AI-Powered Student Performance Prediction and Academic Risk Analytics in Higher Education: A Class-Decomposed XGBoost Framework

1st Ensophea Toch
School of Graduate Studies
Paragon International University
Phnom Penh, Cambodia
etoch@paragoniu.edu.kh

2nd Sophanidate Toem
Faculty of Science and Technology
ACLEDA University of Business
Phnom Penh, Cambodia
detmarian6@gmail.com

3rd Pakrigna Long
Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia
Phnom Penh, Cambodia
long.pakrigna@itc.edu.kh

***Abstract*—Student dropout remains a significant challenge in higher education, resulting in substantial academic, financial, and institutional consequences. This paper presents a machine learning framework for early identification of students at risk of dropout, combining XGBoost with an engineered set of academic-progression ratio features and a Class Decomposition (CD) strategy: the majority (non-dropout) class is partitioned into k homogeneous sub-clusters via k-means prior to training, so the classifier is fit as a multi-class problem and the resulting sub-cluster probabilities are merged back into the original two classes at inference time. The proposed framework is evaluated on the publicly available *Predict Students' Dropout and Academic Success* benchmark dataset under a binary dropout/non-dropout framing. At the default 0.5 decision threshold, the proposed XGBoost model reaches an AUC-ROC of 0.940 and an accuracy of 90.0% on a held-out test set. Hyperparameters are selected via randomized search optimizing AUC under stratified cross-validation, and an alternative operating threshold can be separately calibrated using out-of-fold training predictions, avoiding test-set leakage. Comparisons with recent studies using the same benchmark dataset show that the proposed framework achieves the strongest reported AUC and accuracy among prior work restricted to this dataset and evaluation protocol, without relying on synthetic oversampling. Beyond raw predictive performance, we formalize the Class Decomposition procedure as a model-agnostic pre-training step (Algorithm 1), discuss deployment as a two-stage early-warning module inside institutional Student Information Systems (SIS), and outline the threats to validity and reproducibility considerations relevant to adopting the pipeline at a new institution. The findings suggest that the proposed pipeline is a practical, auditable foundation for institutional early-warning systems.**

***Index Terms*—student dropout prediction, educational data mining, XGBoost, class decomposition, academic risk analytics, student information systems**

I. Introduction

Fewer than 40% of students who begin tertiary education in OECD countries complete their degree within the standard timeframe [1], at substantial cost to students, institutions, and the wider economy [2]. Reactive intervention, after a student has already failed courses or withdrawn, is markedly less effective than proactive outreach triggered by early risk indicators [3]. As institutions increasingly digitize academic records through Learning Management Systems (LMS) and Student Information Systems (SIS), the administrative data required for early-warning analytics is already being collected – the remaining gap is a validated, deployable modeling pipeline that institutions can attach to that data with minimal re-engineering.

Machine learning is increasingly used to identify at-risk students from administrative records [4]–[7], but three gaps recur in this literature. First, feature engineering choices are often ad hoc and institution-specific, limiting reproducibility. Second, standard classifiers treat the majority (non-dropout) class as a single homogeneous group, even though it mixes structurally different sub-populations (e.g., graduated versus still-enrolled students), which can blur the decision boundary around the minority dropout class. Third, few studies discuss how a validated model would actually be integrated into the software systems – SIS, LMS, advising dashboards – that institutions already operate, leaving a gap between benchmark performance and deployable analytics. This paper addresses all three gaps with a documented ratio-based feature pipeline, a Class Decomposition (CD) strategy that explicitly models majority-class heterogeneity, and an explicit discussion of SIS-level integration, evaluated on the Predict Students' Dropout and Academic Success dataset [8].

Contributions: (1) a documented, reproducible set of ratio-based academic-progression features – first- and second-semester approval and evaluation ratios, a grade-trend indicator, and an overall approval ratio – built directly from curricular-unit counts; (2) a formalized, model-agnostic Class Decomposition procedure (Algorithm 1) that partitions the majority class into k KMeans sub-clusters before training, converting the task into a multi-class problem whose sub-cluster probabilities are merged back into the original two classes at inference time; (3) an XGBoost classifier reaching AUC 0.940 and accuracy 90.0% at the default 0.5 threshold, benchmarked exclusively against other work using the same dataset and evaluation protocol; and (4) an explicit discussion of deployment as a two-stage early-warning module within institutional SIS/LMS infrastructure, together with threats-to-validity and

reproducibility considerations for cross-institutional adoption.

II. Related Work

Classical retention theory frames dropout as weak academic and social integration, later extended to treat financial precarity as an independent driver [9], [10]. Educational data mining surveys report that academic-performance features consistently outperform demographic ones, and that ensemble tree methods outperform single classifiers [11], [12]; neural approaches have shown only marginal gains over gradient-boosted trees on structured educational data, at higher computational cost [13]. Clustering has also been used to structure heterogeneous student sub-populations for downstream prediction [14]; we build on this idea, but apply it to decompose the majority class itself rather than to engineer an additional input feature.

Handling class imbalance. Resampling methods such as SMOTE have been applied on this exact dataset [19], but synthetic oversampling can distort the original feature distribution and, in a deployed SIS context, complicates auditability since synthetic records have no corresponding student. Class decomposition instead partitions the already-adequately-sized majority class into homogeneous sub-groups prior to training, sharpening the decision boundary around the minority class without duplicating or synthesizing any sample – every training row remains traceable to a real student, which matters for institutional accountability and for any downstream fairness audit.

Prior work on this dataset. The dataset [8] originates from a Portuguese polytechnic and has been cited over 160 times, so we restrict quantitative comparison to studies using it. Realinho et al. [8], the dataset's creators, benchmarked Random Forest, XGBoost, LightGBM, and CatBoost, reporting 86.8% accuracy under a comparable binary framing. Romero and Liao [17] compared Lasso, a GAM, Random Forest, XGBoost, and a single-layer neural network on the same 4,424-record binary task; XGBoost was strongest, with second-semester grades and credited units the leading predictors, consistent with our findings, though that study reports F1 rather than AUC or accuracy. Ridwan and Priyatno [18] applied XGBoost with variance-threshold and mutual-information feature selection, reporting 90% accuracy, with an 80:20 split as the best operating point. Silkhi et al. [19] combined five boosting algorithms with SMOTE on the same 4,424-record dataset, leading with ROC-AUC as their primary criterion and reporting a best result of AUC = 0.935 (LightGBM, SMOTE-balanced). Almeida et al. [20] evaluated Random Forest, XGBoost, TabNet, and two graph neural networks (GCN, GraphSAGE) on the original three-class formulation; the best GNN configuration edged out XGBoost by a small margin the authors attribute to graph-construction strategy rather than a general GNN advantage. Because that study keeps all three classes rather than the binary framing used here and in [8], [17]–[19], its numbers are discussed qualitatively rather than placed in Table IV. Together, this body of work confirms gradient-boosted trees as the strongest practical baseline on this dataset, and, to our knowledge, no prior study on this dataset combines this ratio-based feature set with a class-decomposition strategy over the majority class, nor discusses its integration into SIS-level early-warning infrastructure.

Analytics infrastructure in higher education. A separate strand of work addresses learning-analytics dashboards and SIS-integrated alerting rather than model accuracy per se, generally treating the underlying classifier as a black box to be swapped in [7]. Our framework is designed to sit in that slot: it consumes only semester-level administrative fields already present in most SIS schemas, and exposes a single dropout probability per student, so it can be adopted without requiring institutions to redesign their existing analytics infrastructure.

III. Dataset

The dataset comprises 4,424 records with 34 predictor variables spanning demographic, socioeconomic, academic, and macroeconomic attributes. The original three-class target (Dropout/Enrolled/Graduate) is collapsed to a binary task, Dropout as the positive class (32.8% prevalence), split into training and held-out test sets via stratified sampling (85:15). Approved units drop slightly from semester 1 to semester 2 ($4.68 \rightarrow 4.41$); 87.9% of students had tuition fees up to date, 24.4% held scholarships, and 11.4% were classified as debtors. These three administrative fields are notable because, unlike academic-performance features, they are directly actionable by an institution (e.g., a payment plan or scholarship review) rather than merely diagnostic, a distinction we return to in Section VII.

IV. Feature Engineering and Class Decomposition

Dropped columns. `Nacionality` and `International` are near-constant (over 97% one value) and are dropped before modeling, as they add noise rather than signal. Course and previous qualification are passed to XGBoost as their original integer category codes rather than one-hot encoded, since tree-based splitting separates categories through learned thresholds without an added dimensionality increase.

Ratio-based feature engineering. Rather than feeding raw curricular-unit counts to the model on their own, we derive ten additional features (Table I) that describe a student's academic progression relative to their own enrollment load, which is far more informative than an absolute count and is naturally scale-free – a property that also aids cross-institutional transfer, since enrollment loads vary across programs and institutions while ratios remain comparable. Together with the 32 retained original columns, this gives a final feature space of 42 columns.

Class Decomposition (CD). The non-dropout class mixes two structurally different populations – students who have already graduated and students who are still enrolled and progressing normally – which we hypothesize dilutes the decision boundary that separates either group from genuine dropouts. Rather than training a standard binary classifier, we first fit k-means to the training-set feature vectors of the majority class only (features standardized for this clustering step, since curricular counts, age, and binary flags live on very different numeric scales), producing k sub-clusters. Each majority-class

TABLE I
ENGINEERED RATIO-BASED FEATURES

Feature	Definition
approval_ratio_1st	CU 1st sem (approved) / CU 1st sem (enrolled)
approval_ratio_2nd	CU 2nd sem (approved) / CU 2nd sem (enrolled)
eval_ratio_1st	CU 1st sem (evaluations) / CU 1st sem (enrolled)
eval_ratio_2nd	CU 2nd sem (evaluations) / CU 2nd sem (enrolled)
grade_trend	CU 2nd sem (grade) − CU 1st sem (grade)
total_approved	CU 1st sem (approved) + CU 2nd sem (approved)
total_enrolled	CU 1st sem (enrolled) + CU 2nd sem (enrolled)
overall_approval_ratio	total_approved / total_enrolled
avg_grade	mean of 1st and 2nd sem. grades
no_eval_flag	1 if either semester has > 0 units without evaluation, else 0

Algorithm 1 Class Decomposition (CD) pre-training step

Require: Training features X, binary labels $y \in \{0, 1\}$ (0 = majority/non-dropout, 1 = minority/dropout), number of sub-clusters k
Ensure: Fitted $(k{+}1)$-class model M
1: $X_{\text{maj}} \leftarrow \{x_i \in X : y_i = 0\}$
2: Standardize X_{maj} to zero mean, unit variance $\rightarrow \tilde{X}_{\text{maj}}$
3: $C \leftarrow \text{KMeans}(\tilde{X}_{\text{maj}}, k)$ ▷ fit on training majority rows only
4: $c_i \leftarrow C.\text{predict}(\tilde{x}_i)$ for each $x_i \in X_{\text{maj}}$
5: $y'_i \leftarrow c_i$ for $x_i \in X_{\text{maj}}$; $y'_i \leftarrow k$ for x_i with $y_i = 1$
6: Fit $M \leftarrow \text{XGBoost}(X, y')$ as a $(k{+}1)$-class classifier
7: **return** M, C ▷ C is retained for re-clustering at a new institution

training row is relabeled with its sub-cluster identity while the minority (dropout) label is left untouched, and XGBoost is trained as a $(k{+}1)$-class classifier (`multi:softprob`). At inference time, the predicted probabilities of all k majority sub-classes are summed back into a single non-dropout probability, giving a merged two-class output; the dropout probability is used as the model's score, and the arg-max of the two merged probabilities gives the predicted label. Because this procedure re-labels existing rows rather than duplicating or synthesizing any sample, it preserves the original data distribution while still sharpening the boundary around the minority class. Algorithm 1 formalizes the procedure so it can be reproduced independently of this specific dataset, and Fig. 1 summarizes the full pipeline.

Computational complexity. The added cost of CD over plain XGBoost is the KMeans fit on the majority-class training rows, $O(n_{\text{maj}} \cdot d \cdot k \cdot t)$ for t Lloyd iterations, which is negligible next to gradient-boosted tree fitting on the full 42-dimensional feature space; at inference time CD adds only a fixed-cost summation over k probabilities per prediction, so real-time scoring inside an SIS pipeline is unaffected.

V. METHODOLOGY

Hyperparameters are selected via randomized search (60 iterations) over the space in Table II, optimizing AUC under 5-fold stratified cross-validation. The number of majority-class sub-clusters, k, is selected separately by sweeping

Algorithm 2 Inference with a CD-trained model

Require: Test point x, fitted model M, sub-cluster count k
Ensure: Dropout probability p_{drop}
1: $(p_0, \ldots, p_{k-1}, p_k) \leftarrow M.\text{predict_proba}(x)$
2: $p_{\text{non-drop}} \leftarrow \sum_{j=0}^{k-1} p_j$
3: $p_{\text{drop}} \leftarrow p_k$
4: **return** p_{drop} ▷ compare against operating threshold τ

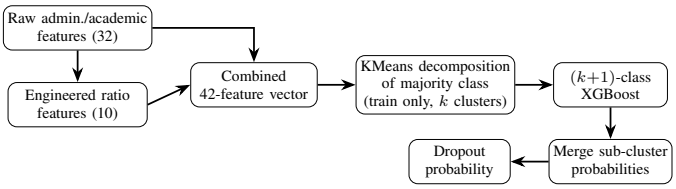


Fig. 1. Feature and modeling pipeline. Majority-class decomposition is applied only to the training set; centroids and the resulting sub-cluster structure are learned once and used to relabel training rows before fitting a single multi-class XGBoost model.

$k \in \{2, 3, 4, 5, 6\}$ under the same 5-fold stratified cross-validation protocol and choosing the value that maximizes mean validation AUC. The final model is refit on the full training set using the selected hyperparameters and k, and evaluated once on the held-out test set.

The headline results in Section VI are reported at the default 0.5 decision threshold. As an alternative, an F1-optimal threshold can be calibrated using out-of-fold predictions computed on the training set only (never touching the test set), then applied once to the held-out test set at evaluation time – this avoids the leakage that would result from tuning a threshold directly against test-set outcomes.

Reproducibility. All splits are stratified and seeded; the randomized hyperparameter search, the k-sweep, and the final refit use a fixed random state so that the reported test-set numbers are exactly reproducible from the same seed. The preprocessing pipeline, feature definitions (Table I), and hyperparameter search space (Table II) are fully specified in this paper to support independent re-implementation, consistent with growing expectations for reproducible educational data mining research.

VI. RESULTS

A. Predictive Performance

Table III presents the performance of the proposed XGBoost with Class Decomposition (XGBoost+CD) framework on the held-out test set. The model achieved an AUC–ROC of 0.940 and an accuracy of 90.0%, indicating excellent discrimination between dropout and non-dropout students.

B. Comparison with Prior Work on the Same Dataset

Table IV is restricted to studies using the identical 4,424-record dataset under a comparable binary framing; the three-class GNN/TabNet results of [20] are excluded for the reason given in Section II and discussed only in prose above. Following this study's focus on AUC and accuracy, F1 is

TABLE II
XGBOOST HYPERPARAMETER SEARCH SPACE (RANDOMIZEDSEARCHCV, 60 ITERATIONS, 5-FOLD CV, SCORING = AUC)

Parameter	Range
n_estimators	[200, 900]
max_depth	[3, 8]
learning_rate	[0.01, 0.16]
subsample	[0.6, 1.0]
colsample_bytree	[0.6, 1.0]
min_child_weight	[1, 8]
reg_lambda	[0.5, 3.5]
gamma	[0.0, 1.0]

TABLE III
PERFORMANCE SUMMARY OF THE PROPOSED FRAMEWORK

Model	Threshold	AUC–ROC	Accuracy
Proposed XGBoost+CD	0.50	0.940	0.900

omitted from the comparison; studies that reported only F1 (e.g., [17]) are noted in prose rather than in the table. Our AUC exceeds the dataset creators' reported figure by 4.2 points and also exceeds the SMOTE-balanced result of [19], while our accuracy matches the figure reported by [18]. As with any comparison across studies, reported gains remain sensitive to the evaluation protocol (split, threshold) as much as to modeling choices; we present our results as competitive with, rather than uniformly superior to, this line of work in every respect. We report AUC and accuracy on a single held-out split, consistent with the comparator studies in Table IV; a paired significance test (e.g., DeLong's test for AUC, or McNemar's test on the confusion matrices) against a matched plain-XGBoost baseline is a natural next step and is discussed as immediate future work in Section VII-D.

VII. DISCUSSION

A. Deployment as a Two-Stage SIS Module

Feature importance is available directly from XGBoost's built-in gain-based scores, without requiring a separate interpretability pass, which simplifies deployment and gives advisors a defensible basis for each flagged case. Because several first-semester features are informative on their own, we propose a two-stage design consistent with the ISHE tracks on learning-systems integration: a lower-confidence alert generated at the end of semester 1 using only first-semester and enrollment-time data, surfaced inside the institution's existing SIS or advising dashboard, followed by the full CD-based assessment at the end of semester 2 once second-semester records are available. Tuition-fee and debtor status are administratively actionable signals distinct from academic progression, reinforcing financial precarity as an independently actionable driver [9], [10]; in practice this suggests routing academically-flagged and financially-flagged students to different advising workflows rather than a single generic alert. Institutions adopting this pipeline should audit outcomes

TABLE IV
COMPARISON WITH PRIOR WORK ON THE SAME DATASET

Study	AUC	Acc.
Realinho et al. [8]	0.898	0.868
Romero & Liao [17]	—	—
Ridwan & Priyatno [18]	—	0.900
Silkhi et al. [19]	0.935	0.814
This work	**0.940**	**0.900**

across demographic sub-groups (e.g., age at enrollment, part-time or mature-entry status) using standard fairness metrics such as demographic parity and equal opportunity [16], since a statistical association in the training data does not by itself justify differential treatment of these groups in practice.

B. Cross-Institutional Applicability

All results pertain to a single Portuguese polytechnic institution. Portability depends on three factors: the retained administrative features are semester-level variables common to most student information systems, so the input schema should transfer with minor renaming; the majority-class sub-cluster structure (k and the corresponding centroids) is institution-specific and must be re-estimated on local training data (Algorithm 1, line 3, is re-run rather than reused); and the classification threshold reflects this institution's class balance and must be recalibrated locally if a threshold other than 0.5 is used. Related studies on comparable but distinct datasets report that models trained at one institution transfer with reduced but non-trivial accuracy to another [6], [13], suggesting this pipeline is a reasonable starting point for a new deployment provided the three re-fitting steps above are followed, rather than a model applied as-is.

C. Threats to Validity

Internal validity. Hyperparameters, k, and the operating threshold are all selected under stratified cross-validation on the training set only, and the test set is touched exactly once at final evaluation, limiting optimistic bias from tuning-on-test leakage. *External validity.* The single-institution, single-country nature of the dataset is the paper's main limitation; performance figures should be read as evidence the approach works on this benchmark, not as a guarantee at a different institution without re-fitting as described in Section VIIB. *Construct validity.* The binary collapse of the original three-class label treats "Enrolled" (still active, outcome unknown) and "Graduate" identically as "non-dropout"; Class Decomposition is precisely the mechanism intended to compensate for this collapse by letting the model see the internal structure it otherwise discards, but the collapse itself remains a modeling choice worth stating explicitly.

D. Limitations and Future Work

The dataset covers a single institution and semester-level aggregates only, without within-semester behavioral signals (e.g., LMS activity); the model captures statistical association, not causation, and the operating threshold will need periodic

recalibration as the enrolled population changes. Comparisons across studies using this dataset remain complicated by differing splits, thresholds, and reported metrics (Table IV), underscoring the need for a shared evaluation protocol. Immediate future work includes: (1) a controlled ablation directly quantifying the AUC/accuracy gain attributable to Class Decomposition over a matched plain-XGBoost baseline on the identical 42-feature set and split, reported with a paired significance test; (2) an empirical, not only discussed, cross-institutional transfer study once a second labeled dataset becomes available; (3) integration of within-semester LMS activity logs where available; and (4) a benchmark against modern tabular architectures (e.g., TabNet, FT-Transformer) under this same shared evaluation protocol.

VIII. Conclusion

This paper presented an XGBoost-based dropout prediction pipeline built around ten engineered ratio features and a formalized Class Decomposition strategy that partitions the majority class into homogeneous sub-clusters before training. On the Realinho et al. dataset, the model reaches an AUC of 0.940 and an accuracy of 90.0% at the default 0.5 threshold, exceeding the AUC and accuracy figures reported by prior work using the same dataset and a comparable binary framing, without relying on synthetic oversampling. Beyond the benchmark result, we position the pipeline as a two-stage module deployable within existing SIS/LMS infrastructure and outline the re-fitting steps, threats to validity, and reproducibility considerations required for adoption at a new institution. Future work should quantify the Class Decomposition strategy's contribution via a controlled ablation with paired significance testing, pursue empirical cross-institutional validation with locally re-fitted clusters and thresholds, and benchmark against modern tabular architectures under a shared evaluation protocol.

IX. Acknowledgment

The authors gratefully acknowledge the encouragement and support provided by ECI Distribution Co., Ltd., Paragon International University, the Institute of Technology of Cambodia, and ACLEDA University of Business, which contributed to the successful completion of this research. The authors also extend their sincere appreciation to colleagues, faculty members, and reviewers for their valuable insights, constructive feedback, and continuous encouragement throughout the research process. Their support significantly contributed to improving the quality, rigor, and overall impact of this study. The authors further appreciate all individuals who provided assistance, resources, and guidance during the development and evaluation of the proposed machine learning framework. Special thanks are extended to the academic and professional communities whose knowledge-sharing and collaboration helped strengthen this research and its contribution to advancing intelligent educational analytics.

References

[1] OECD, *Education at a Glance 2022: OECD Indicators*. Paris: OECD Publishing, 2022. https://doi.org/10.1787/3197152b-en

[2] V. Tinto, *Leaving College: Rethinking the Causes and Cures of Student Attrition*. Chicago, IL: Univ. of Chicago Press, 1987. https://eric.ed.gov/?id=ED371658

[3] S. Swecker, C. Fifolt, and L. Searby, "Academic advising and first-generation college students," *NACADA Journal*, vol. 33, no. 1, pp. 46–53, 2013. https://doi.org/10.12930/NACADA-13-192

[4] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern. C*, vol. 40, no. 6, pp. 601–618, 2010. https://doi.org/10.1109/TSMCC.2010.2053532

[5] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting student dropout: A case study," in *Proc. 2nd Int. Conf. Educational Data Mining*, 2009, pp. 41–50. https://www.educationaldatamining.org/EDM2009/uploads/proceedings/dekker.pdf

[6] N. Mduma, K. Kalegele, and D. Machuve, "A survey of ML approaches for student dropout prediction," *Data Science Journal*, vol. 18, no. 1, p. 14, 2019. https://doi.org/10.5334/dsj-2019-014

[7] J. Berens, K. Schneider, S. Görtz, S. Oster, and J. Burghoff, "Early detection of students at risk," *J. Educational Data Mining*, vol. 11, no. 3, pp. 1–41, 2019. https://doi.org/10.5281/zenodo.3594771

[8] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," *Data*, vol. 7, no. 11, p. 146, 2022. https://doi.org/10.3390/data7110146

[9] A. F. Cabrera, A. Nora, and M. B. Castaneda, "College persistence," *J. Higher Education*, vol. 64, no. 2, pp. 123–139, 1993. https://doi.org/10.1080/00221546.1993.11778419

[10] D. Hossler, J. Schmit, and N. Vesper, *Going to College*. Baltimore, MD: Johns Hopkins Univ. Press, 1999. https://www.scirp.org/reference/referencespapers?referenceid=2633181

[11] C. Romero and S. Ventura, "Data mining in education," *WIREs Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12–27, 2013. https://doi.org/10.1002/widm.1075

[12] C. Márquez-Vera et al., "Early dropout prediction using data mining," *Expert Systems*, vol. 33, no. 1, pp. 107–124, 2016. https://doi.org/10.1111/exsy.12135

[13] J. Kabathova and M. Drlik, "Towards predicting students' dropout using different ML techniques," *Applied Sciences*, vol. 11, no. 7, p. 3130, 2021. https://doi.org/10.3390/app11073130

[14] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017. https://doi.org/10.1109/ACCESS.2017.2654247

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785

[16] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *NeurIPS*, vol. 29, 2016, pp. 3315–3323. https://arxiv.org/abs/1610.02413

[17] S. Romero, X. Liao, N. Rathnayake, and M. Salas-Velasco, "Statistical and machine learning models for predicting university dropout and scholarship impact," *PLOS ONE*, vol. 20, no. 6, p. e0325047, 2025. https://doi.org/10.1371/journal.pone.0325047

[18] A. Ridwan and A. Priyatno, "Predict students' dropout and academic success with XGBoost," *J. Education and Computer Applications*, vol. 4, no. 2, pp. 1–8, 2024. https://jeca.aks.or.id/jeca/article/view/13

[19] H. Silkhi, B. Bakkas, and K. Housni, "Boosting academic success: Predictive modeling on imbalanced student data," in *Innovative Approaches and Applications for Sustainable Development (I2ASD 2025)*, Advances in Science, Technology & Innovation. Cham: Springer, 2026, pp. 289–292. https://doi.org/10.1007/978-3-031-97825-8_58

[20] P. G. Almeida, G. A. L. Silva, V. Santos, G. Moreira, P. Silva, and E. Luz, "Deep learning for school dropout detection," arXiv:2508.14057, 2025. https://arxiv.org/abs/2508.14057