LMS-Based Early Warning in an AI-Permissive Software Engineering Course

Andrea Vázquez-Ingelmo[1,*], Francisco José García-Peñalvo[1], Alicia García-Holgado[1] and M. Puerto Paule-Ruiz[2]

[1]Department of Computer Science, University of Salamanca, Salamanca, Spain
[2]Department of Computer Science, University of Oviedo, Oviedo, Spain

Abstract

We examine whether LMS engagement data from the first weeks of a semester can flag at-risk students in a Software Engineering course operating under an explicit AI usage policy. Using Moodle logs and first-milestone grades from two consecutive post-policy cohorts ($n = 126$), we trained classifiers to predict milestone failure from behavioural features computed over windows of 7 to 28 days. A logistic regression model on a 28-day window achieved AUC = 0.71 (sensitivity = 0.98, specificity = 0.40, 5-fold CV), approximately 27 days before the milestone deadline. Timing features (how quickly students first access resources) outperform access counts, and task-related resources are the most predictive type, while theoretical materials carry little signal. However, cross-cohort transfer is poor (AUC $\approx$ 0.52), and fail rates differ sharply between cohorts (20% vs 48%). We discuss these findings as indicative of the limited shelf-life of early-warning models when student workflows are reshaped by evolving AI tools, and suggest that AI policies should consider yearly recalibration of any associated monitoring instrument.

Keywords

early-warning systems, learning analytics, AI policy, cross-cohort generalisation, LMS engagement, software engineering education

1. Introduction

LMS-based early-warning systems are a well-established approach in learning analytics [3]. Simple engagement indicators, such as logins, resource access timing, or submission patterns, have been shown to flag at-risk students with useful lead time. However, this evidence was developed before the widespread adoption of generative AI tools in higher education. Many computing courses now operate under policies that explicitly allow AI use subject to documentation requirements [5, 6], and it is an open question whether classical LMS indicators retain their diagnostic value in this setting.

This paper presents a descriptive study of one such course: Software Engineering I at the University of Salamanca, which has operated under an explicit AI usage policy since 2023–24. We ask whether behavioural features derived from the first weeks of Moodle activity can predict first-milestone failure, and whether the resulting models transfer across consecutive cohorts. We address two research questions:

- **RQ1.** Can LMS engagement features from a short early window predict first-milestone failure within an AI-permissive course?
- **RQ2.** Does the predictive model generalise across consecutive cohorts operating under the same AI policy?

The study is deliberately small-scale and single-course. We do not estimate the effect of the AI policy itself and we do not measure individual AI tool use. The interest is practical: given a course already running under an AI policy, what can responsibly be expected from LMS-based monitoring.

LASI Spain 26: Learning Analytics Summer Institute Spain 2026, Zamora, May 07–08, 2026
*Corresponding author.
andreavazquez@usal.es (A. Vázquez-Ingelmo)

Table 1
Model comparison at 28-day window ($n = 126$, fail rate = 0.365, 5-fold CV). Operating point at Youden-J optimum.

Model	AUC	Sens.	Spec.	PPV	NPV	J
Baseline (majority)	0.50	0.00	1.00	—	0.63	0.00
Logistic Regression	**0.71**	**0.98**	0.40	0.48	**0.97**	**0.38**
Random Forest	0.66	0.91	0.38	0.46	0.88	0.29
Gradient Boosting	0.58	0.50	0.68	0.47	0.70	0.18

2. Course Context and Data

Data come from the 2023–24 and 2024–25 cohorts of Software Engineering I, a core second-year course with project-based pedagogy. Throughout, an explicit AI usage policy was in effect: students could use AI-assisted tools provided they documented prompts, reasoning, and generated outputs in their submission portfolios. The policy and assessment framework were stable across both years.

Analysis is restricted to students in the continuous-assessment modality with a recorded first-milestone grade, yielding $n = 126$ (54 in 2023–24; 73 in 2024–25). The pooled fail rate (grade < 5.0 on a 0–10 scale) is 36.5%, but differs substantially between cohorts: 20.4% in 2023–24 and 47.9% in 2024–25. Moodle logs were pseudonymised (Art. 4(5) GDPR) using salted SHA-256 hashing, in accordance with Regulation (EU) 2016/679 [1].

3. Method

Features. For each student we compute behavioural aggregates over a window of W days from the course start, per resource type (theoretical materials, videos, rubrics, tools, tasks, forums, Zoom sessions): average time to first access (ATFA), average time since last access to window end (ATLA), and total number of accesses (TNA). Resource availability is inferred from teacher access patterns since Moodle does not record publication dates. Non-access is encoded as zero. Administrative activities (modality selection polls) were excluded as they reflect enrolment logistics rather than learning engagement.

Classifiers and evaluation. The target is binary milestone failure (grade < 5). We compare a majority-class baseline, logistic regression (balanced weights), random forest (300 trees, balanced), and gradient boosting (200 trees, max depth 3). Performance is estimated by 5-fold stratified cross-validation; given the moderate sample size ($n = 126$), 5 folds were chosen as a bias–variance trade-off to ensure sufficiently large validation splits while maintaining stable estimates. We report AUC and sensitivity/specificity at the Youden-J operating point [2]. Cross-cohort transfer is tested by training on one year and evaluating on the other without retuning. We repeat the analysis for $W \in \{7, 14, 21, 28\}$ days.

4. Results

4.1. RQ1: Predictive Performance

Table 1 summarises performance at the 28-day window, which yielded the best results. Logistic regression achieves AUC = 0.71 and very high sensitivity (0.98) at the cost of low specificity (0.40). In practical terms, this model catches nearly all students who will fail the milestone but generates a substantial number of false positives. Gradient boosting trades sensitivity for specificity (AUC = 0.58, specificity = 0.68) but at a lower overall AUC.

Table 2 shows that performance improves monotonically with window length for logistic regression. A 7-day window already operates above chance (AUC = 0.57) but provides limited discrimination. The

Table 2
AUC by window length (Logistic Regression, 5-fold CV). Lead time measured from window end to H1 deadline.

Window (days)	Lead time (days)	AUC
7	~48	0.57
14	~41	0.58
21	~34	0.65
28	**~27**	**0.71**

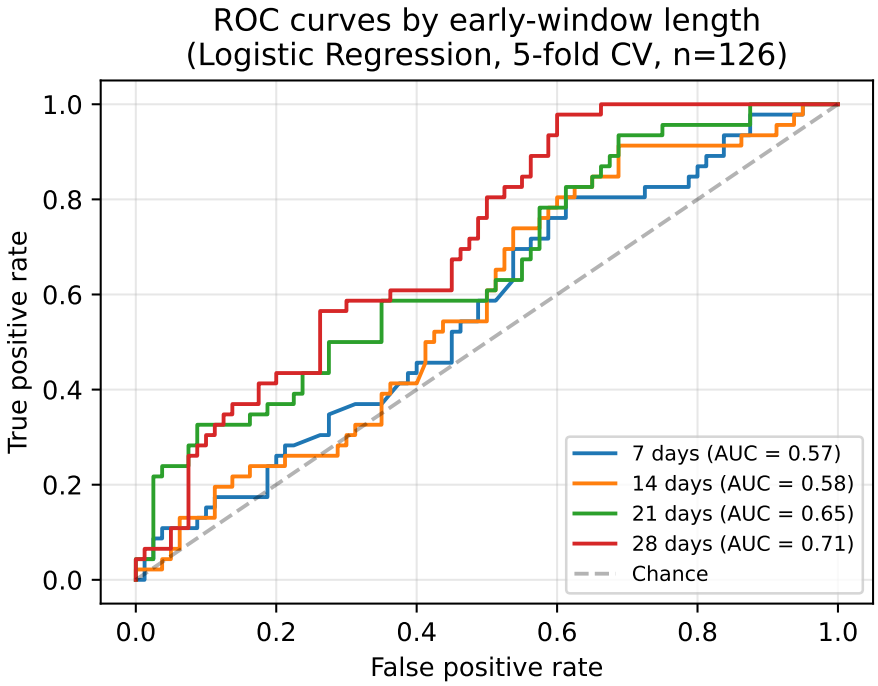

Figure 1: ROC curves by window length (Logistic Regression, 5-fold CV, $n = 126$).

Table 3
Cross-cohort transfer (Logistic Regression, 28-day window).

Condition	n_{train}	n_{test}	AUC
2023–24 → 2024–25	54	73	0.52
2024–25 → 2023–24	73	54	0.66

28-day window offers the best predictive accuracy at a lead time of approximately 27 days before the milestone deadline.

Permutation importance on the 28-day model identifies theoretical-materials timing (ATFA and ATLA) and task access timing as the dominant features. Access counts contribute comparatively little signal. Among resource types, tasks are the strongest single predictor (AUC = 0.60 in isolation), while theoretical materials, forums, and videos contribute minimally.

4.2. RQ2: Cross-Cohort Transfer

Cross-cohort transfer is poor (Table 3). Training on 2023–24 and testing on 2024–25 yields AUC = 0.52, near chance; the reverse direction gives AUC = 0.66, but this is likely driven by the base-rate shift (the 2024–25 cohort has a much higher fail rate). Within each cohort individually, 5-fold CV produces AUC values close to 0.50, indicating that cohort-specific subsamples are too small for the classifier to learn

stable patterns.

5. Discussion

Early warning is feasible but modest. A 28-day logistic regression model offers AUC = 0.71 with high sensitivity and roughly 27 days of lead time. In practice this means the model catches most students who will eventually fail the milestone, at the cost of also flagging a number of students who would have passed. This trade-off is acceptable for low-cost interventions (e.g., an email reminder or an invitation to tutoring) but not for consequential decisions. The moderate AUC is unsurprising given the small sample and the substantial difference in base rates between cohorts.

What predicts and what does not. Timing of first access is more informative than frequency of access, and task-related resources (deliverable specifications, assignment briefs) outperform theoretical materials. This is largely a calendar effect: in the first four weeks, the course is oriented toward the first team deliverable, so accessing task specifications early signals organised engagement. The limited contribution of theoretical materials could additionally reflect that some students consult AI tools for conceptual content rather than opening lecture slides on the LMS, though we did not measure AI usage and cannot confirm this.

Cross-cohort instability. Both cohorts operated under the same AI policy, with the same teaching team and materials, yet the model trained on one year shows limited transferability to the other, with performance close to chance level (AUC $\approx$ 0.52). The small per-cohort samples are an obvious contributor; any decision boundary learned from 54 or 73 students is inherently fragile. However, the marked shift in fail rates (20% to 48%) also suggests genuine year-to-year changes in student engagement, plausibly related to the rapid evolution of AI tool capabilities and students' familiarity with them between the two academic years.

A broader difficulty is that we have very limited visibility into how students actually use AI. We do not know which tools they use, how often, or for what purposes; and even the documentation required by the course policy (prompts, reasoning, outputs) is collected as part of the deliverable portfolio rather than as structured log data amenable to quantitative analysis. It is therefore hard to tell whether a student who shows low LMS activity is disengaged, or is simply doing much of their learning through AI-mediated channels that leave no trace in Moodle.

It is also plausible that students increasingly access course materials outside the LMS altogether, e.g. downloading files once and working locally, sharing resources through group chats, or asking an AI tool to explain content that they would previously have looked up in the platform. To the extent that this happens, the LMS becomes a partial and potentially shrinking window onto student behaviour, and models trained on LMS features inherit that limitation.

From a practical standpoint, this would argue for refitting early-warning models annually rather than treating them as stable instruments, and for reporting cross-cohort transfer alongside within-cohort performance so that instructors know how much to trust the predictions. Whether collecting AI usage documentation through the LMS itself (rather than in deliverable annexes) could restore some observability is an open question that goes beyond the scope of this study.

5.1. Limitations

The study uses a single course at a single institution with a small sample, which constrains generalisability. Individual AI tool usage was not measured; interpretations involving AI-mediated learning pathways are speculative. Per-cohort subsamples ($n = 54$, $n = 73$) are small for supervised learning, and the cross-cohort transfer estimates carry wide uncertainty. The administrative activity exclusion (modality-selection polls) was identified post hoc and applied uniformly, but other similar artefacts may remain undetected.

6. Conclusion

Four weeks of Moodle data can predict first-milestone failure in an AI-permissive Software Engineering course at a moderate level (AUC = 0.71), with timing of task access as the dominant signal. However, the model does not transfer across consecutive cohorts, and fail rates shift substantially between years under the same policy. These findings suggest that early-warning models in AI-permissive settings have a limited shelf-life and should be recalibrated annually. The broader implication is that AI usage policies, which typically focus on student obligations, would benefit from also addressing how institutions maintain the validity of learning monitoring over time.

Acknowledgments

This work was supported by the Ministry of Science, Innovation and Universities and co-funded by the European Regional Development Fund (ERDF) under project PID2024-155586OB-I00 (MCIU/AEI/10.13039/501100011033). Additional support was provided by the Government of the Principality of Asturias through grant GRU-GIC-24-070.

Declaration on Generative AI

During the preparation of this work, the author(s) used a large language model in order to: assist with data analysis, language polishing, and LaTeX formatting. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] European Parliament and Council, "Regulation (EU) 2016/679 (General Data Protection Regulation)," Official Journal of the European Union, 2016.

[2] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[3] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an early warning system for educators: A proof of concept," *Computers & Education*, vol. 54, no. 2, pp. 588–599, 2010.

[4] D. Gašević, S. Dawson, T. Rogers, and D. Gasevic, "Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success," *The Internet and Higher Education*, vol. 28, pp. 68–84, 2016.

[5] UNESCO, *Guidance for Generative AI in Education and Research*. Paris: UNESCO, 2023.

[6] M. Alier, F. J. García-Peñalvo, M. J. Casañ, J. A. Pereira, and F. Llorens-Largo, "Safe AI in Education Manifesto," version 0.4.0, 2024. [Online]. Available: https://manifesto.safeaieducation.org