SAMYRAD 2026

SMART ALGORITHMS & MODELS: YOUR REACH TO ADVANCED DIALOGUES · 5–6 OCTOBER 2026 · SEVILLE, SPAIN
SAMYRAD OPEN-ACCESS PROCEEDINGS | PUBLISHER: INTELLIGENT NUDGE

PAPER • OPEN ACCESS

Incorporating Domain Knowledge into Student Risk Prediction: An Ablation Study Using Higher Education Enrollment Records

Sara A. Khan
SaraKhan@uowdubai.ac.ae

Faculty of Engineering and Information Systems, University of Wollongong in Dubai, Dubai, UAE

Article DOI: 10.68286/samyrad2026.332

ABSTRACT
This paper examines how domain knowledge can be used to predict student academic risk using higher-education administrative data. The study uses 37,583 anonymized student enrollment records from Peru and applies an ablation-style experimental architecture. The target variable is modeled using a two-stage approach: (1) risk occurrence and (2) risk severity. Four feature conditions are tested: Raw-only (student profile variables), Cluster-only (unsupervised cluster labels), and Hybrid configurations with or without domain-based weighting. The domain-based weighting scheme scales feature groups using theory-backed weights. Results show that the Raw-only model establishes a credible baseline with an error profile suited for early intervention. Standalone Cluster-only models performed poorly across both stages. Under 5-fold cross-validation, the weighted Hybrid model detected significantly more at-risk students than Raw-only (McNemar p = 0.0045), but not significantly more than the unweighted Hybrid model (p = 0.25). When concluding the severity predictions, a students' risk severity was directly correlated to their number of enrolled courses. Therefore, the severity stage largely recovered course load as the severity prediction. Once applied, weighting reduced this reliance on course load. Overall, theory-driven weighting is a transparent, testable mechanism for incorporating domain knowledge, with a modest effect on at-risk detection. Future work should evaluate severity on data where it varies independently of course load and calibrate decision thresholds against alert fatigue.

KEYWORDS domain knowledge, ablation framework, regression models, student records.

How to cite: Khan, S. A. (2026). Incorporating Domain Knowledge into Student Risk Prediction: An Ablation Study Using Higher Education Enrollment Records. SAMYRAD 2026 Open-Access Proceedings. Intelligent Nudge. https://doi.org/10.68286/samyrad2026.332

Introduction

Despite constant advancements in higher education, institutions continue to face student dropout and academic failure, creating a need for early-warning methods that can identify at-risk students before this risk develops into disengagement or withdrawal [1].

Educational Data Mining (EDM) allows educators to analyze student records through statistical, machine learning, and data mining methods [2]. Prior work shows that behavioral engagement, attendance, study habits, and administrative or demographic attributes can all contribute to a student's risk profile [3][4][5]. However, predictive pipelines often treat all these variables as equally meaningful and assume that any patterns

discovered in cleaned data are useful without modifications. Such an assumption can obscure contextual factors that matter with individuals embedded in social and institutional environments such as higher education institutions [6][7].

Clustering is frequently introduced to facilitate the grouping of students with similar profiles into categories [8]. Theoretically, clustering helps capture latent structures within data. Practically, when applied in a purely data-driven manner, clustering might lead to the formation of conclusions not completely aligned with established knowledge (i.e., the field of student learning and academic success). Evaluating schedule choice based on chronotype helps illustrate this problem. Despite its relevance and importance in academic performance and students' retention levels, chronotype is rarely incorporated explicitly into predictive models of academic risk, leaving a gap in understanding students in higher education [9] [10]. This makes it possible for the same recorded behavior to have different implications for students operating under different environments.

This study evaluates domain knowledge as a modeling intervention rather than as post-hoc narrative context. It follows the theory-guided data-science principle that established knowledge can be encoded directly within a computational pipeline [11]. Bean's student-attrition framework is used to assign fixed weights to feature groups according to their theoretical proximity to attrition [12]. The contribution is tested through a controlled ablation design in which the dataset, cross-validation folds, preprocessing, estimators, and metrics remain constant while only cluster-derived features and domain-informed scaling are changed. Risk is separated into occurrence, whether at least one course is at risk, and severity, the quantity of at-risk courses while risk is present. This separation is intended to support identification and intervention prioritization.

The paper addresses three main questions:

RQ1: *To what extent can cleaned raw student profile attributes predict academic risk outcomes at the individual level?*

RQ2: *What is the effect of incorporating clustering into academic-risk prediction when compared to prediction using cleaned raw student profile data alone?*

RQ3: *What is the effect of incorporating domain knowledge into clustering (through domain-informed feature weights) on predictive outcomes, compared to purely data-driven clustering and unweighted hybrid feature sets?*

Related work

Educational Data Mining and Domain Knowledge

Student-retention research has increasingly moved from isolated academic explanations toward multi-dimensional accounts of student attrition. Tinto emphasized a focus on academic and social integration [13], whereas Bean emphasized the effect of environmental, organizational, and background determinants in shaping a student's intent to leave [12]. These theories provide a structured basis for interpreting administrative variables such as tuition behavior, study mode, schedule, enrollment, and student background. Demographic and pre-admission features have also been shown to support student-performance prediction; particularly when behavioral logs through learning platforms are unavailable [14].

These demographic and pre-admission factors can further be categorized into varying levels of priority based on the accumulated knowledge of experts (i.e., domain knowledge). The use of domain knowledge is important because high predictive performance does not itself establish explanatory meaning [15]. Theory-guided modeling makes assumptions explicit and allows them to be tested as part of a model pipeline. In the present study, the weights are not optimized to maximize a metric; their use aims to incorporate a literature-grounded hierarchy whose behavioral effect can be compared under controlled conditions. Educational prediction commonly draws from either behavioral activity in higher education learning-platforms or administrative and demographic records. While platform behavior can form a detailed activity log, the findings become limited to what students do inside the observed system [16]. Administrative records provide less behavioral detail yet remain available before and outside a student's platform engagement. This makes them particularly relevant to early screening, provided that their interpretation remains sensitive to institutional conditions and does not reduce demographic associations to causal explanations [11].

Clustering and Ablation-Style Evaluation

Supervised classifiers remain common in EDM, while clustering can often be used to discover student profiles before classification [17]. Hybrid architectures have reported improvements in a model's predictive accuracy when clustering is able to add information which is not already represented within a cleaned dataset [18] [19]. For example, a cluster label derived from the same raw attributes may function mainly as a significantly compressed summary of existing information. Its descriptive usefulness cannot guarantee an improved individual-level prediction.

Breaking down a system into its component parts to test each component in isolation (i.e., ablation) provides a direct way to evaluate the influence of incorporating this distinction. Component-wise ablation has been used in knowledge-tracing models to isolate the effect of embeddings or modules [20], while feature-level ablation quantifies whether raw or derived variables contribute independent information to a prediction model [21]. This is particularly relevant when clustering might result in cluster groups that do not align with the domain knowledge developed in a scientific field [22]. The present design therefore treats raw attributes, cluster labels, their combination, and theory-driven weighting as separate, testable conditions to identify the effect of each condition on predictive outcomes.

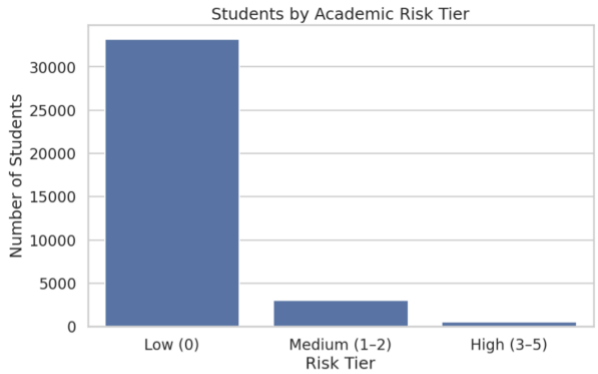
A central focus of this study is in distinguishing between a variable's descriptive and predictive utility. A cluster can be internally coherent and still be weak as a predictor. Conversely, a supervised model can discriminate risk without producing easily isolated student archetypes. Hybrid EDM studies often combine the two functions, but the resulting gain must be attributed

carefully: improvements may arise for various reasons including the utilization of complementary data, different inductive biases, or additional model complexity [17]–[19]. Keeping the preprocessing and evaluation conditions fixed allows this study to focus on the effect of the testable conditions alone. The desire to preserve interpretability motivated the use of linear occurrence and count models. While more complex architectures would have allowed this study to capture nonlinear interactions, they would have limited the ability to determine an exact source of the deviations in model performance [23]. Logistic and Poisson regression provide probability and count outputs that can be inspected directly, while the ablation conditions expose how the same evaluation metrics respond to models with raw, clustered, and theory-weighted datasets.

Methodology

Dataset and Preprocessing

The source dataset contains 37,583 anonymized enrollment records from Peru for 2023 [24]; as a publicly accessible, anonymized, secondary dataset, a formal ethics approval was not required. Each record includes mixed categorical and numeric attributes. The outcome (i.e., the number of at-risk courses) is a discrete count from 0 to 5. Two targets were derived: risk occurrence (count ≥ 1) and risk severity (the count, modeled only for records with observed risk). A reporting variable grouped counts into low (0), medium (1–2), and high (3–5) risk tiers. For all 3,597 at-risk records, the severity variable count equaled the number of enrolled courses, so severity was determined by course load in this dataset. Columns with more than 50% missingness (i.e., type of educational institution, institution status, and educational institution name) were removed. Listwise deletion was used to remove 736 records. The resulting dataset contained 36,847 records and 17 columns. Gender labels were normalized from a mix of numeric and categorical coding. The final class distribution was strongly imbalanced with 33,187 records in the low tier, 3,056 in the medium tier, and 541 in the high tier, as shown in Fig. 1.



risk_tier	count
Low (0)	33187
Medium (1–2)	3056
High (3–5)	541

Fig. 1. Distribution of students across academic risk tiers after cleaning.

Performance metrics represent 5-fold stratified cross-validation means (stratified by risk occurrence, random seed 42), applied through shared indices so that every ablation condition used identical records. To ensure differences in at-risk detection between models exceeded what would be expected by chance, McNemar's exact test was applied to the out-of-fold predictions. A stratified 70/30 split was kept for Figs. 2–4. Numeric features were passed through, while categorical features were one-hot encoded with unknown categories ignored. Preprocessing was implemented within scikit-learn Pipeline and Column Transformer objects fitted on each training fold. To maintain identical transformations and prevent leakage in encoding, the cluster label was assigned from predictors before splitting.

Clustering and Ablation Conditions

Clustering was used only as a feature-construction step; the outcome was excluded. Hierarchical clustering required a full pairwise Gower distance matrix, whose memory grows with the square of the number of records. This would have required about 5.4 GB for all 36,847 records. Running it on the full dataset exceeded the memory available in a standard Google Colab session. Therefore, it was run on a random sample of 15,000 records (0.9 GB) to select k (k = 3 to 8). Silhouette scores were low (0.294 at k = 3, 0.281 at k = 4, 0.226 at k = 5, 0.1866 at k = 6), indicating weak but non-random structure relative to a random-label baseline, which explains the limited predictive value of the cluster label; k = 4 was retained for granularity. K-prototypes was then used to assign the final four-cluster categorical label to all records. It agreed poorly with the hierarchical partition (adjusted Rand index = 0.235) but was stable between sample and full-data fits (0.700). Thus, the label could act as a plausible partition rather than for the purpose of forming distinct segments.

Four conditions were evaluated: (1) Raw-only, using enrollment, gender, program major, shift schedule, study mode, department, number of enrolled courses, tuition-payment indicators for March 2022 and March 2023, and benefit discounts; (2) Cluster-only, using the cluster label alone; (3) Hybrid unweighted, combining raw attributes with the cluster label; and (4) Hybrid weighted, applying theory-informed scaling to the encoded Hybrid matrix before fitting.

Bean-Based Feature Weighting

The weighting intervention grouped variables into four pillars derived from Bean's attrition model [12]. Environmental variables (i.e., tuition-payment indicators and benefit discounts) and Academic variables (i.e., enrollment and number of enrolled courses) received a weight of 0.50. Organizational variables—shift schedule and study mode—received a 0.30. Background variables—gender, department, and program major—received a 0.20. One-hot expanded columns inherited the weight of their source feature. A sparse-safe Function Transformer applied the scale factors after encoding and before model fitting. The 'sensitivity variant' values represent a theory-driven design choice rather than a parameter set optimized for the best metrics. Two alternatives breaking the Environmental/Academic tie (Table 1) tested sensitivity to this choice.

Table 1. Domain-informed feature weights

Pillar (weight)	Mapped variables	Sensitivity variant
Environmental (0.50)	Tuition payment 2022/2023; benefit discounts	0.60 / 0.40
Academic (0.50)	Enrollment; number of enrolled courses	0.40 / 0.60
Organizational (0.30)	Shift schedule; study mode	0.30
Background (0.20)	Gender; department; program major	0.20

Predictive Models and Evaluation

Stage 1 modeled risk occurrence with logistic regression. The estimator used a maximum of 2,000 iterations and balanced class weights so errors on the minority at-risk class contributed more strongly to training. Models were evaluated using accuracy, at-risk precision, at-risk recall, F1, confusion matrices, and receiver operating characteristic—area under curve (ROC-AUC). Because the use case prioritizes early intervention, recall and false negatives were prioritized over accuracy alone.

Stage 2 modeled risk severity with Poisson regression (2,000 iterations, L2 alpha = 0.1) among students with observed risk; since severity equals course load here, it tests how closely each configuration recovers this relationship. MAE and RMSE quantified continuous error, the Pearson χ^2 statistic divided by residual degrees of freedom (χ^2/df) checked the Poisson mean–variance assumption (above 1 indicates overdispersion, below 1 under dispersion), and rounded predictions were checked for exact and ±1/±2 agreement. The stages were combined as expected at-risk courses = P(risk) × predicted severity, with severity predicted for every student.

As nonlinear reference points outside the ablation, Random Forest, XGBoost, and CatBoost were trained on the Raw-only features with the same folds, using class-balanced weighting and, for XGBoost and CatBoost severity models, a Poisson objective.

All analysis was executed in Python within Google Colab using pandas and NumPy for data handling, matplotlib and seaborn for visualization, scikit-learn for preprocessing and modeling, XGBoost and CatBoost for gradient-boosted benchmarks, stats models for McNemar's test, and SciPy for sparse matrix operations. The same fitted preprocessing objects were reused within each pipeline during testing. This was important for the weighted condition because one-hot expansion changes the number and order of columns. Therefore, weights were mapped to encoded columns only after the training preprocessor had established the feature space.

To provide a holistic view of results, the evaluation incorporates multiple metrics into predicting both the likeliness (i.e., risk classification) and severity of risk. For classification, ROC-AUC checks the accuracy of the model in differentiating between safe and at-risk students, while recall and precision show the quality and probability of identifying at-risk students. For severity, MAE and RMSE look at continuous instead of rounded values, while agreement metrics help confirm the practical reliability of predicted scores for advisors in higher education.

Results

Risk Occurrence

Table 2 summarizes the occurrence results as means across the five cross-validation folds. The Raw-only model established a credible baseline with ROC-AUC = 0.7631 and recall = 0.7067. Given the severe class imbalance in the dataset (approximately 9 safe students for every at-risk student), the observed precision of 0.2010 represents an enrichment over random chance. While false positives are elevated due to the low base rate (10%), the model successfully captures 70.7% of true at-risk students. This trade-off may be appropriate for a screening layer in which a supportive review follows the automated flag.

Table 2. Risk OCCURRENCE RESULTS (5-FOLD MEANS; SD ≤ 0.018)

Model	Acc.	Recall	Prec.	AUC
Raw	0.6966	0.7067	0.2010	0.7631
Cluster	0.8122	0.3228	0.2061	0.6298
Hybrid-U	0.6961	0.7125	0.2017	0.7631
Hybrid-W	0.6948	0.7162	0.2016	0.7635

Cluster-only achieved the highest accuracy (0.8122) but the weakest intervention profile. Recall fell to 0.3228 and ROC-AUC to 0.6298, showing that a broad cluster label could not preserve enough individual-level information to distinguish which students inside a cluster were at risk. The accuracy increase is therefore attributable largely to the majority no-risk class rather than to improved risk detection.

The confusion matrices from the illustrative 70/30 split make the difference concrete. Raw-only identified 756 at-risk students and missed 323, while generating 3,074 false positives and 6,883 true negatives. Cluster-only identified only 346 at-risk students and missed 733, although its majority-class predictions produced 8,594 true negatives and only 1,363 false positives. Fig. 2 serves as a visual representation of the aforementioned values. The accuracy advantage appears to come from a more conservative alert policy which substantially increased the number of struggling students left undetected.

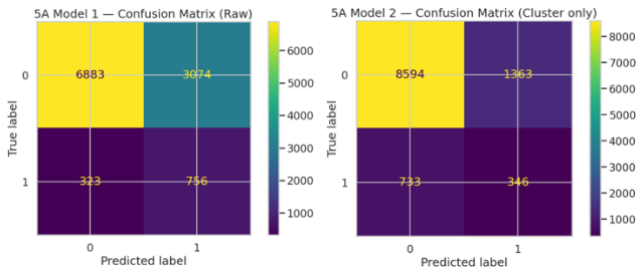


Fig. 2. Confusion matrices for the Raw and Cluster-only occurrence model.

Adding the cluster label to the raw attributes produced almost no change: the unweighted Hybrid model reached ROC-AUC =

0.7631 and recall = 0.7125. The weighted Hybrid model, by contrast, shifted the decision profile slightly further toward higher sensitivity. Recall increased to 0.7162 while ROC-AUC remained essentially unchanged at 0.7635. Accuracy decreased slightly to 0.6948, consistent with the model accepting more false positives to reduce false negatives. Across the out-of-fold predictions, the weighted Hybrid model detected 34 more at-risk students than Raw-only (McNemar p = 0.0045) but only 13 more than the unweighted Hybrid model, which was not statistically significant (p = 0.2503). On the illustrative 70/30 split, Fig. 3 shows that weighting changed the Hybrid confusion matrix from 326 false negatives and 753 true positives to 313 false negatives and 766 true positives.

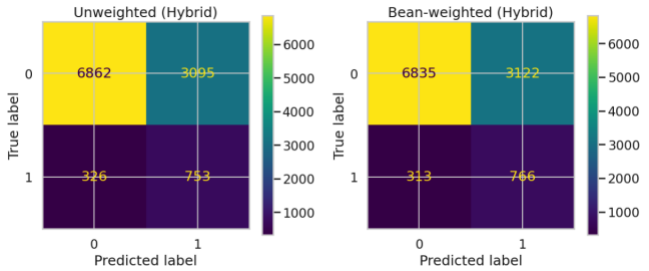


Fig. 3. Confusion matrices for the unweighted and Bean-weighted Hybrid occurrence models.

Risk Severity

Table 3 reports severity error (5-fold means) among at-risk students, measured as MAE and RMSE between the predicted and actual number of at-risk courses. Because this number equals each student's course load, lower error indicates how closely a configuration reproduces the number of enrolled courses rather than an independent measure of risk. Raw-only and Hybrid unweighted were almost identical (MAE 0.1234 vs. 0.1232; RMSE 0.1583 vs. 0.1585), so the cluster label added little beyond the profile variables from which it was derived.

Table 3. Risk severity results (5-fold means; sd ≤ 0.035)

Model	MAE	RMSE
Raw	0.1234	0.1583
Cluster	0.6051	0.7275
Hybrid-U	0.1232	0.1585
Hybrid-W	0.2081	0.2304

Cluster-only performed worst, with MAE = 0.6051 and RMSE = 0.7275. Its predictions effectively approximated an average severity for each cluster and could not represent variations within clusters, because the cluster label does not contain course load. The weighted Hybrid model showed a more complex pattern. Continuous error increased to MAE = 0.2081 and RMSE = 0.2304, consistent with the weighting reducing the influence of the course-load variable, yet every rounded prediction fell within ±1 course in all five folds. On the 70/30 split, approximately 97.3% of rounded predictions exactly matched the observed integer count, and train and test MAE were nearly identical (0.208 and 0.209), which does not indicate obvious overfitting.

Weighting also redistributed severity estimates rather than uniformly increasing them. On the 70/30 split, median predicted severity rose from 1.0938 to 1.2204, while the mean decreased slightly from 1.5604 to 1.5477. The intervention therefore moved central predictions upward while reducing some higher-end estimates. This explains why integer-level agreement remained high even though the continuous residuals increased.

The end-to-end expected-risk construction preserves the distinction between risk classification and intensity, offering a record-level summary that institutions can inspect. Because severity equals course load in this dataset, the expected count approximately equals the probability of risk multiplied by the number of enrolled courses, so a student with a heavier course load may warrant earlier review even if both students have the same predicted probability of risk. Retaining the component outputs avoids the loss of information that would occur if the system exposed only a single rounded value of risk. The expected count is therefore best viewed as an ordering aid rather than a replacement for probability, severity, and contextual review.

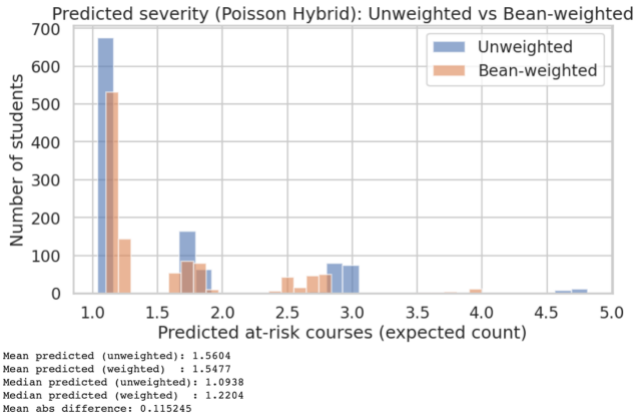


Fig. 4. Distribution of predicted severity for the unweighted and Bean-weighted Hybrid Poisson models.

Weighting Sensitivity and Nonlinear Benchmarks

Breaking the Environmental/Academic tie in either direction left occurrence results unchanged (recall 0.7162–0.7175; ROC-AUC 0.7635 throughout). Severity MAE depended on it (0.1752 with Academic higher, 0.2081 with equal weights, 0.2611 with Environmental higher) since the Academic variables include course load.

The nonlinear benchmarks (Table 4) achieved higher occurrence discrimination than logistic regression (ROC-AUC 0.8096-0.8158 vs 0.7631) and reproduced severity almost exactly (MAE ≤ 0.0009), the latter because tree-based models map course load directly onto the count.

Table 4. Nonlinear benchmarks on raw-only features (5-fold means)

Model	AUC	Severity MAE
Logistic / Poisson	0.7631	0.1234
Random Forest	0.8096	0.0005
XGBoost	0.8149	0.0003
CatBoost	0.8158	0.0009

Discussion

What the Ablation Reveals

Across both stages, raw student attributes carried most of the predictive signal, consistent with Bean's model, whose environmental, organizational, and background determinants map directly onto these attributes [12]. Occurrence discrimination was moderate, and the low severity error reflects severity equaling course load. Cluster labels alone were insufficient for individual prediction, consistent with their weak structure, and added little to the raw dataset. This does not imply that clustering lacks value: the four clusters can still support profiling, communication, or exploratory analysis. A single label is simply too coarse to replace the underlying attributes when deciding which individual student should be flagged.

The result also demonstrates why accuracy is an inadequate headline metric for imbalanced early-warning systems. A model can achieve high accuracy by predicting the majority class and still miss the minority class (i.e., students who are at risk). The Cluster-only occurrence model is the clearest example: its 0.8122 accuracy appears favorable until its 0.3228 recall and 0.6298 ROC-AUC are considered. Ablation makes this failure mode visible because the folds, preprocessing, and evaluation metrics are kept constant.

The near-equivalence of Raw-only and Hybrid unweighted clarifies the role of derived features: a cluster label generated from the same profile matrix summarizes patterns the linear models can already access. Larger gains would require a derived variable capturing external, longitudinal, or theoretically constrained information unavailable in the raw data, so cluster labels are best used as descriptive context unless ablation shows independent predictive value.

Domain Knowledge as a Modeling Intervention

The principal effect of Bean-based weighting was behavioral rather than a uniform gain. In occurrence prediction, the weighted Hybrid model had the highest recall without changing ranking ability and detected significantly more at-risk students than Raw-only ($p = 0.0045$), but its gain over Hybrid unweighted was not significant ($p = 0.2503$), so the effect of weighting alone could not be distinguished from chance. Domain knowledge can nonetheless be encoded as an explicit, reproducible pipeline component [25] [26]. In Bean's terms, the weights emphasize environmental pulls such as tuition payment and benefit discounts over more distal background attributes [12]. The intervention changed the cost profile in a direction aligned with early intervention: more struggling students were detected than by the Raw-only model, at the expense of a small increase in overall at-risk detections.

The severity result requires cautious interpretation. The weighted Poisson model had higher MAE and RMSE than Raw-only and Hybrid unweighted. Since severity equals course load here, this difference reflects how much weight each configuration gives to the enrolled courses rather than predictive insight. Error was lowest when the Academic pillar, which contains course load, was weighted higher. This dataset therefore cannot show whether weighting improves severity estimates; the stage's practical value is combining risk probability with course load to prioritize cases.

Practical Implications

The results support an early-warning architecture in which administrative records provide the core individual-level signal, clustering supports descriptive segmentation, and domain weighting is treated as a policy-sensitive calibration step. The occurrence model should operate to support decision making rather than facilitating automated judgment. Its precision of about 0.20 (four in five flags are false alarms) makes this stage the main source of alert fatigue. This is acceptable because a flag triggers a low-cost supportive review. Thresholds can then be calibrated to the number of advisors or interventions available; allowing the institution to function based on capacity and prevent alert fatigue [27]. A practical deployment could therefore use a tiered workflow: a low threshold for an initial supportive check, a higher threshold for resource-intensive intervention, and the expected count (risk probability × course load) to order cases by priority. Such a workflow would preserve human review and allow thresholds to update when institutional capacity (e.g., available staff, venue, or educational resources) changes.

The role of schedule-related variables also remains contextual. Chronotype is not directly measured in the dataset, and shift schedule remains a proxy for the institutional timing conditions under which students study. Prior work on social jet lag and performance shows why such variables may deserve interpretation beyond their raw statistical coefficient [16]. The present weighting scheme is one transparent way to express that interpretation, but it does not establish causal effects. Model predictions should therefore augment, not replace, human advisor review.

Limitations and future work

The study uses one anonymized 2023 cohort from a single repository, so the results may not generalize across institutions, policies, economic conditions, or academic years. The at-risk course count is an operational proxy for academic struggle and may reflect course difficulty, grading practice, or assessment design. Additionally, it also translates to course load for at-risk students, so partial risk is not captured. The models estimate associations rather than causal drivers of dropout. The approximately 9:1 class imbalance also means that accuracy must be interpreted alongside recall, precision, and ROC-AUC.

Logistic and Poisson regression were chosen for transparency and controlled comparison, not maximum performance. Nonlinear benchmarks reached higher occurrence discrimination (ROC-AUC up to 0.8158), indicating unmodeled interactions. Across all students, the at-risk course count is overdispersed (variance-to-mean ratio ≈ 1.77), driven by the large share of students with no at-risk courses; a single count model fitted to every student would therefore require negative binomial regression [28] or a hurdle or zero-inflated model [29]. The present two-stage design follows the hurdle approach by modeling occurrence separately and severity only among at-risk students. Within the severity stage, however, the divergence between MAE and rounded agreement is not due to overdispersion: Pearson χ^2/df was 0.013-0.034, indicating strong underdispersion, as expected when severity equals course load. Evaluating severity therefore requires data in which it varies independently of course load.

The pillar weights were literature-grounded but not optimized; occurrence results were robust to the Environmental/Academic ratio, whereas severity error was not. Broader alternatives, such as 0.6/0.3/0.1 and 0.5/0.3/0.2, should be tested across multiple cohorts. Decision thresholds should be selected against institutional capacity, and fairness analysis could compare false-negative rates across relevant subgroups. Because demographic variables can improve prediction while also creating ethical and deployment concerns, protected attributes require explicit governance and robustness checks [30]. Additional longitudinal features, including payment stability and schedule consistency, may improve interpretability and generalization.

Conclusion

This study used a two-stage ablation framework to separate risk occurrence from risk severity and to test raw attributes, cluster labels, their combination, and Bean-based feature weighting under controlled conditions. Raw student-profile variables provided a strong baseline. Cluster labels alone performed poorly and added little incremental predictive value when combined with the same attributes from which they were derived. Domain-informed weighting did not produce a universal accuracy gain, but it shifted occurrence prediction toward higher recall while maintaining ROC-AUC, detecting significantly more at-risk students than the Raw-only model, although its gain over the unweighted Hybrid model was not statistically significant. In severity prediction, weighting increased continuous error but retained high rounded agreement; because severity equaled course load for every at-risk student, this stage mainly recovered course load, and the higher error reflects the weighting reducing the influence of that variable. The findings position domain knowledge as a transparent modeling intervention whose effect should be evaluated against the institutional decision objective, rather than assumed to improve every metric. Future work should evaluate severity on data in which it varies independently of course load, compare count models suited to bounded, under dispersed counts, and calibrate support thresholds to balance institutional capacity (e.g., available staff, venue, or educational resources) against alert fatigue.

References

[1] M. O. González-Morales, D. López-Aguilar, P. R. Álvarez-Pérez, and P. A. Toledo-Delgado, "Dropping out of higher education: Analysis of variables that characterise students who interrupt their studies," Acta Psychologica, vol. 252, Art. no. 104669, 2025, doi: 10.1016/j.actpsy.2024.104669.

[2] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Trans. Syst., Man, Cybern. C, vol. 40, pp. 601–618, 2010, doi: 10.1109/TSMCC.2010.2053532.

[3] A. Almalawi, B. Soh, A. Li, and H. Samra, "Predictive models for educational purposes: A systematic review," Big Data Cogn. Comput., vol. 8, no. 12, 2024, doi: 10.3390/bdcc8120187.

[4] A. Hysaj, G. Farouqa, SA Khan, and L. Hiasat, "Reducing Stress Through Formative Assessments: A Case of the Digital Platform," *Lecture Notes in Computer Science*, pp. 486–500, Jan. 2023, doi: 10.1007/978-3-031-35927-9_33.

[5] G. Farouqa *et al.*, "Undergraduate Students' Journey with AI in the United Arab Emirates," *Lecture Notes in Computer Science*, pp. 155–171, 2025, doi: 10.1007/978-3-031-93539-8_11.

[6] I. Schoon and J. Heckhausen, "Conceptualizing individual agency in the transition from school to work: A social-ecological developmental perspective," Adolescent Res. Rev., vol. 4, 2019, doi: 10.1007/s40894-019-00111-3.

[7] J.-A. Botha and I. Potgieter, "Socio-demographical variables as predictors of academic self-directedness," Int. Rev. Res. Open Distrib. Learn., vol. 25, no. 4, pp. 38–51, 2024, doi: 10.19173/irrodl.v25i4.7516.

[8] Y. Lu, S. Yeom, J. Maktoubian, M. M. Rahman, and S.-H. Kim, "Improve student risk prediction with clustering techniques: A systematic review in education data mining," Educ. Sci., vol. 15, no. 12, 2025, doi: 10.3390/educsci15121695.

[9] T. Enright and R. Refinetti, "Chronotype, class times, and academic achievement of university students," Chronobiol. Int., vol. 34, pp. 1–6, 2017, doi: 10.1080/07420528.2017.1281287.

[10] L. S. Kalbacher and C. Randler, "Exploring the nexus between sleep, chronotype, and non-cognitive predictors in university students: Implications for academic success," Chronobiol. Int., vol. 41, no. 8, pp. 1165–1175, 2024, doi: 10.1080/07420528.2024.2383396.

[11] A. Karpatne et al., "Theory-guided data science: A new paradigm for scientific discovery from data," IEEE Trans. Knowl. Data Eng., vol. 29, no. 10, pp. 2318–2331, 2017, doi: 10.1109/TKDE.2017.2720168.

[12] J. P. Bean, "The synthesis of a theoretical model of student attrition," 1981.

[13] V. Tinto, "Stages of student departure: Reflections on the longitudinal character of student leaving," J. Higher Educ., vol. 59, no. 4, pp. 438–455, 1988, doi: 10.2307/1981920.

[14] M. Bilal, M. Omar, W. Anwar, R. H. Bokhari, and G. S. Choi, "The role of demographic and academic features in a student performance prediction," Sci. Rep., vol. 12, Art. no. 12508, 2022, doi: 10.1038/s41598-022-15880-6.

[15] López-Meneses, E., Mellado-Moreno, P. C., Herrerías, C. G., & Pelícano-Piris, N. (2025). Educational Data Mining and Predictive Modeling in the Age of Artificial Intelligence: An In-Depth Analysis of Research Dynamics. *Computers*, *14*(2). https://doi.org/10.3390/computers14020068.

[16] B. Smarr and A. Schirmer, "3.4 million real-world learning management system logins reveal the majority of students experience social jet lag correlated with decreased performance," Sci. Rep., vol. 8, 2018, doi: 10.1038/s41598-018-23044-8.

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

[18] D. Hooshyar, Y.-M. Huang, and Y. Yang, "A three-layered student learning model for prediction of failure risk in online learning," Human-Centric Comput. Inf. Sci., vol. 12, pp. 357–372, 2022, doi: 10.22967/HCIS.2022.12.028.

[19] S. Fida, N. Masood, N. Tariq, and F. Qayyum, “A novel hybrid ensemble clustering technique for student performance prediction,” J. Universal Comput. Sci., vol. 28, no. 8, pp. 777–798, 2022, doi: 10.3897/jucs.73427.
[20] S. Pandey and G. Karypis, “A self-attentive model for knowledge tracing,” 2019.
[21] I. Lazic, C. Barà, M. Iovino, S. Stramaglia, N. Jakovljevic, and L. Faes, “Information-theoretic quantification of high-order feature effects in classification problems,” arXiv:2507.04362, 2025, doi: 10.48550/arXiv.2507.04362.
[22] A. Jasinska-Piadlo et al., “Data-driven versus a domain-led approach to k-means clustering on an open heart failure dataset,” Int. J. Data Sci. Anal., vol. 15, no. 1, pp. 49–66, 2023, doi: 10.1007/s41060-022-00346-9.
[23] Kim, S., Choi, E., Jun, Y.-K., & Lee, S. (2023). Student Dropout Prediction for University with High Precision and Recall. *Applied Sciences*, *13*(10). https://doi.org/10.3390/app13106275
[24] M. M. Diaz, “Peru student enrollment data (2023),” Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/miguelmallqui17/peru-student-enrollment-data-2023
[25] M. Melchior, “Incorporating domain knowledge for learning interpretable features,” Archives of Data Science, Series A, vol. 8, no. 2, 2022, doi: 10.5445/IR/1000150142.
[26] T. Miller, I. Durlik, A. Łobodzińska, L. Dorobczyński, and R. Jasionowski, “AI in context: Harnessing domain knowledge for smarter machine learning,” Appl. Sci., vol. 14, no. 24, 2024, doi: 10.3390/app142411612.
[27] F. Voutsas, J. Violos, and A. Leivadeas, “Mitigating alert fatigue in cloud monitoring systems: A machine learning perspective,” Comput. Netw., vol. 250, Art. no. 110543, 2024, doi: 10.1016/j.comnet.2024.110543.
[28] Y. Tiara, M. Aidi, E. Erfiani, and R. Rachmawati, “Overdispersion handling in Poisson regression model by applying negative binomial regression,” BAREKENG, vol. 17, pp. 417–426, 2023, doi: 10.30598/barekengvol17iss1pp0417-0426.
[29] M. Flynn and L. A. F. Fcas, “More flexible GLMs: Zero-inflated models and hybrid models,” Casualty Actuarial Society E-Forum, 2009.
[30] R. Yu, H. Lee, and R. F. Kizilcec, “Should college dropout prediction models include protected attributes?” arXiv:2103.15237, 2021, doi: 10.48550/arXiv.2103.15237.