EARLY RISK WARNING IN A SOFT SKILLS COURSE WITH CALIBRATED MULTICLASS XGBOOST: A PREDICT → TIER → INTERVENE FRAMEWORK

Tang Thi Vinh*[1], Tran Van San[2]

* Corresponding author:
Email: tranvansan@hitu.edu.vn

[1] Email:tangthivinh@hitu.edu.vn
[1,2] Ho Chi Minh City Industry and Trade College
20 Tang Nhon Phu street, Phuoc Long ward,
Ho Chi Minh City, Vietnam

Received: 21/10/2025
Revised: 19/11/2025
Accepted: 05/12/2025
Published: 20/02/2026

Abstract: We propose a three-step *Predict* → *Tier* → *Intervene* framework for a college Soft Skills course. Weekly T1–T7 activity traces, attendance/bonus, mini-tasks, and peer/self reports are standardized into individual–and group–level features. A multiclass XGBoost model with *isotonic calibration* and *group-aware* splits yields well–calibrated probabilities that are mapped to RED/YELLOW/GREEN alerts and class summaries of tier mix and at-risk students/groups. On a held-out semester, the system achieves *Accuracy = 0.772, Macro–F1 = 0.520, AUPRC(C) = 0.739, Brier = 0.1008, ECE = 0.0577*. The model clearly separates high from average achievers and–despite the rarity of tier C–provides effective risk ranking in precision–recall analysis. The pipeline is low–cost (Google Forms/Sheets plus concise Python), transparent, and reproducible, supporting timely, tiered interventions.

Keywords: *Educational Data Mining; early warning; tiered instruction; soft skills; probability calibration; XGBoost.*

CẢNH BÁO SỚM HỌC LỰC TRONG HỌC PHẦN KĨ NĂNG MỀM BẰNG XGBOOST ĐA LỚP HIỆU CHỈNH XÁC SUẤT: KHUNG DỰ BÁO → PHÂN TẦNG → CAN THIỆP

Tăng Thị Vinh*[1], Trần Văn San[2]

* Tác giả liên hệ:
Email: tranvansan@hitu.edu.vn

[2] Email: tranvansan@hitu.edu.vn
[1,2] Trường Cao đẳng Công thương
Thành phố Hồ Chí Minh,
20 Tăng Nhơn Phú, phường Phước Long,
Thành phố Hồ Chí Minh, Việt Nam

Nhận bài: 21/10/2025
Chỉnh sửa xong: 19/11/2025
Chấp nhận đăng: 05/12/2025
Xuất bản: 20/02/2026

Tóm tắt: Tác giả đề xuất khung Dự báo → Phân tầng → Can thiệp cho học phần Kĩ năng mềm. Nhật kí hoạt động tuần T1-T7, điểm chuyên cần/thưởng, bài tập ngắn và báo cáo đồng học/tự đánh giá được chuẩn hoá thành đặc trưng ở mức cá nhân và nhóm. Mô hình XGBoost đa lớp kèm hiệu chỉnh isotonic và chia tách group–aware tạo xác suất tin cậy, từ đó suy ra cảnh báo ĐỎ/VÀNG/XANH và tổng hợp lớp/nhóm (phối trộn tầng, danh sách rủi ro). Trên một học kì giữ lại để kiểm định, hệ thống đạt Accuracy = 0.772, Macro-F1 = 0.520, AUPRC(C) = 0.739, Brier = 0.1008, ECE = 0.0577; mô hình tách rõ nhóm thành tích cao so với trung bình và dù lớp C hiếm vẫn xếp hạng rủi ro hiệu quả theo precision - recall. Quy trình chi phí thấp (Google Forms/Sheets thêm ít mã Python), minh bạch, dễ tái lập và hỗ trợ can thiệp phân tầng kịp thời.

Từ khóa: *Khai phá dữ liệu giáo dục (EDM), cảnh báo sớm, dạy học phân tầng, kĩ năng mềm, hiệu chỉnh xác suất, XGBoost.*

1. Đặt vấn đề

Bối cảnh Việt Nam cho thấy nhu cầu cấp thiết: khoảng 80% sinh viên đại học, cao đẳng thiếu hụt kĩ năng mềm (ước chừng 1,6-2 triệu người cần can thiệp sớm); trong đó 78% tự nhận yếu giao tiếp, 48% thiếu sáng tạo, 35% yếu làm việc nhóm và 21% gặp khó khăn về quản lí thời gian cũng như tương tác xã hội (Tạp chí Giáo dục Thành phố Hồ Chí Minh, 2024). Từ bức tranh quy mô lớn này, yêu cầu trọng tâm không phải "can thiệp nhiều hơn" mà là can thiệp đúng lúc, đúng đối tượng dựa trên bằng chứng chuyển các dữ liệu hành vi học tập thành xác suất rủi ro đã hiệu chỉnh rồi ánh xạ thành mức ưu tiên (ĐỎ/VÀNG/XANH) dưới ràng buộc ngân sách để phân bổ nguồn lực sư phạm một cách minh bạch và hiệu quả.

Khai phá Dữ liệu Giáo dục (Educational Data Mining - EDM) trích xuất tri thức từ dữ liệu hành vi - tương tác của người học để hỗ trợ quyết định sư phạm dựa trên bằng chứng (Baker & Yacef, 2009; Romero & Ventura, 2010). Ứng dụng trọng tâm là dự báo sớm nhằm kích hoạt can thiệp kịp thời (cá nhân hóa, cố vấn, điều chỉnh hoạt động), dựa trên nguồn

dữ liệu LMS, điểm số và nhật kí hoạt động (Romero & Ventura, 2010). Trên dữ liệu Moodle đa học kì, Angeioplastis các cộng sự (2025) so sánh k-NN, cây quyết định, RF, hồi quy logistic và mạng nơ-ron ở cả thiết lập nhị phân và đa lớp. Kết quả nhất quán với thực tiễn EDM: 1) Thiết lập nhị phân ổn định hơn đa lớp; 2) Sàng lọc đặc trưng theo tương quan/cấu trúc cải thiện hiệu năng rõ rệt; 3) Cây quyết định và k-NN bền vững ở dự báo giữa kì (Romero & Ventura, 2010; Angeioplastis và các cộng sự, 2025). Về định tính, RF/GBDT dẫn đầu trên dữ liệu bảng nhỏ - trung bình; SVM/k-NN theo sát, còn Naive Bayes kém ổn định khi mất cân bằng lớp (Romero & Ventura, 2010).

Dữ liệu lớp học thường ở dạng "bảng", có quy mô nhỏ - trung bình, khuyết giá trị, mất cân bằng lớp và trôi theo thời gian. Vì vậy, tập đặc trưng tối ưu không cố định (Romero & Ventura, 2010). Khung Tiến hóa Tập hợp Đặc trưng Động (Dynamic Feature Ensemble Evolution for Enhanced Feature Selection, DE - FS) kết hợp nhiều tiêu chí (tương quan/MI/χ^2/embedded) với ngưỡng thích nghi để duy trì độ chính xác và giảm dư thừa khi phân phối thay đổi (Malik và các cộng sự, 2025). Hướng này đặc biệt phù hợp bối cảnh kĩ năng mềm, nơi hành vi - tương tác biến thiên theo giai đoạn và nhóm. Bên cạnh RF/XGBoost/CatBoost, các mô hình nền tảng cho dữ liệu bảng (Tabular Foundation Models) cho thấy tiềm năng trên dữ liệu nhỏ - trung bình: TabPFN là một "thuật toán dự báo tổng quát" học qua tiền huấn luyện quy mô lớn, cho suy luận nhanh và ít phụ thuộc tinh chỉnh (Hollmann và các cộng sự, 2025), còn Real-TabPFN tiếp tục nâng hiệu năng bằng tiền huấn luyện trên dữ liệu thực tế tuyển chọn (Garg và các cộng sự, 2025). Xu hướng chuyển từ thủ công hoá đặc trưng sang học biểu diễn tự động cũng được ghi nhận (Jiang và các cộng sự, 2025).

Phần lớn công trình EDM dừng ở tối đa hoá độ chính xác dự báo, trong khi việc gắn dự báo với hành động sư phạm còn hạn chế (Baker & Yacef, 2009). Với kĩ năng mềm (giao tiếp, hợp tác, giải quyết xung đột, ra quyết định), mục tiêu đa chiều và tín hiệu giàu ngữ cảnh (LMS, hoạt động trên lớp, nộp thu hoạch, tự/đồng học báo cáo) đòi hỏi cơ chế chuyển hoá dự báo thành thiết kế can thiệp cụ thể. Công trình này kế thừa quy trình mô hình hoá trên dữ liệu lớp học (Angeioplastis và các cộng sự, 2025), tích hợp DE-FS (Malik và các cộng sự, 2025) để thích nghi đặc trưng theo thời gian, áp dụng đánh giá group-aware (trên khoá LỚP/NHÓM) và hiệu chỉnh xác suất (Platt/Isotonic) nhằm đảm bảo xác suất dự báo dùng được trong lớp học.

Mục tiêu, khoảng trống và đóng góp:

Khoảng trống. Nghiên cứu EDM hiện chủ yếu tối đa hóa độ chính xác dự báo, nhưng còn thiếu: 1) Cơ chế chuyển xác suất → hành động sư phạm dưới ràng buộc ngu ồn lực; 2) Đánh giá không rò rỉ bối cảnh bằng chia tách theo LỚP/NHÓM; 3) Hiệu chỉnh xác suất đa lớp để đặt ngưỡng cảnh báo khi lớp C hiếm.

Câu hỏi nghiên cứu:

1) XGBoost đa lớp kết hợp isotonic calibration và group-aware split có tạo xác suất đáng tin (Brier/ECE thấp) cho vận hành giữa kì/cuối kì?

2) Xác suất đã hiệu chỉnh có cải thiện *xếp hạng rủi ro lớp C* (AUPRC) trong bối cảnh lớp hiếm?

3) Có thể ánh xạ xác suất → RED/YELLOW/GREEN bằng ngưỡng (τ_{yel}, τ_{red}) hoặc top-k theo ρ_{max} để can thiệp "đúng lúc, đúng đối tượng"?

Khung tiếp cận. Áp dụng khung 4 lớp (dữ liệu → đầu vào → mô hình → ứng dụng) cho dự báo sớm và can thiệp phân tầng (chi tiết ở Mục 2).

Đóng góp:

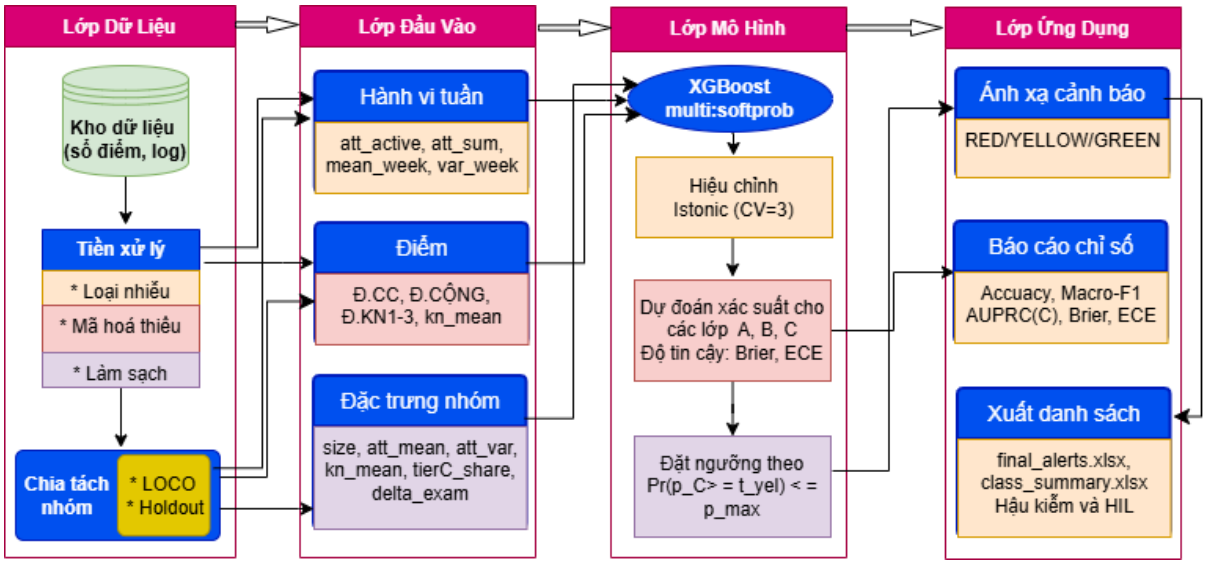
Quy trình hoàn chỉnh và kiểm soát rò rỉ: chuẩn hoá cột; mã hoá T1 đếnT7 thành đặc trưng hành vi {att_active, att_sum, mean_week, var_week}; bổ sung kn_mean, group_*, tierC_share; tách *group-aware* theo LỚP/NHÓM, chỉ dùng thông tin trước mốc t^*.

Mô hình hoá và đóng gói suy luận: XGBoost đa lớp (multi: softprob) với hiệu chỉnh xác suất isotonic (CalibratedClassifierCV, *cv*=3), hai cấu hình đặc trưng theo giai đoạn F_{mid} (không dùng Đ.QT) và F_{full} = $F_{mid} \cup$ {Đ.QT}, mô hình được huấn luyện trên bảng dữ liệu 869 dòng 32 cột và *đóng gói* (model/encoder/feature list) cho vận hành.

Kết quả và chuyển hoá thành hành động: đánh giá thực nghiệm trên bảng dữ liệu 80 hàng và 20 cột của 3 lớp gồm 26 nhóm đạt Accuracy = 0.772, Macro-F1 = 0.520, AUPRC(C) = 0.739, Brier = 0.1008, ECE = 0.0577; đầu ra các file bảng *class_summary.xlsx* và *final_alerts.xlsx* ưu tiên *who/where* với các mức RED/YELLOW/ GREEN, hỗ trợ *giảng dạy phân tầng* trong bối cảnh kĩ năng mềm.

Cách tiếp cận này đưa EDM từ dự báo để biết sang dự báo để hành động, với chi phí triển khai thấp (Python) và khả năng tái lập cho các học phần tương tự.

 https://doi.org/10.15625/2615-8957/12610202



Hình 1: *Quy trình Dự báo sớm - Phân tầng - Can thiệp (4 lớp).*

2. Phương pháp nghiên cứu

Mục tiêu là chuyển dự báo sớm thành hành động sư phạm trong học phần Kĩ năng mềm. Chúng tôi đề xuất quy trình gồm 4 lớp: 1) Dữ liệu: làm sạch; Group-aware split (LỚP/NHÓM). 2) Đầu vào: đặc trưng T1-T7, điểm, đặc trưng nhóm. 3) Mô hình: XGBoost + isotonic (CV=3); ngưỡng ρ_max; RED/YELLOW/GREEN. 4) Ứng dụng: xuất danh sách, hậu kiểm, báo cáo chỉ số, HIL.

2.1. Các kí hiệu, độ đo và các định nghĩa

Kí hiệu chung: $D=\{(x_i,y_i)\}_{i=1}^{n}$, với nhãn $y_i \in Y=\{A,B,C\}$. Mô hình cho phân phối $p_i(c)=\Pr(Y=c\mid x_i)$ với $c\in Y$, $\sum_c P_i(c)=1$. Dự đoán nhãn: $\hat{y}_i=\arg\max_{c\in y} p_i(c)$. Lớp dương khi đánh giá PR: c* = C (lớp nguy cơ). Hàm chỉ thị 1[·] bằng 1 nếu mệnh đề đúng, ngược lại bằng 0.

Các độ đo phân loại Precision, Recall, F1 Macro-F1, Weighted-F1, Accuracy theo từng lớp: Với một lớp $c\in Y$,

$$TP_c=\sum_{i=1}^{n}1[y_i=c\wedge\hat{y}_t=c],\ FP_c=\sum_{i=1}^{n}1[y_i\neq c\wedge\hat{y}_i=c],$$

$$FN_c=\sum_{i=1}^{n}1[y_i=c\wedge\hat{y}_i\neq c]$$

$$Precision_c=\frac{TP_c}{TP_c+FP_c},\ Pecall_c=\frac{TP_c}{TP_c+FN_c},$$

$$F1_c=\frac{2Precision_cRecall_c}{Precision_c+Recall_c}.$$

$$Marco-F1=\frac{1}{|y|}\sum_{c\in Y}F1_c,$$

$$Weighted-F1=\sum_{c\in Y}\frac{N_c}{n}F1_c,$$

$$Accuracy=\frac{1}{n}\sum_{i=1}^{n}1[y_i=\hat{y}_i],$$

trong đó $N_c=\sum_i 1[y_i=c]$ là *support* của lớp c.

Độ đo dựa trên xác suất Brier đa lớp (Brier, 1950)*:* Đặt vecto one-hot $e(y_i)\in\{0,1\}^{|Y|}$ với $e(y_i)_c=1[y_i=c]$. Brier score (trung bình MSE) cho đa lớp:

$$Brier=\frac{1}{n}\sum_{i=1}^{n}\sum_{c\in Y}(e(y_i)_c-p_i(c))^2$$

Độ đo theo xác suất cực đại ECE (Guo và các cộng sự, 2017)*:* Gọi $q_i=\max_c p_i(c)$ là độ tin cậy của dự đoán, và $correct_i=1[y_i=\hat{y}_i]$. Chia [0,1] thành M khoảng $(\beta_{m-1},\beta_m]$. Với mỗi bin $B_m=\{i:q_i\in(\beta_{m-1},\beta_m]\}$,

$$acc(B_m)=\frac{1}{|B_m|}\sum_{i\in B_m}correct_i,\quad conf(B_m)=\frac{1}{|B_m|}\sum_{i\in B_m}q_i,$$

$$ECE=\sum_{m=1}^{M}\frac{|B_m|}{n}|acc(B_m)-conf(B_m)|.$$

Precision–Recall và AUPRC cho lớp C: Đặt $z_i = 1[y_i = C]$ và điểm xếp hạng $s_i = p_i(C)$. Khi quét ngưỡng τ từ 1 xuống 0:

$$\hat{z}_i(\tau)=1[s_i\geq\tau],\ Precision(\tau)=\frac{\sum_i 1[\hat{z}_i(\tau)=1\wedge z_i=1]}{\sum_i 1[\hat{z}_i(\tau)=1]},$$

$$Recall(\tau)=\frac{\sum_i 1[\hat{z}_i(\tau)=1\wedge z_i=1]}{\sum_i 1[z_i=1]},$$

Đường cong PR là tập các điểm (Recall(τ), Precision(τ)). *Average Precision* (xấp xỉ AUPRC) là tổng có trọng số theo ΔRecall sau khi sắp s_i giảm dần:

$$AP = \sum_{k=1}^{n} (Recall_k - Recall_{k-1}) Precision_k$$

Baseline kì vọng của AUPRC bằng *prevalence* của lớp C:

$$AUPRC_{baseline} = \frac{1}{n} \sum_{i=1}^{n} 1[y_i = C] = \frac{N_c}{n}$$

Ma trận nhầm lẫn (Confusion Matrix): Với thứ tự lớp cố định y = (A,B,C) hoặc (C,B,A), ma trận đếm $M \in \mathbb{N}^{|y| \times |y|}$:

$$M_{r,c} = \sum_{i=1}^{n} 1[y_i = r \wedge \hat{y}_i = c], \quad r, c \in Y$$

Tỉ lệ theo hàng (dễ so sánh giữa các lớp thật):

$$\tilde{M}_{r,c} = \frac{M_{r,c}}{\sum_{c'} M_{r,c'}} \quad \text{(quy ước } 0/0 = 0)$$

2.2. *Tiền xử lí dữ liệu*

Nguồn dữ liệu: Bộ dữ liệu tổng hợp từ sổ điểm và bản ghi (log) hoạt động; mỗi dòng là một cặp *sinh viên–lớp–nhóm*. Tập hiện có 869 bản ghi, 20 cột: LỚP, NHÓM, MSSV, LỚP CN; bảy cột từ tuần T1 đến tuần T7; các cột điểm Đ.CC, Đ.CỘNG, Đ.KN1 đến Đ.KN3, Đ.QT, Đ.THI, Đ.MH. Có 134 nhóm theo khoá (LỚP, NHÓM) (xem Bảng 1).

Mã hoá hành vi theo tuần: Từ tuần 1 đến tuần 7 là T1: T7 và được chuẩn hoá: {1, . . . , 7} giữ nguyên; P → 0 (vắng có phép); T→ –1 (đi trễ); V→ –3 (vắng không phép); –1, –2 giữ nguyên; rỗng→NaN. Hai đặc trưng cơ bản được sử dụng để định lượng sự tham gia của sinh viên trong 7 tuần là *att_active* là số tuần hoạt động (đếm số tuần có điểm tương tác dương) dùng để phản ánh tần suất/sự hiện diện tối thiểu. Trong khi đó, *att_sum* là tổng điểm hoạt động hoặc hiện diện (tổng điểm tích lũy) dùng để đo lường tổng khối lượng tham gia hoặc đóng góp của sinh viên:

$$att\_active = \sum_{w=1}^{7} 1[T_w > 0], \quad att\_sum = \sum_{w=1}^{7} T_w \qquad (1)$$

Hai đặc trưng *mean_week* (trung bình hoạt động hàng tuần) và *var_week* (phương sai hoạt động hàng tuần) được tạo ra để đánh giá tính ổn định và nhất quán trong hành vi của sinh viên qua 7 tuần:

$$mean\_week = mean(T_{1:7}), \quad var\_week = var(T_{1:7}) \quad (2)$$

Khi tổng hợp, NaN ở T1:T7 được coi như 0 để phản ánh không hoạt động.

Xử lí điểm và dán nhãn: Đ.KN1: Đ.KN3, Đ.CC, Đ.CỘNG, Đ.QT, Đ.THI, Đ.MH được ép kiểu số; đặc trưng *kn_mean* = *mean*(Đ.KN1:Đ.KN3). *Tier3_true* là 3 cấp độ A/B/C suy từ Đ.MH và chỉ dùng cho *hậu kiểm*.

Bảng 1: *Tập dữ liệu gốc gồm 869 dòng × 20 cột, được trích tiêu biểu.*

STT	LỚP	NHÓM	MSSV	T1	...	T7	Đ.CC	Đ.CỘNG	Đ.KN3	Đ.QT	Đ.THI	Đ.MH
1	22203917	2317-N01	2123170302	2	…	1	10	3.5	6.5	9.2	8.0	8.5
2	22203917	2317-N01	2123170654	3	…	1	10	5.5	6.5	9.3	8.5	8.8
3	22203917	2317-N01	2123170316	2	…	1	9	4.0	6.5	8.6	7.0	7.6
4	22203917	2317-N01	2123170305	2	…	1	7	4.0	6.5	8.2	7.0	7.5
5	22203917	2317-N01	2123170301	2	…	1	10	3.5	6.5	8.8	7.5	8.0
…	…	…	…	…	…	…	…	…	…	…	…	…
865	22204260	4260-N06	2124270081	2	…	P	7	1.5	6.5	7.3	7.0	7.1
866	22204260	4260-N06	2124270110	4	…	NaN	7	3.5	8.0	8.3	7.5	7.8
867	22204260	4260-N06	2124270113	2	…	NaN	10	2.5	9.0	9.1	8.3	8.6
868	22204260	4260-N06	2124270078	2	…	NaN	10	3.5	6.3	8.7	8.3	8.5
869	22204260	4260-N06	2124270087	V	…	NaN	4	1.5	6.5	6.7	7.0	6.9

(Ghi chú: NaN = Thiếu dữ liệu; V = Vắng không phép; P = Vắng có phép)

 https://doi.org/10.15625/2615-8957/12610202

Đặc trưng nhóm: Trên khoá (LỚP, NHÓM) các tham số chính được tính toán bao gồm: quy mô nhóm là *group_size*, trung bình tổng hoạt động là *group_att_mean = mean(att_sum)*; phương sai hoạt động, phản ánh sự đồng đều là *group_att_var = var(att_sum)* (điền 0 nếu nhóm đơn); trung bình kĩ năng của nhóm là *group_kn_mean = mean(kn_mean)*, tỉ lệ sinh viên có nguy cơ thực tế (Lớp C) trong nhóm là *tierC_share = Pr(Tier3_true = C)*; còn tham số *delta_exam = Đ.THI – Đ.QT* là một đặc trưng hành vi quan trọng nó phản ánh sự chênh lệch về hiệu suất của sinh viên giữa quá trình học tập (quá trình) và thời điểm kiểm tra cuối cùng (thi), sau đó nối ngược về từng sinh viên (xem Bảng 2).

2.3. Cấu hình đặc trưng

Chúng tôi tuân thủ nghiêm ngặt nguyên tắc không rò rỉ theo mốc thời gian t*: mô hình chỉ dùng thông tin quan sát được trước (hoặc tại) t*. Hai cấu hình đặc trưng:

Fmid={att_active,att_sum,mean_week,Đ.CC, Đ.CỘNG,Đ.KN1:Đ.KN3,kn_mean,group_size, group_att_mean,group_att_var,group_kn_mean, tierC_share}

$F_{full} = F_{mid} \cup \{Đ.QT\}$ (*Đ.QT* chỉ dùng khi đã chốt).

Trong đó, các đặc trưng cấp nhóm bao gồm: *group_size*, *group_att_mean*, *group_att_var*, *group_kn_mean*, và *tierC_share* (tỉ lệ hạng C trong nhóm).

2.4. Chia dữ liệu và kiểm soát rò rỉ

Để đảm bảo ước lượng và đánh giá độc lập theo bối cảnh giảng dạy, mọi phép chia dữ liệu đều giữ nguyên biên giới nhóm theo LỚP (hoặc LỚP/NHÓM) để bảo toàn độc lập bối cảnh giảng dạy; tức là các bản ghi cùng một khóa nhóm LỚP hoặc LỚP/NHÓM không bao giờ bị tách sang hai phía huấn luyện (Train) và kiểm thử (Test) trong cùng một lần chia. Chúng tôi sử dụng GroupKFold và GroupShuffleSplit (Pedregosa và các cộng sự, 2011) với khóa nhóm LỚP (hoặc LỚP/NHÓM). Ở thiết lập GroupKFold (OOF/LOCO), tập các nhóm duy nhất được chia thành K phần rời nhau theo nhóm; mỗi vòng "bỏ một nhóm" làm tập đánh giá (Validation) và huấn luyện trên phần còn lại, sau đó ghép dự báo out-of-fold (OOF) từ tất cả các vòng để tính chỉ số– tương đương Leave-One-Group-Out (LOCO) theo lớp/nhóm. Với GroupShuffleSplit (holdout 80/20), chúng tôi thực hiện một lần tách huấn luyện/kiểm thử = 80/20 theo nhóm, bảo toàn biên giới LỚP (hoặc LỚP/NHÓM), phù hợp khi cần một phép đo "giữ-lại" (Holdout) đơn giản nhưng vẫn tránh rò rỉ bối cảnh. Khi lớp C hiếm, cần kiểm tra điều kiện $\exists i : y_i = C$ trong tập kiểm thử của phép holdout; nếu điều kiện không thỏa, thước đo AUPRC(C) trên holdout không còn ý nghĩa và khi đó nên báo cáo theo

Bảng 2: *Tập dữ liệu sau xử lí gồm 869 dòng × 32 cột; được trích tiêu biểu.*

STT	att_ active	att_ sum	mean_ week	var_week	kn_mean	...	delta_ exam	Tier3_ true	group_ size	group_att_ var	tierC_ share
1	5	7.0	1.4	0.3	7.833333	…	-1.2	A	5	4.8	0.0
2	6	11.0	1.833333	0.966667	7.333333	…	-0.8	A	5	4.8	0.0
3	5	8.0	1.6	0.3	7.166667	…	-1.6	B	5	4.8	0.0
4	5	5.0	0.833333	4.166667	7.166667	…	-1.2	B	5	4.8	0.0
5	5	7.0	1.4	0.3	7.333333	…	-1.3	A	5	4.8	0.0
…	…	…	…	…	…	…	…	…	…	…	…
865	2	0.0	0.0	7.0	6.833333	…	-0.3	B	7	15.809524	0.0
866	3	4.0	1.0	8.666667	7.5	…	-0.8	B	7	15.809524	0.0
867	5	7.0	1.4	0.3	7.266667	…	-0.4	A	7	15.809524	0.0
868	4	5.0	1.25	0.25	8.0	…	-0.8	A	7	15.809524	0.0
869	3	-3.0	-0.6	4.8	6.833333	…	0.3	B	7	15.809524	0.0

(Ghi chú: Tier3_true: A (≥ 8.0), B (6.5–7.9), C (< 6.5))

OOF/LOCO. Cuối cùng, để chống rò rỉ, mọi bước chuẩn hóa/biến đổi/chọn đặc trưng và hiệu chỉnh xác suất (Calibrated probability) đều được ước lượng (Fit) trên tập huấn luyện của từng lần chia (Fold/Split), sau đó mới áp dụng biến đổi (Transform) lên tập đánh giá hoặc tập kiểm thử tương ứng.

2.5. Mô hình hoá, hiệu chỉnh xác suất và chọn ngưỡng cảnh báo:

Mô hình. Sử dụng *XGBoost đa lớp* (Chen & Guestrin, 2016) với tham số *objective* = *multi:softprob* để dự báo xác suất $p_i(c)$ cho ba hạng $c \in \{A,B,C\}$ trên các đặc trưng đã nêu (chia tách group-aware theo *LỚP/NHÓM*).

Hiệu chỉnh xác suất. Áp dụng *Isotonic Calibration* (Zadrozny & Elkan, 2002), một phương pháp phi tham số, để ước lượng hàm đơn điệu g* (theo hồi quy isotonic) để ánh xạ điểm đầu vào s_i (logit hoặc xác suất thô) sang xác suất đã hiệu chỉnh:

$$p_i^{cal} = g^*(s_i),\ \ g^* = \arg\min_{g\,\text{đơn điệu}} \sum_{i=1}^{n}(z_i - g(s_i))^2$$

trong đó z_i là nhãn nhị phân ở bài toán một-chống-phần-còn-lại (hoặc thực hiện theo từng lớp/one-vs-rest trong đa lớp). Hàm g* giúp giảm sai lệch hiệu chỉnh, thể hiện qua ECE và Brier thấp hơn. Triển khai bằng CalibratedClassifierCV(cv=3, method="isotonic"). Lựa chọn này giúp cải thiện các thước đo dựa trên xác suất (Brier, ECE) và tạo đầu ra xác suất ổn định để đặt ngưỡng cảnh báo. Do lớp C hiếm, chúng tôi ưu tiên báo cáo AUPRC(C) thay vì AUC-ROC; các kết quả đối chiếu giữa OOF/LOCO và Holdout được tóm tắt ở Bảng 3.

Phân tầng cảnh báo. Từ xác suất đã hiệu chỉnh $p_i^{cal}(C)$, định nghĩa ba mức: *RED* (Đỏ) là mức nguy cơ cao cần can thiệp khẩn/cá nhân hoá, *YELLOW* (Vàng) là mức cảnh báo sớm, theo dõi chặt chẽ + can thiệp nhẹ, còn *GREEN* (Xanh) là mức đạt yêu cầu, không cần can thiệp. Việc phân tầng dựa trên cặp ngưỡng (τ_{yel},τ_{red}) và đặc trưng tham gia *att_active* (số tuần sinh viên có hoạt động tích cực). Cụ thể: ngưỡng vàng τ_{yel} là mức xác suất từ đó trở lên sẽ kích hoạt cảnh báo YELLOW, còn τ_{red} ngưỡng đỏ là mức xác suất cao hơn để kích hoạt RED. Các ngưỡng được chọn dưới ràng buộc ngân sách can thiệp ρ_{max} (tỉ lệ tối đa sinh viên bị gắn cờ):

$\Pr(p_C^{cal} \geq \tau_{red}) \leq \rho_{max}$ trên OOF/LOCO.

Quy tắc ánh xạ mẫu:

1) RED nếu $p_C^{cal}(C) \geq \tau_{red}$ và *att_active_i* = *Tue.*

2) YELLOW nếu $\tau_{yel} \leq p_i^{cal}(C) < \tau_{red}$.

3) GREEN nếu $p_i^{cal}(C) < \tau_{yel}$.

Có thể hiệu chỉnh ngưỡng theo ρ_{max}: chọn τ_{red} sao

Bảng 3: *Tóm tắt chỉ số huấn luyện và phân loại theo lớp.*

Thiết lập	Macro-F1	AUPRC(C)	Brier	ECE
OOF/LOCO (Toàn cục)	0.803	0.872	0.083	0.016
OOF/LOCO (Mean fold)	0.758	0.868	---	---
Apparent (Train-on -train)	0.918	1.0	0.051	0.072
Phân loại theo lớp (Apparent)				
Lớp	Precision	Recall	F1-score	Support
A	0.892	0.935	0.913	479
B	0.91	0.857	0.883	378
C	1.0	0.917	0.957	12
Accuracy	0.901			
Macro avg	0.934	0.903	0.918	869
Weighted avg	0.902	0.901	0.901	869

(Ghi chú: OOF/LOCO tránh rò rỉ; Apparent thường lạc quan; AUPRC(C) đo lớp hiếm; PR(C) dùng OOF/LOCO nếu holdout thiếu lớp C).

 https://doi.org/10.15625/2615-8957/12610202

cho Có thể hiệu chỉnh ngưỡng theo ngân sách can thiệp ρ_{max}: chọn τ_{red} sao cho

$$\frac{1}{n}\sum_{i=1}^{n} 1[p_i^{cal}(C) \geq \tau_{yel}] \leq \rho_{max}.$$

Thực nghiệm mặc định sử dụng $(\tau_{yel}, \tau_{red}) = (0.40, 0.60)$. Ở tuần cuối, có thể hạ ngưỡng (Ví dụ: $\tau_{yel} = 0.30$) để tăng khả năng phát hiện lớp C (Recall(C)), dù phải đánh đổi thêm nguồn lực – chi tiết xem Bảng 4.

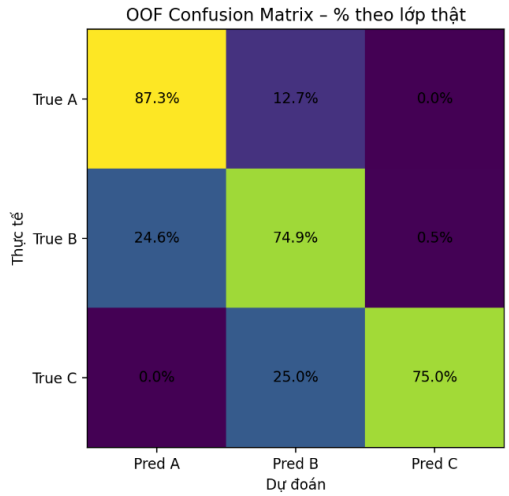
2.6. Triển khai và đầu ra vận hành:

Mô hình, bộ mã hoá nhãn và danh sách đặc trưng lần lượt được đóng gói tại *best_cal_xgb.joblib, label_encoder.joblib và feature_list.joblib*. Khi suy diễn, hệ thống sinh *final_alerts.xlsx* (cảnh báo phân tầng) và *class_summary.xlsx* (tổng hợp theo lớp/nhóm) để phục vụ giám sát và phân bổ nguồn lực, kèm *predictions_with_truth.xlsx* và *new_eval_metrics.txt* cho hậu kiểm khi có nhãn; các ví dụ được đối chiếu tại Bảng 8. Nhờ xác suất đã hiệu chỉnh tốt (*ECE* nhỏ), ngưỡng cảnh báo có thể áp dụng trực tiếp và tái lập giữa các kì như đã minh hoạ Hình 15.

Bảng 4: *Tóm tắt hiệu năng (đã hiệu chỉnh isotonic) và báo cáo theo lớp.*

Chỉ số tổng hợp				
Accuracy	**Macro-F1**	**AUPRC(C)**	**Brier**	**ECE**
0.772	0.52	0.739	0.1008	0.0577
Báo cáo phân loại (F1)				
Lớp	F1			
A	0.76			
B	0.8			
C	0.0			

(Ghi chú: Số liệu trên tập kiểm định vận hành; lớp C hiếm dẫn tới F1(C)=0 ở ngưỡng hiện tại)



Hình 2: *Ma trận nhầm lẫn OOF trên tập huấn luyện (LOCO).*

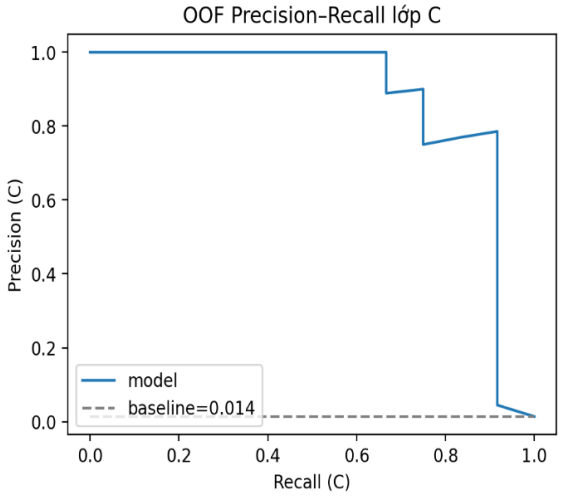
3. Kết quả nghiên cứu

Dựa trên mô hình đã huấn luyện, chúng tôi đánh giá hiệu năng theo ba kịch bản bổ sung cho nhau (OOF/LOCO, holdout 80/20 group-aware và kiểm định vận hành.

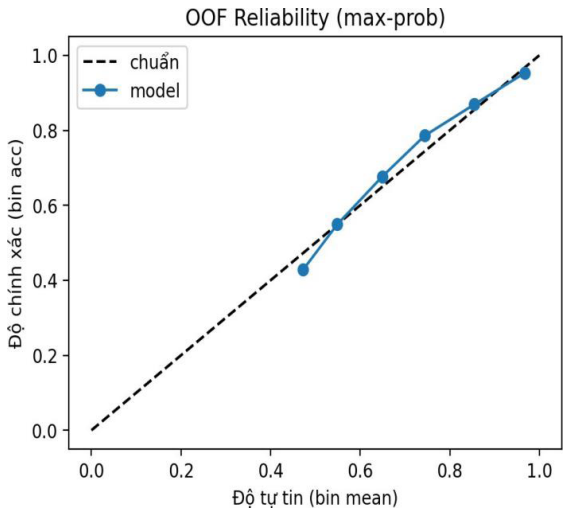
3.1. Đánh giá bằng OOF/LOCO (thước đo chính)

Chia *group-aware* theo LỚP/NHÓM giúp ngăn rò rỉ theo bối cảnh giảng dạy. Kết quả tổng hợp (xem Bảng 3) cho thấy *Macro-F1* = 0.803, *AUPRC(C)* = 0.872, *Brier* = 0.083, *ECE* = 0.016 (toàn cục từ OOF); trung bình theo fold đạt *Macro-F1* = 0.758 và *AUPRC(C)* = 0.868. So với phép đo *apparent* (Train - On - Train) có chỉ số cao hơn đáng kể, chênh lệch này xác nhận tính lạc quan khi đo trên chính dữ liệu huấn luyện và củng cố việc chọn OOF/LOCO làm chuẩn báo cáo.

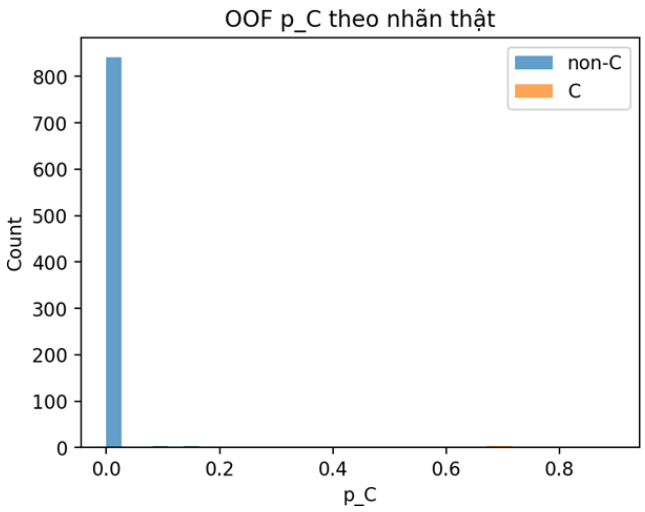
Về trực quan, ma trận nhầm lẫn OOF (xem Hình 2) cho thấy nhầm lẫn chủ đạo giữa A và B; rò rỉ sang C rất nhỏ. Đường PR lớp C (xem Hình 3) vượt rõ rệt đường cơ sở theo tần suất, cho thấy năng lực xếp hạng rủi ro với lớp hiếm. Biểu đồ hiệu chỉnh (xem Hình 4) bám sát đường chéo, ECE nhỏ cho phép dùng trực tiếp xác suất đã hiệu chỉnh để đặt ngưỡng cảnh báo. Phân phối p_C (xem Hình 5) tập trung thấp ở non-C, gợi ý chiến lược hạ ngưỡng hoặc chọn top-k khi cần tăng Recall(C).



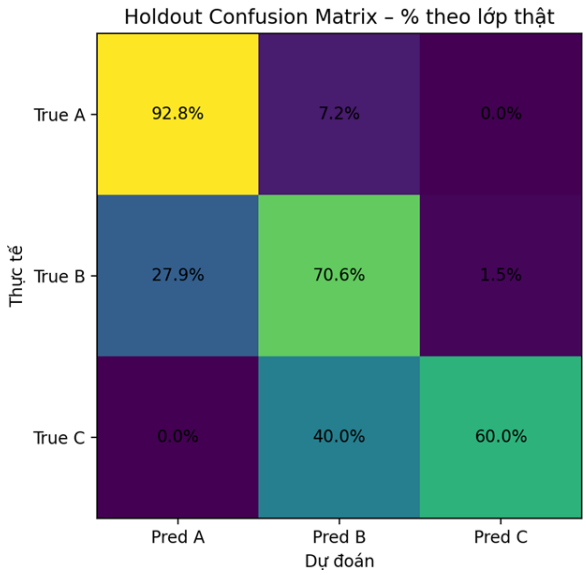
Hình 3: *Đường cong Precision–Recall cho lớp C (OOF/LOCO).*



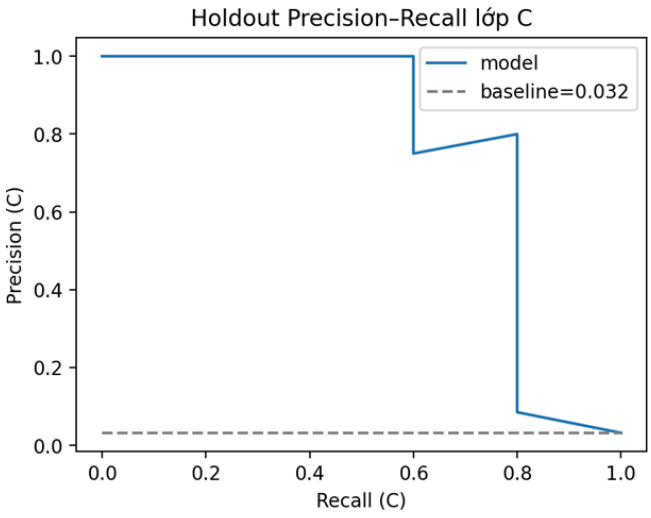
Hình 4: *Biểu đồ hiệu chỉnh (OOF/LOCO); ECE tính theo xác suất cực đại.*



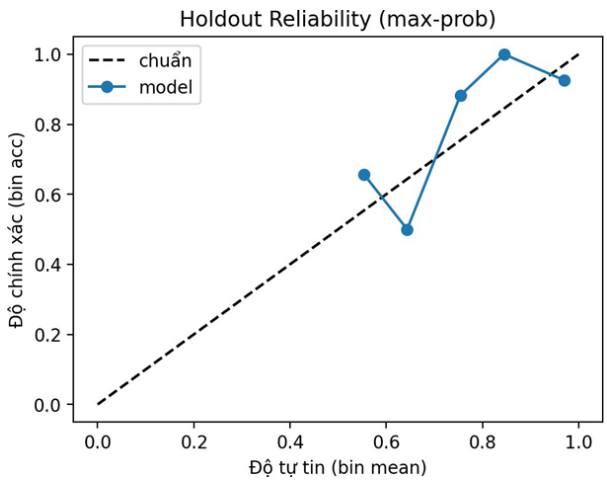
Hình 5: *Phân phối p_C theo nhãn thật (C với non-C) cho dự báo OOF/LOCO.*



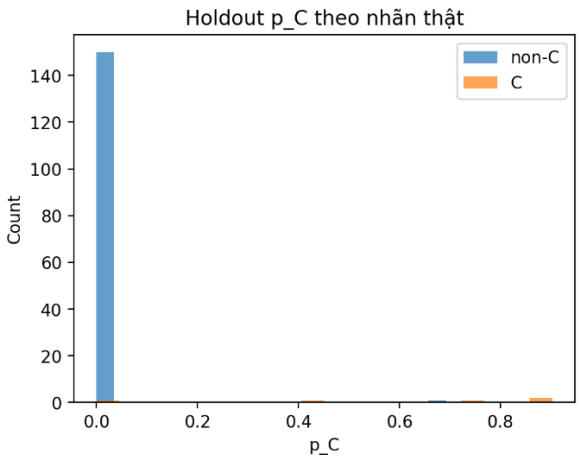
Hình 6: *Ma trận nhầm lẫn Holdout.*



Hình 7: *Đường cong Precision–Recall cho lớp C trên holdout.*



Hình 8: *Biểu đồ hiệu chỉnh trên holdout (ECE theo xác suất cực đại).*



Hình 9: *Phân phối p_C theo nhãn thật trên holdout.*

 https://doi.org/10.15625/2615-8957/12610202

Kết luận OOF/LOCO: Mô hình ổn định và xác suất được hiệu chỉnh tốt trên toàn bộ dữ liệu ngoài fold (*Macro-F1* = 0.76 ± 0.12, *AUPRC(C)* = 0.87 ± 0.19, *Brier* = 0.083, *ECE* = 0.052). Do đó, OOF/LOCO được dùng làm thước đo chính để so sánh mô hình và báo cáo *PR(C)*.

3.2. Phân tích độ vững theo holdout 80/20 (giữ LỚP)

Để kiểm tra thêm ngoài mẫu, chúng tôi dùng GroupShuffleSplit 80/20 theo LỚP/NHÓM. Hai tình huống thực tế có thể xảy ra: 1) *Holdout có lớp C trong test:* một tách thoả điều kiện dương C cho *Macro-F1* = 0.767, *AUPRC(C)* = 0.766, *Brier* = 0.086, *ECE* = 0.093. So với OOF/LOCO, *Macro-F1* hơi thấp nhưng cùng trật tự; PR(C) vẫn vượt cơ sở, xác nhận mô hình ổn định ở A/B và giữ được năng lực xếp hạng lớp C (xem Hình 6, Hình 7, Hình 8, Hình 9); 2) *Holdout không có lớp C trong test:* ở một tách khác, *Macro-F1* = 0.828, *Brier* = 0.079, *ECE* = 0.065 nhưng *AUPRC(C)* không tính được (NaN) do thiếu dương C, khiến PR(C) mất ý nghĩa. Trường hợp này được gắn cờ và *không* dùng để kết luận về lớp C.

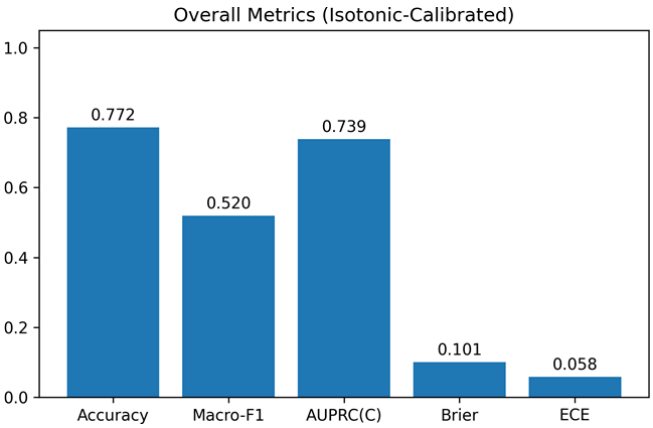
Khuyến nghị sử dụng: Khi đánh giá PR(C) trên holdout, cần phân tầng/lặp tách để đảm bảo có dương C trong test; nếu không thoả, dùng OOF/LOCO làm chuẩn báo cáo PR(C). Trong vận hành, có thể chọn ngưỡng theo ràng buộc tỉ lệ gắn cờ tối đa; nếu cần ưu tiên *Recall(C)*, hạ ngưỡng hoặc áp dụng *top-k* theo năng lực can thiệp như đã thảo luận ở Phần 4.

3.3. Đánh giá trên tập kiểm định

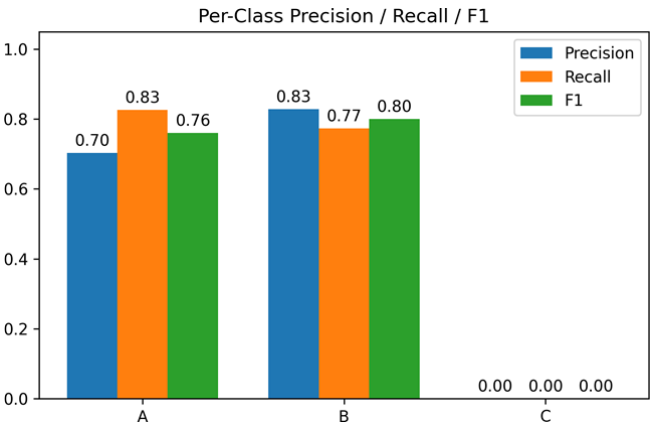
Tổng quan: Trên tập kiểm định gồm 180 bản ghi (3 lớp, 26 nhóm), sau hiệu chỉnh xác suất bằng isotonic, mô hình đạt *Accuracy* = 0.772 và *Macro-F1* = 0.520. Hai chỉ số hiệu chỉnh xác suất ở mức tốt: *Brier* = 0.101 và *ECE* = 0.058, cho thấy xác suất đầu ra đáng tin cậy để sử dụng đặt ngưỡng. Hai lớp chính A/B ổn định với điểm *F1* lần lượt khoảng 0.76 và 0.80; lớp C hiếm nên ở ngưỡng vận hành hiện tại, *F1(C)* = 0.000 (xem Bảng 4; đối chiếu chỉ số tổng hợp tại Hình 10).

Hiệu năng lớp C theo xếp hạng: Mặc dù *F1(C)* bằng 0 tại ngưỡng vận hành hiện tại, *AUPRC(C)* đạt 0.739-vượt xa đường cơ sở theo tần suất của C (khoảng 2.8%). Điều này cho thấy mô hình *xếp hạng rủi ro* tốt đối với lớp C: các ca C thực sự có xu hướng được đẩy lên vùng xác suất p_C cao hơn phần còn lại nhưng chưa vượt qua ngưỡng cảnh báo đang đặt khá thận trọng (xem Hình 11 và Hình 12; đối chiếu Bảng 4). Hàm ý vận hành là nên đặt ngưỡng dựa theo đường PR (hoặc chọn *top-k* theo p_C) để tăng *Recall(C)* trong giới hạn nguồn lực.

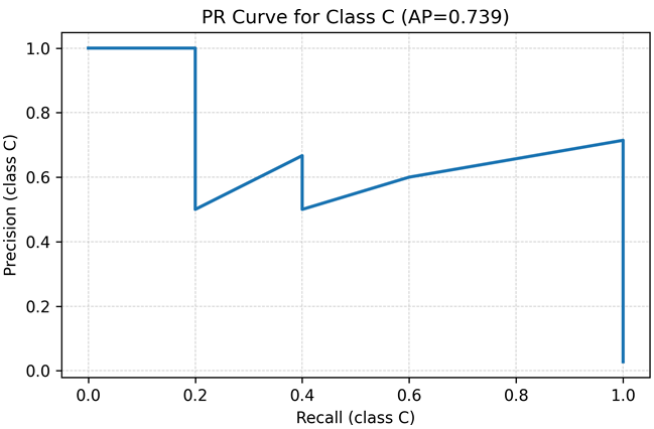
Đọc lỗi từ ma trận và báo cáo lớp: Mô hình phân biệt A và B tương đối tốt nhưng còn chồng lấn: ước tính tỷ lệ đúng theo hàng khoảng 82.6% ở A và 77.4% ở B; lỗi chủ đạo là nhầm A thành B (khoảng 17.4%) và B



Hình 10: *Chỉ số tổng hợp sau hiệu chỉnh isotonic trên tập kiểm định.*



Hình 11: *Precision/Recall/F1 theo từng lớp A/B/C. Hai lớp A/B đạt F1 cao và cân bằng; F1(C)= 0 do lớp C hiếm và ngưỡng cảnh báo đang đặt ưu tiên độ chính xác.*



Hình 12: *Đường cong Precision - Recall cho lớp C.*

thành A (khoảng 22.6%). Với lớp C, cả 5 trường hợp đều bị dự đoán về B, phản ánh việc ngưỡng hiện tại chưa "kích hoạt" phân tách cho nhóm rủi ro thấp nhất. Quy đổi sang số lượng, khoảng 12 sinh viên lớp A bị kéo lên B và khoảng 24 sinh viên lớp B bị kéo xuống A. Mẫu hình lỗi này phù hợp với việc phân phối xác suất p_C tập trung gần 0 ở phần lớn mẫu (xem Hình 13 và Bảng 5; tham chiếu thêm phân phối p_C tại Hình 14).

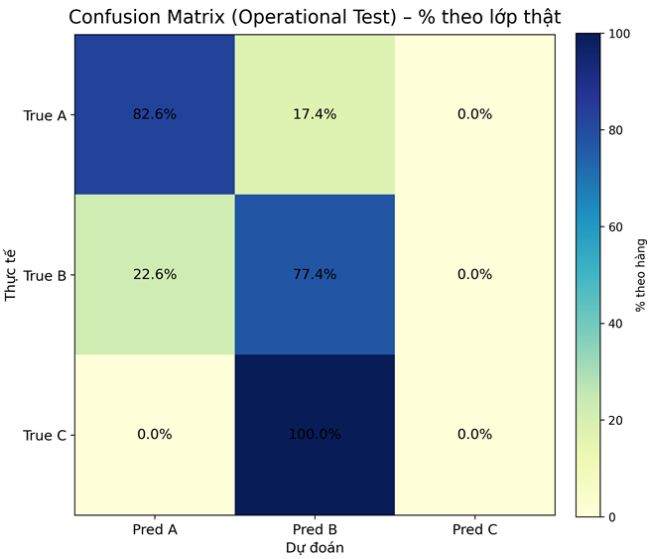
Cân bằng ngưỡng cảnh báo và nguồn lực can thiệp: Với cấu hình ưu tiên độ chính xác hiện tại (gắn cờ yellow/red khi $p_C \geq 0.40/0.60$), tỉ lệ cảnh báo ở mức thấp (đa số green). Để tăng *Recall(C)* trong vận hành, có thể: (1) hạ ngưỡng p_C, hoặc (2) chọn *top-k* theo năng lực can thiệp (5–10% tổng số sinh viên). Vì *ECE* nhỏ, xác suất sau hiệu chỉnh có thể dùng trực tiếp để đặt ngưỡng minh bạch và tái lập giữa các kỳ (xem Hình 15 và minh hoạ danh sách cảnh báo tại Bảng 6).

Hàm ý vận hành: Bảng 7 cho thấy phân bố nghiêng mạnh về A/B; xác suất p_C hiếm khi khác 0, phản ánh sự hiếm của lớp C và nguồn lực hỗ trợ hạn chế, chiến lược *top-k* theo p_C (Ví dụ 9–18 sinh viên cho $N = 180$) là thực dụng để bao phủ rủi ro mà vẫn kiểm soát khối lượng công việc. Đồng thời, nên tăng cường các đặc trưng động theo thời gian (chuỗi vắng/late có trọng số, động lượng tham gia, phương sai nhóm) nhằm cải thiện phân tách A/B và đẩy những ca C tiềm năng lên vùng xác suất cao hơn. Báo cáo phân tầng theo lớp/nhóm giúp phân bổ mentoring mục tiêu hiệu quả hơn (xem Bảng 8).

Bảng 5: *Báo cáo phân loại trên 180 mẫu (3 lớp, 26 nhóm)*

Lớp	Precision	Recall	F1	Support
A	0.704	0.826	0.76	69
B	0.828	0.774	0.8	106
C	0.0	0.0	0.0	5
Accuracy			0.772	
Macro avg	0.511	0.533	0.52	180
Weighted avg	0.758	0.772	0.762	180

(Chú thích: Macro avg là trung bình đều giữa các lớp; Weighted avg có trọng số theo support, phản ánh mất cân bằng lớp).



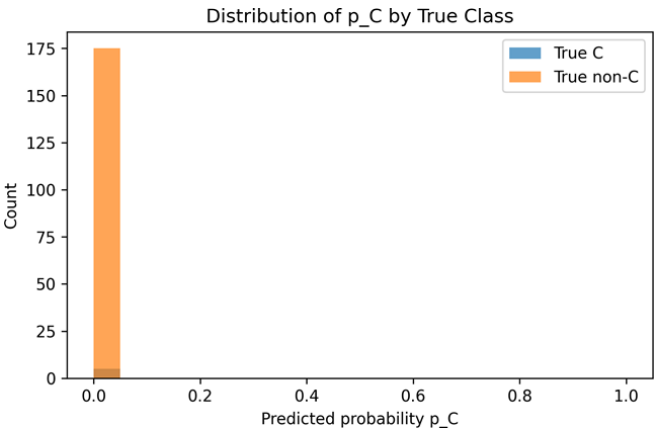
Hình 13: *Ma trận nhầm lẫn trên tập kiểm định*

4. Thảo luận

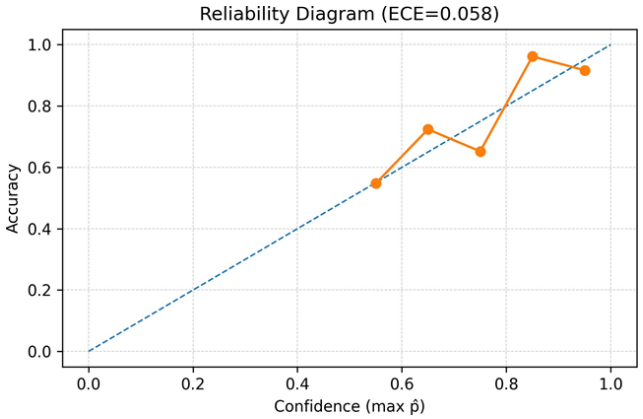
Các kết quả cho thấy mô hình xếp hạng rủi ro tốt ngay cả khi lớp C hiếm; xác suất sau hiệu chỉnh isotonic ổn định đủ để dùng trực tiếp trong vận hành.

4.1. Hàm ý sư phạm và khuyến nghị vận hành

Mục tiêu chung: Biến xác suất đã hiệu chỉnh thành hành động lớp học theo hướng *giảng dạy phân tầng*: ưu tiên can thiệp sớm cho nhóm rủi ro (C), củng cố



Hình 14: *Phân phối p_C theo nhãn thật*



Hình 15: *Biểu đồ hiệu chỉnh (Reliability) theo xác suất cực đại*

 https://doi.org/10.15625/2615-8957/12610202

Bảng 6: *Cảnh báo cuối kì (final_alerts; 180 dòng × 12 cột) – trích xuất tiêu biểu*

STT	MSSV	LỚP	NHÓM	Tier_ pred	p_A	p_B	p_C	att_ active	att_ sum	kn_mean	alert
1	2123260067	2203603	3603-N01	A	0.72865	0.27135	0.0	1	1	7.666667	GREEN
2	2123260007	2203603	3603-N01	A	0.72865	0.27135	0.0	2	2	7.666667	GREEN
3	2123260006	2203603	3603-N01	B	0.0	1.0	0.0	1	-5	6.833333	GREEN
4	2123260066	2203603	3603-N01	A	0.984127	0.015873	0.0	2	3	8.166667	GREEN
5	2123260021	2203603	3603-N01	A	0.797149	0.202851	0.0	2	2	7.833333	GREEN
…	…	…	…	…	…	…	…	…	…	…	…
176	2121190046	22212119	2119-N08	B	0.011068	0.988932	0.0	2	-1	6.833333	GREEN
177	2121190002	22212119	2119-N08	B	0.043855	0.956145	0.0	1	-2	7.0	GREEN
178	2121190010	22212119	2119-N08	B	0.096695	0.903305	0.0	2	-1	7.0	GREEN
179	2121190001	22212119	2119-N08	B	0.096695	0.903305	0.0	1	-1	7.0	GREEN
180	2121190036	22212119	2119-N08	B	0.030351	0.969649	0.0	2	-1	7.0	GREEN

(Ghi chú: Alert được ánh xạ từ xác suất đã hiệu chỉnh: RED nếu p_C ≥ 0.60 hoặc dự đoán C với att_active ≤ 2; YELLOW nếu p_C ≥ 0.40; còn lại là GREEN).

Bảng 7: *Phân tầng cảnh báo theo lớp (predictions_wit_truth.xlsx; 180 dòng × 12 cột) - trích tiêu biểu.*

STT	LỚP	NHÓM	MSSV	Đ.QT	Đ.THI	Đ.MH	Tier3_ true	Tier_ pred	p_A	p_B	p_C
1	2203603	3603-N01	2123260067	8.4	7.5	7.9	B	A	0.725328	0.274672	0.0
2	2203603	3603-N01	2123260007	8.7	8.0	8.3	A	A	0.72865	0.27135	0.0
3	2203603	3603-N01	2123260006	6.5	7.0	6.8	B	B	0.0	1.0	0.0
4	2203603	3603-N01	2123260066	9.2	8.0	8.5	A	A	0.889249	0.110751	0.0
5	2203603	3603-N01	2123260021	8.8	8.0	8.3	A	A	0.797149	0.202851	0.0
…	…	…	…	…	…	…	…	…	…	…	…
176	22212119	2119-N08	2121190046	7.4	7.5	7.4	B	B	0.011068	0.988932	0.0
177	22212119	2119-N08	2121190002	7.5	7.0	7.3	B	B	0.043855	0.956145	0.0
178	22212119	2119-N08	2121190010	7.7	7.5	7.6	B	B	0.043855	0.956145	0.0
179	22212119	2119-N08	2121190001	7.7	8.0	7.8	B	B	0.093356	0.906644	0.0
180	22212119	2119-N08	2121190036	7.5	7.0	7.3	B	B	0.030351	0.969649	0.0

(Ghi chú: Các cột p_A, p_B, p_C là xác suất đã hiệu chỉnh, phản ánh mức độ tin cậy khi phân vào từng tầng).

Bảng 8: *Tổng hợp theo lớp (trích từ class_summary.xlsx).*

LỚP	N	%A	%B	%C	Top5_by_p_C (MSSV)	Nhóm_rủi_ro
2203603	62	41.9	58.1	0.0	2123260145, 2123260022, 2123260196, 2123260049, …	3603-N09, 3603-N06, 3603-N08, 3603-N03
22203902	62	43.5	56.5	0.0	2123120121, 2123120163, 2123120503, 2123120181, …	3902-N09, 3902-N06, 3902-N01, 3902-N05, 3902-N07
22212119	56	50.0	50.0	0.0	2121190061, 2121190030, 2121190059, 2121190039, …	2119-N06, 2119-N02, 2119-N01

(Ghi chú: Top5_by_p_C liệt kê MSSV có xác suất lớp C cao nhất trong mỗi lớp; Nhóm_rủi_rô là các NHÓM xuất hiện trong top-k).

đều đặn cho nhóm trung bình (B) và tận dụng vai trò dẫn dắt của nhóm cao (A).

Nhóm nguy cơ (C): Ưu tiên gặp trực tiếp cá nhân hoặc nhóm nhỏ; tập trung vào nhiệm vụ ngắn, rõ mục tiêu (ví dụ: "3-2-1" - ba điểm chính, hai câu hỏi, một cam kết tuần), kèm trình bày ngắn và tiêu chí đánh giá thu gọn để theo dõi tiến bộ. Vận hành ngưỡng bất đối xứng (dễ kích hoạt hơn với các ca có tham gia thấp), đồng thời xem *cụm nhóm* rủi ro cao để phân công giảng viên/phụ giảng. Khối lượng điển hình: 5-10% sinh viên/lớp, khoảng 60-90 phút/tuần.

Nhóm trung bình (B): Duy trì nhịp luyện tập có phản hồi nhanh: đóng vai theo kịch bản, báo cáo ngắn theo khung "Tình huống - Hành động - Kết quả", checklist kĩ năng hợp tác. Gắn cờ yellow/green, yêu cầu bằng chứng hàng tuần với tỉ trọng điểm nhỏ để khuyến khích tần suất. Khối lượng: 45 - 60 phút/tuần.

Nhóm cao (A): Giao vai trò dẫn dắt mini–project ngắn, hỗ trợ kèm cặp B/C theo NHÓM, chia sẻ thực hành tốt. Tổ chức phản hồi đồng đẳng có định dạng để không tăng tải giảng viên. Khối lượng: khoảng 30 phút/tuần.

Tính khả thi trong học kì: Tổng thời lượng bổ sung cho giảng viên khoảng 2–3 giờ/tuần/lớp, tập trung cho ca red/yellow. Báo cáo định kì sử dụng hai đầu ra: *final_alerts* (theo sinh viên) và *class_summary* (theo lớp/nhóm) để cân bằng giữa mức độ nhắm mục tiêu, tính khả thi vận hành và mục tiêu tăng bao phủ lớp C.

4.2. Ràng buộc, giám sát và bảo trì mô hình

Theo dõi độ tin cậy: Giám sát Brier và ECE theo tuần. Khi ECE tăng đáng kể, ưu tiên hiệu chỉnh lại xác suất (Recalibration) hoặc cập nhật mô hình cuối kì.

Độ vững và bất định: Do lớp C hiếm, nên kèm khoảng tin cậy (bootstrap) cho Macro–F1 và AUPRC(C), đồng thời báo cáo độ nhạy theo ngưỡng để minh bạch đánh đổi giữa *precision* và *recall*.

Công bằng giữa các nhóm: Đối chiếu chỉ số theo LỚP/NHÓM để phát hiện lệch can thiệp có hệ thống. Khi cần, áp dụng điều chỉnh ngưỡng theo nhóm hoặc ràng buộc trong huấn luyện.

Quy trình "Human-in-the-loop": Các ca red/yellow được giảng viên duyệt trước khi áp dụng biện pháp mạnh; phản hồi của giảng viên được dùng làm tín hiệu học tăng cường cho phiên bản tiếp theo.

Dữ liệu và riêng tư: Tích luỹ qua nhiều học kì để ổn định ước lượng, song hành với ẩn danh hoá và tối thiểu hoá thông tin nhận diện.

4.3. Nhận xét bổ sung từ dữ liệu dự báo

Phân bố dự đoán: Dữ liệu thực nghiệm nghiêng mạnh về A/B; xác suất cho C hiếm khi vượt ngưỡng, phù hợp tần suất C thấp. Mô hình ổn định ở A/B nhưng còn nhầm lẫn chéo khi đặc trưng hành vi/điểm quá trình chồng lấn.

Giảm nhầm lẫn A - B: Bổ sung đặc trưng động theo thời gian (xu hướng, dao động, đúng hạn), đặc trưng nhóm (phương sai nội nhóm) và tín hiệu nhiệm vụ nhỏ có tiêu chí đánh giá; có thể áp dụng ràng buộc đơn điệu hoặc trọng số mất mát.

Ưu tiên phát hiện C: Khi mục tiêu là nâng *recall* cho C, hạ ngưỡng cảnh báo hoặc áp dụng chiến lược *top-k* theo ngân sách can thiệp (tỉ lệ sinh viên tối đa cần xử lí mỗi kì). Trong báo cáo, ưu tiên đường cong PR cho C thay vì F1 đơn điểm để phản ánh đúng năng lực xếp hạng khi lớp hiếm.

4.4. Tổng hợp định hướng triển khai

- Dùng xác suất đã hiệu chỉnh để đặt ngưỡng minh bạch, có giám sát theo tuần bằng Brier/ECE.

- Áp dụng *top-k* hoặc hạ ngưỡng có kiểm soát nhằm tăng bao phủ C, kèm duyệt của giảng viên trước can thiệp.

 https://doi.org/10.15625/2615-8957/12610202

- Bổ sung đặc trưng động theo thời gian và nhóm để giảm nhầm A–B.

- Duy trì đánh giá group–aware (OOF/LOCO) làm chuẩn, kết hợp báo cáo bất định và công bằng theo lớp/nhóm cho các quyết định vận hành.

5. Kết luận

Bài báo giới thiệu một khung *Dự báo → Phân tầng → Can thiệp* cho học phần Kĩ năng mềm, nhấn mạnh tính khả thi triển khai trong lớp học thực và khả năng tái lập qua các kì. Các kết quả cho thấy mô hình có năng lực xếp hạng rủi ro tốt ngay cả khi lớp C hiếm, và xác suất sau hiệu chỉnh isotonic đủ tin cậy để chuyển hoá thành quyết định vận hành.

- *Đóng góp chính:* 1) Xây dựng pipeline dữ liệu–mô hình theo thực hành tốt của EDM, kiểm soát rò rỉ theo thời gian và đánh giá *group-aware* trên cặp khoá LỚP/NHÓM; 2) Thiết kế đặc trưng kết hợp hành vi theo tuần (T1–T7), điểm liên tục và tín hiệu nhóm, với hai chế độ dùng linh hoạt cho dự báo giữa kì và tổng kết cuối kì; 3) Mô hình hoá đa lớp bằng XGBoost và hiệu chỉnh xác suất isotonic, cho đầu ra xác suất ổn định để ánh xạ thành quy tắc cảnh báo; 4) Đóng gói triển khai thực dụng (mã mô hình, bộ mã hoá nhãn, danh sách đặc trưng) cùng báo cáo theo lớp/nhóm giúp chuyển dự báo thành danh sách hành động ở mức "ai/ở đâu".

Ý nghĩa thực tiễn. Khung đề xuất tạo một lộ trình mạch lạc từ *dự báo* đến *hành động sư phạm*: dự báo sinh viên theo tầng, tổng hợp theo lớp/nhóm để phân bổ nguồn lực, và kích hoạt can thiệp sớm với ngưỡng minh bạch dựa trên xác suất đã hiệu chỉnh. Cách tiếp cận này phù hợp bối cảnh dữ liệu giáo dục nhỏ - trung bình, nơi tính tin cậy của xác suất và quy trình vận hành đóng vai trò quyết định.

Hướng phát triển: 1) Chuẩn hoá quy tắc chọn ngưỡng trên đường PR theo ràng buộc nguồn lực (tỉ lệ gắn cờ tối đa); 2) Bổ sung báo cáo bất định (Bootstrap) cho AUPRC và ECE, đồng thời giám sát *drift* theo thời gian; 3) Khảo sát các mô hình nền tảng cho dữ liệu bảng (TabPFN/Real-TabPFN) để tăng độ bền vững trên tập nhỏ - trung bình; 4) Hoàn thiện quy trình *human-in-the-loop* giữa mô hình và giảng viên/trợ giảng nhằm bảo đảm can thiệp phù hợp ngữ cảnh.

Tài liệu tham khảo

Angeioplastis, A., Aliprantis, J., Konstantakis, M. & Tsimpiris, A. (2025). Predicting student performance and enhancing learning outcomes: A data-driven approach using educational data mining techniques. *Computers, 14*(3), 83. https://doi.org/10.3390/computers14030083

Baker, R. S. & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), pp.3–17.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), pp.1–3.

Chen, T. & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16),* pp.785–794. https://doi.org/10.1145/2939672.2939785.

Garg, A., Ali, N., Hollmann, N., Purucker, L., Müller, S. & Hutter, F. (2025). Real-TabPFN: Improving tabular foundation models via continued pre-training with real-world data. *In Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data*.

Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. (2017). On calibration of modern neural networks. *In Proceedings of the 34th International Conference on Machine Learning,* pp. 1321–1330.

Hollmann, N., Hütter, S., Schirrmeister, R. T., et al. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*. Advance online publication.

Malik, S., Patro, S. G. K., Mahanty, C., Hegde, R., Naveed, Q. N., Lasisi, A., Buradi, A., Emma, A. F. & Kraiem, N. (2025). Advancing educational data mining for enhanced student performance prediction: A fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution. *Scientific Reports, 15*, p.8738. https://doi.org/10.1038/s41598-025-92324-x

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, pp.2825–2830.

Romero, C. & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), pp.601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Tạp chí Giáo dục Thành phố Hồ Chí Minh. (2024). Báo động 80% sinh viên thiếu hụt kĩ năng mềm: Cao đẳng Việt Mĩ nỗ lực đổi mới đào tạo. *Giáo dục Thành phố Hồ Chí Minh.* https://giaoduc.edu.vn/bao-dong-80-sinh-vien-thieu-hut-ky-nang-mem-cao-dang-viet-my-no-luc-doi-moi-dao-tao/

Zadrozny, B. & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.694–699.