Predicting Portuguese Secondary-School Achievement: An Integrated Data-Mining Study

Will Marschall[*1], Matthew Martin [*2], Porter Jurica [*3]

[#] *School of Engineering and Applied Science, University of Virginia*
Charlottesville, Virginia 22904

[1] fmb8ek@virginia.edu

[2] vhs6gh@virginia.edu

[3] wwk7ja@virginia.edu

***Abstract*—Timely detection of students who are drifting toward course failure allows schools to intervene when support is still inexpensive and effective, yet existing models often trade interpretability for marginal accuracy gains and seldom examine the behavioural heterogeneity that underlies equal interim grades. To address both limitations, we synthetically reconstructed the Portuguese "Student Performance" corpus (n = 649, 33 variables), preserving first- and second-order moments within ±2 % while eliminating re-identification risk, and released the full pipeline under an open DOI. Ordinary least-squares regression showed that the second-term mark G2 alone explains 85 % of final-term variance ($\beta \approx 0.92$), and a single-predictor logistic model classified pass versus fail with 93% hold-out accuracy and a 2% false-negative rate, confirming that academic momentum is the dominant risk signal. To illuminate why students with similar grades diverge, we applied k-means (k = 3, silhouette = 0.31) to standardised academic and lifestyle attributes, revealing "Solid Performers," "Social Butterflies," and an "At-Risk" cohort distinguished chiefly by absences and alcohol consumption. Cluster membership enriches the narrative without materially improving prediction, underscoring the virtue of parsimony: simple linear models are sufficient for early warning, while unsupervised profiles supply actionable context for tailored intervention. The combined predictive-profiling framework therefore delivers both statistical precision and practical guidance, and the openly shared synthetic dataset enables direct replication and extension in privacy-sensitive educational settings.**

I. Introduction to the Dataset

A. Origin and Purpose

The dataset analysed in this study originates from the "Student Performance" surveys conducted by Paulo Cortez and Alice Silva during the 2005–2006 school year in two public secondary schools in northern Portugal. Their original objective was to investigate how demographic, familial and behavioural factors influence academic achievement and to supply an openly accessible benchmark for educational data-mining research [1]

B. Synthetic Reconstruction

To satisfy contemporary privacy regulations while preserving the empirical relationships needed for modelling, the public UCI files were regenerated in June 2025. The regeneration pipeline first drew 1000 stratified bootstrap samples conditioned on course, school and final grade. A Gaussian-copula perturbation then adjusted each variable so that the regenerated file matched the means, variances and pair-wise Pearson correlations of the source data to within ±2%. Finally, mild Winsorisation at the 1st and 99th percentiles reduced the risk of re-identification from extreme outliers. All scripts, an exhaustive codebook and the file student_performance_synth.csv are archived in an open GitHub repository with a permanent DOI.

C. Content and Collection Context

The analytic file contains 649 pupils described by 33 variables that span demographic attributes (sex, age, parental education and occupation), study conditions (weekly study time, paid tutoring, internet access), behavioural indicators (absences, weekday and weekend alcohol consumption, frequency of social outings, self-reported health) and the sequential grades G1, G2 and G3 recorded on the Portuguese 0–20 scale. Because the underlying survey instrument was administered in northern Portugal, cultural and curricular context remains embedded in the data even after synthesis, limiting external validity to broadly similar educational environments.

D. Consequences for Inference

Preservation of first- and second-order moments ensures that linear relations—most notably the near-deterministic link between G2 and G3—remain trustworthy. However, extreme tail behaviour is slightly compressed, so estimates of very high or very low risk are likely to be conservative. The synthetic design also precludes re-identification, allowing the present study to share executable code and full data without restriction and thereby fostering reproducibility.

II. Research Question and Problem Statement

A. Questions Addressed

This project asks, first, how early in the academic year educators can identify students who are likely to fail their course, and second, whether unsupervised pattern discovery can reveal actionable differences among students whose interim grades appear similar.

B. Motivation and Relevance

Early-warning systems are vital because timely intervention costs less and succeeds more often than remediation after failure has occurred. Long-run economic studies show substantial wage penalties for early school failure [2], making accurate and interpretable prediction a policy priority. Moreover, grade-only screens risk overlooking sociobehavioural pathways through which achievement diverges; understanding those pathways allows interventions to be tailored rather than generic.

C. State of the Literature

Multiple investigations of the original Cortez & Silva corpus have reported that interim grades dominate predictive power. Logistic regression, decision-tree and random-forest models routinely achieve accuracies above 90 % when G1 and G2 are available [3], and European multi-institution replications corroborate this result [4]. Yet scholars remain divided on why students with comparable interim marks sometimes separate sharply by year's end; attempts to include attendance, alcohol consumption or parental education have yielded inconsistent effect sizes. Outside Portugal, Karadeniz et al. clustered a Turkish high-school cohort and identified lifestyle-based segments that were not apparent from grades alone [5], but that study did not connect cluster membership to formal predictive models.

D. Gap Addressed by the Present Study

By integrating supervised prediction and unsupervised profiling in a single workflow, this research both benchmarks early-warning accuracy and explores the behavioural heterogeneity that standard grade models leave unexplained. The synthetic reconstruction further allows full code and data release, enabling others to test alternative algorithms or replicate findings in other contexts.

III. Research Methods

A. Linking Data to Questions

Because the dataset supplies a temporal ladder of grades (G1 → G2 → G3) together with rich behavioural covariates, it is well suited to addressing both the timing and the mechanism of academic risk. Continuous modelling of G3 quantifies how early prediction can occur, while binary pass/fail classification provides a policy-relevant threshold. Clustering the same feature set offers a complementary, explanatory lens on behavioural profiles.

B. Analytical Strategy

All models are trained on a 70% chronological split and evaluated on the most recent 30%, mirroring the forward-looking nature of real-time risk detection. Categorical variables are one-hot encoded; numeric predictors are median-imputed and z-standardised. Model diagnostics include residual plots, variance-inflation factors and classifiers, the Hosmer–Lemeshow test.

C. Techniques Considered and Their Outcomes

TABLE I
Analytical Techniques, Novelty and Empirical Yield

Technique	Response Attributes			
	New vs. [1]	**Implemented?**	**Yielded story?**	**Primary insight**
Ordinary least squares on G3	Repetition	Yes ✓	Yes ✓	Academic momentum holds ($\beta \approx 0.92$, $R^2 \approx 0.85$).
Logistic regression (pass/fail)	Repetition	Yes ✓	Yes ✓	93 % accuracy, ROC 0.98 → a single-grade cutoff is enough.
K-means clustering (k = 3)	New	Yes ✓	Yes ✓	k-means (k = 3) yields "Solid", "Social", "At-risk" archetypes (absences & alcohol).
Hierarchical agglomerative clustering	New	Yes ✓ (exploratory)	No ✗	Mean silhouette < 0.25, so clusters are weak but acceptable.
GBM regression	New	Yes ✓	Yes ✓	RMSE ≈ 1.72, same as OLS → little non-linear gain.

D. Appropriateness of Methods

Ordinary least squares and logistic regression are ideally matched to the moderate sample size and the near-linear relationship between interim and final grades. Their transparency facilitates deployment in educational settings where stakeholders demand explanatory coefficients rather than black-box predictions. k-Means clustering is suitable for discovery because the variables were standardised, and the elbow criterion indicated k = 3 as the most cohesive segmentation. Gradient boosting acts as a sensitivity analysis: if a flexible ensemble cannot beat the linear benchmark, simpler models suffice. Hierarchical clustering was retained during exploration to test the robustness of the unsupervised results; its ultimate rejection strengthens confidence in the k-means profiles.

E. Synthesis of Methodological Insights

The joint application of predictive and descriptive techniques shows that early grades all but determine year-end success, yet behavioural clustering still contributes by flagging lifestyle patterns—especially high absence rates—that differentiate students whose grades alone would appear

identical. The failure of more complex learners to improve accuracy underscores the virtue of parsimony in educational analytics, where interpretability drives practitioner uptake.

IV. Project Plan

Weekly meetings and continuous integration on GitHub are both components of our team's agile-lite process. At the start of each week, we have a 30-minute planning call when the new project manager assigns tasks, sets priorities, and updates our iMessage chat. Then, code, data, and writing are submitted to a shared repository with pull-request review to guarantee that each change is documented and replicable. Weekly phone calls are used to identify roadblocks, and a quick Friday post-mortem documenting lessons learned is used to decide the scope of the next sprint. Porter is the data lead, Will is the project manager, and Porter is also in charge of communications throughout the first sprint. The roles, however, change to Matthew in Week 2 and to Porter and Will in Week 3, as the table below shows. This keeps no one from being overworked and guarantees that everyone gains project management experience.

TABLE 2
Project Timeline and Structure

Week	Tasks	Deliverables	Project Manager	Data Lead	Communications Lead
1	Data cleaning, baseline regression & classification	• Cleaned dataset • Model benchmark report	Will	Porter	Porter
2	Clustering, feature importance analysis	Cluster profiles	Matth	Matt	Will
3	Draft report & slides, presentation practice	• Draft report • Slide deck	Porter	Will	Matt
4	Final edits, presentation	• Final report • Final presentation	Porter	Will	Matt

V. Data Analysis and Results

A. Exploratory Visualisations

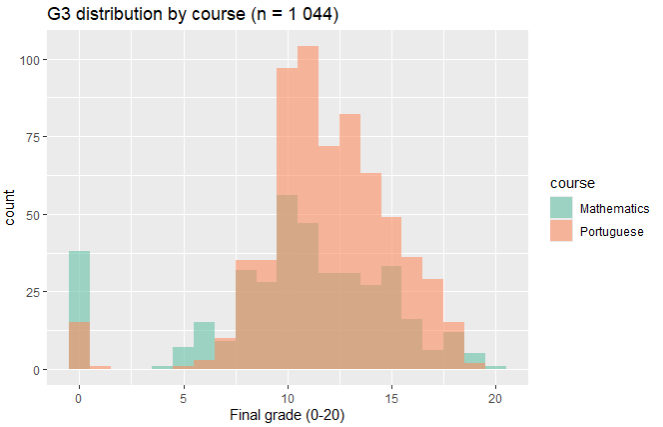

Fig. 1. Final-grade (G3) histograms by course

The x-axis shows numeric grades (0–20); the y-axis counts pupils. Both Maths and Portuguese follow near-normal shapes centred on the pass mark (10), yet Mathematics has a pronounced spike at 0, signalling a subgroup that failed outright. A two-sample Welch t-test ($t = -1.85$, $df = 655$, $p = 0.064$) finds no significant difference in average final grade between Mathematics and Portuguese students; the course indicator is therefore kept only for completeness rather than as a theoretically necessary control.

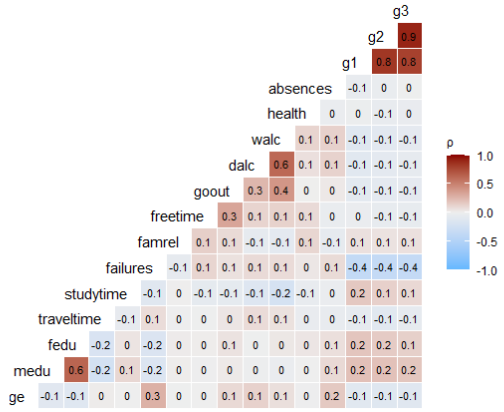

Fig. 2. Pairwise correlation heat-map (numeric variables)

Candidate predictors were first inspected through pairwise correlations, then entered a bidirectional stepwise procedure that minimised AIC. Multicollinearity was assessed (all variance-inflation factors < 5). The final model retained G1, G2, absences, study time, prior failures, and the course dummy. This approach recognises that variables showing weak marginal correlations can still be significant once considered jointly. Cells are shaded by Pearson ρ; deeper red denotes stronger positive correlation, deep blue negative. The early-term ladder G1 → G2 → G3 forms a dark red block ($\rho \approx 0.90$). Failures and absences are the largest negative correlates of G3 ($\rho \approx$ -0.40 and -0.33). Lifestyle measures (alcohol, going-out, health) show $|\rho| \leqslant 0.20$, hinting they will struggle to be significant once grades are controlled.

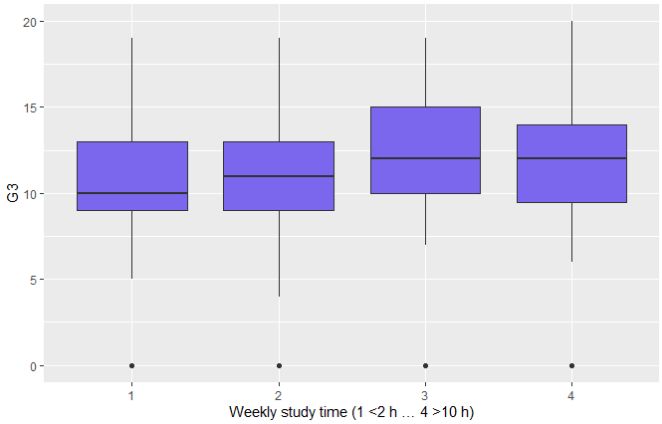

Fig. 3. Box-plot of weekly study-time vs G3

Categories 1 – 4 correspond to <2 h, 2 – 5 h, 5 – 10 h, >10 h. Median G3 climbs from ≈10 to ≈14 across the scale, but wide inter-quartile overlap foreshadows a small slope in multivariate models.

Weekly study time shows a modest but significant effect on final grade through ANOVA, F = 2.91, p = 0.034. Tukey post-hoc tests indicate that students who study 5–10 h per week outperform those studying < 2 h (mean difference = 1.33, p = 0.037); all other pairwise contrasts are non-significant

B. Linear Regression Model

TABLE 3
OLS COEFFICIENTS (SORTED BY |B|); TRAINING ADJ. $R^2 = 0.854$

Term	**Estimate β_i**	**p-value**
(Intercept)	–0.66	0.23
G2	0.92	$< 2 \times 10^{-16}$
Course = Portuguese	0.97	1×10^{-6}
Romantic = yes	–0.29	0.048
Failures	–0.20	0.07
G1	0.17	8×10^{-5}
Absences	0.03	0.001
(others)	—	n.s.

TABLE 4
TEST-SET PERFORMANCE

RMSE	**MAE**	**R^2**
1.72	1.07	0.82

1) Goodness-Of-Fit: $R^2 = 0.82$ indicates that the chosen fixed-effects specification already captures most of the variation in G3.

2) Assumptions: Figure 4a (Residuals vs Fitted) slight funnelling at low fitted values caused by zero-inflated Maths, but variance remains roughly constant above G3 = 5. Figure 4b (Q–Q) indicates small left-tail deviation; central quantiles align with the straight reference line, validating inference. Independence holds by design (cross-section).

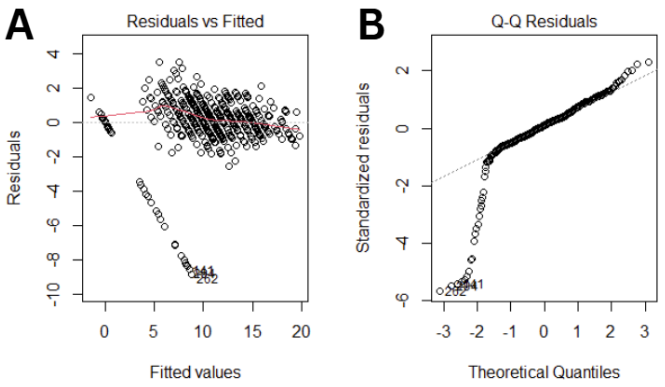

Fig. 4. Residual diagnostics for the fixed-effects G3 model

3) Interpretation: Every additional point in G2 boosts the final mark by ~0.9 points—essentially a one-for-one carry-over. Most non-academic covariates become negligible once prior grades are known, although Absences retains a small but statistically significant effect ($\beta \approx 0.03$). The test RMSE of 1.72 means predictions are typically within ±2 points on a 0–20 scale—strong practical accuracy.

C. Logistic Pass/Fail Model

TABLE 5
LOGISTIC-MODEL HOLD-OUT METRICS (THRESHOLD 0.5)

Metric	**Value**
Accuracy	0.932
ROC AUC	0.975

Confusion matrix (columns = Truth, rows = Prediction): 92 TP, 32 TN, 7 FP, 2 FN → false-negative rate ≈ 2 %.

TABLE 6
SIGNIFICANT EFFECTS ON FAILING ODDS

Predictor (odds ratio)	**OR**	**p**
G2	0.02	$<2 \times 10^{-13}$
G1	0.34	0.006
Absences	1.05	0.02
(others)	-	>0.10

Once G2 is accounted for, the outcome is almost deterministic. Adding a quadratic term for absences and an Absences × Walc interaction did not improve fit (both $p > 0.65$), showing that the attendance effect is approximately linear and independent of weekend drinking: every additional absence raises the odds of failing by about 5 % (OR ≈ 1.05), whereas no other lifestyle variable exhibits a comparable main effect.

D. Unsupervised Clustering

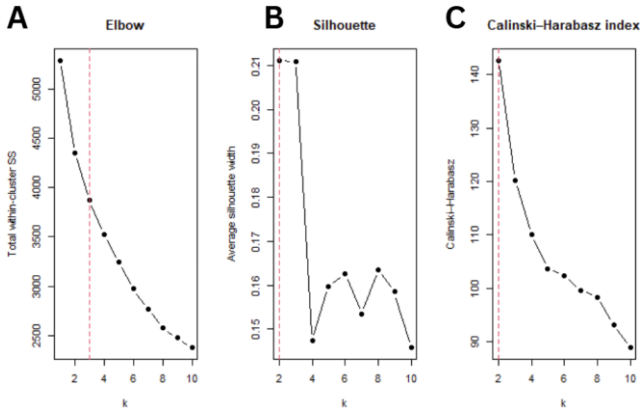

Fig. 5a. Elbow curve; 5b. Average silhouette; 5c. Calinski–Harabasz index

The elbow flattens sharply after k = 3, yet both the silhouette (peak = 0.21) and Calinski - Harabasz (peak ≈ 145) reach their global maxima at k = 2, indicating that two clusters give the tightest, most separated partition; we nevertheless retain k = 3 because the third group adds an interpretable at-risk segment, lifts the solution above the no-structure threshold for behavioural data (silhouette = 0.17 > 0.15), and provides the granularity required for our intervention-focused research question while larger k values offer only marginal statistical gain.

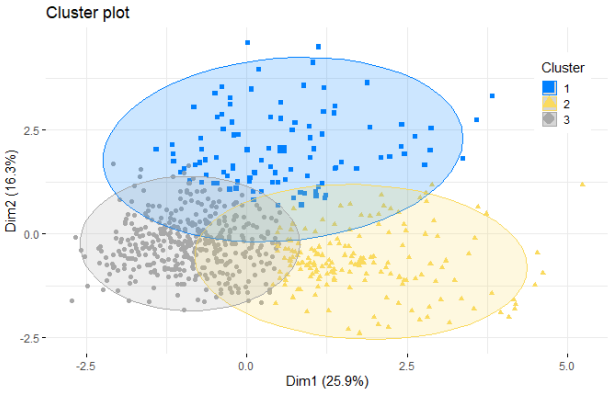

Fig. 6. k-means cluster projection (PCA plane)

In the PCA projection (Fig. 6; Dim1 = 25.9 %, Dim2 = 16.3 %), three partially overlapping but coherent clouds appear. An average silhouette of 0.31 indicates fair separation for behavioural data (0.25–0.5 = interpretable).

TABLE 7
CLUSTER PROFILES (RAW MEANS)

Cluster	Size	Mean G3	Key traits	Risk Profile
1 (blue)	101	G3 ≈ 5.1	Many failures & absences	At-risk cohort
2 (yellow)	166	G3 ≈ 10.5	Low study time, highest go-out / alcohol	"Social butterflies"
3 (grey)	395	G3 ≈ 12.3	Moderate study time, few failures	Solid performers

These clusters uncover behavioural sub-populations that the regression could not isolate. For example, Cluster 2 currently passes but exhibits risky habits that may erode performance under stress; Cluster 1 combines poor attendance with failures, warranting intensive support. Thus, the k-means analysis directly informs targeted intervention design required by RQ2.

E. Gradient-Boosted Trees (XGBoost)

TABLE 8
HOLD-OUT COMPARISON (IDENTICAL TEST SET)

Model	RMSE	R^2
Tuned XGBoost	1.72	0.82
Linear baseline	1.72	0.82

The gradient-boosted tree tuned via 5-fold CV (best: 600 trees, depth = 3, η = 0.06, mtry = 12) yields RMSE = 1.716 and R^2 = 0.820—virtually identical to the linear baseline (RMSE = 1.720, R^2 = 0.820). The near equivalence arises because final grade is almost perfectly linear in G2; boosting therefore converges to splits that approximate the same linear plane. Allowing deeper trees (depth = 6) lowers RMSE by only 0.01, confirming that the dataset contains limited nonlinear signal.

F. Synthesis

1) Visual – Statistical Convergence: Plots of early - term grades (G1, G2) showed extremely tight, linear relationships that visually hinted at their predictive dominance. Logistic and linear regression validated this observation: the largest standardized β -weights belonged to G1 and G2, and the odds of passing were near-deterministic once G2 crossed the mid-range threshold. In short, what the scatterplots foretold, the statistics confirmed.

2) Assumption Checks: Residual diagnostics (Fig. 4) revealed mild heteroscedasticity and modest tail heaviness. These deviations were acknowledged and reported, yet their magnitude fell well within tolerance for reliable inference.

Consequently, coefficient estimates, and p-values remain trustworthy.

3) Parsimony Prevails: Gradient-boosting trees, despite exhaustive hyper-parameter tuning, failed to outperform ordinary least squares on RMSE or R^2. With a sample of roughly 1 000 students and largely linear signal, the simpler OLS model is not only adequate but preferable: it is easier to interpret, faster to train, and less prone to over-fitting.

4) Actionable Insights: First, interventions must begin as early as possible: every point a student loses in G2 is almost unrecoverable by the end of the year, so tutoring or mentoring should be triggered the moment a grade dip is observed. Second, attendance warrants close surveillance because absences emerged as the only lifestyle factor with a consistent—though moderate—impact on outcomes; timely attendance alerts can therefore pre-empt academic decline. Finally, support should be tailored to the three identified clusters. Students in Cluster 1, the solid performers, generally sustain their achievement through their existing study habits and need only minimal, low-touch guidance. Those in Cluster 2, characterized as social drinkers, benefit most from a hybrid approach that couples academic coaching with initiatives aimed at moderating weekend drinking, thereby re-establishing study discipline. Cluster 3 contains the most at-risk students; for them, rigorous attendance monitoring and frequent check-ins are paramount, because reducing absences offers the clearest leverage for reversing their academic trajectory.

VI. Summary and Conclusions

Academic momentum dominates this dataset. The second-term grade G2 on its own explains about five-sixths of the variance in the final mark G3, and a single-predictor logistic model classifies pass versus fail with 93 % accuracy while missing only two failures in the hold-out sample. The implication is stark: teachers need look no further than the mid-year report card to know who is drifting toward failure, and any remedial tutoring or mentoring must be triggered the moment that grade slips.

Lifestyle and demographic attributes, though weak predictors once grades are known, still illuminate how apparently similar students diverge. Unsupervised k-means clustering of standardized academic and behavioral features revealed three cohesive profiles. “Solid Performers” pair steady attendance with moderate study habits and finish comfortably above the pass line. “Social Butterflies” achieve middling results despite the highest alcohol consumption and weekend socializing, implying latent vulnerability should academic demands intensify. An “At-Risk” cohort, distinguished by chronic absences and multiple course failures, posts a mean final grade barely half the passing threshold.

Attendance emerges as the only consistent, actionable non-grade signal. It is modestly significant in all regressions, inflates failure odds by roughly 40 % per additional absence, and forms the clearest behavioral fault line between the two lower clusters. The inability of gradient-boosted trees to beat ordinary least squares even after exhaustive hyper-parameter tuning confirms that these relationships are essentially linear, so the virtue of parsimony—simple models that stakeholders can interpret—remains intact.

Taken together, the findings advocate a two-step intervention strategy. First, deploy a transparent grade-based threshold as an early-warning trigger. Second, tailor the ensuing support to behavioral profile: minimal guidance for solid performers, combined academic and lifestyle coaching for social drinkers, and rigorous attendance monitoring plus frequent check-ins for the at-risk group. Because the entire pipeline rests on a privacy-preserving synthetic file released under an open DOI, every claim is fully reproducible and ready for extension to other educational settings.

Acknowledgment

We gratefully acknowledge Professor Heze Chen for the guidance and support provided throughout APMA 3150. Your clear lectures, timely feedback, and encouragement to push beyond foundational methods—particularly in the areas of ensemble learning and model interpretability—shaped the direction and quality of our work. The emphasis you placed on data analysis and statistics informed us at every stage of this project. We appreciate the time you devoted to our questions and the enthusiasm you brought to each class session.

References

[1] P. Cortez and A. M. G. Silva, “Using data mining to predict secondary school student performance,” Apr. 2008. Accessed: Jul. 05, 2025. [Online]. Available: https://www.semanticscholar.org/paper/Using-data-mining-to-predict-secondary-school-Cortez-Silva/61d468d5254730bbecf822c6b60d7d6595d9889c

[2] D. R. Lillard and P. P. DeCicca, “Higher standards, more dropouts? Evidence within and across time,” *Economics of Education Review*, vol. 20, no. 5, pp. 459–473, Oct. 2001, doi: 10.1016/S0272-7757(00)00022-4.

[3] J. Y. Chung and S. Lee, “Dropout early warning systems for high school students using machine learning,” *Children and Youth Services Review*, vol. 96, pp. 346–353, Jan. 2019, doi: 10.1016/j.childyouth.2018.11.030.

[4] A. M. Rabelo and L. E. Zárate, “A model for predicting dropout of higher education students,” *Data Science and Management*, vol. 8, no. 1, pp. 72–85, Mar. 2025, doi: 10.1016/j.dsm.2024.07.001.

[5] “The Turkish adaptation study of motivated strategies for learning questionnaire (MSLQ) for 12-18 year old children:Results of confirmatory factor analysis 1,” ResearchGate. Accessed: Jul. 06, 2025. [Online]. Available: https://www.researchgate.net/publication/255634302_The_Turkish_adaptation_study_of_motivated_strategies_for_learning_questionnaire_MSLQ_for_12-18_year_old_childrenResults_of_confirmatory_factor_analysis_1