõ họ, tên)



Phạm Thị Tố Nga

CÁN BỘ HƯỚNG DẪN
(Ký và ghi rõ họ, tên)



Phạm Ngọc Hùng

**XÁC NHẬN CỦA CƠ SỞ ĐÀO TẠO/
CONFIRMATION FROM THE TRAINING UNIVERSITY**

THE ABSTRACT OF DOCTORAL THESIS

a) Preliminary Information

- Author: Pham Thi To Nga
- Title of the doctoral thesis: Several Methods for Ensuring Fairness in Machine Learning Systems in Education
- Major: Computer Science
- Code: 9480101.01
- Training Institution: University of Engineering and Technology, Vietnam National University, Hanoi

b) Content Summary

- **Research Objectives and Scope:** The dissertation aims to propose effective methods for ensuring fairness in machine learning (ML) systems applied to educational data, which are typically tabular and contain multiple sensitive features. The focus is to enhance fairness simultaneously across several sensitive features while maintaining model performance, and to clarify the trade-off relationship between fairness and accuracy to support optimal model selection. Three main methods are developed: (1) FairEdu – a preprocessing method that removes dependencies between sensitive attributes and non-sensitive features; (2) DPF (Data Partitioning Fairness) – a data-level balancing method that uses synthetic data generation (SDG) to equalize the representation of intersectional groups; and (3) FairEduPlus – a hybrid framework integrating both approaches to achieve two-dimensional fairness (distributional and relational). In addition, (4) a trade-off index is proposed to systematically quantify the balance between fairness and predictive performance. The research objects are ML models commonly used in educational applications, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and basic Neural Networks, trained on datasets containing multiple sensitive variables such as gender, ethnicity, age, health status, debt condition, and residence.

- **Research Methods:** The dissertation combines theoretical analysis with empirical experiments, comparing the proposed methods against the most recent state-of-the-art approaches worldwide. This is a modern and reliable research methodology, widely adopted in contemporary AI and data science studies.

- **Main Results and Conclusions:** The dissertation investigates and develops fairness-enhancing methods for machine learning models applied in education, where data are tabular, imbalanced, and contain multiple sensitive features. Four major results are achieved: (1) FairEdu Method: A multivariate regression-based preprocessing technique that removes statistical dependencies between sensitive and non-sensitive attributes, effectively improving fairness across multiple sensitive variables. (2) DPF (Data Partitioning Fairness): A method leveraging synthetic data generation (SDG) to rebalance the distribution among intersectional subgroups, thereby increasing the representativeness

of the training data. (3) FairEduPlus Method: A comprehensive, two-dimensional approach integrating FairEdu and DPF to simultaneously adjust data features and group distributions, achieving overall fairness without significant performance loss. (4) Trade-off Index: A new quantitative measure that evaluates the balance between model fairness and predictive accuracy. Experiments on four educational datasets demonstrate that the proposed methods substantially reduce fairness disparities (measured by DI, SPD, AOD, and EOD) while maintaining high model performance. In terms of contribution, the dissertation introduces a systematic methodological framework for fairness assessment in ML, with potential extensions to other sensitive domains such as healthcare and recruitment. However, the proposed methods have been validated primarily on tabular data and still require further automation to be applied widely in large-scale, real-world systems.

Hanoi, July 10, 2026

PHD STUDENT

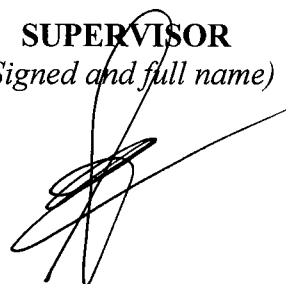
(Signed and full name)



Pham Thi To Nga

SUPERVISOR

(Signed and full name)



Pham Ngoc Hung

CONFIRMATION FROM THE TRAINING UNIVERSITY