JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Prescriptive Learning Analytics for Student Dropout: Integrating Temporal Velocity and Counterfactual Explanations in Longitudinal Data

Nurul Hidayat *, Lasmedi Afuan, and Helmi Roichatul Jannah

Department of Informatics, Faculty of Engineering, Universitas Jenderal Soedirman, Purwokerto, Central Java 53122, Indonesia;
e-mail : nurul@unsoed.ac.id; lasmedi.afuan@unsoed.ac.id; helmi.roichatul@unsoed.ac.id
* Corresponding Author : Nurul Hidayat

Abstract: Student dropout in higher education remains a persistent socioeconomic challenge, yet many predictive models reported in the literature are methodologically compromised by randomized cross-validation schemes that introduce temporal data leakage and artificially inflate predictive performance. This study proposes a longitudinal prescriptive learning analytics framework integrating three complementary methodological components: a Leave-One-Cohort-Out (LOCO) temporal validation protocol, a hybrid SMOTE-ENN class balancing strategy, and temporal velocity feature engineering derived from Learning Management System (LMS) behavioral trajectories. The framework was evaluated on a longitudinal dataset comprising 464,739 enrollment records and 77 features. Five predictive algorithms—XGBoost, LightGBM, CatBoost, Random Forest, and Logistic Regression—were comparatively assessed on a strictly isolated blind holdout cohort (2022), with CatBoost emerging as the champion estimator, achieving a PR-AUC of 0.8859, a Macro F1-Score of 0.9143, and the lowest Brier Score (0.0221), thereby demonstrating superior calibration and discriminative capability under severe class imbalance (93:7 ratio). Comprehensive ablation analysis revealed that temporal velocity features function not merely as additive predictors, but as a structural prerequisite enabling Synthetic Minority Oversampling Technique with Edited Nearest Neighbors (SMOTE-ENN) to generate high-quality synthetic boundary instances; removing these features reduced minority-class precision from 0.8302 to 0.6721. To operationalize predictive outputs into actionable intervention pathways, Diverse Counterfactual Explanations (DiCE) were implemented under a three-tier causal constraint architecture on 96 borderline high-risk students, generating 384 feasible intervention scenarios exclusively targeting forward-looking behavioral velocity metrics without constraint violations. Collectively, these findings advance the paradigm of prescriptive learning analytics by providing educational institutions with interpretable risk diagnostics and operationally feasible intervention guidance grounded in empirically validated behavioral and temporal dynamics.

Keywords: Class Imbalance; Counterfactual Explanations; Early Warning Systems; Educational Data Mining; Explainable Artificial Intelligence; Learning Analytics; Student Dropout Prediction; Temporal Data Leakage.

Received: March, 28th 2026
Revised: May, 11th 2026
Accepted: May, 12th 2026
Published: May, 21st 2026

1. Introduction

Student dropout remains a persistent challenge that threatens economic sustainability, social mobility, and the long-term effectiveness of higher education systems worldwide [1], [2]. The ongoing digital transformation in education has accelerated a paradigm shift from descriptive academic reporting toward proactive and prescriptive intervention strategies [3], [4]. In this context, the integration of Educational Data Mining (EDM) and Learning Analytics has emerged as a strategic foundation for extracting actionable pedagogical insights from large-scale student data repositories [5], [6]. Real-time predictive analytics leveraging

multimodal student interaction data is increasingly recognized as a critical mechanism for identifying academic risk before failure materializes [7], [8].

Despite the substantial predictive performance reported in prior studies, much of the existing literature still relies on randomized cross-validation protocols that inherently disregard the chronological structure of educational data [9], [10]. Such randomized evaluation schemes introduce temporal data leakage, whereby future information is inadvertently utilized to validate past observations, resulting in overly optimistic performance estimates [4], [11]. Moreover, conventional predictive frameworks often exhibit limited resilience to concept drift, particularly in the post-pandemic educational landscape characterized by evolving behavioral and cultural patterns across student cohorts [8], [12]. These methodological limitations underscore the necessity of rigorous longitudinal validation protocols capable of assessing model robustness under future cohort distributions [2], [13].

The development of reliable early-warning systems is further complicated by severe class imbalance, where students at risk of dropping out represent only a small fraction of the overall population [14]. Conventional oversampling strategies frequently distort minority-class distributions and generate overlapping decision boundaries that reduce predictive reliability [15], [16]. In addition, many existing studies remain constrained to Learning Management System (LMS) clickstream analysis while overlooking physical-spatial behavioral indicators that may capture broader dimensions of student engagement [2], [17]. Consequently, integrating temporal behavioral velocity signals derived from LMS trajectories with physical attendance proxies offers the potential to construct a more comprehensive representation of academic disengagement dynamics [2].

Another major barrier to the operational adoption of predictive analytics in educational environments is the black-box nature of highly accurate machine learning models [18], [19]. Predicting academic failure without providing interpretable causal justification offers limited practical value for institutional stakeholders seeking to design targeted interventions [20], [21]. This lack of transparency constrains predictive systems to reactive diagnostic tools rather than enabling their evolution into prescriptive decision-support systems capable of facilitating individualized educational interventions [8], [22]. Accordingly, the integration of Explainable Artificial Intelligence (XAI) becomes essential for translating complex predictive behavior into interpretable feature-level insights that can support evidence-based pedagogical decision-making and personalized academic support strategies [23], [24].

To address these limitations, this study proposes a longitudinal prescriptive learning analytics framework explicitly designed to mitigate concept drift while generating operationally feasible intervention recommendations [2], [4]. The proposed framework incorporates a Leave-One-Cohort-Out (LOCO) temporal validation strategy to evaluate predictive robustness under future cohort conditions [8], [12]. Furthermore, this study integrates a Synthetic Minority Oversampling Technique with Edited Nearest Neighbors (SMOTE-ENN) preprocessing mechanism within a multidimensional behavioral representation combining LMS interaction trajectories and physical-spatial attendance proxies to establish more reliable minority-class decision boundaries [14], [25]. To improve interpretability and intervention feasibility, the framework combines SHAP-based feature attribution with Diverse Counterfactual Explanations (DiCE), enabling the generation of individualized behavioral intervention scenarios rather than merely reporting static risk probabilities [21], [24]. The main contributions of this study are summarized as follows:

- A longitudinal student dropout prediction framework based on LOCO validation is proposed to mitigate temporal data leakage and improve robustness against cohort-level concept drift in educational data.
- This study demonstrates, through ablation and interaction analysis, that temporal behavioral velocity features are not merely additive predictors but play a structural role in enabling SMOTE-ENN to generate more discriminative minority-class boundaries under severe class imbalance conditions.
- A constrained counterfactual explanation framework integrating SHAP and DiCE is introduced to generate behavior-oriented and operationally feasible intervention scenarios for high-risk students while preserving causal and temporal consistency.

The remainder of this paper is organized as follows. Section 2 reviews related studies on student dropout prediction, temporal validation strategies, class imbalance learning, and explainable learning analytics. Section 3 presents the proposed longitudinal predictive

framework, including data preprocessing, temporal velocity feature engineering, hybrid balancing, predictive modeling, and explainability components. Section 4 discusses the experimental setup and empirical results, including ablation studies, interaction analysis, counterfactual intervention analysis, and error analysis. Finally, Section 5 concludes the paper and outlines potential directions for future research..

2. Literature Review

2.1 Dynamics of Educational Data Mining and Learning Analytics

Educational Data Mining (EDM) and Learning Analytics (LA) have evolved from retrospective descriptive reporting toward proactive predictive paradigms aimed at enabling early educational intervention. This transition has been accelerated by the increasing availability of large-scale longitudinal educational datasets that continuously record multimodal student interactions across digital learning environments [2]. Prior studies have shown that integrating heterogeneous behavioral signals—such as LMS activity logs, academic progression indicators, and physical attendance proxies—can provide a more comprehensive representation of student engagement trajectories compared with relying solely on static demographic information [2], [5], [6].

Within this context, predictive modeling is commonly formulated as a binary classification problem in which the objective is to estimate the probability of student dropout during an ongoing academic cycle. Logistic Regression remains widely adopted in educational prediction tasks due to its interpretability, probabilistic formulation, and robustness under high-dimensional feature spaces [26], [27]. In several EDM studies, this model is frequently employed as a baseline benchmark to evaluate the incremental benefits of more sophisticated machine learning architectures. The logistic regression function used to estimate dropout probability is formally expressed in Equation (1).

$$P(y = 1|x) = \frac{1}{1 + e^{-\left(\beta_0 + \beta^T \mathrm{x}\right)}} \tag{1}$$

where β denotes the weight coefficients learned from the training data [26], [27]. To improve model stability and reduce multicollinearity among overlapping temporal LMS features, L1 and L2 regularization strategies are commonly incorporated within this formulation.

2.2 Ensemble Predictive Modeling and the Challenge of Class Imbalance

Recent advances in EDM increasingly rely on ensemble learning architectures, particularly gradient boosting frameworks such as XGBoost, LightGBM, and CatBoost, due to their ability to model complex non-linear interactions within heterogeneous educational data [1], [23], [28]. These algorithms iteratively construct weak learners that minimize residual prediction errors from previous iterations, allowing them to capture latent behavioral patterns associated with academic disengagement. Prior research has consistently reported that boosting-based models outperform conventional linear classifiers in educational risk prediction tasks involving high-dimensional academic, socioeconomic, and behavioral indicators [23], [28]. Despite these performance advantages, educational prediction problems are frequently characterized by severe class imbalance, where dropout students constitute only a small minority relative to retained students [14]. Such imbalance can bias predictive models toward majority-class dominance, resulting in poor minority-class sensitivity and unreliable early-warning capability. To address this issue, numerous studies have adopted hybrid resampling techniques that combine oversampling and noise reduction mechanisms [15], [25].

Among these approaches, SMOTE-ENN has emerged as an effective strategy because it simultaneously generates synthetic minority-class instances while refining noisy decision boundaries. Specifically, Synthetic Minority Over-sampling Technique (SMOTE) creates artificial minority samples to improve class representation, whereas Edited Nearest Neighbors (ENN) removes observations that are inconsistently classified by their nearest neighbors [15], [25]. Alternative approaches such as SMOTE-Tomek primarily focus on eliminating Tomek links between classes to reduce overlap [15], [25]. Existing literature suggests that hybrid balancing strategies are particularly beneficial for educational datasets with overlapping behavioral distributions and sparse minority regions. Importantly, resampling procedures must be

executed strictly within the training folds during cross-validation to prevent synthetic data leakage into testing partitions, thereby preserving the validity of model evaluation.

2.3 Explainable Artificial Intelligence and Temporal Feature Engineering

The increasing adoption of machine learning within educational decision-making has intensified the need for transparent and interpretable predictive systems [20]. Although ensemble algorithms often achieve superior predictive performance, their black-box nature limits their operational usefulness for academic stakeholders seeking actionable intervention strategies [18], [19]. Consequently, Explainable Artificial Intelligence (XAI) techniques are increasingly integrated into EDM frameworks to improve interpretability and support evidence-based pedagogical decision-making [20], [23].

Among contemporary XAI methods, SHapley Additive exPlanations (SHAP) has become widely adopted because it provides consistent feature attribution grounded in cooperative game theory [23]. Several studies employ SHAP to identify dominant behavioral and academic predictors contributing to student dropout risk [18], [19], [29]. The Shapley value formulation used to quantify the marginal contribution of individual features is expressed in Equation (2).

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\ (|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S)) \tag{2}$$

Through this mechanism, feature contributions can be interpreted at both global and individual levels, enabling educational institutions to better understand the behavioral factors associated with academic failure. Recent studies have further explored causality-aware explainability frameworks, including the integration of structural causal model ontologies with Local Interpretable Model-Agnostic Explanations (LIME), to improve fairness and reduce explanation bias in educational decision systems [30]. However, most existing XAI applications remain primarily descriptive and are rarely extended toward prescriptive intervention analysis.

To address this limitation, recent research has begun incorporating Counterfactual Explanations to simulate alternative behavioral trajectories capable of altering predictive outcomes [31]. Counterfactual reasoning is particularly attractive in educational settings because it transforms predictive analytics into actionable intervention guidance by identifying minimal behavioral changes associated with reduced dropout risk. In parallel, temporal feature engineering approaches have demonstrated that behavioral velocity indicators derived from LMS trajectories provide earlier and more sensitive signals of academic disengagement compared with static aggregation metrics [32]. These temporal dynamics are especially relevant in longitudinal educational environments characterized by evolving engagement patterns and cohort-level behavioral drift.

2.4 Related Work and Research Gap

Prior studies on student dropout prediction can generally be grouped into several methodological categories. A dominant body of work focuses on conventional machine learning classification using static demographic variables, socioeconomic attributes, and terminal academic performance indicators to predict dropout outcomes [28], [33], [34]. Another line of research emphasizes the integration of Explainable Artificial Intelligence frameworks to identify globally important predictive features and improve interpretability within educational risk models [21], [22], [35]. More recent studies have explored temporal behavioral modeling using LMS interaction sequences and time-series learning approaches to better capture evolving student engagement patterns.

Despite reporting high predictive performance, much of the existing literature remains methodologically constrained by randomized train-test partitioning strategies, particularly conventional K-Fold cross-validation applied to aggregated historical cohorts [4], [10]. Such evaluation protocols unintentionally introduce temporal data leakage, allowing future behavioral patterns to contaminate past training distributions and thereby producing overly optimistic performance estimates. In addition, many existing studies fail to explicitly evaluate model robustness against cohort-level concept drift, especially within post-pandemic educational environments characterized by rapidly changing learning behaviors [8], [12].

Another important limitation concerns the scope of existing explainability frameworks. Most prior XAI implementations remain restricted to descriptive feature importance analysis and rarely provide operationally feasible intervention guidance at the individual level [21], [22]. Similarly, although several studies incorporate LMS behavioral features, relatively few investigate the interaction between temporal behavioral dynamics, class imbalance handling, and longitudinal validation protocols within a unified predictive framework.

Accordingly, this study establishes a distinct methodological positioning by proposing a strictly isolated temporal learning architecture designed to eliminate temporal data leakage through a deterministic LOCO validation protocol. The framework is trained exclusively on historical cohorts (2018–2021) and evaluated on a blind future cohort (2022) to assess robustness under realistic deployment conditions. Furthermore, this research integrates temporal velocity feature engineering, hybrid SMOTE-ENN balancing, and constrained counterfactual explanations within a unified prescriptive learning analytics framework. By combining SHAP diagnostics with DiCE, the proposed approach extends beyond conventional risk prediction toward operationally feasible behavioral intervention recommendations for vulnerable students.

3. Methodology

3.1. Data Acquisition, Unit of Analysis, and Target Class Definition

This study utilizes the University Student Dropout: A Longitudinal Dataset [36]. The unit of analysis is defined at the student enrollment level for each academic cohort (caca). Dataset integration was performed using composite relational identifiers (dni_hash and caca), resulting in a consolidated analytical matrix comprising 464,739 enrollment records and 77 features. To address the severe class imbalance inherent in the dataset, the target labels were explicitly resolved such that label A (Abandono) represents the positive class corresponding to Dropout (6.2%), whereas label B represents the negative class corresponding to Graduate (93.8%).

3.2. Early-Warning Prediction Point and Target Leakage Prevention

To ensure the validity of the proposed early-warning framework, the prediction point was strictly constrained to the end of the first academic year. Accordingly, the observation window only included student interaction data recorded between September (Year 1) and August (Year 2), thereby preventing the incorporation of information temporally proximate to the final academic outcome.

Target leakage mitigation was conducted through a two-stage filtering strategy. First, deterministic feature elimination was applied to remove attributes directly associated with post-outcome academic status, including impagado_curso_mat, as well as high-cardinality identifiers such as dni_hash, tit_hash, and asi_hash. Second, statistical filtering was performed by excluding features exhibiting an absolute Pearson correlation coefficient greater than 0.90 relative to the target variable. To further ensure the absence of trivially predictive variables, a Zero-R validation mechanism was implemented, preventing the inclusion of any single feature capable of independently achieving an accuracy of ≥99%.

3.3. Preprocessing and Temporal Feature Engineering

The preprocessing strategy was designed through a bifurcated imputation pipeline tailored to the characteristics of the missing data patterns. Structural missingness within Wi-Fi attendance sensor features was imputed using a sentinel value of −999, allowing tree-based ensemble models to construct dedicated partitioning structures for absent observations. In contrast, features exhibiting Missing Completely at Random (MCAR) characteristics were imputed using median substitution for numerical variables and mode substitution for categorical variables. Importantly, all imputation parameters were estimated exclusively from the training set to prevent data leakage during validation.

Categorical attributes were transformed using ordinal encoding, where unseen categories encountered during testing were explicitly mapped to a penalty value of −1 to preserve inference stability under future cohort distributions. Feature engineering primarily focused on capturing temporal behavioral dynamics through velocity-based representations of student engagement. LMS interaction activities were temporally segmented into the first and second

halves of the academic semester, enabling the extraction of behavioral progression trends rather than relying solely on cumulative interaction counts. The temporal velocity vector v was computed as follows:

$$v = \frac{\sum x_{\text{end}} + \epsilon}{\sum x_{\text{start}} + \epsilon} \tag{3}$$

where $\epsilon = 10^{-5}$ denotes a stabilizing constant introduced to avoid zero-division instability. To mitigate the influence of extreme outliers and dimensionality explosion, Winsorization was applied at the 1st and 99th percentiles prior to feature filtering. Subsequently, a zero-variance filtering procedure was employed, reducing the multidimensional feature space from 389 to 225 informative attributes characterized by non-zero standard deviation (σ>0).

3.4 Leave-One-Cohort-Out (LOCO) Validation and Class Balancing

To avoid the performance inflation commonly induced by randomized shuffling procedures, this study implemented a temporally aware validation strategy based on LOCO evaluation. The primary training tensor was constructed exclusively from historical cohorts (2018–2021), while the 2022 cohort was strictly isolated as a blind holdout dataset to evaluate predictive robustness under future cohort conditions.

To address the severe class imbalance ratio (93:7), a hybrid SMOTE-ENN resampling framework was adopted. SMOTE was configured with k_neighbors=5 to synthesize minority-class dropout instances, whereas ENN utilized n_neighbors=3 to remove ambiguous majority-class samples located near unstable decision boundaries. Prior studies have demonstrated that such hybrid balancing mechanisms are particularly effective in educational datasets characterized by overlapping behavioral distributions and sparse minority regions [15], [25]. To rigorously prevent synthetic data contamination, the entire resampling process was encapsulated within a pipeline architecture (e.g., imblearn.pipeline), ensuring that augmentation procedures were executed exclusively during the training phase (fit()) and never exposed to validation or testing partitions during cross-validation.

3.5. Model Architecture, Hyperparameter Optimization, and Calibration

Five predictive estimators—XGBoost, LightGBM, CatBoost, Random Forest, and Logistic Regression—were comparatively evaluated to assess their effectiveness under longitudinal educational prediction settings. These models were selected because they collectively represent both interpretable linear approaches and state-of-the-art ensemble learning architectures widely adopted in EDM research. Hyperparameter optimization was conducted using the Tree-structured Parzen Estimator (TPE) algorithm implemented within the Optuna optimization framework, with the objective function explicitly oriented toward predictive metric maximization.

Table 1. Training configuration and hyperparameter search space (Optuna-TPE).

Algorithm	Hyperparameter Search Space
XGBoost	learning_rate: [0.01, 0.2] (log); max_depth: [3,10]; n_estimators: [100,1000] (step 100); gamma: [10^{-8},1.0] (log); min_child_weight: [1,10]; subsample, colsample_bytree: [0.5,1.0]; reg_alpha, reg_lambda: [10^{-8},100.0] (log)
LightGBM	learning_rate: [0.01,0.2] (log); max_depth: [3,12]; num_leaves: [20, min(100, 2^{max_depth})]; n_estimators: [100,1000] (step 100); min_child_samples: [5,100]; feature_fraction: [0.4,1.0]; subsample: [0.5,1.0]; reg_alpha, reg_lambda: [10^{-8},100.0] (log)
CatBoost	learning_rate: [0.01,0.2] (log); depth: [4,10]; iterations: [100,1000] (step 100); l2_leaf_reg: [10^{-8},100.0] (log)
Random Forest	n_estimators: [100,500] (step 100); max_depth: [5,25]; min_samples_split: [2,20]; min_samples_leaf: [1,20]; max_features: {sqrt, log2, None}
Logistic Regression	C: [10^{-4},10^{2}] (log); penalty: {l2, None}; solver: saga; max_iter: 400

To preserve temporal consistency during optimization, internal validation employed a 3-fold Rolling-Origin Cross-Validation strategy rather than randomized partitioning. This approach ensured that each validation fold respected the chronological ordering of student cohorts, thereby reducing the risk of temporal leakage during model selection. The complete hyperparameter search configuration is summarized in Table 1.

To prevent the selection of models exhibiting unstable performance across temporal folds, a Variance Penalty Guardrail mechanism was incorporated during optimization. This formulation penalizes models with excessive cross-fold variability, thereby prioritizing temporal stability in addition to predictive performance. The adjusted evaluation score is defined as follows:

$$\text{Score}_{adj} = F_1 - 2.0 \cdot \left(\sigma_{F_1} - 0.05\right), \quad \text{if} \quad \sigma_{F_1} > 0.05 \tag{4}$$

This penalty mechanism was specifically designed to discourage overfitting to particular temporal sub-periods and to improve generalization under future cohort distributions. Following hyperparameter optimization, the best-performing model was calibrated using Isotonic Regression on a dedicated 20% calibration holdout subset. Calibration was accepted only when the resulting reduction in discriminative capability remained within a predefined tolerance threshold ($\Delta\text{AUC} \leq 0.010$), thereby ensuring improved probabilistic reliability without substantially compromising class separability.

3.6. Ablation Study Design

Following best practices in predictive model evaluation, a comprehensive ablation framework was designed to systematically quantify both the independent and interaction-level contributions of the principal components within the proposed pipeline. The optimal configuration (S_{FULL}) was iteratively retrained under six controlled structural deconstruction scenarios:

- A_{NO_LOCO}: replacing the LOCO temporal validation protocol with conventional randomized cross-validation to evaluate the impact of temporal data leakage;
- B_{NO_SMOTE}: disabling the hybrid resampling mechanism to assess the model's native handling of severe class imbalance;
- $C_{NO\_VELOCITY}$: removing temporal behavioral velocity features to determine the contribution of longitudinal engagement dynamics;
- $D_{\text{NO\_LOCO\_NO\_SMOTE}}$: jointly removing temporal validation and class balancing mechanisms;
- $E_{\text{NO\_SMOTE\_NO\_VELOCITY}}$: jointly removing imbalance handling and temporal behavioral dynamics; and
- $F_{\text{NO\_LOCO\_NO\_VELOCITY}}$: jointly removing temporal validation and behavioral velocity representations.

The first three scenarios were designed to isolate the individual contribution of each principal component, whereas the remaining combinatorial scenarios were intended to investigate structural interdependencies across the broader pipeline architecture. This interaction-oriented ablation strategy enables the analysis of whether specific mechanisms operate independently or synergistically. In particular, these experiments provide insight into the extent to which temporal behavioral velocity features support the effectiveness of SMOTE-ENN balancing and whether temporal validation protocols influence the stability of behavior-driven predictive representations under future cohort distributions.

In addition to the structural ablations, a feature-level robustness analysis was conducted to distinguish genuine early-warning capability from potential dependence on near-target administrative proxies. In this experiment, the optimal model was retrained using a strictly constrained feature space in which all explicit administrative status indicators (e.g., active enrollment status) and direct academic debt variables were intentionally excluded. Under this "Behavioral-Only" configuration, the model relied exclusively on temporal LMS engagement velocity features and spatial attendance proxies. The objective of this experiment was to evaluate whether latent behavioral deterioration signals could predict dropout risk independently of definitive institutional administrative records, thereby providing stronger evidence that the proposed framework captures early-stage disengagement dynamics rather than merely exploiting late-stage academic proxies.

3.7. Evaluation Metrics

Given the severe imbalance in the class distribution, conventional evaluation metrics such as Accuracy and ROC-AUC may produce misleadingly optimistic estimates of predictive performance due to their sensitivity to majority-class dominance. In educational early-warning systems, the primary objective is not merely maximizing overall classification accuracy, but accurately identifying vulnerable students belonging to the minority dropout class. Consequently, this study adopts the Area Under the Precision-Recall Curve (PR-AUC) as the primary evaluation metric because it provides a more reliable assessment of classifier performance under highly skewed distributions by emphasizing the trade-off between minority-class precision and recall.

To provide complementary perspectives on predictive performance, additional evaluation metrics were also employed. Macro F1-Score was included to evaluate the balance between precision and recall across both classes while reducing majority-class bias. Dropout-specific Precision was used to assess the reliability of positive dropout predictions, which is particularly important in educational intervention settings where excessive false alarms may lead to unnecessary institutional actions. Dropout-specific Recall (Sensitivity) was additionally measured to quantify the model's ability to correctly identify at-risk students, minimizing missed intervention opportunities. Finally, the Brier Score was incorporated to evaluate probabilistic calibration quality, ensuring that predicted probabilities remain reliable for downstream prescriptive decision-making and counterfactual intervention analysis. All final performance evaluations were conducted exclusively on the strictly isolated blind holdout dataset corresponding to the 2022 cohort, thereby ensuring temporally realistic assessment conditions and preventing contamination from historical training distributions.

3.8. Explainable Artificial Intelligence and Diverse Counterfactual Explanations (DiCE)

To improve algorithmic transparency and interpretability, SHapley Additive exPlanations (SHAP) values were computed using a randomized subsample of 1,000 instances extracted from the Blind Holdout Tensor. The consistency and mathematical integrity of the generated explanations were validated using absolute additivity constraints to ensure that the explanation residuals remained within the prescribed tolerance threshold. As a progression toward prescriptive learning analytics, the DiCE framework was implemented on the best-performing CatBoost model to generate individualized behavioral intervention scenarios. To preserve operational feasibility and causal consistency, the counterfactual search space was governed by a three-tier constraint architecture.

First, the simulation cohort was restricted to borderline high-risk students with factual dropout probabilities satisfying: $0.60 \leq P(Dropout) < 0.80$. This interval was selected to represent the intervention window in which behavioral modification remains potentially actionable.

Second, the features_to_vary parameter was strictly constrained to forward-looking behavioral indicators, specifically LMS engagement velocity metrics and current-semester interaction activities. Historical academic accumulation variables and demographic attributes were explicitly treated as immutable features to preserve temporal and causal validity.

Third, monotonic directional constraints ($\Delta \geq 0$) were imposed on all modifiable engagement features, thereby preventing the generation of counterfactual recommendations that unrealistically reduce learning activity while simultaneously lowering dropout risk. The matricula_activa attribute was additionally modeled as a categorical intervention variable, where the only permissible transition was from inactive to active status ($0 \rightarrow 1$).

Finally, all permitted behavioral perturbations were bounded within one standard deviation of the activity distribution observed among retained students. This restriction was introduced to ensure that the generated counterfactual recommendations remained behaviorally plausible and operationally achievable within realistic educational intervention settings.

4. Results and Discussion

This section presents a comprehensive evaluation of the proposed longitudinal early-warning framework. The analytical progression is organized sequentially, beginning with the characterization of the temporal cohort partitioning strategy, followed by comparative

predictive performance analysis, ablation-based robustness evaluation, interpretability assessment, and finally the simulation of prescriptive educational intervention scenarios.

4.1. Dataset Characteristics and Temporal Class Distribution

The principal methodological challenges in student dropout prediction are severe class imbalance and the risk of temporal data leakage arising from inappropriate validation protocols. To simulate realistic deployment conditions, the dataset was partitioned using the LOCO strategy. Historical cohorts from 2018–2021 were exclusively utilized for training, while the 2022 cohort was strictly isolated as a blind future test set. The statistical characteristics of the temporal partitioning scheme are summarized in Table 2.

Table 1. Dataset partition statistics under the LOCO temporal validation protocol.

Data Partition	Cohort Year	Total Samples (N)	Graduated / Active Students	Dropout Students	Dropout Ratio (%)	Number of Features
Training Set	2018–2021	305,566	286,744	18,822	6.16	225
Blind Test Set	2022	159,173	148,384	10,789	6.78	225

As shown in Table 2, the minority dropout class represents only 6.16% of the historical training population and 6.78% of the future holdout cohort. The relatively stable dropout proportion across temporal cohorts suggests that dropout constitutes a persistent structural phenomenon rather than a transient cohort-specific anomaly. Nevertheless, the severe imbalance ratio remains sufficiently extreme to bias predictive models toward majority-class dominance if no balancing intervention is applied.

Accordingly, the implementation of the hybrid SMOTE-ENN strategy during training was essential to improve minority-class representation while preserving decision boundary integrity. Unlike naive oversampling approaches, the hybrid balancing mechanism simultaneously augments minority instances and removes noisy neighborhood overlaps, thereby improving separability between dropout and retained student populations. To visually illustrate the strict temporal segregation between historical and future cohorts, the LOCO partitioning strategy is depicted in Figure 1.

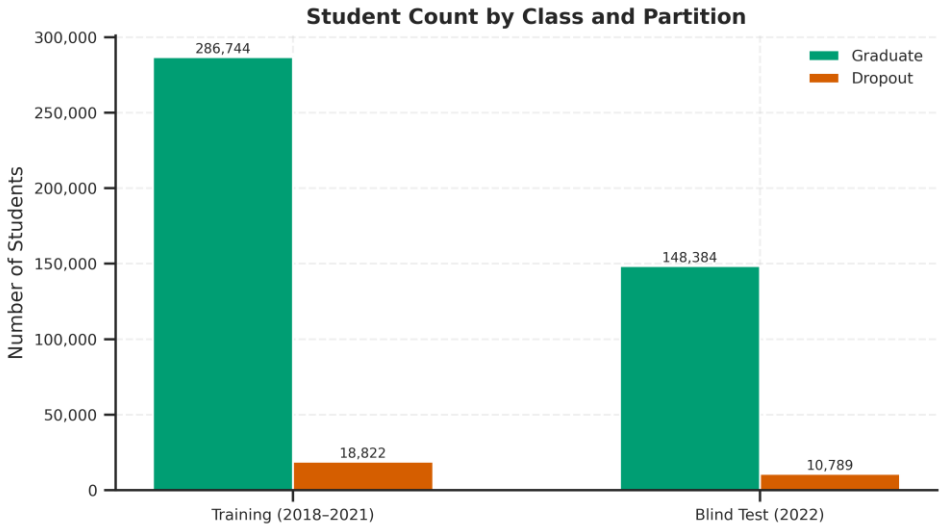

Figure 1. Temporal cohort distribution illustrating the strict isolation between the training cohorts (2018–2021) and the blind holdout cohort (2022) under the LOCO validation framework.

The explicit temporal separation illustrated in Figure 1 establishes the methodological foundation for preventing future information leakage. Consequently, all predictive performance evaluations reported in this study are derived exclusively from the unseen 2022 cohort, ensuring temporally realistic assessment conditions that more closely resemble real-world institutional deployment scenarios.

4.2. Comparative Model Performance

The proposed framework was evaluated across a diverse set of predictive architectures ranging from interpretable linear models to advanced ensemble-based learning algorithms. Given the severe class imbalance, conventional evaluation metrics such as Accuracy and ROC-AUC may produce artificially optimistic performance estimates due to the dominance of true negatives. Therefore, PR-AUC and Macro F1-Score were designated as the principal evaluation metrics because they provide a more informative assessment of minority-class predictive capability and class balance performance. The comparative performance of all evaluated models on the blind holdout cohort is presented in Table 3.

Table 3. Comparative predictive performance on the blind holdout cohort (2022).

Model	PR-AUC	ROC-AUC	Macro F1-Score	Precision (Dropout)	Recall (Dropout)	Brier Score (↓)
CatBoost	0.8859	0.9573	0.9143	0.8302	0.8507	0.0221
LightGBM	0.8809	0.9410	0.8147	0.5373	0.8628	0.0313
XGBoost	0.8640	0.9496	0.7890	0.4820	0.8603	0.0597
Random Forest	0.8255	0.9605	0.8342	0.9905	0.5242	0.0259
Logistic Regression	0.5707	0.9108	0.7255	0.7544	0.3502	0.0413

As summarized in Table 3, CatBoost emerged as the best-performing model across the primary evaluation criteria, achieving the highest PR-AUC (0.8859) and Macro F1-Score (0.9143). These results indicate that CatBoost provides the most balanced trade-off between minority-class sensitivity and prediction reliability under highly imbalanced longitudinal conditions. Although Random Forest achieved the highest ROC-AUC (0.9605) and an exceptionally high dropout precision (0.9905), its Recall for dropout detection remained substantially lower (0.5242). This indicates that the model adopted an overly conservative decision boundary, correctly identifying only a limited subset of at-risk students. In contrast, CatBoost maintained a substantially better balance between Recall (0.8507) and Precision (0.8302), successfully identifying the majority of vulnerable students while preserving a manageable false-positive rate.

The probabilistic reliability of the evaluated models was further assessed using the Brier Score. CatBoost achieved the lowest calibration error (0.0221), indicating that its predicted dropout probabilities closely aligned with empirical outcome frequencies. This property is particularly important in prescriptive educational systems where intervention prioritization depends not only on classification accuracy but also on reliable probability estimation. The comparative PR-AUC and ROC-AUC curves are illustrated in Figure 2.

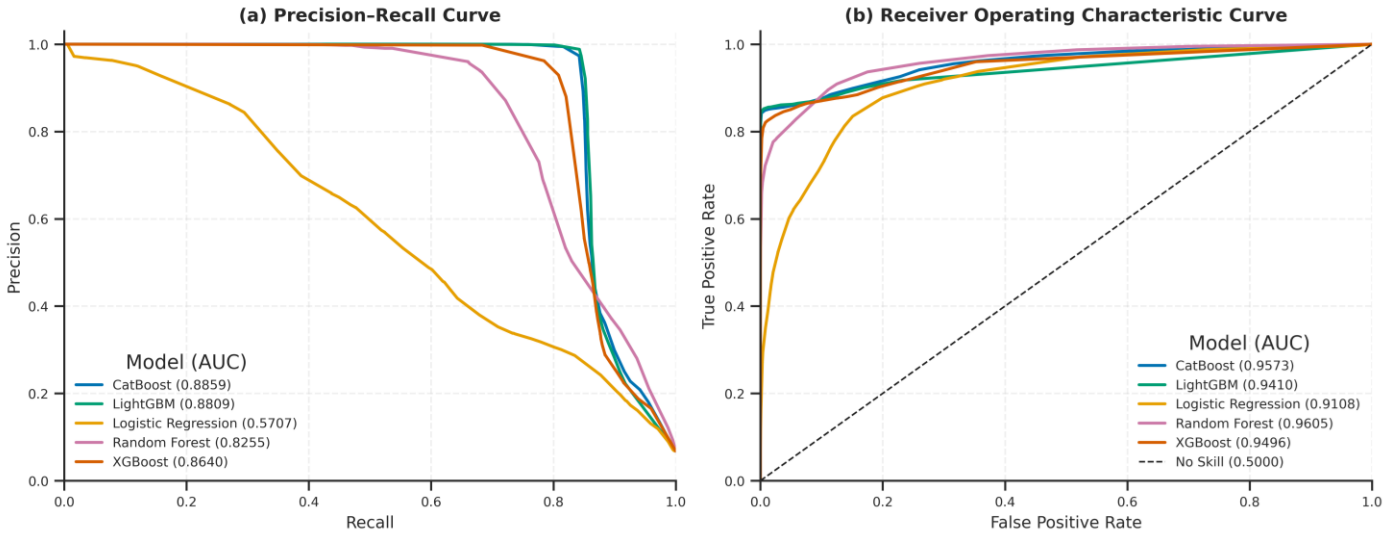

Figure 2. Comparative PR-AUC and ROC-AUC performance curves highlighting the superiority of the CatBoost model in identifying minority-class dropout patterns relative to other baseline estimators.

The superior performance of CatBoost can likely be attributed to its ordered target encoding mechanism and robust resistance to overfitting when processing high-dimensional categorical educational data. Furthermore, its ability to effectively integrate synthetic minority representations generated by SMOTE-ENN without substantially degrading structural generalization suggests that CatBoost is particularly well-suited for longitudinal educational datasets characterized by sparse minority regions and heterogeneous behavioral interactions.

4.3. Ablation Study

The methodological contribution of this study is fundamentally grounded in the integration of three principal components: the LOCO temporal validation protocol, hybrid SMOTE-ENN class balancing, and temporal behavioral velocity feature engineering. To verify that these components constitute empirical necessities rather than arbitrary architectural choices, a comprehensive ablation analysis was conducted. The complete framework configuration, denoted as S_{FULL}, was systematically compared against both isolated and combinatorial structural deconstruction scenarios. The resulting performance metrics are summarized in Table 4.

Table 4. Ablation study results across isolated and interaction-based structural deconstruction scenarios.

Scenario	PR-AUC	Macro F1-Score	Precision (DO)	Recall (DO)	Brier Score (↓)
S_{FULL}	0.8859	0.9143	0.8302	0.8507	0.0221
A_{NO_LOCO}	0.9657	0.9550	0.9064	0.9255	0.0084
B_{NO_SMOTE}	0.8865	0.9004	0.7873	0.8445	0.0267
$C_{NO\_VELOCITY}$	0.8815	0.8654	0.6721	0.8523	0.0344
$D_{\text{NO\_LOCO\_NO\_SMOTE}}$	0.9775	0.9663	0.9415	0.9325	0.0061
$E_{\text{NO\_SMOTE\_NO\_VELOCITY}}$	0.8843	0.8868	0.7371	0.8514	0.0288
$F_{\text{NO\_LOCO\_NO\_VELOCITY}}$	0.9651	0.9550	0.9119	0.9198	0.0084

The results presented in Table 4 reveal several important structural characteristics of the proposed framework. Most notably, all scenarios excluding the LOCO validation protocol (A_{NO_LOCO}, $D_{\text{NO\_LOCO\_NO\_SMOTE}}$, and $F_{\text{NO\_LOCO\_NO\_VELOCITY}}$) exhibited substantial artificial performance inflation, with PR-AUC values increasing to the range of 0.9651–0.9775. This phenomenon provides strong empirical evidence of temporal data leakage under randomized validation settings, where future behavioral distributions unintentionally contaminate historical training partitions. Consequently, the model effectively memorizes cohort-specific patterns rather than learning genuinely transferable dropout representations.

Beyond validating the necessity of temporal isolation, the ablation analysis also reveals important interaction effects between the SMOTE-ENN balancing mechanism and temporal behavioral velocity features. The relative degradation across evaluation metrics is further visualized in the heatmap presented in Figure 3.

Referring jointly to Figure 3 and Table 4, the isolated removal of temporal velocity features ($C_{NO\_VELOCITY}$) produced the most severe deterioration in minority-class precision, decreasing from 0.8302 to 0.6721. Interestingly, the simultaneous removal of both SMOTE-ENN and velocity features ($E_{\text{NO\_SMOTE\_NO\_VELOCITY}}$) resulted in higher Precision (0.7371) and Macro F1-Score (0.8868) compared with the velocity-only ablation scenario.

This non-linear degradation pattern suggests that temporal behavioral velocity features are not merely additive predictors. Rather, they appear to provide the structural distinctiveness necessary for SMOTE-ENN to synthesize minority-class samples that remain mathematically separable from overlapping majority regions. In the absence of dynamic behavioral signals, synthetic oversampling may inadvertently generate noisy or overlapping minority representations, thereby reducing classification precision. These findings indicate that temporal behavioral trajectories play a synergistic role in stabilizing class boundary formation under highly imbalanced educational data conditions.

More importantly, this result suggests that the effectiveness of data-level balancing strategies in educational prediction tasks is strongly dependent on the representational quality of

the underlying feature space. In highly static representations, synthetic oversampling may amplify overlap between vulnerable and retained students, whereas temporally dynamic behavioral features provide the separability structure necessary for generating informative minority-class samples. This finding highlights that feature engineering and imbalance handling should not be treated as independent optimization stages, but rather as mutually dependent components within longitudinal learning architectures.

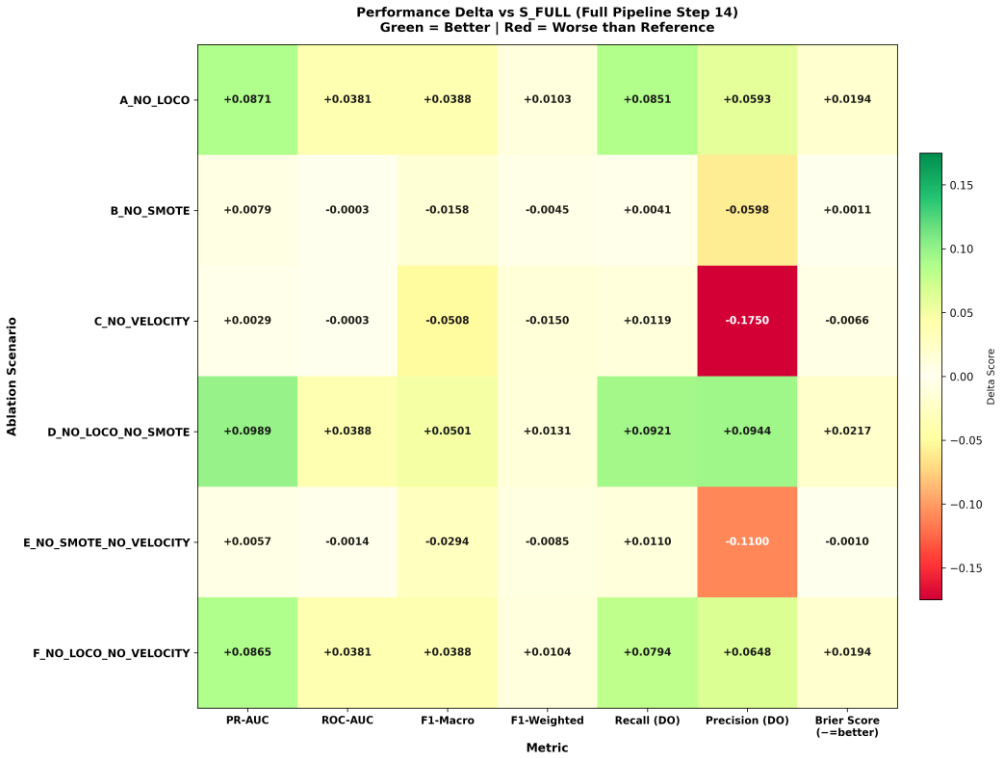

Figure 3. Heatmap illustrating relative performance degradation across isolated and combinatorial ablation scenarios with respect to S_{FULL} configuration.

To further evaluate whether the proposed framework genuinely captures early-stage disengagement dynamics rather than relying exclusively on administrative proxies, an additional "Behavioral-Only" robustness analysis was conducted. In this constrained configuration, all explicit administrative status variables and direct academic debt indicators were removed, forcing the model to rely exclusively on LMS interaction velocity and spatial attendance proxies. Although predictive performance declined relative to the full model configuration, the resulting classifier still maintained predictive capability substantially above random baseline performance. This finding suggests that latent behavioral deterioration patterns contain meaningful early-warning signals even in the absence of definitive institutional administrative indicators.

4.4. Global Interpretability and Feature Significance

Despite their predictive strength, high-capacity ensemble models often face adoption barriers in educational environments due to their limited interpretability. To improve transparency and facilitate evidence-based institutional decision-making, an Explainable Artificial Intelligence (XAI) framework based on SHapley Additive exPlanations (SHAP) was integrated into the proposed system. The top-ranked features according to their global SHAP contribution scores are presented in Table 5. As summarized in the global SHAP ranking, the most influential predictors are dominated by an interaction between administrative status indicators and longitudinal academic progression variables. The feature matricula_activa emerged as the strongest global predictor, indicating that active enrollment status constitutes the most proximal institutional signal associated with student retention. This finding aligns with operational educational realities, where failure to maintain active registration frequently precedes formal attrition events.

Table 5. Top ten features ranked by mean absolute SHAP contribution.

Rank	Feature Name	Domain Category	Mean Absolute SHAP Value
1	matricula_activa	Administrative / Other	0.07947
2	cred_pend_sup_tit	Academic Progress	0.07239
3	cred_sup_tit	Academic Progress	0.06535
4	cred_mat_normal	Academic Progress	0.03378
5	rendimiento_cuat_b	Academic Progress	0.02827
6	cred_mat1	Academic Progress	0.01822
7	curso_mas_alto	Academic Progress	0.01721
8	cred_sup_2o	Academic Progress	0.01533
9	anyo_ingreso	Historical / Other	0.01411
10	curso_mas_bajo	Academic Progress	0.01116

The remaining dominant predictors are primarily associated with accumulated academic debt and progression imbalance, particularly the volume of pending credits (cred_pend_sup_tit and cred_sup_tit) and semester-level academic performance indicators such as rendimiento_cuat_b. Collectively, these variables suggest that the predictive framework learns progressive academic deterioration trajectories rather than isolated performance events. To further examine the directional influence and distributional behavior of these variables, the SHAP summary plot for the CatBoost model is presented in Figure 4.

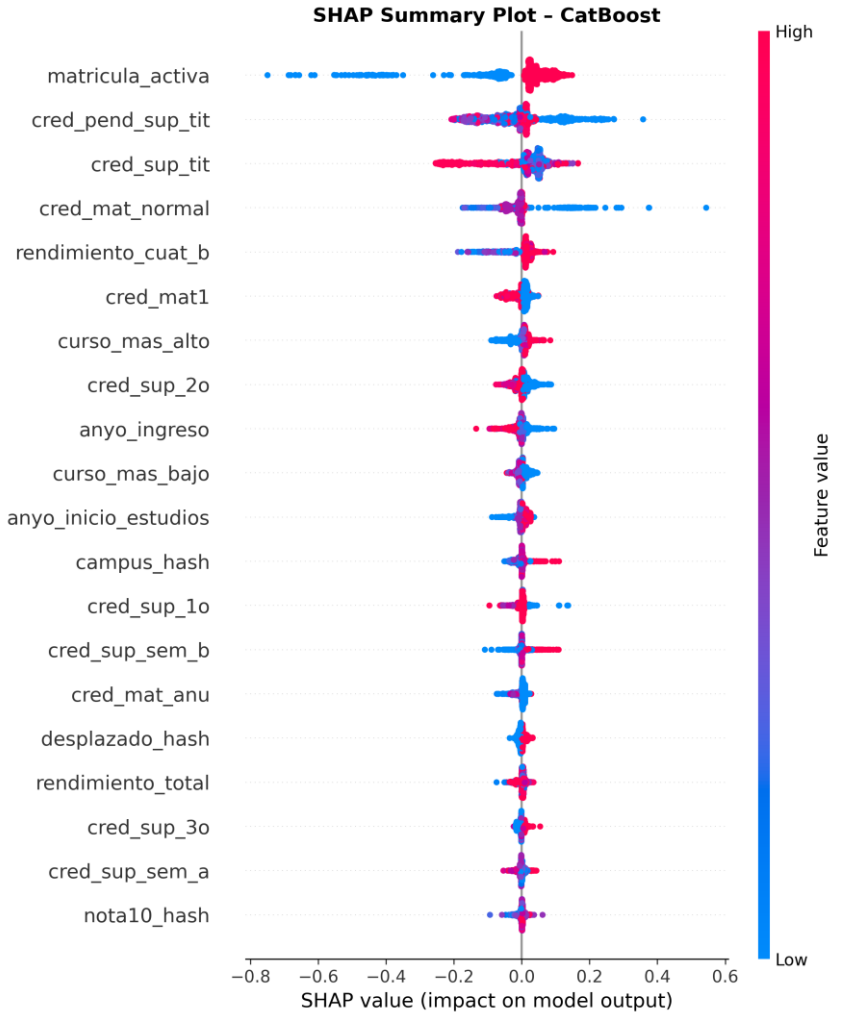

Figure 4. SHAP summary plot illustrating the distributional impact and directional contribution of features toward dropout prediction in the CatBoost model.

As illustrated in Figure 4, the SHAP decomposition not only quantifies global feature importance but also explicitly captures the directional relationship between feature values and dropout probability. Lower values of matricula_activa (blue distribution), corresponding to inactive enrollment conditions, strongly shift predictions toward elevated dropout risk through positive SHAP contributions. Similarly, high values of pending credit accumulation variables, particularly cred_pend_sup_tit, substantially increase predicted vulnerability to academic attrition.

Importantly, the interpretability results demonstrate that the predictive behavior of the proposed framework remains consistent with established pedagogical and institutional logic. Rather than relying on opaque statistical correlations, the model appears to capture coherent academic progression patterns associated with disengagement and dropout vulnerability, thereby improving the operational trustworthiness of the proposed early-warning system. Interestingly, although administrative and academic progression variables remain dominant predictors, the SHAP distributions also reveal that dropout risk is not determined by a single isolated feature, but rather emerges from cumulative interaction patterns across behavioral, academic, and institutional dimensions. This observation reinforces the necessity of multidimensional longitudinal modeling in educational analytics, particularly in environments characterized by evolving cohort behavior and post-pandemic learning adaptation.

4.5. Strict Early-Warning Capability Analysis

To rigorously evaluate whether the proposed framework functions as a genuine early-warning system rather than merely reflecting administrative status proxies, an additional robustness analysis was conducted. As demonstrated in the global interpretability analysis (Table 5), several dominant predictors—including matricula_activa and academic debt indicators such as cred_pend_sup_tit—represent highly proximal institutional signals associated with eventual attrition. Although these variables provide substantial predictive power, they may also reflect late-stage administrative realities rather than early behavioral disengagement dynamics.

To isolate the model's capability to detect authentic early-warning signals, the CatBoost classifier was retrained under a strictly constrained "Behavioral-Only" feature configuration. All administrative status indicators and direct academic progression variables were intentionally excluded. Under this setting, the model relied exclusively on temporal LMS behavioral velocity features, digital interaction trajectories, and Wi-Fi-based spatial attendance proxies. The comparative results between the complete framework and the constrained behavioral-only configuration are summarized in Table 6.

Table 6. Comparative evaluation of full-feature and behavioral-only configurations.

Feature Configuration	PR-AUC	ROC-AUC	Macro F1-Score	Precision (DO)	Recall (DO)	Brier Score (↓)
Full Features	0.8859	0.9573	0.9143	0.8302	0.8507	0.0221
Behavioral-Only Features	0.3077	0.7887	0.6532	0.3914	0.3144	0.0624

As expected, removing explicit administrative and academic progression indicators resulted in a substantial reduction in predictive certainty relative to the full-feature baseline. PR-AUC declined to 0.3077, while minority-class Precision decreased to 0.3914. From a methodological perspective, however, this degradation is structurally informative rather than problematic. Despite the exclusion of highly deterministic academic signals, the behavioral-only configuration maintained an ROC-AUC of 0.7887, substantially exceeding the random baseline threshold (0.5000). This finding provides empirical evidence that latent behavioral deterioration—captured through LMS engagement velocity and spatial attendance dynamics—contains meaningful predictive information prior to formal institutional status changes. In other words, the proposed framework does not merely reproduce administrative classifications; rather, it captures measurable behavioral momentum associated with early-stage disengagement trajectories.

From an operational perspective, this result is particularly important because behavioral interaction signals are often observable substantially earlier than definitive academic failure

indicators. Consequently, the framework demonstrates potential utility not only for retrospective risk identification, but also for proactive intervention scheduling during the early phases of disengagement. Nevertheless, the observed performance decline also indicates that administrative and academic progression variables still contribute substantial predictive information, suggesting that fully behavior-only early-warning systems remain challenging in heterogeneous educational environments.

4.6. Counterfactual-Based Intervention Prescriptions

Transforming predictive insights into actionable educational interventions represents the ultimate objective of the proposed prescriptive learning analytics framework. Rather than merely reporting dropout probabilities, the system leverages DiCE to generate "virtual recovery" scenarios that identify feasible behavioral modifications capable of shifting students from high-risk regions toward safer predictive boundaries. Under the constraint architecture defined in the methodology, DiCE was applied to a cohort of borderline high-risk students satisfying: $0.60 \leq P(Dropout) < 0.80$. This process generated 384 feasible counterfactual scenarios across 96 student instances, with up to four alternative intervention pathways produced per student. Across all generated scenarios, the algorithm successfully identified risk-reducing interventions in 372 out of 384 cases (96.9%). The mean predicted dropout probability decreased from 0.6076 to approximately 0.485 across successful counterfactuals, corresponding to a mean absolute reduction of 12.7 percentage points (Median = 12.1 pp, Min = 0.0 pp, Max = 34.4 pp).

Importantly, these constrained counterfactual recommendations produced moderate yet operationally plausible reductions in predicted risk, contrasting sharply with the unrealistically large probability shifts frequently observed in unconstrained counterfactual simulations. Aggregate feature-level perturbation statistics across all feasible scenarios are summarized in Table 7.

Table 7. Aggregate statistics of constrained counterfactual interventions for borderline high-risk students (96 students; 384 counterfactual scenarios).

Feature	Description	Frequency	Mean Δ	Median Δ	Min Δ	Max Δ
n_resource_days_velocity	Resource access day velocity	281	+0.31	+0.23	+0.06	+0.70
resource_events_velocity	Resource interaction event velocity	275	+0.36	+0.23	+0.01	+1.02
pft_assignment_submissions_velocity	Assignment submission velocity	208	+0.23	+0.23	+0.23	+0.23
pft_days_logged_velocity	Platform login day velocity	201	+0.31	+0.37	+0.03	+0.50
pft_visits_velocity	Platform visit velocity	180	+0.36	+0.39	+0.19	+0.90
n_wifi_days_velocity	Wi-Fi connection day velocity	152	+10.59	+3.58	+0.20	+52.45
pft_total_minutes_velocity	Total platform engagement time velocity	148	+0.23	+0.18	+0.01	+0.91
pft_events_velocity	Total platform event velocity	144	+0.18	+0.18	+0.09	+1.00

Note: Aggregate statistics were derived from 384 counterfactual scenarios generated under strict monotonic directional constraints ($\Delta \geq 0$). Frequency denotes the number of scenarios in which a feature was modified. The absence of negative perturbation values across all engagement features confirms adherence to the imposed pedagogical validity constraints.

As shown in Table 7, the generated interventions exclusively target forward-looking behavioral velocity indicators, while historical academic debt and demographic variables remain absent from the modification space. The two most frequently modified variables—n_resource_days_velocity (281 occurrences) and resource_events_velocity (275 occurrences)—suggest that insufficient acceleration in learning resource engagement, rather than merely low absolute engagement volume, constitutes the dominant modifiable risk factor identified by the framework.

Importantly, the required intervention magnitudes remain relatively moderate. For example, the median perturbation for both n_resource_days_velocity and resource_events_velocity is approximately +0.23 units, corresponding to behavioral activity ranges commonly observed among lower-risk students. These findings provide empirical support for the operational feasibility of the generated recommendations. Nevertheless, the n_wifi_days_velocity feature exhibited substantially higher variance, with a mean perturbation of +10.59 compared with a median of +3.58. This discrepancy indicates a positively skewed distribution driven by a limited subset of students exhibiting significantly larger engagement deficits. Such variability highlights the importance of interpreting counterfactual outputs as individualized directional guidance rather than universally standardized intervention prescriptions.

More broadly, the counterfactual analysis demonstrates that dropout vulnerability emerges as a dynamic behavioral process rather than a fixed academic state. Consequently, the proposed framework extends beyond static risk classification by providing institutions with behavior-oriented intervention pathways grounded in temporally constrained and pedagogically plausible behavioral adjustments.

4.7. Error Analysis and Failure Patterns

To better understand the operational limitations of the proposed framework and identify opportunities for future refinement, a detailed error analysis was conducted on the 2022 blind holdout cohort (N=159,173). Specifically, the analysis focused on False Negatives (FN; n=1,722), representing undetected dropout students, and False Positives (FP; n=1,637), representing incorrectly flagged retained students. Comparative feature statistics across prediction outcomes are summarized in Table 8.

Table 8. Mean feature values across prediction outcomes.

Feature	Description	TP (Correct DO)	FN (Missed DO)	TN (Correct Active)	FP (False DO)
matricula_activa	Active administrative enrollment status	0.00	0.73	0.77	0.99
cred_sup	Total accumulated passed credits	251.12	284.45	352.33	420.08
cred_sup_normal	Normal passed credits	142.59	157.85	224.63	272.09
rend_total_antepenultimo	Historical academic performance index	96.61	96.28	96.53	99.86
pft_events_velocity	LMS interaction velocity	0.91	0.92	0.93	1.00

False Negatives constitute the most critical failure mode within educational early-warning systems because these students remain undetected despite eventually disengaging from the institution. As shown in Table 8, correctly identified dropout students (TP) exhibited fully inactive enrollment status (matricula_activa = 0.00), whereas False Negatives retained substantially higher administrative activity (matricula_activa = 0.73). This indicates that most undetected dropout students remained formally active at the prediction cutoff despite subsequently withdrawing from the institution.

In addition, False Negatives demonstrated slightly stronger academic progression indicators relative to True Positives (cred_sup_normal: 157.85 vs. 142.59), suggesting that these students had not yet exhibited sufficiently severe deterioration in the dominant predictive features. This pattern reveals that the framework still retains partial structural dependence on explicit institutional indicators and encounters difficulty identifying "silent dropout" trajectories—students who remain formally enrolled and academically capable while internally disengaging from the learning process.

These findings suggest that future predictive architectures should incorporate more granular behavioral deterioration indicators, particularly temporal changes in LMS engagement dynamics, rather than relying predominantly on definitive administrative transitions. False Positives represent a different operational challenge because they may divert institutional intervention resources toward students who are not genuinely at risk. The FP profile

in Table 8 reveals substantially higher accumulated credits relative to correctly retained students (cred_sup: 420.08 vs. 352.33), while simultaneously maintaining near-perfect enrollment activity (matricula_activa = 0.99) and strong LMS engagement velocity (pft_events_velocity = 1.00).

This pattern suggests the presence of "graduation confusion," where the predictive model associates very high academic accumulation with imminent institutional departure, thereby conflating graduation trajectories with dropout trajectories. Because both graduating and dropout students may disappear from subsequent enrollment records, the model occasionally interprets imminent graduation as potential attrition risk. Operationally, this phenomenon may lead institutions to allocate retention interventions toward students who are actively progressing toward graduation. Accordingly, future work should consider refining the binary target structure to explicitly distinguish graduation from attrition outcomes or incorporating post-processing heuristic filters that exclude students already meeting institutional graduation thresholds from the intervention recommendation pool.

4.8 Discussion

The experimental findings establish the CatBoost architecture as the most effective predictive estimator within the proposed longitudinal early-warning framework, achieving a PR-AUC of 0.8859 and a Macro F1-Score of 0.9143 on the blind holdout cohort. The superior performance of gradient boosting architectures in handling heterogeneous educational datasets is consistent with prior findings reported in [17] and [37], which demonstrated the effectiveness of ensemble boosting algorithms for educational prediction tasks characterized by complex behavioral and academic interactions. Furthermore, the prioritization of PR-AUC over ROC-AUC directly addresses the concerns raised in [38], where ROC-AUC was shown to produce overly optimistic performance interpretations under severe class imbalance conditions.

Beyond predictive performance, the present study emphasizes temporal robustness as a central methodological requirement in educational analytics. In contrast to prior studies that predominantly rely on randomized partitioning schemes, the proposed framework adopts a strictly isolated LOCO validation protocol to mitigate temporal data leakage and evaluate future cohort generalization [11], [13]. The ablation analysis demonstrated that removing the LOCO protocol resulted in substantial artificial performance inflation, particularly in PR-AUC and minority-class precision. This finding strongly suggests that many previously reported high-performing educational prediction systems may partially reflect cohort memorization rather than genuine longitudinal generalization capability. Consequently, temporally isolated validation should be regarded not merely as an optional evaluation strategy, but as a methodological prerequisite for realistic deployment-oriented educational prediction research.

The combinatorial ablation analysis further provides important insight into the structural interaction between temporal behavioral velocity features and hybrid imbalance handling. The substantial precision degradation observed under the isolated removal of velocity features ($C_{NO\_VELOCITY}$), relative to the simultaneous removal of both velocity and SMOTE-ENN ($E_{\text{NO\_SMOTE\_NO\_VELOCITY}}$), indicates that temporal dynamics provide the separability structure necessary for synthetic minority generation to remain effective. In static feature spaces, oversampling mechanisms may inadvertently amplify overlap between vulnerable and retained students, thereby degrading minority-class precision. Conversely, temporal behavioral trajectories provide discriminatory boundaries that stabilize the synthetic sample generation process. These findings suggest that feature engineering and imbalance handling should not be treated as independent optimization stages, but rather as mutually dependent components within longitudinal educational learning architectures.

The behavioral-only robustness analysis additionally highlights a fundamental trade-off within predictive learning analytics: the balance between predictive lead time and diagnostic certainty. Administrative and academic progression variables provide strong predictive signal because they closely reflect institutional outcome states; however, these variables often emerge relatively late within the disengagement trajectory, limiting the available intervention window. In contrast, behavioral velocity indicators derived from LMS interactions and spatial attendance proxies capture earlier manifestations of disengagement, albeit with higher uncertainty. The principal contribution of the S_{FULL} architecture therefore lies in synthesizing

these complementary dimensions: behavioral dynamics enable earlier detection capability, whereas structural academic variables stabilize final probabilistic estimation without entirely sacrificing intervention lead time.

From an interpretability perspective, the integration of SHAP-based feature attribution and DiCE extends the framework beyond conventional black-box predictive analytics. Previous studies have repeatedly emphasized the limitations of opaque educational AI systems that provide risk predictions without actionable explanatory context [22], [31]. In response, the proposed framework not only quantifies dropout probability but also generates behavior-oriented intervention pathways grounded in interpretable feature dynamics. Consequently, the system supports a transition from passive institutional observation toward transparent, personalized, and proactive educational intervention strategies [35].

The error analysis, however, reveals several important structural limitations with direct operational implications. The observed "graduation confusion" phenomenon—where highly active senior students with large accumulated credit loads are occasionally classified as dropout risks—highlights a limitation of binary target formulations that implicitly conflate institutional exit caused by graduation with attrition-related withdrawal. Simultaneously, the reduced sensitivity toward "silent dropouts," many of whom remained administratively active at the prediction cutoff, indicates that the model still retains partial dependence on lagging institutional indicators. These findings demonstrate that strong aggregate predictive metrics alone do not guarantee contextual correctness at the individual level. Future educational prediction frameworks should therefore explicitly disentangle graduation trajectories from dropout trajectories through refined target engineering and post-processing filtering mechanisms.

An additional methodological refinement introduced in this study concerns the constrained counterfactual architecture applied within the DiCE framework. Unconstrained counterfactual generation is inherently vulnerable to exploiting spurious correlations learned by the predictive model, potentially producing pedagogically implausible recommendations—such as reducing educational engagement to lower dropout probability. To mitigate this issue, the proposed framework enforced monotonic directional constraints ($\Delta \geq 0$) across all engagement-related features while restricting modifications exclusively to forward-looking behavioral velocity indicators. The complete absence of negative perturbation values across all generated feature modifications empirically confirms that the imposed constraint architecture successfully eliminated contradictory recommendations.

Nevertheless, two residual limitations remain important. First, 12 out of 384 counterfactual scenarios (3.1%) produced no measurable reduction in predicted risk, suggesting that behavioral interventions alone may be insufficient for certain student profiles and that broader institutional support mechanisms may still be required. Second, the relatively moderate mean risk reduction (12.7 percentage points) reflects the intentionally constrained intervention setting focused on borderline high-risk students, where realistic behavioral perturbations are expected to produce incremental rather than dramatic probability shifts. Consequently, the constrained DiCE framework should be interpreted primarily as a targeted decision-support mechanism for early-stage intervention guidance rather than a universal risk mitigation solution.

5. Conclusions

This study proposed and validated a temporally aware prescriptive learning analytics framework for student dropout prediction using longitudinal educational data. The primary objective was to develop an early-warning architecture capable of mitigating temporal data leakage while preserving predictive reliability under severe class imbalance conditions. To address these challenges, the framework integrated three principal components: Leave-One-Cohort-Out (LOCO) temporal validation, hybrid SMOTE-ENN balancing, and temporal behavioral velocity feature engineering derived from LMS and spatial attendance trajectories.The experimental results demonstrate that the proposed framework achieved strong predictive performance on the strictly isolated 2022 blind holdout cohort, with CatBoost emerging as the optimal estimator (PR-AUC = 0.8859; Macro F1-Score = 0.9143). The ablation analysis further confirmed that the LOCO protocol substantially reduced the risk of artificially inflated evaluation caused by temporal leakage, while the interaction analysis revealed that temporal behavioral velocity features play a critical role in stabilizing the effectiveness of SMOTE-ENN under highly imbalanced conditions.

Beyond predictive performance, the study also demonstrated that meaningful early-warning signals can be extracted from behavioral interaction trajectories prior to formal institutional status changes. Although predictive performance declined under the behavioral-only configuration, the resulting model retained discriminatory capability substantially above random baseline performance, suggesting that LMS engagement dynamics and spatial attendance velocity contain measurable indicators of early disengagement.The integration of Explainable Artificial Intelligence (XAI) further extended the framework toward prescriptive educational analytics. SHAP analysis provided transparent interpretation of the dominant predictive factors, while the constrained DiCE framework generated behavior-oriented intervention pathways focused exclusively on forward-looking engagement variables. The generated recommendations remained operationally plausible under monotonic pedagogical constraints and consistently avoided contradictory intervention directions.

Nevertheless, several limitations remain. The error analysis revealed that the model occasionally confuses high-credit graduating students with dropout trajectories, indicating limitations in the current binary target definition. In addition, "silent dropout" cases remain difficult to identify because many students retain administratively active status despite progressive disengagement. Furthermore, the present framework remains dependent on structured institutional records and does not yet incorporate psychosocial or motivational indicators that may further improve early detection sensitivity.

Future research should therefore focus on refining target engineering to explicitly distinguish graduation from attrition trajectories, integrating richer real-time behavioral indicators, and incorporating Natural Language Processing approaches to capture latent psychosocial disengagement signals from unstructured educational interactions. Additional multi-institutional validation studies are also necessary to evaluate the generalizability and robustness of the proposed framework across diverse educational environments.

Author Contributions: Conceptualization: N.H. and L.A.; Methodology: N.H. and L.A.; Software: L.A.; Validation: N.H., L.A. and H.R.J.; Formal analysis: N.H.; Investigation: L.A.; Resources: N.H.; Data curation: H.R.J.; Writing—original draft preparation: N.H., L.A. and H.R.J; Writing—review and editing: N.H.; Visualization: N.H.; Supervision: N.H.; Project administration: L.A.; funding acquisition: N.H. and L.A. All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Acknowledgments: The authors would like to express their sincere gratitude to all parties who contributed to the completion of this research. Special appreciation is addressed to the institution and colleagues who provided valuable support, guidance, and constructive feedback throughout the research process. The authors also thank the respondents and all participants involved in data collection and validation activities.

Conflicts of Interest: The authors declare no conflict of interest

References

[1] B. Duro, A. Gomes, F. B. Correia, A. R. Borges, and J. Bernardino, "Machine Learning and Deep Learning for Dropout Prediction in Higher Education: A Review," *Computers*, vol. 15, no. 3, p. 164, Mar. 2026, doi: 10.3390/computers15030164.

[2] A. Igualde-Sáez *et al.*, "University Student Dropout: A Longitudinal Dataset of Demographic, Socioeconomic, and Academic Indicators," *Data*, vol. 10, no. 10, p. 162, Oct. 2025, doi: 10.3390/data10100162.

[3] I. Elbouknify *et al.*, "AI-based identification and support of at-risk students: A case study of the Moroccan education system," *ArXiv*. Apr. 09, 2025. [Online]. Available: http://arxiv.org/abs/2504.07160

[4] A. Shaikhanova, O. Kuznetsov, K. Iklassova, A. Tokkuliyeva, and L. Sugurova, "Interpretable Predictive Modeling for Educational Equity: A Workload-Aware Decision Support System for Early Identification of At-Risk Students," *Big Data Cogn. Comput.*, vol. 9, no. 11, p. 297, Nov. 2025, doi: 10.3390/bdcc9110297.

[5] Y. Lin, H. Chen, W. Xia, F. Lin, Z. Wang, and Y. Liu, "A Comprehensive Survey on Deep Learning Techniques in Educational Data Mining," *Data Sci. Eng.*, vol. 10, no. 4, pp. 564–590, Dec. 2025, doi: 10.1007/s41019-025-00303-z.

[6] R. Paul, S. Sarker, H. El Aouifi, S. Hussain, A. K. Baruah, and S. Gaftandzhieva, "Analyzing dropout of students and an explainable prediction of academic performance utilizing artificial intelligence techniques," *Front. Educ.*, vol. 10, Dec. 2025, doi: 10.3389/feduc.2025.1698505.

[7] W. Chango, J. A. Lara, R. Cerezo, and C. Romero, "A review on data fusion in multimodal learning analytics and educational data mining," *WIREs Data Min. Knowl. Discov.*, vol. 12, no. 4, Jul. 2022, doi: 10.1002/widm.1458.

[8] W. Dai *et al.*, "Learning Analytics for Early Identification of At-Risk Students and Feedback Intervention," *J. Learn. Anal.*, vol. 12, no. 3, pp. 102–125, Nov. 2025, doi: 10.18608/jla.2025.8735.
[9] L. Sasse *et al.*, "Overview of leakage scenarios in supervised machine learning," *J. Big Data*, vol. 12, no. 1, p. 135, May 2025, doi: 10.1186/s40537-025-01193-8.
[10] E. Tiukhova *et al.*, "Explainable Learning Analytics: Assessing the stability of student success prediction models by means of explainable AI," *Decis. Support Syst.*, vol. 182, p. 114229, Jul. 2024, doi: 10.1016/j.dss.2024.114229.
[11] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, "Data leakage inflates prediction performance in connectome-based machine learning models," *Nat. Commun.*, vol. 15, no. 1, p. 1829, Feb. 2024, doi: 10.1038/s41467-024-46150-w.
[12] A. Turkmenbayev, E. Abdykerimova, S. Nurgozhayev, G. Karabassova, and D. Baigozhanova, "The application of machine learning in predicting student performance in university engineering programs: a rapid review," *Front. Educ.*, vol. 10, Sep. 2025, doi: 10.3389/feduc.2025.1562586.
[13] I. K. Nti and S. Ramanayake, "Explainable machine learning for student dropout prediction and tailored interventions in online personalized education," *Discov. Artif. Intell.*, vol. 6, no. 1, p. 288, Feb. 2026, doi: 10.1007/s44163-026-01016-6.
[14] I. K. R. Arthana, "Optimizing Dropout Prediction in University Using Oversampling Techniques for Imbalanced Datasets," *Int. J. Inf. Educ. Technol.*, vol. 14, no. 8, pp. 1052–1060, 2024, doi: 10.18178/ijiet.2024.14.8.2133.
[15] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.
[16] C. H. Cho, Y. W. Yu, and H. G. Kim, "A Study on Dropout Prediction for University Students Using Machine Learning," *Appl. Sci.*, vol. 13, no. 21, p. 12004, Nov. 2023, doi: 10.3390/app132112004.
[17] M. Rebelo Marcolino *et al.*, "Student dropout prediction through machine learning optimization: insights from moodle log data," *Sci. Rep.*, vol. 15, no. 1, p. 9840, Mar. 2025, doi: 10.1038/s41598-025-93918-1.
[18] A. Bettahi, F.-Z. Belouadha, and H. Harroud, "A Modular and Explainable Machine Learning Pipeline for Student Dropout Prediction in Higher Education," *Algorithms*, vol. 18, no. 10, p. 662, Oct. 2025, doi: 10.3390/a18100662.
[19] W.-C. Choi, C.-T. Lam, P. C.-I. Pang, and A. J. Mendes, "A Systematic Literature Review of Explainable Artificial Intelligence (XAI) for Interpreting Student Performance Prediction in Computer Science and STEM Education," in *Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education V. 1*, Jun. 2025, pp. 221–227. doi: 10.1145/3724363.3729027.
[20] H. Khosravi *et al.*, "Explainable Artificial Intelligence in education," *Comput. Educ. Artif. Intell.*, vol. 3, p. 100074, 2022, doi: 10.1016/j.caeai.2022.100074.
[21] L. C. Nnadi, C. P. Isiwu, D. Ding, D. M. Muepu, and Y. Watanobe, "Multi-Level Explainable AI for Predicting Student Depression Risk: Global, Subgroup, and Individual Insights," *IEEE Access*, vol. 14, pp. 6271–6286, 2026, doi: 10.1109/ACCESS.2026.3652631.
[22] M. Nagy and R. Molontay, "Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 2, pp. 274–300, Jun. 2024, doi: 10.1007/s40593-023-00331-8.
[23] W. Hidayatulloh, F. Mahardika, and D. I. Junaedi, "Explainable Artificial Intelligence-Based Model for Student Academic Performance Prediction," *J. Inf. Syst. Explor. Res.*, vol. 4, no. 1, pp. 31–40, Feb. 2026, doi: 10.52465/joiser.v4i1.624.
[24] W. Kim, C. Lee, and H. Kim, "KTCF: Actionable Recourse in Knowledge Tracing via Counterfactual Explanations for Education," *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 45, pp. 38726–38735, Mar. 2026, doi: 10.1609/aaai.v40i45.41216.
[25] N. Mduma, "Data Balancing Techniques for Predicting Student Dropout Using Machine Learning," *Data*, vol. 8, no. 3, p. 49, Feb. 2023, doi: 10.3390/data8030049.
[26] B. Bouihi, A. Bousselham, E. Aoula, F. Ennibras, and A. Deraoui, "Prediction of Higher Education Student Dropout based on Regularized Regression Models," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 6, pp. 17811–17815, Dec. 2024, doi: 10.48084/etasr.8644.
[27] J. K. Hoyos Osorio and G. Daza Santacoloma, "Predictive Model to Identify College Students with High Dropout Rates," *Rev. Electrónica Investig. Educ.*, vol. 25, pp. 1–10, May 2023, doi: 10.24320/redie.2023.25.e13.5398.
[28] A. Villar and C. R. V. de Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," *Discov. Artif. Intell.*, vol. 4, no. 1, p. 2, Jan. 2024, doi: 10.1007/s44163-023-00079-z.
[29] E. Arslan, S. Gaftandzhieva, A. Gorgani Firouzjaei, J. Hassannataj Joloudari, and R. Doneva, "Ex-ADA: a SHAP-based explainable AdaBoost framework for predicting at-risk students," *Front. Educ.*, vol. 10, Jan. 2026, doi: 10.3389/feduc.2025.1728070.
[30] B. I. Igoche, O. Matthew, P. Bednar, and A. Gegov, "Integrating Structural Causal Model Ontologies with LIME for Fair Machine Learning Explanations in Educational Admissions," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 65–85, Jun. 2024, doi: 10.62411/jcta.10501.
[31] G. Ramaswami, T. Susnjak, and A. Mathrani, "Supporting Students' Academic Performance Using Explainable Machine Learning with Automated Prescriptive Analytics," *Big Data Cogn. Comput.*, vol. 6, no. 4, p. 105, Sep. 2022, doi: 10.3390/bdcc6040105.
[32] J. Cheng, Z.-Q. Yang, J. Cao, Y. Yang, and X. Zheng, "Predicting Student Dropout Risk With A Dual-Modal Abrupt Behavioral Changes Approach," *ArXiv*. May 16, 2025. [Online]. Available: http://arxiv.org/abs/2505.11119
[33] M. Delogu, R. Lagravinese, D. Paolini, and G. Resce, "Predicting dropout from higher education: Evidence from Italy," *Econ. Model.*, vol. 130, p. 106583, Jan. 2024, doi: 10.1016/j.econmod.2023.106583.
[34] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting Student Dropout and Academic Success," *Data*, vol. 7, no. 11, p. 146, Oct. 2022, doi: 10.3390/data7110146.
[35] Z. Liu, X. Zhou, and Y. Liu, "Student Dropout Prediction Using Ensemble Learning with SHAP-Based Explainable AI Analysis," *J. Soc. Syst. Policy Anal.*, vol. 2, no. 3, pp. 111–132, Aug. 2025, doi: 10.62762/JSSPA.2025.321501.
[36] A. Igualde-Sáez *et al.*, "StudentDropoutDataset," *Zenodo*, Oct. 2025. https://doi.org/10.5281/zenodo.17239943
[37] F. E. Arévalo-Cordovilla and M. Peña, "Evaluating ensemble models for fair and interpretable prediction in higher education using multimodal data," *Sci. Rep.*, vol. 15, no. 1, p. 29420, Aug. 2025, doi: 10.1038/s41598-025-15388-9.
[38] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.