International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 04, No. 01, page 21 - 34
Submitted 9/2/2026; Accepted 28/3/2026; Published 29/3/2026

Data Infrastructure Application in Education: An Integrated Architecture for Secure Learning Analytics and Student Performance Prediction

Dinesh Pranav Mukerjea
Department of Business Administration, Dhaka International University, Dhaka-1212, BANGLADESH
e-mail: din.muker@diu.ac

Corresponding Autor: Dinesh Pranav Mukerjea

Abstract
Data infrastructure has become a strategic backbone of contemporary education because digital learning environments continuously generate student traces that can be transformed into actionable evidence for teaching, advising, and institutional planning. Yet the practical value of educational data depends on much more than storage capacity. Institutions must integrate heterogeneous sources, manage raw and curated data simultaneously, enforce privacy constraints, and deliver analytics outputs that are operationally useful and ethically defensible. This study develops a layered educational data infrastructure architecture that connects raw learning data, extract-transform-load processes, governance mechanisms, curated analytics repositories, and machine-learning services. This paper includes a reproducible empirical evaluation using the real xAPI-Edu-Data benchmark collected from the Kalboard 360 learning management environment. Three machine-learning models are compared under a common preprocessing pipeline, and an ablation analysis quantifies the incremental value of integrated behavioral, parental, and contextual features. The best-performing model achieves a test macro-F1 of 0.797 and a macro one-vs-rest ROC-AUC of 0.919, while the ablation study shows that the full integrated feature set clearly outperforms demographic-only and behavior-only alternatives. The paper contributes structured architecture, mathematical formalization of integrated learning analytics, and empirical evidence that richer, better-governed data pipelines produce more useful predictive signals for educational decision support.

Keywords—Data infrastructure, Learning analytics, Educational data mining, Data integration, Data lake, Student performance prediction, Educational privacy.

1 Introduction

The digitalization of education has shifted institutions from episodic record keeping toward continuous, platform-mediated data production. Learning management systems (LMS), student information systems, parental portals, assessment engines, discussion forums, and virtual classroom tools now produce diverse streams of student traces that can be transformed into actionable evidence for teaching, advising, and planning. The growth of learning analytics has therefore been driven not only by advances in statistical modeling, but also by the emergence of data infrastructures capable of collecting, integrating, and operationalizing educational evidence at scale [1–6]. In practical terms, this means that educational analytics is not just a modeling problem. It is equally a systems problem involving ingestion, standardization, meta- data, governance, and secure access.

The acceleration of online and hybrid learning during and after the COVID- 19 period made that infrastructural dependency more visible. Institutions that could observe only grades and attendance had difficulty interpreting asynchronous participation, resource access, remote engagement, and intervention timing, whereas institutions with richer digital ecosystems could form a more complete view of student activity [6, 7]. This change also revealed that

educational data are no longer predominantly tabular or transactional. They are increasingly mixed, combining structured student records with semi-structured event logs, clickstreams, xAPI traces, viewing sequences, and interaction histories. Consequently, questions of storage, integration, and governance are inseparable from pedagogical questions about student success.

The studies showed that real value of educational data depends on how heterogeneous sources are harmonized, how raw and curated representations are separated, how privacy and consent are protected, and how analytics-ready features are constructed from operational traces [8–13]. Without those mechanisms, institutions may accumulate large volumes of data yet remain unable to use them responsibly, reproducibly, or reliably.

At the same time, data-rich education creates ethical and legal obligations. Student interaction logs can reveal habits, attendance patterns, disengagement signals, or family context that are analytically useful but highly sensitive. If such traces are integrated carelessly, misinterpreted, or used without proportional safeguards, the same infrastructure intended to improve education can become a source of surveillance, bias, and loss of trust [14– 19]. Therefore, a rigorous discussion of educational data infrastructure must examine not only storage and integration, but also security, transparency, role-based access, and human oversight.

This paper addresses four contributions. First, it proposes a layered educational data infrastructure architecture that connects raw landing zones, ETL processes, curated analytics repositories, governance controls, and decision-support services. Second, it formalizes the logic of integrated educational analytics through mathematical expressions that clarify how fragmented student traces are transformed into predictive representations. Third, it validates the value of integration empirically using the real xAPI-Edu-Data benchmark from the Kalboard 360 learning environment [20, 21]. Fourth, it quantifies novelty through ablation analysis, showing that full cross-source feature integration materially outperforms narrower feature blocks. In this sense, the paper argues that educational data infrastructure should be designed as an analytical and ethical system rather than as a passive repository.

2 Related Work and Research Gap

Learning analytics literature has matured into a substantial interdisciplinary field spanning educational research, human-computer interaction, data mining, and institutional decision support. Foundational work defined learning analytics as the measurement, collection, analysis, and reporting of data about learners and their contexts [1–3]. Later studies expanded that discussion by examining adoption, dashboards, personalized learning, and institutional implementation challenges [4, 5, 22–25]. Collectively, this body of work shows that educational analytics is not merely a predictive exercise; it is a socio-technical process in which data acquisition, interpretation, and action are tightly coupled.

The second line of work focuses on student performance prediction and educational data mining. Surveys consistently report that models are strongest when they combine static background information with dynamic learner behavior, attendance, and contextual signals [26–28]. Ensemble methods, explainable models, and neural approaches have all been used to forecast achievement, progression, or dropout risk. Yet many of these studies treat data engineering as a preliminary setup task rather than as a research contribution in its own right. The model receives center stage, while the upstream transformation that produced the usable feature set remains weakly documented.

A third-stream addresses privacy, ethics, and student agency. Researchers have argued that analytics should operate under explicit principles of transparency, student benefit, proportionality, and privacy by design [14–18]. These concerns matter because educational data are produced within relationships of dependency and evaluation; students cannot be treated as neutral data subjects in the same way as anonymous clickstream users in commercial settings. Recent work on cloud privacy and secure storage further highlights the need for architecture that incorporates pseudonymisation, encryption, controlled access, and auditable data movement [19].

The infrastructure literature adds a different but complementary perspective. Data-lake, metadata, ETL, and integration studies explain how heterogeneous data are ingested, transformed, and curated for downstream use [8–13]. However, these studies are rarely tailored to educational intervention contexts. They explain how to manage heterogeneity, but not how to align infrastructure with educational interpretability, early-warning needs, or student-facing accountability. The field therefore still contains a gap between educational aspiration and technical implementation.

Table 1 summarizes that gap more explicitly. Existing learning analytics work often assumes that integrated data already exists. Existing infrastructure work often lacks education-specific validation. This paper bridges those streams by treating data integration as both a conceptual architecture problem and an empirical performance factor.

The central research question of the paper is therefore not only whether educational data is useful, but how infrastructure should be organized so that educational data become usable, trustworthy, and analytically productive. This framing leads directly to the empirical hypothesis tested later in the paper: integrated educational feature sets

should outperform narrower, fragmented feature blocks in student performance prediction. If that hypothesis is supported, then the value of infrastructure is not rhetorical but measurable.

Table 1: Positioning of the present study against related work.

Stream	What prior work establishes	Remaining limitation	How this study extends it
Learning analytics foundations [1–5]	Analytics concepts, dashboards, and adoption	Limited infrastructure engineering detail	Links architecture, governance, and predictive evaluation
Student performance prediction [20, 21, 26–28]	Strong benchmarking tradition for educational models	Integration often treated as background preprocessing	Quantifies the value of feature integration through ablation
Privacy and ethics [14–19]	Normative guidance for responsible analytics	Often not tied to operational pipeline design	Embeds governance and least-privilege access in the proposed architecture
Data-lake and ETL research [8–13]	Principles for storage, metadata, and transformation	Rarely evaluated in education-specific decision support settings	Provides an education-specific layered architecture plus reproducible experiment

3 Problem Formulation

3.1 Fragmentation, data lakes, and analytical under-utilisation

Educational institutions typically operate multiple operational systems that were not originally designed for shared analytics. Administrative systems store enrollment and demographic records; LMS platforms record content access, clicks, submissions, and discussion activity; communication systems capture announcements and responses; and parent-facing systems reflect satisfaction, support, or survey participation. When these sources remain disconnected, decision-makers receive fragmented evidence. A lecturer may see assignment grades but not resource use, an advisor may observe absence patterns but not behavioral participation, and administrators may see aggregate pass rates without understanding the interaction structure behind them.

Data lakes emerged as a practical response to this problem because they can ingest heterogeneous data with minimal up-front schema constraints [8, 9]. In educational settings, this allows institutions to preserve raw event traces, semi-structured logs, administrative extracts, and snapshots of adjacent support systems in a common environment. Yet flexibility alone does not create knowledge. If metadata, lineage, and curation are weak, a data lake can become a "data swamp" in which data exist but are not interpretable or reusable [10]. This is especially problematic in education because intended users of analytics such as lecturers, counselors, and quality-assurance staff are usually not data engineers.

The fragmentation problem can be formalized. Let raw educational data arrive from n heterogeneous systems:

$$X_{\text{raw}} = \{X_1, X_2, \ldots, X_n\} \tag{1}$$

If the systems remain disconnected, then the effective information available for decision-making is

$$I_{frag} = \sum_{i=1}^{n} g(X_i), \tag{2}$$

Where g() denotes source-specific interpretation. Because cross-source relationships are not explicitly represented, interactions among sources remain weakly observed or entirely absent. By contrast, integrated analytics aims to construct

$$X_{int} = \Phi(X_1, X_2, \ldots, X_n) \tag{3}$$

Where Φ() denotes extraction, cleaning, semantic alignment, joining, and feature construction. The central problem is therefore not only capacity, but the absence of a robust transformation process that can convert heterogeneity into coherent analytical evidence.

3.2 Privacy, security, and educational trust

The second major problem concerns privacy and security. Educational data often includes personally identifiable information, attendance records, performance histories, support interactions, and behavioral traces that can reveal disengagement or social vulnerability. When such information is integrated across systems, the potential value of analytics increases, but so does the potential impact of unauthorized access, opaque inference, or unjustified intervention [14–17]. A secure educational infrastructure must therefore optimize a dual objective: maximize analytical usefulness while minimizing exposure and unfair use.

From a systems perspective, overall breach exposure can be conceptualized as

$$R = f(V, E, A, G) \tag{4}$$

Where V denotes technical vulnerabilities, E the number of exchange interfaces, A the accessibility of sensitive attributes, and G the maturity of governance controls. If data are exchanged via unmanaged spreadsheets, insecure staging areas, or weakly controlled exports, then E and A increase while G remains low. The result is higher systemic risk.

The issue is not limited to malicious intrusion. Educational harm can also arise from over-collection, ambiguous model outputs, or interventions that students do not understand. Privacy by design, encryption, and role-based access control are therefore necessary but not sufficient. Institutions also require transparent purposes for collection, retention limits, audit trails, and human accountability for model-triggered decisions [14–18]. In a trustworthy infrastructure, security is not a single software feature; it is a property of the whole pipeline.

3.3 Research objective and analytical hypothesis

The conceptual argument above leads to a concrete empirical claim: integrated educational feature sets should outperform narrower or fragmented feature blocks in student performance prediction. This hypothesis is consistent with educational data mining research showing that learner outcomes are shaped jointly by demographic context, behavioral engagement, attendance, and family support [20, 21, 26, 27]. If that is correct, then infrastructure is not an abstract organizational convenience. Better integration should translate into measurably better predictive fidelity.

Accordingly, the objective of this study is twofold:

i. To design a secure, layered data infrastructure model for educational analytics, and
ii. To evaluate whether integrated feature representations improve student-performance prediction relative to more limited alternatives.

The experimental study is not intended to claim universal causality across all institutions. Rather, it serves as a reproducible proof of concept showing that integration matters materially when educational data are modeled for decision support.

4 Proposed Integrated Data Infrastructure Framework

4.1 Architectural layers

Figure 1 presents the proposed architecture. The design begins with educational data sources, including LMS events, student information systems, attendance records, xAPI-compatible learning traces, and parent-related signals. Raw data first enters a landing zone or data lake where fidelity is preserved, and schema constraints are kept intentionally light. This design choice is important because raw events may later be re-used for alternative feature engineering, fairness analysis, or retrospective audit.

A second layer performs integration and transformation. Here, ETL processes are responsible for profiling source quality, harmonizing identifiers, validating ranges, imputing missing values where necessary, normalizing categories, and constructing analytics-ready features [11–13]. The separation between raw and transformed zones is methodological as well as operational. Raw data supports traceability and reproducibility, whereas curated data supports institutional reporting and predictive modeling. Conflating these zones makes lineage difficult to verify and complicates governance.

A third layer stores curated and semantically aligned data in a warehouse or feature mart optimized for analysis. This layer supports dashboards, early warning systems, model training, and longitudinal trend analysis. Above it sits analytics services, including educator-facing dashboards, ranked risk lists, and cohort-level summaries. Importantly, governance and security are treated as transversal rather than peripheral layers. Metadata, lineage, role-based access control, pseudonymisation, encryption, and audit logging should attach to the pipeline end-to-end [14–16, 19]. This architecture ensures that analytical readiness is achieved without sacrificing accountability.

4.2 Mathematical formalisation of integrated educational analytics

To make the role of infrastructure analytically explicit, the integrated representation of student *i* can be written as

$$x_i = h\left(x_i^{(d)} \parallel x_i^{(b)} \parallel x_i^{(p)}\right) \tag{5}$$

Where $x_i^{(d)}$ denotes demographic and academic background variables, $x_i^{(b))}$ behavioral engagement variables, and $x_i^{(p)}$ parental and attendance context. The operator denotes feature concatenation after preprocessing and semantic alignment. Equation 5 embodies the paper's central infrastructural claim: useful educational insight depends on linking features that originate in different operational subsystems.

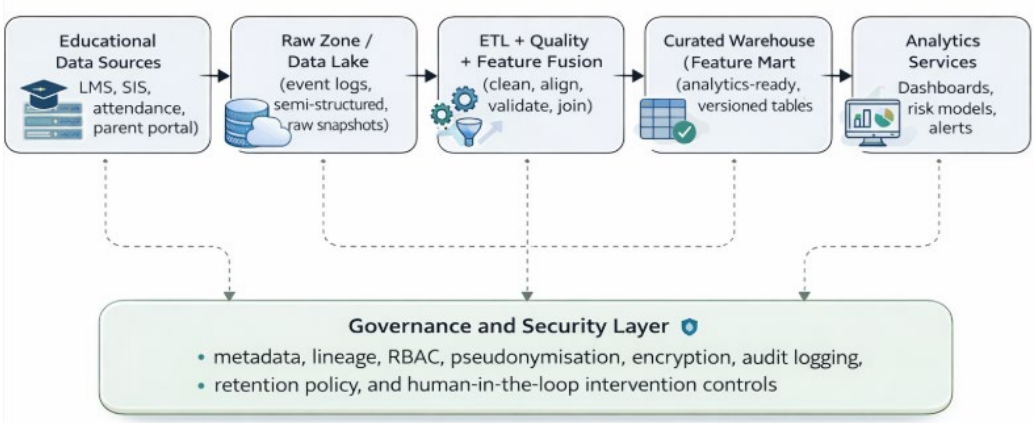

Figure 1: Proposed integrated data infrastructure for educational analytics. The architecture separates raw preservation from curated transformation and embeds governance across the full pipeline.

For multiclass performance prediction with target yi ∈ L, M, H the posterior probability of class c may be expressed as

$$P(y_i = c \mid x_i) = \frac{\exp(z_{ic})}{\sum_{k=1}^{C} \exp(z_{ik})} \tag{6}$$

Where z_{ic} denotes the class-specific score assigned to students *i*. Although the empirical study compares several model families rather than relying on a single classifier, Equation 6 expresses the general learning objective: map an integrated student representation to a calibrated probability distribution over educational outcome class.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} \log P(y_i = c \mid x_i) \tag{7}$$

Algorithm 1 Secure ETL and feature fusion for educational analytics

Require: Raw sources $\{X_1, \ldots, X_n\}$, identifier schema $\mathcal{I}$, governance policy $\mathcal{G}$
Ensure: Curated analytical table X_{int}
Initialize empty curated store $X_{\text{int}} \leftarrow \emptyset$
for each source X_j **do**
    Validate schema, ranges, and timestamp integrity
    Standardize identifiers using $\mathcal{I}$
    Attach metadata: lineage, source version, ingestion time
    Apply policy checks from $\mathcal{G}$ (role, sensitivity, retention)
    Clean anomalies; encode categories; normalize numeric fields
end for
Join source tables on standardized identifiers and time windows
Construct derived features (engagement, attendance, parental context)
Pseudonymise analytical identifiers and store curated output
return X_{int}

4.3 Algorithms for secure ETL and operational risk scoring

Algorithm 1 describes the secure ETL and feature-fusion workflow that operationalizes Equations 1–5. The key idea is that integration is treated as a controlled transformation rather than a blind merger. Source-specific records are profiled, validated, aligned through shared identifiers, and subjected to governance checks before being emitted to the curated analytical layer.

Algorithm 2 then shows how the curated analytical table can support operational early warning while preserving human oversight. The model produces class probabilities and a ranked risk score. However, any intervention still requires a domain review step rather than fully automated action. This detail is important because a technically accurate alert is not automatically an educationally appropriate intervention.

4.4 Governance requirements for deployment

A submission-ready architecture for education must specify governance requirements explicitly. First, identifiers and source mappings should be versioned and documented so that linkage is reproducible and auditable. Second, attribute exposure should follow least privilege principles: instructors may require class-level risk summaries without access to all raw behavioral records, while data engineers may require ingestion access without pedagogical authority to trigger interventions. Third, models should be monitored not only for predictive drift but also for governance drift, including changes in source quality, category definitions, and consent conditions.

Algorithm 2 Human-in-the-loop risk scoring and alert generation

Require: Curated feature matrix X_{int}, trained classifier $h(\cdot)$, threshold τ
Ensure: Ranked alert list $\mathcal{A}$
$\mathcal{A} \leftarrow \emptyset$
for each student record $\mathbf{x}_i$ **do**
    Estimate class probabilities $\mathbf{p}_i \leftarrow h(\mathbf{x}_i)$
    Compute risk score $r_i = 1 - p_i(H)$
    if $r_i \geq \tau$ **then**
        Append $(i, r_i, \mathbf{p}_i)$ to $\mathcal{A}$
    end if
end for
Sort $\mathcal{A}$ by descending risk score
Route $\mathcal{A}$ to authorized educator/advisor for contextual review
return $\mathcal{A}$

These requirements are particularly important when parental, attendance, and behavioral variables are combined. Such variables can improve prediction but also heighten ethical sensitivity. Therefore, purpose limitation, transparent

retention rules, and documentation of intervention logic should be built into the infrastructure itself. Technically elegant architecture without these controls would remain insufficient for trustworthy educational use.

5 Experimental Design

5.1 Dataset and task definition

To complement conceptual architecture with empirical evidence, the study uses the real xAPI-Edu-Data benchmark derived from the Kalboard 360 learning management environment [20, 21]. The dataset contains 480 student records and 17 columns spanning demographics, academic stage, parental context, behavioral engagement, and target performance class. Its continued use in recent educational prediction research makes it a suitable benchmark for testing whether cross-source integration improves predictive performance [26–28].

Table 2: Feature groups used in the empirical evaluation.

Feature block	Variables	Count
Demographic and academic background	gender, NationalITy, PlaceofBirth, StageID, GradeID, SectionID, Topic, Semester, Relation	9
Behavioral engagement	raisedhands, VisITedResources, AnnouncementsView, Discussion	4
Parental and attendance context	ParentAnsweringSurvey, ParentschoolSatisfaction, StudentAbsenceDays	3
Target label	Class (L, M, H)	1

The prediction target is the three-level performance label Class with categories Low (L), Medium (M), and High (H). In the benchmark used here, the class distribution is moderately imbalanced but not extreme: 127 low-performing students, 211 medium-performing students, and 142 high-performing students. This makes macro-averaged metrics preferable to raw accuracy alone because educationally important minority patterns can otherwise be obscured.

The dataset structure is especially appropriate for the present paper because it reflects the integration problem discussed conceptually. Demographic and academic variables approximate records that may live in administrative systems. Behavioral variables resemble event traces captured by LMS platforms. Parental response and attendance variables represent contextual data that may come from adjacent support systems. Thus, even though the benchmark is compact, it mirrors the broader claim that educational analytics depends on linking multiple categories of evidence.

5.2 Preprocessing, models, and evaluation protocol

A rigorous experiment should make the preprocessing pipeline explicit. The study therefore applies a unified transformation strategy across models. Categorical variables are imputed with the most frequent category and encoded with one-hot representation. Numeric variables are median imputed and standardized. Although the dataset contains no missing values in its downloaded benchmark form, the imputation steps are retained so the pipeline is robust and reusable.

The evaluation protocol uses a stratified 80/20 train-test split with random state 42 to preserve class proportions in both partitions. Robust assessment is then performed using five-fold stratified cross-validation on the training split. This arrangement follows common benchmarking practice: cross-validation estimates expected model behavior under resampling, while the hold-out test set provides an untouched point of comparison.

Three models are compared: Logistic Regression as a transparent linear baseline, Random Forest as a strong bagged-tree ensemble, and Extra Trees as a more randomized ensemble. These models were selected for substantive reasons. Logistic Regression provides interpretability and a conventional baseline. Random Forest is widely used in educational prediction because of its robustness to mix feature types. Extra Trees can capture threshold effects and

interactions while reducing variance through aggressive randomization, making it a strong candidate when integrated features are heterogeneous.

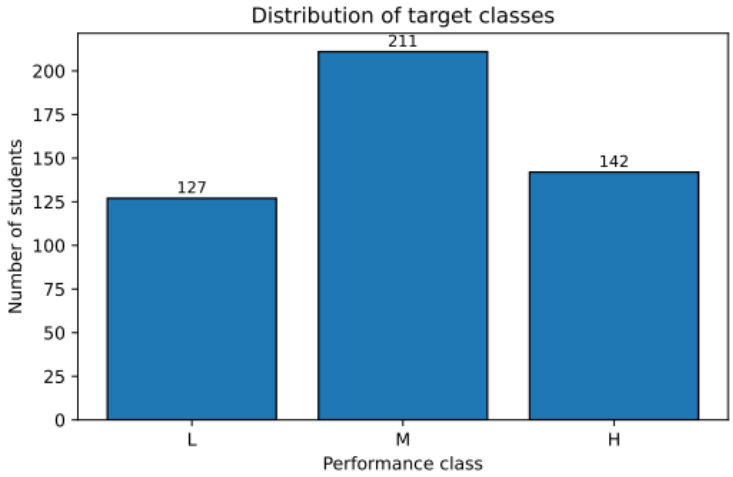

Figure 2: Distribution of target classes in the xAPI-Edu-Data benchmark.

5.3 Novelty validation through ablation

Beyond model comparison, the study conducts a feature-block ablation using the best-performing model. Four configurations are tested: demographic- only, behavioral-only, behavioral plus parental, and the full integrated set. This design is central to the paper's novelty claim. Instead of asking only which algorithm performs best, the ablation asks how much analytical value is produced by integration itself.

Table 2: Feature groups used in the empirical evaluation.

Item	Specification
Dataset	xAPI-Edu-Data benchmark from Kalboard 360 [20, 21]
Instances / predictors	480 students / 16 predictors + 1 target
Task	Three-class student performance prediction (L, M, H)
Train-test split	80/20 stratified split, random_state = 42
Cross-validation	5-fold stratified CV on training split
Preprocessing	Median imputation + scaling for numeric features; most-frequent imputation + one-hot encoding for categorical features
Models compared	Logistic Regression, Random Forest, Extra Trees
Metrics	Accuracy, macro-precision, macro-recall, macro-F1, macro ROC-AUC (one-vs-rest)

The end-to-end experimental workflow is shown in Figure 3. It begins with dataset ingestion and unified preprocessing, proceeds through model training and evaluation, and concludes with ablation-based novelty proof. This makes the paper more than a conceptual essay: the claim that infrastructure matters are tested quantitatively under a reproducible protocol.

6 Results

6.1 Comparative model performance

Table 4 presents the main predictive results. The Extra Trees model delivered the strongest hold-out macro-F1 (0.797) and macro one-vs-rest ROC- AUC (0.919), while Random Forest matched it in accuracy (0.792) but trailed

slightly in balanced class performance. Logistic Regression served as a useful baseline but underperformed the ensemble models across all major metrics. The grouped bar chart in Figure 4 makes the same pattern visible at a glance. These results are important for two reasons. First, they show that the integrated educational feature space contains non-linear interaction structure that tree ensembles can exploit more effectively than a linear baseline. Second, they indicate that evaluating educational prediction only by accuracy is insufficient. Random Forest and Extra Trees are tied in test accuracy, yet Extra Trees provides the strongest macro-F1, which is more informative when minority or pedagogically sensitive classes matter. This design is central to the paper's novelty claim. Instead of asking only which algorithm performs best, the ablation asks how much analytical value is produced by integration itself.

Future studies may integrate additional algorithms such as SVM or Random Forest to improve classification accuracy and explore multi-label sentiment analysis to identify multiple sentiments within a single opinion.

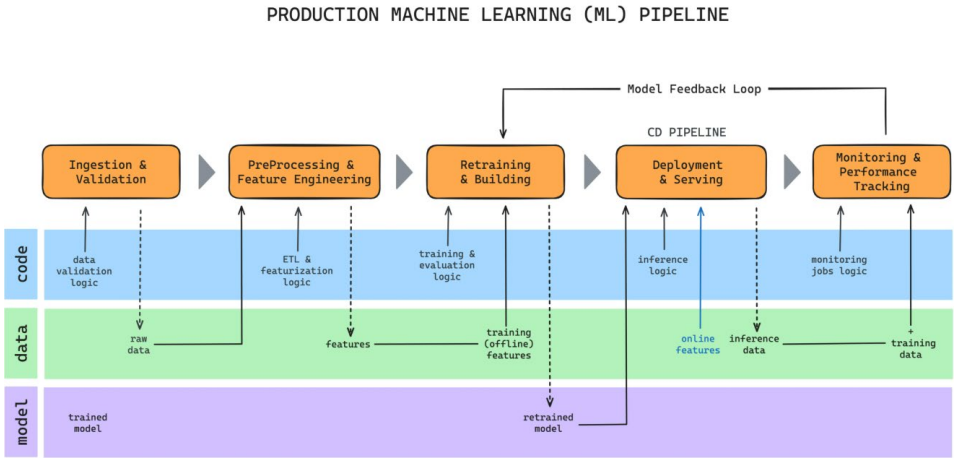

Figure 3: End-to-end experimental pipeline used to convert integrated educational data into validated prediction results.

Table 4: Model comparison on cross-validation and hold-out test data.

Model	CV Acc.	CV Prec.	CV Rec.	CV F1	Test Acc.	Test Prec.	Test Rec.	Test F1	Test ROC-AUC
Extra Trees	0.776	0.778	0.773	0.773	0.792	0.813	0.788	0.797	0.919
Random Forest	0.786	0.791	0.783	0.783	0.792	0.805	0.782	0.787	0.917
Logistic Regression	0.737	0.737	0.73	0.733	0.74	0.754	0.758	0.757	0.887

6.2 Ablation analysis: quantifying the value of integration

The ablation study in Table 5 provides the clearest empirical answer to the paper's main question. Demographic features alone yielded only 0.573 macro-F1 on the test split. Behavioral variables alone improved macro-F1 to 0.653. Adding parental and attendance context increased macro-F1 to 0.710. The full integrated feature set then achieved the strongest result, with 0.797 macro-F1 and 0.919 macro ROC-AUC.

This progression is analytically meaningful. It shows that no single feature block is sufficient for robust student performance prediction. Demographic variables provide only a coarse approximation of learner context. Behavioral variables capture platform activity but not the wider support environment. Once parental and attendance signals are added, performance improves materially. The full integrated set clearly outperforms every partial alternative. This is the empirical novelty proof of the manuscript: better integration creates better predictive information.

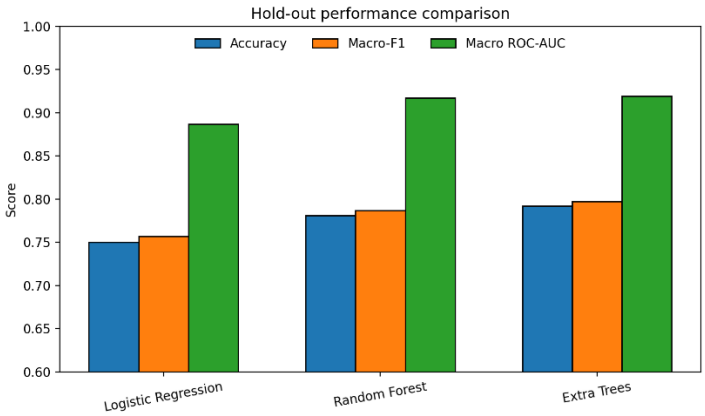

Figure 4: Hold-out performance comparison across the three evaluated models.

Table 5: Ablation study using the best-performing model (Extra Trees).

Feature Group	Features	CV Acc.	CV F1	Test Acc.	Test F1	Test ROC-AUC
Full integrated set	16	0.787	0.782	0.792	0.797	0.919
Behavioral + parental	7	0.708	0.697	0.729	0.71	0.867
Behavioral only	4	0.648	0.638	0.677	0.653	0.798
Demographic only	9	0.562	0.543	0.594	0.573	0.717

6.3 Class-wise behavior and confusion analysis

The confusion matrix in Figure 5 and the class-wise scores in Table 6 show that the best-performing model handles the low-performing class particularly well, achieving 0.875 precision, 0.840 recall, and 0.857 F1. This is operationally important because early-warning systems are most valuable when they can identify weaker-performing students with acceptable precision. The medium class remains the most ambiguous, which is unsurprising because it lies between the two extremes and shares characteristics with both. From an educational intervention perspective, this confusion pattern is plausible. In many settings, the distinction between medium and high performance reflects degree rather than kind; both groups may show healthy engagement and manageable absence patterns. By contrast, low-performing students often exhibit more distinct combinations of weak attendance and limited interaction, making them easier to detect. This again underscores the argument: such patterns become visible only when attendance, engagement, and contextual signals are jointly available.

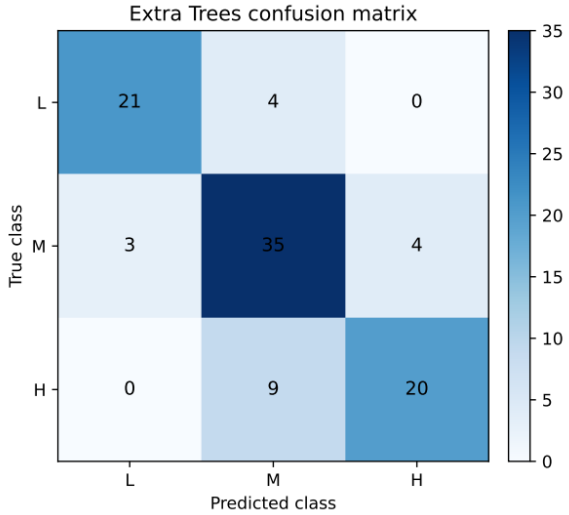

Figure 5: Confusion matrix for the Extra Trees model on the hold-out test split.

6.4 Feature importance and explanatory signals

Figure 6 summarizes the strongest predictors in the Extra Trees model after aggregating one-hot encoded importance values back to their original variables. Student absence is the most influential signal, followed by resource visitation, topic, raised hands, parental survey response, and relation-to-parent variables. The prominence of absence and activity variables aligns well with educational intuition and prior xAPI-based prediction studies [20, 21]. This ranking also strengthens the broader thesis of the paper. The strongest predictors arise at the intersection of behavior, context, and support environment rather than from static background variables alone. In infrastructural terms, that means predictive value depends directly on whether institutions can combine traces that are usually stored in different operational contexts. If any one of these streams is missing or delayed, the utility of the resulting analytics pipeline is diminished.

Table 6: Class-wise precision, recall, and F1-score for the best-performing model

Class	Support	Precision	Recall	F1-score
Low (L)	25	0.875	0.84	0.857
Medium (M)	42	0.729	0.833	0.778
High (H)	29	0.833	0.69	0.755

7 Discussion

The empirical findings show that the most useful educational data infrastructures are not those that merely accumulate the largest amount of raw data. They are infrastructures that convert heterogeneous evidence into curated, governed, analytics-ready representations. The ablation results make this explicit: integration across feature blocks produces substantial gains in predictive quality. In other words, the infrastructure layer influences model quality before any algorithm is selected.

The results clarify the relationship between educational analytics and data architecture. Much of the learning analytics literature rightly focuses on dashboards, student success, and intervention. Yet those applications depend on upstream data engineering and governance. The present manuscript therefore argues for a tighter coupling between pedagogical and infrastructural reasoning. Early-warning systems should not be evaluated only by predictive score; they should also be evaluated by whether their data foundations are reproducible, semantically coherent, and ethically governed [14–17]. This point becomes more important as institutions scale analytics from isolated pilots to institution-wide services.

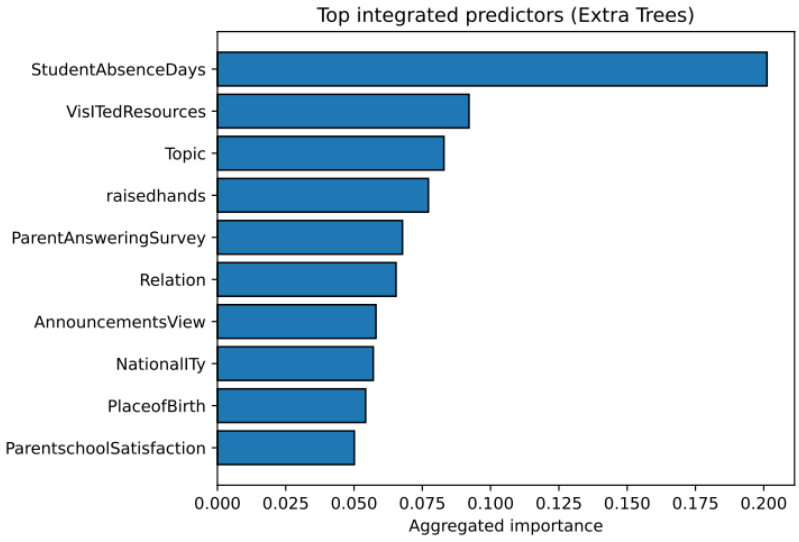

Figure 6: Top aggregated feature importances identified by the Extra Trees model.

The study has practical implications for institutions with limited resources. The benchmark used here is not a massive enterprise repository; it is a relatively compact but structurally rich educational dataset. Even so, integrated

features yield clear performance gains. This suggests that institutions do not need hyperscale infrastructure before they can benefit from better data architecture. What they need is disciplined design: stable identifiers, raw-to-curated separation, metadata, ETL quality control, controlled access, and repeatable evaluation. A smaller but well-designed pipeline may be more valuable than a larger but weakly curated data lake.

The findings matter for interpretability and trust. The importance of absence, resource usage, and classroom interaction is understandable to educators and advisors, which improves the likelihood of operational acceptance. However, the presence of parental and contextual variables also underscores the need for restraint. A predictor can be useful without being appropriate for unrestricted access. Therefore, explainability must be paired with governance. The relevant question is not simply which variables matter, but who is allowed to see them and for what purpose.

8 Practical Implications for Educational Institutions

Several actionable implications follow from the study. Institutions planning analytics initiatives should begin with data inventory and semantic mapping rather than model procurement alone. If identifiers, event definitions, update frequencies, and access policies are inconsistent, later modeling work will remain fragile regardless of algorithmic sophistication. Raw zones and curated zones should be separated explicitly so that auditability and institutional usability are both preserved.

Security should be embedded into the pipeline rather than applied only at the dashboard stage. Encryption in transit and at rest, audit logging, pseudonymisation for model development, and role-based access control should be treated as baseline infrastructure requirements rather than optional enhancements [19]. Similarly, evaluation practice should reflect educational priorities: macro-level metrics and class-wise diagnostics are often more informative than accuracy alone when the main purpose is timely intervention for lower-performing students.

Finally, institutions should embed model output in human decision processes with clear oversight. Ranked risk lists should inform educators and advisors, not replace pedagogical judgment. Prediction without contextual review can create overconfidence and encourage inappropriate intervention. Mature educational analytics program therefore combines technical pipelines with governance, documentation, and professional accountability.

9 Limitations, Reproducibility, and Future Research

The benchmark is real and widely used, but it represents a single learning environment and a modest sample size relative to large institutional deployments. The reported metrics should therefore be interpreted as evidence of infrastructural plausibility rather than universal guarantees. External validation on larger repositories such as OULAD would strengthen generalisability [29, 30].

The current experiment focuses on tabular, already-curated benchmark data. Real institutional data infrastructures often handle noisier event streams, asynchronous updates, missing identifiers, and policy constraints not fully captured in benchmark corpora. A third limitation is that the present analysis compares strong baseline models but does not yet include temporal sequence models, multimodal fusion, fairness auditing, or causal evaluation. These extensions are important for mature deployment, especially when interventions may affect subgroups differently.

Future work should proceed in at least four directions. First, the proposed architecture should be validated on larger, multi-table educational repositories. Second, privacy-preserving learning analytics strategies, including differential privacy and federated approaches, should be examined for institutionally sensitive settings. Third, temporal and sequence-aware models should be used to capture progression rather than static summary features. Fourth, future studies should evaluate not only predictive quality but also governance effectiveness, student acceptance, and intervention outcomes.

10 Conclusion

This study reconceptualizes data infrastructure in education as a combined problem of architecture, integration, governance, and analytics readiness. The proposed layered architecture spanning data sources, raw landing zones, ETL, curated warehouses, governance controls, and analytics services offer a practical and theoretically grounded model for institutions seeking to operationalize learning analytics responsibly.

The empirical results strengthen that conceptual argument. Extra Trees achieved the best overall hold-out macro-F1, and the ablation study showed that the full integrated feature set clearly outperformed demographic-only, behavior-only, and partially integrated alternatives. These findings indicate that integration creates measurable analytical value. More broadly, they show that infrastructure decisions shape educational analytics outcomes before modeling choices are even made.

Educational data infrastructure should therefore not be treated as passive plumbing. It is a strategic layer that conditions what institutions can know, how reliably they can know it, and how ethically they can act on it. Institutions that invest in integrated, governed, and secure data architectures will be better positioned to support early intervention, evidence-based planning, and trustworthy learning analytics. Institutions that neglect these foundations may accumulate data without achieving educational intelligence.

BIBLIOGRAPHY

[1]. G. Siemens, Learning analytics: The emergence of a discipline, American Behavioral Scientist 57 (10) (2013) 1380–1400. doi:10.1177/ 0002764213498851.

[2]. M. A. Chatti, A. L. Dyckhoff, U. Schroeder, H. Thüs, A reference model for learning analytics, International Journal of Technology Enhanced Learning 4 (5-6) (2012) 318–331. doi:10.1504/IJTEL.2012.051815.

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

[4]. L. Marquez-Vera, et al., Adoption of learning analytics in higher education institutions: A systematic literature review, British Journal of Educational Technology (2024). doi:10.1111/bjet.13385.

[5]. K. Verbert, E. Duval, J. Klerkx, S. Govaerts, J. L. Santos, Learning analytics dashboard applications, American Behavioral Scientist 57 (10) (2013) 1500–1509. doi:10.1177/0002764213479363.

[6]. A. Al-Fraihat, M. Joy, R. Masa'deh, J. Sinclair, Evaluating e-learning systems success: An empirical study, Computers in Human Behavior 102 (2020) 67–86. doi:10.1016/j.chb.2019.08.004.

[7]. T. Basilaia, D. Kvavadze, Transition to online education in schools during a SARS-CoV-2 coronavirus pandemic in georgia, Pedagogical Research 5 (4) (2020). doi:10.29333/pr/7937.

[8]. P. N. Sawadogo, J. Darmont, On data lake architectures and metadata management, Journal of Intelligent Information Systems 56 (2021) 97– 120. doi:10.1007/s10844-020-00608-7.

[9]. S. Azzabi, Z. Alfughi, A. Ouda, Data lakes: A survey of concepts and architectures, Computers 13 (7) (2024) 183. doi:10.3390/ computers13070183.

[10]. D. Boukraâ, M. Bala, S. Rizzi, Metadata management in data lake environments: A survey, Journal of Library Metadata (2024). doi: 10.1080/19386389.2024.2359310.

[11]. A. Halevy, A. Rajaraman, J. J. Ordille, Data integration: The teenage years, Proceedings of the VLDB Endowment 2 (2) (2009) 9–16. doi: 10.14778/1687553.1687555.

[12]. J. Noverlita and H. Surbakti, Streamlining stock price analysis: Hadoop ecosystem for Machine Learning Models and big data analytics, International Journal of Information Technology and Computer Science, vol. 15, no. 5, pp. 25–34, Oct. 2023. doi:10.5815/ijitcs.2023.05.03.

[13]. M. Lenzerini, Data integration: A theoretical perspective, in: Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2002, pp. 233–246. doi: 10.1145/543613.543644.

[14]. S. Slade, P. Prinsloo, Learning analytics: Ethical issues and dilemmas, American Behavioral Scientist 57 (10) (2013) 1510–1529. doi:10.1177/ 0002764213479366.

[15]. A. Pardo, G. Siemens, Ethical and privacy principles for learning analytics, British Journal of Educational Technology 45 (3) (2014) 438–450. doi:10.1111/bjet.12152.

[16]. T. Hoel, W. Chen, Privacy and data protection in learning analytics should be a feature, not a bug, Research and Practice in Technology Enhanced Learning 13 (2018) 25. doi:10.1186/s41039-018-0086-8.

[17]. C. Lawson, C. Beer, D. Rossi, T. Moore, J. Fleming, Identification of 'at risk' students using learning analytics: The ethical dilemmas of intervention strategies in a higher education institution, Educational Technology Research and Development 64 (5) (2016) 957–968. doi:10.1007/s11423-016-9459-0.

[18]. W. Weng, Exploring the ethical topic of learning analytics, Educational Technology Research and Development 69 (2021) 339–341. doi:10. 1007/s11423-020-09873-3.

[19]. P. Yang, N. Xiong, J. Ren, Data security and privacy protection for cloud storage: A survey, IEEE Access 8 (2020) 131723–131740. doi: 10.1109/ACCESS.2020.3009876.

[20]. E. A. Amrieh, T. Hamtini, I. Aljarah, Preprocessing and analyzing educational data set using X-API for improving student's performance, in: 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, 2015, pp. 1–5. doi:10.1109/AEECT.2015. 7360581.

[21]. E. A. Amrieh, T. Hamtini, I. Aljarah, Mining educational data to predict student's academic performance using ensemble methods, International Journal of Database Theory and Application 9 (8) (2016) 119–

136. doi: 10.14257/ijdta.2016.9.8.13.
[22]. M. H. de Menéndez, R. Morales-Menendez, H. E. Díaz, J. C. Arámburo-Lizárraga, Learning analytics: State of the art, Journal of Computing in Higher Education 34 (2022) 547–565. doi:10.1007/ s12528-022-00930-0.
[23]. E. T. Khor, N. F. M. Noor, S. M. Yusof, A systematic review of the role of learning analytics in personalized learning, Education Sciences 14 (1) (2024) 51. doi:10.3390/educsci14010051.
[24]. N. A. Johar, et al., Learning analytics on student engagement to enhance learning performance: A systematic review, Sustainability 15 (10) (2023) 7849. doi:10.3390/su15107849.
[25]. D. Hooshyar, et al., Learning analytics in supporting student agency: A systematic review, Sustainability 15 (18) (2023) 13662. doi:10.3390/ su151813662.
[26]. W. Xiao, P. Ji, J. Hu, A survey on educational data mining methods used for predicting students' performance, Engineering Reports 4 (5) (2022). doi:10.1002/eng2.12482.
[27]. W. Xiao, P. Ji, J. Hu, A state-of-the-art survey of predicting students' performance using artificial neural networks, Engineering Reports (2023). doi:10.1002/eng2.12652.
[28]. R. Alamri, B. Alharbi, Explainable student performance prediction models: A systematic review, IEEE Access 9 (2021) 33132–33143. doi:10.1109/ACCESS.2021.3061368.
[29]. J. Kuzilek, M. Hlosta, Z. Zdrahal, Open university learning analytics dataset, Scientific Data 4 (2017) 170171. doi:10.1038/sdata.2017. 171.
[30]. J. Kuzilek, M. Hlosta, Z. Zdrahal, Open university learning analytics dataset, UCI Machine Learning Repository (2015). doi:10.24432/ C5KK69.