**International Journal of Information Technology and Computer Science Applications (IJITCSA)**
**p-ISSN: 2964-3139 e-ISSN: 2985-5330**
Vol. 04, No. 01, page 47 - 58
Submitted 18/2/2026; Accepted 1/4/2026; Published 13/4/2026

# A Lakehouse-Oriented Big Data Infrastructure for Educational Analytics: Integrating Administrative and Assessment Data for Early Student Risk Prediction

**Bhairav Kaphle[1]* and Biswajit Shrestha[1]**
[1]Digital Technology, Madan Bhandari University of Science and Technology (MBUST), NEPAL
e-mail: bhairavphle@gmail.com

**Corresponding Autor:** Bhairav Kaphle



## Abstract
Educational institutions increasingly depend on heterogeneous digital systems, yet many analytics initiatives remain fragmented across student information, registration, assessment, and learning platforms. This paper proposes a lakehouse-oriented big data infrastructure for educational analytics and validates it through a reproducible early-risk prediction study using the Open University Learning Analytics Dataset (OULAD). The study integrates five public OULAD tables student information, course registration, assessment metadata, student assessment submissions, and course presentation metadata into temporally valid feature tables aligned to the student–module–presentation level. We define a windowed feature engineering framework that constructs actionable indicators such as submission rate, weighted completion score, average submission lag, and assessment coverage gap at 30%, 50%, 70%, and 100% of the course timeline. Two supervised classifiers, logistic regression and random forest, are evaluated under a stratified 80/20 protocol. The results show that administrative data alone provides weak discrimination (AUC ≈ 0.673), whereas integrated mid-course assessment evidence substantially improves performance. At the 50% course window, the random-forest model achieves an AUC of 0.947, F1 of 0.879, and recall of 0.829; even at the 30% window the model already reaches an AUC of 0.904. These findings demonstrate that the value of educational prediction depends not only on model choice but also on data integration architecture. The paper contributes (i) a lakehouse-oriented reference architecture for higher-education analytics, (ii) a temporally constrained feature engineering strategy for early-warning systems, and (iii) an empirical ablation showing that multi-source integration yields large and operationally meaningful gains.

**Keywords**—Educational data integration, Learning analytics, Student risk prediction, Lakehouse architecture, Higher education, Reproducible analytics.

## 1 Introduction

Universities now operate a dense ecosystem of digital systems that continuously generate heterogeneous data namely student information systems, registration platforms, assessment platforms, learning management systems, advising portals, and increasingly, analytics dashboards. This expansion has made educational decision-making simultaneously more data-rich and more operationally complex. Learning analytics has long argued that such data can support timely and evidence-based interventions, especially for identifying students who are likely to fail or withdraw before the end of a course [1–5]. Yet many institutions still struggle to move from isolated reporting toward integrated, production-grade data infrastructures that combine governance, scalability, and analytical usefulness [12, 15, 17–18].

The literature motivated this study correctly recognized that educational applications require the storage, integration, and analysis of large and diverse institutional data. However, it remained largely conceptual, offered no operational data model, and did not validate its claims with a reproducible experiment. The present manuscript reconstructs that idea into a stronger research contribution by linking two strands of literature that are often discussed separately. The first is the literature on educational data mining and learning analytics, which focuses on predictive models, risk indicators, student success, and ethical intervention design [6–9]. The second is the literature on modern data platforms, which focuses on ingestion, governance, metadata, scalable storage, and the convergence of data lakes and warehouses into lakehouse-style architectures [16–18].

Our central argument is that educational prediction quality is not only a modeling question. It is also an infrastructure question. If student data remain fragmented across operational silos, then important signals become unavailable, delayed, inconsistent, or difficult to govern. In contrast, an integrated architecture that aligns administrative records with temporally available assessment evidence can materially improve predictive performance while preserving auditability and deployment realism. This paper therefore asks the following research question: *"How can a lakehouse-oriented educational data infrastructure integrate heterogeneous academic data sources in a temporally valid manner to support early student risk prediction?"*

To answer this question, we use the Open University Learning Analytics Dataset (OULAD), a widely used public benchmark that contains 32,593 student-course records, 22 module presentations, and linked tables for demographics, registration, assessments, and activity traces [10–11]. Rather than treating the dataset as a flat file, we explicitly model it as a multi-source educational data platform. We construct integrated feature tables over successive temporal windows and evaluate whether predictive performance improves when administrative data are augmented with progressively richer assessment evidence. This design allows the paper to speak to both systems and analytics audiences. It demonstrates how an educational data platform can be structured and also quantifies the gain from integration.

The contribution of this study is threefold. First, we propose a lakehouse-oriented reference architecture for educational analytics that organizes heterogeneous institutional data into Bronze, Silver, and Gold layers while incorporating governance, privacy, and monitoring requirements. Second, we define a temporally constrained feature engineering framework that transforms administrative and assessment data into early-warning indicators such as submission rate, weighted completion score, lateness, and assessment coverage gap. Third, we validate the approach empirically on OULAD and show that integrated mid-course evidence substantially outperforms administrative-only baselines. In our experiment, an administrative-only baseline yields an area under the ROC curve (AUC) of only 0.673, whereas the integrated random-forest model reaches an AUC of 0.947 by the 50% course window and 0.904 already at the 30% window.

This paper is structured as follows. Section 2 reviews the state of the art on educational analytics, interoperability, and modern data platforms. Section 3 describes the dataset, the proposed architecture, and the temporally valid feature engineering procedure. Section 4 details the experimental setup. Section 5 reports the results and discusses their practical implications. The final section concludes with limitations and future work.

## 2 Related Work

### *2.1 Educational data mining and learning analytics*

Educational data mining (EDM) and learning analytics (LA) emerged from a common interest in understanding and improving learning processes through data, although they emphasize somewhat different traditions and objectives. Siemens and Long [1] framed learning analytics as a response to the increasing digitization of education and the corresponding availability of trace data. Ferguson [2] further highlighted the drivers and challenges of LA, including prediction, personalization, and the need for responsible data use. Greller and Drachsler [3] proposed one of the early conceptual frameworks for LA, identifying stakeholders, objectives, data, instruments, external constraints, and internal limitations. More recently, Romero and Ventura [5] emphasized the convergence of EDM and LA in practical applications while also noting the persistent gap between methodological development and institutional deployment.

A major strand of this literature addresses academic success, course completion, and dropout prediction. Alyahyan and Düştegör [6] reviewed predictive studies in higher education and concluded that while many models achieve promising accuracy, the real challenge lies in early prediction, data quality, and intervention relevance. Cantabella et al. [7] showed that student behavior in learning management systems can provide highly informative features for performance analysis. Vaarma et al. [8] demonstrated that multi-source data from transcripts, demographics, and LMS systems can support dropout prediction in university settings. Rabelo et al. [9] similarly

reported strong performance from ensemble models for identifying higher-education dropouts. Across these studies, one pattern is clear: predictive success improves when behavioral and administrative signals are combined rather than analyzed in isolation.

However, the literature also shows that most predictive studies remain workflow-centric rather than infrastructure-centric. Many papers describe preprocessing, feature engineering, and model fitting, but fewer explain how such processes should be operationalized within institutional data platforms. Samuelsen et al. [12] reviewed studies that integrate multiple data sources for learning analytics and found that the LMS is the most common source, yet genuine multisource integration remains limited. This is a critical gap because educational institutions rarely operate from a single canonical data source. Student records, course events, assessment histories, and digital interaction traces are usually stored in different operational systems with different temporal granularities and governance controls.

### *2.2 Interoperability and semantic integration in educational systems*

Interoperability is essential if educational analytics is to scale beyond single-system reporting. Dodero et al. [13] discussed the trade-off between interoperability and data collection performance in web-based learning environments, showing that analytics quality depends on how effectively distributed systems can exchange semantically meaningful events. Masud et al. [14] proposed semantic-data approaches for collaborative e-learning environments, emphasizing interoperability and distributed metadata management. More recently, Paneque et al. [15] introduced the e-LION semantic model to consolidate multiple e-learning knowledge bases, thereby enriching downstream analysis. Taken together, these studies suggest that the problem is not only the volume of educational data, but also the alignment of schemas, entities, events, and temporal meaning.

This issue is particularly relevant in higher education because a student is represented differently across systems. In one database the student is a registration record; in another, an assessment submission; in another, a behavioral trace or advising event. Without integration, institutional analytics often become either incomplete or misleading. A realistic educational data platform must therefore support relational linking, temporal harmonization, metadata management, and quality validation.

### *2.3 Modern data platforms: warehouse, lake, and lakehouse*

From the data engineering perspective, educational analytics increasingly resembles other enterprise analytics domains that must reconcile heterogeneous ingestion pipelines with governed analytical outputs. Traditional data warehouses remain effective for curated, schema-controlled, business-ready marts, but they can be restrictive for exploratory and semi-structured data. Data lakes improve flexibility but often create governance and consistency challenges. The lakehouse paradigm has emerged as an attempt to combine the strengths of both low-cost raw data storage plus transactional consistency, schema evolution, and reliable analytical serving [16–18].

Schneider et al. [17] argue that lakehouses should be understood as a distinct architectural approach rather than merely a marketing label, because they address concrete requirements around data management, openness, and analytical reuse. Harby et al. [18] further compare data lake, warehouse, and lakehouse approaches and show experimentally that lakehouse systems can improve analytical responsiveness while retaining flexibility. Although this literature rarely focuses on education specifically, its relevance is immediate. Educational institutions require exactly the same qualities: reproducible ingestion, multi-table integration, quality controls, and the ability to serve both operational dashboards and research-oriented analyses.

### *2.4 Ethics, privacy, and deployment realism*

Predictive analytics in education cannot be separated from ethical and governance concerns. Slade and Prinsloo [19] argued that learning analytics requires principled attention to transparency, consent, and institutional responsibility. Prinsloo and Slade [20] further noted that analytics systems shape how institutions see students and allocate interventions; therefore, technical choices have normative consequences. Ifenthaler and Schumacher [21] showed that students themselves have nuanced perspectives on privacy and the use of educational data. These concerns imply that a robust educational data platform should embed governance mechanisms rather than treat them as post hoc compliance issues.

Literature therefore motivates an integrated research need. Educational prediction is well studied. Interoperability is recognized as important. Lakehouse platforms are gaining maturity. Yet there remains limited work that explicitly connects these themes into a reproducible design-and-evaluation study for higher education. This paper addresses that gap by proposing an educational lakehouse reference architecture and validating it with time-aware multisource prediction.

# 3 Dataset, Problem Formulation, and Proposed Architecture

## *3.1 Dataset and analytical objective*

The empirical study uses OULAD, a public higher-education dataset introduced by Kuzilek et al. [10]. OULAD is particularly suitable for this work because it was explicitly designed as a multi-table learning analytics resource rather than a single flat benchmark. In its full form, it links student demographics, registration histories, assessment outcomes, and virtual learning environment interactions [10–11]. For the present experiment we use five raw tables that are sufficient for a reproducible early-warning pipeline: *studentInfo.csv*, *studentRegistration.csv*, *tudentAssessment.csv*, *assessments.csv*, and *courses.csv*. This configuration allows us to study multisource integration based on administrative and assessment data while keeping the experimental package lightweight and directly reproducible.

Table 1 summarizes the raw input used in the study. The final modeling unit is the student–module–presentation tuple, which corresponds to one student enrolled in one course presentation. The dataset contains 32,593 such records, spanning 28,785 unique students, 7 modules, and 22 module presentations. The outcome distribution is non-trivial: 31.16% of records are Withdrawn and 21.64% are Fail, yielding an at-risk rate of 52.80% when Fail and Withdrawn are combined.

Table 1. Raw tables used in the experiment.

| Table | Size | Semantics |
|---|---|---|
| studentInfo | 32,593 rows | Demographics, prior attempts, studied credits, disability status, and final result |
| studentRegistration | 32,548 rows | Registration timing for each student–module–presentation tuple |
| studentAssessment | 173,912 rows | Student-level submissions, scores, and banked status for assessments |
| assessments | 206 rows | Assessment metadata, types, due dates, and weights |
| courses | 22 rows | Module presentation lengths used to normalize temporal windows |

The analytical objective is early risk prediction. We define the binary outcome as

$$y_s = \mathbb{I}\{\text{final\_result}_s \in \{\text{Fail,Withdrawn}\}\} \quad (1)$$

Where s indexes a student–course record. This definition is operationally meaningful because both failure and withdrawal typically trigger institutional concern, whereas Pass and Distinction correspond to non-risk outcomes. The resulting task is to estimate the probability that a currently enrolled student will belong to the at-risk class before the course is completed.

## *3.2 Lakehouse-oriented reference architecture*

Figure 1 presents the proposed architecture. The architecture is designed to address three common weaknesses of educational analytics deployments: fragmented source systems, ad hoc feature construction, and poor traceability between raw events and analytical outputs. The proposed platform follows a lakehouse pattern with layered storage and explicit governance.

In the Educational Data Sources, raw data is ingested from operational systems without aggressive transformation. This layer preserves source fidelity and supports auditability. For educational contexts, in the Lakehouse Layers, Bronze data may include student information system extracts, registration records, assessment logs, learning management system events, advising records, and optional sensor or engagement data where institutionally appropriate. Schema checks, PII masking, and event-level quality rules are applied at ingestion time.

The Silver layer performs entity resolution and relational harmonization. At this stage, student, module, presentation, assessment, and timestamp fields are standardized, and cross-table relationships are made analytically usable. In our OULAD implementation, this layer resolves the student–module–presentation unit and aligns assessment schedules with student submissions and registration timing. Silver tables thus become the trusted integrated core from which features can be built.

The Gold layer exposes analytical marts and feature tables for dashboards, model training, and intervention workflows. A key design principle is temporal validity: Gold features must be generated only from evidence available up to the prediction window. This prevents leakage and makes offline evaluation a more faithful approximation of deployment conditions.

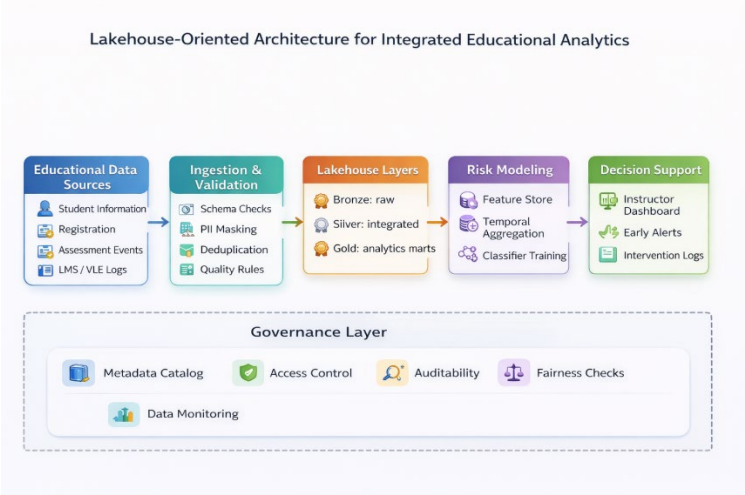


Figure 1: Lakehouse-oriented architecture for integrated educational analytics.

### *3.3 Temporal feature engineering*

Let $s = (i, m, p)$ denote a student $i$ enrolled in module $m$ during presentation $p$. Our integrated feature vector at prediction window $\tau$ is defined as

$$\mathbf{x}_s^{(\tau)} = \phi\left(D_s^{\text{admin}}, D_s^{\text{reg}}, D_s^{\text{assess},\tau}\right) \tag{2}$$

where $D_s^{\text{admin}}$ contains demographic and academic background information, $D_s^{\text{reg}}$ contains registration timing and course length, and $D_s^{\text{assess},\tau}$ contains assessment evidence whose due dates fall within the observed proportion $\tau$ of the course timeline. The temporal windows used in the experiment are $\tau \in \{0.3, 0.5, 0.7, 1.0\}$, corresponding to 30%, 50%, 70%, and 100% of the assessment timeline relative to module presentation length. This windowing strategy is important because early-warning systems are only useful if they can act before the end of the course.

For each student-course tuple, we construct the following integrated indicators

$$SR_s^{(\tau)} = \frac{1}{N_s^{(\tau)}} \sum_{j=1}^{N_s^{(\tau)}} \mathbb{I}\{q_{sj} \text{ observed}\} \tag{3}$$

Where $SR_s^{(\tau)}$ is the submission rate and $N_s^{(\tau)}$ is the number of scheduled non-exam assessments observed by the window.

$$WCS_s^{(\tau)} = \frac{\sum_{j=1}^{N_s^{(\tau)}} w_j \, \tilde{q}_{sj}}{\sum_{j=1}^{N_s^{(\tau)}} w_j}, \qquad \tilde{q}_{sj} = \begin{cases} q_{sj}, & \text{if submitted} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Where $WCS_s^{(\tau)}$ is the weighted completion score, $w_j$ is the assessment weight, and $q_{sj}$ is the student score. This formulation penalizes missing submissions, making it more informative than the mean score over submitted assessments alone.

$$Lag_s^{(\tau)} = \frac{1}{M_s^{(\tau)}} \sum_{j=1}^{N_s^{(\tau)}} \mathbb{I}\{q_{sj} \text{ observed}\}\left(d_{sj}^{\text{submit}} - d_j^{\text{due}}\right) \tag{5}$$

Where $Lag_s^{(\tau)}$ captures average submission delay and $M_s^{(\tau)}$ is the number of observed submissions. We also derive the late ratio, banked ratio, assessment coverage gap, number of expected assessments, and number of submitted assessments.

Administrative features include gender, region, highest education, IMD band, age band, disability, number of previous attempts, studied credits, registration lead time, module identity, presentation identity, and module length. Table 2 groups the final variables used in modeling.

Table 2. Feature groups used in the experiment.

| Feature group | Variables |
|---|---|
| Administrative | Gender, region, highest education, IMD band, age band, disability, number of previous attempts, studied credits |
| Registration | Registration lead time, module presentation length, module code, presentation code |
| Assessment behavior | Expected assessments, submitted assessments, submission rate, mean score, weighted completion score, mean lag, late ratio, banked ratio, assessment coverage gap |

### *3.4 Predictive models and training objective*

Two supervised classifiers are evaluated: logistic regression and random forest. Logistic regression provides an interpretable linear baseline, while random forest offers a stronger nonlinear learner that can capture interactions and threshold effects without strict distributional assumptions [22, 25–26]. For a probabilistic classifier $f_\theta$, prediction can be written as

$$\hat{p}_s^{(\tau)} = f_\theta\left(\mathbf{x}_s^{(\tau)}\right) \tag{6}$$

With the logistic regression model minimizing the binary cross-entropy objective

$$\mathcal{L}(\theta) = -\frac{1}{n}\sum_{s=1}^{n}[y_s \log \hat{p}_s + (1 - y_s)\log(1 - \hat{p}_s)] \tag{7}$$

All preprocessing steps are embedded in the modeling pipeline to avoid train–test contamination. Numeric features are imputed with medians and standardized; categorical features are imputed with the mode and one-hot encoded. The random forest uses class-balanced subsampling to mitigate skews in the class distribution.

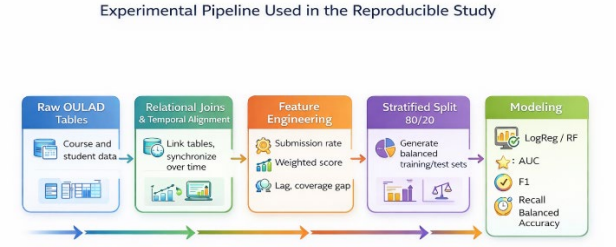


Figure 2. Experimental pipeline used in the reproducible study.

### *3.5 Algorithmic workflow*

Algorithm 1 summarizes the full workflow from raw tables to evaluated models. The key design choice is that assessment features are generated only from assessments whose scheduled dates are observable by the chosen window. This constraint distinguishes deployment-realistic early-warning modeling from retrospective full-information classification.

***Algorithm 1. Time-aware educational data integration and prediction***

1. **Input:** Raw tables $T_1 =$ studentInfo, $T_2 =$ studentRegistration, $T_3 =$ studentAssessment, $T_4 =$ assessments, $T_5 =$ courses
2. **for** each student–module–presentation tuple $s$ **do**:
   a. Build base administrative record from $T_1$, $T_2$, and $T_5$
   b. **for** each temporal window $\tau \in \{0.3, 0.5, 0.7, 1.0\}$ **do**
      i. Select scheduled assessments from $T_4$ with due date $\leq \tau \times$ module length
      ii. Left-join observed submissions from $T_3$
      iii. Compute $SR_s^{(\tau)}$, $WCS_s^{(\tau)}$, $Lag_s^{(\tau)}$, late ratio, and coverage gap
      iv. Concatenate administrative and temporal features into $\mathbf{x}_s^{(\tau)}$
3. **end for**
4. **end for**
5. Create stratified 80/20 train-test split
6. Fit logistic regression and random forest on each feature table
7. Evaluate AUC, F1, precision, recall, balanced accuracy, and accuracy
8. Export Gold-layer feature tables and metrics for dashboards and intervention services

## 4 Experimental Setup

### *4.1 Design rationale*

The experiment is structured as an ablation study on data integration. The baseline condition uses only administrative and registration variables. The integrated conditions then add temporally available assessment features for 30%, 50%, 70%, and 100% of the course timeline. This setup tests whether performance gains arise from multisource integration rather than simply from selecting a stronger classifier.

The train–test protocol is a stratified 80/20 split with a fixed random seed. Metrics are reported on the held-out test set. Because the task is operational rather than purely statistical, we report multiple measures: AUC for ranking quality [23], F1 as a balance of precision and recall, balanced accuracy to reduce sensitivity to class skew, and conventional accuracy for completeness. Precision and recall are reported because educational interventions often face a trade-off between missing vulnerable students and over-alerting advisors [24].

### *4.2 Implementation details*

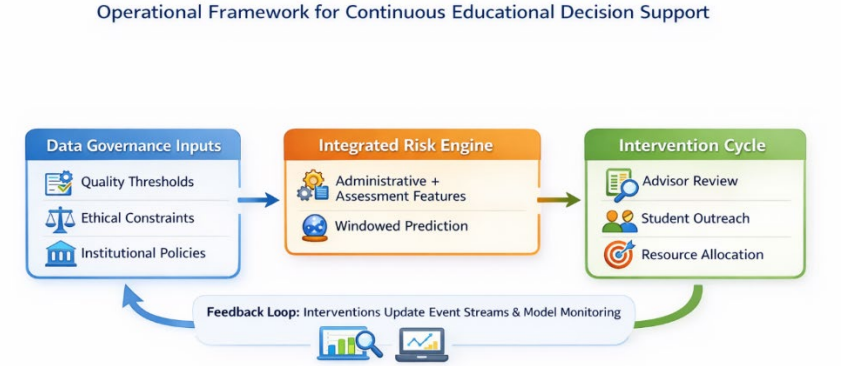


Figure 3. Operational framework connecting governance, integrated risk scoring, and intervention cycles.

The empirical pipeline is implemented in Python using pandas and scikit-learn. The full notebook and processed feature tables are included in the accompanying artifact package. The random forest uses 120 trees, maximum depth 18, minimum leaf size 3, and class-balanced subsampling. Logistic regression uses balanced class weights and a standardized feature pipeline. Missing values are handled through median imputation for numeric variables and most-frequent imputation for categorical variables.

The study intentionally focuses on five OULAD tables. Although OULAD also includes VLE interaction data, the present design aims to demonstrate that substantial predictive gains are already attainable from the integration of administrative, registration, and assessment streams. This choice is relevant for institutions that do not yet have mature LMS event pipelines but already possess reliable registration and assessment records.

## 5 Results and Discussion

### *5.1 Descriptive overview*

Figure 4 shows the class distribution. Pass is the largest category, followed by Withdrawn, Fail, and Distinction. The resulting at-risk class is therefore substantial rather than rare, which makes the task realistic for institutional intervention settings. Module-level variation is also non-negligible, suggesting that model features should preserve course context rather than flatten all presentations into a generic student record.

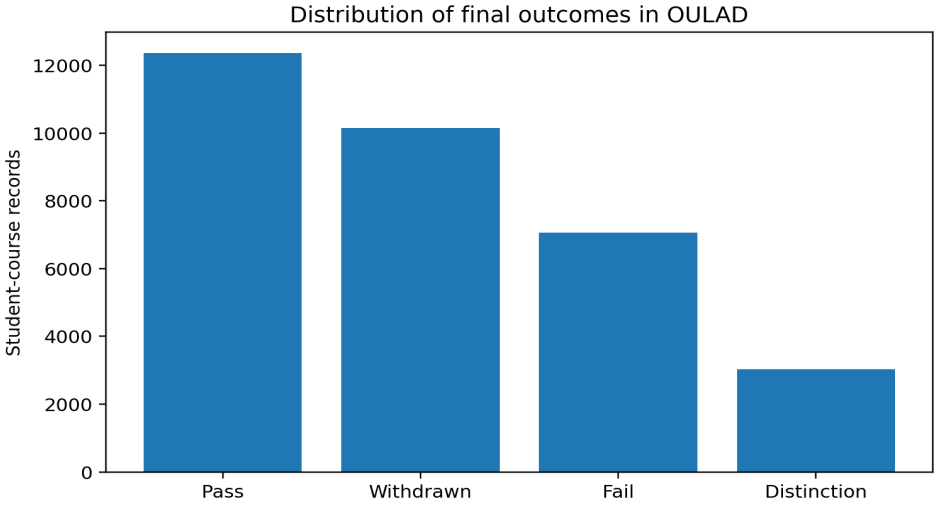


Figure 4. Distribution of final outcomes in the OULAD student-course records.

### *5.2 Predictive performance across temporal windows*

Table 3 reports the main results. Three patterns are especially important. First, the administrative-only baseline is weak for both models, with AUC values near 0.67. This indicates that demographic and registration information alone are insufficient for reliable early identification. Second, even a 30% window of integrated assessment evidence sharply improves performance: the random forest reaches an AUC of 0.904. Third, the 50% window already delivers highly actionable performance, with the random forest attaining an AUC of 0.947, F1 of 0.879, and recall of 0.829. This is a particularly important result because a model that only becomes accurate near course completion has limited practical value.

Table 3. Test-set performance by model and observed proportion of the assessment timeline

| Window | Model | AUC | F1 | Precision | Recall | Balanced Acc. | Accuracy |
|---|---|---|---|---|---|---|---|
| Baseline | Logistic Regression | 0.672 | 0.634 | 0.658 | 0.612 | 0.628 | 0.627 |
| Baseline | Random Forest | 0.673 | 0.637 | 0.648 | 0.626 | 0.623 | 0.623 |
| 30% | Logistic Regression | 0.900 | 0.817 | 0.894 | 0.752 | 0.827 | 0.822 |
| 30% | Random Forest | 0.904 | 0.824 | 0.903 | 0.757 | 0.833 | 0.829 |
| 50% | Logistic Regression | 0.939 | 0.870 | 0.924 | 0.822 | 0.873 | 0.870 |
| 50% | Random Forest | 0.947 | 0.879 | 0.936 | 0.829 | 0.883 | 0.880 |

| Window | Model | AUC | F1 | Precision | Recall | Balanced Acc. | Accuracy |
|---|---|---|---|---|---|---|---|
| 70% | Logistic Regression | 0.961 | 0.901 | 0.940 | 0.865 | 0.901 | 0.899 |
| 70% | Random Forest | 0.968 | 0.913 | 0.953 | 0.875 | 0.914 | 0.912 |
| 100% | Logistic Regression | 0.972 | 0.922 | 0.958 | 0.888 | 0.922 | 0.920 |
| 100% | Random Forest | 0.980 | 0.935 | 0.970 | 0.902 | 0.935 | 0.933 |

Figure 5 visualizes the AUC trend. The shape of the curve is informative in itself: performance increases monotonically with additional integrated evidence, but the steepest gain occurs between the baseline and the first two windows. In other words, the data integration benefit is not marginal. It is transformational. By mid-course, the model has already crossed into a performance range that is plausible for decision support.

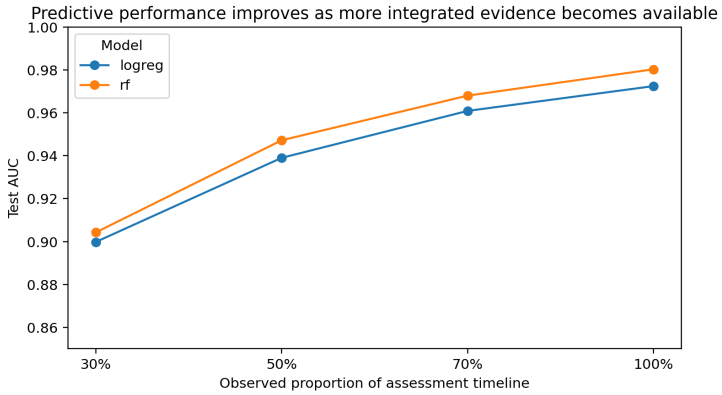


Figure 5. Predictive performance improves as more integrated evidence becomes available.

### *5.3 Why integration matters*

The improvement from baseline to integrated models demonstrates the substantive value of data integration. Administrative variables such as prior attempts, studied credits, or neighborhood deprivation can describe structural background conditions, but they do not directly capture how a student is participating in the current course. Once temporally aligned assessment data are introduced, the model gains access to behavioral signals that are more proximal to eventual failure or withdrawal. These signals include whether students submit scheduled work, whether they submit on time, and how much weighted course progress they actually complete.

This finding aligns with the literature showing that richer behavioral data improve educational prediction [6–8]. However, our results extend that literature by showing that the gain can be framed explicitly as an integration effect. The baseline and integrated models use the same prediction task, the same split, and largely the same modeling machinery. What changes is the architecture of the feature space. This makes the case for investing in better educational data infrastructure, not merely better algorithms.

### *5.4 Feature importance and operational interpretation*

Figure 6 reports the top random-forest feature importances at the 50% window. Submission rate, weighted completion score, assessment coverage gap, number of submitted assessments, and mean score dominate the ranking. This pattern is conceptually coherent. Students who fail or withdraw are not only those who score lower; they are often those who disengage from the assessment process itself. The model therefore responds to a mixture of achievement and participation signals.

From an intervention perspective, this is useful because the important variables are actionable. An advisor can respond to missing or late submissions much earlier than to final grade outcomes. Likewise, a course leader can inspect whether a specific module produces unusually large coverage gaps or late-submission patterns. The framework thus supports not only student-level prediction but also module-level quality assurance.

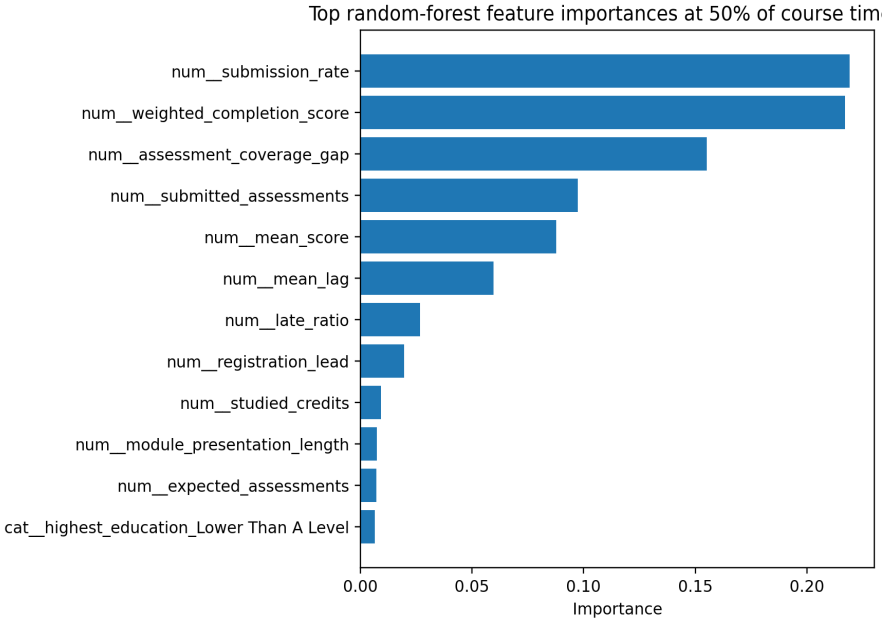


Figure 6. Top random-forest feature importances at the 50% window.

### *5.5 Model comparison*

The random forest consistently outperforms logistic regression at every integrated window, although the margin is moderate rather than dramatic. This suggests that nonlinearities and interactions exist in the data, but a well-constructed linear model remains competitive. For institutions with stricter interpretability requirements, logistic regression may therefore be a viable deployment option, particularly at the 50% and 70% windows. For institutions prioritizing maximal ranking quality, the random forest is preferable.

Figure 7 shows ROC curves for the random-forest model across the four windows. The widening separation from the diagonal baseline confirms that additional integrated evidence yields progressively better discrimination. Yet the 30% curve is already strong, which indicates that institutions do not need to wait until the end of a semester to obtain useful alerts.

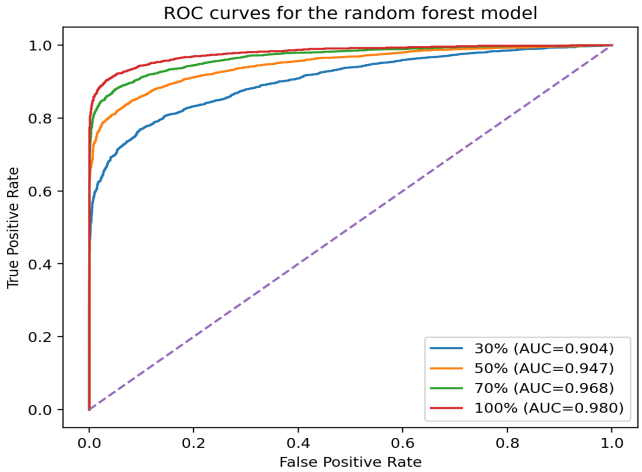


Figure 7. ROC curves for the random-forest model across timeline windows.

### *5.6 Implications for educational big data infrastructure*

The results support three practical implications. First, educational analytics platforms should prioritize reliable integration of registration and assessment systems before pursuing more complex sources. Our experiment shows that these sources alone already deliver large gains when properly aligned. Second, temporal validity should be enforced as a first-class design principle. Feature stores that leak end-of-course information into early-warning

models may look impressive offline but fail in real use. Third, Gold-layer analytics marts should expose both predictive outputs and human-readable indicators such as submission rate and coverage gap, because intervention workflows require operational interpretability.

The proposed lakehouse design also improves reproducibility. By separating Bronze, Silver, and Gold concerns, institutions can trace predictions back to raw records, update features incrementally, and version analytical outputs. This is especially important in educational settings where models may influence support allocation, escalation procedures, or policy decisions.

### *5.7 Threats to validity and limitations*

Several limitations should be acknowledged. First, the experiment uses public OULAD data, which improves reproducibility but may not capture the full diversity of institutional contexts. Second, the current empirical package does not include VLE clickstream features, even though the architecture is designed to support them. This means the paper demonstrates strong multisource integration, but not the maximum possible breadth of educational telemetry. Third, we evaluate a fixed train–test split rather than multi-institution external validation. Fourth, the study focuses on predictive quality and feature interpretability, not causal intervention impact.

These limitations do not negate the main finding. They clarify its scope. The contribution is a reproducible demonstration that integrating administrative and assessment data within a temporally valid infrastructure meaningfully improves early student risk prediction. Future work should extend the framework with VLE traces, advising notes, streaming event data, fairness auditing, and prospective deployment studies.

## 6 Conclusion

This paper transformed a conceptual discussion of big data infrastructures in education into a reproducible design-and-evaluation study. We proposed a lakehouse-oriented reference architecture for educational analytics, formalized a temporally valid feature engineering framework, and validated the approach on a public higher-education benchmark. The empirical results show that administrative data alone are inadequate for accurate early-warning prediction, whereas the integration of assessment evidence yields large and operationally significant gains. In particular, the random-forest model improved from an AUC of 0.673 in the administrative-only baseline to 0.947 by the mid-course window.

The main message is straightforward: effective educational analytics depends on infrastructure as much as on algorithms. Institutions that want actionable early-warning systems need governed integration pipelines, not just isolated models. By linking educational data mining with modern lakehouse design, the paper offers a framework that is both analytically effective and operationally plausible. This makes it suitable as a foundation for future work on production-grade, ethically governed, and intervention-aware learning analytics systems.

BIBLIOGRAPHY

[1]. G. Siemens and P. Long, “Penetrating the fog: Analytics in learning and education,” EDUCAUSE Review, vol. 46, no. 5, pp. 30–40, 2011.
[2]. R. Ferguson, “Learning analytics: drivers, developments and challenges,” International Journal of Technology Enhanced Learning, vol. 4, no. 5–6, pp. 304–317, 2012.
[3]. W. Greller and H. Drachsler, “Translating learning into numbers: A generic framework for learning analytics,” in Proc. 2nd International Conference on Learning Analytics and Knowledge, 2012, pp. 42–57.
[4]. C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” IEEE Transactions on Systems, Man, and Cybernetics, Part C, vol. 40, no. 6, pp. 601–618, 2010.
[5]. C. Romero and S. Ventura, “Educational data mining and learning analytics: An updated survey,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 3, p. e1355, 2020.
[6]. E. Alyahyan and D. Düştegör, “Predicting academic success in higher education: literature review and best practices,” International Journal of Educational Technology in Higher Education, vol. 17, no. 3, 2020.
[7]. M. Cantabella, R. Martínez-España, B. Ayuso, J. A. Yáñez, and A. Muñoz, “Analysis of student behavior in learning management systems through a big data framework,” Future Generation Computer Systems, vol. 90, pp. 262–272, 2019.
[8]. M. Vaarma et al., “Predicting student dropouts with machine learning,” International Journal of Educational Research, 2024.
[9]. A. M. Rabelo et al., “A model for predicting dropout of higher education students,” Smart Learning

Environments, 2025.
[10]. J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," Scientific Data, vol. 4, p. 170171, 2017.
[11]. E. Howard, "ouladFormat R package: Preparing the Open University Learning Analytics Dataset for analysis," arXiv preprint arXiv:2501.08366, 2025.
[12]. J. Samuelsen, W. Chen, and B. Wasson, "Integrating multiple data sources for learning analytics—review of literature," International Journal of Educational Technology in Higher Education, vol. 16, no. 11, 2019.
[13]. J. M. Dodero et al., "Trade-off between interoperability and data collection in learning analytics systems," Computers & Education, vol. 106, pp. 44–57, 2017.
[14]. M. Masud, X. Huang, J. Yong, and others, "Collaborative e-learning systems using semantic data interoperability and distributed metadata management," Computers in Human Behavior, vol. 72, pp. 298–310, 2017.
[15]. M. Paneque et al., "e-LION: Data integration semantic model to enhance learning analytics in multi-source e-learning ecosystems," Expert Systems with Applications, vol. 213, p. 119245, 2023.
[16]. M. Armbrust et al., "Delta Lake: High-performance ACID table storage over cloud object stores," Proceedings of the VLDB Endowment, vol. 13, no. 12, pp. 3411–3424, 2020.
[17]. J. Schneider et al., "The Lakehouse: State of the Art on Concepts and Technologies," SN Computer Science, vol. 5, 2024.
[18]. A. A. Harby et al., "Data Lakehouse: A survey and experimental study," Information Systems, 2025.
[19]. S. Slade and P. Prinsloo, "Learning analytics: Ethical issues and dilemmas," American Behavioral Scientist, vol. 57, no. 10, pp. 1510–1529, 2013.
[20]. P. Prinsloo and S. Slade, "An elephant in the learning analytics room: The obligation to act," in Proc. Seventh International Conference on Learning Analytics & Knowledge, 2017, pp. 46–55.
[21]. D. Ifenthaler and C. Schumacher, "Student perceptions of privacy principles for learning analytics," Educational Technology Research and Development, vol. 64, no. 5, pp. 923–938, 2016.
[22]. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[23]. T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
[24]. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, p. e0118432, 2015.
[25]. M. Kuhn and K. Johnson, Applied Predictive Modeling. New York, NY, USA: Springer, 2013.
[26]. B. Boehmke and B. Greenwell, Hands-On Machine Learning with R. Boca Raton, FL, USA: CRC Press, 2019.
[27]. R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in Learning Analytics. New York, NY, USA: Springer, 2014, pp. 61–75.
[28]. O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, "The current landscape of learning analytics in higher education," Computers in Human Behavior, vol. 89, pp. 98–110, 2018.
[29]. E. López-Meneses et al., "Educational Data Mining and Predictive Modeling in the Era of Artificial Intelligence," Computers, vol. 14, no. 2, p. 68, 2025.
[30]. T. Nguyen et al., "Data quality management in big data: Strategies, tools, and AI-enabled directions," 2025.
[31]. M. M. Ncube and P. Ngulube, "Leveraging learning analytics to personalise academic library services for enhanced student success: A systematic review," The Journal of Academic Librarianship, 2025.
[32]. S. Boujmiraz et al., "Predicting student performance: A comprehensive review," Machine Learning with Applications, 2026.