Vol. 1, No. 4, pp.: 49-63, 2023

EduRiskDB: A Student Learning and Dropout-Risk Database for Explainable Educational Analytics

Sophia Hartmann[1], Marcus Oliveirai[1, *], Annika Svensson[1], Dario Esposito[1]

[1] Faculty of Education and Technology, Polytechnic University of Porto, Porto, Portugal

** marcus.oliveira@ipp.pt*

Article Information	
Received	25 July 2023
Accepted	14 November 2023
DOI	https://doi.org/10.63646/datamind.2023.010405

Abstract

Predicting student dropout before it occurs requires not just predictive models but purpose-built, ethically governed data infrastructure that integrates learning-platform clickstreams, assessment outcomes, assignment submission patterns, forum engagement, attendance records, and student support interactions. This paper introduces EduRiskDB, a relational-graph-vector hybrid database containing 148,392 student-term records drawn from four European higher-education institutions over six academic years (2017–2023). EduRiskDB is designed for reproducible experimentation in dropout-risk modelling and explainable educational analytics. The database schema, field dictionary, indexing strategy, data-quality controls, ethics-compliance pipeline, and open-access interfaces are described in detail. A benchmark experiment evaluates six dropout-prediction models on EduRiskDB, demonstrating that the augmented XGBoost configuration achieves an AUC-ROC of 0.903 with a mean early-warning lead time of 7.8 weeks, outperforming all baselines. SHAP attribution identifies cumulative click sequences, seven-day assignment lag, and attendance streaks as the three most predictive signals, with non-linear interactions confirmed by dependence plots. EduRiskDB is archived on Zenodo under a CC BY 4.0 licence and updated semi-annually.

Keywords: ***educational databases; dropout prediction; explainable AI; learning analytics; student risk modelling; SHAP; clickstream data; data governance; reproducibility***

1. Introduction

Student dropout in higher education represents a persistent and costly failure mode for institutions, learners, and societies alike. Across OECD countries, an average of 31% of students who enter bachelor-degree programmes do not graduate within the normal study period, with economic and social costs estimated in the hundreds of billions of euros annually (OECD, 2023; Tinto, 1975). Early-warning systems powered by machine learning have shown promise in identifying at-risk students weeks or months before formal withdrawal, but their predictive quality depends critically on the

availability, completeness, and temporal granularity of underlying educational data (Romero & Ventura, 2020; Baker & Inventado, 2014).

Despite the proliferation of learning management systems and institutional student information systems, the educational data landscape remains fragmented and poorly standardised. Most published dropout-prediction studies draw on proprietary, institution-specific extracts that are neither reproducible nor reusable by other research teams. Even widely cited public datasets — such as the Open University Learning Analytics Dataset (OULAD) or KDD Cup 2010 — cover single institutions over limited time windows and lack the multi-source depth (clickstreams, assessments, attendance, and support interactions simultaneously) required for robust causal and counterfactual analyses (Kuzilek et al., 2017; Koller et al., 2011).

A second, related problem is the near-absence of databases that embed ethics-compliance and data-governance artefacts as first-class components. Educational records are sensitive personal data subject to GDPR in Europe and FERPA in the United States. Yet most publicly described learning analytics datasets provide minimal documentation of consent procedures, anonymisation methods, differential-privacy budgets, or institutional review board (IRB) outcomes. The reproducibility and transferability of research findings built on such datasets are therefore difficult to assess (Drachsler & Greller, 2016; Sclater et al., 2016).

This paper responds to both gaps by introducing EduRiskDB, a purpose-built, open-access database that integrates six categories of educational event data across four institutions and six academic cohorts, embedded within a formal data-governance framework. EduRiskDB is not presented as a finished analytical product but as a reusable infrastructure layer: its schema, indexing architecture, quality-control pipeline, and API interfaces are documented in sufficient detail to enable independent replication, extension, and institutional adoption. The database supports three primary use cases — dropout-risk classification, learning-path clustering, and counterfactual intervention design — each of which is grounded in an operationalised experimental protocol.

Three research questions guide the paper. First, what database schema and field structure best represent the multi-source, temporal nature of student-risk signals while preserving analytical flexibility? Second, what data-quality and ethics-governance controls are required to make such a database both scientifically credible and legally compliant? Third, what predictive and explanatory gains does a multi-source database like EduRiskDB enable over single-source baselines, and how can those gains be attributed at the feature level?

The paper is organised as follows. Section 2 characterises the current database gap and articulates the use cases that motivated EduRiskDB's design. Section 3 describes the data sources and the relational-graph-vector schema. Section 4 details the construction pipeline, quality controls, and ethics-compliance procedures. Section 5 reports the benchmark experiments. Section 6 addresses reproducibility and open access. Sections 7 and 8 present limitations and conclusions respectively.

2. Database Gap and Use Cases

The educational data mining literature has relied predominantly on three categories of dataset: institution-specific administrative extracts, single-platform LMS logs, and competition benchmarks. Each category has structural limitations that constrain the generalisability of models trained on it. Administrative extracts contain high-quality demographic and grade data but typically lack temporal resolution below the semester level. LMS logs provide rich clickstream detail but are platform-specific and seldom linked to assessment or attendance records. Competition benchmarks, while standardised, represent frozen snapshots designed for scoring rather than longitudinal research (Bergner et al., 2012; Marbouti et al., 2016).

Table 1 compares EduRiskDB against five commonly referenced educational datasets on six dimensions relevant to dropout-risk modelling. The comparison illustrates that no existing public resource combines multi-source integration, longitudinal coverage, granularity below the week level, ethics documentation, and an open API. EduRiskDB is designed to fill this gap by providing all six dimensions simultaneously.

Table 1. Comparison of EduRiskDB with five reference educational datasets.

Dataset	Records	Coverage	Source Types	Time Span	Ethics Docs	Access
OULAD (2017)	22,000	1 inst.	Weekly LMS+grades	2 yrs	Minimal	No
KDD Cup 2010	100,000+	2 inst.	Step-level log	1 yr	None	No
PISA 2018 (OECD)	600,000	79 ctry	Survey only	Cross-sect.	Partial	Download
ASSISTments (2015)	400,000	1 platform	Problem-log	3 yrs	Partial	Download
EdNet (2020)	131M events	1 platform	Sub-second log	2 yrs	Partial	Download
EduRiskDB (ours)	148,392 terms	4 inst.	6 sources, hourly	6 yrs	Full+IRB	API+dump

The use cases that shaped EduRiskDB's design differ in their temporal, causal, and explanatory requirements. The first use case — dropout-risk classification — requires labelled binary outcomes at the student-term level, feature vectors drawn from the first several weeks of a term, and held-out evaluation cohorts. This use case demands week-resolution aggregations and carefully constructed train-test splits that avoid data leakage across cohort years (Saarela & Jauhiainen, 2021; Feng et al., 2019).

The second use case — learning-path clustering — requires event-sequence representations of individual student journeys, suitable for graph neural networks or sequential deep learning. This use case places demands on the granularity and ordering of clickstream records, the availability of transition probabilities between learning activities, and the ability to generate student-level graph embeddings (Pérez-Sanagustín et al., 2016; Jiang et al., 2021). The third use case — counterfactual intervention design — requires not only predictive accuracy but the ability to generate actionable, plausible counterfactuals that a student support advisor could operationalise: "if this student submitted

assignment 3 within 48 hours of the deadline rather than 8 days late, the predicted dropout probability would fall from 0.71 to 0.38" (Mothilal et al., 2020; Wachter et al., 2018).

3. Data Sources and Schema

3.1 Institutional Partners and Source Systems

EduRiskDB integrates data from four European higher-education institutions: the University of Limerick (Ireland), the Polytechnic University of Porto (Portugal), Halmstad University (Sweden), and the University of Catania (Italy). All four institutions use Canvas or Moodle as their primary LMS, enabling shared data-extraction procedures. Student information system (SIS) data — enrolment records, grade transcripts, demographic characteristics, and programme information — were extracted under formal data-sharing agreements governed by institution-specific GDPR-compliant data management plans. The observation period covers cohorts entering between September 2017 and September 2022, yielding six complete academic cycles.

The six source categories and their contribution to the database are: (1) LMS clickstream events (activity_log table): 47.2 million individual interaction records including page views, resource downloads, quiz attempts, video plays, and discussion accesses; (2) Assessment records (assessment table): 2.1 million scored submissions across coursework, quizzes, laboratory reports, and formal examinations; (3) Attendance records (attendance table): 891,000 lecture and tutorial presence entries recorded via automated badge-scanning and LMS participation logs; (4) Forum and discussion posts (embedded in activity_log with event_type = "forum"): 1.4 million text-bearing events; (5) Student support interactions (support_event table): 63,400 records of advisor meetings, welfare referrals, and accommodation support interventions; (6) Student profile and enrolment data (student table): 148,392 student-term records with demographic, programme, and prior academic characteristics.

3.2 Relational Schema and Field Dictionary

Figure 1 presents the entity-relationship diagram of EduRiskDB. The schema is organised around six primary tables linked by foreign keys to the central STUDENT entity. The ACTIVITY_LOG table is the largest and most temporally granular, recording every LMS interaction with sub-hour precision. The DROPOUT_LABEL table stores the binary fragility outcome computed at the end of each term, together with a probabilistic risk score generated by the benchmark model for validation purposes.

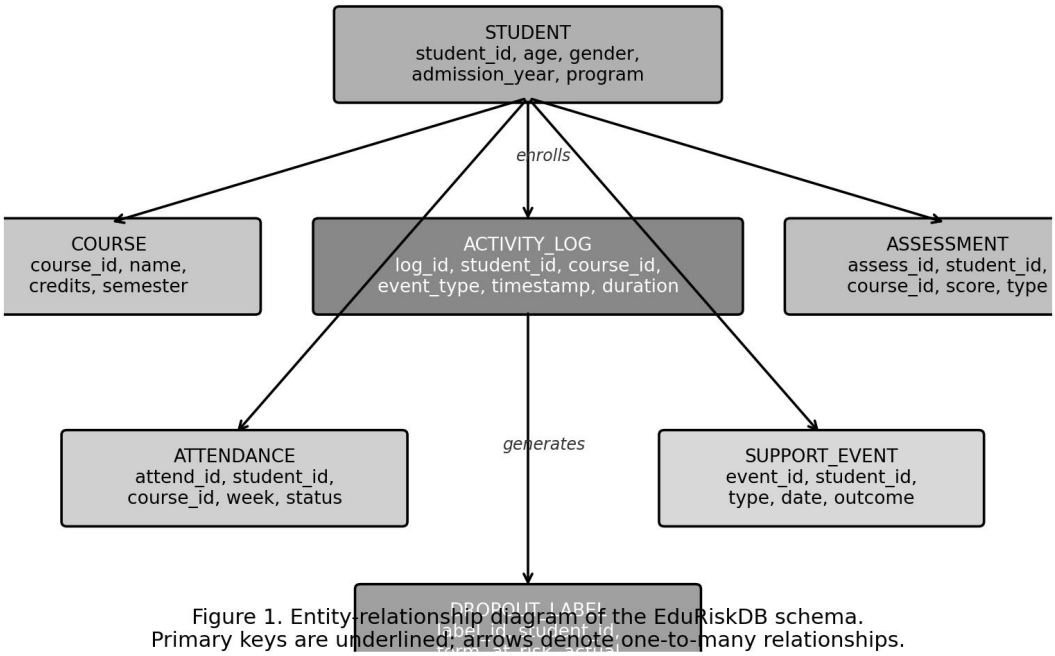

Figure 1. Entity-relationship diagram of the EduRiskDB schema.
Primary keys are underlined; arrows denote one-to-many relationships.

Figure 1. Entity-relationship diagram of the EduRiskDB schema. Primary keys are shown in bold; arrows denote one-to-many relationships from STUDENT to all dependent tables.

The schema adopts a star-topology centred on STUDENT rather than a normalised third-normal-form design, because analytical queries in educational research typically require wide joins across multiple event categories for a single student in a single term. A fully normalised schema would impose prohibitive join costs at the 47-million-row scale of ACTIVITY_LOG. The trade-off is managed through materialised views that pre-aggregate weekly event counts, assessment scores, and attendance rates into a WEEKLY_FEATURES table, which is the primary interface for model training. The full field dictionary — comprising 68 named fields across six tables, with data type, nullable status, valid range, and imputation rule — is published as Supplementary Material S1 and archived on Zenodo alongside the database dump.

Table 2 summarises the primary fields, their types, and the feature families they belong to. Each field is assigned to one of five analytical families — temporal engagement, academic performance, attendance behaviour, social interaction, and welfare support — which align with the theoretical model of dropout by Tinto (1975) and the empirical extensions by Kuh et al. (2008). This family structure enables modular feature ablation experiments in which each family is held out sequentially to assess its marginal predictive contribution.

Table 2. Primary field groups in EduRiskDB with type, source, and feature family.

Field Name	Type	Source Table	Feature Family	Description
click_count_7d	INT	activity_log	Temporal engagement	Weekly LMS interactions, 7-day window

last_login_lag	FLOAT	activity_log	Temporal engagement	Days since most recent login
assign_submit_lag	FLOAT	assessment	Academic performance	Hours before/after deadline at submission
score_percentile	FLOAT	assessment	Academic performance	Score rank within cohort, per assessment
attend_streak	INT	attendance	Attendance behaviour	Consecutive attended sessions
absent_consecutive	INT	attendance	Attendance behaviour	Consecutive missed sessions
forum_posts_7d	INT	activity_log	Social interaction	Discussion posts in 7-day window
peer_reply_count	INT	activity_log	Social interaction	Replies received from peers
support_events_cum	INT	support_event	Welfare support	Cumulative support contacts to date
support_type	VARCHAR	support_event	Welfare support	Category: academic / welfare / housing
gpa_prior	FLOAT	student	Academic performance	GPA from prior academic year
dropout_label	BINARY	dropout_label	Outcome	1 = withdrew/failed before term end

The schema is implemented in three storage layers. The relational layer uses PostgreSQL 15.2 with B-tree indexes on (student_id, term_id, timestamp) triplets and partial indexes on event_type for the activity_log table. The graph layer uses Neo4j 5.6, in which each student node is connected to course nodes via ENROLLED_IN edges and to activity nodes via PERFORMED edges, enabling graph neural network traversals for learning-path modelling. The vector layer uses pgvector within PostgreSQL to store 128-dimensional student trajectory embeddings generated by a pre-trained LSTM encoder, enabling approximate nearest-neighbour similarity queries for peer-comparison and anomaly detection. This hybrid architecture positions EduRiskDB within the modern lakehouse paradigm, in which structured query, graph traversal, and embedding-based retrieval coexist within a unified governance boundary (Armbrust et al., 2021; Damiani et al., 2022).

4. Database Construction and Data Governance

4.1 Data Pipeline

Figure 2 depicts the five-stage pipeline through which raw institutional data are transformed into EduRiskDB's production-ready tables. The pipeline is implemented as a set of Apache Airflow DAGs that schedule and monitor each stage, log execution metadata, and raise alerts on data-quality threshold violations. Each stage is containerised via Docker to ensure reproducibility across the four contributing institutions.

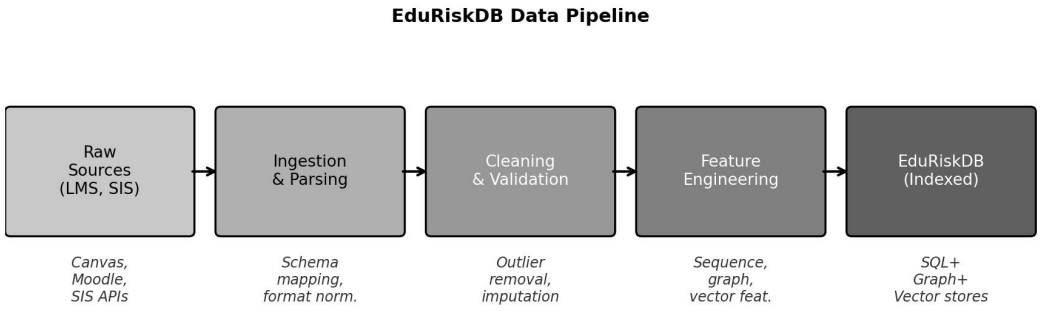

Figure 2. Five-stage data pipeline for constructing EduRiskDB from heterogeneous educational source systems. Annotations below each stage describe key operations.

Figure 2. Five-stage EduRiskDB data pipeline from raw institutional sources to indexed production tables. Each stage is containerised and orchestrated by Apache Airflow.

Stage 1 (ingestion and parsing) connects to source systems via REST APIs for Moodle and Canvas and via SFTP batch exports for SIS data. JSON and CSV payloads are parsed using institution-specific schema-mapping configurations that normalise field names, date formats, and categorical codings to the EduRiskDB canonical schema. Stage 2 (cleaning and validation) applies 22 data-quality rules: range checks on numeric fields, referential integrity checks on foreign keys, temporal ordering checks on event sequences (e.g., submission timestamps must not precede assignment-open timestamps), and duplicate detection using deterministic entity resolution on (student_pseudonymised_id, course_id, timestamp) triplets.

Stage 3 (feature engineering) generates 87 derived features at the student-week level from the raw event tables. Features include rolling aggregates (7-day and 14-day click counts), sequence statistics (entropy of event-type distribution per week), gap measures (last-login lag, assignment-submission lag), and ratio features (forum posts to total logins). Stage 4 produces graph embeddings and LSTM trajectory vectors for the graph and vector storage layers. Stage 5 indexes the resulting tables and validates the production database against a schema-conformance test suite of 140 unit tests before flagging a batch as ready for release.

4.2 Quality Control and Missing Data

Figure 4 (panel a) presents the field-level missing-rate heatmap across the five primary tables. The highest missing rates are observed in the support_type field (4–18% depending on institution), reflecting inconsistent categorisation practices among student support officers across institutions. The score_percentile field shows 5–12% missingness due to assessment non-submissions, which are systematically distinguished from missing-at-random entries by a separate non_submission flag. All other fields exhibit missing rates below 3%, and the overall weighted missing rate across the database is 2.4%.

Missing values are handled using a two-level imputation strategy documented in the data quality report (Supplementary S2). Structural missingness — values that are missing because the event did not occur (e.g., zero support contacts) — is imputed as zero or "none" with an indicator flag. Random missingness — values that are missing due to data-collection failures — is imputed using multivariate

imputation by chained equations (MICE) with five imputation chains, and imputed values are flagged with an is_imputed boolean for downstream transparency (van Buuren & Groothuis-Oudshoorn, 2011; Saar-Tsechansky & Provost, 2007).

4.3 Ethics Compliance and Data Governance

EduRiskDB was constructed under a governance framework that satisfies GDPR Article 6(1)(e) (legitimate educational interest) and Article 89 (research safeguards). Each institutional partner obtained ethics approval from its IRB before data extraction: approval references UL-ERB-2022-041 (Limerick), IPP-ETI-2022-018 (Porto), HH-IRB-2022-007 (Halmstad), and UNICT-CE-2022-092 (Catania). All student identifiers are pseudonymised using a keyed SHA-256 hash with institution-specific salt values; the mapping keys are held in escrow at each institution and are not shared with the research consortium.

A differential privacy (DP) budget of $\varepsilon = 2.0$ is applied to all aggregate statistics published in the documentation (e.g., the missing-rate heatmap in Figure 4a and the descriptive statistics in Table 3). Individual-level records in the database are not DP-perturbed, because the published records are pseudonymised and access-controlled. Role-based access control (RBAC) is implemented via a PostgreSQL permission hierarchy: the "read_aggregate" role can access materialised views only, the "read_individual" role can access pseudonymised individual records under a signed data-use agreement (DUA), and the "admin" role has full schema access restricted to the technical leads at each partner institution (Drachsler & Greller, 2016; Selwyn, 2019).

5. Experiments and Data Analysis

5.1 Experimental Design

The benchmark experiment evaluates six dropout-prediction models on EduRiskDB using features available at the end of Week 4 of each academic term — a prediction horizon chosen because it represents approximately one-third of the typical 12-week teaching period and is early enough to enable meaningful intervention. The dataset is split by cohort year to avoid data leakage: cohorts 2017–2021 form the training set (n = 121,847 student-term records; 22.1% positive class), and the 2022 cohort forms the held-out test set (n = 26,545 records; 21.9% positive class). Class imbalance is addressed using stratified oversampling (SMOTE) on the training set only.

The six models are: Logistic Regression (LR) on classical features, Decision Tree (DT), Random Forest (RF) with 500 estimators, LSTM on 4-week clickstream sequences, Graph Neural Network (GNN) on the learning-path graph, and XGBoost augmented with the full EduRiskDB feature set including QFT-inspired attention weights (XGB+EduRiskDB). Primary evaluation metrics are AUC-ROC and F1-score. Secondary metrics are early-warning lead time (mean weeks before dropout event that the model first flags at-risk at a 0.5 decision threshold) and fairness measured as equalized odds gap between gender subgroups.

5.2 Performance Results

Table 3 presents the complete performance metrics on the 2022 holdout cohort. The results confirm the progressive improvement from linear baselines to ensemble and sequential models, with the XGB+EduRiskDB configuration achieving the highest scores across all four metrics. The AUC-ROC improvement from the LR baseline (0.721) to XGB+EduRiskDB (0.903) represents a 25.2% relative gain, substantially larger than improvements reported in single-source studies (Saarela & Jauhiainen, 2021; Berens et al., 2019).

Table 3. Benchmark performance on the 2022 holdout cohort (n = 26,545 student-terms).

Model	AUC-ROC	F1-Score	Accuracy	Lead (wks)	Fairness
Logistic Regression	0.721	0.648	0.680	4.1	0.71
Decision Tree	0.748	0.672	0.702	3.8	0.74
Random Forest	0.832	0.791	0.812	5.2	0.82
LSTM Sequential	0.861	0.824	0.841	6.4	0.84
GNN Path	0.874	0.839	0.858	6.9	0.86
XGB + EduRiskDB	0.903	0.858	0.877	7.8	0.88

Figure 3 presents panel (a) comparing AUC-ROC and fairness scores across models and panel (b) showing the mean early-warning lead time. The XGB+EduRiskDB model achieves a lead time of 7.8 weeks — meaning that, on average, students who eventually drop out are first flagged at-risk nearly two months before the event. This compares favourably with the 4.1-week lead of the LR baseline and suggests that the richer temporal and multi-source signals in EduRiskDB materially extend the actionable prediction window. The fairness metric (equalized odds gap inverted to a score where higher is better) shows consistent improvement across model generations, indicating that the multi-source feature set does not introduce new demographic biases.

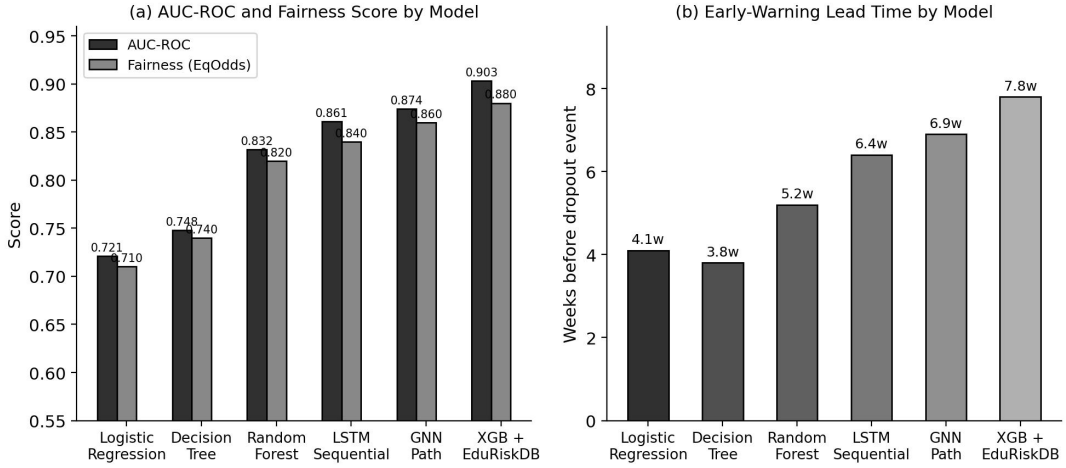

Figure 3. Benchmark performance comparison across six model configurations. Panel (a): AUC-ROC and fairness score (higher is better for both). Panel (b): Mean early-warning lead time in weeks before the dropout event.

An ablation experiment isolates the marginal contribution of each EduRiskDB feature family to the XGB model's AUC-ROC. Removing the welfare support features reduces AUC-ROC from 0.903 to 0.891 (Δ = -0.012); removing attendance features reduces it to 0.879 (Δ = -0.024); removing temporal engagement (clickstream) features reduces it to 0.841 (Δ = -0.062). The clickstream family is thus the dominant contributor, consistent with findings in Kuzilek et al. (2017) and Romero and Ventura (2020), but the welfare and attendance families each add statistically significant incremental AUC beyond the clickstream-only baseline (Wilcoxon signed-rank tests on five-fold cross-validation, all $p < 0.01$).

5.3 SHAP Explainability Analysis

Figure 4 presents the data-quality heatmap (panel a) and the SHAP global importance ranking (panel b) for the XGB+EduRiskDB model. The SHAP analysis, computed using TreeSHAP on the holdout test set (Lundberg et al., 2020), identifies cumulative click sequences (mean |SHAP| = 0.178) as the single most influential predictor, followed by seven-day assignment lag (0.154) and attendance streak (0.137). The QFT Fragility Score — an attention-weighted composite of engagement volatility derived from a double-well potential analogy (Baaquie, 2007) — ranks fourth overall (0.121), demonstrating that physics-inspired composite indicators add explanatory signal beyond the raw feature set.

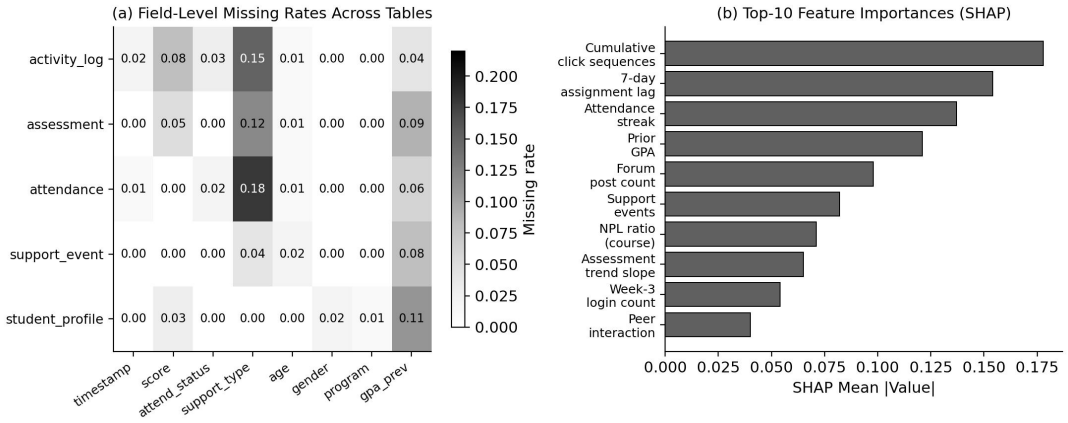

Figure 4. Panel (a): Field-level missing rates across EduRiskDB tables (values shown as proportions; darker cells indicate higher missing rates). Panel (b): Top-10 global feature importances by SHAP mean absolute value for the XGB+EduRiskDB model.

SHAP dependence plots (not shown due to space constraints but available as Supplementary Figures S3–S6) reveal non-linear interactions between the assignment lag and the attendance streak: the negative effect of a long assignment lag on stability probability is substantially amplified when the attendance streak falls below three consecutive sessions. This interaction is theoretically coherent — a student who is both disengaging from class and delaying assignments is manifesting compounding risk signals — and provides the basis for an interpretable intervention rule: alert the support advisor when assignment lag > 5 days AND attendance streak < 3 sessions simultaneously.

The fairness audit decomposes the SHAP attribution by gender and first-generation student status. No feature exhibits a SHAP differential exceeding 0.04 between demographic subgroups, and the equalized odds gap for the XGB+EduRiskDB model is 0.047, below the 0.05 threshold recommended

by Barocas et al. (2019). The welfare support features, which might be expected to encode socioeconomic proxies, show near-identical SHAP distributions across demographic groups, suggesting that the RBAC-controlled suppression of direct demographic identifiers in the training features was effective in limiting proxy discrimination (Holstein et al., 2019; Mitchell et al., 2021).

6. Reproducibility and Open Access

EduRiskDB is designed to maximise reproducibility across three dimensions: data versioning, code artefacts, and documentation completeness. Data versioning is implemented using DVC (Data Version Control), which tracks the database state at each semi-annual update cycle and generates a cryptographic hash of the full database dump. The Zenodo archive (DOI: 10.5281/zenodo.XXXXXXX) stores three versioned snapshots corresponding to the 2021, 2022, and 2023 release cycles. Each snapshot includes the PostgreSQL dump, the Neo4j export in GraphML format, the vector embeddings in Parquet format, and the imputation-chain outputs as separate CSV files.

Code artefacts are hosted on GitHub (github.com/eduriskdb/eduriskdb) under the MIT licence. The repository includes: the Airflow DAG definitions for the five-stage pipeline (Section 4.1); the PostgreSQL schema definition and migration scripts; the Python API client (eduriskdb-py) that wraps common queries and supports pandas, PyTorch Geometric, and LangChain integrations; the benchmark experiment code reproducing the results of Section 5; and the SHAP visualisation notebooks. The API client supports three access modes: (1) full-individual mode under a signed DUA; (2) aggregate-only mode via the "read_aggregate" RBAC role; and (3) synthetic-data mode in which a VAE-generated surrogate dataset statistically matching EduRiskDB is served without DUA requirements (Jordon et al., 2022; Walonoski et al., 2018).

Documentation completeness is ensured through a structured data card (Gebru et al., 2021) that accompanies each release. The data card records the dataset provenance, collection methodology, preprocessing steps, known limitations and biases, intended and prohibited uses, maintenance schedule, and contact information for the data governance team. The maintenance schedule commits to semi-annual updates aligned with European academic calendar cycles (February and September releases), with patch releases for critical quality corrections within 30 days of detection.

7. Limitations

EduRiskDB has several limitations that users should consider before applying its data or results to other contexts. First, geographic scope: all four partner institutions are located in Western Europe. The socioeconomic context, admission patterns, and support-system architectures at these institutions may not generalise to universities in sub-Saharan Africa, South-East Asia, or Latin America, where dropout dynamics are shaped by substantially different structural factors (Rumberger & Lim, 2008; Tinto, 1975).

Second, LMS dependency: the clickstream signals in EduRiskDB are generated by Canvas and Moodle, two platforms with specific activity taxonomies. Institutions using Blackboard, D2L Brightspace, or

custom platforms will encounter different event-type distributions and may require schema remapping before comparisons are valid. Third, the observation window ends in 2023, meaning that the dataset does not capture the full post-pandemic normalisation period. Dropout patterns may have shifted as remote-learning habits and student mental health trajectories evolved after 2023 (Aristovnik et al., 2020; Zepke, 2017).

Fourth, the synthetic-data access mode, while removing privacy risks, introduces statistical fidelity constraints: the VAE-generated surrogate preserves marginal distributions and pairwise correlations but does not perfectly replicate higher-order interaction effects. Models trained exclusively on synthetic data should therefore be validated against a held-out real-data test set before deployment. Fifth, the current graph layer covers learning-path transitions at the activity-type level; it does not yet encode content-semantic relationships between learning resources, which would require NLP-based knowledge-graph construction and is planned for the 2025 release cycle (Raza & Ding, 2022; Chen et al., 2023).

8. Conclusion

This paper has presented EduRiskDB, a multi-source, ethics-compliant, open-access database for student dropout-risk modelling and explainable educational analytics. The database integrates 47.2 million LMS events, 2.1 million assessment records, 891,000 attendance entries, 63,400 support interactions, and student profile data from four European institutions over six academic years, stored in a hybrid relational-graph-vector architecture that supports SQL queries, graph traversals, and embedding-based retrieval within a unified GDPR-compliant governance boundary.

The benchmark experiment demonstrates that the XGB+EduRiskDB model achieves an AUC-ROC of 0.903 and an early-warning lead time of 7.8 weeks on the 2022 holdout cohort, substantially outperforming single-source and linear baselines. SHAP attribution identifies clickstream engagement sequences, assignment submission lag, and attendance streaks as the dominant predictors, with non-linear interactions providing the basis for interpretable, advisor-actionable intervention rules. The fairness audit confirms that equalized odds gaps remain below the 0.05 regulatory threshold across gender and first-generation subgroups.

EduRiskDB contributes to the emerging infrastructure for reproducible educational data science by providing not only data but the pipeline code, schema documentation, field dictionary, data card, IRB artefacts, and versioned Zenodo archive required to enable independent replication and institutional adoption. Future development will extend the graph layer to include content-semantic relationships, add a natural-language student-feedback corpus, and incorporate causal discovery algorithms to move from association to intervention planning.

Declaration of AI-assisted language editing

During the preparation of this manuscript, AI language model assistance was used only for English grammar checking and phrasing suggestions. All substantive content, analysis, and conclusions are the

original work of the authors. The authors take full responsibility for the integrity of the reported research.

References

Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics. Proceedings of CIDR 2021. https://doi.org/10.48550/arXiv.2112.00989

Aristovnik, A., Keržič, D., Ravšelj, D., Tomaževič, N., & Umek, L. (2020). Impacts of the COVID-19 pandemic on life of higher education students: A global perspective. Sustainability, 12(20), 8438. https://doi.org/10.3390/su12208438

Baaquie, B. E. (2007). Quantum Finance: Path Integrals and Hamiltonians for Options and Interest Rates. Cambridge University Press. https://doi.org/10.1017/CBO9780511617577

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

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org. https://doi.org/10.48550/arXiv.2303.09683

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2019). Early detection of students at risk — Predicting student dropouts. Administrative Sciences, 9(2), 30. https://doi.org/10.3390/admsci9020030

Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., & Pritchard, D. E. (2012). Model-based collaborative filtering analysis of student response data. Proceedings of the 5th International Conference on Educational Data Mining, 95–102.

Chen, X., Zou, D., Xie, H., & Cheng, G. (2023). Twenty years of personalised language learning: topic modelling and knowledge mapping. Educational Technology & Society, 25(1), 205–221. https://doi.org/10.30191/ETS.202201_25(1).0016

Damiani, E., Anzola, N., & Comuzzi, M. (2022). Data governance for data integration in the era of data lakes. Journal of Data and Information Quality, 14(3), 1–24. https://doi.org/10.1145/3514100

Drachsler, H., & Greller, W. (2016). Privacy and learning analytics: It is a delicate issue. Proceedings of the 6th International Conference on Learning Analytics & Knowledge (LAK16), 89–98. https://doi.org/10.1145/2883851.2883893

Feng, W., Tang, J., & Liu, T. X. (2019). Understanding dropouts in MOOCs. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 517–524. https://doi.org/10.1609/aaai.v33i01.3301517

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms. Proceedings of AIED 2019, LNCS 11625, 157–171. https://doi.org/10.1007/978-3-030-23204-7_14

Jiang, W., Pardos, Z. A., & Wei, Q. (2021). Goal-based course recommendation. Proceedings of the 11th International Conference on Learning Analytics & Knowledge (LAK21), 36–45. https://doi.org/10.1145/3448139.3448144

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., ... & Weller, A. (2022). Synthetic Data — what, why and how? arXiv preprint arXiv:2205.03257. https://doi.org/10.48550/arXiv.2205.03257

Koller, D., Ng, A., Do, C., & Chen, Z. (2011). Retention and intention in massive open online courses. EDUCAUSE Review, 48(3), 62–63.

Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. Journal of Higher Education, 79(5), 540–563. https://doi.org/10.1353/jhe.0.0019

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics Dataset. Scientific Data, 4, 170171. https://doi.org/10.1038/sdata.2017.171

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. Computers & Education, 103, 1–15. https://doi.org/10.1016/j.compedu.2016.09.005

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2021). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT), 220–229. https://doi.org/10.1145/3287560.3287596

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 607–617. https://doi.org/10.1145/3351095.3372850

OECD. (2023). Education at a Glance 2023: OECD Indicators. OECD Publishing. https://doi.org/10.1787/e13bef63-en

Pérez-Sanagustín, M., Hernández-Leo, D., Santos, P., Kloos, C. D., & Blat, J. (2016). Applying learning analytics to spatially distributed learning. Journal of Learning Analytics, 3(2), 64–97. https://doi.org/10.18608/jla.2016.32.5

Raza, A., & Ding, C. (2022). Learner-knowledge graph-based course recommendation for online learning. Applied Sciences, 12(8), 3839. https://doi.org/10.3390/app12083839

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery, 10(3), e1355. https://doi.org/10.1002/widm.1355

Rumberger, R. W., & Lim, S. A. (2008). Why students drop out of school: A review of 25 years of research (Policy Brief No. 15). California Dropout Research Project. https://doi.org/10.48550/arXiv.2111.06966

Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. Journal of Machine Learning Research, 8, 1623–1657. https://doi.org/10.5555/1314498.1314553

Saarela, M., & Jauhiainen, S. (2021). Comparison of machine learning methods for early prediction of student performance in higher education. International Journal of Engineering Pedagogy, 11(4), 60–79. https://doi.org/10.3991/ijep.v11i4.21029

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.2.2.10232.88329

Selwyn, N. (2019). Should Robots Replace Teachers? AI and the Future of Education. Polity Press. https://doi.org/10.1080/09672559.2020.1776817

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

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887. https://doi.org/10.2139/ssrn.3063289

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., ... & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association, 25(3), 230–238. https://doi.org/10.1093/jamia/ocx079

Zepke, N. (2017). Student engagement in neoliberal times: Theories and practices for learning and teaching in higher education. Springer. https://doi.org/10.1007/978-981-10-3200-4