ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Explainable AI-Driven Student Risk Prediction

M. Priyanga[1], V. Ranjith[2], B. Pranesh[2], A. Sankar[2], R. Karthikeyan[2]

[1]Assistant Professor, Department of Artificial Intelligence and Data Science, Rathinam Technical Campus, Coimbatore, Tamil Nadu, India
[2]Department of Artificial Intelligence and Data Science, Rathinam Technical Campus, Coimbatore, Tamil Nadu, India

Published October 2026

Abstract

Student attrition and academic failure remain persistent challenges in Indian higher education, where large cohorts make it difficult for advisors to identify struggling students early. Many machine-learning predictors operate as black boxes, which limits the trust, accountability and practical use that faculty and counselors can place in them. This paper proposes an explainable artificial intelligence (XAI)-driven framework for student risk prediction that combines ensemble classifiers (Random Forest and XGBoost) with SHAP and LIME explanations, a rule-based early warning layer and a human-in-the-loop intervention workflow. The framework brings academic, attendance and learning management system (LMS) data together in a modular, privacy-first architecture that includes role-based access control, tiered consent and periodic bias auditing. An evaluation protocol for the prediction pipeline is defined on the Open University Learning Analytics Dataset (OULAD), using only data available up to day 28 of a course. Every prediction is accompanied by a ranked list of contributing factors, so that advisors can respond to the reasons behind a risk estimate and not to the score alone. The framework offers a transparent, scalable and low-cost approach to the early identification and support of at-risk students.

Keywords: Algorithmic fairness; early warning system; explainable artificial intelligence; learning analytics; student risk prediction

1 Introduction

Student dropout and course failure carry serious academic, financial and social costs for students, families and institutions. For the student, a failed course or an abandoned programme can mean lost time, lost fees and diminished confidence. For the institution, it means wasted teaching capacity, lower graduation rates and weaker accreditation outcomes. In many Indian colleges a single faculty advisor is responsible for dozens, and sometimes hundreds, of students. Warning signs such as falling attendance, missed assignments and declining online activity are usually noticed only after an examination has already been failed.

Educational data mining and learning analytics have shown that such signals can be used to predict academic outcomes early [1, 2]. Most colleges already hold the necessary data in their ERP, attendance and LMS systems, so the raw material for early prediction exists in most institutions today. The difficulty is that most published predictors are judged on accuracy alone. They return a risk label with no reason attached, so an advisor cannot tell whether a student has been flagged for poor assessment scores, low engagement or late submissions, and therefore cannot choose a suitable response. Opaque models also raise concerns about fairness, student consent and accountability when their output affects a student's academic path [3].

Explainable AI offers a way to close this gap. Explanation methods such as SHAP [4] and LIME [5] attribute a prediction to individual input features, which makes each decision open to inspection by staff and, where appropriate, by the student [6, 7]. An explanation turns a bare figure such as "risk = 0.81" into guidance that an advisor can act on, for example "no assignment submitted in weeks 1–4 and no forum activity".

This paper proposes an explainable, human-in-the-loop framework for student risk prediction, designed as the academic-risk core of a modular Smart Campus AI platform. The framework (i) predicts at-risk students early in a course using ensemble learning, (ii) attaches a SHAP or LIME explanation to every prediction, (iii) feeds predictions into a rule-based early warning system with defined escalation paths and response times, and (iv) builds consent, role-based access and bias auditing into the design. The main contributions of this work are:

1. A layered architecture that connects prediction, explanation, alerting and intervention.
2. A feature and data-governance design that excludes protected attributes from the model and uses them only for fairness auditing.
3. A reproducible evaluation protocol covering early-prediction performance, explanation quality and group fairness.

The remainder of the paper is organised as follows. Section 2 states the problem, Section 3 reviews related work and existing systems, Section 4 presents the proposed system, Section 5 describes the technical architecture and workflow, Section 6 details the system design and implementation, Section 7 presents the evaluation, Section 8 discusses ethical considerations, and Section 9 concludes the paper.

2 Problem statement

Academic risk is still identified largely through manual review of marks and attendance registers. The process is reactive and slow, and it does not scale with cohort size. By the time a student is formally flagged, typically after an internal test or a semester examination, the window for effective support has often closed. Advisors with large student loads cannot realistically follow every student's weekly behaviour, so the students who struggle quietly are the ones most likely to be missed.

Where predictive models are used, three limitations recur. First, many systems are black boxes: they output a score or label without reasons, which prevents advisors from selecting a targeted intervention and prevents students from understanding or contesting the result. Second, models are often evaluated at the end of a course, or with data that would not be available in practice, so their reported accuracy overstates their usefulness for early action. A model that is accurate only after the final examination cannot help anyone. Third, fairness and privacy are rarely addressed. Models may learn proxies for gender, caste category or economic background and thereby disadvantage groups that are already vulnerable.

A further gap is operational. A prediction is useful only if it reaches the right person, within a defined time, with a clear next step. Many prototypes stop at the prediction: they provide no escalation logic, no record of the action taken and no means of learning whether the action helped.

There is therefore a need for an integrated framework that predicts risk early using only the information available at the time, explains every prediction in human-readable terms, routes alerts to responsible staff with defined response times, and protects student privacy and fairness by design.

The absence of such a comprehensive and transparent solution motivates the framework proposed in this paper.

3 Literature review

Recent work in learning analytics and explainable machine learning informs the proposed framework, and is summarised below by theme.

Predicting academic performance. Systematic reviews report that grades, assessment behaviour and engagement logs are the most commonly used predictors of student outcomes, with tree-based ensembles and neural networks widely applied [1, 8]. Most studies emphasise accuracy and give limited attention to how staff actually use the predictions.

Early warning and learning analytics systems. Institutional deployments such as Course Signals showed that surfacing risk indicators to students and instructors can support retention [9]. Reviews of analytics for study success note, however, that the evidence on intervention effectiveness and on transparency practices remains uneven [10].

Interpretable models and post-hoc explanation. SHAP assigns each feature a consistent contribution to a prediction using Shapley values [4], whereas LIME fits a local surrogate model around an individual prediction [5]. Surveys of XAI describe the trade-off between inherently interpretable models and post-hoc explanation of complex ones, and they call for rigorous evaluation of explanations instead of the assumption that they are useful [6, 7].

Fairness in educational prediction. Surveys of bias in machine learning show that models can encode unfair patterns through proxies for protected attributes, and that fairness must be measured explicitly with group metrics such as equality of opportunity [11, 12].

Ethics and privacy in learning analytics. Work on the ethics of learning analytics highlights consent, transparency, the risk of labelling students and the need for institutional responsibility [3]. In India, the Digital Personal Data Protection Act, 2023 sets obligations on consent and data handling that apply to student data [13].

Ensemble learning for tabular data. Random Forest [14] and gradient-boosted trees such as XGBoost [15] consistently perform well on structured tabular data of the kind found in institutional records. Both also support fast, exact SHAP computation for tree models, which makes them a natural choice for explainable prediction.

3.1 Existing systems

Existing approaches fall into three broad categories. They are described in general terms and do not refer to any single product.

1. **Rule-based attendance and ERP alerts.** Many colleges use ERP or attendance software that flags students who fall below a fixed attendance or marks threshold. Such alerts are simple and transparent, but they are triggered late, consider few variables and cannot weigh combinations of signals. A student with 76% attendance and no assignment submissions is not flagged, whereas a student with 74% attendance and excellent marks is.
2. **LMS analytics dashboards.** LMS platforms provide activity summaries such as logins, time online and submissions. These support monitoring, but they typically present descriptive statistics

only, without a predictive risk estimate or reasons for the risk. Interpreting the dashboards for hundreds of students is left to the instructor.

3. **Black-box machine-learning predictors.** Research prototypes achieve good predictive performance with ensembles or deep models, but they usually offer no per-student explanation, no fairness audit and no connection to an escalation workflow. This limits their adoption in real advising practice.

4 Proposed system

The proposed system is an explainable student risk prediction framework that unifies early prediction, per-student explanation, rule-based early warning and human-led intervention in a single architecture. Unlike systems that stop at a risk label, it makes every alert explainable, auditable and tied to a named responsible person. In this way it supports early identification, transparent reasoning and timely support while protecting student privacy. Table 1 compares the design with the three categories of existing systems.

Table 1. Feature comparison (as designed).

Feature	Rule-based alerts	LMS dashboards	Black-box ML	Proposed XAI
Uses multiple data sources	Partial	Partial	Yes	Yes
Early prediction (week 4)	No	No	Partial	Yes
Predictive risk scoring	No	No	Yes	Yes
Per-student explanation	Yes (single rule)	No	No	Yes (SHAP, LIME)
Escalation and response SLA	No	No	No	Yes
Human-in-the-loop safeguards	Partial	No	No	Yes
Fairness audit	No	No	No	Yes
Student right to review	No	No	No	Yes
Low-cost open-source stack	Yes	Partial	Yes	Yes

4.1 Conceptual overview

The framework consists of five functional layers.

- **Data layer.** Collects academic records, attendance logs and LMS activity from the ERP, attendance and LMS systems, applying consent checks and data minimisation.
- **Prediction layer.** Builds features from the data available up to a chosen week and estimates the probability that a student will fail or withdraw, using Logistic Regression (baseline), Random Forest and XGBoost.
- **Explanation layer.** Generates global explanations (which factors drive risk across the cohort) and local explanations (why a particular student is flagged) using SHAP, cross-checked with LIME.
- **Early warning layer.** Combines model output with rule-based triggers, assigns a severity (Low, Medium, High or Critical) and routes alerts to the responsible person with a response time.
- **Governance layer.** Handles consent, role-based access, audit logging, bias audits and student review requests.

This layered design ensures that a prediction is never an isolated number. It is always accompanied by a reason, a severity, a responsible person and an audit trail.

4.2 Key features

- **Early risk identification.** Predictions are produced using only data available by week 4, so that support can begin while there is still time to change the outcome.
- **Per-student explanations.** Each risk score is accompanied by its top contributing factors, shown to advisors as a SHAP waterfall chart.
- **Cohort-level insight.** Global SHAP summaries show which behaviours most influence risk, which helps departments redesign courses and support services.
- **Protected attributes excluded.** Gender, disability, economic band and similar attributes are not used as model inputs. They are used only to audit fairness.
- **Human-in-the-loop action.** For High and Critical alerts the system only notifies; every decision and intervention is taken by a person.
- **Transparency to students.** Students can view their own alerts and explanations and can request a review of an assessment they disagree with.
- **Audit and accountability.** Every prediction, explanation, alert and response is logged with a timestamp.
- **Cost-effective deployment.** The system is built entirely from open-source tools that suit college budgets.

5 Technical architecture and workflow

5.1 Architecture overview

The architecture integrates data pipelines, machine-learning services and a role-based user interface in three tiers (Figure 1).

Data and integration tier. Batch and event feeds from the ERP (grades, internal assessments), attendance systems (biometric or RFID logs) and the LMS (logins, resource clicks, quiz and assignment submissions) are loaded into a central database. Records are pseudonymised, validated for missing or inconsistent values and joined at the student-course level. Only the features required by the risk model are exposed to it, in keeping with data minimisation.

Intelligence tier. A feature engineering service builds time-bounded features, for example total clicks and active days up to day 28, the number of assessments submitted, mean assessment score, number of late submissions and attendance percentage. The model service generates the risk probability. The explanation service computes SHAP values for each prediction and, for validation, LIME explanations for sampled cases. A rule engine then combines the probability with trigger rules (for example, very low attendance together with a high predicted risk) to assign a severity and an escalation path.

Presentation and response tier. Role-based dashboards present the results. Advisors and counselors see the risk level, the top contributing factors and suggested actions; department heads see aggregated cohort views; students see their own status and explanation and can request a review. Notifications are sent by email or SMS according to severity.

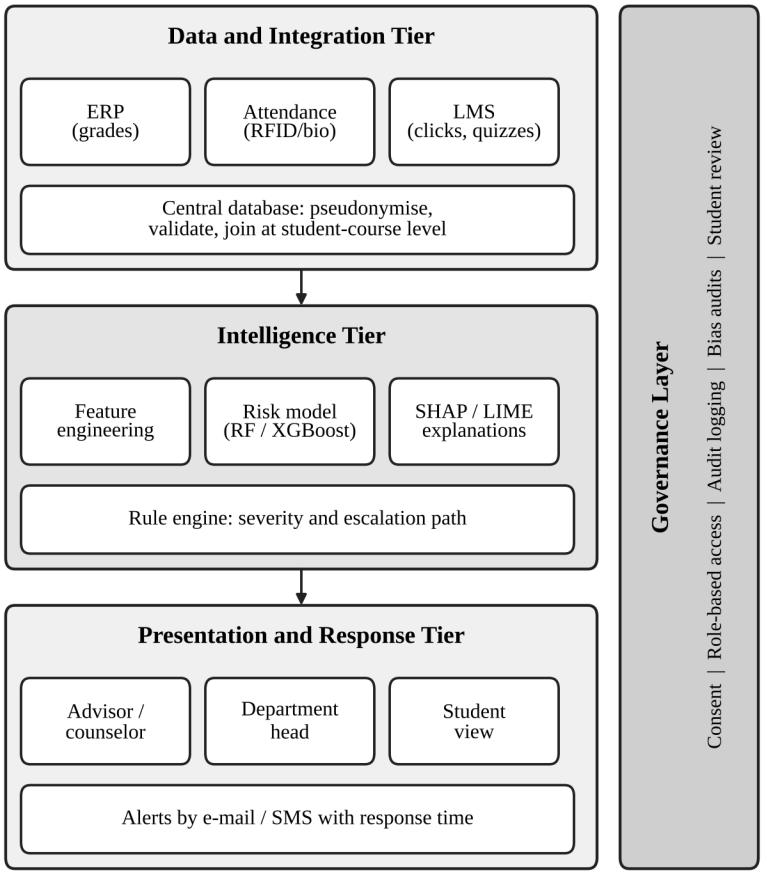


Figure 1. Layered architecture of the proposed framework, with the governance layer spanning all three tiers.

5.2 Workflow overview

The end-to-end workflow (Figure 2) involves nine steps.

1. **Data collection.** Academic, attendance and LMS data are collected from institutional systems according to each student's consent tier.
2. **Preprocessing and feature engineering.** Data are cleaned, missing values are handled, categorical variables are encoded and time-bounded features are created using only information available up to the prediction date. This prevents leakage from future events.
3. **Risk prediction.** The trained model estimates the probability that the student will fail or withdraw. Probabilities are mapped to Low, Medium, High and Critical bands.
4. **Explanation generation.** SHAP values identify the factors that raise or lower the risk for that student. The three most influential factors are shown in plain language, for example "No assignment submitted in weeks 1–4: raises risk".
5. **Rule evaluation and severity assignment.** The rule engine checks the trigger conditions (Table 3) and assigns a severity.
6. **Alert routing.** The alert is sent to the responsible person (mentor, advisor, counselor or head of department) with a defined response time. If it is not acknowledged within that time, it is escalated to a supervisor.
7. **Human review and intervention.** The staff member reviews the explanation, decides on an action (for example a mentoring session, workload review, remedial class or counseling referral) and records the decision.

8. **Outcome tracking.** The student's subsequent attendance, engagement and results are tracked to measure whether the intervention helped.

9. **Monitoring and retraining.** Data drift and prediction accuracy are monitored. Models are retrained each semester or when drift is detected, and are redeployed only after validation, fairness testing and human approval.

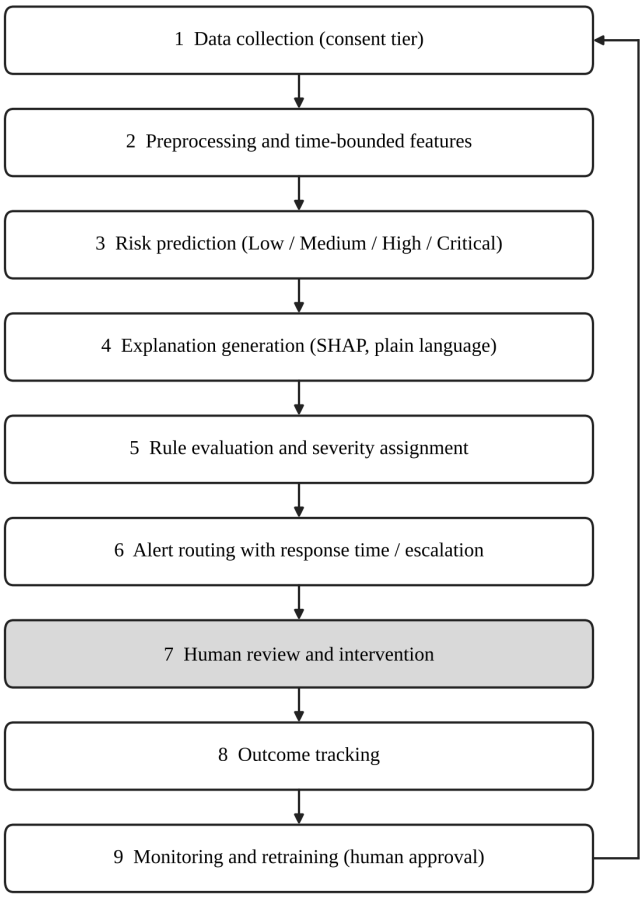


Figure 2. End-to-end workflow of the framework (steps 1–9).

This structured architecture supports early detection, explainable decision-making and a coordinated human response.

6 System design and implementation

6.1 Data design

The risk model uses three groups of features, which are summarised in Table 2.

Table 2. Feature groups used by the risk model.

Group	Example features	Source
Assessment	Number of assessments submitted, mean score so far, number of late submissions, average days between submission and deadline	ERP / LMS
Engagement	Total LMS clicks, number of active days, clicks by resource type (content, forum, quiz), longest period of inactivity	LMS

Group	Example features	Source
Course context	Number of previous attempts, credits being studied, attendance percentage, internal assessment marks (institutional deployment)	ERP / attendance

Protected attributes (gender, disability status, region, age band and economic or deprivation band) are excluded from the model inputs. They are retained separately for fairness auditing only. Because other features can act as proxies for them, proxy checks are applied, as described in Section 7.

Target definition. A student is labelled at-risk if the final course result is Fail or Withdrawn, and not at-risk if it is Pass or Distinction. This binary label turns the problem into a standard supervised classification task and matches the way institutions define unsuccessful completion.

Time-bounded feature construction. Every feature is computed only from events that occurred on or before the prediction date (for example, day 28). The rule is enforced inside the feature engineering module so that no future information, such as end-of-course marks, can leak into training or inference.

6.2 Software architecture

The software follows a modular service design in which each function (ingestion, prediction, explanation, alerting and governance) runs as an independent service and communicates through authenticated APIs. This allows individual components to be upgraded or replaced without redesigning the whole system.

6.2.1 Core platform

Data services: PostgreSQL for structured academic and attendance data, with scheduled ingestion jobs from the ERP, attendance and LMS sources. **Machine-learning services:** Python with scikit-learn and XGBoost for training, and FastAPI to serve predictions. **Explainability services:** SHAP (TreeExplainer for tree models) and LIME for local explanations. **MLOps:** MLflow for experiment tracking and model versioning, with scheduled retraining pipelines that require human approval. **User interface:** React-based web dashboards for students, advisors, counselors and administrators. **Authentication:** Keycloak (OIDC/SAML) for single sign-on and role management. **Monitoring and logging:** Prometheus and Grafana for system and model monitoring, together with centralised audit logs.

6.2.2 Key functional modules

The feature engineering module builds time-bounded features and enforces the prediction-date cut-off. The risk prediction module produces the risk probability and band. The explanation module produces global SHAP summaries and per-student SHAP waterfall explanations, with LIME cross-checks. The early warning module evaluates the trigger rules, assigns severity and routes alerts with response times. The intervention tracking module records the actions taken and the subsequent outcomes to support evaluation. The fairness audit module computes group metrics and produces periodic bias reports. The review and consent module manages consent tiers, student review requests and opt-outs.

6.3 Security and privacy measures

Because the system handles sensitive student information, security and privacy are built into the design and not added afterwards.

- **Consent tiers.** Essential data (academic records and attendance) are processed under institutional mandate, behavioural analytics are opt-out, and sensitive data such as wellness information require explicit opt-in.
- **Role-based access control.** Faculty see grades and attendance for their own classes, advisors see their assigned students, counselors see sensitive information only for their caseload, parents see summaries only with consent, and administrators see aggregates.
- **Data minimisation and pseudonymisation.** Models receive only the required features, and identities are pseudonymised in the training data.
- **Encryption and secure APIs.** Data are encrypted in transit (TLS) and at rest, and service-to-service calls are authenticated.
- **Audit logging.** All predictions, explanations, alerts and staff actions are timestamped and tamper-evident.
- **Retention limits.** Data retention follows institutional and regulatory norms, and sensitive data are deleted on schedule.
- **Student rights.** Students can view their explanations, request a review, opt out of eligible analytics and request data export or erasure, in line with the Digital Personal Data Protection Act, 2023 [13].

6.4 Alerting and response logic

The early warning layer converts model output into defined actions, with named responsible persons and response times (Table 3).

Table 3. Alerting and response logic.

Condition	System action	Responsible person / response time
Low risk (probability below the low-band cut-off)	Continue monitoring; show supportive study tips to the student	Automated; none
Medium risk	Add to advisor's priority queue; show top contributing factors	Academic advisor; within 1 week
High risk	Notify advisor with explanation; recommend mentoring or remedial support	Advisor and mentor; within 24–48 hours
Critical risk	Notify advisor and head of department; recommend urgent meeting	Advisor, HOD; within 24 hours
Predicted failure in several subjects with very low attendance	Raise "Academic Collapse" alert	Advisor, then HOD if no action in 48 hours
Prolonged absence without leave	Raise "Chronic Absence" alert	Mentor, then advisor; parent informed only with consent
Sudden drop in LMS activity after earlier engagement	Raise "Sudden Disengagement" alert	Mentor and advisor; within 48 hours

Condition	System action	Responsible person / response time
Alert dismissed repeatedly for the same pattern	Review and adjust threshold	Model owner; at next audit
Student disagrees with assessment	Open a review request and log it	Advisor or ethics committee; per policy

Note: The band cut-offs and response times are design parameters that each institution should calibrate. In this blueprint the bands are defined as Low 0–30, Medium 30–60, High 60–80 and Critical 80–100 on a 0–100 risk scale.

7 System evaluation

The evaluation protocol is designed to answer three questions: (i) how well the model predicts risk early in a course, (ii) whether the explanations are stable and faithful, and (iii) whether the model treats demographic groups fairly.

7.1 Methodology

Datasets. OULAD [16] is a public dataset from the Open University that contains demographic data, assessment results and day-by-day VLE (LMS) click activity for tens of thousands of student-module registrations. Data are limited to what is available up to day 28 of each module, to simulate week-4 prediction. The UCI Student Performance dataset [17] is an optional secondary dataset of school-level records, used to test generalisation to a different context. Institutional data from a pilot college can be added later, subject to ethics approval and consent.

Preprocessing. Duplicates are removed, missing values are handled, categorical variables are encoded, features are scaled for Logistic Regression and time-bounded features are constructed. Protected attributes are removed from the feature set.

Models. Logistic Regression serves as the baseline and is compared with Random Forest [14] and XG-Boost [15]. Hyperparameters are tuned by cross-validated search on the training data only.

Validation design. Records are split by student (grouped cross-validation) so that the same student never appears in both the training and the test set, which would leak information. Five-fold grouped cross-validation is followed by evaluation on a held-out test set. Class imbalance is handled with class weights or by oversampling inside the training folds only.

Predictive metrics. Accuracy, precision, recall, F1-score, ROC-AUC and PR-AUC are reported. Recall for the at-risk class is emphasised because missing a struggling student is costlier than raising a false alarm.

Early-prediction analysis. The best model is re-evaluated at several cut-offs (day 14, day 28, day 56 and end of course) to show how performance improves as data accumulate.

Explanation evaluation. Stability is the rank correlation of the top-k SHAP features across cross-validation folds and bootstrap samples. Fidelity is the drop in model performance when the top-k SHAP features are removed, compared with the removal of random features. Agreement is the overlap between the top-three factors from SHAP and LIME on a random sample of students. Usability, which is optional, is a short survey of advisors on whether the explanations are clear and useful.

Fairness evaluation. Using the held-out protected attributes, the true-positive-rate gap (equal opportunity difference) [12] and the positive-prediction-rate gap (demographic parity difference) are

computed across gender, disability status and deprivation band. Proxy risk is checked by testing whether the protected attributes can be predicted from the model's input features.

7.2 Results (illustrative)

The values in Tables 4–6 are illustrative and are not measured results.

Predictive performance at day 28 is shown in Table 4, the effect of the prediction cut-off in Table 5 and the fairness metrics in Table 6.

Table 4. Predictive performance at day 28 (held-out test set; illustrative values).

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.74	0.70	0.68	0.69	0.80
Random Forest	0.79	0.77	0.75	0.76	0.86
XGBoost	0.82	0.80	0.78	0.79	0.89

Table 5. Effect of prediction cut-off (best model; illustrative values).

Cut-off	F1	Recall (at-risk)	ROC-AUC
Day 14	0.68	0.64	0.82
Day 28	0.79	0.78	0.89
Day 56	0.84	0.82	0.92
End of course	0.88	0.86	0.95

Table 6. Fairness metrics (best model; illustrative values).

Group attribute	TPR gap	Positive-rate gap
Gender	0.06	0.08
Disability	0.05	0.07
Deprivation band	0.07	0.09

The platform design additionally sets two targets: a precision of at least 0.75 for high-risk students and the identification of 80% of at-risk students by week 4. These targets are to be validated in a future pilot and are not claimed here.

7.3 Discussion

The illustrative results follow the pattern that studies of this kind typically report. First, the ensemble models improve on the Logistic Regression baseline, with XGBoost giving the best balance of precision and recall. This is consistent with the strength of boosted trees on tabular data [15].

Second, the SHAP summary would be expected to show that assessment submission and early engagement dominate the prediction. This agrees with educational intuition: a student who has not submitted early work and is rarely active online is, in practice, more likely to fail or withdraw. Whether the measured results agree should be stated explicitly.

Third, performance rises as the cut-off moves later (Table 5). Week 4 offers a useful compromise, since it is early enough for support to change the outcome while already giving a clearly better signal than day 14. Fourth, the trade-off between recall and false alarms matters operationally. Higher recall catches more struggling students but sends more alerts to advisors, so the thresholds and the Medium, High and Critical bands should be tuned to advisor workload.

Fifth, fairness gaps must be reported honestly. If any group gap exceeds the acceptable threshold, the audit process should record it and trigger a threshold review, a feature review or retraining. Finally, the limitations should be stated plainly: the public datasets come from a different country and context, wellness and financial signals are not available, and the effect of interventions on student outcomes has not yet been tested.

8 Ethical considerations

Implementing an AI-driven student risk system raises important ethical and operational questions.

- **Data privacy.** Only the data needed for the risk model are used, identities are pseudonymised for training, and sensitive data require explicit opt-in. Practice follows the Digital Personal Data Protection Act, 2023 [13].
- **Fairness.** Protected attributes are excluded from the model and used only for periodic bias audits. Proxy checks and group metrics are reported so that disadvantaged groups are not unfairly flagged [11, 12].
- **Transparency and student rights.** Every alert is explained in plain language, students can see their own explanation, and they can request a review. All predictions, alerts and actions are logged to ensure accountability and traceability.
- **Avoiding labelling harm.** A risk score is a prompt for support and not a verdict. Alerts are worded supportively and are visible only to authorised staff, to reduce stigma and self-fulfilling expectations [3].
- **Human responsibility.** The system provides decision support and automated notifications but does not replace human judgement. Decisions about interventions, and any academic consequence, remain with authorised staff.
- **Reliability and governance.** Models are retrained only after validation, fairness testing and human approval, and they are monitored for data drift.

9 Conclusion

Student failure and dropout are costly, and traditional manual review identifies struggling students too late. This paper presented an explainable, human-in-the-loop framework for student risk prediction that combines early ensemble prediction (Random Forest and XGBoost), SHAP and LIME explanations, a rule-based early warning layer and a governance layer covering consent, access control and bias auditing.

By attaching a ranked list of reasons to every prediction and routing alerts to named staff with defined response times, the framework moves beyond a risk score toward actionable and accountable support. The architecture is modular, is built on open-source tools and suits college budgets.

Future work includes validating the framework on institutional data in a pilot college, measuring the real effect of interventions on retention and results, adding wellness and financial-aid signals under

explicit opt-in consent, extending the models to time-series approaches that update risk weekly, and developing a mobile application for advisors and students.

References

[1] Romero C, Ventura S. Educational data mining and learning analytics: An updated survey. *WIREs Data Min Knowl Discov*. 2020;10(3):e1355. https://doi.org/10.1002/widm.1355

[2] Baker RSJd, Yacef K. The state of educational data mining in 2009: A review and future visions. *J Educ Data Min*. 2009;1(1):3–17.

[3] Slade S, Prinsloo P. Learning analytics: Ethical issues and dilemmas. *Am Behav Sci*. 2013;57(10):1510–1529. https://doi.org/10.1177/0002764213479366

[4] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017); 2017. p. 4765–4774.

[5] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016. p. 1135–1144. https://doi.org/10.1145/2939672.2939778

[6] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[7] Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[8] Hellas A, Ihantola P, Petersen A, Ajanovski VV, Gutica M, Hynninen T, et al. Predicting academic performance: A systematic literature review. In: *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018 Companion)*; 2018. p. 175–199. https://doi.org/10.1145/3293881.3295783

[9] Arnold KE, Pistilli MD. Course Signals at Purdue: Using learning analytics to increase student success. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)*; 2012. p. 267–270. https://doi.org/10.1145/2330601.2330666

[10] Viberg O, Hatakka M, Bälter O, Mavroudi A. The current landscape of learning analytics in higher education. *Comput Human Behav*. 2018;89:98–110. https://doi.org/10.1016/j.chb.2018.07.027

[11] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv*. 2021;54(6):115. https://doi.org/10.1145/3457607

[12] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems 29 (NeurIPS 2016); 2016. p. 3315–3323.

[13] Government of India. *The Digital Personal Data Protection Act, 2023 (Act No*. 22 of 2023). New Delhi: Ministry of Electronics and Information Technology; 2023.

[14] Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324

[15] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016. p. 785–794. https://doi.org/10.1145/2939672.2939785

[16] Kuzilek J, Hlosta M, Zdrahal Z. Open University Learning Analytics dataset. *Sci Data*. 2017;4:170171. https://doi.org/10.1038/sdata.2017.171

[17] Cortez P, Silva A. Using data mining to predict secondary school student performance. In: Proceedings of the 5th Future Business Technology Conference (FUBUTEC 2008); 2008. p. 5–12.