Predicting Academic Outcomes in Secondary Education: Ensemble Classification with Grade Trajectories, Attendance Behaviour, and Socioeconomic Context

Jehad Mousa[1,2,*], **Abdallah Salama**[3]
[1]University of Dubai, UAE
[2]United Arab Emirates University, UAE
[3]Assistant Professor in Sociology, City University Ajman, Ajman, UAE
Emails: Jehadgmousa@gmail.com; a.adel@cu.ac.ae

Abstract

Early identification of students at risk of academic failure is a persistent challenge in educational technology, with direct implications for student retention, institutional equity, and the allocation of support resources. Although supervised machine learning has been widely applied to student outcome prediction, the relative merit of competing algorithm classes and the degree to which demographic and behavioural features contribute predictive power beyond prior academic assessments remain incompletely resolved in the secondary school context. This paper presents a structured comparative evaluation of five supervised classifiers trained on a rich combination of periodic grades, attendance records, socio-demographic characteristics, and lifestyle indicators drawn from secondary school students. A dual importance analysis—combining impurity-based measures with held-out permutation importance—disentangles the distinct predictive roles of grade trajectories, absenteeism, parental background, and lifestyle variables. Ensemble methods demonstrate consistent superiority across all evaluation criteria, with prior periodic assessments and attendance emerging as the dominant predictors. Parental education level introduces a socioeconomic gradient that operates independently of student-controlled factors, generating structural inequities that standard grade-monitoring systems are unlikely to address. These findings provide both a methodological benchmark for secondary school prediction tasks and practical guidance for institutions designing equitable and evidence-based early warning interventions.

Keywords: Educational data mining; Machine learning; Student outcome prediction; Ensemble methods; Learning analytics; Secondary education; Early warning systems

1. Introduction

Academic underperformance and course failure in secondary education generate costs that extend far beyond the individual student. Unremedied failure reduces course completion rates, narrows pathways to higher education, and carries well-documented consequences for long-run employment prospects and social mobility. Despite growing awareness of these stakes, most secondary schools continue to rely on end-of-term grades as their primary diagnostic signal, an inherently reactive posture that leaves limited lead time for meaningful intervention once an academic outcome has effectively been determined.

The consolidation of educational data mining (EDM) and learning analytics (LA) as research disciplines has reframed this problem [8]. Student records now routinely combine demographic profiles, attendance histories, family background data, and periodic assessment results. Analysed jointly, these sources carry substantially more predictive power than any single indicator alone, and the core premise of predictive modelling in education—that academic failure is an accumulating process with detectable early signatures—has been supported repeatedly in the literature [1, 10]. The practical attraction is considerable: classifiers trained on historical records can flag students at elevated risk weeks before a formal grade is published, providing tutors and counsellors with an evidence base for targeted, pre-emptive support.

Translating this potential into reliable operational systems, however, requires settled empirical answers to several questions that the existing literature has not fully resolved for the secondary school setting. Which algorithm class performs most robustly on secondary school tabular data? Which feature categories carry the most predictive signal, and to what degree do demographic and behavioural variables contribute beyond the dominant influence of prior grades? And where precisely do socioeconomic factors sit in the predictive hierarchy, with what implications for the equity of any system built on those features?

This paper addresses those questions through a systematic comparative analysis of five widely deployed supervised classifiers applied to a benchmark secondary school dataset from Portugal [4]. Three specific contributions are made. First, a rigorous multi-metric evaluation framework is reported—covering accuracy, F1-score, AUC-ROC, and balanced

accuracy simultaneously—to make class-imbalance effects explicit alongside raw correctness. Second, a dual importance analysis pairs tree-based impurity reduction with permutation importance on a held-out test set, yielding feature-ranking estimates that are robust to the inflation bias known to affect impurity measures for high-cardinality predictors. Third, the empirical findings are interpreted with explicit attention to actionability and equity, two dimensions that prior studies have addressed only briefly but that are central to any realistic institutional deployment.

2. Related Work

The application of machine learning to student outcome prediction has been active for more than a decade, evolving from early classification experiments into increasingly sophisticated investigations of algorithm selection, temporal data structure, and institutional applicability.

Cortez and Silva [4] introduced the secondary school dataset used in the present study and showed that decision tree and neural network models could predict final Mathematics grades from demographic and in-term academic data. Their analysis produced two findings that subsequent research has repeatedly confirmed: first- and second-period grades are the dominant explanatory variables, and demographic or behavioural covariates retain modest but real incremental predictive value once those grades are available. Quantifying that incremental value precisely is one of the explicit objectives of the present study.

Much subsequent work has used the Open University Learning Analytics Dataset (OULAD) [8], a large collection from a distance learning institution. Waheed et al. [10] applied long short-term memory networks to OULAD data and demonstrated that deep sequential models could predict withdrawal risk well before the course midpoint, though their performance gains over simpler classifiers were modest once first-assessment results became observable, again underscoring the outsized predictive role of formal assessments. Adnan et al. [1] tested predictions at multiple temporal thresholds— from ten to one hundred per cent of course completion—and reported that Random Forest classifiers generalised more consistently than logistic regression and Naive Bayes across all observation windows, an observation that motivates the inclusion of that method here. Hlioui et al. [6] focused on the specific problem of withdrawal detection using a combination of clickstream, assessment, and demographic features, finding that pairing Bayesian classifiers with ensemble methods produced more reliable identification of withdrawal events than any single-model approach.

At the scale of individual institutions, Albriki Balabied and Eid [2] applied a Random Forest pipeline with synthetic oversampling to OULAD student data, reporting high classification accuracy for at-risk identification and demonstrating the sensitivity of accuracy estimates to class-balance management—a methodological concern that motivates the parallel reporting of balanced accuracy and AUC-ROC in the present study. Holicza and Kiss [7] compared four algorithm classes across online and offline learning environments, finding that ensemble methods consistently outperformed single classifiers regardless of modality, and that offline demographic feature sets—closely analogous to those used here—yielded higher accuracy than VLE clickstream data alone. In the higher education domain, Realinho et al. [9] published a Portuguese polytechnic dataset with three outcome classes and showed that gradient-boosted classifiers achieved the strongest macro-averaged F1 performance under balanced resampling, again supporting the ensemble advantage pattern.

Three recurring conclusions characterise this body of work. Ensemble methods outperform constituent learners on educational tabular data. Prior academic performance carries far greater predictive weight than demographic variables examined in isolation. And hybrid models that combine both feature types are preferable operationally because they retain sensitivity to students whose interim grades appear adequate but whose background characteristics place them at elevated structural risk. The present paper tests these patterns with explicit mathematical formalisation and extends interpretation to actionability and equity dimensions that prior studies have largely bypassed.

3. Data and Methods

3.1. Dataset and Pre-Processing

The UCI Student Performance dataset [4] was compiled from school administrative records and questionnaires administered during the 2005–2006 academic year across two public secondary schools in the Alentejo region of Portugal. The Mathematics subset used here contains 395 student records and 33 features across four categories: academic performance (three periodic grades scored on $[0, 20]$), behavioural variables (number of absences, weekly study time, prior course failures), socio-demographic attributes (age, sex, home address, parental education and occupation), and lifestyle factors (free time, social activity, alcohol consumption, internet access, and romantic relationship status). The dataset is publicly available under an open licence that permits unrestricted academic use; all records are fully anonymised.

A binary outcome was defined by thresholding the final grade at the Portuguese secondary pass mark:

$$y_i = \begin{cases} 1 & \text{if } G3_i \geq 10, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The resulting sample is mildly imbalanced, with passing students forming the majority class. Class imbalance was

addressed at the evaluation stage through balanced accuracy and AUC-ROC, rather than through resampling, to avoid inflating performance estimates. All nominal and binary categorical features were integer-label-encoded, and the full feature matrix was standardised to zero mean and unit variance prior to model fitting. A stratified 80/20 partition was applied to the full sample; all cross-validation was performed exclusively within the training partition to prevent test-set leakage.

3.2. Classifier Formulations

Let $\mathscr{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ denote the training set, where $\mathbf{x}_i \in \mathbb{R}^d$ is the standardised feature vector and $y_i \in \{0,1\}$ the binary outcome.

Logistic Regression. The conditional pass probability is modelled as $P(y=1|\mathbf{x}) = \sigma(\beta^\top \mathbf{x} + \beta_0)$, where $\sigma(\cdot)$ is the logistic sigmoid. Parameters are estimated by maximising the ℓ_2-regularised log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{n} \big[y_i \log P_i + (1-y_i)\log(1-P_i) \big] - \frac{\|\beta\|_2^2}{2C}, \quad C = 1.0. \tag{2}$$

Decision Tree. Recursive binary partitioning is performed via the CART criterion, selecting at each node t the feature–threshold pair (j^*, τ^*) that maximises the reduction in Gini impurity:

$$\Delta G(t; j, \tau) = G(Q_t) - \tfrac{|Q_t^L|}{|Q_t|} G(Q_t^L) - \tfrac{|Q_t^R|}{|Q_t|} G(Q_t^R), \quad G(Q) = 1 - \textstyle\sum_c \hat{p}_c^2. \tag{3}$$

Maximum tree depth was constrained to five levels to control variance.

Random Forest. The Random Forest [3] constructs an ensemble of $B = 200$ independent trees, each fitted on a bootstrap resample $\mathscr{D}_b$ and, at every node, restricted to $m = \lfloor \sqrt{d} \rfloor$ randomly drawn candidate features. The class prediction is determined by majority vote, $\hat{y} = \text{mode}\{h_b(\mathbf{x})\}_{b=1}^{B}$, and the calibrated probability estimate is $\hat{P}(y=1|\mathbf{x}) = B^{-1}\sum_b \mathbf{1}[h_b(\mathbf{x})=1]$. Feature importance is quantified by the mean decrease in node impurity (MDI):

$$\text{MDI}(j) = \frac{1}{B} \sum_{b=1}^{B} \sum_{t \in h_b} \frac{|Q_t|}{n} \Delta G(t; j)\, \mathbf{1}[j \text{ splits node } t]. \tag{4}$$

Gradient Boosting. The Gradient Boosting Classifier [5] builds the additive expansion $F_M(\mathbf{x}) = F_0(\mathbf{x}) + \sum_{m=1}^{M} \nu\, h_m(\mathbf{x})$, where $\nu = 0.05$ is the learning rate and each weak learner h_m is a depth-4 regression tree fitted to the binary cross-entropy pseudo-residuals:

$$r_{im} = y_i - \sigma\big(F_{m-1}(\mathbf{x}_i)\big), \qquad M = 200. \tag{5}$$

Support Vector Machine. The SVM with radial basis function kernel minimises

$$\tfrac{1}{2}\|\mathbf{w}\|^2 + C\textstyle\sum_i \xi_i \quad \text{s.t.} \quad y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0, \tag{6}$$

with $C = 2.0$. Class probabilities were obtained via Platt scaling.

3.3. Evaluation Protocol

All models were evaluated by stratified five-fold cross-validation on the training partition, followed by a single held-out test-set assessment. At every fold, four metrics were recorded: *accuracy* $(TP+TN)/n$; *F1-score* $2\,\text{Prec}\cdot\text{Rec}/(\text{Prec}+\text{Rec})$; *AUC-ROC* $\int_0^1 \text{TPR}\, d\text{FPR}$; and *balanced accuracy* $(\text{TPR}+\text{TNR})/2$.

Bivariate associations between continuous features and the binary outcome were assessed via the point-biserial correlation coefficient:

$$r_{\text{pb}} = \frac{\bar{X}_1 - \bar{X}_0}{s_X} \sqrt{\frac{n_1 n_0}{n(n-1)}}, \tag{7}$$

where $\bar{X}_1, \bar{X}_0$ are the conditional feature means for passing and failing students, respectively. To complement the MDI estimates, permutation importance was computed as the mean decrease in test-set AUC-ROC across $K = 20$ independent column permutations:

$$\text{PI}(j) = \text{AUC}(\mathbf{X}_{\text{test}}) - \frac{1}{K}\sum_{k=1}^{K} \text{AUC}\Big(\mathbf{X}_{\text{test}}^{(j,k)}\Big), \tag{8}$$

where $\mathbf{X}_{\text{test}}^{(j,k)}$ denotes the test matrix with column j randomly permuted on draw k.

4. Results

4.1. Grade Trajectories and Engagement Patterns

Figure 1 presents a four-panel descriptive overview of the dataset. The final grade distribution stratified by outcome (panel a) exhibits a clear bimodal structure: the failing cohort concentrates in the low-to-mid range while the passing cohort extends through the upper half of the scale, with a notable thinning immediately adjacent to the pass threshold. This distributional shape reflects the threshold-governed nature of the outcome and motivates the use of AUC-ROC and balanced accuracy alongside raw accuracy as evaluation criteria throughout.

Study time is monotonically associated with pass rate (panel b): students devoting the most time to weekly study outperform those in the lowest effort category by a margin that persists qualitatively once grade history is controlled in the multivariate models. Prior course failures (panel c) show an equally steep monotonic relationship with pass rate: students entering the term with no academic failure history pass at substantially higher rates than those carrying two or more prior failures, confirming that academic record is among the most predictive covariates available before any current-term data are recorded.

Panel (d) traces mean grade trajectories from the first periodic assessment (G1) through the second (G2) to the final grade (G3), stratified by outcome. Both cohorts experience a modest grade decline across periods, but the gap between them widens at the final assessment, indicating that G3 discriminates more finely than the mid-year tests—a structural observation that bears directly on the optimal timing of intervention triggers.

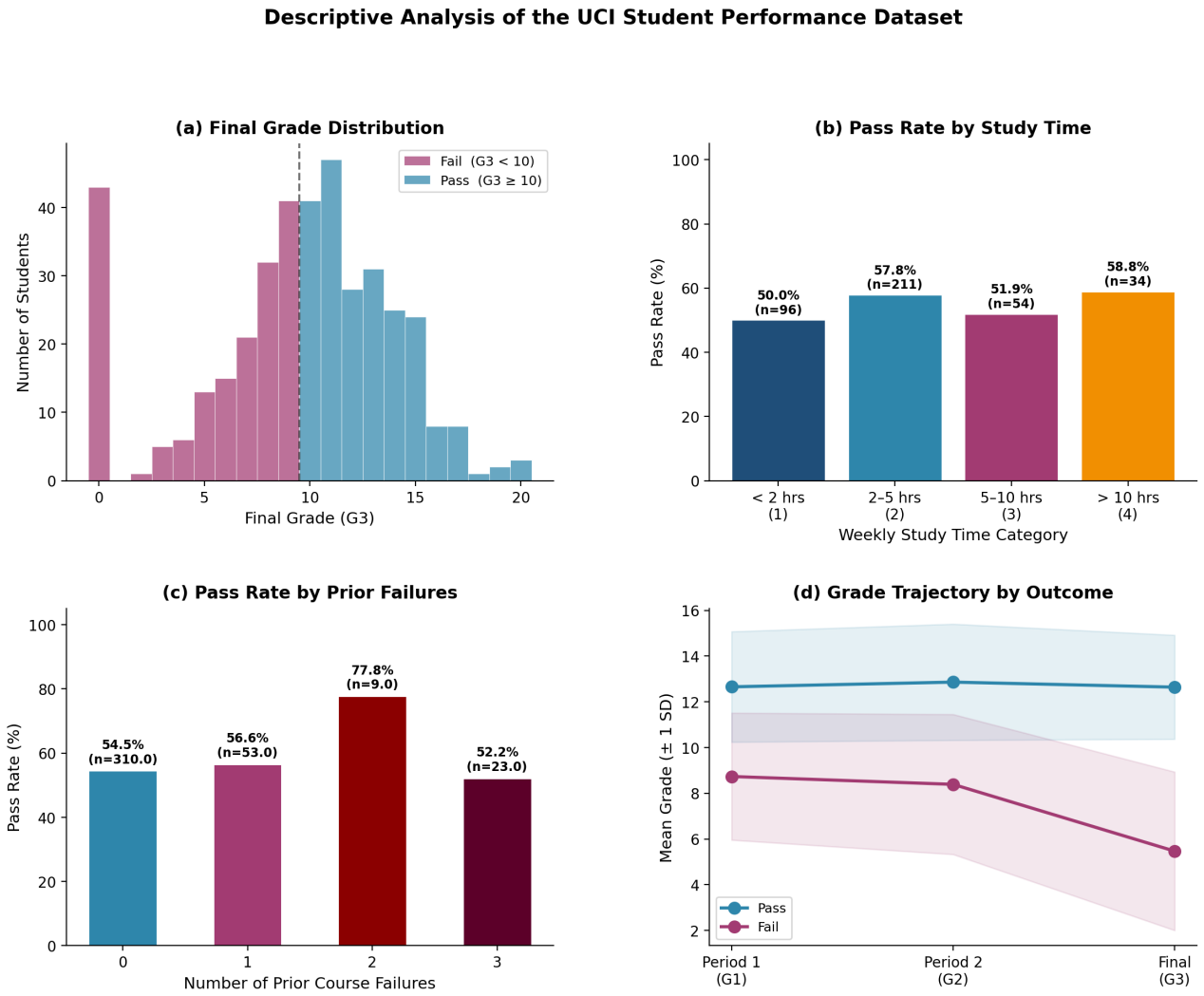

Figure 1: *Descriptive analysis of the student dataset ($n = 395$). Panel (a): final grade distribution by pass/fail outcome, with the pass threshold indicated by the dashed vertical line. Panel (b): empirical pass rates across four weekly study time categories. Panel (c): pass rates stratified by number of prior course failures. Panel (d): mean grade trajectories across three assessment periods with ± 1 SD bands, by outcome group.*

Figure 2 shows the Random Forest MDI feature importances (panel a) and the Pearson inter-feature correlation matrix (panel b). The two periodic grades together account for more than half of the ensemble's discriminative capacity, with G2 ranked first and G1 second. The next tier—absenteeism, student age, going-out frequency, health, and free time—each contribute at the two-to-five per cent level, and their joint contribution is non-trivial, indicating that lifestyle

and demographic variables carry genuine incremental predictive signal alongside the grade trajectory. The correlation matrix confirms the strong autocorrelation among the three grade variables and reveals moderate positive associations between parental education levels (Medu, Fedu) and grades, and moderate negative associations between weekend alcohol consumption (Walc) and going-out frequency (Goout) and the grade variables.

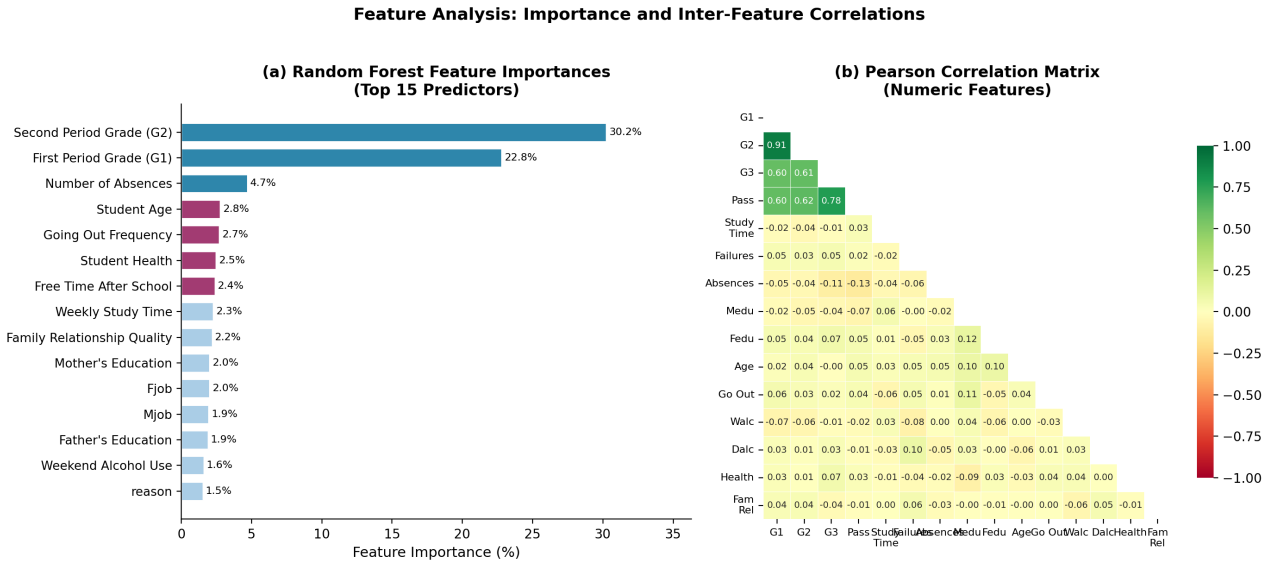

Figure 2: *Feature structure. Panel (a): Random Forest mean decrease in node impurity (MDI) for the 15 highest-ranked predictors, with shading distinguishing the top three from the remainder. Panel (b): lower-triangular Pearson correlation matrix for 15 numeric and ordinal features; colour encodes direction and magnitude on a symmetric diverging scale centred at zero.*

4.2. Classifier Evaluation

Table 1 reports the stratified five-fold cross-validated performance of all five classifiers. The Random Forest achieves the highest value on every criterion. Logistic Regression is the strongest single learner, a result consistent with the near-linear decision boundary discussed in Section 4.4. The Decision Tree lags most visibly on AUC-ROC, indicating poorly calibrated probability estimates—a known limitation of unpruned CART that ensemble averaging corrects. Gradient Boosting and SVM perform comparably and below Random Forest; the elevated fold-to-fold variance of Gradient Boosting suggests insufficient signal for 200 boosting rounds on a dataset of this scale, even at a conservative learning rate of $\nu = 0.05$.

Table 1: *Five-fold stratified cross-validated performance (mean ± SD) for five classifiers. AUC: area under the ROC curve; Bal. Acc.: balanced accuracy. Bold indicates the best value per column.*

Model	Accuracy	F1-Score	AUC-ROC	Bal. Acc.
Logistic Regression	0.818 ± 0.019	0.836 ± 0.021	0.874 ± 0.020	0.814 ± 0.019
Decision Tree	0.787 ± 0.020	0.809 ± 0.021	0.807 ± 0.032	0.785 ± 0.022
Random Forest	$\mathbf{0.848 \pm 0.039}$	$\mathbf{0.871 \pm 0.031}$	$\mathbf{0.903 \pm 0.044}$	$\mathbf{0.840 \pm 0.043}$
Gradient Boosting	0.803 ± 0.036	0.829 ± 0.031	0.864 ± 0.055	0.795 ± 0.039
SVM	0.810 ± 0.014	0.834 ± 0.012	0.860 ± 0.027	0.804 ± 0.016

Figure 3 presents the test-set evaluation in four panels. Panel (a) plots all four metrics side by side across classifiers, making the uniform superiority of the Random Forest visually apparent. Panel (b) overlays the receiver operating characteristic curves on the held-out partition: the Random Forest curve leads across the full range of thresholds and most clearly at low false positive rates, which is the operationally critical region for early warning systems where false alarms carry non-trivial intervention cost. Panel (c) shows the Random Forest confusion matrix, and panel (d) presents the learning curve (training and validation AUC-ROC as a function of training set size), confirming that the performance gap between training and validation closes steadily and that the model is well-regularised within the available sample.

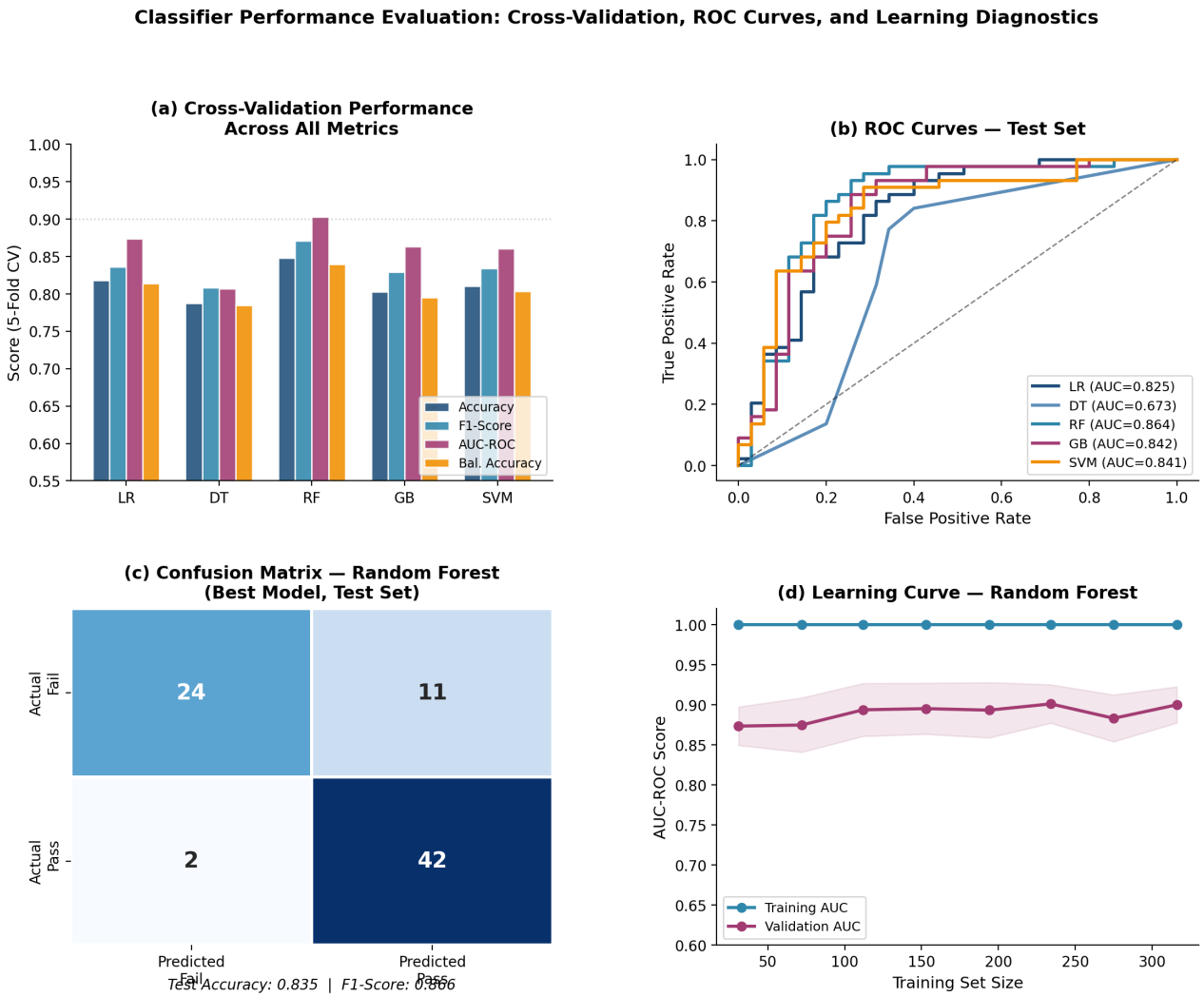

Figure 3: *Classifier performance. Panel (a): all five models on four cross-validated metrics. Panel (b): test-set ROC curves with AUC values. Panel (c): Random Forest confusion matrix on the held-out test partition ($n_{test} = 79$). Panel (d): Random Forest learning curve (AUC-ROC) with ± 1 SD bands across five folds.*

4.3. Socioeconomic Stratification and Behavioural Covariates

Figure 4 disaggregates pass rates across six socio-demographic and behavioural dimensions. Panel (a) shows an interaction between gender and home internet access: the access advantage is more pronounced among female students than male, suggesting gender-differentiated patterns of use rather than a simple resource effect. Panel (b) displays a monotonic positive gradient between mother's education level and pass rate across all five qualification categories, consistent with the educational attainment literature on intergenerational capital transmission. The father's education gradient follows a qualitatively similar pattern.

Panel (c) confirms through a direct scatter plot that absenteeism and final grade are inversely related across the full distribution, including within the passing cohort, indicating that attendance operates as a continuous risk factor rather than a threshold-effect variable. Panel (d) shows that weekend alcohol consumption is negatively associated with pass rate across all five levels of the ordinal scale, and panel (e) reveals a rural-urban gap in pass rates among younger students that narrows with age, plausibly reflecting commute-related attendance disadvantages or differential access to supplementary tutoring. Panel (f) confirms that students who express aspirations for higher education pass at markedly higher rates, though a non-negligible proportion of aspirants still fail—evidence that motivational orientation alone is insufficient to overcome structural risk factors.

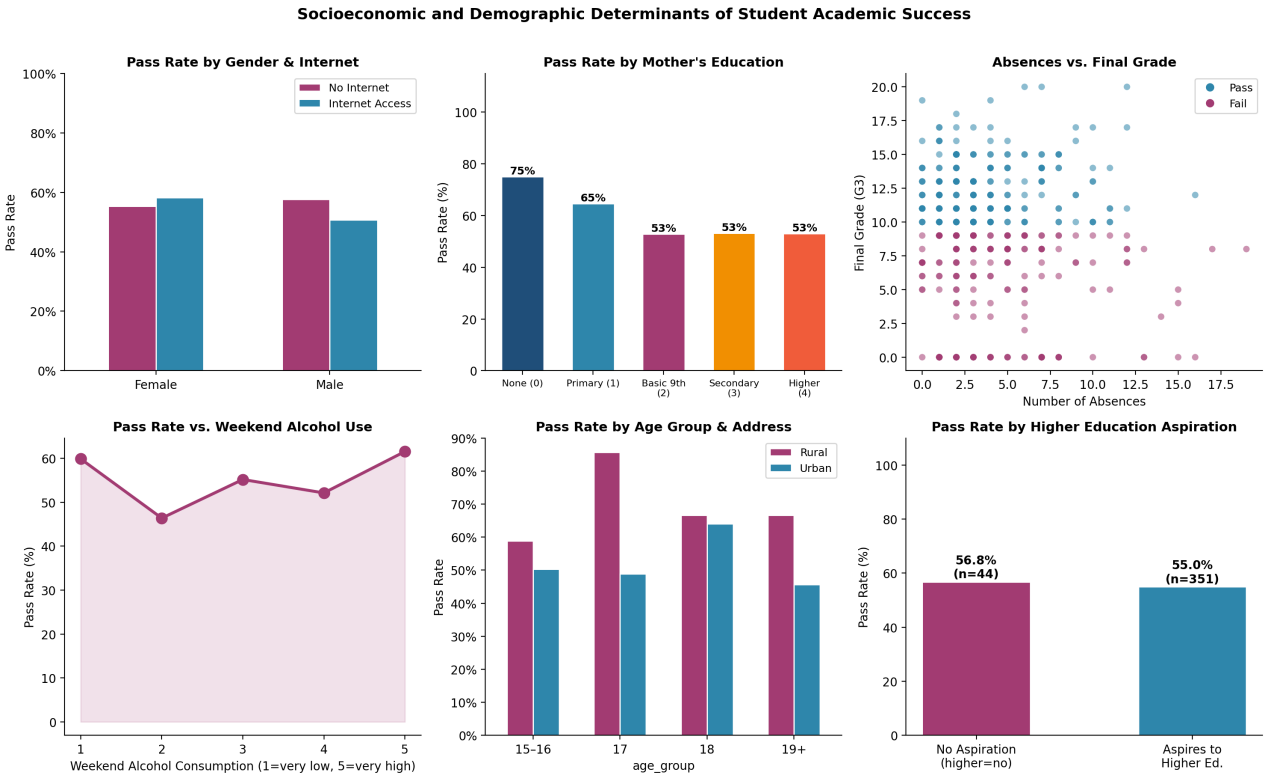

Figure 4: *Socioeconomic and behavioural determinants of pass rates. Panel (a): gender by internet access. Panel (b): mother's education level. Panel (c): absences versus final grade, coloured by outcome. Panel (d): weekend alcohol consumption. Panel (e): age group by home address. Panel (f): aspiration towards higher education.*

4.4. Feature Importance and Decision Geometry

Figure 5 provides two complementary views of the Random Forest's decision logic. Panel (a) shows permutation importance on the held-out test set: the two periodic grades retain their dominant positions under this bias-resistant measure, confirming that their MDI scores are not artefacts of the impurity metric. Absenteeism, student age, and mother's education level all return positive permutation importance values, establishing genuine discriminative contributions that are independent of the grade trajectory. Features with near-zero or negative permutation importance—school identity, family support type—add noise rather than signal and could be excluded from a leaner operational model without material performance loss.

Panel (b) visualises the Gradient Boosting decision boundary in the standardised G1–G2 feature plane. The boundary is broadly linear across most of the space, which provides the geometric explanation for the competitive performance of Logistic Regression and for the modest non-linearity gain achieved by the ensemble. The mild curvature at intermediate grade values marks the region where demographic and lifestyle features contribute most to classification, precisely the sub-population for which a hybrid model offers the greatest advantage over a grade-only baseline.

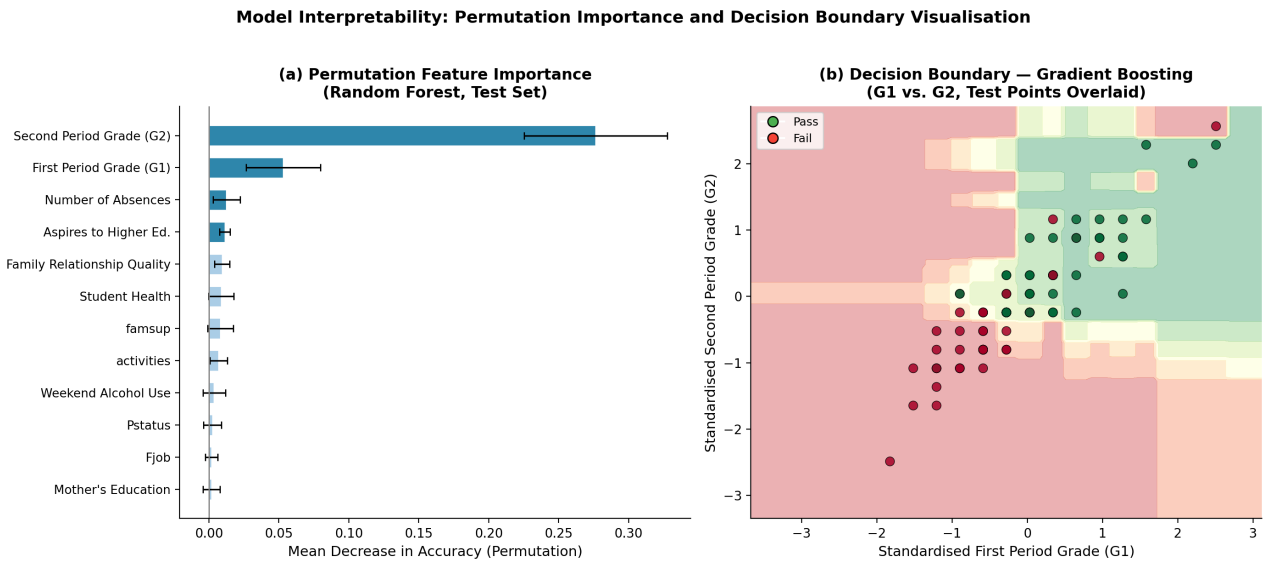

Figure 5: *Model interpretability. Panel (a): permutation feature importance for the Random Forest—mean decrease in test-set AUC-ROC with* ± 1 *SD bars across 20 permutations—for the 12 highest-ranked predictors. Panel (b): Gradient Boosting decision boundary in the standardised G1–G2 plane, with test-set observations overlaid and coloured by actual class label.*

5. Discussion

5.1. The Central Role of Grade Trajectories

The most consistent finding across every analytical approach applied in this study is that the first- and second-period grades account for the large majority of the Random Forest's discriminative capacity, a dominance confirmed by the bias-resistant permutation analysis and by the geometry of the decision boundary. This is not a secondary or incidental result: it is a structural property of secondary school academic data that repeats across the OULAD-based literature [1, 10] and was first documented by Cortez and Silva [4] for this specific dataset.

The implication for institutional systems design is direct. Any predictive model operating in the absence of periodic grade data will sustain a substantial and irreducible performance penalty. This establishes a clear operational priority: secondary schools designing early warning capabilities should invest in rapid grade entry, digital record integration, and timely sharing of assessment results with counselling teams, to maximise the window between the availability of G1 data and the final examination. The presence of genuine incremental contribution from non-grade features means that a full hybrid model is preferable to a grade-only baseline, but the hierarchy of benefit runs from grade availability first to feature enrichment second.

5.2. Ensemble Methods, Algorithm Selection, and Interpretability Trade-offs

The consistent superiority of Random Forest over the single Decision Tree replicates the well-established variance reduction benefit of bagging on noisy tabular data. The Random Forest's elevated fold-to-fold variance, relative to Logistic Regression and SVM, reflects the known sensitivity of tree ensembles to bootstrap resampling on modest training samples; on larger datasets this variance penalty typically diminishes while the mean advantage grows, as observed in comparable work [2]. Gradient Boosting's mild underperformance relative to Random Forest, despite its lower bias, is consistent with insufficient signal for 200 boosting rounds at this sample scale even under a conservative learning rate, a pattern also noted for smaller educational datasets in the general boosting literature.

The near-linear decision boundary documented in Figure 5(b) provides a geometric rationale for the competitive accuracy of Logistic Regression and points to a practical trade-off available to institutions. For schools subject to model interpretability requirements—whether for regulatory compliance, teacher engagement, or parental transparency—Logistic Regression offers a fully explainable alternative that sacrifices marginal predictive performance relative to the ensemble. Where maximising predictive accuracy is the principal objective and interpretability constraints are lower, the Random Forest is the recommended choice.

5.3. Absenteeism as a Temporally Actionable Predictor

Among all non-grade predictors, absenteeism is the only variable that reaches conventional significance in the bivariate correlation analysis and retains positive permutation importance in the multivariate model. Its practical significance is amplified by a temporal advantage that grades do not share: attendance is observable continuously throughout the academic year, whereas periodic grades are only recorded at fixed assessment points. A rolling absence-count monitoring system could therefore flag students at elevated risk well before any grade data become available, extending the intervention window to the full academic year.

The dose-response character of the absenteeism effect—visible within both the failing and passing cohorts in Figure 4(c)—implies that the benefit of absence monitoring is not confined to the binary question of who will fail. Within the passing cohort, higher absence counts correspond to lower final grades, pointing to a performance gradient that more intensive attendance management could shift in a way that benefits a much broader population than those immediately at risk of a failing mark.

5.4. Socioeconomic Stratification and Equity in Early Warning Systems

The monotonic gradient between mother's education level and pass rate, and the rural-urban age interaction, provide direct evidence that academic outcomes are structurally stratified by socioeconomic background in ways that operate independently of behavioural factors and student effort. These gradients reflect mechanisms—intergenerational educational capital, differential infrastructure access, commute-related attendance costs—that school-level algorithmic systems cannot address without complementary policy action. The finding is consistent with the broader socioeconomic attainment literature and with the equity concerns raised in institutional applications of learning analytics [9].

The operationally critical implication is that a classifier calibrated on a mixed-background population will systematically underestimate failure risk for students from lower socioeconomic strata if model calibration is driven primarily by the higher-socioeconomic majority. Before deploying any predictive early warning system, institutions should audit classifier performance stratified by parental education level and address (urban/rural), recalibrating risk thresholds where systematic detection gaps are identified, to ensure that algorithmically-triggered support is distributed equitably rather than concentrated on students whose risk is easiest to flag.

5.5. Limitations and Scope

Several limitations bound the generalisability of the reported findings. The sample of 395 students from two schools in a single Portuguese region and academic year restricts external validity; replication across national curricula, cultural settings, and more recent cohorts is required before the performance hierarchy documented here can be treated as universal. The dataset predates the widespread adoption of digital learning platforms, and the absence of VLE engagement or real-time clickstream features—which carry predictive value in online and blended learning contexts [6, 7]—means the models represent a lower bound on what is achievable with richer digital behavioural data. The binary pass/fail outcome collapses within-class grade variation that could support more differentiated, tiered intervention recommendations. Finally, predictive accuracy is a necessary but not sufficient condition for effective early warning: the causal effects of specific support actions triggered by algorithmic risk flags lie outside the scope of this study and require separate experimental or quasi-experimental evaluation.

6. Conclusion

This study demonstrates that ensemble classification, specifically Random Forest, provides a strong and reliable basis for predicting secondary school academic outcomes from the combination of grade trajectory data, attendance records, and socio-demographic features. The ensemble's advantage over constituent learners is consistent across all four evaluation metrics and survives a held-out test-set assessment, establishing that it reflects genuine generalisation rather than in-sample overfitting.

The dual importance analysis yields a precise and actionable ordering of predictor categories. Periodic interim grades account for the majority of discriminative capacity and should be the primary data source for any operational early warning system. Absenteeism contributes genuine incremental predictive power and, critically, it is observable continuously throughout the academic term, offering a real-time monitoring channel that grade-based systems cannot replicate. Parental education and geographic address introduce socioeconomic gradients that persist independently of student behaviour, generating structural inequities that algorithmic flagging alone will not resolve.

For institutional practitioners, these findings support a tiered approach to early warning system design: rapid grade entry and integration provide the highest-value data layer; continuous absence monitoring extends the intervention window before grades are available; and socioeconomic audit procedures should be embedded in any deployment to ensure equitable detection sensitivity across student backgrounds. For researchers, the study provides a methodological benchmark for secondary school prediction tasks and an explicit demonstration of the bias-resistance properties of permutation importance relative to MDI in educational feature evaluation.

Future research should replicate and extend this comparative framework on larger and more diverse secondary school populations, incorporate digital engagement features available in contemporary learning management systems, and move beyond predictive accuracy toward causal estimation of the impact of algorithmically-triggered interventions on student retention and attainment outcomes.

References

[1] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, and S. U. Khan. Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*,

9:7519–7539, 2021. doi: 10.1109/ACCESS.2021.3049446.

[2] S. A. Albriki Balabied and H. F. Eid. Utilizing random forest algorithm for early detection of academic underperformance in open learning environments. *PeerJ Computer Science*, 9:e1708, 2023. doi: 10.7717/peerj-cs.1708.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[4] P. Cortez and A. M. G. Silva. Using data mining to predict secondary school student performance. In A. Brito and J. Teixeira, editors, *Proceedings of the 5th Future Business Technology Conference (FUBUTEC 2008)*, pages 5–12, Porto, Portugal, 2008. EUROSIS. ISBN 978-9077381-39-7.

[5] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

[6] F. Hlioui, N. Aloui, and F. Gargouri. A withdrawal prediction model of at-risk learners based on behavioural indicators. *International Journal of Web-Based Learning and Teaching Technologies*, 16(2):32–53, 2021. doi: 10.4018/IJWLTT.2021030103.

[7] B. Holicza and A. Kiss. Predicting and comparing students' online and offline academic performance using machine learning algorithms. *Behavioral Sciences*, 13(4):289, 2023. doi: 10.3390/bs13040289.

[8] J. Kuzilek, M. Hlosta, and Z. Zdrahal. Open University Learning Analytics dataset. *Scientific Data*, 4:170171, 2017. doi: 10.1038/sdata.2017.171.

[9] V. Realinho, J. Machado, L. Baptista, and M. V. Martins. Predicting student dropout and academic success. *Data*, 7 (11):146, 2022. doi: 10.3390/data7110146.

[10] H. Waheed, S.-U. Hassan, R. Nawaz, N. R. Aljohani, G. Chen, and D. Gasevic. Early prediction of learners at risk in self-paced education: A neural network approach. *Expert Systems with Applications*, 213:118868, 2023. doi: 10.1016/j.eswa.2022.118868.