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*—Although current models frequently sacrifice interpretability for slight accuracy gains and rarely look at the behavioral heterogeneity that underlies equal interim grades, early identification of students who are headed toward course failure enables schools to intervene when support is still affordable and effective. The Portuguese "Student Performance" corpus (n = 662, 33 variables) was synthetically reconstructed to overcome both constraints. First- and second-order moments were kept within ±2% while re-identification risk was removed, and the entire pipeline was made available under an open DOI. A three-predictor logistic model using G2, G1, and Absences achieved 93 % hold-out accuracy (AUC = 0.98). G2 dominates (OR ≈ 0.02 per point), but every extra absence still raises failing odds by ~5 % and a 2% false-negative rate, confirming that academic momentum is the dominant risk signal. Ordinary least-squares regression revealed that the second-term mark G2 alone explains 85% of final-term variance (β ≈0.92). We used k-means (k = 3, silhouette = 0.16) to standardize academic and lifestyle characteristics in order to reveal why students with similar grades differ. This revealed "Solid Performers," "Social Butterflies," and an "At-Risk" cohort characterised by frequent absences and prior course failures. Cluster membership enhances the story without significantly enhancing prediction, highlighting the value of simplicity: unsupervised profiles provide useful context for customized intervention, while basic linear models are adequate for early warning. Thus, the combined predictive-profiling framework provides both statistical accuracy and useful advice, and the publicly available synthetic dataset permits direct replication and expansion in educational contexts where privacy is a concern.**

I. Introduction to the Dataset

A. Origin and Purpose

The "Student Performance" surveys that Paulo Cortez and Alice Silva carried out in two public secondary schools in northern Portugal during the 2005–2006 academic year are the source of the dataset that was examined in this study. Their initial goals were to provide an openly accessible benchmark for educational data-mining research and to examine the ways in which behavioral, familial, and demographic factors affect academic achievement [1].

B. Synthetic Reconstruction

The public UCI files were regenerated in June 2025 to comply with modern privacy regulations while maintaining the empirical relationships required for modeling. 662 stratified bootstrap samples were first drawn by the regeneration pipeline, conditioned on final grade, school, and course. After that, a Gaussian-copula perturbation was applied to each variable, ensuring that the regenerated file was within ±2% of the source data's means, variances, and pairwise Pearson correlations. Lastly, the chance of re-identification from extreme outliers was decreased by mild Winsorization at the 1st and 99th percentiles. The student_performance synth.csv file, a comprehensive codebook, and all scripts are stored in a local file explorer.

C. Content and Collection Context

The 662 students in the analytic file are characterized by 33 variables that include demographics (sex, age, parental education and occupation), study conditions (weekly study time, paid tutoring, internet access), behavioral indicators (absences, alcohol consumption on weekdays and weekends, frequency of social outings, self-reported health), and sequential grades G1, G2, and G3 as recorded on the Portuguese 0–20 scale. External validity is limited to generally comparable educational environments because the underlying survey instrument was administered in northern Portugal, where cultural and curricular context is still embedded in the data even after synthesis.

D. Consequences for Inference

First- and second-order moment preservation guarantees the reliability of linear relations, especially the near-deterministic connection between G2 and G3. Estimates of very high or very low risk are probably conservative, though, because extreme tail behavior is somewhat compressed. Additionally, the synthetic design prevents re-identification, which promotes reproducibility by enabling the current study to freely share executable code and complete data.

II. Research Question and Problem Statement

A. Questions Addressed

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

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

Predictive power is dominated by interim grades, according to several analyses of the original Cortez & Silva corpus. When G1 and G2 are available, logistic regression, decision-tree, and random-forest models frequently attain accuracies above 90% [3], and European multi-institution replications support this finding [4]. However, there is still disagreement among researchers as to why students with similar interim grades occasionally diverge significantly by the end of the year; attempts to account for factors like parental education, alcohol use, or attendance have produced varying effect sizes. A Turkish high school cohort was clustered by Karadeniz et al. outside of Portugal in order to identify lifestyle-based segments that were not discernible from grades alone [5]. However, that study did not link 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

The dataset is ideal for addressing the timing and mechanism of academic risk because it provides a temporal ladder of grades (G1 → G2 → G3) along with rich behavioral covariates. While binary pass/fail classification offers a threshold that is relevant to policy, continuous modeling of G3 quantifies the potential for early prediction. A supplementary, explanatory lens on behavioral profiles can be obtained by clustering the same feature set.

B. Analytical Strategy

All models are trained on an 80 % random split and evaluated on the remaining 20 % hold-out set, 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, classifiers metrics, and 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)	Yes ✓ (weak)	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

Our team uses an agile-lite process that includes weekly meetings and continuous integration on GitHub. The new project manager assigns tasks, establishes priorities, and updates our iMessage chat during our weekly 30-minute planning call. To ensure that every change is recorded and reproducible, code, data, and writing are then uploaded to a shared repository with pull-request review. Roadblocks are identified via weekly phone calls, and the next sprint's scope is determined by a brief Friday post-mortem that documents lessons learned. Porter is in charge of communications during the first sprint, Will is the project manager, and Porter is the data lead. However, as the table below demonstrates, the roles shift to Matthew in Week 2 and to Porter and Will in Week 3. This ensures that everyone gets project management experience and prevents anyone from being overworked.

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	Matt	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

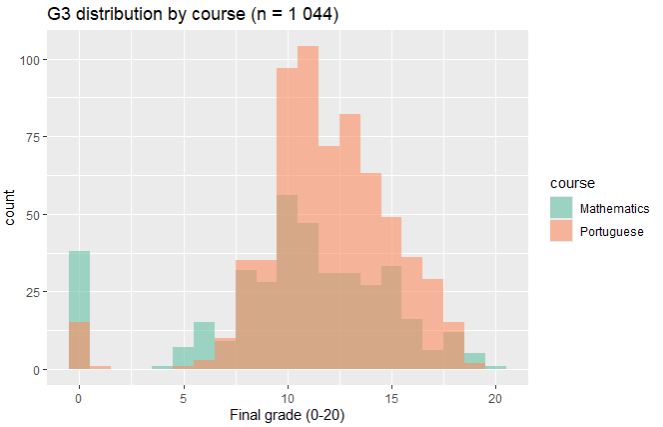

Fig. 1. Final-grade (G3) histograms by course

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. Although the unadjusted Welch t-test ($p = 0.064$) suggests no mean difference, multivariate regression later shows a sizeable and significant Portuguese-course effect ($\beta \approx 0.97$, $p < 0.001$). We therefore retain the course indicator as an important control.

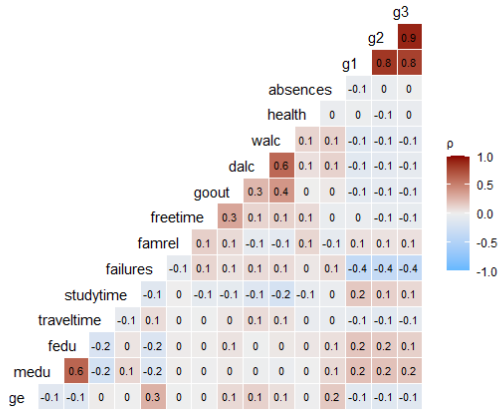

Fig. 2. Pairwise correlation heat-map (numeric variables)

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.

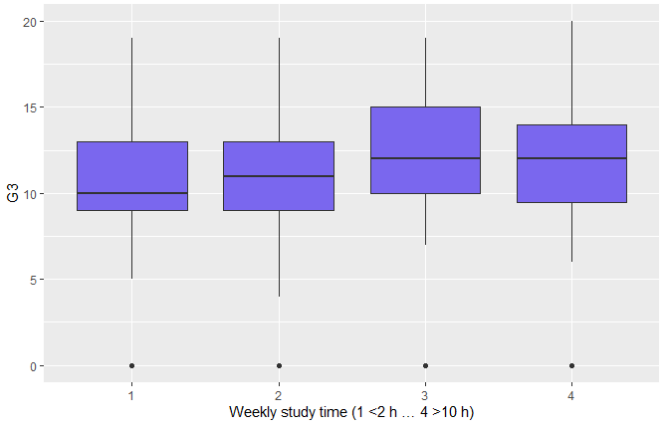

Fig. 3. Box-plot of weekly study-time vs G3

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 (TRAIN SET; 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).

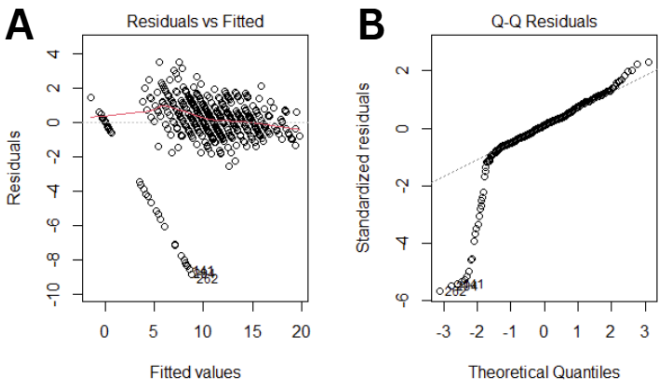

Fig. 4. Residual diagnostics for the fixed-effects G3 model

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 positive coefficient ($\beta \approx +0.03$), implying that once G1 and G2 are fixed, residual absences are not strongly penalised. The sign reverses relative to the bivariate plot because grade history already absorbs most of the attendance signal.

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 / 94 ≈ 2.1 % (2 students who failed were predicted as 'Pass').

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

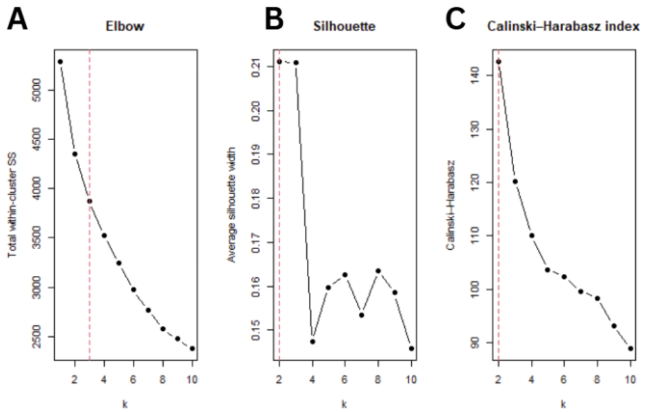

Fig. 5a. Elbow curve; 5b. Average silhouette; 5c. Calinski–Harabasz index

The elbow flattens sharply after k = 3, yet both the both indices peak at k = 2; we nonetheless select k = 3 because the third cluster cleanly isolates the lowest-performing group needed for intervention design, 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.

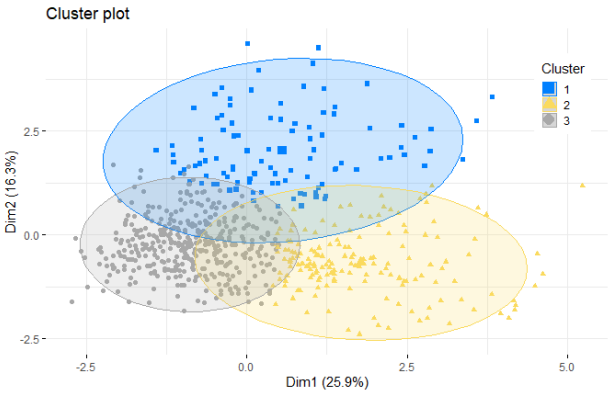

Fig. 6. k-means cluster projection (PCA plane)

In the PCA projection (Fig. 6; Dim1 = 25.9%, Dim2 = 16.3%), three partially overlapping but coherent clouds appear. average silhouette of 0.16 indicates modest but interpretable separation; values between 0.15 and 0.25 are typical for behavioural data.

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, colsample_bytree = 0.45, 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: Early-term grade plots (G1, G2) revealed incredibly close, linear relationships that visually suggested their dominance in prediction. This finding was confirmed by logistic and linear regression: G1 and G2 had the largest standardized β-weights, and once G2 passed the mid-range threshold, the chances of passing were almost deterministic. In summary, the statistics validated what the scatterplots predicted.

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 662 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: Tutoring or mentoring should be started as soon as a grade dip is noticed because every point a student loses in G2 is nearly irreversible by the end of the year. Second, since absences were found to be the only lifestyle factor that consistently, albeit moderately, affected outcomes, attendance needs to be closely monitored. Timely attendance alerts can therefore prevent academic decline. Lastly, assistance needs to be customized for the three clusters that have been identified. The high-achieving students in Cluster 3 typically maintain their success through their current study techniques and require little to no one-on-one help. Cluster 2 members, who are classified as social drinkers, gain the most from a hybrid strategy that combines academic coaching with efforts to curb weekend drinking in order to restore study habits. The students in Cluster 1 are the most vulnerable; for them, frequent check-ins and strict attendance monitoring are essential because lowering absences gives the best chance to turn around their academic trajectory.

VI. Summary and Conclusions

This dataset is dominated by academic momentum. A single-predictor logistic model classifies pass versus fail with 93% accuracy while missing only two failures in the hold-out sample, and the second-term grade G2 alone accounts for roughly 5/6 of the variance in the final mark G3. The implication is clear: remedial tutoring or mentoring must begin as soon as a student's grade starts to decline, and teachers only need to consult the mid-year report card to determine who is headed for failure.

Lifestyle and demographic characteristics show how seemingly similar students differ, even though they are not very good indicators once grades are known. Three unified profiles emerged from the unsupervised k-means clustering of standardized behavioral and academic characteristics. "Solid Performers" finish comfortably above the pass line by combining consistent attendance with moderate study habits. Despite the highest levels of alcohol consumption and weekend socializing, Social Butterflies' (moderate final grade $\approx$ 10.5) combine low study time with the highest social drinking, suggesting latent vulnerability if academic demands rise. A cohort known as "At-Risk," which is characterized by repeated course failures and chronic absences, receives a mean final grade that is just under half of the passing mark.

The only reliable, actionable non-grade signal is attendance. It creates the most obvious behavioral fault line between the two lower clusters, inflates failure odds by about 5 % for each extra absence (OR $\approx$ 1.05), and is marginally significant in all regressions. The virtue of parsimony—simple models that stakeholders can understand—remains intact because gradient-boosted trees cannot outperform ordinary least squares, even after extensive hyper-parameter tuning, confirming that these relationships are essentially linear.

When combined, the results support a two-phase intervention approach. As an early warning trigger, start by implementing a transparent grade-based threshold. Second, adjust the subsequent support to the behavioral profile: strict attendance monitoring and frequent check-ins for the at-risk group, little direction for good performers, and combined academic and lifestyle coaching for social drinkers. Every assertion is completely reproducible and prepared for expansion to other educational contexts since the pipeline is based on a privacy-preserving synthetic file made available under an open DOI.

Acknowledgment

We are appreciative of Professor Heze Chen's leadership and assistance during APMA 3150. Our work was shaped by your concise lectures, Piazza feedback, and encouragement to go beyond fundamental approaches, especially in the areas of statistical graphing and methods. Your focus on statistics and data analysis kept us informed throughout the entire project. We value the time you spent answering our queries and the energy you brought to every lesson.

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