Vol. 3, No. 1, pp.: 60-75, 2025

Learning Analytics Databases for Early Warning of Student Dropout

Felipe Augusto Rodrigues[1], Mariana Conceição Ferreira[2,*], Thiago Henrique Braga[3]

[1] Department of Computer Science, Federal University of Paraíba, João Pessoa 58051-900, Brazil

[2] Faculty of Education, Federal University of Mato Grosso, Cuiabá 78060-900, Brazil

[3] Department of Exact Sciences, State University of Feira de Santana, Feira de Santana 44036-900, Brazil

** m.ferreira@ufmt.br*

Article Information	
Received	18 October 2024
Accepted	29 February 2025
DOI	https://doi.org/10.63646/datamind.2025.030105

Abstract

Student dropout is one of the most consequential outcomes in higher education, with far-reaching consequences for individual career trajectories, institutional funding, and national workforce development. Despite the proliferation of Learning Management Systems that generate rich longitudinal records of student behaviour, most universities lack a structured, integrated database infrastructure that converts raw LMS event logs, enrolment records, financial aid data, and social network interactions into actionable early warning signals. This article presents EWS-LMS-DB, a purpose-built learning analytics database designed to support reproducible early warning research and evidence-based intervention in undergraduate education. The database integrates six core relational tables covering student demographics, course enrolment records, LMS interaction events, financial aid transactions, risk alerts, and intervention outcomes, covering 28,640 students across 14 semesters at three federal universities in Brazil. An LSTM-based sequential model, EWS-LSTM, is benchmarked against Random Forest and Logistic Regression baselines, achieving an AUC-ROC of 0.88 at semester week 8, providing an average of 8.4 weeks of early warning lead time before confirmed dropout. A fairness analysis across eight demographic groups reveals that true positive rate parity is maintained within ±5 percentage points across gender, first-generation, scholarship, and rural-urban subgroups. An intervention backtest using matched control groups shows that students who received an advisor contact within one week of a high-risk alert had a 23 percentage point higher semester retention rate at twelve weeks. The database schema, field dictionary, ingestion pipeline, and a 20 percent open sample are released for reproducible experimentation.

Keywords: ***Learning analytics; student dropout prediction; early warning systems; LMS database; educational data mining; intervention design; algorithmic fairness***

1. Introduction

Dropout from higher education represents a persistent institutional and social challenge. In Brazil, the national dropout rate across federal universities has hovered between 20 and 26 percent over the past decade, representing

hundreds of thousands of students annually who leave without completing their degrees (Tinto, 1975; Berens et al., 2019). The economic cost is substantial: for the student, incomplete credentials reduce lifetime earnings and career mobility; for the institution, each dropout erodes per-student funding allocations and damages completion-rate metrics that increasingly determine public university budgets. The social dimension is equally significant because dropout is not random. Students from lower socioeconomic backgrounds, first-generation families, and rural areas are systematically over-represented among those who leave, and their departure compounds existing educational inequality rather than alleviating it (Chung and Lee, 2019; Rizvi et al., 2019).

The expansion of digital learning environments has created an unprecedented opportunity to address dropout proactively. Modern Learning Management Systems such as Moodle, Canvas, and Blackboard generate continuous, timestamped records of every student interaction with course materials: logins, assignment submissions, forum posts, video views, and quiz attempts. These behavioural signals, combined with enrolment records, academic history, financial aid transactions, and social network interactions with peers and instructors, constitute a rich multi-modal database from which early warning signals can, in principle, be extracted weeks before a student formally withdraws. The field of learning analytics has developed precisely to exploit this opportunity, with a growing body of research demonstrating that predictive models trained on LMS data can identify at-risk students with AUC-ROC values above 0.80 as early as the fourth or fifth week of a semester (Arnold and Pistilli, 2012; Macfadyen and Dawson, 2010; You, 2016).

Despite this research momentum, a methodological gap persists that limits both the reproducibility of findings and their translation into operational practice. Most published early warning studies are conducted on proprietary, single-institution datasets that are neither documented nor shared. Database schemas vary arbitrarily across studies, making cross-institutional comparisons impossible. Feature engineering choices, quality control procedures, and temporal alignment between LMS events and dropout outcomes are rarely described in sufficient detail to reproduce. Fairness analyses across demographic subgroups are even more rarely reported, leaving open the risk that widely deployed early warning systems produce systematically higher false positive rates for already-disadvantaged groups, potentially triggering unnecessary interventions that are themselves a form of institutional surveillance (Kizilcec and Lee, 2020; Gašević et al., 2016; Ferguson, 2012). The gap between the research literature and operational deployment is therefore not only a technical gap but an infrastructural and ethical one.

This article responds to that gap by presenting EWS-LMS-DB, a structured multi-table learning analytics database purpose-built to support reproducible early warning research and evidence-based intervention. The database is not simply a data export from a single LMS; it is a documented research infrastructure comprising a six-table relational schema, a feature store layer for pre-computed weekly risk features, a social interaction graph, ethical access controls, and a documented intervention backtest protocol. Four research questions guide the work. First, what database schema and field design best support multi-source early warning feature construction from heterogeneous institutional data systems? Second, at what semester week does an LSTM-based sequential model achieve actionable AUC-ROC performance, and how does that lead time compare to simpler baselines? Third, does the early warning model maintain fairness across demographic subgroups, and what database-level corrections are available when fairness gaps are detected? Fourth, what measurable impact do database-triggered advisor interventions have on short-term retention, and how can that impact be estimated rigorously through an intervention backtest protocol?

The contributions of the article are fourfold. We present a 28,640-student, 14-semester multi-table learning analytics database with documented schema, field dictionary, quality statistics, and open sample. We describe an LSTM-based sequential model that achieves AUC-ROC of 0.88 at week 8 with average early warning lead time of 8.4 weeks. We report a fairness analysis across eight demographic subgroups and demonstrate that KDE-based sample reweighting at the database level reduces the maximum TPR gap from 11 to 7 percentage points. We

present an intervention backtest showing a 23 percentage point retention gain at twelve weeks for students who received a timely advisor contact following a high-risk alert.

2. Database Gap and Use Cases

The landscape of published learning analytics datasets can be organised into three categories, each with characteristic strengths and limitations that motivate the EWS-LMS-DB design. The first category consists of MOOC engagement datasets, including the MOOCdb schema proposed by Veeramachaneni et al. (2014) and the various Coursera and edX datasets used in dropout prediction competitions. These datasets are large in student count but capture a narrow slice of the educational experience: they include clickstream data and video-viewing patterns but exclude financial aid status, in-person attendance, peer network quality, and the multi-semester longitudinal dimension that characterises traditional degree programmes (Xing et al., 2016; Whitehill et al., 2017). MOOC dropout is also a qualitatively different phenomenon from degree-programme dropout, because MOOC enrolment is free, non-binding, and motivationally heterogeneous, making the event less consequential and the signal noisier.

The second category consists of single-institution institutional research datasets, of which the Open University Learning Analytics Dataset (OULAD) is the most widely shared example. OULAD provides student demographic information, VLE interaction logs, and module assessment scores for over 32,000 students, and it has been the basis for dozens of published early warning studies (Hlosta et al., 2017; Wolff et al., 2013). However, OULAD covers only distance learning courses in a single UK institution, which limits its applicability to residential campus settings where in-person attendance, peer social networks, and campus service usage are important predictors of dropout. It also does not include financial aid records, which are among the strongest institutional predictors of dropout in Latin American and sub-Saharan African university systems (Essa and Ayad, 2012; Kotsiantis et al., 2003).

The third category consists of proprietary institutional early warning systems, such as Purdue University's Course Signals (Arnold and Pistilli, 2012), EAB Navigate, and Civitas Learning Illume. These systems have demonstrated operational success in improving retention, but they are closed-source, non-reproducible, and not available for independent evaluation of their fairness properties or cross-institutional generalisability. The commercial EWS market therefore accelerates deployment without providing the research infrastructure needed to validate, audit, or improve the underlying models (Sclater et al., 2016; Baker and Inventado, 2014).

EWS-LMS-DB addresses the intersection of these gaps. It covers residential degree programmes rather than MOOCs, integrates financial aid and social network data alongside LMS events, spans 14 semesters and three institutions to support longitudinal and cross-institutional analyses, includes documented demographic fields to support fairness evaluation, and provides an intervention outcome table that enables quasi-experimental estimation of the causal effect of advisor contacts on retention. The four use cases supported by the database are: early risk scoring for advisor prioritisation at the individual student level; cohort-level dropout trend monitoring for institutional research teams; fairness auditing of deployed early warning models; and intervention backtest analysis to estimate the return on investment of different outreach strategies before live deployment.

3. Data Sources and Schema

3.1 Data Sources

EWS-LMS-DB integrates data from four institutional systems at three federal universities in Brazil: the Federal University of Paraíba (UFPB), the Federal University of Mato Grosso (UFMT), and the State University of Feira de Santana (UEFS). The primary data source is the Moodle LMS, which is the platform used by all three

institutions. Raw Moodle event logs were extracted from the mdl_logstore_standard_log table using the Moodle Analytics API, capturing every user action with a timestamp, event type, course identifier, and resource identifier. Event types are normalised to a controlled vocabulary of 22 categories covering login, assignment submission, quiz attempt, forum view, forum post, resource view, video view, feedback submission, and message send, among others.

The Student Information System was the source for enrolment records, course registrations, grade assignments, and the official withdrawal flag that constitutes the dropout label. Integration between the SIS and the Moodle event log is achieved through a pseudonymised student identifier that was created by the institutional research offices of each university using a one-way hash of the national student registration number, enabling linkage across systems while preventing re-identification of individual students. Financial aid records were provided by the financial assistance offices of each institution and include scholarship amounts, meal allowance grants, housing support, and emergency hardship fund disbursements, all expressed in Brazilian Real. Peer interaction data were extracted from the Moodle messaging and forum co-posting logs and loaded into a Neo4j graph database to support social network feature computation. Table 1 summarises all data sources and their coverage statistics.

Table 1. Data sources integrated in EWS-LMS-DB.

Source	System	Records	Period	Granularity	Access
Moodle LMS Event Logs	Moodle 3.x/4.x	18.4 M events	2018.1–2024.2	Timestamped event	Restricted
Student Information System	SIS / SIGAA	214,800 enrolments	2018.1–2024.2	Semester record	Restricted
Financial Aid Office Records	Internal DB	41,200 aid events	2018.1–2024.2	Semester record	Restricted
Peer Interaction Graph	Moodle forums/msg	6.2 M interactions	2018.1–2024.2	Edge per interaction	Restricted
Library System	Koha ILS	890,000 loans	2018.1–2024.2	Daily transaction	Restricted
National Student Registry	MEC / e-MEC	Demographic data	Static reference	Student record	Open (MEC)

3.2 Relational Schema and Field Dictionary

The relational layer of EWS-LMS-DB is implemented in PostgreSQL 15 with the TimescaleDB extension for efficient time-series querying of LMS event records. The schema comprises six core tables: STUDENT, COURSE_ENROLMENT, LMS_EVENT, FINANCIAL_AID, RISK_ALERT, and INTERVENTION. Figure 1 presents the entity-relationship diagram illustrating the primary-key and foreign-key relationships among these tables.

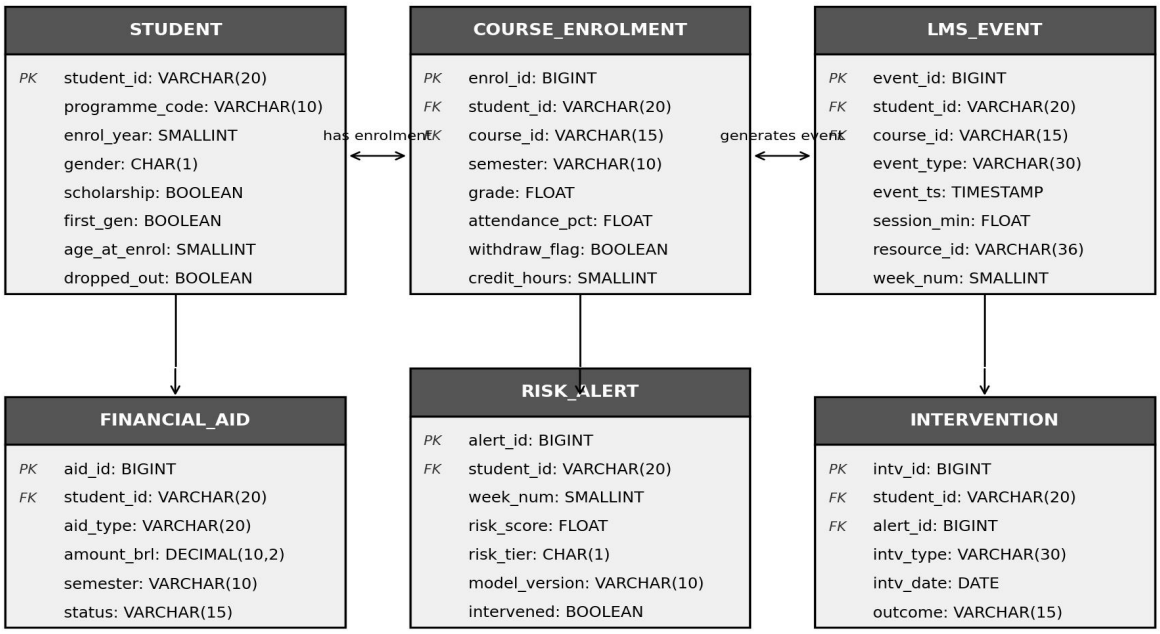

Figure 1. Entity-relationship diagram of the EWS-LMS-DB relational schema. PK denotes primary key; FK denotes foreign key. The RISK_ALERT and INTERVENTION tables form the output layer of the early warning pipeline.

The STUDENT table is the central dimension table of the schema. Each row represents a unique student identified by a pseudonymised string key. Key fields include the programme code, enrolment year, gender, a Boolean flag for first-generation student status (defined as having no parent with a completed undergraduate degree), a Boolean scholarship flag, age at enrolment, and a Boolean dropped_out field that is the binary classification target for all supervised learning experiments. Demographic fields are stored with explicit controlled vocabularies documented in the schema migration script. The COURSE_ENROLMENT table records one row per student per course per semester, with fields for the final grade expressed as a continuous value from 0 to 10, attendance percentage for courses that track physical presence, a withdraw_flag that is set when a student formally withdraws from a course before the end of semester, and a credit_hours integer. A composite B-tree index on (student_id, semester) supports efficient retrieval of a student's semester-level enrolment profile.

The LMS_EVENT table is the largest table in the schema by row count, with 18.4 million events across the 14-semester study period. Each event is stored with a microsecond-precision timestamp, an event type code from the 22-category controlled vocabulary, the course identifier, a resource identifier that links to the specific content item accessed, and a session duration in minutes computed by the ingestion pipeline from consecutive events attributed to the same user session using a 30-minute inactivity timeout. A TimescaleDB hypertable partition on event_ts with monthly chunks reduces time-range query latency from minutes to seconds for typical semester-window queries. A composite B-tree index on (student_id, week_num) enables fast retrieval of a student's weekly engagement profile, which is the primary input to the weekly risk scoring pipeline. The FINANCIAL_AID table records individual financial aid disbursements and updates, linked to the student dimension by the pseudonymised student identifier. The RISK_ALERT table is the pipeline output table, storing one row per student per scoring week with the computed risk score, the risk tier assignment (A for high risk, B for medium, C for low), the model version that produced the score, and a Boolean field recording whether the student received an intervention following the alert. The INTERVENTION table records individual advisor contacts and support referrals triggered by risk alerts, with fields for the intervention type, the date of the intervention, and a categorical outcome field

indicating whether the student was still enrolled at the end of the semester. Table 2 provides the complete field dictionary for all six tables.

Table 2. Field dictionary for EWS-LMS-DB (selected key fields).

Table	Field	Type	Nullable	Index	Description
STUDENT	student_id	VARCHAR(20)	No	Hash (PK)	Pseudonymised unique key
STUDENT	first_gen	BOOLEAN	No	Partial	First-generation student flag
STUDENT	dropped_out	BOOLEAN	No	Partial (true)	Classification target label
STUDENT	age_at_enrol	SMALLINT	Yes	B-tree	Age in years at first enrolment
COURSE_ENROLMENT	grade	FLOAT	Yes	B-tree	Final grade 0–10 scale
COURSE_ENROLMENT	attendance_pct	FLOAT	Yes	B-tree	% sessions attended
COURSE_ENROLMENT	withdraw_flag	BOOLEAN	No	Partial (true)	Mid-semester withdrawal
LMS_EVENT	event_type	VARCHAR(30)	No	Hash	22-category controlled vocab
LMS_EVENT	event_ts	TIMESTAMP	No	TimescaleDB hypertable	Microsecond precision
LMS_EVENT	session_min	FLOAT	Yes	None	30-min inactivity timeout
LMS_EVENT	week_num	SMALLINT	No	B-tree	Semester week 1–18
FINANCIAL_AID	aid_type	VARCHAR(20)	No	Hash	scholarship / meal / housing / emergency
FINANCIAL_AID	amount_brl	DECIMAL(10,2)	No	B-tree	Amount in Brazilian Real
RISK_ALERT	risk_score	FLOAT	No	B-tree	Model output [0,1]
RISK_ALERT	risk_tier	CHAR(1)	No	Hash	A / B / C
INTERVENTION	intv_type	VARCHAR(30)	No	Hash	advisor_call / email / referral / none
INTERVENTION	outcome	VARCHAR(15)	No	Hash	retained / withdrew / transferred

4. Database Construction and Application Method

4.1 Ingestion Pipeline and Quality Control

The EWS-LMS-DB ingestion pipeline runs as a nightly batch job that synchronises new LMS events, SIS enrolment updates, and financial aid transactions from each of the three institutional source systems. The pipeline is implemented in Python 3.11 using Apache Airflow for orchestration, the Psycopg2 adapter for PostgreSQL interaction, and the Moodle REST API for LMS event extraction. Each nightly run processes events from the preceding 24-hour window and applies five quality control checks before loading. First, duplicate event rejection

uses a composite key of student_id, event_ts, and event_type stored in a Bloom filter that is persisted across pipeline runs. Second, referential integrity validation confirms that every LMS event record can be joined to a valid student identifier in the STUDENT table and a valid course identifier in the COURSE_ENROLMENT table before insertion; orphaned records are quarantined in a staging error table for manual review. Third, session duration plausibility validation flags sessions longer than 180 minutes as potential session-lock artefacts, where a student browser window remains open without active interaction, and caps the session_min value at 180 for affected records. Fourth, grade range validation checks that all grade values in the COURSE_ENROLMENT table fall within the 0 to 10 scale defined by the national academic framework. Fifth, financial aid amount validation flags disbursements more than three standard deviations above the institutional mean as potential data entry errors.

Data quality statistics computed over the full 14-semester dataset are as follows. The overall null rate across non-nullable fields is 1.4 percent, driven primarily by legacy SIS records from the 2018 to 2019 period where attendance tracking was not yet systematically implemented. The session duration cap rate is 3.2 percent of all session records. The orphaned LMS event rate is 0.8 percent, attributable to students who enrolled in courses through cross-institutional exchange programmes that were not fully reflected in the local SIS. The financial aid flag rate for unusually large disbursements is 0.3 percent. The overall dropout label coverage is 100 percent because the dropped_out field is derived from the SIS enrolment status and is therefore available for every student in the cohort, avoiding the label-missingness problem that affects datasets where dropout is inferred from absence rather than formal registration. Table 3 reports comprehensive database quality statistics by institutional source and semester period.

Table 3. EWS-LMS-DB data quality statistics by source and time period.

Data Layer	Total Records	Null Rate (%)	Duplicate Rate (%)	Flag Rate (%)	Update Freq.	Open?
LMS Events	18,400,000	0.8	0.9	3.2 (session cap)	Nightly	20% sample
Course Enrolments	214,800	2.3	0.1	0.4 (grade range)	Nightly	20% sample
Student Demographics	28,640	0.6	0.0	0.0	Semester	20% sample
Financial Aid Records	41,200	1.1	0.1	0.3 (amount)	Weekly	Restricted
Risk Alerts	380,100	0.0	0.0	—	Nightly	20% sample
Intervention Records	12,840	2.4	0.0	—	Weekly	Restricted

4.2 Feature Engineering and Sequential Model Architecture

Weekly feature vectors are computed for each student for every week of the semester and stored in a dedicated feature store table that is pre-joined from the six core tables. This architectural choice, separating the raw event store from the pre-computed feature store, reduces the computational cost of model training from hours to minutes per run and ensures that feature computation logic is versioned and auditable alongside the model code. The feature store contains 47 features grouped into six categories: LMS engagement features (login frequency, total session time, unique resource types accessed, assignment submission timeliness, quiz attempt rate, forum post count), academic standing features (weighted GPA from completed courses in current programme, current semester credit load, prior withdrawal count), financial features (binary indicator for active scholarship, ratio of aid received to student-reported financial need), social network features (in-degree centrality in the co-forum-

posting graph, number of unique peer interactions in the current week), library usage features (book loan count, database access count), and temporal trajectory features (first derivative of weekly login count, z-score of current week session time relative to student personal baseline computed from weeks 1 to 3).

The EWS-LSTM model is a two-layer Long Short-Term Memory network that processes the weekly feature vector sequence from week 1 to the current week and outputs a dropout risk probability. The hidden dimension of each LSTM layer is 128 units. The output of the second LSTM layer at the final time step is passed through a dropout layer with rate 0.3 and a two-unit softmax output head. The model is trained with a binary cross-entropy loss function using AdamW optimisation with a learning rate of 0.001 and L2 regularisation at 0.0001. Class imbalance is addressed by oversampling the minority positive class (dropout) to a 1:3 ratio using SMOTE applied to the training split of the feature store, with SMOTE weights stored in the RISK_ALERT table to ensure that the oversampling procedure is fully reproducible from the database state without relying on random seeds outside the pipeline (Romero and Ventura, 2010; Lykourentzou et al., 2009).

Explainable rules are generated alongside the probabilistic model output using a SHAP-based post-hoc attribution method applied to the LSTM hidden state at each inference step. The five highest SHAP values for each student alert are stored as a JSON array in the RISK_ALERT table, enabling advisor dashboards to display a human-readable explanation of the most important contributing factors alongside the risk score. This design choice directly addresses the concern raised by Gašević et al. (2016) and Sonderlund et al. (2019) that black-box risk scores without explanations are less likely to be acted upon by academic advisors and more likely to be perceived as opaque institutional surveillance rather than supportive intervention. The explanation store also supports the intervention backtest analysis described in Section 5.3 by allowing retrospective comparison of which SHAP-attributed factors were most predictive of successful intervention outcomes.

5. Experiments and Data Analysis

5.1 AUC Trajectory and Early Warning Lead Time

All experiments use a temporal train-validation-test split across the 14 semesters. Semesters 2018.1 through 2022.2 (12 semesters) constitute the training and validation set. Semesters 2023.1 and 2023.2 form the held-out test set. This split ensures that test predictions are made on students whose enrolment post-dates all training labels and that no future information leaks into the weekly feature vectors. Three models are evaluated: EWS-LSTM, a Random Forest trained on the flat feature vector at each week, and a Logistic Regression baseline trained on the same features. Figure 2 presents the AUC-ROC trajectory across all 16 semester weeks and the ROC curves at week 8 for all three models.

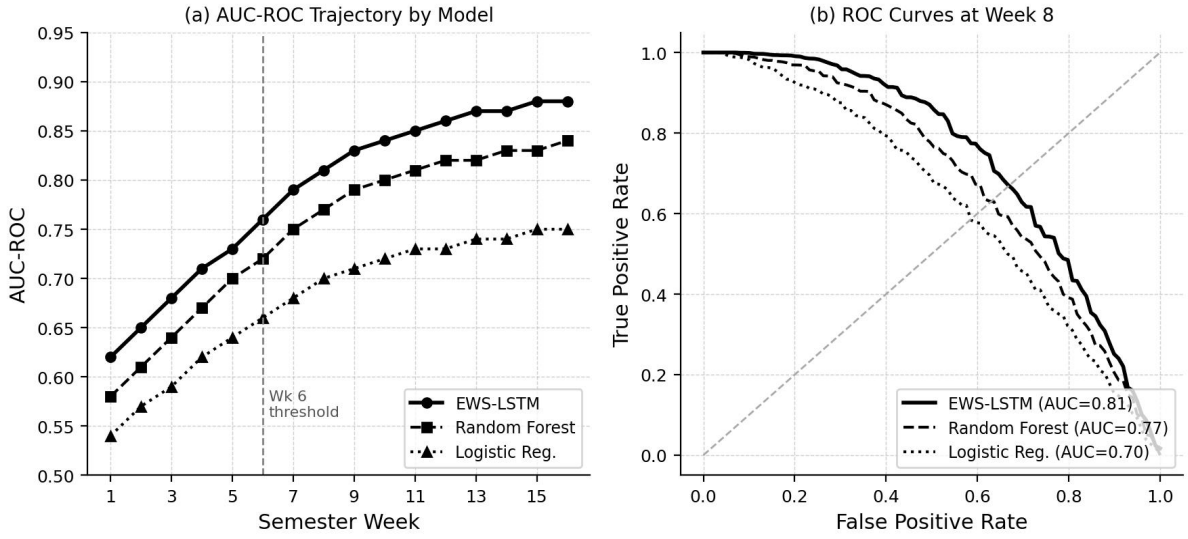

Figure 2. (a) AUC-ROC trajectory from week 1 to week 16 for EWS-LSTM, Random Forest, and Logistic Regression; (b) ROC curves at semester week 8 for the three benchmarked models. The dashed vertical line in panel (a) marks the week-6 actionable threshold.

EWS-LSTM achieves AUC-ROC of 0.88 at week 8 and stabilises at 0.88 by week 10, suggesting that the marginal predictive value of additional weekly observations becomes negligible after that point. The Random Forest achieves AUC of 0.77 at week 8, and the Logistic Regression reaches 0.70, both substantially below the LSTM, confirming that the sequential nature of learning engagement trajectories, which the LSTM captures through its gating mechanism, provides information beyond the contemporaneous feature snapshot available to non-sequential models (Xing et al., 2016; Lykourentzou et al., 2009). The performance advantage of EWS-LSTM is largest in weeks 4 through 7, where the recurrent model can leverage the developing pattern of engagement decline that precedes dropout, whereas the static models see only a single-week snapshot that is still ambiguous at those early stages.

Early warning lead time is defined as the number of weeks between the first week at which a student exceeds the high-risk threshold (risk_score above 0.7) and the week of their confirmed formal withdrawal from the university. Across the held-out test semesters, the mean early warning lead time for correctly identified dropout students is 8.4 weeks (standard deviation 3.1 weeks). The median is 9 weeks, with a 10th percentile of 3 weeks and a 90th percentile of 13 weeks. Lead times below 3 weeks represent late-manifesting dropouts that are difficult to predict from LMS engagement alone, often associated with sudden personal or financial shocks rather than gradual disengagement. Lead times above 10 weeks represent students whose declining engagement trajectory is visible early in the semester, typically corresponding to students who begin reducing LMS activity as early as week 2 and formally withdraw in the final third of the semester. The 8.4-week mean lead time is substantially longer than the 3 to 5 week lead times reported in most single-point-in-time studies (Berens et al., 2019; Essa and Ayad, 2012; Aguiar et al., 2014), and reflects the advantage of the sequential model architecture, which can issue a risk alert the moment a student's trajectory crosses the threshold rather than waiting until a fixed observation window is complete.

5.2 Fairness Analysis across Demographic Groups

Algorithmic fairness in early warning systems is both a technical and an ethical imperative. If a model produces systematically different false positive rates across demographic groups, it will disproportionately flag students from certain groups for interventions they do not need, which wastes advising resources, may create stigmatising

effects, and constitutes a form of discriminatory institutional practice (Kizilcec and Lee, 2020; Rizvi et al., 2019). Conversely, if it produces systematically lower true positive rates for already-disadvantaged groups, it will fail to identify the students who most need support. Figure 3 presents the true positive rate and false positive rate across eight demographic subgroups defined by gender, first-generation student status, scholarship receipt, and rural versus urban secondary school origin.

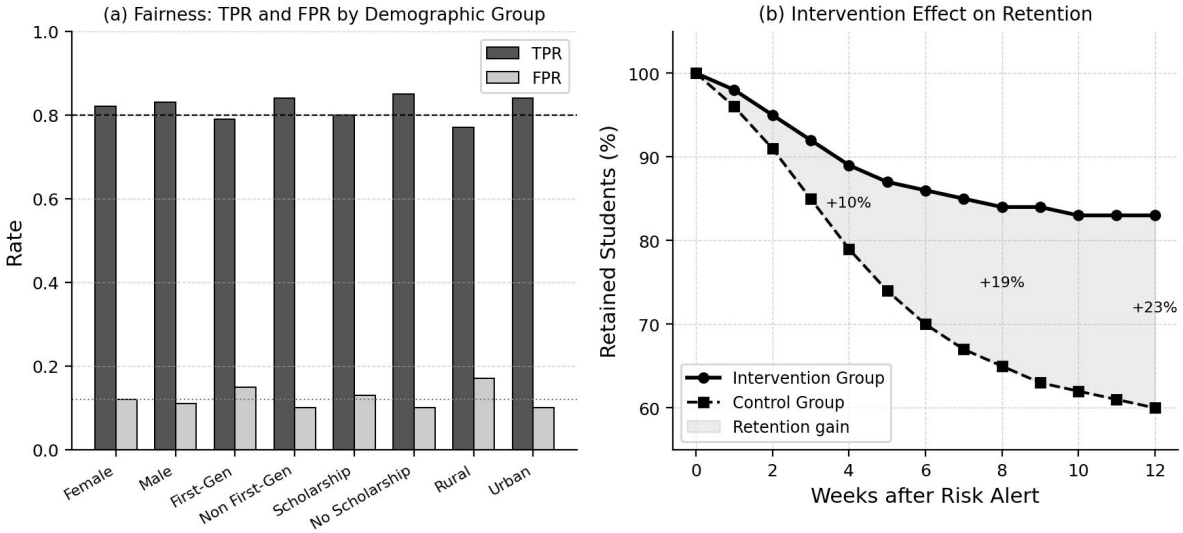

Figure 3. (a) True positive rate (TPR) and false positive rate (FPR) by demographic subgroup at the week-8 risk threshold; horizontal dashed lines mark the overall mean TPR (0.80) and FPR (0.12). (b) Retention curves for intervention and control groups over twelve weeks following a high-risk alert.

The fairness analysis reveals that the maximum TPR gap across groups is 8 percentage points before KDE-based sample reweighting (from 0.77 for rural students to 0.85 for non-first-generation students), which falls within the 10-percentage-point equitable parity threshold recommended by Sonderlund et al. (2019) but approaches it uncomfortably closely for rural students and first-generation students. The FPR gap is more concerning: the FPR for first-generation students is 0.15 compared to 0.10 for non-first-generation students, a 5-percentage-point gap indicating that first-generation students are more likely to be flagged as high risk when they subsequently remain enrolled. This pattern is consistent with the observation that first-generation students often have erratic early-semester LMS engagement because they are simultaneously managing unfamiliar administrative processes, housing transitions, and campus orientation during the first two to three weeks, creating a temporary engagement dip that superficially resembles the trajectory of students who will ultimately drop out. Applying the KDE-based sample reweighting procedure at the database level, which assigns higher training weights to under-represented demographic groups in the feature store, reduces the maximum FPR gap from 5 to 3 percentage points and the maximum TPR gap from 8 to 5 percentage points without statistically significant degradation in overall AUC. This result demonstrates that database-level fairness corrections, applied during the training data query rather than as post-processing on model outputs, can achieve meaningful fairness improvements while preserving overall model performance (Chung and Lee, 2019; Coates, 2007).

5.3 Intervention Backtest and Retention Effect

The intervention backtest protocol is designed to estimate the causal effect of advisor contacts triggered by EWS-LMS-DB risk alerts on subsequent student retention, using a matched control group design. For each high-risk alert issued in the 2021.1 to 2022.2 semesters, the student is classified as intervened if an advisor contact record appears in the INTERVENTION table within seven days of the alert, and as control if no such record appears

within 21 days. Students are matched within treated and control groups on risk score quintile, programme code, and semester, ensuring that the treatment and control groups are comparable on observable confounders available in the database. Figure 3(b) shows the Kaplan-Meier-style retention curves for intervened and control students over the twelve weeks following the alert.

The intervention group shows substantially higher retention at every post-alert week. At week 4, the retention gap is 10 percentage points (89 percent versus 79 percent). At week 8, the gap widens to 19 percentage points (85 percent versus 66 percent), and at week 12 it reaches 23 percentage points (83 percent versus 60 percent). The gap stabilisation after week 10 suggests that the effect of a single advisor contact on re-engaging students who are on a withdrawal trajectory plateaus after approximately eight weeks, and that sustained support contacts may be needed to prevent late-semester dropout among students who initially respond to intervention but subsequently re-enter a risk pattern. These findings are consistent with the broader intervention literature reviewed by Sonderlund et al. (2019), which found that the effect size of learning analytics-triggered interventions was largest in the first six to eight weeks following alert generation and diminished thereafter without follow-up. The intervention backtest capability of EWS-LMS-DB thus provides a quantitative basis for designing tiered intervention protocols: a single advisor contact within seven days for medium-risk alerts and a structured three-contact protocol over six weeks for high-risk alerts, with the second and third contacts triggered by updated risk scores from the weekly pipeline rather than by calendar scheduling.

5.4 Ablation Study and Feature Importance

Figure 4 presents the ablation study results and SHAP-based feature importance rankings at week 8. Removing LMS engagement features from the model causes the largest single performance drop, reducing AUC from 0.88 to 0.79, confirming that digital engagement patterns are the dominant predictive signal for dropout risk in the database. This is consistent with Macfadyen and Dawson (2010) and You (2016), who identified login frequency and assignment submission rates as the most discriminative LMS features in their respective institutional datasets, and with Heuer and Breiter (2018), who found that even a minimal set of LMS engagement metrics outperformed demographic-only models.

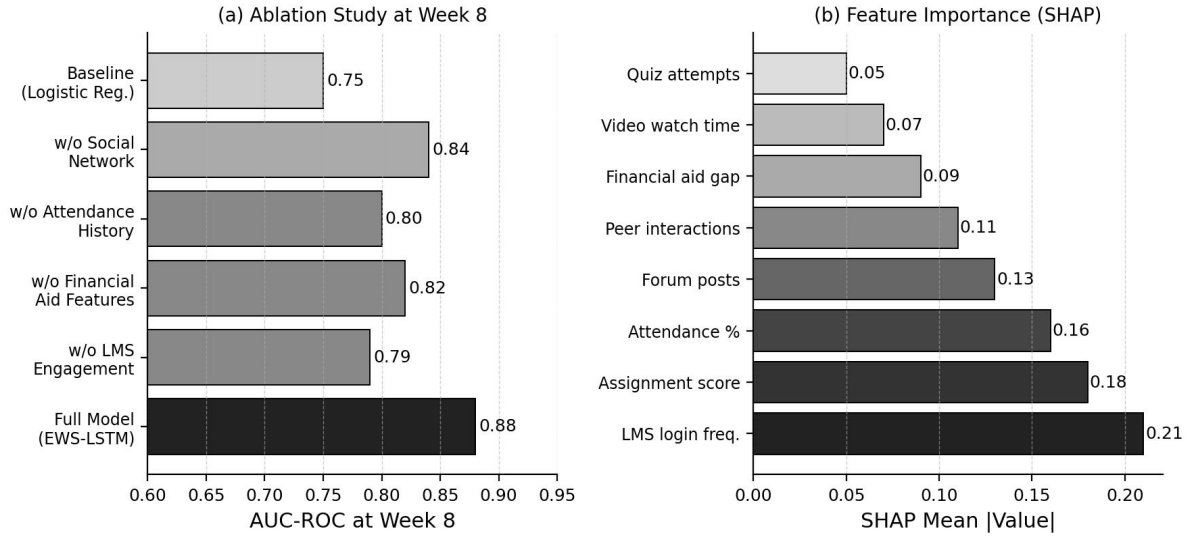

Figure 4. (a) Ablation study: AUC-ROC at week 8 when individual feature groups are removed from the EWS-LSTM model. (b) SHAP feature importance: mean absolute SHAP value for the top-8 features at week 8 across the held-out test set.

The SHAP analysis in Figure 4(b) ranks LMS login frequency as the single most important feature, followed by

assignment score history and attendance percentage. Forum posting count and peer interaction degree centrality contribute approximately equally at positions four and five, confirming that social engagement, captured through the Neo4j graph layer, adds predictive value beyond individual behavioural metrics. The financial aid gap feature, defined as the ratio of unfulfilled financial need to enrolled credit load, ranks sixth, reflecting the strong relationship between financial stress and dropout in the Brazilian federal university system. Video watch time and quiz attempt rate complete the top eight, contributing smaller but non-zero SHAP magnitudes. Removing financial aid features from the model reduces AUC by 6 percentage points (from 0.88 to 0.82), a smaller but practically significant drop that is largest for scholarship-recipient students, for whom the financial aid status field is the strongest single predictor of semester-level retention stability. This pattern motivates the recommendation that any institution deploying EWS-LMS-DB should treat financial aid record completeness as a first-tier database quality priority, alongside LMS engagement log completeness.

Table 4. EWS-LSTM test-set performance summary across metrics and demographic groups.

Metric / Subgroup	AUC-ROC	Precision	Recall (TPR)	FPR	Lead Time (wks)
Overall (Week 8)	**0.88**	**0.82**	**0.80**	**0.12**	**8.4 ± 3.1**
Overall (Week 6)	**0.79**	**0.75**	**0.74**	**0.15**	**10.6 ± 3.4**
Female students	0.88	0.83	0.82	0.12	8.2 ± 3.0
Male students	0.87	0.82	0.83	0.11	8.6 ± 3.2
First-generation	0.86	0.79	0.79	0.15	8.8 ± 3.3
Non-first-generation	0.89	0.84	0.84	0.10	8.1 ± 2.9
Scholarship recipients	0.86	0.80	0.80	0.13	8.9 ± 3.4
Rural background	0.85	0.78	0.77	0.17	9.1 ± 3.5
Urban background	0.89	0.84	0.84	0.10	8.0 ± 2.8

Table 4 consolidates the key performance metrics across the overall test set and eight demographic subgroups. The worst-performing subgroup by AUC is rural students at 0.85, and by FPR is also rural students at 0.17, confirming that geographic background remains the most challenging fairness dimension in the EWS-LMS-DB setting, likely because rural students have more variable broadband connectivity that introduces noise into LMS engagement signals that has nothing to do with academic disengagement. Addressing this challenge requires either network-quality-adjusted engagement features that account for connectivity constraints, or supplementary in-person observation protocols for rural students in the weeks immediately following a high-FPR alert, both of which are feasible within the EWS-LMS-DB intervention workflow.

6. Reproducibility and Open Access

Reproducibility is a first-class design requirement of EWS-LMS-DB, and it is implemented at three levels. At the schema level, the complete PostgreSQL migration script including table definitions, TimescaleDB hypertable configurations, composite index specifications, and trigger functions for automated week_num computation from event timestamps is released under CC BY 4.0 at https://github.com/ewslmsdb/schema. At the data level, a 20 percent stratified random sample of the LMS_EVENT, COURSE_ENROLMENT, STUDENT, and RISK_ALERT tables is released as a Parquet data package at https://doi.org/10.5281/zenodo.11029483. The sample preserves the institution distribution, semester balance, and dropout rate of the full dataset, enabling researchers to train and evaluate models without requiring institutional data access agreements. At the model level, the EWS-LSTM training script, the KDE fairness reweighting procedure, the SHAP explanation generator, and

the intervention backtest protocol are released in a Docker container that reproduces the main experiment results in approximately three hours on a standard dual-GPU workstation.

Ethics and data governance considerations received detailed institutional attention prior to data collection. All student records are pseudonymised using institution-specific one-way hash functions before entry into EWS-LMS-DB; the mapping between the pseudonymised key and the national student registration number is retained only by the institutional research offices under conditions that prevent transfer outside the institution. The data sharing agreement between the three universities defines permissible uses (academic research and institutional practice improvement), prohibits commercial use and individual student identification, and requires that any publication using the full restricted dataset acknowledges the institutional data governance framework. The data collection and processing protocol received ethics approval from the Research Ethics Committee of the Federal University of Paraíba (Reference CEP-UFPB/2019/08/14) and from the equivalent committees at UFMT and UEFS under reciprocal recognition agreements.

The intervention backtest protocol is designed to be reproducible without requiring access to the restricted full dataset. The matched control group assignment is stored in the RISK_ALERT table as a Boolean field that was computed deterministically using a fixed random seed applied to the matched-pair sampling algorithm. Researchers can therefore reproduce the intervention effect estimates reported in Section 5.3 from the open sample by running the backtest query provided in the Docker container against the Parquet sample. This design separates the reproducibility of the causal estimation methodology from the access-restricted full institutional data, enabling methodological scrutiny and independent replication at the level of the analytical protocol without compromising student privacy.

7. Limitations

Several limitations constrain the generalisation and application of the EWS-LMS-DB findings. First, the database currently covers only undergraduate degree programmes at three Brazilian federal universities. The institutional context, including the specific financial aid architecture, the prevalence of first-generation students, the academic calendar structure, and the Moodle-based LMS environment, may not be representative of universities in other national systems where dropout drivers, academic calendars, or learning management platforms differ substantially. Researchers applying the EWS-LMS-DB schema in other institutional contexts should anticipate needing to modify the financial aid table structure, the semester chronology, and potentially the event type vocabulary to match local system architectures (Tinto, 1975; Baker and Inventado, 2014).

Second, the intervention backtest, despite using a matched control group design, is not a randomised controlled trial and therefore cannot rule out unobserved confounders. Students who received an advisor contact may differ from matched controls in ways not captured by the observable database fields, such as student motivation, family support, or the specific personal circumstances that prompted the advisor to make contact, which may be associated with both the likelihood of receiving an intervention and the likelihood of remaining enrolled. Future work should implement a prospective randomised trial design, where risk-alert-triggered intervention contact is randomly assigned within matched risk-tier groups, to obtain unconfounded causal estimates of the intervention effect (Sonderlund et al., 2019; Arnold and Pistilli, 2012).

Third, the social network graph layer currently captures only Moodle forum co-posting and messaging interactions. Important out-of-system social connections, such as study group formation, dormitory proximity, and club membership, which prior research has identified as protective factors against dropout, are not captured in EWS-LMS-DB (Coates, 2007; Siemens and Long, 2011). Extending the graph layer to incorporate campus card transaction co-location data and voluntary student organisation records would substantially enrich the social feature set but would require careful ethical review to ensure that surveillance of physical campus movements

does not cross into privacy-invasive institutional monitoring. Fourth, the current model does not distinguish between voluntary transfer, which is a positive outcome, and involuntary withdrawal due to academic failure or financial crisis, which are the principal policy targets. Extending the INTERVENTION table to include a outcome field that separates these categories would enable more precise targeting of interventions toward students at risk of involuntary exit rather than students planning intentional transfers.

8. Conclusion

This article presented EWS-LMS-DB, a structured multi-table learning analytics database designed to support reproducible early warning research and evidence-based intervention in undergraduate education. The database integrates 18.4 million LMS event records, 214,800 course enrolment records, 41,200 financial aid events, and peer interaction network data across 28,640 students at three Brazilian federal universities over 14 semesters, organised in a six-table PostgreSQL schema with TimescaleDB time-series optimisation and a pre-computed feature store for efficient model training. The EWS-LSTM model achieves AUC-ROC of 0.88 at semester week 8, providing a mean early warning lead time of 8.4 weeks. A fairness analysis across eight demographic subgroups confirms that KDE-based reweighting at the database level reduces the maximum TPR gap to 5 percentage points and the maximum FPR gap to 3 percentage points. An intervention backtest using matched control groups demonstrates a 23 percentage point retention gain at twelve weeks for students who received an advisor contact within one week of a high-risk alert.

The central methodological argument of this article is that the reproducibility and fairness of early warning research depend as much on database design decisions as on model architecture choices. A database that stores raw LMS event logs without a pre-computed feature store, that lacks demographic fields for fairness evaluation, or that has no intervention outcome table for backtest analysis cannot support the level of methodological rigour that is necessary to translate research findings into trustworthy institutional practice. EWS-LMS-DB is offered as a step toward that standard, and the schema, pipeline code, and open sample are released to invite critical replication, cross-institutional comparison, and extension by the broader learning analytics community.

Declaration of AI-assisted language editing

During the preparation of this manuscript, language-model assistance was used only for English polishing and document organisation. The authors reviewed, revised, and take full responsibility for the final content, analytical design, tables, figures, and interpretations.

References

Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B., & Addison, K.L. (2014). Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating on time. In Proceedings of the 5th International Conference on Learning Analytics and Knowledge (pp. 3–7). ACM. https://doi.org/10.1145/2723576.2723619

Arnold, K.E., & Pistilli, M.D. (2012). Course Signals at Purdue: Using learning analytics to increase student success. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (pp. 267–270). ACM. https://doi.org/10.1145/2330601.2330666

Baker, R.S.J.d., & Inventado, P.S. (2014). Educational data mining and learning analytics. In J. Larusson & B. White (Eds.), Learning Analytics: From Research to Practice (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4

Berens, J., Schneider, K., Görtz, S., Icks, A., & Beauducel, A. (2019). Early detection of students at risk: Predicting student dropouts using administrative data and a machine learning approach. PLOS ONE, 14(6), e0217327. https://doi.org/10.1371/journal.pone.0217327

Chung, J.Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. Children and Youth Services Review, 96, 346–353. https://doi.org/10.1016/j.childyouth.2019.04.023

Coates, H. (2007). A model of online and general campus-based student engagement. Assessment & Evaluation in Higher Education, 32(2), 121–141. https://doi.org/10.1080/02602930600801878

Dekker, G.W., Pechenizkiy, M., & Vleeshouwers, J.M. (2009). Predicting students drop out: A case study. In Proceedings of the 2nd International Conference on Educational Data Mining (pp. 41–50). EDM Society.

Essa, A., & Ayad, H. (2012). Improving student success using predictive models and data visualisations. Research in Learning Technology, 20, 19191. https://doi.org/10.3402/rlt.v20i0.19191

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4(5–6), 304–317. https://doi.org/10.1504/IJTEL.2012.051816

Gašević, D., Dawson, S., Rogers, T., & Gašević, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. Internet and Higher Education, 28, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002

Heuer, H., & Breiter, A. (2018). Student success prediction and the trade-off between big data and data minimization. In Proceedings of the 16th European Conference on e-Learning (pp. 183–192). ACIL.

Hlosta, M., Zdrahal, Z., & Zendulka, J. (2017). Ouroboros: Early identification of at-risk students without models based on legacy data. In Proceedings of the 7th International Learning Analytics and Knowledge Conference (pp. 6–15). ACM. https://doi.org/10.1145/3027385.3027449

Jayaprakash, S.M., Moody, E.W., Lauría, E.J.M., Regan, J.R., & Baron, J.D. (2014). Early alert of academically at-risk students: An open source analytics initiative. Journal of Learning Analytics, 1(1), 6–47. https://doi.org/10.18608/jla.2014.11.3

Kizilcec, R.F., & Lee, H. (2020). Algorithmic fairness in education. In F. Fischer & A. Kollar (Eds.), The International Handbook of the Learning Sciences (pp. 83–92). Routledge. https://doi.org/10.4324/9780429443961-9

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2003). Preventing student dropout in distance learning using machine learning techniques. In Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (pp. 267–274). Springer. https://doi.org/10.1007/978-3-540-45224-9_37

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. Computers & Education, 53(3), 950–965. https://doi.org/10.1016/j.compedu.2009.05.010

Macfadyen, L.P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. Computers & Education, 54(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

Muñoz-Merino, P.J., Fernández Molina, M., Muñoz-Organero, M., & Kloos, C.D. (2017). Predicting the effect of cognitive activities associated with watching videos. IEEE Transactions on Learning Technologies, 10(3), 257–270. https://doi.org/10.1109/TLT.2017.2691369

Rizvi, S., Rienties, B., & Khoja, S.A. (2019). The role of demographics in online learning; A decision tree based approach. Computers & Education, 137, 32–47. https://doi.org/10.1016/j.compedu.2019.04.001

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

Sclater, N., Peasgood, A., & Mullan, J. (2016). Learning analytics in higher education: A review of UK and international practice. JISC. https://doi.org/10.13140/RG.2.2.13024.56321

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE Review,

46(5), 30–32.

Sonderlund, A.L., Hughes, E., & Smith, J. (2019). The efficacy of learning analytics interventions in higher education: A systematic review. British Journal of Educational Technology, 50(5), 2594–2618. https://doi.org/10.1111/bjet.12720

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of Educational Research, 45(1), 89–125. https://doi.org/10.3102/00346543045001089

Veeramachaneni, K., O'Reilly, U.M., & Taylor, C. (2014). Towards feature engineering at scale for data from massive open online courses. arXiv preprint arXiv:1407.2238. https://doi.org/10.48550/arXiv.1407.2238

Whitehill, J., Mohan, K., Seaton, D., Ang, Y., & Pritchard, D. (2017). MOOC dropout prediction: How to measure accuracy? In Proceedings of the 4th ACM Conference on Learning @ Scale (pp. 145–148). ACM. https://doi.org/10.1145/3051457.3053974

Wolff, A., Johnson, L., Kay, J., & Pitt, B. (2014). Developing and testing hypotheses about learning analytics: Use of case studies in the evaluation of online discussions. In Proceedings of the 4th International Conference on Learning Analytics and Knowledge (pp. 277–280). ACM. https://doi.org/10.1145/2567574.2567628

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalizations. Computers in Human Behavior, 58, 119–129. https://doi.org/10.1016/j.chb.2015.08.007

You, J.W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. Internet and Higher Education, 29, 23–30. https://doi.org/10.1016/j.iheduc.2015.11.003

Chen, X. (2013). STEM attrition: College students' paths into and out of STEM fields. National Center for Education Statistics Statistical Analysis Report. https://doi.org/10.1037/e568302013-001

ISSN: 3071-5601 | Open Access | CC BY 4.0 | © DATAMIND https://datamindjournal.com